Phylogeography of an Amazonian Cichlid supports strong structuration by water current and past evolution by vicariance associated to the Amazon's formation.

Nicolas Leroux[1], François-Étienne Sylvain[1], Eric Normandeau[1], Aleicia Holland[2], Adalberto Luis Val[3], and Nicolas Derome[1]

[1]Universite Laval Institut de Biologie Integrative et des Systemes
[2]La Trobe University School of Life Sciences
[3]Instituto Nacional de Pesquisas da Amazonia

November 24, 2021

Abstract

Amazonia is characterized by very heterogeneous riverscapes dominated by two drastically divergent water types: black (ion-poor, dissolved organic carbonate rich and acidic) and white (nutrient rich and turbid) waters. Recent phylogeographic and genomic studies have associated the ecotone formed by these environments to ecologically driven speciation in fish species. With the objective of better understanding the evolutionary forces behind the Amazonian Teleostean diversification, we sampled 240 Mesonauta festivus from 12 sites on a wide area of the Amazonian basin. These sites included three confluences of black and white water environments to seek for repeated evidences of ecological speciation at these ecotones. Our genetic dataset of 41,268 SNPs is contrasting with previous results and supports a low structuring power of water types. Conversely, we detected a strong pattern of isolation by unidirectional downstream water current and evidence of past events of vicariance potentially linked to the Amazon River formation and salt-water incursions that occurred 2.5 Mya. Using a combination of population genetic, phylogeographic analysis and environmental association models, we decomposed the spatial variance from the environmental genetic variance specifically to assess which evolutive forces have shaped inter-population differences in M. festivus' genome. Our sampling design comprising four major Amazonian rivers and three confluences of black and white water rivers supports the possibility that past studies potentially confounded ecological speciation with a site effect unrepresentative of the full Amazonian watershed. While ecological speciation admittedly played a role in Amazonian fish species diversification, we argue that neutral evolutionary processes explain most of the divergence between M. festivus populations.

UNIVERSITÉ
LAVAL

Introduction

The neotropical region hosts the most diverse freshwater fish fauna in the world, with estimates varying from 6,000 to 8,000 fish species (Reis 2013). Yet, even though its knowledge is central for their conservation (Beheregaray et *al.* 2015), this megadiverse region is understudied and studies describing the evolutionary history of South American fish are still scarce. Multiple speciation hypotheses have been constructed based on a mix of geologic, phenotypic and genetic data (Hubert et *al* . 2007, Ríos-Villamizar et *al* . 2013, Bragança and Costa 2018). From these, ecological speciation, the evolution of new species from populations affected by an ecologically driven divergent natural selection (Rundle & Nosil 2005; Schluter 2001; 2009), allopatric speciation, and the evolution of new species by vicariance events leading to the accumulation of mutations (Yamaguchi & Iwasa 2013), are promising hypotheses.

The first hypothesis, ecological speciation, suggests that speciation of Amazonian fish species is based on the presence of very heterogeneous Amazonian aquatic ecosystems resulting from two water types with contrasting chemistry. White water has a cloudy appearance caused by a large amount of silicate in suspension. Also, it has a lower amount of dissolved organic carbonate, a circumneutral pH and a higher quantity of ions than black water which is rich in DOC, acidic and ion-poor. For instance, conductivity at 20°C for black water is close to 8 µS/cm versus 70 µS/cm for white water. In the same way, the pH for black water is usually lower than five, compared to seven for white water (Sioli 1984, Val and Almeida-Val 1995, Ríos-Villamizar et *al* . 2013, Holland et *al* . 2017). These variable parameters could represent physiological challenges for fish species (Ríos-Villamizar et *al.*2013). The high dissimilarity of the two water types could lead to natural selection against migrants, a post-zygotic isolation which could reduce gene flow between white and black water sites (Rundle and Nosil 2005). Additionally, the different selective pressures in each environment should lead to directional and discrete modifications in small and specific regions of the genome (Schluter 2001, Rundle and Nosil 2005).

In the Amazon, the ecological speciation hypothesis is mainly based on recent genetic studies conducted at the confluence of the Amazon's main tributaries, the Rio Negro (black water) and the Rio Solimões (white water) (Cooke et *al.* 2012; 2012b; 2014; Beheregaray et *al* . 2015). For instance, Beheregaray et *al.* (2015) detected strong natural selection with gene flow based on the observation of a strong neutral genetic divergence between *Triportheus albus* populations living in divergent Amazonian water types. Also, a genome scans analysis (Cooke et *al.* 2012a) and a reproductive isolation experiment (Pires et *al.* 2018) have suggested that multiple fish species from these same two major rivers have a strong genetic divergence as a result of possible ecological speciation associated with the waters contrasting environmental conditions. The conclusions of these studies tend to disagree with the well-established hypothesis of evolution of Amazonian fish by allopatry (Abreu et *al.* 2020; Roxo et*al.* 2012). However, the sampling designs of these papers (Cooke et *al.* 2012; 2012b; 2014; Beheregaray et *al* . 2015) are all very similar. They are sampling the same confluence of rivers which could lead to repeated misinterpretations of results based on a site-specific effect rather than an ecological process representative of the full Amazonian watershed (Smart 2017).

On the other hand, speciation by allopatry is a much older and less speculative evolutionary hypothesis based on speciation events which have occurred in lakes or rivers isolated by geologic and sea-level fluctuation events since the Pliocene (Hubert & Renno 2006). Typically, this evolutionary hypothesis assumes that isolated populations are evolving independently by genetic drift, leading to genome wide random variations of allele frequencies in separated populations. Also, this genetic divergence is linear with time, meaning that populations separated by older geologic processes should be increasingly different. The Amazon has a complex but well-studied geologic past which has been the source of very informative and robust evolutionary hypotheses (Albert et *al.* 2018; Hoorn et *al.* 2010; Potter 1997). Geologic studies revealed that the Amazonian region was characterized by a series of connected mega-lakes from the late Miocene until ˜2.5 Mya (Campbell et *al.* 2006). The modern Amazon formed after the breaching of the eastern rim of the sedimentary basin resulting from the Andean uplift, which led to the now-established connection with the Atlantic Ocean (Campbell et *al.* 2006). The water drainage engendered strong erosion and sediment accumulations, leading to the formation of thousands of small lakes and rivers (Moreira et *al.* 2020). These geological records are strongly supported by terrestrial diversification processes (Lynch Alfaro et *al.* 2015; Ribas et *al.* 2012) and coincide with important sea-level fluctuation events caused by the interglacial cycles of the last 2.75 Ma, leading to temporary salt-water incursions (Berger & Loutre 2010; Grant et *al* . 2019). Likewise, sedimentation and saltwater incursions are natural processes that can cause isolation of freshwater bodies, a potential vector of evolution by allopatry (Lovejoy et*al.* 2006; Latrubesse & Franzinelli 2005). Consequently, using mitochondrial molecular clock estimates in several fish species, i.e., Discus (*Symphysodon sp.*) (Farias & Hrbek 2008),*Steindachnerina sp.* (Venere et *al.* 2008) and the miniature killifish (*Fluviphylax*) (Bragança & Costa 2018), events of speciation by allopatry were associated with the Amazon formation and with interglacial cycles.

Differentiating the effects of ecological speciation and allopatry is a complex challenge in the Amazon since the water's chemistry is naturally linked to its geologic past. In this sense, the transition from black to white water (and *vice versa*) always takes origin from a geologic event that shaped today's watercourses. To decipher each evolutionary force effect, it is necessary to compare multiple black and white water boundaries and look for repeated evidence of ecological speciation. In the presence of ecological speciation mediated by water type, the similar environmental shift between the populations should lead to a similar directional genetic signal located on specific loci and/or some genome wide variations in allele frequencies caused by the reduced gene flow between different water types (Rundle and Nosil 2005). On the contrary, evolution by allopatry should show a time-related linear genetic divergence, where populations who were separated by an older geologic process are more strongly differentiated genome wide. It is also difficult to distinguish the molecular signatures of ecological speciation and allopatric evolution since the Amazonian watershed connectivity varied a lot through of past climate oscillations, saltwater incursions and geologic events (Nores 2004, Thom et *al* . 2020). Additionally, the high connectivity between rivers facilitates the migration and tends to weaken the genetic signal coming from older geologic processes.

Recent advances in genomics and bioinformatics have led to the development of more precise methods to describe evolutionary processes. While mitochondrial DNA and microsatellites are very informative about the time of the divergence of populations (Guichoux et *al.* 2011), the identification of thousands of genome-wide single nucleotide polymorphism (SNP) markers by a Genotyping By Sequencing (GBS or RADseq) approach is less biased and gives precise information on samples' genotypes and population structure (Nielsen et *al.* 2011). Also, this genome-wide analysis allows the detection and quantification of associations between genotypes and the environment underlying events of ecological (Li & Wang 2017). Multiple environmental association study (EAS) models have been recently developed (Forester et *al* . 2016; Caye et *al* . 2019; Coop et *al* . 2010; Gautier 2015). These offer the advantage of correcting for the neutral genetic structure of populations, a feature that was not available in genome scan models. These complex and diversified models lead to lower false discovery rates (FDR) and higher rates of true positives, leading to more reliable results when considering large datasets (Capblancq et *al.* 2018; de Villemereuil et *al.* 2014; Forester et *al.* 2018; Rellstab et *al.* 2015).

Our study aimed to test the validity of the two different speciation hypotheses (ecological or allopatric) in relation to populations of*Mesonauta festivus* (Heckel 1840), one of the most widespread Teleosteans in the Amazon basin (Pires et *al.* 2015).*Mesonauta festivus* is a small cichlid ubiquitous to the Amazon basin. With its sedentary behaviour and limited long-distance migration ability, *M. festivus* has a very low upstream migration potential but a moderate downstream migration potential since it can let itself be carried by the current (Pires et *al* . 2015). Additionally,*M. festivus* shows strong parental care investment and a high environmental tolerance, allowing the species to live under different abiotic conditions, potentially facilitating the establishment of new populations (Pires et *al.* 2015). Here, we used a powerful sampling design consisting of 12 sites spread throughout the Amazon basin and comprising three black and white water river confluences to (1) trace the phylogeographic history of *M. festivus* in the Amazonian watershed; (2) describe the importance of the diverging water characteristics on the genotype of *M. festivus* and (3) assess which scenario, either ecological speciation and/or allopatric speciation, better explains the genetic structure of *M. festivus*within the Amazon basin. This integrative approach helped us decipher the impact of the water's physicochemical characteristics and historical vicariance events as structuring factors for populations of *M.*

festivus across the Amazonian watershed.

Material and methods

Experimental design

We sampled 240 *M. festivus* at 12 sites distributed across three major Amazonian rivers (i.e., Rio Branco, Rio Negro and Rio Solimões) [Fig. 1]. These sites comprise ecosystems with drastically divergent physicochemical parameters: five black water sites (BAR, NEG, CEM, ANA and TEF) and seven white water sites (PIR, SOL, MAN, JAR, JAC, CAT, BRA) [Table 1]. These sites comprise three confluence zones between black and white water sites : (1) Six sampling sites were close to the confluence of two major Amazonian tributaries, the Rio Negro and Rio Solimões (ANA, CEM, CAT, JAC, JAR and MAN); (2) three sites were close to the confluence of the Rio Branco and Rio Negro (BAR, NEG and BRA) and (3) three sites were close to the confluence of the Rio Tefé and Rio Solimões (TEF, SOL and PIR) [Fig. 1].

Field trips were conducted from September to December 2018–2019 during the dry season. Sites coordinates were recorded using the Global Positioning System (GPS) and the fishing perimeters were assessed in the field according to each site's local geography. We estimated the watercourse distance separating each sampling site using Google Earth pro. A multiparameter YSI Professional+Series meter (YSI Inc/Xylem Inc USA) was used to characterize water physicochemical properties (conductivity and pH) on sites. Two litres of water was also sampled at each site 30 cm below the surface, the depth where *M. festivus*were fished (Pires et *al.* 2015), to characterize other water parameters in the laboratory. At the laboratory, dissolved organic carbon (DOC) was quantified and characterized, and the concentration of nutrients (NO_2^-, NO_3^-, silicates), free ions (Ca, Na, Cl, Mg, K) and 12 metals (Al, V, Cr, Mn, Fe, Co, Ni, Cu, Zn, As, Cd and Pb) were determined according to the method detailed in Sylvain et*al* . (2019, 2020). All environmental parameters are the means of three measurements and can be found in the supplementary information section.

Twenty fish specimens were collected at each sampling site using a combination of small seine net-fishing and line fishing (fish collection permit number 29837-18). Shortly after collection, fish were euthanized using a classic MS-222 protocol to dissect a fin clip which was kept in NAP conservation buffer to preserve DNA integrity (Camacho-Sanchez et*al.* 2013). Samples were frozen at -20 degC right after the dissection and until DNA extraction.

DNA Extraction and Sequencing

A total of 240 fin samples were collected and processed. Genomic DNA was extracted using the QIAGEN DNeasy(r) Blood and Tissues kit following the manufacturer's instructions. Genomic DNA was then purified with AMPure beads (Beckman Coulter Genomics), according to the manufacturer's instructions. Double strand DNA was quantified using the Qubit(r) 2.0 Fluorometer. GBS libraries and sequencing were done at the *IBIS Plateforme d'Analyse Genomique* using *PstI and MspI* as restriction enzymes. Barcodes were ligated to the digested DNA fragments and libraries were sequenced using *Ion proton* technology.

SNPs Calling

The raw sequence reads were trimmed with Cutadapt (Martin 2011) in order to remove the adapter sequences and sequence quality was assessed using FastQC. The sequences were extracted and trimmed to 80 pb using process_radtags in STACKS V1.5 (Catchen et *al.* 2013). After trimming with Cutadapt (Martin 2011) and extracting with process_radtags, samples had an average of 2.91 (sd = 0.91) million reads (n = 231). Nine samples were discarded during the filtration process because of low-quality sequences and/or insufficient coverage. The STACKS programs for *de novo* SNP discovery were then run with the following filters and parameters. Firstly, we ran *Ustacks*considering a population size of "p = 1", a minimum stack depth of "m

= 3”, a distance allowed between stacks of “M = 3”, a maximum number of mismatches allowed of “N = 5” and activating the “disable-gapped”, “H” and “deleverage” options. These parameters have proven to be robust and lead to a low false discovery rate for*de novo* SNPs discovery (Paris et *al.* 2017). Afterward, we allowed a distance of “n = 1” between catalogue loci in *Cstacks*and applied the following filtration and transformation steps:*Sstacks* to match samples against the catalogue, *tsv2bam* to transpose the data so it is stored by locus, *gstacks* to call variant sites in each population and ran *populations* using a “p = 4”, a “r = 0.6” and the option “vcf” to produce a variant call format (VCF) file . These filtration steps led to 428,095 putative SNPs across *M. festivus'* genome. SNPs with heterozygosity rates higher than 0.5 were filtered out. The resulting VCF file was further filtered using the script “05_filter_vcf_fast.py” from stacks workflow (https://github.com/enormandeau/stacks_workflow) with the following parameters: a minimum allele coverage to keep genotypes of “m = 3”, a minimum percent of genotype data per population of “percent_genotype = 60”, a maximum number of populations that can fail percent genotypes of “max_pop_fail = 0” and a minimum number of samples with rare alleles of “min_mas = 2”. In total, 41,268 SNPs were conserved after these filtration steps, keeping only non-duplicated loci and removing SNPs in linkage. Overall missing data for the dataset was 11.70%. Also, we screened for high relatedness between samples and estimated the mean heterozygosity rate at each site using the function “genetic_diff” from the library *vcfR* in R (Knaus and Grunwald 2017) [Table 1].

Data analysis

Mesonauta festivus Phylogeographic History

Based on this 41,268 SNPs dataset (n = 231), we used ADMIXTURE, a program that estimates individual ancestry from SNP datasets (Alexander & Lange 2011; Pritchard et *al.* 2000), to calculate the posterior membership probability of each sample considering two to eight potential genetic clusters in the dataset [Fig. S1]. Then, a combination of cross-validation error values from ADMIXTURE [Fig. S2], the goodness of fit (BIC) measures from the “find.cluster” function from*Adegenet* (Jombart 2008) in R [Fig. S3] and the analysis of multiple membership probability plots from ADMIXTURE [Fig. S1], also considering biological significance according to our sampling design were used to select the number of genetic clusters present in the data. According to this clustering, we represented the ADMIXTURE estimated posterior membership probability of each sample merged by sampling sites in a pie chart map [Fig. 2]. In addition, pairwise fixation indexes (Fst) were estimated assuming four genetic clusters with the “stamppFst” function from *StAMPP* (Pembleton et *al.*2013*)* in R. We produced a heatmap to illustrate the Fst values obtained [Fig. 3].

We conducted a multiple linear regression on distance matrices (MRM) analysis (Lichstein 2007), an extension of the partial Mantel analysis, considering a pairwise “Fst/(1-Fst)” matrix as the dependent variable, as they are more adapted to detect scenarios of isolation by distance (Rousset 1997), and three explanatory matrices: (1) a matrix with pairwise river course distances between sites for isolation by distance; (2) a matrix indicating whether pairs of sites are from the same water type (same = 0, different = 1) for isolation by ecology and (3) a matrix indicating whether pairs of sites are connected by downstream water flow (flow-connected = 0, flow-unconnected = 1) for isolation by unidirectional downstream water currents. The linear MRM was ran with 1000 permutations to assess the significance and the power of the three explanatory variables at explaining the genetic dissimilarity between the sampled sites. To visualize the results from the MRM analysis, we produced three simple Mantel tests (Mantel 1967) using the function “mantel.randtest” from the package *ade4* in R to look for one by one linkage between the pairwise Fst matrix and the three explanatory matrices used in the MRM [Fig. S4].

Environmental association study

Water Type Comparison Using Environmental Parameters

At each site, water type was assessed based on visual observations and literature research prior to the environmental parameters' characterization. Using the library *factoextra* in *R* , we calculated a principal component analysis (PCA) on normalized environmental parameters, normalized as deviations from the mean, between sites. It was used to assess which environmental variables were associated with certain water types [Fig. S5]. The EAS required environmental variables not to be strongly correlated with each other. For this reason, we selected 5 environmental parameters not strongly associated with each others which can differentiate black and white water sites in our study. The selected environmental parameters were the concentration of silicate in suspension (µmol/L), concentration of dissolved organic carbonate (DOC) (mg/L), conductivity (µS/L), productivity (µg of Chla/L) and aluminum concentration (µg/L). Again, using the library *factoextra* , we produced a biplot to visualize the importance of these environmental parameters in the differentiation of the black and white water types [Fig. 4].

Detecting Genetic-Environment Associations

Since EAS models do not accept missing data in their genotype files, we used the ADMIXTURE most probable posterior membership probability result of each sample to impute the missing genotypes according to the procedure described on GitHub: enormandeau/stacks_workflow (Alexander & Lange 2011; Pritchard et *al.* 2000). We used three recently developed approaches of EAS, which are correcting for neutral genetic variation, to detect associations between our 41,268 SNPs and the diverging environmental parameters of the 5 black and 7 white water sites sampled. For the three EAS methods, we used the imputed SNPs dataset as the dependent variable and used the selected normalized environmental parameters and the water type directly, as a qualitative variable (black water = 1; white water = -1), as the explanatory variables to detect associations between the SNPs and the environment.

First, a *constrained ordination redundancy analysis* (RDA) (Forester et *al.* 2016; Legendre & Legendre 2012) was performed to detect covariation between loci and environmental predictors. We performed the RDA using the genotype matrix as the dependent variable and the environmental parameters as explanatory variables. The covariation and multicollinearity between environmental parameters were verified, using the function “vif.cca” from *vegan* (Oksanen et*al.* 2019) in *R* , to ensure that the variance inflation factor was below 5 for each variable. We verified for the significance of each constrained axis and plotted the RDA to look how the samples at each site are clustering according to their respective environmental parameters [Fig. S6]. Afterward, we selected the SNPs which were very significantly associated to a given constrained axis by selecting outliers SNPs using 3.5 standard deviation cut-offs (two-tailed p-value = 0.00046). These putatively under selection SNPs were then associated to the environmental variable they are the most strongly correlated to. We plotted the selected SNPs to visualize their distribution according to the selected constrained axes [Fig S7].

Second, we used *Baypass 2.1* (Gautier 2015), a Bayesian hierarchical model based on the *BayEnv* model (Coop et *al.*2010), to detect linear associations between environmental predictors and genetic markers. We performed the core model of Baypass 2.1 with the following parameters (-npop 12, -gfile imputed SNPs dataset, -efile environmental parameters normalized as deviations from the mean, -burnin 10000 and -pilotlenght 2500). Using a heat map, we compared the resulting covariance matrix of population allele frequencies [Fig. S8] to the Fst values heat map [Fig. 3] to ensure that the neutral genetic variation was adequately computed by Baypass. Afterward, SNPs strongly associated with environmental predictors (eBPis > 1.5) in the “betail_reg_out” file were selected as putatively under selection for a given environmental variable.

Finally, we used a latent factor mixed model (LFMM2) (Caye et *al.*2019), a least-squares estimation approach to detect associations between environmental parameters and genotypes. We performed the LFMM using the function “lfmm2” from the library *LEA* (Frichot and Francois 2015, Caye et *al.* 2019) in R. We ran the function “lfmm2.test” with the genomic control activated and considering a linear model. We corrected for the neutral genetic structure by considering four latent factors based on the four genetic clusters previously detected. We verified the validity of the model using a histogram of p-values for each explanatory variable, aiming for a flat histogram with a peak near zero [Fig. S9]. Afterward, we adjusted the p-value associated to each SNP using a Bonferroni correction and selected significant SNPs (p-value < 0.05) as putatively under selection.

These three programs have proven to be robust and correct for the neutral genetic structure (Capblancq et *al.* 2018; de Villemereuil et *al.* 2014; Forester et *al.* 2018; Rellstab et *al.*2015). Nonetheless, we decided to reduce the false positive rate by selecting only loci detected by at least two of the three methods using the function "calculate_overlap" from the package *VennDiagram*(Chen and Boutros 2011) in R (de Villemereuil et *al.* 2014) [Fig. 6]. Selected markers associated with environmental variables were then extracted from the full-genotype dataset and visualized using a PCA generated with *Vegan* (Oksanen et *al.* 2019) in R [Fig. 6].

Results

Phylogeographic analysis

Admixture results considering different number of clusters (2-8) [Fig. S1] provide compelling information about the genetic structure of the 12 sites sampled. When using [K = 2], the sites from Rio Negro and Rio Solimões are strongly differentiated. Only CAT and JAR, two sites located downstream from the Rio Negro, have a weak admixture with Rio Negro sites. When using [K = 3], the two sites located upstream of the Rio Negro (BAR and NEG) are differentiated, but still show some admixture with sites downstream (CEM and ANA). Nevertheless, BRA does not share posterior membership probability with BAR and NEG even though these sites are close geographically. When using [K = 4], sites located far upstream of the Rio Solimões (SOL and TEF) are differentiated from the rest of the Rio Solimões sites. But PIR, a site even further upstream the Rio Solimões, is not differentiated from other downstream lakes. Each subsequent increase of the K-value up to eight led to the differentiation of a new site, [K = 5] differentiates MAN, [K = 6] differentiates BRA, [K = 7] differentiates PIR and [K = 8] differentiates JAR.

According to the results from the cross-validation error values from ADMIXTURE [Fig. S2], the optimal number of clusters for our genetic data is three. Nevertheless, this cross-validation value 0.20966 is close to the one obtained with four clusters 0.20997. The "find.cluster" function from *Adegenet* led to a similar result since its goodness of the fit (BIC) values reduced more slowly at the fourth cluster [Fig. S3]. Additionally, the posterior membership probability plots [Fig. S1] stopped forming biologically significant clusters after [K = 4], differentiating only one sampling site at the time when adding more clusters. For this reason, we completed the phylogeographic analysis assuming that our full SNPs dataset is represented by four genetic clusters.

According to the pie chart map [Fig. 2], some sites that are close to each other are showing a strong genetic differentiation with each others. For instance, there is a *genetic gap* , a disproportional genetic distance compared to the river course distance separating sites, at the confluence of the Rio Negro and Rio Solimões. Additionally, there is another similar *genetic gap* between BRA and NEG. These*genetic gaps* are detectable in every neutral population structure analysis that we have produced; the pie chart map [Fig. 2], the high pairwise Fst values [Fig. 3] and the admixture barplots [Fig S1]. These two *genetic gaps* represent sites that are separated by a short river course distance, but that are isolated by a strong downstream water flow and are from drastically different environmental conditions, as shown by the differences in physicochemical properties recorded at these sites [Table 1]. However, BRA (white water) has a higher relatedness with ANA (black water) than NEG (black water) has [Fig 2 and 3]. This is despite BRA and NEG being at similar river course distances from ANA. This migration pattern, migrating downstream preferentially from white water to black water, is inverse to the one observed at the Rio Negro-Solimões confluence, where the white water sites are more closely related to other white water sites. Similarly, ANA and CEM, sites from the Rio Negro, share common posterior memberships with sites downstream (JAR and CAT) [Fig. 2] even though they are from drastically divergent environments.

We detected a third *genetic gap* at the confluence of the Rio Tefé and Solimões. Effectively, the genetic distance between TEF and PIR is disproportionately big when compared to the small geographic distance separating the sites. Additionally, SOL (white water) is more closely related to TEF (black water) than PIR (white water). This result is detectable in the pie chart map [Fig. 2], where PIR shares a common posterior membership with other downstream Rio Solimões sites, and TEF and SOL are clustered apart. This result is also present in the pairwise Fst heatmap [Fig. 3], where SOL and TEF are almost identical. Howerver, according to the pairwise Fst heatmap, TEF and SOL are not much more genetically distant to other downstream Rio Solimões sites than PIR [Fig. 2].

The multiple regression on distance matrices (MRM) analysis detected a significant association between the pairwise Fst matrix and both the river course distance (p-value = 0.021) and the connectivity (p-value = 0.001) matrices. The relation between the genetic distance and the water type similarity matrix was not significant (p-value = 0.571). When using both the river course distance and the connectivity matrices, 59.23 % of the dependant matrix is explained by the linear model produced. According to the one-by-one Mantel tests [Fig. S4], the pairwise genetic distances between sites are moderately correlated with the pairwise river course distances (correlation coefficient of 0.54 with a p-value of 0.004). In the same way, there is a strong correlation between the pairwise genetic distances and downstream water flow connectivity (correlation coefficient of 0.71 with a p-value of 0.001) and a non-significant correlation between genetic distances and the water type similarity matrix (correlation coefficient of 0.25 with a p-value > 0.05).

Environmental Association Study

As seen in the physicochemical parameters biplot using the five selected environmental parameters [Fig. 4], differences in water physicochemical characteristics can differentiate the two water types. Black water sites were characterized by higher DOC and Al concentrations and lower pH, while white water sites had higher amounts of silicate in suspension, as well as higher conductivity and Chl a concentration [Table 1 and Fig. 4].

All six axes of the RDA were significant (p-value < 0.05) and used for the detection of associations between the genotypes and environmental predictors. The corrected sum of the variance explained by the environmental predictors in the redundancy analysis is 4.93 %. Sample representation in the RDA according to the explanatory variables was unrelated to their respective genetic clusters [Fig. S6]. A total of 584 unique SNPs were associated to the environmental predictors in the RDA. From these, 45 were associated to aluminum concentration, 29 to productivity, 74 to conductivity, 44 to DOC concentration, 357 to silicate concentration and 35 directly to water types. For the LFMM2, a total of 367 unique SNPs had a significant p-value after the Bonferroni correction. From these, 13 were associated to aluminum concentration, 215 to productivity, 107 to conductivity, 4 to DOC concentration, 117 to silicate concentration and 24 directly to water type. For Baypass2, the neutral genetic structure estimated by the program [Fig. S8] is concordant with the Fst heatmap previously produced [Fig. 3]. A total of 307 unique SNPs had an eBPis superior to 1.5 and were considered as putatively under selection. From these, 178 were associated to aluminum concentration, 63 to productivity, 60 to conductivity, 5 to DOC concentration, 21 to silicate concentration and 15 directly to water type. From these SNPs, 172 were found in at least 2 methods and kept for the following analyses [Fig. 5].

Yet, the 172 selected SNPs resulting from our EAS are not structuring the samples according to their water type. According to the PCA using the water type associated SNPs [Fig. 6], samples are clustering according to their watershed of origin [Fig. 6B] and not according to their water type [Fig. 6C]. Samples from the two main Amazonian watersheds are well differentiated by PC1, which retains 26.56 % of the variation in the genetic matrix. Additionally, BRA (white water) is clustering with black water sites from the Rio Negro (i.e., ANA, CEM, NEG and BAR). In contrast, TEF and SOL (respectively black and white water sites) seemed to be isolated from the other Solimões River sites, which is concordant with our previous results [Fig. 2 and 3]. When compared to a PCA using the full 41,268 SNPs [Fig. S10], the general clusters stay the same. The only major difference is in the clustering of SOL and TEF with the other sites from the Solimões watershed and the higher dispersion of the sites from Rio Negro along PC2. Again, the differences in water type between sites do not seem to be the main structuring factors in the data.

Discussion

Using a genotyping by sequencing approach, we provide a unique and robust dataset supporting the structuration of *M. festivus*populations by evolution through vicariance events. Our holistic sampling design, combining 12 wisely positioned sites, 231 *M. festivus* samples and an optimized SNPs calling bioinformatic pipeline render remarkably different conclusions than what other similar studies obtained (Cooke et *al* . 2009; 2014; 2012a; b; Beheregaray et*al* . 2015). Using a wider sampling design and a combination of phylogeographic and environmental association studies, we challenge the previous ecological speciation hypothesis that genetic structure is linked to diverging water types and show strong evidence that water type has in fact a low-structuring power on *M. festivus* populations. Our results support a much more important influence of vicariance events associated with the Amazon's formation and isolation by unidirectional downstream water current on structuring populations in this clade.

Phylogeographic analysis

According to the cross-validation error values from ADMIXTURE [Fig. S2], the BIC values from "find.cluster" [Fig. S3] and the posterior membership probability plots [Fig. S1], four genetic populations were detected across our 12 sampling sites. The geographic structuring of these four genetic populations can be used to infer the phylogeographic history of *M. festivus* , its colonization potential and to infer on the evolutionary history of closely related clades.

The first genetic group detected is formed by five Rio Solimões sites: CAT, JAR, JAC, MAN and PIR. Despite the small river course distance separating them, it appears that an important genetic distance separates CAT, JAR, JAC and MAN from the sites sampled in the Rio Negro [Fig. 2]. This first *genetic gap* , between the Rio Negro and Solimões, has already been assessed in a series of scientific papers based on similar sampling designs (Cooke et *al* . 2009; 2012a; b; 2014; Beheregaray et *al* . 2015). In these studies, this *genetic gap* was associated with ecological speciation, an evolutionary process caused by the presence of an ecotone like the black and white water confluence. Strikingly, fish sampled at CAT are more closely related to PIR than to CEM, located respectively at 631 and 71 km from CAT [Fig. 2]. Here, the genetic distances seem to be unrelated to the geographic distance, but strongly correlated to other variables such as water type differences and isolation by unidirectional downstream currents. While this result is in concordance with the ecological speciation hypothesis and the results from Cooke et *al* . (2009; 2014; 2012a; b), it ultimately does not lead us to the same evolutionary conclusions.

Indeed, we observed a second and similar genetic gap at the confluence of the Negro and Branco Rivers [Fig. 2], forming our second genetic group composed of the sites BAR and NEG. This genetic divergence has not been recorded in previous Amazonian population genetic studies, demonstrating the power of using a larger sampling design. While it might be interpreted as evidence of the divergence between the black (NEG) and white (BRA) water types, this conjecture is refuted by the higher relatedness between BRA and ANA (black water) than between NEG and ANA in the pie chart map and the Pairwise Fst heatmap [Fig. 2-3]. This is despite the drastic environmental shift separating ANA and BRA, and both sites being at a similar river course distance from ANA, respectively 215 and 225 km for NEG and BRA. Although Branco River is sometimes considered as a clear-water river during the rainy season (Junk et *al* . 2015; Filoso & Williams 2000), this is only for the upper part of the river, which flows through rocky ground. In fact, the lower part of the Branco River shows white water characteristics during most of the year [Table 1] (circumneutral pH, lower DOC concentration, high ion concentration and high inorganic suspended solids) (Ríos-Villamizar et *al.* 2013; Venticinque et *al.*2016). To reiterate, Branco River (white water) is more closely related to ANA (black water) than NEG (black water). This result supports that gene flow is stronger between BRA and ANA and that water type changes are not an important migration barrier for *M. festivus* at these sites. These genetic patterns are similar when comparing BRA and NEG with CEM [Fig. 2 and 3]. These sites: BRA, ANA and CEM, are forming the third genetic population observed.

Here, the strong genetic divergence between NEG and BRA could be caused by the strong unidirectional downstream water current that prevents the migration of *M. festivus* from the Rio Branco to upstream of the Rio Negro and *vice versa* . This downstream biased gene flow could lead to a partial isolation of upstream *M. festivus* populations, which admix in downstream rivers. Ultimately, this should lead to a downstream increase in intraspecific genetic diversity at the confluence of large rivers where multiple upstream and isolated by current genetic populations meet (Paz-Vinas et *al.* 2015). Systematically, the sites located downstream of the main rivers (i.e., CEM, ANA, CAT and JAR) did have a higher mean heterozygosity rate than their upstream relatives (i.e., NEG, BRA and MAN) [Table 1]. In fact, the sites with the lowest heterozygosity were the ones located the most upstream (i.e., BRA, BAR, PIR and TEF). Similarly, freshwater fish diversity hotspots in the Amazon watershed are usually located at the crossing of large rivers where these multiple watercourses meet. For instance, the highest total species richness is found at the confluence of the Negro and Solimões Rivers and other large river confluences show similar diversity patterns, for example the Tapajós-Amazon Rivers confluence and Branco-Negro Rivers (Oberdorff et *al.* 2019). This is supporting the hypothesis of evolution by allopatry, where neutral evolutionary processes in fish populations located in isolated rivers were followed by a reconnection of the waterways and a mix of the newly formed and reproductively isolated species (Oberdorff et al. 2019).

The fourth and last genetic population is formed by two sites located in the upper Rio Solimões, near the Rio Téfé: SOL and TEF. There is a clear*genetic gap* , the third one, between this genetic population (TEF-SOL) and PIR, a site located only 6 km upstream of SOL. This*genetic gap* is easy to observe in the pie chart map [Fig. 2], but the pairwise Fst heatmap gives more precise information about its nature [Fig. 3]. According to the Fst heatmap, this genetic population does not arbour an important genetic dissimilarity with the other Rio Solimões sites [Fig. 2]. Yet, the pairwise Fst Values between SOL and PIR are similar to Fst values with sites located more than 500 km downstream the Rio Solimões. What is striking is the near absence of dissimilarity between TEF and SOL in the pairwise Fst heatmap, while PIR shows high pairwise Fst values when compared with these two sites. Therefore, TEF (black water) seems to lead to higher gene flow into SOL (white water) than PIR (white water) does. While this result goes against the ecological speciation hypothesis, the dissimilarity between SOL and PIR could, again, be caused by isolation by strong downstream water currents. We selected the site SOL as a site located right between PIR and TEF, two sites characterized by drastically divergent environmental conditions [Table 1]. It is possible that the river characteristics between TEF and SOL are more favourable to gene flow coming from TEF than from PIR for *M. festivus* . Either way, this again supports that water type is not an important migration barrier for *M. festivus* since the populations from different water types are more genetically alike to each others than the site from the same water type.

According to the MRM analysis results, most of the gene flow between the sampled sites happened between sites connected by downstream water flow, with higher amounts of gene flow happening between sites located close to each others, irrespectively of the water type at each site. In this sense, the water type at each site had a non-significant and low power at explaining the genetic distance between sites. On the contrary, the connectivity between sites was very significant and inverse to the genetic distance between sites. In the same way, sites separated by a small pairwise river course distance were more genetically related. According to these results, the divergent water types do not significantly affect the migration rate between sites for *M. festivus* . Conversely, the downstream water flows connectivity and the distance between rivers seemed to play a much more important role at structuring *M. festivus* genetic populations.

The three genetic gaps observed at the Rio Negro-Branco confluence, the Rio Negro-Solimões confluence and the TEF-PIR confluence could be caused by a combination of ecological speciation, isolation by strong water unidirectional current and past evolution by allopatry between watersheds that recently reconnected. While the diverging environmental conditions must certainly have at least some effects on fish evolution, due to differences in productivity, food availability, species assemblages, environmental pressures, physiological demands, etc., we argue that isolation by strong downstream water current and past isolation by geological processes played a much more important role in shaping the genetic structure of *M. festivus* populations.

Environmental Association Study

While we did not detect an important impact of the water types on the structure of *M. festivus* genetic populations, the effect of an ecologically driven change in specific allelic frequencies in presence of gene flow could still be detected using genotype to environment association models. We conducted a complete environment to genotype association study (EAS) aiming to identify SNPs strongly associated with selected environmental variables and directly to water type. We chose to use DOC concentration, chlorophyll a, conductivity, silicate in suspension and dissolved aluminum concentration since other authors have previously employed these to characterize black and white waters (Junk et *al.* 2011, 2015). While a low pH has proven to be a major characteristic of the black water environment (Ríos-Villamizar et*al.* 2013), its strong covariation with other parameters required its exclusion from the analyses. The aforementioned parameters have a very good power at differentiating black and white water sites in our study [Fig. 4]. If water type is responsible for a strong ecological speciation in our system, we expect to detect a strong pattern of differentiation between sites of different water types at the 172 SNPs associated with water type and its associated physicochemical parameters [Fig. 5]. Also, samples should no cluster according to their watershed of origin.

In a PCA, the clustering of samples according to the SNPs associated to the environment [Fig. 6] is similar to the result obtained using the full dataset [Fig. S10], which is compatible with the scenario of neutral divergence in allopatry. Likewise, the dominant influence of neutral evolutionary forces, mutations and genetic drift in conditions of low gene flow between certain populations, contrasts with the low influence of directional evolutionary forces in the genetic structure of environmentally associated SNPs. When clustered by watershed, samples are very well differentiated in the PCA plot [Fig. 6B], while clustering the samples by their water type gives a much more admixed PCA plot [Fig. 6C]. This is strong evidence that the presence of divergent water types is not one of the main evolutionary factors and that neutral evolutionary processes have a much stronger impact on the differentiation of these populations. This is even though we corrected for the neutral genetic structure in the three EAS methods. Since we only sampled a fraction of *M. festivus'* genome and did not have access to a reference genome to map for SNPs associated with genes of interest, this result does not rule out the possibility that other genes could be positively selected in a specific water type. Doing an analysis with a reference genome could lead to the discovery of key genes that affect the fitness of *M. festivus* individuals in each environment. When combined with our previous analyses, the EAS results provide very strong evidence that evolution by ecological speciation did not have an important influence on *M. festivus* population structure in Amazonia.

Refuting Previous Assumptions of Strong Ecological Speciation

The recently accepted assumption that ecological speciation was a major *Teleostean* evolution driver in the tropics seems to have been derived from observations of a site effect in papers with similar experimental designs (Cooke et *al* . 2009; 2012; 2012a; 2015). Our results support that, in these papers, the structuring effect caused by the water type was potentially confounded with the effect of a strong unidirectional water flow at the sampled sites. Furthermore, adding new sites to these past studies could lead them to different conclusions. For instance, if we consider only the sites that were used by Cooke et*al* . (2009; 2012; 2012a; 2015) in our results, it leads us to very similar results and conclusions, and this is even though this series of previous studies focused on different species that are genetically distant from *M. festivus* . For these papers, Cooke et *al* . sampled the *genetic gap* happening at the confluence of the Rio Negro and Solimões. Our genetic data behaves identically to theirs at these sites and we detected an extensive genetic distance between the Rio Negro and Solimões populations. It is only after adding the data from sampling sites at two other black and white water confluences that we detected evidence that adaptative divergence to specific water types was not the main driver of population structure in *M. festivus* . There seems to be a real possibility that adding new sites to Cooke's studies would lead to results like ours. The application of a wide sampling design and a deep sequencing approach have previously resolved fine-scale phylogeographic patterns in other teleostean species (Fairweather et *al.* 2018; Fang et *al.* 2018).

A Demographic Scenario Based on the Amazon's Geological History

Our study detected three main *genetic gaps* at the confluences of: (1) the Negro and Solimões Rivers, (2) Branco and Negro Rivers, and (3) Lago Tefé and Solimões Rivers. What do these three genetic gaps have in common? First, they all represent a confluence of black and white water rivers. Second, gene flow is always going downstream, which could mean the water current is the primary driver of gene flow for *M. festivus* . Considering its physiology and sedentary behaviour (Pires et*al.* 2015), strong current velocity represents an important environmental gene flow barrier for *M. festivus* . Additionally, there have been anecdotes of fishers seeing *M. festivus* floating in high current while imitating a dead leaf (Pires et *al* . 2015), which could explain why there are sources of downstream gene flow over very long distances. Lastly, the three confluences result from the formation of the modern Amazon during the late Pliocene, between 2.5 Ma and 700 Ka (Campbell et *al.* 2006; Ribas et *al* . 2012).

The formation of the modern Amazon is relatively recent and has shown to be one of the main causes of the Amazonian Animalia terrestrial diversity boom (Albert et *al.* 2018; Araújo-Silva et *al.*2017; Lynch Alfaro et *al.* 2015). Numerous authors support the Pleistocene refugia hypothesis, assuming that Amazonia diversity is partly due to geographic isolation caused by geologic processes and salt-water incursion events (Farias and Hrbek 2008, Braganca and Costa 2018, da Rocha and Kaefer 2019). Amazonian cichlids are much older, dating from the separation from their African sister clade, and have evolved mostly in riverine ecosystems (Concheiro Perez et *al.*2007). For this reason, *M. festivus* , a ubiquitous Cichlid species, is a good model to understand the impact of the formation of the modern Amazon on fish evolution. The modern transcontinental Amazon got established 2.5 Mya, when the west and east territories were divided in two (Campbell et *al.* 2006). The water drainage led to sediment accumulation, creating thousands of small lakes and rivers. The Negro and Branco Rivers were formed approximately between 1.0 to 0.7 Mya and Tefe River approximately between 2.0 to 1.0 Mya (Ribas et *al.*2012). *M. festivus* was present well before the formation of these rivers and probably colonized the new downstream environments as they formed. As previously mentioned, *M. festivus* has a very good colonization potential of new environments and show strong proof of downstream migration. In this sense, the founding population of *M. festivus* probably came from the Andes, in the west, and admixed in multiple downstream rivers, irrespective of their water type. Likewise, the *genetic gaps* detected at the confluences of black and white water rivers are probably related to the formation of the rivers (geologic history) rather than the confounding effect of the water type. As a result, the strong genetic divergence between the *M.festivus*populations of the Negro and Solimoes rivers probably stems from geological processes that led to a neutral evolution by allopatry.

M. festivus low upstream migration potential (Pires et *al.*2015) must have slowed the admixture process after the reconnection of the waterways as observed in the present watershed structures. Considering this and the similarity between our results and those from Cooke et *al* . (2009; 2012; 2012a; 2015), multiple Amazonian fish species could have evolved in these same vicariant conditions, potentially leading to the genesis of many new species, who could have dispersed after the connection of the waterways. In order to test whether some population divergence times are linked to important geological events, it would be interesting to do a demographic analysis with a molecular clock estimate based on *M. festivus*mitochondrial DNA from sites positioned close to past major geologic processes.

Conclusion

Our study aimed to investigate the support for the two main speciation hypotheses (ecological and allopatric) used to explain the evolution of*M. festivus* within the Amazon basin. We showed strong evidence that the divergent physicochemical characteristics between black and white water have a weak structuring power on *M. festivus*populations in the Amazonian watershed. Furthermore, our results challenge the recently suggested ecological speciation hypothesis explaining fish diversification in Amazonia. Unlike previous studies focusing on a single confluence between black and white water, our extensive sampling design comprising 12 populations of *M. festivus* detected a genetic structure congruent with isolation by unidirectional downstream water current, past geologic events, and waterways connectivity shift. While the Brazilian Amazon supports one of the richest fish faunas on Earth, our comprehension of the evolutionary processes which shaped its biodiversity is still lacking. Understanding the origin of such richness would help us protect its diversity. Our study not only constitutes a step forward in understanding these important processes but also provides a conceptual framework that should benefit the sampling designs of future investigations on this matter.

ACKNOWLEDGEMENTS

We gratefully acknowledge help from all members of the Derome and Val laboratory. Thank you to the *IBIS Plateforme d'Analyse Genomique*for the sequencing and to ICMBIO/Instituto Chico Mendes de Conservacao da Biodiversidade for in situ support for fish collection and for issuing the permit to transport biological samples (permit number 29837/18). This research was part of the ADAPTA project at INPA and was supported by the INCT ADAPTA (CNPq/FAPEAM) and INPA/MCTI grants to ALV (ORCID: 0000-0002-3823-3868), the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grant (grant #6333) and the Canada-Brazil Awards Joint Research Project to ND (ORCID: 0000-0002-2509-6104), and the NSERC BESC M Alexander-Graham-Bell, the FRQNT Master Scholarship grant, the MITACS GlobaLink grant, the Ressources Aquatiques Quebec (RAQ) internship grant and the Andre Darveau grant to NL (ORCID: 0000-0003-1198-9435). Additionally, we would like to thank Martin Laporte, Tommy Pontbriand and Jeremie Boucher-Fontaine for their help reviewing the manuscript.

AUTHOR CONTRIBUTIONS

F.S., N.L., E.N., N.D. and A.V. designed the experiment. F.S., N.L., A.H., N.D. and A.V. organized sampling expeditions. F.S., N.L., A.H. and N.D. sampled fish during field expeditions. N.L., F.S., A.H. and N.D. processed samples in the laboratory (fish dissections and DNA extractions). N.L. and E.N. performed bioinformatic analyses. N.L. wrote the manuscript. All authors reviewed the manuscript.

ETHICAL APPROVAL

This study was carried out in accordance with the recommendations of the Ethics Committee for the Use of Animals of the *Instituto Nacional de Pesquisas da Amazonia* (INPA). The permit (number 29837-18 as of 23 March 2021) was approved by the Ethics Committee for the Use of Animals of INPA.

DATA AVAILABILITY STATEMENT

The scripts and the datasets used for the statistical analysis of this project are freely available on the Open Science Framework (DOI 10.17605/OSF.IO/2J5FH). Raw sequence reads are deposited in the SRA (BioProject XXX) and metadata are also stored in the SRA (BioProject XXX) using the MIxS package MIGS.eu.5.0.

References

Abreu, J. M. S., A. C. S. Saraiva, J. S. Albert, and N. M. Piorski. 2020. Paleogeographic influences on freshwater fish distributions in northeastern Brazil. Journal of South American Earth Sciences 102:102692.

Albert, J. S., P. Val, and C. Hoorn. 2018. The changing course of the Amazon River in the Neogene: center stage for Neotropical diversification. Neotropical Ichthyology 16:e180033.

Alexander, D. H., and K. Lange. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. BMC Bioinformatics 12:246.

Araujo-Silva, L. E., L. S. Miranda, L. Carneiro, and A. Aleixo. 2017. Phylogeography and diversification of an Amazonian understorey hummingbird: paraphyly and evidence for widespread cryptic speciation in the Plio-Pleistocene. Ibis 159:778–791.

Beheregaray, L., G. Cooke, N. Chao, and E. Landguth. 2015. Ecological speciation in the tropics: insights from comparative genetic studies in Amazonia. FRONTIERS IN GENETICS 5.

Berger, A., and M. F. Loutre. 2010. Modeling the 100-kyr glacial–interglacial cycles. Global and Planetary Change 72:275–281.

Braganca, P. H. N., and W. J. E. M. Costa. 2018. Time-calibrated molecular phylogeny reveals a Miocene–Pliocene diversification in the Amazon miniature killifish genus *Fluviphylax* (Cyprinodontiformes: Cyprinodontoidei). Organisms Diversity & Evolution 18:345–353.

Camacho-Sanchez, M., P. Burraco, I. Gomez-Mestre, and J. A. Leonard. 2013. Preservation of RNA and DNA from mammal samples under field conditions. Molecular Ecology Resources 13:663–673.

Campbell, K. E., C. D. Frailey, and L. Romero-Pittman. 2006. The Pan-Amazonian Ucayali Peneplain, late Neogene sedimentation in Amazonia, and the birth of the modern Amazon River system. Palaeogeography Palaeoclimatology Palaeoecology 239:166–219.

Capblancq, T., K. Luu, M. G. B. Blum, and E. Bazin. 2018. Evaluation of redundancy analysis to identify signatures of local adaptation. Molecular Ecology Resources 18:1223–1233.

Catchen, J., P. A. Hohenlohe, S. Bassham, A. Amores, and W. A. Cresko. 2013. Stacks: an analysis tool set for population genomics. Molecular Ecology 22:3124–3140.

Caye, K., B. Jumentier, J. Lepeule, and O. Francois. 2019. LFMM 2: Fast and Accurate Inference of Gene-Environment Associations in Genome-Wide Studies. Molecular Biology and Evolution 36:852–860.

Chen, H., and P. C. Boutros. 2011. VennDiagram: a package for the generation of highly-customizable Venn and Euler diagrams in R. BMC bioinformatics 12:35.

Concheiro Perez, G. A., O. Říčan, G. Ortí, E. Bermingham, I. Doadrio, and R. Zardoya. 2007. Phylogeny and biogeography of 91 species of heroine cichlids (Teleostei: Cichlidae) based on sequences of the cytochrome b gene. Molecular Phylogenetics and Evolution 43:91–110.

Cooke, G. M., N. L. Chao, and L. B. Beheregaray. 2009. Phylogeography of a flooded forest specialist fish from central Amazonia based on intron DNA: the cardinal tetra *Paracheirodon axelrodi* . Freshwater Biology 54:1216–1232.

Cooke, G. M., N. L. Chao, and L. B. Beheregaray. 2012a. Marine incursions, cryptic species and ecological diversification in Amazonia: the biogeographic history of the croaker genus *Plagioscion*(Sciaenidae). Journal of Biogeography 39:724–738.

Cooke, G. M., N. L. Chao, and L. B. Beheregaray. 2012b. Divergent natural selection with gene flow along major environmental gradients in Amazonia: insights from genome scans, population genetics and phylogeography of the characin fish *Triportheus albus* . Molecular Ecology 21:2410–2427.

Cooke, G. M., N. L. Chao, and L. B. Beheregaray. 2012c. Natural selection in the water: freshwater invasion and adaptation by water colour in the Amazonian pufferfish. Journal of Evolutionary Biology 25:1305–1320.

Cooke, G. M., E. L. Landguth, and L. B. Beheregaray. 2014. Riverscape Genetics Identifies Replicated Ecological Divergence Across an Amazonian Ecotone. Evolution 68:1947–1960.

Coop, G., D. Witonsky, A. Di Rienzo, and J. K. Pritchard. 2010. Using Environmental Correlations to Identify Loci Underlying Local Adaptation. Genetics 185:1411–1423.

Fairweather, R., I. R. Bradbury, S. J. Helyar, M. de Bruyn, N. O. Therkildsen, P. Bentzen, J. Hemmer-Hansen, and G. R. Carvalho. 2018. Range-wide genomic data synthesis reveals transatlantic vicariance and secondary contact in Atlantic cod. Ecology and Evolution 8:12140–12152.

Fang, B., J. Merila, F. Ribeiro, C. M. Alexandre, and P. Momigliano. 2018. Worldwide phylogeny of three-spined sticklebacks. Molecular Phylogenetics and Evolution 127:613–625.

Farias, I. P., and T. Hrbek. 2008. Patterns of diversification in the discus fishes (*Symphysodon spp* . Cichlidae) of the Amazon basin. Molecular Phylogenetics and Evolution 49:32–43.

Filoso, S., and M. R. Williams. 2000. The hydrochemical influence of the Branco River on the Negro River and Anavilhanas archipelago, Amazonas, Brazil. Archiv fur Hydrobiologie:563–585.

Forester, B. R., M. R. Jones, S. Joost, E. L. Landguth, and J. R. Lasky. 2016. Detecting spatial genetic signatures of local adaptation in heterogeneous landscapes. Molecular Ecology 25:104–120.

Forester, B. R., J. R. Lasky, H. H. Wagner, and D. L. Urban. 2018. Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations. Molecular Ecology 27:2215–2233.

Frichot, E., and O. Francois. 2015. LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6:925–929.

Gautier, M. 2015. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. bioRxiv:023721.

Grant, G. R., T. R. Naish, G. B. Dunbar, P. Stocchi, M. A. Kominz, P. J. J. Kamp, C. A. Tapia, R. M. McKay, R. H. Levy, and M. O. Patterson. 2019. The amplitude and origin of sea-level variability during the Pliocene epoch. Nature 574:237–241.

Guichoux, E., L. Lagache, S. Wagner, P. Chaumeil, P. Leger, O. Lepais, C. Lepoittevin, T. Malausa, E. Revardel, F. Salin, and R. J. Petit. 2011. Current trends in microsatellite genotyping. Molecular Ecology Resources 11:591–611.

Holland, A., C. M. Wood, D. S. Smith, T. G. Correia, and A. L. Val. 2017. Nickel toxicity to cardinal tetra (*Paracheirodon axelrodi*) differs seasonally and among the black, white and clear river waters of the Amazon basin. Water Research 123:21–29.

Hoorn, C., F. P. Wesselingh, H. ter Steege, M. A. Bermudez, A. Mora, J. Sevink, I. Sanmartin, A. Sanchez-Meseguer, C. L. Anderson, J. P. Figueiredo, C. Jaramillo, D. Riff, F. R. Negri, H. Hooghiemstra, J. Lundberg, T. Stadler, T. Sarkinen, and A. Antonelli. 2010. Amazonia Through Time: Andean Uplift, Climate Change, Landscape Evolution, and Biodiversity. Science 330:927–931.

Hubert, N., F. Duponchelle, J. Nunez, C. Garcia-Davila, D. Paugy, and J.-F. Renno. 2007. Phylogeography of the piranha genera*Serrasalmus* and *Pygocentrus* : implications for the diversification of the Neotropical ichthyofauna. Molecular Ecology 16:2115–2136.

Hubert, N., and J.-F. Renno. 2006. Historical biogeography of South American freshwater fishes. Journal of Biogeography 33:1414–1436.

Jari Oksanen and F. Guillaume Blanchet and Michael Friendly and Roeland Kindt and Pierre Legendre and Dan McGlinn and Peter R. Minchin and R. B. O'Hara and Gavin L. Simpson and Peter Solymos and M. Henry H. Stevens and Eduard Szoecs and Helene Wagner. 2019. vegan: Community Ecology Package.

Jombart, T. 2008. adegenet: a R package for the multivariate analysis of genetic markers. BIOINFORMATICS 24:1403–1405.

Junk, W. J., M. T. F. Piedade, J. Schongart, M. Cohn-Haft, J. M. Adeney, and F. Wittmann. 2011. A Classification of Major Naturally-Occurring Amazonian Lowland Wetlands. Wetlands 31:623–640.

Junk, W. J., F. Wittmann, J. Schoengart, and M. T. F. Piedade. 2015. A classification of the major habitats of Amazonian black-water river floodplains and a comparison with their white-water counterparts. Wetlands Ecology and Management 23:677–693.

Knaus, B. J., and N. J. Grunwald. 2017. vcfr: a package to manipulate and visualize variant call format data in R. Molecular Ecology Resources 17:44–53.

Latrubesse, E. M., and E. Franzinelli. 2005. The late Quaternary evolution of the Negro River, Amazon, Brazil: Implications for island and floodplain formation in large anabranching tropical systems. Geomorphology 70:372–397.

Legendre, P., and L. Legendre. 2012. Numerical Ecology. Page (P. Legendre and L. Legendre, Eds.). Elsevier.

Li, Y.-H., and H.-P. Wang. 2017. Advances of genotyping-by-sequencing in fisheries and aquaculture. Reviews in Fish Biology and Fisheries 27:535–559.

Lichstein, J. W. 2007. Multiple regression on distance matrices: a multivariate spatial analysis tool. Plant Ecology 188:117–131.

Lovejoy, N. R., J. S. Albert, and W. G. R. Crampton. 2006. Miocene marine incursions and marine/freshwater transitions: Evidence from Neotropical fishes. Journal of South American Earth Sciences 21:5–13.

Lynch Alfaro, J. W., J. P. Boubli, F. P. Paim, C. C. Ribas, M. N. F. da Silva, M. R. Messias, F. Rohe, M. P. Merces, J. S. Silva Junior, C. R. Silva, G. M. Pinho, G. Koshkarian, M. T. T. Nguyen, M. L. Harada, R. M. Rabelo, H. L. Queiroz, M. E. Alfaro, and I. P. Farias. 2015. Biogeography of squirrel monkeys (genus *Saimiri*): South-central Amazon origin and rapid pan-Amazonian diversification of a lowland primate. Molecular Phylogenetics and Evolution 82:436–454.

Mantel, N. 1967. The Detection of Disease Clustering and a Generalized Regression Approach. Cancer Research 27:209–220.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17:10–12.

Moreira, L. S., P. Moreira-Turcq, R. C. Cordeiro, B. Turcq, K. C. Aniceto, M. Moreira-Ramirez, A. P. S. Cruz, S. Caquineau, and V. C. Silva. 2020. Climate and hydrologic controls on late Holocene sediment supply to an Amazon floodplain lake. Journal of Paleolimnology 64:389–403.

Nielsen, R., J. S. Paul, A. Albrechtsen, and Y. S. Song. 2011. Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics 12:443–451.

Nores, M. 2004. The Implications of Tertiary and Quaternary Sea Level Rise Events for Avian Distribution Patterns in the Lowlands of Northern South America. Global Ecology and Biogeography 13:149–161.

Oberdorff, T., M. S. Dias, C. Jezequel, J. S. Albert, C. C. Arantes, R. Bigorne, F. M. Carvajal-Valleros, A. De Wever, R. G. Frederico, M. Hidalgo, B. Hugueny, F. Leprieur, M. Maldonado, J. Maldonado-Ocampo, K. Martens, H. Ortega, J. Sarmiento, P. A. Tedesco, G. Torrente-Vilara, K. O. Winemiller, and J. Zuanon. 2019. Unexpected fish diversity gradients in the Amazon basin. Science Advances 5.

Paris, J. R., J. R. Stevens, and J. M. Catchen. 2017. Lost in parameter space: a road map for stacks. Methods in Ecology and Evolution 8:1360–1373.

Paz-Vinas, I., G. Loot, V. M. Stevens, and S. Blanchet. 2015. Evolutionary processes driving spatial patterns of intraspecific genetic diversity in river ecosystems. Molecular Ecology 24:4586–4604.

Pembleton, L. W., N. O. I. Cogan, and J. W. Forster. 2013. StAMPP: an R package for calculation of genetic differentiation and structure of mixed-ploidy level populations. Molecular Ecology Resources 13:946–952.

Pires, T., D. Campos, C. Ropke, J. Sodre, S. Amadio, and J. Zuanon. 2015. Ecology and life-history of *Mesonauta festivus* : biological traits of a broad ranged and abundant Neotropical cichlid. Environmental Biology of Fishes 98:789–799.

Pires, T. H. S., E. A. Borghezan, V. N. Machado, D. L. Powell, C. P. Ropke, C. Oliveira, J. Zuanon, and I. P. Farias. 2018. Testing Wallace's intuition: water type, reproductive isolation and divergence in an Amazonian fish. Journal of Evolutionary Biology 31:882–892.

Potter, P. E. 1997. The Mesozoic and Cenozoic paleodrainage of South America: a natural history. Journal of South American Earth Sciences 10:331–344.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. Genetics 155:945.

Reis, R. E. 2013. Conserving the freshwater fishes of South America: Conserving South American Freshwater Fishes. International Zoo Yearbook 47:65–70.

Rellstab, C., F. Gugerli, A. J. Eckert, A. M. Hancock, and R. Holderegger. 2015. A practical guide to environmental association analysis in landscape genomics. Molecular Ecology 24:4348–4370.

Ribas, C. C., A. Aleixo, A. C. R. Nogueira, C. Y. Miyaki, and J. Cracraft. 2012. A palaeobiogeographic model for biotic diversification within Amazonia over the past three million years. Proceedings of the Royal Society B: Biological Sciences 279:681–689.

Rios-Villamizar, E. A., M. T. F. Piedade, J. G. Da Costa, J. M. Adeney, and W. J. Junk. 2013. Chemistry of different Amazonian water types for river classification: a preliminary review. Pages 17–28. New Forest, UK.

da Rocha, D. G., and I. L. Kaefer. 2019. What has become of the refugia hypothesis to explain biological diversity in Amazonia? Ecology and Evolution 9:4302–4309.

Rousset, F. 1997. Genetic Differentiation and Estimation of Gene Flow from F-Statistics Under Isolation by Distance. Genetics 145:1219–1228.

Roxo, F. F., C. H. Zawadzki, M. A. Alexandrou, G. J. Costa Silva, M. C. Chiachio, F. Foresti, and C. Oliveira. 2012. Evolutionary and biogeographic history of the subfamily Neoplecostominae (Siluriformes: Loricariidae). Ecology and Evolution 2:2438–2449.

Rundle, H. D., and P. Nosil. 2005. Ecological speciation. Ecology Letters 8:336–352.

Schluter, D. 2001. Ecology and the origin of species. Trends in Ecology & Evolution 16:372–380.

Schluter, D. 2009. Evidence for Ecological Speciation and Its Alternative. Science 323:737–741.

Sioli, H. 1984. Water chemistry of the Amazon basin: The distribution of chemical elements among freshwaters. Pages 167–199 *in* H. Sioli, editor. The Amazon: Limnology and landscape ecology of a mighty tropical river and its basin. Springer Netherlands, Dordrecht.

Smart, P. 2017. Redundant publication and salami slicing: the significance of splitting data. Developmental Medicine & Child Neurology 59:775–775.

Sylvain, F.-E., A. Holland, E. Audet-Gilbert, A. L. Val, and N. Derome. 2019. Amazon fish bacterial communities show structural convergence along widespread hydrochemical gradients. Molecular Ecology 28:3612–3626.

Sylvain, F.-E., A. Holland, S. Bouslama, E. Audet-Gilbert, C. Lavoie, A. L. Val, and N. Derome. 2020. Fish Skin and Gut Microbiomes Show Contrasting Signatures of Host Species and Habitat. Applied and Environmental Microbiology 86:e00789-20.

Thom, G., A. T. Xue, A. O. Sawakuchi, C. C. Ribas, M. J. Hickerson, A. Aleixo, and C. Miyaki. 2020. Quaternary climate changes as speciation drivers in the Amazon floodplains. Science Advances 6:eaax4718.

Val, A. L., and V. M. F. Almeida-Val. 1995. Fishes of the Amazon and Their Environment. Springer Berlin Heidelberg.

Venere, P. C., I. L. Souza, L. K. S. Silva, M. B. D. Anjos, R. R. D. Oliveira, and P. M. Galetti. 2008. Recent chromosome diversification in the evolutionary radiation of the freshwater fish family Curimatidae (Characiformes). Journal of Fish Biology 72:1976–1989.

Venticinque, E., B. Forsberg, R. Barthem, P. Petry, L. Hess, A. Mercado, C. Canas, M. Montoya, C. Durigan, and M. Goulding. 2016. An explicit GIS-based river basin framework for aquatic ecosystem conservation in the Amazon. Earth System Science Data 8:651–661.

de Villemereuil, P., E. Frichot, E. Bazin, O. Francois, and O. E. Gaggiotti. 2014. Genome scan methods against more complex models: when and how much should we trust them? Molecular Ecology 23:2006–2019.

Yamaguchi, R., and Y. Iwasa. 2013. First passage time to allopatric speciation. Interface Focus 3:20130026.

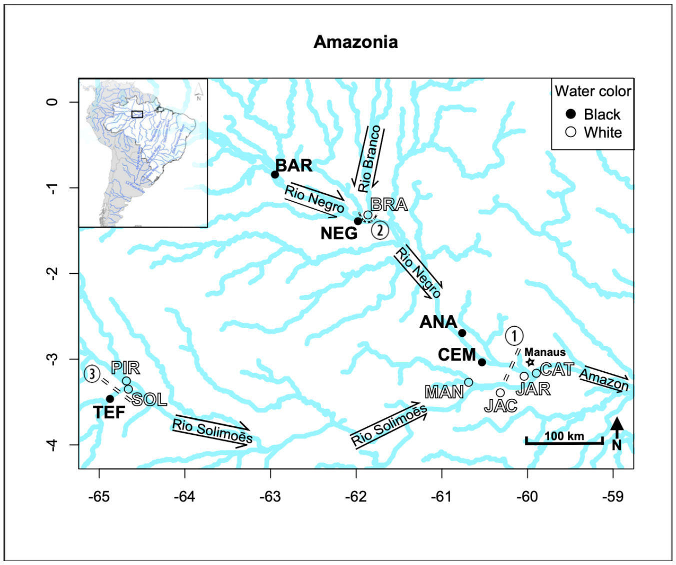

Figure 1: Location of the sampling sites (n = 12) in the Amazon basin. The 4 major tributaries of the Amazon are labeled, and the water flow directions are identified using arrows. Water is always flowing toward the East, in the direction of the Amazon. Sites have been identified using their 3-letter acronyms and point colors are consistent with the described water type at a given site.

Hosted file

`Table 1.docx` available at https://authorea.com/users/447881/articles/546790-phylogeography-of-an-amazonian-cichlid-supports-strong-structuration-by-water-current-and-past-evolution-by-vicariance-associated-to-the-amazon-s-formation

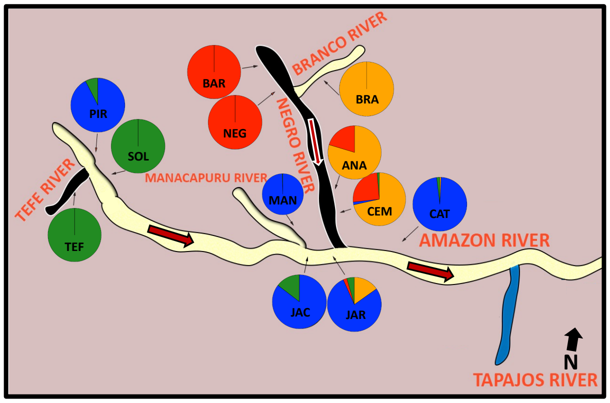

Figure 2: Posterior membership probability pie chart map of *Mesonauta festivus* populations based on the ADMIXTURE results from 41,268 SNPs and 231 samples in 12 sites and considering 4 genetic clusters. Pie chart colors represent the proportion of samples present at a given site that are associated with one of the four genetic clusters. The river flow is indicated by burgundy arrows. This schema is not to scale. For BAR, CAT, JAR, MAN, CEM, BRA, TEF, SOL and PIR (n = 20), for ANA (n = 19), and for JAC and NEG (n = 16).

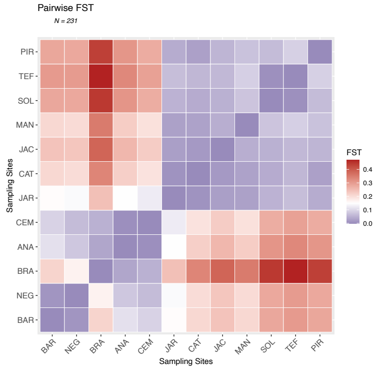

Figure 3: Pairwise Fst/(1-Fst) heatmap of the Fst values between the 12 sampling sites of *Mesonauta festivus* in the Amazon basin. Fst values were estimated using the *stamppFst* function in *R* with 100 bootstraps and assuming 4 genetic clusters. Fst values represent the genetic differentiation between each pair of sites, a value of one meaning that the populations are completely differentiated and a value of zero meaning that they are identical. Sites were ordered according to their watershed of origin. For BAR, CAT, JAR, MAN, CEM, BRA, TEF, SOL and PIR (n = 20), for ANA (n = 19), and for JAC and NEG (n = 16).

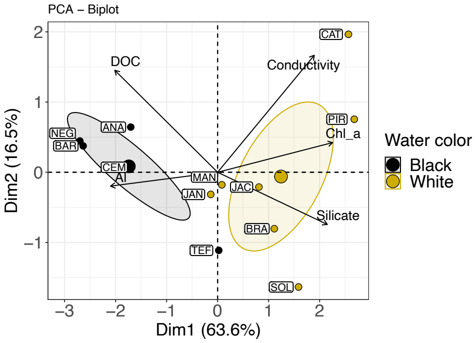

Figure 4: Biplot of the five selected environmental variables that explains most of the differences between the three water types sampled at the 12 sites of the study. Colored ellipses represent a 95% confidence interval for the two water types, with the larger circles representing the centroids (mean values) of each ellipse. The labels identify each site sampled (n = 12) with a three letters acronym. Vectors length represents each variable contribution to the PCA.

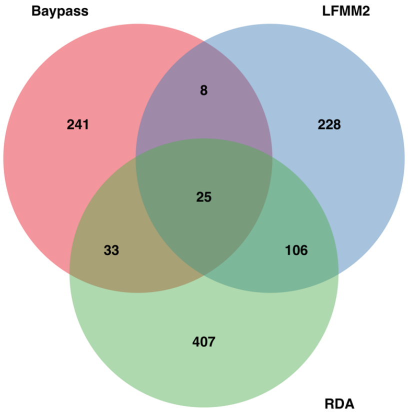

Figure 5: Venn diagram representing the number of SNPs putatively associated to environmental variables detected with each of the environmental association study methods. In total, Baypass detected 307 SNPs putatively under selection, while LFMM2 and RDA respectively detected 367 and 584 SNPs. This Venn diagram was produced using the library VennDiagram in R.

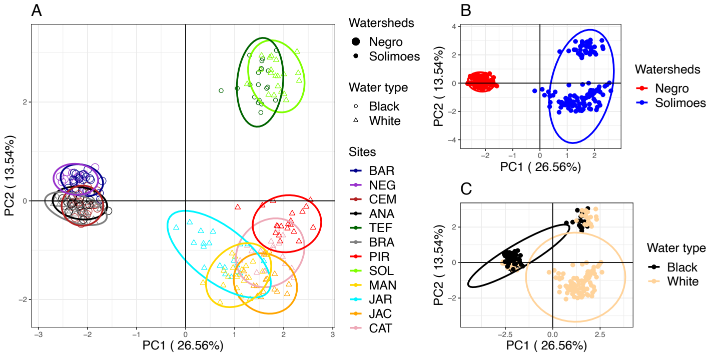

Figure 6: Principal component analysis, from black and white water sites (12 sites (n = 231)), based on the 172 SNPs associated to water type or to environmental variables associated with diverging water types. The percentage of the variance explained by the principal component's axes are written in parentheses. Ellipses represent the 95% confidence interval of the groupings position in space. In A, the ellipses represent the sampling site, the point size represents the site's watershed and the point shape the site's water type. In B, samples are clustered by watershed, and, in C, they are clustered by water type. See in figure S10 PCAs with the 41,268 SNPs, which shows a very similar clustering pattern.