Heredity (2017) 119, 287–294

www.nature.com/hdy

ORIGINAL ARTICLE

Genomic signatures of paleodrainages in a freshwater fish along the southeastern coast of Brazil: genetic structure reflects past riverine properties

AT Thomaz[1], LR Malabarba[2] and LL Knowles[1]

Past shifts in connectivity in riverine environments (for example, sea-level changes) and the properties of current drainages can act as drivers of genetic structure and demographic processes in riverine population of fishes. However, it is unclear whether the same river properties that structure variation on recent timescales will also leave similar genomic signatures that reflect paleodrainage properties. By characterizing genetic structure in a freshwater fish species (*Hollandichthys multifasciatus*) from a system of basins along the Atlantic coast of Brazil we test for the effects of paleodrainages caused by sea-level changes during the Pleistocene. Given that the paleodrainage properties differ along the Brazilian coast, we also evaluate whether estimated genetic diversity within paleodrainages can be explained by past riverine properties (i.e., area and number of rivers in a paleodrainage). Our results demonstrate that genetic structure between populations is not just highly concordant with paleodrainages, but that differences in the genetic diversity among paleodrainages correspond to the joint effect of differences in the area encompassed by, and the number of rivers, within a paleodrainage. Our findings extend the influence of current riverine properties on genetic diversity to those associated with past paleodrainage properties. We discuss how these findings may explain the inconsistent support for paleodrainages in structuring divergence from different global regions and the importance of taking into account past conditions for understanding the high species diversity of freshwater fish that we currently observe in the world, and especially in the Neotropics.
Heredity (2017) **119,** 287–294; doi:10.1038/hdy.2017.46; published online 2 August 2017

INTRODUCTION

The properties of a riverine drainage are known to structure genetic variation among fish populations because of the constraints this habitat imposes on movement patterns. For example, theoretical models reveal how specific attributes of a river's architecture act as a driver of genetic divergence (for example, Morrissey and de Kerckhove, 2009; Thomaz *et al.*, 2016). Likewise, empirical studies identify genetic structure associated with shifts in species distribution in the past (for example, Neuenschwander *et al.*, 2008), especially for coastal fishes where Pleistocene sea-level changes provided connections among rivers that are not present today (Thomaz *et al.*, 2015). However, it is unclear whether the same properties of river architecture that structure variation on recent timescales will also leave similar genomic signatures (i.e., patterns of genetic variation among individuals/populations) that reflect paleodrainage architecture. In particular, although regional structuring of genetic variation reflective of the isolation among different paleodrainages due to changes in sea level have been documented in some cases (for example, Chakona *et al.*, 2013; Unmack *et al.*, 2013; Thomaz *et al.*, 2015), the impact of the properties of the paleodrainages themselves on patterns of genetic variation has not yet been tested. Specifically, because of the connections paleodrainages provide among currently isolated rivers during periods of sea-level retreat, the properties of paleodrainages themselves may be reflected in regional measures of genetic diversity.

We address this question using genomic analyses in the freshwater fish *Hollandichthys multifasciatus* (Characiformes: Characidae), which is endemic to drainages along the southeastern Atlantic coast of Brazil. Specifically, we test the extent to which (i) structuring of genetic variation reflects past riverine connections (i.e., connections among currently isolated rivers within a paleodrainage) during the most extreme sea-level retreat on the Pleistocene, the last glacial maximum (LGM, 24–18 ka), and given that the architecture of paleodrainages differs along the Brazilian coast (Figure 1), we (ii) test whether there are corresponding differences in the genetic diversity across paleodrainages that reflect the properties of paleodrainages themselves. We examine these questions using an approach that does not presuppose that genetic structure will be partitioned by paleodrainage boundaries. That is, we do not *a priori* define paleodrainages to ask whether there is a significant effect on genetic structure (as with a F_{ST} analysis; Thomaz *et al.*, 2015). Because multiple drainages are sampled within paleodrainages (except for four northern paleodrainages; Figure 1), the genetic divergence associated with paleodrainages and their respective properties are not reducible to a single drainage

[1]Department of Ecology and Evolutionary Biology, University of Michigan, Ann Arbor, MI, USA and [2]Departamento de Zoologia, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Brazil
Correspondence: Dr AT Thomaz, Department of Ecology and Evolutionary Biology, University of Michigan, 2089 Museums Building, 1109 Geddes Avenue, Ann Arbor, MI 48109, USA.
E-mail: thomaz@umich.edu
Received 24 January 2017; revised 22 June 2017; accepted 23 June 2017; published online 2 August 2017

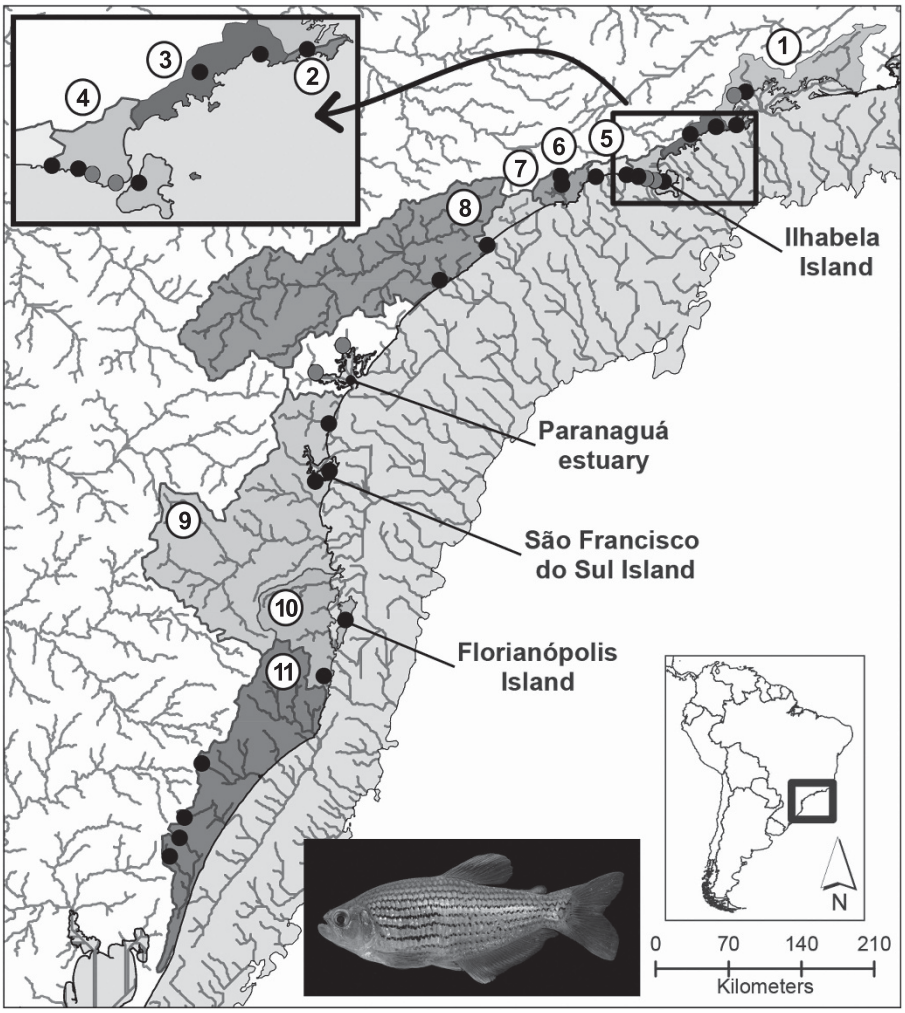

Figure 1 Map of the 11 studied paleodrainages that formed during sea-level retreats of the LGM along the southeastern coast of Brazil, with an image of *H. multifasciatus* (99.5 mm standard length). The paleodrainage area is shown in different colors and populations sampled for genomic analyses are marked by black dots. Note that one dot in paleodrainage 10 represents three populations on Florianópolis Island. The gray shaded area marks the exposed area during the sea-level retreat in the LGM. The gray dots identify populations excluded from analyses (see Materials and Methods for detail; also Thomaz *et al.*, 2015). A full color version of this figure is available at the *Heredity* journal online.

(or properties of a single drainage; Thomaz *et al.*, 2015). Moreover, we do not assume that paleodrainages are the only potential factors influencing patterns of genetic variation. Instead, we apply a series of hierarchical analyses to infer genetic clusters that can accommodate a complex history in which multiple factors may be operating at different temporal and spatial scales (i.e, recent versus deeper past, and local versus regional barriers; Massatti and Knowles, 2014). As such, our study provides not only the first analysis (that we are aware of) of the effects of paleodrainage properties on patterns of genetic diversity but also our approach highlights potential methodological issues that might bias or contribute to some of the inconsistencies in past studies on the role of paleodrainages in structuring divergence among fish populations. Moreover, this historical perspective provides a complement to investigations of the effects of contemporary river architecture on genetic variation (Morrissey and de Kerckhove, 2009; Paz-Vinas *et al.*, 2015; Thomaz *et al.*, 2016), although our specific study does not address the effects of contemporary river architecture.

MATERIALS AND METHODS

Sampling and RADseq genomic data generation and processing

Genomic data were generated for 182 individuals across the entire distribution of *H. multifasciatus*. Sampled individuals were collected from 28 rivers (hereafter, referred to as populations; Figure 1) that span 12 paleodrainages; however, only individuals from 23 populations and 11 paleodrainages were analyzed (see below); for a brief description of how paleodrainages were identified from bathymetric data see text in the Supplementary Material. Ethanol-preserved tissues used in the study are cataloged in the ichthyology collections at the Universidade Federal do Rio Grande do Sul, Museu the Ciências e Tecnologia, Pontifícia Universidade Católica do Rio Grande do Sul and Museu the História Natural Capão da Imbuia (see complete list in Supplementary Table 1).

Genomic DNA was extracted from body muscle using DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, USA) or modified salt-precipitation protocol (Medrano *et al.*, 1990), and two double-digest reduced representation libraries were constructed following the protocol of Parchman *et al.* (2012). Briefly, the DNA was double-digested with two restriction enzymes (*Eco*RI and *Mse*I), followed by a ligation step and amplification by PCR, where unique barcodes (10 bp) and Illumina adapters were added to the digested DNA. PCR products were cleaned to select fragments between 350 and 450 bp by gel extraction

(QIAquick Gel Extraction Kit—Qiagen). The two libraries were sequenced for 100 bp in two HiSeq2000 lanes (Illumina, San Diego, CA, USA) at the University of Michigan DNA core facility (Ann Arbor, MI, USA), producing 325 million reads in total (146 and 179 million in each library).

The pipeline STACKS version 1.35 (Catchen *et al.*, 2013) was used to demultiplex and process the genomic sequences. One mismatch in the adapter sequence (--adapter_mm) and a barcode distance of two was used in *process radtags* to allow barcode rescue (--barcode_dist); adapter sites were removed using Seqtk (Heng Li, https://github.com/lh3/seqtk) by deleting 5 bp in the 5′-end (-b 5). Individuals from the two libraries were then pooled together and individuals with <500K sequences were excluded. The resulting 239 million retained reads from 166 individuals (average of 1 422 655 ± 615 385 sequences per individual) was run in USTACKS with the following settings: a minimum depth coverage of five (-m 5), the *Removal algorithm* (-r) and the *Deleveraging algorithm* (-d), with model type equal bounded (--model_type), and an error bound for ϵ of 0.1 (--bound_high), which generated data with a mean coverage of 13.7 (±5.7). A catalog of genomic sequences was built in CSTACKS, allowing for two mismatches between sample tags (-n 2), and loci for each individual were identified using SSTACKS under default options.

From SSTACKS output we directly run the POPULATIONS module (with parameters: -r 0 -p 2 -m 5 --min_maf 0 --max_obs_het 0.5). The resulting output was processed in R version 3.3.1 (R Core Team, 2016) to eliminate single-nucleotide polymorphisms (SNPs) from the five last base pairs in the 3′-end of each locus, as well as loci with exceedingly high genetic diversity as such high values are suggestive of sequencing and assembly errors (i.e., $\theta > 0.024$, representing loci in the upper 95% quantile of the distribution of genetic diversity; Supplementary Figure S1). In addition, five populations were excluded because of limited data (i.e., three populations with less than two individuals after data processing) or ambiguities with paleodrainage assignment (i.e., two populations associated with the Paranaguá estuary; Thomaz *et al.*, 2015). The resulting data set contained a total of 517 874 SNPs in 196 845 loci (maximum of 10 SNPs per locus), with a genotyping rate of 0.29, for 149 individuals sampled in 23 populations from 11 paleodrainages (see Supplementary Table S1 for number of reads per individual). All STACKS modules were run under parallel execution with eight threads in the University of Michigan flux.

Because the robustness of different methods of analysis to missing data varies, we generated two data sets with different levels of missing data. Specifically, one data set included loci present in at least 10 populations and 75% of individuals within a population (i.e., 149 individuals and 62 549 SNPs in 18 407 loci, with a genotyping rate of 0.55) and was used to calculate genetic diversity summary statistics for each paleodrainage in STACKS (i.e., π and H_{EXP} averaged across populations within a given paleodrainage; Supplementary Table S2). F_{ST} values and its significances were calculated in Arlequin 3.5.2.2 (Excoffier and Lischer, 2010) with 10 000 replicates with a Bonferroni correction for multiple comparisons. The other data set included loci with a maximum of 25% missing data per unlinked SNP (hereafter referred to simply as SNPs) per individual (i.e., 116 individuals and 6574 SNPs, with a genotyping rate of 0.89) and was filtered using the toolset PLINK v.1.07 (Purcell *et al.*, 2007).

Because the degree of divergence among individuals affects the proportion of shared loci in RADseq data (Huang and Knowles, 2014), in addition to the two aforementioned data sets with individuals from the entire geographic range (hereafter referred to as the full data sets), we also processed the genetic data using two subsets of individuals to minimize the effect of missing data. Specifically, we processed individuals from the northern and southern regions separately (hereafter referred to as the northern and southern data sets, respectively), thereby increasing the amount of SNPs retained in each subset because of a fairly deep divergence separating northern and southern groups (Thomaz *et al.*, 2015).

Tests of genetic structure associated with paleodrainages

To evaluate whether there was a correspondence between population genomic structure and paleodrainages without conditioning upon paleodrainage membership, inferences of genetic structure were made using STRUCTURE 2.3.4 (Pritchard *et al.*, 2000). The full data set was analyzed with *K*-values ranging from 1 to 12 (maximum number of paleodrainages+1). An iterative approach was then used to explore potential hierarchical genetic structure (i.e., genetic structure that might be present within initial clusters identified by STRUCTURE; Ryan *et al.*, 2007; Massatti and Knowles, 2014). Specifically, STRUCTURE analyses were run for a subset of individuals contained within genetic clusters and individuals were assigned probabilistically to genetic clusters, where the number of *K*-values analyzed ranged from $K = 1$ to the number of paleodrainages+1, depending the data subset. These analyses were conducted using the northern and southern data sets to take advantage of inclusive loci to each of the two geographic areas (as described above). Ten independent runs were performed for each STRUCTURE analyses using the 'Admixture model' and 'Allele Frequencies Correlated' model for 300 000 Markov chain Monte Carlo iterations and 100 000 of burn-in, except for a few cases in which 500 000 Markov chain Monte Carlo and 200 000 of burn-in were performed to ensure convergence. The ΔK of Evanno *et al.* (2005) implemented in STRUCTURE HARVESTER (Earl and vonHoldt, 2012) was used to identify the *K* number of genetic clusters that best fit the data, with the assignment of individuals in proportion to their putative ancestral history presented graphically using the CLUMPAK pipeline (Kopelman *et al.*, 2015).

Estimates of divergence times

Divergence times between neighboring paleodrainages were estimated using a composite-likelihood method based on the site frequency spectrum (SFS) as implemented in FASTSIMCOAL2 (Excoffier and Foll, 2011; Excoffier *et al.*, 2013) to evaluate whether they were consistent with a Pleistocene divergence, and specifically, a divergence time during the LGM. We used a python script to remove all missing data to calculate the joint SFS between each neighboring paleodrainage pair (available from Papadopoulou and Knowles, 2015), based on the vcf file from STACKS with a single SNP per locus. Five individuals from each paleodrainage were used to calculate the SFS, except for two paleodrainages (paleodrainage 3 and 8; Figure 1) where only three individuals were available. Divergence times were estimated assuming no migration between paleodrainages from polymorphic loci (i.e., using the '*removeZeroSFS*' option in FASTSIMCOAL2). This assumption of no migration might result in underestimates of divergence times, however we note that the STRUCTURE analyses do not provide strong evidence of substantial admixture. Moreover, it is the relative similarity in the estimated divergence, not the absolute timing of divergence *per se*, that is particularly relevant to interpreting the relationship between paleodrainage properties and genetic diversity (i.e., general similarities in divergence times control for the potential confounding effect of different genetic diversities that could have resulted if the times to accumulate genetic diversity differed among paleodrainages).

To improve the accuracy of parameter estimates from the SFS (following the recommendations of the program; Excoffier and Foll, 2011), we calculated the effective population size of one paleodrainage (N_1) directly from the empirical data (i.e., specifically, from the nucleotide diversity (π) of fixed and variable sites). The other parameters of the divergence model (N_2, ancestral population size N_{ANC} and divergence time T_{DIV}) were estimated based on the SFS, with a mutation rate, μ, of 2.24×10^{-8}. This mutation rate was estimated from the regression formula for cellular organisms (Lynch, 2010) based on a genome size of 1500 mb for *Hollandichthys* (which is based on the average genome size of Characidae 'clade C', where *Hollandichthys* is currently positioned; Thomaz *et al.*, 2010; www.genomesize.com), with one generation per year. A total of 40 FASTSIMCOAL2 runs were conducted for each paleodrainage pair with 100 000–250 000 simulations per likelihood estimation based upon a stopping criterion of 0.001, and 10–40 expectation-conditional maximization (ECM). A parametric bootstrap was used to estimate 95% confidence intervals on the model parameters. Specifically 100 simulated SFS with the same number of individuals, loci and parameters from the maximum composite-likelihood estimate were used to re-estimate demographic parameters (as with the estimates of the empirical data, 40 FASTSIMCOAL2 runs was performed per simulated data set with the same criteria for likelihood estimation).

Tests of relationship between genetic diversity and paleodrainage properties

To test whether patterns of genetic diversity (i.e., π and H_{EXP}) correspond to paleodrainage properties, we estimated two properties: land area and number of isolated rivers within a paleodrainage. The relationship between genetic diversity and these paleodrainage properties were evaluated using generalized linear models (i.e., linear regression and covariance analyses) with the function *lm* in the basic stats package in R. For the four models (i.e., area, number of rivers, area+number of rivers and area*number of rivers), the corrected Akaike information criterion was used for model comparison using the function *aictab* in the R package *AICcmodavg* (Mazerolle, 2016).

The paleodrainage property of land area was characterized based on the current exposed land area (i.e., excluding the submerged area) in ArcGIS 10 based on the paleodrainages map (see text in Supplementary Material for a brief summary of details regarding the identification of paleodrainages based on topographic relief contours; Thomaz *et al.*, 2015). Note that total paleodrainage area was also calculated. However, because it was highly correlated with current exposed area ($R^2 = 0.97$; P-value < 0.001; Supplementary Figure S2A), and as all inferences about genetic diversity are based on sampled populations from the exposed area, we only present results on the current exposed area (and hereafter is referred for simplification as area).

The number of isolated rivers in a paleodrainage (i.e., those that are not currently connected) was used as a measure of complexity, in the sense that more rivers translate into more opportunities for the retention of genetic differences. The number of rivers in a paleodrainage was calculated using hydrological data and maps based on shuttle elevation derivatives at multiple scale (United States Geological Survey) maps. Grids with an upstream catchment area of $\geqslant 1000$ cells were defined as rivers, which for the region is $\sim 8\ km^2$.

RESULTS

Tests of genetic structure associated with paleodrainages

STRUCTURE analyses identified genetic clusters that corresponded to paleodrainage membership without using prior information about the geographic location of individuals (i.e., without conditioning on paleodrainage; Table 1 and Figure 2). At each level of the analysis for each subset of data an additional paleodrainage break was identified given the hierarchical structure of genetic variation. Moreover, probabilistic assignment of individuals to the respective genetic clusters revealed little evidence of admixture; admixture was inferred between two of seven sampled populations from paleodrainages 9 and 10 (Figure 2).

There was one exception in which the genetic break did not correspond to a paleodrainage boundary, in addition to the previously documented pronounced biogeographic division between northern and southern populations (Supplementary Figure S3; see also Thomaz *et al.*, 2015, based on mitochondrial DNA). Specifically, there was an unexpected genetic break between Ilhabela and São Sebastião 1 populations in the paleodrainage 4 (Figure 2). Note that as there was not significant structuring associated with paleodrainage 4 it was not included in the subsequent STRUCTURE analyses aimed at detecting additional structure within regional groups.

Pairwise genetic differentiation (F_{ST}) varied almost one order of magnitude (0–0.95 mean $= 0.67 \pm 0.21$; Supplementary Table S3). This broad range reflected the hierarchical structuring of genetic variation (Figure 2). Specifically, there is a pronounced differentiation between comparisons of populations between the southern and northern regional groups (mean $= 0.77 \pm 0.12$) relative to lower levels of differentiation between paleodrainages (mean $= 0.56 \pm 0.23$) within the respective northern and southern regions, or among populations from the same paleodrainage (mean $= 0.41 \pm 0.32$).

Estimates of divergence times

Divergence time estimates corroborate the hierarchical structure of genetic variation with an older regional divergence between the northern and southern regions versus relatively recent divergence times among geographically adjacent paleodrainages not separated by this geographic split (Figure 3). Specifically, the divergence between the northern and southern regions was estimated around 80 ka, whereas divergence time estimates between paleodrainages pairs are generally centered on the LGM, ranging between 12 and 44 ka (Figure 3). In most cases, estimates of the ancestral population sizes were larger than the current populations, except for the paleodrainage pairs 2–3 and 7–8 (Supplementary Table S4). Also, note that the most recent divergence time is estimated between paleodrainages 4 and 5, and one of the largest ancestral population sizes is estimated between paleodrainages 3 and 4 (Supplementary Table 4). These parameter estimates are likely biased because paleodrainage 4 violates the assumption that divergence times between paleodrainages predate divergence among populations within a paleodrainage (Figure 2).

Tests of relationship between genetic diversity and paleodrainage properties

Irrespective of which measure of genetic diversity was used (i.e., π or H_{EXP}; Table 2), the linear models identified a significant association between genetic diversity and the joint effect of paleodrainage area and number of rivers within a paleodrainage (Figure 4; Supplementary Figure S4). Specifically, despite the additional model complexity, when both paleodrainages properties are analyzed together (i.e., considering the covariance between area and the number of rivers within a paleodrainage), model comparison based on Akaike

Table 1 Summary of STRUCTURE results for a series of sequential analyses to account for the hierarchical nature of divergence (see Figure 2 for detailed plots of the probable ancestry of each individual)

Paleodrainages analyzed	*Loci*	*Individuals*	*Genotyping rate*	*First* K	ΔK	*Second* K	ΔK
All (1–11)	6574	117	0.89	2	7218.7	4	1799.0
North (1–8)	8638	70	0.91	2	7270.5	6	118.4
1, 2, 3	8126	22	0.94	3	1120.0	2	2.23
5, 6, 7, 8	8204	36	0.91	3	910.03	4	697.42
7, 8	7459	12	0.91	2	509.4	—	—
South (9–11)	9105	51	0.89	2	5053.6	3	396.1
9, 10	7387	23	0.9	2	2651.3	4	6.4

For each analysis (i.e., row), the first and second most probable *K*-values identified using Evanno method are reported along with the correspondent Δ*K*. The total number of loci and individuals analyzed are given, as well as the total individual genotyping rate.

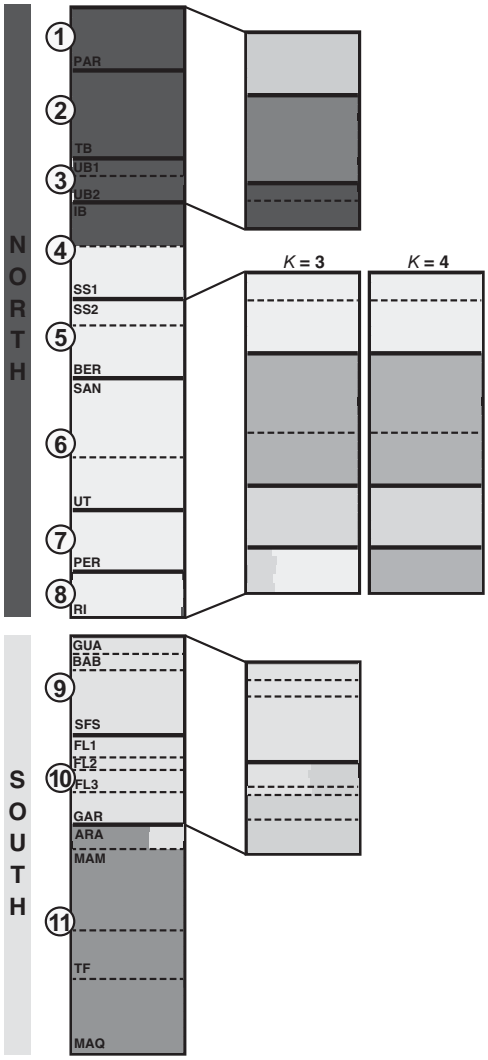

Figure 2 Results from hierarchical STRUCTURE analyses depicting the hierarchical nature of genetic structure (i.e., each block corresponds to separate analyses, with different colors identifying the different numbers of inferred *K* genetic clusters). Thick black lines and numbers in circles demarcate paleodrainages and dashed lines the populations within a paleodrainage, whose names are listed on the left, arranged from north (Paraty-PAR) to south (Maquiné-MAQ). Color pattern in the hierarchical runs corresponds to the individual paleodrainages on the map in Figure 1. The posterior probabilities of the ancestry of each individual are shown (i.e., the relative proportion of different colors). A full color version of this figure is available at the *Heredity* journal online.

information criterion scores suggests a significantly better fit compared to analyses based on each riverine property separately, or when considering a possible interaction between the paleodrainage properties (Table 2); area is correlated with the number of rivers in each paleodrainage ($R^2 = 0.39$; *P*-value = 0.04; Supplementary Figure 2B). Among the models tested, the number of rivers within a paleodrainage was the worst fit, and by itself was not significant; however, this may reflect reduced statistical power given the restricted range of differences in this variable across paleodrainages (Supplementary Table S5).

DISCUSSION

Studies have clearly demonstrated the role of paleodrainages in structuring patterns of genetic variation, where genetic divergence accumulates due to the relative isolation of rivers from different paleodrainages compared with the past connections forged among rivers within a paleodrainage, although these effects appear to vary (for example, Chakona *et al.*, 2013; Unmack *et al.*, 2013; Thomaz *et al.*, 2015). Our work adds another empirical example, and it extends this influence to dimensions that have not yet been studied. Specifically, inferences about the structuring of genetic variation by paleodrainage are (i) detected without a prior classification of paleodrainage membership of populations, in contrast to tests like F_{ST} analyses in which the groups are defined *a priori*, and (ii) we show that the paleodrainage properties themselves affect genetic diversity (i.e., the presumed connections among currently isolated rivers during periods of sea level influence regional patterns of genetic diversity). Below we discuss why considering potential contributors of processes at different spatial and temporal scales (i.e., regional versus local, and current versus past history) might explain some of the enigmatic results about the relative importance of specific factors in structuring populations of fish, as demonstrated in terrestrial environments (for example, Papadopoulou and Knowles, 2016), as well as processes of fish diversification that might underlie regional and/or taxonomic differences in richness patterns (for example, Tedesco *et al.*, 2012; Dias *et al.*, 2014).

Paleodrainage effects on genetic variation

With regards to the methodologies used to detect the contribution of paleodrainages, our results highlight how the criteria applied for such inferences may influence the conclusions (Papadopoulou and Knowles, 2016). For example, our results show a strong correspondence of genomic structure in *Hollandichthys* with paleodrainages boundaries (i.e., in 10 of the 11 paleodrainages, with the only exception in paleodrainage 4), without *a priori* classification of populations to paleodrainage (i.e., the genetic clustering of populations sampled within a paleodrainage reflects shared ancestry under the

Table 2 Comparison of the relative effect of area and number of rivers per paleodrainage on patterns of genetic variation based on the AICc; models are listed in order of their predictive value for analysis based either of the population genetic summary statistics, π or H_{EXP}

Sum. stat.	*Model*	*No. of parameters*	R^2	R^2-*adj*	P-*value*	*AICc*	*ΔAICc*	*Model prob.*
π	Area+river	4	0.81	0.76	<0.01	−57.25	0.00	0.94
	Area	3	0.43	0.37	0.03	−50.62	6.63	0.03
	Area*river	5	0.81	0.72	0.01	−49.96	7.28	0.02
	River	3	0.01	−0.11	0.84	−44.50	12.75	0.00
H_{EXP}	Area+river	4	0.79	0.74	<0.01	−60.09	0.00	0.93
	Area	3	0.40	0.33	0.04	−53.77	6.32	0.04
	Area*river	5	0.79	0.70	0.01	−52.79	7.31	0.02
	River	3	0.01	−0.10	0.78	−48.25	11.85	0.00

Abbreviations: AICc, corrected Akaike information criterion; prob., probability; Sum. stat. summary statistics.

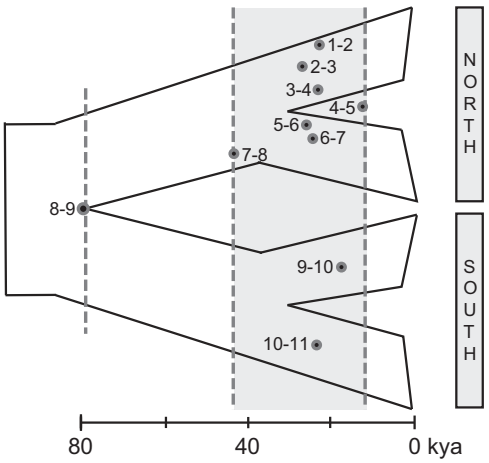

Figure 3 Schematic representation of divergence time point estimates between geographically adjacent pairs of paleodrainages (represented schematically as a general population split in either the northern or southern areas), along with the divergence time estimate between the northern and southern regions (i.e., between paleodrainages 8 and 9). The scale bar at the bottom of the figures shows estimates of the absolute divergence times (see Materials and Methods for details regarding the mutation rate used for dating).

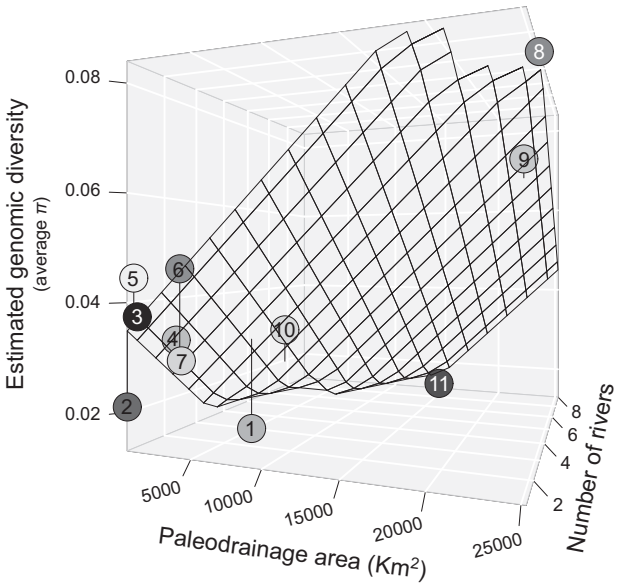

Figure 4 General linear model fit between paleodrainage properties (i.e., area and number of rivers) and nucleotide diversity (π; $R^2 = 0.81$, P-value < 0.01; Supplementary Figure S4 for results based on H_{EXP}). The colored dots and numbers correspond to the individual paleodrainages on the map in Figure 1. A full color version of this figure is available at the *Heredity* journal online.

presumed genetic equilibrium being modeled here). However, the detected genetic structure above the level of individual populations is not limited to paleodrainage boundaries (Figure 2; Table 1). For example, the northern–southern split between the Paranaguá estuary populations (Supplementary Figure S3; Thomaz *et al.*, 2015) predates the divergences reflective of paleodrainage structure (Figure 3). By applying a series of hierarchical-independent STRUCTURE analyses to accommodate this complex history of divergence, the genetic structure associated with paleodrainages becomes clear (Figure 2). In other words, the effects of paleodrainages are clear when accounting for the complexity of the history of *Hollandichthys*, but could have been overlooked by framing the question about structuring of genetic variation by paleodrainages as a binary 'yes' or 'no' question.

Similar arguments about potentially misleading conclusions might be made based on how DNA sequences are analyzed. For example, for recent divergence histories, a correspondence between clades in a gene tree and paleodrainage boundaries or the distribution of haplotypes across populations within paleodrainages (for example, Chakona *et al.*, 2013; Unmack *et al.*, 2013) are very conservative criteria for inferences about the role of paleodrainages in structuring genetic variation. The lack of monophyly may simply reflect that there has not been sufficient time for the sorting of ancestral polymorphism (Hudson and Coyne, 2002; Knowles, 2009). Likewise, the lack of shared haplotypes among rivers within a paleodrainage does not discount the possible role of paleodrainages; it simply identifies structure associated with current isolated rivers (as do our analyses; Figure 2; Supplementary Table S3). Because of overestimation of divergence times when based directly on pairwise sequence differences (Edwards and Beerli, 2000), such estimates are also unlikely to coincide with Pleistocene driven sea-level shifts that define paleodrainage boundaries. In other words, conclusions about the role of paleodrainages associated with Pleistocene sea-level changes might be sensitive to how tests are conducted and interpreted given the time frame of these historical events (Knowles, 2009). With relatively larger ancestral population sizes than current effective population sizes estimated for paleodrainages in *Hollandichthys* (Supplementary Table S4), the much more recent divergences estimated here (Figure 3) compared to previous estimates based on mitochondrial DNA (Thomaz *et al.*, 2015) are not unexpected given these divergence estimates reported here take into account gene divergences that predate population divergence (Carstens and Knowles, 2007). Migration (which was not modeled here) would result in underestimates of divergence times; however, there is little to no evidence of admixture among paleodrainages (Figure 2).

Besides methodological biases, differences in the detected effects of paleodrainages on genetic variation across studies might also reflect differences in specific properties of a local region. For example, the availability and stability of environments over time are known to affect the current genetic diversity of species in terrestrial organisms (for example, Pleistocene refugia theory; He *et al.*, 2013; Massatti and Knowles, 2016). In a similar way, our findings demonstrate that genetic diversity within paleodrainages is a function of its properties, with higher genetic diversities observed in larger and more branched paleodrainages (i.e., more constituent rivers). Note the similarity in divergence times among paleodrainages (except the north–south break; Figure 3) means that this pattern cannot be explained by differing times of accumulation of genetic diversity among paleodrainages.

This strong genomic signature urges the incorporation of information about past river structure (Neuenschwander *et al.*, 2008), rather than just considering the properties of current rivers. As with the detected effects of paleodrainage area and river number demonstrated here (Figure 4), additional properties of rivers in the past (which were assumed to be constant in space and time here) might also contribute to differences in genetic diversity among paleodrainages. For example, the effect of water flow intensity, river shape and environment (i.e., geomorphology) are known to differ regionally and affect the distribution of genetic diversity (Morrissey and de Kerckhove, 2009; Albert *et al.*, 2011; Paz-Vinas *et al.*, 2015; Thomaz *et al.*, 2016), which

make them potentially interesting to explore with respect to paleodrainages. However, this would require new developments, as with recent advances for incorporating environmental variables to study the effects of current river properties (for example, Domisch *et al.*, 2015). The impacts of such methodological developments are likely to extend beyond, deepening our knowledge of the effect of shifts in riverine properties over time.

Insights into species diversification of freshwater fish

Vicariance has a clear role in structuring species diversity patterns of riverine fish, reflecting a life history constrained to the rivers that predisposes fish in particular to becoming geographic isolated (Lundberg *et al.*, 2000; Albert *et al.*, 2011). Nonetheless, dispersal is also recognized to have a role in shaping richness patterns. Specifically, the distribution of fish species across multiple basins may be explained by the following: (i) river captures in which a river tributary changes its direction and start flowing to the neighbor basin; or (ii) dispersal associated with temporary connections.

As our study (for example, Figures 2 and 3) and others show (for example, Chakona *et al.*, 2013; Unmack *et al.*, 2013; Thomaz *et al.*, 2015), dispersal associated with temporary connections that were forged between currently isolated rivers in past drainages (i.e., paleodrainages) when sea levels repeatedly decreased may contribute to the spatial structuring and timing of divergence. Nonetheless, it might be argued that this mechanism (i.e., dispersal across drainages via past connections that opened during periods of low sea level) may be relatively species-specific (Waters and Burridge, 2016) unlike river capture and vicariance, which tends to affect communities as a whole (Burridge *et al.*, 2007; Albert *et al.*, 2011). For example, *Hollandichthys* is associated with the presence of riparian forest (Bertaco and Malabarba, 2013), and consequently is distributed in lower land tributaries, which might make downstream dispersal more likely during the cycles of sea-level retreat, given the geographic proximity to the temporary river connections that existed among drainages in the past. However, for fish inhabiting different portions of the rivers (i.e., headwaters, as opposed to lowland tributaries), such temporary connections forged by sea-level retreat might not have been accessible. If such divergence processes act in a species-specific manner, these temporary connections might be helpful to explain differences in species diversity across landscapes (i.e., discord across taxa), and consideration of the species-specific ecologies might explain why the geographic distribution of particular constituents of the ichthyofauna may differ (Waters and Burridge, 2016).

Although the links between the processes structuring genetic variation within species to those structuring species diversity patterns can be tenuous (Kisel and Barraclough, 2010; Rosenblum *et al.*, 2012; Papadopoulou and Knowles, 2017), there are some noteworthy parallels, but also discordances, between our findings and species diversity studies in freshwater fishes (Vellend and Geber, 2005; Fourtune *et al.*, 2016). For example, genetic diversity does not only reflects drainage area (Figure 4), but species richness-area relationships have been largely observed for current and past drainages over the world and for the study region, the Neotropics (Albert *et al.*, 2011; Dias *et al.*, 2014). On the other hand, our focus here was on the recent evolutionary past, and this temporal scale does not correspond to the divergence times estimated for fish species diversification, which often predates the Pleistocene (Lundberg *et al.*, 2000). This does not necessarily mean that other mechanisms did not contribute to species diversity patterns in the more distant past. However, because the sea-level changes during the LGM were some of the most extreme events and temporally matches with most of the point estimates for divergence times, these recent events would over-ride divergences associated with the more distant past (Papadopoulou and Knowles, 2015) if the geography of such divergence patterns were generally coincident with those of the LGM (for an exception, we note the regional split between the northern and southern regions, Supplementary Figure 3, which does not coincide with the boundary of recognized paleodrainages; Thomaz *et al.*, 2015). This argument is also predicated on the presumed importance of divergence associated with geographic isolation, and it does not address whether other evolutionary processes (for example, selection) might have played more or less of an important role in the past relative to the present.

Of the populations of *Hollandichthys* studied here, individuals from paleodrainage 11 (Figure 1) have recently been described as a putative new species, *H. taramandahy* (Bertaco and Malabarba, 2013). The strong structuring of genetic variation in *Hollandichthys* may be indicative of a putative species boundary, and consequently, suggest that paleodrainages may be responsible for long-term isolation that culminates in speciation. However, the degree of genomic differentiation for this putative species is similar to those observed between the populations from other paleodrainages in *Hollandichthys*, as is estimates of its divergence (i.e., ~24 ky, Figure 3). It is unknown whether any of a set of geographically isolated regions/populations will become new species (Sukumaran and Knowles, 2017), but the clear morphological characters (Bertaco and Malabarba, 2013) of the proposed new taxon may suggest that this lineage has preceded beyond what might be expected from geographic isolation at the microevolutionary level (i.e., population isolation; Rosenblum *et al.*, 2012). Further analyses that tests for morphological differences across individuals in the other paleodrainages are required to determine whether the differentiation observed in the new putative species *H. taramandahy* (Bertaco and Malabarba, 2013) is statistically equivalent to other divergences separating paleodrainage populations that are not recognized as different species (for example, Solis-Lemus *et al.*, 2014; Huang and Knowles, 2016).

CONCLUSIONS

Our study not only highlights the effect of Pleistocene paleodrainages on patterns of genetic variation in a freshwater fish species along basins of the Brazilian Atlantic coast but the findings also may help explain why support for paleodrainages as drivers of divergence across taxa and continents have not been consistent. Specifically, the properties that impact population isolation and connectivity in riverine systems may be linked to those of past paleodrainages, not necessarily the current landscape. Given these phenomena occur over short evolutionary timescales, we also highlight how biases in the test applied and/or interpretation of results can contribute to ambiguities regarding the effects of past river landscapes, as well as how the development of new tools for modeling riverine environments that parallel those from the terrestrial realm will promote more refined hypotheses that could help explain differences in genetic variation across regions and/or taxa.

DATA ARCHIVING

Scripts and data are available from the Dryad repository: http://dx.doi.org/10.5061/dryad.7hr7f.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This work was funded by the Hubbs, Carl L. and Laura C. Fellowship from University of Michigan Museum of Zoology (UMMZ), by the Ichthyology Student Award from UMMZ (to ATT), by the NSF Dissertation Improvement Grant DEB-15-01301 (to LLK and ATT) and by CNPq—Brazil 307890/2016-3 (to LRM). We thank all people who generously contributed samples for this study—specifically, C Lucena (MCP) and V Abilhoa (MHNCI), as well as all the people who contributed to fieldwork—specifically, V Bertaco, F Carvalho, J Ferrer, F Jerep, G Neves and J Wingert. Finally, we thank TP Carvalho and K Marske, and three anonymous reviewers who provided helpful comments that improved the quality of the manuscript, and TP Carvalho also for the *H. multifasciatus* picture.

Albert JS, Petry P, Reis RE (2011). Major biogeographic and phylogenetic patterns. In: Albert JS, Reis RE (eds). *Historical Biogeography of Neotropical Freshwater Fishes*. University of California Press: Berkley, CA, USA, pp 21–57.

Bertaco VA, Malabarba LR (2013). A new species of the characid genus *Hollandichthys* Eigenmann from coastal rivers of southern Brazil (Teleostei: Characiformes) with a discussion on the diagnosis of the genus. *Neotrop Ichthyol* **11**: 767–778.

Burridge CP, Craw D, Waters JM (2007). An empirical test of freshwater vicariance via river capture. *Mol Ecol* **16**: 1883–1895.

Carstens BC, Knowles LL (2007). Shifting distributions and speciation: species divergence during rapid climate change. *Mol Ecol* **16**: 619–627.

Catchen J, Hohenlohe P, Bassham S, Amores A, Cresko W (2013). Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**: 3124–3140.

Chakona A, Swartz ER, Gouws G (2013). Evolutionary drivers of diversification and distribution of a southern temperate stream fish assemblage: testing the role of historical isolation and spatial range expansion. *PLoS One* **8**: 1–13.

Dias MS, Oberdorff T, Hugueny B, Leprieur F, Jézéquel C, Cornu JF *et al.* (2014). Global imprint of historical connectivity on freshwater fish biodiversity. *Ecol Lett* **17**: 1130–1140.

Domisch S, Amatulli G, Jetz W (2015). Near-global freshwater-specific environmental variables for biodiversity analyses in 1 km resolution. *Sci Data* **2**: 150073.

Earl DA, vonHoldt BM (2012). STRUCTURE HARVESTER: a website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conserv Genet Resour* **4**: 359–361.

Edwards S, Beerli P (2000). Perspective: gene divergence, population divergence, and the variance in coalescence time in phylogeographic studies. *Evolution* **54**: 1839–1854.

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

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M (2013). Robust demographic inference from genomic and SNP data. *PLoS Genet* **9**: 1–17.

Excoffier L, Foll M (2011). Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics* **27**: 1332–1334.

Excoffier L, Lischer HE (2010). Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Mol Ecol Res* **10**: 564–567.

Fourtune L, Paz-Vinas I, Loot G, Prunier JG, Blanchet S (2016). Lessons from the fish: a multi-species analysis reveals common processes underlying similar species-genetic diversity correlations. *Freshw Biol* **61**: 1830–1845.

He Q, Edwards DL, Knowles LL (2013). Integrative testing of how environments from the past to the present shape genetic structure across landscapes. *Evolution* **67**: 3386–3402.

Huang H, Knowles LL (2014). Unforeseen consequences of excluding missing data from next-generation sequences: simulation study of RAD sequences. *Syst Biol* **65**: 357–365.

Huang JP, Knowles LL (2016). The species versus subspecies conundrum: quantitative delimitation from integrating multiple data types within a single Bayesian approach in Hercules beetles. *Syst Biol* **65**: 685–699.

Hudson RR, Coyne JA (2002). Mathematical consequences of the genealogical species concept. *Evolution* **56**: 1557–1565.

Kisel Y, Barraclough TG (2010). Speciation has a spatial scale that depends on levels of gene flow. *Am Nat* **175**: 316–334.

Knowles LL (2009). Statistical phylogeography. *Ann Rev Ecol Evol Syst* **40**: 593–612.

Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015). Clumpak: a program for identifying clustering modes and packaging population structure inferences across K. *Mol Ecol Resour* **15**: 1179–1191.

Lundberg JG, Kottelat M, Smith GR, Stiassny ML, Gill AC (2000). So many fishes, so little time: an overview of recent ichthyological discovery in continental waters. *Ann Mo Bot Gard* **87**: 26–62.

Lynch M (2010). Evolution of the mutation rate. *Trends Genet* **26**: 345–352.

Massatti R, Knowles LL (2014). Microhabitat differences impact phylogeographic concordance of codistributed species: genomic evidence in montane sedges (*Carex* L.) from the Rocky Mountains. *Evolution* **68**: 2833–2846.

Massatti R, Knowles LL (2016). Contrasting support for alternative models of genomic variation based on microhabitat preference: species-specific effects of climate change in alpine sedges. *Mol Ecol* **25**: 3974–3986.

Mazerolle MJ (2016). AICcmodavg: Model selection and multimodel inference based on (Q) AIC(c). R package version 2.0-4. Available at: http://CRAN.R-project.org/package = AICcmodavg (accessed on January 2017).

Medrano JF, Aasen E, Sharrow L (1990). DNA extraction from nucleated red blood cells. *Biotechniques* **8**: 43.

Morrissey MB, de Kerckhove DT (2009). The maintenance of genetic variation due to asymmetric gene flow in dendritic metapopulations. *Am Nat* **174**: 875–889.

Neuenschwander S, Largiadèr CR, Ray N, Currat M, Vonlanthen P, Excoffier L (2008). Colonization history of the Swiss Rhine basin by the bullhead (*Cottus gobio*): inference under a Bayesian spatially explicit framework. *Mol Ecol* **17**: 757–772.

Papadopoulou A, Knowles LL (2015). Genomic tests of the species-pump hypothesis: recent island connectivity cycles drive population divergence but not speciation in Caribbean crickets across the Virgin Islands. *Evolution* **69**: 1501–1517.

Papadopoulou A, Knowles LL (2016). Toward a paradigm shift in comparative phylogeography driven by trait-based hypotheses. *Proc Natl Acad Sci USA* **113**: 8018–8024.

Papadopoulou A, Knowles LL (2017). Linking micro- and macroevolutionary perspectives to evaluate the role of Quaternary sea-level oscillations in island diversification. *Evolution*. in review.

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle C (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol Ecol* **21**: 2991–3005.

Paz-Vinas I, Loot G, Stevens VM, Blanchet S (2015). Evolutionary processes driving spatial patterns of intraspecific genetic diversity in river ecosystems. *Mol Ecol* **24**: 4586–4604.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D *et al.* (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**: 559–575.

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

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/ (accessed on January 2017).

Rosenblum EB, Sarver BA, Brown JW, Des Roches S, Hardwick KM, Hether TD *et al.* (2012). Goldilocks meets Santa Rosalia: an ephemeral speciation model explains patterns of diversification across time scales. *Evol Biol* **39**: 255–261.

Ryan PG, Bloomer P, Moloney CL, Grant TJ, Delport W (2007). Ecological speciation in South Atlantic island finches. *Science* **315**: 1420–1423.

Solis-Lemus C, Knowles LL, Ané C (2014). Bayesian species delimitation combining multiple genes and traits in a unified framework. *Evolution* **69**: 492–507.

Sukumaran J, Knowles LL (2017). The multiple coalescent delimits structure not species. *Proc Natl Acad Sci USA* **144**: 1607–1612.

Tedesco PA, Leprieur F, Hugueny B, Brosse S, Dürr HH, Beauchard O *et al.* (2012). Patterns and processes of global riverine fish endemism. *Glob Ecol Biogeogr* **21**: 977–987.

Thomaz AT, Christie MR, Knowles LL (2016). The architecture of river networks can drive the evolutionary dynamics of aquatic populations. *Evolution* **70**: 731–739.

Thomaz AT, Malabarba LR, Bonatto SL (2010). The phylogenetic placement of *Hollandichthys* Eigenmann 1909 (Teleostei: Characidae) and related genera. *Mol Phylogenet Evol* **57**: 1347–1352.

Thomaz AT, Malabarba LR, Bonatto SL, Knowles LL (2015). Testing the effect of palaeodrainages versus habitat stability on genetic divergence in riverine systems: study of a Neotropical fish of the Brazilian coastal Atlantic Forest. *J Biogeogr* **42**: 2389–2401.

Unmack PJ, Hammer MP, Adams M, Johnson JB, Dowling TE (2013). The role of continental shelf width in determining freshwater phylogeographic patterns in southeastern Australian pygmy perches (Teleostei: Percichthyidae). *Mol Ecol* **22**: 1683–1699.

Vellend M, Geber MA (2005). Connections between species diversity and genetic diversity. *Ecol Lett* **8**: 767–781.

Waters JM, Burridge CP (2016). Fine-scale habitat preferences influence within-river population connectivity: a case-study using two sympatric New Zealand Galaxias fish species. *Freshw Biol* **61**: 51–56.

Supplementary Information accompanies this paper on Heredity website (http://www.nature.com/hdy)