TABLE S1 | Detailed description of the filtering scheme applied to the RADseq dataset of the three armored catfish species in this study. The order of the rows indicates the order in which filters were applied from the bioinformatics processing step to the filtering step to select neutral SNPs. [1] Each species started the quality filtering step from a total of 25,676 raw SNPs. [2] The genotypes AK407_S9 and AK454_S69 (belonging to *Baryancistrus longipinnis* species) were eliminated of analysis due to missing data (> 75%), resulting in a final total of 24 samples for the subsequent steps. [3] The genome scans of these species were not applied due to the low number of samples collected. *Quality filters are described in the Materials and Methods. The number of SNPs retained in bold in each species at the end of each quality filtering step were assumed to be neutral SNPs. N_i: Initial sample size (number of individuals entering the neutral dataset filtering); N_f: Final sample size (number of individuals after all neutral SNP filtering for that dataset).

Filtering Steps	Count / Result	Description / Parameters
1. Bioinformatics Processing (All Species)		
Raw Data	47.584.483 Reads	Initial total sequencing reads.
In silico Digestion	45.244.760 Reads	Reads after in silico restriction enzyme digestion simulation.
Quality Trimming	45.235.280 Reads	Reads retained after quality score (Q20) and head cropping (5bp) filtering.
Length Filtering	39.869.024 Reads	Reads retained after filtering for a minimum length of 144 base pairs.
Total Remaining Reads	83.79% Reads	Percentage of reads remaining after all processing steps
2. De novo Assembly & Initial SNP Filtering (All Species) [1]		
Initial Assembly	475.061 raw SNPs	Raw SNPs identified after de novo assembly using Stacks: -m 3 (min identical raw reads to form a stack); -M 2 (mismatches allowed between loci when processing the catalog).
SNP Filtering	25.676 raw SNPs	SNPs retained after filtering for minor allele frequency. single SNP per locus. maximum observed heterozygosity. and proportion of missing data: –min-maf 0.025; –write-single-snp; –max-obs-het 0.7; -r 0.5.
3. Neutral Dataset (*B. longipinnis*) [2]		
Quality Filter*	5.143 SNPs (N_i = 26)	SNPs retained based on read depth and missing data thresholds.
HWE Filter	5.131 SNPs	SNPs retained after Hardy-Weinberg Equilibrium (HWE) filtering.
LD Filter (Between Contigs)	3.478 SNPs	SNPs retained after linkage disequilibrium (LD) filtering between contigs.
FST Outliers (sNMF)	**2.758 SNPs** (N_f = 24)	SNPs remaining after identification and removal of FST outlier loci detected by sNMF.
4. Neutral Dataset (*B. niveatus*) [3]		
Quality Filter*	6.088 SNPs (N_i = 15)	SNPs retained based on read depth and missing data thresholds.
HWE Filter	6.088 SNPs	SNPs retained after Hardy-Weinberg Equilibrium (HWE) filtering.
LD Filter (Between Contigs)	**5.040 SNPs** (N_f = 15)	SNPs retained after linkage disequilibrium (LD) filtering between contigs.
5. Neutral Dataset (*Baryancistrus* sp. "Araguaia") [3]		
Quality Filter*	3.701 SNPs (N_i = 6)	SNPs retained based on read depth and missing data thresholds.
HWE Filter	3.701 SNPs	SNPs retained after Hardy-Weinberg Equilibrium (HWE) filtering.
LD Filter (Between Contigs)	**3.508 SNPs** (N_f = 6)	SNPs retained after linkage disequilibrium (LD) filtering between contigs.

TABLE S2 | List of specimens of the three armored catfish Baryancistrus species used in genomic analyses with the number of raw reads (RawData) after RAD sequencing. Average coverage depth and missing data values (in percentage) at the end of each quality filtering step (neutral datasets each species). Each sample was classified by species using a morphological classification key and RADSeq (Fig. 2). *These are erroneous IDs using morphological classification keys that were identified and renamed from molecular information.

Sample ID [1]	RawData (Reads)	Coverage	Missing Data (%)	Morphological ID	RADSeq ID
AK027_S62	900.945	6.55	0.54	*B. niveatus*	*B. longipinnis**
AK085_S45	833.879	7.35	0.53	*B.* sp. "Araguaia"	*B. longipinnis**
AK127_S26	845.088	7.41	0.48	*B. longipinnis*	*B. longipinnis*
AK129_S18	849.067	7.30	0.51	*B.* sp. "Araguaia"	*B. longipinnis**
AK452_S52	911.211	6.54	0.69	*B. longipinnis*	*B. longipinnis*
AK453_S77	966.835	9.18	0.11	*B. longipinnis*	*B. longipinnis*
AK455_S41	759.689	7.82	0.24	*B. longipinnis*	*B. longipinnis*
AK456_S43	1.774.793	12.52	0.04	*B. longipinnis*	*B. longipinnis*
AK464_S85	1.394.290	10.01	0.08	*B. longipinnis*	*B. longipinnis*
AK465_S82	939.589	7.98	0.24	*B. longipinnis*	*B. longipinnis*
AK466_S56	933.052	7.10	0.38	*B. longipinnis*	*B. longipinnis*
AK467_S67	1.108.911	9.60	0.08	*B. longipinnis*	*B. longipinnis*
AK468_S15	810.703	11.16	0.13	*B. longipinnis*	*B. longipinnis*
AK487_S68	967.227	12.90	0.08	*B. longipinnis*	*B. longipinnis*
AK493_S47	866.530	12.67	0.13	*B. longipinnis*	*B. longipinnis*
AK494_S66	916.388	11.89	0.06	*B. longipinnis*	*B. longipinnis*
AKA119_S5	1.214.243	11.04	0.15	*B.* sp. "Araguaia"	*B. longipinnis**
AKA261_S84	1.135.738	9.94	0.16	*B. niveatus*	*B. longipinnis**
AKA263_S32	1.088.360	14.50	0.19	*B. niveatus*	*B. longipinnis**
AKA264_S87	981.057	11.89	0.18	*B. niveatus*	*B. longipinnis**
AKA269_S23	938.877	8.41	0.23	*B.* sp. "Araguaia"	*B. longipinnis**
AKA465_S71	699.558	7.56	0.26	*B. longipinnis*	*B. longipinnis*
AKA466_S2	1.062.569	9.48	0.11	*B. longipinnis*	*B. longipinnis*
AKA467_S29	1.003.936	9.36	0.16	*B. longipinnis*	*B. longipinnis*
AKRT015_S39	945.697	7.43	0.21	*B. niveatus*	*B.* sp. "Araguaia"*
AKRT030_S6	1.191.172	10.83	0.05	*B. niveatus*	*B.* sp. "Araguaia"*
AKRT035_S35	1.311.571	11.752	0.09	*B.* sp. "Araguaia"	*B.* sp. "Araguaia"
AKRT044_S40	3.106.590	20.79	0.05	*B.* sp. "Araguaia"	*B.* sp. "Araguaia"
AKRT045_S24	1.049.033	9.69	0.07	*B. niveatus*	*B.* sp. "Araguaia"*
AKRT098_S19	1.233.087	6.10	0.59	*B. niveatus*	*B.* sp. "Araguaia"*
AK137_S65	814.367	7.11	0.33	*B. niveatus*	*B. niveatus*
AK213_S79	779.951	6.95	0.31	*B. niveatus*	*B. niveatus*
AK214_S14	754.039	6.23	0.74	*B. niveatus*	*B. niveatus*
AK276_S33	767.224	6.38	0.56	*B. niveatus*	*B. niveatus*
AK289_S88	896.493	7.13	0.30	*B. niveatus*	*B. niveatus*
AK333_S73	1.003.329	6.75	0.56	*B. niveatus*	*B. niveatus*
AK365_S61	881.455	6.37	0.57	*B. niveatus*	*B. niveatus*
AK471_S78	768.915	7.29	0.24	*B. niveatus*	*B. niveatus*
AK472_S51	713.000	7.66	0.21	*B. niveatus*	*B. niveatus*
AKA262_S44	1.030.088	9.30	0.12	*B. niveatus*	*B. niveatus*
AKA265_S30	1.090.317	10.27	0.06	*B. niveatus*	*B. niveatus*
AKA267_S3	1.145.981	10.26	0.06	*B. niveatus*	*B. niveatus*
AKA268_S21	1.059.118	8.63	0.12	*B. niveatus*	*B. niveatus*
AKRT052_S8	1.303.540	10.40	0.05	*B. niveatus*	*B. niveatus*
AKRT097_S58	928.376	10.72	0.17	*B. niveatus*	*B. niveatus*

(a) Laboratory methods for ddRADseq

RADSeq libraries were generated following the ddRADseq methodology of double digest restriction site-associated DNA sequencing (Peterson *et al*., 2012, adapted by Campos *et al*., 2017) using the enzymes EcoRI and MspI. Genomic DNA (200 ng) of each sample was digested at 37°C for 3 h using 10 U of EcoR1 (six-base recognition site, GAATTC) and 5 U of MspI (four-base recognition site, CCGG), and 1x CutSmart Buffer (New England Biosciences), made up to a total volume of 40 µL with PCR-grade H2O. Digested sample purification was carried out using the Agencourt AMPure XP Beads PCR Purification Kit® (Beckman Coulter, CA, USA). Pairs of hybridized adapters corresponding to each enzyme were then ligated to the digested DNA in a final reaction volume of 40 µL, consisting of 31,5 µL of digested DNA, 0,5 µL of T4 DNA Ligase (Promega), 2,0 µL of each adapter, and 4,0 µL of 1x T4 DNA Ligase Buffer (Promega, EUA). The mixture was then incubated at 23°C for 30 min, followed by heat killing at 65°C for 10 min, and again purified using Agencourt AMPure XP beads.

In the indexing reaction, polymerase chain reaction (PCR) was used in each sample to attach the i5 and i7 index primers (Nextera DNA CD indexes - 96 indexes, 96 samples, IIllumina, San Diego, CA, USA). The PCR was composed of 15 µL of ligation products, 25 µL of ligation products GoTaq ®Master Mixes (Promega, EUA) and 5 µL each of the i5 and i7 primers. The PCR had as step first a cycle of 72°C for 3 min, which was followed by 16 cycles of denaturing at 95°C for 30 sec and 95°C for 10 sec, annealing at 55°C for 30 sec, and extension at 72°C for 30 sec. Each of the PCRs was purified with Agencourt AMPure XP beads, the concentration of each sample was adjusted to 25 ng/µl based on the Qubit reading (ng/µL), and the weighted average fragment size (bp) was checked on a 1% agarose gel.

All samples were pooled and purified again using Agencourt AMPure XP beads. For size selection, fragments between 300–500 bp were selected using a 1% agarose gel, and then size-selected sample pools were purified with a Wizard® SV Gel and PCR Clean-Up System kit (Promega, EUA) following the manufacturer's instructions. For size selection, fragments between 300–500 bp were selected using a 1% agarose gel, and then size-selected sample pools were purified with a Wizard® SV Gel and PCR Clean-Up System kit (Promega, EUA) following the manufacturer's instructions. The final library was quantified using Qubit and real-time PCR (qPCR) and then submitted to Illumina sequencing (see material and methods section).

(b) Sample size sufficiency analysis

To assess whether our sample sizes were sufficient to yield reliable estimates of genetic diversity and differentiation, we applied the Sample Size Impact (SaSii) framework, a resampling-based approach implemented in R (Scaketti *et al*., 2025). Rather than prescribing a universal minimum sample size, SaSii helps define dataset-specific thresholds by simulating how genetic metrics stabilize with increasing sample size.

The method evaluates sampling sufficiency by comparing subsets of increasing size (starting from N = 5 individuals) to the full dataset, across multiple repetitions. For each sample size, it calculates summary statistics and tracks how closely these match the full data. We performed independent analyses for *Baryancistrus longipinnis* (N = 24)

and *B. niveatus* (N = 15), using high-quality SNP datasets filtered by a minor allele frequency (MAF) ≥ 0.05. Parameters used: Starting number of individuals = 5; Number of repetitions per sample size = 50; Threshold for defining common alleles = 0.05. We extracted and visualized the following metrics from SaSii output: (a) Fraction of common alleles detected; (b) Mean difference from original allele frequencies; (c) Mean expected heterozygosity (He); and (d) Mean pairwise F_{ST}. This analysis supports sample design and interpretation by identifying the minimum number of individuals needed to recover 95% of common alleles and achieve sTABLE estimates of He and F_{ST}.

(c) Morphology check of genetic outliers

To further explore the discrepancies identified in the genomic analyses, we conducted a targeted morphological assessment of individuals classified as genetic outliers. Although a comprehensive phenotypic analysis is reserved for a forthcoming study, this preliminary evaluation aimed to determine whether observed morphological inconsistencies could be attributed to factors such as incomplete maturity or phenotypic plasticity. Notably, a size difference was observed between genetic outliers and concordant individuals (outliers tended to be larger, with a mean length of ~127 mm compared to ~115 mm for concordant individuals, see **S14**: Additional information on Results). This size variation may contribute to inconsistencies in morphological identification, as diagnostic traits can change throughout ontogeny (Rapp Py-Daniel *et al*., 2011, potentially leading to misclassification under morphology-based taxonomy.

To contextualize and support the observed mismatches between initial morphological identification and final molecular classification, we carried out complementary morphometric analyses. Recognizing that ontogeny can influence the expression of diagnostic features, a Principal Component Analysis (PCA) was performed using relative morphological measurements, *i.e*., all linear traits expressed as proportions of Standard length. This approach helps mitigate the influence of absolute body size and highlights shape differences. PCA was conducted using the *prcomp* function from the *stats* package in R (R Core Team), with scaling enabled. The analysis included individuals from all three *Baryancistrus* species identified in our study. Importantly, group classification was based on genomic concordance, with individuals labeled as either "*concordant*" or "*outlier*" according to their morphological versus genetic assignment. The PCA was thus not intended to distinguish species per se, but rather to assess the morphological consistency of genomic classifications. The variables included in the analysis were: Cleithrum width, Interorbital distance, Lower lip width, Head width, Pre-dorsal length, Head length, Distance between pectoral and pelvic fins, Post-anal length, Body width at dorsal fin origin, Orbit diameter, Snout length, Internarial distance, Dentary length, and Premaxillary length.

Following PCA, Wilcoxon rank-sum tests were applied to each relative morphological variable to identify which traits significantly differed between genomic groups. These non-parametric tests were implemented using the *wilcox.test* function from the stats package in R. This combined multivariate and univariate strategy offers a robust framework for assessing whether morphological patterns corroborate or contradict genetic structure, supporting the validity of our molecular-based taxonomic reassignment.

(d) Population structure

In both sNMF and DAPC, assumption-free methods, the number of ancestral populations (k) was allowed to vary between 1 and 10, according to previous studies carried out on multiple Amazonian species (Lanes *et al.*, 2018; Carvalho *et al.*, 2019; Silva *et al.*, 2020). For sNMF, ten run replicates for each k-value were performed under different alpha values (α = 1, 10, 100, 500) using a burn-in of 200 iterations, and the best k was chosen based on cross-entropy (Frichot *et al.*, 2014). For DAPC, the number of clusters was assessed by the "*find.cluster*" function, and the best-supported number of genetic clusters was inferred from the Bayesian Information Criterion (BIC) using all available principal components (100% of the variance) (Jombart *et al.*, 2010).

(e) Spatial genetic structure analyses

To investigate Isolation by Distance (IBD) and evaluate the suitability of different geographical distance representations in fluvial environments, we employed two complementary approaches: Landscape Mixed-Effects Models (MLPE) and Mantel tests. Prevosti's inter-individual genetic distance was used as the response variable in both methods. In MLPE models, genetic distance was modeled as a function of geographical distances (Euclidean or riverine) as fixed effects, with random effects terms for the identity of originating and destination populations, controlling for the non-independence of paired observations (GeneticDist | GeographicDistance + (1|pop1) + (1|pop2)). Model selection was based on the Akaike Information Criterion corrected for small sample sizes (AICc). Marginal R^2 (R^2_m) was calculated for each model (using the *r.squaredGLMM* function from the *MuMIn* package), indicating the proportion of genetic variance explained exclusively by the fixed effects.

Mantel tests were performed in parallel between all pairs of individuals to estimate the correlation between a genetic distance matrix and a geographic distance matrix. Specifically for IBD, we correlated Prevosti's genetic distance matrix with both Euclidean and riverine geographic distance matrices using the mantel.randtest function from the R adegenet package (Jombart, 2008) with 999 replicates. Prevosti's genetic distance matrix was calculated using the *provesti.dist* function of the R package poppr (Kamvar *et al.*, 2014). Euclidean geographic distance matrices were generated using the dist function of the R package adegenet. The statistical significance of these correlations was assessed through permutations. The comparison between Euclidean and riverine distances was included in response to reviewer suggestions. Results from MLPE models, particularly the lower AICc values consistently observed for riverine distance across all species and regions, indicated that riverine distance was the most suiTABLE representation for landscape genetics analyses (Tab. **S8**). This metric was primarily used in subsequent evaluations of the influence of multiple landscape factors.

The fine-scale spatial genetic structure within each species was assessed by quantifying spatial autocorrelation in genetic relatedness statistics (Aij) (Yang *et al.*, 2010). These analyses were performed using the local polynomial fitting (LOESS) of pairwise genetic relatedness (for both indexes) for different pairwise fluvial geographical distances (https://github.com/rojaff/Lplot; Bruno *et al.*, 2008). The genetic relatedness statistic (Aij) was calculated with the Relatedness function implemented in the *r2vcftools* package.

As a final step, we investigated the presence of barriers to gene flow by conducting separate analyses for each species and region. We conducted geographic barrier

detection to assess potential constraints to gene flow and evaluate evidence for isolation by barrier (IBB) in the TO-AR basin. We applied Monmonier's method (Monmonier, 1973) using a K-Nearest Neighbors-based spatial connectivity network implemented in the R package adegenet (Jombart, 2008) to identify spatial genetic discontinuities. Specifically, this approach maximizes genetic distances along a network of geographically connected localities to delineate genetic boundaries. The analysis involved calculating a pairwise Prevosti's genetic distance matrix, projecting sample coordinates into the UTM coordinate system to ensure spatial accuracy, and generating a Delaunay triangulation to establish spatial connections (*e.g.*, K = 4 nearest neighbors). The resulting barriers were overlaid on the hydrographic network and sampling locations. Their consistency with physical features of the landscape (*e.g.*, rapids, dams, tributaries, or other environmental discontinuities) was used to infer potential IBB effects.

All analyses were conducted in the R statistical environment, utilizing the *lme4*, *lmerTest*, *MuMIn*, *ade4*, *adegenet*, *poppr*, and *r2vcftools* packages.

(f) Buffer optimization and spatial scale strategies

Defining the appropriate spatial scale is a critical component in landscape genetics studies, as it directly influences the ability to detect meaningful relationships between landscape features and genetic structure. To address this issue robustly in our study, we employed and compared three distinct buffer definition strategies for each landscape variable (habitat amount in 1985, habitat amount in 2020, and habitat loss) and for each genetic response variable (expected heterozygosity – *He*; and inbreeding coefficient – *F*):

(1) Variable-optimized buffer strategy: For each combination of landscape variable and genetic response, we performed buffer size optimization. This process involved evaluating a wide range of buffer sizes, encompassing both 'Micro' (1 km, 2 km, 3 km, 4 km, and 5 km) and 'Macro' (10 km, 15 km, and 20 km) spatial scales. The best-performing model, indicated by the lowest AICc value for each landscape variable (historical, loss, and current) within each scale (Micro or Macro) and for each genetic variable (*He* or *F*), was selected for a final competition among all optimized models. This approach allows the landscape influence scale to be data-driven, reflecting specific ecological processes.

(2) Fixed micro buffer strategy: To complement the optimized approach and explore the influence of processes at a finer spatial scale, we applied a series of fixed buffer sizes at the 'micro' scale (1 km, 2 km, 3 km, 4 km, and 5 km). The best model, based on the lowest AICc value for each genetic variable (*He* or *F*) and applied exclusively to the current landscape variable (habitat amount in 2020), was then replicated as a standard for the other landscape variables (habitat amount in 1985 and habitat loss) within this fixed buffer strategy. This scale was selected considering the species' limited geographic distribution, restriction to rheophilic environments, and low dispersal capability. These factors increase their susceptibility to habitat disturbance, thereby affecting their genetic variability.

(3) Fixed macro buffer strategy: Similarly, to assess the influence of broader-scale processes, the same procedure described for the micro scale was applied using fixed buffer sizes at the 'macro' scale (10 km, 15 km, and 20 km). The best-performing macro buffer, based on the lowest AICc value for each genetic variable (*He* or *F*) and optimized for the current landscape variable (habitat amount in 2020) was then used as a standard for the other landscape variables (habitat amount in 1985 and habitat loss) within this macro buffer strategy. This scale aimed to capture large-scale processes potentially affecting genetic patterns, such as regional fragmentation driven by agricultural expansion, urban development, and deforestation. For example, Fig. **S12**, illustrates the landscape raster for 2020, where the study area for *B. longipinnis* (Regions 1 and 2) is specifically defined by a 20 km buffer around the sampling points.

The comparison of results obtained through these three strategies, as detailed in Tab. **S13**, enabled a comprehensive assessment of landscape effect scales on the genetic patterns of the studied species, providing insights into the sensitivity of model outcomes to spatial scale choices.

REFERENCES

- **Bruno C, Macchiavelli R, Balzarini M.** Non-parametric smoothing of multivariate genetic distances in the analysis of spatial population structure at fine scale. Theor Appl Genet. 2008; 117:435–47. https://doi.org/10.1007/s00122-008-0788-1
- **Campos M, Conn JE, Alonso DP, Vinetz JM, Emerson KJ, Ribolla PEM.** Microgeographical structure in the major Neotropical malaria vector *Anopheles darlingi* using microsatellites and SNP markers. Parasites Vectors. 2017; 10:76. https://doi.org/10.1186/s13071-017-2014-y
- **Carvalho CS, Lanes ECM, Silva AR, Caldeira CF, Carvalho-Filho N, Gastauer M *et al.*** Habitat loss does not always entail negative genetic consequences. Front Genet. 2019; 10:1011. https://doi.org/10.3389/fgene.2019.01101
- **Frichot E, Mathieu F, Trouillon T, Bouchard G, François O.** Fast and efficient estimation of individual ancestry coefficients. Genetics. 2014; 196(4):973–83. https://doi.org/10.1534/genetics.113.160572
- **Jombart T.** adegenet: a R package for the multivariate analysis of genetic markers. Bioinformatics. 2008; 24(11):1403–05. https://doi.org/10.1093/bioinformatics/btn129
- **Jombart T, Devillard S, Balloux F.** Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. BMC Genet. 2010; 11:94. https://doi.org/10.1186/1471-2156-11-94
- **Kamvar ZN, Tabima JF, Grünwald NJ.** *Poppr*: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. PeerJ. 2014; 2:e281. https://doi.org/10.7717/peerj.281
- **Lanes EC, Pope NS, Alves R, Carvalho Filho NM, Giannini TC, Oliveira G *et al.*** Landscape genomic conservation assessment of a narrow-endemic and a widespread morning glory from Amazonian savannas. Front Plant Sci. 2018; 9:532. https://doi.org/10.3389/fpls.2018.00532
- **Monmonier MS.** Maximum-difference barriers: an alternative numerical regionalization method*. Geogr Anal. 1973; 5:245–61. https://doi.org/10.1111/j.1538-4632.1973.tb01011.x

- **Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE *et al.*** Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE. 2012; 7(5):e37135. https://doi.org/10.1371/journal.pone.0037135
- **Scaketti M, Sujii PS, Alves-Pereira A, Schwarcz KD, Francisconi AF, Moro MS *et al.*** Sample size impact (SaSii): an R script for estimating optimal sample sizes in population genetics and population genomics studies. *PLoS ONE. 2025;* 20(2):e0316634. https://doi.org/10.1371/journal.pone.0316634
- **Silva AR, Resende-Moreira LC, Carvalho CS, Lanes ECM, Ortiz-Vera MP, Viana PL *et al.*** Range-wide neutral and adaptive genetic structure of an endemic herb from Amazonian Savannas. AoB Plants. 2020; 12:plaa003. https://doi.org/10.1093/aobpla/plaa003
- **Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR *et al.*** Common SNPs explain a large proportion of the heritability for human heigth. Nat Genet. 2010; 42:565–69. https://doi.org/10.1038/ng.608

TABLE S4 | Description of sampling sites for the three armored catfish Baryancistrus species included in genomic analyses, detailing sampling locality and region, number of samples, geographical coordinates, and main anthropic impacts. [1] Region: Region 1 and Region 2 are microregions near the Tucuruí Dam and around the city of Marabá, respectively (details of the map can be seen in Fig. 1), where Region_Outside is the location of sampling outside of Region 1 and Region 2 (collected sites unique to the Tocantins River). [3] Main Anthropic Impact: Description of the main anthropic impacts that are expanding strongly in each location, which may pose threats to the life of rheophilic species. During fieldwork at the sampled sites, the type of impact was characterized. Habitat loss over time can be visually assessed in Fig. **S12**. N: number of samples collected per site.

Species	Site Code (Region)[1]	N	Longitude	Latitude	Main Anthropic Impact[3]
B. longipinnis	Pedral do Lourenço (Region_1)	3	-49.364	-4.953	Hydroelectric projects
	Pedral do Lourenço (Region_1)	12	-49.336	-4.992	Hydroelectric projects
	Marabá (Region_2)	5	-49.073	-5.312	Pastures and/or agriculture activities
	Cabras Island (Region_2)	4	-48.959	-5.294	Pastures and/or agriculture activities
B. niveatus	Pedral do Lourenço (Region_1)	2	-49.336	-4.992	Hydroelectric projects
	Marabá (Region_2)	4	-49.073	-5.312	Urban expansion
	Itacaiúnas (Region_2)	1	-49.141	-5.355	Urban expansion
	Cabras Island (Region_2)	1	-48.959	-5.294	Pastures and/or agriculture activities
	Ilha do Jaú (Region_2)	3	-48.854	-5.337	Pastures and/or agriculture activities
	Cajú Amigo (Region_2)	2	-48.788	-5.341	Pastures and/or agriculture activities
	Porto Franco (Region_Outside)	1	-47.386	-6.286	Pastures and/or agriculture activities
	Peixe (Region_Outside)	1	-48.444	-12.225	Pastures and/or agriculture activities
Baryancistrus sp. "Araguaia"	Itaguatins (Region_Outside)	1	-47.479	-5.769	Pastures and/or agriculture activities
	Porto Franco (Region_Outside)	4	-47.386	-6.286	Pastures and/or agriculture activities
	Peixe (Region_Outside)	1	-48.444	-12.225	Pastures and/or agriculture activities

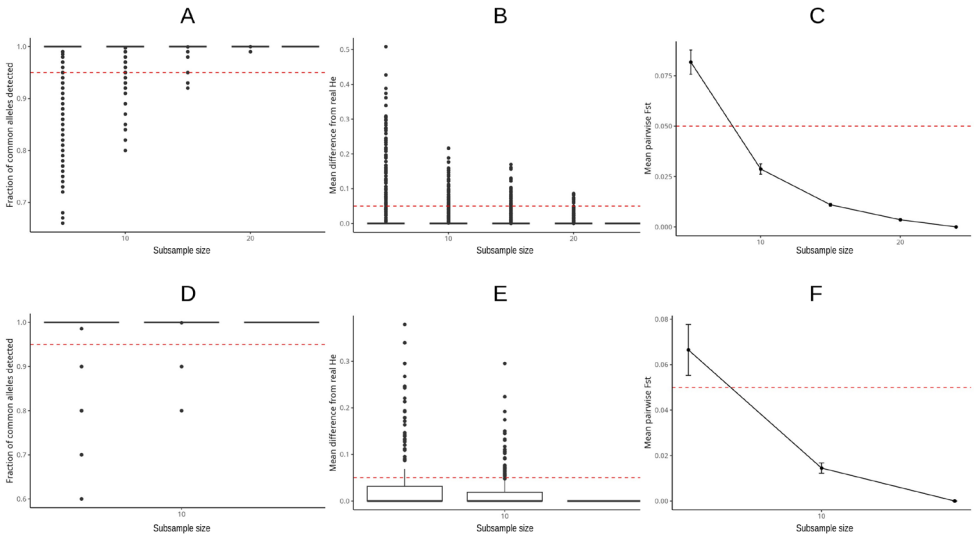

FIGURE S5 | Sample size sufficiency analysis for *Baryancistrus longipinnis* (N = 24; panels **A–C**) and *B. niveatus* (N = 15; panels D–F), based on the SaSii package (Scaketti et al., 2025). Each panel shows the relationship between the number of individuals sampled and: (**A**, **D**) the fraction of common alleles detected, (**B**, **E**) the mean difference in expected heterozygosity (He), (**C**, **F**) the mean pairwise F_{ST} values. The horizontal red dashed line at 0.05 represents the empirical threshold used to define convergence of genetic estimates. This value does not correspond to a statistical *p*-value, but indicates the point at which subsampled estimates deviate less than 5% from the full dataset, suggesting sufficient sampling effort.

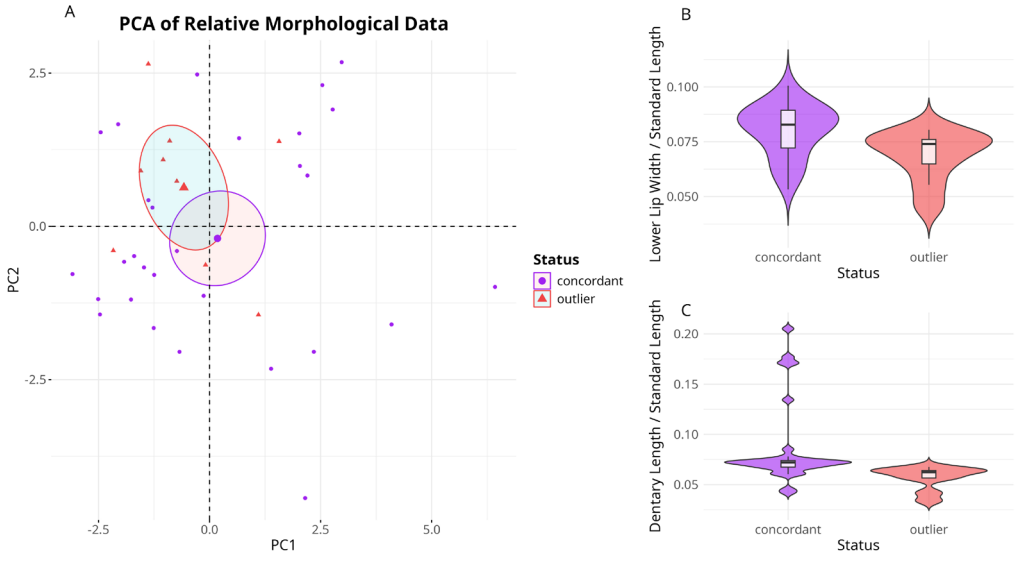

FIGURE S6 | Principal component analysis (PCA) of size-standardized morphological variables and distribution of key morphological ratios by genomic status in Baryancistrus species. (**A**) PCA based on 14 size-standardized variables (ratios of raw measurements to standard length) for individuals classified as concordant (purple circles) or outlier (red triangles) based on genomic data. Ellipses indicate 95% confidence intervals. Although PCA did not reveal a clear morphological separation between groups, two traits exhibited significant differences: (**B**) relative lower lip width and (**C**) relative dentary length (Wilcoxon rank-sum test, p = 0.0092 and p < 0.001, respectively), suggesting potential morphological signals associated with genomic divergence.

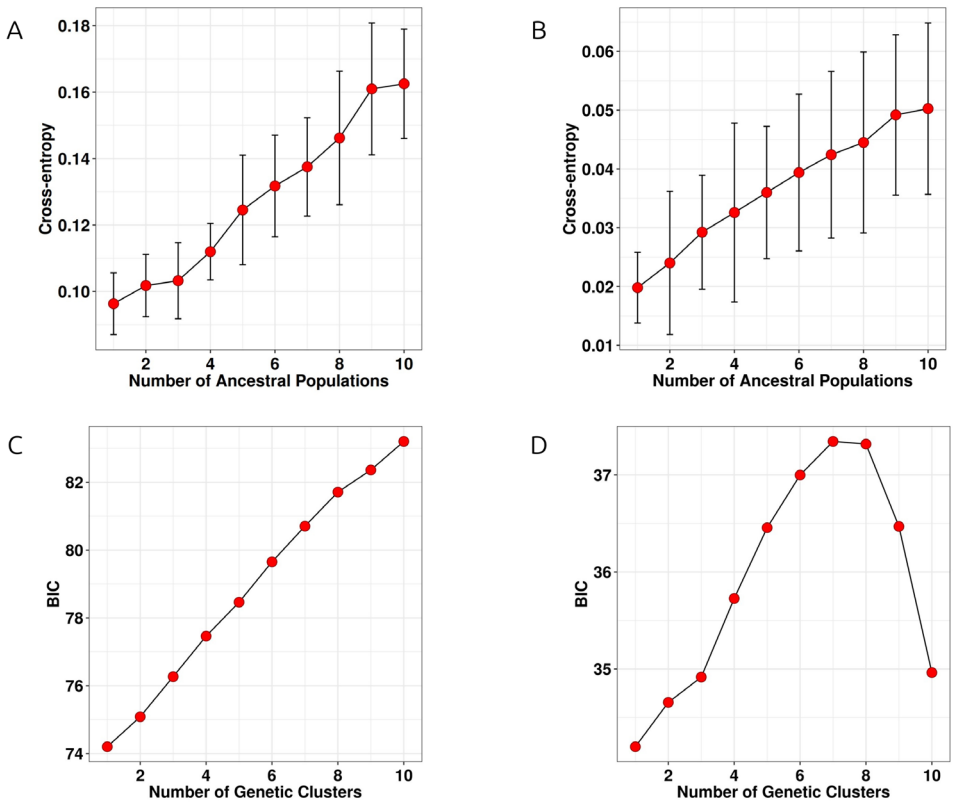

FIGURE S7 | Plots showing the optimal number of genetic clusters (k) for *Baryancistrus longipinnis* (**A**, **C**) and *B. niveatus* (**B**, **D**). The optimal k choice is based on mean ± sd cross-entropy from sparse nonnegative matrix factorization (sNMF) using neutral markers (*upper panels*). Discriminant analysis of principal components (DAPC) where the minimum value of Bayesian information criterion (BIC - indicates best-supported number of genetic cluster - *lower panels*).

TABLE S8 | Results of Isolation by Distance (IBD) analyses for *Baryancistrus longipinnis* and *B. niveatus* in Regions 1 and 2, comparing Mixed-Effects Models (MLPE) and Mantel test approaches. Provesti's inter-individual genetic distance was modeled against Euclidean and riverine geographic distances using both MLPE and simple Mantel tests. [a] For MLPE models, R^2_m refers to the marginal R^2, representing the variance explained by the fixed effects only. [b] Statistical significance levels: p <0.05 (*), p <0.01 (**), and p <0.001 (***).

Species	Region	Distance Type	Method	Slope (MLPE *β*)	AICc	R^2_m [a]	Mantel's r	*p-value* [b]
B. longipinnis	Region 1	Euclidean	MLPE	1.84×10^{-7}	-47.82	0.002	-	0.666
			Mantel	—	—	—	0.058	0.523
		Riverine	MLPE	6.45×10^{-7}	-48.77	0.002	—	0.666
			Mantel	—	—	—	0.058	0.493
	Region 2	Euclidean	MLPE	-6.92×10^{-7}	-46.23	0.18	—	0.117
			Mantel	—	—	—	-0.096	0.814
		Riverine	MLPE	6.45×10^{-7}	-49.12	0.002	—	0.666
			Mantel	—	—	—	-0.096	0.835
B. niveatus	Region 2	Euclidean	MLPE	-4.05×10^{-8}	-45.04	0.003	—	0.617
			Mantel	—	—	—	-0.334	0.995
		Riverine	MLPE	-2.56×10^{-8}	-50.88	0.008	—	0.248
			Mantel	—	—	—	-0.065	0.623

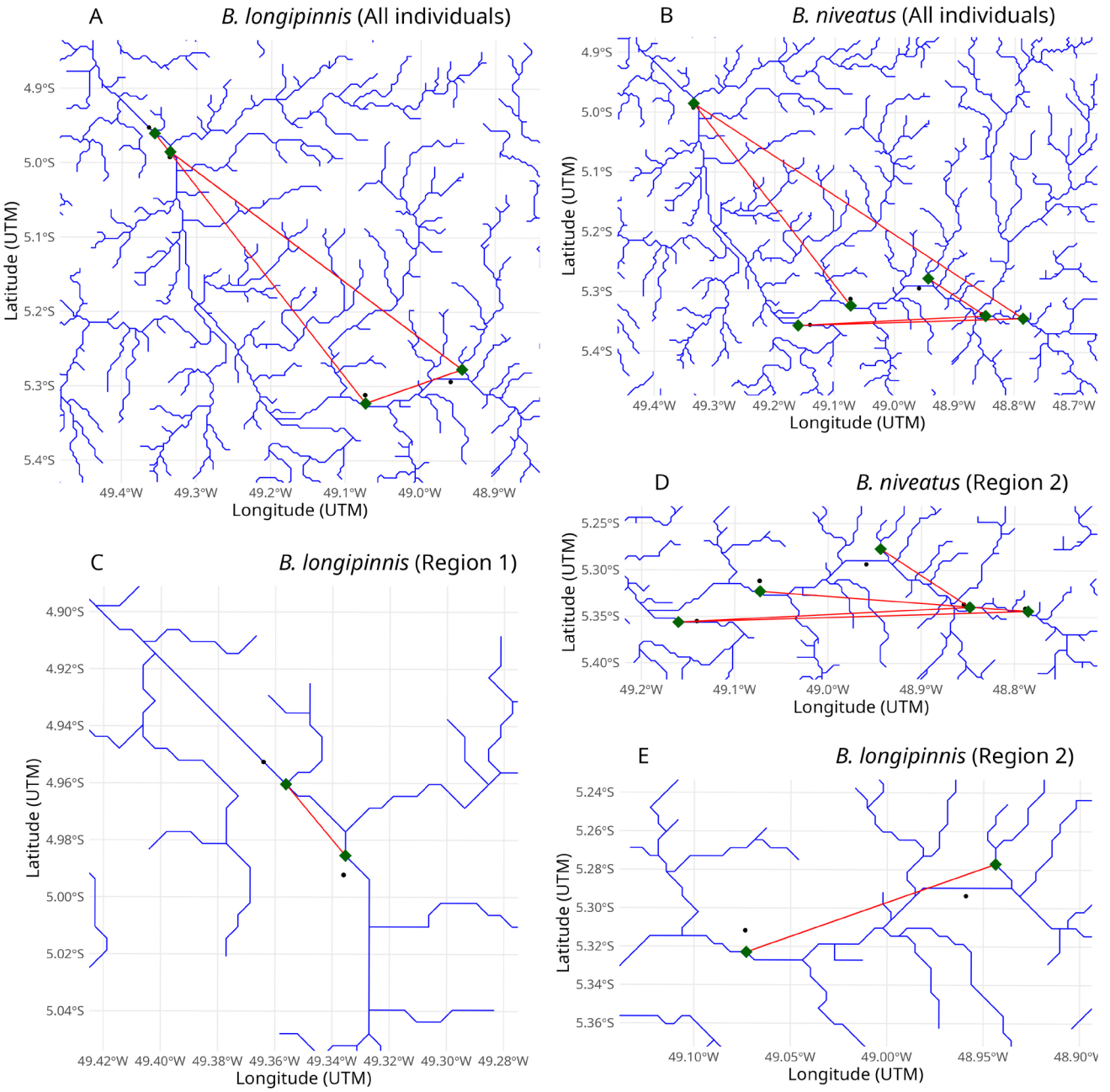

FIGURE S9 | Visualization of genetic discontinuities based on Monmonier's maximum difference algorithm applied to *Baryancistrus longipinnis* (**A**, **C**, **E**) and *B. niveatus* (**B**, **D**). Red lines represent edges in the riverine K-nearest neighbor graph where the highest pairwise genetic distances were observed, incorporating river-based connectivity derived from pairwise topological distances along a validated river network (blue lines). Although no statistically significant or consistent barriers were detected, these edges indicate zones of maximum genetic dissimilarity among sampled locations (green points). Regional analyses (panels **C**–**E**) were performed to investigate finer-scale patterns. Samples from Porto Franco and Peixe were excluded from these analyses due to their large geographic distance from Regions 1 and 2, which could distort the spatial network topology.

TABLE S10 | Coefficients from the best-fitting univariate MLPE models (ΔAICc ≤ 2) explaining genetic distances in *Baryancistrus longipinnis* and *B. niveatus* across regions. Only univariate models are shown, as multivariate models did not meet the ΔAICc ≤ 2 criterion for model support. Parameter estimates for each predictor are presented with standard errors (SE), t-values, p-values, and 95% confidence intervals (CI). Confidence intervals that do not include zero are indicated with an asterisk. [a]Predictor abbreviations and ecological interpretation: **IBD** = Isolation by distance (riverine distance between sampling points); **IBE** = Isolation by environment (Euclidean distances based on PCA of 19 bioclimatic variables); **Slope** = Terrain slope (steepness in degrees); **RipHab** = Riparian habitat amount, calculated as the difference between natural and anthropogenic land cover within a buffer (1,500 m for *B. longipinnis* and 3,000 m for *B. niveatus*) around riparian zones; **NDTI** = Normalized Difference Turbidity Index, used as a proxy for sediment load or water turbidity; and **WST:** Water Surface Temperature. **Significance levels:** Statistical significance of each predictor is indicated by asterisks based on t-tests from the fixed effects of the linear mixed models: p <0.05 (*), p <0.01 (**), and p <0.001 (***).

Species	Region	Predictor [a]	Estimate	SE	*t*-valeu	*p*-valeu	CI
B. longipinnis	Region 1	HabitatRip	-1.7 x 10-3	1.0 x 10-3	-1.49	0.140	[-0.004 / 6 x 10^{-4}]
		Altitude	-1.9 x 10-3	1.7 x 10-3	-1.15	0.251	[-0.005 / 0.001]
		IBE	-1.5 x 10-3	1.6 x 10-3	-0.95	0.342	[-0.005 / 0.002]
		IBD	1.2 x 10-3	2.7 x 10-3	0.44	0.660	[-0.004 / 0.007]
		Slope	-1.4 x 10-3	1.2 x 10-3	-1.23	0.220	[-0.004 / 9 x 10^{-4}]
	Region 2	HabitatRip	-4.1 x 10-3	2.3 x 10-3	-1.75	0.091	[-0.009 / 7 x 10^{-4}]
		IBE	-4.0 x 10-3	2.4 x 10-3	-1.65	0.109	[-0.009 / 0.001]
		IBD	-3.8 x 10-3	2.3 x 10-3	-1.61	0.117	[-0.009 / 0.001]
		Altitude	1.2 x 10-3	3.4 x 10-3	0.34	0.736	[-0.006 / 0.008]
		WST	3.0 x 10-3	3.1 x 10-3	0.08	0.935	[-0.006 / 0.007]
B. niveatus	Region 2	NDTI	7 x 10-4	4 x 10-4	1.86	0.068	[-1 x 10^{-4} / 0.002]
		Slope	-6 x 10-4	3 x 10-4	-1.91	0.062	[-0.001 / 0.000]
		HabitatRip	-1 x 10-3	5 x 10-4	-1.91	0.062	[-0.002 / 0.000]

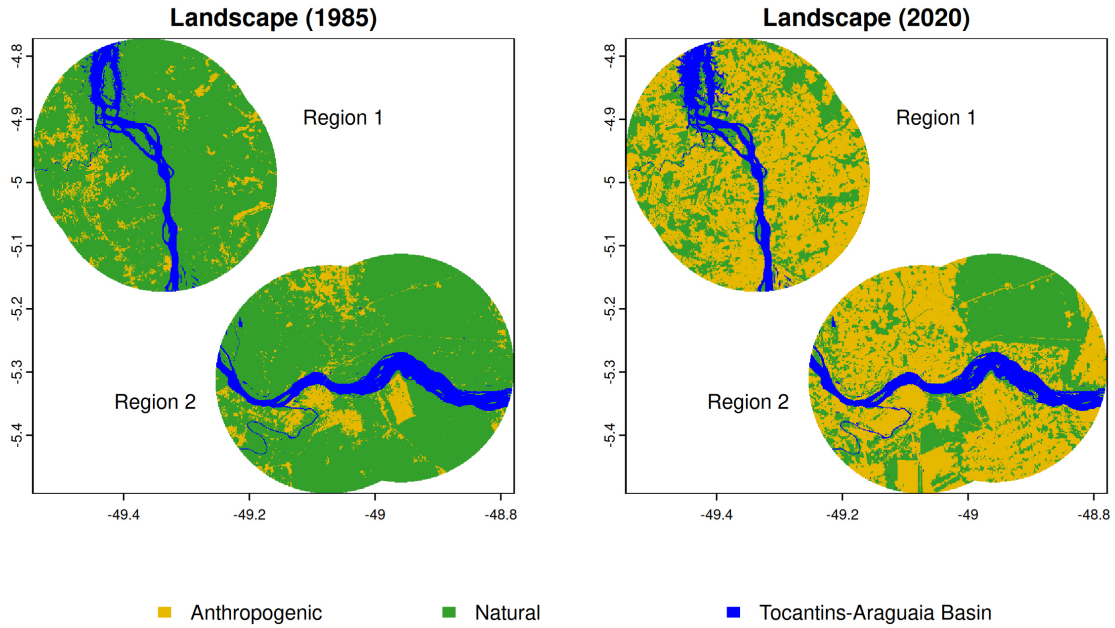

FIGURE S11 | Comparison of landscape cover in the Tocantins-Araguaia basin study area in 1985 and 2020 (cropped to the *Baryancistrus* longipinnis regions of collection, using a 20 km buffer). The panels display the distribution of natural and anthropogenic areas, including rivers, in Regions 1 and 2. Yellow areas represent anthropogenic cover, green areas indicate natural cover, and blue areas correspond to the water bodies of the Tocantins-Araguaia basin.

TABLE S12 | MLPE model selection results for *Baryancistrus longipinnis* and *B. niveatus* in Regions 1 and 2, evaluating the influence of historical habitat amount (1985), current habitat amount (2020), and habitat loss as predictors of genetic parameters (expected heterozygosity – *He* and inbreeding coefficient – *F*). Models were compared across different regions and spatial buffer strategies (variable-optimized, fixed micro, and fixed macro). [a] Likelihood Ratio Tests (LRTs) were conducted to evaluate whether each predictor significantly improved the model's log-likelihood. Statistical significance is indicated as follows: p <0.05 (*), p <0.01 (**), and p <0.001 (***). The best-supported models, based on AICc, are highlighted in bold.

Species/ Region	Response Variable	Buffer Strategy	Landscape Variable [a]	Buffer (km)	LogLik	AICc	ΔAICc	AICc Weight
B. longipinnis / Region 1	*He*	1. Variable-Optimized	**Habitat amount in 1985***	**20**	**31.91**	**-51.93**	**0.00**	**0.36**
			None (null model)	**-**	**29.48**	**-50.78**	**1.05**	**0.22**
			Habitat amount in 2020	**20**	**30.99**	**-49.98**	**1.84**	**0.14**
			Habitat amount in 2020	3	30.91	-49.81	2.02	0.13
			Habitat loss	20	30.66	-49.31	2.51	0.10
			Habitat loss	3	29.69	-47.39	4.44	0.04
		2. Fixed (Micro: 3 km)	**None (null model)**	**-**	**29.48**	**-50.78**	**0.00**	**0.49**
			Habitat amount in 2020	**3**	**30.91**	**-49.81**	**0.97**	**0.30**
			Habitat amount in 1985	3	30.05	-48.10	2.68	0.13
			Habitat loss	3	29.69	-47.39	3.39	0.09
		3. Fixed (Macro: 20 km)	**Habitat amount in 1985***	**20**	**31.91**	**-51.83**	**0.00**	**0.47**
			None (null model)	**-**	**29.48**	**-50.78**	**1.05**	**0.28**
			Habitat amount in 2020	**20**	**30.99**	**-49.98**	**1.84**	**0.19**
			Habitat loss	20	29.77	-47.55	4.28	0.06

TABLE S12 | (Continued)

Species/ Region	Response Variable	Buffer Strategy	Landscape Variable [a]	Buffer (km)	LogLik	AICc	ΔAICc	AICc Weight
B. longipinnis / Region 1	*F*	1. Variable-Optimized	**Habitat amount in 1985****	**20**	**8.85**	**-5.70**	**0.00**	**0.45**
			Habitat amount in 2020**	**20**	**7.93**	**-3.86**	**1.84**	**0.18**
			Habitat amount in 2020	3	7.84	-3.69	2.02	0.16
			Habitat loss	20	7.60	-3.19	2.51	0.13
			Habitat loss	3	6.63	-1.27	4.44	0.05
			None (null model)	-	4.24	-0.30	5.41	0.03
		2. Fixed (Micro: 3 km)	**Habitat amount in 2020****	**3**	**7.84**	**-3.69**	**0.00**	**0.52**
			Habitat amount in 1985*	**3**	**6.99**	**-1.98**	**1.71**	**0.22**
			Habitat loss	3	6.63	-1.27	2.42	0.16
			None (null model)	-	4.24	-0.30	3.39	0.10
		3. Fixed (Macro: 20 km)	**Habitat amount in 1985****	**20**	**8.85**	**-5.70**	**0.00**	**0.57**
			Habitat amount in 2020**	**20**	**7.93**	**-3.86**	**1.84**	**0.23**
			Habitat loss	20	7.60	-3.19	2.51	0.16
			None (null model)	-	4.24	-0.30	5.41	0.04
B. longipinnis / Region 2	*He*	1. Variable-Optimized	**None (null model)**	**-**	**12.53**	**-14.26**	**0.00**	**0.50**
			Habitat amount in 1985*	**10**	**15.61**	**-13.22**	**1.04**	**0.30**
			Habitat amount in 2020	2	14.69	-11.37	2.89	0.12
			Habitat loss	20	13.93	-9.87	4.40	0.06
			Habitat loss	5	13.13	-8.27	6.00	0.03
		2. Fixed (Micro: 2 km)	**None (null model)**	**2**	**12.53**	**-14.26**	**0.00**	**0.76**
			Habitat amount in 2020	2	14.69	-11.37	2.89	0.18
			Habitat amount in 1985	2	12.89	-7.79	6.48	0.03
			Habitat loss	2	12.74	-7.48	6.78	0.03
		3. Fixed (Macro: 20 km)	**None (null model)**	**20**	**12.53**	**-14.26**	**0.00**	**0.66**
			Habitat amount in 1985	20	15.03	-12.03	2.23	0.22
			Habitat loss	20	13.93	-9.87	4.4	0.07
			Habitat amount in 2020	20	13.64	-9.28	4.98	0.05
	F	1. Variable-Optimized	**None (null model)**	**-**	**4.02**	**2.75**	**0.00**	**0.56**
			Habitat amount in 1985*	**10**	**6.87**	**4.27**	**1.52**	**0.26**
			Habitat amount in 2020	2	5.94	6.12	3.37	0.10
			Habitat loss	20	5.19	7.62	4.47	0.05
			Habitat loss	5	4.39	9.22	6.47	0.02
		2. Fixed (Micro: 2 km)	**None (null model)**	**-**	**4.02**	**2.75**	**0.00**	**0.80**
			Habitat amount in 2020	2	5.94	6.12	3.37	0.15
			Habitat amount in 1985	2	4.15	9.70	6.95	0.02
			Habitat loss	2	4.00	10.01	7.26	0.02
		3. Fixed (Macro: 20 km)	**None (null model)**	**-**	**4.02**	**2.75**	**0.00**	**0.71**
			Habitat amount in 1985	20	6.27	5.46	2.71	0.18
			Habitat loss	20	5.19	7.62	4.87	0.06
			Habitat amount in 2020	20	4.90	8.20	5.45	0.05

TABLE S12 | (Continued)

Species/ Region	Response Variable	Buffer Strategy	Landscape Variable [a]	Buffer (km)	LogLik	AICc	ΔAICc	AICc Weight
B. niveatus / Region 2	*He*	1. Variable-Optimized	**None (null model)**	**-**	**22.18**	**-34.94**	**0.00**	**0.70**
			Habitat amount in 2020	1	23.23	-31.79	3.15	0.14
			Habitat amount in 1985	20	22.65	-30.64	4.30	0.08
			Habitat loss	4	21.84	-29.01	5.93	0.04
			Habitat amount in 1985	2	21.82	-28.98	5.96	0.04
		2. Fixed (Micro: 1 km)	**None (null model)**	**-**	**22.18**	**-34.94**	**0.00**	**0.79**
			Habitat amount in 2020	1	23.23	-31.79	3.15	0.16
			Habitat amount in 1985	1	21.19	-27.72	7.22	0.02
			Habitat loss	1	21.19	-27.71	7.23	0.02
		3. Fixed (Macro: 10 km)	**None (null model)**	**-**	**22.18**	**-34.94**	**0.00**	**0.91**
			Habitat amount in 2020	10	21.84	-29.01	5.93	0.05
			Habitat amount in 1985	10	21.22	-27.79	7.18	0.03
			Habitat loss	10	20.98	-27.3	7.64	0.02
	F	1. Variable-Optimized	**Habitat amount in 2020****	**2**	**8.29**	**-1.92**	**0.00**	**0.55**
			Habitat amount in 1985*	**20**	**7.38**	**-0.10**	**1.82**	**0.22**
			None (null model)	-	4.31	0.81	2.73	0.14
			Habitat loss	20	5.89	2.89	4.81	0.05
			Habitat loss	5	5.58	3.51	5.43	0.04
		2. Fixed (Micro: 2 km)	**Habitat amount in 2020****	**2**	**8.29**	**-1.92**	**0.00**	**0.73**
			None (null model)	-	4.31	0.81	2.73	0.19
			Habitat amount in 1985	2	5.55	3.58	5.49	0.05
			Habitat loss	2	5.16	4.35	6.27	0.03
		3. Fixed (Macro: 20 km)	**Habitat amount in 1985***	**20**	**7.38**	**-0.10**	**0.00**	**0.48**
			None (null model)	**-**	**4.31**	**0.81**	**0.91**	**0.31**
			Habitat loss	20	5.89	2.89	2.99	0.11
			Habitat amount in 2020	20	5.85	2.97	3.07	0.10

S13 | Additional information on Results.

(a) Sampling sufficiency for genetic metrics

The SaSii analysis indicated that for *Baryancistrus longipinnis* (N = 24), approximately 8 individuals were sufficient to detect at least 95% of the common alleles (MAF ≥0.05), with the detection curve reaching full saturation at 10 individuals. Mean differences in expected heterozygosity (*He*) compared to the full dataset fell below 0.05 with around 15 individuals, and pairwise F_{ST} values also stabilized near 0.05 from this sample size, indicating reliable estimates of genetic diversity and differentiation beyond this threshold.

For *B. niveatus* (N = 15), 95% of the common alleles were similarly recovered with 8 individuals, and saturation was reached at 10 individuals. The mean difference in *He* dropped below 0.05 with approximately 8 individuals, and F_{ST} values stabilized between 8 and 10 individuals. These results suggest that even smaller sample sizes are sufficient to capture consistent patterns of genetic structure in this species.

It is important to note that the threshold value of 0.05 used here does not represent a statistical significance level (*p-value*), but rather an empirical benchmark indicating

when genetic metrics derived from subsampled data converge closely to those obtained from the full dataset. This approach aligns with findings from Scaketti *et al.* (2025), where studies on diverse plant and animal species using SNP data consistently showed that, despite varying initial sample sizes (ranging from 12 to 53 samples), the average resampling indicated robust genetic diversity and differentiation estimates could be achieved with values as low as 10 to 20 samples. Together, these findings indicate that the sampling effort was adequate to provide robust estimates of genetic diversity and differentiation, consistent with empirical expectations from other species evaluated using the SaSii framework. All results are summarized in Fig. **S5**.

(b) Morphological evidence supporting genomic classification

To assess whether morphometric variation reflects genomic divergence in *Baryancistrus* species, we performed a principal component analysis (PCA) based on 14 size-standardized morphological variables. Each morphological variable was standardized relative to its *standard length* to control for body size, applying a standard allometric correction approach. The PCA was followed by *Wilcoxon rank-sum tests* comparing individuals classified as *concordant* or *outlier* according to genomic analyses. Among the 14 relative measurements, two showed statistically significant differences between genomic groups: (1) Relative lower lip width (*lower lip width / standard length*): $p = 0.0092$; and (2) Relative dentary length (*dentary length / standard length*): $p < 0.001$. These results are depicted in Fig. **S6**, panels (**B**) and (**C**), and indicate the presence of subtle but significant morphological signals potentially associated with genomic divergence, despite the limited visual separation observed in multivariate space.

Additionally, we compared standard length between genomic *concordant* and *outlier* individuals to evaluate whether size differences might contribute to morphological inconsistencies. Outliers exhibited a numerically higher, albeit not statistically significant, average standard length (mean = 127 mm, median = 131 mm) compared to concordant individuals (mean = 115 mm, median = 97.3 mm; *Wilcoxon rank-sum test*, $p = 0.686$). Although this numerical trend appears contrary to the hypothesis that misclassifications stem from immature individuals, the lack of statistical significance suggests that morphological misidentification may not be strictly associated with smaller or larger size in a statistically discernible manner, but rather with other factors, such as intraspecific variation or ontogenetic trajectory mismatches. A formal ontogenetic analysis will be necessary to resolve this.

REFERENCE

- **Scaketti M, Sujii PS, Alves-Pereira A, Schwarcz KD, Francisconi AF, Moro MS *et al.*** Sample size impact (SaSii): an R script for estimating optimal sample sizes in population genetics and population genomics studies. PLoS ONE. 2025; 20(2):e0316634. https://doi.org/10.1371/journal.pone.0316634

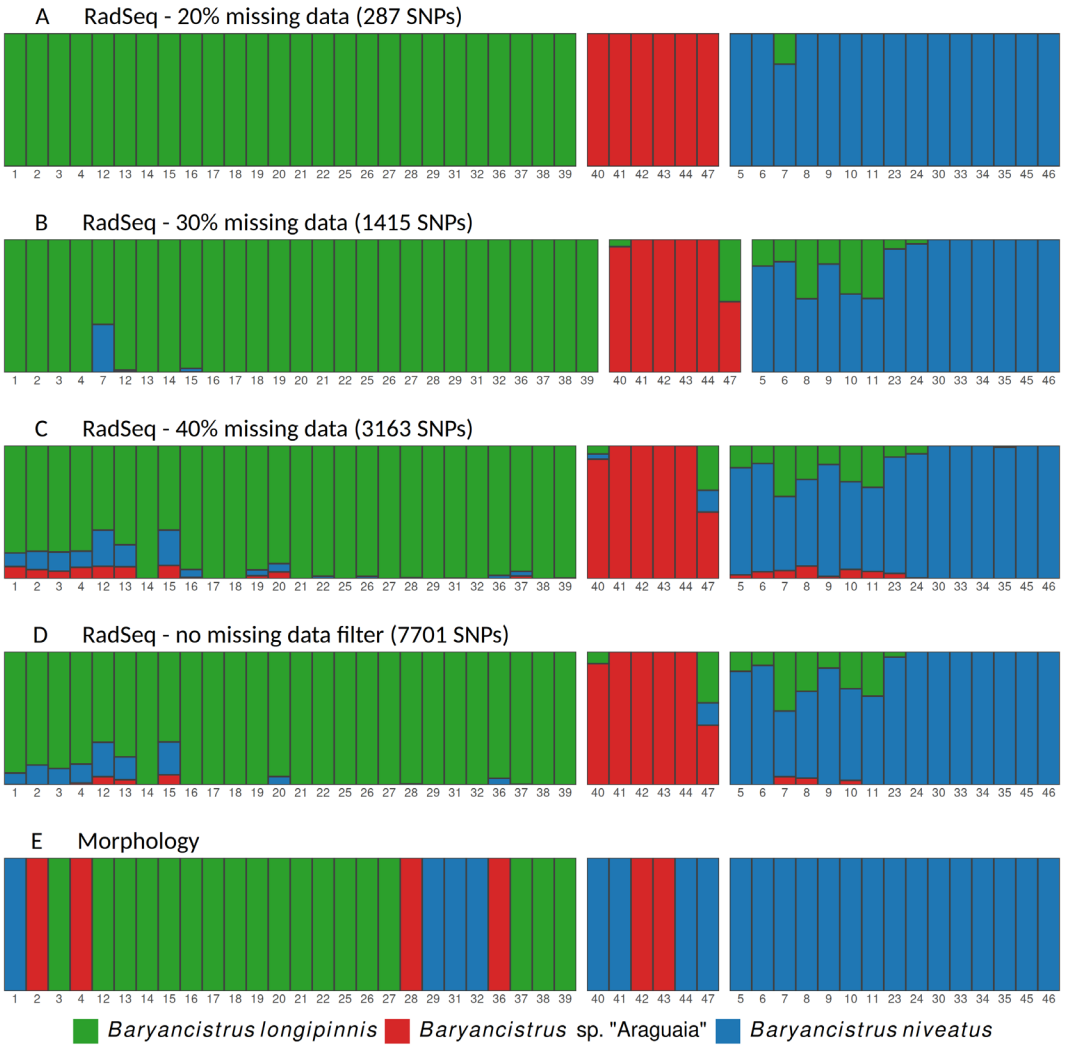

FIGURE S14 | Population genomic structure of the three *Baryancistrus* species. Each vertical bar represents an individual, and each color corresponds to a different species assignment: *Baryancistrus* sp. "Araguaia" (red), *B. longipinnis* (green), and *B. niveatus* (blue). Ancestry coefficients were inferred using the sNMF function (K = 1 to 10), based on datasets with progressively relaxed missing data thresholds: (A) 20% missing data (287 SNPs), (B) 30% (1415 SNPs), (C) 40% (3163 SNPs), and (D) no missing data filter (7701 SNPs). Only filters for sequencing depth (3–100) and missing data were applied in this preliminary step. Panel (E) shows the initial morphological species assignments based on Oliveira (2018). Comparisons across panels reveal cases of mismatch between morphological and genetic assignments, which were further investigated and supported by additional genomic and morphological analyses (see S14, Section b: Morphological evidence supporting genomic classification).

Neotropical Ichthyology

OPEN ACCESS

HOW TO CITE THIS ARTICLE

- **Leonardo MG, Lanes ECM, Araújo FAA, Cruz VP, Pavan SE, Monteiro ILP, Quaresma TF, Oliveira C, Akama A**. Where are we going to live? Population genomics emits warning signals about endangered Amazonian rheophilic loricariids under habitat loss. Neotrop Ichthyol. 2026; 24(2):e250148. https://doi.org/10.1590/1982-0224-2025-0148

SBI Official Journal of the Sociedade Brasileira de Ictiologia