

# DNA barcode and minibarcode identification of freshwater fishes from Cerrado headwater streams in Central Brazil

Justin C. Bagley<sup>1,2,3</sup>  | Pedro De Podestà Uchôa de Aquino<sup>1</sup> |  
María Florencia Breitman<sup>4</sup> | Francisco Langeani<sup>2</sup> | Guarino R. Colli<sup>1</sup>

<sup>1</sup>Departamento de Zoologia, Instituto de Ciências Biológicas, Universidade de Brasília, Brasília, DF, Brazil

<sup>2</sup>Departamento de Zoologia e Botânica, UNESP, Universidade Estadual Paulista, IBILCE, São José do Rio Preto, SP, Brazil

<sup>3</sup>Department of Biology, Virginia Commonwealth University, Richmond, VA, USA

<sup>4</sup>Departamento de Ecologia, Instituto de Ciências Biológicas, Universidade de Brasília, Brasília, DF, Brazil

## Correspondence

Justin C. Bagley, Department of Biology, University of Missouri-St. Louis, One University Boulevard, 223 Research Building, St. Louis, MO 63121-4499, USA.  
Email: bagleyj@umsl.edu

## Funding information

Research was funded by a Ciência Sem Fronteiras (Science Without Borders) postdoctoral fellowship from the Brazilian Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq, Processo 314724/2014-1 and 306566/2014) to J.C.B. and F.L., by CNPq and the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) to F.L., by Coordenação de Apoio à Formação de Pessoal de Nível Superior (CAPES), CNPq, Fundação de Apoio à Pesquisa do Distrito Federal (FAPDF) to G.R.C. and by the USAID PEER programme (under cooperative agreement AID-OAA-A-11-00012).

## Abstract

The extraordinary species diversity of the Neotropical freshwater fish fauna is world renown. Yet, despite rich species diversity, taxonomic and genetic resources for its Cerrado ichthyofauna remain poorly developed. We provide a reference library of 149 DNA barcodes for 39 species/lineages of Cerrado headwater stream fishes from the Brazilian Distrito Federal and nearby areas and test the utility of distance-based criteria, tree-based criteria and minibarcodes for specimen identification. Mean Kimura 2-parameter genetic distances within species to orders ranged 1.8–12.1%. However, mean intraspecific v. congeneric-interspecific distances (0.9–1.3%) overlapped extensively and distance-based barcoding failed to achieve correct identifications due to c. 4–12.1% error rates and 19.5% ambiguous identifications related to the presence of singletons. Overlap was reduced and best-match success rates improved drastically to 83.5% when *Characidium* barcodes representing potential misidentifications or undescribed species were removed. Tree-based monophyly criteria generally performed similarly to distance methods, correctly differentiating up to c. 85% of species/lineages despite neighbour-joining and Bayesian tree errors (random lineage-branching events, long-branch attraction). Five clusters (*Ancistrus aguaboensis*, *Characidium* spp., *Eigenmannia trilineata*, *Hasemania hansenii* and *Hypostomus* sp. 2) exhibited deep intraspecific divergences or para-/polyphyly and multiple Barcode Index Number assignments indicative of putative candidate species needing taxonomic re-examination. Sliding-window analyses also indicated that a 200 bp minibarcode region performed just as well at specimen identification as the entire barcode gene. Future DNA barcoding studies of Distrito Federal–Cerrado freshwater fishes will benefit from increased sampling coverage, as well as consideration of minibarcode targets for degraded samples and next-generation sequencing.

## KEYWORDS

Brazil, Cerrado, cytochrome c oxidase subunit 1, DNA barcoding, freshwater fishes, minibarcode

## 1 | INTRODUCTION

The greatest diversity of freshwater fishes worldwide, including approximately c. 4600–13,000 species, or c. 35% of all freshwater

fishes, is concentrated in the rivers and lakes of the South American Neotropics (Albert *et al.*, 2011; Albert & Reis, 2011a, 2011b). Taking into account this exceptional diversity, as well as increasing human threats to biomes and native biota, the Neotropical region has become

a priority area for biodiversity research and conservation efforts (Klink & Machado, 2005; Myers, Mittermeier, Mittermeier, da Fonseca, & Kent, 2000; Prance, 2006) and for freshwater fishes in particular (Abell *et al.*, 2008).

Within the Neotropics, freshwater fish communities of the Cerrado biodiversity hotspot in central Brazil (Hoffmann & Jackson, 2000; Myers *et al.*, 2000) are recognised as highly diversified (c. 1200 species, c. 40% of Brazilian fish species), endemic and threatened by human activities (Abell *et al.*, 2008; Albert *et al.*, 2011; Klink & Machado, 2005; Langeani *et al.*, 2009). In particular, fishes of the Araguaia-Tocantins, upper Paraná and São Francisco river basins are respectively 44%, 48% and 59% endemic and each of these basins contains 181–346 species (Albert *et al.*, 2011; Langeani *et al.*, 2009). However, Cerrado fishes from these and proximal drainages (e.g., Xingu arm of the Amazon Basin) face many environmental challenges, including major anthropogenic effects. For example, >55% of the Cerrado's natural landscapes have been converted for agriculture, pastureland, or other human use by 2010, resulting in a rate of Cerrado deforestation greater than that of the Amazon forest (Beuchle *et al.*, 2015; Klink & Machado, 2005). Surprisingly, only c. 9% of the Cerrado biome is under any legal protection, about 7% when considering areas covered by native vegetation, compared with c. 28% of Brazilian Amazonia (Françoso *et al.*, 2015; Pacheco *et al.*, 2018). It is also worrying that climate and land-use changes have promoted increases in temperature and rainfall extremes and, by 2100, projections point to an increase in dry spells and warm days and nights (Magrin *et al.*, 2014). These changes, in turn, may promote increased runoff and soil erosion and major morphological and hydrological changes in large river systems of the Cerrado (Anache *et al.*, 2018; Arnell & Gosling, 2013; Coe *et al.*, 2011). Upper portions of major Cerrado water courses, for example, are already completely dry during the pronounced regional dry season in areas such as the Tocantins River; it is expected, therefore, that major droughts will only aggravate the problem and reduce populations of freshwater fish and their connectivity even further in coming decades. Freshwater habitats of the Cerrado also suffer increasing threats from dam construction and silting and destruction of riparian areas at the expense of agriculture and livestock, combined with a paucity of protected areas, all of which sadly have become common to virtually all continental waters of Brazil (Agostinho *et al.*, 2005; Klink & Machado, 2005; Nogueira *et al.*, 2010).

Notwithstanding their great diversity and need of conservation protection, the basic biology and taxonomy of Cerrado freshwater fishes, as well as the patterns and processes responsible for generating and maintaining their outstanding diversity, remain little known (Buckup *et al.*, 2007; Langeani *et al.*, 2007, 2009). Recent studies elucidating new details on the ecology and distribution of fish species of river basins in eastern Brazil and the Cerrado are encouraging (Aquino *et al.*, 2009; Aquino & Colli, 2017; Langeani *et al.*, 2009). However, there are few studies in general and many have emphasised inventorying taxonomic groups or fish communities from restricted geographical areas (Aquino *et al.*, 2009; Casatti & Castro, 1998). Accurate identification of fish specimens remains challenging due to overlap in species morphological characters, wide variation in their

geographical distributions (Junk *et al.*, 2007; Reis *et al.*, 2003), the need for systematic revisions to elucidate and stabilise taxonomy (Langeani *et al.*, 2009), as well as the paucity of genetic studies and resources (but see Pereira *et al.*, 2013; Bagley *et al.*, in revision). Moreover, the sheer number and morphological diversity of native lineages of tetras (Characiformes) and catfishes (Siluriformes), the two most diverse orders of Cerrado fishes (Aquino & Colli, 2017), is overwhelming, requiring tedious study as well as consultation with taxonomic experts working on different families and clades within these orders.

The advent of molecular methods for DNA sequencing has spawned new approaches for the effective identification and discovery of species based on variation at mitochondrial DNA cytochrome c oxidase subunit 1 (*cox1*) gene, known as DNA barcoding (Hebert *et al.*, 2003). Through development of a *cox1* sequence library tied to expert-identified samples, DNA barcoding permits subsequent identification of unknown material to species level, usually with ≥90% success rates (reviewed in Collins *et al.*, 2012), including studies of freshwater fishes (Díaz *et al.*, 2016; Hubert *et al.*, 2008; Ward *et al.*, 2005). Results from studies comparing *cox1* variation across samples from multiple collection sites throughout species ranges have also proven DNA barcoding to be adept at identifying morphologically cryptic species, including undescribed biodiversity (Bickford *et al.*, 2006; Hebert *et al.*, 2004; Hubert *et al.*, 2008). For this reason, DNA barcoding has assumed an essential position in integrative taxonomy (Padial *et al.*, 2010; Puillandre *et al.*, 2012). Thus, the DNA barcoding approach presents a key, but currently broadly unapplied tool for the development of genetic resources for taxonomic identification of nominal species and discovery of cryptic species (species limits) of Cerrado freshwater fishes.

This study presents a DNA barcoding library and analysis of freshwater fishes from headwater streams from Cerrado upland areas within and adjacent to the Brazilian Distrito Federal (capital), an area that is characteristic of higher elevation Cerrado habitats facing drought seasonality and anthropogenic effects. A variety of studies have elucidated finer-scale environmental determinants of fish distributions and community composition within this area of central-western Brazil, providing baseline ecological and taxonomic information (Aquino *et al.*, 2009; Aquino & Colli, 2017; Araújo & Tejerina-Garro, 2007; Benedito-Cecilio *et al.*, 2004; Claro-García & Shibatta, 2013; Couto & Aquino, 2011; Fialho *et al.*, 2007; Langeani *et al.*, 2007; Ribeiro *et al.*, 2008; Viana, 1989). However, these studies have been limited by vagaries of current morphospecies taxonomy, as modern molecular genetic studies of freshwater fish species from the region have been limited. Although two recent DNA barcoding papers have focused on understanding fish diversity and identifications in the lower and upper Paraná River, one of these studies did not sequence any taxa from the Distrito Federal (Díaz *et al.*, 2016), while the second focused mainly on lower reaches of the upper Paraná Basin and included only a handful of samples and species from the Distrito Federal (Pereira *et al.*, 2013). Another study focused on DNA barcoding of São Francisco Basin freshwater fishes, but likewise, included no samples from the Distrito Federal (de Carvalho *et al.*, 2011). Yet another study focused specifically on testing whether one widespread

species, *Piabina argentea* Reinhardt 1867 (also sampled in the present study), contained multiple undescribed forms (Pereira *et al.*, 2011). Herein, we sequence 39 species/lineages of fish from the Distrito Federal and nearby areas to provide a reference DNA barcoding library for this system and test two hypotheses: first, that DNA barcoding techniques can accurately identify local freshwater fish species in the face of taxonomic uncertainty and second, that genetic variation within nominal species will be lower on average than that witnessed between species. We also evaluate the utility of minibarcode regions for DNA barcoding in this system, because archive and frozen material from Neotropical fish collections are often degraded and sequencing shorter fragments can improve chances of DNA amplification in such cases (Meusnier *et al.*, 2008).

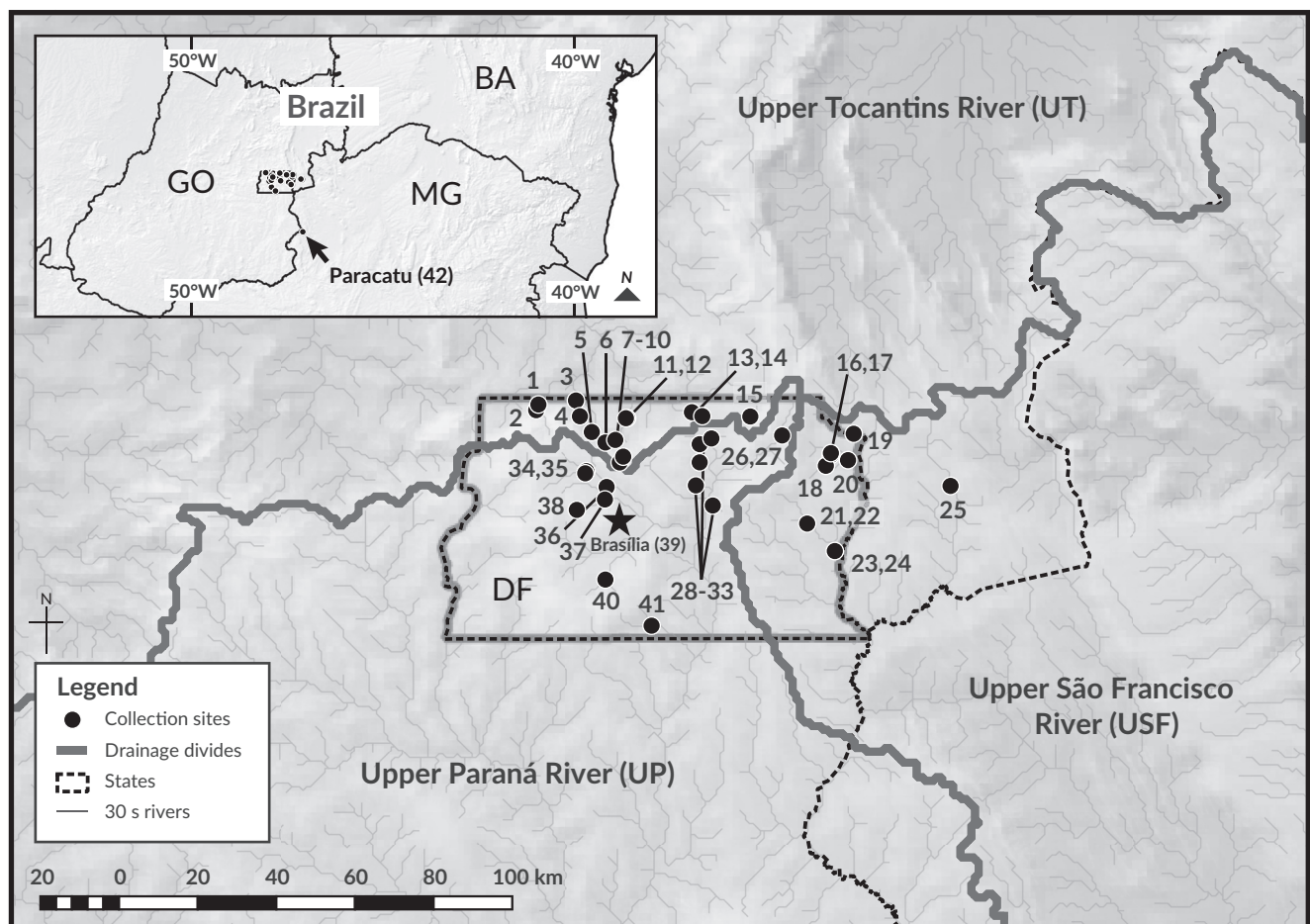
## 2 | MATERIALS AND METHODS

Permission to undertake 2015 field sampling and material transport were granted through SISBIO permits #48145-1 to J.C.B. and #48111-4 to F.L.N. Sampling procedures and permitting for years

2008–2011 are detailed in Aquino *et al.* (2009) and Couto and Aquino (2011) and included SISBIO (Sistema de Autorização e Informação em Biodiversidade) permit #42573-1.

### 2.1 | Taxon sampling and DNA sequencing

Fishes were captured from 42 sites across the Brazilian Distrito Federal and nearby areas (Figure 1 and Table S1) between 2009 and 2015 using  $3 \times 1$  m seines with 2 mm mesh diameter. Fin clips or muscle plugs were taken from anaesthetised fishes in the field, stored in 95–99% ethanol and returned to the laboratory on ice, along with voucher material preserved in 10% formalin. Vouchers for all species are deposited in two ichthyological collections: the Coleção Ictiológica da Universidade de Brasília (CIUnB) in Brasília and the Coleção de Peixes do Departamento de Zoologia e Botânica da Universidade Estadual Paulista, São José do Rio Preto (DZSJRP). Tissue samples were obtained for molecular analyses from a total of 152 individuals representing 39 species/lineages (nominal taxa plus candidate species



**FIGURE 1** Map of freshwater fish sampling localities across the study area, including the Brazilian Distrito Federal (DF) and surrounding areas of the Cerrado biome. (●), Sampling sites; (★), site 41 ReBIO do Gama within the Brasília metropolitan area. The three main river basins: UP, upper Paraná River; USF, upper São Francisco River; UT, upper Tocantins River (UT). Inset map shows the location of site 42 Paracatu, as well as the position of the DF relative to other states in central-western Brazil: BA, Bahia; GO, Goiás; MG, Minas Gerais

listed in Aquino *et al.* (2009), Langeani *et al.* (2007) and Aquino and Colli (2017)).

Initial taxonomic determinations were made to genus or species level in the field and followed up as needed by detailed identifications under a confocal microscope, or by shipping samples to taxonomic experts for identification. Identifications are compliant with Fish-BOL ([www.fishbol.org](http://www.fishbol.org)) collaborators' reliability levels 1 through 3 (Steinke & Hanner, 2011; Ward, 2012). We follow the family and subfamily classification scheme in Reis *et al.* (2003).

We extracted whole genomic DNA from tissue samples using Invitrogen PureLink Genomic Purification kits (Invitrogen; [www.invitrogen.com](http://www.invitrogen.com)) and DNA was eluted from spin columns using two consecutive spins into 75–100 µl of elution buffer. Success of DNA extractions was confirmed using 1% agarose gels and final sample DNA concentration and purity were quantified using a NanoDrop 2000c Spectrophotometer (Thermo Scientific; [www.thermoscientific.com](http://www.thermoscientific.com)). We amplified the *cox1* gene using fish-specific primers *Fish-F1* and *Fish-R1* described in Ward *et al.* (2005) and in rare cases we used the universal barcode primers *Lco-1490* and *Hco-2198* of Hebert *et al.* (2003). Final concentrations of PCR components per 25 µl reaction included c. 25 ng template DNA, 0.25 µM of each primer, 0.625 units of Taq DNA polymerase, 0.1 mM per deoxynucleotide triphosphate (dNTP), 2.5 µl of 10X reaction buffer and 2.5 mM MgCl<sub>2</sub>. Our *Fish-F1*–*Fish-R1* PCR protocol involved initial denaturation at 95°C for 2 min, followed by 35 cycles of 95°C for 30 s, 54°C for 30 s and 72°C for 1 min 30 s, with a final extension at 72°C for 7 min. The *Lco-1490*–*Hco-2198* PCR protocol involved annealing temperatures of 45°C and 50°C and is given in Hebert *et al.* (2003). Forward and reverse sequences were obtained through cycle sequencing using BigDye Terminator 3.1 cycle sequencing kit chemistry using 1/16th reaction size and the manufacturer's instructions (Applied Biosystems; [www.appliedbiosystems.com](http://www.appliedbiosystems.com)). Sequenced products were purified using cocktails of exonuclease I combined with shrimp alkaline phosphatase (ExoSAP-IT; USB Corp., Thermo Fisher Scientific) and then run on Applied Biosystems 3500 and 3730xl automated capillary sequencers. Edited sequence data, unedited sequence chromatograms and sample metadata including collections information were uploaded to the Barcode of Life Data (BOLD) database ([www.boldsystems.org](http://www.boldsystems.org)) under our Barcoding Freshwater Fishes of the Distrito Federal (BFFDF) accession. Additionally, aligned DNA sequence data were submitted to GenBank (accession numbers MK464024–MK464172; Table S1). We also provide our DNA sequence alignment in a Mendeley Data accession (<https://doi.org/10.17632/9pr3c3p33g.1>).

## 2.2 | Data analysis

Our mtDNA *cox1* gene sequences contained no gaps; consequently, they were straightforwardly edited and aligned by-eye in GENEIOUS R10 (Biomatters Ltd.; [www.geneious.com](http://www.geneious.com)). All DNA sequences at least 500 bp in length were trimmed to include only the 652 bp gene region prior to analyses. Descriptive statistics for Kimura's (Kimura, 1980; R Core Team, 2017; [www.r-project.org](http://www.r-project.org)) based on mean within-

and between-group K2P distances for species, genera, families and orders estimated in MEGA7 (Kumar *et al.*, 2016).

To assess barcode success at specimen identification using distance methods, we used SPECIESIDENTIFIER 1.8 (Meier *et al.*, 2006) to simulate DNA barcoding identification by querying sequences in the library against one another and quantifying the proportion of correctly *versus* incorrectly identified specimens based on best match (BM) and best close match (BCM) criteria. K2P genetic distances were used as the basis for all query comparisons and for threshold calculations during BCM. Exploratory analysis was used to identify an appropriate pairwise K2P distance threshold for species identification in our sample, based on overlap range of intraspecific and congeneric-interspecific distances and a 5% error margin from the upper end of the range. We then conducted BCM simulations over a series of threshold values ranging from the traditional BOLD value of 1% (Ratnasingham & Hebert, 2007), to the 2% value widely used for freshwater fishes (Ward *et al.*, 2009) and up to 15%, including the intraspecific threshold value estimated above. To understand the effect of singletons on our results, we used the R package spider (Brown *et al.*, 2012) to conduct similar BM and BCM analyses while excluding singletons. We also evaluated the smallest and most extreme intraspecific and congeneric-interspecific pairwise distance relationships among species in SpeciesIdentifier (<https://github.com/gaurav/taxondna/releases/tag/1.8>). We used spider and the BarcodingR package (Zhang *et al.*, 2017) to calculate and plot the DNA barcode gap, or amount of separation between intraspecific and congeneric-interspecific distances (Meier *et al.*, 2006; Meier *et al.*, 2008; Meyer & Paulay, 2005). Finally, we clustered barcodes in SpeciesIdentifier such that each sequence had at least one match below the intraspecific threshold distance and we evaluated the percentage of threshold violations and the contents of each cluster.

We compared our distance-only analyses above to the performance of using tree-based monophyly criteria to assess whether DNA barcodes could distinguish species/lineages in the Distrito Federal fish library. We used PAUP\* 4.0a (build 159; Swofford, 2002) to infer a single best neighbour-joining (NJ) tree showing clustering relationships among samples based on K2P distances and we assessed confidence in individual nodes of the tree using 500 NJ bootstrap pseudo-replicates. We also inferred a phylogeny of the barcodes using Bayesian inference analysis in MrBayes 3.2 (Ronquist *et al.*, 2012). Prior to running MrBayes, we estimated the optimal *cox1* codon partitioning scheme (*versus* concatenation) and DNA substitution model in PartitionFinder 2.0 (Lanfear *et al.*, 2012; Lanfear *et al.*, 2014) while unlinking branch lengths, only considering models available in MrBayes and using the corrected Akaike information criterion (AICc) statistic for model selection. Subsequently, we ran MrBayes for 10 million generations (sampling every 1000th generation and saving branch lengths) while specifying the scheme and models selected in PartitionFinder. After checking the posterior traces of MrBayes parameters for proper run convergence and sampling (e.g., stationarity) in Tracer 1.6 (Rambaut *et al.*, 2013) and confirming convergence of the potential scale reduction factor to a value of 1.0, we summarised the posterior distribution of trees and parameters in a



Bayesian consensus topology in MrBayes while excluding the first 25% of trees as burn-in. We rooted our NJ and Bayesian consensus trees with two barcodes of introduced *Lepomis gibbosus* (L. 1758) (Centrarchidae) sampled herein, which represent distant outgroups (Table S1). We conducted two tree-based identification assessments on each topology. First, we deemed barcodes successful at differentiating a given species/lineage in the library from others when barcodes of that species formed a monophyletic cluster with all conspecific barcodes with  $\geq 70\%$  bootstrap support or  $\geq 0.95$  Bayesian posterior probability. Singletons were counted as monophyletic so long as they did not nest within samples of another species/lineage. Second, we applied Hebert *et al.*'s (2003) criteria, which only require monophyly but deem singletons as ambiguous and thus unidentified samples. Misidentifications were assessed when conspecific barcodes occurred in multiple clusters (Hebert *et al.*, 2003; Meier *et al.*, 2006). We tempered our monophyly results by testing the null hypothesis that patterns of monophyletic clustering of our barcode samples might reflect chance branching, which coalescent theory predicts can occur with greater frequency when (intraspecific) sample sizes are small, as for many taxa in our study. We tested the random-branching null hypothesis by calculating Rosenberg's (2007) probability of monophyly for every node in a rooted NJ tree reconstructed from K2P distances using the rosenberg function available in spider and a significance level of  $\alpha = 0.05$ .

Despite overlap in intraspecific and interspecific distances in our barcode library (see §3), we conducted analyses to identify primary species delimitation hypotheses using the popular barcode index numbers (BIN) approach based on refined single linkage (RESL) analysis (Ratnasingham & Hebert, 2013), which is automated through the operational taxonomic unit (OTU) Pipeline on BOLD. This method assigns sequences to BINs independent of prior taxonomic assignment and has been shown to outperform several other methods in taxonomic performance and computational efficiency (Ratnasingham & Hebert, 2013). The BINs were compared with nominal taxonomy using the BIN discordance report functionality of BOLD. Additionally, we evaluated trees and distance patterns, including summaries output by the Pairwise Explorer and Extreme Pairwise modules of SpeciesIdentifier and classified genetic divergences among conspecific and congeneric comparisons as moderate at  $>2\%$  (typical fish barcode threshold; Ward *et al.*, 2009) and extreme at  $>5.3\%$  (SpeciesIdentifier cutoff; see §3). Species or lineages exhibiting extreme pairwise divergences were compared to BINs and we conducted additional analyses of genetic structure for these groups of samples. First, we estimated statistical parsimony networks among *cox1* haplotypes from these species sequences in TCS 1.2.1 (Clement *et al.*, 2000) while using a 95% connection limit, identifying ancestral haplotypes and using congeneric samples as outgroups. Second, we used Arlequin 3.5.1.2 (Excoffier *et al.*, 2005) to estimate  $F_{ST}$  values for pairwise comparisons of clusters within these species (independent networks), while testing significance of the  $F$ -statistics using 10,000 permutations.

A number of samples available to us in the present study were not sequenced due to low-quality preservation, or poor amplification of DNA extracts possibly due to degradation in storage. As this will likely

be a general feature of Neotropical fish material, we decided to conduct a minibarcode analysis. We used the slideAnalyses function in apider to determine the shortest window, among 50, 100, 150 and 200 bp windows surveyed at codon intervals across the *cox1* gene, that could best differentiate among Distrito Federal freshwater fish species. We defined the best minibarcode as the sliding-window position and length that yielded a proportion of identical clades  $>85\%$  for shallow clusters (no deeper than the median node depth) as well as the lowest proportion of zero pairwise non-conspecific distances. If necessary, we broke ties by choosing the window minimising the mean\_distance (mean distance in the matrix) statistic, which is desirable with small sample sizes. The single best window identified through this procedure was then subjected to BM and BCM analyses in SpeciesIdentifier similar to those described above for the full barcode sequence set. Following Doña *et al.* (2015), we used DNASP 5 (Librado & Rozas, 2009) to evaluate the suitability of regions flanking the best window for development of fish-specific primers to amplify the minibarcode, based on DNA polymorphism of sliding windows across the gene.

### 3 | RESULTS

We obtained *cox1* sequences of 652 bp in length for a sample of 152 individual fish spanning 39 species/lineages (37 native, plus 2 exotic) representing five orders and 13 families (Table S1). After removing sequences from three individuals (*Characidium fasciatum* Reinhardt 1867, JCB 112 (Table S1); *Plabina argentea* Reinhardt 1867, CIUnB 985\_1; *Planaltina myersi* Böhlke 1954, CIUnB 912\_1; Table S1) with greater than c. 30% missing data ( $< 500$  bp of data), the final alignment contained 149 barcodes representing the same number of species. Our sampling covered c. 35% (39/110 species/lineages) of the known stream ichthyofauna of the Distrito Federal including 35–70 species/lineages identified in Aquino and Colli (2017), plus the introduced pumpkinseed sunfish *L. gibbosus* (Centrarchidae) and several candidate species (Aquino *et al.*, 2009; P.P.U.A., J.C.B. and F.L.N., unpubl. Data; Table 1). Given sequences were the same length and contained no insertions or deletions (indels) or stop codons, we could assume that we did not mistakenly amplify nuclear copies of mtDNA, or nuclear mitochondrial DNA segments (NUMT; Funk & Omland, 2003). The shortest sequence was 519 bp in length and the mean sequence length was 645 bp (Figure S1) and base composition was CT-biased as expected for freshwater fish *cox1* (A, 24.5; C, 26.7; G, 18.1; T, 30.7%).

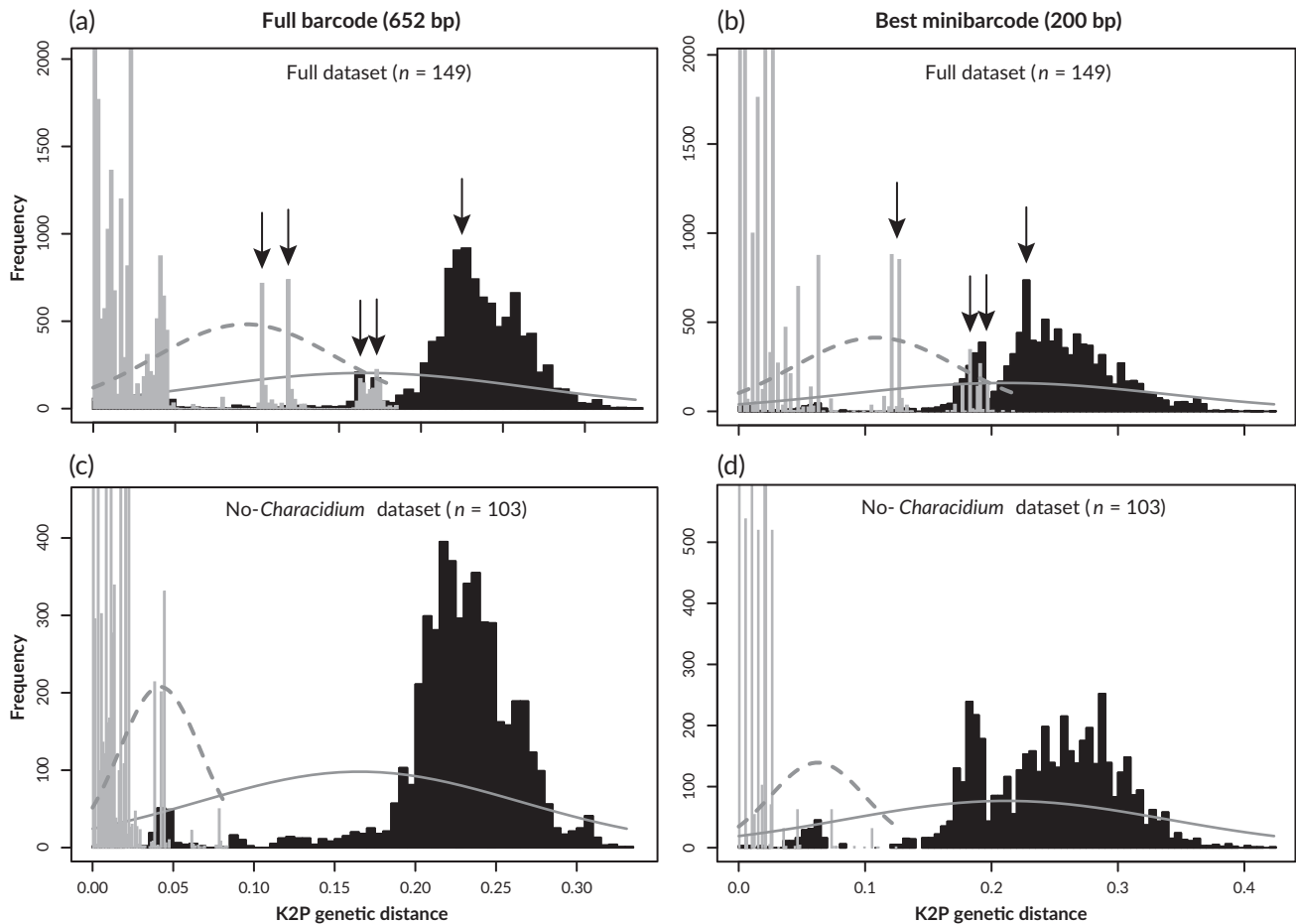
The mean K2P distances within species, genera, families and orders were 1.3% (Table 1), 1.8%, 9.7% and 12.1%, respectively and thus exhibited c. seven-fold higher genetic distances within orders as compared with species (Figure S2). Despite substantial genetic variation, SpeciesIdentifier analysis showed that intraspecific and interspecific distances overlapped widely, from 0% up to 17.8%, with all congeneric-interspecific distances falling in this interval. The region of 90% overlap (5% error in each tail) spanned 0.15% to 12.49%, the latter of which represents the critical intraspecific threshold value. These

**TABLE 1** Cerrado freshwater fish species sampled in the present study, along with their status, sample sizes (*n*), and mean intraspecific Kimura (1980) 2-parameter (K2P) genetic distances. Supraspecific taxonomic ranks apply to all subsequent rows in the table until the next rank is given

Order	Family	Subfamily	Species	Status	<i>n</i>	Mean K2P distance
Characiformes	Crenuchidae	–	<i>Characidium fasciatum</i>	Native	11	5.502%
–	–	–	<i>Characidium gomesi</i>	Native	9	–
–	–	–	<i>Characidium zebra</i>	Native	26	1.797%
–	Characidae	Stethaprioninae	<i>Hasemania hanseni</i>	Native	13	0.494%
–	–	–	<i>Hasemania</i> sp.	Native	2	0%
–	–	–	<i>Moenkhausia aurantia</i>	Native	2	0.441%
–	–	–	<i>Astyanax fasciatus</i>	Native	2	0.220%
–	–	Cheirodontinae	<i>Kolpotocheiroidon theloura</i>	Native	1	–
–	–	Stevardiinae	<i>Bryconamericus turiuba</i> Langeani	Native	3	2.567%
–	–	–	<i>Creagrutus atrisignum</i>	Native	1	–
–	–	–	<i>Hysteronotus megalostomus</i>	Native	2	0%
–	–	–	<i>Knodus moenkhausii</i>	Native	2	0%
–	–	–	<i>Piabina argentea</i>	Native	9	0.512%
–	–	–	<i>Planaltina myersi</i>	Native	2	–
–	Erythrinidae	–	<i>Hoplias malabaricus</i>	Native	1	–
–	Serrasalminae	–	<i>Serrasalmus maculatus</i>	Native	1	–
Siluriformes	Trichomycteridae	Trichomycterinae	<i>Ituglanis goya</i> Datovo,	Native	1	–
–	Callichthyidae	Corydoradinae	<i>Aspidoras eurycephalus</i>	Native	1	–
–	–	–	<i>Aspidoras fuscoguttatus</i>	Native	2	0%
–	Loricariidae	Otothyriinae	<i>Microlepidogaster longicollis</i>	Native	2	0.220%
–	–	–	<i>Rhinolekos</i> sp.	Native	2	0%
–	–	Loricariinae	<i>Harttia punctata</i>	Native	1	–
–	–	–	<i>Rineloricaria latirostris</i>	Native	2	0%
–	–	Hypostominae	<i>Hypostomus ancistroides</i>	Native	3	0%
–	–	–	<i>Hypostomus</i> sp. 1	Native	1	0%
–	–	–	<i>Hypostomus</i> sp. 2	Native	24	1.362%
–	–	–	<i>Ancistrus aguaboensis</i>	Native	3	6.028%
–	Heptapteridae	–	<i>Cetopsorhamdia iheringi</i>	Native	2	2.231
–	–	–	<i>Heptapterus</i> sp.	Native	1	–
–	–	–	<i>Imparfinis borodini</i>	Native	1	–
–	–	–	<i>Phenacorhamdia unifasciata</i>	Native	4	0%
–	–	–	<i>Rhamdia quelen</i>	Native	1	–
–	Pimelodidae	–	<i>Pimelodella</i> sp.	Native	2	2.929%
Gymnotiformes	Sternopygidae	–	<i>Eigenmannia trilineata</i>	Native	2	7.038%
Cyprinodontiformes	Rivulidae	–	<i>Melanorivulus pictus</i>	Native	1	–
–	Cynolebiidae	Cynolebiinae	<i>Simpsonichthys boitonei</i>	Native	1	–
–	Poeciliidae	Poeciliinae	<i>Phalloceros harpagos</i>	Native	4	0%
–	–	–	<i>Poecilia reticulata</i>	Exotic	2	0%
Perciformes	Centrarchidae	–	<i>Lepomis gibbosus</i>	Exotic	2	0%

results indicated lack of a DNA barcode gap, which was confirmed visually in BarcodingR by extensive overlapping distance distributions (Figure 2a) and by plotting boxplots of the distribution of pairwise K2P distances in each 50 bp sliding window at a per-codon frequency in spider, as shown by the extensive overlap in Figure S3. Accordingly,

success was low to moderate using distance-only criteria. Identification success under the BM approach was 68.5% (*n* = 102), with 29 ambiguous queries (19.5%) and 18 incorrect (12.1%). Identification success under BCM was slightly worse, ranging from 57.7% to 68.5% across the 1% to 15% threshold values in our sensitivity analysis and



**FIGURE 2** DNA barcode-gap results for samples of freshwater fishes from the Brazilian Distrito Federal and surrounding areas of the Cerrado biome. Barplots show the relative distributions of intraspecific genetic distances (■) and interspecific genetic distances (■) based on Kimura's (1980) 2-parameter (K2P) model a, c for the full barcode alignment and b, d the best minibarcode given in Table 1. (a), (b) Results for the full dataset; (c), (d) results when problematic *Characidium* samples were excluded from the analysis. (i), Sets of extreme pairwise distances removed during no-*Characidium* analyses. Extremely long vertical bars are truncated for ease of presentation, but had frequencies ranging in the thousands

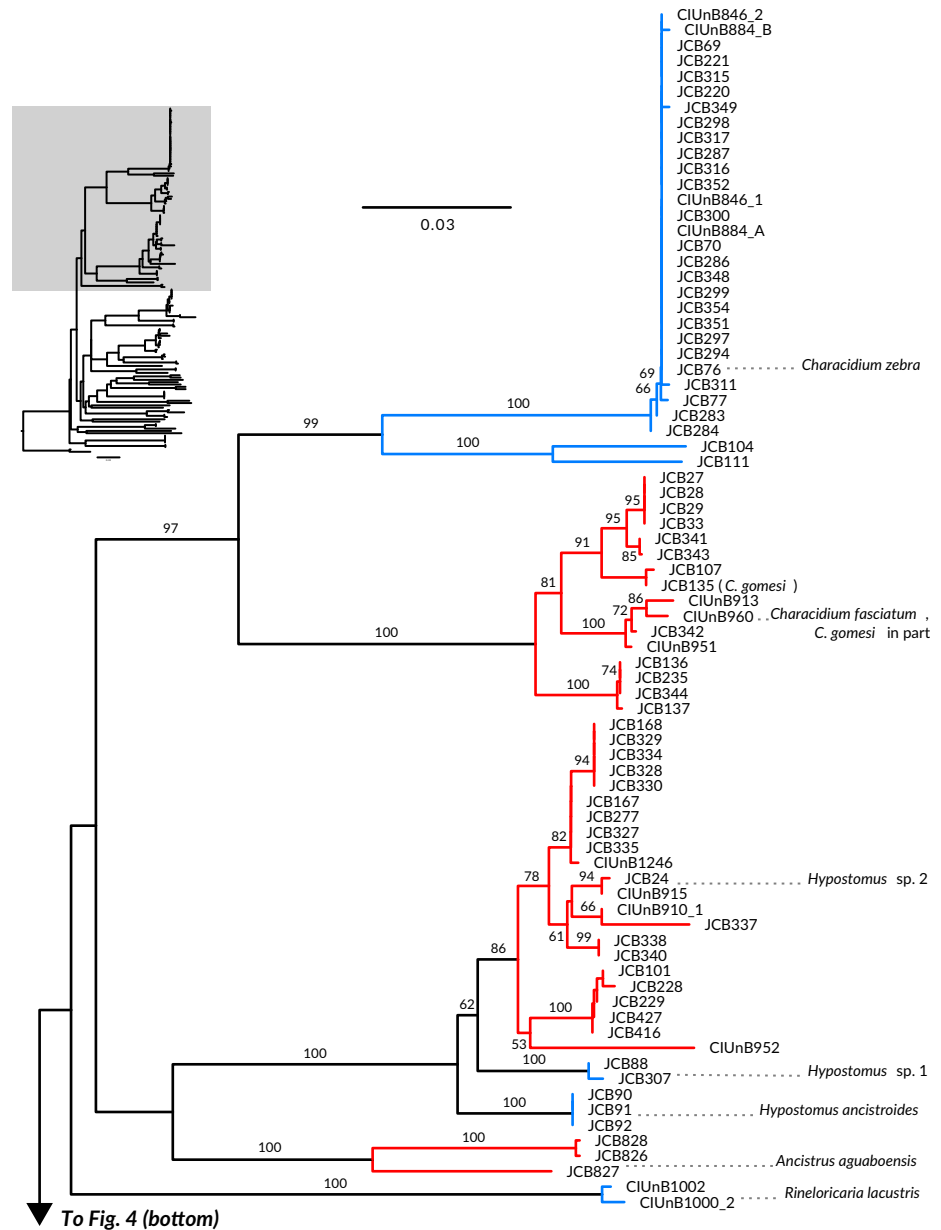
being 68.5% at the intraspecific threshold value (12.49%). However, while ambiguous identifications were the same, BCM yielded a lower incorrect rate of only 10 queries (6.7%) and only 8 sequences had no match closer than 12.49%. Clustering samples at the threshold value yielded 27 clusters, only 18 (66.7%) of which corresponded to a single species/lineage in the library, suggesting that clustering by closest match did not outperform our best match approaches above. Conducting BCM in spider at the intraspecific threshold value above gave results similar to those obtained with SpeciesIdentifier, placing success at 67.8%, with 31 ambiguous queries (including no id queries). Our threshold-level BCM results did not change when singletons were excluded in SpeciesIdentifier, which again gave a 67.8% success rate.

From comparisons of K2P distances, we determined that a sizable portion of overlap in intraspecific and interspecific distances reported above was caused by a single problematic set of samples, those of the genus *Characidium* Reinhardt 1867 (Figure 2). When *Characidium* were removed from the matrix and distances were recalculated, a barcode gap was not obtained (Figure 2c) but the degree of overlap of intra- versus interspecific distances in SpeciesIdentifier declined markedly to 4%

with a region of 90% overlap of only 0.96% (3.97% to 4.94%), re-establishing the cut-off threshold at a lower value of 4.94%. Therefore, we re-ran the entire distance-based analysis on the no-*Characidium* library in SpeciesIdentifier. We found that success rate was greatly improved when *Characidium* were excluded, with identification under BM reaching 83.5% ( $n = 86$ ), with a massive drop in ambiguous queries to 2 (1.94%) but a similar incorrect rate (14.6%). While BCM performance remained lower than that of BM, the BCM results improved to an 80.6% success rate with nearly the same ambiguous and incorrect queries as BM. Clustering sequences at the new threshold yielded 34 clusters, with 27 (71.1%) of clusters corresponding to species/lineage labels. Thus, in the absence of taxonomically problematic samples, clustering was notably improved but did not outperform BM.

Our NJ tree (Figures 3 and 4) had a branch-length distribution suggesting a long timescale of evolution, leading to short internode distances near the root among clusters with low bootstrap support values. This is consistent with Aquino and Colli's (2017) taxonomy-based super tree for Distrito Federal fish communities, which they

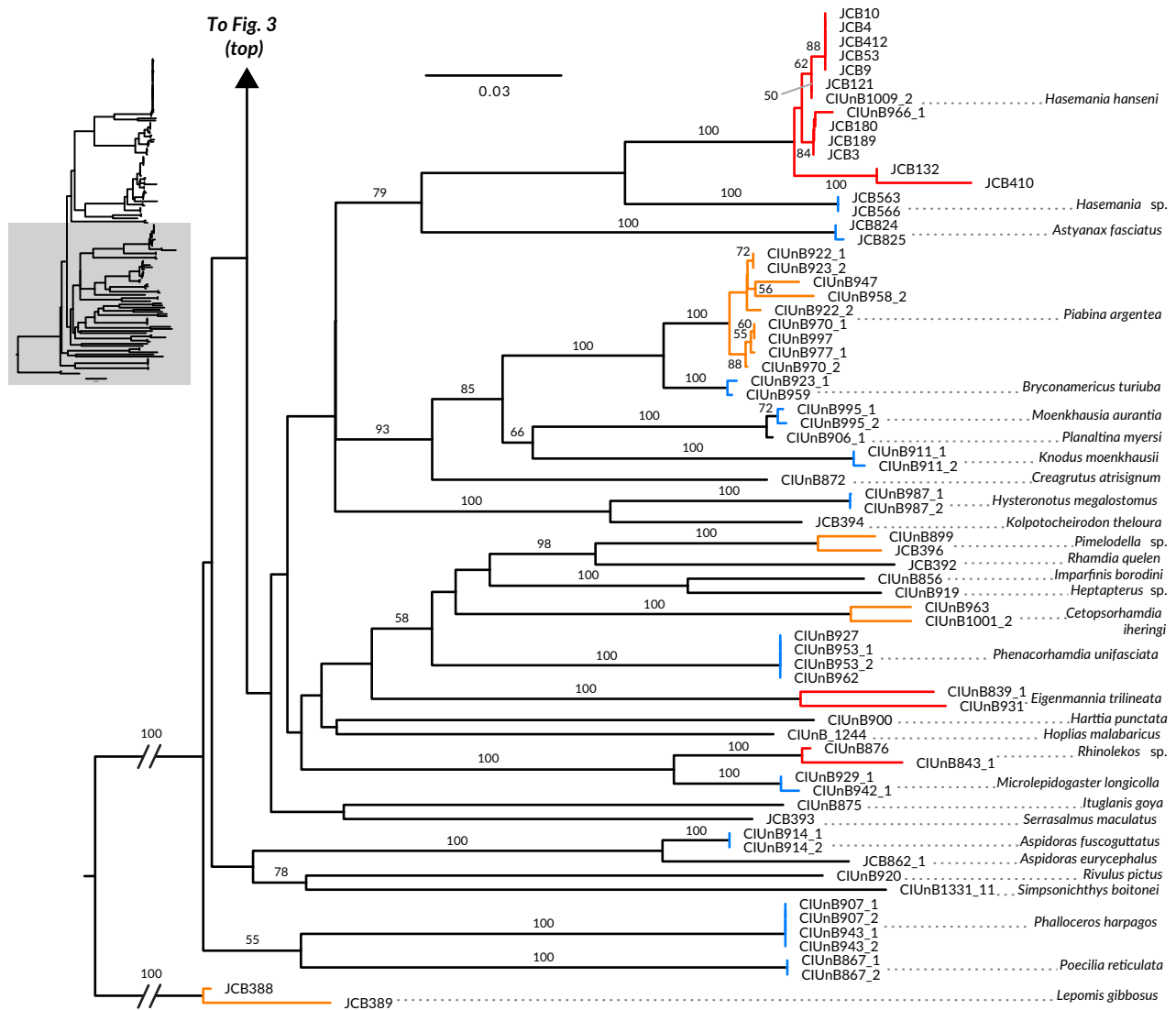
**FIGURE 3** Upper portion of neighbour-joining tree (area indicated by grey shading at top-left) of 149 *cox1* sequences from 39 species/lineages of freshwater fishes from the Brazilian Distrito Federal and surrounding areas of the Cerrado biome, based on K2P genetic distances. Branches corresponding to categories of intraspecific distances are highlighted different colours ((—), < 2%; (—), 2.0–5.3%; (—) > 5.3%). Bootstrap proportion measures of nodal support >50% are shown along branches of the tree



calibrated to a root depth of 210 million years ago (Mya) using multiple fossil and phylogenetic node calibrations. Instances of putative long-branch attraction were identified involving the characiform species *Serrasalmus maculatus* Kner 1858 and *Hoplias malabaricus* (Bloch 1794), both singletons with long branches that grouped incorrectly with catfish species (Siluriformes) but without significant nodal support (Figure 4). Many families formed monophyletic clusters (10/13, or 77%), but the three largest families were para or polyphyletic, with characiform lineages nested within Siluriformes, which itself formed a paraphyletic grade toward the root interspersed with characiform, gymnotiform and cyprinodontiform lineages. For example, clusters of conspecific *Eigenmannia trilineata* López & Castello 1966 (Gymnotidae) and *Melanorivulus pictus* (Costa 1989) + *Simpsonichthys boitoni* Carvalho 1959 (Rivulidae) erroneously clustered with catfish lineages. The Bayesian inference topology from MrBayes was very

similar to that of the NJ tree, with a similar pattern of short internodes–polytomies near the root with low Bayesian posterior probability support, the same patterns of long-branch attraction and similar para or polyphyletic patterns of characiform families (Figure S4). Given sufficient similarity between the two topologies, we focus most of the remainder of our discussion on the NJ tree. Despite the limitations discussed above, identifications based on tree-based criteria applied to the NJ topology were successful in c. 85% of cases ( $n = 33/39$  species/lineages, 84.6%) under the first method, with much lower success of 64.1% of cases when all 13 singletons were counted as ambiguous (Hebert *et al.*, 2003). Still, the rate of misidentifications under both scenarios was low, at 2.6% and 0%, respectively. Rosenberg's monophyly test results are presented over nodes of the NJ tree in Figure S5, which demonstrates that clustering of nodes in the region of the tree with the most singleton and two-sample species





**FIGURE 4** Lower portion of neighbour-joining tree (area indicated by grey shading at top-left) of 149 *cox1* sequences from 39 species/lineages of freshwater fishes from the Brazilian Distrito Federal and surrounding areas of the Cerrado biome, based on K2P genetic distances. Branches corresponding to categories of intraspecific distances are highlighted different colours (—), < 2%; (—), 2.0–5.3%; (—) > 5.3%). Bootstrap proportion measures of nodal support > 50% are shown along branches of the tree

had the greatest proportion of nodes that could be monophyletic at random (results were the same during attempts to account for small sample size by increasing  $\alpha$  to 0.10 or 0.20; J.C.B., pers. obs.). In light of our distance-based results, we removed *Characidium* samples and built a second NJ tree (Figure S6). Rescoring monophyly over this tree revealed an c. 85% (86.1%) success rate under the simplest method, but a c. 64% success rate when counting the increased proportion of singletons as ambiguous.

A total of 148 sequences (99.3%) met quality criteria for BIN assignment on BOLD and these were assigned into 46 BINs (Table S1), or OTUs forming hypotheses of candidate species limits. Most BINs were concordant ( $n = 24$ ), 19 BINs were based on singletons and only 3 BINs were categorised as discordant, containing two or more taxonomic designations at species or higher rank. Discordant BINs were each due to a single conflicting species and discordant

BINs due to higher taxonomic-rank conflicts included BOLD: AAE2970 ( $n = 2$  *Moenkhausia aurantia* Bertaco, Jerep & Carvalho 2011,  $n = 1$  *P. myersi*) and BOLD:ABZ1565 ( $n = 1$  *Bryconamericus turiuba* Langeani, Lucena, Pedrini & Tarelho-Pereira 2005,  $n = 1$  *P. argentea*). Subsequent voucher checks revealed that the specimens associated with the first of these conflicts have correct identifications; however, the second of these conflicts resulted from a single mis-identification and the correct species assignment for specimens in BIN BOLD:ABZ1565 is *P. argentea*. The only discordant BIN due to a species-level conflict was BOLD:ADR3920 ( $n = 1$  *Characidium gomesi* Travassos 1956,  $n = 1$  *C. fasciatum*) found in the problematic genus *Characidium*, a case that could reflect inadequate taxonomy.

A higher number of BINs than nominal species or threshold clusters could indicate the presence of undescribed species or genetic structure; consistent with this, we found 7 species/lineages that were

assigned to multiple BINs (Table S1), or putative candidate species. Five of these species/lineages were also classified into the extreme divergence category during distance comparisons, as shown by high-light colours along tip branches in Figures 3 and 4. These included maximum intraspecific distances of 17.8%, 7.8%, 6.8%, 6.2% and 5.4% among *Characidium* spp. (maximum intraspecific distance within single lineage: 13.9%, *C. fasciatum*), *Ancistrus aguaboensis* Fisch-Muller, Mazzoni & Weber 2001, *Hypostomus* sp. 2, *Eigenmannia trilineata* and *Hasemania hansenii* (Fowler 1949), barcodes, respectively. Sample sizes were too small to conduct analyses of population genetic structure in *A. aguaboensis* and *E. trilineata*. However, population structure results for *Characidium* spp. revealed that the five main clusters identified from the NJ and Bayesian topologies formed six distinct parsimony networks (Figure S7), three of which were significantly different with large pairwise  $F_{ST}$  estimates of 0.75–1 ( $p < .001$  to  $p < .05 \pm .001$ ). Highly differentiated clusters exhibited phylogeographic structure with the *Characidium zebra* Eigenmann 1909 + *C. fasciatum* cluster (cluster 1; BIN: BOLD:ADR3189) mainly from the Upper Tocantins basin, a cluster of *C. fasciatum* plus *C. gomesi* in part (cluster 5; BIN: BOLD:ADR3921) from the upper Paraná Basin and a *C. fasciatum* cluster (cluster 4; BIN: BOLD:AAO5269) from the upper São Francisco Basin (Figure 3a and Figure S7). Results for *Hasemania* Ellis 1911 confirmed that *H. hansenii* and *Hasemania* sp. form two parsimony networks separated by 12 mutations (BINs: BOLD:AAO6055, BOLD:AAY2122; Figure S8) and are highly differentiated based on a significant pairwise  $F_{ST}$  of 0.34 ( $p < .05 \pm .001$ ). A phylogeographic break was observed in this lineage between *Hasemania* sp., which were exclusively from the southernmost upper Paraná site at Paracatu, MG (site 42; Figure 1) and all other *Hasemania* samples (northern sites, all three basins). *Hypostomus* Lacépède 1803 clusters also exhibited significant population structure, with *Hypostomus* sp. 2 forming four unconnected parsimony networks (clades 1–4; BINs: imperfectly corresponding to BOLD:AAD0323, BOLD:ACE2903 and BOLD:ACZ7798) and *Hypostomus ancistroides* (Ihering 1911) and *Hypostomus* sp. 1 each forming a unique network (concordant BINs BOLD:AAW9386 and BOLD:ADR4025, respectively; Figure S9). These clusters were significantly differentiated at pairwise  $F_{ST}$ , with the most robust estimates based on the largest sample sizes being the  $F_{ST}$  of 0.79 ( $p = <.0001$ ) distinguishing *Hypostomus* sp. 2 (parsimony network clade 1) from *Hypostomus* sp. 1 and an  $F_{ST}$  of 0.82 ( $p < .001 \pm .0003$ ) distinguishing *Hypostomus* sp. 2 from *H. ancistroides*. The *Hypostomus* results also highlight a distinct pattern of phylogeographical differentiation, with *Hypostomus* sp. 1 and *H. ancistroides* samples only in the upper Tocantins Basin and *Hypostomus* sp. 2 in all three sampled basins.

Minibarcoding analysis with different sliding-window sizes revealed optimum windows in similar positions along the *cox1* gene, clustering within the final 200 bp of the alignment (Table 2). The proportion of zero non-conspecific distances was greater than zero in all cases, consistent with the lack of a DNA barcode gap in the original full-length barcodes. The best window was a 200 bp window starting from *cox1* nucleotide position 379 and going through position 579. In BM and BCM tests, this 200 bp window minibarcoding exhibited no barcode

gap (Figure 2b) but correctly classified 66.4% of queries (both tests), which was very similar to the level of identification success in the BM and BCM analyses of the full barcode sequences. The 12.75% pairwise intraspecific distance threshold value estimated for the minibarcoding region was also nearly identical to that of the full barcodes. Removing problematic *Characidium* samples from the minibarcoding alignment again improved the barcode gap (Figure 2d) and re-running the distance analyses improved results drastically, increasing BM and BCM success rates to 80.6% and 75.7%, respectively, at a new threshold value of 3.09%. Analysis of polymorphic sites, as indicated by nucleotide diversity ( $\pi$ ) and Watterson's theta ( $\theta_W$ ), revealed that this minibarcoding corresponded to a relatively conserved region (Figure 5) positioned such that a 5' primer could be designed in the transition area to conserved sites and either reverse primer used during PCR amplification (see §2) could be used as the 3' primer.

## 4 | DISCUSSION

Most studies of DNA barcoding in freshwater fishes report specimen identification success rates  $\geq 90\%$  and as high as 99–100%, over broad spatial and taxonomic scales from continents (April *et al.*, 2011; Hubert *et al.*, 2008; Pereira *et al.*, 2013; Ward *et al.*, 2009) to large regions (Benzaquem *et al.*, 2015; Díaz *et al.*, 2016; Pereira *et al.*, 2011). In addition to clear patterns of monophyly on NJ clustering trees, identification success is often attributed to the presence of a DNA barcode gap caused by lower divergence within species than between species, supporting the application of arbitrary distance thresholds of around c. 2% in freshwater fish species (Ward *et al.*, 2009). Fish barcoding studies also frequently recover limited spatial-genetic structure between populations within species (Ward, 2009). These results contradict patterns of mtDNA variation, genetic structure and taxonomic uncertainty across various spatiotemporal scales in the metazoan literature, including evidence from 30 years of phylogeography studies showing that cryptic genetic structuring within species is widespread (Avice, 2000; Bagley & Johnson, 2014; Bickford *et al.*, 2006). The implied 0–10% failure rate is also much less than the estimated 23% incidence of para or polyphyletic patterns in Funk and Omland's (2003) review of mtDNA genetic results across a wide diversity of animal taxa. This suggests that, while barcoding of fishes has been and continues to be a successful endeavour, DNA barcoding success rates have probably been overestimated, as indicated by studies of fishes (Bremner *et al.*, 2016) as well as dipterans, butterflies and marine gastropods (Meyer & Paulay, 2005; Meier *et al.*, 2006, 2008; Wiemers & Fiedler, 2007). The underlying issue is that failing to include allopatric species (*i.e.*, sister taxa) can lead barcode libraries from broad geographical scales to overestimate interspecific variation, inflating barcode success (Meyer & Paulay, 2005). In a recent analysis of spiders, expanding the geographical scale up to discrete temperate continents increased rather than decreased DNA barcoding and gap-analysis success (Čandek & Kuntner, 2015). However, the generality of this finding for the Neotropics seems doubtful, especially given its

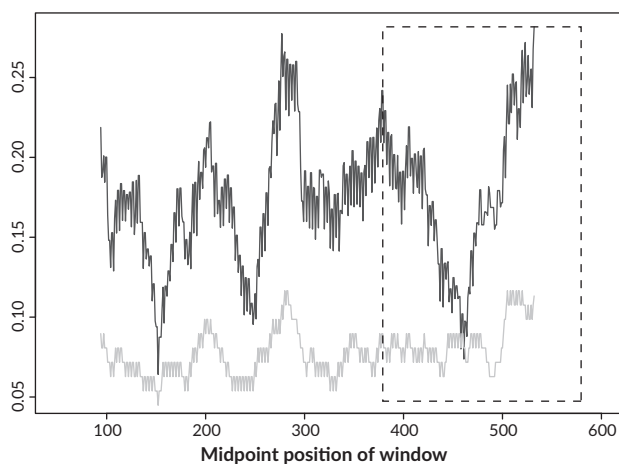
**TABLE 2** Results of sliding-window analyses in spider (Brown *et al.*, 2012). Results are presented across four different sliding-window sizes for the best window positions meeting the criterion of having a proportion of identical neighbour-joining (NJ) clades shared between full-length data and the window of >85% of identical clades for shallow nodes (shallower than the midpoint of root depth). Results from the overall best window are shown in bold font. Mean distance refers to the mean of distances in the matrix of pairwise K2P distances calculated from sequences within the window

Window size (bp)	Position (starting nucleotide)	Mean distance	Proportion zero non-conspecific distances	Proportion of identical clades shared
50	none	–	–	–
100	502	0.28	0.24	0.90
150	487	0.25	0.23	0.88
200	<b>379</b>	<b>0.23</b>	<b>0.21</b>	<b>0.95</b>

large continuous habitats and distinct, yet historically connected biomes (Werneck *et al.*, 2012; Ledo & Colli, 2017) and river drainages (Menezes *et al.*, 2008; Lima & Ribeiro, 2011; Aquino & Colli, 2017; Bagley *et al.* (n.d.)). Our results indicate a perhaps more realistic scenario: that distance and tree-based criteria successfully identify c. 64.1–85% of Cerrado headwater fish species and candidate species based on a DNA barcode library of 149 sequences representing c. 35% taxonomic coverage of the regional freshwater fish assemblage. Indeed, the midpoint of this success range is 74.6%, which closely matches the para or polyphyly rate in Funk and Omland (2003). Below, we discuss our findings and show that the two most likely factors contributing to incorrect and ambiguous identifications in our study are unidentified phylogeographical structure (*e.g.*, cryptic species or lineages) and inadequate taxonomy.

The barcode gap is a topic of contention and debate in DNA barcoding studies, as no single threshold value applies to all species (Virgilio *et al.*, 2012) and threshold approaches could fail in c. 20% of cases due to para or polyphyly (this study; Meyer & Paulay, 2005) and

incomplete sampling (Ratnasingham & Hebert, 2007). Among the most striking results of our study was that we found evidence for overlapping intraspecific and interspecific genetic distances (even after discarding the upper 5% of distances) causing no barcode gap in the reference DNA barcode library for Distrito Federal fishes (Figure 2), over a relatively smaller regional scale of c. 5000 km<sup>2</sup> (Figure 1). Breman *et al.* (2016) encountered a similar degree of barcode overlap in a barcoding study of littoral cichlid species from Lake Tanganyika, which has a total area of c. 32,900 km<sup>2</sup> but a much smaller littoral zone, determined by shallow waters along the c. 1800 km shoreline (Yohannes, 2009). In addition to similarities in area between the two studies, we examined a similar percentage of the regional fish assemblage; 52% here, as compared with c. 50% in Breman *et al.* (2016). Breman *et al.* (2016) attributed the observed lack of a barcode gap to ill-informed taxonomy and interspecific hybridization and introgression of sympatric cichlid species during adaptive radiation. We attribute the lack of a barcode gap in our study largely to the substantial phylogeographical structuring and genetic divergence within five species/lineages for which we had relatively large sample sizes from multiple sites across the study area: *Characidium* spp., *A. aguaboensis*, *Hypostomus* sp. 2, *E. trilineata* and *H. hanseni*. This is illustrated by clustering, parsimony networks (Figures S7–S9),  $F_{ST}$  results and assignment of the barcodes of these species/lineages into multiple BINs (Table S1). We interpret these novel patterns of genetic divergences above the fish threshold of 2% (Ward *et al.*, 2009) as indicating putative candidate species warranting further study. Distance analyses were also specifically affected by inclusion of *Characidium* barcodes, which exhibited the highest degree of para or polyphyly of any species/lineages we investigated (Figure 3 and Figure S4) and also formed the largest number of BINs, with nine BINs in the genus (Table S1). *Characidium* is the most speciose genus within Crenuchidae yet remains in a state of taxonomic flux: the genus lacks a clear hypothesis of phylogenetic relationships (Buckup, 1993) and has experienced a string of recent species descriptions (Mendonça & Netto-Ferreira, 2015). New putative undescribed species of *Characidium* also occur within the Distrito Federal (Aquino & Colli, 2017) and these may have been misidentified in our study, causing species labels to improperly match cryptic species/lineages with previously unclear boundaries. Thus, we consider *Characidium* species from the DF to warrant genetic and taxonomic re-examination, preferably



**FIGURE 5** Distributions of DNA polymorphism across the DNA barcode gene of freshwater fish from the Brazilian Distrito Federal and surrounding areas of the Cerrado biome, as inferred from nucleotide diversity ( $\pi$ , —,  $\pi$ ) and the population mutation parameter (Watterson's theta, —,  $\theta_W$ ) in DNASP (Librado & Rozas, 2009). Each data point shows results from a 20 bp window, the midpoint of which is plotted along the x-axis. □, The best 200 bp minibarcode region

using genome-wide markers (e.g., single nucleotide polymorphisms; Peterson *et al.*, 2012) or an integrative taxonomy framework combining data from DNA sequences and morphological characters (Padial *et al.*, 2010; Puillandre *et al.*, 2012; Recknagel *et al.*, 2013). Detailed morphological and genetic studies of the *Characidium* lineages and BINs we have uncovered seems likely to yield new species descriptions, increasing the alpha diversity of the regional fish fauna.

Under these circumstances, we found that distance-based methods that rely on a barcode gap, including BM and BCM approaches (Meier *et al.*, 2006, 2008), initially exhibited low identification success (68.5%) when *Characidium* samples were included. The upper c. 12% error rates in our initial analysis matched the lower end of the 15–27% error rates for littoral African cichlids (Bremar *et al.*, 2016). However, our outright error rates are much lower than the 17–20% error (false negative) rates obtained by Meyer and Paulay (2005) in cownies, or the low level of DNA barcoding success that Toffoli *et al.* (2008) found in *Potamotrygon* Garman 1877 freshwater stingrays from across Brazil, whose DNA barcodes could not discriminate 40% of nominal *Potamotrygon* spp. Paz *et al.* (2014) found that DNA barcodes from rosy tetras (genus *Hyphessobrycon* Durbin 1908) of the Amazon Basin exhibited para or polyphyletic patterns or phylogeographical structure indicative of potentially cryptic species-level diversity, which obfuscated DNA barcode identifications in 40% of nominal species (four species) sampled. Although our error rates were lower, general lack of a DNA barcode gap ensured that it would always be impossible to successfully identify a fraction of our samples using distance-based methods. However, this fraction was much higher when *Characidium* sequences were included in the analysis and dropped to statistically acceptable levels (c. 2%) when *Characidium* barcodes were removed. Success rates also improved drastically to c. 85% after removing *Characidium* barcodes. These findings underscore the importance of identifying and removing barcodes representing potentially taxonomically problematic or misidentified samples when conducting DNA barcoding, as highlighted by previous studies (Bremar *et al.*, 2016; Meier *et al.*, 2006; Tautz *et al.*, 2003). Moreover, we found that distance-based methods performed similarly to tree-based approaches using simple monophyly and nodal support criteria, which do not depend on presence–absence of a barcode gap but instead reflect marker informativeness, coalescent sorting and species limits. Indeed, our simplest monophyly criterion yielded results similar to the distance analysis (84.6% success at identification), as did using monophyly criteria classifying singletons as ambiguous (cf. Hebert *et al.*, 2003). Still, the performance of the simple monophyly criterion proved more robust to errors in the NJ clustering and Bayesian tree topologies (e.g., *Characiformes* polyphyly associated with long-branch attraction; Figures 3 and S4).

Notably, our results are not affected by several issues that plagued recent barcoding studies, such as a high proportion of identical barcodes or consensus barcodes shared between different species (Meier *et al.*, 2006; Tautz *et al.*, 2003); low sequence divergence due to unusually recent or rapid diversification (Bremar *et al.*, 2016; Toffoli *et al.*, 2008) and associated incomplete lineage sorting (Wiemers & Fiedler, 2007); or collection of samples from only one or

a few sites (Hebert *et al.*, 2003, 2004; Triantafyllidis *et al.*, 2011). Instead, only one species (2.6% of taxa) in our study, *C. gomesi*, shared an identical *cox1* haplotype with another species and only with a congener, *C. fasciatum* (BIN: BOLD:ADR3920). We also observe a very deep 17.8% genetic divergence in *Characidium*, which roughly corresponds to at least 9 million years of divergence if we assume a standard 1–2% pairwise fish mtDNA clock rate (Burridge *et al.*, 2008). Unlike previous studies (Wiemers & Fiedler, 2007; Paz *et al.*, 2014), this result seems unrelated to incomplete lineage sorting, which is mostly associated with recent speciation and large effective population sizes (Avice, 2000; Funk & Omland, 2003). We also generally observed high levels of intraspecific genetic divergence, with the presence of phylogeographical structure and cryptic lineages driving us to reject the barcoding null hypothesis that genetic variation within nominal species will be lower on average than that witnessed between species. For example, intraspecific divergences within *A. aguaboensis* and *E. trilineata*, are c. 3 to 4.5fold greater on average than those within other species. Although our intraspecific sampling was limited overall (Table 1), extensive overlap in intra and interspecific divergences in our study cannot be explained completely by small sample sizes. First, the overall range of genetic divergence in interspecific comparisons is large (0% to >30%) and congeneric-interspecific distances were high on average, at around 8.7% over 10 species with valid comparisons (Figure 2 and Table S2). Second, the lowest congeneric-interspecific divergences occurred among *Characidium* species sharing identical or similar haplotypes and these formed the single most problematic distance set, with the next smallest congeneric-interspecific distances being c. 3% to 5% for *Hypostomus* species/lineages; values widely considered to represent interspecific differentiation in fishes (Hubert *et al.*, 2008; Ward, 2009; Ward *et al.*, 2005, 2009). This also provides additional justification for our removal of *Characidium* barcodes in the reanalysis and corresponding results discussed above, as well as the need for future taxonomic reassessment.

Whereas amplification of barcodes with sufficient sequence length and overlap from degraded museum material has been an issue for DNA barcoding, our results support the idea that minibarcoding provides a means of obtaining similar accuracy to standard DNA barcoding using Sanger or high-throughput sequencing (Doña *et al.*, 2015; Hajibabaei *et al.*, 2006; Meusnier *et al.*, 2008). Our sliding window results identified a 200 bp minibarcode region of the *cox1* gene for Cerrado freshwater fishes that performed equally well at specimen identification as compared with the entire DNA barcode (Table 2) and seems well positioned for the development of degenerate primers for PCR (Figure 5). Perhaps with the resolution of para or polyphyletic patterns and species limits in lineages such as *Characidium* through careful taxonomic studies and greater taxonomic and numerical sampling of the regional fish assemblage, future barcoding studies will infer a clear barcode gap for this system and surrounding areas of the Cerrado. Regardless, our progress on the minibarcode front suggests that future studies of barcoding in Neotropical freshwater fishes will benefit from capitalising on the final c. 200 bp region of the DNA barcode as a target for rapid development of DNA barcode libraries



for hundreds or thousands of samples (hundreds of species) simultaneously using second-generation next-generation sequencing platforms and beyond. Combined with new barcoding techniques for dietary composition and metabarcoding of parasite and microbiome communities using mtDNA *cox1* and 18 s ribosomal DNA markers, high-throughput DNA barcode library development may allow not only high-resolution inferences of fish species limits, but also of ecological and coevolutionary dynamics (Bohmann *et al.*, 2011; Zhan *et al.*, 2014). Indeed, such integration could pave the way for broader, synthetic assessments of ecosystem composition and function to improve conservation and management (Gibson *et al.*, 2014) in the Cerrado biodiversity hotspot.

## ACKNOWLEDGEMENTS

We thank Thiago B. d'Araujo Couto, Ingrid Pinheiro Paschoaletto and Yan F. Figueira-Soares for assistance during fieldwork for this project.

## AUTHOR CONTRIBUTIONS

J.C.B. and P.P.U.A. conceived and designed the study, conducted fieldwork and organised museum collections. J.C.B. and M.F.B. generated the data. J.C.B. wrote the draft of the manuscript and P.P.U.A., M.F.B., F.L.N. and G.R.C. provided edits and comments on the manuscript. J.C.B., F.L.N. and G.R.C. obtained funding for the study. All authors read and approved the final manuscript.

## ORCID

Justin C. Bagley  <https://orcid.org/0000-0001-6737-8380>

## REFERENCES

- Abell, R., Thieme, M. L., Revenga, C., Bryer, M., Kottelat, M., Bogutskaya, N., ... Stiassny, M. L. (2008). Freshwater ecoregions of the world: A new map of biogeographic units for freshwater biodiversity conservation. *Bioscience*, 58, 403–414.
- Agostinho, A. A., Thomaz, S. M., & Gomes, L. C. (2005). Conservação da biodiversidade em águas continentais do Brasil. *Megadiversidade*, 1, 70–78.
- Albert, J. S., Petry, P., & Reis, R. E. (2011). Major biogeographic and phylogenetic patterns. In J. S. Albert & R. E. Reis (Eds.), *Historical biogeography of Neotropical freshwater fishes* (pp. 21–57). Berkeley, CA: University of California Press.
- Albert, J. S., & Reis, R. E. (Eds.). (2011a). *Historical biogeography of Neotropical freshwater fishes*. Berkeley, CA: University of California Press.
- Albert, J. S., & Reis, R. E. (2011b). Introduction to Neotropical freshwaters. In J. S. Albert & R. E. Reis (Eds.), *Historical biogeography of Neotropical freshwater fishes* (pp. 3–20). Berkeley, CA: University of California Press.
- Anache, J. A. A., Flanagan, D. C., Srivastava, A., & Wendland, E. C. (2018). Land use and climate change impacts on runoff and soil erosion at the hillslope scale in the Brazilian Cerrado. *Science of the Total Environment*, 622, 140–151.
- April, J., Mayden, R. L., Hanner, R. H., & Bernatchez, L. (2011). Genetic calibration of species diversity among North America's freshwater fishes. *Proceedings of the National Academy of Sciences*, 108, 10602–10607.
- Aquino, P. D. P. U., & Colli, G. R. (2017). Headwater captures and the phylogenetic structure of freshwater fish assemblages: A case study in Central Brazil. *Journal of Biogeography*, 44, 207–216.
- Aquino, P. D. P. U., Schneider, M., Martins-Silva, M. J., Fonseca, C. P., Arakawa, H. B., & Cavalcanti, D. R. (2009). Ictiofauna dos córregos do Parque Nacional de Brasília, bacia do Alto Rio Paraná, Distrito Federal, Brasil Central. *Biota Neotropica*, 9, 217–230.
- Araújo, N. B., & Tejerina-Garro, F. L. (2007). Composição e diversidade da ictiofauna em riachos do Cerrado, bacia do ribeirão Ouvidor, alto rio Paraná, Goiás, Brasil. *Revista Brasileira de Zoologia*, 24, 981–990.
- Arnell, N. W., & Gosling, S. N. (2013). The impacts of climate change on river flow regimes at the global scale. *Journal of Hydrology*, 486, 351–364.
- Avice, J. C. (2000). *Phylogeography: The history and formation of species*. Cambridge, MA: Harvard University Press.
- Bagley, J. C., Aquino, P. D. P. U., Hrbek, T., Hernandez, S. G., Langeani, F., & Colli, G. R. Using ddRAD-seq phylogeography to test for genetic effects of headwater river capture in suckermouth armored catfish (Loricariidae: *Hypostomus*) from the central Brazilian shield. *Molecular Ecology*. in revision
- Bagley, J. C., & Johnson, J. B. (2014). Phylogeography and biogeography of the lower Central American Neotropics: Diversification between two continents and between two seas. *Biological Reviews*, 89, 767–790.
- Benedito-Cecilio, E., Minte-Vera, C. V., Zawadzki, C. H., Pavanelli, C. S., Rodrigues, F. H. G., & Gimenes, M. F. (2004). Ichthyofauna from the Emas National Park region: Composition and structure. *Brazilian Journal of Biology*, 64, 371–382.
- Benzaquem, D. C., Oliveira, C., da Silva Batista, J., Zuanon, J., & Porto, J. I. R. (2015). DNA barcoding in pencilfishes (Lebiasinidae: *Nannostomus*) reveals cryptic diversity across the Brazilian Amazon. *PLoS One*, 10, e0112217.
- Beuchle, R., Grecchi, R. C., Shimabukuro, Y. E., Seliger, R., Eva, H. D., Sano, E., & Achard, F. (2015). Land cover changes in the Brazilian Cerrado and Caatinga biomes from 1990 to 2010 based on a systematic remote sensing sampling approach. *Applied Geography*, 58, 116–127.
- Bickford, D., Lohman, D. J., Sodhi, N. S., Ng, P. K. L., Meier, R., Winker, K., ... Das, I. (2006). Cryptic species as a window on diversity and conservation. *Trends in Ecology and Evolution*, 22, 148–155.
- Bohmann, K., Monadjem, A., Noer, C. L., Rasmussen, M., Zeale, M. R., Clare, E., ... Gilbert, M. T. (2011). Molecular diet analysis of two African free-tailed bats (Molossidae) using high throughput sequencing. *PLoS One*, 6, e21441.
- Breman, F. C., Loix, S., Jordaens, K., Snoeks, J., & Van Steenberge, M. (2016). Testing the potential of DNA barcoding in vertebrate radiations: the case of the littoral cichlids (Pisces, Perciformes, Cichlidae) from Lake Tanganyika. *Molecular Ecology Resources*, 16, 1455–1464.
- Brown, S. D. J., Collins, R. A., Boyer, S., Lefort, M.-C., Malumbres-Olarte, J., Vink, C. J., & Cruickshank, R. H. (2012). SPIDER: An R package for the analysis of species identity and evolution, with particular reference to DNA barcoding. *Molecular Ecology Resources*, 12, 562–565.
- Buckup, P. A. (1993). Phylogenetic interrelationships and reductive evolution in Neotropical characidiin fishes (Characiformes, Ostariophysi). *Cladistics*, 9, 305–341.
- Buckup, P. A., Menezes, N. A., & Ghazzi, M. S. (2007). *Catálogo das espécies de peixes de água doce do Brasil*. Rio de Janeiro: Museu Nacional do Rio de Janeiro.
- Burridge, C. P., Craw, D., Fletcher, D., & Waters, J. M. (2008). Geological dates and molecular rates: fish DNA sheds light on time dependency. *Molecular Biology and Evolution*, 25, 624–633.
- de Carvalho, D. C., Oliveira, D. A., Pompeu, P. S., Leal, C. G., Oliveira, C., & Hanner, R. (2011). Deep barcode divergence in Brazilian freshwater fishes: the case of the São Francisco River basin. *Mitochondrial DNA*, 22(suppl 1), 80–86.



- Casatti, L., & Castro, R. M. C. (1998). A fish community of the São Francisco River headwater riffles, southeastern Brazil. *Ichthyological Exploration of Freshwaters*, 9, 229–242.
- Claro-García, A., & Shibatta, O. A. (2013). The fish fauna of streams from the upper rio Tocantins basin, Goiás State, Brazil. *Check List*, 9, 28–33.
- Clement, M., Posada, D., & Crandall, K. A. (2000). TCS: A computer program to estimate gene genealogies. *Molecular Ecology*, 9, 1657–1659.
- Coe, M. T., Latrubesse, E. M., Ferreira, M. E., & Amsler, M. L. (2011). The effects of deforestation and climate variability on the streamflow of the Araguaia River, Brazil. *Biogeochemistry*, 105, 119–131.
- Collins, R. A., Boykin, L. M., Cruickshank, R. H., & Armstrong, K. F. (2012). Barcoding's next top model: An evaluation of nucleotide substitution models for specimen identification. *Methods in Ecology and Evolution*, 3, 457–465.
- Couto, T. B. D. A., & Aquino, P. D. P. U. (2011). Structure and integrity of fish assemblages in streams associated to conservation units in Central Brazil. *Neotropical Ichthyology*, 9, 445–454.
- Díaz, J., Villanova, G. V., Brancolini, F., del Pazo, F., Posner, V. M., Grimberg, A., & Arranz, S. E. (2016). First DNA barcode reference library for the identification of South American freshwater fish from the lower Paraná river. *PLoS One*, 11, e0157419.
- Doña, J., Diaz-Real, J., Mironov, S., Bazaga, P., Serrano, D., & Jovani, R. (2015). DNA barcoding and minibarcoding as a powerful tool for feather mite studies. *Molecular Ecology Resources*, 15, 1216–1225.
- Excoffier, L., Laval, G., & Schneider, S. (2005). Arlequin version 3.0: an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online*, 1, 47–50.
- Fialho, A. P., Oliveira, L. G., Tejerina-Garro, F. L., & Gomes, L. C. (2007). Fish assemblage structure in tributaries of the Meia Ponte River, Goiás, Brazil. *Neotropical Ichthyology*, 5(1), 53–60.
- Françoso, R. D., Brandão, R., Nogueira, C. C., Salmons, Y. B., Machado, R. B., & Colli, G. R. (2015). Habitat loss and the effectiveness of protected areas in the Cerrado Biodiversity Hotspot. *Natureza & Conservação*, 13, 35–40.
- Funk, D. J., & Omland, K. E. (2003). Species-level paraphyly and polyphyly: frequency, causes and consequences, with insights from animal mitochondrial DNA. *Annual Review of Ecology and Systematics*, 34, 397–423.
- Gibson, J., Shokralla, S., Porter, T. M., King, I., van Konynenburg, S., Janzen, D. H., ... Hajibabaei, M. (2014). Simultaneous assessment of the macrobiome and microbiome in a bulk sample of tropical arthropods through DNA metasytematics. *Proceedings of the National Academy of Sciences*, 111, 8007–8012.
- Hajibabaei, M., Janzen, D. H., Burns, J. M., Hallwachs, W., & Hebert, P. D. (2006). DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, 103, 968–971.
- Hebert, P. D., Cywinska, A., & Ball, S. L. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London B*, 270, 313–321.
- Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., & Francis, C. M. (2004). Identification of birds through DNA barcodes. *PLoS Biology*, 2, e312.
- Hoffmann, W. A., & Jackson, R. B. (2000). Vegetation-climate feedbacks in the conversion of tropical savanna to grassland. *Journal of Climate*, 13, 1593–1602.
- Hubert, N., Hanner, R., Holm, E., Mandrak, N. E., Taylor, E., Burrige, M., ... Zhang, J. (2008). Identifying Canadian freshwater fishes through DNA barcodes. *PLoS One*, 3, e2490.
- Junk, W. J., Soares, M. G., & Bayley, M. (2007). Freshwater fishes of the Amazon River basin: their biodiversity, fisheries and habitats. *Aquatic Ecosystem Health & Management*, 10, 153–173.
- Kimura, M. (1980). A simple method of estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution*, 16, 111–120.
- Klink, C. A., & Machado, R. B. (2005). Conservation of the Brazilian Cerrado. *Conservation Biology*, 19, 707–713.
- Kumar, S., Stecher, G., & Tamura, K. (2016). MEGA7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*, 33, 1870–1874.
- Lanfear, R., Calcott, B., Ho, S. Y. W., & Guindon, S. (2012). PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular Biology and Evolution*, 29, 1695–1701.
- Lanfear, R., Calcott, B., Kainer, D., Mayer, C., & Stamatakis, A. (2014). Selecting optimal partitioning schemes for phylogenomic datasets. *BMC Evolutionary Biology*, 14, 82.
- Langeani, F., Castro, R. M. C., Oyakawa, O. T., Shibatta, O. A., Pavanelli, C. S., & Casatti, L. (2007). Diversidade da ictiofauna do Alto Rio Paraná: Composição atual e perspectivas futuras. *Biota Neotropica*, 7, 1–17.
- Langeani, F., Buckup, P. A., Malabarba, L. R., Py-Daniel, L. H. R., Lucena, C. A. S., Rosa, R. S., ... Gomes-Filho, G. (2009). Peixes de água doce. *Congresso Brasileiro de Zoologia (27: 2008: Curitiba)*, 1, 211–230.
- Librado, P., & Rozas, J. (2009). DnaSP v5: A software for comprehensive analysis of DNA polymorphism data. *Bioinformatics*, 25, 1451–1452.
- Lima, F. C. T., & Ribeiro, A. C. (2011). Continental-scale tectonic controls of biogeography and ecology. In J. S. Albert & R. E. Reis (Eds.), *Historical biogeography of neotropical freshwater fishes* (pp. 145–164). Berkeley, CA: University of California Press.
- Magrin, G. O., Marengo, J. A., Boulanger, J.-P., Buckeridge, M. S., Castellanos, E., Poveda, G., ... Vicuña, S. (2014). Central and South America. In V. R. Barros, C. B. Field, D. J. Dokken, M. D. Mastrandrea, K. J. Mach, T. E. Bilir, et al. (Eds.), *Climate change 2014: Impacts, adaptation and vulnerability. Part B: Regional aspects. Contribution of Working Group II to the fifth assessment report of the intergovernmental panel on climate change* (pp. 1499–1566). New York, NY: Cambridge University Press.
- Meier, R., Shiyang, K., Vaidya, G., & Ng, P. K. L. (2006). DNA barcoding and taxonomy in Diptera: A tale of high intraspecific variability and low identification success. *Systematic Biology*, 55, 715–728.
- Meier, R., Zhang, G., & Ali, F. (2008). The use of mean instead of smallest interspecific distances exaggerates the size of the barcoding gap and leads to misidentification. *Systematic Biology*, 57, 809–813.
- Meusnier, I., Singer, G. A., Landry, J. F., Hickey, D. A., Hebert, P. D., & Hajibabaei, M. (2008). A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, 9(1), 214.
- Menezes, N. A., Ribeiro, A. C., Weitzman, S. H., & Torres, R. A. (2008). Biogeography of Glandulocaudinae (Teleostei: Characiformes: Characidae) revisited: Phylogenetic patterns, historical geology and genetic connectivity. *Zootaxa*, 1726, 33–48.
- Mendonça, M. B., & Netto-Ferreira, A. L. (2015). New species of *Characidium* (Characiformes: Crenuchidae) from the rio Tapajós and rio Xingu drainages, Pará, Brazil. *Zootaxa*, 4021, 187–194.
- Meyer, C. P., & Paulay, G. (2005). DNA barcoding: Error rates based on comprehensive sampling. *PLoS Biology*, 3, e422.
- Myers, N., Mittermeier, R. A., Mittermeier, C. G., da Fonseca, G. A., & Kent, J. (2000). Biodiversity hotspots for conservation priorities. *Nature*, 403, 853–858.
- Nogueira, C., Buckup, P. A., Menezes, N. A., Oyakawa, O. T., Kasecker, T. P., Neto, M. B. R., & da Silva, J. M. C. (2010). Restricted-range fishes and the conservation of Brazilian freshwaters. *PLoS One*, 5, e11390.
- Pacheco, A. A., Neves, A. C. O., & Fernandes, G. W. (2018). Uneven conservation efforts compromise Brazil to meet the Target 11 of convention on biological diversity. *Perspectives in Ecology and Conservation*, 16, 43–48.
- Padial, J. M., Miralles, A., De la Riva, I., & Vences, M. (2010). The integrative future of taxonomy. *Frontiers in Zoology*, 7, 16.

- Paz, F. P. C., da Silva Batista, J., & Porto, J. I. R. (2014). DNA barcodes of rosy tetras and allied species (Characiformes: Characidae: Hyphessobrycon) from the Brazilian Amazon basin. *PLoS One*, 9, e98603.
- Pereira, L. H., Hanner, R., Foresti, F., & Oliveira, C. (2013). Can DNA barcoding accurately discriminate megadiverse Neotropical freshwater fish fauna? *BMC Genetics*, 14, 20–34.
- Pereira, L. H. G., Maia, G. M. G., Hanner, R., Foresti, F., & Oliveira, C. (2011). DNA barcodes discriminate freshwater fishes from the Paraíba do Sul River Basin, São Paulo, Brazil. *Mitochondrial DNA*, 22(Suppl 1), 71–79.
- Peterson, B. K., Weber, J. N., Kay, E. H., Fisher, H. S., & Hoekstra, H. E. (2012). Double digest RADseq: An inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. *PLoS One*, 7, e37135.
- Prance, G. T. (2006). Tropical savannas and seasonally dry forests: An introduction. *Journal of Biogeography*, 33, 385–386.
- Puillandre, N., Lambert, A., Brouillet, S., & Achaz, G. (2012). ABGD, automatic barcode gap discovery for primary species delimitation. *Molecular Ecology*, 21, 1864–1877.
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. Retrieved from <http://www.R-project.org/>
- Rambaut, A., Suchard, M. A., Xie, W., & Drummond, A. J. (2013). *Tracer*, version 1.6. Retrieved from <http://tree.bio.ed.ac.uk/software/tracer/>
- Ratnasingham, S., & Hebert, P. D. N. (2007). bold: The Barcode of Life supporting information table system (<http://www.barcodinglife.org>). *Molecular Ecology Notes*, 7, 355–364.
- Ratnasingham, S., & Hebert, P. D. N. (2013). A DNA-based registry for all animal species: the Barcode Index Number (BIN) system. *PLoS One*, 8, e66213.
- Recknagel, H., Kusche, H., Elmer, K. R., & Meyer, A. (2013). Two new endemic species in the Midas cichlid species complex from Nicaraguan crater lakes: *Amphilophus tolteca* and *Amphilophus viridis* (Perciformes, Cichlidae). *aqua*, 19, 207–224.
- Reis, R. E., Kullander, S. O., & Ferraris, C. J., Jr. (2003). *Check list of the freshwater fishes of south and Central America*. Porto Alegre: Edipucrs.
- Ribeiro, M. C. L. B., Perdigão, V. S. J., Rosa, J. W. C., & Santos, M. V. (2008). A singularidade do fenômeno de Águas Emendadas: uma explicação biogeográfica. In F. O. Fonseca (Ed.), *Águas Emendadas* (pp. 294–310). Brasília: Seduma.
- Ronquist, F., Teslenko, M., Van Der Mark, P., Ayres, D. L., Darling, A., Höhna, S., ... Huelsenbeck, J. P. (2012). MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic Biology*, 61, 539–542.
- Rosenberg, N. A. (2007). Statistical tests for taxonomic distinctiveness from observations of monophyly. *Evolution*, 61, 317–323.
- Steinke, D., & Hanner, R. (2011). The FISH-BOL collaborators' protocol. *Mitochondrial DNA*, 22, 10–14.
- Swofford, D. L. (2002). *PAUP\*. Phylogenetic analysis using parsimony (\*and other methods)*. Version 4.
- Tautz, D., Arctander, P., Minelli, A., Thomas, R. H., & Vogler, A. P. (2003). A plea for DNA taxonomy. *Trends in Ecology and Evolution*, 18, 70–74.
- Toffoli, D., Hrbek, T., Araújo, M. L. G. D., Almeida, M. P. D., Charvet-Almeida, P., & Farias, I. P. (2008). A test of the utility of DNA barcoding in the radiation of the freshwater stingray genus *Potamotrygon* (Potamotrygonidae, Myliobatiformes). *Genetics and Molecular Biology*, 31, 324–336.
- Triantafyllidis, A., Bobori, D., Koliamitra, C., Gbandi, E., Mpanti, M., Petriki, O., & Karaiskou, N. (2011). DNA barcoding analysis of fish species diversity in four north Greek lakes. *Mitochondrial DNA*, 22(suppl 1), 37–42.
- Viana, J. P. (1989). *Estrutura da comunidade dos peixes do Ribeirão Santana (Brasília, DF) ao longo de gradientes ambientais*. (MSc Thesis). Universidade de Brasília, Brasília.
- Virgilio, M., Jordaens, K., Breman, F. C., Backeljau, T., & De Meyer, M. (2012). Identifying insects with incomplete DNA barcode libraries, African fruit flies (Diptera: Tephritidae) as a test case. *PLoS One*, 7, e31581.
- Ward, R. D. (2009). DNA barcode divergence among species and genera of birds and fishes. *Molecular Ecology Resources*, 9, 1077–1085.
- Ward, R. D. (2012). FISH-BOL, a case study for DNA barcodes. In R. D. Ward (Ed.), *DNA Barcodes* (pp. 423–439). Totowa, NJ: Humana Press.
- Ward, R. D., Hanner, R., & Hebert, P. D. N. (2009). The campaign to DNA barcode all fishes, FISH-BOL. *Journal of Fish Biology*, 74, 329–356.
- Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R., & Hebert, P. D. (2005). DNA barcoding Australia's fish species. *Philosophical Transactions of the Royal Society B*, 360, 1847–1857.
- Werneck, F. P., Gamble, T., Colli, G. R., Rodrigues, M. T., & Sites, J. W., Jr. (2012). Deep diversification and long-term persistence in the South American 'dry diagonal': Integrating continent-wide phylogeography and distribution modeling of geckos. *Evolution*, 66, 3014–3034.
- Yohannes, O. (2009). *Water resources and inter-riparian relations in the Nile Basin: The search for an integrative discourse*. Stony Brook, NY: SUNY Press.
- Zhan, A., Bailey, S. A., Heath, D. D., & Macisaac, H. J. (2014). Performance comparison of genetic markers for high-throughput sequencing-based biodiversity assessment in complex communities. *Molecular Ecology Resources*, 14, 1049–1059.
- Zhang, A. B., Hao, M. D., Yang, C. Q., & Shi, Z. Y. (2017). BarcodingR: an integrated R package for species identification using DNA barcodes. *Methods in Ecology and Evolution*, 8, 627–634.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Bagley JC, de Aquino Pedro De Podestà Uchôa, Breitman MF, Langeani F, Colli GR. DNA barcode and minibarcode identification of freshwater fishes from Cerrado headwater streams in Central Brazil. *J Fish Biol.* 2019;95:1046–1060. <https://doi.org/10.1111/jfb.14098>