1 Introduction

Biodiversity is threatened by problems arising from human population growth [1,2], and species identification is an important step for the preservation of biodiversity [2,3]. However, species identification is a difficult task, since many controversies have persisted to the present day, ranging from the definition of the species concept to the definition of the threshold for speciation [4]. Many types of methods have been proposed as the best tools for species identification, with genetic tools considered as an alternative approach [5–12].

The integration of research areas for species identification can bring great benefits in biodiversity characterization in regions where such information is lacking, avoiding the loss of endemic species that may perhaps occur in these regions, enabling the creation of conservation areas. Examples of this characterization include the checklist of ichthyofauna from the Ribeira de Iguape River Basin [13–15] and the Coastal Rivers of São Paulo [13,16]. These two water bodies are located in two of the most urbanized and densely populated regions of Brazil, but data on the number of freshwater fish species in both regions are incomplete.

Although molecular tools have provided new insights into several areas of study, such as evolutionary biology, phylogenetic systematics, and population genetics, only in 2003 it was proposed that a short segment of 648 nucleotides from the mitochondrial gene cytochrome c oxidase I (COI) could standardize species identification [6,7]. The methodology called “DNA Barcode” subsequently has been widely used, and many studies have already validated it as a tool for species identification in selected geographical areas [17–28].

However, many questions have been raised about the use of genetic markers for species delimitation. Population genetic parameters such as gene flow and population size and phylogenetic parameters such as the pattern and
timing of lineage diversification are being very actively studied, because the limit of genetic divergence within and between species is a factor that must be considered [25,29–32]. “Phylogenetic analysis is traditionally concerned with estimating relationships among species, while population genetics is concerned with adaptation and understanding the population-level forces that change allele frequencies” Carstens et al.[4]. This brings us to a question: to what extent do genetic differences correspond to individual variation within a population or a divergence between the two species? Thus, for species identification, it is necessary to use different methodologies for data analysis so that the conflicting data can be treated with greater care [4]. More recent studies have used analysis of multispecies coalescent models such as the GMYC [33] for the determination of species. In addition, the Barcode of Life Database provides a tool for analysis that allows us to identify different molecular MOTUs without prior morphological identification, known as the BIN [34].

Based on this information, the aim of this study was to use the methodology of DNA Barcode to identify fish species of the Ribeira de Iguape River Basin and Coastal Rivers of São Paulo, and to evaluate the efficiency of GMYC and BIN analysis in order to determine the number of MOTUs in these areas.

2 Methods

2.1 Sampling

For this study, 805 specimens of freshwater fishes (89 morphospecies) were used, including 490 samples from the Ribeira de Iguape Basin and 315 from Coastal Rivers (Figure 1). All specimens used were identified by specialist PhD Osvaldo Takeshi Oyakawa and identification-keys. After identification, the specimens were deposited in the fish collection of the Laboratory of Biology and Genetic of Fish (LBP), Department of Morphology, Biosciences Institute, UNESP, Botucatu, São Paulo, Brazil. (see Table S1 of the Supplementary Material). Specimen data, including geospatial coordinates of collection sites, are recorded in the publicly accessible BOLD projects titled “Fishes from Brazilian Coastal Rivers; Fishes from Brazilian Coastal Rivers part II” (project codes: FBCR; FBCRB), this projects were merged and analyzed as one.

2.2 DNA Extraction, PCR amplification, and Sequencing

DNA barcoding was performed at the Laboratório de Biologia e Genética de Peixes (LBP), UNESP, Botucatu, Brazil, and at the Canadian Centre for DNA Barcoding.
(CCDB), Canada. Total genomic DNA was isolated from fins or muscle tissue of each specimen using one of two different methods: with a DNeasy Tissue Kit (Qiagen), according to the manufacturer’s instructions; or with a vertebrate lysis buffer containing proteinase K digested overnight at 56°C and subsequent extraction using a membrane-based approach on a Biomek NX (www.pall.com) liquid handling station using AcroPrep96 (www.beckman.com) and 1-mL filter plates with 10 mm PALL glass fiber media [35] according to the CCDB protocol. A segment of approximately 648 bp from the 5’ end of the mitochondrial cytochrome c oxidase subunit I (COI) gene was amplified by the polymerase chain reaction (PCR) using different combinations of primers: FishF1, FishR1, FishF2, FishR2 [36], the Mi13-tailed primer cocktails C_FishF1t1 – C_FishR1t1 and C_VF1L Ft1 – C_VR1L Rt1 [37], and the pair 15698-Asn [38] and H7271-COI [39]. A total volume of 12.5 μL of the PCR mixture included 1.25 μL of 10× buffer (10 mM Tris-HCl + 15 mM MgCl₂), 0.5 μL dNTPs (200 mM each), 0.5 μL each 5 mM primer, 0.05 μL Platinum® Taq Polymerase (Invitrogen), 1 μL template DNA (12 ng), and 8.7 μL ddH2O. The PCR reactions consisted of 30–40 cycles, 30 s at 95°C, 15–30 s at 48–54°C (according to each combination of primers), and 45 s at 72°C. At LBP, all PCR products were first visually identified on a 1% agarose gel and then purified using ExoSap-IT® (USB Corporation) following instructions of the manufacturer. The purified PCR products were sequenced using the Big Dye™ Terminator v 3.1 Cycle Sequencing Ready Reaction Kit (Applied Biosystems) in a final volume of 7.0 μL containing 1.4 μL of template, 0.35 μL of primer (10 μM), 1.05 μL of 5× buffer, 0.7 μL of BigDye mix and water. The cycle sequencing conditions included initial denaturation at 96°C for 2 min followed by 35 cycles of denaturation at 96°C for 45 s, annealing at 50°C for 60 s, and extension at 60°C for 4 min. The PCR sequencing products were purified with EDTA/sodium acetate/alcohol following the protocol suggested in the BigDye™ Terminator v.3.1 Cycle Sequencing kit’s manual (Applied Biosystems) and loaded on an automatic sequencer 3130-Genetic Analyzer capillary sequencer (Applied Biosystems) At the CCDB, PCR products were labeled using the BigDye Terminator v.3.1 Cycle Sequencing Ready Reaction kit (Applied Biosystems) using standard methods [40] and were bidirectionally sequenced using an ABI3730 capillary sequencer. Sequence data, trace files, primer details, and collection localities for specimens are available within the projects FBRCR and FBCRB on BOLD (http://v3.boldsystems.org/). Sequences have also been deposited in GenBank (Accession numbers at Table S1 of the Supplementary Material).

2.3 Sequence analysis

Consensus sequences from forward and reverse strands were obtained using GENEIOUS PRO 5.4.2 [41]. Alignments were generated using MUSCLE [42] under default parameters. After alignment, the matrix was checked visually for any obvious misalignments and potential cases of sequencing errors. A quality control step was included in our workflow to detect contamination, paralogous copies, or pseudogenes. Then, the presence of stop codons was checked using GENEIOUS PRO 5.4.2.

Nucleotide variation and genetic distances were examined using MEGA 6.0 [43]. The best nucleotide evolution models for the COI gene were evaluated using MODELTEST 3.7 [44].

2.4 GMYC molecular species hypothesis

The lognormal relaxed molecular clock tree was estimated using BEAST v1.6.2 [45], since GMYC requires an ultrametric tree, but the tree was not time calibrated. The nucleotide evolutionary model used to estimate the ultrametric tree was the GTR+I+G model with gamma distribution (estimated by the program MODELTEST 3.7). A random tree was used as a starting tree for the Markov chain Monte Carlo searches. Eight chains were run simultaneously for 10,000,000 generations and a tree was sampled every 100th generation. The distribution of log-likelihood scores was examined to determine the stationary phase for each search and to decide if extra runs were required to achieve convergence, using the program Tracer 1.5 [46]. All sampled topologies beneath the asymptote (10,000,000 generations) were discarded as part of a burn-in procedure, and the remaining trees were used to construct a 50% majority-rule consensus tree in TreeAnnotator v1.6.2 [47].

For species delimitation and consequent identification of independent evolutionary units (IUE) we used the GMYC model. We performed the single threshold method with standard parameters (interval = c(0,10)). This analysis was developed in the program R version 3.0.0 [48] with the package “splits” (Species Limits by Threshold Statistics) (http://r-forge.r-project.org/projects/splits) and default parameters.

2.5 Analytical Tools of Barcode of Life Database (BOLD)

The Barcode of Life Database (BOLD) offers analytical tools for identifying and delimiting species. Thus, in order to test the consistency of the data obtained by Bayesian
analysis, two further analyses were performed.

To evaluate the existence of a “barcoding gap” [49], we used the distance matrix generated on the Barcode of Life Database (BOLD) (available on http://www.boldsystems.org), performed under a K2P model and preliminary alignment with MUSCLE algorithm.

Beyond that, the BIN was used to estimate the number of species directly from the barcode records (the full BIN database is available at BOLD system), following the conditions specified in research performed by Ratnasingham and Herbert [34].

3 Results

We analyzed 805 samples from 89 morphologically distinct groups (morphospecies) belonging to 64 genera, 24 families, and 8 orders. The number of specimens analyzed ranged from 1 to 56 per morphospecies (mean = 9). The consensus sequences of COI from every sample have a mean size of 654 bp (minimum 600 pb) with no stop codons, insertions, or deletions detected. After alignment and editing a matrix was obtained with 600 characters and a nucleotide composition of 24.12% adenine, 27.28% cytosine, 18.2% guanine, and 30.4% thymine. The MODELTEST analyses select the GTR + I + G as best nucleotide evolution model for the dataset.

The mean Kimura-2-Parameter (K2P) genetic divergence ranged from 0% to 8.7% (mean = 0.8%) for intra-specific comparisons and from 0% to 24.5% (mean = 10.9%) for congeneric comparisons, establishing a barcode gap of about almost 14 times between congeneric and intra-specific variation. The analyses of the distribution of K2P divergence values showed that 88.7% of the intra-specific comparisons were less than 2%; however, 14.6% of the divergence values between congeners were also less than 2%. The pairwise distance comparison showed that the range of genetic divergence between distinct morphospecies overlaps the variation within some morphospecies (Figure 2), but the average of genetic divergence found between the nearest neighbor distance (NND) is almost 14 times the average found within each morphospecies (synthesis of divergence genetic comparison shown in Figure 2).

Although there were 89 previously recognized morphospecies, the molecular approaches indicate that this number may be underestimated. The GMYC analysis identified an intra/inter-specific threshold time of −0.01112289 and recognized 99 (confidence interval: 95–102) independent evolution units (interpreted as distinct species or OTUs [33]), while the BIN approach indicated the existence of 104 BIN (interpreted as distinct species or MOTUs) [34] (Figure 3). By cross-tabulating the results of all approaches, we identified 109 MOTUs by at least one methodology. In most of the cases (89 MOTUs) the genetic approaches are in concordance with morphological identification.

The present analyses show that the different *Pimelodella* morphospecies cannot be identified by our genetic data (Figure 3) (genetic divergence ranging from 0.3 to 1.1% and an average of 0.5% between the two morphospecies). Other groups that show similar patterns are *Gymnotus* (ranging from 0 to 5.4%, average 4.2%) and *Phalloceros* (ranging from 1.7 to 9.7%, average 6.1%), whose morphospecies could not be distinguished by genetic analyses. However, in contrast to the *Pimelodella*

![Figure 2](image-url) A) Histogram plotting the distribution of the mean intra-specific distances for each species. B) Histogram plotting the distribution of the nearest neighbor distances for each species. The distribution details are summarized below each histogram.
Figure 3 Part 1: Summarized Bayesian tree of the all samples obtained with COI data. The asterisks in the node branches represent the posterior probability greater than 95%. The taxa names were inferred by morphological identification. The vertical columns represent the status of the identification of the MOTUs, where the left side represents the GMYC analysis identifications, while the right side represents the BIN analysis identification. White regions in the columns represent MOTUs identified by all methodologies. Black marks represent MOTUs with conflicts in the identifications by at least one methodology. Red marks represent the clusters with morphospecies that are no reciprocally monophyletic (these clusters are shown in detail in Figure 4).
**Figure 3 part 2**

Legend is the same of the Figure 3 part 1.
The existence of more than one MOTU within each morphospecies.

4 Discussion

From the data obtained in this study, it is possible to confirm the importance of interdisciplinary approaches in order to obtain more accurate data on species identification. A preliminary morphological analysis identified 89 species in our database. Comparison with the genetic data revealed that the number of species varied results, these two groups have well-structured clusters, although these clusters are not composed of only one morphospecies (Figure 4). However, some morphospecies such Symbranchus marmoratus (genetic divergence ranging from 0 to 5.5%, average 3.4%), Listrura camposi (ranging from 0 to 16.1%, average 8.7%), Characidium lanei (ranging from 0 to 4.6%, average 2.8%), Cyphochlarax santacatarinai (ranging from 0 to 179%, average 5.8%), and Mimagoniates microlepis (ranging from 0 to 5.1%, average 1.6%) have two or more structured clusters (Figure 5), and in these cases, genetics analyses revealed

Figure 4 Detail of the clusters with morphospecies not monophyletic (with red marks in Figure 3). Asterisks represent nodes with a posterior probability higher than 95%. (A) Details of the Gymnotus clade. (B) Details of the Pimelodella clade. (C) Details of the Phalloceros clade.

Note: In B, the scale bars represent 0.001 (or 0.1%) of GTR genetic divergence, while in A and C, the scale bar represents 0.01 (or 1%).
To compare the data obtained in this study with the previous studies on molecular species identification by DNA Barcode, we performed a genetic analysis based on the K2P model. Thus, we estimated the average intra/inter-specific genetic divergence of 0.81% and 10.94%, respectively, where divergence between species is 13.5-fold greater than divergence within them. Other studies using the technique of DNA barcode for species identification according to the geographical region show lower values to those found in the present study, as follows: Valdez-Moreno et al. [18] and Pereira et al. [28], who studied neotropical fishes from Central Mexico and South America, obtained intra/inter-specific genetic distance relationships of 0.45% and 0.36%, respectively.

Although DNA barcode is accurate in identifying most of the studied species, groups with a complex taxonomic
history present discordance in MOTUs delimitation [25,50] because some species are not distinguished by genetic approaches due to lack of reciprocal monophyly between them. All studied groups that we found problems to delimit the species and show overlap between intraspecific and interspecific values are present at the figure 3 (red marks and black marks).

In the case of Gymnotus and Phalloceros, sample identification at the species level is a difficult task because of the similarity in morphological characteristics, especially in juvenile forms [15]. However, these two genera have well-defined internal clusters in our analyses, suggesting that a regional taxonomic revision is still necessary to check the problem of species delimitation in these genera.

However, this explanation cannot be used to account for the non-reciprocal monophyly observed in Pimelodella. The species P. transitoria and P. kronei are apparently easily distinguished morphologically because P. kronei show characteristics of cave fish (little pigmentation and no eyes), while P. transitoria is a typical surface fish [15]. To better investigate the relationship between these two species, other types of markers are necessary.

The different methodologies of analysis also indicate another divergence. In some groups, such Synbranchus marmoratus, Cyphocharax santacatarinae, Characidium lanei, Listrura camposi and Mimagoniates microlepis, we observed a genetic divergence between haplotypes, and both genetic analyses (BIN and GMYC) suggest the existence of more than one MOTUs in the same morphospecies. This result is similar to that found in other studies of fish barcodes [25,28,50]. According to De Queiroz [51], sometimes when a speciation process occurs, morphologic differentiations take time to become perceptible, but with evolution, mutations can accumulate and genetic discriminations is possible without morphological differentiation; these entities are also known as cryptic species. Evolution can act in the morphology of the species for the maintenance of body shape and the result of this selective pressures is the maintaining of the shape, while genetically the lineages can have a high rate of evolution for COI, which could give an illusion that the time of divergence is greater than it really is [11], and the reverse is also true, once COI may evolve slowly among different lineages, as well as, groups can evolve faster morphologically.

Some of the morphospecies in this study are found in all independent drainages (Ribeira basin and all small drainages along the coast that were sampled); among these, some have a high genetic divergence, while others have a low genetic divergence. This may be a reflection of a complex history of evolution of the basins involved in this study [16,52–54].

5 Conclusion

The last ichthyofauna checklist of the Ribeira de Iguape Basin and coastal rivers from São Paulo identified 96 species [52], and in the present study, we found a similar value of 89 to 109 MOTUs, corroborating the hypothesis that DNA barcode can be an effective tool for species identification. However, the present data indicate the existence of unnamed species and species complexes that require further investigation.

Acknowledgments: The authors wish to thank Oswaldo T. Oyakawa for the help with the morphological identifications, Renato Devidé, Bruno F. Melo, and Fábio F. Roxo for their help during the sampling expeditions, and BIO (Biodiversity Institute of Ontario) at the University of Guelph for the exchange program and sequencing facilities.

This research was financially supported by Brazilian agencies such as FAPESP (Fundação de Amparo a Pesquisa do Estado de São Paulo) and CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico).

References


Supplemental Material: The online version of this article (DOI: 10.1515/dna-2015-0015) offers supplementary material.