Molecular systematics of the Cyprinoidea (Teleostei: Cypriniformes), the world’s largest clade of freshwater fishes: Further evidence from six nuclear genes

Wei-Jen Chen *, Richard L. Mayden
Department of Biology, Saint Louis University, 3507 Laclede Ave., St. Louis, MO 63103, USA

1. Introduction

With over 210 genera and 2010 described species, the family Cyprinidae is currently the largest family of freshwater fishes (Nelson, 2006). Over the years, the Cyprinidae has been divided into different “groupings” for either taxonomic convenience or to represent presumed natural groups; usually these groupings have been recognized at or below the level of subfamily (Cavender and Coburn, 1992; Howes, 1991; Nelson, 2006). Howes (1991) recognized seven such subgroupings of the Cyprinidae, including the Alburninae, Cyprininae, Rasborinae, Cultrinae, Acheilognathinae, Tincinae, Leuciscinae, and Gobioninae. Cavender and Coburn (1992) recognized two, the Cyprininae and Leuciscinae, the former including those cyprinids referred to as barbins, labeonins and cyprinins, and the later including those referred to as tincins, rasborins, gobionins, acheilognathins, cultrins, xenocyprins, leucisins and phoxinins. The genus Psilorhynchus (a grouping of small, ventrally flattened fishes adapted for benthic life in fast flowing water) is either considered to be the sole member of the family Psilorhynchidae (Conway and Mayden, 2007; Nelson, 2006; Ramaswami, 1952) following Hora (1925) or as the sole member of the cyprinid subfamily Psilorhynchinae (Chen, 1981; Nelson, 1994).

Previous systematic analyses investigating monophyly and inter-relationships of the Cyprinidae have focused largely on morphology or mitochondrial gene/genome sequences. Chen et al. (1984) was the first such study to propose a “phylogenetic” hypotheses of cyprinid inter-relationships based on morphological data. Later Cavender and Coburn (1992) reanalyzed the data matrix of Chen et al. (1984), recovering a tree of equal length but of a different topology to that recovered by Chen et al. (1984) (Fig. 1B). Cavender and Coburn (1992) also proposed an alternative phylogeny for the Cyprinidae based on the analysis of their own 47 morphological characters (Fig. 1A). Despite these early morphological phylogenetic investigations of the Cyprinidae, some uncertainty regarding the basal lineage of cyprinds and the placement of the enigmatic genus Tinca remains (Fig. 1).

Recent systematic investigations of the Cyprinidae have utilized a molecular phylogenetic approach, with mitochondrial sequence data being most readily utilized (e.g., Cunha et al., 2002; Gilles et al., 2001; He et al., 2008a; Liu and Chen, 2003; Okazaki et al., 2001; Saitoh et al., 2006; Simons et al., 2003; Xiao et al., 2001; Zardoya and Doadrio, 1998). Although the taxonomic sampling of these studies was limited to certain subgroupings or to species from geographic regions of interests to the authors each of these investigations provided valuable insight into the evolution of these morphologically diverse fishes. Two studies presented their hypotheses with the samplings covering approximately all cyprinid subfamilies (Gilles et al., 2001; Liu and Chen, 2003). Both of these studies resolved Tinca as more closely related to members
of the subfamilies Acheilognathinae, Gobioninae, and Cultrinae than to members of the Rasborinae or Cyprininae (Fig. 1C and D).

More recently, using whole mitochondrial genomes, Saitoh et al.'s (2006) provided a robust phylogenetic hypothesis for the main cypriniform lineages for the first time. In their hypothesis, two reciprocal monophyletic groups were resolved: Cobitoidea and Cyprinioidea (or Cypriniidae). Their hypothetical relationships of main cyprinid clades were summarized in Fig. 1E. Even if a larger number of characters (e.g., complete mitochondrial genome data) was used, it has limitations to reaching a resolution of certain relationships within the family and some of critical and/or ambiguous classified taxa such as Psilorhynchus and Leptobarbus were missing in the analysis (Fig. 1). Moreover, the resulting hypotheses require further testing with more intensive taxonomic sampling and additional independent molecular markers.

Herein, we employ a multiple nuclear gene (or phylogenomic) approach with a diverse set of 49 cyprinid species to infer evolutionary relationships of the major clades within the Cypriniformes. We placed particular emphasis on resolving the phylogenetic position of the enigmatic genera Psilorhynchus (the stone carps), Tinca (the tench), and Leptobarbus (the mad barb or sultan fish) in relation to other cyprinid fishes. These enigmatic genera have received little attention from molecular systematists and have been difficult to place within the current cypriniform classification, likely because of morphological divergence. DNA sequence data were generated from six nuclear gene loci (tRNA, Rhodopsin, IRBP, EGR1, EGR2B, and EGR3). These gene markers have recently been shown to be phylogenetically informative in reconstructing the phylogenetic relationships of ray-finned fishes, particularly among fishes of the order Cypriniformes (Chen et al., 2008).

2. Materials and methods

2.1. DNA data collection

A total of 54 samples were included for investigation. The analytical dataset was composed of DNA sequences of 6 targeted nuclear loci obtained from 2 Psilorhynchus species, Tinca tinca, Leptobarbus hoevenii, 45 other diverse specimens of cyprinids from all recognized subfamily groups, and 5 outgroups from the superfamilies Cobitoidea. Several sequences used in this study have been previously described in Mayden et al. (2008) and Chen et al. (2008). Methods for collecting new DNA data from the specimens and/or gene loci followed the procedures outlined in Chen et al. (2008). The GenBank accession numbers of corresponding gene sequences used in this study are listed in the Table 1.

2.2. Phylogenetic analysis

Phylogenetic analyses were based on a partitioned Maximum Likelihood (ML) method and partitioned Bayesian approach (BA) for two different types of character matrices as implemented in the parallel version of RAxML (version 7.0.4) (Stamatakis, 2006) and MrBayes (version 3.1.1) (Huelsenbeck and Ronquist, 2001), respectively. The first matrix was composed of all available characters without employing a particular weighting scheme. As phylogenetic analyses of protein-coding genes can be biased from homoplasy at third codon positions due to multiple substitutions in transitions (Saitoh et al., 2006) and/or because of base composition biases across taxa (Chen et al., 2003; Lockhart et al., 1994), a second matrix (partial RY-coding matrix) was prepared according to the results obtained from absolute saturation tests (Philippe et al., 1994) and from χ² tests of base composition stationarity performed using PAUP*–version 4.0b10 (Swofford, 2002). As outlined in our previous study, no clear saturation plateau on substitutions in transitions at the third codon position of six nuclear genes used here was detected by comparing the sequences recovering all main lineages of cypriniform species (see Fig. 2 in Chen et al., 2008). However, the tests of base composition stationarity revealed that the Rhodopsin dataset exhibits significant base composition bias across taxa when analyzed using variable sites only and the sites at third codon position for the tests. Thus, we compiled an operational dataset in which the nucleotides A and G and the nucleotides T and C at the third codon position of Rhodopsin were converted into purine (R) and pyrimidine (Y), respectively.

Search for optimal ML trees and Bayesian analyses were performed by a high performance cluster computing facility (with 32 nodes) located at Saint Louis University. We used mixed model analysis, which allows an individual model of nucleotide substitution to be estimated independently from each partition for the analyses. Partitions were assigned with respect to the codon positions of each nuclear protein-coding gene. Likelihood ratio tests (Goldman, 1993), as implemented in MrModeltest 2.2 (Nylander, 2004), were used to choose models for each gene coding position
in Partitioned BA. The parameters for running MrBayes were set as follows: "iset nst = 6" (GTR), "iset nst = 2" (HKY), "iset nst = 1" (FB1), "rates = invgamma" (G + I), or "rates = gamma" (G), "unlink" (unlinking of model parameters across data partitions), and "priet ratepr = variable" (rate multiplier variable across data partitions).

Two independent Bayesian searches were conducted for each data set. Four independent MCMC chains were performed with 3,000,000 replicates, sampling one tree per 100 replicates for each run. The distribution of log likelihood scores was examined to determine stationarity for each search and to decide if extra runs were required to achieve convergence in log likelihoods among runs or searches. We discarded initial trees with non-stationary log likelihood values as part of a burn-in procedure, and combined the remaining trees that resulted in convergent log likelihood scores from both independent searches. These trees were used to construct a 50% majority rule consensus tree. For ML search with the mixed model of nucleotide substitution we used a GTR + G + I model (with four discrete rate categories) for each partition because RAxML only provides GTR related models (GTR + G + I and GTR + CAT approximation) of rate heterogeneity for nucleotide data (Stamatakis, 2006). ML tree search was conducted by performing 100 distinct runs using the default algorithm of the program from complete random trees (-d option) as a starting tree for each run. The final tree was determined by a compar-
ison of likelihood scores under GTR + G + I model among suboptimal trees obtained per run.

Nodal support was assessed using the bootstrap (BS) procedure (Felsenstein, 1985) under Maximum Parsimony (MP) and Maximum Likelihood (ML) criterion, based on 1000 pseudo-replicates and the resulting a posteriori probabilities from partitioned Bayesian analysis >=0.95 and resulting MP bootstraps > than 80%. The targeted taxa in this study, *Psilorhynchus*, *Tinca* and *Leptobarbus* are marked in bold. The bars and symbols on the right indicate traditional classification of taxa in Cypriniformes at family/subfamily level. A suggested revised classification, based on robust molecular evidence from this study, is revealed by gray shadow rectangles on the topology.

Fig. 2. Phylogenetic tree depicting relationships of the major clades resolved within the family Cyprinidae (or the superfamily Cyprinoidea). Relationships were obtained using partitioned ML analysis of 5733 aligned nucleotides from six nuclear gene loci. ML score of the tree is ~53906.500279. Branch lengths are proportional to inferred character substitutions under GTR + G + I model. Numbers on branches are ML bootstrap values; those below 50% are not shown. Bold branches on topologies indicate statistically robust nodes with a posteriori probabilities from partitioned Bayesian analysis >=0.95 and resulting MP bootstraps > than 80%. The targeted taxa in this study, *Psilorhynchus*, *Tinca* and *Leptobarbus* are marked in bold. The bars and symbols on the right indicate traditional classification of taxa in Cypriniformes at family/subfamily level. A suggesting revised classification, based on robust molecular evidence from this study, is revealed by gray shadow rectangles on the topology.
3. Results and discussion

3.1. Characteristics of sequence data and inferred phylogenetic tree

A total of 5733 bp were aligned for the exon regions of six nuclear genes for 54 taxa (including five outgroups) sampled in this study. The length of aligned sequences from each locus was 1497 bp (RAG1), 819 bp (RH), 849 bp (IRBP), 846 bp (EGR1), 816 bp (EGR2B), and 906 bp (EGR3). Of the 3573 nucleotides, 2497 were variable sites in which 1996 were parsimony informative. The second or partial RY-coding matrix presented 2407 variable sites in which 1901 were parsimony informative. Relationships of taxa derived from partitioned ML and Bayesian analyses of DNA sequences based on matrix 1 and 2 were nearly identical with slightly differences in relationships where nodal supports are weak; only the ML tree derived from the second (partial RY-coding) matrix is presented herein (Fig. 2). As shown, most of resulting clades were highly supported by partitioned MLBS, MPBS and a posteriori probabilities from partitioned BA (Fig. 2). Accordingly, 10 fully resolved clades or major lineages (as represented by gray shadow rectangles on the topology of Fig. 2) have emerged from the analyses in the present dataset.

3.2. Phylogenetic relationships of the major clades of cyprinid fishes

In all resulting phylogenies (all analyses), Cyprinidae is revealed as a paraphyletic group with respect to the Psilorhynchinae. Within the Cyprinidae except for Rasborinae, the currently recognized cyprinid subfamilies were found to represent monophyletic groupings with strong nodal supports. As shown in ML tree with partial RY-coding analysis (Fig. 2), the Cyprininae and Psilorhynchinae are sister-groups to each other. These two clades together form the basal sister-group to the other cyprinid taxa shown in the tree. Species currently placed within the Rasborinae appear to have five distinct origins among cyprinids. Among 19 rasborine species sampled in this study, 14 species group together in a robust clade with strong nodal supports. As shown in ML tree with partial RY-coding analysis (Fig. 2), the Cyprininae and Psilorhynchinae are sister-groups to each other. These two clades together form the basal sister-group to the other cyprinid taxa shown in the tree.

Species currently placed within the Rasborinae appear to have five distinct origins among cyprinids. Among 19 rasborine species sampled in this study, 14 species group together in a robust clade or Rasborinae-Z clade sensu Conway et al. (2008). This rasborine clade can be subdivided into three strongly supported subgroups. The first subgroup includes species of Luciosoma, Barilius and Aspidorhyncha, which forms the sister-group to the subgroup containing species of Rasbora, Trigonostigma and Horadandia in the partial RY-coding ML analysis (Fig. 2). The remaining subgroup, which includes species of Esomus, Danionella, Microrasbora, Devario and Danio, forms the sister-group to a clade composed of two other subgroups. In equal-weighting ML and BA analyses and partial RY-coding BA analysis, the first subgroup is resolved as the most basal lineage within the rasborine clade (not shown). This result corroborates to the findings from a recent study investigating inter-relationships among 31 rasborine taxa with RAG1 sequence data (Conway et al., 2008) but none of the studies with involved analyses resolve confidently the inter-relationships among these three mentioned rasborine subgroups in terms of statistical nodal supports. Otherwise, most of intra-relationships within the subgroups are well resolved in this study (Fig. 2).

Four of five remaining “rasborine” taxa sampled in this study (Opsarichthys, Zaccho, Macrochirichthys, and Aphyocypris) appear in three different placements in the tree and are more closely related to members of the Cultrinae and 2 other cyprinids with uncertain classification (Yosshianicus and Paralaubuca). We refer to this well-supported monophyletic group as the “cultrine” clade. These results, indicating a closer evolutionary affinity among certain rasborine species and cultrine species, are consistent with the findings of other previous phylogenetic analyses based on either mitochondriald or nuclear DNA sequence data (e.g. Saitoh et al., 2006; Conway et al., 2008; He et al., 2008b; Mayden et al., 2008).

The last rasborine species sampled in this study, Tanichthys albonubes, is nested within the terminal clade of the tree and separate from all other rasbornes. Members from this clade represent many species endemic to Eurasia and North America from the cyprinid subfamilies Acheilognathinae, Gobioninae, Tincinae, and Leuciscinae (Fig. 2). Within the terminal clade, several subgroups were supported as monophyletic, notably at subfamily level. In all analyses, a sister-group relationship between the Leuciscinae and Gobioninae was recovered, but received only weak nodal support. This relationship is supported by previous molecular studies (Gilles et al., 2001; Liu and Chen, 2003) (Fig. 1), and in most of the resulting phylogenies based on varied analytical methods and datasets (four nuclear loci and whole mt-genomic data) in a recent study (Mayden et al., 2009). Finally, another taxon of our interest, Leptobarbus, forms the sister-group to the cultrine clade plus the terminal clade of cyprinids described above. This relationship is strongly supported (100% for MLBS, MPBS, and a posteriori probabilities) (Fig. 2).

Overall, most of the phylogenetic relationships among cyprinid fishes presented here are well resolved using the DNA data from six nuclear loci (5733 bp). The results presented are largely congruent with the resulting cypriniform phylogeny using whole-genome data (14,563 bp) (Saitoh et al., 2006). All of these steady molecular evidences currently established are challenging the morphological hypotheses and the classification of this group requires a further revision (see below: Section 3.4).

3.3. Systematics of Tinca, Psilorhynchus, and Leptobarbus

The systematic status of the genus Tinca, and Psilorhynchus is historically chaotic. As Howes (1991) stated, the inclusion of monotypic genus Tinca in any of the cyprid subgroups is a taxonomic problem. For instance, two of the available morphological studies attempting to resolve the phylogenetic placement of Tinca among other cyprinids disagreed with each other (Cavender and Coburn, 1992; Chen et al., 1984) (Fig. 1A and B). Accordingly, erection of the subfamily Tincinae for Tinca became an optimal solution, yet its relationships remain uncertain. Early molecular hypotheses derived from mitochondrial sequence data rejected the morphological hypotheses and showed a closer evolutionary affinity of Tinca with achariognathines, gobionines, leuciscines, and cultrines (Fig. 1C and D). Recent mt-genomic analysis identifies the phylogenetic position of the Tincinae, which should appear to be the sister-taxon to the Leuciscinae (Fig. 1E). Our resulting phylogeny, cannot further confirm this particular molecular hypothesis, as the corresponding nodes for those concerned relationships are weakly supported (Fig. 2). Nonetheless the hypothesis (as shown in Saitoh et al., 2006; Liu and Chen, 2003) (Fig. 1D and E) implicating that cultrine taxa have closely affinity to the Tincinae and to the remaining cyprinids (excluding cyprinines and “rasbornes”) is less likely. Indeed, we resolve that Tinca is a member of the terminal clade of cyprinids (Fig. 1F; Fig. 2), the monophyly of which is highly supported.

Regarding the placement of the genus Psilorhynchus within the Cypriniformes, over the last two centuries this genus has been placed with different loach families (either Balitoridae or Cobitidae) or within the Cyprinidae (see Conway and Mayden, 2007; Šlechtová et al., 2007). The only morphological hypothesis applying phylogenetic analysis with relevant taxa from all families of the Cypriniformes showed that Psilorhynchus is the sister-taxon to a clade including cobitids and balitorids (Conway and Mayden 2007), Saitoh et al. (2006) did not include this taxon in their analysis. An alternative study using whole mt-genomic data with 17 representatives from the order Cypriniformes He et al. (2008a) identified Psilorhynchus and the Cyprinidae as sister-group to each other. Our present study confirms this hypothesis.
Finally, prior to our study, no systematic study, inclusive of the genus _Leptobarbus_, had been conducted. _Leptobarbus_ was included as member of the “_Damioninae_ (possibly following _Gosline_, 1975)” in the generic distribution list of cyprinids occurring in South East Asia by _Rainboth_ (1991). In this study, we discover the systematic status of the genus _Leptobarbus_ to be of intermediate position in the cyprinid tree (Fig. 2). Four described species are presented in the genus. Those are species native to South Asia from Thailand to Sumatra and Borneo. They may reach up to about 60 cm long. Interestingly, this huge taxon is the sister-group of an extremely diverse group comprising a large number of cyprinid species with highly diverse body shapes and sizes, occurring in different ecosystems.

3.4. Concluding remarks with taxonomy implication

According to the solid molecular evidence presented here, which supports the existence of ten monophyletic groups of cyprinids (or cyprinoid) fishes, we tentatively suggest the following revisions to the current cypriniform classification. Five of the subfamilies of the Cyprinidae (Cyprinae, Acheilognathinae, Tincinae, Leuciscinae, and Gobioninae) (Howes, 1991; Nelson, 2006) that have been previously recognized and widely believed to represent monophyletic groupings should be elevated from subfamily status to family status under the superfamily Cyprinioidea. The family group name _Psilorhynchidae_, which has been previously accepted by numerous authors (Conway and Mayden, 2007; Nelson, 2006; Ramaswami, 1952) should be retained. Erection of new cyprinid families (Leptobarbiidae and Tanichthyidae) for two distinct lineages revealed from our analyses containing the species from _Leptobarbus_ and _Tanichthys_, respectively is recommended. We would also need to give the relevant family names for the clades discovered in this study. There are two clades (rashorine and cultrine clades) containing, respectively, major rashorine species and cultrine species plus their putative allies such as _Opsariichthys_, _Zacco_, _Paralabuca_, _Macroichthys_, _Yaoshanichthys_, _Aphyocypris_, _taxa from “_Xenocyprinae_”, and _taxa from “_Squaliobarbinae_” that are identified in this study (Fig. 2) and in some other molecular studies (He et al., 2008b; Liu and Chen, 2003; Mayden et al., 2008; Saitoh et al., 2006). Finally, ten families herein are suggested to be recognized for the Cyprinioidea with respect to seven families for its reciprocal superfamily Cobitoidea (Slechtová et al., 2007).

Acknowledgments

We greatly thank Kevin L. Conway for providing valuable comments for this work and for manuscript-improvement. We thank Maski Miya, Tetsuya Sado, and Shunping He for the specimens. This research is part of an ongoing international Tree of Life initiative on the order Cypriniformes USA NSF awards EF 0431326.

References


References


