Gene 423 (2008) 125–134

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Phylogenetic utility of two existing and four novel nuclear gene loci in reconstructing Tree of Life of ray-finned fishes: The order Cypriniformes (Ostariophysi) as a case study

Wei-Jen Chen [a,*], Masaki Miya [b], Kenji Saitoh [c], Richard L. Mayden [a]

[a] Department of Biology, Saint Louis University, 3507 Laclede Avenue, St. Louis, MO 63103-2010, USA
[b] Department of Zoology, Natural History Museum and Institute, Chiba, 955-2 Aoba-cho, Chuo-ku, Chiba 260-8682, Japan
[c] Tohoku National Fisheries Research Institute, 3-27-5 Shinhama, Shiogama, Miyagi 985-0001, Japan

ARTICLE INFO

Article history:
Received 23 January 2008
Received in revised form 17 June 2008
Accepted 17 July 2008
Available online 25 July 2008

Keywords:
Nuclear gene
Phylogenomic approach
PCR primers
Teleostei
Polyploidy
Gene duplication

ABSTRACT

After the completion of several entire genome projects and a remarkable increase in public genetic databases in the recent years the results of post-genomic analyses can facilitate a better understanding of the genomic evolution underlying the diversity of organisms and the complexity of gene function. This influx of genomic information and resources is also beneficial to the discipline of systematic biology. In this paper, we describe a set of 6 previous and 22 new PCR/sequencing primers for RAG1, Rhodopsin and four novel nuclear markers from IRBP, EGR1, EGR2B and EGR3 that we developed through an approach making use of public genetic/genomic data mining for one of the ongoing tree of life projects aimed at understanding the evolutionary relationships of the planet's largest clade of freshwater fishes — the Cypriniformes. The primers and laboratory protocols presented here were successfully tested in 33 species comprising all cypriniform family and subfamily groups. Phylogenetic performance of each gene, as well as their implications in the investigation of the evolution of cypriniform fishes were assessed and discussed.

1. Introduction

Advances in molecular biology have lead to a great accumulation of genetic resources for vastly different research fields. These molecular tools have galvanized a move in systematic biology to analyze multiple independent gene loci, expanding to the genomic scale (i.e., "phylogenomics"), and this has become a practical approach for molecular systematics (Chen et al., 2004; Philippe et al., 2005). In adopting the phylogenomic approach our aspiration is to increase the accuracy of inferences by reducing stochastic errors with increasing sample size (number of independent loci) and ultimately gaining a better representation of the whole genome (Cummings et al., 1995; Chen et al., 2004). In recent years this ever increasing desirable approach to assembling the tree of life of living organisms has brought us closer to channeling efforts of Darwin's dream into a reality (Delsuc et al., 2005). Despite considerable effort across the planet towards developing a tree of life, there remain two important impediments to accomplishing this goal. First, computational needs remain somewhat problematic in dealing with the increasing number of taxa and character data. Second, for many taxonomic groups our efforts to reconstruct evolutionary relationships continue to be significantly hindered by a lack of suitable nuclear markers and/or specific PCR primers. While the mitochondrial and plastid genomes serve as fundamental sources of important evolutionary and phylogenetic information, a tremendous amount of data from the nuclear genome is not really or readily available for systematic and evolutionary studies. One such example includes the fishes of the order Cypriniformes, one of the largest groups of ray-finned fishes. Cypriniformes, contains many culturally, economically (e.g., carps) and scientifically important species (e.g., the model organism species *Danio rerio*, zebrafish), and is the planet's largest monophyletic group of freshwater fishes, with over 400 genera and 3000 recognized species (estimates to 5000) species native to Asia, Europe, Africa, and North America. Because of this diversity and the diversity in their morphologies, ecologies, physiologies, distributions, and other life history aspects, research outcomes from systematic and evolutionary studies of this group have tremendous potential for complementary and valuable information in

Abbreviations: BP, Bootstrap proportion; bp, base pair; CI, Consistency index; EGR, early growth response; FSGD, fish-specific genome duplication; IRBP, interphotoreceptor retinoid-binding protein; MP, Maximum Parsimony; mt, mitochondrial; NJ, Neighbor-Joining; RAG, recombination-activating gene; RI, Retention index; TBR, tree bisection-reconnection.

* Corresponding author. Fax: +1 314 977 3658.
E-mail address: wjchen.actinops@gmail.com (W.-J. Chen).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.07.016

comparative biology (Mabee et al., 2007), which may be beneficial for disease control (in human and in fishes), conservation and aquaculture (Mayden et al., 2007; Schilling and Webb, 2007).

With the availability of 'universal' PCR primers and standard laboratory protocols (Kocher et al., 1989; Palumbi, 1996; Miya and Nishida, 1999; Miya and Nishida, 2000; Miya et al., 2006), recent efforts in molecular systematics of the Cypriniformes have focused on the use of mitochondrial DNA sequences (Simons and Mayden, 1998; Gilles et al., 2001; Cunha et al., 2002; Durand et al., 2002; Tang et al., 2006) or sequences from whole mitochondrial genomes (Saitoh et al., 2006; He et al., 2008a). However, the exclusive use of mitochondrial genes as markers for phylogenetic reconstruction may be problematic because of inherent attributes associated with these genes or the genomes such as hybridization or introgression, independence of genes, and maternally inherited genomes. Therefore, it is possible that a resulting "gene tree" or "phylogeny" based on mitochondrial DNA sequences may not reflect a "species tree". Analyses from "separate characters sets" from the nuclear genome provides another, alternative opportunity for assessing reliability of phylogenetic hypotheses and for elucidating the evolutionary history of organisms based on the comparison of gene trees derived from independent gene loci (Chen et al., 2003). Unfortunately, only a few molecular investigations studying relationships of cypriniform fishes have been conducted using nuclear gene sequence data. This limitation exists simply because only a few nuclear genes that provide consistent and reliable results are available today. Most the studies have relied on sequences from a single nuclear gene dataset ((Šlechtová et al., 2007) [RAG1]; (Wang et al., 2007) [RAG2]; (He et al., 2008b) [S7]). To our knowledge, only two very recent studies have been based on a 'multi'-locus approach using sequences from a combination of mt-DNA and two nuclear genes (Mayden et al., 2007; Perdices et al., 2008). Regrettably, the number of represented gene sequences from nuclear loci in these papers is significantly smaller by comparison with the represented sequences from mt-genes and possible nuclear loci (see Table 1 in Mayden et al., 2007). This discrepancy is likely due to failures in amplification of nuclear gene fragments using previously published primers of Rhodopsin and RAG1 for ray-finned fishes (Chen et al., 2003; López et al., 2004). The use of primers for reliable alternative nuclear gene loci, or eventually developing new cypriniform-specific primers (and other specific or generalized primers for teleost taxa), is critically important to ensure the accomplishment of large international-scale projects focused on assembling the tree of life initiative (ATOL) and, more specifically, in fish systematics the Cypriniformes Tree of Life initiative (CToL) (www.cypriniformes.org) underway at our laboratories and others around the world.

Herein, we focus on the following objectives: (1) developing additional reliable nuclear markers and PCR/sequencing primers for cypriniform species; (2) establishing standard protocols that are adaptable to high throughput PCR/sequencing for systematic studies of Cypriniformes and other teleosts; (3) testing PCR performances using the existing and newly determined PCR primer sets for RAG1, Rhodopsin and four new nuclear markers (IRBP, EGR1, EGR2B and EGR3) on a diverse sampling from all major cypriniform lineages and other groups from the Ostariophysi; and (4) evaluating the utility of these six nuclear loci for phylogenetic and evolutionary studies of the Cypriniformes.

2. Materials and methods

2.1. Development of new nuclear gene markers

The selection of "good" gene markers adaptable to different levels of phylogenetic studies is a challenging task. In reconstructing the Cypriniformes Tree of Life, because of the diversity of the group and their hypothesized age, molecular markers should (1) be conservative enough to retain the phylogenetic signal during 48.6–55.8 MY (first cypriniform fossil discovered from the Early Eocene, but it appears that the divergence time of this group could be much earlier than this date) (Agassiz, 1843 (1833–1843); Sytchevskaya, 1986; Rüber et al., 2007); (2) contain considerable phylogenetically informative variation, but not too divergent and thus creating excess homoplasy and causing difficulties for the assessment of primary homology (sequence alignment); (3) be easy to amplify and sequence from diverse samples across a large spectrum of species diversity; (4) be of sufficient length (ideally >800 bp); and (5) be 'single-copy' genes in the nuclear genome. In considering these five aspects as criteria, we have searched for phylogenetically useful gene markers using the following two strategies.

First, we looked for nuclear markers that have been widely used for phylogenetic purposes in ray-finned fishes and other vertebrate lineages. Interestingly, in addition to Rhodopsin, RAG1 and RAG2 mentioned above, there are very few useful nuclear markers available that meet this selection criterion. For example, ribosomal DNA from 28S and S7 ribosomal RNA gene (including intron region) are pioneer nuclear loci (Lê et al., 1989; Chow and Hazama, 1998) and have become popular recently in molecular studies of teleost fishes (number of sequence for teleosteans presented in Genbank records on Nov. 2, 2007 for 28S and S7 is 1374 and 1479, respectively). As has been the case with mitochondrial genes as markers, frequently used genes have grown in popularity because of accessibility of 'universal' PCR primers and a lessening of technical hurdles. However, when either of these two genes are employed in analyses one is faced with a great number of possible alignments for a high number of sequences (e.g., 1000 sequences per gene for the CToL project) because assessment of primary homology could only minimally be based on secondary structure of the molecule (if any) with manual adjustments and/or automatic multiple alignment methods, an extremely time-consuming effort and a problem that will very likely lead to some degree of erroneous assessment of positional homology. As such, we have focused our searches on developing markers from exon regions of single-copy nuclear genes, assuming that multiple alignments of protein-coding genes will be relatively easy and straightforward because of conserved functional constraints of genes and triplet codes for amino acids. Unfortunately, most of the markers that have been widely used for the studies of vertebrate phylogeny (especially for tetrapod groups) such as c-mos (Saint et al., 1998), c-myr (Mohammad-Ali et al., 1995), and Tmo4C4 (Lovejoy and Collete, 2001) are relatively short in length (<600 bp). Sequencing such a short gene fragment is, in our opinion, not cost-effective.

Using the above criteria we identified one candidate marker — a gene encoding interphotoreceptor retinoid-binding protein (IRBP). IRBP mediates the transfer of all-*trans* retinol and 11-*cis* retinal between the pigmented epithelium and the photoreceptors (Pepperberg et al., 1993). The human IRBP gene is ~9.5 kbp and consists of one long exon (exon 1) plus three short exons separated by three introns (Fong et al., 1990). The human IRBP exon 1 is 3051 bp but only 1194 bp for the zebrafish due to a subsequent loss of a partial protein-coding region in the middle of exon 1 during the evolution of the ray-finned fishes (Rajendran et al., 1996; Nickerson et al., 2006). Because the length of the fragment is sufficiently large, the gene region from IRBP exon 1 has become widely used as a marker for phylogenetic studies of mammalian relationships (Schneider et al., 1996; Smith et al., 1996; Stanhope et al., 1996; DeBry and Sagel, 2001; Jansa and Weksler, 2004; Gaubert and Cordeiro-Estrela, 2006). In fishes, only one very recent instance has been reported for the use of this gene in phylogenetics of the Acanthomorpha (Dettaï and Lecointre 2008). Finally, it should be noted that the teleost genome apparently contains two copies of the IRBP gene arranged head-to-tail (Nickerson et al., 2006), in which the first copy of IRBP1 is without introns. The molecular marker used here and in Dettaï and Lecointre (2008) corresponds to the previously reported IRBP gene in zebrafish (Rajendran et al., 1996) or IRBP2. IRBP2 is, most likely, represented ubiquitously in all teleost genomes while IRBP1 has been lost in the

genomes of some teleost lineages such as medaka and sticklebacks (Nickerson et al., 2006; Dettaï and Lecointre, 2008).

Our second strategy for developing novel nuclear gene loci involved genome-scale mining (e.g., Li et al., 2007). First, we extracted a set of sequences (mRNA) from nuclear protein-coding genes of zebrafish (*Danio rerio*) and/or cypriniforms from Genbank. Our BLAST search of these sequences against whole-genome sequences of *Danio rerio* from the ENSEMBL database was conducted to extract candidate nuclear genes with exon sequences >800 bp in length. Starting with the sequences of the candidate genes, we conducted a broad search in the genomic databases (Genbank and complete databases of four other fish models, medaka (*Oryzias latipes),* stickleback (*Gasterosteus aculeatus),* and two puffer species (*Takifugu rubripes*, *Tetraodon nigroviridis*) for constructing datasets with orthologous (and paralogus) sequences of human, mouse, *Xenopus*, 5 model fishes and other ray-finned fishes (if available). Based on the retrieved sequences, phylogenetic trees were reconstructed to assess gene homology and identify potential gene duplication events (if any). After several runs of filtering based on the selection criteria described above, we successfully identified three novel nuclear markers from the early growth response (EGR) gene family — EGR1, EGR2B, and EGR3. EGR genes also belong to a family of zinc-finger transcription factor genes and their encoding proteins act as nuclear effectors of extracellular signals (Müller et al., 1991; Decker et al., 2003). As a consequence of two putative whole-genome duplications during the evolution of vertebrates, four EGR gene copies (EGR1–4) occur in vertebrate genomes (Knight et al., 2000; Schilling and Knight, 2001; Decker et al., 2003; Burmeister and Fernald, 2005). Based on our broad BLAST search in genomic databases and pre-phylogenetic analysis, four copies of EGR genes were found in the teleost genome that are homologous to the mammalian EGR1–3. However, there are no homologous (or significantly similar) sequences of mammalian EGR4 found in the teleost genomes. However, an extra copy of EGR2 was discovered in teleost genomes (Sun et al., 2002), which may have resulted from the fish-specific genome duplication (FSGD) (Taylor et al., 2001; Christoffels et al., 2004; Meyer and Van de Peer, 2005) (Supplementary Data 1). EGR genes in vertebrates are characterized as having a short exon 1, an intron and a long exon 2. The length of exon 2 of EGR genes from 5 model fish species varies from 744 to1323 bp, depending on the taxon and gene copies. Exon 2 of EGR2A is shorter than others. When comparing sequences of the 5′ end of exon 2 of EGR2A across taxa, the degree of sequence variation in this gene is high, rending it difficult to design a suitable forward primer. As such, this gene locus was not chosen as one of our targeted markers.

2.2. Primer design

An initial set of PCR and sequencing primers (Table 1) were designed with reference to the aligned sequences (mainly from the model fishes *Danio rerio*, *Oryzias latipes*, *Gasterosteus aculeatus*, *Takifugu rubripes*, and *Tetraodon nigroviridis*) retrieved from Genbank and genomic databases. An on-line tool, PRIMER 3 (http://biotools.umassmed.edu/bioapps/primer3_www.cgi) was used for designing primers. Resulting (initial) primers were used in obtaining sequences of each targeted gene from a few representative cypriniform taxa including *Sewellia lineolata* (Balitoridae: Balitorinae) and *Ischikauia steenackeri* (Cyprinidae: Cultrinae). Those sequences were used as a template to either redesign a set of cypriniform-specific nested primers for PCR or implemented for a primer-walking procedure to determine unknown flanking sequences at both ends of gene region of our targeted gene markers. For the latter procedure, a genome walking strategy (Siebert et al., 1995) as implemented in Universal GenomeWalker Kit (BD Biociences) was employed to obtain the outer sequence of Rhodopsin, IRBP, EGR1 and EGR2B for *S. lineolata* and *I. steenackeri*. The detailed protocol is described in the manufacture's manual (Universal GenomeWalker™, BD Biosciences). Sequences from the longer gene fragments from *S. lineolata*, *I. steenackeri* and *D. rerio* (latter sequence retrieved from complete genome database) were later used as references for designing the outer primers for amplifying and sequencing our targeted gene regions for all cypriniform taxa. Finally, another set of cypriniform-specific nested primers for PCR/sequencing was designed when necessary. The procedures presented herein will help guarantee to have the same size (more or less) for targeted gene regions of the markers as originally defined throughout the entire primer developing phase. A more classical procedure for developing only nested primers will lead to only shorter and shorter amplified fragments when more and more nested primers are designed for PCR/sequencing. Primers used for this study are listed in Table 1.

Table 1
PCR/Sequencing primer information

Locus/primer[a]	Primer sequence (5′–3′)	Source
RAG1		
R1 2533F	CTGAGCTGCAGTCAGTACCATAAGATGT	López et al., 2004
R1 4078R	*TGAGCCTCCATGAACTTCTGAAGRTAYTT*	López et al., 2004
R1 4090R	*CTGAGTCCTTGTGAGCTTCCATRAAYTT*	López et al., 2004
R1 4061R	*AATACTTGGAGGTGTAGAGCCAGT*	Chen et al., 2007
Rhodopsin		
RH 28F[c]	TACGTGCCTATGTCCAAYGC	This study
RH 1039R[b]	*TGCTTGTTCATGCAGATGTAGA*	Chen et al., 2003
RH 193F[b]	CNTATGAATAYCCTCAGTACTACC	Chen et al., 2003
RH 233F[d]	ATATGCCTGCCTGGCYGCTTAC	This study
IRBP		
IRBP 109F[b]	AACTACTGCTCRCCAGAAAARC	This study
IRBP 1001R[b]	*GGAAATGCATAGTTGTCTGCAA*	This study
IRBP 76F[c]	CTTRTTGTGGATATGGCAAAAAT	This study
IRBP 1162R[c]	*TGGTGGWCTTYAGGCACTTGT*	This study
IRBP 101F[d]	TCMTGGACAAYTACTGCTCACC	This study
IRBP 1068R[d]	*AGATCAKGYTGTATTCCCCACTA*	This study
EGR1		
E1 290F[b]	TMTCTTACACAGGCCGYTTCAC	This study
E1 1126R[b]	*CTTTYTCTGCTTTCTTGTCCTTCT*	This study
E1 228F[c]	GAAATTTCCATGGASAAACCTCT	This study
E1 1389R[c]	*AGAACTGTAGATGTTGCTGRCCAC*	This study
E1 333F[d]	CAGYAACAGTCTRTGGGCTGAG	This study
E1 1104R[d]	*CGCAGGTGGATCTTRGTGTG*	This study
EGR2B		
E2B 278F[b]	AGTTTTCCATCGACTCSCAGTA	This study
E2B 1117R[b]	*AGGTGGATTTTGGTGTGTCTYTT*	This study
E2B 205F[c]	ACTTRTCCTATYCCAGCAGCTT	This study
E2B 1108R[d]	*TTTTGTGTGTCTCTTTCTYTCGTC*	This study
E2B 287F[d]	TTGACTCSCAGTATCCAGGTAAC	This study
EGR3		
E3 161F[b]	AATATCATGGACYTGGGNATGG	This study
E3 1136R[b]	*GGYTTCTTGTCCTTCTGTTTSAG*	This study
E3 254F[d]	GTCACCTAYYTGGGSAAGTTT	This study

[a] Reverse primers in italics; Abbreviations of genes: RAG1, recombination activation gene 1; RH, Rhodopsin; IRBP, interphotoreceptor retinoid-binding protein gene; EGR, Early growth response protein gene.
[b] Initial primer.
[c] Outer primer.
[d] Nested primer.

2.3. Specimens

A total of 37 samples were examined, comprising all cypriniform family and subfamily groups; an additional eight samples of other Ostariophysi species were used for this study (Supplementary data 2). The classification following Nelson (2006) was used as a guideline for our choices of samples. Our investigation was not intended to choose any particular taxon for the analyses. Rather, the central criterion employed was the quality of the genomic DNA as PCR failure can

simply be due to poor quality of original genomic DNA rather than other parameters such as the specificity of the primers to the template.

2.4. Suggested protocol for molecular work

Our suggested protocol for molecular laboratory work for these new genes is described as follows. Tissue extraction was performed using Qiagen DNAeasy extraction kit (Qiagen, Valencia, CA) according to the manufacturer's instructions. Extracted DNA quantity was measured by Spectrophotometer (Eppendorf). Conditions for amplification (PCR) were as follows: GoTaq® Flexi DNA Polymerase (0.5 units) (Promega), 1x reaction buffer, 2 mM of $MgCl_2$, 200 μM of each dNTP, 0.2 μM of each primer, and 20–50 ng of genomic DNA in a 25 μl of final reaction volume. Thermocycler conditions for PCR were: initial denaturing step at 95 °C for 4 min followed by 35 cycles of 95 °C (for 40 s), annealing Tm (for 40 s), and 72 °C (for 1–1.5 min. depending on size of fragments), and then a final extension step of 72 °C (for 7 min) before a 4 °C soak. Tm was 55 °C for all gene loci except for the RAG1 (Tm = 53 °C). When the PCR result was weak, a double amount of DNA Polymerase (and/or genomic DNA) was added in the PCR pre-mix to improve PCR productivity. Alternatively, a high fidelity *Taq*, Takara *Ex Taq* (Takara Bio Inc.) can be used. This *Taq* is very efficient for amplifying longer gene fragments such as RAG1 and for a PCR using a degenerated primer (e.g., amplification of Rhodopin using the forward primer, RH 193F). Finally, the PCR cleanup procedure followed the AMPure magnetic bead cleanup protocol (Agencourt Bioscience Corporation) and resuspended in 30 μL of sterile water. Sequences were then determined by Macrogen Inc. (Seoul, South Korea) using ABI 3730xl analyzer (Applied Biosystems).

2.5. Sequence data and phylogenetic analysis

DNA sequences were edited and managed using Se-Al v2.0a11 (Rambaut, 1996). Descriptive statistics of comparing sequences and tests of homogeneity of base frequencies across taxa using chi-square tests were performed using PAUP*-version 4.0b10 (Swofford, 2002). The same program (PAUP*) was used for evaluating the phylogenetic performance for each data set (each gene) in terms of levels of homoplasy, as measured by indices CI and RI, and in terms of node robustness estimated using the bootstrap procedure (Felsenstein, 1985) under Maximum Parsimony (MP) criterion and NJ distance method (Saitou and Nei, 1987) with uncorrected p-distance. Bootstrap analyses were based on 1000 pseudo-replicates. Number of nodes with bootstrap support of 80% or higher was counted from resulting bootstrap consensus trees of each dataset. For the MP analyses, optimal trees were obtained by heuristic searches with random stepwise addition sequences followed by TBR swapping for 10 replications (Swofford, 2002).

To gain further insight on 'phylogenetic' signals presented in each data set across taxa, a protocol called repeated-bootstrap components (Chen et al., 2003) was employed. This protocol consisted of scoring the repeated clades found in the listing of bootstrap bipartitions from MP analysis produced by PAUP* for each separate gene partition (repeated-bootstrap components) using the computer program developed by the first author (available upon request to WJC). Resulting repeated-bootstrap components were mapped onto the MP tree derived from simultaneous analysis of the combined dataset. The corresponding bootstrap values for each data partition were displayed in the form of a histogram for each node on the tree.

3. Results

3.1. PCR performance

PCR performance was tested using the existing and newly determined PCR primers described in Table 1 for RAG1, Rhodopsin and four new nuclear markers (IRBP, EGR1, EGR2B, EGR3) on 37 samples of cypriniform species, and 8 samples of other ostariophysans species. The results of PCR assays, based on different combinations of forward and reverse PCR primers for our gene markers, are shown in the table of Supplementary data 2. Using existing primers for RAG1 described in López et al. (2004) and Chen et al. (2007) we successfully amplified this gene fragment across a wide spectrum of cypriniform diversity; however, two new forward Rhodopsin primers (RH 28F and RH 233F) were needed for the complete PCR recovery of the Rhodopsin gene locus for the cypriniform taxa. In general, we observed approximately 90% success using a single pair of primers for a particular gene locus for our test cypriniform samples; we observed 100 % PCR success by combining these with a set of different primer pairs. With regard to the assays for the outgroup taxa, except for IRBP, many of the primer pairs used in amplifying the gene regions of our targeted loci for cypriniforms perform equally well for other ostariophysans (Supplementary data 2) and for other teleosts, especially species of the Percomorpha (data not shown).

3.2. Characteristics of sequence data and nucleotide substitution patterns

Sequences of each locus from a common set of 30 cypriniform taxa (those with an asterisk in the table of Supplementary data 2) were successfully obtained using the PCR primers (Table 1) for sequencing. These sequences (Genebank accession numbers: EU409606–EU409791), plus sequences from *Danio rerio* (U71093; L11014; X85957; NM_131248; NM_130997; scaffold2320.1) were used in describing sequence variation among taxa and assessing the phylogenetic performance. Sequences from all gene loci were unambiguously aligned manually except for a small part of the 5′ end region of the amplified fragment of EGR1. This region contains a teleost-specific insert of ~20 residues of a serine/threonine-rich domain (Burmeister and Fernald, 2005). No internal indels were found among the aligned sequences of RAG1, Rhodopsin, and IRBP. A few gaps needed to be introduced in adjusting sequence alignment of EGR genes (Table 2).

The length of the aligned sequences and other descriptive statistics, as implemented in PAUP*, for each gene dataset is summarized in Table 2. Patterns of nucleotide variations differ among genes. EGR2B is one of the most conserved genes, whereas IRBP is the most variable one among the genes examined. As compared to two selected loci from the mitochondrial genome (fragment consisting of complete 12S, tRNA-Val, and partial 16S genes and Cytochrome *b* gene) and 10 new markers developed for higher-level phylogenetics of ray-finned fishes in a very recent study by Li et al. (2007), our nuclear markers come out to contain more phylogenetic information than half of the markers described in the latter paper do; however, our genes appear to be less variable than all of the mitochondrial genes do (the ribosomal genes represented here are purported to the most slowly evolving genes of the mitochondrial genome) (Table 3). The phylogenetic information inherent in the genes outlined herein was evaluated simply through sequence length and pairwise distances between the sequences of *Semotilus atromaculatus* (Cyprinidae: Leuciscinae) and *Danio rerio* (Cyprinidae: Rasborinae), the taxa common to all relevant gene datasets.

With regard to the patterns of gene evolution, as with protein-coding genes in general, the third codon position of our gene loci accumulate most of the nucleotide substitutions as compared to codon positions one and two (Fig. 1). Excess of nucleotide substitutions at third codon position is likely to result in a higher levels of homoplasy relative to the first and second positions, especially when one examines sequences from more distantly related species. To evaluate this, we performed the absolute saturation tests (Philippe et al., 1994) on transitions and transversions for each gene separately at the third codon position (Fig. 2). With the exception of EGR1 and 2B which exhibit a slight saturation on transitions at the third codon position (Figs. 2G and I), we did not detect any clearly diagnostic

Table 2
Descriptive statistics of sequences and phylogenetic performance from each locus

Locus	Length[a] (in bp)	Indels	Mean pairwise difference[b] (range)	No. parsimony-informative sites	No. variable sites (in %)	Base frequencies homogeneity[c] (3 pos.)	CI (RI)[d]	No. node with >80% MPBP (NJBP)[e]
RAG1	1497	No	0.121 (0.014–0.195)	501	620 (41.42%)	0.999 (0.672)	0.43 (0.61)	18 (20)
Rhodopsin	819	No	0.138 (0.010–0.203)	276	344 (42.00%)	0.000* (0.000)*	0.46 (0.67)	15 (20)
IRBP	849	No	0.142 (0.008–0.226)	330	446 (52.53%)	0.999 (0.906)	0.46 (0.64)	17 (16)
EGR1	852	A few	0.101 (0.006–0.158)	243	321 (37.68%)	0.999 (0.102)	0.48 (0.67)	15 (14)
EGR2B	819	Few	0.072 (0.004–0.136)	172	254 (31.01%)	1.000 (0.816)	0.50 (0.65)	10 (12)
EGR3	912	Few	0.085 (0.009–0.146)	206	289 (31.67%)	0.999 (0.031)*	0.51 (0.63)	14 (18)

[a] Calculated from length of aligned DNA nucleotide sequences in base pair (bp).
[b] Calculated from uncorrected p-distance.
[c] p value from Chi-square test of homogeneity of base frequencies across taxa. The test was performed either with all nucleotides or the nucleotides in the third codon position (value between parentheses). Asterisk sign indicates that the data are significantly rejected by Chi-square test.
[d] Consistency index (CI) and Retention index (RI).
[e] Node support was assessed using the bootstrap procedure (BP) through Maximum Parsimony (MP) method and neighbor-joining (NJ) method (using uncorrected p-distance).

saturation plateau as can be seen in mitochondrial protein-coding genes (Saitoh et al., 2006).

Phylogenetic analyses of protein-coding genes may also be biased from potential homoplasy (particularly at third codon positions) due to base composition bias across taxa (Lockhart et al., 1994; Chen et al., 2003). Interestingly, a significant bias was only observed for the Rhodopsin dataset and the third codon position of EGR3 (Table 2). These observations and tests indicate that the nuclear markers from our study may contain a stronger phylogenetically informative signal, even at the third codon positions, than other gene loci (e.g., mitochondrial protein-coding genes) that are currently employed for many systematic studies of ray-finned fishes.

3.3. Independence of genes

The benefit from analyses of multiple nuclear gene loci in phylogenetic inferences of compared species is that of assessment of reliability based on the proportion of the concordant gene genealogical results obtained from the genes independently (Chen et al., 2003) (see also the Introduction). Therefore, independence of genes is one of the critical evaluations for the phylogenetic utility of developed gene markers. The genes used in this study can be considered independence genes in terms of their physical locations in the genomes and/or their functional constraints and selective pressures (Slowinski and Page, 1999). According to the available information obtained from ENSEMBLE databases of whole genomic sequences of model fishes, where known, these gene loci are located on different chromosomes in the fish genomes (Table 4). Furthermore, gene functions are completely different among RAG1 (function implemented in immune system of vertebrates), Rhodopsin (encoding a visual pigment), IRBP (encoding a protein to function as an intercellular transporter in the vertebrate eye) and EGR genes (regulatory genes).

3.4. Phylogenetic performance

The results of our evaluations of the phylogenetic performance of these genes, as inferred from indices of homoplasy and bootstrap values, were shown in Table 2. None of the data exhibit extremely high homoplasy (CI and RI were always higher than 0.4 and 0.6 respectively). As compared to a systematic study for the Cypriniformes using whole mitogenome genome sequences (16 ingroup taxa included with a similar taxonomic coverage as our study) (He et al., 2008a), the RI value measured from our datasets (RI=0.65 on average) (Table 2) are much higher than those resulting from the mitogenome dataset of He et al. (2008a) (RI=0.38).

Table 3
Comparison of the pairwise distance between the sequences of *Semotilus atromaculatus* (Cyprinidae: Leuciscinae) and *Danio rerio* (Cyprinidae: Rasborinae) for different gene markers

Locus	Length (in bp)	Pairwise distance	Source
RAG1	1497	0.111	This study
Rhodopsin	819	0.160	This study
IRBP	849	0.139	This study
EGR1	852	0.118	This study
EGR2B	819	0.081	This study
EGR3	912	0.107	This study
ENC1	810	0.083	Li et al. (2007)
GLYT	819	0.105	Li et al. (2007)
MYH6	735	0.141	Li et al. (2007)
PLAGL2	669	0.055	Li et al. (2007)
PTR	699	0.077	Li et al. (2007)
RYR3	821	0.107	Li et al. (2007)
SERB2	988	0.055	Li et al. (2007)
SH3PX3	705	0.077	Li et al. (2007)
TBR1	648	0.144	Li et al. (2007)
ZIC1	858	0.044	Li et al. (2007)
12 S-16 S partial	2716	0.162	Simons and Mayden (1998)
Cytochrome b	1133	0.237	Dowling et al. (2002)

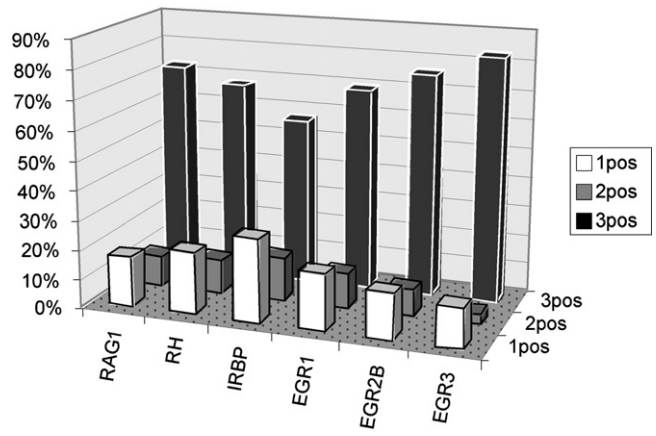

Fig. 1. Distribution of variable nucleotide sites among three codon positions from the protein-coding gene sequences in this study.

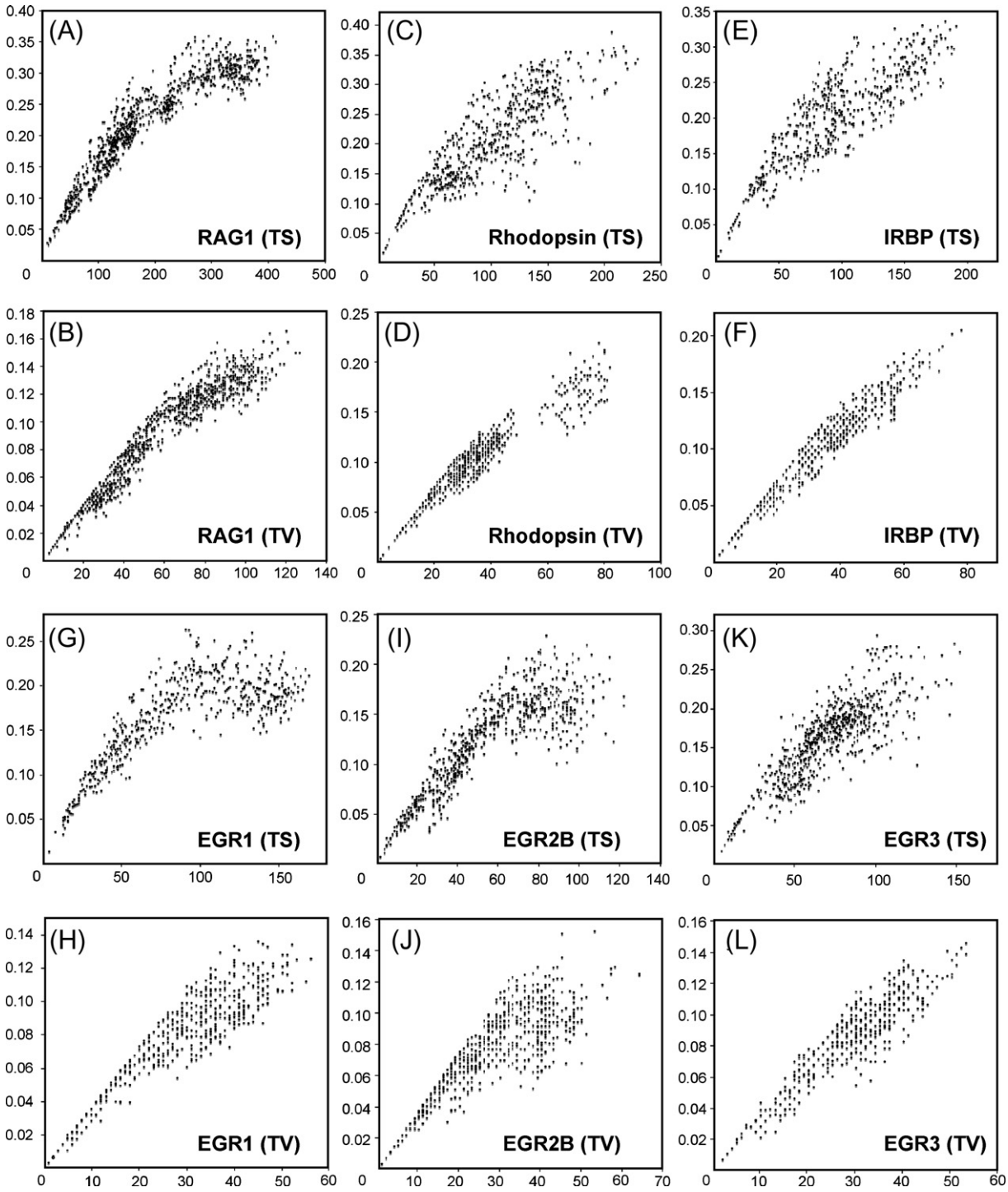

Fig. 2. Absolute saturation tests (Philippe et al., 1994) for the nuclear genes employed in this study. Plots show transitions (TS) at third codon positions of each gene locus (A, C, E, G, I, K) and transversions (TV) at third codon positions of each gene locus (B, D, F, H, J, L). X-axis: number of substitutions among all pairs of terminals inferred from the Maximum Parsimony tree; *Y*-axis: mean character differences (adjusted for missing data or gaps) between two sequences.

With regard to the phylogenetic performance of these genes, as interpreted by the number of robust nodes in bootstrap consensus trees, analyses from the RAG1 dataset resulted in the best overall score among all the data sets (Table 2). When the individual gene data sets were combined for a global analysis the number of supported nodes from MP and NJ analyses increased to 22 and 21, respectively, or about 70% of the nodes of a fully resolved tree, indicating an additive pattern of phylogenetic signal among the genes from each dataset in a global analysis. In He et al. (2008a) only 56% of the nodes were recovered with MP bootstrap values of 80% or higher. It should also be noted that the number of nucleotide characters in studies using whole mitogenome sequences (e.g., 14,768 bp in the dataset of He et al., 2008a) is roughly 2.5 times greater than the dataset we used (combination of 6 nuclear loci). Despite the results of these comparisons we do not argue against the use of mitogenome sequences in phylogenetic studies. Rather, economically speaking the collection of a large data set of mitogenome sequences may not be feasible in a single laboratory with a limited budget and the alternative gene markers presented here are equally appropriate for a particular systematics or evolutionary study.

Table 4
Location of gene loci used in this study in model fish genomes[a]

	RAG1	Rhodopsin	IRBP	EGR1	EGR2B	EGR3
Model fish genomes						
Danio rerio	Chr. 25[b]	Chr. 7	Chr. 12	Chr. 14	Chr. 12	NA[c]
Oryzias latipes	Chr. 6	Chr. 7	NA	Chr. 10	Chr. 19	Chr. 9
Gasterosteus aculeatus	Chr. 19	Chr. 12	Chr. 5	Chr. 4	NA	Chr. 13
Takifugu rubripes	NA	NA	NA	NA	NA	NA
Tetraodon nigroviridis	Chr. 13	Chr. 9	Chr. 2	Chr. 1	NA	Chr. 12

[a] Location information obtained from ENSEMBL database by BLAST search of the corresponding sequences of genes for the species against whole-genome sequences.
[b] Gene location on a denoted chromosome (Chr.), e.g., Chr. 25 indicates gene is located on chromosome number 25.
[c] Location information not available (NA).

Finally, further investigation of the phylogenetic performance from each gene data set was conducted using the repeated-bootstrap components protocol. If bootstrap values are regarded as general measure of hierarchical signal (Hillis and Bull, 1993), under the criterion of repeatability (Chen et al., 2003), the histogram mapped on the tree can be interpreted as the contribution to phylogenetic signal of each data partition in support of a corresponding node (Fig. 3). This information may also be useful for future studies of particular cypriniform groups by focusing on signal-rich genes when the target taxonomic samples become available. In general, given the results shown in the Fig. 3 the phylogenetic signal is homogeneously distributed across different data partitions throughout the tree. Likewise, it appears that the contribution of the EGR3 data is rather weak in some case and the RAG1 data set contains a stronger phylogenetic signal. Regarding particular parts of tree, the RAG1, IRBP and EGR1 data seem to perform equally well for recovering several major clades within the Cypriniformes. These same clades have also been corroborated in taxonomically congruent findings between morphology and molecular studies (Cavender and Coburn, 1992; Smith, 1992; Šlechtová et al., 2007; Saitoh et al., 2006; He et al., 2008a). For instance, the monophyly of family/subfamily of Catostomidae, Cobitinae, Cyprinidae, Gobioninae, Cyprininae and Leuciscinae sensu Cavender and Coburn (1992) is recovered with strong nodal supports, while naturalness of traditionally recognized Balitoridae, Cobitidae and Rasborinae is still challenged, as outlined by recent molecular studies (Šlechtová et al., 2007; Saitoh et al., 2006; He et al., 2008a). Although the rhodopsin data perform well for resolving many branches within the tree, it contains relatively poor to no signal to

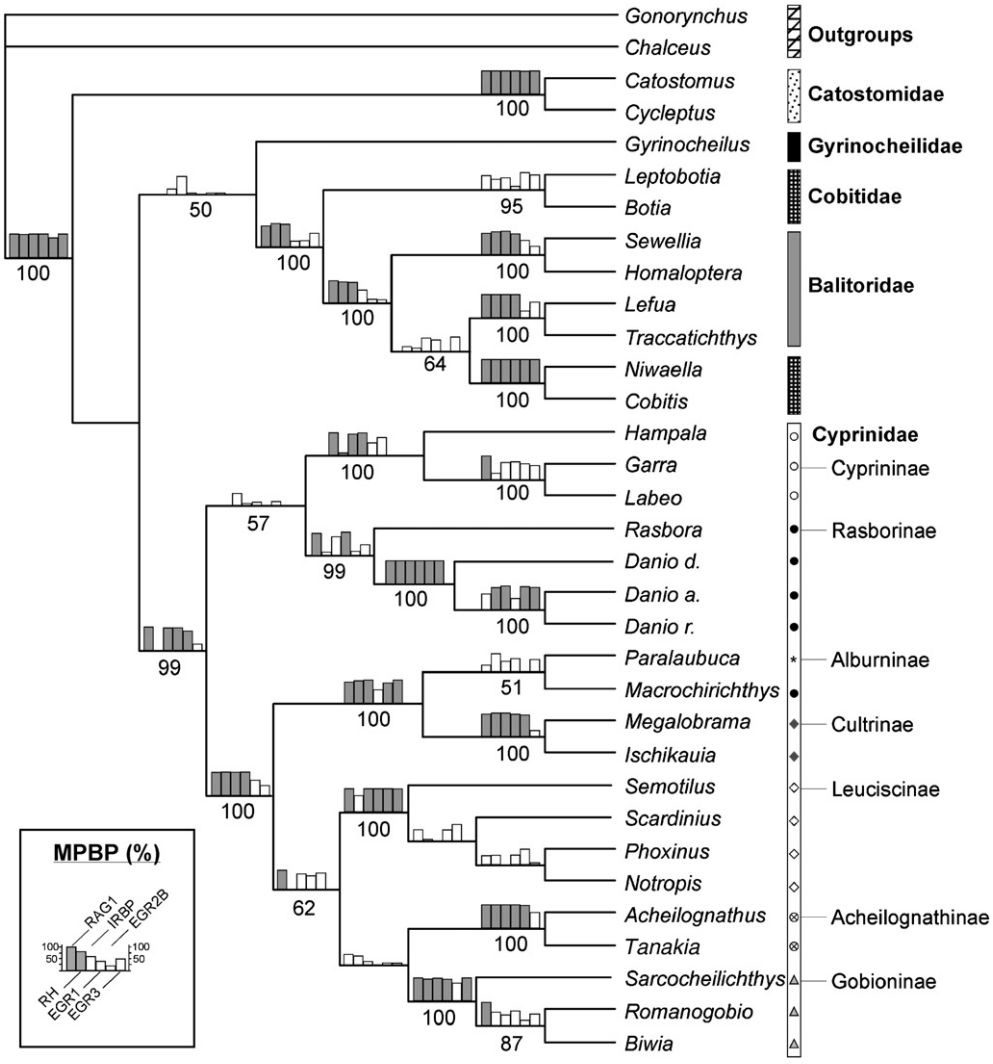

Fig. 3. Maximum Parsimony (MP) tree of the Cypriniformes from simultaneous analysis of the combined dataset (6 nuclear genes: 5817 bp). Only single MP tree with length of 8452 was obtained. Numbers below the branches present global bootstrap (BP) values in percentage from simultaneous analysis. Values below 50% are not shown. Small histograms over branches are the MPBPs from repeated-bootstrap components (only the ones congruent with this MP tree are shown). They are BPs taken from each of the six listings of repeated BP bipartitions obtained from separate MPBP analyses, displayed for RAG1, RH, IRBP, EGR1, EGR2B, and EGR3 respectively. Gray histograms indicate that the resulting separate MPBPs are equal to or higher than 80.

recover the monophyly of the family Cyprinidae, the subfamilies Cyprininae or Rasborinae (less *Macrochirichthys*), or a clade grouping Leuciscinae, Acheilognathinae, and Gobioninae (Fig. 3). None of the gene markers used here provide a promising phylogenetic signal for resolving the internal relationships of the Leucisinae or the inter-relationships among Leuciscinae, Acheilognathinae and Gobioninae.

4. Discussion

Evolution of nuclear genomes is more complicated than that of mitochondrial genomes with a single parental inheritance and a lack of recombination (at least in animal mt-genomes). One of the critical issues in phylogenetic inference when using nuclear markers is the potential uncertainty regarding the orthology of the sequences analyzed (Martin and Burg, 2002). That is, multiple (or paralogous) copies of a targeted gene could be amplified in PCR from whole genomic DNA and researchers may be comparing paralogous rather than orthologous genes in a phylogenetic analysis. Therefore, inferences from analyses of datasets containing a mixture of orthologous and paralogous copies will yield spurious results of taxon relationships. This particular complication should be a major concern in phylogenetic inferences of many ray-finned fishes. Many putative single-copy genes in vertebrates have been hypothesized to have been duplicated during the evolution of ray-finned fishes with the postulated FSGD (Christoffels et al., 2004). This event(s) is postulated to have occurred roughly 320 MYA, before the divergence of most teleost species (Hoegg et al., 2004). In fact, with the genealogies derived from preliminary analyses of each gene, three of the six nuclear loci presented here (Rhodopsin, IRBP, EGR2) are known to have extra copies in teleost genomes that likely originated during the FSGD event. This indicates that although the use of a 'single-copy' gene is important or minimally a preference for phylogenetic inferences, at this point in time such a requirement seems impossible for phylogenetic studies of teleost fishes using nuclear genes. Care, however, must be taken when working with nuclear loci to make sure that where duplication has been identified in a group the orthologous genes are being compared.

Practically, to minimize chances of sampling paralogous copies in analyses, we used the following strategy to guarantee orthology among sequences for our nuclear loci. First, we avoided selecting gene markers that showed a high degree of nucleotide sequence similarity (>70%) with sequence from detectable paralogous copies with the help of the broad Blast search of candidate genes against the complete genome databases of fish model systems. Second, when designing primers, we selected primer-binding sites that differed among putative paralogous genes. This can be easily achieved since, for instance, divergence between EGR2B and paralogous gene (EGR2A) is far greater than the divergence observed among all teleost EGR2B sequences because the duplication event leading to the separation of EGR2A and EGR2B occurred before the diversification of all teleostean fishes. However, this procedure cannot prevent a possible problem resulting from a mixture of orthologous and paralogous copies due to more recent duplication events.

The occurrence of recent whole-genome duplication events (so-called polyploidy) in fishes, especially in the Cypriniformes has been documented (Tsigenopoulos et al., 2002; David et al., 2003; Leggatt and Iwama, 2003; Saitoh, 2003; Le Comber and Smith, 2004). Lim et al. (1997) reported two distinct copies of the rhodopsin gene for the common carp, *Cyprinus carpio*. In a screen from a goldfish (*Carassius auratus auratus*) bacterial artificial chromosome genomic library Luo et al. (2006) detected three copies of the nuclear gene RAG1, with two sequences being nearly identical. This gene has been purported to be a single-copy nuclear gene in vertebrates including ray-finned fishes. Not surprisingly, both of these instances are typical cyprinids with polyploid genomes (Ohno et al., 1967; Yu et al., 1987; Larhammar and Risinger, 1994; David et al., 2003). Fortunately, these kinds of whole-genome duplications are most likely lineage-specific events for the Cypriniformes (Larhammar and Risinger, 1994; David et al., 2003), and would have limited impact on higher-level phylogenetic inferences. In these instances, even if one had sequenced genes from such polyploid individuals or sequenced 'by mistake' a paralogous gene from other taxa, the resulting sequence profiles would have shown this as excessive divergence and it is detectable in early stages of analysis. In fact, in a few of the sequences obtained for our analyses we did observe instances of one to several polymorphic sites (determined by a mixture of double-base callings on the sequence chromatograms) along a sequence. Observed polymorphic sites on sequences could result from either PCR/sequencing of both alleles from a heterozygous diploid individual or a PCR/sequencing of all or some of the copies of genes from a polyploid individual. In the latter case, one would expect to observe the retention of polymorphic sites from a majority of the nuclear gene data sets and/or the sequences of concern to have an excess in the number of polymorphic sites (> 1% of sequence nucleotides generally according to our observations). In most of the sequences obtained in this study we only found two cases corresponding to this particular situation, both in species of the family Catostomidae where genome duplication or a tetraploid origin has been previously hypothesized (Uyeno and Smith, 1972). In the Blue Sucker, *Cycleptus elongates*, polymorphic sites were present in four of the six nuclear loci, among which RAG1 and EGR1 possessed an excess number of polymorphic sites. In the Spotted Sucker, *Minytrema melanops*, we failed to unambiguously determine sequences for EGR1 and 2B due to the presence of an excess number of polymorphic sites in the sequence chromatograms. A few polymorphic sites were observed in available sequences from two of the other four nuclear loci. The presence of multi-paralogous copies of nuclear genes in the genomes of the catostomid species examined could be a consequence of a single whole-genome duplication event before divergence of all catostomid fishes, followed by the subsequent loss of a gene (entirely or partially) in some lineages (Ferris and Whitt, 1977; Buth and Mayden, 2001). In addition, Clements et al. (2004) has demonstrated the presence of two distinct growth hormone genes in the Smallmouth Buffalo, *Ictiobus bubalus* (Catostomidae), confirming this suspicion. However, a thorough assessment using multi-locus approach with relevant taxonomic sampling is still required for future studies in genomic evolution of catostomid species.

After all, although gene or genomic duplication events as described herein, and their occurrence impacting phylogenetic inferences, may not frequently exist in the great diversity of cypriniform fishes, the interpretation of phylogenetic results derived from nuclear genes as markers should be treated with caution, especially for the Catostomidae and the subfamily Cyprininae (family Cyprinidae) where tetraploids have also been hypothesized. Ultimately, however, as we addressed earlier, a multi-locus approach is the best strategy for resolving troublesome evolutionary relationships for a given taxonomic group and for detecting the origins of conflicting inferences. A focus on repeated clades obtained from different genes trees (e.g., Chen et al., 2003) seems to be the most conservative way at this point in time for assessing the reliability of phylogenetic inferences, at least in the early stages of such studies. It is highly unlikely that an erroneous clade resulting from mistaken orthology would be repeatedly obtained through independent gene trees. This is not a justification for orthology, but stresses that a careful analysis through multi-gene data sets cannot be challenged by undetected paralogies.

Finally, genetic introgression between species is widely recognized in cypriniform as well as in some other fishes (DeMarais et al., 1992; Saitoh et al., 2004; Costedoat et al., 2007). The process of this type of gene transfer is usually achieved via hybridization and subsequent backcrossing with genetic recombination. This can occur between species that are closely related and between species that are not closely related. Interestingly, introgressive hybridization has long been hypothesized to play an important role in the evolutionary

diversification of living organisms whereby lineages have been hypothesized to benefit through the incorporation of new genetic variations (Anderson, 1949; Dowling and Secor, 1997; Gerber et al., 2001). This naturally-occurring process may also have a significant impact in the conservation of species when a rare or endangered species is genetically assimilated by a more common species (DeMarais et al., 1992; Rhymer and Simberloff, 1996; Rieseberg, 1998; Costedoat et al., 2007). In addition to the value of multiple nuclear genes in systematic studies, as the nuclear genome of a bisexual organism is derived from the both parents, the use of analyses of nuclear gene loci (in combination of a mitochondrial marker) as established herein may provide additional information of extreme value for conservation and restoration efforts of species.

Acknowledgments

We are grateful to Drs. Henry L. Bart Jr. and Kevin Tang for the discussions and comments on this manuscript. This research was supported by the National Science Foundation's Cypriniformes Tree of Life Initiative as part of the NSF Assembling the Tree of Life Initiative (EF 0431326). Research support for MM and KS is by the Grants-in-Aid from the Ministry of Education, Culture, Sports, Science and Technology, Japan (Grant No. 17207007).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2008.07.016.

References

Agassiz, L., 1833–1843. Recherches sur les poissons fossiles. Neuchâtel and Soleure, Petitpierre.
Anderson, E., 1949. Introgressive Hybridization. John Wiley, New York.
Burmeister, S.S., Fernald, R.D., 2005. Evolutionary conservation of the egr-1 immediate-early gene response in a teleost. J. Comp. Neurol. 481, 220–232.
Buth, D.G., Mayden, R.L., 2001. Allozymic and Isozymic evidence for polytypy in the North American Catostomid genus *Cycleptus*. Copeia 2001, 899–906.
Cavender, T.M., Coburn, M., 1992. Phylogenetic relationships of North American Cyprinidae. In: Mayden, RL (Ed.), Systematics, historical ecology and North American freshwater fishes. Standford University Press, Stanford, pp. 328–378.
Chen, W.-J., Bonillo, C., Lecointre, G., 2003. Repeatability of clades as criterion of reliability: a case study for molecular phylogeny of Acanthomorpha (Teleostei) with larger number of taxa. Mol. Phylogenet. Evol. 26, 262–288.
Chen, W.-J., Orti, G., Meyer, A., 2004. Novel evolutionary relationship among four fish model systems. Trends Genet. 20, 424–431.
Chen, W.-J., Ruiz-Carus, R., Ortí, G., 2007. Relationships among four genera of mojarras (Teleostei: Perciformes: Gerreidae) from the western Atlantic and their tentative placement among percomorph fishes. J. Fish Biol. 70 (sb), 202–218.
Chow, S., Hazama, K., 1998. Universal PCR primers for S7 ribosomal protein gene introns in fish. Mol. Ecol. 7, 1255–1256.
Christoffels, A., Koh, E.G., Chia, J.M., Brenner, S., Aparicio, S., Venkatesh, B., 2004. Fugu genome analysis provides evidence for a whole-genome duplication early during the evolution of ray-finned fishes. Mol. Biol. Evol. 21, 1146–1151.
Clements, M.D., Bart, H.L.J., Hurley, D.L., 2004. Isolation and characterization of two distinct growth hormone cDNAs from the tetraploid smallmouth buffalofish (*Ictiobus bubalus*). Gen. Comp. Endocrinol. 136, 411–418.
Costedoat, C., Pech, N., Chappaz, R., Gilles, A., 2007. Novelties in hybrid zones: crossroads between population genomic and ecological approaches. PLoS ONE 2, e357.
Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12, 814–822.
Cunha, C., Mesquita, N., Dowling, T.E., Gilles, A., Coelho, M.M., 2002. Phylogenetic relationships of Eurasian and American cyprinids using cytochrome b sequences. J. Fish Biol. 61, 929–944.
David, L., Blum, S., Feldman, M.W., Lavi, U., Hillel, J., 2003. Recent duplication of the common carp (*Cyprinus carpio* L.) genome as revealed by analyses of microsatellite loci. Mol. Biol. Evol. 20, 1425–1434.
DeBry, R.W., Sagel, R.M., 2001. Phylogeny of Rodentia (Mammalia) inferred from the nuclear-encoded gene IRBP. Mol. Phylogenet. Evol. 19, 290–301.
Decker, E.L., Nehmann, N., Kampen, E., Eibel, H., Zipfel, P.F., Skerka, C., 2003. Early growth response proteins (EGR) and nuclear factors of activated T cells (NFAT) form heterodimers and regulate proinflammatory cytokine gene expression. Nucleic Acids Res. 31, 911–921.
Delsuc, F., Brinkmann, H., Philippe, H., 2005. Phylogenomics and the reconstruction of the tree of life. Nat. Rev. Genet. 6, 361–375.
DeMarais, B.D., Dowling, T.E., Douglas, M.E., Minckley, W.L., Marsh, P.C., 1992. Origin of *Gila seminuda* (Teleostei: Cyprinidae) through introgressive hybridization: implications for evolution and conservation. Proc. Natl. Acad. Sci. U.S.A. 89, 2747–2751.
Dettaï, A., Lecointre, G., 2008. New insights into the organization and evolution of vertebrate IRBP genes and utility of IRBP gene sequences for the phylogenetic study of the Acanthomorpha (Actinopterygii: Teleostei). Mol. Phylogenet. Evol. in press.
Dowling, T.E., Secor, C.L., 1997. The role of hybridization and introgression in the diversification of animals. Annu. Rev. Ecol. Syst. 28, 593–619.
Dowling, T.E., Tibbets, C.A., Minckley, W.L., Smith, G.R., 2002. Evolutionary relationships of the Plagopterins (Teleostei: Cyprinidae) from cytochrome b sequences. Copeia 665–678.
Durand, J., Tsigenopoulos, C.S., Unlü, E., Berrebi, P., 2002. Phylogeny and biogeography of the family Cyprinidae in the Middle East inferred from cytochrome b DNA — evolutionary significance of this region. Mol. Phylogenet. Evol. 22, 91–100.
Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39, 783–791.
Ferris, S.D., Whitt, G.S., 1977. Loss of duplicate gene expression after polyploidisation. Nature 265, 258–260.
Fong, S.L., Fong, W.B., Morris, T.A., Kedzie, K.M., Bridges, C.D., 1990. Characterization and comparative structural features of the gene for human interstitial retinol-binding protein. J. Biol. Chem. 265, 3648–3653.
Gaubert, P., Cordeiro-Estrela, P., 2006. Phylogenetic systematics and tempo of evolution of the Viverrinae (Mammalia, Carnivora, Viverridae) within feliformians: implications for faunal exchanges between Asia and Africa. Mol. Phylogenet. Evol. 41, 266–278.
Gerber, A.S., Tibbets, C.A., Dowling, T.E., 2001. The role of introgressive hybridization in the evolution of the *Gila robusta* complex (Teleostei: Cyprinidae). Evolution 55, 2028–2039.
Gilles, A., Lecointre, G., Miquelis, A., Loerstcher, M., Chappaz, R., Brun, G., 2001. Partial combination applied to phylogeny of European cyprinids using the mitochondrial control region. Mol. Phylogenet. Evol. 19, 22–33.
He, S., Gu, X., Mayden, R.L., Chen, W.-J., Conway, K.W., Chen, Y., 2008a. Phylogenetic position of the enigmatic genus *Psilorhynchus* (Ostariophysi: Cypriniformes): evidence from the mitochondrial genome. Mol. Phylogenet. Evol. 47, 419–425.
He, S., Mayden, R.L., Wang, X., Wang, W., Tang, K.L., Chen, W.-J., Chen, Y., 2008b. Molecular phylogenetics of the family Cyprinidae (Actinopterygii: Cypriniformes) as evidenced by sequence variation in the first intron of s7 ribosomal protein-coding gene: further evidence from a nuclear gene of the systematic chaos in the family. Mol. Phylogenet. Evol. 46, 818–829.
Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42, 182–192.
Hoegg, S., Brinkmann, H., Taylor, J.S., Meyer, A., 2004. Phylogenetic timing of the fish-specific genome duplication correlates with the diversification of teleost fish. J. Mol. Evol. 59, 190–203.
Jansa, S.A., Weksler, M., 2004. Phylogeny of muroid rodents: relationships within and among major lineages as determined by IRBP gene sequences. Mol. Phylogenet. Evol. 31, 256–276.
Knight, R.D., Panopoulou, G.D., Holland, P.W., Shimeld, S.M., 2000. An amphioxus Krox gene: insights into vertebrate hindbrain evolution. Dev. Genes Evol. 210, 518–521.
Kocher, T.D., Thomas, W.K., Meyer, A., Edwards, S.V., Pääbo, S., Villablanca, F.X., Wilson, A.C., 1989. Dynamics of mitochondrial DNA evolution in animals: Amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. U.S.A. 86, 6196–6200.
Larhammar, D., Risinger, C., 1994. Molecular genetic aspects of tetraploidy in the common carp *Cyprinus carpio*. Mol. Phylogenet. Evol. 3, 59–68.
Le Comber, S.C., Smith, C., 2004. Polyploidy in fishes: patterns and processes. Biol. J. Linn. Soc. 82, 431–442.
Lê, H.L.V., Perasso, R., Billard, R., 1989. Phylogénie moléculaire préliminaire des “poissons” basée sur l'analyse de séquences d'ARN ribosomique 28 S. C. R. Acad. Sci. Paris 309, 493–498.
Leggatt, R.A., Iwama, G.K., 2003. Occurrence of polyploidy in fishes. Rev. in Fish Biol. and Fish. 13, 237–246.
Li, C., Ortí, G., Zhang, G., Lu, G., 2007. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol. Biol. 7, 44.
Lim, J., Chang, J.L., Tsai, H.J., 1997. A second type of rod opsin cDNA from the common carp (*Cyprinus carpio*). Biochim. Biophys. Acta 1352, 8–12.
Lockhart, P.J., Steel, M.A., Hendy, M.D., Penny, D., 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11, 605–612.
López, J.A., Chen, W.-J., Ortí, G., 2004. Esociform phylogeny. Copeia 2004, 449–464.
Lovejoy, N.R., Collete, B.B., 2001. Phylogenetic relationships of New World needlefishes (Teleostei: Belonidae) and the biogeography of transitions between marine and freshwater habitats. Copeia 2001, 324–338.
Luo, J., Lang, M., Salzburger, W., Siegel, N., Stölting, K.N., Meyer, A., 2006. A BAC library for the goldfish *Carassius auratus auratus* (Cyprinidae, Cypriniformes). J. Exp. Zool. (Mol. Dev. Evol.) 306B, 567–574.
Mabee, P.M., Arratia, G., Coburn, M., Haendel, M., Hilton, E.J., Lundberg, J.G., Mayden, R.L., Rios, N., Westerfield, M., 2007. Connecting evolutionary morphology to genomics using ontologies: a case study from Cypriniformes including zebrafish. J. Exp. Zool. (Mol. Dev. Evol.) 308B, 655–668.
Martin, A.P., Burg, T.M., 2002. Perils of paralogy: using HSP70 genes for Inferring organismal phylogenies. Syst. Biol. 51, 570–587.
Mayden, R.L., Tang, K.L., Conway, K.W., Freyhof, J., Chamberlain, S., Haskins, M., Schneider, L., Sudkamp, M., Wood, R.M., Agnew, M., Bufalino, A., Sulaiman, Z., Miya, M., Saitoh, K., He, S., 2007. Phylogenetic relationships of *Danio* within the order Cypriniformes: a framework for comparative and evolutionary studies of a model species. J. Exp. Zool. (Mol. Dev. Evol.) 308B, 642–654.
Meyer, A., Van de Peer, Y., 2005. From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). BioEssays 27, 937–945.

Miya, M., Nishida, M., 1999. Organization of the mitochondrial genome of a deep-sea fish *Gonostoma gracile* (Teleostei: Stomiiformes): First example of transfer RNA gene rearrangements in bony fishes. Mar. Biotechnol. 1, 416–426.
Miya, M., Nishida, M., 2000. Use of mitogenomic information in teleostean molecular phylogenetics: a tree-based exploration under the maximum-parsimony optimality criterion. Mol. Phylogenet. Evol. 17, 437–455.
Miya, M., Saitoh, K., Wood, R., Nishida, M., Mayden, R.L., 2006. New primers for amplifying and sequencing the mitochondrial ND4/ND5 gene region of the Cypriniformes (Actinopterygii: Ostariophysi). Ichthyol. Res. 53, 75–81.
Mohammad-Ali, K., Eladari, M.E., Galibert, F., 1995. Gorilla and orangutan c-myc nucleotide sequences: inference on hominoid phylogeny. J. Mol. Evol. 41, 262–276.
Müller, H.J., Skerka, C., Bialonski, A., Zipfel, P.F., 1991. Clone pAT 133 identifies a gene that encodes another human member of a class of growth factor-induced genes with almost identical zinc-finger domains. Proc. Natl. Acad. Sci. U.S.A. 88, 10079–10083.
Nelson, J.S., 2006. Fishes of the World, 4 ed. John Wiley and Sons, Inc., Hoboken, NJ.
Nickerson, J.M., Frey, R.A., Ciavatta, V.T., Stenkamp, D.L., 2006. Interphotoreceptor retinoid-binding protein gene structure in tetrapods and teleost fish. Mol. Ver. 12, 1565–1585.
Ohno, S., Muramoto, J., Christian, L., Atkin, N.B., 1967. Diploid-tetraploid relationship among Old World members of the fish family Cyprinidae. Chromosoma (Berl.) 23, 1–9.
Palumbi, S.R., 1996. Nucleic acids II: the polymerase chain reaction. In: Millis, D.M., Mortiz, C., Mable, B.K. (Eds.), Molecular Systematics. Sinauer Associates, pp. 205–247. Sunderland, MA.
Pepperberg, D.R., Okajima, T.L., Wiggert, B., Ripps, H., Crouch, R.K., Chader, G.J., 1993. Interphotoreceptor retinoid-binding protein (IRBP). Molecular biology and physiological role in the visual cycle of rhodopsin. Mol. Neurobiol. 7, 61–85.
Perdices, A., Bohlen, J., Doadrio, I., 2008. The molecular diversity of adriatic spined loaches (Teleostei, Cobitidae). Mol. Phylogenet. Evol. 46, 382–390.
Philippe, H., Lartillot, N., Brinkmann, H., 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22, 1246–1253.
Philippe, H., Sorhannus, U., Baroin, A., Perasso, R., Gasse, F., Adoutte, A., 1994. Comparison of molecular and paleontological data in diatoms suggests a major gap in the fossil record. Mol. Phylogenet. Evol. 7, 247–265.
Rajendran, R.R., Van Niel, E.E., Stenkamp, D.L., Cunningham, L.L., Raymond, P.A., Gonzalez-Fernandez, F., 1996. Zebrafish interphotoreceptor retinoid-binding protein: differential circadian expression among cone subtypes. J. Exp. Biol. 199, 2775–2787.
Rambaut, A., 1996. Sequence alignment editor version 1.0 α1. Available from http://evolve.zoo.ox.ac.uk/Se-Al/Se-Al.html.
Rhymer, J.M., Simberloff, D., 1996. Extinction by hybridization and introgression. Annu. Rev. Ecol. Syst. 27, 83–109.
Rieseberg, L.H., 1998. Molecular ecology of hybridization. In: Carvalho, G.R. (Ed.), Advances in Molecular Ecology. IOS Press, Amsterdam, Netherlands, pp. 243–265.
Rüber, L., Kottelat, M., Tan, H.H., Ng, P.K.L., Britz, R., 2007. Evolution of miniaturization and the phylogenetic position of *Paedocypris*, comprising the world's smallest vertebrate. BMC Evol. Biol. 7, 38.
Saint, K.M., Austin, C.C., Donnellan, S.C., Hutchinson, M.N., 1998. C-mos, a nuclear marker useful for squamate phylogenetic analysis. Mol. Phylogenet. Evol. 10, 259–263.
Saitoh, K., 2003. Mitotic and meiotic analyses of the 'large race' of *Cobitis* striata, a polyploid spined loach of hybrid origin. Folia Biol. 51 (suppl.), 101–105.
Saitoh, K., Kim, I.S., Lee, E.H., 2004. Mitochondrial gene introgression between spined loaches via hybridogenesis. Zoolog. Sci. 21, 795–798.
Saitoh, K., Sado, T., Mayden, R.L., Hanzawa, N., Nakamura, K., Nishida, M., Miya, M., 2006. Mitogenomic evolution and interrelationships of the Cypriniformes (Actinopterygii: Ostariophysi): the first evidence toward resolution of higher-level relationships of the world's largest freshwater fish clade based on 59 whole mitogenome sequences. J. Mol. Evol. 63, 826–841.
Saitou, N., Nei, M., 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. 4, 406–425.
Schilling, T.F., Knight, R.D., 2001. Origins of anteroposterior patterning and Hox gene regulation during chordate evolution. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 356, 1599–1613.
Schilling, T.F., Webb, J., 2007. Considering the zebrafish in a comparative context. J. Exp. Zool. (Mol. Dev. Evol.) 308B, 515–522.
Schneider, H., Sampaio, I., Harada, M.L., Barroso, C.M., Schneider, M.P., Czelusniak, J., Goodman, M., 1996. Molecular phylogeny of the New World monkeys (Platyrrhini, Primates) based on two unlinked nuclear genes: IRBP intron 1 and epsilon-globin sequences. Am. J. Phys. Anthropol. 100, 153–179.
Siebert, P.D., Chenchik, A., Kellogg, D.E., Lukyanov, K.A., Lukyanov, S.A., 1995. An improved PCR method for walking in uncloned genomic DNA. Nucleic Acids Res. 23, 1087–1088.
Simons, A.M., Mayden, R.L., 1998. Phylogenetic relationships of the western North American phoxinins (Actinopterygii: Cyprinidae) as inferred from mitochondrial 12 S and 16 S ribosomal RNA sequences. Mol. Phylogenet. Evol. 9, 308–329.
Šlechtová, V., Bohlen, J., Tan, H.H., 2007. Families of Cobitoidea (Teleostei; Cypriniformes) as revealed from nuclear genetic data and the position of the mysterious genera *Barbucca, Psilorhynchus, Serpenticobitis* and *Vaillantella*. Mol. Phylogenet. Evol. 44, 1358–1365.
Slowinski, J.B., Page, R.D.M., 1999. How should species phylogenies be inferred from sequence data? Syst. Biol. 48, 814–825.
Smith, G.R., 1992. Phylogeny and biogeography of the Catostomidae, freshwater fishes of North America and Asia. In: Mayden, R.L. (Ed.), Systematics, historical ecology and North American freshwater fishes. Standford University Press, Stanford, pp. 778–813.
Smith, M.R., Shivji, M.S., Waddell, V.G., Stanhope, M.J., 1996. Phylogenetic evidence from the IRBP gene for the paraphyly of toothed whales, with mixed support for Cetacea as a suborder of Artiodactyla. Mol. Biol. Evol. 13, 918–922.
Stanhope, M.J., Smith, M.R., Waddell, V.G., Porter, C.A., Shivji, M.S., Goodman, M., 1996. Mammalian evolution and the interphotoreceptor retinoid binding protein (IRBP) gene: convincing evidence for several superordinal clades. J. Mol. Evol. 43, 83–92.
Sun, Z., Shi, K, Su, Y., Meng, A., 2002. A novel zinc finger transcription factor resembles krox-20 in structure and in expression pattern in zebrafish. Mech. Dev. 114, 133–135.
Swofford, D.L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4., 4 ed. Sinauer Associates, Sunderland, Massachusetts.
Sytchevskaya, E.K., 1986. Palaeogene freshwater fish fauna of the USSR and Mongolia. The Joint Soviet-Mongolia Paleontological Expedition 29, 1–157.
Tang, Q., Liu, H., Mayden, R.L., Xiong, B., 2006. Comparison of evolutionary rates in the mitochondrial DNA cytochrome b gene and control region and their implications for phylogeny of the Cobitoidea (Teleostei: Cypriniformes). Mol. Phylogenet. Evol. 39, 347–357.
Taylor, J.S., Van de Peer, Y., Braasch, I., Meyer, A., 2001. Comparative genomics provides evidence for an ancient genome duplication event in fish. Philos. Trans. R. Soc. Lond., B, Biol. Sci. 356, 1661–1679.
Tsigenopoulos, C.S., Ráb, P., Naran, D., Berrebi, P., 2002. Multiple origins of polyploidy in the phylogeny of southern African barbs (Cyprinidae) as inferred from mtDNA markers. Heredity 88, 466–473.
Uyeno, T., Smith, G.R., 1972. Tetraploid origin of the karyotype of catostomid fishes. Science 175, 644–646.
Wang, X., Li, J., He, S., 2007. Molecular evidence for the monophyly of East Asian groups of Cyprinidae (Teleostei: Cypriniformes) derived from the nuclear recombination activating gene 2 sequences. Mol. Phylogenet. Evol. 42, 157–170.
Yu, X., Zhou, T., Li, K., Li, Y., Zhou, M., 1987. One the karyosystematics of cyprinid fishes and a summary of fish chromosome studies in China. Genetica 72, 225–236.