RESEARCH ARTICLE

Complete mitochondrial genome of threatened mahseer *Tor tor* (Hamilton 1822) and its phylogenetic relationship within Cyprinidae family

A. PAVAN-KUMAR[1]*, SUDHANSHU RAMAN[1], PRAKASH G. KORINGA[2], NAMRATA PATEL[2], TEJAS SHAH[2], RAJEEV K. SINGH[3], GOPAL KRISHNA[1], C. G. JOSHI[2], P. GIREESH-BABU[1] and APARNA CHAUDHARI[1]

[1]*Fish Genetics and Biotechnology Division, ICAR-Central Institute of Fisheries Education (Deemed University) ICAR, Mumbai 400 061, India*
[2]*Department of Animal Biotechnology, College of Veterinary Science and Animal Husbandry, Anand Agricultural University, Anand 388 110, India*
[3]*ICAR-National Bureau of Fish Genetic Resources, Canal Ring Road, Dilkusha PO, Lucknow 226 002, India*

Abstract

The mahseers (Tor, Neolissochilus and Naziritor) are an important group of fishes endemic to Asia with the conservation status of most species evaluated as threatened. Conservation plans to revive these declining wild populations are hindered by unstable taxonomy. Molecular phylogeny studies with mitochondrial genome have been successfully used to reconstruct the phylogenetic tree and to resolve taxonomic ambiguity. In the present study, complete mitochondrial genome of *Tor tor* has been sequenced using ion torrent next-generation sequencing platform with coverage of more than 1000×. Comparative mitogenome analysis shows higher divergence value at *ND1* gene than *COI* gene. Further, occurrence of a distinct genetic lineage of *T. tor* is revealed. The phylogenetic relationship among mahseer group has been defined as *Neolissochilus hexagonolepis* ((*T. sinensis* (*T. putitora, T. tor*), (*T. khudree*, *T. tambroides*)).

[Pavan-Kumar A., Raman S., Koringa P. G., Patel N., Shah T., Singh R. K., Krishna G., Joshi C. G., Gireesh-Babu P. and Chaudhari A. 2016 Complete mitochondrial genome of threatened mahseer *Tor tor* (Hamilton 1822) and its phylogenetic relationship within Cyprinidae family. *J. Genet.* **95**, xx–xx]

Introduction

Tor mahseer, *Tor tor* (Hamilton 1822) is one of the most important food and game fishes of India. It inhabits rivers of 'Tor region' (600–1200 m) and plains with rocky bottom (Singh and Kumar 2000). This species has been reported from the Indus, Ganga (including sub-Himalayan range), Brahmaputra, Godavari and Krishna riverine systems and forms major fishery in Narmada river (Desai 2003; Jayaram 2005; Lal *et al.* 2013). However, due to overfishing, habitat destruction and other anthropogenic activities, the natural populations of this species are declining and the conservation status is evaluated as 'near threatened' (IUCN 2015). For formulating effective management and conservation measures, organisms need to be identified/delimited accurately. But conversely, the systematics of mahseer fishes are highly ambiguous due to phenotype plasticity leading to taxonomic instability (Mohindra *et al.* 2007; Khare *et al.* 2014).

Taxonomically, mahseers are defined as carps with big head and scales, two pairs of barbles, fleshy lips that are continuous at the angles of the mouth with an interrupted fold or groove across the lower jaw (Desai 2003). Initially, Hamilton (1822) classified mahseers under the genus '*Cyprinus*', but later, Gray (1834) proposed new genus '*Tor*' for these species. Rainboth (1996) classified Tor species under the family Cyprinidae, subfamily Cyprininae, tribe Cyprinini. Thai *et al.* (2007) placed Tor species under tribe Barbini (interchangeably referred to as subfamily Barbinae by various authors). The tribe name became more contentious after Yang *et al.* (2010) showed the polyphyletic nature of tribe Barbini within Cyprininae.

In India, information on mahseer genetic resources is inconclusive and several studies using limited markers (COI and D-loop) showed misapplication of different species names to diverse morphotypes of conspecific individuals (Laskar *et al.* 2013; Khare *et al.* 2014). As a first step to address the taxonomic ambiguity of mahseer species, several studies have briefly described complete mitochondrial genomes of *T. putitora* (Sati *et al.* 2014), *T. khudree* (Raman *et al.* 2015) and

*For correspondence. E-mail: pavankumar@cife.edu.in.

Keywords. mahseer; next generation sequencing; phylogeny; mitochondrial genome.

T. tor (Kumar *et al.* 2015) with limited phylogenetic analysis. However, molecular phylogeny with more number of characters would generate accurate phylogenetic tree than limited number of characters. Kumar *et al.* (2015) reported complete mitochondrial genome of *T. tor*, but their study was limited to brief description of mitogenome and partial phylogeny. Further, molecular phylogeny with conspecific individuals from different geographical locations would give insight on speciation and evolution of genetic lineages. Additionally, comparative analysis of mitogenome could provide leads on pattern of amino acid usage and gene evolution among mahseer fishes. With this back ground, the present study was carried with an objective of characterizing complete mitochondrial genome of *T. tor* to infer its phylogenetic position within mahseer group.

Materials and methods

Sample collection

Two samples of *T. tor* (Hamilton 1822) (figure 1 in electronic supplementary material at http:www.ias.ac.in/jgenet/) were collected from Narmada river (Hoshangabad, 77°74′E, 22°76′N) during the month of April 2015. The species were identified at the field by observing morphological and meristic characters (Desai 2003). Voucher specimens were prepared by preserving the specimen in absolute alcohol with proper labelling. Fins were collected under aseptic conditions and preserved in absolute alcohol and kept at −80°C for further analysis. For phylogenetic and comparative analyses, reported mitochondrial DNA sequences of mahseer species were downloaded from NCBI, GenBank.

DNA extraction and long PCR amplification

Total genomic DNA was isolated from the fins using standard phenol/chloroform method. Whole mitochondrial genome was amplified in two amplicons using two sets of reported primers L-1231-Leu: 5′-GGTCTTAGGAACCAAAAACTCTTGGTGCAA-3′ and S-LA-16S-H: 5′-TGCACCATTAGGATGTCCTGATCCAACATC-3′; H-12321-Leu: 5′-TTGCACCAAGAGTTTTTGGTTCCTAAGACC-3′ and S-LA-16S-L: 5′-GATGTTGGATCAGGACATCCCAATGGTGCA-3′ (Kim and Lee 2004). Long range polymerase chain reaction (PCR) was carried out in a 20 μL reaction mixture containing 9.7 μL sterile distilled water, 2 μL of 10× clean α-Taq reaction buffer, 4 μL of 5× HQ buffer, 2 μL of dNTPs mix, 1 μL of each primer and 0.3 μL of clean α-Taq DNA polymerase (Krishgen Biosystems, USA) and 2 μL of template DNA using a Bio-Rad iCycler PCR machine in 0.2 mL PCR tubes. The thermocycler was programmed for initial denaturation at 95°C for 3 min, followed by 40 cycles of 95°C for 20 s and annealing and extension temperature at 68°C for 15 min. The PCR products were visualized on 1.0% agarose gels and the amplicons were purified by gel extraction kit (Qiagen, Valencia, USA) following the manufacturer's protocol.

Amplicon sequencing by ion torrent NGS platform

DNA (amplicon) quality and quantity were measured using Qubit® 2.0 fluorometer. Around 500 ng of amplified product was digested using ultrasonification to get the desired fragment size of 400 bp. The fragment's quality and quantity were verified by Bioanalyzer and were purified by AMPure bead purification. Adapter ligation, size selection, nick repair and amplification were performed as described in the ion torrent protocol. Library preparation was carried out using the Ion Xpress™ Fragement Library kit (ThermoFisher Scientfic, MA, USA), with 500 ng of DNA and a different barcode adaptor was used for each library (barcode series 53 and 54 for *T. tor* specimens 1 and 2, respectively). After appropriate dilution, the two libraries were pooled in equimolar concentration. Emulsion PCR and enrichment steps were carried out using the Ion Xpress™ Template kit and associated protocol. Libraries were sequenced using 314 chip and the reads obtained from ion torrent sequencer were grouped according to their barcodes into two separate files.

Data analysis

After checking the quality, raw reads were trimmed, aligned to contigs and the resulted contigs were mapped against *T. putitora* reference mitogenome (GenBank: NC_021755.1), using CLC Genomic Workbench ver. 5.0. Assembled mitogenome was annotated using MITOFISH (MitoAnnotator) online mitochondrial genome annotation server (Iwasaki *et al.* 2013). The ribosonal RNA (rRNA) gene boundaries were confirmed by alignment with other reported fish mitogenomes. The transfer RNA (tRNA) structures were predicted using 'tRNAscan' web server (Schattner *et al.* 2005) with a search mode 'tRNAscan only' using vertebrate mitochondrial genetic code. The origin of replication on L strand O_L region was predicted using online 'Mfold' software (Zuker 2003). The sequences were analysed for base composition, variation using MEGA6 (Tamura *et al.* 2013). The DAMBE5 software was used to test the sequence substitution saturation (Xia 2013), and DnaSp5 software was used to calculate the divergence between mitogenomes using sliding window analysis (Rozas 2009). Codon usage analysis was done using MEGA6 and the heatmap of codon usage was prepared using online tool 'CIMminer' (http://discover.nci.nih.gov/cimminer/home.do). The evolutionary models were estimated using jModelTest 2 software (Guindon and Gascuel 2003; Darriba *et al.* 2012). PAUP ver. 4.0 (Swofford 2003) was used to reconstruct the phylogenetic tree using parsimony (MP), maximum likelihood (ML) and neighbour joining (NJ) methods. Nonparametric bootstrap support for each node of the tree was estimated using 100 heuristic bootstrap replicates. Bayesian inference (BI) was implemented with MR BAYES ver. 3.1.2 (Huelsenbeck and Ronquist 2003) with the prior probability of a flat Dirichlet distribution for the substitution rates and stationary nucleotide frequencies. The Bayesian analysis was performed for five million generations with sampling every 1000 generations.

Results

The mitochondrial genome of *T. tor* was amplified into two overlapping amplicons of 8–9 kb size (figure 2 in electronic supplementary material). These amplicons were purified and processed for sequencing on ion torrent NGS platform. Ion torrent sequencer has generated 347,885 and 137,605 reads with a mean length of 238 bp for *T. tor* specimens 1 and 2, respectively (figure 3 in electronic supplementary material). These sequences were assembled into contigs and mapped against reference sequence (*T. putitora*, NC021755.1) using CLC Genomics workbench. The coverage depth was 4213× and 1669× for *T. tor* specimens 1 and 2, respectively (table 1).

Genome content and organization

The total length of *T. tor* mitochondrial genome was 16571 bp with 13 protein-coding genes, 22 tRNAs, two rRNAs and a noncoding control region (figure 4 in electronic supplementary material). These sequences were submitted to NCBI GenBank with accession numbers of KR868704–KR868705. Most of the genes were encoded on the heavy strand (H-strand), whereas only ND6 and eight tRNA (glutamine, alanine, asparagine, cysteine, tyrosine, serine (UCN), glutamic acid and proline) genes were encoded on the light strand (L-strand). Gene overlapping was observed between ATPase 8 and ATPase 6 (7 bp); ND4L and ND4 (7 bp); ND5 and ND6 (4 bp); CytB and tRNA^{Thr} (17 bp); and tRNA^{Ile} and tRNA^{Gln} (2 bp). Around 15 intergenic spacer sequences were found with a total size of 29 bp (table 2).

The overall base composition was found to be A: 31.91, C: 27.38, G: 15.64 and T: 25% with a high A+T content. The overall GC-skews and AT-skews of the H-strand were −0.274 and 0.126, respectively, and it confirmed that the heavy strand compositional bias towards C over G nucleotides and a slight excess of A over T nucleotides.

Protein-coding genes

As observed in all other vertebrates, 13 protein coding genes were also present in *T. tor* mitochondrial genome. These genes include seven subunits of the NADH ubiquinone oxidoreductase complex (ND1-6, ND4L), three subunits of the cytochrome c oxidase (COI–III), one subunit of the ubiquinol cytochrome b oxidoreductase complex (Cyt b), and two subunits of ATP synthases (ATP-6 and ATP-8). The total length of these genes was 11,408 bp, accounting for 68.8% of the whole mitogenome.

All the 12 genes encoded on the H-strand showed a marked similarity in nucleotide composition with an antiguanine bias, whereas ND6 encoded on the L-strand exhibited an opposite trend (33.9% for G). Among the protein-coding genes, A+T content varied from 52.1 (ND1) to 61.8% (ATPase 8). Comparative analyses of the base composition at each codon base position of the 12 protein-coding genes showed that the codon third base position possessed the highest A+T content (63%) with a strong bias against the use of G (6.8%).

The typical initiation codon ‘ATG’ was used by 12 of 13 protein coding genes, while *COI* gene has ‘GTG’ as the start codon. A different pattern of codon usage was observed for stop codons such as incomplete codons to stop the protein translation (Oh *et al.* 2007). Six of 13 protein coding genes are terminated with incomplete codons of T− − (CytB, ND2, ND3, ND4, COII and COIII) and the remaining six genes (ND1, COI, ATPase 6, ND4L, ND5 and ND6) have TAA termination codon. ATPase 8 gene has termination codon of ‘TAG’. Among all the protein-coding genes, *ND5* gene (1820 bp) is the longest one, while the shortest is ATPase 8 (164 bp).

A total of 3799 codons were identified including all protein-coding genes. The most and least frequently used amino acids are leucine (16.48%) and cysteine (0.66%), respectively. Analysis of overall codon usage data indicated that codons ending with T or A were used more frequently than those ending with C or G. This observation is in agreement with the overall high A+T content (56%) of the entire mitogenome. In addition, the most frequently used codon is CTA (258/3799), while CGG is the least frequently used codon (4/3799; table 3).

Transfer and ribosomal RNA genes

The mitogenome of *T. tor* contained 22 typical tRNA genes interspersed between the protein-coding genes and rRNA genes. Twenty tRNA genes code for 20 different amino acids while there is an additional tRNA gene for serine (GCU)

Table 1. Details of reference-guided sequence mapping for *T. tor*.

	Species	Count	Percentage of reads	Average length	Number of bases	Coverage
Total reads	*T. tor* specimen 1 (KR868704)	347,885	–	238	82,796,630	–
	T. tor specimen 2 (KR868705)	137,605	–	238	32,749,990	–
Reads after trimming	*T. tor* specimen 1 (KR868704)	337,753	97.08	238	80,385,214	–
	T. tor specimen 2 (KR868705)	127,412	93	238	30,324,056	–
Mapped reads	*T. tor* specimen 1 (KR868704)	248,956	84.37	245	69,814,220	1669×
	T. tor specimen 2 (KR868705)	115,229	90.44	240	27,654,960	4213×
Not mapped reads	*T. tor* specimen 1 (KR868704)	52,797	15.63	203	18,025,791	–
	T. tor specimen 2 (KR868705)	12,138	9.56	224	2,718,912	–
Reference	*T. putitora*	–	–	16,576	16,576	–

Table 2. Organization of the complete mitochondrial genome of *T. tor*.

Gene	Position		Size		Codon		Gene overlapping/	Strand
	From	To	Nucleotide (bp)	No. of amino acids	Start	Stop	intergenic nucleotides	
$tRNA^{Phe}$ (trnF-gaa)	1	69	69	–	–	–	0	H
12S rRNA	70	1025	955	–	–	–	0	H
$tRNA^{Val}$ (trnV-uac)	1026	1097	71	–	–	–	0	H
16S rRNA	1098	2774	1676	–	–	–	0	H
$tRNA^{Leu}$ (trnL-uaa)	2775	2850	75	–	–	–	1	H
ND1	2852	3826	974	324	ATG	TAA	4	H
$tRNA^{Ile}$ (trnl-gau)	3831	3902	71	–	–	–	−2	H
$tRNA^{Gln}$ (trn-Q-uug)	3901	3971	70	–	–	–	1	L
$tRNA^{Met}$ (trn-M-cau)	3973	4041	68	–	–	–	0	H
ND2	4042	5086	1044	–	ATG	T– –	0	H
$tRNA^{trp}$ (trnW-uca)	5087	5158	71	–	–	–	2	H
$tRNA^{Ala}$ (trnA-ugc)	5161	5229	68	–	–	–	1	L
$tRNA^{Asn}$ (trnN-guu)	5231	5303	72	–	–	–	0	L
O_L	5304	5339	36	–	–	–	0	L
$tRNA^{Cys}$ (trnC-gca)	5340	5405	65	–	–	–	2	L
$tRNA^{Tyr}$ (trnY-gua)	5408	5478	70	–	–	–	1	L
COI	5480	7030	1550	516	GTG	TAA	0	H
$tRNA^{ser}$ (trnS-uga)	7031	7101	70	–	–	–	1	L
$tRNA^{Asp}$ (trnD-guc)	7103	7174	71	–	–	–	6	H
COII	7181	7871	690	234	ATG	T– –	0	H
$tRNA^{Lys}$ (trnK uuu)	7872	7947	75	–	–	–	1	H
ATPase 8	7949	8113	164	54	ATG	TAG	−7	H
ATPase 6	8107	8789	682	227	ATG	TAA	0	H
COIII	8790	9573	783	279	ATG	T– –	0	H
$tRNA^{Gly}$ (trnG ucc)	9574	9645	71	–	–	–	0	H
ND3	9646	9994	348	116	ATG	T– –	0	H
$tRNA^{Arg}$ (trn-R-ucg)	9995	10064	69	–	–	–	0	H
ND4L	10065	10361	296	98	ATG	TAA	−7	H
ND4	10355	11735	1380	475	ATG	T– –	0	H
$tRNA^{His}$ (trn-H-gug)	11736	11804	68	–	–	–	0	H
$tRNA^{Ser}$ (trn-S-gcu)	11805	11873	68	–	–	–	1	H
$tRNA^{Leu}$ (trnL-uag)	11875	11947	72	–	–	–	3	H
ND5	11951	13774	1823	607	ATG	TAA	−4	H
ND6	13771	14292	521	173	ATG	TAA	3	L
$tRNA^{Glu}$ (trn-E-uuc)	14296	14364	68	–	–	–	1	L
Cytb	14366	15526	1160	386	ATG	T– –	−17	H
$tRNA^{Thr}$ (trnT-ugu)	15510	15581	71	–	–	–	0	H
$tRNA^{Pro}$ (trnP-ugg)	15581	15650	69	–	–	–	1	L
D loop/control region	15651	16571	920	–	–	–	–	–

and leucine (CUN). The size of tRNA genes varied from 65 (tRNA^{Cys}) to 75 bp (tRNA^{Leu}). Except tRNA^{Ser} (UGC), the remaining tRNAs have showed typical clover leaf secondary structure. The majority of predicted tRNA secondary structures had a common feature with 7 bp in the amino acid stem, 5 bp in the TΨC stem, 5 bp in the anticodon stem, and 4 bp in the DHU stem (figure 5 in electronic supplementary material).

The length of 12S and 16S rRNA genes was found to be 955 and 1676 bp, respectively. The two ribosomal RNA genes are separated by tRNA^{Val} and this entire gene cassette is flanked by tRNA^{Phe} and tRNA^{Leu}. The base composition of the 12S and 16S rRNAs genes are A, 35; G, 20.6; C, 25.4; T, 19.0%. The A+T content of rRNA genes is 54%, which is slightly lower than the protein-coding gene (56%), but is higher in comparison with tRNA genes (55.7%).

Noncoding regions

Two noncoding regions; an origin of O$_L$ and control region were observed in *T. tor* mitochondrial genome. The O$_L$ (36 bp) was located within the cluster of five tRNA genes (WANCY) region between tRNA^{Asn} and tRNA^{Cys}. This region showed to fold into a stable stem–loop secondary structure with 13 bp in the stem and 10 bp in the loop (figure 6 in electronic supplementary material). Alignment of this region with other cyprinidae family species showed an overrepresentation of pyrimidines (C, T) in the 5′ side of sequence and a conserved motif of 'GCGGG' at 3′ side (figure 7 in electronic supplementary material). The AT content of O$_L$ was 38.8%.

Even though, D-loop (control region) is a highly variable region, it contains few conserved sequences such as termination

Table 3. Codon usage in *T. tor* mitochondrial protein-coding genes.

Amino acid	Codon	Number/frequency (%)	Amino acid	Codon	Number/frequency (%)
Phenylalanine (Phe)	TTT	85/2.24	Tyrosine (Tyr)	TAT	70/1.84
	TTC	141/3.71		TAC	43/1.13
Leucine (Leu)	TTA	118/3.11	Stop	TAA	4/0.11
	TTG	21/0.55		TAG	1/0.03
	CTT	82/2.16	Histidine (His)	CAT	24/0.63
	CTC	100/2.63		CAC	81/2.13
	CTA	258/6.79	Glutamine (Gln)	CAA	92/2.42
	CTG	47/1.24		CAG	8/0.21
Isoleucine (Ile)	ATT	155/4.08	Asparagine (Asn)	AAT	36/0.95
	ATC	138/3.63		AAC	85/2.24
Methinone (Met)	ATA	131/3.45	Lysine (Lys)	AAA	71/1.87
	ATG	43/1.13		AAG	7/0.18
Valine (Val)	GTT	58/1.53	Aspartic acid (Asp)	GAT	23/0.61
	GTC	39/1.03		GAC	53/1.39
	GTA	100/2.63	Glutamic acid (Glu)	GAA	86/2.26
	GTG	22/0.58		GAG	17/0.45
Serine (Ser)	TCT	24/0.63	Cysteine (Cys)	TGT	6/0.16
	TCC	60/1.58		TGC	19/0.50
	TCA	96/2.53	Tryptophan (Trp)	TGA	111/2.92
	TCG	7/0.18		TGG	9/0.24
Proline (Pro)	CCT	26/0.68	Arginine (Arg)	CGT	11/0.29
	CCC	58/1.53		CGC	11/0.29
	CCA	121/3.18		CGA	49/1.29
	CCG	9/0.24		CGG	4/0.11
Threonine (Thr)	ACT	35/0.92	Serine (Ser)	AGT	16/0.42
	ACC	124/3.26		AGC	34/0.89
	ACA	146/3.84	Stop	AGA	0/0.00
	ACG	7/0.18		AGG	0/0.00
Alanine (Ala)	GCT	58/1.53	Glycine (Gly)	GGT	32/0.84
	GCC	132/3.47		GGC	41/1.08
	GCA	129/3.39		GGA	132/3.47
	GCG	10/0.26		GGG	43/1.13

associated sequence (TAS) and conserved sequence blocks (CSB). The TAS was identified by the presence of unique sequence 'TACATAAAAYYYAAT' in the D-loop region. However, in the present study, *T. tor* showed a partial conserved region 'TACATAATAT' in the D-loop. Likewise CSBI, CSBII and CSBIII were also identified based on conserved sequences. Additionally dinucleotide repeats $(TA)_{13}$ were identified in the control region (figure 8 in electronic supplementary material). The AT content (68%) of D-loop region was higher than the average value of complete mitochondrial genome.

Comparative analysis among mahseer fishes

The mitogenomes of seven mahseer species were compared to find the most variable gene fragment and the pattern of codon usage. Apart from the D-loop, more variable sites were observed in *ND1* and minimum in *COIII* genes (table 1 in electronic supplementary material). However, this variation was much less at protein level due to the degeneracy of the genetic code.

Sliding window analysis showed an unexpectedly large pair-wise nucleotide difference between *T. tor* (KP795444) and *T. tor* (present study: KR868704) (figure 1). Heat map analysis showed a remarkable difference in codon usage between *Neolissochilus hexagonolepis* and other species of 'Tor'. Within Tor species, *T. tambroides* and *T. sinensis* showed similar trend while a slight difference was observed between *T. tor* (present study: KR868704) and *T. tor* (KP795444) (figure 2). The mitochondrial D-loop region showed tandem repeats of the 'TA' with varying repeat numbers. *T. tambroides* showed minimum number of repeats (11), while maximum number of repeats (14) was found in *T. sinensis*. Remaining species of *Tor* showed 13 repeats of 'TA' in the D-loop region.

Phylogenetic analysis

Twelve mitochondrial protein-coding genes excluding ND6 were concatenated (~10,800 bp) and used for phylogenetic analysis. Base frequency of concatenated sequences was not homogeneous across the taxa ($P = 0.00$). Sequence saturation analysis showed the increase of frequency of both transitions and transversions linearly along with the divergence value (figure 3). The index of substitution saturation (ISS = 0.3598) was significantly lower than the critical index of substitution saturation ($ISS_c = 0.7599$). Based on the log-likelihood ratios and Akaike information criterion

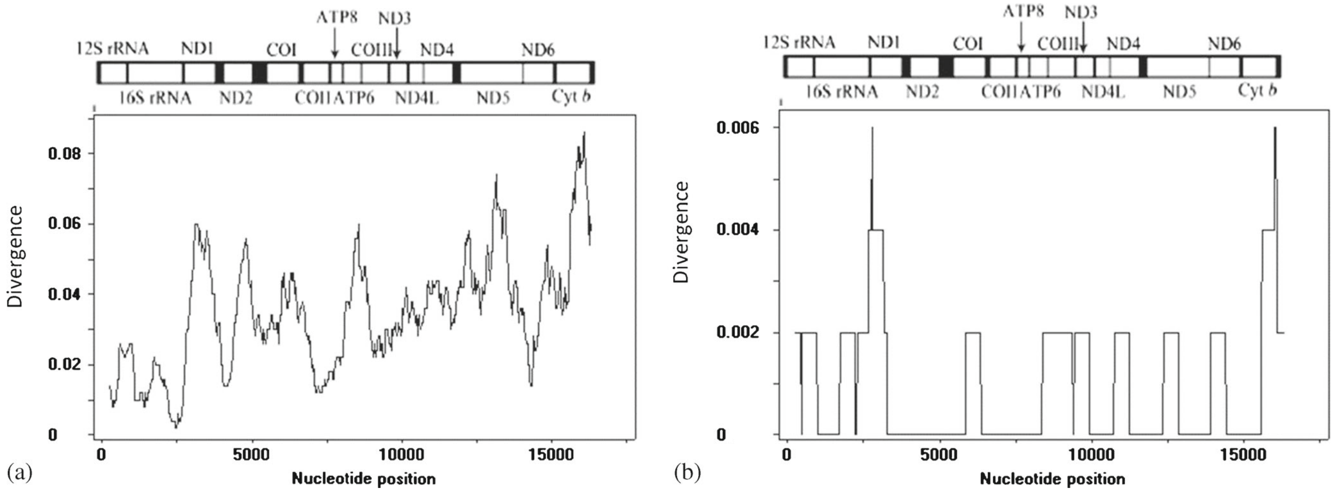

Figure 1. Plot of divergences among mtDNA sequences. (a) The sliding window analysis calculates the divergence between *T. tor* specimens (present study: KR868704 and reported specimen: KP795444). (b) Divergence plot between *T. tor* specimens (present study specimens: KR868704-05). The bar at the top illustrates the position of protein-coding genes and rRNAs, and the tRNAs are represented as black boxes.

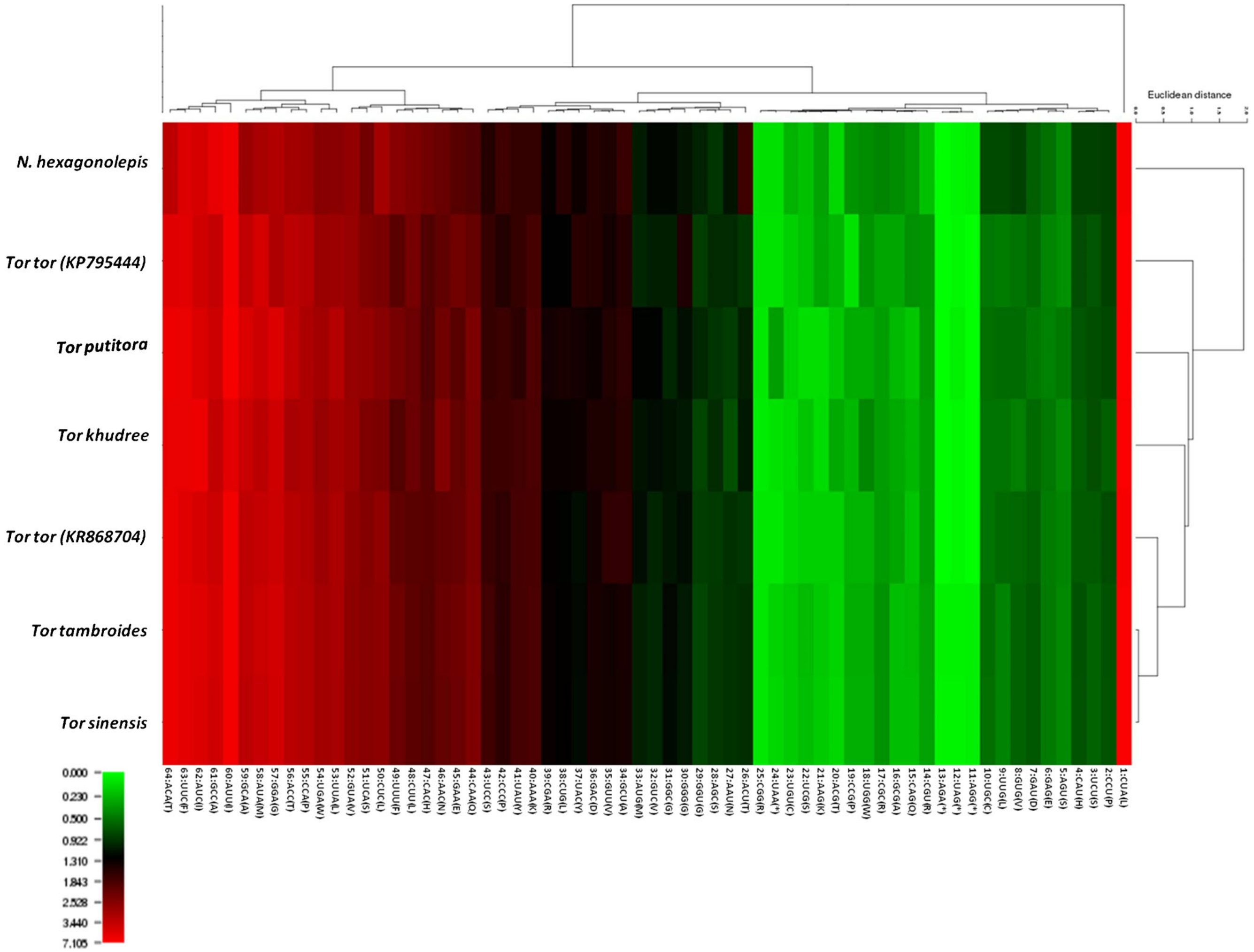

Figure 2. Heatmap of codon usage.

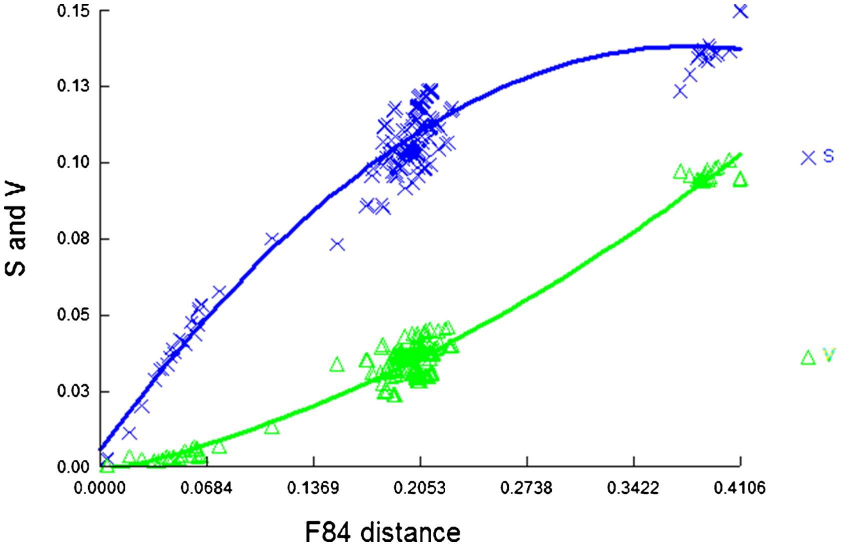

Figure 3. Transition (S) and transversion (V) saturation plots for dataset (protein-coding genes).

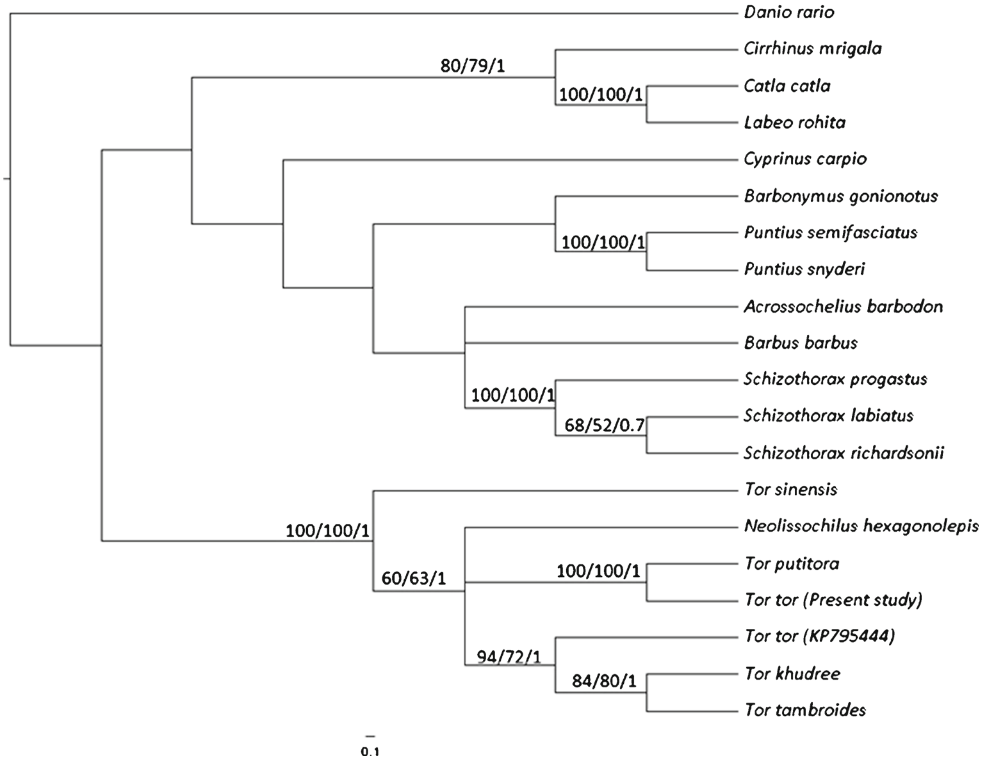

Figure 4. Phylogenetic analysis of mahseers and other cyprinids using concatenated mitochondrial protein-coding genes. The phylogenetic relationship was analysed by maximum parsimony (MP), maximum likelihood (ML) and Bayesian inferences (BI) methods. Tree topology produced by different methods is similar. Bootstrap values are in MP/ML/BI order.

(AIC) values, evolutionary model ‘general time reversible with gamma distribution (GTR+I+G)’ was selected for phylogeny tree reconstruction using maximum likelihood and Bayesian inference methods. Since the base frequency was heterogeneous across taxa, distance matrix based methods using LogDet transformation were also used for phylogeny reconstruction through neighbour-joining (NJ) method.

Except NJ, all other character based methods (MP, ML and BI) revealed similar tree topologies with significant bootstrap values (figure 4). In all tree topologies, mahseers (genus *Tor* and *Neolissochilus*) formed a separate clade within cyprininae subfamily. However, the close relatives of this group were not found as the sample size was less (all species of cyprininae subfamily were not included in

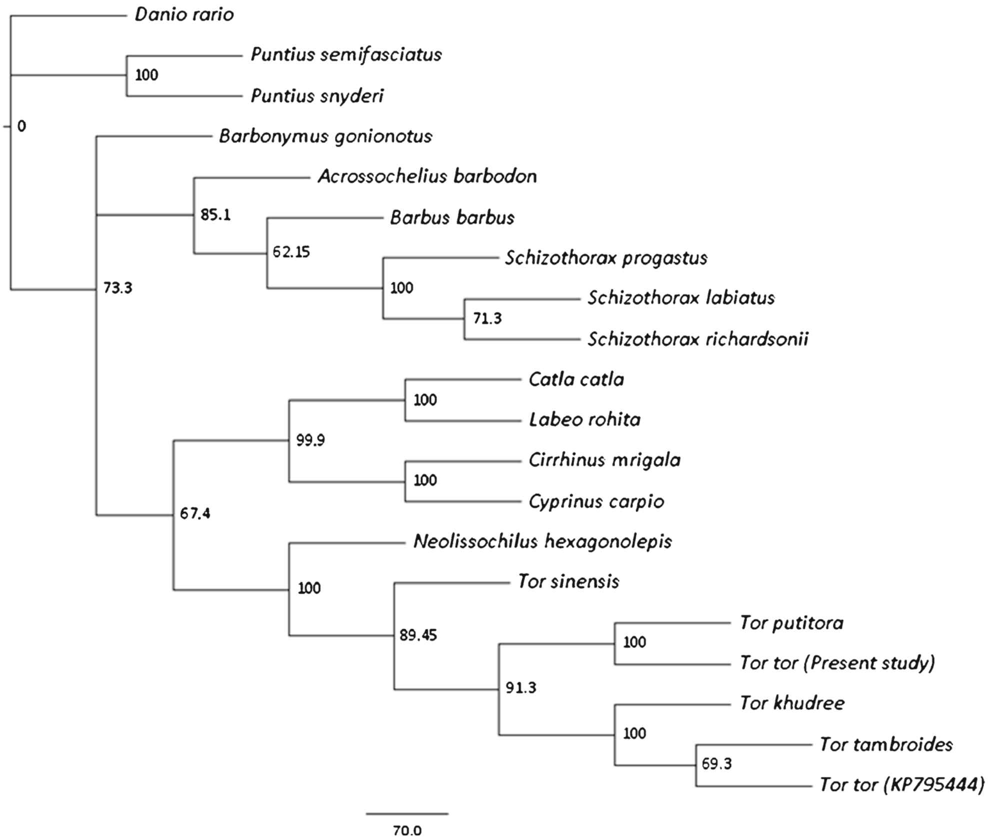

Figure 5. Phylogenetic analysis of mahseers and other cyprinids using concatenated mitochondrial protein-coding genes. The phylogenetic relationship was analysed by NJ method based on LogDet distance values.

the analysis). Within mahseer group, *T. putitora* and *T. tor* (present study: KR868704); *T. tambroides* and *T. khudree* were formed as sister species. *N. hexagonolepis* was formed as a separate taxon within the clade of tor species with low bootstrap value.

The phylogeny tree constructed by NJ method showed different tree topologies especially for mahseer group (figure 5). In this tree topology, *N. hexagonolepis* formed as a basal and distinct taxon/sister group to species of *Tor*. Further, sister group relationship was observed between *T. putitora* – *T. tor* (present study: KR868704) and *T. tambroides* – *T. tor* (KP795444) with less bootstrap value. *T. khudree* formed as a sister species to the group of *T. tor* and *T. tambroides* with significant bootstrap value.

Discussion

Traditionally, fish mitochondrial genome was sequenced through Sanger sequencing by cloning different fragments of mitochondrial genome. However, these methods are cumbersome and are time consuming. High throughput sequencing technologies with bioinformatics pipeline have revolutionized the genomics research. Using next-generation sequencing technologies, it is now possible to sequence mitochondrial genome of multiple species in a single reaction through multiplexing (Tang *et al.* 2014). Several NGS platforms such as Roche 454 FLX, Illumina and ion torrent have been used to sequence mitochondrial genomes of fishes (Sahoo *et al.* 2015a; Kushwaha *et al.* 2015). In the present study, the amplicons (8–9 kb) of mitochondrial DNA were sequenced with high coverage using ion torrent platform.

The mitochondrial genome size is not equal among mahseer group and in the present study, the mitochondrial genome size of *T. tor* is less than *T. tambroides* (16,690 bp) and higher than *N. hexagonolepis* (16,563 bp) and *T. tor* (16,554 bp; KP795444). This variation could be due to the presence/absence of intergenic regions and the length of control region (Chen *et al.* 2013; Marshall *et al.* 2013).

The gene size and arrangement are similar to that of other vertebrates (Campbell *et al.* 2014; Norfatimah *et al.* 2014). The nucleotide composition of the *T. tor* mitogenome is biased towards A and T, with A (31.91%) being the most abundant nucleotide while G (15.64%) is the least abundant. The A+T content at codon's third base position of the protein-coding genes is relatively high (63%) and this observation is in congruence with other previous studies (Cheng *et al.* 2012; Norfatimah *et al.* 2014). Incomplete stop codons were observed for some of the protein-coding genes and the presence of such noncanonical stop codons is common in vertebrate mitogenomes (Wei *et al.* 2013). These truncated stop codons would be converted into a fully functional TAA stop codon via posttranscriptional polyadenylation (Ojala *et al.* 1981).

The most frequently used codon is CTA and this codon usage bias might be directly proportional to the available tRNA abundance (Lee and Kocher 1995; Xia *et al.* 2007; Cheng *et al.* 2010). The predicted tRNAs structure is similar to that of other vertebrates. However, as reported previously for bony fishes, $tRNA^{Ser}$ (UGC) did not show the typical clover leaf-like structure due to absence of a discernible DHU stem (Oh *et al.* 2007).

The O_L showed conserved stem-loop structures, which are reported to be associated with the transition from RNA to DNA synthesis during mtDNA replication (Hixson *et al.* 1986). The TAS observed in mitochondrial D-loop region would play a key role in terminating the synthesis of the heavy strand (Cheng *et al.* 2010). The conserved sequence blocks (CSB-1, CSB-2 and CSB-3) are thought to be associated with positioning RNA polymerase for priming both replication and transcription (Clayton 1991; Shadel and Clayton 1997).

The comparative analysis of mitochondrial genomes of mahseer fishes showed less divergence value at cytochrome c oxidase subunit I locus than the ND1 region. This significant observation warrants that to delineate the species of this group, *COI* gene alone may not be sufficient. High divergence values at *ND1* gene were reported for both vertebrates and invertebrates (Rach *et al.* 2008; Pilgrim *et al.* 2012). Comparison of mitogenomes between the present study *T. tor* species and reported *T. tor* (KP795444, Kumar *et al.* 2015) showed significant divergence value and it is confirmed by sliding window and codon usage analysis. The systematics of mahseer fishes is so ambiguous and often species identification based on morphological characters may cause misidentification. We hypothesize that the reported species might be a different species/subspecies with similar morphology as *T. tor* species.

Previously, most of the phylogeny studies of mahseer species have been based on limited genes/markers (Nguyen *et al.* 2009; Laskar *et al.* 2013). After the first description of the genus of *Tor* (Gray 1834), Rainboth (1985) erected new genus *Neolissochilus* on the basis of absence of median lobe, number of gill rakers and pharyngeal arches. Several studies have shown the monophyletic nature of *Neolissochilus* genus and sister group relationship with the species of *Tor* (Nguyen *et al.* 2009; Laskar *et al.* 2013; Khare *et al.* 2014; Raman *et al.* 2015). In the present study also, the codon usage analysis showed the distinct codon usage pattern for *N. hexagonolepis* compared to other species of *Tor* genus. However, in the phylogeny studies, except NJ, all other methods generated tree topologies, where the *N. hexagonolepis* formed as a distinct taxon within *Tor* species clade with low bootstrap value. Several previous studies on mitochondrial genome of mahseer fishes have also reported the same relationship with maximum likelihood method (Kumar *et al.* 2015; Sahoo *et al.* 2015b). It has been reported that sophisticated models with large number of parameters may not always generate accurate tree topology especially with large dataset. Sometimes simple tree building methods based on distance (logdet) can generate reliable tree topologies (Russo *et al.* 1996; Tamura *et al.* 2004).

In all tree topologies, sister group relationship was observed between *T. putitora* and *T. tor* and it is in accordance with the previous studies (Laskar *et al.* 2013; Khare *et al.* 2014). Sahoo *et al.* (2015b) have also shown sister group relationship between *T. tor* (KP795444) and *T. putitora*, but with less bootstrap value (<75%). Interestingly, *T. tor* (KP795444) reported by Kumar *et al.* (2015) has not clustered with either *T. tor* (present study) or *T. putitora*. The reason could be that the *T. tor* species (KP795444) may be a misidentified species / a distinct genetic lineage (subspecies) and requires additional markers (nuclear and morphological) for further classification of this taxon. Several studies using complete mitochondrial genome sequences have reported occurrence of different species / genetic lineages within a species complex (Morin *et al.* 2010). Cheng *et al.* (2012) found unexpected divergence between two *Collichthys niveatus* mitogenomes and proposed that species misidentification could be the reason for this high divergence value. As reported in previous studies, another sister group relationship was observed between *T. khudree* and *T. tambroides* (Nguyen *et al.* 2009; Raman *et al.* 2015). The biogeographical analysis also reported that these two species have originated during lower Pleistocene period as a sequence of dispersal and vicariance events (Nguyen *et al.* 2009; Khare *et al.* 2014).

In conclusion, the present study reported the complete mitochondrial genome sequence of *T. tor* species using ion torrent platform. The comparative and phylogenetic analyses showed a genetically distinct specimen of *T. tor.* The present study data could be useful for further exploration of mitogenomics of other *Tor* species for better taxonomic understanding.

Acknowledgements

This research was funded by the Department of Science and Technology, Science and Engineering Research Board project no. YSS/2014/000302. The authors are thankful to former Director Dr W. S. Lakra, ICAR-CIFE, for the guidance and support.

References

Campbell M. A., López J. A., Satoh T. P. and Chen W. J. 2014 Mitochondrial genomic investigation of flatfish monophyly. *Gene* **551**, 176–182.

Chen X., Ai W., Xiang D., Chen Y. and Chen S. 2013 Complete mitogenome of the pale-edged stingray *Dasyatis zugei* (Myliobatiformes: Dasyatidae). *Mitochondr. DNA* **24**, 196–198.

Cheng Y. Z., Xu T. J., Shi G. and Wang R. X. 2010 Complete mitochondrial genome of the miiuy croaker *Miichthys miiuy* (Perciforms, Scienidae) with phylogenetic consideration. *Mar. Genomics* **3**, 201–209.

Cheng J., Ma G. Q., Song N. and Gao T. X. 2012 Complete mitochondrial genome sequence of bighead croaker *Collichthys niveatus* (Perciformes, Sciaenidae): a mitogenomic perspective on the phylogenetic relationships of Pseudosciaeniae. *Gene* **491**, 210–223.

Clayton D. A. 1991 Nuclear gadgets in mitochondrial DNA replication and transcription. *Trends Biochem. Sci.* **16**, 107–111.

Darriba D., Taboada G. L., Doallo R. and Posada D. 2012 jModelTest 2: more models, new heuristics and parallel computing. *Nat. Methods* **9**, 772.

Desai V. R. 2003 Synopsis of biological data on the Tor Mahseer *Tor tor* (Hamilton 1822). FAO Fisheries Synopsis. Food and Agriculture Organization of the United Nations, Rome, Italy, pp. 1–31.

Gray J. E. 1834 *The Illustrations of Indian zoology, chiefly selected from the collection of general Hardwick*, pp. 96. Parbury, Allen and Co., London, UK.

Guindon S. and Gascuel O. 2003 A simple, fast and accurate method to estimate large phylogenies by maximum-likelihood. *Syst. Biol.* **52**, 696–704.

Hamilton F. 1822 *An account of the fishes found in the River Ganges and its branches*, pp. 405. Robinson and Co., Edinburgh and London.

Hixson J. E., Wong T. W. and Clayton D. A. 1986 Both the conserved stem-loop and divergent 5′-flanking sequences are required for initiation at the human mitochondrial origin of light-strand DNA replication. *J. Biol. Chem.* **261**, 2384–2390.

Huelsenbeck J. P. and Ronquist F. 2003 MR BAYES 3: bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**, 1572–1574.

IUCN 2015 The IUCN red list of threatened species. Version 2015-3. http://www.iucnredlist.org. Downloaded on 9 September 2015.

Iwasaki W., Fukunaga T., Isagozawa R., Yamada K., Maeda Y., Satoh T. P., Sado T. *et al.* 2013 MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. *Mol. Biol. Evol.* **30**, 2531–2540.

Jayaram K. C. 2005 The deccan mahseer fishes: their eco-status and threat percepts. Records of Zoological Survey of India. Occasional paper no. 238.

Khare P., Mohindra V., Barman A. S., Singh R. V. and Lal K. K. 2014 Molecular evidence to reconcile taxonomic instability in mahseer species (Pisces: Cyprinidae) of India. *Org. Diver. Evol.* **14**, 307–326.

Kim I. C. and Lee J. S. 2004 The complete mitochondrial genome of the rockfish *Sebastes schlegeli* (Scorpaeniformes, Scorpaenidae). *Mol. Cells.* **17**, 322–328.

Kumar R., Goel C., Sahoo P. K., Singh A. K. and Barat A. 2015 Complete mitochondrial genome organization of *Tor tor* (Hamilton 1822). *Mitochondr. DNA* **28**, 1–2.

Kushwaha B., Kumar R., Agarwal S., Pandey M. and Nagpure N. S. 2015 Assembly and variation analyses of *Clarias batrachus* mitogenome retrieved from WGS data and its phylogenetic relationship with other catfishes. *Meta Gene*, doi:10.1016/j.mgene.2015.06.004.

Lal K. K., Singh R. K., Pandey A., Gupta B. K., Mohindra V. and Punia P. 2013 Distribution Records of Tor Mahseer *Tor tor* (Hamilton 1822) from southern India. *J. Appl. Ichthyol.* **29**, 1–5.

Laskar B. A., Bhattacharjee M. J., Dhar B., Mahadani P., Kundu S., Sankar S. and Ghosh S. K. 2013 The species dilemma of northeast Indian mahseer (Actinopterygii: Cyprinidae): DNA barcoding in clarifying the riddle. *PLoS One* **8**, e53704.

Lee W. J. and Kocher T. D. 1995 Complete sequence of a sea lamprey (*Petromyzon marinus*) mitochondrial genome: early establishment of the vertebrate genome organization. *Genetics* **139**, 873–887.

Marshall H. D., Baker A. J. and Grant A. R. 2013 Complete mitochondrial genomes from four subspecies of common chaffinch (*Fringilla coelebs*): new inferences about mitochondrial rate heterogeneity, neutral theory, and phylogenetic relationships within the order Passeriformes. *Gene* **517**, 37–45.

Mohindra V., Khare P., Lal K. K., Punia P., Singh R. K., Barman A. S. and Lakra W. S. 2007 Molecular discrimination of five Mahseer species from Indian peninsula using RAPD analysis. *Acta Zool. Sinica* **53**, 725–732.

Morin P. A., Archer F. I., Foote A. D., Vilstrup J. and Allen E. E. 2010 Complete mitochondrial genome phylogeographic analysis of killer whales (*Orcinus orca*) indicates multiple species. *Genome Res.* **20**, 908–916.

Nguyen T. T. T., Na Nakorn U., Sukmanomon S. and Chen Z. M. 2009 A study on phylogeny and biogeography of mahseer species (Pisces: Cyprinidae) using sequences of three mitochondrial DNA gene regions. *Mol. Phylogenet. Evol.* **48**, 1223–1331.

Norfatimah M. Y., Teh L. K., Salleh M. Z., Mat Isa M. N. and SitiAzizah M. N. 2014 Complete mitochondrial genome of Malaysian Mahseer (*Tor tambroides*). *Gene* **548**, 263–269.

Oh D. J., Kim J. Y., Lee J. A., Yoon W. J., Park S. Y. and Jung Y. H. 2007 Complete mitochondrial genome of the rabbitfish *Siganus fuscescens* (Perciformes, Siganidae). *DNA Seq.* **18**, 295–301.

Ojala D., Montoya J. and Attardi G. 1981 tRNA punctuation model of RNA processing in human mitochondria. *Nature* **290**, 470–474.

Pilgrim B. L., Perry R. C., Barron J. L. and Marshall H. D. 2012 Nucleotide variation in the mitochondrial genome provides evidence for dual routes of postglacial recolonization and genetic recombination in the northeastern brook trout (*Salvelinus fontinalis*). *Genet. Mol. Res.* **26**, 3466–3481.

Rach J., DeSalle R., Sarkar I., Schierwater B. and Hadrys H. 2008 Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. *Proc. R. Soc. Lond. B Biol. Sci.* **275**, 237–247.

Rainboth W. J. 1985 Neolissochilus, a new genus of south Asian cyprinid fishes. *Beaufortia* **35**, 25–35.

Rainboth W. J. 1996 Fishes of the Cambodian Mekong. FAO Species Identification Field Guide for Fishery Purposes, pp. 265. United Nations Food and Agriculture Organization, Rome.

Raman S., Pavan-Kumar A., Koringa P. G., Patel N., Shah T., Singh R. K. *et al.* 2015 Ion torrent next-generation sequencing reveals the complete mitochondrial genome of endangered mahseer *Tor khudree* (Sykes 1839). *Mitochondrial DNA*, doi: 10.3109/19401736.2015.1060455.

Rozas J. 2009 DNA sequence polymorphism analysis using DnaSP. In *Bioinformatics for DNA sequence analysis*, pp. 337–350 (ed. Posada) Methods in Molecular Biology Series. Humana Press, NJ, USA.

Russo C. A., Takezaki N. and Nei M. 1996 Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* **13**, 525–536.

Sahoo L., Kumar S., Das S. P., Patnaik S., Bit A., Sundaray J. K., Jayasankar P. and Das P. 2015a Complete mitochondrial genome sequence of *Heteropneustes fossilis* obtained by paired end next generation sequencing. *Mitochondrial DNA*, doi: 10.3109/19401736.2015.1033710.

Sahoo P. K., Goel C., Kumar R., Dhama N., Ali S., Sarma D., Nanda P. and Barat A. 2015b The complete mitochondrial genome of threatened chocolate mahseer (*Neolissochilus hexagonolepis*) and its phylogeny. *Gene* **570**, 299–303.

Sati J., Goel C., Kumar R., Ali S., Patiyal R. S., Singh V. K., Sahoo P. K. and Barat A. 2014 Complete mitochondrial genome organization of *Tor putitora*. *Mitochondrial DNA* **25**, 278–279.

Schattner P., Brooks A. N. and Lowe T. M. 2005 The tRNA scan-SE, snoscan and snoGPS webservers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res.* **33**, 686–689.

Shadel G. S. and Clayton D. A. 1997 Mitochondrial DNA maintenance in vertebrates. *Annu. Rev. Biochem.* **66**, 409–435.

Singh H. R. and Kumar N. 2000 Some aspects of ecology of hill streams; stream morphology, zonation, characteristics, and adaptive features of ichthyofauna in Garhwal Himalaya. In *Modern trends in fish biology research* (ed. J. S. Datta Munshi), pp. 1–18. Narendra Publishing House, New Delhi, India.

Swofford D. L. 2003 PAUP: phylogenetic analysis using parsimony (and other methods), version 4.0b10. Sinauer Associates, Sunderland, MA, USA.

Tamura K., Nei M. and Kumar S. 2004 Prospects for inferring very large phylogenies by using the neighbor-joining method. *Proc. Natl. Acad. Sci. USA*, **101**, 11030–11035.

Tamura K., Stecher G., Peterson D., Filipski A. and Kumar S. 2013 MEGA6: molecular evolutionary genetic analysis, Version 6.0. *Mol. Biol. Evol.* **30**, 2725–2729.

Tang M., Tan M., Meng G., Yang S., Su X., Liu S. *et al.* 2014 Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Res.* doi: 10.1093/nar/gku917.

Thai B. T., Si V. N., Phan P. D. and Austin C. M. 2007 Phylogenetic evaluation of subfamily classification of the Cyprinidae focusing on Vietnamese species. *Aquat. Living Resour.* **20**, 143–153.

Wei T., Jin X. X. and Xu T. J. 2013 The first complete mitochondrial genome from Bostrychus genus (*Bostrychus sinensis*) and partitioned Bayesian analysis of Eleotridae fish phylogeny. *J. Genet.* **92**, 247–257.

Xia J., Xia K., Gong J. and Jiang S. 2007 Complete mitochondrial DNA sequence, gene organization and genetic variation of control regions in *Parargyrops edita*. *Fish Sci.* **73**, 1042–1049.

Xia X. 2013 DAMBE5: a comprehensive software package for data analysis in molecular biology and evolution. *Mol. Biol. Evol.* **30**, 1720–1728.

Yang L., Mayden R. L., Sado T., He S., Saitoh K. and Miya M. 2010 Molecular phylogeny of the fishes traditionally referred to Cyprinini sensu stricto (Teleostei: Cypriniformes). *Zool. Scr.* **39**, 527–550.

Zuker M. 2003 Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415.

Received 26 November 2015, in revised form 18 January 2016; accepted 28 March 2016
Unedited version published online: 30 March 2016
Final version published online: 22 November 2016

Corresponding editor: N. G. PRASAD