PUBLICATIONS

Publications in peer-reviewed journals

1. Chakraborty, M., & Ghosh, S. K. (2014). An assessment of the DNA barcodes of Indian freshwater fishes. Gene, **537**(1), 20-28.
2. Chakraborty, M., & Ghosh, S. K. (2014). Unraveling the sequence information in *COI* barcode to achieve higher taxon assignment based on Indian freshwater fishes. Mitochondrial DNA, (0), 1-3.
3. Chakraborty, M., Ghosh, S. K., Khomdram B. Revealing the genetic diversity of *Clarias batrachus* using DNA barcode. Journal of environment and socio-biology.**10**(1):25-32, 2013
4. Talukdar, F.R., Ghosh, S.K., Laskar, R.S., Mahadani, P., Chakraborty, M., Dhar, B. and Bhattacharjee, M.J. Implication of nucleotide substitution at third codon position of the DNA barcode sequences. Journal of environment and socio-biology.:**10**(1):55-63, 2013.
5. Ghosh, S.K., Bhattacharjee, M.J., Devi, M.K., Ahanthem, M., Kundu, S., Mahadani, P., Dhar, B., Khondram, B., Chakraborty, M., Rahman, F., Mondal, R., Hansa, J., Laskar, R., Mazumdar, Sarkar, P., Rajbonshi, S., Chakraborty, A., Das, M., Ghosh, P.R., Das, K.C and Laskar, B.A. DNA Barcoding: Digital Taxonomy of Biorsources. Strategic Physiological Research for Sustainable Animal Biodiversity.

Manuscript communicated

Barcode-motif: a novel approach for DNA barcode based species identification. Chakraborty, M., Ghosh, S.K. PlosOne (Under review).

Book Chapter contribution

Mahadani, P., Devi, M.K., Das, M., Chakraborty, M., Rahman, F., Hansa, J., Ghosh, S.K. Bioinformatics in DNA Barcoding . A text book on DNA Barcoding. ISBN 81-922989-4-8.

Gene 537 (2014) 20–28

Contents lists available at ScienceDirect

Gene

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

An assessment of the DNA barcodes of Indian freshwater fishes

Mohua Chakraborty, Sankar Kumar Ghosh *

Department of Biotechnology, Assam University, Silchar 788011, Assam, India

ARTICLE INFO

Article history:
Accepted 18 December 2013
Available online 28 December 2013

Keywords:
Indian freshwater fishes
DNA barcode
COI gene
BOLD and Genbank
Neighbor Joining tree
Cryptic species complex

ABSTRACT

Freshwater fishes in India are poorly known and plagued by many unresolved cryptic species complexes that masks some latent and endemic species. Limitations in traditional taxonomy have resulted in this crypticism. Hence, molecular approaches like DNA barcoding, are needed to diagnose these latent species. We have analyzed 1383 barcode sequences of 175 Indian freshwater fish species available in the databases, of which 172 sequences of 70 species were generated. The congeneric and conspecific genetic divergences were calculated using Kimura's 2 parameter distance model followed by the construction of a Neighbor Joining tree using the MEGA 5.1. DNA barcoding principle at its first hand approach, led to the straightforward identification of 82% of the studied species with 2.9% (S.E = 0.2) divergence between the nearest congeners. However, after validating some cases of synonymy and mislabeled sequences, 5% more species were found to be valid. Sequences submitted to the database under different names were found to represent single species. On the other hand, some sequences of the species like *Barilius barna*, *Barilius bendelisis* and *Labeo bata* were submitted to the database under a single name but were found to represent either some unexplored species or latent species. Overall, 87% of the available Indian freshwater fish barcodes were diagnosed as true species in parity with the existing checklist and can act as reference barcode for the particular taxa. For the remaining 13% (21 species) the correct species name was difficult to assign as they depicted some erroneous identification and cryptic species complex. Thus, these barcodes will need further assay and inclusion of barcodes of more specimens from same and sister species.

1. Introduction

India is rich in fishery resources and comprises of around 2508 fish species (Eschmeyer and Fricke, 2012) of which 856 are freshwater inhabitants (Froese and Pauly, 2012; Menon, 1999). Indian freshwater fishes represent about 8.9% of the known fish species of the world and occupy the ninth position in terms of freshwater fish diversity (Leveque et al., 2008). Distribution of freshwater fishes in India is widely varied with the north eastern zone alone claiming 40% of the reported species (Ponniah and Sarkar, 2000). However, the actual number of fish species found in India is still not accurately known because of taxonomic impediments (Hoagland, 1996) arising due to lack of exploration, indiscernibility among some alike species, and ambiguity in taxonomic keys (Pethyagoda and Kottelat, 1994). As a result, many species have been considered as cryptic some of which may also be latent (Darshan et al., 2010; Pethyagoda and Kottelat, 1994). Furthermore, due to lack of proper morphological description with respect to sexual dimorphism, allopatric population, etc., the statuses of a few nominal species have been contentious (Darshan et al., 2010; Kottelat and Lim, 1995; Ng and Hadiaty, 2009; Roberts, 1992; Vishwanath and Linthoingambi, 2007). Therefore, for legible characterization of Indian freshwater fishes, there is an urgent need of species scrutiny using advanced molecular methods.

Among the molecular methods used for taxon diagnosis, DNA barcoding has been extensively used for species identification as well as species discovery in various groups of organism (Hajibabaei, 2012; Mendonca et al., 2012). It has gained wide popularity with over 1000 publications in less than a decade's time. Effectiveness of DNA barcoding has now been validated for many groups of animals (Waugh, 2007), both invertebrates and vertebrates (Hajibabaei et al., 2006a,b; Hebert et al., 2004; Hernandez-Davila et al., 2012), with fishes being one of the most extensively studied groups among them (Becker et al., 2011; Ward, 2012). Many successful nationwide studies on ichthyofaunal diversity have been undertaken using this method for both marine and freshwater fishes (Carvalho et al., 2011; de Oliveira Ribeiro et al., 2012; Mabragana et al., 2011; Wang et al., 2012). These studies have resolved several cases of crypticism and diagnosed many latent species (Carvalho et al., 2011; Pereira et al., 2011). Furthermore, these studies have also generated a large scale of barcode data that are available in BOLD (Ratnasingham, 2007) and NCBI. However, the reference library of barcodes is still incomplete as many geographical locations, particularly in Asia, are yet to be exhaustively covered. In India, in addition to the absence of an updated compiled checklist of freshwater fishes, the identification keys for many valid species have not been updated since, the study of Talwar and Jhingran, 1991 (Talwar and Jhingran, 1991;

Abbreviations: K2P, Kimura 2 parameter; COI, Cytochrome C oxidase 1; NJ, Neighbor Joining; ML, Maximum Likelihood; S.E, Standard Error; NUMTs, nuclear DNA originating from mitochondrial DNA sequences; EDTA, ethylenediaminetetraacetic acid; dNTPs, deoxynucleotide triphosphates; MEGA, molecular evolutionary genetic analysis; BOLD, barcode of life datasystem; BLAST, basic local alignment search tool.

* Corresponding author. Tel.: +91 9435372338.
E-mail address: drsankarghosh@gmail.com (S.K. Ghosh).

0378-1119/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2013.12.047

Jayaram, 1999). During the last decade, many new species have been described in India, particularly from the north-eastern region (Dishma and Vishwanath, 2012; Nath and Dey, 1989; Selim and Vishwanath, 2002; Vishwanath and Linthoingambi, 2005) and Western Ghat (Devi et al., 2010; Pethyagoda and Kottelat, 1994). In this context, there is an urgent need for the assessment of Indian freshwater fish species through DNA barcoding. Many sequences of Indian freshwater fishes have been submitted to the database and few studies have addressed problems specific to certain groups (Benziger et al., 2011; Bhattacharjee et al., 2012; Laskar et al., 2013). However, a comprehensive assessment of DNA barcodes of Indian freshwater fishes has not yet been done though a similar study has been done for the Indian marine fishes (Lakra et al., 2011). Furthermore, in order to ensure that the reference barcode library is accurate, it is important to detect the ambiguities in the base collection of barcodes at an early stage.

In this study, the entire DNA barcode data of Indian freshwater fishes, so far available in the database, were analyzed. Among them, DNA barcodes of freshwater species from the north-eastern part of India were developed. The study re-evaluated and elucidated the actual species status of the majority of the studied species and particularly helped to flag the species whose statuses have been doubtful. It will open up scopes of research in the ambiguous fish groups and will also contribute as a reference material for future studies.

2. Materials and methods

2.1. Data acquisition of DNA barcode sequences of Indian freshwater fishes

The Public Data Portal of BOLD (Ratnasingham, 2007) and Core Nucleotide database of GenBank were searched for *COI* (Cytochorome C Oxidase 1) barcode sequences of Indian freshwater fishes. The data were retrieved using Boolean operator ‘AND’ with two terms under a different context (taxonomic: Order and geographic: India) thereby extracting records that only matched both the terms. Sequences from both the databases were compiled together and duplicate records were removed, to finally get a set of 1413 barcode sequences for 179 species. Sequences of length >600 bp, with no missing nucleotides or gaps, were included, thereby reducing the possibility of NUMTs (nuclear DNA originating from mitochondrial DNA sequences) (Zhang and Hewitt, 1996), and aligned using Clustal Omega (Sievers et al., 2011). Suspected erroneous sequences, with highly unlikely positions (species clustering with different family or order) or having extreme branch lengths were omitted, based on a Neighbor-Joining tree. The *COI* coding DNA sequence were translated using MEGA 5.1 and aligned with the available COI amino acid sequences to ensure the presence of an open reading frame (Tamura et al., 2011). The sequences were trimmed at either ends to exclude any gaps and a final set of 503 bp long 1383 consensus barcode sequences for 175 species were used for analysis. Among them, 172 sequences for North-East Indian freshwater fishes were developed following the protocol as below.

2.2. DNA extraction and PCR assay

The extraction of DNA was performed with Phenol–Chloroform–Isoamylalcohol method (Sambrook and Russell, 2001). 20 mg of anal fin tissue was taken aseptically and dissolved in 500 μL of TES buffer (50 mM TrisHCl, 25 mM EDTA and 150 mM NaCl) in a microcentrifuge tube. The *COI* gene (655 bp) was amplified using the set of published primers (Ward et al., 2005) i.e. FishF1-5′TCAACCAACCACAAAGACATTG GCAC 3′and FishR1-5′TAGACTTCTGGGTGGCCAAAGAATCA 3′in a Veriti Mastercycler (Applied Biosystems Inc., CA, USA). The amplification reactions were performed in a total volume of 25 μl comprising 1X PCR buffer, 2 mM $MgCl_2$, 10 pmol of each primer, 0.25 mM of each dNTPs, 0.25 U of high-fidelity Taq polymerase (Applied Biosystems Inc., CA, USA) and 100 ng of DNA template. The thermal profile of the PCR reaction was as follows: An initial denaturation at 94 °C for 2 min, 30 cycles at denaturation temperature of 94 °C for 45 s, annealing temperature of 50 °C for 45 s and elongation temperature of 72 °C for 1 min, and concluded with a final elongation step at 72 °C for 8 min followed by a hold at 4 °C. The PCR-amplified products were analyzed in 1% agarose gels containing ethidium bromide staining (10 mg/ml) and the single uniform band was then purified using QIAquick^R Gel extraction kit (QIAGEN, USA), following manufacturer's instructions. The amplicons were bidirectionally sequenced in an automated DNA sequencer (ABI 3500, Applied Biosystems Inc., CA, USA), through the sequencing service of Bose Institute (Kolkata, India).

The *COI* barcode sequences generated were submitted to BOLD under 3 separate projects i.e., “DNA barcoding of freshwater catfishes of North-East India,” “DNA barcoding of Mahseer fishes from North-East India” and “DNA barcoding of ornamental fishes of North-East India,” containing the detailed sampling and taxonomic information. Most of the data used in this study were retrieved from the database while fish samples for DNA extraction were collected from wild habitats of North-East India. Department of Biotechnology, Government of India has given the necessary permission to conduct research on fish DNA barcoding of North-East India, vide letter number BT/HRD/01/002/2007. Hence, no ethical permission is required for this study as collection of fishes is a routine practice for commercial purpose in this region.

2.3. Data analysis

Genetic distances were calculated using Kimura's two parameter (K2P) models (Kimura, 1980), as implemented in MEGA 5.1 (Tamura et al., 2011) to quantify sequence divergences among individuals. K2P based method has been argued to be not the best suited method for species delimitation (Srivathsan and Meier, 2012). Despite this, we chose this model as, studies have shown, differences in distance between best considered model and K2P model estimates were usually minimal, and identification success rates were largely unaffected by model choice (Collins et al., 2012). Interspecific K2P distances were calculated for species with at least 2 sequences, and intraspecific K2P distances were calculated between species in the entire data set. Genetic distances were analyzed at species, genus and family level in MEGA 5.1. Neighbor joining (NJ) analysis was performed for *COI* using the K2P distance model as recommended by Hebert et al.(2003) using MEGA 5.1. A second phylogenetic tree was constructed using Maximum likelihood (ML) method. Node supports for both the trees were evaluated with 1000 bootstrap pseudo-replicates. The details regarding the sequences generated and retrieved from the database is given in Table S1. Correctly delimited species based on cohesive clustering by the conspecifics in the NJ and ML trees were considered as true species. The maximum conspecific and minimum congeneric divergences between these species were used to define the species boundaries as elaborated in the “Results” section. The divergence between the minimum congeneric and maximum conspecific divergence is the lowest divergence between congeners. This divergence value has been assumed as the threshold level of species delineation and thereby considered as a barcoding gap in this study. This process of species delimitation is preferred over other methods and has been used by many computational methods to partition a sequence alignment dataset into candidate species (Puillandre et al., 2012; Zhang et al., 2013)

3. Results

A total of 1383 mitochondrial *COI* barcode sequences of 175 species belonging to 10 orders, 34 families, 77 genera (Table S1) and constituting almost 20% of Indian freshwater fishes were retrieved. Among them, 172 barcode sequences, representing 70 different species, were generated from North-East India. No insertion, deletion or stop codons were observed in any sequence. The absence of stop codon as well as coherent partial amino acid codes confirmed them to be a partial fragment of mitochondrial *COI* gene. In the dataset, for most species, multiple

Table 1
Summary of genetic divergences (K2P model) for each taxonomic level of comparison. The divergence showed an increase with the hierarchical increase of taxon comparison. Mean interspecific divergence showed 6.2 folds higher than mean intraspecific divergence however there is significant overlapping in the distribution of intra- and interspecific divergence.

Comparison within	Taxa(n)	Mean	Min	Max	S.E
Species	175	1.6	0	18.87	0.1
Genus	77	7.16	0	21.42	1.1
Family	34	15.66	11.51	32.23	1.9
Order	10	25.32	20.42	45.41	2.3

specimens were used to document intraspecific variability (with an average of 5 specimens per species). However, 30 species were represented by single specimen only.

The sequence analysis revealed a hierarchical increase in K2P mean divergence across all the taxon from within species (1.6%, S.E = 0.1) to within genus (9.925%, S.E = 2.7), within family (15.66%, S.E = 1.9) and within orders (25.32%, S.E = 2.3) and is presented in Table 1. Conspecific divergence between sequences of same species varied in the range 0%–18%. While, congeneric divergence, between sequences of different species under same genus, varied from 0% to 21%. This overlap in the distribution of conspecific and congeneric means caused hindrance in defining the threshold value for species boundary. To resolve the problem, we explored the NJ and ML tree (Figs. S1, S2) and found 3 general cases, as shown in Fig. 1:

1) Sequences, with same species name, exhibiting cohesive clustering by the conspecies and distinct clustering by the congeners with high bootstrap support (90–100%).
2) Sequences, with same species name, not exhibiting cohesive clustering by the conspecies.
3) Sequences, with a different species name, exhibiting cohesive clustering.

Among the 3 cases, only the first group abided by the first principle of DNA barcoding (same named species should cluster cohesively and distinctly from the rest) and represented 82% of the total 175 species. These sequences were considered to represent true species.

In this set, the highest conspecific mean divergence was shown by *Mastacembelus armatus* (2.3%, S.E = 0.1), and the lowest congeneric divergence was shown between *Tor khudree and Tor mosal* (2.9%, S.E = 0.2). These values were used to define the species boundary in this study as shown in Fig. 2. All the remaining species of this group showed conspecific divergence lower than 2.3% and congeneric divergence higher than 2.9%.

The remaining two groups, that constituted 18% of the studied species, were considered problematic.

Sequences, with same species name, formed 2 or more distinct subclusters and showed divergence above maximum conspecific value. Here 17 species i.e., *Puntius conchonius* (n = 2), *Osteobrama cotio cotio* (n = 4), *Pterygoplichthys pardalis* (n = 3), *Devario devario* (n = 2), *Clupisoma garua* (n = 3), *Garra hughi* (n = 5), *Channa marulius* (n = 6), *Glossogobius giuris* (n = 2), *Barilius tileo* (n = 6), *Lates calcarifer* (n = 94), *Heteropneustes fossilis* (n = 14), *Clarias batrachus* (n = 18), *Labeo bata* (n = 17), *Channa orientalis* (n = 11), *Puntius filamentosus* (n = 4), *Barilius bendelisis* (n = 40), and *Barilius barna* (n = 29) formed separate or dispersed cluster and are mentioned in Table 2.

C. batrachus, represented by 18 sequences (mean = 6.6%, S.E = 0.6), formed 2 distinct clusters with 6 sequences (FJ459456-59, JQ667517-18) being separated from the remaining 12 sequences of the species by 11.5% mean distance (Fig. 3a). However, divergences within the individual clusters were 1.2% (S.E = 0.2) and 0.6% (S.E = 0.2) respectively. *H. fossilis* showed conspecific divergence of 2.6% (S.E = 0.3) and one sequence of *H. fossilis* (HQ009491) clustered away from the rest of the 13 sequences with mean K2P divergence between them being 10.5% (S.E = 0.3). This single sequence of *H. fossilis* clustered with *Heteropneustes microps* with a congeneric distance of 0.2% (S.E = 0.1) (Fig. 3b). With the exclusion of the single aberrant sequence, the remaining sequences of *H. fossilis* formed close cluster with a conspecific divergence of 0. 9% (S.E = 0.1). *L. calcarifer*, represented by 94 sequences and conspecific divergence of 2.6% (S.E = 0.2), formed 2 different clusters with 4 sequences (HQ219138–HQ219141) clustering separately from the remaining 90 sequences of the same species. Similarly, *Labeo bata* with a conspecific mean of 5.4% (S.E = 0.7) formed 2 distinct clusters with one cluster of 6 sequences (EU30664-67, JQ713847, FJ459423), clustering away with mean divergence between the 2 clusters being 11.5% (S.E = 0.6) and mean distance within the clusters being 0% and 2% respectively (Fig. 3c). In *Barilius* genus, 69 sequences of the 2 species *B. bendelisis* and *B. barna* were indistinguishable in the NJ and ML tree as their respective representative sequences did not form unique clusters. Rather, they collectively formed 6 clusters and were designated as *Barilius* cluster1 (accession numbers FJ459411, FJ459412, FJ459418-22, JN965193, JN965195, JN965196, JX1054820-65 of *B. bendelisis* and accession numbers EU417797-99, HM042258-63, HM042170-81 of *B. barna*), *Barilius* cluster2 (accession numbers EU822331-33, HM042230-35 of *B. bendelisis*), *Barilius* cluster3 (accession number HM042236-41 of *B. bendelisis*), *Barilius* cluster4 (accession

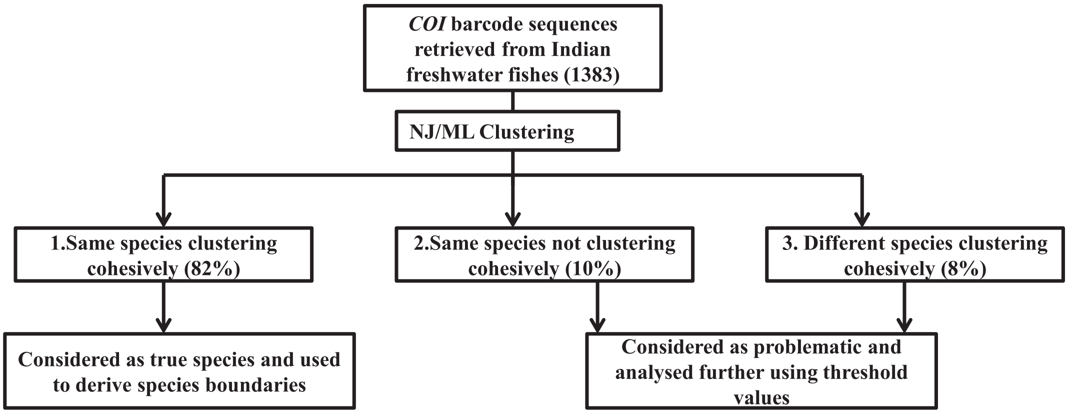
COI barcode sequences retrieved from Indian freshwater fishes (1383)
NJ/ML Clustering
1.Same species clustering cohesively (82%)
2.Same species not clustering cohesively (10%)
3. Different species clustering cohesively (8%)
Considered as true species and used to derive species boundaries
Considered as problematic and analysed further using threshold values

Fig. 1. The folwchart describes the categorization of the species into three groups based on clustering pattern in NJ and ML tree. Sequences having same species name clustering cohesively and distinct from others are categorized as Group 1. Sequences having same species name but forming distinct clusters are categorized as Group 2. Sequences having different species name and forming cohesive cluster are categorized as Group 3. Species representing first group are considered as true species and used to derive threshold to define species boundaries while the remaining two groups are considered problematic.

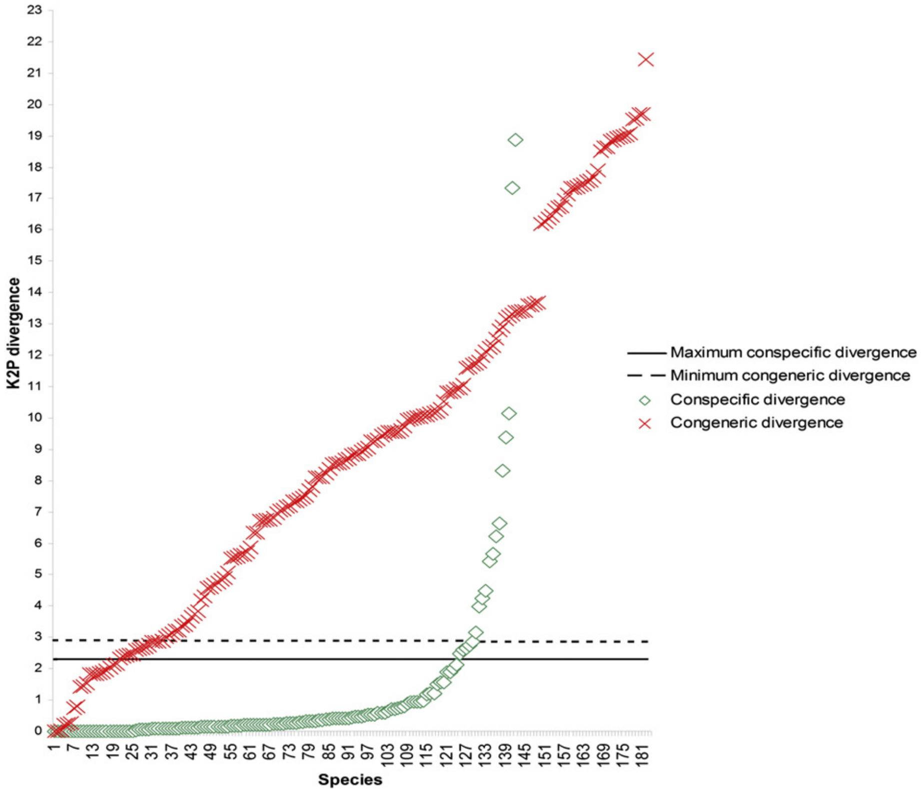

Fig. 2. Distribution of conspecific and congeneric K2P mean divergence of 175 species of Indian freshwater fishes (arranged in ascending order). The maximum conspecific divergence (2.3%, black solid line) and minimum congeneric divergence (2.9%, black dashed line) represent the threshold level of conspecific and congeneric divergence respectively. Data series marked by 'cubes' represent conspecific divergence of 142 species, which were represented by more than one sequence. 82% of the total 175 species showed divergence below 2.3% and represented true species. Data series marked by 'cross' represents divergence between congeneric pairs of species, those points that lie below threshold conspecific line consists of either synonymous or misidentified species and discussed in Table 3.

number HM042248-53 of *B. bendelisis*), *Barilius* cluster5 (accession number HM042242-47 of *B. bendelisis*), *Barilius* cluster6 (accession number HM042164-69 of *B. barna*) in the NJ and MLtree. The sequences within each of the 6 cohesive clusters of *Barilius* showed mean K2P divergence below the value of maximum conspecific divergence while the mean divergences between the sequences of separate clusters were above the value of minimum congeneric divergence (Table 4). Sequences of *C. orientalis*, *P. pardalis*, *P. filamentosus* seemed to form subclusters under single node. However, nodes parallel to these nodes gave rise to different species. Moreover, sequences of these species showed high mean intraspecific divergence of 2.8%, 2.7% and 3.1% respectively. For the remaining 11 species, the number of representative sequences was not sufficient to draw any conclusion.

In Group 3, sequences, with different species name, (that are expected to cluster separately) clustered cohesively and showed divergence below maximum conspecific divergence [*Labeo dussumieri* versus *Labeo rajasthanicus* (Fig. 3c), *Poecilia sphenops* versus *Poecilia velifera*, *Aspidoparia morar* versus *Aspidoparia jaya*, *Mystus vittatus* versus *Mystus horai*, *Mystus tengara* versus *Neotropius atherinoides*, *Macrognathus aral* versus *Macrognathus aculeatus* (Fig. 3e)] and are detailed in Table 3. *P. sphenops* versus *P. velifera* and *L. dussumieri* versus *L. rajasthanicus* exhibited a divergence of 0%. Moreover, pairs with interspecific divergence between 1.14% and 2.3% were not clustered together, but formed dispersed clusters supported with low bootstrap value e.g. *Tor tor* together with *Tor putitora*, *Tor macrolepis*, *Tor mussullah*, *Tor mosal mahanadicus*; *Poecilia latipinna* with *P. sphenops* and *P. velifera*.

Table 2
Sequences with same species name that formed two or more distinct sub-clusters and showed divergence above maximum conspecific value.

	Species name	% distance	S.E	Number of sequence	Number of clusters	Bootstrap
1	*Lates calcarifer*	2.60	0.2	94	2	100, 100
2	*Heteropneustes fossilis*	2.66	0.3	14	2	100, 100
3	*Pterygoplichthys pardalis*	2.73	0.5	3	2	100, 100
4	*Channa orientalis*	2.87	0.4	11	2	100, 99
5	*Puntius filamentosus*	3.16	0.5	4	2	100, 100
6	*Barilius bendelisis*	3.96	0.5	40	Dispersed clusters	78, 100
7	*Barilius barna*	4.24	0.5	29	Dispersed clusters	78, 100
8	*Osteobrama cotio cotio*	4.48	0.7	4	2	100, 100
9	*Labeo bata*	5.43	0.7	17	2	100, 100
10	*Devario devario*	5.66	1	2	2	100, 100
11	*Clupisoma garua*	6.21	0.9	3	2	100, 100
12	*Clarias batrachus*	6.63	5	18	2	100, 100
13	*Garra hughi*	8.30	1	5	2	98, 82
14	*Barilius tileo*	9.36	1.2	6	2	100, 100
15	*Channa marulius*	10.13	1.1	6	2	100, 100
16	*Glossogobius giuris*	17.32	2.1	2	2	100, 100
17	*Puntius conchonius*	18.87	2	2	2	100, 100

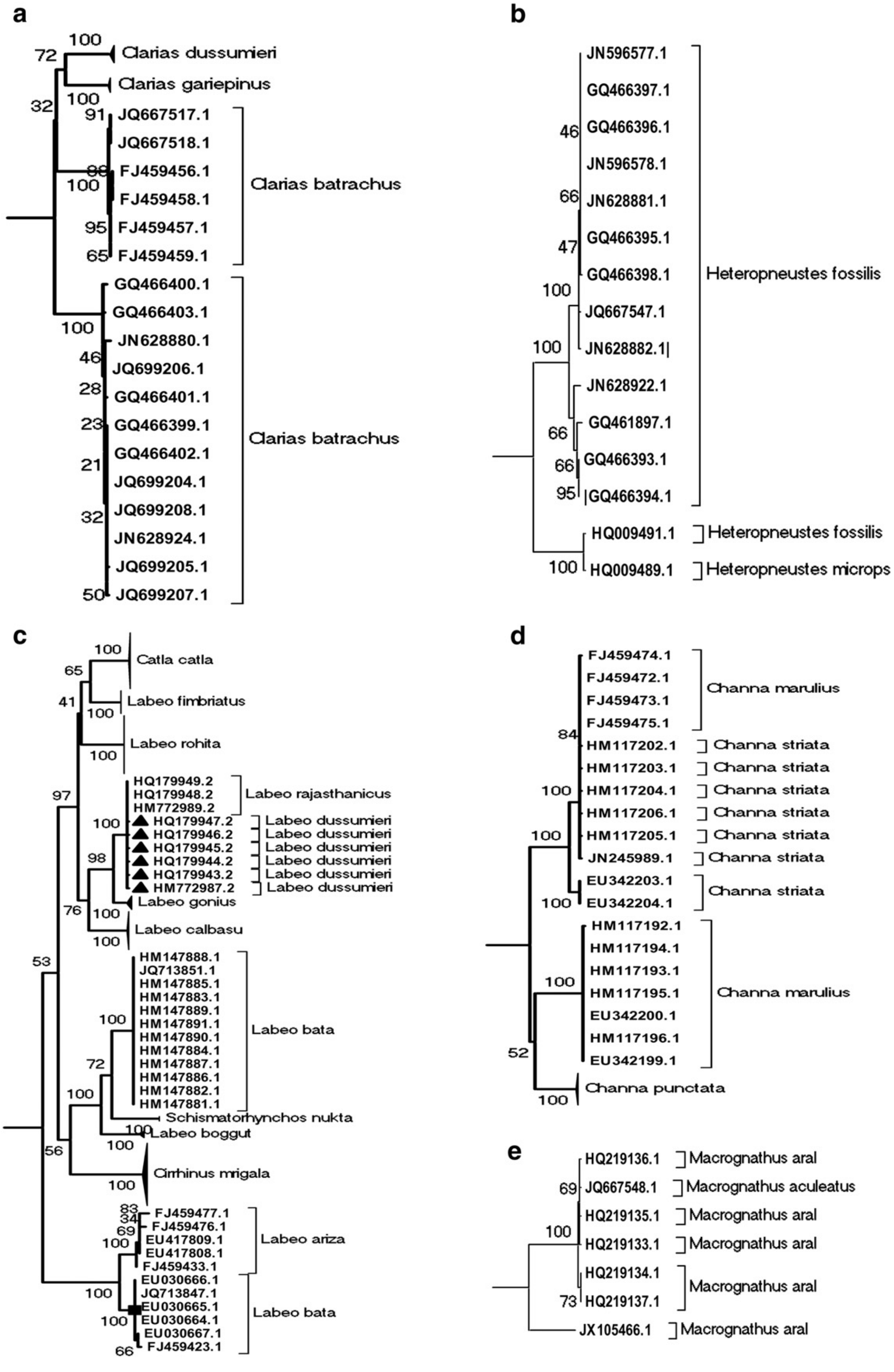

Fig. 3. Sections of Neighbor-Joining tree (Fig. S1) showing the problematic groups. Sequences having same species name forms separate clusters *Clarias batrachus* (a), *Heteropneustes fossilis* (b), *Labeo bata* (c), *Channa marulius* (d) *and Macrognathus aral* (e). Sequences having different species name clusters together i.e. *Heteropneustes fossilis and Heteropneustes microps* (b), *Labeo dussumieri and Labeo rajasthanicus* (c), *Channa marulius and Channa striata* (d), *Macrognathus aral* and *Macrognathus aculeatus* (e).

Table 3
Sequences with different species name that clustered cohesively and showed divergence below maximum conspecific divergence.

Serial Number	Species 1	Species 2	% distance	S.E
1	*Labeo dussumieri* (6)	*Labeo rajasthanicus* (3)	0.00	0
2	*Poecilia sphenops* (1)	*Poecilia velifera* (4)	0.00	0
3	*Tor mosal mahanadicus* (11)	*Tor macrolepis* (5)	0.04	0
4	*Aspidoparia morar* (2)	*Aspidoparia jaya* (1)	0.20	0
5	*Tor putitora* (38)	*Tor macrolepis* (5)	0.22	0.1
6	*Tor mosal mahanadicus* (11)	*Tor putitora* (38)	0.26	0
7	*Mystus vittatus* (8)	*Mystus horai* (1)	0.43	0.1
8	*Mystus tengara* (4)	*Neotropius atherinoides* (7)	0.74	0.1
9	*Macrognathus aral* (6)	*Macrognathus aculeatus* (1)	0.79	0.2

4. Discussion

Species identification through DNA barcoding is based upon the principle that interspecific divergence sufficiently outscores intraspecific divergence and the biological species can be clearly demarcated by a threshold value, which corresponds to the divergence between the nearest neighbors within a group (Hebert et al., 2003). However, despite extensive application of DNA barcoding throughout the last decade, no universal standard threshold has been defined for interspecies demarcation. A prime reason being that mitochondrial DNA (mtDNA) rates of evolution vary between and within species and between different groups of species resulting in broad overlaps of intra and interspecific distances (Rubinoff et al., 2006). Most of the DNA barcoding studies have rather used a threshold that was specifically estimated for the dataset under study (deWaard et al., 2011; Kim et al., 2012; Lijtmaer et al., 2011). Recent studies have preferred the use of difference between minimum congeneric and maximum conspecific divergence to define the barcoding gap (April et al., 2011; Bhattacharjee et al., 2012) and has found it to be more efficient over the use of mean of intra and interspecific sequence variability (Meier et al., 2008). In this study, 136 species showed cohesive clustering between conspecies in the NJ and ML tree and have been considered as true species. However, few species like *Badis badis*, *Schizothorax progastus*, *Channa gachua*, *Puntius sarana*, *Macrognathus aral*, *Puntius chelynoides*, *Tor malabaricus*, *Channa striata*, *Epalzeorhynchos bicolor*, *Acanthocobitis botia* and *Mastacembelus armatus* formed subclusters under single node. These straightforward cases exhibit a mean divergence of 0.45% (S.E = 0.2) which is close to such studies elsewhere like 0.73% for North American freshwater fishes (April et al., 2011), 0.6% for Cuban freshwater fishes (Lara et al., 2010) and 0.39% for Australian marine fishes (Ward et al., 2005). Besides these straightforward cases that correspond to 82% of the studied species (Group 1), some conspecific sequences exhibited interspecific divergence and vice versa, indicating a mismatch of nomenclature and DNA barcode.

Some conspecific sequences (Group 2) have clustered separately in NJ and ML trees and have shown divergence above the threshold value. These groups may comprise of either some erroneously identified species or latent species. This high divergence may be an indicator of unidentified candidate species within named species. Previous DNA barcoding studies in other taxonomic groups have successfully identified cryptic species diversity within single known species (Mat Jaafar et al., 2012; Rougerie et al., 2012), for example ten undescribed species embedded within a single known species of skipper butterfly was delineated using DNA barcodes (Hebert et al., 2004). In some problematic cases, the number of representative sequences in the dataset was too few to be interpreted.

L. calcarifer, comprising of 94 sequences, formed 2 different clusters with 4 sequences clustering separately. The similarity search result of the 4 sequences using BLASTN have shown 99% (E-value = 0.00) match with *Pampus argenteus* while the remaining 90 sequences showed 100% match with *L. calcarifer* sequences of other countries. This clearly has shown that the 4 sequences marked as *L. calcarifer* are mislabeled, while the remaining 90 sequences represent true *L. calcarifer* species. The case of *C. batrachus* has been clarified as mislabeled in the database (Wong et al., 2011). Sequences of *B. bendelisis* and *B. barna* have formed 6 dispersed clusters; the mean conspecific divergence within each cluster was 0.8% and the sequences representative of each cluster maintained the interspecies threshold. The genus *Barilius* is comprised of at least 20 indigenous species from India (Froese and Pauly, 2012), among them only 5 species have been barcoded so far. Therefore, the sequences of *B. bendelisis* and *B. barna* altogether may represent 6 species (Fig. 4) which may either belong to the already described species or may represent latent species. Interestingly, *H. microps* (type locality: Dambuwa, Sri Lanka) is distinguished from its nominal congeners *H. fossilis* only by its caudal and anal fins being confluent (vs. separate) (Günther, 1864). It has long been considered as synonymous species of *H. fossilis* (Ferraris, 2007). Pethiyagoda et al., stated that *H. microps* is a result of anomalous fin regeneration in *H. fossilis*, injury being one of the possible cause and considered *H. microps* a junior synonym of *H. fossilis* (Pethiyagoda and Lanka, 1991). However, there are also reports of these two species being sympatric thus reducing chances of hybridization between the two species.

Table 4
Mean divergence within and between clusters of *Barilius* genus.

Clusters of species	Mean divergence within and between clusters of *Barilius* genus										
Barilius cluster1	**1.31**										
Barilius cluster2	4.78	**0.65**									
Barilius cluster3	4.67	3.20	**0**								
Barilius cluster4	6.01	4.41	2.96	**0.16**							
Barilius cluster5	9.94	8.15	6.89	6.89	**0**						
Barilius cluster6	9.21	11.92	10.75	12.39	16.30	**0**					
Barilius vagra	13.35	16.28	16.08	16.31	18.84	17.61	**0.09**				
Barilius gatensis	20.35	22.81	22.35	23.05	23.94	22.08	15.07	**0.18**			
Opsarius bakeri	17.62	19.60	19.31	20.02	21.61	19.21	16.53	13.08	**0.88**		
Opsarius canarensis	18.16	20.18	19.70	20.63	23.94	19.81	16.84	12.21	4.99	**n/c**	
Barilius tileo	20.53	23.41	23.70	24.93	26.18	22.60	18.46	16.77	17.61	16.12	**12.73**

Diagonals represent mean divergence (in bold) within the group, and the remaining cells represent pair-wise divergence between the groups.

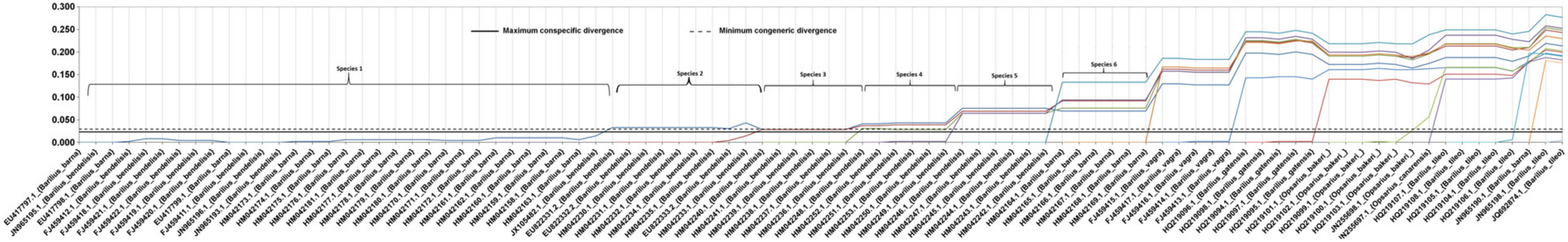

Fig. 4. Divergence summary (1:1, computed with K2P model) of all *COI* barcode sequences (90) belonging to the genus *Barilius* and *Opsarius*. The Y-axis represents K2P divergence value while the species are represented along the X-axis. The maximum conspecific divergence (2.3%, black solid line) and minimum congeneric divergence (2.9%, black dashed line) represent the threshold level of conspecific and congeneric divergence respectively. Sequences that represented the species such as *Barilius vagra*, *Barilius gatensis*, *Barilius tileo* (except sequences of accession number JN965198, JQ692874) *Opsarius bakeri* and *Opsarius canarensis* showed divergence below maximum conspecific value with their conspecific sequences and above minimum congeneric value with respect to all other congeneric sequences and are considered as true species. Based on the thresholds it is found that conspecific sequences of *Barilius bendelisis* and *Barilius barna* do not reveal conspecific divergence however altogether represents 6 species (shown as species 1–6 in the figure) and thereby masks some latent species.

In our study, one individual of *H. fossilis* has clustered away from rest of the 13 individuals of *H. fossilis* with divergence in congeneric range and has clustered with *H. microps* with low divergence, which creates the contention that *H. microps* is a distinct species and the doubtful sequence of *H. fossilis* is mislabeled. Sequences of *Channa orientalis* have formed 3 subclusters. Among them, two clusters (comprising of FJ459480–FJ459484 and JX105470–JX105474) clustered under a single node. However, one sequence (JQ667514) has clustered distinctly, with 5.8% mean distance from the remaining 10 sequences of *Channa orientalis*. This sequence belongs to a different geographical location and is the sole representative of that location. Hence, it is unclear whether this sequence belongs to a different species or represents a distant haplotype of the same species. Similarly, the distantly cluster sequences of *P. filamentosus* and *P. pardalis* have shown divergence of 5.9% and 8.2% with their respective conspecific sequences. These single isolated sequences, therefore, might be erroneous and further evaluation and inclusion of more sequences is required.

In Group 3, a total of 15 nominal species (Table 3) have exhibited fairly low interspecies divergence, caused by the inclusion of either some synonymous or misidentified species. There are some established cases of synonymy like, *T. macrolepis* and *T. mosal mahanadicus* as a synonym of *T. putitora* (Laskar et al., 2013). Moreover, inadequate taxonomic details often result in identifying species located in different geographical location as two different species. One such case of synonymy has been addressed recently where *M. horai* was reported as a junior synonym of *M. vittatus* (Bhattacharjee et al., 2012). Similarly, *L. rajasthanicus*is known only from its type locality, Jasiamand lake in Rajasthan and has never been reported after its initial documentation (Talwar and Jhingran, 1991). In our study, cohesive clustering of all 3 sequences of *L. rajasthanicus* with *L. dussumieri* (0% divergence) suggests that the former might be a junior synonym of the latter. *C. marulius* (n = 11) with conspecific divergence of 10% has formed 2 distinct clusters, one of which (accession numbers FJ459472–FJ459475) has clustered closely with *C. striata* (categorized as true species in 82% straightforward cases) with a divergence of 0.15% (S.E = 0.02) (Fig. 3d). Whereas, the divergence between the remaining sequences of *C. marulius* and *C. striata* was above the congenerics threshold. Such high intraspecific divergence in these *C. marulius* has also been reported earlier (Benziger et al., 2011) indicating that a few *C. striata* specimens have been erroneously identified as *C. marulius*, while the remaining 7 sequences represented true *C. marulius* sequence. Similarly, the other 4 cases of low interspecific divergence were may be due to erroneous identification that needs revision.

Thus confusion regarding 17 of the 32 problematic species has been resolved, thereby leading to the categorization of 87.4% of the studied species as true and valid in parity with the existing checklist.

Our survey has revealed that DNA barcoding of freshwater fish resources of India is far from being comprehensive with only 20% of the recorded species being barcoded until now. With the available DNA barcode data, 88% (approximately) of the species have been identified in parity with existing checklist. Thus, the species level demarcation based on the K2P divergence and NJ and ML based phylogenetic clustering of *COI* sequences is worthy. However, this study has detected some cases of erroneous identification, and the presence of some latent species, which have resulted in an incoherent reference barcode library of freshwater fishes of Indian subcontinent. Therefore, there is a need to revisit the specimens whose barcode data are erroneous and further taxonomic inquiry is recommended for species whose statuses are found to be doubtful. Thus, this study reflects the current quantitative and qualitative status of DNA barcoding of Indian freshwater fishes. It will also provide direction to future studies by highlighting the fish groups, which need to be barcoded. Further, the development of a comprehensive DNA barcode library of the nationwide fish resources would help to describe some new species and to understand the endemic fish species resources.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2013.12.047.

Conflict of interest

The authors report no conflict of interest.

Acknowledgment

Our humble acknowledgement goes to the Department of Biotechnology, Government of India for providing infrastructure facilities (BT/HRD/01/002/2007).

References

April, J., Mayden, R.L., Hanner, R.H., Bernatchez, L., 2011. Genetic calibration of species diversity among North America's freshwater fishes. Proc. Natl. Acad. Sci. U. S. A. 108, 10602–10607.

Becker, S., Hanner, R., Steinke, D., 2011. Five years of FISH-BOL: brief status report. Mitochondrial DNA 22 (Suppl. 1), 3–9.

Benziger, A., et al., 2011. Unraveling a 146 years old taxonomic puzzle: validation of Malabar snakehead, species-status and its relevance for channid systematics and evolution. PLoS One 6, e21272.

Bhattacharjee, M.J., Laskar, B.A., Dhar, B., Ghosh, S.K., 2012. Identification and re-evaluation of freshwater catfishes through DNA barcoding. PLoS One 7, e49950.

Carvalho, D.C., Neto, D.A., Brasil, B.S., Oliveira, D.A., 2011. DNA barcoding unveils a high rate of mislabeling in a commercial freshwater catfish from Brazil. Mitochondrial DNA 22 (Suppl. 1), 97–105.

Collins, R.A., Boykin, L.M., Cruickshank, R.H., Armstrong, K.F., 2012. Barcoding's next top model: an evaluation of nucleotide substitution models for specimen identification. Methods Ecol. Evol. 3, 457–465.

Darshan, A., Anganthoibi, N., Vishwanath, W., 2010. Redescription of the striped catfish *Mystus carcio* (Hamilton) (Siluriformes: Bagridae). Zootaxa 2475, 48–54.

de Oliveira Ribeiro, A., Caires, R.A., Mariguela, T.C., Pereira, L.H., Hanner, R., Oliveira, C., 2012. DNA barcodes identify marine fishes of Sao Paulo State, Brazil. Mol. Ecol. Resour. 12, 1012–1020.

Devi, K.R., Indra, T.J., Knight, J.D.M., 2010. Puntius rohani (Teleostei: Cyprinidae), a new species of barb in the *Puntius filamentosus* group from the southern Western Ghats of India. J. Threat. Taxa 2 (9), 121–1129.

deWaard, J.R., Hebert, P.D., Humble, L.M., 2011. A comprehensive DNA barcode library for the looper moths (Lepidoptera: Geometridae) of British Columbia, Canada. PLoS One 6, e18290.

Dishma, M., Vishwanath, W., 2012. Barilius profundus, a new cyprinid fish (Teleostei: Cyprinidae) from the Koladyne basin, India. J. Threat. Taxa 4, 2363–2369.

Eschmeyer, W.N., Fricke, R. (Eds.), 2012. Catalog of Fishes Electronic Version (Available: http://research.calacademy.org/ichthyology/catalog/fishcatmain.asp Accessed:2012 Dec 28).

Ferraris, C.J., 2007. Checklist of Catfishes, Recent and Fossil (Osteichthyes: Siluriformes), and Catalogue of Siluriform Primary Types. Magnolia Press.

Froese, R., Pauly, D. (Eds.), 2012. FishBase. World Wide Web Electronic Publication (Available:www.fishbase.org Accessed: 2012 Dec 28).

Günther, A., 1864. Catalogue of the Physostomi, Containing the Families Siluridae, Characinidae, Haplochitonidae, Sternoptychidae, Scopelidae, Stomiatidae in the Collection of the British Museum. order of the Trustees.

Hajibabaei, M., 2012. The golden age of DNA metasystematics. Trends Genet. 28, 535–537.

Hajibabaei, M., Janzen, D.H., Burns, J.M., Hallwachs, W., Hebert, P.D., 2006a. DNA barcodes distinguish species of tropical Lepidoptera. Proc. Natl. Acad. Sci. U. S. A. 103, 968–971.

Hajibabaei, M., Singer, G.A., Hickey, D.A., 2006b. Benchmarking DNA barcodes: an assessment using available primate sequences. Genome 49, 851–854.

Hebert, P.D., Cywinska, A., Ball, S.L., deWaard, J.R., 2003. Biological identifications through DNA barcodes. Proc. Biol. Sci. 270, 313–321.

Hebert, P.D., Penton, E.H., Burns, J.M., Janzen, D.H., Hallwachs, W., 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly Astraptes fulgerator. Proc. Natl. Acad. Sci. U. S. A. 101, 14812–14817.

Hernandez-Davila, A., Vargas, J.A., Martinez-Mendez, N., Lim, B.K., Engstrom, M.D., Ortega, J., 2012. DNA barcoding and genetic diversity of phyllostomid bats from the Yucatan Peninsula with comparisons to Central America. Mol. Ecol. Resour. 12, 590–597.

Hoagland, K.E., 1996. The Taxonomic Impediment and the Conventionon Biodiversity. ASC News 24 (61–62), 66–67.

Jayaram, K.C., 1999. The freshwater fishes of the Indian region. Narendra Pub. House.

Kim, D.W., et al., 2012. DNA barcoding of fish, insects, and shellfish in Korea. Genomics Inform. 10, 206–211.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.

Kottelat, M., Lim, K.K.P., 1995. Freshwater fishes of Sarawak and Brunei Darussalam: a preliminary annotated checklist. Sarawak Mus. J. 48, 227–258.

Lakra, W.S., et al., 2011. DNA barcoding Indian marine fishes. Mol. Ecol. Resour. 11, 60–71.

Lara, A., et al., 2010. DNA barcoding of Cuban freshwater fishes: evidence for cryptic species and taxonomic conflicts. Mol. Ecol. Resour. 10, 421–430.

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

Leveque, C., Oberdorff, T., Paugy, D., Stiassny, M.L.J., Tedesco, P.A., 2008. Global diversity of fish (Pisces) in freshwater. Hydrobiologia 198, 545–567.

Lijtmaer, D.A., Kerr, K.C., Barreira, A.S., Hebert, P.D., Tubaro, P.L., 2011. DNA barcode libraries provide insight into continental patterns of avian diversification. PLoS One 6, e20744.

Mabragana, E., Diaz de Astarloa, J.M., Hanner, R., Zhang, J., Gonzalez Castro, M., 2011. DNA barcoding identifies Argentine fishes from marine and brackish waters. PLoS One 6, e28655.

Mat Jaafar, T.N., Taylor, M.I., Mohd Nor, S.A., de Bruyn, M., Carvalho, G.R., 2012. DNA barcoding reveals cryptic diversity within commercially exploited Indo-Malay Carangidae (Teleosteii: Perciformes). PLoS One 7, e49623.

Meier, R., Zhang, G., Ali, F., 2008. The use of mean instead of smallest interspecific distances exaggerates the size of the "barcoding gap" and leads to misidentification. Syst. Biol. 57, 809–813.

Mendonca, A., Cunha, A., Chakrabarti, R., 2012. Natural Resources, Sustainability and Humanity: A Comprehensive View. Springer.

Menon, A.G.K., 1999. Check list—Fresh water fishes of India. Records of the Zoological Survey of India. Misc. Publ. 175, 1–366.

Nath, P., Dey, S.C., 1989. Two new species of the genus Amblyceps Blyth from Arunachal Pradesh, India. J. Assam Sci. 32, 1–6.

Ng, H.H., Hadiaty, R.K., 2009. Ompok brevirictus, new catfish (Teleostei: Siluridae) from Sumatra. Zootaxa 2232, 50–60.

Pereira, L.H., Pazian, M.F., Hanner, R., Foresti, F., Oliveira, C., 2011. DNA barcoding reveals hidden diversity in the Neotropical freshwater fish Piabina argentea (Characiformes: Characidae) from the Upper Parana Basin of Brazil. Mitochondrial DNA 22 (Suppl. 1), 87–96.

Pethiyagoda, R., Lanka, W.H.T.o.S., 1991. Freshwater Fishes of Sri Lanka. Wildlife Heritage Trust of Sri Lanka.

Pethyagoda, R., Kottelat, M., 1994. Three new species of fishes of the genera Osteochilichthys (Cyprinidae), Travancoria (Balitoridae) and Horabagrus (Bagridae) from the Chalakudy River, Kerela, India. J. South Asian Nat. Hist. 1, 97–116.

Ponniah, A.G., Sarkar, U.K., 2000. Fish Biodiversity of North East India. National Bureau of Fish Genetic Resources, North Eastern Council.

Puillandre, N., Lambert, A., Brouillet, S., Achaz, G., 2012. ABGD, Automatic Barcode Gap Discovery for primary species delimitation. Mol. Ecol. 21, 1864–1877.

Ratnasingham, S.H.P., 2007. BOLD: The Barcode of Life Data System. (http://www.barcodinglife.org) Mol. Ecol. Notes 7, 355–364.

Roberts, T.R., 1992. Revision of the striped catfishes of Thailand misidentified as Mystus vittatus, with description of two new species (Pisces: Bagridae). Ichthyol. Explor. Freshwater 3, 77–88.

Rougerie, R., Naumann, S., Nassig, W.A., 2012. Morphology and molecules reveal unexpected cryptic diversity in the enigmatic genus Sinobirma Bryk, 1944 (Lepidoptera: Saturniidae). PLoS One 7, e43920.

Rubinoff, D., Cameron, S., Will, K., 2006. A Genomic Perspective on the Shortcomings of Mitochondrial DNA for "Barcoding" Identification. J. Hered. 97, 581–594.

Sambrook, J., Russell, D.D.W., 2001. Molecular Cloning: A Laboratory Manual. Cold Spring Harbor Laboratory Press.

Selim, K., Vishwanath, W., 2002. A new cyprinid fish species of Barilius Hamilton from the Chatrickong River, Manipur, India. J. Bombay Nat. Hist. Soc. 99, 267–270.

Sievers, F., et al., 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol. Syst. Biol. 7.

Srivathsan, A., Meier, R., 2012. On the inappropriate use of Kimura-2-parameter (K2P) divergences in the DNA-barcoding literature. Cladistics 28, 190–194.

Talwar, P.K., Jhingran, A.G., 1991. Inland fishes of India and adjacent countries. Oxford & IBH Pub. Co.

Tamura, K., et al., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol. 10, 2731–2739.

Vishwanath, W., Linthoingambi, I., 2005. A new sisorid catfish of the genus Glyptothorax Blyth from Manipur, India. J. Bombay Nat. Hist. Soc. 102 (2), 201–203.

Vishwanath, W., Linthoingambi, I., 2007. Redescription of catfishes Amblyceps arunachalensis Nath and Dey and Amblyceps apangi Nath and Dey (Teleostei: Amblyciptidae). Zoos Print J. 22, 2662–2664.

Wang, Z.D., Guo, Y.S., Liu, X.M., Fan, Y.B., Liu, C.W., 2012. DNA barcoding South China Sea fishes. Mitochondrial DNA 23, 405–410.

Ward, R.D., 2012. FISH-BOL, a case study for DNA barcodes. Methods Mol. Biol. 858, 423–439.

Ward, R.D., Zemlak, T.S., Innes, B.H., Last, P.R., Hebert, P.D., 2005. DNA barcoding Australia's fish species. Philos. Trans. R. Soc. B-Biol. Sci. 360, 1847–1857.

Waugh, J., 2007. DNA barcoding in animal species: progress, potential and pitfalls. Bioessays 29, 188–197.

Wong, L.L., et al., 2011. DNA barcoding of catfish: species authentication and phylogenetic assessment. PLoS One 6, e17812.

Zhang, D.X., Hewitt, G.M., 1996. Nuclear integrations: challenge for mitochondrial DNA markers. Trends Ecol. Evol. 11 (11), 247–251.

Zhang, J., Kapli, P., Pavlidis, P., Stamatakis, A., 2013. A general species delimitation method with applications to phylogenetic placements. Bioinformatics btt499.

http://informahealthcare.com/mdn
ISSN: 1940-1736 (print), 1940-1744 (electronic)

Mitochondrial DNA, Early Online: 1–3
© 2014 Informa UK Ltd. DOI: 10.3109/19401736.2013.855923

SHORT COMMUNICATION

Unraveling the sequence information in *COI* barcode to achieve higher taxon assignment based on Indian freshwater fishes

Mohua Chakraborty and Sankar Kumar Ghosh

Department of Biotechnology, Assam University, Silchar, Assam, India

Abstract

Efficacy of cytochrome c oxidase subunit I (*COI*) DNA barcode in higher taxon assignment is still under debate in spite of several attempts, using the conventional DNA barcoding methods, to assign higher taxa. Here we try to understand whether nucleotide and amino acid sequence in *COI* gene carry sufficient information to assign species to their higher taxonomic rank, using 160 species of Indian freshwater fishes. Our results reveal that with increase in the taxonomic rank, sequence conservation decreases for both nucleotides and amino acids. Order level exhibits lowest conservation with 50% of the nucleotides and amino acids being conserved. Among the variable sites, 30–50% were found to carry high information content within an order, while it was 70–80% within a family and 80–99% within a genus. High information content shows sites with almost conserved sequence but varying at one or two locations, which can be due to variations at species or population level. Thus, the potential of *COI* gene in higher taxon assignment is revealed with validation of ample inherent signals latent in the gene.

Keywords

BOLD, *COI* gene, DNA barcode, Genbank, higher taxon assignment, Indian freshwater fishes, R_{seq}, sequence conservation

History

Received 13 July 2013
Revised 3 October 2013
Accepted 5 October 2013
Published online 10 January 2014

Introduction

DNA barcode has proven to be an effective tool in species identification and resolution of various taxonomic impediments. Hebert et al. (2003a), while proposing barcoding of all animal life using *COI* gene, advocated that diversity in the nucleotides and coded amino acids of the 5′ section of this gene, was sufficient to reliably place species into higher taxonomic categories along with discriminating the closely allied species. Over the years, the efficiency of *COI* gene in accurate species identification has been proved for various groups of animals (Dinca et al., 2011; Sass et al., 2007; Saunders, 2005; Ward, 2012). However, its efficacy in assigning species to their higher taxonomic ranks viz. genus, family, order, is still dubious (Wilson et al., 2011). Many methods of higher taxon assignment using DNA barcodes have been proposed (Bergmann et al., 2013; Erickson & Driskell, 2012; Rach et al., 2008; Ward et al., 2005) but, none could be standardized.

Diagnostic attributes of DNA barcode is an outcome of two phenomena. First, flexibility acquired from various combinations of four nucleotides at each position. Second, constriction, governed by several constraints like codon bias, degenerate nature of genetic code and transition – transversion bias (DeSalle et al., 2005; Hebert et al., 2003a, b). Both these cumulatively define unique characteristic of each group of closely related members. Thus, variation pattern of nucleotide and amino acid content of *COI* gene across various taxonomic ranks might have inherent information to enable classification at each level.

Here, we analyze the DNA sequence and translated protein sequence of the barcode region of *COI* from 160 species of Indian freshwater fishes. We choose this group, as fishes are one of the highest explored groups of organism using barcodes (Ward, 2012) and many incongruities exist regarding higher level assignments in Indian freshwater fishes (Talwar & Jhingran, 1991). Our results reveal that, with increase in taxonomic rank, sequence conservation decreases for both nucleotides and amino acids. Order level exhibits lowest conservation and 30–50% of the variable sites, were found to carry high information content, while it was 70–80% within a family and 80–99% within a genus. High information content represents sites with nearly conserved sequence varying at one or two locations, which can be due to variations at species or population level. Thus, the potential of *COI* gene in higher taxon assignment is revealed with validation of ample inherent signal latent in the gene.

Materials and methods

COI barcode sequences of all Indian freshwater fish species available in the Public Data Portal of BOLD (Ratnasingham & Hebert, 2007) and GenBank were retrieved. Among them, three orders, Siluriformes, Cypriniformes and Perciformes were selected for the study as they were well represented by barcodes at different taxonomic ranks. The accession number of the sequences along with their taxonomic identity has been provided in the Supplementary material. Sequences greater than 600 bp, with no missing nucleotides or gaps, were included, thereby reducing the possibility of NUMTs (Zhang & Hewitt, 1996). The sequences were aligned using Clustal Omega (Sievers et al., 2011) and MEGA 5.1 (Tamura et al., 2011). These sequences were of varying length so they had to be trimmed at either end to exclude any gaps in the alignment file. This also reduced the possibility of including very low-frequency nucleotide variants (nVLFs) arising from sequencing errors (Stoeckle & Kerr, 2012). Finally a set of 510 bp long consensus sequences was obtained.

Information content at each nucleotide or amino acid position from the sequences was calculated using Schneider & Stephens

Correspondence: Prof. Sankar Kumar Ghosh, Department of Biotechnology, Assam University, Silchar 788011, Assam, India. Tel: 91 9435372338. E-mail: drsankarghosh@gmail.com

(1990) method. Here the degree of sequence conservation per site R_{seq} is defined as:

$$R_{\text{seq}} = \log_2 N - \left(-\sum_p \log_2 p\right)$$

where N is the number of options per site (4 for DNA, 20 for proteins) and p is the observed frequency of each nucleotide base or amino acid at a particular position.

At each taxonomic rank, the site with the highest possible value of R_{seq} (2 for nucleotides and 4.32 for amino acids) represented the site conserved at that level. For those that represented lower R_{seq} values, we further looked into lower taxonomic rank to locate where the saturation took place.

Results and discussion

A total of 1306 sequences belonging to 22 families, 71 genera and 160 species of Indian freshwater fishes (Supplementary material) were analyzed to reveal the variation pattern of sequence conservation of *COI* gene across various taxonomic ranks. Figure 1 shows, as we move higher in taxonomic hierarchy (i.e. from species to order) nucleotide conservation decreases. The three orders studied here shows 47–71% conservation in their nucleotide composition followed by 77–98% within their families and 77–99% within the genera. Families having only one or few genera showed high range of conservation (>90%) which is actually a reflection of the conservation in the underlying genus.

The degree of sequence conservation was studied for each of these variable sites to estimate the amount of sequence information available in each taxonomic rank thereby facilitating their identification. Information content of each of these sites, measured in terms of R_{seq} values, were found to increase from order to species level (Table 1). All the three orders had a significant proportion of variable sites (90%) with R_{seq} value lying in the range of 0.50–1.99 and 5% sites in the range of 0.00–0.49. Thus, at order level, at least 50% of the sites had conserved nucleotide base pairs and 90% of the variable sites had possibility of occurrence of two nucleotides. Furthermore, in 74% sites, amino acids were variable between the different orders studied. However, within the orders, 52% variable amino acid sites were present in Cypriniformes, 49% sites in Siluriformes and 29% sites in Perciformes. Most of the sites had possibility of occurrence of two amino acids at a given position.

At family level, percentage of conserved and nearly conserved nucleotide sites (sites with R_{seq} value 1.50–1.99) rose to around 70–80% and at genus level, it further increased to 80–99%.

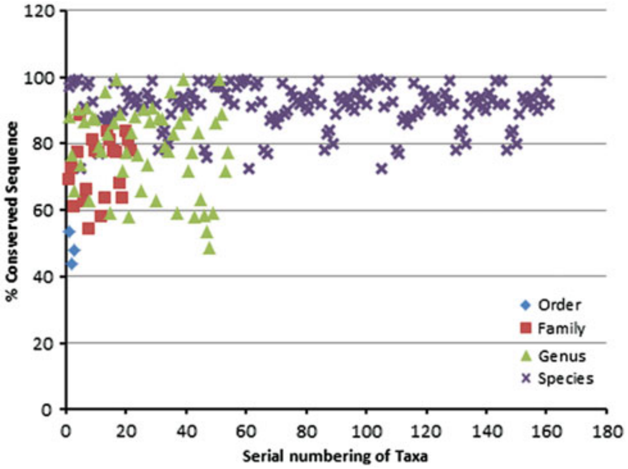

Figure 1. Pattern of distribution of sequence conservation across various taxonomic ranks of Indian freshwater fishes. The horizontal axis represents serial arrangements of different taxa. Different markers mark different taxonomic ranks viz: orders are marked by diamonds, families by squares, genera by triangles and species by crosses. Vertical axis represents percentage of conserved sequence in each taxon. Clustering of the markers in the scatter diagram represents the variation pattern of conserved sequences in the representative taxon.

Table 1. Frequency distribution of R_{seq} value of variable nucleotides and amino acids in various taxonomic ranks of Indian freshwater fishes.

Taxa			Nucleotides R_{seq} values					Amino acid R_{seq} values				
Order	Family	Genus	0–0.49	0.5–0.99	1–1.49	1.5–1.99	2–2.49	3.12–3.41	3.42–3.71	3.72–4.01	4.02–4.31	4.32–4.61
Cypriniformes			23	62	81	120	0	0	3	5	82	0
	Balitoridae		6	23	118	4	135	0	3	6	79	2
		Acanthocobitis	0	0	20	2	264	1	1	1	0	87
		Nemacheilus	0	0	2	0	284	0	0	0	0	90
		Schistura	0	2	68	0	216	0	0	0	0	90
		Balitora	0	0	1	0	285	0	0	0	0	90
	Cyprinidae		20	61	79	124	2	1	0	3	1	85
		Labeo	3	8	84	27	164	0	0	0	7	83
		Tor	0	1	12	44	229	1	0	0	14	75
		Puntius	18	40	92	53	83	2	2	8	7	71
		Barilius	2	23	92	77	92	7	3	19	7	54
		Neolissichilus	0	1	14	7	264	1	0	0	2	87
		Garra	0	8	72	26	180	0	3	2	0	85
		Osteobrama	0	4	83	1	198	0	6	1	0	83
Siluriformes			26	55	83	73	0	0	2	5	42	0
	Schilbeidae		5	13	78	42	99	0	1	1	3	44
		Eutropiichthys	0	0	59	2	176	0	0	2	0	47
	Siluridae		5	16	82	37	97	1	1	0	1	46
		Ompok	5	2	94	18	118	0	2	0	1	46
	Bagridae		23	46	96	34	38	1	2	6	8	32
		Sperata	0	0	50	0	187	0	1	0	0	48
		Mystus	13	44	101	31	48	1	1	4	8	35
	Sisoridae		2	25	64	109	37	1	1	0	20	27
		Glyptothorax	0	5	76	53	103	1	1	0	6	41
Perciformes			24	61	95	86	1	4	3	1	41	1
		Channa	16	41	105	53	53	5	0	18	8	18
		Lates	0	0	5	36	227	0	0	1	4	45

Moreover, a major share of the sites had a bias towards fewer numbers of nucleotide. Sites having variable amino acid sequence varied from as low as 3% in Siluridae to as high as 51% in Cyprinidae. However, in Cyprinidae, of the 88 variable sites, 79 sites had R_{seq} value close to 4.31 indicating that the variation was mostly due to presence of point mutations between species or population. At genus level, the percentage of variable amino acid sites further decreased to a lower range (0–11%). Thus, at family level, approximately 60–70% sequence information was carried by nucleotide and amino acid sequences cumulatively. While at genus level sequence information content rose to an approximate value of 80–90%.

Thus, sequence conservation and variable site information of nucleotides and amino acids together reflected the potential of the *COI* gene in categorizing the species to their respective higher taxonomic ranks. Potential of *COI* gene in higher taxon assignment and phylogeny was also observed in some previous studies (Hebert et al., 2003a; Rach et al., 2008; Wilson et al., 2011). Our results, in concordance to previous studies, reveal that, a pattern variation exists at each level of taxonomic hierarchy that endows the *COI* gene with inherent potential to discriminate species as well as higher taxa. Moreover, the bias towards binary pattern of variation of both nucleotides and amino acids observed in this study was also reported in other taxonomic groups such as aves (Stoeckle & Kerr, 2012). This suggests that, use of *COI* barcode in higher taxon assignment can be extended to other taxonomic groups as well. Higher taxon assignment with DNA barcodes will thus be advantageous with novel methods that will explore the information content latent in the gene.

Declarations of interest

The authors report no conflicts of interest. The authors alone are responsible for the content and writing of this article. Our humble acknowledgement goes to the Department of Biotechnology, Government of India for providing infrastructure facilities (BT/HRD/01/002/2007).

References

Bergmann T, Rach J, Damm S, Desalle R, Schierwater B, Hadrys H. (2013). The potential of distance-based thresholds and character-based DNA barcoding for defining problematic taxonomic entities by CO1 and ND1. Mol Ecol Resour 13:1069–81.

Desalle R, Egan MG, Siddall M. (2005). The unholy trinity: Taxonomy, species delimitation and DNA barcoding. Philos Trans R Soc Lond B Biol Sci 360:1905–16.

Dinca V, Zakharov EV, Hebert PD, Vila R. (2011). Complete DNA barcode reference library for a country's butterfly fauna reveals high performance for temperate Europe. Proc Biol Sci 278:347–55.

Erickson DL, Driskell AC. (2012). Construction and analysis of phylogenetic trees using DNA barcode data. Methods Mol Biol 858: 395–408.

Hebert PD, Cywinska A, Ball SL, Dewaard JR. (2003a). Biological identifications through DNA barcodes. Proc Biol Sci 270:313–21.

Hebert PD, Ratnasingham S, Dewaard JR. (2003b). Barcoding animal life: Cytochrome c oxidase subunit 1 divergences among closely related species. Proc Biol Sci 270:S96–9.

Rach J, Desalle R, Sarkar IN, Schierwater B, Hadrys H. (2008). Character-based DNA barcoding allows discrimination of genera, species and populations in Odonata. Proc Biol Sci 275:237–47.

Ratnasingham S, Hebert PD. (2007). BOLD: The Barcode of Life Data System (http://www.barcodinglife.org). Mol Ecol Notes 7: 355–64.

Sass C, Little DP, Stevenson DW, Specht CD. (2007). DNA barcoding in the cycadales: Testing the potential of proposed barcoding markers for species identification of cycads. PLoS One 2:e1154.

Saunders GW. (2005). Applying DNA barcoding to red macroalgae: A preliminary appraisal holds promise for future applications. Philos Trans R Soc Lond B Biol Sci 360:1879–88.

Schneider TD, Stephens RM. (1990). Sequence logos: A new way to display consensus sequences. Nucleic Acids Res 18:6097–100.

Sievers F, Wilm A, Dineen DG, Gibson TJ, Karplus K, Li W, Lopez R, et al. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7: 539.

Stoeckle MY, Kerr KC. (2012). Frequency matrix approach demonstrates high sequence quality in avian BARCODEs and highlights cryptic pseudogenes. PLoS One 7:e43992.

Talwar PK, Jhingran AG. (1991). Inland fishes of India and adjacent countries. New Delhi, India: Oxford & IBH Publishing Company.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–9.

Ward RD. (2012). FISH-BOL: A case study for DNA barcodes. Methods Mol Biol 858:423–39.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD. (2005). DNA barcoding Australia's fish species. Philos Trans R Soc Lond B Biol Sci 360:1847–57.

Wilson JJ, Rougerie R, Schonfeld J, Janzen DH, Hallwachs W, Hajibabaei M, Kitching IJ, et al. (2011). When species matches are unavailable are DNA barcodes correctly assigned to higher taxa? An assessment using sphingid moths. BMC Ecol 11:18.

Zhang DX, Hewitt GM. (1996). Nuclear integrations: Challenge for mitochondrial DNA markers. Trends Ecol Evol 11:247–51.

Supplementary material available online

Supplementary material

J. Environ. & Sociobiol. : **10**(1) : 25-32, 2013

REVEALING GENETIC DIVERSITY OF *CLARIAS BATRACHUS* USING DNA BARCODE

Mohua Chakraborty, S.K.Ghosh,* Bishal Dhar and Bijoya Khomdram Devi

DNA barcode and Genomics Laboratory, Department of Biotechnology, Assam University, Silchar-788011, Assam, India

ABSTRACT

Clarias batrachus is a species of freshwater catfish widely used for human consumption. Increasing demand of this species worldwide, coupled with its ability to survive in wide range of environmental conditions, has led to its introduction in many countries. This widespread translocation and distribution of *C. batrachus* have resulted in the rise of wide variety of haplotypes of this species. Nevertheless, there has been constant decrement in their population density in the last two decades. Thus, to endeavour conservation of the species we employ molecular technique of DNA barcoding in solving the standing problem of crypticism and haplotype sharing of the species. A better knowhow of the genetic makeup of the unique identifier region, that is, the 648 basepair region of COI DNA barcode will help to differentiate among closely related species and identify endemic species. In this study, a comparative analysis of *C. batrachus* from different regions in India and other parts of world shows presence of distinct haplotypes in different geographical locations. We also present a descriptive study of the various species of *Clarias* genus that have been barcoded in India till date. Our results also solve the dilemma of considering some species as synonymy of *C. batrachus.*

Key words : *Clarias batrachus, cytochrome c oxidase subunit I (COI), genetic diversity, Kimura's two parameter (K2P), mean divergence, neighbour joining tree.*

INTRODUCTION

The walking catfish, *Clarias batrachus*, is a species of freshwater air breathing catfish so named for its ability to "walk" across dry land, to find food or suitable environments. They normally inhabit in swamps, marshy and derelict waters (Lakra and Sarkar, 2007). The walking catfish is a native of South East Asia including Malaysia, Thailand, eastern India, Sri Lanka, Bangladesh, Burma, Indonesia, Singapore and Brunei (Kuang, 1986; Shrestha, 1978) and is also found in Philippines. Owing to its capability to adapt to wide range of environmental conditions, candidates of the genus *Clarias* has extensively travelled to many continents, adapting itself successfully and is now found throughout Asia and Africa.

* Telephone No: +919435372338, Fax: +91 3842270802, Email ID: drsankarghosh@gmail.com

Such introduction of new population of *C. batrachus* in addition to its existing wildtype species has resulted in enriching the ecosystem with wide variety of haplotypes of *C. batrachus*. However, their population has decreased significantly during the last two decades, owing mainly to imprudent development and reckless fishing. Conservation of this species is therefore the need of the hour. But, despite centuries of taxonomic inquiry, problems inherent to species identification continue to hamper the conservation of this species as *C. batrachus* for a long period of time (Ng and Hadiaty, 2011; Talwar and Jhingran, 1991). Many such populations of *C. batrachus* were originally identified as distinct species but later they were amalgamated with *C. batrachus* or considered as synonym of this species (Day, 1958; Talwar and Jhingran, 1991), though sufficient molecular data supporting this debate is lacking. There is also a significant lag in understanding the genetic diversity of the species. In such scenario, introduction of this species in a new environment may severely affect the existing population. Therefore, there is a need to understand the genetic composition of natural population of *C. batrachus* using molecular techniques to evaluate the latent genetic effects induced by hatchery operations. Among molecular methods, DNA barcoding has served as an effective tool in solving many distorted views of biodiversity (Hebert and Cywinska, 2003a; Hebert *et al.*, 2003b). Indeed, DNA barcoding surveys, using partial cytochrome c oxidase subunit I (COI) sequences, have revealed cryptic diversity across the animal kingdom (Dudgeon *et al.*, 2012; Janzen *et al.*, 2011; Mat Jaafar *et al.*, 2012; Puckridge *et al.*, 2012). For instance, previous DNA barcoding studies in other taxonomic groups have found as many as nine undescribed species embedded within a single known species of skipper butterfly (Hebert and Penton, 2004). Ward *et al.* (2005) had sequenced (barcoded) 655 bp region of the mitochondrial COI of 207 species of fish and recommended that COI barcode of all fish species could be generated from the same primer and all fish species could be differentiated by their COI sequence. Bucklin *et al.*(2011) calculated an average retrieval of 2 % new species in larger fish DNA barcoding studies, and they extrapolated this rate to about 600 overlooked or cryptic species to await discovery through similar studies. From the 31,000 species currently listed in the Catalog of Fishes, about 4000 have been described as new during the past 10 years (2000–2009), with 500 added in 2008 and 300 in 2009 (Eschmeyer *et al.*, 2012).

Thus, DNA barcode can be used as an effective tool to understand the genetic diversity of *C. batrachus.* In this study, we evaluate genetic diversity of Indian *C. batrachus* using mitochondrial cytochrome c oxidase subunit I (COI) sequences as DNA barcodes. Using distance based methods we attempt to resolve some of the key areas of discrepancies underlying the current taxonomic classification of this species. Moreover, our study will help to identify geographically isolated population of *C.batrachus*, which will further help in sustainable regulation of fishing practice of this economically important species.

MATERIAL AND METHODS

Data acquisition

A total of 50 COI barcode sequences of *C. batrachus* were mined from BOLD and GenBank. Among them, 12 sequences of *C. batrachus* were retrieved from BOLD and 38 sequences from NCBI. The two sets of sequences were checked for redundancy and a collection of 38 sequences of *C. batrachus* was used for final analysis. Further, these databases were mined for available sequences of other species of *Clarias* genus in Indian subcontinent and six *Clarias dussumieri* and five *Clarias gariepinus* were retrieved. Sequences were included provided they had length greater than 540 bp with no missing nucleotides or gaps. All sequences were aligned using ClustalX. Probable erroneous sequences (with highly unlikely positions or extreme branch lengths, based on a neighbour-joining tree calculated with all sequences) were identified and omitted. The COI coding DNA sequence alignments were guided by pre-aligned COI protein sequences using MEGA5 software.

Data analysis

Genetic distances were calculated to quantify sequence divergences among individuals using Kimura's two parameter (K2P) models (Kimura, 1980), as implemented in MEGA 5.1 (Tamura *et al.,* 2011). Interspecific K2P distances were calculated for those species with at least two sequences, and intraspecific K2P distances were calculated between species in the entire data set. Genetic distances were analysed at species level in MEGA 5.1. Neighbour joining analyses were also conducted independently for COI using K2P distance model using MEGA 5.1. Node support was evaluated with 1000 bootstrap pseudoreplicates.

RESULTS AND DISCUSSION

We obtained 18 DNA barcode sequences of Indian *C. batrachus* deposited in NCBI, among which two sequences are also deposited in BOLD. These sequences belonged to specimens collected from different geographic regions in India. Of the 18 *C. batrachus* barcodes, five specimens belonged to Alibagh coast, Mumbai, Maharashtra and another four were from Mathabhanga, West Bengal. Further, two sequences were generated in our lab which were collected from Lala, Assam and these are the only barcode sequences of *C. batrachus* to be submitted in BOLD (Bhattacharjee *et al.*, 2012), because the remaining 7 sequences in NCBI geographic location is not specified. We also retrieved 20 other COI sequences of *C. batrachus* from different countries that have been barcoded till now (Table 1).

Table 1. Summary of *Clarias batrachus* sequences analysed

Name of species	Geographic location	NCBI	BOLD
Clarias batrachus	Thailand	13	0
Clarias batrachus	Philippines	5	4
Clarias batrachus	Vietnam	1	1
Clarias batrachus	India	18	2
Clarias fuscus	China	1	0
Unspecified		0	5
	Total	38	12

Distance analysis of COI barcode sequence based on K2P method of *C. batrachus* species from India showed large conspecific mean divergence (6.85 ± 0.76) % thus indicating either the presence of different haplotypes of *C. batrachus* in India or presence of some mislabelled sequence. Geographically isolated populations of same species often show high genetic deviation. To explore the possibility of presence of distinct haplotypes of *C. batrachus,* the sequences were grouped according to different geographic locations (within India), from which they have been collected and genetic variation within and between the groups were analysed (Table 2 and Table 3). We observed low conspecific mean genetic distance for *C. batrachus* species of a particular geographic location. Mean genetic distance of *C. batrachus* from Alibagh coast, Mumbai,

Table 2. Conspecific mean genetic divergence of different species of *Clarias* genus of different geographical locations

Species	Geographic location	Mean% Dist	S.E
Clarias batrachus	Thailand	0.12	0. 08
Clarias batrachus	Philippines	0	0
Clarias batrachus	Vietnam	6.85	0.07
Clarias batrachus	India (all sequences)	N/A	N/A
Clarias batrachus	India (not available)	5.66	0.62
Clarias batrachus	India (Maharashtra ,Mumbai)	0.19	0.10
Clarias batrachus	India (Mathabhanga, West Bengal)	0.31	0.18
Clarias batrachus	India (Lala, Assam)	1.51	0.47
Clarias dussumieri	India (not available)	0.30	0.19
Clarias gariepinus	India (not available)	0.31	0.46

Table 3. Mean genetic distance between different samples of *Clarias* genus of different geographical locations

Sample 1	Sample 2	Mean% Dist	S. E.
Clarias batrachus India	*Clarias batrachus* Thailand	12.29	1.20
Clarias batrachus India	*Clarias batrachus* Philippines	11.37	1.16
Clarias batrachus India	*Clarias batrachus* Vietnam	8.95	1.04
Clarias batrachus Thailand	*Clarias batrachus* Vietnam	13.80	1.60
Clarias batrachus Philippines	*Clarias batrachus* Vietnam	12.33	1.52
Clarias batrachus Thailand	*Clarias batrachus* Philippines	2.51	0.63
Clarias batrachus India	*Clarias batrachus* India (Lala, Assam)	1.80	0.40
Clarias batrachus India	*Clarias batrachus* India (Mumbai, Maharashtra)	1.85	0.50
Clarias batrachus India (Lala, Assam)	*Clarias batrachus* India (Mumbai, Maharashtra)	0.93	0.30
Clarias batrachus India	*Clarias batrachus* India	12.69	1.50
Clarias batrachus India (Lala, Assam)	*Clarias batrachus* India	13.51	1.50
Clarias batrachus India (Mumbai, Maharashtra)	*Clarias batrachus* India	13.11	1.50
Clarias batrachus India	*Clarias batrachus* India (Mathabhanga, West Bengal)	12.92	1.50
Clarias batrachus India (Lala, Assam)	*Clarias batrachus* India (Mathabhanga, West Bengal)	13.63	1.50
Clarias batrachus India (Mumbai, Maharashtra)	*Clarias batrachus* India (Mathabhanga, West Bengal)	13.34	1.51
Clarias batrachus India	*Clarias batrachus* India (Mathabhanga, West Bengal)	0.87	0.30

Maharashtra, Mathabhanga, West Bengal and Lala, Assam were found to be 0.5%, 0.31% and 1.5% respectively, while remaining 7 individuals for which no location has been specified, show an overall mean of 0.15%. The mean distance between populations of Lala, Assam and Mathabhanga, West Bengal were found to be as high as 13.639 %. Many other interpopulation divergences also showed similar high range of conspecific divergence that lies in the range of congeneric divergence. However, not all populations exhibited high divergence, *e.g.,* species from Lala, Assam and Alibagh coast, Maharashtra shared narrow divergence of 0.934%, thereby indicating that geographic isolation can be partly but not solely responsible for the observed high genetic divergence within Indian *C. batrachus* species.

Altogether, six sequences of Indian *C. batrachus* cluster separately from all other sequences of *C. batrachus* and diverge from the other Indian *C. batrachus* sequences by 13.21 ± 1.25%. Moreover, with exclusion of these six sequences, conspecific divergence of remaining Indian *C. batrachus* sequences lowers to 1.21 ± 0.25%. This indicates the presence of cryptic species diversity within Indian *C. batrachus* and that the six sequences might represent a different species. In previous taxonomic classification, many species showed conflict with *C. batrachus* and some are considered to be synonyms (Ferraris, 2007) like *C. assamensis, C. punctatus, C. marpus, C. magur, C. fuscus, etc.* (Day, 2011). Among these species many were traditionally found in India and some were introduced later (Talwar and Jhingran, 1991). Thus, barcode analysis indicates that discrepancy in conspecific mean divergence of Indian *C. batrachus* may be due to the presence of one or more of these species wrongly identified as *C. batrachus.* A comparison of DNA barcodes of the aberrant *C. batrachus* with barcode sequence of all other representative species of *Clarias* genus will verify whether these sequences belong to some previously described species or represent a latent or different species. However, *C. dussumieri* (six sequences in NCBI, 0 in BOLD) and *C. gariepinus* (five sequences in NCBI, 0 in BOLD) are the only species of *Clarias* genus other than *C. batrachus* for which barcode sequences are available from India. These two species cluster distinctly from both the clusters of *C. batrachus* and show divergence from *C. batrachus* in the range of congeneric divergence, thus validating that the aberrant sequences of *C. batrachus* is not any of the two barcoded species of *C. batrachus*. Further, taxonomic investigation and inclusion of more samples along with a comprehensive barcoding of all species of *Clarias* genus is required to resolve this issue.

Comparing barcodes of *C. batrachus* of different countries it reveals that populations of different countries also show high divergence between them (Table 3). *C. batrachus* sequences from Thailand and Philippines showed low conspecific divergence of 0 % and 0.121 ± .08% and formed single cohesive cluster in Neighbour joining tree. Further, sequences of the two countries form two distinct clusters (Fig. 1) with divergence between them being 2.5% (Table 3). This indicates that sequences of *C. batrachus* from these two countries represent unique haplotype specific for each country. However, narrow divergence in Thailand and Philippines (Table 2) is not conclusive of the diversity of *C. batrachus* in these countries as small sample size and inadequate sampling within these countries may result in the present observation. Barcode of a synonym species of *C. batrachus* from China of the name *C. fuscus* was retrieved from GenBank, and it was found to cluster closely with *C. batrachus* from Philippines (Fig 1) in the NJ tree with 0 % congeneric distance. This

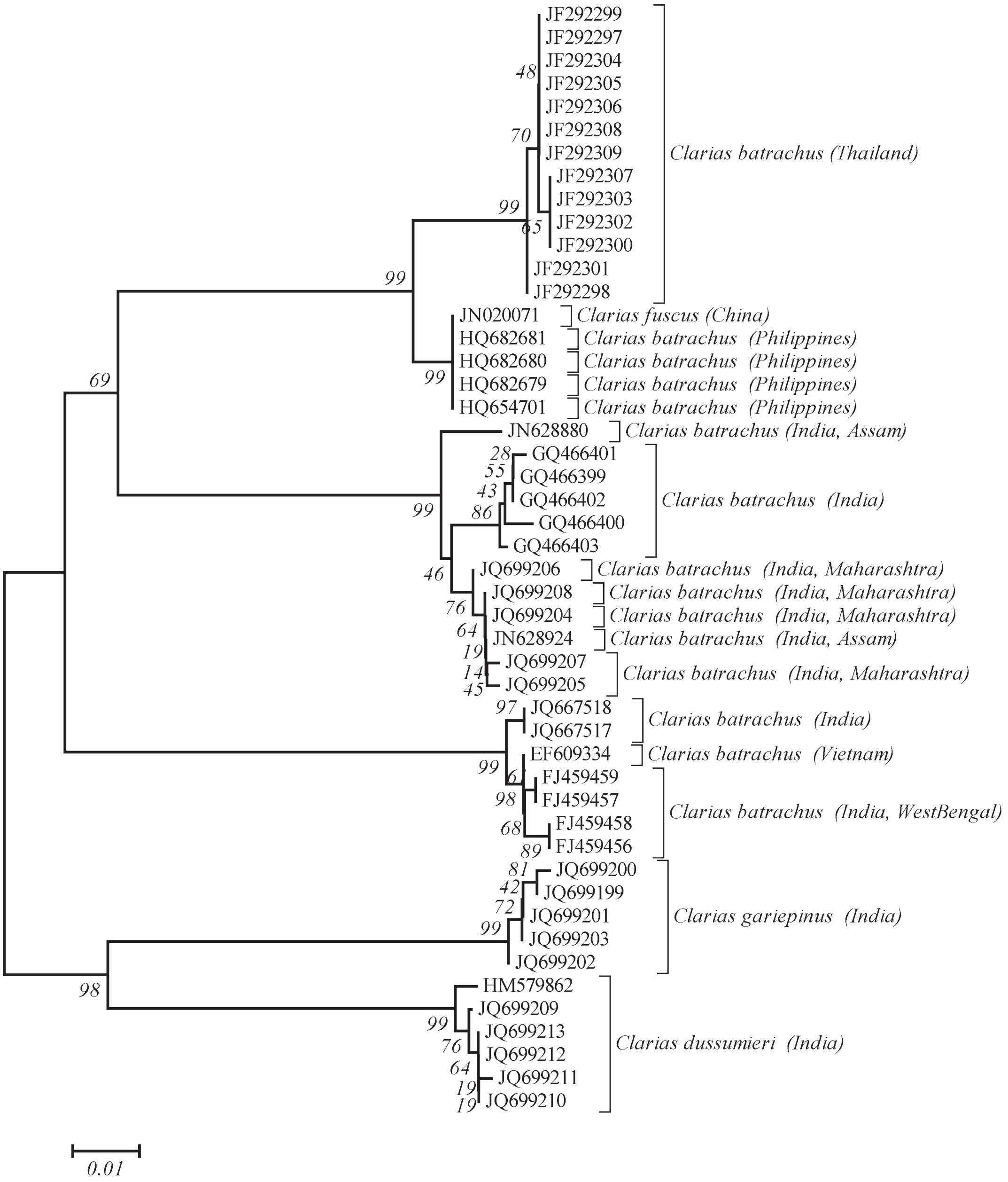
JF292299
JF292297
JF292304
JF292305
JF292306
JF292308
JF292309
JF292307
JF292303
JF292302
JF292300
JF292301
JF292298
48
70
99
65
99
Clarias batrachus (Thailand)
JN020071
Clarias fuscus (China)
HQ682681
HQ682680
HQ682679
HQ654701
Clarias batrachus (Philippines)
Clarias batrachus (Philippines)
Clarias batrachus (Philippines)
Clarias batrachus (Philippines)
99
69
JN628880
Clarias batrachus (India, Assam)
28
GQ466401
55
GQ466399
43
GQ466402
86
GQ466400
GQ466403
Clarias batrachus (India)
99
46
JQ699206
Clarias batrachus (India, Maharashtra)
JQ699208
Clarias batrachus (India, Maharashtra)
76
JQ699204
Clarias batrachus (India, Maharashtra)
64
JN628924
Clarias batrachus (India, Assam)
19
JQ699207
14
JQ699205
45
Clarias batrachus (India, Maharashtra)
97
JQ667518
JQ667517
Clarias batrachus (India)
EF609334
Clarias batrachus (Vietnam)
99
61
FJ459459
98
FJ459457
68
FJ459458
89
FJ459456
Clarias batrachus (India, WestBengal)
81
JQ699200
42
JQ699199
72
JQ699201
99
JQ699203
JQ699202
Clarias gariepinus (India)
98
HM579862
JQ699209
99
JQ699213
76
JQ699212
64
JQ699211
19
19
JQ699210
Clarias dussumieri (India)
0.01

Fig 1 Neighbour joining tree (based on K2P parameter) representing clustering of Indian *Clarias batrachus* species with C. *batrachus* species of different countries and other species of *Clarias* genus barcoded in India till date. *Regional information within a country (if available) are included in parentheses.

supports the contention that *C. fuscus* is a synonymy of *C. batrachus.* Nevertheless, inclusion of more sequences of individuals sharing morphological symmetry with the specimen of *C. fuscus* is required to validate that these two species are synonymy and thus the two species can be amalgamated into single named species.

Further, even after exclusion of the six aberrant sequences of Indian *C. batrachus*, conspecific divergence of Indian *C. batrachus* sequences with the same named species of other countries shows high mean divergence of 10.51 ± 1.25%. Similarly as observed above, all other sequences of *C. batrachus* show high conspecific mean divergence between populations of different countries. These observations suggest presence of unique haplotypes of *C. batrachus* with high genetic divergence between populations of different countries. This may be the result of high rate of introduction of *C. batrachus* species throughout the world, which have resulted in enriching the ecosystem with wide variety of haplotypes of *C. batrachus.*

CONCLUSION

This study reveals the presence of high genetic diversity of *C. batrachus* species across the world. Species of *C. batrachus* shows high range of divergence between different countries, which is easily traceable by COI barcodes. However, there is a significant lack of proper sampling of barcode sequences of all recorded populations of this economically important species in India and other countries which when done will definitely resolve the existing dilemma in this species. This is essential for proper regulation of import, export and introduction of a population to a new environment thereby ensuring sustainable fishing practice.

REFERENCES

Bhattacharjee, M. J., Laskar, B. A., Dhar, B. and Ghosh, S. K. 2012. Identification and re-evaluation of freshwater catfishes through DNA Barcoding. *PloS One*, **7**(11): e49950

Bucklin, A., Steinke, D. and Blanco-Bercial, L. 2011. DNA barcoding of marine metazoa. *Annual Review of Marine Science*, **3**: 471-508.

Day, F. 1958. *The fishes of India: being a natural history of the fishes known to inhabit the seas and fresh waters of India, Burma and Ceylon.* Dawson.

Dudgeon, C. L., Blower, D. C., Broderick, D., Giles, J. L., Holmes, B. J., Kashiwagi, T., Kruck, N. C., Morgan, J. A., Tillett, B. J. and Ovenden, J. R. 2012. A review of the application of molecular genetics for fisheries management and conservation of sharks and rays. *Journal of Fish Biology*, **80**: 1789-1843.

Eschmeyer, W. N., Fricke, R. (Eds.) 2012. *Catalog of Fishes electronic version*, Available:http: //research.calacademy.org/ichthyology/catalog/fishcatmain.asp Accessed: 2012 Dec 28.

Ferraris, C. J. 2007. *Checklist of Catfishes, Recent and Fossil (Osteichthyes: Siluriformes), and Catalogue of Siluriform Primary Types*. Magnolia Press.

Hebert, P. D. and Cywinska, A. 2003a. Biological identifications through DNA barcodes. *Proceedings. Biological Sciences / The Royal Society*, **270** (1512): 313-321.

Hebert, P. D. and Penton, E. H. 2004. Ten species in one: DNA barcoding reveals cryptic species in the neotropical skipper butterfly *Astraptes fulgerator*. *Proceedings of the National Academy of Sciences of the United States of America*, **101**(41): 14812-14817.

Hebert, P. D., Ratnasingham, S. and de Waard, J. R. 2003b. Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings. Biological sciences / The Royal Society*, **270 (**Suppl 1): 96-99.

Janzen, D. H., Hallwachs, W., Burns, J. M., Hajibabaei, M., Bertrand, C. and Hebert, P. D. 2011. Reading the complex skipper butterfly fauna of one tropical place. *PloS One*, **6** (8): e19874.

Kuang, D. H. and Ni, Y. 1986. *The freshwater and estuaries fishes of Hainan Island.* Guangdong Science and Technology Press.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution,* **16** (2): 111-120.

Lakra, W. S. and Sarkar, U. K. 2007. *Fresh water diversity of Central India*. National Bureau of Fish Genetic Resources (ICAR), Lucknow, India.

Mat Jaafar, T. N., Taylor, M. I., Mohd Nor, S. A., de Bruyn, M. and Carvalho, G. R. 2012. DNA Barcoding reveals cryptic diversity within commercially exploited Indo-Malay Carangidae (Teleosteii: Perciformes). *PloS One*, **7** (11):e49623.

Ng, H. H., and Hadiaty, R. K. 2011. *Clarias microspilus*, a new walking catfish (Teleostei: Clariidae) from northern Sumatra, Indonesia. *Journal of Threatened Taxa,* **3** (3): 1577-1584.

Puckridge, M., Andreakis, N., Appleyard, S. A. and Ward, R. D. 2012. Cryptic diversity in flathead fishes (Scorpaeniformes: Platycephalidae) across the Indo-West Pacific uncovered by DNA barcoding. *Molecular Ecology Resources*, **13** (1): 32-42.

Shrestha, J. 1978. Fish fauna of Nepal. *Journal of Natural History Museum Tribhuvan University,* **5** (1-4): 33-43.

Talwar, P. K. and Jhingran, A. G. 1991. Inland fishes of India and adjacent countries. *Oxford & IBH Publishing Co., New Delhi, Bombay, Calcutta*. 1-2.

Tamura, K., P. D., Peterson, N., Stecher, G., Nei, M. and Kumar, S. 2011. MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum Parsimony methods. *Molecular Biology and Evolution,* **10**: 2731-2739.

Ward, R. D., Zemlak, T. S., Innes, B. H., Last, P. R. and Hebert, P. D. 2005. DNA barcoding Australia's fish species. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **360**: 1847-1857.

J. Environ. & Sociobiol. : **10**(1) : 55-63, 2013

IMPLICATION OF NUCLEOTIDE SUBSTITUTION AT THIRD CODON POSITION OF THE DNA BARCODE SEQUENCES

Fazlur Rahman Talukdar, Sankar K. Ghosh*, Ruhina S. Laskar, Pradosh Mahadani, Mohua Chakraborty, Bishal Dhar and M. Joyraj Bhattacharjee

Department of Biotechnology, Assam University, Silchar-788011, Assam, India

ABSTRACT

DNA markers (barcode) differentiate species based on their nucleotide sequence diversity among various species. In this study we analyzed the rate and pattern of nucleotide substitution and their consequent influence on the amino acid substitution patterns of the sequences used as barcode mitochondrial COI, cyt *b* and the exon 1 of nuclear IRBP gene for animals from 15 different species of vertebrates. The analysis shows unlike other codon positions, nucleotide substitution at the third codon position does not show strong correlation with the amino acid substitution, for the three gene sequences. Furthermore, COI gene shows a very low percentage of amino acid variability (15.38%) inspite of high percentage of variation in its nucleotide sequence (40.76%) as well as a significantly ($p<0.0001$) low level of amino acid sequence divergence than the other gene fragments under study. Interestingly among the compared sequences, a significantly conserved amino acid substitution pattern seems to be a unique feature of barcode region of the COI gene making it a more efficient marker for species identification. Hence, it was concluded that the property of species identification of these sequences is based upon the variable nature of third codon position.

Key words : *Species-specific marker, DNA barcoding, nucleotide substitution, codon positions, amino acid substitution*

INTRODUCTION

Diversity of animal kingdom provides a challenging task for taxonomists to identify and characterize the millions of species thriving on earth. In addition to morphology-based identification systems, recent scientific investigations have revealed that DNA-based markers are promising and might provide more accurate and easy identification and characterization of species. DNA markers distinguish species based on sequence diversity in small segments of DNA among different species and the DNA used for this purpose can be both protein coding and protein non-coding fragments (Kress and Erickson, 2012).

* E-mail : drsankarghosh@gmail.com.

Hebert *et al.* in 2003 proposed a DNA barcoding system for animals, which was appreciated and widely accepted. Hebert's barcoding system is based upon sequence diversity in *cytochrome c oxidase subunit I* (COI, approximate gene length 1600 bp) of mitochondrial genome. The diversity is observed in the stretch of nucleotide sequence nearly 650-700 bp from the 5' end of this gene that is sufficient to reliably place species in their appropriate belonging groups. Barcoding based on mitochondrial COI gene has been implemented for a wide range of animals (Luo *et al.*, 2011).

Apart from the widely accepted mitochondrial COI barcode sequences, some other genes and gene fragments are also claimed to be suitable for species identification in animals. The nuclear gene, *interstitial retinol-binding protein 3* (IRBP) is commonly used in animals as a DNA phylogenetic marker (Barbosa *et al.*, 2012). The exon 1 of this gene has been used to infer the phylogeny of placental mammal orders (Madsen *et al*., 2001), major clades of rodentia (Blanga-Kanfi *et al*., 2009), primates (Poux *et al*., 2006) even for identification of carnivora species (Yonezawa *et al*., 2009). In another instance, Johns and Avise (1998) demonstrated that closely related species of vertebrates regularly show more than 2% divergence at another mitochondrial gene, the *cytochrome b* (cyt *b*). The cyt *b* gene is used for species identification by PCR-RFLP based (Latrofa *et al*., 2012) as well as sequence variability based identification (Chen *et al*., 2012). It is also used for inferring deep phylogenetic relationships (Lecompte *et al*., 2008). However, in a recent study it has been reported that COI, IRBP and cyt *b* are together used as DNA barcode to decipher the phylogeny based delimitation of species boundary of Rattini tribe (Pages *et al*., 2010).

These genes differentiate specifically each species based on the nucleotide sequence variation within the specified segment of the DNA. However, the polypeptide encoded by these regions of the gene fragments play functionally active roles in various metabolic pathways in every species. So, the three dimensional structure and functional integrity of the polypeptide must be maintained throughout the species in order to ensure normal metabolism of an organism, and there must exist a mechanism which minimizes variations at the amino acid level, despite of the high difference in the nucleotide sequence among species. This mechanism is speculated to be due to the degeneracy of codons.

In this study, we aimed to determine the rate and pattern of base substitution and their consequent influence on the amino acid substitution patterns of different genes and gene fragments used for species identification, as well as to compare the characteristics that make them reliable as identification marker. To achieve this objective, we have chosen three genes or gene fragments, *viz.*, mitochondrial cyt *b* gene, barcode region of the COI gene and the exon 1 of nuclear gene IRBP. Our aim was to analyse how the functional polypeptide encoded by these DNA sequences maintains their structural integrity inspite of having nucleotide variations enough for resolving species identification discrepancies. We assessed the substitution rate and pattern at nucleotide and amino acid level by considering different parameters, such as, percentage of variation, amount of four-fold degenerate codons, and sequence divergence of DNA and polypeptide across species that reflects the mechanism responsible for allowing species identification. Further, we also established the correlation between the effects of base substitution rate at each of the three codon positions with the subsequent amino acid substitution rate for each of the genes under study.

MATERIALS AND METHODS

The barcode region of mitochondrial COI gene, cyt *b* gene and the exon 1 of the nuclear IRBP gene sequences and their respective polypeptide sequences of 15 species under this study were taken from NCBI genome website (http://www.ncbi.nlm.nih.gov/genomes/organelles). The species under this study were selected from diverse classes of chordates, such as, pisces (fishes), amphibians, reptiles, aves (birds) and mammals living in different habitats (species names and Accession numbers are given in data accessibility section).

Multiple alignments of the selected sequences were performed using ClustralX, and the nucleotide and amino acid sequence divergence were calculated by MEGA version 4.1 and the percentage of nucleotide and amino acid variation as well as the number of four-fold degenerate sites were obtained by using MEGA version 4.1 (Tamura *et al.*, 2007). The nucleotide substitution per site, as well as the as the amino acid differences among all the compared species was calculated using *p*-distance parameters (number of base differences per site for all the sequences, by pair-wise analysis between all the species). Standard error was estimated by 500 bootstrap replicates. Numbers of synonymous and non-synonymous substitution rates were calculated using Nei-Gojobori Method in MEGA version 4.1.

Correlation coefficient was calculated between the nucleotide substitution rates at different codon positions (vertebrate mitochondrial genetic code table was used for COI and cyt *b* gene and standard genetic code table for IRBP gene) and the amino acid substitution rate per site for all the sequences.

RESULTS

Percentage of nucleotide and amino acid sequence variation

Comparison of nucleotide and amino acid sequence variation over the selected species revealed an exceptionally conserved pattern of amino acid variation in the barcode region of the COI gene. There were a total of 287 variable sites, which account for 40.76% of 704 nucleotides in the COI barcode region, whereas, this variation accounted for around 47.72% (609 out of 1276) in the IRBP gene and 50.56% (578 out of 1143) in the cyt *b* gene following multiple alignment. The resulting change in the amino acid composition is highly masked in the COI gene as it is evident from a amino acid variation percentage of only 15.38% (36 out of 234 amino acid residues) as compared to 41.64% in IRBP gene and 41.05% in cyt *b* gene (Fig. 1). The amino acid conservation rate of the barcode sequence is around 2.6 fold higher than the cyt *b* and IRBP genes, and as a rule, this must get reflected in the underlying mechanisms that allow this conservation pattern to occur.

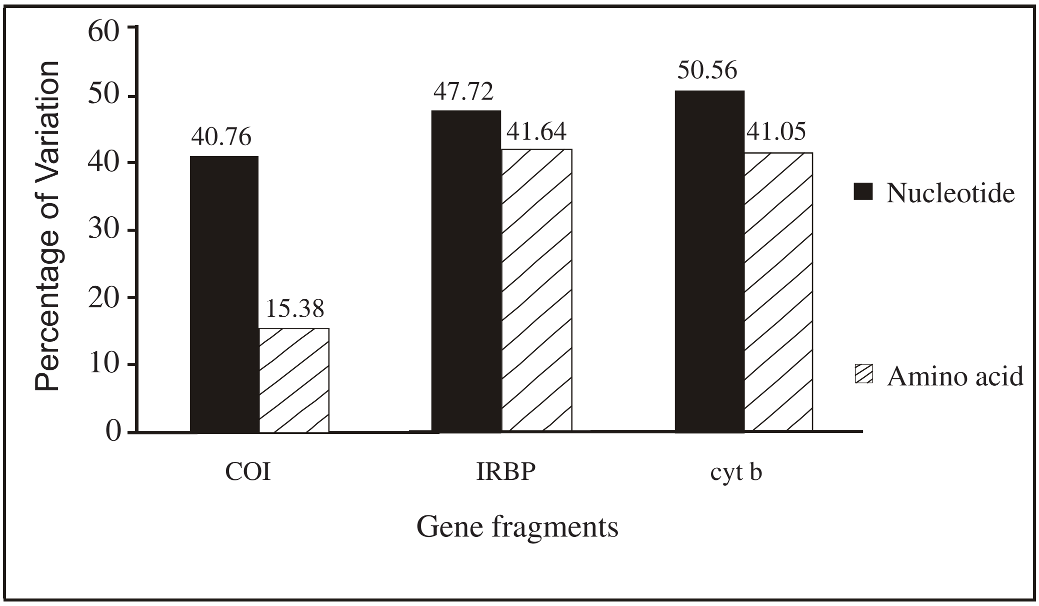

Fig. 1. Nucleotide and amino acid sequence variation: Percentage of sequence variation in the gene fragments and their corresponding amino acid sequences of COI, IRBP and cyt *b* genes

Analysis of evolutionary divergence per site

Estimation of evolutionary divergence between the compared species was conducted in terms of *P*-distance. It was found (Table 1) that, although there is no significant difference in the nucleotide substitution rates (*P*-distance mean 0.19±0.01, p<1.5) among the three gene sequences, the amino acid substitution rate was significantly lower (mean 0.15±0.01 in cyt *b* and IRBP versus 0.03±0.008 in barcode, p<0.0001) in the COI region as compared to the IRBP and cyt *b* gene.

Synonymous and non-synonymous substitution rate

The number of synonymous differences per synonymous site (d_S) and non-synonymous differences per non-synonymous site (d_N) averaging over all sequence pairs is calculated for barcode region of the COI gene, exon 1 of the rbp3 gene and the cyt *b* gene. Interestingly, the synonymous substitution rate was found to be highest in the barcode region of the COI gene (d_S= 0.68±0.02) as compared to IRBP (d_S=0.40±0.01) and cyt *b* (d_S= 0.61±0.01) gene sequences. Moreover, the non-synonymous substitution rate was significantly lower in the COI region (d_N=0.02±0.004, p<0.0001, t= 7.44), whereas it was found to be equal in the other two compared gene sequences (d_N= 0.08±0.007). As all the genes under study are functionally active, the synonymous substitution (d_S) is expected to be higher than the non-synonymous substitutions (d_N) and so the ù, defined as (d_N/d_S) should be lower than 1. The ratio of non-synonymous to synonymous substitution 'ù'

(d_N/d_S) was much lower in the COI region, around 6.5 and10 fold lower than cyt*b* and rbp3 gene sequence respectively (Table 1).

Table1. Comparison of different parameters among the gene sequences under study: evolutionary divergence (p-distance) of nucleotide and amino acid sequences; the rate of synonymous (d_S) and non-synonymous (d_N) substitution; the ratio of non-synonymous to synonymous substitution (ù=d_N/d_S) and the percentage of four fold degenerate codons of the gene fragments under study namely COI, cytb and IRBP

Gene fragments	**Nucleotide *p*-distance**	**Amino acid *p*-distance**	**Synonymous substitution/ site (d_S)**	**Non-synonymous substitution/ site (d_N)**	**ù=d_N/d_S**	**Percentage of four fold degenerate codons**
COI	0.19±0.01	0.03±0.01	0.68±0.02	0.02±0.004	0.02	43.61
Cyt*b*	0.21±0.01	0.15±0.01	0.61±0.01	0.08±0.007	0.13	32.28
IRBP	0.17±0.01	0.15±0.01	0.40±0.01	0.08±0.007	0.20	41.17

Analysis of codon degeneracy

During the process of protein/polypeptide synthesis, most of amino acids are usually encoded by more than one codon, commonly known as degeneracy of codons, which consist of two degrees, two fold, four fold and six fold degenerate codons (2 codons, 4 codons and 6 codons encoding same amino acid). From the analysis of average amino acid composition among the compared species it was revealed that, the percentage of four fold degenerate amino acids, such as, Ala, Pro, Gly, Thr is quite high in the barcode region of the COI gene. This result is also evident by the observation that the percentage of four-fold degenerate codons is comparatively higher (43.16%) in the COI gene than the other two gene sequences IRBP and cyt *b* (41.17% and 32.28% respectively) as shown in Table 1.

Correlation between the nucleotide base substitution rates at different codon positions and amino acid substitution rates

Correlation co-efficient between substitution rates of nucleotide bases at different codon positions and the corresponding substitution rate of the corresponding amino acids was calculated. It was found that there is no significant ($p = 1.42$) correlation between the base substitution at third codon position and amino acid substitution in the barcode region of the COI gene. Whereas, other gene fragment under this study showed significant ($p< 0.00001$) but low positive correlation for the third codon position (IRBP: r_3= 0.93; cyt *b*: r_3= 0.79). However, the base substitution at the first and second positions of the COI gene showed a strong positive correlation (COI:r_1= 0.85, r_2=0.90, $p< 0.00001$) with amino acid substitution, as the other two sequences (IRBP: r_1= 0.98, r_2= 0.97 $p< 0.00001$; cyt *b*: r_1= 0.96, r_2= 0.95 $p= 0.00001$, Fig. 2).

Thus, it can be concluded that, substitution at second codon position showed highest correlation with the amino acid substitution followed by the first codon position and least

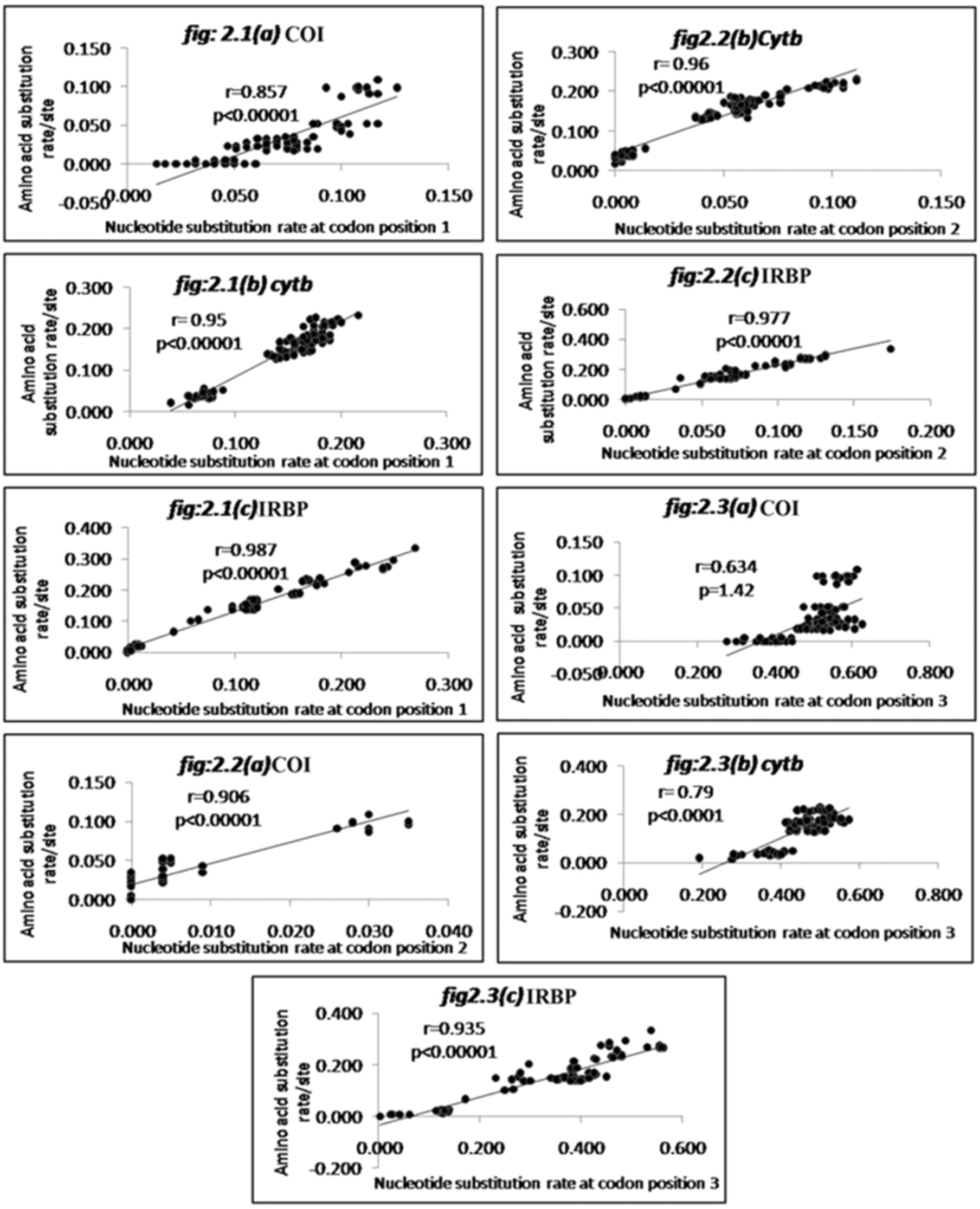

Fig. 2.1. Correlation between amino acid substitution rate and nucleotide substitution rate at codon position 1 for the gene fragments under study. (a) **COI:** Correlation coefficient (r)=0.857, p<0.0001, (b) **cyt *b:*** r=0.95, p<0.00001, (c) **IRBP:** r=0.987, p<0.00001.

Fig. 2.2. Correlation between amino acid substitution rate and nucleotide substitution rate at codon position 2 for the gene fragments under study. (a) **COI:** Correlation coefficient (r)=0.90, p<0.0001, (b) **cyt *b:*** r=0.96, p<0.00001, (c) **IRBP:** r=0.97, p<0.00001.

Fig. 2.3. Correlation between amino acid substitution rate and nucleotide substitution rate at codon position 3 for the gene fragments under study. (a) **COI:** Correlation coefficient (r)=0.63, p<1.42, (b) **cyt *b:*** r=0.79, p<0.0001, (c) **IRBP:** r=0.93, p<0.00001.

in case of third codon position. This observation was true for all the three gene sequences. Another interesting feature to be noticed is that the substitution rate at the second codon position is very less as compared to the first codon position and on the other hand, the third codon position showed the highest substitution rate.

DISCUSSION

From the current study, it was revealed that the barcode region of the COI gene shows a higher degree of conservation in the amino acid substitution pattern than the other gene sequences used for barcoding, *viz.*, IRBP and cyt *b*. The barcode region of COI has a lower amino acid sequence variation (15.38%) instead of almost 40.76% variation in the nucleotide sequence, which implies a minimum effect of nucleotide variation on the amino acid sequence. On the other hand, IRBP and cyt *b* showed 47.72% and 50.56% nucleotide sequence variation and 41.64% and 41.05% amino acid sequence variation respectively, indicating that the variation of nucleotide sequence is reflected by the consequent variation in amino acid sequence.

The estimation of evolutionary divergence (*p*-distance) based upon the nucleotide and amino acid sequence showed an almost equal pattern of divergence at both the levels for IRBP and cyt *b* genes. Surprisingly, in case of COI gene the sequence divergence at the nucleotide level was not reflected at the amino acid level, with a significantly greater extent ($p<0.0001$) of conservation in the amino acid sequence encoded by the gene fragment. This may be due to the higher synonymous substitution in the barcode region of the COI gene ($d_S = 0.68 \pm 0.02$) as compared to IRBP ($d_S = 0.40 \pm 0.01$) and cyt *b* ($d_S = 0.61 \pm 0.01$), as well as significantly low non-synonymous substitution rate in the COI gene ($d_N = 0.02 \pm 0.004$, $p < 0.0001$). As the gene fragments under this study are functionally active, the synonymous substitution (d_S) is expected to be higher than that of non-synonymous substitutions (d_N) in order to maintain functional constraint. This hypothesis was proved to be correct as the ù, defined as (d_N/d_S) was found much lower than 1, a value true for functionally active genes. Moreover, in the COI gene the ratio of non-synonymous to synonymous substitution 'ù' (d_N/d_S) was found to be approximately 6.5 and 10 fold lower than for cyt *b* and IRBP gene sequence respectively. This indicates a very high extent of synonymy in the COI gene. In addition, the four fold degenerate codons are more frequent in the barcode region of the COI gene than in cyt *b* or IRBP gene.

Substitution at second codon position usually leads to a significant change in amino acid sequence as this codon position is highly correlated with amino acid substitution. However, the second codon position exhibits very high stability among all the compared species and in all the three gene sequences under this study (IRBP: $r_2 = 0.97$ $p < 0.00001$; cyt *b*: $r_2 = 0.95$ $p< 0.00001$; and barcode of COI: 0.90 $p < 0.00001$).

The correlation co-efficient between substitution rates of bases at different codon positions and corresponding amino acids revealed the fact that the variation at third codon position largely forms the basis for using any coding DNA sequence as a species-specific identification marker, because the variation at this position has the least influence on the amino acid substitution. This has been earlier assumed true for barcode sequence of the COI gene and was reported by Hebert *et al.* (2004) and Ward and Holmes (2007). Our study also showed similar results for all the three sequences under this study. However, in case of barcode region of COI, base substitution at the third codon position was not significantly correlated with amino acid substitution. This observation indicates that the greater level of liberty for change provided to the nucleotide at third codon position of the barcode sequence of the COI gene helps in the maintenance of a functionally active polypeptide sequence encoded by this region. It also may confer an advantage to use the barcode sequence of COI gene for discriminating species more efficiently than any other sequence used as species identification marker.

All these findings point towards an interesting phenomenon of flexibility in base substitution at third codon position that allows nucleotide sequence variation among species with little or no influence on the functional polypeptide encoded by the sequence. Kimura in 1983 has already reported the high rate of substitution of the third codon position followed by the first and second codon position (Kimura, 1980), but the mechanism behind this is yet to be resolved. It is important to shed light on some of these unanswered questions, such as, what allows the high mutation rate or fast evolution rate in third codon position of a gene? How does the second codon position of genes remains almost conserved in all the species? What mechanism works behind these site-specific mutations? When these mysteries will be resolved, a far deeper understanding of the phenomena of evolution and speciation are likely to be achieved.

REFERENCES

Barbosa, S., Pauperio, J., Searle, J. B. and Alves, P. C. 2012. Genetic identification of Iberian rodent species using both mitochondrial and nuclear loci: application to noninvasive sampling. *Molecular Ecology Resources*., **13**(1) : 43-46.

Blanga-Kanfi, S., Miranda, H., Penn, O., Pupko, T., De Bry, R. and Huchson, D. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. *BMC Evolutionary Biology,* **9**: 71.

Chen, R., Jiang, L. Y. and Qiao, G. X. 2012. The effectiveness of three regions in mitochondrial genome for aphid DNA barcoding: a case in lachininae. *PloS One,* **7**(10): e46190.

Hebert, P. D., Cywinska, A., Ball S. L. and deWaard J. R. 2003. Biological identifications through DNA barcodes. *Proceedings. Biological Sciences / The Royal Society,* **270**(1512): 313-321.

Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S. and Francis, C. M. 2004. Identification of birds through DNA Barcodes. *PLoS Biology,* **2**(10): e312.

Johns, G. C. and Avise, J. C. 1998. A comparative summary of genetic distances in the vertebrates from the mitochondrial cytochrome b gene. *Molecular Biology and Evolution,* **15**(11)**:** 1481-1490.

Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution,* **16**(2): 111-120.

Kress. W. J. and Erickson D. L. 2012. DNA barcodes: methods and protocols. *Methods in Molecular Biology,* **858** : 3-8.

Latrofa, M. S., Annoscia, G., Dantas-Torres, F., Traversa, D. and Otranto, D. 2012. Towards a rapid molecular identification of the common phlebotomine sand flies in the Mediterranean region. *Veterinary Parasitology,* **184**(2-4) : 267-270.

Lecompte, E., Aplin, K., Denys, C., Catzeflis, F., Chades, M. and Chevert, P. 2008. Phylogeny and biogeography of African Murinae based on mitochondrial and nuclear gene sequences, with a new tribal classification of the subfamily. *BMC Evolutionary Biology,* **8** : 199.

Luo, A., Ai Bing, Z., Simon, YWHO., Weifun Xu, Yanzhou, Z., Weifang Shi, Stephen., L. Cameron and Chaodong, Zhu, 2011. Potential efficacy of mitochondrial genes for animal DNA barcoding: a case study using eutherian mammals. *BMC Genomics,* **12**: 84.

Madsen, Ole, Mark Scally, Christophe J. Douady, Diana, J. Kao, Ronald W. DeBry, Ronald Adkins, Heather M. Amrine, Michael J. Stanhope, Wilfried W. de Jong and Mark S. Springer., 2001. Parallel adaptive radiations in two major clades of placental mammals. *Nature,* **409**(6820) : 610-614.

Pages, M., Yannick, C., Vincent, H., Surachit, W., Jean-Francois, C., Jean-Pierre, H., Serge, M. and Johan, M., 2010. Revisiting the taxonomy of the Rattini tribe: a phylogeny-based delimitation of species boundaries. *BMC Evolutionary Biology,* **10 :** 184.

Poux, C., Chevret, P., Huchon, D., de Jong, W. W. and Douzery, E. J. 2006. Arrival and diversification of caviomorph rodents and platyrrhine primates in South America. *Systematic Biology,* **55**(2): 228-244.

Tamura, K., Dudley, J., Nei, M. and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Molecular Biology and Evolution,* **24**(8): 1596-1599.

Ward, R. D. and Holmes, B. H. 2007. An analysis of nucleotide and amino acid variability in the barcode region of cytochrome coxidase I (COI) in fishes. *Molecular Ecology Notes,* **7**(6): 899-907.

Yonezawa, T., Kohno, N. and Hasegawa, M. 2009. The monophyletic origin of sea lions and fur seals (Carnivora; Otariidae) in the Southern hemisphere. *Gene,* **441**(1-2): 89-99.

Biodiversity Conservation
Volume 8 (No. 1) March, 2011

DNA Barcoding: Digital Taxonomy of Bioresources

Sankar Kumar Ghosh*

Maloyjo Joyraj Bhattacharjee, Ksh. Miranda Devi, Monika Ahanthem, Shantanu Kundu, Pradosh Mahadani, Bishal Dhar, Khomdram Bijoya Devi, Mohua Chakraborty, Fazlur Rahman, Rosy Mondal, Jagdish Hansa, Ruhina S Laskar, Tarikul Huda Mazumder, Priyanka Sarkar, Shiba Rajbongshi, Arijit Chakraborty, Mridul M Das, Probal Ranjan Ghosh, Kulendra Chandra Das & Boni Amin Laskar

Department of Biotechnology, Assam University, Silchar-788011
Email: drsankarghosh@gmail.com
*Supervisor and presenting author

Biodiversity assessments and implementation of conservation actions worldwide are hampered by slow progress in taxonomic research, termed the Taxonomic Impediment[1]. The existing workforce of taxonomists cannot cope with the overwhelming need for basic field surveys, species descriptions and systematic revisions to provide basic information for conservation planning. In addition, few taxonomists are able to distinguish critically between more than 1000 taxa[2]. The reality facing taxonomy is that there may be more species that remain to be discovered on Earth than those that have already been described[3]. It is estimated that 1.4-1.8 million species have been described[4-6] out of a possible total of approximately 7-15 million6. Species and populations are going extinct at an alarming but poorly understood rate[7,8]. Many species may be going to extinct before they can be identified or described. This presents a problem for conservation planning and prioritization, obviously because species that have not been identified cannot be protected effectively and is evident throughout the world. Taxonomic expertise is lacking even for major and commercially important groups. The few taxonomists who are working in developing countries, home to more than 95% of globally described species; find it difficult to access basic taxonomic information such as species descriptions[9]. Where taxonomic keys are available, they are rarely revised and often inadequate to identify specimens unambiguously to the species level[10]. There have been strident calls on the taxonomy community to embrace new technologies and to form networks to speed up the description of biodiversity[11,12] and to improve our ability to identify species[2, 13-16]. How this could be achieved without compromising rigorous taxonomic research principles is, however, questioned[17]. There is legitimate concern that too little money is being spent on morphological taxonomy compared to molecular studies[18]. A purely DNA taxonomy approach (e.g.Blaxter[19]), not to be confused with DNA barcoding[21] is too simplistic in our view. DNA taxonomy can, however, provide additional characters for discrimination, especially in cases where other characters vary among species and are thus difficult to interpret. Therefore we do not agree with de Carvalho *et at.*[17] that a molecular approach will do little to address the real problems in taxonomy. In our experience, most traditional taxonomists welcome the opportunity to refer to molecular data, but DNA taxonomy should not replace taxonomic research that can be based on multiple character data sets. Nucleotide sequence divergence in short, standardized gene regions as DNA barcodes can be used to identify known species and facilitate the discovery of new ones[13,21]. Mitochondrial DNA is a valuable marker in population genetic or phylogeographic studies because it is maternally inherited, evolves rapidly and recombination is rare or absent[22]. Therefore, a part of the mitochondrial cytochrome oxidase subunit I (COI) has been chosen as a standard gene region for barcoding animals. Although there was no prior reason for choosing COI among the 13 protein-coding mitochondrial genes[13] it has the advantage of having robust universal primers that can recover the 5' end of COI of most animal species. Barcoding of species has become cheaper through the technological advances made by other molecular programmes, especially the human genome project[23]. Rapid barcoding and comparison with the growing database of COI sequences[24] will increase the speed of identification of newly collected or unknown specimens. This may focus taxonomists' attention on unidentified lineages, and with the addition of morphological and other taxonomically relevant data, could lead to a faster rate of species description. Kress et al. (2005)[25] suggest that the use of the COI sequence is not appropriate for most species of plants because of a much slower rate of cytochrome c oxidase I gene evolution in higher plants than in animals. A series of experiments was then conducted to find a more suitable region of the genome for use in the DNA barcoding of flowering plants (or the larger group of land plants)[26]. One 2005 proposal was the nuclear internal transcribed spacer region and the plastid trnH-psbA intergenic spacer[25] other researchers advocated other regions such as matK[26]. In 2009, a collaboration of a large group of plant DNA barcode researchers proposed two chloroplast genes, rbcL and matK, taken together, as a barcode for plants[4]. Jesse Ausubel, a DNA barcode researcher not involved in that effort, suggested that standardizing on a sequence was the best way to produce a large database of plant sequences, and that time would tell whether this choice would be sufficiently good at distinguishing different plant species[25]. The Consortium for the Barcode of Life (CBOL) was established as a growing coalition of biodiversity organizations interested in developing barcoding as a global standard for DNA-based species identification. CBOL promotes FISH-BOL as an international campaign to barcode all marine and freshwater fish. FISH-BOL consists of ten regional working groups representing Africa, Australia, Oceania/Antarctica, the Americas (North, Central and South America), Europe and Asia (India, North East Asia, and South East Asia). Each of these regional working groups is charged with the task of organizing support and participation to barcode the ichthyofauna within their region, based on expert-identified voucher specimens. The main purpose of barcoding is to identify species reliably, increase the rate of species discovery and raise the profile of taxonomic research[25].

Methods:

Sample collection:

Effective DNA barcoding depends on the quality of the biological material. Simple sampling protocol will ensure proper preservation of biological samples for DNA studies.

For mammals, fish, birds and large invertebrates

> Freeze whole individual or part of the tissue specimens in plastic bags or a cryo-vial; use a write-on label to record vessel/expedition name/code, locality or station number, latitude and longitude, date, species name and collector's name. Store labelled specimens in freezer. Photograph the whole specimen before discarding, and cross reference the digital photo to the tissue sample code. It is essential that species-diagnostic characters can be seen on the photograph. Avoid formalin work areas for handling specimens.

For small fish and invertebrates

> Use 96% pure ethanol (~80-85% for fragile arthropods) for fixation and preservation. Do not use denaturised alcohol. Label all samples with locality, coordinates, date and collector. Record specimen collection data on a waxy paper label, use pencil. Add label to ethanol filled jar. Record the vessel/expedition name/code, locality or station number, date, species name, and name of scientist making the identification.

If ethanol is impractical or unavailable in large amounts, the samples (or specimens) can be sub-sampled in ethanol (as above). Cross reference labelling with unique identifiers is important to link subsamples or tissue samples with primary samples. Specimens that must be kept dry for morphological studies (such as butterflies) should be kept frozen (at -20°C or lower temperatures) or quickly dried in an oven or incubator.

Specimen Preservation: Following DNA extraction the samples are preserved in 10% formalin to be used as voucher specimen in future.

PCR amplification& Genome analysis:

It can be carried out using standard protocol.PCR fragments are then sequenced in an automated DNA sequencer. Nucleic acid sequences were subjected to BLASTn searches at the National Center for Biotechnology Information (NCBI), (http://www.ncbi.nlm.nih.gov/blast) and they were aligned using ClustalW software (http://www.ebi.ac.uk/clustalw). The final analyzed sequences will be deposited in GenBank, under the accession numbers. Kimura 2-parameter genetic distances will then be determined using the computer program Molecular Evolutionary Genetics Analysis (MEGA Version 4.1) to get the level of within and among species variation. It will further help to trace the phylogenetic relationship among groups.

Application:

The barcode of life provides an additional master key to knowledge about a species. Compiling a public library of sequences linked to named specimens, plus faster and cheaper sequencing, will make this new barcode key increasingly practical and useful. The additional powers that barcode offers includes:

1. Works with fragments. Barcoding can identify a species from bits and pieces. When established, barcoding will quickly identify undesirable animal or plant material in processed foodstuffs and detect commercial products derived from regulated species. Barcoding will help reconstruct food cycles by identifying fragments in stomachs and assist plant science by identifying roots sampled from soil layers.
2. Works for all stages of life. Barcoding can identify a species in its many forms, from eggs and seed, through larvae and seedlings, to adults and flowers.
3. Unmasks look-alikes. Barcoding can distinguish among species that look alike, uncovering dangerous organisms masquerading as harmless ones and enabling a more accurate view of biodiversity.
4. Reduces ambiguity. Written as a sequence of four discrete nucleotides - CATG - along a uniform locality on genomes, a barcode of life provides a digital identifying feature, supplementing the more analog gradations of words, shapes and colors. A library of digital barcodes will provide an unambiguous reference that will facilitate identifying species invading and retreating across the globe and through centuries.
5. Makes expertise go further. The bewildering diversity of about 2 million species already known confines even an expert to morphological identification of only a small part of the plant and animal kingdoms. Foreseeing millions more species to go, scientists can equip themselves with barcoding to speed identification of known organisms and facilitate rapid recognition of new species.
6. Democratizes access. A standardized library of barcodes will empower many more people to call by name the species around them. It will make possible identification of species whether abundant or rare, native or invasive, engendering appreciation of biodiversity locally & globally.
7. Opens the way for an electronic handheld field guide, the Life Barcoder. Barcoding links biological identification to advancing frontiers in DNA sequencing, miniaturization in electronics, and computerized information storage. Integrating those links will lead to portable desktop devices and ultimately to hand-held barcoders. Imagine the promise of a schoolchild with a barcoder in hand learning to read wild biodiversity, the power granted to a field ecologist surveying with a barcoder and global positioning system, or the security imparted by a port inspector with a barcoder linked to a central computer!
8. Sprouts new leaves on the tree of life. Since Darwin, biologists seeking a natural system of classification have drawn genealogical trees to represent évolutionary history. Barcoding the similarities and differences among the nearly 2 million species already named will provide a wealth of genetic detail, helping to draw the tree of life on Earth. Barcoding newly discovered species will help show where they belong among known species, sprouting new leaves on the tree of life.
9. Demonstrates value of collections. Compiling the library of barcodes begins with the multimillions of specimens in museums, herbaria, zoos and gardens, and other biological repositories. The spotlight that barcoding shines on these institutions and their collections will strengthen their ongoing efforts to preserve Earth's biodiversity.
10. Speeds writing the encyclopaedia of life. Compiling a library of barcodes linked to voucher specimens and their binomial names will enhance public access to biological knowledge, helping to create an on-line encyclopaedia of life on Earth, with a web page for every species of plant and animal.

And Many More.....

Global initiative: Consortium for the barcode of life (CBOL)

The Consortium for the Barcode of Life (CBOL) is an international initiative devoted to developing DNA barcoding as a global standard for the identification of biological species. Established in 2004 through support from the Alfred P. Sloan Foundation, CBOL promotes barcoding through Working Groups, networks, workshops, conferences, outreach, and training. CBOL has 200 Member Organizations from 50 countries and operates from a Secretariat Office located in the Smithsonian Institution's National Museum of Natural History in Washington, DC.

Major CBOL projects:

ABBI, the All Birds Barcoding Initiative, is a campaign to collect DNA barcodes from 5 or more individuals of all of the approximately 10,000 bird species in the world. The ABBI DNA barcode library will help speed discovery of new species, open new avenues for scientific investigation, and provide a forensic tool for identifying specimens, including for example tissue fragments from bird-airplane collisions and avian blood samples from biting insects that harbour West Nile virus or other human disease agents.

Bee-BOL, the Bee Barcode of Life Initiative, is a global effort to coordinate the assembly of a standardized reference sequence library for all ~20,000 bee species. Bee-BOL is creating a valuable public resource in the form of an electronic database containing DNA barcodes, images, and geospatial coordinates of examined specimens. The database contains linkages to voucher specimens, information on species distributions, nomenclature, authoritative taxonomic information, collateral natural history information and literature citations.

ECBOL is an information and coordination hub on DNA barcoding in Europe organized within EDIT, the European Institute of Taxonomy and maintained by CBS, the Centraalbureau voor Schimmelcultures in Utrecth. The ECBOL initiative (Calibrating European Biodiversity using DNA Barcodes) is a network of European researchers and is seeking to obtain funding from the coordination and maintenance of a Network of European Leading Labs.

FISH-BOL, the Fish Barcode of Life campaign, is collecting barcodes from at least five specimens representing the 30,000+ species of marine, freshwater and estuarine fish of the world. Like ABBI, FISH-BOL has a central Steering Committee and Regional Working Groups.

MarBOL is an international campaign to obtain at least 50,000 barcode records of marine species by October 2010. MarBOL is led by an international Steering Committee and an affiliated project of the Census of Marine Life (CoML).

MBI, the Mosquito Barcode Initiative is another "demonstration project" aimed at producing a global operational system for identifying mosquitoes in two years. MBI plans to barcode at least five specimens from 80% of the 3200 known mosquito species. Disease-bearing species and their closest relatives will be the highest priority.

TBI, the Tephritid Barcode Initiative is a two-year "demonstration project" that will create an operational system for identifying fruit flies around the world. TBI will barcode at least five representatives of all tephritid fruit flies that are either (1) agricultural pests, (2) beneficial species used for biological control of other pests, (3) closely related to pests or beneficial species; and (4) representative species from other families of tephritids. TBI plans to obtain barcodes from approximately 2000 species of the estimated 4500 known tephritid species.

CBOL's International Network for Barcoding Invasive and Pest Species (INBIPS) is made up of researchers and representatives of government agencies. They're exchanging information and developing plans for global barcoding projects.

CBOL's Plant Working Group (PlantWG) focuses on the urgent need to reach a consensus on the standard barcoding regions for plants.

There are many other major barcoding initiatives organized by other barcoding groups which include International barcode of life initiative (iBOL) and Quarantine barcode of life (QBOL).

Our Initiative:

India being a hotspot of diversity borne a diverse variety of flora and fauna however, the limitations inherent to morpho-taxonomy latent's the actual status. The advent of DNA barcoding represents an important step to monitor Northeast India biodiversity. In our approach, more than 350 mitochondrial COI DNA barcode of diverse flora and fauna have been submitted to Barcode gene bank of India, particularly in Northeast which includes Fishes, Turtles and Tortoises, Birds, Economic and Endemic animals like Manipuri pony, Indian Rhinoceros, Golden Langur etc. and Medicinal plants. The analysis of the sequences showed a congeneric and conspecific barcode gap and species label identification was relatively straightforward. The cases of latent species, introgression and haplotypes diversity was observed among certain species, which earlier either revealed an ambiguous taxonomic state or misperception. The study further came with the possible discovery of new species under certain taxa. It provided an independent means to reassess the existing phylogeny which also showed the misplacement of certain organisms under certain taxa. The initiative is highlighted as poster presentation at the third international conference on DNA barcoding held at Mexico City 28.

Fish barcoding initiative: The Northeast India is known for its large repository of freshwater fishes which is 33% of the total freshwater fishes found in India. In our initiative we concentrated on Catfish and Ornamental fish barcoding as both the groups hold traded value and had a major demand in international market and needs molecular label authentication to avoid adulteration and species replacement.

Catfish barcoding: A total of 82 barcode sequences were developed for 27 species under 17 genera and 10 families. The analysis and comparison match with database revealed straightforward identification for only 20% of the species and rest returned no match. Hence, most of the sequences were noble sequences and an addition to the existing reference database. The analysis further showed a positive role of Transversion mutation in species differentiation and extraction of mini and character based barcode for the species which will help in technical overhaul of barcode process29.

Ornamental fish barcoding: The northeast India has is a hub for the trade of ornamental fishes to the global market however; no strict law is there to control or regulate the trade. The major concern being increased demand had lead to over exploitation of many fishes which are now either critically endangered or endangered. The generation of barcode sequences will help to regulate the trade and hence, a total of 80 barcode sequences were developed for 45 endemic species of fishes. The result of the above project is highlighted in the 22nd Pacific Science Congress at Kuala Lumpur, Malaysia30.

Turtle and Tortoise DNA barcoding: The critically endangered Testudines species widely distributed in Asia are on the verge of extinction by anthropogenic and environmental threats. Phylogeny of turtle and tortoise are perplexing when morphological traits were evaluated with molecular data's. Our data generated barcode sequences includes 5 softshell turtles, 6 hardshell turtles and 2 tortoise species exclusively found in Northeast India. The study is a novel approach to generate species specific DNA barcode tags of endangered Testudines and authenticate the existing phylogenic construction with a strong support for proper systematic, aiming conservation of turtle biodiversity.

Medicinal Plant DNA barcoding: More than 35 Accession No. of plant barcode sequences were generated from 20 different medicinal plants of Northeast India. 15 sequences are novel (first time in NCBI) sequence in Genbank database. The sequences showed indels could be a possible marker for several species of plants.

Other Vertebrates DNA barcoding from India: Many other endemic vertebrates like Indian rhinoceroses, Golden Langur, Manipuri Pony, Himalayan Siri breed of cow, birds etc. which had their last remaining strong holds in Northeast India were also barcoded for their partial mitochondrial COI gene. The sequences generated had direct application in many fields discussed below.

- Poaching: The population of many endemic and economic animals like Golden Langur, Indian Rhinoceros, Turtles, Birds, Fishes etc. of India is going towards a bottleneck due to heavy exploitation. The animals were mostly targeted for their ivory, skin, meat etc. however, the regulation to monitor these happenings was inadequate because of the identification challenges of animals source. As barcoding is a DNA based technique and DNA could be extracted from any biological source, it will greatly help to monitor the poaching of animal's products and meat etc.
- Food adulteration: Food adulteration is a global concern now. Our initiative in this respect involves a PCR-RFLP based marker to monitor and check meat and milk adulteration. The marker is well capable of discriminating the source of the milk and

meat from a mixture and thus has a major application to check adulteration.

- Phylogeograpic studies: A major application of this study also involves the phylogeographic studies. The sequences of the mitochondrial D-loop part acts as breed specific marker and on monitoring the migration range of the animals based on assessing the number of total and isolated haplotypes within a population. The above finding reflects a diversity range of a population which depicts the ecological status of the organism.
- New species: Based on DNA barcoding we proposed new species in Turtle, Mahseer fish, catfishes and ornamental fishes from Northeast India.
- Cross-species primers for Barcode development

Development of individual barcode for human- Proposed powerful marker for the UID (Unique Identity) Project in India

The 12-digit unique number (**UID card-AADHAAR**) will store basic demographics and biometric information - photograph, ten fingerprints and iris - of each individual in a central database and will issue for all residents in India. Established in February 2009, and will own and operate the Unique Identification Number database containing biometrics data. The UID will link a person's Passport Number, Driving License, PAN card, Bank Accounts, Address, Voter ID, etc and all this information will be checked through a database. So, for example, if someone has different addresses on PAN and driving license, is liable to get caught. Those who will opt out of this program will have much inconvenience in doing business, operating bank accounts and other offices which will require a UID. The UID (Unique Identity) project is undertaken by the Government of India with the agenda of implementing the envisioned Unique Identification card to about all the population of India. The present programme undertaken is indeed novel in its approach but with some limitations. The identification of a mutilated body either by disaster such as Mumbai blast or a terrible accident such as Gyaneshwari Express train accident etc., is mainly based on some specific assumption and several disputes of claim and disclaim had made the situation worsen. To address it, in addition to the biometric data, the inclusion of DNA based identification marker in the panel might prove to be very useful. Here, we propose the use of the mitochondrial DNA (mtDNA) D-Loop hypervariable regions, owing to its high mutation rate, for unique individual identification as DNA barcode (Acc. No. JN603607-29). We sequenced and analyzed 25 human D-Loop hypervariable regions and observed significant variation ($p = 0.04$) in terms of total nucleotide differences among the sequences. Moreover, our analysis revealed the presence of at least one or more indels in each of the sequences, which makes them unique from each other. **So, d-loop sequences are unique sequences among human: a possible marker for personalized identification of human.**

Future perspective

These projects are designed to identify species through a composite database based on the mitochondrial COI region of as many species as possible, to provide a hitherto unavailable 'horizontal genomics' perspective[31].This COI database can form the foundation for more detailed phylogeographic studies. In their planning phases, preliminary analyses of phylogenetic studies will be able to benefit from this species-rich COI database in important ways. One such benefit will be in selecting in- or outgroups suitable for molecular phylogenetic analyses in a particular study.The identification of species remains, however, the main objective of barcoding. Owing to maternal inheritance, mtDNA reflects only a portion of the ancestry of a species. Additional information is necessary to test whether divergent mtDNA lineages constitute separate species[32-34].We therefore view provisional species assignments (that is, not formal descriptions) based only on mtDNA as working hypotheses. As new collections are made, individuals suspected of being unique (especially of equivocal taxonomic status) can be barcoded. If they are divergent or form part of a unique lineage, they can then be flagged for further taxonomic investigation. As PCR and sequencing technology advances- such as nanolitre-scale sequencing technology developments[35]-the evolution of a hand-held barcoder becomes more feasible. This will markedly increase our capacity to identify species and understand the distribution of indigenous and alien species and their unique genetic lineages, especially in those regions where there are few trained taxonomists and sequencing facilities. It may also spark more interest in biodiversity, taxonomy and conservation issues if the general public have easier access to species information through rapid species identification[2, 36].

Reference:

1. Hoagland K.E. (1996). The Taxonomic Impediment and the Convention on Biodiversity. ASC News 24, 61-62, 66-67.
2. Costa F.O. and Carvalho G.R. (2007). The Barcode of Life Initiative: synopsis and prospective societal impacts of DNA barcoding of fish. Genomics Soc. Policy 3, 29-40.
3. May R.M. and Beverton R.J.H. (1990). How many species? Philos. Trans.: Biol. Sci. 330, 293-304.
4. Stork N.E. (1988). Insect diversity: facts, fiction and speculation. Biol. J. Linn. Soc. 35, 321-337.
5. Southwood T.R.E. (1978). The components of diversity. In Diversity of Insect Faunas, eds L.A. Mound and N. Waloff, pp. 19-40. Blackwell, Oxford.
6. Mace G.M. (2004). The role of taxonomy in species conservation. Phil. Trans. R. Soc. Lond. B 359, 711-719.
7. Thomas C.D., Cameron A., Green R.E., Bakkenes M., Beaumont L.J. et al. (2004). Extinction risk from climate change. Nature 427, 145-148.
8. Hughes J.B., Daily G.C. and Ehrlich P.R. (1997). Population diversity: its extent and extinction. Science 278, 689-692.
9. Agosti D. (2006). Biodiversity data are out of local taxonomists' reach. Nature 439, 392-392.
10. Balakrishnan R. (2005). Species concepts, species boundaries and species identification: a view from the Tropics. Syst. Biol. 54, 689-693.
11. Wheeler Q.D., Raven P.H. and Wilson E.O. (2004). Taxonomy: impediment o expedient? Science 303, 285.
12. Polaszek A. (2005). A universal register for animal names. Nature 437, 477.
13. Hebert P.D.N., Cywinska A., Ball S.L. and De Waard J.R. (2003). Biological identifications through DNA barcodes. Proc. R. Soc. Lond. B 270, 313 - 321.
14. Schander C. and Willassen E. (2005). What can biological barcoding do for marine biology? Mar. Biol. Res. 1, 79 - 83.
15. Schindel D.E. and Miller S.E. (2005). DNA barcoding a useful tool for taxonomists. Nature 435, 14-17.
16. Miller S.E. (2007). DNA barcoding and the renaissance of taxonomy. Proc. Natl Acad. Sci. USA 104, 4775-4776.
17. de Carvalho M.R., Bockmann F.A., Amorim D.S., deVivo M., de Toledo-Piza M. et al. (2005). Revisiting the taxonomic impediment. Science 307, 353.
18. Ebach M.C. and Holdrege C. (2005). DNA barcoding is no substitute for taxonomy. Nature 434, 697-697.
19. Blaxter M.L. (2004). The promise of a DNA taxonomy. Phil. Trans. R. Soc. B 359, 669-679.
20. DeSalle R. (2007). Phenetic and DNA taxonomy; a comment on Waugh. Bioessays 29, 1289-1290.
21. Gómez A., Wright P.J., Lunt D.H., Cancino J.M., Carvalho G.R. et al. (2007). Mating trials validate the use ofDNAbarcoding to reveal cryptic speciation of a marine bryozoan taxon. Proc. R. Soc. Lond. B 274, 199-207.
22. Moritz C., Dowling T.E. and BrownW.M. (1987). Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annu. Rev. Ecol. Syst. 18, 269-292.
23. Collins F.S., Morgan M. and Patrinos A. (2003). The Human Genome Project: lessons from large-scale biology. Science 300, 286-290.
24. Ratnasingham S. and Hebert P.D.N. (2007). BOLD: the Barcode of Life data system (http://www.barcodinglife.org). Mol. Ecol. Notes 7, 355-364.
25. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (June 2005). "Use of DNA barcodes to identify flowering plants". Proc. Natl. Acad. Sci. U.S.A. 102 (23): 8369-74.
26. Kress WJ, Erickson DL (2008). "DNA barcodes: Genes, genomics, and bioinformatics". PNAS 105 (8): 2761-2762.
27. Jesse H. Ausubel (August 4, 2009). "A botanical macroscope". Proceedings of the National Academy of Sciences 106 (31): 12569.
28. Ghosh S.K., Ghosh P.R., Trivedi S., Das P.J., Chetry A.J., Mahadani P., Bhatcharjee M.J., Ahenthem M., Khondram B., Kshetrimayum M., Das K.C., Tiwary B.K., Choudhary A., and Charkraborty C.S., (2009) DNA Barcode of Life : Species specific DNA Sequence from East and Northeast Indian Biodiversity.3rd international conference of Barcode of Life conference, Maxico city.p.p-193.
29. Bhattacharjee., M.J., Ghosh S. K., Perusing DNA Barcode Contrasts Species of Catfish Genus Mystus and Eutropiichthys of Northeast India: Discern Mini-Barcode. PLoS One (Under review).

30. Bhattacharjee M. J., Khomdram B., Ghosh S. K., (2011) DNA barcoding of Ornamental Fishes of Northeast India, Proceedings 22nd Pacific Science Congress, Kuala Lumpur, Malaysia. P.p-37.
31. Ward R.D., Zemlak T.S., Innes B.H., Last P.R., Hebert P.D.N. (2005). DNA barcoding Australia's fish species. Philosophical Transactions of the Royal Society, Series B 360: 1847-1857.
32. Hubert N., Hanner R., Holm E., Mandrak N.E., Taylor E., Burridge M., Watkinson D., Dumont P., Curry A., Bentzen P., Zhang J., April J. and Bernatchez L. (2008). Identifying Canadian freshwater fishes through DNA barcodes. Plos ONE 3, e2490-2490.
33. Venkatesh B., Dandona N. and Brenner S. (2006). Fugu genome does not contain mitochondrial pseudogenes. Genomics 87, 307-310.
34. Waters J.M. and Wallis G.P. (2001). Cladogenesis and loss of the marine life-history phase in freshwater galaxiid fishes (Osmeriformes: Galaxiidae).Evolution 55, 587-597.
35. Blazej R.G., Kumaresan P. and Mathies R.A. (2006). Microfabricated bioprocessor for integrated nanoliter-scale Sanger DNA sequencing. Proc. Natl Acad. Sci. USA 103, 7240-7245.
36. P. (2007). The book of life goes online. Genomics Soc. Policy 3, 48-51.