Preliminary study: Species identification of *Rasbora sumatrana* through the cytochrome oxidase subunit I DNA barcoding marker

M Y Norfatimah, M A R Akram, W Z A Lyena, Z N Aili and A Farizan

Faculty of Applied Sciences, Universiti Teknologi MARA, 40450 Shah Alam, Selangor Malaysia
Corresponding author: norfatimah9778@uitm.edu.my

**Abstract.** *Rasbora sumatrana* or locally known as “seluang” in Malaysia is a pelagic fish. Unlike its other congeneric species is not categorized as an ornamental fish. Samples of the presumed *R. sumatrana* was obtained from an abandoned swamp in Hulu Langat, Selangor. Initial identification of the samples was conducted based on meristic and morphometric analyses, which showed congruence with *R. sumatrana* when referenced to FishBase. To complement the analysis, a molecular study was conducted based on the cytochrome oxidase subunit I (COI) DNA barcoding marker. Confirmation of the species identity was done by BLAST results with six voucher sequences of *R. sumatrana* and other *Rasbora* species from GenBank. The relationship of *Rasbora* spp. was analysed by phylogenetic and Automatic Barcode Gap Discovery (ABGD) analysis. Contradicting with morphological and meristic data analysis, both Maximum-Likelihood and Maximum Parsimony phylogenetic tree showed that the currently studied samples did not cluster with *R. sumatrana*. We believe that they may represent cryptic diversity and are tentatively classified as *Rasbora* spp.

1. Introduction

DNA Barcoding of fishes and other species in Animalia kingdom has brought enormous advantages to us. DNA barcoding not only opens a new branch of bioinformatics but also brings new insights into biophysics, developmental biology, comparative biology, ecology and evolution, mycology, and biodiversity [1]. DNA barcoding implementation in fisheries industry does help developing country like Indonesia to enhance their economic status [3]. In Malaysia, there are a lot of studies on fishes that implement DNA barcoding. Since the COI gene is a conserved gene throughout many species, it therefore has a considerable of impact to DNA barcoding since it serves as an identification tags for the discovery of closely related species for example in parasite [4], and discovery of new species on *Danio h坦amanthis* [5]. The DNA barcoding analysis COI gene amplification helps us to identify a lot of new species.

*Rasbora sumatrana* is a freshwater fish and can only be found in Asia, including Peninsular Malaysia, Sumatra, Western Borneo, Southeast Thailand and Southwest Cambodia [6]. Even though this fish is native in Malaysia, but there are very limited data were available on this species when this study was initiated. Due to their small size, they usually become a target to the predator fish. As a member of minnows and carps family [2], *R. sumatrana* is being sold as an ornamental fishes feed in some Southeast Asia country or as bait for angling activities. This exploitation occurs not simply because of their small size, but also due to their active behaviour in water, which can attract game fish such as bass and pike.
Genus of *Rasbora* consists of 78 species and each species has its own unique characteristics. *R. paviei* or *R. paviana* can be found at river and lake channels in Vietnam [8]. *R. paviana* can also be categorized as pelagic fish, but it only feeds exogenous insects. Many studies have been done on this species such as in the discovery of nematodes species in the gut [9]. Another species *R. daniconius*, also freshwater and pelagic species but with special characteristic which is have an ability to live in a brackish water. It can live in the water with high salt concentration without any problems of dehydration. This ability has led to an analysis of the number of eggs produced by this species. *R. tawaresis*, is found primarily in the Province of Aceh, Indonesia, in Lake Tawar. This freshwater and pelagic fish have been recognized by IUCN Red List as a vulnerable species, which means that this species is declining in 16 number yearly. As it only can be found in Lake Tawar, Aceh, Indonesia, specific research on the spawning season had been conducted [9].

*Rasbora sumatrana* is a pelagic fish that prefers to live on the water surface and cannot live more than 100 meters below water surface because of extremely high pressure that they cannot deal with. This makes them prone to danger, such as ingesting unnecessary things into their digestive path. Mostly, they tend to ingest microplastic and mesoplastic [10]. Based on study by [10], microplastic and mesoplastic were discovered in gastrointestinal tract of freshwater benthopelagic fishes, significantly lower compared to sea benthopelagic fishes. Microplastic and mesoplastic are produced from larger plastics that degrade into a smaller compound. A lot of freshwater organisms, such as fish, freshwater prawns and insects have been found to consume microplastic and mesoplastic, as these types of plastics are small in size and can be easily transported through water [10]. Confronted with this threat, *R. sumatrana* is likely to face extinction. However, the status has been updated on the databases such as IUCN red list.

This preliminary study on DNA barcoding analysis of *R. sumatrana* from Malaysia will help to update information in international or local databases. Updated data may increase the visibility of *R. sumatrana* in the economic sector, as the data obtained from the DNA barcoding analysis will provide actual information on genetic makeup of this species.

### 2. Methodology

*Rasbora sumatrana* samples of eleven individuals were collected at a swamp in Kg Sungai Lui at district of Hulu Langat in Selangor, Malaysia. Latitude and longitude coordinates for the sampling location are 3.148670, 101.916834. After meristic and morphometric analysis, the fins of all fish were cut and stored in 95% ethanol. The morphometric and meristic analysis were calculated based on characterization by [11] and [6]. Another physical character for *Rasbora* spp., which is the lip characterization of lachrymal groove on the lips has been set and discussed by [6].

Total genomic isolation was performed for every 25 mg of individual fish fin by referring to the standard protocol of DNeasy® Blood and Tissue Kit. DNA barcoding was performed by PCR amplification of COI region using a pair of primers, FishF 5′–TCAACCAAACCAAGACATTGGCAC–3′ and FishR1 5′–TAGACTTCTGGGTTGGCCAAAGAATCA–3′. Master mix of 25 µl of total volume each was prepared for polymerase chain reaction by mixing the sterile distilled water, 0.5 × PCR buffer, 0.2mM dNTPs, 1.25mM MgCl2, 1.0 µl of each forward and reverse 5 µl template DNA and 0.04U of *Taq* polymerase. Meanwhile, PCR thermocycling program was initially pre-denatured at 94°C for 5 minutes and carried out with 35 cycles of denaturation at 94°C for 30 seconds, annealing at 52°C for 40 seconds and extension 72°C for 2 minutes. The PCR purification was conducted by using QIAquick® PCR purification kit before proceed with sequencing process.

The sequence data of two representatives samples of *R. sumatrana* from this study and five other individual data from similar genus of standard COI gene including *R.sumatrana* (EF452882.1), *R. vulgaris* (HM224243.1), *R. hobelmani* (HM224229.1), *R. paviana* (KP263428.1), *R. tawaresis* (HM100250.1) and one outgroup from genus *Alburnus*, which is the *A. alburnus* (MG806821.1) were used for data analysis. All sequences were trimmed and aligned; followed by a phylogenetic analysis using MEGA 7.0 software. Meanwhile, the trimmed sequences together with the outgroup sequence
were tested using saturation test available on the DAMBE 7 software provided by XiaLab from University of Ottawa http://dambe.bio.uottawa.ca/DAMBE/dambe.aspx.). Automatic Barcode Gap Discovery (ABGD) test [13] was also conducted to observe the ‘barcode gap’ between the family members.

3. Results and Discussion

3.1. Meristic and morphometric data analysis of expected Rasbora sumatrana

The image of all samples was captured according to the standard of specimen voucher prior to the collection of meristic and morphometric data. Altogether there are about ten specific features that were counted for meristic data and compared with standard obtained from Fishbase.org. The features are the number of anal fins, anal fins soft rays, caudal fins, dorsal fins number, dorsal fins soft rays, lateral line, pectoral fins number, pelvic fins number, scales around caudal peduncle and the scales on the lateral line. The data are shown in Table 1.

Nine features of each sample were evaluated for morphometric data. The features include eye diameter, fork length, head length, pre-anal length, pre-dorsal length, pre-orbital length, pre-pectoral length, pre-pelvic length and standard length. The morphometric data for all samples are shown in Table 2. Compared to the standard meristic and morphometric data for *R. sumatrana* from Fishbase.org, it is found that the samples presumed to be *R. sumatrana* because the mean calculated from the samples is ±3.0 to the standard which is a small range of differences [13].

| Meristic Data | Mean From Sample | Standard Data |
|---------------|------------------|---------------|
| Anal fins number | 1 | 1 |
| Anal fins soft rays | 7 | 8 |
| Caudal fins | Forked | Forked |
| Dorsal fins | 1 | 1 |
| Dorsal fins number | 8 | 9 |
| Lateral line | 1 | 1 |
| Scales on lateral line | 25 | 27 |

3.2. DNA Barcoding of Rasbora spp. Data Analysis

The amplification size of COI region was successfully amplified with an overall size of approximately 500 bp total size per region. Genetic distance analysis shows that both samples, RS_3 and RS_6, are similar to each other with 0.0% difference. The distance between RS_3 and RS_6 to *R. sumatrana* (EF452882.1) is more than 2%. Both RS_3 and RS_6 were similar to the other four species in *Rasbora*.
spp.. These include *R. vulgaris* (HM224243.1), *R. hobelmani* (HM224229.1), *R. paviana* (KP263428.1) and *R. tawarensis* (HM100250.1).

All eight individuals were analysed using two types of phylogenetic trees which are the maximum likelihood tree (data not shown here) and the maximum parsimony tree (Figure 1). Higher bootstrap value indicates higher chance that these individuals are similar. The ABGD approach was conducted to confirm the results of the barcode gap analysis described above. The initial partition at a prior intraspecific divergence of (P) (P = 0.0077–0.0359) was chosen, as suggested by [14], that the prior intraspecific divergence should be between 1% till 3% and that number of OTUs should be similar to other analyses. These results are also concordant with the earlier analyses of the phylogenetic tree (figure 1), which revealed that that the current studied samples did not cluster with *R. sumatrana*. We believe that they may represent cryptic diversity and are tentatively classified as *Rasbora* spp. The detailed results showed the following taxon delimitation; (1) RS_3, RS_6 to *R. sumatrana* (EF452882.1), *R. vulgaris* (HM224243.1), *R. hobelmani* (HM224229.1), *R. paviana* (KP263428)

![Figure 1](image.png)

**Figure 1.** Evolutionary relationships of taxa inferred using the Maximum Parsimony methods sample of *Rasbora* spp.

Based on the morphometric (Table 1) and meristic (Table 2), these evidences suggest that all the current samples in this study were similar to *R. sumatrana* with less than ±3% difference for each data, which led this study to proceed with molecular techniques for validation test and a preliminary test for DNA barcoding study of *R. sumatrana* from Malaysia. The meristic and morphometric data are the first parameters to measure the species for each sample belong to, since these data are standard and accepted by the biodiversity community [13]. In addition, for the high accuracy of species identification, the parameters measured in morphometric analysis should be added, especially on the characteristic of the inner part of their body cavity, because small differences on physical characteristic of individual will result of different species [14]. According to the finding by [6], the size of lateral line and size of pigmented region on anal fins of *Rasbora* sp. may also give a different result on exact species for the samples, although other characteristics and physical features of them were similar to each other.

Mitochondrial DNA was chosen in this study because mitochondrial DNA are inherited by an individual through their maternal line [18] and therefore produces a conserved region suitable for barcoding studies. The Mitochondrial DNA also been proven to have a small mutation rate compared to the genomic DNA, with only about one base mutating every 1000 years [15]; [16]. This gives the advantage of mitochondrial DNA to be used as molecular marker, especially for species identification purposes. Theoretically, the COI gene region will give the result on the general lineage of each sample [17].

The genetic distance test identifies the relationship between the tested individual and the test found that RS_3 and RS_6 were similar to each other with 0.000 differences. Based on the genetic distance data, RS_3 and RS_6 were quite far from *R. sumatrana* with the percentage of different between them is 2% but relationship of RS_3 and RS_6 are much closer to *R. vulgaris*, *R. hobelmani*, *R. paviana* and *R. tawarensis* with the percentage of differences between them is less than 1%. The genetic distance between individual determines the distance between them, the lower percentage means that the individual is close to each other [18].
Therefore, to verify the results another test was conducted, namely phylogenetic tree and ABGD analysis, but the phylogenetic tree needs preliminary test which are model test on Kimura-2-Parameter (K2 + G) and the saturation test using Kimura 80 (K80) model which identify the point mutation on nucleotide of the tested individual. The Kimura-2-Parameter measures the transition distance and the transversion rate occurring in the nucleotide [19]. The rate of transition occurs more than the rate of transversion, and this is consistent with the idea that rate of transition should be higher than the rate of transversion [20].

The transition will produce synonymous mutations whose base changes do not affect the type of amino acid produced [21]. This will give the individual the advantages to of living normally without a genetic disease problem. The saturation test comparing the rate of transition and transversion to the K80 genetic distance model also found that the transition is higher than the rate of transversion [22]. This saturation test, which produces a divergent graph shape, is an indicator that there are common genes in the individual that originate from the same ancestor. This common gene is called homology [23].

Phylogenetic trees built for this grouping were Maximum Parsimony and Maximum-Likelihood phylogenetic trees because these are character-based tree [10]. Both trees are constructed using trimmed nucleotide sequence [10]. The Maximum Parsimony is a widely used for species identification because it skips several unnecessarily evolutionary steps [24] while the Maximum-Likelihood is a tree that can produce reproducible phylogeny information on the individual tested [10].

Both trees in this study proved that RS_3 and RS_6 are similar to each other, but their relationship toward R. sumatrana standard sequence from GenBank is far. They appear to be similar to the group of R. vulgaris and R. tawarensis in Maximum-Parsimony tree while RS_3 and RS_6 are 34% similar to the R. paviana in Maximum Likelihood. In this preliminary study, the number of sample size in this study are limited:- two samples, with at least 30 samples should be used so that the common shared traits among them can be used in the classification using Maximum Parsimony or Maximum-Likelihood phylogenetic tree [25]. Another possible source of error is the number of standard sequences of R. sumatrana used by GenBank. As this species does not have much scientific research done on it, there is only one sequence for COI genes in the database, which will affect the specificity of the data obtained from this study.

To prove this, a further test, known as ABGD analysis, is carried out using online ABGD software. This test is a powerful tool to find the relationship between individual tests and the date representation on a graph [12]. The ABGD analysis will determine the number of groups that exist from the uploaded data and give the percentage of similarity between them. In this study, the accepted value is only from 0.5% to 2.0%, as we are using the Kimura 80 model that suggest only this range of data will provide correct information on each species.

From the analysis, there are only two groups exist in the tested sample. The first group is for Alburnus alburnus, the out group for this study and another group is for all individuals under Rasbora spp. Since the ABGD analysis is looking for gaps between individuals, this outcome shows that there is no gap between COI sequence of RS_3, RS_6, R. sumatrana, R. vulgaris, R. paviana, R. hobelmani and R. tawarensis. Thus, it is proven that COI gene can only become a barcoding marker until the classification of the genera but cannot differentiate small differences between species [26].

4. Conclusion
Rasbora sumatrana can be identified through meristic and morphometric data obtained during the sampling process by comparing it with the standard data retrieved from fishbase.org. The amplification process for standard barcoding gene, the CO1 gene was successfully performed using universal primer. Furthermore, both phylogenetic trees and ABGD analysis concordantly reveal specific group for them which is Rasbora spp. In future, the DNA Barcoding analysis for R. sumatrana can be conducted extensively by using samples from other sampling sites with larger sample size and by using the other alternative barcoding marker.

Acknowledgments
We are indebted to the laboratory staff of Molecular Biology lab and Dean of Faculty of Applied Sciences, Universiti Teknologi MARA Shah Alam, Malaysia for invaluable knowledge specifically for the experience, opportunity, and funding with our research group.

References
[1] Wilson J J, Sing K W and Jaturas N 2018 DNA Barcoding: Bioinformatics Workflows For Beginners. In Reference Module in Life Sciences (Elsevier)
[2] Helfman G S 2013 Fishes, Biodiversity af. In S. A. Levin (Ed.), Encyclopedia of Biodiversity Second Edition Academic Press 456–476 pp
[3] Abdullah A and Rehbein H 2017 DNA barcoding for the species identification of commercially important fishery products in Indonesian markets. International Journal of Food Science & Technology 52 266–274
[4] Lehrer V, Bañuls A L, Léger N, Rioux J A and Depaquit J 2017 Phlebotomus (Paraphlebotomus) chabaudi and Phlebotomus riouxi: closely related species or synonyms? Parasite (Paris, France) 24, 47
[5] Kullander S O, Rahman M M, Norén M and Mollah A R 2018 Laubuka tenella, a new species of cyprinid fish from southeastern Bangladesh and southwestern Myanmar (Teleostei, Cyprinidae, Danioninae) ZooKeys 742 105–126
[6] Lumbantobing D N 2014 Four new species of Rasbora of the Sumatrana group (Teleostei: Cyprinidae) from northern Sumatra, Indonesia Zootaxa 3764 1–25
[7] Moravec F and Kamchou K. 2012 Description of Rhabdochona (Globochona) rasborae sp. (Nematoda: Rhabdochonidae) from the freshwater cyprinid fish Rasbora paviana Tirant in Southern Thailand Folia Parasitologica 59 209–215.
[8] Stolbunov I A, Than N T H and Pavlov D S 2011 Morphological and behavioral variation of Rasbora paviei (Cyprinidae, Cypriniformes) from lotic and limnic habitats (Central Vietnam). Journal of Ichthyology 51 352.
[9] Muchlisin Z A, Musman M and Azizah M S 2010 Spawning seasons of Rasbora tawarensis (Pisces: Cyprinidae) in Lake Laut Tawar, Aceh Province, Indonesia Reproductive Biology and Endocrinology 8.
[10] Jabeen K, Su L, Li J, Yang D, Tong C, Mu, J and Shi H 2017 Microplastics and mesoplastics in fish from coastal and fresh waters of China Environmental Pollution 221 141–149.
[11] Taki, Y 1974b Fishes of the Lao Mekong Basin. (Laos: U.S. Agency for International Development Mission to Laos Agriculture Division)
[12] Puillandre N, Lambert A, Brouillet S and Achaz G 2012 ABGD, Automatic Barcode Gap Discovery for primary species delimitation: ABGD, Automatic Barcode Gap Discovery. Molecular Ecology 21 1864–1877
[13] Begg G A and Waldman J R 1999 A holistic approach to fish stock identification Fisheries Research 43 35–44.
[14] Ratnasingham S and Hebert P D 2007 BOLD: The Barcode of Life Data System (http://www.barcodinglife.org) Molecular Ecology Notes 7 355–364
[15] Ladoukakis E D and Zouros E 2017b Evolution and inheritance of animal mitochondrial DNA: rules and exceptions Journal of Biological Research-Thessaloniki 24 2
[16] Haag-Liautard C, Coffey N, Houle D, Lynch M, Charlesworth B and Keightley P D 2008. Direct Estimation of the Mitochondrial DNA Mutation Rate in Drosophila melanogaster. PLoS Biology 6
[17] Chao Z, Liao J, Liang Z, Huang S, Zhang L and Li J 2014 Cytochrome C oxidase subunit I barcodes provide an efficient tool for Jinqian Baihua She (Bungarus parvus authentication) Pharmacognosy Magazine 10 449–457
[18] Nei M 1972 Genetic Distance between Populations The American Naturalist 106 283–292
[19] Collins R A, Boykin L M, Cruickshank R H and Armstrong K F 2012 Barcoding’s next top model: an evaluation of nucleotide substitution models for specimen identification Methods in
Ecology and Evolution 3 457–465

[20] Anne Z Y, Yoder A D and Primates L 1999 Estimation of the Transition/Transversion Rate Bias and Species Sampling

[21] Kristofich J, Morgenthaler A B, Kinney WR, Ebmeier C C, Snyder D J, Old W M, Copley S D 2018 Synonymous mutations make dramatic contributions to fitness when growth is limited by a weak-link enzyme PLOS Genetics 14 e1007615

[22] Xia X and Xie Z 2001 DAMBE: Software Package for Data Analysis in Molecular Biology and Evolution Journal of Heredity 92 371–373.

[23] Philippe H, Brinkmann H, Lavrov D V, Littlewood D T J, Manuel M, Wörheide G and Baurain D 2011 Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough PLOS Biology 9 e1000602.

[24] Kannan L and Wheeler W C 2012 Maximum Parsimony on Phylogenetic networks. Algorithms for Molecular Biology: AMB, 7, 9

[25] Subramanian S 2016 The effects of sample size on population genomic analyses –implications for the tests of neutrality. BMC Genomics 17

[26] Chee S Y 2015 Limitations of cytochrome oxidase I for the barcoding of Neritidae (Mollusca: Gastropoda) as revealed by Bayesian analysis Genetics and Molecular Research 14 5677–5684