Research Article

SOHAG JOURNAL OF SCIENCES

https://doi.org/10.21608/sjsci.2024.312092.1219

Identification of Hypervariable Domains in 12SrDNA for Enhanced Fish Species Barcoding from the Red Sea Regions of Egypt and Yemen

Asmaa Galal-Khallaf[1], Aya Ibrahim Elberri[1], Omir AbdelWahab Rabie[1], Taha S.S. Baker[2], Abdelaziz A.A. El-Sayed[3,4], Hamdy A. M. Soliman[5,*], Khaled Mohammed-Geba[1]

[1] *Department of Zoology, Faculty of Science, Menoufia University, Shebin El-Kom, Egypt.*
[2] *Department of Biology, College of Education, Ataq Shabowah University, Yemen.*
[3] *Department of Zoology, Faculty of Science, Zagazig University, Zagazig, Egypt.*
[4] *Department of Zoology, Faculty of Science, Islamic University of Madinah, Saudi Arabia.*
[5] *Department of Zoology, Faculty of Science, Sohag University, Egypt.*
[*] *Email: hamdy_soliman @science.sohag.edu.eg*

Received: 13th August 2024, **Revised:** 7th December 2024, **Accepted:** 19th December 2024
Published online: 7th February 2025

Abstract: Ribosomal RNA (rRNA) genes, particularly 12SrDNA, have proven to be effective barcodes for fish species identification due to their secondary structure domains that exhibit interspecific variability. This study focused on identifying hypervariable domains within the 12SrDNA gene that contribute to its barcoding efficiency. Fish samples from Egypt and Yemen, representing various orders and families, were analyzed using 12SrDNA and COI gene sequencing. Four hypervariable domains (D2, D3, D5, D7) within 12SrDNA showed the highest divergence among species. Concatenating these domains significantly increased interspecific genetic distances. COI and 12SrDNA barcodes revealed 3 misidentified species (*Plectorhinchus sordidus, Mulloidichthys flavolineatus*, and *Lethrinus nebulosus.*), emphasizing the importance of focusing on specific hypervariable domains for accurate species identification.

Keywords: COI gene; DNA barcoding; Hypervariable domains, Fish species; 12SrDNA, Red Sea.

1. Introduction

DNA barcoding, first introduced by Henert *et al.* [1], has revolutionized molecular taxonomy and biodiversity studies. This technique, which involves the PCR-based amplification for the 5`barcode region of the mitochondrial cytochrome oxidase subunit I (COI) gene, followed by sequencing and comparison with standard databases, and has proven effective for accurate fish species identification. Applications range from assessing cryptic diversity in deep-sea fish [2], to ensuring traceability in fish markets [3], and creating comprehensive national fish inventories [4].

Despite the success of COI, 12SrDNA has gained popularity in fish barcoding due to its specificity and fewer amplification issues in some species [5-8]. (or the presence of barcodes of misidentified species in some genetic databases in other cases [9, 10]. The Red Sea, known for its rich marine biodiversity and significant fisheries in Yemen and Egypt, has become a focus of DNA barcoding studies to monitor and protect its unique ecosystems [11-16]. However, little attention has been given to the 12SrDNA gene's potential as a discriminatory barcode for fish iHe

Given the ecological and economic importance of the Red Sea and the fluctuations in fisheries production, this study aims to explore DNA sequence variability within the 12SrDNA gene's secondary structures in fish species from the northern Red Sea (Egypt) and the southern Red Sea (Yemen) [17-19]. This study seeks to identify whether specific domains within the 12SrDNA can serve as efficient barcodes for species discrimination in this vital marine environment. This can be a step ahead towards achievement of the 14th goal of the United Nations Specific Developmental Goals (SFGs), that is concenrned with the study and protection of life below water.

2. Materials and methods:

2.1. Collection of Samples

Fish samples were collected from two key regions: Suez City, Egypt, and Sanaa City, Yemen. Five individuals from each of the five species were obtained from Egypt, and five individuals from each of the six species were sourced from Yemen. These species belonged to three orders—Mugiliformes, Eupercaria incertae sedis, and Mulliformes—and six families, including Labridae, Mugilidae, and Lutjanidae (Fig. 1).

Collected scales were rinsed, dried, and transported to the Molecular Biology and Biotechnology Laboratory at Menoufia University for DNA extraction and barcoding. Ethical approval for animal handling was granted by the Institutional Animal Care and Use Committee (IACUC) at Menoufia University under the number MUFSFGE122.

2.2. DNA barcoding procedures:

2.2.1. DNA extraction

Genomic DNA was extracted from each dried scale

individually using 5 % Chelex®100 sodium form resin (Sigma-Aldrich, Madrid, Spain) in TE buffer (pH 8) according to the protocols described by [20, 21]. Proteinase K (1.6 U, ThermoFisher) was added to each tube. Samples were incubated at 60 ℃ with repeated vortexing every 30 min for a total of 2hours. The samples were then boiled in a 100 ℃ Thermal block (dry bath, Benchmark Scientific, USA) for 20 minutes. Finally, the samples were stored at 4 ℃ until PCR amplification of target genes.

2.2.2. Amplification of barcode fragment

The sequence of the barcode region of the mitochondrial 12SrDNA gene in all samples was amplified using universal primers 12SA: 5′-AAACTGGGATTAGATACCCCAC-3′ and 12SF: 5′- GAGGGTGACGGGCGGGCGGTG-3′ [22].

Also, the 5′ barcode region sequence of the mitochondrial COI gene in each sample was amplified by PCR using the set of primers described by Ward *et al.* [23], namely: wCoI-Fw: 5′- TCAACCAACCACAAAGACATTGGCAC-3′, and wCoI-Rv: 5′- TAGACTTCTGGGTGGCCAAAGAATCA-3′.The amplification reactions were set up as 100 ng of template DNA, 1x MyTaq™ Red Mix (Bioline), 0.5 µM of each primer, and 200 ng mL^{-1} of bovine serum albumin (BSA), to a total volume of 25 µL. PCRs were carried out in the thermal cycler TC512 (Techne, UK). The PCR program included an initial preheating/polymerase activating step at 95 ℃ for 5 min for both genes. For the COI gene, PCR involved 40 cycles of amplification with the following steps: denaturation at 95°C for 30 seconds, annealing at 56°C for 30 seconds, and extension at 72°C for 30 seconds. The same number of cycles was used for the 12SrDNA gene but with annealing at 52°C for 30 seconds and an extended extension time of 90 seconds at 72°C. Both protocols concluded with a final extension step at 72°C for 10 minutes. The amplified products were visualized using 1 % agarose gel electrophoresis stained by 0.5 µg $µL^{-1}$ of ethidium bromide. PCR products were then sent to Macrogen Inc., South Korea, for sequencing, applying the conventional Sanger chain termination sequencing method.

2.2.3. Analysis of DNA barcodes

Mitochondrial 12SrDNA and COI gene sequences were reviewed and manually trimmed whenever necessary. Sequences edition was carried out using Chromas Lite software version 2.6.5 (Technelysium- Pty Ltd, available from the URL http://technelysium.com.au/), they were then compared to international DNA barcode databases, including GenBank and BOLD, for species identification.

2.2.4. Species authentication using phylogenetic analysis

2SrDNA sequences were aligned using CLUSTALW in MEGA11, and reference sequences were obtained from GenBank (https://www.ncbi.nlm.nih.gov/nuccore/). The best nucleotide substitution model was selected using the ModelTest algorithm, and a neighbor-joining phylogenetic tree was constructed with 1,000 bootstrap replicates.

2.2.5. Delimiting hypervariable regions in 12SrDNA sequences

Sequences were uploaded to the RNAcentral 2021 database (https://rnacentral.org/) for 2D structural visualization using the R2DT framework [24]. This framework allows the identification of secondary structures' domains based on the previous building of domain-specific covariance models, and integrates them as domain-specific templates that enable the prediction of different features and secondary structures in the 2D structure of query rDNA [24]. The reference sequence from Crossostoma lacustre (GenBank accession M91245.1) was used to identify hypervariable regions. Alignments were conducted using MEGA11, and the best nucleotide substitution model was determined using jModelTest 2 [25]. Interspecific genetic pairwise distances were calculated for both full and concatenated 12SrDNA sequences.

3. Results and Discussion:

3.1. Amplification and analysis of DNA sequences:

Amplification of the 12SrDNA gene in all samples resulted in PCR products of about 450 base pairs (bp). Trimming of non-informative and background nucleotide sequences resulted in about 400 bp-length nucleotide sequences. Meanwhile, COI, amplicon sizes were about 700 bp in length, of which manual trimming resulted in datasets of about 650 bp-length for all sequences.

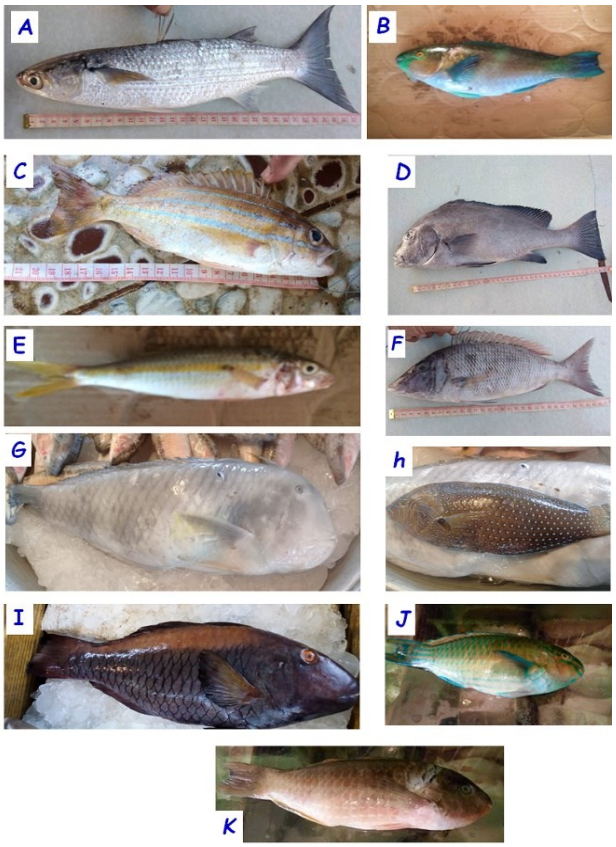

Fig. 1: Fish collected in the current study for analysis of 12SrDNA-COI gene barcodes. The fishes collected from the Sanaa City fish market (Yemen) were identified as *A: Valamugil seheli, B: Scarus ghobban, C: Lutjanus kasmira, D: Plectorhinchus pictus, E: Mulloidichthys vanicolensis,* and F: *Lethrinus nebulosus.* The fishes collected from the Suez City fish market (Egypt) were identified as G: *Iniistius pavo,* H: *Anampses caeruleopunctatus,* I: *Cetoscarus bicolor,* J: *Scarus ghobban,* and K: *S. collana*.

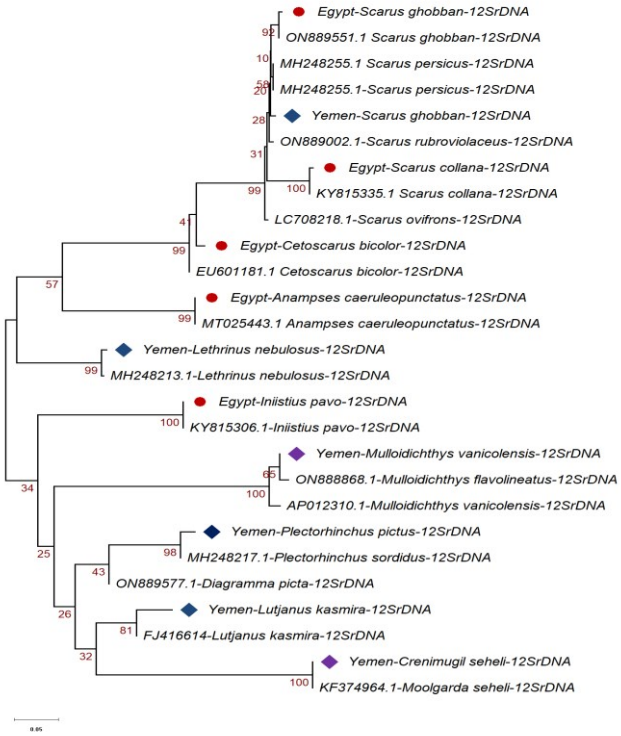

Fig. 2: Neighbor-Joining (NJ) phylogenetic tree between sampled fish in the current study from Egyptian and Yemeni waters and references for conspecifics in different areas in the world that were downloaded from the GenBank database. Bootstrap values are shown above the tree. Kimura-2-parameter was automatically selected by Mega11, with G value of 0.27.

3.1.1. 12SrDNA mitochondrial gene identities

In general, Yemeni samples showed more 12SrDNA sequence variations from their GenBank references than was the case for the Egyptian ones, except for *Valamugil seheli.* The samples collected for *V. seheli* exhibited 12SrDNA sequence identity between 98.92 % and 100 % with different GenBank references for *Crenimugil seheli* accession number (acc. no. KF374963.1 and KF374964.1, respectively). *Scarus ghobban*, however, showed mixed identities, between 98.64% with both *S. ghobban* and *S. persicus* (ON889551.1 and MH248255.1, respectively). *Lutjanus kasmira* samples showed 100 % sequence identity with GenBank references for *L. fulviflamma* (acc. no. MH248214.1), 98.66 % with *L. carponotatus* (acc. no. NC_044104.1), and 98.4 % with *L. ophuysenii* (acc. no. AB972219.2). *Plectorhinchus pictus* showed the highest 12SrDNA identity (i.e., 99 %) with the sequences with (acc. no. MH248217.1), which was *P. sordidus*, but only 95 % ID with *P. pictus* (acc. no. ON889577.1). Likewise, samples collected and sold as *Mulloidichthys vanicolensis* showed the highest 12SrDNA identity (i.e., 99.73 %) with *M. flavolineatus* (acc. no. ON888874.1, ON888868.1), but lower ID (i.e., 97.58 %-97.85 %) with *M. vanicolensis* (acc. no. AP012310.1- ON889210.1). *Lethrinus nebulosus* samples collected in the current study showed mixed identities (*i.e.*, 98.93 %) with *L. nebulosus* and *L. lentjan* (acc. no. LC645339.1, ON888683.1, respectively). Hence, 12SrDNA alone was not adequate to discriminate this species.

For the fish species collected from Egypt, *Iniistius pavo* showed 100 % 12SrDNA sequence identity with the GenBank sequence for the same species, with the (acc. no. KY815306.1). *Anampses caeruleopunctatus* showed 99.4-100% 12SrDNA sequences identities with GenBank references, for example, the ones with(acc. no. AJ810124.1 and JN935298.1). *Cetoscarus bicolor* showed the highest sequence identity, *i.e.,* 98.37% and 98.14% with references for the same species in GenBank with (acc. no. EU601181.1 and AY081070.1, respectively). Interestingly, the Egyptian samples of *Scarus ghobban* showed some difference with the same species collected from Yemen, having 100 % 12SrDNA sequence identity with ON889551.1 (Indonesia), and MW630158.1 (Fiji),- all of, which were *S. ghobban*. Lower identity, despite being very close, was found with *S. persicus* (acc. no. MH248255.1). *Scarus collana* showed 100 % identity with the sequence available in GenBank under (acc. no. KY815335.1).

The constructed NJ phylogenetic tree showed consistent results for species identity with GenBank comparisons. All Egyptian and Yemeni samples claded with their conspecifics from different places in the world with high probability values (Fig. 2). Based mainly on barcode identities and NJ phylogenetic analysis, the 12SrDNA sequences for the obtained species were deposited in the GenBank database with the (acc. no. PQ669144-PQ669154).

3.1.2. COI

The sequences of the COI gene were more consistent with the initial identification for most of the species collected from Yemeni and Egyptian fish markets in the current study, except for *Plectorhinchus* and *Mulloidichthys.* For *V. seheli,* COI barcode sequence identity was 99-100 % with GenBank references for the same species, such as the ones with (acc. no. KF010468.1 and MT888994.1). Yemeni and Egyptian *Scarus ghobban* samples showed a wide range of identities with references of the same species in the GenBank database, for example, 97.42% with South China specimen of *S. ghobban* (acc. no. OK347602.1), and 99.6-100% with specimens from the Arabian Gulf (acc. no. HQ149929.1 and HQ149930.1, respectively). *Lutjanus kasmira* showed 99.49-99.68 % with GenBank references for the same species ex. (acc. no. HQ658118.1, MZ606158.1). The samples collected as *Plectorhinchus pictus* showed 99.84-100% sequence identities as *P. sordidus* (acc. no. JX042284.1, MH331826.1). Only 84% sequence identity was observed with *P. pictus* (acc. no. KY371972.1). Specimens collected as *Mulloidichthys vanicolensis* showed 99.8-100% sequence identities with *M. flavolineatus* (acc. no. MN733607.1-KY371761.1), but only 93 % with *M. vanicolensis* (acc. no. JF493910.1). *Lethrinus nebulosus* showed 98-99.9% sequence identity with GenBank references with (acc. no. MN511939.1 and MT328991.1, respectively). Yet, they only showed 91-93% identity with *L. lentjan* GenBank COI references, such as the ones with (acc. no. KJ920116.1 and MW498679.1). *Iniistius pavo* barcode COI regions sequences showed as low as 99.46 % and as much as 100 % identity with the same species collected from different areas in the Indian Ocean and the Red Sea, for example, these with (acc. no. KU944587.1 and MT888972.1).

Anampses caeruleopunctatus COI sequences identified in the current study showed 99.24-100% identity with several GenBank references ex. (acc. no. KU892941.1, MK657001.1). *Cetoscarus bicolor* COI sequences showed 99.05%-100 % identities with different GenBank references for the same species, for example, the ones with (acc. no. KU892951.1 and KC970464.1). *Scarus collana* showed 100 % COI sequence identity with MW872754.1 and MF124032.1. Based mainly on barcode identities with their GenBank refences, the COI sequences for the obtained species were deposited in the GenBank database with the (acc. no. PQ672284-PQ672294).

3.2. Species comparison using 12SrDNA hypervariable domains sequences

Secondary structures in the 12SrDNA gene region in different Egyptian and Yemeni Red Sea species sequenced in the current study were identified compared to *Crossostoma lacustre* (Teleosts: Siluriformes) due to the availability of its 12SrDNA full sequence and secondary domains structures on RNACentral database (Fig. 3). The zone usually amplified by the universal primer pairs of Palumbi (1996), *i.e.,* 12SA and 12SF primers, was located in the 3` area of the gene (Fig. 3). The partial nucleotide sequences of these 3` areas of the 12SrDNA in the assessed fishes exhibited hypervariability between species (Figs. 3-4).

In the 3`-barcode region of the 12SrDNA, 7 domains could be identified as D1-D7 (Figs. 3-4). However, only 4 of them showed clear interspecific variations at the level of stretches of adjacent nucleotides, which were D2, D3, D5, and D7 (Fig. 3). The D3 itself showed the least or no interspecific variations. However, a right-side finger from this domain exhibited this species-specific variation (Figs. 4a-g). For *I. pavo,* the D3 side extension (D3-SE), D5 and D7 sequences were 5`-TACCCTACAATG-3`, 5`-TTGTCCCTT-3`, and 5`-ACTAGTC-3`. For the Egyptian samples of *S. ghobban*, D3-SE, D5, and D7 sequences were 5`-AGTGAAACACTG-3`, 5`-TTGCTCATT-3`, and 5`-ACCCATA-3`, respectively. For the Egyptian samples of *S. ghobban*, D3-SE, D5, and D7 sequences were 5`-AGTGAAACACTG-3`, 5`-TTGCTCATT-3`, and 5`-ACCCATA-3`, respectively. For *S. collana*, D3-SE, D5 and D7 sequences were 5`-AGTGGGACACTG-3`, 5`-TTGCTCATT-3`, and 5`-ACCCATA-3`, respectively. For *C. bicolor*, D3-SE, D5 and D7 sequences were 5`-AGTGAAACACTG-3`, 5`-TTGCTTATT-3`, and 5`-ATCAACA-3`, respectively. For *A. caeruleopunctatus*, D3-SE, D5 and D7 sequences were 5`-ATAGCAATATTG-3`, 5`-TTGCCTAGT-3`, and 5`-GATAGAA-3`, respectively. Despite the similarities among D3-SE, D5, and D7 in *S. ghobban*, *S. collana*, and *C. bicolor*, their D2 were showed more interspecific hypervariability, having the sequences of 5`-GTAAGTGGCGAATAGAGAGCCCCACTGAAATTGGCC C-3` (for the Egyptian *S. ghobban)*, 5`-GTAAGTGGGGAGTAGAGAGCCCCGCTGAAATCGGCC C-3`, and 5`-GTAAGCGGGGAATAGAGAGCCCCACTGAAACCGGCC C-3`, respectively. In *A. caeruleopunctatus,* D2 sequence was 5`-GTAAGCAGGGAATAGAGAGCCCCGCTGAAACCGGCC C-3` (Fig. 4).

Table 1: Genetic pairwise distances among 12SrDNA 3` barcode region sequences of the collected Red Sea fish species in the study. The kimura-2 parameter was identified as the best substitution model fitting these data.

	Egypt-Scarus collana	Egypt-Scarus ghobban	Yemen-Scarus ghobban	Egypt-Cetoscarus bicolor	Egypt-Anampses caeruleopunctatus	Egypt-Iniistius pavo	Yemen-Lutjanus kasmira	Yemen-Plectorhinchus pictus	Yemen-Mulloidichthys vanicolensis	Yemen-Lethrinus nebulosus
Egypt-Scarus collana										
Egypt-Scarus ghobban	0.06									
Yemen-Scarus ghobban	0.05	0.01								
Egypt-Cetoscarus bicolor	0.12	0.09	0.11							
Egypt-Anampses caeruleopunctatus	0.41	0.41	0.38	0.26						
Egypt-Iniistius pavo	0.57	0.53	0.56	0.59	0.35					
Yemen-Lutjanus kasmira	0.61	0.48	0.54	0.41	0.35	0.38				
Yemen-Plectorhinchus pictus	0.60	0.49	0.59	0.50	0.44	0.32	0.20			
Yemen-Mulloidichthys vanicolensis	0.60	0.67	0.64	0.57	0.47	0.45	0.38	0.32		
Yemen-Lethrinus nebulosus	0.41	0.39	0.43	0.35	0.39	0.37	0.22	0.30	0.43	
Yemen-Crenimugil seheli	0.79	0.62	0.70	0.54	0.58	0.49	0.32	0.36	0.53	0.32

Table 2: Genetic pairwise distances among 12SrDNA 3` concatenated hypervariable domains in the 12SrDNA 3` barcode region (i.e., D2, D3, D5, and D7) sequences of the collected Red Sea fish species in the current study. Jukes-Cantor was identified as the best substitution model fitting these data.

	Egypt-Scarus collana	Egypt-Scarus ghobban	Yemen-Scarus ghobban	Egypt-Cetoscarus bicolor	Egypt-Anampses caeruleopunctatus	Egypt-Iniistius pavo	Yemen-Lutjanus kasmira	Yemen-Plectorhinchus pictus	Yemen-Mulloidichthys vanicolensis	Yemen-Lethrinus nebulosus
Egypt-Scarus collana										
Egypt-Scarus ghobban	0.07									
Yemen-Scarus ghobban	0.06	0.01								
Egypt-Cetoscarus bicolor	0.17	0.14	0.14							
Egypt-Anampses caeruleopunctatus	0.56	0.54	0.49	0.30						
Egypt-Iniistius pavo	0.85	0.79	0.79	0.92	0.55					
Yemen-Lutjanus kasmira	0.73	0.60	0.60	0.46	0.46	0.72				
Yemen-Plectorhinchus pictus	0.72	0.66	0.73	0.66	0.68	0.60	0.32			
Yemen-Mulloidichthys vanicolensis	0.87	0.91	0.86	0.73	0.82	0.80	0.53	0.48		
Yemen-Lethrinus nebulosus	0.59	0.57	0.57	0.45	0.50	0.53	0.25	0.38	0.69	
Yemen-Crenimugil seheli	1.03	0.76	0.76	0.67	0.84	0.76	0.44	0.53	0.84	0.50

For *Crenimugil seheli*, the D3-SE, D5 and D7 sequences were 5`-CCTAAGCAGG-3`, 5`-TTG TTC CTA C-3`, and 5`-ACT AGT A-3` (Fig. 4a), respectively. For Yemeni *Scarus ghobban*, D3-SE, D5, and D7 sequences were the same as in the Egyptian samples of the same species (Fig. 4b). For *Plectorhinchus* sp., the sequences were D3: 5`-ACTGTATCA GT-3`, D5: 5`-TTGTTTTTCC-3`, and D7: 5`-CCTTACA-3` (Fig. 4c). For *Lutjanus kasmira*, the sequences were D3: 5`-CCTGATTAC AGT-3`, D5: 5`-TTGTTTTTT C-3`, D7: 5`-ACTTATA-3` (Fig. 4d). For *Mulloidichthys* sp., the sequences were D3: 5`-GGTACAATAGC-3`, D5: 5`-CTGTTAATCC-3`, D7: 5`-AATCATA-3` (Fig. 4e).

In *Lethrinus nebulosus*, the sequences were D3: 5`-ACC CTT TAC GGT-3`, D5: 5`-TTG TTC ATC C-3`, and D7: 5`-TCCAATA-3` (Fig. 4f). `. For the Egyptian samples of *S. ghobban*, D3-SE, D5 and D7 sequences were 5`-AGTGAAACACTG-3`, 5`-TTGCTCATT-3`, and 5`-ACCCATA-3`, respectively (Fig. 4g). For *I. pavo,* the D3 side extension (D3-SE), D5 and D7 sequences were 5`-TACCCTACAATG-3`, 5`-TTGTCCCTT-3`, and 5`-ACTAGTC-3 (Fig. 4h). For *A. caeruleopunctatus*, D3-SE, D5 and D7 sequences were 5`-ATAGCAATATTG-3`, 5`-TTGCCTAGT-3`, and 5`-GATAGAA-3`, respectively (Fig. 4i). For *C. bicolor*, D3-SE, D5 and D7 sequences were 5`-

AGTGAAACACTG-3`, 5`-TTGCTTATT-3`, and 5`-ATCAACA-3`, respectively. For *S. collana*, D3-SE, D5 and D7 sequences were 5`-AGTGGGACACTG-3`, 5`-TTGCTCATT-3`, and 5`-ACCCATA-3`, respectively (Fig. **4k**)

Finally, genetic pairwise distances were identified, based on aligning fish sequences of 12SrDNA full barcode DNA fragment and 12SrDNA D3, D5, and D7 concatenated domains; and COI barcode fragment (Tables **1-2**). The best substitution models found for them were the Kimura-2 parameter and Jukes-Cantor, respectively. For the 12S rNDA full barcode, the distances ranged between 0.178, between *Plectorhinchus* sp. and *L. kasmira,* to 0.539, i.e., between *C. seheli* and *S. ghobban* (Table **1**). However, restricting the genetic pairwise comparisons to only concatenated D3, D5, and D7 domains increased, to almost one-fold, the values to the range between 0.252, i.e., between *Plectorhinchus* sp. and *L. kasmira,* to 1.012, i.e., between *C. seheli* and *S. ghobban* (Table **2**). This study successfully identified four hypervariable domains within the 3' end of the 12SrDNA gene that efficiently discriminated among 11 fish species commonly found in the Red Sea and Eastern African fisheries.

Species discrimination are based on GenBank comparison of sequences and NJ tree phylogenetic analysis. The phylogenetic analysis identified the Kimura 2-parameter (K2P) model as the best to describe substitutions in the 12SrDNA barcode, whereas Jukes-Cantor (JC69) was the best for COI barcodes. The JC69 is the simplest nucleotide substitution model. It assumes equal nucleotide frequencies, and that any nucleotide can change to any other with equal probability [**26-27**]. The K2P model agrees with JK9 in assuming equal base frequencies, but it considers transition (purines/purine or pyrimidine/pyrimidine substitutions) and transversion rates (purine/pyrimidine substitutions) among all sites, and it identifies that they occur at the same rate [**27-28**]. The utility of ribosomal RNA genes as DNA barcodes has been well-documented, often showing performance equal to or superior to the COI gene for species discrimination. For instance, the 16S rDNA gene has been demonstrated to be a better phylogenetic marker than COI in hydrozoans [**29**], and the 12SrDNA gene has shown teleost-specific barcoding patterns in fish [**30**].

Additionally, 12SrDNA is increasingly recognized as a potential key tool for detecting fish species in environmental DNA (eDNA) [**31 - 34**]. Despite this, the lack of comprehensive morpho-genetic studies combining 12SrDNA-based DNA barcoding with traditional fish identification has led to gaps in barcoding databases, resulting in the loss of operational taxonomic units (OTUs) [**16, 35**]. Therefore, expanding the 12SrDNA database with individually collected species remains crucial. It is now well-established that not all domains within ribosomal RNA gene sequences contribute equally to species discrimination; only hypervariable domains possess significant species-discriminative power. This concept has been debated across different taxa, with specific domains being highlighted for their utility in species identification. For example, the bacterial 16S rDNA gene has long been accepted to encompass 9 hypervariable domains [**23, 36**].

However, many studies showed that the hypervariable domains (V4-V6) [**37**] and V1-V3 [**38**] show the highest species discrimination power. Similarly, in higher organisms, D2, D5, and D8 domains of the 28S rDNA gene were the most effective for species differentiation in mites [**39**], whereas D2 and D3 domains were the most adequate for inferring phylogenetic relationships within the hymenopteran genus *Encarsia* [**40**]. Furthermore, variations in the central domain of the 12SrDNA gene and the D-domain of the 16S rDNA gene specifically were characterized for 11 cod fish species [**41**].

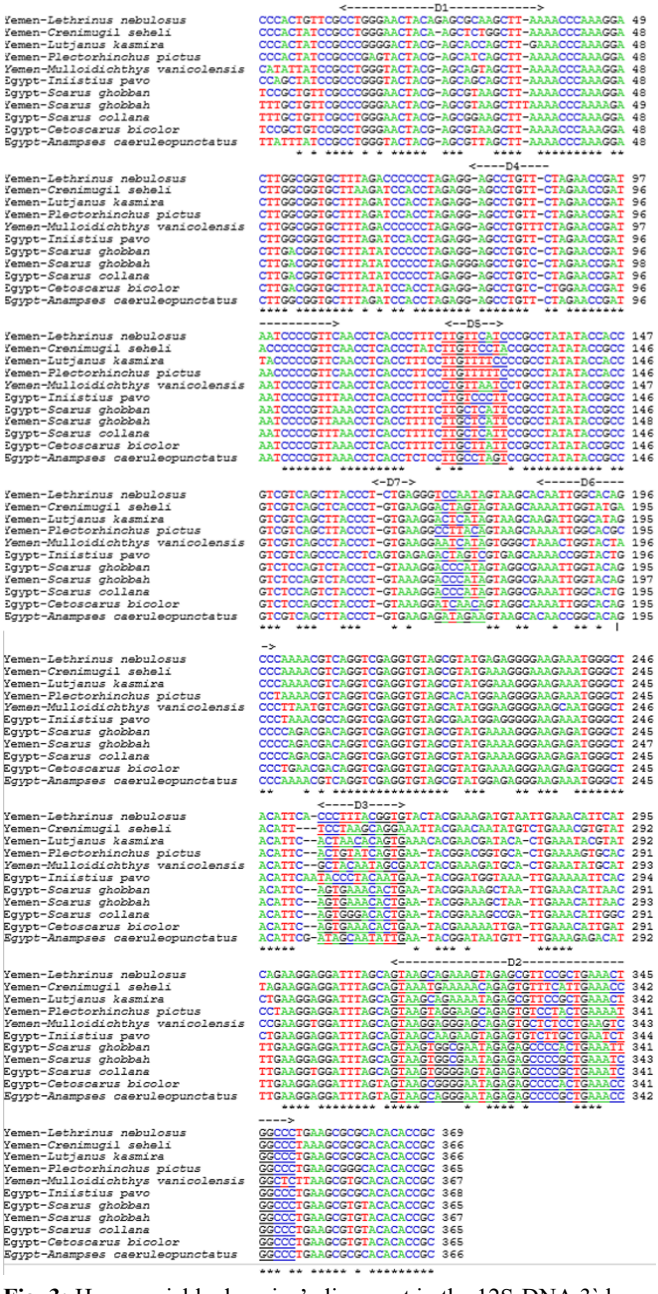

Fig. 3: Hypervariable domains' alignment in the 12SrDNA 3` barcode region in a: Iniistius pavo, b: *Anampses caeruleopunctatus*, c: Cetoscarus bicolor, d: Egyptian *Scarus ghobban*, e: *S. collana*, f: *C. seheli*, g: *S. ghobban*, h: *Plectorhinchus sp*., i: *Lutjanus kasmira*, j: *Mulloidichthys sp*., k: *Lethrinus nebulosus*. The full 12SrDNA sequence of *C. lacustre* (l) was retrieved from the RNAcentral database (https://rnacentral.org/) and used as a reference for secondary structures in this gene.

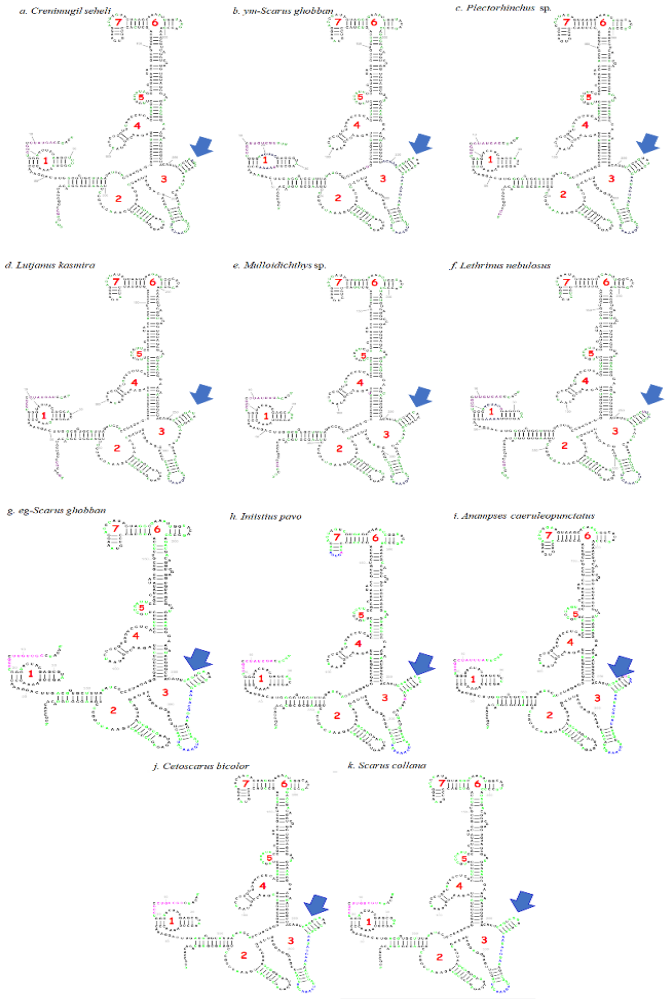

Fig. 4: 2-dimensional (2D) RNA structure of the sequenced 12SrDNA barcode regions as visualized in R2DT framework (https://rnacentral.org/r2dt) for: a) *C. seheli*, b) Yemeni *S. ghobban*, c) *Plectorhinchus sp*., d) *L. kasmira*, e) *Mulloidichthys sp*, f) *L. nebulosus*, g) Egyptian *S. ghobban*, h) *Iniistius pavo*, i) *Anampses caeruleopunctatus*, j) *Cetoscarus bicolor*, and *k) Scarus collana*.

Focusing on a limited number of hypervariable domains for sequencing enhances accuracy in species discrimination [39, 41, 42], whereas sequencing larger fragments can increase error rates due to non-informative domains [38]. Therefore, a strategic approach to species discrimination should involve targeting specific hypervariable domains and limiting sequencing to these regions.

Some of the collected species in the current study appeared to be mis-identified. A single species, i.e. *S. ghobban*, showed more than 99 % sequence identity with *S. perscius* indicating potential misidentification. Similarly, *L. nebulosus*, which was collected as *Lethrinus lentjan*; *Mulloidichthys flavolineatus* collected as *M. vanicolensis*; and *Plectorhinchus sordidus* collected as *P. pictus* were misidentified. Morphological similarities, particularly in species like *Mulloidichthys flavolineatusis* and *M. vanicolensis* can lead to inaccurate species identification and, accordingly, fish stock management [43]. This similarity encouraged the use of internal markers for species discrimination, such as otolith morphologies and ultrastructure [44]. However, the presence of hybrids, particularly among reef fishes, could also contribute to these identification challenges in the Yemeni and Southern Arabian waters For example, a hybrid between butterflyfish species *Chaetodon collare* Bloch, 1787 and *Chaetodon lunula* (Lacepede, 1802) was identified in the waters of Socotra island in the Gulf of Aden, where alleles of both species appeared in the nuclear genome, but only *C. lunula* sequences were the genetic signatures pronounced in the maternal, i.e., mitochdondrial, genome of that hybrid [45]. Likewise, putative hybrids of the clownfish species *Amphiprion bicinctus*, native to the Red Sea, and *Amphiprion omanensis*, native to the Arabian Sea, were recorded in the intermediate Yemeni waters of Socotra [46].

Regarding the ecological status of the misidentified species *S. persico,* it is listed as "Least Concern" (LC) in the IUCN Red List of species within the Arabian Gulf, but "Vulnerable (VU)" outside it [47]. *Scarus ghobban* is categorized as (LC) outside the Gulf, but "Endangered" within it. Proper barcoding and morphogenetic-based continuous surveillance are crucial for the conservation of these species. *Plectorhinchus sordidus* and *L. nebulosus* are both categorized as LC in the Arabian Gulf [48, 49], While, *Mulloidichthys flavolineatus* is categorized under "Data Deficient" in the Arabian Gulf [50], and *P. pictus*, however, is "Near Threatened" [51]. This study highlights the need for continued research and the importance of accurate species identification for effective conservation and fisheries management in the Red Sea and beyond.

4. Conclusion

This study underscores the significance of specific hypervariable domains within the 12SrDNA gene sequence that enhance its effectiveness as a teleost barcode with strong species discrimination power. A notable limitation was the lack of reference sequences in the GenBank database, which highlights the need to continued work in this region of the Indian Ocean waters and its related environments, mainly the Red Sea. Additionally, integrating more morpho-genetic data for the studied species could improve the management of fisheries South Asia and East Africa, especially considering the presence of hybrid species abundance in these areas.

CRediT authorship contribution statement

Asmaa Galal-Khallaf, Khaled Mohammed-Geba, Taha Baker, Abdelaziz A.A. El-Sayed, Hamdy A. M. Soliman: Design of the study. Asmaa Galal Khallaf, Aya Ibrahim Elberri, Omir AbdelWahab Rabie: DNA extraction, PCR, electrophoresis, sequences' analyses, other bioinformatic analyses. T.B., Abdelaziz A.A. El-Sayed, Hamdy A. M. Soliman Collection of Yemeni and Egyptian samples. Asmaa Galal Khallaf, Khaled Mohammed-Geba, Taha S.S. BakerAbdelaziz A.A. El-Sayed, Hamdy A. M. Soliman: preparation of figures and tables, formal analyses, and Writing, reviewing the manuscript. All authors have read and agreed to the published version of the manuscript.

Data availability statement

Data for this work are available upon demand.

Declaration of competing interest

The authors declare that they have no conflict of interest with the results of the current work.

Acknowledgments

The authors are thankful to the editor and reviewers for their valuable comments towards improving this paper.

References

[1] P. D. Hebert, A. Cywinska, S. L. Ball, & J. R. DeWaard, *Series B: Biological Sciences*, 270 (2003) 313-321.
[2] A.Teramura, K. Koeda, A. Matsuo, M. P. Sato, H. Senou, H. C. Ho, & S. Hirase, *Marine Ecology Progress Series*, 701 (2022) 83-98.
[3] A. Galal-Khallaf, A. Ardura, K. Mohammed-Geba, Y. J. Borrell & E. Garcia-Vazquez, *Food control*, 46 (2023) 441-445.
[4] L. Zang, S. Schäffer, D. Daill, T. Friedrich, W. Gessl, M. Mladinić & S. Koblmüller, *Plos one*, 17 (2022) e0268694.
[5] J. A. Bitencourt, P. R. Affonso, R. T. Ramos, H. Schneider, & I. Sampaio, *Molecular Phylogenetics and Evolution*, 178 (2023) 107631.
[6] Z. Li, P. Jiang, L. Wang, L. Liu, M. Li, & K. Zou, *China. Ecological Indicators,* 147 (2023) 109915.
[7] E. Decru, N.Vranken, H. Maetens, A. Mertens De Vry, A. Kayenbergh, J. Snoeks, & M. Van Steenberge, *Hydrobiologia,* 849 (2022) 1743-1762.
[8] K. Mohammed-Geba, E. M. Abbas, H. O. Ahmed, M. A. Shalabi, E. S. A. Hamed, F. A. A. Razek, & T. Soliman, *Zootaxa*, 5092 (2022) 559-575.
[9] G. Riccioni, I. Domaizon, A. Gandolfi, M. Pindo, A. Boscaini, M. Vautier & J. Wanzenböck, *Advances in Oceanography and Limnology,* 13 (2022) 10017.
[10] N. L. Lira, S. Tonello, R. L. Lui, J. B. Traldi, H. Brandão, C. Oliveira, & D. R. Blanco, *Molecular Biology Reports,* 50 (2023) 1713-1726.
[11] D. Tesfamichael, P. Rossing, & H. Saeed,. *Fisheries Centre Research Reports*, 20 (2012) 105-152.
[12] S., Maiyza, S. F. Mehanna, & I. A. El-karyoney, *Egyptian Journal of Aquatic Biology and Fisheries,* 24 (2020) 441-452.
[13] A. M. Al-Fareh*, LSE Middle East Centre Report,* (2018) 1-36.
[14] M. Kabil, AbdAlmoity, E. A. K. Csobán & L. D. Dávid*, Plos one,* 17 (2022) e0268047.
[15] M. M. Al-Rshaidat, A. Snider, S. Rosebraugh, A. M. Devine, T. D. Devine, L. Plaisance & M. Leray, *Genome,* 59 (2016) 724-737.
[16] A. Galal-Khallaf, A. G. Osman, A. El-Ganainy, M. M. Farrag, E. Mohammed-AbdAllah, M. A. Moustafa & K. E. Mohammed-Geba, *Food control*, 46 (2023) 441-445.
[17] L. Rabaoui, L.Yacoubi, D. Sanna, M. Casu, F. Scarpa, Y. J. Lin, & M. A. Qurban, *Journal of Fish Biology*, 95 (2019) 1286-1297.
[18] E. N. Rachmilovitz, O. Shabbat, M. Yerushalmy & B. Rinkevich, *Journal of Marine Science and Engineering*, 10 (2022) 1917.
[19] M. A. M. Afifi, M. Sarhan, H. M. Khalaf-Allah, A. N. Alabssawy, M. M. M. Abbas, F. Abdel-Hamid, & M. A. M. El-Tabakh, *Zoologischer Anzeiger,* 304 (2023) 84-93.
[20] A. Estoup, C.R. Largiader, E. Perrot, D. Chourrout, *Mol. Mar. Biol. Biotechnol*. 5 (1996) 295-298.
[21] J.N.Wolff, N.J Gemmell, *Biol. Reprod*., 79 (2008) 247e252.
[22] S.R. Palumbi, Molecular Systematic, 2nd ed. Sinauer Associates Inc., USA, 1996
[23] R. D. Ward, T.S. Zemlak, B.H. Innes, P.R. Last, P.D. Hebert, *Philos. Trans. R. Soc*., 360 (2005) 1847-1857.
[24] B. A. Sweeney, D. Hoksza, E. P. Nawrocki, C. E. Ribas, F. Madeira, J. J. Cannone, A. I. Petrov, *Nature communications*, 12 (2021) 1-12.
[25] D .Darriba, G. L. Taboada, R. Doallo, D. Posada, *Nat Methods* 9 (2012) 772.
[26] T. H. Jukes, C. R. Cantor, *New York Academic Press*, (1969) 21-132.
[27] A. N. Egan, K. A. Crandall, C. W. Fox & J. B. Wolf. *Evolutionary Genetics: Concepts and Case Studies*, 1 (2006) 426-436.
[28] M. Kimura, *J Mol Evol*, 16 (1980) 111-120.
[29] L. Zheng, J. He, Y. Lin, W. Cao, & W. Zhang, *Acta Oceanologica Sinica*, 33 (2014) 55-76.
[30] E. Quéméré, M. Aucourd, V. Troispoux, S. Brosse, J. Murienne, R. Covain, M. Galan, *Environmental DNA*, 3 (2021) 889-900.
[31] RNAcentral 2021, *Nucleic acids research,* 49 (2021) D212-D220.
[32] G. Kumar, A. M. Reaume, E. Farrell, & M. R. Gaither, *PloS one,* 17 (2022) e0266720.
[33] R. Ragot & R. Villemur, *Environmental Monitoring and Assessment*, 194 (2022) 1-13.
[34] R. Chakroun, S. Abdellatif, & T. Villemur, *Internet of Things*, 19 (2022) 100510.
[35] A. Galal-Khallaf, A. Ardura, K. Mohammed-Geba, Y. J. Borrell, & E. Garcia-Vazquez*, Food control*, 46 (2017) 441-445.
[36] I. Bakke & S. Johansen, *Molecular Phylogenetics and Evolution*, 25 (2002) 87-100.
[37] B. Yang, Y. Wang, & P. Y. Qian, *BMC Bioinformatics*, 17 (2016) 1-8.
[38] H. K. Allen, D. O. Bayles, T. Looft, J. Trachsel, B. E. Bass, D. P. Alt & T. A. Casey, *BMC Research Notes*, 9 (2016) 1-6.
[39] Y. Zhao, W. Y. Zhang, R. L. Wang, & D. L. Niu, *Parasites & vectors*, 13 (2020) 1-12.
[40] S. Schmidt, F. Driver & P. De Barro, *Organisms Diversity & Evolution*, 6 (2006) 127-139.
[41] I. Bakke & S. Johansen, *Molecular Phylogenetics and Evolution*, 25 (2002) 87-100.
[42] Y. M. Hung, W. N. Lyu, M. L. Tsai, C. L. Liu, L. C. Lai, M. H. Tsai, & E. Y. Chuang, *Computers in Biology and Medicine*, 145 (2022) 105416.
[43] F. Uiblein, *Smithiana Bulletin*, 13 (2011) 51-73
[44] A. G. M. Osman, M. M. Farrag, *Iranian Journal of Fisheries Sciences*, 19 (2020) 814-832.
[45] J. D. DiBattista, L. A. Rocha, J. P. A. Hobbs, S. He, M. A. Priest, T. H. Sinclair-Taylor & M. L. Berumen, *Journal of Biogeography*, 42 (2015) 1601-1614.
[46] P. Saenz-Agudelo, J. D. Dibattista, M. J. Piatek, M. R. Gaither, H. B. Harrison, G. B. Nanninga, & M. L. Berumen, *Molecular Ecology*, 24 (2015) 6241-6255.
[47] J.H. Choat, *The IUCN Red List of Threatened Species,* (2015) e.T190765A57137283.
[48] S. Alam, J. Bishop, B. Russell, S. Hartmann, Y. Iwatsuki, K.E. Carpenter, E. Abdulqader, F. Kaymaram, K. Al-Khalaf, Jassim Kawari, & Q. Alghawzi, *The IUCN Red List* of *Threatened Species,* (2015) e.T194435A57127795.
[49] B. Russell, Y. Iwatsuki, K. E. Carpenter & S. Hartmann, *The IUCN Red List of Threatened Species,* (2015) e.T16720181A57143372.
[50] F. Kaymaram, E. Abdulqader, S. Hartmann, M. Almukhtar, M. Al-Husaini, S. Alam, & Q. Alghawzi, *The IUCN Red List of Threatened Species*, (2015) e.T50903119A57159328.
[51] Y. Iwatsuki, S. Hartmann, K.E. Carpenter, B. Russell, E. Abdulqader, J. Bishop, F. Kaymaram, S. Alam, K. Al-Khalaf, A. Jassim Kawari, & Q. Alghawzi, *The IUCN Red List of Threatened Species,* (2015) e.T46086124A57127737.