Zoological Studies **60:**76 (2021)
doi:10.6620/ZS.2021.60-76

Zoological Studies

Open Access

Metabarcoding of Fish Larvae in the Merbok River Reveals Species Diversity and Distribution Along its Mangrove Environment

Norli Fauzani Mohd Abu Hassan Alshari[1], Siti Zuliana Ahmad[1], Azali Azlan[1], Youn-Ho Lee[3], Ghows Azzam[1,*], and Siti Azizah Mohd Nor[1,2]

[1]*School of Biological Sciences, Universiti Sains Malaysia, 11800, Penang, Malaysia. *Correspondence: E-mail: ghows@usm.my (Azzam) E-mail: nfauzani0693@gmail.com (Mohd Abu Hassan Alshari); liazuliana11@gmail.com (Ahmad); azlanazali01@gmail.com (Azlan); sazizah@usm.my (Nor)*
[2]*Institute of Marine Biotechnology, Universiti Malaysia Terengganu, 21030 Kuala Terengganu, Terengganu, Malaysia. E-mail: s.azizah@umt.edu.my (Nor)*
[3]*Korean Institute of Ocean Science and Technology, Republic of Korea. E-mail: ylee@kiost.ac.kr (Lee)*

Received 24 November 2020 / Accepted 11 October 2021 / Published 22 December 2021
Communicated by Ryuji Machida

The Merbok River (north-west of Peninsular Malaysia) is a mangrove estuary that provides habitat for over 100 species of fish, which are economically and ecologically important. Threats such as habitat loss and overfishing are becoming a great concern for fisheries conservation and management. The identification of larval fish in this estuarine system is important to complement information on the adults. This is because the data could inform the spawning behaviour, reproductive biology, selection of nursery grounds and migration route of fish. Such information is invaluable for fisheries and aquatic environmental monitoring, and thus for their conservation and management. However, identifying fish larvae is a challenging task based only on morphology and even traditional DNA barcoding. To address this, DNA metabarcoding was utilised to detect the diversity of fish in the Merbok River. To complete the study, the fish larvae were collected at six sampling sites of the river. The extracted larval DNA was amplified for the Cytochrome Oxidase subunit 1 (*COI*) and 12S ribosomal RNA (12S rRNA) genes based on the metabarcoding approach using shotgun sequencing on the next-generation sequencing (NGS) Illumina MiSeq platform. Eighty-nine species from 65 genera and 41 families were detected, with *Oryzias javanicus*, *Oryzias dancena*, *Lutjanus argentimaculatus* and *Lutjanus malabaricus* among the most common species. The lower diversity observed from previous morphological studies is suggested to be mainly due to seasonal variation over the sampling period between the two methods and limited 12S rRNA sequences in current databases. The metabarcode data and a validation Sanger sequencing step using 15 species-specific primer pairs detected three species in common: *Oryzias javanicus*, *Decapterus maruadsi* and *Pennahia macrocephalus*. Several discrepancies observed between the two molecular approaches could be attributed to contaminants during sampling and DNA extraction, which could mask the presence of target species, especially when DNA from the contaminants is more abundant than the target organisms. In conclusion, this rapid and cost-effective identification method using DNA metabarcoding allowed the detection of numerous fish species from bulk larval samples in the Merbok River. This method can be applied to other sites and other organisms of interest.

Key words: Fish larvae, Mangrove estuary, Merbok River, DNA metabarcoding, Next-generation sequencing.

Citation: Alshari NFMAH, Ahmad SZ, Azlan A, Lee YH, Azzam G, Nor SAM. 2021. Metabarcoding of fish larvae in the Merbok River reveals species diversity and distribution along its mangrove environment. Zool Stud **60:**76. doi:10.6620/ZS.2021.60-76.

BACKGROUND

In their various stages of life cycles, fish communities provide valuable insights into the ecological conditions of their habitats and furnish information to manage fishery resources (Moser and Smith 1993; Moser 1996; Kidwai and Amjad 2001). However, a prerequisite for such investigations is their precise identification. Acknowledging the shortcomings of traditional approach for species identification, molecular techniques are increasingly used to facilitate the identification process (Lewis et al. 2016). The DNA barcoding approach introduced by Hebert et al. (2003) based on species variation in the mitochondrial cytochrome oxidase subunit 1 (*COI*) gene is regarded as the gold standard for molecular identification. It has been widely successful in discriminating most animal specimens to the species level, including identifying fish species, whether whole or using specific parts of an individual (Ko et al. 2013; Lewis et al. 2016; Azmir et al. 2017; Collet et al. 2018). However, sorting and identifying minute larval ichthyoplankton specimens needed for individual-based DNA barcoding is time- and cost-consuming.

DNA metabarcoding, which applies the next generation sequencing (NGS) approach, is a rapid and cost-effective approach to processing bulk samples, damaged and fragmented specimens, and possibly degraded DNA (*e.g.*, ichthyoplankton, soil, water, and feces) (Taberlet et al. 2012) for biodiversity assessment and ecological studies (Coissac et al. 2012; Cristescu 2014; Lobo et al. 2017). DNA metabarcoding could provide an accurate taxonomic and biodiversity assessment of organisms in their native habitats, which are critical for their management. Based on this technique, several studies focussing on bulk ichthyoplankton specimens have successfully assigned ichthyoplankton to the species level (Maggia et al. 2017; Mariac et al. 2018; Nobile et al. 2019; Ratcliffe et al. 2021).

Considering the threats of overharvesting and habitat degradation, more active and stringent steps must be taken to manage areas to support sustainable fisheries for the local community. While regulations are in place to manage the adult fishes, nothing is known on the diversity and distribution of larvae. This information is vital for fisheries managers to understand the species utilizing the area and the locations they inhabit as their nursery grounds. With this knowledge, fisheries managers can take measures to protect the specific sites. Thus, to complement the management efforts on the adult fishes, more comprehensive and holistic management strategies can be implemented through this study using the DNA metabarcoding method.

This study investigates the diversity and distribution of fish larvae in a mangrove estuarine area in the northern part of Peninsular Malaysia known as the Merbok River. The main river connects small rivers or tributaries within the Merbok Permanent Forest Reserve (MPFR). Facing the Strait of Malacca, this ca. 4000 hectare mangrove area is recognised as one of the world's mangrove species diversity hotspots, harbouring more than half of the global species (Mazlan et al. 2005). The Merbok River, similar to other mangrove estuarine areas, is an important ecosystem for fisheries resources, in addition to its highly diverse natural floral resources (Jusoff and Taha 2008). Previous studies of the Merbok River have recorded a combined total of 120 fish species through morphological identification of the adult specimens (Mansor et al. 2012a b). The 35 km Merbok River that runs through a gradient of freshwater in the upper reaches to the more saline coastal waters flows through agricultural, aquaculture and residential areas. The land conversion in the MPFR area for these activities, including the infrastructure development, could negatively impact the faunal and floral communities that occupy the mangrove ecosystems, such as reducing catch from fisheries (Manson et al. 2005; Jusoff and Taha 2008). In addition, based on this study, we hypothesise that the diversity and abundance of fish larvae is higher in the coastal lower reaches of the river than in the upper reaches.

MATERIALS AND METHODS

Study area

The fish larvae samples were collected from a mangrove estuary in the Merbok Permanent Forest Reserve (MPFR) in northwestern Peninsular Malaysia. The estuary is locally known as Merbok River. It lies between latitude 100°20'57.33" and longitude 5°40'53.74" facing the Straits of Malacca and between latitude 100°30'24.56" and longitude 5°42'13.46" at the upper reaches (Mansor et al. 2012b). Small tributaries connect the 35 km estuary with freshwater discharged into the estuary from small tributaries, especially at the upper part of the river. The Merbok River has high salinity along the lower zone and decreases up the river, the former due to its proximity to the coastal area, while the upper zone has freshwater inflow. The Merbok River is surrounded by 39 true mangrove species (Ong et al. 2015) dominated by *Rhizophora apiculata* and *Bruguiera parviflora* along the 35 km stretch of the main river (Mansor et al. 2012a). The upper zone of the river is surrounded by mangrove forests near residential areas, fishing villages, agricultural fields, shrimp

and oyster farms. The middle zone is surrounded by mangrove forests and some fish aquaculture activities. The lower zone is surrounded by mangrove forests, palm oil plantations, and its shrimp and fish facilities and land development activities for tourist attractions. Thus, the whole area is anthropogenically important due to its high mangrove diversity.

Sample collection and preservation

Fish larvae samples were collected during the primary wet season in August 2016 at six sampling locations along the tributaries of the Merbok River (Fig. 1). The sampling localities were from the freshwater upper zone (St1: Lalang River and St2: Semeling River), middle (St3.1: Keluang River and St3.2: Teluk Wang River), and lower brackish/marine zone (St4.1: Gelam River and St4.2: Terus River), in concordance to the sites of previous studies which divided the sampling sites according to these three zonations (Mansor et al. 2012b; Fatema et al. 2014). This allowed us to compare biodiversity assessments among the studies based on different approaches (morphological vs. metabarcoding). Furthermore, the decreasing salinity gradient from the upper to lower zone provides an excellent insight into larval diversity based on their salinity tolerance and nursing grounds. The collection of fish larvae samples was standardized by scooping five times in the same (or approximately) spot near the mangrove roots at the riverbank area by using a modified hand scoop net of 500 µm mesh size (radius: 30 cm) (Arshad et al. 2012; Wibowo and Sloterdijk 2015). The samples were kept in separate 50 mL bottles filled with water from the sampling sites and kept cool on ice during transport to the Molecular Ecology Research Laboratory, Universiti Sains Malaysia (USM), Penang. The filtered samples were then rinsed in distilled water and pooled in five replicate tubes for each site filled with 70% ethanol prior to the DNA extraction process.

Water parameters were recorded for each sampling site to assess the habitat type (FishBase category) at the point of sampling. The water parameters were measured using the following equipment: Secchi disk and tape were used to measure water depth (WD), and turbidity (TURB), SCT Meter YSI Model 33 (YSI Inc., USA) was used to measure water temperature (TEMP) and salinity (SAL), while YSI 550A (YSI Inc., USA) was used to measure water pH and dissolved oxygen (DO). No ecological analysis was intended as this was a one-off sampling measurement.

DNA Barcoding referencing of fish species

Specimens of 22 adult fish species without available molecular sequences of the 12S rRNA gene in the public databases were obtained from local wet markets for analysis. Samples were identified based on the FAO species identification guide book (Carpenter and Niem 2001). Ikan Laut Malaysia (Atan et al. 2010) and Fishes of Malaysia (Ambak et al. 2012). Each specimen was photographed, and whole specimens were permanently stored in 70% ethanol in the Molecular Ecology Research Laboratory, USM. The pectoral fin clips of each species (one to three specimens) were

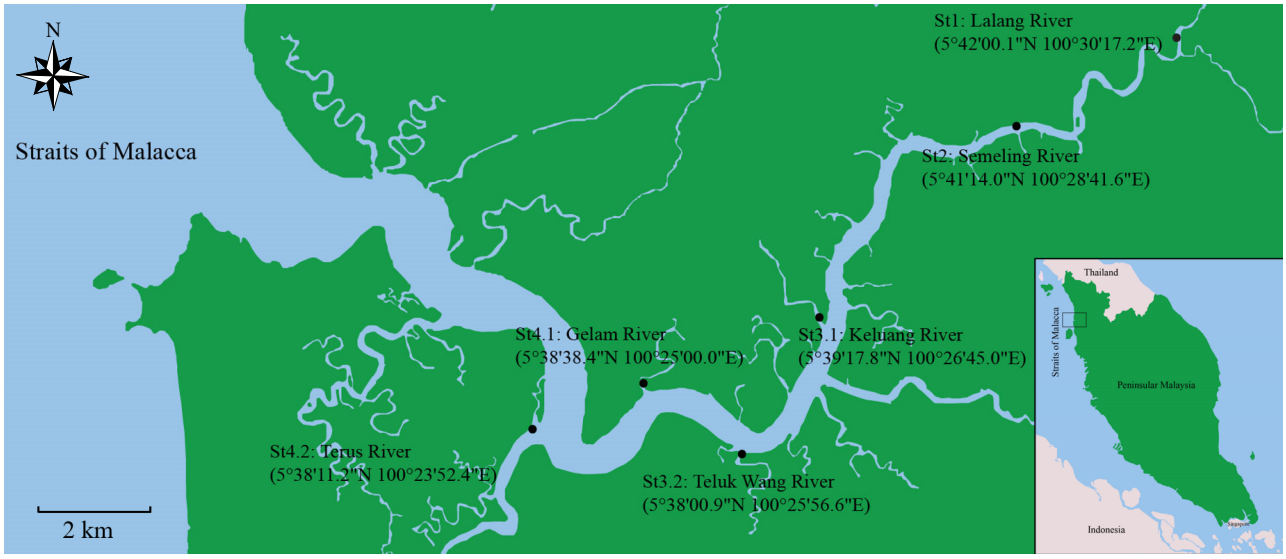

Fig. 1. Merbok River with six sampling stations, divided into three zones: upper [St1: Lalang River (5°42'00.1"N 100°30'17.2"E), St2: Semeling River (5°41'14.0"N 100°28'41.6"E)], middle [St3.1: Keluang River (5°39'17.8"N 100°26'45.0"E) and St3.2: Teluk Wang River (5°38'00.9"N 100°25'56.6"E)] and lower [St4.1: Gelam River (5°38'38.4"N 100°25'00.0"E) and St4.2: Terus River (5°38'11.2"N 100°23'52.4"E)].

preserved in 96% ethanol for molecular identification. The combined report from Mansor et al. (2012a b c) recorded a total of 120 morphologically identified adult fish species in the Merbok River, of which 68 species (Mansor et al. 2012b) were classified according to their habitat category estuarine (E), marine (M), marine-estuarine dependent (MED), freshwater-estuarine dependent (FED) and freshwater (F) while the remaining 52 species had not been previously classified. The genomic DNA of adult specimens (22 species without 12S rRNA reference sequences) was extracted using the modified hexadecyltrimethylammonium bromide (CTAB) protocol (Grewe et al. 1993) from approximately 1.0 mm of the preserved fin clip. A segment of the 12S rRNA gene was amplified from the extracted DNA using the primer pairs MiFish-U-F 5'-GTC GGT AAA ACT CGT GCC AGC-3' and MiFish-U-R 5'-CAT AGT GGG GTA TCT AAT CCC AGT TTG-3' (Miya et al. 2015). The 25 µL PCR reaction mix contained 2.5 µL of 10X $MgCl_2$ free PCR buffer, 2.0 µL of 50mM $MgCl_2$ 1.0 µL of 10mM dNTP, 0.5 µL of each 5µM forward and 5µM reverse primers, 0.25 µL of 5U/µL *Taq* polymerase (iNtRON, Gyeonggido, Korea), 1.0 µL of DNA template and 16.75 µL of double-distilled water. The thermal conditions were: a pre-denaturation step of 2 minutes at 95°C; followed by 35 cycles of 20 seconds at 94°C; 15 seconds at 47.9°C and 15 seconds at 72°C; followed by a final extension of 5 minutes at 72°C and then stored at 4°C. Sequncing of the PCR products was done at the First Base Laboratories Sdn. Bhd. (Selangor, Malaysia) on an ABI3730XL capillary sequencer (Applied Biosystems, USA). To aid molecular confirmation of each species, samples from the same specimens were also analysed with the *COI* gene based on the following primer pair: FishF1 5'-TCA ACC AAC CAC AAA GAC ATT GGC AC-3' and FishR1 5'-TAG ACT TCT GGG TGG CCA AAG AAT CA-3' (Ward et al. 2005). The thermal conditions were: a pre-denaturation step of 4 minutes at 95°C; followed by 35 cycles of 30 seconds at 94°C, 50 seconds at 47.9°C and 1 minute at 72°C; followed by a final extension of 7 minutes at 72°C and then stored at 4°C. The sequencing protocol was the same as for the 12S rRNA gene.

Forward and reverse sequences were trimmed and aligned using MEGA7 software (Kumar et al. 2016). The *COI* sequences were then compared to the Barcoding of Life Database (BOLD) System. Its comprehensive features including morphological information (photographic record) and other supporting data for species identification permit effective cross referencing to the 12S rRNA gene sequence for the same sample (and species). The newly generated 12S rRNA gene sequence of each species was submitted to NCBI (GenBank) (https://www.ncbi.nlm.nih.gov) under accession numbers KY379960-KY379968, KY778751-KY778754, MG729393, MG729396, MG729397, MG748713, MG748714, MK330865-MK330867.

DNA metabarcoding

Genomic DNA extraction and amplification

The genomic DNA extraction of the larval specimens was conducted following the protocol of the adult specimens with some modifications: 1) the larval specimens that were preserved in five replicate tubes for each location containing 70% ethanol were first cut and minced; 2) then, the minced samples were pooled into six separate labelled 1.5 mL microcentrifuge tubes based on sampling stations (St1, St2, St3.1, St3.2, St4.1, St4.2). The number of individuals varied among sites, but as earlier mentioned, the volume was standardised for all sites by maintaining a uniform number of scoops (5X). The extracted DNA was purified using Wizard® SV Gel and PCR Clean-Up System kit (Promega, USA) following the manufacturer's instruction to remove excess inhibitor that could potentially inhibit the amplification of mitochondrial DNA (mtDNA). The purity and quantity of the extracted and purified DNA were measured using UV spectrophotometer Q3000 (Quawell, USA) before and after purification. The mitochondrial genome amplification and enrichment step were then conducted on the purified DNA of each pooled sample extract using REPLI-g Mitochondrial DNA kit (Qiagen, USA) following the provided protocol. The amplification of the whole mitochondrial genome was aimed to get complete mitogenomes of almost all fish species in one shot. This is to reduce the cost for sequencing and analysis compared to individual mitogenomes. After the amplification steps, samples St3.1 and St3.2 were pooled and was labelled as sample St3. At the same time, samples St4.1 and St4.2 were also pooled and labelled as sample St4 for the library preparation step in the Illumina MiSeq NGS platform (refer to Results for pooling clarification). Successfully amplified samples were sent for pre-processing and next-generation sequencing at the Shanghai Majorbio Pharmaceutical Technology Co., Ltd. (Shanghai, China).

Library preparation and sequencing

The NGS shotgun sequencing was conducted on an Illumina MiSeq (Illumina, San Diego, USA) with paired-end 250 bp insert size. The library preparation was done to add adapter sequences onto the ends of the DNA fragments. The steps involved in library

preparation were; 1) fragmentation, 2) end-repair, 3) A-tailing, 4) ligation and 5) paired-end sequencing. Firstly, DNA samples were sheared into approximately 400 to 500 bp fragments using an ultrasonicator, Covaris M220 (https://covaris.com/products/afa-ultrasonication/m-series/). Then, the sheared DNA fragments were purified using QIAquick PCR Purification Kit (Qiagen, Germany). The fragmented DNA was then end-repaired, and the 5'-end were phosphorylated. Next, the blunt 3'-ends were A-tailed by adding an adenine (A) base to form an overhang. During the A-tailing, the overhang A-tail allows adapters containing thymine (T) base to pair with the DNA fragments. The A-tailing of the 3'-ends is important to facilitate ligation of the DNA template to the sequencing adapters. The ligase enzyme covalently links the adapter and DNA fragments during adapter-fragment ligation. The ligated DNA products were then PCR amplified using TruSeq™ DNA Sample Prep Kit (Illumina, California, USA) to enrich the DNA ligation products. Finally, the genomic DNA library was assessed by electrophoresis, nanodrop and qubit as a part of the library assurance (QA) and quality control (QC) procedures. The genomic library with satisfactory QA and QC was continued with the cluster generation and sequencing. NGS data pre-processing steps were conducted on each raw sequence read which involved quality control procedures to filter sequence reads with low-quality and remove of the adapter sequences prior to analysis of sequence reads of each sample. All the above procedures from library preparation to sequencing (1–5) and NGS data pre-processing were conducted at the Shanghai Majorbio Pharmaceutical Technology Co., Ltd. (Shanghai, China).

Bioinformatics procedure

The data generated from the shotgun sequencing were then analysed using several bioinformatics software and run in the Linux platform. Quality analysis of the MiSeq reads was done using FastQC available from https://www.bioinformatics.babraham.ac.uk/projects/fastqc/. Adapters and low-quality reads were filtered and trimmed using Trimmomatic (Bolger et al. 2014) using the following parameters: ILLUMINCLIP (to perform adapter removal): TruSeq2-PE.fa:2:30:10; LEADING (to cut bases at the start of the read):3; TRAILING (to cut bases at the end of the read):3; SLIDINGWINDOW (to perform sliding window trimming):4:28; MINLEN (the minimum length specified to cut the reads):100. The clean paired-end reads obtained after quality trimming with an average length of 100 to 250 bp and average GC content of 44% to 45% proceeded to be *de novo* assembled for scaffold formation. Following the default parameter settings, the *de novo* assembly was done using MEGAHIT (v.1.0.2) assembler software (Li et al. 2015). The parameters used were: i) the min-count: 2; ii) k-min: 21; iii) k-max: 99; iv) k-step: 20; and v) min-contig-len: 200 (Table S1).

The assembled scaffolds were divided into taxonomic classified reads and taxonomic unclassified reads using Kraken 2 software (Wood et al. 2019). The reads with taxonomic classification were further blast on a mitochondrial genome reference database of *COI* and 12S rRNA genes (RefSeq) of 35,655 current fish sequences downloaded from NCBI (GenBank) in the FASTA file format for BLAST analysis with scaffolds of each sample. The BLAST analysis on *COI* and 12S rRNA gene reference databases was performed by using 'megablast' using several criteria (blast identity: ≥ 97% (Mariac et al. 2018; Fujii et al. 2019), word size: 28, e-value: 0.0001) for species-level assignment and diversity analysis. The scaffolds were realigned with the sequences of the identified species and reference sequence of 120 fish species from Merbok River to confirm the annotation and taxonomic classification. Only scaffolds with ≥ 97% similarity with the reference sequences were assigned to species.

Metabarcoding results verification

Species-specific primer design

To verify the metabarcoding results of fish larvae identification, species-specific primer pairs were developed for 15 fish species randomly selected based on the DNA metabarcoding results (Table S2). These primer pairs targeted the *COI* gene region because of its well-developed reference database in both the NCBI and BOLD systems compared to other genes. The sequences of these 15 species were downloaded from the two databases, and primer development was conducted through an online tool, Primer3Plus (http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi) (Untergasser et al. 2007). All primer pairs were designed following the standard criteria for primer design, such as the primer length (18 to 22 bp), product size (200 to 630 bp), GC content (45% to 65%), and melting temperature (Tm: 50°C to 65°C).

PCR amplification of larval samples using newly designed primer pairs

PCR amplification was conducted on the pooled genomic DNA of the four sampling stations (St1, St2, St3, and St4) using the 15 newly designed species-specific primers. The 25 μL PCR reaction contained 16.75 μL of double-distilled water, 2.5 μL of 10×PCR

buffer, 2.0 µL of $MgCl_2$, 1.0 µL of dNTP, 0.5 µL of each forward and reverse primer and 0.25 µL of *Taq* polymerase (iNtRON, Gyeonggido, Korea) and 1.0 µL DNA template of pooled samples. The same PCR conditions were applied for each primer pair: pre-denaturation step of 2 minutes at 95°C; 35 cycles of 30 seconds at 94°C, 30 seconds at 45°C to 55°C and 50 seconds at 72°C; final extension step of 10 minutes at 72°C and stored at 4°C. Successfully amplified PCR products were sent to First BASE Laboratories Sdn. Bhd. for Sanger-sequencing on the ABI3730XL sequencer (Applied Biosystem, USA).

Diversity analysis

The read abundance of fish larvae was used to analyse the diversity within the four stations (alpha diversity). The data were tabulated with the size bins combined across the samples, square root-transformed by measuring diversity indices (*i.e.*, Shannon, Margalef, Menhinick, Evenness, and Equitability). For larval fish diversity among stations, Bray-Curtis similarity was conducted on the relative abundance to assess and visualise the Merbok River's beta diversity, displayed through a two-dimensional nonmetric-multidimensional scaling (NMDS) ordination based on their similarity (%). The alpha and beta diversity analyses were conducted using PRIMER7 and PERMANOVA+ (version 7; Primer-E, Ivybridge, UK).

RESULTS

General water condition of the Merbok River

As only a single measurement was taken, the water quality assessment was only a snapshot of the general water conditions and was used to classify the stations into habitat types (freshwater, estuarine, marine or combinations of these). Based on salinity, the stations were classified as mesohaline (salinity range 5.0–17.9 ppt) in the upper zone (St1 and St2) and polyhaline (salinity range: 18.0–29.0) in the middle and lowest zones (St3.1, St3.2, and St4.1, St4.2). St3.1 and St3.2 were combined and renamed St3, and similarly St4.1, St4.2 were also combined and renamed St4. The pooling was done with the potential for capturing higher diversity and considering of the relatively short distance within the combined sets and their similar water quality characteristics. Water parameters were recorded for each sampling site (Table 1); water depth, turbidity, temperature, salinity, pH, and dissolved oxygen.

Fish larvae assignment and diversity based on the metabarcoding method

The Illumina MiSeq platform sequencer generated 3,123,982, 2,668,052, 2,388,913 and 2,566,647 paired-end raw reads from each of the four samples; St1, St2, St3 and St4, respectively. After sequence quality trimming, the final paired-end reads were 1,400,112, 1,581,822, 1,422,667 and 1,642,143 for sample St1, St2, St3 and St4, respectively. These high-quality and cleaned reads were assembled into a total of 1,939 scaffolds, 3,486 scaffolds, 1,900 scaffolds and 1,932 scaffolds for samples St1, St2, St3, and St4, respectively. The *de novo* assembly analysis revealed a minimum scaffold length of 200 bp, maximum scaffold lengths of 6,419 bp to 6,753 bp and average scaffold length of 610 bp to 758 bp (Table S1).

The BLAST analysis annotated a total of 1,658 (18%) and 1,367 (15%) scaffolds to the *COI* and 12S rRNA genes, respectively. Scaffolds annotated to *COI* and 12S rRNA were further used in the BLAST analysis for taxonomic assignment of the fish larvae with an acceptable limit of blast identity at ≥ 97%. The

Table 1. The environmental parameters of the Merbok: water depth, turbidity, salinity, pH, temperature, and dissolved oxygen during the time of sampling

Parameters	Locations					
	Lalang River (St1) 5°42'00.1"N 100°30'17.2"E	Semeling River (St2) 5°41'14.0"N 100°28'41.6"E	Keluang River (St3) 5°39'17.8"N 100°26'45.0"E	Teluk Wang (St4) 5°38'00.9"N 100°25'56.6"E	Gelam River (St5) 5°38'38.4"N 100°25'00.0"E	Terus River (St6) 5°38'11.2"N 100°23'52.4"E
Water depth (cm)	38.5	65.3	124.0	125.3	98.6	113.5
Turbidity (cm)	38.5	65.3	124.0	124.3	88.2	92.5
Salinity (ppt)	10	10.3	19	22	23	24
pH	5.8	5.8	5.9	6.3	6.3	6.4
Temperature (°C)	27.3	28.8	30.6	31.3	31.2	31.1
Dissolved oxygen (mg/L)	4.40	4.98	6.70	6.50	7.15	7.77

combined results of BLAST analysis of 2,014 scaffolds annotated to *COI* (1,071 scaffolds), and 12S rRNA (943 scaffolds) genes (Table S3) revealed a total of 89 species, 65 genera, and 41 families in the Merbok River. Among these species, 88 species were identified by the *COI* gene, while the 12S rRNA gene identified 78 species. Although this study standardized the sampling replicates for each site and standardized pooling of the DNA samples for amplification and NGS, a low annotation rate still occurred after the assembly. The total amount of mitochondrial DNA in the samples is unknown and uneven for pooled taxa, together with the presence of nuclear DNA from the larvae samples and non-target DNA (contaminants) that may be present in the samples such as from the gut of the larvae. This could affect the proportion of mtDNA in the total DNA extracts and the annotation (Tang et al. 2014).

Species detection through metabarcoding of larval fish (89 species) was lower than previously recorded morphologically identified adult species (120 species). The number of species detected by the metabarcoding approach were: 12 (St1), 26 (St2), 46 (St3), and 76 (St4). Six species were detected at all stations: *Oryzias javanicus*, *Oryzias dancena*, *Oreochromis niloticus*, *Oreochromis aureus*, *Lutjanus malabaricus*, and *Siganus fuscescens*. In terms of habitat category, the first four of these common species are freshwater-estuarine (FE), while *Lutjanus malabaricus* and *Siganus fuscescens* are marine-estuarine (ME) species.

Another six species were recorded at three of the four locations. Among these, one species was detected in St1, St2 and St3: *Oryzias melastigma* (FE). In comparison, the other five species were detected in St2, St3, and St4: *Netuma thalassina* (MFE), *Alepes djedaba* (marine habitat, M), *Lutjanus argentimaculatus* (MFE), *Pennahia pawak* (M) and *Terapon jarbua* (MFE). A much higher number, 40 species, were detected at two of the four sampling stations. *Osphronemus goramy* (F) was detected at St1 and St2 only. Four species were detected in St2 and St3: *Ambassis gymnocephalus* (MFE), *Elops hawaiensis* (MFE), *Clarias batrachus* (F), and *Mastacembelus erythrotaenia* (F), while six species were detected in St2 and St4: *Mystus cavasius* (FE), *Mystus vittatus* (FE), *Gerres oyena* (ME), *Hyporhamphus quoyi* (MFE), *Lutjanus johnii* (ME), and *Johnius carouna* (MFE). The rest (33) of the two-site species were detected in St3 and St4 only, these two sites being nearest to the coast.

Thirty-seven species were site specific, detected in only a single sampling station. Four species were only detected at St1: *Brachygobius xanthomelas* (F), *Traypauchen vagina* (ME), *Trichogaster pectoralis* (F), and *Monopterus albus* (FE). Two species were site-specific to St2: *Macrognathus aculeatus* (FE) and *Liza planiceps* (MFE). One species was detected only at St3: *Pennahia argentata* (M) habitat species. The remaining 30 species were detected only at St4. The larvae occurrence generally parallel the expected habitat with related freshwater species at the upper stations and marine related ones at the lower stations. However, many species were also common in several stations which is not unexpected as a considerable number of the recorded species are multi-habitat tolerant according to FishBase. Details on the occurrence of species at the sampling stations and habitat category, as detected by *COI*/12S rRNA gene, are shown in table 2. The relative abundance of fish larvae among sampling sites is shown in figure 2.

Detection of non-target species

This study detected non-target species from the remaining 7,243 scaffolds reads that were not taxonomically classified as fish species (Fig. 3). Most of the reads are classified as bacteria (5,021 reads) known as fish-associated bacteria (from the phyla Proteobacteria, Actinobacteria, Firmicutes and Bacteriodetes), reads that taxonomically remained unassigned (1642 reads), other eukaryote (507 reads) (*e.g.*, shrimps and molluscs), and archaea (73 reads).

Comparison of larval fish diversity among four stations along the Merbok River

The beta diversity of the fish larvae among different sampling sites and different genes was compared using Bray-Curtis similarity plotted in the two-dimensional non-metric multidimensional scaling (NMDS) (Fig. 4). As expected, the *COI* and 12S rRNA genes were clustered together according to each station. Based on the NMDS, two major clusters with 36% similarity were formed; St1 and St2 were grouped in a cluster with 52% similarity, while St3 and St4 were grouped in a cluster with 62% similarity. In the St1 and St2 clusters, two clusters formed show species diversity identified using *COI* and 12S rRNA genes with 89% similarity between both genes in St1 and 90% similarity between both genes from St2. In the St3 and St4 clusters, the clustering was similar to the St1 and St2. The *COI* and 12S rRNA genes were grouped in a cluster with 94.9% similarity, while the *COI* and 12S rRNA genes in St4 were clustered with 95% similarity (Fig. 4).

Validation of larval fish species

Only nine of the 15 newly designed *COI* primers were successfully amplified. These primers detected five species and, unexpectedly, also a shrimp species.

Of these, only three of the five species in this validation step were detected in the DNA metabarcoding analysis. Surprisingly, none of the primers were specifically designed for these three species. The species detected were *Oryzias javanicus* (99%), *Decapterus maruadsi* (98%), and *Pennahia macrocephalus* (96%). The other two species, *Ambassis marianus* (99%) and an unknown species with the closest match to *Carangoides chrysophrys* at 83%, were not detected in the DNA metabarcoding analysis. More unexpectedly, a shrimp species, *Acetes sibogae* of family Sergestidae, (98%) was also amplified.

DISCUSSION

Accuracy of diversity estimates using metabarcoding

This study reports the utilization of the DNA metabarcoding approach to assess the larval fish distribution and diversity in a biodiverse mangrove river system. It is a pioneering application of this technique in a Malaysian aquatic system and further supports its reliability for biodiversity assessment and potential future applications. In general, larvae were distributed in the Merbok River according to the

Table 2. Presence/absence of larval fish species along the Merbok River based on metabarcoding analysis of *COI* (▲) and 12S rRNA (◊) genes and habitat category of each species, where F: freshwater; FE: freshwater estuarine; MFE: marine, freshwater estuarine; M: marine; and ME: marine estuarine

No.	Family	Species	Habitat category	St1	St2	St3	St4
1.	Adrianichthyidae	*Oryzias javanicus*	FE	▲ ◊	▲ ◊	▲ ◊	▲ ◊
2.	Adrianichthyidae	*Oryzias melastigma*	FE	▲ ◊	▲ ◊	▲ ◊	
3.	Adrianichthyidae	*Oryzias dancena*	FE	▲ ◊	▲ ◊	▲ ◊	▲ ◊
4.	Ambassidae	*Ambassis gymnocephalus*	FE		▲	▲	
5.	Ariidae	*Netuma thalassina*	MFE		▲	▲	▲
6.	Bagridae	*Mystus cavasius*	FE		▲ ◊		▲ ◊
7.	Bagridae	*Mystus vittatus*	FE		▲ ◊		▲ ◊
8.	Carangidae	*Alepes djedaba*	M		▲ ◊	▲ ◊	▲ ◊
9.	Carangidae	*Alepes kleinii*	M			▲ ◊	▲ ◊
10.	Carangidae	*Atule mate*	ME			▲ ◊	▲ ◊
11.	Carangidae	*Caranx tille*	ME			▲ ◊	▲ ◊
12.	Carangidae	*Caranx ignobilis*	ME			▲ ◊	▲ ◊
13.	Carangidae	*Carangoides equula*	M				▲ ◊
14.	Carangidae	*Carangoides malabaricus*	M			▲ ◊	▲ ◊
15.	Carangidae	*Decapterus macarellus*	M			▲ ◊	▲ ◊
16.	Carangidae	*Decapterus maruadsi*	M			▲ ◊	▲ ◊
17.	Carangidae	*Megalaspis cordyla*	ME			▲ ◊	▲ ◊
18.	Carangidae	*Selaroides leptolepis*	ME			▲ ◊	▲ ◊
19.	Carangidae	*Trachinotus blochii*	ME				▲
20.	Chaetodontidae	*Chaetodon trifasciatus*	M			▲ ◊	▲ ◊
21.	Cichlidae	*Oreochromis niloticus*	F	▲ ◊	▲ ◊	▲ ◊	▲ ◊
22.	Cichlidae	*Oreochromis aureus*	F	▲ ◊	▲ ◊	▲ ◊	▲ ◊
23.	Clariidae	*Clarias batrachus*	F		▲ ◊	▲ ◊	
24.	Clupeidae	*Anodontostoma chacunda*	MFE			▲ ◊	▲ ◊
25.	Clupeidae	*Sardinella gibbosa*	M			▲ ◊	▲ ◊
26.	Clupeidae	*Escualosa thoracata*	MFE				▲ ◊
27.	Cynoglossidae	*Cynoglossus bilineatus*	ME				▲ ◊
28.	Eleotridae	*Oxyeleotris marmorata*	FE				▲ ◊
29.	Elopidae	*Elops hawaiensis*	MFE		▲ ◊	▲ ◊	
30.	Engraulidae	*Thryssa dussumieri*	ME			▲	▲
31.	Engraulidae	*Thryssa hamiltonii*	MFE			▲ ◊	▲ ◊
32.	Engraulidae	*Thryssa kammalensis*	ME			▲ ◊	▲ ◊
33.	Engraulidae	*Setipinna taty*	ME			▲ ◊	▲ ◊
34.	Engraulidae	*Stolephorus commersonnii*	ME				▲
35.	Ephippidae	*Ephippus orbis*	M			▲	▲

Table 2. (Continued)

No.	Family	Species	Habitat category	St1	St2	St3	St4
36.	Ephippidae	*Platax teira*	M			▲	▲ ◊
37.	Gerreidae	*Gerres filamentosus*	MFE				▲ ◊
38.	Gerreidae	*Gerres oyena*	ME		▲ ◊		▲ ◊
39.	Gobiidae	*Acentrogobius caninus*	MFE				▲ ◊
40.	Gobiidae	*Brachygobius xanthomelas*	F	◊			
41.	Gobiidae	*Trypauchen vagina*	ME	▲ ◊			
42.	Gymnuridae	*Gymnura poecilura*	M				▲ ◊
43.	Hemiramphidae	*Hyporhamphus quoyi*	MFE		▲		▲
44.	Latidae	*Lates calcarifer*	MFE			▲ ◊	▲ ◊
45.	Leiognathidae	*Gazza minuta*	ME				▲ ◊
46.	Lutjanidae	*Lutjanus argentimaculatus*	MFE		▲ ◊	▲ ◊	▲ ◊
47.	Lutjanidae	*Lutjanus malabaricus*	ME	▲ ◊	▲ ◊	▲ ◊	▲ ◊
48.	Lutjanidae	*Lutjanus johnii*	ME		▲ ◊		▲ ◊
49.	Lutjanidae	*Lutjanus russellii*	ME			▲ ◊	▲ ◊
50.	Mastacembelidae	*Mastacembelus erythrotaenia*	F		▲ ◊	▲ ◊	
51.	Mastacembelidae	*Macrognathus aculeatus*	FE		▲ ◊		
52.	Megalopidae	*Megalops cyprinoides*	MFE			▲ ◊	▲ ◊
53.	Mugilidae	*Liza planiceps*	MFE		▲ ◊		
54.	Mugilidae	*Moolgarda cunnesius*	MFE				▲ ◊
55.	Osphronemidae	*Osphronemus goramy*	F	▲ ◊	▲ ◊		
56.	Osphronemidae	*Trichogaster pectoralis*	F	▲ ◊			
57.	Platycephalidae	*Platycephalus indicus*	ME				▲ ◊
58.	Polynemidae	*Eleutheronema tetradactylum*	MFE			▲ ◊	▲ ◊
59.	Pristigasteridae	*Ilisha elongata*	ME				▲ ◊
60.	Pristigasteridae	*Opithopterus tardoore*	ME				▲ ◊
61.	Scatophagidae	*Scatophagus argus*	MFE				▲ ◊
62.	Sciaenidae	*Dendrophysa russelii*	MFE				▲ ◊
63.	Sciaenidae	*Johnius borneensis*	MFE				▲ ◊
64.	Sciaenidae	*Johnius carouna*	MFE		▲ ◊		▲ ◊
65.	Sciaenidae	*Johnius belangerii*	ME				▲ ◊
66.	Sciaenidae	*Pennahia argentata*	M			▲ ◊	
67.	Sciaenidae	*Pennahia macrocephalus*	M			▲ ◊	▲ ◊
68.	Sciaenidae	*Pennahia pawak*	M		▲	▲	▲
69.	Sciaenidae	*Otolithes ruber*	ME				▲ ◊
70.	Scombridae	*Auxis thazard*	M				▲ ◊
71.	Scombridae	*Euthynnus affinis*	M				▲ ◊
72.	Serranidae	*Epinephelus sexfasciatus*	M			▲ ◊	▲ ◊
73.	Serranidae	*Epinephelus tukula*	M				▲ ◊
74.	Siganidae	*Siganus canaliculatus*	ME			▲ ◊	▲ ◊
75.	Siganidae	*Siganus fuscescens*	ME	▲ ◊	▲ ◊	▲ ◊	▲ ◊
76.	Siganidae	*Siganus guttatus*	ME			▲ ◊	▲ ◊
77.	Sillaginidae	*Sillago aeolus*	M			▲ ◊	▲ ◊
78.	Sillaginidae	*Sillago sihama*	ME				▲ ◊
79.	Sphyraenidae	*Sphyraena barracuda*	ME			▲ ◊	▲ ◊
80.	Sphyraenidae	*Sphyraena jello*	ME				▲ ◊
81.	Stromatidae	*Pampus argenteus*	M				▲ ◊
82.	Synbranchidae	*Monopterus albus*	FE	▲			
83.	Terapontidae	*Terapon jarbua*	MFE		▲ ◊	▲ ◊	▲ ◊
84.	Tetraodontidae	*Tetraodon nigroviridis*	FE				▲ ◊
85.	Tetraodontidae	*Lagocephalus wheeleri*	M				▲
86.	Tetraodontidae	*Takifugu oblongus*	ME				▲ ◊
87.	Tetraodontidae	*Lagocephalus lunaris*	ME				▲ ◊
88.	Toxotidae	*Toxotes chatareus*	FE			▲ ◊	▲ ◊
89.	Triacanthodidae	*Triacanthodes anomalus*	M				▲ ◊

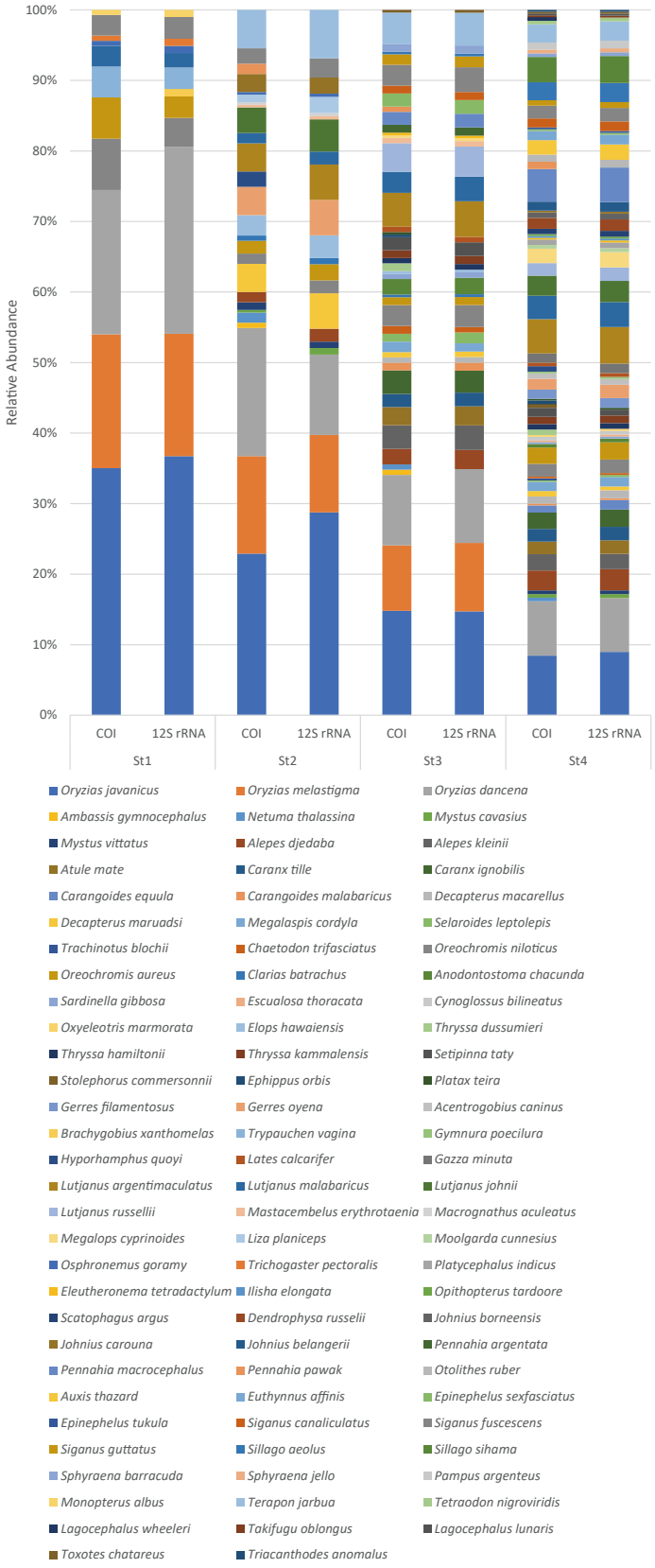

Fig. 2. The relative abundance of fish larvae species detected among sampling sites (St1, St2, St3 and ST4) based on the number of scaffolds reads of the *COI* and 12S rRNA genes.

expected zonation, but not exclusively, with freshwater-estuarine species predominating the upper stations (St1 and St2), marine-estuarine and marine species equally distributed in the middle zone (St3), and marine-estuarine species dominating the lower zone (St4), although fully marine species were also abundant. The species diversity detected in this study was lower than the morphologically identified adult specimens reported in previous studies (references). A total of 91 species (metabarcoding and Sanger sequencing of designed primer) were detected in the current study compared to 120 morphologically identified adults in earlier studies (Mansor et al. 2012a b c). The study also elucidated the alpha and beta diversities of fish larvae in the river. Furthermore, a comparison between the two approaches showed an overlap of only 47 species (28.7%). More than half of the species morphologically documented in previous studies were not detected in the current study, most probably due to different sampling seasons.

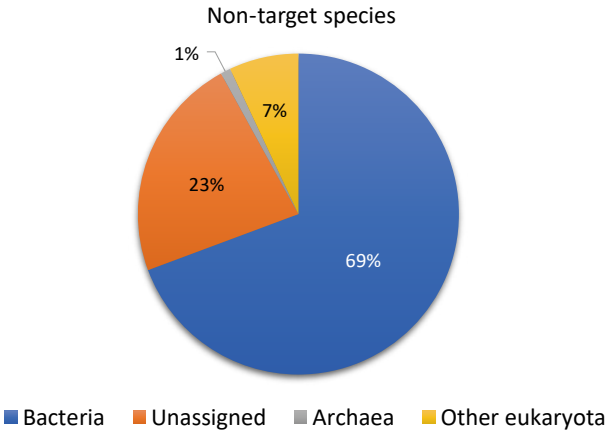

Fig. 3. Abundances of non-target DNA reads.

The discordance between the current and previous studies may be explained by variation in seasonal abundance of the species. Larvae samples were collected during the rainy season, presumed to be the spawning

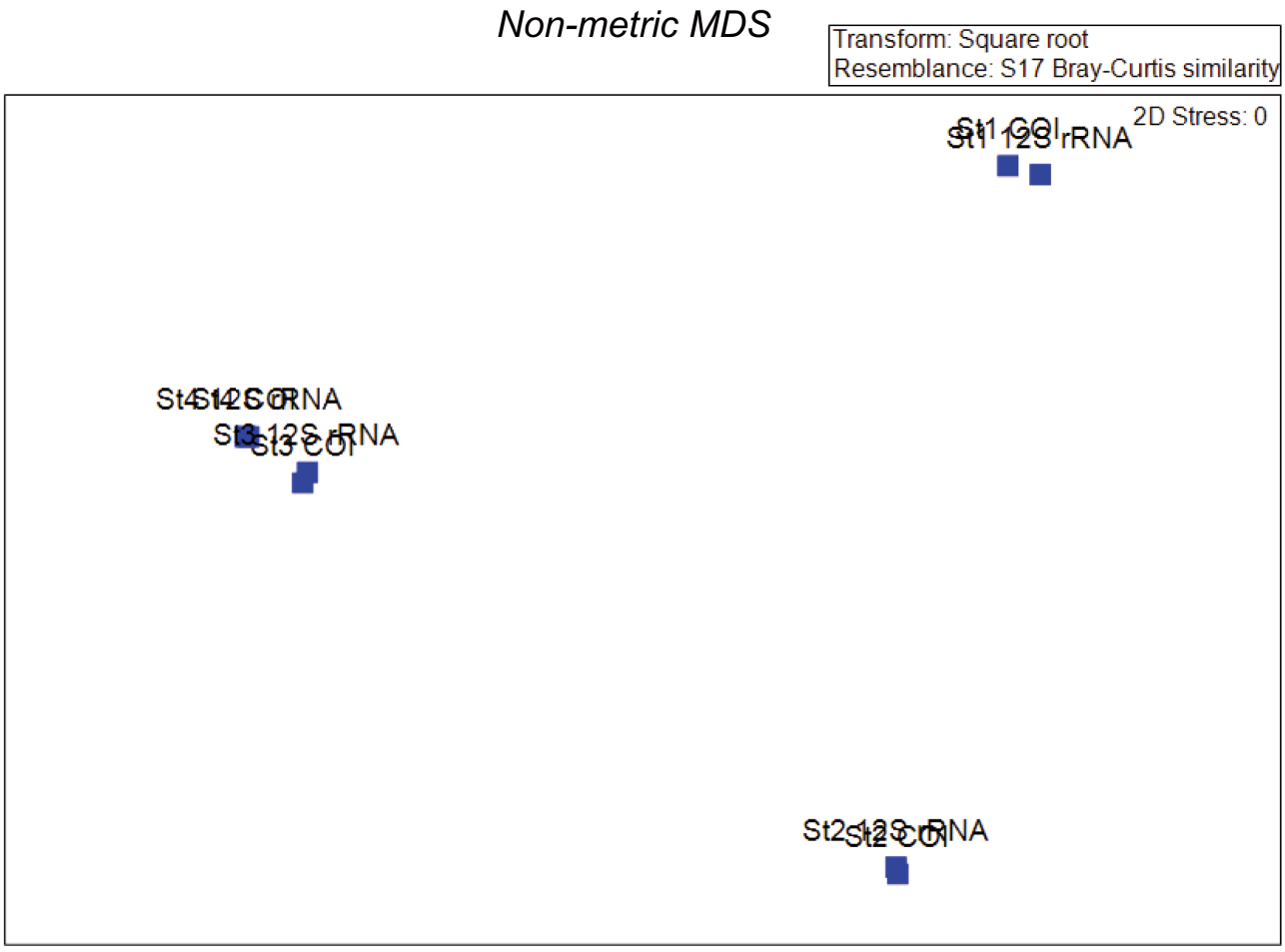

Fig. 4. Bray-Curtis similarity plotted by two-dimensional non-metric multidimensional scaling (NMDS) showing the similarity in larval fish species diversity among all four stations of the Merbok River identified using the *COI* and 12S rRNA genes.

peak for most fishes with higher plankton availability as food for fish larvae (Ikejima et al. 2003; Chew and Chong 2011; Ooi 2012). On the other hand, the earlier studies applied a whole-year sampling strategy, mainly from fishermen catches, which increased the probability of higher collection. For instance, the abundance of engraulids (*T. dussumieri*, *T. hamiltoni*, *T. kammalensis*, *S. taty*, and *S. commersonnii*) during the larvae sampling period in August strongly signified that this is the spawning season for members of family Engraulidae, which paralleled the findings by Ooi and Chong (2011). The catfish, *Arius argyropleuron* was not detected by the metabarcoding approach, although it was the most abundant fish species recorded in earlier studies (Mansor et al. 2012a b c). This species has a major spawning peak in April and a minor peak in July, not coinciding with our sampling period. Valdez-Moreno et al. (2010) compared adult and larval data. They found only 34 matches between the two groups, while another 75 records of species from larval data were not matched to the adult species. This they attributed to the seasonal diversity of larvae, influenced by species-specific spawning time. On the other hand, we recorded 44 species that had not been assigned in previous studies of adult populations. The detection of these species by metabarcoding highlights the usefulness of DNA metabarcoding to uncover species undetected by morphological assessments (Emilson et al. 2017). The inclusion of these metabarcoded fish species has generated a more comprehensive list of species present in the Merbok River with complementary data from fish larvae. Our data also provide insights into the spawning time and habitats of the identified species. Species misidentifications in previous studies may have also contributed to these discrepancies, although this is not expected to be a major reason since adult specimens have well-defined characteristics. Thus, we believe the other factors also played a major role in the differences of findings, including technical and bioinformatics issues.

Issues related to sample handling and bioinformatic analysis

One of the factors that could affect the accuracy of taxonomic assignment and biodiversity estimates is the selection of markers. Most metabarcoding studies on fish larvae have utilised the typical DNA barcoding marker of the *COI* gene (Maggia et al. 2017; Mariac et al. 2018; Nobile et al. 2019). The 12S rRNA gene utilised in this study has a proven record of delivering species-level identification of fish in metabarcoding investigations of eDNA and larval fish (Thomsen et al. 2012; Miya et al. 2015; Sato et al. 2017; Ratcliffe et al. 2021; Kim et al. 2021). However, the fish database of 12S rRNA genes is still incomplete, which could have led to missing species. Thus, a comprehensive and precise reference database of the DNA marker is an important prerequisite to obtain an accurate diversity assessment (Taberlet et al. 2012; Clarke et al. 2014; Bucklin et al. 2016; Weigand et al. 2019) and to avoid false positive and false negative species identification results due to a poor reference database (Bucklin et al. 2016). Although we rectified this by generating reference sequences of the 12S rRNA for species that had been morphologically cataloged in the area but with no available voucher sequences in the databases, the discrepancies were significant. This is likely due to several other factors that may be affecting the larval supply during sampling, which are larval distribution in small and isolated areas that are difficult to reach the survival of the larvae before and after arrival in the nursery grounds, and predation that occurs during the larval settlement in the nursery grounds (Pineda et al. 2010).

The high number of reads annotated to non-target taxa *i.e.*, non-fish species, a likely consequence of contamination, were observed. Bacteria had the highest composition, followed by archaea and other eukaryotes such as shrimps and molluscs. While the detection of these non-target organisms shows the versatility of metabarcoding to detect non-target organisms, this generality could affect the accuracy of metabarcoding for biodiversity estimates. In many cases, contaminants could occur naturally from the host or in the environment where the samples are collected (McKnight et al. 2019). Ficetola et al. (2015) and Liu et al. (2020) attributed contamination as one of the factors that could affect the accuracy of metabarcoding and influence false-positive and false-negative detection. The presence of these contaminants during sampling and DNA extraction and inadequate bioinformatic analyses could mask the presence of target organisms or species, especially when DNA from the contaminants is more abundant than the target organisms. We adhered to a strict protocol throughout the process; clean equipment and closed containers were used to prevent cross-contamination among sites, but breakthrough contamination could still occur. This was revealed by the unexpected detection of the native shrimp species *Acetes sibogae.* Although great precautions were taken, the samples may have been contaminated with *A. sibogae* tissues during sampling or in the laboratory. The abundance and higher affinity of DNA templates of this shrimp species to P12 and P13 primers could also explain the detection of *A. sibogae* in St3 and St4. An eDNA study conducted by Thomsen et al. (2012) targeting fish species at The Sound of Elsinore,

Denmark also unexpectedly detected four species of birds that occasionally cross the sampling area during migration.

The risk of missing target species due to contamination can be overcome by increasing the number of technical replicates (Ficetola et al. 2015). However, a high number of replicates leads to increased sequencing costs. Our study pooled five sampling replicates to increase the chances of identifying more species. However, such a strategy is also associated with several disadvantages. It may dilute the DNA of rare species or low abundant species present in the bulk samples, resulting in non-detection of these species and further loss of these rare fish lineage information (Kelly et al. 2014; Shaw et al. 2016; Sato et al. 2017). Sato et al. (2017) stated that the pooling of samples in eDNA metabarcoding is unsuitable if the objective of the study is to assess species richness and alpha diversity of species. To improve the species detection and to avoid false positive and false negative due to the pooling of samples, it is advised that the specimens be sorted and extracted individually according to their size before pooling the extracted DNA together or to cut the specimens in similar size for pool DNA extraction (Ji et al. 2013; Elbrecht et al. 2017), which must be considered in future investigations.

Another plausible reason for the failure to detect targeted species is the low concentration or low affinity of the target species DNA templates to the tested primer pairs, which could be outcompeted by higher affinity DNA templates (Lobo et al. 2017). This was evident when we amplified the abundant *Oryzias javanicus* even though it was not a target species of the newly designed primers. The lower DNA concentration of target species versus non-target species may cause a false negative in the DNA metabarcoding data, as Smith (2017) noted. Furthermore, mismatches between primers and target templates can prevent certain species from being amplified by PCR and prevent that species from being detected (Bru et al. 2008; Deagle et al. 2014; Elbrecht and Leese 2015; Piñol et al. 2015). Lobo et al. (2017) suggested that newly designed primers need to be tested and optimized on individual specimens or using assembled mixtures prior to large-scale analysis of bulk samples. This should be done to prevent mismatches between primers and target templates in bulk samples of metabarcoding. These factors must be taken into account in future investigations.

The large metabarcoding dataset is one of the challenges and difficulties in the bioinformatic analysis for precise taxonomic assignment. One of the major steps in bioinformatic analysis is the trimming of raw sequence reads. The step involves the removal of sequences containing excessive ambiguous or low-confidence base calls (Bokulich et al. 2013; Edgar and Flyvbjerg 2015) to improve the accuracy of the reads. During the trimming process, the parameters must be carefully considered to remove sequencing errors and reads effectively, which can affect downstream diversity and abundance analysis, and loss of reads of low-abundance taxa (Piper et al. 2019). The best-hit classification using alignment-based tools such as BLAST is a more widely used method for taxonomic assignment compared to other methods such as the sequence composition method and phylogenetic method (Piper et al. 2019). However, this simple classification method is prone to over-classifying the query sequence resulting in incorrect species-level taxonomy, especially when the reference data is mislabelled, absent or incomplete, yielding false-positive and false-negative results (Koski and Golding 2001). To overcome this issue that comes from a lack of reference data, the sequence may still be assigned to a higher taxonomic rank with good support for example, the family level (Porter and Hajibabaei 2018).

Suitability and cost-efficiency of metabarcoding for biomonitoring of fish larvae

Studies on fish larvae generate important insights on spawning locations and seasons, reproductive biology, nursery grounds, migratory routes of fishes as well as for biomonitoring of habitats (Kidwai and Amjad 2001; Frantine-Silva et al. 2015; Maggia et al. 2017; Mariac et al. 2018; Nobile et al. 2019; Ratcliffe et al. 2021). DNA-metabarcoding coupled with a comprehensive reference database of fish species is now recognised as an efficient cost-effective technique for large-scale studies involving environmental and bulk samples, when laboratory facilities are available (Maggia et al. 2017; Nobile et al. 2019). Besides, it does not require high-level taxonomic expertise for individual identification (Kacev et al. 2018). A recent morphological checklist of the adult fish survey by Zainal Abidin et al. (2021) of the Merbok River and nearby landing sites recorded an additional 75 species (with an overlap of 12 species of larvae) from those of Mansor et al. (2012a b c). We believe that more species are yet to be documented through a more rigorous sampling procedure coupled with DNA metabarcoding.

Knowledge of species and population distribution patterns is critical in strategizing biodiversity conservation effort (Thomsen and Willerslev 2015). The current metabarcoding data have furnished helpful information on species identities and distributions along with an anthropogenically important river system and the alpha- and beta-diversity estimates. The utilisation of eDNA studies in conservation strategies have

been highlighted in various organisms; Hajibabaei et al. (2011) on freshwater benthic macroinvertebrate, Calvignac-Spencer et al. (2013) on carrion flies (blow and flesh flies) to monitor mammal diversity and Ji et al. (2013) on the diversity of insects and birds. Our DNA metabarcoding analysis has provided strong evidence that the Merbok River still supports a diversity of fish species, a piece of welcome news for the local community. However, a more holistic survey on larval and adult fish, including non-commercial species in Merbok River, should be conducted, including temporal and ecological studies. These morphological and molecular databases will be beneficial for strategizing the management and conservation of fisheries in this area. DNA metabarcoding has proven a rapid and cost-effective identification tool for the Merbok River, which could be a model for similar research in other aquatic ecosystems in Malaysia.

CONCLUSIONS

We detected 89 species of fish larvae through metabarcoding, with two additional species identified in the validation using newly designed primers, making a total of 91 identified species with > 97% species identity based on the existing databases. Although lower in species richness compared to the morphologically identified species of adult specimens in previous studies, we argue that the probability of low species richness is due to the lower sampling effort, which only focused on a single season (rainy season), and possible technical issues in the sampling, laboratory and bioinformatics analyses that should be addressed in future projects. Nevertheless, the current findings further support the suitability of DNA metabarcoding as a cost-effective approach for investigating species distribution and diversity in this region. In addition, it contributed novel data not recorded in previous studies. We suggest a more holistic fish larvae survey in the Merbok River by considering seasonal changes and increased sampling sites. This would enable more comprehensive data to understand the patterns in fish larvae ecology and distribution within the estuarine mangrove habitats and thus their conservation.

Acknowledgments: We would like to thank Universiti Sains Malaysia (1001/PBIOLOGI/870018), Universiti Malaysia Terengganu (UMT/IMB/53307) and DRMREEF-KIOST & IOCWESTPAC for funding this project. Our special thanks to our colleagues in the Molecular Ecology Research Laboratory, USM for their assistance during the research.

Authors' contributions: SAMN and LYH acquired research funding. NFMAHA, SZA, SAMN, LYH and MGMA designed the experiment. NFMAHA and SZA performed the field works and laboratory works. NFMAHA and AA analysed the data. NFMAHA drafted the manuscript. All authors contributed to and participated in revising and approving the manuscript.

Competing interests: There is no conflict of interests among authors regarding the publication of this paper.

Availability of data and materials: The 12S rRNA genes sequence of each species was submitted to NCBI (GenBank) under accession numbers: KY379960-KY379968, KY778751-KY778754, MG729393, MG729396, MG729397, MG748713, MG748714, MK330865-MK330867.

Consent for publication: Not applicable.

Ethics approval consent to participate: Not applicable.

REFERENCES

Ambak MA, Mansor MI, Zakaria MZ, Mazlan AG. 2012. Fishes of Malaysia, Univeristi Malaysia Terengganu, Kuala Terengganu, Terengganu, Penerbit UMT.

Arshad AB, Ara R, Amin SMN, Daud SK, Ghaffar MA. 2012. Larval fish composition and spatio-temporal variation in the estuary of Pendas River, southwestern Johor, Peninsular Malaysia. Coastal Marine Science **35(1):**96–102.

Atan Y, Jaafar H, Majid ARA. 2010. Ikan Laut Malaysia: glosari nama sahih spesies ikan, Dewan Bahasa dan Pustaka.

Azmir I, Esa Y, Amin S, Md Yasin I, Md Yusof F. 2017. Identification of larval fish in mangrove areas of Peninsular Malaysia using morphology and DNA barcoding methods. J Appl Ichthyol **33(5):**998–1006. doi:10.1111/jai.13425.

Bokulich NA, Subramanian S, Faith JJ, Gevers D, Gordon JI, Knight R, Mills DA, Caporaso JG. 2013. Quality-filtering vastly improves diversity estimates from Illumina amplicon sequencing. Nat Methods **10(1):**57–59. doi:10.1038/nmeth.2276.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30(15):**2114–2120. doi:10.1093/bioinformatics/btu170.

Bru D, Martin-Laurent F, Philippot L. 2008. Quantification of the detrimental effect of a single primer-template mismatch by real-time PCR using the 16S rRNA gene as an example. Appl Environ Microbiol **74(5):**1660–1663. doi:10.1128/aem.02403-07.

Bucklin A, Lindeque PK, Rodriguez-Ezpeleta N, Albaina A, Lehtiniemi M. 2016. Metabarcoding of marine zooplankton: prospects, progress and pitfalls. J Plankton Res **38(3):**393–400. doi:10.1093/plankt/fbw023.

Calvignac-Spencer S, Merkel K, Kutzner N, Kühl H, Boesch C, Kappeler PM, Metzger S, Schubert G, Leendertz FH. 2013. Carrion fly-derived DNA as a tool for comprehensive and cost-effective assessment of mammalian biodiversity. Mol Ecol **22(4):**915–924. doi:10.1111/mec.12183.

Carpenter KE, Niem VH. 2001. FAO species identification guide for fishery purposes. The living marine resources of the Western Central Pacific.

Chew LL, Chong VC. 2011. Copepod community structure and abundance in a tropical mangrove estuary, with comparisons to coastal waters. Hydrobiologia **666(1):**127–143. doi:10.1007/s10750-010-0092-3.

Clarke LJ, Soubrier J, Weyrich LS, Cooper A. 2014. Environmental metabarcodes for insects: in silico PCR reveals potential for taxonomic bias. Mol Ecol Resour **14(6):**1160–1170. doi:10.1111/1755-0998.12265.

Coissac E, Riaz T, Puillandre N. 2012. Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol **21(8):**1834–1847. doi:10.1111/j.1365-294X.2012.05550.x.

Collet A, Durand JD, Desmarais E, Cerqueira F, Cantinelli T, Valade P, Ponton D. 2018. DNA barcoding post-larvae can improve the knowledge about fish biodiversity: an example from La Reunion, SW Indian Ocean. Mitochondrial DNA Part A **29(6):**905–918. doi:10.1080/24701394.2017.1383406.

Cristescu ME. 2014. From barcoding single individuals to metabarcoding biological communities: towards an integrative approach to the study of global biodiversity. Trends Ecol Evol **29(10):**566–571. doi:10.1016/j.tree.2014.08.001.

Deagle BE, Jarman SN, Coissac E, Pompanon F, Taberlet P. 2014. DNA metabarcoding and the cytochrome *c* oxidase subunit I marker: not a perfect match. Biol Lett **10(9):**1–4. doi:10.1098/rsbl.2014.0562.

Edgar RC, Flyvbjerg H. 2015. Error filtering, pair assembly and error correction for next-generation sequencing reads. Bioinformatics **31(21):**3476–3482. doi:10.1093/bioinformatics/btv401.

Elbrecht V, Leese F. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass—sequence relationships with an innovative metabarcoding protocol. PLoS ONE **10(7):**1–16. doi:10.1371/journal.pone.0130324.

Elbrecht V, Vamos EE, Meissner K, Aroviita J, Leese F. 2017. Assessing strengths and weaknesses of DNA metabarcoding-based macroinvertebrate identification for routine stream monitoring. Methods Ecol Evol **8(10):**1265–1275. doi:10.1111/2041-210X.12789.

Emilson CE, Thompson DG, Venier LA, Porter TM, Swystun T, Chartrand D, Capell S, Hajibabaei M. 2017. DNA metabarcoding and morphological macroinvertebrate metrics reveal the same changes in boreal watersheds across an environmental gradient. Sci Rep **7:**1–11. doi:10.1038/s41598-017-13157-x.

Fatema K, Maznah WW, Isa MM. 2014. Spatial and temporal variation of physico-chemical parameters in the Merbok Estuary, Kedah, Malaysia. Trop Life Sci Res **25(2):**1–19.

Ficetola GF, Pansu J, Bonin A, Coissac E, Giguet-Covex C, De Barba M, Gielly L, Lopes CM, Boyer F, Pompanon F, Rayé G. 2015. Replication levels, false presences and the estimation of the presence/absence from eDNA metabarcoding data. Mol Ecol Resour **15(3):**543–556. doi:10.1111/1755-0998.12338.

Frantine-Silva W, Sofia SH, Orsi ML, Almeida FS. 2015. DNA barcoding of freshwater ichthyoplankton in the Neotropics as a tool for ecological monitoring. Mol Ecol Resour **15(5):**1226–1237. doi:10.1111/1755-0998.12385.

Fujii K, Doi H, Matsuoka S, Nagano M, Sato H, Yamanaka H. 2019. Environmental DNA metabarcoding for fish community analysis in backwater lakes: A comparison of capture methods. PLoS ONE **14(1):**1–17. doi:10.1371/journal.pone.0210357.

Grewe PM, Krueger CC, Aquadro CF, Bermingham E, Kincaid HL, May B. 1993. Mitochondrial DNA variation among lake trout (*Salvelinus namaycush*) strains stocked into Lake Ontario. Can J Fish Aquat Sci **50(11):**2397–2403. doi:10.1139/f93-264.

Hajibabaei M, Shokralla S, Zhou X, Singer GA, Baird DJ. 2011. Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. PLoS ONE **6(4):**1–7. doi:10.1371/journal.pone.0017497.

Hebert PD, Cywinska A, Ball SL. 2003. Biological identifications through DNA barcodes. P Roy Soc Lond B Bio **270(1512):**313–321. doi:10.1098/rspb.2002.2218.

Ikejima K, Tongnunui P, Medej T, Taniuchi T. 2003. Juvenile and small fishes in a mangrove estuary in Trang province, Thailand: seasonal and habitat differences. Estuar Coast Shelf Sci **56(3-4):**447–457. doi:10.1016/S0272-7714(02)00194-4.

Ji Y, Ashton L, Pedley SM, Edwards DP, Tang Y, Nakamura A, Kitching R, Dolman PM, Woodcock P, Edwards FA. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecol Lett **16(10):**1245–1257. doi:10.1111/ele.12162.

Jusoff K, Taha D. 2008. Managing sustainable mangrove forests in Peninsular Malaysia. Journal of Sustainable Development **1(1):**88–96.

Kacev D, Gillett D, de Carvalho AF, Cash C, Walther S, Thompson A, Thompson L, Bowlin N, Goodwin K, Stein ED. 2018. Assessment of Ichthyoplankton Metabarcoding for Routine Monitoring.

Kelly RP, Port JA, Yamahara KM, Crowder LB. 2014. Using environmental DNA to census marine fishes in a large mesocosm. PLoS ONE **9(1):**1–11. doi:10.1371/journal.pone.0086175.

Kidwai S, Amjad S. 2001. Abundance and distribution of ichthyolarvae from upper pelagic waters of the northwestern Arabian Sea during different monsoon periods, 1992–1994. ICES J Mar Sci **58(3):**719–724. doi:10.1006/jmsc.2000.1057.

Kim AR, Yoon TH, Lee CI, Kang CK, Kim HW. 2021. Metabarcoding analysis of ichthyoplankton in the East/Japan Sea using the novel fish-specific universal primer set. Front Mar Sci **8:**1–15. doi:10.3389/fmars.2021.614394.

Ko HL, Wang YT, Chiu TS, Lee MA, Leu MY, Chang KZ, Chen WY, Shao KT. 2013. Evaluating the Accuracy of Morphological Identification of Larval Fishes by Applying DNA Barcoding. PLoS ONE **8(1):**1–7. doi:10.1371/journal.pone.0053451.

Koski LB, Golding GB. 2001. The closest BLAST hit is often not the nearest neighbor. J Mol Evol **52(6):**540–542. doi:10.1007/s002390010184.

Kumar S, Stecher G, Tamura K. 2016. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol **33(7):**1870–1874. doi:10.1093/molbev/msw054.

Lewis LA, Richardson DE, Zakharov EV, Hanner R. 2016. Integrating DNA barcoding of fish eggs into ichthyoplankton monitoring programs. Fish Bull **114:**153–165. doi:10.7755/FB.114.2.3.

Li D, Liu CM, Luo R, Sadakane K, Lam TW. 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics **31(10):**1674–1676. doi:10.1093/bioinformatics/btv033.

Liu M, Clarke LJ, Baker SC, Jordan GJ, Burridge CP. 2020. A practical guide to DNA metabarcoding for entomological ecologists. Ecol Entomol **45(3):**373–385. doi:10.1111/een.12831.

Lobo J, Shokralla S, Costa MH, Hajibabaei M, Costa FO. 2017. DNA metabarcoding for high-throughput monitoring of estuarine macrobenthic communities. Sci Rep **7(1):**1–13. doi:10.1038/s41598-017-15823-6.

Maggia M, Vigouroux Y, Renno JF, Duponchelle F, Desmarais E, Nunez J, García-Dávila C, Carvajal-Vallejos F, Paradis E, Martin J. 2017. DNA metabarcoding of Amazonian ichthyoplankton swarms. PLoS ONE **12(1):**1–14. doi:10.1371/journal.pone.0170009.

Manson FJ, Loneragan NR, Skilleter GA, Phinn SR. 2005. An

evaluation of the evidence for linkages between mangroves and fisheries: a synthesis of the literature and identification of research directions. Oceanogr Mar Biol **43:**493–524.

Mansor MI, Abdul Basri MN, Mohd Zawawi MZ, Yahya K, Nor SAM. 2012a. Length-weight relationships of some important estuarine fish species from Merbok Estuary, Kedah. Journal of Natural Sciences Research **2(2):**8–19.

Mansor MI, Mohammad-Zafrizal MZ, Nur-Fadhilah M, Khairun Y, Wan-Maznah WO. 2012b. Temporal and spatial variations in fish assemblage structures in relation to the physicochemical parameters of the Merbok estuary, Kedah. Journal of Natural Sciences Research **2(7):**110–127.

Mansor MI, Wan Maznah WO, Khairun Y, Tan SH. 2012c. Persekitaran, Aktiviti dan Sumber Perikanan Artisanal Di Sungai Merbok, Kedah. Universiti Sains Malaysia. Unpublished report.

Mariac C, Vigouroux Y, Duponchelle F, García-Dávila C, Nunez J, Desmarais E, Renno JF. 2018. Metabarcoding by capture using a single *COI* probe (MCSP) to identify and quantify fish species in ichthyoplankton swarms. PLoS ONE **13(9):**1–15. doi:10.1371/journal.pone.0202976.

Mazlan A, Zaidi C, Wan-Lotfi W, Othman B. 2005. On the current status of coastal marine biodiversity in Malaysia. Indian J Mar Sci **34(1):**76–87.

McKnight DT, Huerlimann R, Bower DS, Schwarzkopf L, Alford RA, Zenger KR. 2019. microDecon: A highly accurate read-subtraction tool for the post-sequencing removal of contamination in metabarcoding studies. Environ DNA **1(1):**14–25. doi:10.1002/edn3.11.

Miya M, Sato Y, Fukunaga T, Sado T, Poulsen JY, Sato K, Minamoto T, Yamamoto S, Yamanaka H, Araki H. 2015. MiFish, a set of universal PCR primers for metabarcoding environmental DNA from fishes: detection of more than 230 subtropical marine species. Roy Soc Open Sci **2(7):**1–33. doi:10.1098/rsos.150088.

Moser HG. 1996. The early stages of fishes in the California Current region. Calif Coop Ocean Fish Invest Atlas **33:**1–1505.

Moser HG, Smith PE. 1993. Larval fish assemblages and oceanic boundaries. B Mar Sci **53(2):**283–289.

Nobile AB, Freitas-Souza D, Ruiz-Ruano FJ, Nobile MLM, Costa GO, De Lima FP, Camacho JPM, Foresti F, Oliveira C. 2019. DNA metabarcoding of Neotropical ichthyoplankton: Enabling high accuracy with lower cost. Metabarcoding and Metagenomic **3:**69–76. doi:10.3897/mbmg.3.35060.

Ong JE, Wan Juliana WA, Yong JWH, Maketab M, Wong YY, Mohd Nasir H. 2015. The Merbok mangrove: present status and the way forward. *In*: Abd Rahim AR, Ku Aman KA, Abu Hassan MN, Abdullah M, Nor Hazliza MB, Latiff A, Editors. Hutan paya laut Merbok, Kedah: Pengurusan hutan, persekitaran fizikal dan kepelbagaiana flora. Kuala Lumpur (Malaysia): Jabatan Perhutanan Semenanjung Malaysia pp. 21–33.

Ooi A, Chong V. 2011. Larval fish assemblages in a tropical mangrove estuary and adjacent coastal waters: Offshore-inshore flux of marine and estuarine species. Cont Shelf Res **31(15):**1599–1610. doi:10.1016/j.csr.2011.06.016.

Ooi AL. 2012. Assemblage, recruitment and ecology of fish larvae in Matang mangrove estuary and adjacent waters, Peninsular Malaysia. PhD thesis. University of Malaya.

Pineda J, Porri F, Starczak V, Blythe J. 2010. Causes of decoupling between larval supply and settlement and consequences for understanding recruitment and population connectivity. J Exp Mar Biol and Ecol **392(1-2):**9–21. doi:10.1016/j.jembe.2010.04.008.

Piñol J, Mir G, Gomez-Polo P, Agustí N. 2015. Universal and blocking primer mismatches limit the use of high-throughput DNA sequencing for the quantitative metabarcoding of arthropods. Mol Ecol Resour **15(4):**819–830. doi:10.1111/1755-0998.12355.

Piper AM, Batovska J, Cogan NO, Weiss J, Cunningham JP, Rodoni BC, Blacket MJ. 2019. Prospects and challenges of implementing DNA metabarcoding for high-throughput insect surveillance. GigaScience **8(8):**1–22. doi:10.1093/gigascience/giz092.

Porter TM, Hajibabaei M. 2018. Automated high throughput animal CO1 metabarcode classification. Sci Rep **8(1):**1–10. doi:10.1038/s41598-018-22505-4.

Ratcliffe FC, Webster TMU, Barreto DR, O'rorke R, De Leaniz CG, Consuegra S. 2021. Quantitative assessment of fish larvae community composition in spawning areas using metabarcoding of bulk samples. Ecol Appl **31:**e02284. doi:10.1002/eap.2284.

Sato H, Sogo Y, Doi H, Yamanaka H. 2017. Usefulness and limitations of sample pooling for environmental DNA metabarcoding of freshwater fish communities. Sci Rep **7(1):**1–12. doi:10.1038/s41598-017-14978-6.

Smith L. Biodiversity monitoring using environmental DNA: Can it detect all fish species in a waterbody and is it cost effecting for routine monitoring? MSc. Thesis. Edith Cowan University.

Shaw JL, Clarke LJ, Wedderburn SD, Barnes TC, Weyrich LS, Cooper A. 2016. Comparison of environmental DNA metabarcoding and conventional fish survey methods in a river system. Biol Conserv **197:**131–138. doi:10.1016/j.biocon.2016.03.010.

Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol **21(8):**2045–2050. doi:10.1111/j.1365-294X.2012.05470.x.

Tang M, Tan M, Meng G, Yang S, Su XU, Liu S, Song W, Li Y, Wu Q, Zhang A, Zhou X. 2014. Multiplex sequencing of pooled mitochondrial genomes—a crucial step toward biodiversity analysis using mito-metagenomics. Nucleic Acids Res **42(22):**1–13. doi:10.1093/nar/gku917.

Thomsen PF, Kielgast J, Iversen LL, Møller PR, Rasmussen M, Willerslev E. 2012. Detection of a diverse marine fish fauna using environmental DNA from seawater samples. PLoS ONE **7(8):**1–9. doi:10.1371/journal.pone.0041732.

Thomsen PF, Willerslev E. 2015. Environmental DNA–An emerging tool in conservation for monitoring past and present biodiversity. Biol Conserv **183:**4–18. doi:10.1016/j.biocon.2014.11.019.

Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. 2007. Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res **35(suppl_2):**W71–W74. doi:10.1093/nar/gkm306.

Valdez-Moreno M, Vásquez-Yeomans L, Elías-Gutiérrez M, Ivanova NV, Hebert PD. 2010. Using DNA barcodes to connect adults and early life stages of marine fishes from the Yucatan Peninsula, Mexico: potential in fisheries management. Mar Freshwater Res **61(6):**655–671. doi:10.1071/MF09222.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD. 2005. DNA barcoding Australia's fish species. Philos T Roy Soc B **360(1462):**1847–1857. doi:10.1098/rstb.2005.1716.

Weigand H, Beermann AJ, Čiampor F, Costa FO, Csabai Z, Duarte S, Geiger MF, Grabowski M, Rimet F, Rulik B, Strand M. 2019. DNA barcode reference libraries for the monitoring of aquatic biota in Europe: Gap-analysis and recommendations for future work. Sci Total Environ **678:**499–524. doi:10.1016/j.scitotenv.2019.04.247.

Wibowo A, Sloterdijk H. 2015. Identifying sumatran peat swamp fish larvae through DNA barcoding, evidence of complete life history pattern. Procedia Chem **14:**76–84. doi:10.1016/j.proche.2015.03.012.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis

with Kraken 2. Genome Biol **20(1):**1–13. doi:10.1186/s13059-019-1891-0.

Zainal Abidin DH, Lavoué S, Alshari NFMAH, Nor SAM, Rahim MA, Akib NAM. 2021. Ichthyofauna of Sungai Merbok Mangrove Forest Reserve, northwest Peninsular Malaysia, and its adjacent marine waters. Check List **17(2):**601–631. doi:10.15560/17.2.601.

Supplementary Materials

Table S1. Summary of assembled scaffolds statistics using MEGAHIT (v1.0.2). (download)

Table S2. List of newly designed species-specific primer pairs with description of its primer length, product size, GC content (%) and melting temperature (°C). The successfully amplified primer pairs are marked as '√'. (download)

Table S3. Number of scaffolds annotated to *COI* and 12S rRNA genes that were assigned to fish larvae species with blast identity at ≥ 97%. (download)