FISH IDENTIFICATION ACROSS MERBOK ESTUARY THROUGH DNA BARCODING AND ENVIRONMENTAL DNA (eDNA) METABARCODING IN TROPICAL MANGROVE ESTUARY

DANIAL HARIZ BIN ZAINAL ABIDIN

UNIVERSITI SAINS MALAYSIA

2022

FISH IDENTIFICATION ACROSS MERBOK ESTUARY THROUGH DNA BARCODING AND ENVIRONMENTAL DNA (eDNA) METABARCODING IN TROPICAL MANGROVE ESTUARY

by

DANIAL HARIZ BIN ZAINAL ABIDIN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

September 2022

ACKNOWLEDGEMENT

In the name of Allah, the Most Gracious, the Most Merciful.

This has been an adventurous rollercoaster ride – happy, sad, exciting, and sometimes crazy - and there are so many people for whom I am immensely grateful. Without my supervisor, Dr. Adelyna (Kak Adel), I would not be where I am today. She believed in me, fed me (with all the delicious cakes!), and supported me in the decisions I made regarding the direction of my work. A big thank you goes to Prof. Siti Azizah who has inspired me since my student days, my dedicated advisor, and a woman who deserves my utmost respect. She has always had my back and I am truly grateful to be part of her legacy. I am also very grateful to Dr. Sébastien Lavoué. His passion for 'fish science' has truly inspired me to become a better scientist.

I would also like to thank my amazing Lab. 308 members, Kak Noorul, Jam, Abang Fadli, Kak Adib, Norli, Zu, Mel, Zafirah, Kak Idah, Firdaus, Fong, Kak Fatimah, and all those who always helped me with any problem, no matter how crazy and ridiculous, and stood by me during the difficult phase of my Ph.D. journey. A big thank you to Kak Masa, Pak Su Merbok, and the team at FRI Batu Maung for their support during the memorable sampling activities. Thanks to all my USM buddies, Kak Ain, Kak Fuzah, Daus, and all my best friends (you know who you are) for the friendship, encouragement, and good meal times! May Allah keep our ukhuwah, always.

I would also like to thank the Ministry of Higher Education Malaysia for providing the MyBrain MyPhD scholarship. My work was also made possible by the funds from the FRGS grant (FRGS/1/2016/STG05/USM/02/2). Special thanks to the 8th International Barcode of Life Conference - iBOL2019 for awarding me with the

travel grant and the eDNA workshop in Trondheim, which really shed light on my Ph.D. project.

Finally, and most importantly, I dedicate this work to my parents, Mama (Hjh. Faridah Osman) and Babah (Hj. Zainal Abidin). They believed in me when others doubted me. Your prayers are my strength and I hope I can repay you both with this success. Thank you to all my family members, Abang, Acu, Kak Wan, Kak Azie, Nia, Rayyan, and Rayqal. I love you guys to the moon and back. I know I would not make it without your support. Hariz made it to the end!

"*Allah does not intend to make difficulty for you, but He intends to purify you and complete His favour upon that you may be grateful*." Quran 5:6

TABLE OF CONTENTS

LIST OF PUBLICATIONS

CONFERENCE PROCEEDINGS

LIST OF TABLES

Page

LIST OF FIGURES

Page

LIST OF PLATES

Page

LIST OF SYMBOLS

α	Alpha
β	Beta
♂	Male
♀	Female

LIST OF ABBREVIATIONS

ABGD	Automatic Barcode Gap Discovery
aDNA	Ancient DNA
BI	Bayesian Inference
BIN	Barcode Index Numbers
BOLD	Barcode of Life Data system
bp	Base pairs
BSA	Bovine Serum Albumin
CO_2	Carbon dioxide
COI	Cytochrome oxidase subunit I
Cyt *b*	Cytochrome b
DW	Disk width
E	East
eDNA	Environmental DNA
ESS	Effective sample size
GMYC	Generalized Mixed Yule Coalescent
HTS	High-throughput sequencing
JC	Jukes Cantor
K2P	Kimura-2 Parameter
ML	Maximum Likelihood
MOTU	Molecular Operational Taxonomic Unit
mtDNA	Mitochondrial DNA
N	North
NMDS	Non-metric multidimensional
PERMANOVA	Permutative multivariate analysis of variance

RESL	Refined Single Linkage
SD	Standard deviation
SE	Standard error
TL	Total length

LIST OF APPENDICES

PENGENALPASTIAN IKAN SEPANJANG MUARA MERBOK MELALUI PENGEKODAN DNA DAN METABARKOD DNA PERSEKITARAN (eDNA) DI MUARA BAKAU TROPIKA

ABSTRAK

Titik panas biodiversiti sentiasa berhadapan dengan kekurangan maklumat taksonomi, terutamanya bagi biota yang dijumpai di ekosistem muara bakau tropika. Ekosistem diversiti-mega muara bakau merupakan salah satu ekosistem yang paling terancam di dunia, kesan daripada aktiviti manusia. Tambahan lagi, pengetahuan tentang taksonomi dan taburan spesies dalam ekosistem ini masih belum lengkap bagi kebanyakan kumpulan, terutamanya takson ikan, seterusnya menghalang pelaksanaan pelan pemuliharaan biodiversiti yang lestari. Untuk mengisi jurang taksonomi di kawasan yang kaya dengan spesies ini, kajian ini menyelidik kepelbagaian komuniti ikan di kawasan muara bakau tropika - Muara Merbok yang terletak di barat laut Semenanjung Malaysia, menggunakan kaedah konvensional dan berasaskan DNA. Penilaian konvensional terhadap kepelbagaian spesies ikan di Muara Merbok dan perairan laut berdekatan telah merekodkan sejumlah 138 spesies ikan yang tergolong dalam dua kelas, 18 order, 47 famili, dan 94 genera. Repositori spesimen telah ditubuhkan di Makmal Rujukan Biodiversiti, USM, untuk membolehkan kajian perbandingan dijalankan dengan lebih lanjut. Seterusnya, pengekodan DNA berasaskan gen COI telah digunakan untuk menilai kepelbagaian ikan di Muara Merbok secara genetik dan melengkapi keputusan morfologi yang diperoleh dalam kajian sebelumnya. Depositori rujukan pengekodan DNA bagi 350 individu ikan telah dikumpul. Daripada 138 spesies yang dikenal pasti melalui morfologi, DNA barkod

telah mengesahkan kesahihan 123 spesies. Kajian DNA barkod juga telah mendedahkan kepelbagaian yang tersembunyi dalam tujuh spesies, manakala divergen antara dua pasangan spesies telah dikesan berada di bawah ambang interspesifik dan perlu dikaji dengan lebih lanjut. Perbandingan dengan senarai spesies terdahulu di dalam dan sekitar kawasan ini menunjukkan bahawa liputan taksonomi di Muara Merbok masih tidak lengkap. Untuk mengatasi kelemahan pendekatan terdahulu (iaitu morfologi dan pengekodan DNA) dalam menggambarkan kepelbagaian diversiti yang menyeluruh bagi kawasan kajian, teknologi pemantauan yang lebih maju - pengekodan DNA persekitaran (eDNA) bersama penjujukan pemprosesan-tinggi (HTS) telah digunakan sebagai pendekatan pemantauan biodiversiti yang lebih berpatutan dari segi kos, pantas dan tidak invasif. Ujian metabarkod eDNA mengesan ~82% daripada famili ikan yang direkodkan semasa tinjauan konvensional di Muara Merbok sepanjang dekad yang lalu. Lebih >100 spesies lagi (iaitu spesies residen, migran atau pelawat yang kerap, sebahagian daripadanya baru direkodkan) yang tinggal di Muara Merbok telah dikesan dalam tempoh masa dua hari sahaja. Teknik metabarkod eDNA juga mengesan beberapa taksa yang penting untuk pemuliharaan, lantas memberi penekanan terhadap penggunaan metabarkod eDNA sebagai alat penilaian biodiversiti yang berkesan ke arah pemuliharaan holistik muara bakau tropika. Secara keseluruhannya, kajian ini mengesahkan bahawa kepelbagaian ikan di Muara Merbok amat tinggi dan memerlukan kajian yang lebih intensif serta meluas untuk mendokumentasikannya.

FISH IDENTIFICATION ACROSS MERBOK ESTUARY THROUGH DNA BARCODING AND ENVIRONMENTAL DNA (eDNA) METABARCODING IN TROPICAL MANGROVE ESTUARY

ABSTRACT

Biodiversity hotspots often face a taxonomic information gap, especially those biotas found in tropical mangrove estuarine ecosystems. These megadiverse ecosystems are currently among the most threatened ecosystems in the world due to high human pressures. Moreover, knowledge of the taxonomy and distribution of species in this ecosystem is still incomplete for many groups, especially fish taxa, which hinders the implementation of sustainable biodiversity conservation plans. To fill the taxonomic gap in this species-rich region, this work investigated the diversity of fish communities in a tropical mangrove estuary - the Merbok Estuary in northwestern Peninsular Malaysia using conventional and DNA-based methods. The conventional assessment of ichthyodiversity in the Merbok Estuary and adjacent marine waters recorded a total of 138 fish species from two classes, 18 orders, 47 families, and 94 genera. A repository of specimens was established at the Biodiversity Reference Laboratory, USM, to enable further comparative studies. Subsequently, DNA barcoding based on the COI gene was used to genetically assess the fish diversity in the Merbok Estuary and complement the earlier morphological results. A DNA barcoding reference library of 350 fish individuals was compiled. Of the 138 species initially identified by morphology, the barcodes of 123 species confirm their validity. The barcoding study has also revealed hidden diversity within seven species, while the divergence between two pairs of valid morphological species is below the interspecific threshold, necessitating further taxonomic studies. A comparison with previous

species lists in and around this region shows that the taxonomic coverage in the Merbok Estuary is certainly not complete. To overcome the limitations of the previous approach (i.e., morphology and DNA barcoding) in describing the overall diversity in the study area, an advanced monitoring technology - environmental DNA (eDNA) metabarcoding coupled with high-throughput sequencing (HTS) was used to provide a cost-effective, rapid, and non-invasive approach to monitoring species diversity. The eDNA metabarcoding assays (i.e., COI and 12S rRNA gene) detected ~82% of the fish families recorded in conventional surveys in the Merbok Estuary over the last decade. A further > 100 species (i.e., residents, migrants, or frequent visitors, some of which were newly recorded) living in the Merbok Estuary were detected in just two days. The metabarcoding assays also detected a few taxa of conservation, highlighting the value of eDNA metabarcoding as an effective biodiversity assessment tool and a promising step towards the holistic conservation of tropical mangrove estuaries. Overall, this study confirms that fish diversity in the Merbok Estuary is extremely rich and requires more intensive and extensive studies to fully document it.

CHAPTER 1

INTRODUCTION

1.1 Overview of study

Biodiversity stands for the variety of life that can be measured at different levels (e.g., genetics, species, and ecosystem) and at different scales (spatial and temporal). Biodiversity constitutes all terrestrial, freshwater, and marine organisms, including animals, plants, fungi, and microorganisms, as well as their variation at genetic, community, and ecosystem levels. The richness, composition, and interactions of an organism with other organisms, as well as abiotic variables, can all be measured in terms of biodiversity richness, composition, and interactions (Cardinale et al., 2012). Changes in these components can affect the ecosystem's resilience and resistance to environmental change (Cardinale et al., 2012). Scientists believe that we are currently facing the sixth mass extinction, with significant biodiversity collapse (Barnosky et al., 2011). This devastating circumstance is mainly due to anthropogenic activities, and the principal reason is associated with changes in land use (Foley et al., 2005; Sala et al., 2000). Trends in biodiversity and ecosystem services have been analysed for terrestrial, freshwater, and marine ecosystems, and observed scenarios consistently point to a significant decline in global biodiversity in the 21st century (Pereira et al., 2010). The livelihoods of people who rely on ecosystem goods and services may be adversely affected if biodiversity declines significantly (Worm et al., 2006). If the trend of decline continues, half of the species on earth could likely be gone forever by 2050 (Leakey & Lewin, 1996; Thomas et al., 2004).

Although land-use change is the leading cause of threats to global biodiversity, several other causes have been identified, including climate change, changes in

atmospheric CO_2, biotic exchange, and nitrogen deposition (Sala et al., 2000). It is an indisputable fact that anthropogenic impacts on the earth's entire biosphere are now so great that scientists are debating the concept of a new geological epoch shaped by human activity, namely the Anthropocene (Zalasiewicz et al., 2011). Analyses by Butchart et al. (2010) have shown that biodiversity has continued to decline worldwide over the last four decades (1970 - 2010) and that we are currently living in the midst of a global wave of anthropogenically induced biodiversity loss (Dirzo et al., 2014). The species extinction scenario is so common that it is currently seen as a limitless form of global crisis (Brook et al., 2008). Although species extinction has always be perceived as a natural phenomenon, it is currently occurring at an unnaturally rapid rate as a result of anthropogenic interventions. Unfortunately, we have already doomed thousands, perhaps millions, of species to extinction. Countless species are disappearing unnoticed before they have even been described by science. In Southeast Asia alone, three plants and eight animal species were classified as 'extinct' by the International Union for the Conservation of Nature and Natural Resources (IUCN) in the 2000s (IUCN, 2003). The rapid loss of biodiversity primarily affects human health and the sustainable future of our planet.

Southeast Asia is widely known as one of the most biodiverse regions on earth. Approximately 20% to 25% of the world's plant and animal species are found in this area, with a high degree of endemism (Gaither & Rocha, 2013; Woodruff, 2010). As one of the four biodiversity hotspot countries in Southeast Asia, Malaysia is also affected by the agonising threat of biodiversity decline (Sodhi et al., 2004). Malaysia is known for its highly endemic ichthyofauna. However, there are only a limited number of comprehensive studies on the biodiversity of one of the most biodiverse countries in the world, especially on fish taxa. Mohsin and Ambak (1996) reported a

total of 710 species of coastal fishes, and Kottelat and Whitten (1996) recorded more than 600 freshwater fish species in this region. In a more recent assessment, Chong et al. (2010) reported a total of 1418 marine and brackish water fish species in Malaysian waters inhabiting various coastal habitats, including the threatened mangrove ecosystems. This number is believed to be an under representation, with the main concern being that more and more species are disappearing even before they are described by science. The threatening impacts are greater in biodiversity hotspots of Southeast Asia (Myers et al., 2000).

Estuaries and coastal wetlands with mangrove ecosystems serve as transition zones between terrestrial, freshwater habitats and the sea (Levin et al., 2001). Estuaries and their surrounding areas are subject to significant cyclical fluctuations in environmental parameters and strong gradients between freshwater and marine environments, shaping the ecosystem's rich biodiversity (Chabrerie et al., 2001). The coastal ecosystem (i.e.,, the estuary in conjunction with the mangroves) provides essential ecosystem services, including coastal protection, nutrient production, and fisheries resources (Wagner & Sallema-Mtui, 2016). The Merbok Estuary in northwestern Peninsular Malaysia hosts one of the largest remaining mangrove forests in the region (~40,000 ha) (Fatema et al., 2014; Ong et al., 2015). In 1951, the Merbok Estuary was designated as a permanent forest reserve - Sungai Merbok Mangrove Forest Reserve. Numerous research studies, including several biodiversity inventories (Hookham et al., 2014; Jamaluddin et al., 2019; Mansor et al., 2012), have been conducted in the Merbok Estuary to determine the value and critical role of this area in providing vital ecosystem services and livelihoods for local people. These studies have demonstrated the great dependence and its importance in supporting the socioeconomic activities of local communities (Sarathchandra et al., 2018).

Unfortunately, this important relationship between humans and nature is often threatened by habitat pollution, destruction, and overfishing (Brown et al., 2019). It is also affected by other factors such as species invasion and climate change (Levin et al., 2001). Knowing these threats and the causes of aquatic biodiversity loss in this important ecosystem, immediate action must be taken to address these resilient impacts. Most countries have enacted regulations to protect endangered species and their habitats. These have been developed by gathering information on the distribution, diversity, and biology of species and conducting regular biological monitoring (biomonitoring) or biodiversity assessments.

In order to understand the biodiversity of a particular ecosystem, assessments are crucial to identify the essential components of the ecosystem. Biodiversity assessments or surveys are critical for monitoring and evaluating the health of ecosystems and the species living in them. Biodiversity assessments are usually conducted using traditional identification methods that classify morphological features and use taxonomic keys to identify species. Morphology-based biodiversity assessment can be invasive, time-consuming and financially expensive. In addition, availability of experienced taxonomists and specially developed keys for species identification are the main challenges in this assessment method. Misidentification of species would eventually lead to inaccurate descriptions of species distribution. It is impossible to develop conservation and long-term management plans for an ecosystem without knowing the species composition. Currently, there are several standard biological monitoring methods that use a specific design for a particular group of organisms that includes a combination of observation and invasive trapping methods. However, these assessment methods are constrained by high costs, a limited number of trained workers, and a significant time commitment (Darling & Mahon, 2011).

Similarly, the invasive nature of conventional capture-based surveys increases the likelihood of predation risk to the organism and threatens the overall ecosystem (Shaw et al., 2017). In addition, the decline in the number of taxonomic experts and the non-standardised qualifications of different taxonomists may reduce the efficiency of conventional methods based on morphological identification and lead to assessment biasness (Shaw et al., 2016).

Biodiversity assessment is undeniably the central aspect of conservation biology. Existing biodiversity assessment methods need to be improved in order to monitor and protect all biodiversity in general, especially with regard to the future of biodiversity and aquatic ecosystem resources. To achieve this, more detailed, comprehensive, and rapid surveys are essential. However, such advances are not possible with traditional methods without increasing costs, time and effort. Therefore, time- and cost-efficient tools need to be developed to overcome conventional biodiversity assessment limitations and achieve satisfactory taxonomic resolution.

DNA barcoding and environmental DNA metabarcoding are among the tools used in biodiversity assessment that have minimal impact on the environment (Bohmann et al., 2014; Schnell et al., 2012). DNA barcoding involves sequencing short fragments of DNA - the mitochondrial (mt) gene cytochrome c oxidase subunit 1 (COI) - to identify a taxon by referencing it to an established database (Hebert et al., 2003a). It is a molecular tool that provides the species (individual specimens) under study with a unique identifier (DNA barcode). Environmental DNA (eDNA) metabarcoding, on the other hand, is a relatively new molecular method used to characterise biological taxa using DNA found in an environmental sample (e.g., water, soil, permafrost) (Dejean et al., 2012). While the concept of DNA barcoding for taxa identification is well established, the use of eDNA has in recent years gained

popularity due to its effectiveness in identifying taxa from bulk environmental samples. Environmental DNA mixtures can consist of DNA from multiple taxa of all life stages, such as vertebrates, invertebrates, bacteria, or algae, and from various sample types such as sediments, soil, faeces, or marine and freshwater (Dejean et al., 2012). This advanced approach extends the concept of DNA barcoding by analysing environmental samples to determine species composition within a sample. Since the first metabarcoding study by Giovannoni et al. (1990), numerous research efforts have been conducted to identify multiple taxa through the metabarcoding method. Both eDNA metabarcoding and DNA barcoding methods should be used as a time- and cost-efficient bioassessment tool that can complement traditional methods of biological surveys. These advances will enable researchers and natural resource managers to better understand, manage and conserve global biodiversity more efficiently.

1.2 Problem statement

The loss of biodiversity in highly diverse mangrove-estuarine ecosystems is of great concern. The ichthyofauna of this critical ecosystem is of great importance for conservation and biological research. However, knowledge of the taxonomy and distribution of species in this ecosystem is still incomplete for many groups, especially fish taxa, which arguably hinders the implementation of sustainable biodiversity conservation plans. This scenario is exacerbated by the difficulty of reliably distinguishing fish taxa in megadiverse faunas, as there are many closely related and physically identical species. Given that many species disappear every year and the vast majority of those that remain cannot be identified, biodiversity monitoring is undoubtedly a discipline of great importance. Only through taxonomy can we gain a fundamental understanding of biodiversity and make informed decisions about its

management. To fill the taxonomic gap in this species-rich ecosystem, this work investigated the diversity of fish communities in a tropical mangrove estuary - the Merbok Estuary based on two methods: (1) conventional methods using morphological identification and (2) DNA-based methods using DNA barcoding and environmental DNA (eDNA) metabarcoding. In this thesis, conventional and DNA-based approaches are applied as effective biomonitoring tools to assist in the management and conservation of endangered and elusive taxa, including several commercially important species in Malaysian fisheries.

1.3 Thesis objectives

The main objective of this work is to assess the diversity of the mangrove-associated fish community in the study area through morphological and DNA-based approaches (DNA barcoding) and to develop a comprehensive DNA barcoding database (i.e., COI sequences). This objective will be extended through the application of the revolutionary bioassessment tool - environmental DNA (eDNA) metabarcoding coupled with high-throughput sequencing (HTS). Nevertheless, detailed objectives of this work are as follows:

1. To morphologically identify and build a checklist of fish communities in Merbok Estuary and its adjacent waters.
2. To assess the fish diversity of Merbok Estuary and its adjacent waters through DNA-taxonomy (DNA barcoding) and establish a localised DNA barcode library.
3. To assess the richness and diversity of fish communities in ecologically diverse landscapes in Merbok Estuary using eDNA metabarcoding and describe the

advantages and limitations of an eDNA metabarcoding survey in characterising biodiversity in a tropical mangrove estuary.

1.4 Thesis organisation

In total, four manuscripts have been prepared in the structuring of this thesis (submission status: in-review; accepted; or published). In order to consummate all of the objectives (as in 1.3), the thesis is divided into the following six chapters.

This introductory chapter, Chapter 1 provides a brief overview of the study, problem statement, research objectives, and thesis outline.

Chapter 2 provides a comprehensive review of the previous and current literature on molecular methods for biodiversity assessment. This chapter focuses on reviewing diversity assessment studies in aquatic ecosystems that used DNA barcoding and eDNA metabarcoding of various taxa, emphasizing on fish diversity. This chapter also introduces the selected study site, the Merbok Estuary, and the development of the eDNA metabarcoding method in biodiversity surveys using high-throughput sequencing (HTS). Part of this chapter has been published in Publication I: "Environmental DNA (eDNA) metabarcoding as a sustainable tool of coastal biodiversity assessment", doi: https://doi.org/10.1007/978-3-030-15604-6_14 (Zainal Abidin & Noor Adelyna, 2020) (Appendix A).

Chapter 3 focuses on the identification of the fish communities that occur in the study area of the Merbok Estuary and adjacent waters using a morphological approach. Fish samples are identified to the most precise taxonomic unit using established keys, and their diversity is assessed. A checklist of species is compiled, and a dedicated tissue and specimen collection section is established at the Biodiversity

Reference Laboratory, Universiti Sains Malaysia in Penang to facilitate further comparative studies. This chapter has been published in Publication II: "Ichthyofauna of Sungai Merbok Mangrove Forest Reserve, northwest Peninsular Malaysia, and its adjacent marine waters", doi: https://doi.org/10.15560/17.2.601 (Zainal Abidin et al., 2021a) (Appendix B).

Chapter 4 describes the reliability of the DNA barcoding method as one of the molecular tools for species identification and inventorying the biodiversity of the study area. In this chapter, a DNA barcoding reference library of fishes from the Merbok Estuary and its adjacent waters is compiled to describe the fish diversity in the study area and to provide a complementary view to the previous morphology-based results (Chapter 3). The output of this chapter has been published in Publication III: "DNA-based taxonomy of a mangrove-associated community of fishes in Southeast Asia", doi: https://doi.org/10.1038/s41598-021-97324-1 (Zainal Abidin et al., 2021b) (Appendix C).

Chapter 5 focuses on the utilisation of a time and cost-efficient bioassessment tool, eDNA metabarcoding, to complement the traditional method of biodiversity survey. The performance of eDNA metabarcoding in characterising fish diversity in the Merbok Estuary is evaluated, and community analyses are conducted using the robust data generated by HTS. This chapter details the field techniques (i.e., the collection and the processing of the aqueous environmental samples prior to the genomic work), the laboratory procedures (i.e., the pre-and post-sequencing), and the bioinformatics pipeline used to analyse the HTS data. Futhermore, the advantages and limitations of an eDNA metabarcoding study of the fish community in the Merbok Estuary are discussed. Findings have been submitted for publication and is currently under revision with the Scientific Reports journal as Publication IV: "Assessing a

Megadiverse but Poorly Known Community of Fishes in a Tropical Mangrove Estuary through Environmental DNA (eDNA) Metabarcoding". The manuscript preprint is available on Researchsquare (https://doi.org/10.21203/rs.3.rs-1350797/v1).

Finally, Chapter 6 summarises the main findings and research outcomes from the previous chapters. This chapter provides a summary of the dissertation results regarding the further development of DNA-based methods in the study of biodiversity in an aquatic ecosystem. Recommendations for future studies on the implementation of DNA-based approaches in routine biodiversity monitoring are provided in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Tropical mangrove estuaries

Estuaries are dynamic transitional environments which are ecologically more complex than freshwater or marine ecosystems (Day et al., 2012). The biodiversity within this system is influenced by river runoff, which transports enormous amounts of nutrients and organic matter, and by marine waters, which allow water renewal, hence the significant fluctuations in salinity (Pasquaud et al., 2015). This ecosystem requires an exhaustive sampling approach to record all its biota and accurately understand their ecological relationships. The biodiversity in estuaries is typically under-explored, leading to poor resource management and a restricted understanding of estuarine environments.

The most important type of vegetation that maintains biodiversity in tropical estuaries is mangroves, forested wetlands situated in the intertidal zone (Twilley et al., 1996). These are areas where the water is turbid and contains fine sediments. The vegetation grows at an incredible rate, and the inhabitants are predominantly deposit feeders (Day et al., 2012). Mangroves make up only a small part of the world's forested environment, but they cover 240,000 km^2 of protected subtropical and tropical coastlines (Twilley et al., 1992). According to FAO (2007) about 6.8 million ha (~34-42%) of the global mangrove area was in Southeast Asia in the 1980s. Since then, this region has lost the most mangroves, around 1.9 million ha, mainly due to the conversion of mangrove forests to other land uses (FAO, 2007). However, the annual loss of mangroves has slowed down, from about 187,000 ha in the 1980s to 102,000 ha in the 2000s (Jusoff, 2013). The largest mangrove areas in Southeast Asia are in

Indonesia (almost 60% of the total area of Southeast Asia), second is Malaysia (~12%), followed by Myanmar (~9%), Papua New Guinea (~9%) and Thailand (~5%) (Jusoff, 2013). Mangrove forests in Malaysia cover about 577,500 ha, with Sabah having the largest at 59% (341,000 ha) of the country's total area, followed by Sarawak with 132,000 ha (23%) and Peninsular Malaysia with 104,200 ha (18%) (Jusoff, 2013).

Estuaries associated with mangroves act as boundaries protecting coastal land from destruction by ocean waves, tsunamis, and storms (Zhang et al., 2012). This valuable ecosystem also provides critical habitat for many recreational and commercial species (e.g., feeding, spawning, mating and nursery grounds) and is characterised by high oceanographic variables (Barbier et al., 2011; Twilley et al., 1996). Tropical mangrove estuaries serve as spawning grounds for a significant proportion of Malaysia's commercial and recreational fisheries, including the sea bass (*Lates calcarifer*) and banana prawn (*Penaeus merguinensis*) (Sasekumar et al., 1992). About 75% of Malaysia's commercial and recreational fishes and shrimps spend at least part of their lives in the mangrove system (Sasekumar et al., 1992). Tropical coastal and estuarine areas are inextricably linked to mangroves, with high fish diversity and complex interactions. An assessment by Blaber (2008) discovered that a single mangrove estuary in the tropical Indo-West Pacific could harbour more than 200 different fish species, while the tropical East Atlantic and Neotropical estuaries support a respectable number of 100 different fish species. These ecosystems are among the most productive on the planet and are therefore heavily exploited by humans (Costanza et al., 1997).

Despite the significant importance of mangroves in providing ecosystem services and ensuring the livelihoods of local communities, this vital ecosystem is threatened by anthropogenic activities and natural impacts. Human activities pose six

major threats to mangroves: (1) conversion to other uses, (2) overharvesting, (3) overfishing, (4) pollution, (5) sedimentation, and (6) changes in flow regimes (Gilman et al., 2008; Spalding, 2010). Together, these six factors constitute the greatest threat to mangroves. Of these, direct conversion to other uses (e.g., urban and industrial areas, aquaculture, and agriculture) is the most critical factor in the destruction of mangroves worldwide (Gilman et al., 2008). Other natural factors that affect mangroves in Peninsular Malaysia include coastal erosion and marine incur notably the devastating tsunami in December 2004 (Jusoff, 2013).

2.1.1 Merbok Estuary as an ideal conservation site of a tropical mangrove estuary

The mangrove forests in Peninsular Malaysia are mainly found along the sheltered coasts, estuaries, rivers, and some offshore islands. One of the largest intact mangrove forests is located in the Merbok Estuary, northwest of Peninsular Malaysia, facing the Strait of Malacca (Figure 2.1). This estuary is associated with a main river that stretches for about 35 km (Mansor et al., 2012a). The width of the estuary varies from about 20 m in the upper reaches to 2 km near the mouth of the open ocean (Mansor et al., 2012a; Ong et al., 1991). The estuary is also fed by smaller rivers with a depth of 3 to 15 m (Ong et al., 1991). The annual rainfall in this area ranges from 200 cm to 250 cm, with the primary maximum rainfall in September - November and the primary minimum in January - February (www.met.gov.my). The estuary was designated as a permanent forest reserve known as the Sungai Merbok Mangrove Forest Reserve, in 1951, and it is the second-largest mangrove forest in Peninsular Malaysia after the Larut Matang Forest Reserve in Perak. The Merbok Estuary and its surroundings constitute a dynamic and productive ecosystem that harbours the world's highest

mangrove species diversity per unit area in a contiguous habitat, with 39 of the world's estimated 70 true mangrove species described (Ong et al., 2015). This area is also an important source of raw materials for the local population. A large part of the population lives in the adjacent estuaries, as these areas offer greater socio-economic potential (Jamaluddin et al., 2019).

Estuaries must be managed appropriately to ensure their preservation. Therefore, a scientific understanding is crucial for the practical and sustainable management of this vital ecosystem incorporating knowledge of geology, hydrology, chemistry, physics and biology (Day et al., 2012). In particular, the Merbok Estuary is characterised by its unique morphology and biodiversity and has been the focus of much scientific interest since the 1990s (Fatema et al., 2014; Lim et al., 1995; Ogawa, 2003; Ong et al., 1991; Ong et al., 2015). Recognising its value and critical role in providing vital ecosystem services and livelihoods for local people, numerous biodiversity inventories have been conducted in the Merbok Estuary (Hookham et al., 2014; Jamaluddin et al., 2019; Mansor et al., 2012a; Mansor et al., 2012b). Hookham et al. (2014) reported the diversity of mangrove trees and their associated gastropods in Sungai Merbok. They assessed the gastropod diversity in the habitat as bioindicators to measure the impact of anthropogenic disturbances on the mangrove ecosystem. A study conducted by Jamaluddin et al. 2019 recorded at least 18 species of shrimps, demonstrating the potential of this mangrove habitat in harbouring immense aquatic biodiversity. Another comprehensive diversity inventory was done by Mansor et al. (2012a); Mansor et al. (2012b). The authors documented up to 81 species (from 36 families) of fish inhabiting the Merbok Estuary and identified the ecosystem as an important nursery ground for fish and shrimp. They also evaluated its utility as an ideal habitat for bivalves, especially oysters and clams (Mansor et al., 2012a). However,

according to their assessment, the Merbok Estuary is polluted by sewage and pesticides from nearby agricultural activities, aquaculture, solid wastes deposited from residential areas, and fish wastes (i.e., trash fish - juveniles or small fish) from local fishing activities (Mansor et al., 2012a).

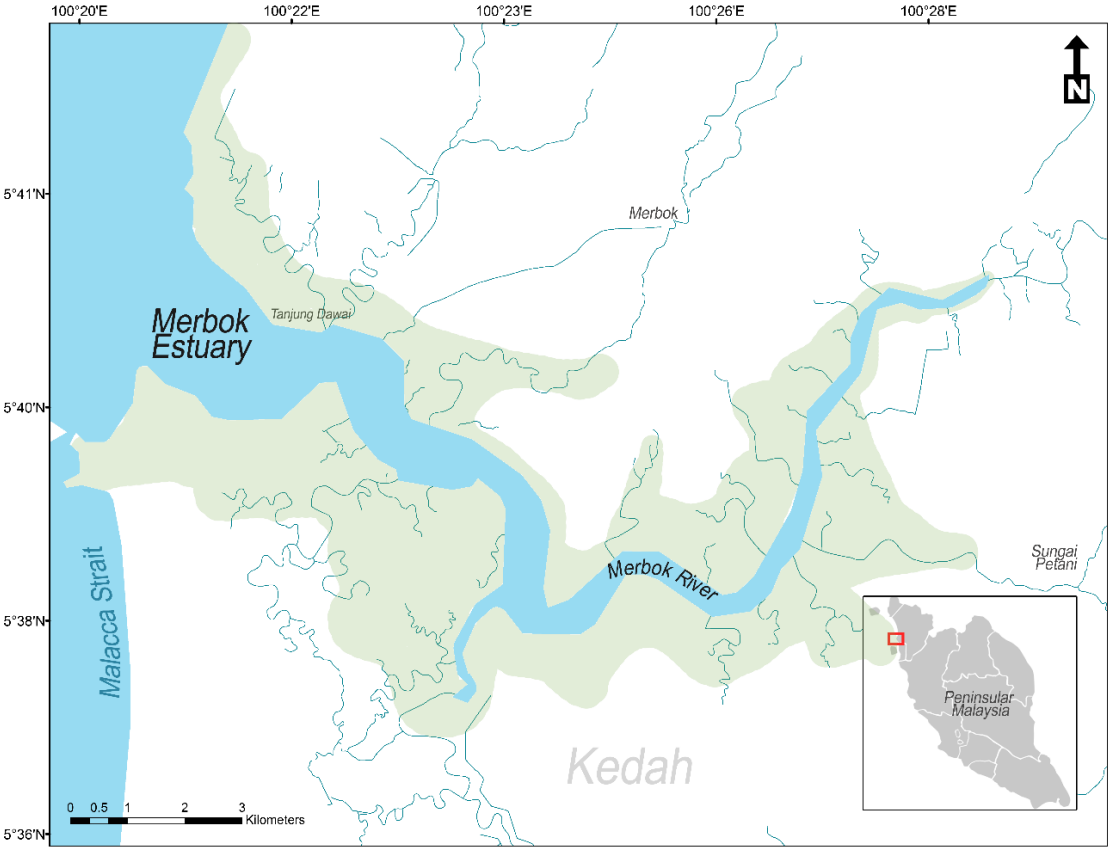

Figure 2.1: The Merbok Estuary as an important mangrove study site. The green shaded region represents the mangrove areas. Inset map shows the location of the study area within Peninsular Malaysia. Map is generated using ArcMap 10.8 and further edited in Adobe Photoshop CC 2019.

2.2 Characterising biodiversity hotspot through biodiversity surveys

Conservation efforts are critical in megadiverse environments like tropical mangrove estuaries, where biodiversity is particularly imperilled. Its primary residents and frequenters, specifically the numerous fish species are a substantial component of biodiversity in this ecosystem. Half of the world's reported species of vertebrates are fishes, with wide array of diversity spanning more than 35,900 species (Fricke et al., 2021; Nelson et al., 2016). Malaysia is located in the biodiversity hotspot of Sundaland and hosts an extraordinary level of diversity and endemism (Myers et al., 2000). For example, Chong et al. (2010) reported the occurrence of more than 1,400 marine and brackish water species in the Malaysian coastal waters, a significant proportion of which live in or frequent the mangrove ecosystems. However, human interference threatens fish biodiversity in the fragile aquatic ecosystem. Even though approximately 400 new fish species have been discovered in the last two decades (Fricke et al., 2021), anthropogenic impacts, such as overfishing, pollution, and habitat degradation have significantly caused a catastrophic loss of fish diversity (Shelton et al., 2018). The decline in fish diversity could be exacerbated and slow to recover if caused by factors such as climate change, eutrophication or the invasion of alien species (Xiong et al., 2022). There is cause to fear that species extinctions would exceed discoveries, particularly in diverse and fragile ecosystems like tropical mangrove estuaries. Therefore, biodiversity surveys through accurate species identification are key to effective conservation plans, especially in this crucial ecosystem. Nevertheless, in many taxa including fishes, conventional approaches for species identification (i.e., morphology-based) have proven to be hindered by a few limitations.

While morphology is the core of taxonomic information, it has various limitations: (1) phenotypic plasticity, in which alteration in their physical appearance as a result of environmental influences may result in misidentification; (2) the presence of cryptic or hidden diversity (individuals that are morphologically similar but genetically distinct) can obscure the detection of multiple taxa within a single morphologically recognised taxon; (3) variation in diagnostic keys within a species due to the influence of life stage or gender; and (4) a lack of expert knowledge (i.e., trained taxonomist), which results in misdiagnoses. In classical taxonomic identification, morphometric characteristics (e.g., body form, allometric features, number of lateral lines or fin rays, and/or colour patterns) are used to classify fish species. Unfortunately, morphological characteristics are often unstable during different developmental phases (i.e., larval, juvenile, or adults). Incomplete samples or rare and cryptic species may not be accurately documented. Even when working with fully intact adults, fish identification can be difficult due to the morphological similarity of congeners during their early life cycles and the discrepancies in the extant literature and taxonomic history (Chen et al., 2021). Therefore, DNA-based methods for species identification have been utilised as an effective tool, complementing the conventional methods**.**

2.3 DNA barcoding: DNA-based taxonomy for species identification

Two decades have passed since the inception of DNA barcoding as a rapid and effective tool for species identification and biodiversity assessment (Hebert et al., 2003a). DNA barcoding utilises small genetic identifiers (i.e., DNA barcodes – typically ~650bp fragment of the 5' region of the mitochondrial gene cytochrome oxidase subunit I (COI) in animals) to classify specimens according to a pre-existing cataloguing and reference database. Since its introduction, this method has been proven efficient to distinguish different species. The generation of DNA barcodes (species-specific sequences) provides diagnostic markers that supplement classical morphological taxonomy and accelerate taxonomic identification and discoveries, but they are not intend to replace it (DeSalle et al., 2005). The success of the DNA barcoding process is contingent on the availability of a reliable and comprehensive reference database. Besides a high-quality reference sequence, additional features such as geographical, morphological and taxonomic metadata are deposited in a dedicated DNA barcoding platform (BOLD - http://v4.boldsystems.org/) (Ratnasingham & Hebert, 2007), enabling precise species-level resolution (Hebert et al., 2003a). The fundamental process of DNA barcoding is illustrated in Figure 2.2. Presently, many species of chordates and arthropods (COI gene), plants (rbcL, matK, 18S), as well as cyanobacteria (16S) and fungi (ITS) can be identified using the public library of standardised DNA barcodes. DNA barcoding is employed in many applications, including ecosystem management, biodiversity conservation, population genetics, phylogeographic analysis, invasion control, forensics, and food safety (Cristescu, 2014). DNA barcoding provides several advantages for species identification, including resolving taxonomic identification disputes in the presence of cryptic or sibling species (Hebert et al., 2004a; Krück et al., 2013). In addition, this method can

facilitate the linking of taxonomic knowledge from different life stages to build a complete profile of a species, as it can identify specimens at all life stages and even damaged specimens which are exceedingly difficult, if not impossible, to identify (Taylor & Harris, 2012).

More advanced technique has now been developed to overcome some of the limitations of the traditional DNA barcoding which require fresh or appropriately preserved samples (frozen/alcohol preserved). Shorter DNA sequences (mini-barcodes) are now being used to identify materials that have been severely degraded, such as ancient museum specimens, processed biological materials such as canned/smoked foodstuff or specimens fixed in preservatives not generally amenable to DNA extraction (e.g., formalin) (Hajibabaei & McKenna, 2012; Meusnier et al., 2008). Alternative universal metazoan COI primers that render an ~313bp of the COI region when combined with the established Folmer reverse primer has also been developed (Leray et al., 2013). Consequently, these shorter DNA segments can be analysed using high-throughput sequencing (HTS) systems, enabling the simultaneous identification of vast numbers of species at a lower cost.

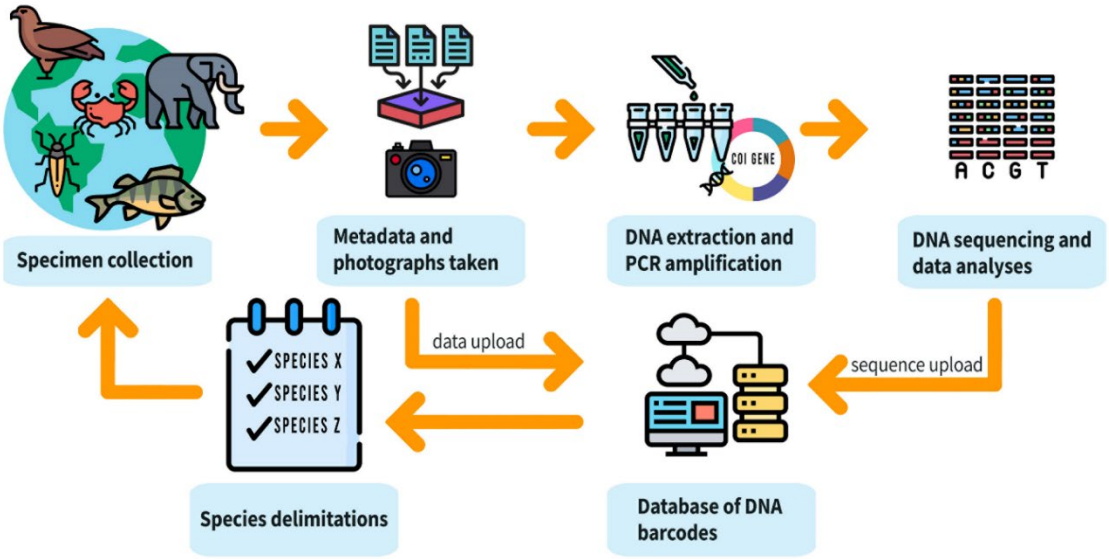

Figure 2.2: Simplified workflow of DNA barcoding – from specimen, to sequence, to species.

However, the rapid advances in DNA barcoding have sparked debate and controversy, especially within the taxonomic community. In the early years after its introduction, the reliability of DNA barcodes over morphological identification was challenged. The use of the COI fragment as a means of species identification was first questioned by Will and Rubinoff (2004), and the concept of DNA barcoding was subjected to various degrees of criticism. Therefore, the "barcoding gap" concept was postulated to increase the trustworthiness of species identification with DNA barcodes (Barrett & Hebert, 2005; Hebert et al., 2004b). The integral principle of DNA barcoding gap relies on the interspecific sequence divergence being higher than the intraspecific divergence (Meyer & Paulay, 2005). The main criteria for selecting COI as the gold standard barcode gene is the characteristic pattern of variation observed across numerous species, with significant divergence and lack of overlap between intraspecific (i.e.,, within the same species) and interspecific (i.e.,, between different species) genetic distances (Hebert et al., 2003b). A significant gap between

intraspecific and interspecific genetic distances is expected in reliable species identification by DNA barcoding (Figure 2.3). The validity of this barcoding gap was initially established in a barcoding study of bird species (Hebert et al., 2004b), where it was discovered that sequence divergence between species was significantly greater than divergence within species. Similar results were also found in the study of Australian fishes (Ward et al., 2005), springtails (Hogg & Hebert, 2004), and spiders (Barrett & Hebert, 2005). In the earliest barcoding study of the fish taxa, a genetic distance threshold of 2% was suggested after analysing 1088 fish species (Ward et al., 2005). Although most interspecies analyses recover a barcode gap, exceptions may occur, especially in recently diverged species (Prosdocimi et al., 2012).

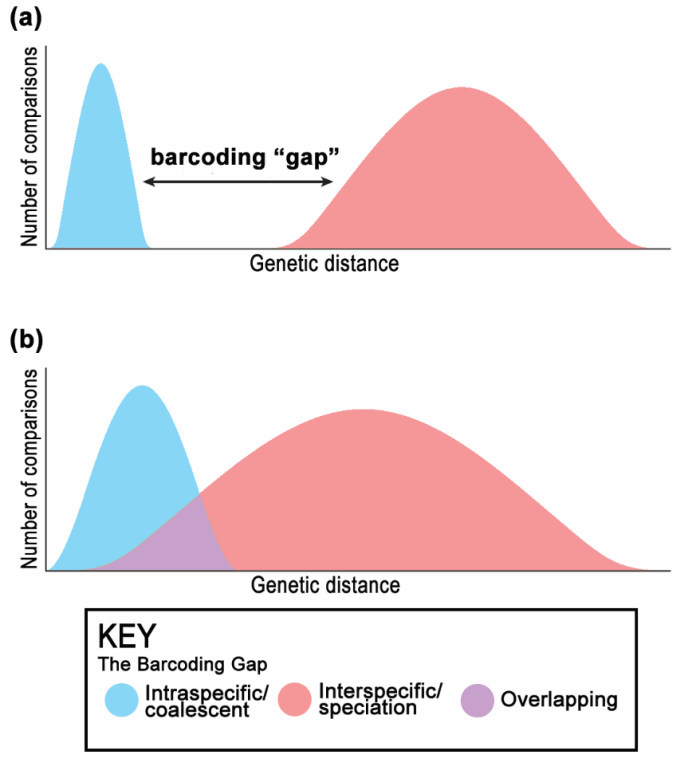

Figure 2.3: Graphical depiction of the barcoding gap concept. (a) the barcoding gap is formed when the intra- and inter-species degrees of divergence are clearly distinguished, and (b) there is no barcode gap with overlapping distributions between intra- and interspecific divergence. Reproduced from Meyer & Paulay (2005).

2.3.1 Bioinformatics of DNA barcoding

The statistical process of DNA barcoding can be categorised into two independent approaches: (1) diagnostic identification of species (i.e., the assignment of an unknown sample (sequence) to a previously described species) and (2) the discovery and identification of a novel sequence for previously unrecorded species (Bucklin et al., 2011). The first strategy presupposes that an individual DNA sequence can be consistently associated with a group of organisms, ideally at the species level within the reference database. However, forming this association could be challenging in some instances for several reasons. These include atypically high intraspecific divergence in the reference sequence and non-parallel gene and species evolutionary history (Bucklin et al., 2011). Most barcoding studies have employed genetic distance analysis to deduce the species identification from the DNA barcodes utilising the NCBI BLAST query (http://blast.ncbi.nlm.nih.gov/Blast.cgi). This technique computes the nearest neighbour of the query sequence using a raw similarity score. While BLAST searches are quick, they possess a few limitations, including a high rate of false-positive results due to the intrinsic flaws of BLAST algorithms (Koski & Golding, 2001). Alternatively, the use of multiple methods and schemes in clustering the generated DNA barcodes provides an efficient approach to identifying putative species.

Presently, there are three different sequence-based methods commonly used to delineate Molecular Operational Taxonomic Units (MOTUs) or putative species based on the analysed DNA barcodes: (1) Refined Single Linkage (RESL), (2) Automatic Barcode Gap Discovery (ABGD), and (3) Generalized Mixed Yule Coalescent (GMYC) (Sholihah et al., 2020; Zainal Abidin et al., 2021). Although these methods

individually may have pitfalls, especially when analysing singletons, in combination they could provide robust results (Ortiz & Francke, 2016). The first analysis, RESL, is performed within the BOLD platform (Ratnasingham & Hebert, 2007) by assigning sequences to specific Barcode Index Numbers (BINs, which are MOTUs). Discordance reports in BOLD were used to compare the BINs with the nominal taxonomy. This method has been shown to outperform several other approaches in terms of taxonomic performance and computational efficiency (Ratnasingham & Hebert, 2013). The ABGD analysis (https://bioinfo.mnhn.fr/abi/public/abgd/abgdweb.html), on the other hand, is computed to census divergence and identify the barcode gap within the analysed species delimitation dataset (Puillandre et al., 2012). Both RESL and ABGD use similar clustering algorithms to distinguish partitions in the genetic divergence of the barcode dataset. However, unlike RESL, ABGD uses an automatic statistical approach to determine an appropriate intra-cluster threshold, resulting in a partitioning that closely resembles the pattern of morphology-based species (Huang et al., 2020). Finally, the GMYC algorithm was developed to delineate species/ MOTU by using a maximum likelihood approach to optimise the shift of gene tree branching patterns from interspecific branches (Yule model) to intraspecific branches (neutral coalescence), thereby identifying sequence clusters corresponding to independently evolving units (Pons et al., 2006). Despite its sound theoretical basis, which includes a rigorous analysis of an ultrametric tree, this approach generally generates more MOTUs than other methods (Hajibabaei et al., 2007), leading to overestimating discovered species/ MOTU.

2.3.2 DNA barcoding in practice

The Consortium for the Barcode of Life Initiative (CBOL) coordinates international efforts to assess global biodiversity through DNA barcoding. Many independent efforts have been conducted in various laboratories globally to establish a comprehensive database of DNA barcodes covering a wide range of taxa. According to the current statistics on BOLD (Ratnasingham & Hebert, 2007), a total of 235,231 animal species have been barcoded from a total of 9,408,667 barcode sequences (as of February 2022). The latest example of the global barcoding initiative is the BIOSCAN project (https://ibol.org/programs/bioscan/), which is expected to provide DNA barcode sequences for over 2 million species (Hobern, 2021; Hobern & Hebert, 2019). Numerous studies have validated the effectiveness of the COI gene as a species identification marker in many groups of taxa such as butterflies (Hajibabaei et al., 2007); spiders (Ashfaq et al., 2019; Barrett & Hebert, 2005); birds (Hebert et al., 2004b; Li et al., 2016); bats (Lim et al., 2017); oysters (Hamaguchi et al., 2017; Suzana et al., 2011); amphibians (Zangl et al., 2020), mammals (Luo et al., 2011); marine invertebrates (Webb et al., 2006); marine and freshwater fishes (Bakar et al., 2018; Fadli et al., 2020; Jaafar et al., 2012; Ward et al., 2008). Biodiversity assessment using DNA barcoding has several advantages, especially in species-rich, difficult-to-access and poorly catalogued habitats. The current loss of biodiversity is particularly pronounced in ecosystem-rich areas (e.g., mangrove, marine, and terrestrial), as many species are expected to become extinct before being taxonomically recorded (Mora et al., 2011). The introduction of DNA barcoding has more or less facilitated in the identification of species to complement the traditional taxonomy.