MOLECULAR IDENTIFICATION, GEOMETRIC MORPHOMETRICS AND PHYLOGENETIC RELATIONSHIP OF COMMERCIALLY IMPORTANT NEMIPTERIDAE FROM MALAYSIAN WATERS AND NEIGHBORING SEAS INFERRED BY mtDNA AND NUCLEAR GENES

by

AYESHA IMTIAZ

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

January 2018

ACKNOWLEDGEMENT

In the name of Allah, most Gracious, most Compassionate

Alhamdulillah, with the doa’ from Mama and Babah, I have completed my PhD thesis with success.

This thesis became a reality also with the kind support and help from many wonderful people and without their support, none of this would be possible. My sincere, utmost delight and profound gratitude goes to my supervisor Dr. Darlina Md. Naim, thank you so much for your continuous guidance, encouragement, absolute patience, invaluable advice, (*occasional nagging*) and especially for believing in me, in which I am able to complete this thesis. May Allah repay you in abundance and bless you with all his grace. I am profoundly grateful to Prof. Dr. Siti Azizah Mohd. Nor, my co-supervisor. Thank you so much for having me in your lab and also for your guidance, invaluable advice and patience especially while looking through my thesis. I am forever grateful to Dr. Aimi and Dr. Tan Min Pao for helping me with the samples providing and the analyses, thank you so much from the bottom of my heart. My special thanks to Dr. Eleanor Adamson from UK, for sharing your data analysis knowledge and expertise during your visit here in USM. Many thanks to Fisheries Research Institutes, Penang and Department of Fisheries Malaysia, Department of Fisheries, Sindh Pakistan for helping me with my sampling.

To my mates of lab 308, thank you so much for teaching me to do the lab work, helping me with my sampling and thesis analyses, companionship over so many lunches and occasional dinner. I will miss our ‘intellectual’ and sometimes silly discussions. From the bottom of my heart I am thankful to Jamsari, Danial,

Faisal, Fong, Hong Chiun, Adibah, Ana, Amirah, Nurul Farhana and everyone that might have cross my path.

My profound gratitude goes to all my in laws family and especially to my husband. Thank you so much for supporting me in my study and especially for putting up with me during the whole 3 years of my study. I extend my appreciation to my beloved Twin Sons, Muhammad Muaz and Muhammad Saad who has sacrificed my time for them during these three years. My deepest thanks to my brothers (Rashid Imtiaz, Umair Imtiaz and Bilal Imtiaz) and the lovely lady – Dr. Afsheen Aman from KIBGE (Karachi Institute of Biotechnology and Genetic Engineering), thank you for your advices and always being there for me. The acknowledgement will remain incomplete without mentioning strong support and prayers of my principal and staff who really deserves the appreciation.

Thank you to TWAS Universiti Sains Malaysia for funding my study under the TWAS fellowship Programme and also for funding my research under the Research University Grant (1001/PBIOLOGI/870018).

TABLE OF CONTENTS

CHAPTER 1 INTRODUCTION

CHAPTER 2 LITERATURE REVIEW

CHAPTER 3 GENERAL LABORATORY METHODS

CHAPTER 4 DNA BARCODING OF THE FAMILY NEMIPTERIDAE IN MALAYSIAN AND SURROUNDING WATERS

LIST OF TABLES

LIST OF FIGURES

Page

LIST OF ABBREVIATIONS

ANOVA	Analysis of variance
BI	Bayesian Inference
bp	Basepair
BOLD	Barcode of life database
COI	Cytochrome oxidase I
CVA	Canonical variate analysis
Cyt b	Cytochrome oxidase b
DA	Discriminant analysis
DNA	Deoxyibonucleic acid
dNTP	Dinucleotide triphosphate
EDTA	Ethylenediamine tetra-acetic acid
ESU	Evolutionary taxonomic unit
GM	Geometric morphometrics
GTR	General time reversible
HKY	Hasegawa-Kishino-Yano
K-80	Kimura-80
MANOVA	Multivariate analysis of variance
MCMC	Marcov Chain monte carlo
mtDNA	Mitochondrial DNA
nDNA	Nuclear DNA
NJ	Neighbor Joining
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
RAPD	Random amplified polymorphic DNA
TNES-Urea	Tris-sodium chloride- EDTA-SDS-Urea

PENGENALPASTIAN MELALUI MOLEKULAR, MORFOMETRIK GEOMETRIK DAN HUBUNGAN FILOGENETIK NEMIPTERIDAE YANG PENTING DARI SEGI KOMERSIAL DI PERAIRAN MALAYSIA DAN PERAIRAN BBERJIRAN (KEJIRANAN) DISIMPULKAN OLEH mtDNA DAN GEN NUKLEAR

ABSTRAK

Kajian ini menggunakan pendekatan genetik dan morfometrik untuk menilai perbezaan molekul dan morfometrik serta hubungan filogenetik spesies komersial yang penting dalam kalangan famili Nemipteridae. Penyiasatan molekul telah dijalankan berdasarkan dua gen mitokondria dan satu gen nuklear pada 210 individu dalam 13 spesies yang diandaikan famili Nemipteridae dari tujuh wilayah laut; Lautan Hindi, Selat Melaka, Laut China Selatan Timur (ESCS), Laut China Selatan Barat (WSCS), Laut China Selatan Utara (NSCS), Laut Sulu (SS) dan, Laut Celebes (CS). Pada bahagian pertama kajian ini, kaedah barkod DNA dengan menggunakan kawasan mitokondria sitokrom oksidase c (COI) (647 bp) telah digunakan untuk mengenal pasti dan untuk menemui potensi spesies baru. Semua sampel dianalisis secara statistik menggunakan perisian MEGA v7.0 dan kesemua sampel terkelompok ke dalam spesies putatif masing-masing. Jarak genetik intraspesifik sebanyak >2% mencadangkan kemungkinan terjadinya spesies kriptik dalam individu *N. japonicus* dan *S. vosmeri*. Perbezaan purata jarak genetik sebanyak 10X telah dicerapi antara purata inter dan intraspesifik pada kesemua taksa kecuali *N. japonicus* dan *S. vosmeri*. Pada bahagian seterusnya dalam kajian ini, data kuantitatif berdasarkan teknik geometrik morfometrik pada 19 petanda homologus ke atas 150 individu yang merangkumi 13 spesies telah dikaji dengan lebih mendalam dengan menggunakan Analisis Variat Pelbagai (MANOVA) dalam pelbagai program tps dan perisian Morpho J. Analisis Komponen Prinsipal (PCA) telah menjana 34 komponen dengan nilai eigen yang sangat rendah iaitu <1, dan oleh itu boleh diabaikan. Walau bagaimanapun, Analisis Variat Berkanonik (CVA) telah menghasilkan 12 variat kanonik, sembilan variat mempunyai nilai eigen >1, CV1 dengan nilai eigen tertinggi menjelaskan 76% adalah variasi bentuk jasad dan boleh mendiskriminasi secara berkesan dalam kalangan 13 taksa yang dianalisis. Walau bagaimanapun, tiada penunjuk spesies kriptik telah dikesan dalam *Nemipterus japonicus* dan *Scolopsis vosmeri* melalui pendekatan ini berbanding dengan analisi barkod DNA. Dalam bahagian akhir, analisis data gen secara individu dan data gabungan daripada dua gen mitokondria (COI dan Cyt b) dan satu gen nuklear (RAGI) ke atas 210 individu dalam 13 spesies telah dianalisis dalam perisian MEGA v7.0 dan perisian MrBayes untuk menjana pohon Kebolehjadian Maksimum dan pohon Inferens Bayes untuk menyelesaikan hubungan filogenetik famili Nemipteridae. Keputusan menunjukkan bahawa semua taksa daripada famili ini secara asalnya adalah monofiletik. Kedua-dua topologi pohon ML dan BI adalah sama dengan hanya sedikit perbezaan pada kedudukan taksa. Set data molekul untuk semua gen (set data individu dan gabungan) jelas menunjukkan bahawa genus *Nemipterus* dan *Pentapodus* adalah berkait rapat manakala genus *Scolopsis* berkait rapat dengan *Parascolopsis.* Gabungan ciri meristik dan ketiga-tiga gen telah mengesahkan bahawa beberapa individu spesies *S. vosmeri* dari Lautan India dan populasi *N. japonicus* dari Laut Cina Selatan adalah spesies kriptik. Data gabungan juga jelas menunjukkan kumpulan kriptik *N. japonicus* dan *S. vosmeri* dan juga mencadangkan rekod baru untuk *S. torquata,* salah satu ahli dalam kompleks *S. vosmeri* di Lautan Hindi dan

Laut China Selatan. Kami mendapati bahawa morfometrik geometri bersamaan dengan barcoding DNA dari segi pengenalpastian spesies. Kegagalan kaedah morfometrik geometrik berbanding kejayaan analisis molekul untuk menemui kepelbagaian spesies kriptik boleh dikaitkan dengan kadar tekanan pemilihan berlainan antara gen dan ciri fenotip dan ekpresi gen (mutasi) tertunda sebagai tindakbalas terhadap perubahan sifat morfologi. Gabungan morfometrik dan data molekul dari kajian ini telah memberikan maklumat yang berguna untuk pengurusan dan pemuliharaan spesies yang penting dari segi komersil dari famili Nemipteridae di perairan Malaysia and perairan kejiranan dengannya.

MOLECULAR IDENTIFICATION, GEOMETRIC MORPHOMETRICS AND PHYLOGENETIC RELATIONSHIP OF COMMERCIALLY IMPORTANT NEMIPTERIDAE FROM MALAYSIAN WATERS AND NEIGHBORING SEAS INFERRED BY mtDNA AND NUCLEAR GENES

ABSTRACT

This study employed genetic and morphometric approaches to assess the molecular and morphometric differentiation as well as phylogenetic relationships among commercially important species of family Nemipteridae. Molecular investigations were conducted based on two mitochondrial and one nuclear genes on 210 individuals within 13 presumed species of family Nemipteridae from seven marine regions; Indian Ocean, Straits of Malacca, East South China Sea (ESCS), West South China Sea (WSCS), North South China Sea (NSCS), Sulu Sea (SS) and Celebes Sea (CS). In the first part of the study, the DNA barcoding method was employed using the standard region of barcoding mitochondrial cytochrome oxidase c (COI) (647 bp) to identify and potentially discovered new species. All samples were statistically analysed in MEGA v7.0 and clustered into fourteen respective putative species. Intraspecific genetic distance of > 2% suggested the potential occurrence of cryptic species within the presumed *N. japonicus* and *S. vosmeri* individuals. An average of 10X differences was observed between mean inter and intra specific genetic distance among all taxa except *N. japonicus* and *S. vosmeri*. In the next part of this study, quantitative data based on geometric morphometrics technique of 19 homologous landmarks were analyzed on 150 individuals within 13 presumed species with Multivariate Analysis of Variance (MANOVA) in various tps programs and Morpho J software. Principal component analysis (PCA) generated 34 components with very low eigenvalue of < 1 and thus of negligible significance. However, Canonical Variate analysis (CVA) generated 12 canonical variates, nine had eigenvalue of >1 with the highest eigenvalue from CV1 that explained 76% of body shape variations and could efficiently discriminate among the 13 taxa analysed. However, no indication of cryptic species was detected in *Nemipterus japonicus* and *Scolopsis vosmeri* in this approach in contrast to the DNA barcoding analysis. In the final section, individual gene analysis and combined (concatenated) data from two mitochondrial (COI and Cyt b) and one nuclear gene (RAGI) were analyzed on 210 individuals within 13 presumed species using Maximum likelihood (ML) and Bayesian Inference methods (BI) to resolve the phylogenetic relationships of the family Nemipteridae. The results showed that all taxa of this family are monophyletic in origin. Both ML and BI tree topologies were similar with only slight differences in positioning among taxa. The molecular dataset of all genes (individual and combined datasets) clearly showed that genera *Nemipterus* and *Pentapodus* are closely related while genus *Scolopsis* is closely related to *Parascolopsis.* Combined meristic characters and the three genes confirmed that several individuals of *S. vosmeri* from the Indian Ocean and *N. japonicus* residents of South China Sea were cryptic. The concatenated data also highlighted the cryptic groups within *N. japonicus* and *S. vosmeri* and also suggested a new record of the newly described *S. torquata,* a member in *S. vosmeri* complex in the Indian Ocean and South China Sea. We found that the geometric morphometrics corresponded with DNA barcoding in terms of species identification. The failure of geometric morphometrics versus success of molecular data to uncover cryptic species diversity may be attributed to different rates of selection pressure between genes and phenotypic characters and also delayed gene (mutations) expression in response to changes in morphological

traits. The combination of morphometric and molecular data from the current study has provided beneficial information for the management and conservation of commercially important species of family Nemipteridae in Malaysian waters and neighboring seas.

CHAPTER 1

GENERAL INTRODUCTION

1.1 Introduction

Malaysia is a biodiversity hotspot with large numbers of endemic and unique fish diversity. Above all, being part of the Indo Malay Archipelago and Sundaland, the Malaysian waters are diverse in both marine as well as freshwater fish fauna. Marine fish is the main protein source of food among peoples in Malaysian region, harvested throughout the year and contributing to 75% of the annual fish catch (Iliyasu *et al*., 2016), Malaysia becomes the 12th major global contributor with 1,458,126 tons of marine fish (FAO, 2016). Tan *et al*. (2015) reported that household use of fish products is 24.7% in Malay communities, 7.4% in Indians and 13.2% by Chinese and other communities (e.g Iban, Kadazan etc). According to the Malaysian National Agro Food Policy (2011-2020), it is projected that the annual fish demand will increase by 13% from 1.74 million tons in year 2013 to 1.93 million tons in year 2020. Therefore, steps are already being taken for a sustainable commercial fisheries production (Yusoff, 2014) but a lot more needs to be done.

The demand of seafood products has increased in recent years giving rise to the flow of more players and introduction of value added fisheries products to attract wider consumer market in the fisheries industry (Surathkal *et al*., 2017). Unfortunately, this has also provided opportunities for unhealthy market practices such as fish species misbranding and false labelling of costly fish species with lower priced species (Rasmussen *et al*., 2009; Miller and Mariani, 2010). This mislabelling of fish has also raised human health concerns. For example, Wong and Hanner

(2008) reported the case of toxic puffer fish being mislabelled as headless monkfish in Chicago, North America. Likewise, Cohen *et al*. (2009) also reported such similar case in fish markets of Thailand. The illegal catching, trade and use of endangered marine fish such as sharks, eels and rays in seafood products show that fisheries conservation and sustainability has been neglected to the point of commercializing overexploited endangered species like eel (Rahman *et al.*, 2015; Asis *et al*., 2016).

In addition to the above, absence of recognisable morphological features in processed fish products or fish fillets make seafood frauds highly prevalent as reported by Wong and Hanner (2008) in North American fish market survey. Cases of substitution of closely related commercially important species have also been reported in the Southeast Asian region attributed to confusion in nomenclature. For instance, some range of species have multiple nomenclatures while on the other hand, a group of species may be labelled with only a single name such as 'tuna' and 'kerisi' (Barbuto *et al*., 2010). The high demand of fish protein diet and exposure of marine fish to a challenging marine environment has increased the importance of advanced fisheries research especially in Malaysian waters, particularly research related to taxonomy, biology, diversity, population structure as well as phylogenetic histories of commercially important marine fish species to ensure continuous supply of fish and fish products. Malaysians are beginning to be conscious about the nutritional value of the food they consumed including fish, and to this end, the correct identification of particular species with its characteristic value is very important.

In like fashion, stock identification of a particular harvested fish group is essential for sustainable fisheries as this attribute is compulsory for evaluating the status of harvested fish. Taylor *et al*. (2011) used the telemetric and otolith count methods and applying the maximum likelihood model for stock identification of tuna populations in Eastern Mediterranean and Western Gulf of Mexico. Multidisciplinary fields of stock identification focus on life history, genealogy, morphology as well as phylogenetics and can be assessed by traditional morphometric methods and also advanced techniques such as geometric morphometrics (GM) (Cardin, 2000) as well as molecular data. Morphological identification technique such as geometric morphometrics is easy, low cost and applicable to clarify not only the differences among species but also differences within population due to various environmental changes (Ibanez *et al*., 2017: Marquez *et al*., 2017).

The family Nemipteridae is a commercially important tropical and subtropical marine food fish group inhabiting the Indo Malayan Archipelago region. The commercial species of this family are locally known by using a single name ‘kerisi’ in Malaysia. It is widely used in the preparation of some popular dishes in Malaysia such as surimi, sushi and fish meatballs due to their good lipid and protein contents besides being used as bait for farmed fish (Galal-Khallaf *et al*., 2016). Commonly known as threadfin breams, they are also commercially important in the Gulf of Suez (FAO, 2011; El-Halfawy and Ramadan, 2014). The Annual Statistics of Malaysia (2014) categorize family Nemipteridae among the highest landing fish and is exported to other countries including Singapore, Japan, China, Italy and Australia.

Many studies have reported on global decrease in population stocks of marine fish (Pontecorvo and Schrank, 2014; Golden *et al*., 2016) including South East Asia (Teh *et al*., 2017). Thus, fish management is crucial to replenish fish resources and this requires the identification of fish populations as a discrete units or fish stocks. The management of fish without stock identification and assessment can lead to severe consequences of depletion in spawning performance, loss of genetic diversity as well as ecological issues (Hilborns and Walter, 2013). Fish in family Nemipteridae are also facing many challenges including environmental pollution, stock depletion and over catching for productions of seafood products (Shyam *et al*., 2017; Zeller *et al*., 2017). Hence, understanding stock identification, stock assessment and genetic structure are needed for management and conservation of this family.

For whole fish specimen, traditional morphometric methods based on meristic counts and conventional morphometric is important and quick for fish identification in the field but the more recent morphometric approach of geometric morphometrics (GM) provides a quantitative approach to differentiate taxa in detail (Bonhomme, 2014). However, difficulties arise when traditional morphometric methods fail to identify samples that are without complete morphological characteristics such as in processed fish and fillet products (Chin *et al*., 2016). To identify these types of fish samples, several protein based analyses methods have been developed specifically for fish species identification (Westermeier, 2016). However, these analytical methods are inefficient in identifying processed fish products because thermolabile fish proteins undergo irreversible denaturation due to heat (Dooley *et al*., 2005). Moreover, it is difficult to differentiate very closely related fish species as they have common protein profiles (Barik *et al*., 2013). The need for a faster approach in taxon delimitation is amplified due to the decreasing number of taxonomists and the limitations of morphological based identification. This is now being partially addressed through advanced molecular identification technique such as DNA barcoding. This approach utilises a standard segment of the mitochondrial DNA as a marker to identify organisms from tissue samples circumventing the need for complete and fresh organisms (Ward *et al*., 2005).

Genes from mitochondrial DNA have the unique ability to track the historical lineages of taxa and their applications have been documented in various research; taxonomy (Henderson *et al*., 2016), cryptic species identification (Azuma *et al*., 2017), population genetics (Lim *et al*., 2016) and stock identification (Shen *et al*., 2016). Cytochrome oxidase c I (COI) and cytochrome oxidase b (Cyt b) sequences from mitochondrial DNA are two of the most frequently used molecular markers now a days (Ward *et al*., 2008; Hubert *et al*., 2012). The COI gene is typically used in DNA barcoding and helps in refining species by identifying unknown specimens using probabilistic algorithms when a set of known species barcode is already established and available in the DNA barcode database called Barcode of Life Database BOLD (Abdo, 2007). Cytochrome oxidase b (Cyt b) has proven to be a robust evolutionary marker among mtDNA protein coding genes, uncovering phylogenies at various taxonomic levels in fishes (Lakra *et al*., 2011). A cytochrome b code is a functionally conserved protein and can be phylogenetically informative in both inter and intra specific studies as it has fast as well as slow evolving segments. However, it is probably most suited for closely related taxa as the nucleotide sequence variation is less saturated by multiple substitutions (Satoh *et al*., 2016).

Equally important, phylogenetic trees are the near exclusive option to relate among species in systematics and taxonomy. Today, the phylogenetic interpretations are the main area of concern for taxonomists and evolutionists in order to search for missing links that can complete the tree of life. Mitochondrial and nuclear genes are molecular markers that are widely applied and successfully proven for phylogenetic inference of vertebrate relationships (Chen *et al*., 2014; Near *et al*., 2014). The choice of gene is very critical for obtaining a successful outcome of the issue being addressed due to the different evolutionary rates of the various genes as i.e.

mitochondrial COI gene is an authentic barcoding gene that is fast evolving and is helpful only for family level phylogenetic comparisons while mitochondrial Cyt b gene is moderately evolving and is efficient for intergeneric and interspecific level of phylogeny comparisons (Hajibabaei *et al*., 2007). Being mitochondrial genes, COI and Cyt b genes have limitations due to maternal inheritance. Therefore, there are risks of incomplete lineages due to missing genetic data from paternal genes. In this regard, the nuclear genes can complement. An example is Recombinant Activating Gene I (RAGI), a slow evolving nuclear gene that can interpret phylogenies efficiently at interspecific level (Chen *et al*., 2014). The utilisation of combined mitochondrial and nuclear markers with different and complementary specific advantages (and weaknesses) will provide a more holistic presentation of the phylogenetic relationships of this all-important fish group.

1.2 Problem Statement

Food safety concerns, environmental (i.e. pollution, seasonal changes) and socio-economic factors (i.e. use of a single name ‘kerisi’ in fish markets, illegal fish trade, over catching practices) are among the few challenges in sustainable utilization of commercial fish of family Nemipteridae to retain their original genetic stocks in biodiversity rich region like Malaysia where fish and rice are ultimate source of food. The stock identification, phylogenetic history and detailed study of species distribution pattern of this family is urgently needed by using recent and authentic approaches for management and conservation of original parental gene stock.

For all the reasons above, the current study is aimed at investigating species diversity including hidden diversity, phylogenetic relationships and morphometric variability on the commercially important family Nemipteridae collected from landing sites of Malaysia and neighbouring oceans. The information obtained from this study will provide the first molecular and morphometric data records for the commercially important family Nemipteridae. This study also gives insights into the stock structure, genetic variations, species complexes and phylogenetic histories of family Nemipteridae and is very valuable to use for long-term management and conservation strategies for the species.

1.3 Objectives

Based on the current issues of Nemipteridae fisheries in Malaysian and surrounding waters the objectives of the project were;

1. To identify commercially important species of family Nemipteridae in Malaysia and its surrounding waters using DNA barcoding of COI gene.
2. To use geometric morphometrics for discrimination among taxa of family Nemipteridae on basis of body shape variations.
3. To determine phylogenetic relationships and distribution of commercially important species of family Nemipteridae by using mitochondrial and nuclear genes.

CHAPTER 2

LITERATURE REVIEW

2.1 South East Asia as a Biodiversity Hotspot

South East Asia is rich in biodiversity, both aquatic and terrestrial. However, it also ranks high in the IUCN red list in terms of endangered species. Geographically located in one of the greatest biodiversity center, Malaysia is the focus of attention for many taxonomists (Marchese, 2015). Surrounded by several seas merging at the junction of the Pacific Ocean and Indian Ocean, the region therefore unsurprisingly supports a plethora of marine organisms including fishes (Bellard *et al*., 2014). Malaysia is divided into two landmasses: Peninsular Malaysia and Borneo (Sabah and Sarawak).

Peninsular Malaysia is an important part of the biogeographical region of Sundaland that was periodically exposed in the past 2-2.5 million years due to lowering of sea level, reaching a low of 30-40 m at some point (Voris, 2000). Sundaland extends over an area of 1,800,000 km^2 including Sumatra, Java and Gulf of Java, Thailand and Gulf of Thailand, Borneo and parts of South China Sea. The Southern and Western boundaries of Sundaland are bordered by the Indian Ocean, the eastern boundary at the Wallace Line, and the northern boundary (although not easy to define is approximately at 9^{o}N). The Wallace line was described in 19th century for the very first time by Alfred Wallace in his book entitled ‘Indo Malayan Archipelago’ (covering an area of 2 million km^2) as an area which divided the warm g pool of Indo Pacific Ocean of the Southeast Asian region from the cooler pool of the Pacific Ocean of the Australian region (Veron *et al*., 2009). Many studies have confirmed that the marine biodiversity present in both regions never crossed their territories except at the transition zone located at the Eastern boundary of Indo Malayan Archipelago in which mixed biodiversity from the Malayan and Australian region can be found (Grudinski *et al*., 2014; Muir *et al*., 2015). The tropical environments of historic Sundaland with high rainfall and the presence of warm pool of Indian Ocean supports the highly renowned 'biodiversity hotspot' status of this region.

Contemporary Peninsular Malaysia is linked to the great landmass of Asia at its northern boundary through Thailand, while Malaysian Borneo is part of the Borneo island, both of which are home to immense coral reef fish diversity (Allen, 2008). Marine coral reef biodiversity tails off further to the West because of increase in turbidity and unavailability of suitable sea shore for reef fringe attachment (McClanahan, 2000). Two major hypotheses have been postulated about patterns of species richness in this important marine zone. Firstly, the 'Centre of Origin' hypothesis. This refers to the sympatric speciation concept about the divergence of species from Indo Malayan Archipelago. In summary, this concept postulates that the origin of speciation occurs at the centre of the Sundaland and then disperses to the periphery. Thus, the expectation of this concept is for the presence of shallow genetic structure at the midpoint ranges of species due to dispersal towards more peripheral regions of the Indo Pacific Ocean as compared with the central region. Secondly, the 'Centre of Overlap' hypothesis, which postulates that vast amount of speciation, occurred during the Pleistocene. At that time, the Indian and Pacific Oceans were intermittently connected and disconnected. During the period of sea level rise, the two areas overlapped resulting in species overlap in the Indo-Malayan region forming an enormous pool of diversity (Gaither and Rocha, 2013) including fish species. Thus, understanding present-day fish distribution is an important indicator to assess the connections between landmasses that were historically connected and in this regard research on the phylogenetics and phylogeography on many species of fish are important tools to achieve this..

The genetic variability within a species plays a key role in a survival and sustainability of trophic levels in marine habitat (Hiddink *et al*., 2007), the higher the diversity, the higher the resilience. This genetic makeup provides the evolutionary potential in a group (Schindler *et al*., 2010). Genetic diversity is mainly influenced by environmental factors, genetic drift and selection pressures (Kovach *et al*., 2013). There have been many reports of the influence of overfishing practices that have resulted in decline of fish populations and hence genetic decline. For example, the Atlantic cod was a commercially important marine fish that was highly caught during the twentieth century that resulted in a decrease of its sexual maturity from >5 years to <3 years and finally collapsed the fish stock (Sterner, 2007). McCain (2015) reported that the decline in Atlantic cod (*Gadus morhua*) was also affected the populations of other fish due to the unavailability of prey. Following that, Mirimin *et al*. (2016) reported the presence of low genetic diversity and low trends of population differentiations in *Argyrosomous japonicus* due to overfishing practices in its distributional range of 2000 km along the coastline of South Africa.

Variations in marine environment can severely lead to decline in coral reef fish especially those distributed in the hotspot regions (Descombes *et al*., 2015). Arai (2015) stated that the environmental and man-made pressures on marine habitat has resulted in 35 marine fish families and 86 marine fish species being categorized as threatened or extinct and further reported that methods of fish practice (blast fishing, trawl fishing, poison fishing) are constant threats along with other anthropogenic and natural pressures for the South China Sea fisheries within the Malaysian water. Thus, it is important to conserve the genetic diversity along with conservation and management of dwindling coral reef diversity in biological hotspot regions.

2.2 Taxonomy and Phylogenetic Considerations in Fish

According to FAO report (2015), more than 50,000 fish species are globally distributed, of which 33,100 fish species have been documented as valid fish species untill April, 2015. The marine fish documentation rate is only 100-150 species per year (Eschmeyer *et al.*, 2010) due to limited expertise in this field. The environmental (i.e. seasonal changes, greenhouse effect, acid rain) and man-made pressures (i.e. overfishing, pollution) are major threats for marine fisheries. The discovery, identification and documentation of marine fish resources are of utmost urgency before the extinction of important yet undiscovered and undocumented fish species. The two reliable web resources, FishBase and Catalog of Fishes provide a growing documentation of global fish and fisheries (Eschmeyer and Fong, 2016; Froese and Pauly, 2016). The documentation, identification and exploration of new fish species are important fields in taxonomy that need rapid research but along with these perspectives, understanding the evolutionary relationships among marine fish taxa is also equally important for insights into the historical dispersal of marine taxa during changes in ocean landscape during geological events and climatic oscillations.

The demand of fish products has increased in recent years giving rise to the flow of more players into this business. This has encouraged economic deception that involves fish species misbranding and false labelling of costly fish species with lower priced species (Rasmussen *et al.*, 2009; Miller and Mariani, 2010). This mislabelling of fish species has also raised food safety concern such as cases of toxic puffer fish samples that mislabelled as 'headless monkfish' or other harmless products (Cohen *et al.*, 2009). Fisheries conservation and sustainability has also been neglected to the point of commercializing overexploited species (Jacquet and Pauly, 2008). There are many cases of substitution of closely related species from other countries or continent among commercialized fish due to an ambiguous nomenclature. Some species are grouped into a single name such as tuna (Lowenstein *et al.*, 2009) while there are also singular species with multiple nomenclatures (Barbuto *et al.*, 2010).

Difficulties can arise with sole dependence on morphological diagnosis for sample identification in the absence of diagnostic characteristics such as in processed fish and fillet products (Sotelo *et al.*, 1992; Unlusayin *et al.*, 2001; Smith *et al.*, 2008). To identify these types of fish samples, several protein based analytical methods have been developed specifically for fish identification (Tepedino *et al.*, 2001: Ochiai *et al.*, 2003). However, these analytical methods can only be applied for identification of raw fish sample because thermo labile fish proteins undergo irreversible denaturation due to heat in a processed fish products (Dooley *et al.*, 2005). Furthermore, it is difficult to differentiate closely related fish species as they have common protein profiles (Bartlett and Davidson, 1991; Smith and Rayment, 1996). Determination of fish population structure is even more challenging because the differences among populations of the same species are substantially minor or even negligible. To address these limitations and to enable precise fish species identification, phylogenetic assessment and population genetic studies, new approaches in morphometrics and DNA based techniques with high thermal stability is successfully used these days (Hajibabaei *et al*., 2007; Dor *et al*., 2014).

2.3 Morphometric approaches in fish identification

In the marine environment, the term ‘stock study’ is used for two types of groups; i) randomly mating intraspecific group that is living in the same habitat constitutes stock identification study, ii) interspecific group of morphologically similar and with non-significant differences living in the same habitat constitutes stock differentiation study (Ihssen *et al.,* 1981; Tzeng, 2004). The identification methods are usually based on differences in phenotypic characters. In a biological context, morphometrics and meristic is the analytical technique to address variations and co-variations in biological forms. Meristic and morphometrics are traditional approaches to identify taxa at species level but are still widely used today due to their ease of application, absolutely no cost once the equipment is available (callipers or the more advanced equipment such as image analyser) and generate very useful data. Meristic approach for species identification uses countable variable of external features (such as fins counts, caudal fin shape, body colour, body shape) whereas morphometrics uses absolute measures, ratios and proportions such as total length, standard length, dry and wet body mass (Russell, 1990; Kumar *et al*., 2015). Russell (1990) used meristic and a few morphometric characters (body length, eye diameter, body depth etc.) to identify marine fishes and his morphological identification key are still referred for identification and discovery of a new species. It is also a fact that the identification of fish species through morphological keys is a time consuming process and also the phenotypic similarities in fish species can lead towards misidentification.

The late 20th century was a momentous period in the history of morphometrics. A more advanced morphometric method known as geometric analysis was introduced or revitalised in the late 1990's (Bookstein, 1985, 1991; Rohlf, 1993; Rohlf and Marcus, 1993; Adams *et al.,* 2004) and combined with powerful statistical analyses and comprehensive approaches of data collection (Sidlauskas *et al.,* 2011) facilitated improvement in understanding of morphological evolution. This was developed through the advancements in statistical analysis in combination with images and geometric methods that analyze variation in coordinate systems in 1980s (Bookstein, 1991; Adams *et al.,* 2004). The use of digitized image (Cadrin and Friedland, 1999), in place of actual specimens was another innovation in morphometrics and combined with traditional multivariate morphometric analysis further revolutionised the field of morphometrics (Collar *et al.,* 2013). Cardin and Friedland (1999) reported that although ontogenetic assessments of taxa are difficult to interpret but use of geometric morphometrics could enhance the traditional species identification techniques.

Geometric morphometrics is based on landmarks geometry and is aimed to understand among species, the ecological and evolutionary changes among species that are the precursors of shape variability in biological forms (Adams and Otárola-Castillo, 2013). Landmarks are homologous points associated with the geometry of the organismal body of closely related taxa (Bookstein, 1991; Gunz *et al.,* 2005). For each specimen, landmarks are chosen on its digital image at specific homologous positions and are expressed on two dimensional or three-dimensional coordinates.

All the specimens should have the same landmarks and incomplete specimens or missing landmarks will be eliminated before further analysis to overcome the misinterpretation of results. The GM analysis allows size and shape to be independently evaluated in the investigation of morphological variation and co-variation.

Geometric morphometrics can efficiently address shape variations because the landmark coordinates are similar in a single (or related) type of taxa and any significant variation in specific range of landmarks help to address the changes in structure (Mitteroecker and Gunz, 2009). Through effective statistical analyses and values generated such as means, principal components, covariate analysis and regression, it is proposed that GM could be useful in evolutionary investigations of ontogeny (Zelditch *et al.*, 2012; Polly *et al.*, 2016). The use of GM methods in morphological and functional studies became increasingly popular at the turn of the century (Adams and Rohlf, 2000; Gunz *et al.*, 2005; Adams *et al.*, 2011).

Costa and Cataudella (2007) explored relationships in body shape and trophic ecology in juveniles of nine species of family Sparidae by GM study on body shape and deduced that mean body shape is different in terms of mouth gap. In carnivore the mouth gap is large and also long body with narrow caudal peduncle. Herbivore is associated with small mouth gap while omnivorous species has a deep body with small mouth gap..

Ibanez *et al*. (2012) differentiated species, genera and populations of family Mugilidae from different locations on the basis of differences in ctenoid scales by use of geometric morphometrics. They used discrimination function analysis and PCA on landmarks data but found GM analysis of form variations of ctenoid scales less effective in discrimination of taxa but most effective to differentiate geographically dispersed populations (collected from Gulf of Mexico and Aegean Sea). However, although, morphology will remain as the foundation of taxonomy, four major limitations have been highlighted by Hebert *et al*., 2003); 1) Incorrect identifications resulting from both phenotypic plasticity and genetic variability in the characters employed for species recognition, 2) cryptic taxa that are common in many groups could be overlooked, 3) often only for a particular life stage or gender those morphological keys are effective and 4) misidentifications are common, as the use of keys often requires a very high level of expertise. In complement with morphological data, these problems have been efficiently addressed by the discovery of molecular methods in species identification.

2.3.1 Statistical methods in morphometric analysis

Several statistical methods are widely employed to describe differences in size and body shape. These include univariate analysis such as ANOVA and multivariate statistical approaches such as Principal Component Analysis (PCA) and Discriminant Function Analysis (DFA).

PCA is a quadrate analysis in biostatistics that is used to convert or reduce large sets of variables in the form of principal components that contain all information about the variables and thus determine the maximum amount of variation. The PCA uses total covariance matrix to transform into 2D or 3D visualization in such a way that the first component usually represents the maximum variation that subsequently reduces in succeeding components (Klingenberg *et al*., 2003). In morphometrics, whether traditional or geometric morphometrics, the total variation along all axes of principal components is assessed by eigenvalues. Eigenvalue > 0.30 shows significance, eigenvalue > 0.40 shows higher significance and eigenvalue > 0.50 shows that a highly significant correlation is present (Nimalathasan, 2009; Lombarte *et al*., 2012). Moran *et al*. (2017) stated that PCA is a comparatively less effective method as compared to DFA to address the phenotypic differences between species because principal components can measure differences at only highly diverged axis. Furthermore, PCA analysis does not require any pre-grouping unlike DFA analysis, the latter facilitating magnification of differences among groups (species). The values from PCA can be used to determine the shape models and to construct a tree or dendrogram to visualize relationships. Many studies have reported the success of PCA to address population structure in marine fish. Lim *et al*. (2016) used PCA analysis to observe population variations in *N. japonicus* at various locations in Malaysia.Similarly, Duong *et al.* (2017) used PCA analysis to differentiate three wild and hybrid species of bighead carps and also reported that the first two PCs accounted for 55% differences that explained the anterior parts of the body (head size, fin lengths and distance between dorsal and caudal fin). Claverie and Wainwright (2014) reported that the body elongation is the main factor of shape variations in the evolution of reef fish that diversified after undergoing through the process of diverse changes in body orientations. This was explained by the first two principal component axis that accounted for 58.3% of total body variations of the reef fishes.

PCA is traditional method of multivariate analysis in statistics that reduces dimensions of dataset and transforms data into new coordinates that are linear combination of old datasets. McGarigal *et al.* (2000) reported that the principal components in PCA are uncorrelated to each other and have been aligned according to order so that the first few principal components in PCA represents maximum variations present in the original dataset. The principal components are based on empirical observations so that the variations in body size are mostly inferable from variations in samples and is the main reason for the high value of the first principal component which shows maximum amount of allometric relation between body shape and body size (Slice and Stitzel, 2004).

The discriminant function analysis (DFA) is also a multivariate analysis. It was introduced by Fisher (1936) for the very first time on one morphotype of iris plant and used four measurements to create discriminant functions that could differentiate three species. The DFA is an authentic analysis in geometric morphometrics that can successfully record variations among two or more different species, geographically isolated populations and also two groups of a species in a clade (Zelditch *et al*., 2004; Klingenberg and Montriro, 2005; Mitteroecker and Bookstein, 2011).

In the past, DFA was confused in terminology with another multivariate analysis, which is Canonical Variate Analysis (CVA) but Klingenberg and Monteiro (2005) referred the use of DFA to discriminate between two groups (two species) and CVA to discriminate between more than two groups in the same analysis. DFA is also based on eigenvalue and covariance matrix like PCA. However, in contrast to PCA, DFA used the covariance matrix among groups with a priori defined groups to discriminate within and among groups by use of Procrustes distances and Mahalanobis distance. Strauss (2010) stated that the eigenvalues in DFA address variance among groups in components of Discriminant Functions and Canonical Variates.

Many studies have addressed the efficiency of DFA to discriminate among and between species from the same or different geographical locations. Pérez Quiñónez *et al*. (2017) discriminated three species of *Opisthonema* in a combined molecular and geometric morphometrics study and reported that the first component (with eigenvalue 85.5%, Wilks Lambda 0.26, P < 0.001) in CVA consisted of three times more discriminating power than the second component (eigenvalue 14.5%, Wilks Lambda 0.75, P < 0.001). Klingenberg *et al*. (2003) documented in a study of *Amphilophus* species complex (*A. citrinellus*, *A. zaliosus*, *A. labiatus*) that three coloured morphs showed significant morphological variations due to body shape (Wiliks Lambda = 0.35, p = 0.046).

2.4 Molecular approach in species discovery and documentation

The challenges of using traditional morphometric methods to identify fish species in consideration of factors such as phenotypic plasticity, developmental stages (egg and larvae), sexually dimorphic taxa, cryptic species as well as sibling species have motivated taxonomists to use molecular approaches for species identification. The effectiveness of various DNA based, immunological based and protein separation as well as characterization based techniques has been widely reviewed in the past.

Allozymes was extensively used during the early 1990's and play a major role in the description of species diversity in fishes (Ward and Grewe, 1994; Utter, 1998). RNA and DNA blotting techniques were also at some point popular to blot specific DNA/RNA sequence in DNA/RNA sample and are still in use but are restricted to genetic engineering (Brown, 2001; Josefsen and Nielsen, 2011).

In quantitative method of immunological assays, the binding of antibodies with specific antigens from different species was used to calculate the immunological distance where the binding capacity was inversely proportional to the genetic distance between taxa (Asensio *et al.,* 2008). However, all these methods do not work well in samples acquired from dead organisms because the protein starts to denature soon after death (Telechea, 2009). In this regard, the use of DNA markers is an advanced molecular approach that works equally well on dead and live body cells. Examples of DNA based molecular markers and approaches include randomly amplified polymorphic DNA (RAPD) (Kazan, 1993), amplified fragment length polymorphism (AFLP) (Maldini *et al*., 2006), single nucleotide polymorphism (SNP) (Jones *et al.,* 2013), mitochondrial DNA and nuclear markers (Hebert *et al.,* 2003; Zhang and Hanner, 2012).

2.4.1 DNA barcoding

One of the increasingly important taxonomic tools for species detection is DNA barcoding. Hebert *et al.* (2003) initiated this method that enables almost any form of organisms worldwide to be identified by using a system that exploits a mitochondrial DNA sequence as a taxon 'barcode'. Mitochondrial genes in fish species are promising markers for their identification (Teletchea, 2009) as compared to the nuclear genes. This is because of several special features inherent in the mitochondrial DNA (mtDNA). This small closed circular DNA with the size range of 15-20 kb that occurs in high copy number in each cell is easily recovered through various extraction methods (Hubert *et al*., 2008). Its maternal inheritance pattern, lack of recombination, rapid mutation rate and small effective population makes it an effective tool for studying phylogeny and genealogy of taxa through multi lineage (Moore and Dowhan, 1995; Sangthong and Jondeung, 2003). Moreover, the evolution rate of mitochondrial genes is fast and hence it exhibits great potential as a barcoding gene to delineate species. Mitochondrial genes have been widely utilized in various applications such as in identification, seafood control (Yancy *et al*., 2008; Miller and Mariani, 2010), fisheries management (Holmes *et al*., 2009; Pinhal *et al.*, 2012), and species delineation (Ward *et al.*, 2009; Hubert *et al.*, 2015). Due to these positive features, mitochondrial genes were used extensively for DNA barcoding, also known as molecular tagging. However, there are several limitations of mitochondrial markers; 1) The mitochondrial genome is small (15-20 kb) in fish compared with nuclear genomes, and therefore only represent a small proportion of the genetic material, 2) While the maternally inherited allows tracing of maternal contributions, the non-Mendelian inheritance makes it unsuitable for many genetic studies, 3) Due to the high mutation rate and small size of mitochondrial genomes, back mutation could readily happen that does not reflect the phylogenetic relationship or evolutionary history.

Cytochrome Oxidase C Subunit I Gene (COI)

According to Kress and Erickson (2008), in order to be qualified as a DNA barcode the selected gene region must first meet three requirements; possess considerable species level genetic variation and divergence, presence of conserved flanking sites and features only short sequence length. Generally, mtDNA cytochrome oxidase subunit I (COI) has been accepted as the barcoding gene for animals while (Hebert *et al.,* 2003) a two-locus barcode (*rbcl* and *matK*) is the accepted DNA barcode (Hollingsworth *et al.,* 2011; Cowan and Fay, 2012) for plants. Researchers have resolved many taxonomic ambiguities in identification and classification of various vertebrate groups using the COI gene (Ward *et al.,* 2005; Rach *et al.,* 2008; Rock *et al.,* 2008).

DNA barcoding is not only useful in species identification but could also be applied in assessment of population structure for example for studies in Snakehead fishes (Jamaluddin *et al.,* 2011), for studies in threadfin breams (Lim *et al.,* 2016), for conservation and management of amazonian commercial fish (Ardura *et al.,* 2010) and coral reef fish species (Hubert *et al.,* 2012). It can also aid in identifying invasive species, for example for studies in Jpanese bluegill sunfish (Takahara *et al.,* 2013) and also to unveil new taxa or cryptic species (Bucklin *et al.,* 2011).

Mitochondrial COI gene was used to study 207 species of marine fishes in Australia (Ward *et al.,* 2005). The vast majority of species were unequivocally identified with clear phylogenetic signals in their data. Further, they categorized the sequence samples into four major taxonomic (higher order) groups comprising of chimaerids, rays, sharks and teleost. Intra-generic species clustered together as did intra-familial genera. The study contributed in the precise identification of Australian fish species, a prerequisite for fish biodiversity conservation and fisheries management. Lakra *et al*. (2011) used DNA barcoding technique to identify 115 marine fish species and used NJ method to infer their phylogenetic relationships that clustered 115 fish species into 79 genera.

The DNA barcoding method has been further refined by development of Next Generation Sequencing (NGS) platforms. For instance, mini-barcode fragments that are shorter fragments (between ~ 100 bp) compared to 650 bp in standard barcodes of the COI gene are easily available (Hajibabaei *et al.*, 2011). Its resolution efficiency (at 90% species resolution) (Muirhead *et al.*, 2008) is comparable to the 97% species resolution using full-length DNA barcode sequences (~650 bp) (Hajibabaei *et al.*, 2005; Hajibabaei *et al.*, 2007; Muirhead *et al.*, 2008). Furthermore, archival specimens and processed biological materials such as canned food, which generally do not have the full-sequence of DNA barcode intact can benefit from the mini-barcode development (Hajibabaei *et al.*, 2006; Muirhead *et al.*, 2008).

DNA Barcode Libraries

DNA barcoding is a golden bullet in delimiting species boundaries and is part of global species documentation project 'the Barcode of Life' (BOL) launched to record global biodiversity in a dedicated platform. The Consortium for the Barcode of Life (CBOL) was inaugurated in 2004 to develop standard protocols of DNA barcoding techniques, statistical analyses and documentation that could be applied worldwide in molecular taxonomic laboratories to form a global DNA library. CBOL was later extended with the launching of International Barcode of Life, (iBOL) (http://www.ibol.org) where an initial of 26 countries collaborated to document eukaryotic biodiversity with a preset target of cataloguing 5 million species barcodes. Various DNA extraction protocols, bioinformatics software for sequence analysis and a global data library have been developed in iBOL.

Specific thematic barcode projects were also launched such as FishBOL (Fish Barcode of Life) to barcode all fish life forms, MarBOL (Marine Barcode of life) to barcode marine biodiversity, BeeBOL (Bee Barcode of Life) to barcode all bee species, MBI (Mosquito Barcode of life) to barcode all species of mosquitoes, Lepidoptera BOL to barcode all species in order Lepidoptera along with some regional projects such as Mexico (MexBOL) to barcode whole biodiversity distributed in Mexico, Norway (NorBOL) to barcode biodiversity in Norway, Europe (ECBOL) to barcode biodiversity of Europe etc.

The world heritage site in Costa Rica (The Área de Conservación Guanacaste) is also part of DNA barcode project that is working on insect barcodes in association with iBOL project (Janzen *et al.*, 2005; Hajibabaei *et al.*, 2006; Janzen *et al.*, 2009). All DNA barcode data are submitted to the Canadian Centre for DNA Barcoding (CCDB) for uploading in the database called BOLD (Barcode of life database System) (Ratnasingham *et al*., 2007). This is a public database consisting of barcodes of global biodiversity and is freely available on line via webpage (http://www.boldsystems.og/index.php/Login/page?destination=MAS_Management_User Console). The user friendly interface of BOLD gives information about location of samples, pictures, barcode sequence and sequence resemblance where researchers can also download barcode sequences for statistical analysis and phylogenetic tree construction (Ratnasingham and Hebert, 2007). BOLD database also provides online statistical barcode analysis and inter and intraspecific genetic differences graphical and tabulated representations (Ratnasingham and Hebert, 2007). Malaysian laboratories are actively participating in this global DNA Barcode initiative building database for the biodiversity of various taxonomic groups and locations in this hotspot region that are available online on BOLD website (http://www.boldsystems.og/index.php/Login/page?destination=MAS_Management_User Console).