CHAPTER I

Introduction

With annual production of around 70,064 tones metric from freshwater aquaculture systems, worth almost RM 345 million in 2007 alone (Dept. of Fisheries, Malaysia), aquaculture has been recognised as one of the major sectors highlighted by the Malaysian Government. Aquaculture industries encompasses diverse systems of farming aquatic organisms in confined areas of fresh or marine water, and are currently used to farm a wide range of taxa such as fish, crustaceans, mollusks, and aquatic plants . Apart from fishes like carp and tilapia, crustaceans such as shrimps and prawns are also increasingly intensively being cultured and harvested to meet the world's demand for seafood products.

In Asia, marine penaeid prawns remain the major crustacean group that is currently being cultured at a commercial level. Nevertheless, production of freshwater prawns of the genus *Macrobrachium* has also seen a dramatic increase (Mather & de Bryun, 2003). As the largest species of the genus, the giant freshwater prawn *Macrobrachium rosenbergii* is by far the most popular and important *Macrobrachium* species for commercial culture, with annual global production reaching over 200,000 tonnes in 2002 (New, 2002). Although this species is indigenous to South and Southeast Asia, parts of Oceania and some Pacific islands, the culture of giant freshwater prawn has gained worldwide popularity (New, 2002).

The development of modern farming techniques for *M. rosenbergii* began in 1960 with the early work by Dr. Shao-Wen Ling, an FAO expert working in Malaysia. This work revealed

that, although primarily a freshwater organism, *M. rosenbergii* larvae require brackish water for survival and their early development (New & Singholka, 1985 as cited in Mather & de Bruyn, 2003). By 1972, successful domestication of the species was accomplished through the concerted efforts of a team led by Takuji Fujimura in Hawaii. The first commercial culture of this prawn was conducted in the Anuenue Fisheries Research Center, Hawaii, using introduced brood stock consisting of only 12 individuals originally brought from Malaysia (Hedgehock et al. 1979 as cited in Mather & de Bryun, 2003). Later on, progeny of these initial brood stock from Hawaii and additional individuals from Southeast Asia were introduced into many regions where *M. rosenbergii* was not indigenous, as a means to establish culture industries in these countries (New, 2010). As a result of the early work on the life cycle in captivity and domestication techniques, commercial *M. rosenbergii* farming is now well established in Hawaii and elsewhere (New, 2002).

Giant freshwater prawn is of high value among cultured species due to its dainty taste and high protein content, both attributes that are favorable to the consumer. On top of that, other factors such as ease of culture and global export potential also contribute to the expanding popularity of mass-rearing of *M. rosenbergii* at the commercial level all around the world (Whangchai et al., 2007). According to FAO (2003) statistics, Malaysia was listed among the top 15 producers of *M. rosenbergii*, with production over 700 million tonnes in 2001. Production in Malaysia has shown rapid increase in recent years, and is predicted to continue to expand as more farming efforts are initiated.

However, after several production cycles in culture, there is a tendency for the performance of farmed *M. rosenbergii* to decline during grow-out. The affected performance can

be seen in terms of traits related to productivity and profitability, such as population growth rate, body weight-at harvest, survival rate, disease resistance, and feed conversion ratio (New, 2002). Corresponding declines in production among commercial stocks have been noticed in a number of countries including Taiwan and Thailand. The significant drop from 16,000 t to just 7,665t in Taiwan's cultured prawn industries during early 1990's was attributed to inbreeding depression (sometimes known as genetic degradation) (Mather & de Bryun, 2003). Inbreeding depression can be defined as the reduced fitness of a population caused by inbreeding that result in decreasing genetic variability from one generation to another. Inbreeding most likely occurs in hatcheries due to the practice of 'recycling of animals' where the brood stocks for subsequent breeding cycles are sourced from grow-out ponds rather than from wild or more genetically diverse populations. When this process is repeated for many generations, the level of genetic diversity in the cultured stock can be dramatically reduced (New, 2002).

In countries where *M. rosenbergii* is not indigenous, the problem of inbreeding may also have been exacerbated when initial introductions to the country only included a very small number of broodstock, (New, 2002). The resulting effects of inbreeding in *M. rosenbergii* culture populations' has caused negative impacts on aquaculture industries as farmers suffered great losses due to declining yields and low productivity of farmed prawn stocks. In addition to inbreeding depression reducing yield, other related problems like viral infection and disease outbreaks in hatcheries have also contributed to high mortality rates, thus jeopardizing the sustainability of *Macrobrachium* culture industries (Nguyen, 2009; Wang et al., 2008).

An additional threat to the culture industry of giant freshwater prawn is the depletion of the wild brood stock (Browdy, 1998 as cited in Lehnert et al., 1999), which may mean in future

that genetically diverse individuals can no longer be sourced from the wild to enhance broodstocks and avoid inbreeding. Many natural populations of *M. rosenbergii* have seen rapid declines due to various human actions such as overexploitation, and environmental problem such as habitat loss, thus reducing the genetic diversity of wild stocks (Mather & de Bruyn, 2003; Bhassu et al., 2009; Bhat et al., 2009). Whilst this species is now believed to be locally extinct as a result of pollution and loss of its natural habitats in Singapore (Ng, 1997 as cited in Mather & de Bruyn, 2003), other countries including Indonesia, Thailand, India, and Malaysia have demonstrated wild stock declines in recent years (New et al. 2000 as cited in Mather & de Bruyn, 2003).

To ensure viability and sustainability of freshwater prawn farming, it is vital to monitor genetic diversity levels in cultured stocks through proper assessments of appropriate genetic parameters. On top of that, conservation of wild brood stocks is also required as they play crucial roles in providing valuable resource of genetic variation that can be exploited to address inbreeding problem in cultivated *M. rosenbergii.* This necessitates appropriate documentation of the genetic diversity characteristics that are present in wild stocks to identify populations that may carry unique genetic attributes, thus enabling conservation efforts to be prioritized. Performance of each wild population can also be evaluated for the purpose of establishing base populations for brood stock development programs.

Genetic markers such as microsatellites can be used as a tool to assess genetic variability present in wild populations, and are a good choice of DNA marker as they can be used to address a range of questions at the intra-specific level, including characterization of diversity, determination of population differentiation, parentage analysis, and even quantitative trait

analysis. Characterizing diversity in wild populations of *M. rosenbergii* using microsatellite markers may provide useful information relevant for formulating effective management strategies and to help devise comprehensive conservation policies for this species.

Previous studies have shown that screening genetic diversity across prawn populations helps in identifying strains with superior characteristics for latter application in possible breeding and selection programs. As reviewed in New, (2005); work done by Buranakanonda in 2002 reported that a few specific strains of *M. rosenbergii* in Thailand, and Myanmar prawns from the Yapil River were found to be the meatiest, thus, representing targets for breeding programmes that may be able to maximise fecundity and survival whilst maintaining desirable meat characteristics. In another case, morphological studies performed by Mather and colleagues in 2002 on *M. australiense* demonstrated that trait variations exists within geographical regions as well as between different regions, and that this variation seems to be related to environmental factors (as cited in New, 2005). Further work on *M. rosenbergii* populations will provide information on genetic patterns of diversity in wild stocks, and perhaps also identify performance differences among populations from different geographical areas. This may potentially enable subsequent improvement of particular strains through selection and cross-breeding programs in culture

This study was carried out to relate and examine the extent of geographical population differentiation among wild populations of *M. rosenbergii* using microsatellite DNA markers.

Therefore, the specific objectives of this study were:

a) To characterize novel polymorphic EST-SSR markers generated from RNA of *M. rosenbergii*
b) To screen a set of EST-SSR loci in *M. rosenbergii* individuals from four wild populations;
c) To characterize the distribution of genetic diversity and level of genetic relatedness within and among the four wild populations using the EST-SSR data set.

CHAPTER II

Literature review

2.1 Population genetics

In its broad sense, population genetics can be defined as the study of naturally occurring genetic differences among organisms (Hartl, 2000). Genetic differences can be polymorphism that occurs between organisms within same species, whereas differences that accumulates across species is known as genetic divergence. A number of factors play essential roles which affect the amount and the types of genetic variation in populations; such as selection, inbreeding, genetic drift, gene flow and mutations. With rapid expansion in today's technology, more breakthroughs have been accomplished, and abundant data obtained from molecular-based experiments are made available to assist in transformation of this field. Thus, while population genetics seeks to understand what causes genetic differences; along with their extent and pattern within and among species, molecular biology can provide a rich repertoire of techniques for identifying these differences (Hartl, 2000).

2.1.1 Molecular markers

The field of population genetics in recent days has changed dramatically from those early studies of genetic variations. In addition to analyzing the easily detected and/or quantifiable variations such as color and morphological variants, the trends have shifted towards efforts to investigate genetic disparity at the molecular level. Many molecular markers have been

employed to this area to estimate the extent and pattern of genetic characteristics of studied populations. Analysis using molecular markers is one of good methods for identifying genetic differentiation among populations and populations structuring (King et al., 2001 as reviewed by Hooshmand, 2008). Molecular markers usually exhibit polymorphic pattern that can be observed in the entire genome of particular organisms, and such variation is the quintessence for population genetic studies (Christiansen, 2008). The level of polymorphisms of selected molecular markers determines their resolving power for further assessment of genetic variations.

2.1.2 Microsatellites (Simple sequence repeats, SSR)

Among the vast range of molecular markers available to carry out population genetic studies, microsatellites have shown to be a good genetic tool of choice to characterize populations due to their high degree of polymorphism (Ellegren, 2004; Liu & Cordes, 2004; Li et al., 2002). Furthermore, microsatellites have found such widespread use in population genetics field because they show extensive genome coverage and display co-dominant Mendelian fashion of inheritance (Ellegren, 2004; Ellis & Burke, 2007; Liu & Cordes, 2004). The ubiquitous microsatellites were estimated to occur as frequently as once in every 10kb in the fishes genome as shown by Wright, 1991 (Liu & Cordes, 2004); and it can be found inside gene coding regions, introns, and in the non-coding regions (Li et al., 2002, Yu & Li, 2008).

Besides all these unique properties, microsatellites has also become marker of interest as it can be easily detected by polymerase chain reaction using primers that can hybridize to unique flanking sequence of the repeated core unit of microsatellites regions (Powell et al., 1996 as reviewed in Yu & Li, 2008). Also known as simple sequence repeats (SSRs) or short

tandem repeats (STR), microsatellites are DNA sequences consisting of tandemly arranged short repeating units of nucleotides. The core units range in size from 1-6 bp; one (mono-), two (di-), three (tri-), four and five (tetra- and penta- respectively). For instance, ACA_4, $GATA_5$, $CATG_8$. Microsatellites can present in multiple copies at many different locations in the genome (Ellegren, 2004; Liu & Cordes, 2004; Li et al., 2002; Hartl, 2000).

Microsatellite markers or SSRs are generally classified as Type II markers, since it is associated with genomic regions that have not been annotated to known genes; unless they are linked to genes of known function (O'Brien, 1991 as cited in Liu & Cordes, 2004). Usually, SSRs Type II markers must be developed *de novo* for each species through genomic library enrichment procedures since it is species-specific markers (Coulibaly et al., 2005). This is because, they occur in non-coding regions of the genome which are not highly conserved (Zane et al., 2002 as cited in Kim et al., 2008). Unfortunately, the *de novo* development of SSRs from genomic DNA is costly and a time-consuming endeavor (Ellis & Burke, 2007). Furthermore, this approach is also hampered by the paucity of resources for taxa that lack economic importance (Ellis & Burke, 2007). Therefore, active research has been focused to find alternative to this conventional strategy.

2.1.3 Expressed Sequence Tags-derived Microsatellites (EST-SSR)

As previously mentioned, interesting features possess by microsatellites have enabled them to be extensively used over the last decade in various applications such as population genetics, for testing ecological and evolutionary hypotheses in natural populations, parentage analysis, and genetic mapping (Zhang & Hewitt, 2003, Selkoe & Toonen, 2006 as cited in Kim et al., 2008; Ellis & Burke, 2007; Li et al., 2002; Yu & Li, 2008). Nevertheless, despite

all the advantages, tedious and lengthy processes involved in its *de novo* developments have become major constraint of utilizing the anonymous SSR markers in upcoming research studies.

With the expansion of genomic technology over the past decade, establishment of publicly available genomic database have been made possible. This development has enabled scientists to exploit vast amount of sequence information such as cDNAs and expressed sequence tags (ESTs) from databases, thus offering an *in silico* approach to develop gene-based SSR markers at virtually no cost and minor efforts (Kim et al., 2008; Ellis & Burke, 2007; Coulibaly et al., 2005). To date, ESTs have been generated for a wide variety of organisms and these ESTs provide a window into genome where microsatellites may be found.

Briefly, ESTs are short DNA sequences (usually 200 to 500 bases long) corresponding to a fragment of randomly picked complimentary DNA (cDNA) clones that are generated by sequencing either one or both ends of the expressed genes. These sequences represent transcribed sequences of the genome or genes, and currently being used as a fast and efficient method of profiling genes expressed in particular cell types, various tissues, or organs from different organisms, under specific physiological conditions, or during specific developmental stages (Liu & Cordes, 2004).

Over the past decade, the wealth of these transcribed DNA sequences for various organisms in ESTs database (for example dbEST: www.ncbi.nlm.nih.gov/dbEST/) has increased tremendously, and they have developed into rich resources for the identification of gene-tagged microsatellites. Using certain computational tools; *MISA, Repeat Finder, SSRIT,*

Sputnik just to name few; the sequence data for ESTs can be downloaded from GenBank (http://www.ncbi.nlm.nih.gov/genbank/) and scanned for the identification of microsatellites, which are typically referred to as EST-SSRs or genic microsatellites (Duran et al., 2009; Li et al., 2010). The detection of SSRs within EST sequences connects the function of a transcript with the presence of a microsatellite, creating a Type I marker which is easy and inexpensive to produce, and can frequently be associated with annotated genes. While ESTs provide means for the identification of genes, microsatellites provide high levels of polymorphism (Serapion et al., 2004). The strategy of developing microsatellites Type I markers (from known genes and ESTs) has been used for a variety of organisms ranging from animals such as catfish *Ictalurus punctatus* (Serapion et al., 2004), salmon *Salmo salar* (Siemon et al., 2005); plants like cotton *Gossypium* spp. (Qureshi et al., 2004), shrub *Jatropha curcas L* (Wen et al., 2010); and even pathogen like fungal wheat *Phaeosphaeria nodorum* (Stukenbrock et al., 2005).

As ESTs represent a copy and informative source of genes that are being expressed, they serve as a powerful tool for gene hunting; by providing a sequence resource that can be exploited for large-scale gene discovery by using comparative genomic approaches alongside model organisms to discover putative functions of cDNA clones (Ayeh, 2008). This information will further assist in better understanding of organism genome structure, gene expression, and its function as well as allowing the study of genome evolution (Serapion et al., 2004; Ayeh, 2008).

Besides that, since ESTs are generated from transcribed regions, polymorphic EST-derived microsatellites (EST-SSR) can be located within or near genes conserved between species. The development of genetic maps containing this marker type will allow researchers

to identify regions of chromosomal synteny between species, integrate physical and genetic linkage maps, and identify potential candidate genes for marker-assisted selection (Walbieser et al., 2003; Wang & Guo, 2007). Microsatellite-based linkage maps for selective breeding purpose have been established in commercial/ model plant and animal species such as *Arabidopsis* (Quesada et al., 2002 as cited in Ayeh, 2008), cotton (Han et al., 2004), channel catfish (Walbieser et al., 2005), shrimp (Alcivar-Warren et al., 2007), and more are in progress.

In addition to that, this Type I marker are also considered very valuable as they are transferrable between species, or even genera; and can often be used as anchor markers in comparative mapping between species and evolutionary studies (Zhan et al., 2008; Varshney et al., 2005; Zhou et al., 2008). This is due to the intrinsic advantages of ESTs derived from genes that are evolutionarily conserved, thus enabling successful cross-species PCR amplification compared to anonymous SSRs. A number of previous studies have demonstrated the transferability of EST-SSRs in related species. For instance, all primer pairs developed in rainbow trout (*Onchorynchus mykiss*) have shown ability to amplify expected size amplicons in at least four of the nine other salmonid species (Coulibaly et al., 2005). Apart from that, Li and co-workers (2010) have developed a set of EST-SSR markers for Pacific oysters (*Crassostera gigas*) and transferability of the markers examined on five other species was successfully achieved for all markers in at least one species.

Apart from those aforementioned applications of EST-SSR, currently this marker has gained more attention from scholars to assess genetic diversity of populations of interest. Practical use of EST-SSR in this area, specifically on *Macrobrachium rosenbergii* species will be highlighted in the next section.

2.2 *Macrobrachium rosenbergii*.

2.2.1 Nomenclature and taxonomy

Macrobrachium rosenbergii, or better known as giant freshwater prawn (de Man) is an economically important species belonging to the genus *Macrobrachium* Bate, 1868 (Crustacea: Palaemonidae). Apart from that, it is also being known as giant river prawn, Malaysian prawn, and more popularly called as 'Udang Galah' by local people in Malaysia. About 200 species of the genus have been described, and by far, this species is the most widely cultured species among all (Nandlal & Pickering, 2005). Previously, this species was known by several generic names including *Palaemon carcinus*, *Palaemon dacqueti*, and *Palaemon rosenbergii*. It was only in 1959 that its present scientific name; *Macrobrachium rosenbergii* (De Man 1879) is universally acknowledged (New, 2002).

The taxonomic classification of *Macrobrachium rosenbergii* is shown below:

'Family tree' of giant river prawn *Macrobrachium rosenbergii*

Kingdom	Animalia - animals
Phylum	Arthropoda - insects, spiders, crustaceans etc.
Subphylum	Crustacea - crabs, lobsters, shrimp, etc.
Class	Malacostraca
Order	Decapoda
Sub-order	Pleocyemata
Infraorder	Caridea, sometimes called Natantia
Superfamily	Palaemonoidea
Family	Palaemonidae
Subfamily	Palaemoninae
Genus	Macrobrachium
Species	*Macrobrachium rosenbergii* – giant river prawn

Figure 2.1: Taxonomy of giant river prawn, *Macrobrachium rosenbergii*
(Adapted from Nandlal & Pickering, 2005)

Macrobrachium rosenbergii have become one of the most important aquaculture candidates due to its dainty taste as well as high protein content, ease of culture, and global export potential (Whangchai et al., 2007). Since the first discovery made by a Food and

Agriculture Organization (FAO) expert, Dr Shao-Wen Ling (New, 2010); on the method of raising the prawn fries to the juvenile stage in the 1970's, the modern culture of this species continue to expand till date (Mather & de Bruyn, 2003; Bhassu et al., 2008). In addition to the previous factors, the extensive farming of *M. rosenbergii* including big and small scale domestication is attributed to several other important factors such as fast growth rate, large in size, better meat quality, omnivorous feeding habit, and the established domestic and export markets in Asia (Nandlal & Pickering, 2005).

Early works by De Man (1879) recognized two forms of *M. rosenbergii* species based on morphometric data, and this finding was later supported by morphological analysis performed by Johnson (1973) who introduced the connotation of two subspecies under *Macrobrachium rosenbergii*, namely eastern form and western form (as reviewed by Holthius & Ng, 2010). Many research works conducted since then have shown to be supporting this foundation. In addition to these two forms, another distinct form or 'race' of giant freshwater prawn also being discovered through allozyme and morphological studies of wild stocks achieved by Malecha and colleagues (1977, 1987), as well as Hedgecock et al. (1979); and later recognized as Australian 'race' (as reviewed by Mather & de Bruyn, 2003).

Morphometric and allozyme data in previous studies by Lindenfelser (1984) concluded that there is a biological boundary between the eastern and western *M. rosenbergii* forms which corresponds approximately to geographical frontier namely Wallace's Line (as cited in de Bruyn et al. 2004), and it is further supported by morphological studies accomplished by Wowor & Ng (2007). Besides that, another separate nuclear DNA-based studies using mitochondrial DNA (de Bruyn et al, 2004) and microsatellite markers (Chand et al., 2005)

also provides evidence that complement earlier research and further corroborating the previous conclusion on the distinct forms of *M. rosenbergii* species.

Different nomenclature has been suggested in review by Holthius and Ng (2010) to distinguish both eastern and western subspecies of *M. rosenbergii*, in which the eastern name remain as *Macrobrachium rosenbergii rosenbergii* (De Man, 1879), while the western name is given as *Macrobrachium rosenbergii dacqueti* (Sunier, 1925). The geographical distribution of both forms, or appropriately described as clades is varied; in which the western subspecies are found in the range that includes the east coast of India, the Bay of Bengal, the Gulf of Thailand, Malaysia, and western Indonesia (east to Borneo and Java), whilst the eastern subspecies resides in the Philippines, the Indonesian region of Sulawesi and Irian Jaya, as well as Papua New Guinea and northern Australia (New, 2002; Holthius & Ng, 2010).

Inconsistencies with the current classification may hamper conservation efforts of the *M. rosenbergii* species from genetic perspective. Nevertheless, the freshwater farmers give less attention pertaining to this, and consider that the exact nomenclature has little relevance, since this species has been transferred within its natural geographical range and been introduced into more countries where it may become established (New, 2002). Regarding this issue, more efforts should be focused to classify the *M. rosenbergii* and its sub-species into precise taxonomic nomenclature (Wowor & Ng, 2007). This is because the present conclusions may create confusions and conceive unknown genetic implications especially for scientific communities such as biologists and aquaculturists in utilizing *M. rosenbergii* as a matter of concern in their future research.

Hence, the taxonomic issue would not be addressed in this study as the subjects are strictly being sampled from Malaysian river systems. Therefore, term *M. rosenbergii* as defined here remains as the term referring to prawn individuals in each population.

2.2.2 The distribution and habitat description

Generally, the natural distribution of *M. rosenbergii* extends from Pakistan in the west to southern Vietnam in the east, across South East Asia, and south to northern Australia, Papua New Guinea, and some Pacific and Indian Ocean Islands (Mather & de Bruyn, 2003). However, this range of distribution has been illustrated to be geographically diverged for different forms or clades of *M. rosenbergii* aforementioned. According to studies achieved by Malecha et al. (1980), the distribution of the species and the division of three different types is displayed in the Figure 2.2 below (as reviewed by Malecha et al., 2010).

Since the first introduction of *M .rosenbergii* broodstock into Hawaii from Malaysia in 1965, it has been introduced into almost every continent for farming purposes following the success of mass production of postlarvae, PL for commercialization (New, 2002). The rearing of *M. rosenbergii* has now been expanded to many countries even in places outside their natural range and these include Bangladesh, Brazil, China, Ecuador in South America, India, Malaysia, Taiwan Province of China, Vietnam, and Thailand (New, 2002). In the year 2000, the number of producers was reported to be more than thirty countries indicating the established market of *M. rosenbergii* across the world regions (New, 2002). Among other listed countries that practice commercialized culture of this species include Hawaii on The USA, Honduras, Mauritius, Costa Rica, Israel, and Mexico due to its potential market values (New, 2002; Bhassu et al., 2009).

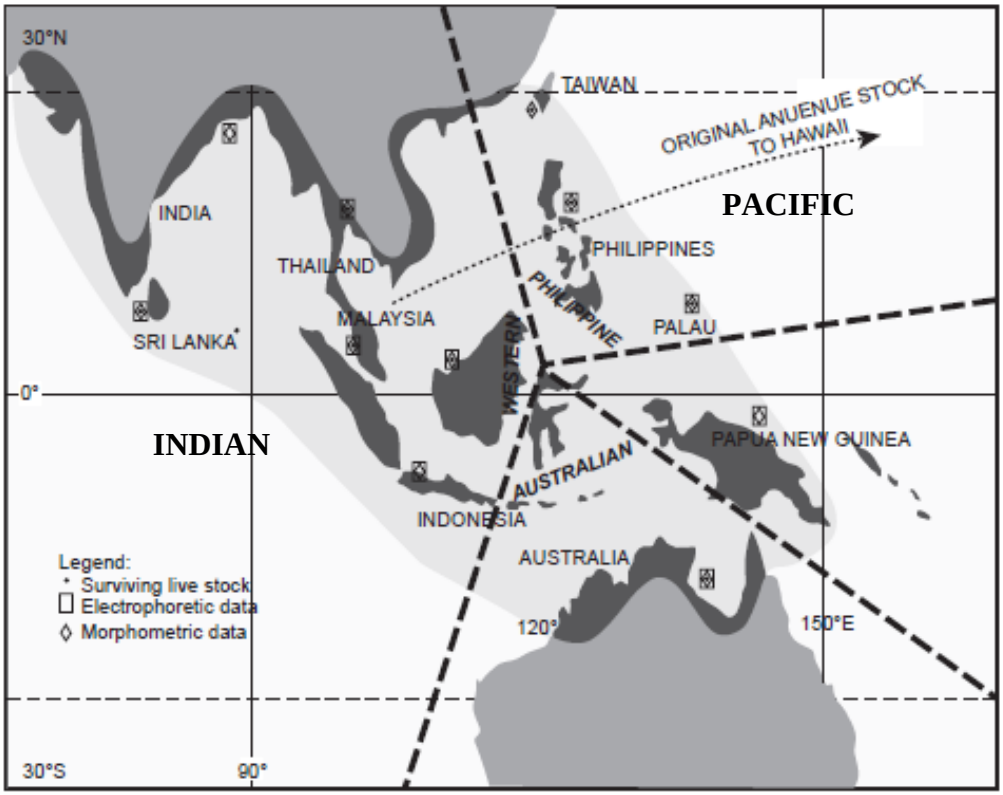


Figure 2.2 : Natural distribution of *M. rosenbergii* (shaded area)
The approximate locations of the 'Western', 'Australian', and 'Philippine Eastern' forms as described by Malecha (1980). Dotted line refers to the origin of the Anuenue stock imported from Penang, Malaysia into Hawaii by Fujimura.
(Picture taken from Malecha et al., 2010)

The giant freshwater prawns, as denoted by its name generally live in turbid freshwater areas, circumtropical marine, and estuarine including lakes, rivers, swamps, irrigation ditches, canals, and ponds (de Bruyn et al., 2004; Bhassu et al., 2009). However, the *M. rosenbergii* larvae were discovered to be required of brackish water for survival and early development (New & Singholka, 1985 as cited in Mather & de Bruyn, 2003). In Malaysia, the population of this freshwater prawn can be easily found in drainage systems in Peninsula Malaysia and Borneo Archipelago, indicating a diverse distribution of its habitats in Malaysia. A numbers of rivers such as Sg Timun in Negeri Sembilan and Klias Wetland in

Sabah are renowned among locals as famous spot whereby the freshwater prawns can be found in abundance especially during the dry season.

2.2.3 Morphological and Biological Characteristics

The Malaysian giant prawn *Macrobrachium rosenbergii* (de Man) has received considerable attention as a fresh-water aquaculture organism. It is the largest species of the genus: the males can reach a total length (from tip of rostrum to tip of telson) of 320mm, the females of 250mm (Brown et al., 2010). According to Brown et al. in *Freshwater Prawns* (2010), *M. rosenbergii* is further described as: "the body is usually of a greenish to brownish grey, sometimes more bluish color, and is darker in larger specimens. There are irregular brown, grey, and whitish streaks, often somewhat placed longitudinally. The lateral ridge of the rostrum may show a red color. An orange spot is present on the articulations between the abdominal somites. The antennae are often blue. The chelipeds may also be blue. All these colors are brighter in the smaller than in the very large specimens," (Figure 2.3).

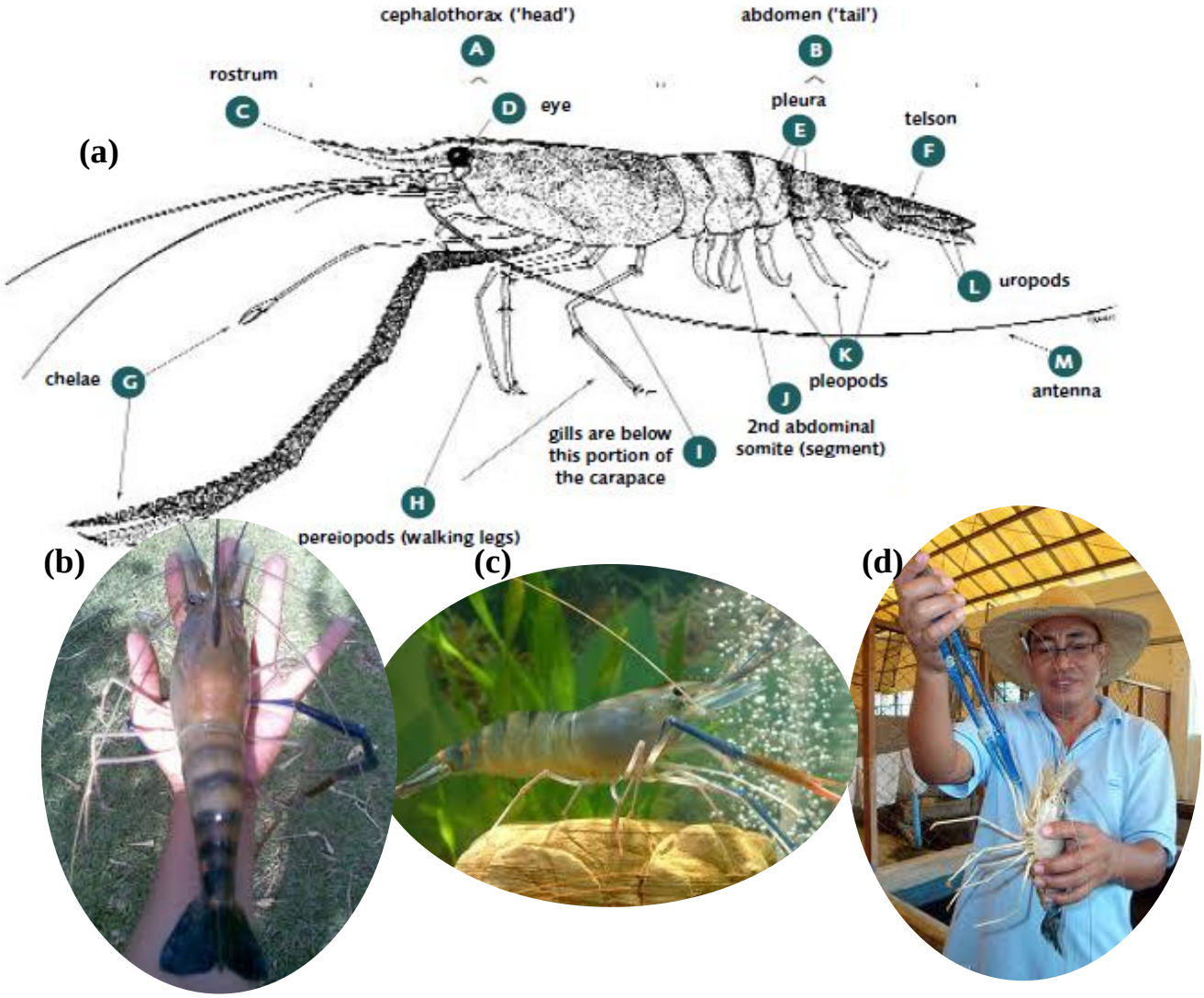


Figure 2.3 : From counterclockwise (a) The external features and morphology of giant freshwater prawn, *Macrobrachium rosenbergii*; (b) Size of adult prawn relative to human hand; (c) Live individual giant freshwater prawn; (d) A man holding an adult male prawn

(All pictures source from Google images, except (a) from New 2002, source: Emanuela D' Antoni)

Like all species of decapods, the prawn body consists of two distinct parts; cephalothorax (A) and abdomen (B)- see Figure 2.3, and is divided into 20 segments (Nandlal & Pickering, 2005; Sharma & Subba, 2005; Brown et al., 2010). Fourteen of these segments are in the cephalothorax and covered by the carapace which is smooth and hard (New, 2002; Nandlal & Pickering, 2005). Six segments are located in the front portion of the head while the rear portion has the rest of the segments, each of which has a pair of appendages (Nandlal & Pickering, 2005).

To identify *M. rosenbergii* from other freshwater prawn species, combination of the following characteristics is usually being observed (New 2002, Nandlal & Pickering, 2005, Brown et al., 2010):

- It is known as the largest *Macrobrachium* species (total body length up to 320mm)
- Adult male has a pair of very long legs (chelipeds)
- The tip of its telson reaches distinctly beyond the posterior spines of the telson
- The rostrum is long and bent in the middle with 11-13 dorsal teeth and 8-10 ventral teeth
- The movable finger of the leg of the adult male is covered by a dense mat of spongy fur
- Distinct black bands on the dorsal side at the junctions of the abdominal segments

Mature *M. rosenbergii* male differentiate into three distinct morphotypes (see figure 2.4), and these distinguishable morphologies were observed to be associated with differences in growth rate of adult prawns (Cohen et al., 1987). Some of the large males had very long second pereiopods (claws) that were deep blue in color and were termed blue claw (BC) males, while some presented as orange-clawed (OC) males that were also large and had long claws (but shorter than BC males) that were usually orange in color. The remaining males

were small and had short claws that were often relatively unpigmented and translucent. These were termed small males (SM). Both SM and BC grew slowly, whilst OC had very high growth rates. However, OC had poor mating success compared to BC which is the most dominant and terrestrial compare to other morphotypes (Ra'anan, 1982, Ra'anan & Sagi, 1985; as cited in Cohen et al. 1987).

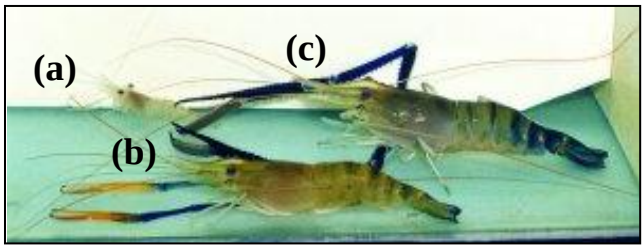


Figure 2.4 : From counterclockwise (a) Small male,SM; (b) Orange-clawed male, OC; (c) Blue-clawed male, BC
(Pictures source from FAO.org)

These freshwater prawns are nocturnal species, which are active at night. Post-larvae and adult *M. rosenbergii* are omnivorous and feed voraciously on a variety of food items. They consume algae, aquatic plants, mollusks, aquatic insects, worm and other crustaceans (John, 1957; Ling 1969 as cited in Brown et al., 2010), whereas the diet of larvae is principally zooplankton (mainly minute crustaceans), very small worms, and the larval stages of other crustaceans (New & Singholka 1985 as cited in Brown et al., 2010). Cannibalistic behavior may occur in the scarcity of food, and/or in overpopulated ponds.

2.2.4 Life cycle

The giant freshwater prawn has four phases in its life cycle: egg, larva (zoea), postlarva (PL) and adult. The time spent in each phase and its growth rate is affected by the environment, especially water temperature and food (Nandlal & Pickering, 2005). The male and females reach first maturity at about 15-35g within 4 to 6 months. Similar to other crustacean, the growing and size development of *M. rosenbergii* occur through moulting process. While it is considered as a freshwater species, the larvae stage depends on brackish water. Once it becomes a juvenile, it will return back and live entirely in freshwater.

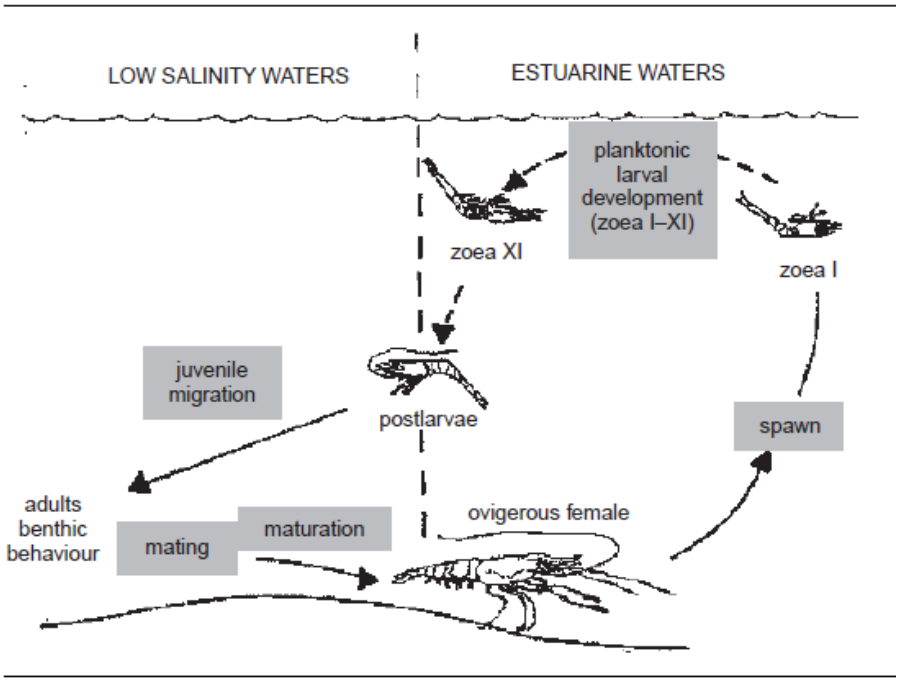


Figure 2.5 : Life cycle of *M. rosenbergii*. Early development and survival of larvae is achieved in low salinity water. Ovigerous (fully mature) females migrate from freshwater to estuarine areas to spawn, where free-swimming larvae hatch from eggs attached to female abdomen. Metamorphosis of larvae into post-larvae take place (after 3-6 weeks) and the newly PL then migrate upstream towards freshwater (Mather & de Bruyn, 2003)

(Pictures taken from Brown et al., 2010)

2.3 The significance of studying genetic diversity of *Macrobrachium rosenbergii*'s wild population using EST-SSRs.

According to New (2005), most global production of freshwater prawn comprise the species of *M. rosenbergii*, despite the expanding productions recorded for other related species under the same genus such as *M. nipponense*, and *M. malcomsonii*. While Vietnam, Thailand, and Taiwan as well as Bangladesh were listed among top producers, Malaysia also manage to be into the list. In Malaysia, the production for local market had showed rapid expansion between 1997 and 2000, but fell back somewhat in 2001 (New, 2005). Many factors contributed to the decline, and one of it was probably the depletion of wild broodstock since the local giant freshwater prawn, GFP industry is strongly dependent on the wild resources. Apart from Malaysia, the trends of GFP industry in Bangladesh also seem to be overwhelming as the export market is also heavily depending on the wild-caught freshwater prawn (New, 2005).

To ensure the uninterrupted supply of broodstock in hatcheries, the juvenile and mature prawn individuals were obtained from rivers and freshwater areas, and then reared in captivity for grow-out operations in farms till several generations before the new batches of wild individuals are seized and farmed again. Some of the prawn farms entrepreneurs also practice commercialization by using wild-caught berried females to harvest the eggs, and those prawns were later been eaten or sold for the food market (Mr. Badrul Nizam, personal communication, 22 May, 2011). The trend of harvesting the undomesticated broodstock is undisputedly jeopardizing the reservoir of prawn diversity and population, thus will subsequently lead to scarcity of sustainable resources for the GFP industry.

As *M. rosenbergii* has been esteemed as delicacies in Asia, its popularity has significantly escalated and the demand keeps growing and growing by day. The need of continuous broodstock supplies for the industry can be achieved by two alternative solution; either using more wild broodstock, or speeding up the domestication of the species via genetic improvement (New, 2002). However, the dependence on wild broodstock affects the planning and management of hatcheries due to seasonal factor that influence the reproductive cycle of *M. rosenbergii*. In addition to that, transportation of wild stock to hatcheries not only will increase the costs but also causes stress to the animal, and eventually lead to low hatching rates (Mohanta, 2000).

The decline of wild resources of GFP at an alarming rate implies that prompt action on proper management and conservation strategies of the wild stock species is pivotal at this current stage, instead of overharvesting them. Furthermore, wild stocks also offer abundance of genetic diversity which is essential in establishing base populations to initiate breeding and domestication of prawn cultures. Again, to achieve the goal in genetic improvement of prawn strains, the proper assessment and documentation on prawn diversity in their natural habitats have to be done (Mather & de Bruyn, 2003). The wild and captive genes within their natural gene pool that are amenable to genetic improvement are important to enhance the productivity of stock in cultured line, thus resulting in commercial success of *M. rosenbergii* farming.

The application of EST-SSRs in studying genetic variation of wild populations of *M. rosenbergii* is one option that can appropriately address this issue. This population genetic study with utilization of newly developed EST-SSRs from the transcribed region of *M. rosenbergii* genome, will add more knowledge and data necessary for future studies in

establishment of breeding program with the aim in improving prawns productivity. The utility of this molecular marker is not only restricted to population genetics. Beyond the diversity studies, EST-SSRs can also be used for mapping quantitative trait loci (QTL) and in marker assisted selection (Erhardt & Weimann, 2007). Recently, many studies have reported the application of EST-derived SSRs for assessment of genetic variation in a wide range of organisms including several commercial aquaculture species such as mentioned earlier.

Through population genetic analyses, a variety of data which could serve as important evolutionary parameters are made available, include the status of genetic variation, the partitioning of this variability within/between studied populations, overall level of inbreeding, effective population sizes and the dynamics of recent population bottlenecks (Ellis & Burke, 2007). The information on genetic diversity provided by utilization of EST-SSRs not only generate estimates of variability parameters which are vital to prawn breeders to identify functional gene for important traits; but also prove useful in discovery of basic evolutionary insights, as well as assists on conservation priority decision for prawn strains that have unique traits (Ellis & Burke, 2007; Erhardt & Weimann, 2007; Duran et al., 2009).

Combining both functional features and considerable high variability characteristics from EST and SSR respectively, these genic-microsatellites serve as most rapid and cost-effective measures of genetic diversity through assay of polymorphism. While visible phenotype reflects crude estimates of functional variants of genes, the transcript-derived marker will provide direct information of novel, functional genes for commercially important traits such as high productivity and resistance towards viral infections (Anon.). Varshney et al. 2005 suggested that the presence of SSR in the transcribed region may play a role in gene expression or function; and/or they could probably be associated with unusual phenotypic

variation. However, the latter assumption has yet to be proven. Owing to this functional marker, evaluation of variation and identification important trait like Quantitative Trait Loci (QTL) can be achieved. When marker-QTL associations are identified, marker-assisted-selection (MAS) can be applied in prawn breeding programs with the aim of improving selection response (Erhardt & Weimann, 2007).

Characterization of genetic variation within natural populations also provides valuable information in the investigation of the origin and domestication of prawn species, as well as useful for estimating genetic relationships and kinships among strains (Varshney et al., 2007; Anon.). Patterns of migrations accomplished by construction of phylogenetic trees throughout their natural distribution provide insight on the prawns' evolutionary relationship. Apart from that, utility of EST-SSRs in population genetic studies helps in identifying geographical areas of admixture among populations of different genetic origins (Anon.).

Besides that, assessment of genetic variation has enabled genetic identification and discrimination of broodstocks that can be achieved through strain comparison, therefore assisting in selecting the best strain for higher productivity of prawn culture. For instance, the comparative work done by Bart & Yen in 2001 on various strains of freshwater prawns revealed the success rate of prawn survival during post-larvae stage, and this will eventually lead to potential improvement that could be gained from selection and/or cross-breeding of those strains (New, 2005). Moreover, the evaluation of genetic variation using SSRs derived from EST source not only displays their utility between populations within a species, but might also be possible between populations across species due to the expectation that the

flanking sequences may be relatively conserved between species, thus making EST-SSRs to be more transferable compared to 'anonymous SSRs' (Varshney et al. 2002; Ellis & Burke, 2007).

Additionally, adequate documentation on the patterns and the extent of genetic information that is present in wild populations is indeed vital in developing effective management strategies and conservation efforts for prawn stocks (Mather & de Bruyn, 2003). As resources for conservation are limited, prioritizations are often necessary to conserve breeds that have plenty of unique traits. By using gametic phase disequilibrium of DNA marker polymorphism, the effective population size (N_e) can be estimated. N_e is an index that is linked to levels of inbreeding and genetic drift in populations, and be able to serve as a critical indicator for assessing degree of endangerment of population (Anon.).

Concisely, vast applications of these newly developed EST-derived SSRs for *M. rosenbergii* remain to be discovered. The initial data provided in this current study substantially help in understanding the molecular relationships among wild stocks, as well as offer foundation for future studies. Through ongoing studies, the ultimate goal in establishing breeding program for future genetic improvement will certainly be realized for the development of sustainable giant freshwater prawns industry not only in Malaysia, but also throughout the globe.

CHAPTER III

Methodology

3.1 Materials

3.1.1 Experimental animals and sources of DNA

A total of 120 wild adult *M. rosenbergii* were collected from four locations in Malaysia, selected for analysis of genetic diversity. Details on these sampling sites including state of origin and geographic information are provided in Table 3.1. The sampling locations are shown in Figure 3.1. Samples of muscle tissue were obtained from each individual and stored in a freezer (-80^{o}C) for preservation prior to DNA extraction.

Table 3.1: Details on prawn individuals (genotypes) and sampling sites for assessment of genetic diversity

Population (Individual code)	State of Origin	Sampling Site	Longitude/ Latitude	Sample size
Sg Timun (B1-B10, J1-J20)	Negeri Sembilan, Malaysia	Sg Timun (Sg Linggi)	2^{0}28'29"N 102^{0}02'05"E	30
Kedah (K1-K30)	Kedah, Malaysia	Sg Muda	5^{0}43'01"N 100^{0}31'46"E	30
Sarawak (S1-S30)	Sarawak, Malaysia	Sg Serian	1^{0}50'05"N 113^{0}54'06"E	30
Terengganu (Te1-Te30)	Terengganu, Malaysia	Sg Penarik	5^{0}37'48"N 102^{0}48'36"E	30

Total genomic DNA was isolated from 100mg muscle tissue samples from each individual using the modified CTAB method described by Doyle & Doyle (1987) (Appendix A). The DNA concentration was measured spectophotometrically by NanoVue^{TM} (GE Healthcare, NJ, USA).

Figure 3.1: The sampling locations

Legend:

1- Sg Timun, Negeri Sembilan
2- Sg Muda, Kedah
3- Sg Penarik, Terengganu
4- Sg Serian, Sarawak

3.2 Methods

3.2.1 Detection of EST-microsatellite markers and primer design

These bioinformatics analysis and primer design were performed by a research assistant, Maizatul Izzah Mohd Shamsudin and further reported in Bhassu *et al.* (in press).

Macrobrachium rosenbergii EST data previously obtained from transcriptome sequencing (in-house; unpublished data) were screened for microsatellites (or short tandem repeats (SSR)) using the iQDD program ((http://primer3.sourceforge.net/, Meglécz et al., 2009). The analysis implemented in iQDD involves three successive stages: sequence cleaning and detection of microsatellites, sequence similarity detection, and microsatellite selection and primer design (Meglécz et al., 2009). In this study, only perfect microsatellites were targeted, and identification of microsatellites was limited to the detection of strings of repeats sequences that contained a minimum of four motif repeats for all di-, tri-, tetra-, penta-, and hexanucleotide motifs.

After microsatellite regions were identified, all-against-all BLAST was carried out to detect sequence similarity, and those that showed multi-hit were omitted from the data analysis; leaving only unique EST sequences. To ascertain the identity of these transcripts, all non-redundant transcripts were annotated against NCBI database using homology search BLASTX tool NCBI (Nawrocki et al., 2009) with masking of low complexity region. A unique set of microsatellite-containing ESTs with annotated gene were obtained prior to the primer design step. Details on this work are presented in a forthcoming paper (Bhassu *et al.*, in press). Primer design was carried out for sequences with minimum number of five repeats using QDD built-in Primer3 program (Meglécz & Martin, 2009).

The major parameters for primer design were set as follows: primer length 20bp, PCR product size 150-320bp, range of annealing temperature 57^{o}C to 63^{o}C with 60^{o}C as the optimum, and GC content 20-80% with 50% as the optimum content.

3.2.2 Overview: validation of microsatellite loci

Primers were synthesized for 60 microsatellite loci and then initial validation was performed to confirm that the microsatellite regions could be amplified from genomic DNA. Validation included optimization of annealing temperature via temperature gradient PCRs; and amplificability of targeted products. This step was carried out using several individuals selected from each studied population. Successful PCR amplification was determined by agarose gel electrophoresis, and primers with no significant amplification (i.e., visual product of expected size) were then discarded from further data collection.

After initial validation, microsatellite markers were screened for consistency of amplification and verification of allele size polymorphism using genomic DNA from 16 individuals of *M. rosenbergii* randomly selected from all four sample populations. In this part of the screening process any primer set that exhibited significant stuttering, possible occurrence of null alleles, or a monomorphic pattern were excluded from further data collection. The degree of polymorphism at each locus was assessed based on the clear resolution of different-sized PCR product on metaphore gels, utilising the set of 16 randomly selected individuals from the different sample sites.

Those loci that appeared to exhibit polymorphism were targeted for assessment of Polymorphic Information Content (PIC, Shete et al., 2000). In this step -FAM fluorophore labeled primers were used during PCR and the resulting product was subject to fragment analysis on an ABI PRISM ® 3130*xl* Genetic Analyzer (Applied Biosystem, USA). Sequencing results were interpreted to score the alleles present at each locus in 32 randomly individuals, and this data was used to calculate the PIC value for each locus. Only microsatellite markers that had passed the initial validation and polymorphism screening and had shown potential for high polymorphism as determined by PIC value were retained for subsequent data collection and analysis of all samples of four wild Malaysian river populations.

3.2.3 PCR conditions and gel electrophoresis

Polymerase Chain Reaction (PCR) amplification for each primer set was performed in a C1000 Thermal Cycler (Bio-Rad) in a total volume of 10µl reaction solution consisting of 2µl of DNA extracted from tissues, 1.5 µl $MgCl_2$ (25 mM), 3.0µl of 1X PCR Buffer (Promega), 0.25µl of each dNTPs (10mM), 0.3µl of *Taq* Polymerase, and 0.5µl of each primer (10mM). The PCR reactions were carried out as follows: initial denaturation at 96^{O}C for 3min, 39 cycles of denaturation at 94^{o}C for 10s, annealing temperature for 10s, and 30s of extension step at 72^{O}C. The program was then completed with a final extension at 72^{O}C for 7min. Initial PCR reactions were performed across an annealing temperature gradient (55-65^{o}C) to determine the best annealing temperature for each primer pair, with subsequent PCR reactions conducted at this optimal temperatures (see Table 3.2).

Following amplification, the presence of PCR products were verified via electrophoresis. 1.0% agarose gel was used in electrophoresis of PCR products for optimization, whilst 4.0% Metaphor® agarose gel was used in electrophoresis of microsatellite PCR products in polymorphism screening. The gel electrophoresis for agarose and Metaphor® were carried out at 70-75V, 150mA using 1xTBE Running Buffer for 45min and 1.5-2hours respectively. The gels were stained with ethidium bromide (10mg/ml) before being visualized under ultraviolet light (Alpha Imager Gel Documentation System, Siber Hegner, Germany).

Table 3.2 : List of primers with optimized annealing temperature (^{o}C)

Gene ID	Primer sequences 5'-3' (Forward and Reverse)	Annealing temperature (^{o}C)	Expected Product Size (bp)	Motif repeat
EST MR 5	5'-TTC CCC AAT GCT TCT TCA TC-3' 5'-ACG CAC CTC CTT GTA TCC AC-3'	55.0	150-177	$(TC)_6$
EST MR 8	5'-ACT TCT TGG CTT CAA GGG CT-3' 5'-TCC AGT CAA AAG AAT TCG CA-3'	55.0	150-200	$(TC)_6$
EST MR 13	5'-TGG ACA TCT TTG CAT AGC CA-3' 5'-CAC ATC GGG GTT ATT TTG GT-3'	61.4	150-160	$(TC)_7$
EST MR 14	5'-CTC TGC TTC GTA AAA TCG CC-3' 5'-GAA CAC TTT TGG CAT GGG AG-3'	61.4	150-163	$(CT)_7$
EST MR 37	5'-GTT ACC AGG TGC CAG GTC C-3' 5'-GCT TCT TGA CCG AGA ACA CC-3'	64.5	150-155	$(GCT)_6$
EST MR 41	5'-TCT CGT GTG ACA TAG GCA GC-3' 5'-GCA GAG AAC AAG ATT TCT ACC TCC-3'	63.3	150-190	$(TCA)_6$
EST MR 51	5'-AGC TGT ACA CCT CTG GCT CG-3' 5'-CTA CGA AAC GCA TGG TTG G-3'	63.3	150	$(CTT)_7$

3.2.4 Fragment Analysis

EST-SSR markers that exhibited potential polymorphism as identified from initial screening (Section 3.2.2) were used to screen 32 samples randomly selected from among all sampled populations for size polymorphism. One primer in each set was labeled with -FAM fluorophore, PCR reactions were performed following the protocol detailed above (Section 3.2.3), and product was run on an ABI PRISM ® 3130*xl* Genetic Analyzer (Applied Biosystem, USA).

In preparation for fragment analysis labeled PCR products were diluted 1:9 or 1:19 in ddH_2O depending on the amount of amplicon present as revealed by gel electrophoresis, and then 1 µl of the diluted mixture was transferred into a new PCR tube and 10 µl of Hidi Formamide loading dye (Analisa Resources (M) Sdn. Bhd.) and 0.2 µl of Genescan 500 LIZ ladder were added into the tube. The mixture was thoroughly mixed by brief vortexing, followed by a brief centrifugation step. Next, the tubes were heated for 5min at 95°C to denature double stranded PCR product, and then kept in ice for exactly 5min, before all samples were transferred into 96-well plate and subjected to fragment analysis using ABI 3130 Genetic Analyzer.

Results of the fragment analysis were interpreted, evaluated and allele sizes were scored using the software packages GeneMapper 4.0 (Applied Biosystems, Foster City, CA, USA) and Peak Scanner v1.0 (Applied Biosystem, USA). Genotyping of each individual at each locus was accomplished by scoring peaks in electropherogram which represent exact alleles sizes in base pairs (bp) of amplified loci.

.

After genotypes had been determined for 32 individuals at each locus, allelic frequencies were used to calculate the Polymorphic Informative Content, PIC (Shete et al., 2000) value for each locus. Only markers with high PIC (more than 0.5), and hence high polymprphism were used to screen variation in the full data set of 120 individuals from all four populations.

The PIC values measure the informativeness of a given DNA marker, and were calculated following the method of Shete et al., 2000 using CERVUS software (http://www.fieldgenetics.com/pages/aboutCervus_Overview.jsp)

$$PIC = 1 - \sum_{i=1}^{k} Pi^2$$

where,

k = total number of alleles detected for a given marker locus

Pi = Frequency of the *i*th allele in the set of genotypes investigated

After loci with high PIC values were identified 6-FAM fluorophore labeled primers were synthesized commercially (First Base, Malaysia). Genomic DNA was amplified using these labeled primers under the PCR reaction conditions described earlier.

3.2.5 Data analysis

Data analysis was undertaken to examine levels of diversity at each locus, and also to determine the extent of population genetic structure present among sampled populations. Prior to data analysis, all raw genotypic data obtained from GeneMapper 4.0 (Applied Biosystems, Foster City, CA, USA) software was collated in Microsoft Excel, and then data files for specific population genetics software were generated using the program CONVERT software version 1.31 (Glaubitz, 2004).

3.2.5.1 Identification And Checking For Scoring Errors

The data set was checked for any genotyping errors that could potentially bias population genetic analysis. These genotyping errors include incorrectly scoring individuals as homozygotes because mutations in the priming site result in non-amplification of specific alleles (null alleles), or because PCR may preferentially amplify shorter alleles in heterozygote individuals (short allele dominance / large allele dropout). Other errors include mis-scoring stutter peaks as true alleles, resulting in an artificial excess of heterozygote genotypes with only one motif repeat difference between alleles. Data was checked for errors using Micro-Checker software (Van Oosterhout et al., 2004). Where evidence was found for the presence of null alleles the frequency of null alleles was estimated (Chakraborty et al., 1992; Brookfield, 1996 as cited in Van Oosterhout et al., 2004) using Micro-checker, and the allele and genotype frequencies of the amplified alleles were recalculated and corrected based on new equation to account for the downward bias resulting from the null alleles (Van Oosterhout et al., 2004), thus permitting their use for further population genetic analysis (Van Oosterhout et al., 2004).

3.2.5.2 Tests For Conformation To Equilibrium Expectations

Conformation to Hardy-Weinberg Equilibrium (HWE) was investigated to find out if the samples constitute collections of randomly mating individuals and/or to see if the EST microsatellites showed evidence for non-neutral evolution (i.e., selection). The principle of HWE states that both allelic and genotypic frequencies in a population will not change over generations in the absence of disturbing factors (selection, genetic drift, gene flow, and mutation) with the condition those individuals in the population exhibit random mating. Tests for HWE are based on computation of expected genotypic frequencies under random mating using Chi-Square (χ^2) Goodness of Fit test. This statistical test aims to determine whether the observed data for each particular locus "sufficiently fits" the expected assumption i.e., the population is in the expected HWE proportions (Hedrick, 2000). Hardy-Weinberg (HWE) Principles can be represented by binomial (with two alleles) or multinomial (with multiple alleles) functions of allelic frequencies.

HWE tests were performed in GENEPOP version 4.0 (Raymond & Rousset, 1995), with significance recalculated following the False Discover Rate procedure (Benjamini & Hochberg, 1995). When any evidence of deviations from HWE was found, subsequent analyses were performed including and excluding data from the deviating locus. Markov chain parameters for all tests were set as follow: dememorization-10000; batches-100; iterations per batch: 1000.

Specifically, the calculations performed to determine whether the observed allelic frequencies were consistent with H-W predictions, χ^2 were calculated as:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O-E)^2}{E}$$

Where ,

O; E = observed and expected number of particular types;

k = number of genotpic classes.

Degree of freedom= k-1

(Hedrick, 2000)

Linkage disequilibrium analysis was performed to check for the presence of any non-random associations of alleles among different loci. Associations between polymorphisms at different loci are measured by the degree of linkage disequilibrium (D). D was first proposed by Lewontin and Kojima (1960). Numerically, it is the difference between observed and expected allelic frequencies (assuming random distributions). If the alleles at a pair of loci are not randomly associated with one another, then there will be a deviation (D) in the expected frequencies, in which case the loci are said to be in linkage disequilibrium (i.e., linked).

Pairs of loci may deviate from linkage equilibrium (i.e be in linkage disequilibrium, LD), due to physical or demographic reasons. For example, LD may indicate that the observed loci might be physically linked to each other by occurring in close proximity on the same chromosome. In this case, recombination is unlikely to lead to independent assortment during meiosis, and alleles at different loci that are closely linked physically will tend to be inherited

together. In contrast, LD may also be observed when two divergent populations are sampled together, as divergent alleles at multiple loci in each population are likely to occur together in single individuals, while divergent alleles drawn from different populations are unlikely to occur together. Thus, the occurrence of LD can provide information about the loci or about the populations sampled. Linkage analysis was performed in GENEPOP software (Markov chain parameters set as follows: dememorisation-10000; batches-20; iterations per batch-5000).

3.2.5.3 Estimating Genetic Diversity

To investigate the level of genetic diversity present at each locus with each population sample several measures of genetic variation were calculated. Allelic and genotypic frequencies, observed number of alleles (A), effective number of alleles (Ne), observed and expected heterozygosity (H_O,H_E) were obtained using software POPGENE version 1.32 (Yeh et al., 1997), and allelic richness was calculated using FSTAT software Ver 2.9.3.2 (Goudet, 1995 http://www2.unil.ch/popgen/softwares/fstat.htm).

Allele frequency is one metric used to quantify genetic variation. It is sometimes synonymously used with gene frequency to measure the commonness of a given allele in a population, that is, the proportion of all alleles of particular gene in the population. On the other hand, genotypic frequency can be defined as the proportion of particular genotype relative to all genotypes at a specific locus in a population.

In a sample of N individuals, N_{ii} and N_{ij} are the numbers of A_iA_i and A_iA_j genotypes observed respectively; whereby A_i, A_j are alleles at the particular locus in the sample. To estimate genotype frequencies, the formula is:

$$P\ (A_iA_j) = N_{ij} / N$$

Therefore, the estimated allelic frequencies of allele Ai for codominant, multiple-alleles system, can be calculated from the sample as:

$$P_i = \frac{N_{ii} + 1/2 \sum_{j=1}^{n} N_{ij}}{N}$$

where $j \neq i$.

Several other measures have been used to describe the genetic variation in a population, but heterozygosity remains as the most widespread measure of variation. It is defined as the probability that a random individual chosen from the population is heterozygous at a locus (Shete et al., 2000), and its value ranges from zero to one. Based on known allele frequency, the expected heterozygosity of a randomly mating population for a particular locus with *n* alleles can be calculated as:

$$H_E = 1 - \sum_{i=1}^{n} P_i^2$$

which is one minus the Hardy-Weinberg homozygosity (Hedrick, 2000).

Effective number of allele, n_e basically the inverse of expected heterozygosity and the formula is given as: $N_E = 1 / 1\text{-}H_E$

Meanwhile, the observed heterozygosity (H_o) for a locus can be estimated using formula:

$$H_O = n(A_iA_j)/ N$$

where, $n(A_iA_j)$ = number of individuals with genotype A_iA_j, $i \neq j$

N = total number of individuals in sample

A_iA_j = alleles at the locus

In most outbreeding populations, the observed heterozygosity is quite close to the theoretical heterozygosity. Deviation of the observed from the expected can be used as an indicator of important population dynamics.

Assessing allelic richness (A) in a set of populations was achieved by estimating the number of alleles expected in samples of specified size using rarefaction approach. This approach uses the frequency distribution of alleles at a locus to estimate the number of alleles that would occur in smaller samples of individuals, and A is standardized to the smallest samples size (N) in a comparison (Petit et al., 1998 in Leberg, 2002). Allelic richness may be useful as an indication of a decrease in population size or of past bottleneck (Nei et al., 1975).

3.2.5.4 Measuring Sub-Population Differentiation

Several factors, such as geographic, ecological, or behavioral differences can result in population subdivision, and consequently change the level of genetic connectedness (gene flow / effective dispersal) among sub-populations. The extent to which sub-populations differ, or degree of genetic differentiation present, can be estimated in a number of ways. One group of methods employs the partition of overall genetic variation present with a group of samples, into components within and among sub-populations (Wright, 1951 as cited in Hedrick, 2000). This approach of partitioning the genetic variation as described by Wright (1951) consists of three F coefficients (Hedrick, 2000). F_{ST} is a measure of the genetic differentiation over subpopulations ($_S$) relative to total population level ($_T$) and described differentiation among sub-populations. The pairwise F_{ST}s can be used as short-term genetic distances (dissimilarities) between populations and are given in the form of a pair-wise matrix between all pairs of sample sub-populations. F_{IS} and F_{IT} are measures of the deviation from Hardy-Weinberg proportions within subpopulations and in the total population, respectively, where ($_I$) represent individuals.

The F_{IS} fixation index relates to an approximation of the deviation of the observed heterozygosity, H_O from the expected, H_E, and is calculated based on Wright model (1951) with the incorporation of more sophisticated model as implemented in Arlequin (Excofifier et al., 2005); with 1023 permutations. If the H_E is higher than H_O, then the value of F_{IS} will be large (positive value), suggesting inbreeding has resulted in a reduction in heterozygotes at a particular location.

The fixation index of F_{ST} measured genetic divergence among populations and can be expressed as :

$$F_{ST} = \frac{H_T - H_S}{H_T}$$

Where : H_S is expected Hardy-Weinberg heterozygosity averaged across all subpopulations

: H_T is expected Hardy-Weinberg heterozygosity for the total population (equivalent to H_E)

F_{ST} ranges from 0 (no differentiation between subpopulations) to 1 (fixation of different alleles in fragments).

An alternative method to compute the fixation index R_{ST} (Slatkin, 1995 as cited in Excofifier et al., 2005), adapted to microsatellite loci by assuming high-rate stepwise mutation model can be calculated as:

$$R_{ST} = \frac{S_T - S_W}{S_T}$$

Where : S_T is the twice the variance in allelic sizes in the total population and S_W is twice the average of the within subpopulation variance in allelic sizes.

Pairwise F_{ST}s were calculated to compare genetic differentiation among the sample populations using Arlequin software. (Excofifier et al., 2005). In addition, an analog of Slatkin's R_{ST} was also employed to exaime inter population differentiation as microsatellites data was specifically being analysed. Significance values for F_{ST}s and R_{ST}s were adjusted following the FDR procedure (Benjamini & Hochberg, 1995).

A third pair-wise index was used to examine inter-population differentiation; *D* diversity indices (Jost, 2008). According to Jost (2008), *D* is actual differentiation that quantifies diversity at a gene locus in terms of effective number of alleles rather than heterozygosity which underlies other measures of genetic differentiation, such as F_{ST}, R_{ST} and G_{ST} (Nei, 1972). These other statistics, when used as a measure of gene diversity, suffer from certain limitations surrounding the handling of mutation and additive heterozygosity. This means that it can sometimes be difficult to compare results obtained from loci with markedly different mutation rates (Ryman & Leimar, 2009); and thus results based on these indices may lead to biased conclusions on the extent of population differentiation (Jost, 2008). *D* can be defined as:

$$\text{D, actual differentiation} = \left(\frac{(Ht-Hs)}{(1-Hs)}\right) . \left(\frac{n}{(n-1)}\right)$$

Where = *Hs*, *Ht* represent expected heterozygosities within sub-population and total population respectively; *n* is number of subpopulation

D was calculated with the web-based application SMOGD (http://www.ngcrawford.com/django/jost/). The degree of congruence among population differentiation measures was assessed by Mantel tests. This statistical test evaluates the correlation between two square matrices (often distance matrices) (Urban, 2003). It is a regression in which the variables are themselves distance or dissimilarity matrices summarizing pairwise similarities among sample locations (Urban, 2003).

The Mantel statistic is based on the simple cross-product term:

$$Z = \sum_{i=1}^{n}\sum_{j=1} X_{ij}.Y_{ij}$$

where = x and y are variables measured at locations i and j ($i \neq j$)
= n is the number of elements in the distance matrice

The significance of Mantel coefficient is evaluated as the result from repeated randomization with null hypothesis assuming no relationship between matrices. Values of Mantel statistics traditionally range from 0 if the compared matrices are not the same, to 1 if they are actually similar. More often, however, the normalized Mantel coefficient, r, is calculated and this ranges from -1 to 1. In this study, the Mantel tests of correlation was performed between F_{ST}, R_{ST}, and D_{ST} using Arlequin software, with no. of permutations for significance:1000.

Later on, population tree based on pair-wise F_{ST} distances was constructed through Unweighed Pair-Group Method of Arithmetic Averages (UPGMA) clustering based on Nei's 1978 unbiased genetic distances using genetic data analysis (GDA) version 1.1 (Lewis and Zaykin, 2001) to illustrate the magnitude of differentiation among populations and subsequently describe the relationship between populations.

3.2.5.5 Inferring Population Structure

Model based Bayesian clustering of genotypic data was carried out to assign individuals into theoretical populations, such that differentiation (F_{ST}) was maximized between population groupings while within populations, conformation to HWE and LD was maintained. This analysis was performed in the software STRUCTURE version 2.3 (Pritchard et al., 2000). Individuals in the studied populations were clustered into *K* new populations regardless of their geographical locations, and probabilities of assignment to each cluster were assigned to each new population. Apart from demonstrating the presence of population structure, this software is also widely applied to identify distinct genetic populations, assigning individuals to populations, and identifying migrants and admixed individuals (Pritchard et al., 2000). The trialed K values ranged from 1 (no structured population) to 4 (each population is structured accordingly), with burning period of 25,000 and No. of MCMC Reps after burning: 75,000.

CHAPTER 1V

Results

4.1 DNA EXTRACTION

Average DNA concentration was recorded as 0.6 µg/µl with purity of $A^{260/280}$ 1.80. Diluted DNA samples with average concentration of 60 ng/nl were further utilized for PCR amplification.

4.2 MICROSTAELLITE PRIMERS AND PRELIMINARY POLYMORPHISM TESTING

Out of all 60 loci initially tested, PCR validation showed that 30 SSR loci were able to be amplified successfully, while the rest of the primers were discarded due to various reasons such as failed amplification, amplification with larger PCR product size, inconsistent amplification or significant stuttering. Out of these 30 successfully amplified loci, 22 potential polymorphic loci were observed from microsatellite banding profiles on the gel images Table 4.1 & Figure 4.1(a), (b), and (c)). This set consisted of eight di- and 14 tri-motif repeat regions.

After initial screening for polymorphism the PIC value was calculated for each primer locus, with PIC values ranging from 0.2858-0.89 (average PIC of 0.5447). A further explanation on primer validation is covered in section 4.3.3 of this chapter. Due to time constraints, only seven out of 22 polymorphic markers with PIC more than 0.5 were selected - see Table 4.4. Others were left out either due to low value of PIC, or because they showed significant stuttering as seen in electropherogram (Figure 4.3(i)).

Table 4.1: Polymorphism screening for 30 successfully amplified EST-SSR loci

Locus	Functional Annotation	Motif	Left_primer	Right_primer	Tm (°C)	PCR product size (bp)	Possible Polymorphic
MR1	Growth	$(GA)_5$	CTCCTTCATCCATCGTCGTT	TGGTGCTGATTCGTCTCTTG	63.3	300-309	Y
MR2	Linkage map	$(CA)_5$	ACACTGGGCAGGGAGTTATG	GGGATGGGTTTGTATGGTTG	63.3	150-167	Y
MR5	Immune	$(TC)_6$	TTCCCCAATGCTTCTTCATC	ACGCACCTCCTTGTATCCAC	55	150-177	Y
MR6	Immune/metabolism from KEGG (IM-K)	$(GT)_6$	CTGTTTTGAGGAACTAATGCAGAA	ACAGGTAGGTCAACAAACGAGTC	50	150-150	N
MR8	IM-K	$(TC)_6$	ACTTCTTGGCTTCAAGGGCT	TCCAGTCAAAAGAATTCGCA	55	150-200	Y
MR11	IM-K	$(CA)_6$	ATGAATGGAATGGGATTTGG	GTGCAACTGCACAATTCTCG	61.4	150-151	N
MR13	Linkage Map	$(TC)_7$	TGGACATCTTTGCATAGCCA	CACATCGGGGTTATTTTGGT	61.4	150-160	Y
MR14	IM-K	$(CT)_8$	CTCTGCTTCGTAAAATCGCC	GAACACTTTTGGCATGGGAG	61.4	150-165	Y
MR18	IM-K	$(TC)_{10}$	GTCCTCACAGCTGGTTCGAT	ATTTAACCCCTCGCCATTCT	65	250-286	Y
MR19	IM-K	$(TC)_9$	TGACCAATTCCCCACTGAAT	ACCTCTTGCAAGGCATTTTG	61.4	250-271	Y
MR23	IM-K	$(TC)_{23}$	CAAAGTGAGATTCATTACGGAGG	GCCTTCATTTGGCATTGAAA	55	150-152	N
MR31	Stress	$(TGG)_5$	CGCCCAAGATCTGATCCAC	ATCTCAACAGTAACATGGACTCAAAC	64.5	150-150	N
MR32	Molting	$(ACC)_5$	AATCGATCATCACCAGCCTC	TTGTTCCAACAGAACCCTCC	64.5	150-184	Y
MR35	Linkage Map	$(ATA)_5$	GCATGAAACCAGCTCATCCT	TCACGTCCATGGTTGATGAT	64.5	200-249	Y
MR36	Immune	$(GAG)_7$	GATGACAGGTGGGGACAGAG	CGTTTATTTTCCCAAGCCAA	64.5	150-180	Y
MR37	Immune	$(GCT)_6$	GTTACCAGGTGCCAGGTCC	GCTTCTTGACCGAGAACACC	64.5	150-155	Y
MR39	IM-K	$(ATT)_6$	GGTGGACTGAGACTGGCTTC	TGACAATGCAGATTCCCAAA	64.5	250-294	N
MR40	IM-K	$(TCC)_6$	TCAGAGATGTATTCCCCACAGA	TCCCCTGATCTTTAAATCCTCC	64.5	150-150	N
MR41	Immune	$(TCA)_7$	TCTCGTGTGACATAGGCAGC	TTGGAAGCAGAGAACAAGATTTC	63.3	150-199	Y
MR44	Immune	$(CTT)_6$	AGGCGAGGATGACTTTTCAA	TCCTGTTGCAGTACGGAGAA	63.3	150-169	Y
MR45	Immune	$(CAC)_6$	CCAATCATTACAATTGGCCC	ATGATGTGGATGCTGAGGTG	63.3	150-196	N
MR47	Stress	$(GGA)_7$	AAGTGGAGGTGGAACAGGTG	CTGAGACGGTCTTCTCCCTG	60.0	300-302	Y
MR51	IM-K	$(CTT)_7$	AGCTGTACACCTCTGGCTCG	CTACGAAACGCATGGTTGG	63.3	150-150	Y
MR52	IM-K	$(TCC)_7$	CCACTCCATTCACTCCCACT	CTACACCACCACAGACACCG	63.3	200-228	Y
MR53	IM-K	$(CAT)_7$	TGGATGTATAACTACGACGACAGC	CCGACTCTTCTTCGTCTTCC	61.4	150-159	Y

MR54	IM-K	$(CAG)_7$	AGCTCTCAACACAGCCGC	CAACATGGAAGACCTACGGG	65	150-150	Y
MR55	IM-K	$(CCT)_6$	GTCCGAGTGGCCTAGGGT	TTGGAATCCAGCTCTGAAGG	65	150-163	Y
MR56	IM-K	$(CCT)_9$	GAGACAAGCCGTGAAGGAAG	AGTGAATGGAGTGGGTGGAG	61.4	150-183	Y
MR57	IM-K	$(AGG)_{13}$	CGCTGTGCTGTACATGACCT	TGGTGTTAGGAACAATGTCG	65	150-150	N
MR58	IM-K	$(TCT)_{24}$	AGTCTCCTAAGACCCCGGAA	TATCGTCGCCATCACTAGCA	65	300-301	Y

Y representing polymorphic, while N representing monomorphic

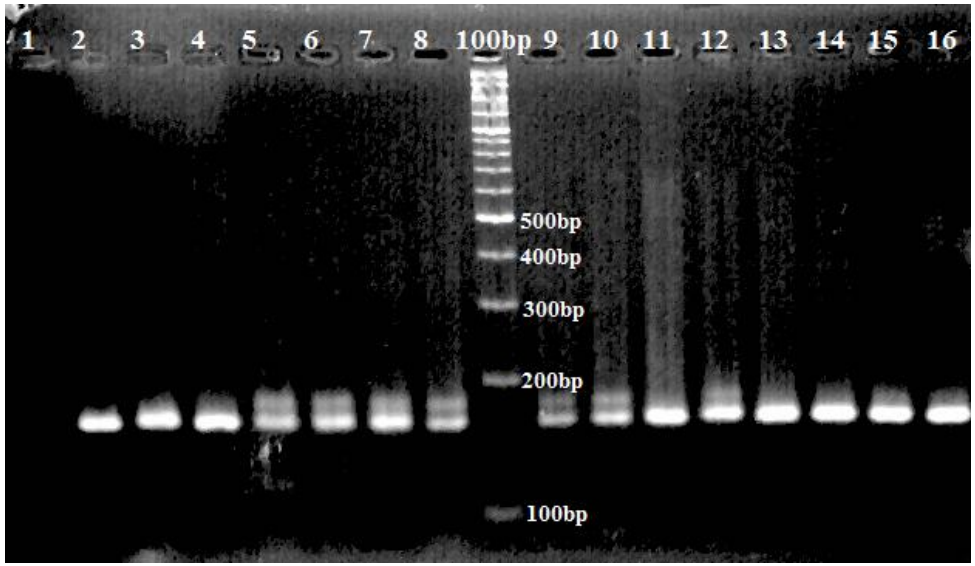


Figure 4.1 (a) Example gel image showing validation of EST MR13 on 16 individuals randomly selected from four wild populations. Polymorphism criteria was revealed with the presence of possibly heterozygous and homozygous individuals. In this image, 100bp DNA ladder was used as a marker.

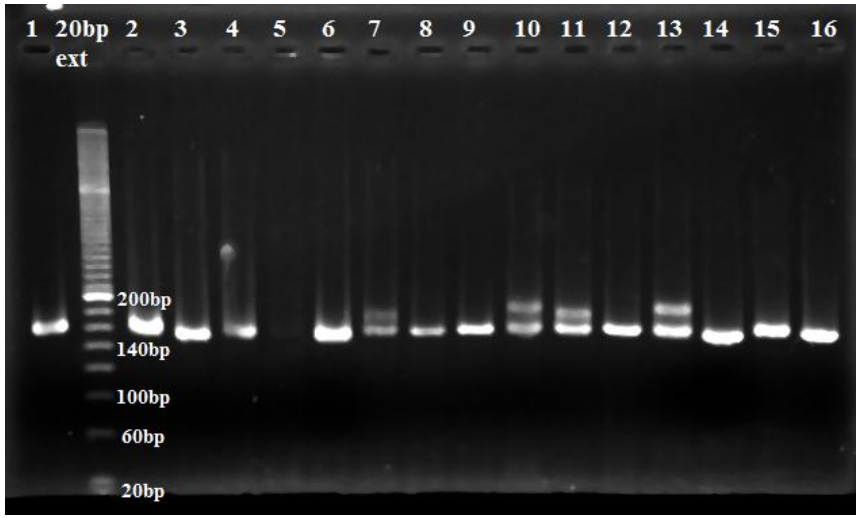


Figure 4.1 (b) Example gel image showing polymorphic banding profile in validation of EST MR 14 on16 individuals. 20bp extended range DNA ladder was used as a marker in this gel image, with increment of 20bp for each band.

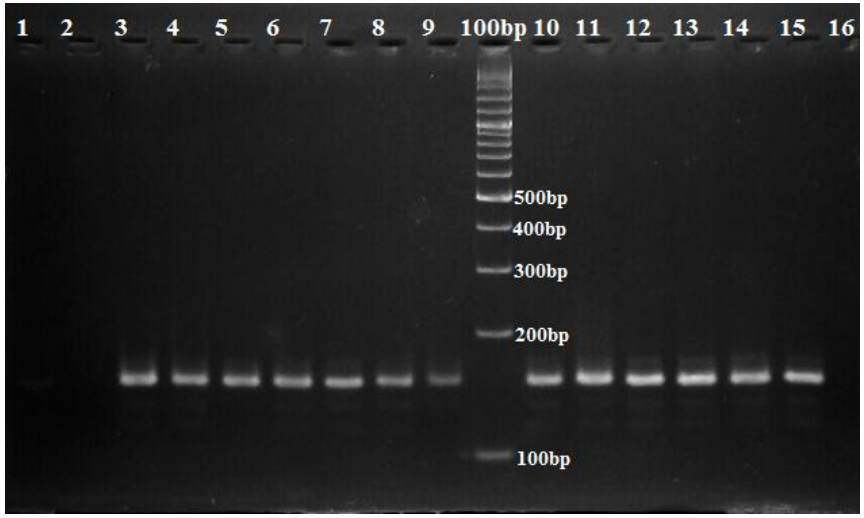


Figure 4.1 (c) Example gel image showing monomorphic pattern of amplified locus. E.g for EST MR40

EST-SSR primers were also characterized according to their annotated protein classes and functional classifications based on EST sequences (See Table 4.1; Appendix B). The distribution of annotated EST-SSR primers that were found to be polymorphic is illustrated in Figure 4.2 based on their functional classification. More than half of the polymorphic loci have been annotated with immune/metabolism function.

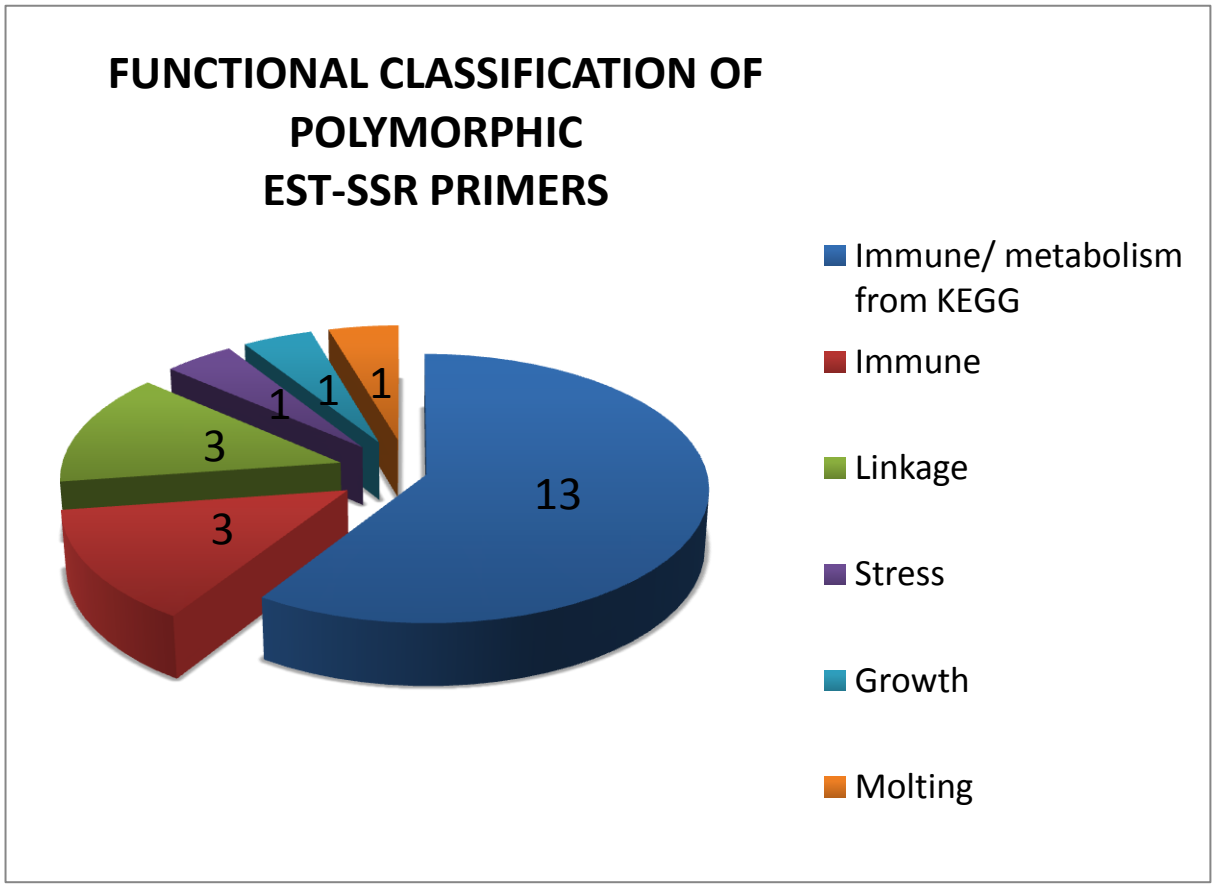


Figure 4.2 : Distribution of 22 polymorphic EST-SSR primers according to functional classification (See Appendix B for functional classification of all 60 primers priory designed)

Conformance of polymorphism pattern on those 22 primers aforementioned revealed the PIC value for each primer locus. Validation of polymorphism and estimation of PIC values for all 22 polymorphic EST-SSR primers were achieved using 32 individuals, and the calculated PIC values are displayed in Appendix F.

A further explanation on primer validation and calculation of PIC value is covered in section 4.3.3 of this chapter. Due to time constrain, only seven out of 22 polymorphic markers with PIC more than 0.5 were picked - see Table 4.4. Others were left out either due to low value of PIC, or showing significant stuttering as seen in electropherogram (Figure 4.3(i)).

4.3 DETERMINATION OF MICROSATELLITE ALLELE SIZES

Full panel of individuals screened by all seven loci was found to have a different multilocus genotype. However, three individuals show no amplification products when PCR was carried out using EST MR 13 (Appendix C). All alleles fell within the expected size range (within 50bps of the original sequence targeted by primer design). Homozygotes and heterozygotes were observed at all loci.

The following pictures (Figures 4.3, (i) & (ii)) display images of peaks observed in electropherogram as displayed by GeneMapper and Peak Scanner.

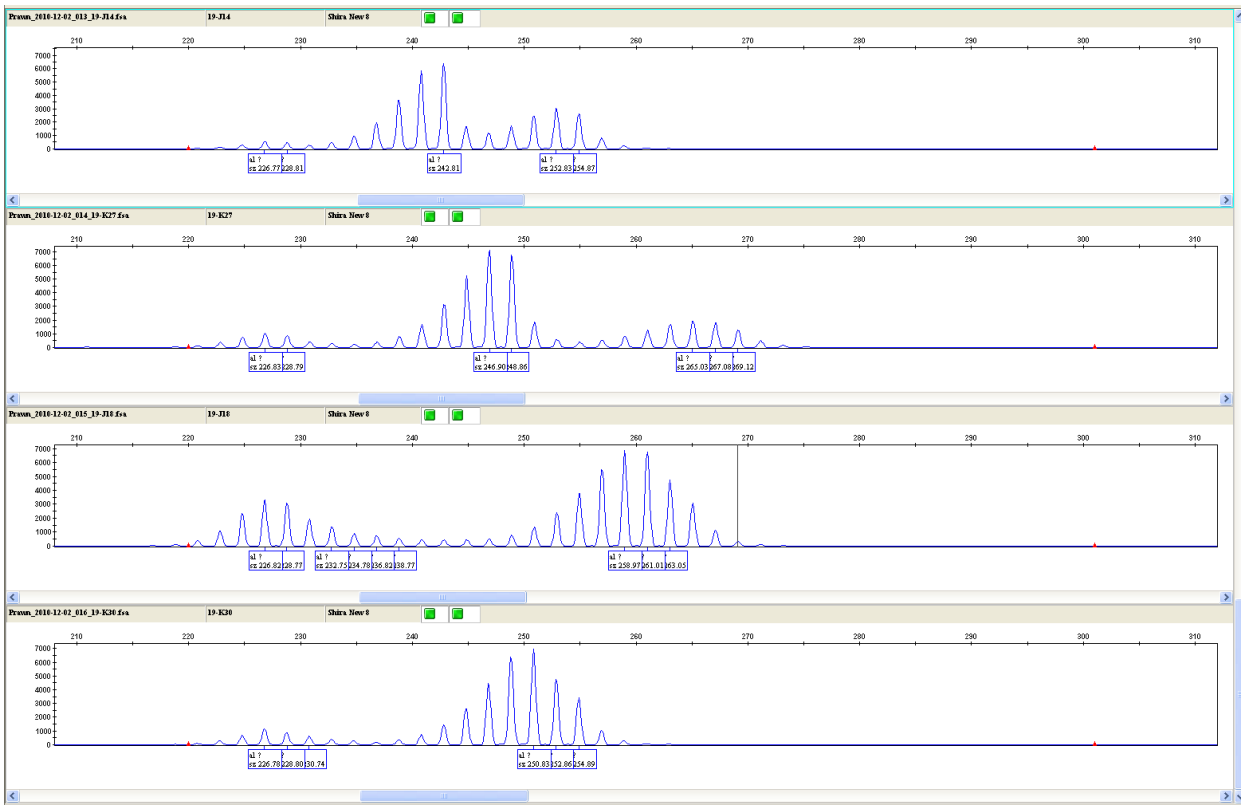


Figure 4.3 (i) : Electropherogram of EST MR19 alleles from four individuals (from top, J14 from Sg Timun; K27 from Kedah; J18 Sg Timun, K30 Kedah). All graphs illustrate the significant stuttering exhibited at the EST MR19 locus. This primer set was omitted from further evaluation. Significant stuttering hinder the correct determination of allele size, thus would have biased the final result of subsequent analysis had the locus been retained. **Picture is taken using GeneMapper v4.0 software.**

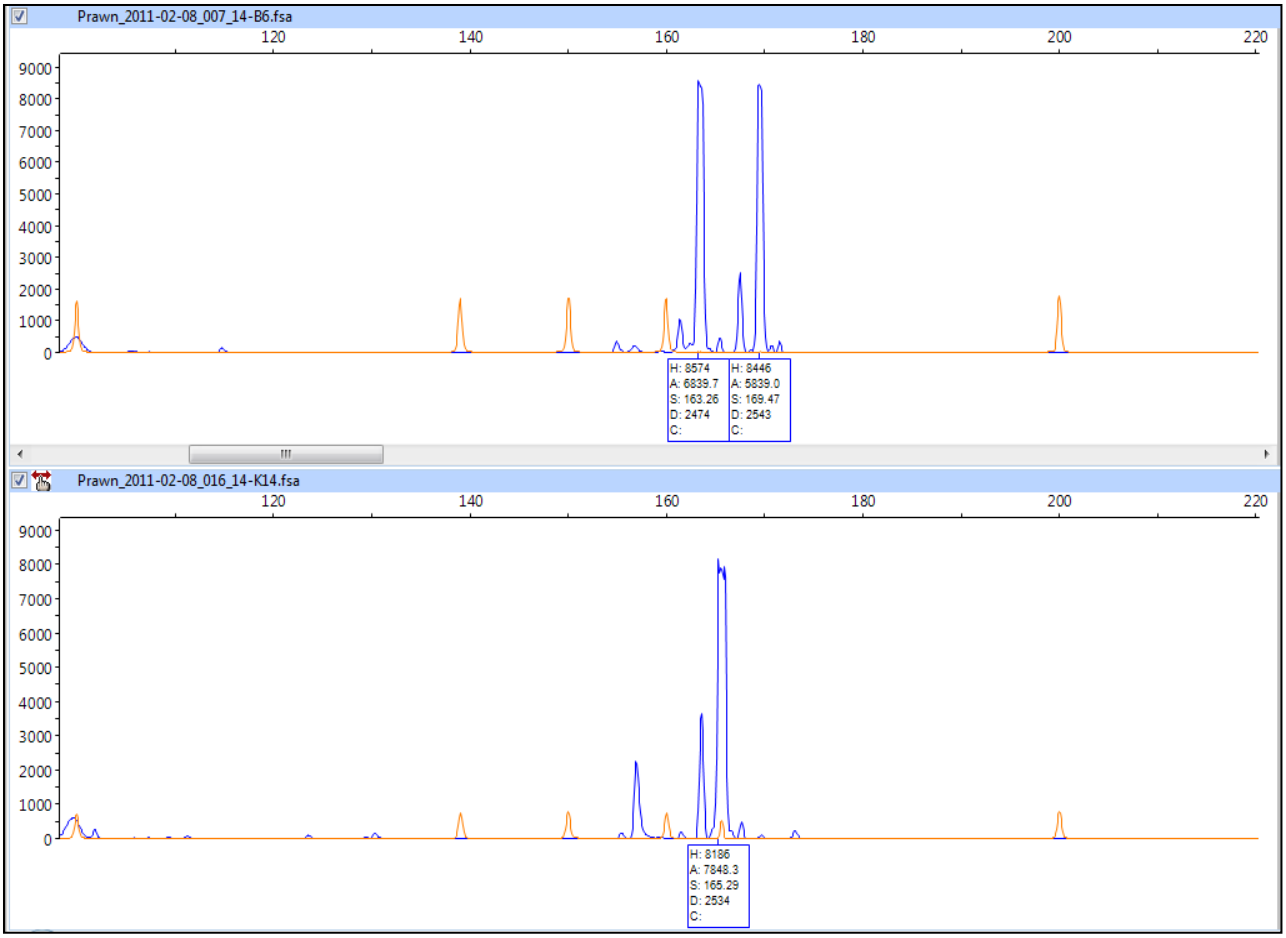


Figure 4.3 (ii) : Electropherogram of EST MR14 alleles for two individuals (upper from Sg Timun, and lower from Kedah) . EST MR14 is a di-repeat locus. The upper graph shows two prominent blue peaks (alleles) of *M. rosenbergii* individual B6 which were sized as 163bp and 169bp respectively, indicating that the individual is heterozygous at this locus. In the bottom graph only one prominent peak can be observed, indicating that individual K14 is homozygous at the EST MR14 locus, with both copies of alleles 165bp in length. **Picture is taken using Peak Scanner v1.0 software.**

4.4 STATISTICAL DATA ANALYSIS

4.4.1 ERROR CHECKING

All individuals genotypes scored in GeneScan were screened for any anomalies. Result obtained from Micro-checker version 2.2.3 analysis (Van Oosterhout et al., 2004) demonstrated no evidence of null allele occurrence, no large allele drop outs, or presence of stutter bands. This indicated that the data was suitable for subsequent genetic analysis.

4.4.2 HARDY-WEINBERG EQUILIBRIUM AND LINKAGE DISEQUILIBRIUM TESTS

The exact probability test for deviations from Hardy-Weinberg Equilibrium, HWE, was performed on all loci for each population. Among the seven loci that were analyzed, EST MR8 showed significant deviations ($p < 0.05$) from HWE for three out of four sampling sites. This implies that one or a combination of factors contribute to the violation of HWE, such as selection, might be significantly impacting on the proportion of genotypes that occurred in that particular locus. Since EST MR8 was found to have not conformed to HWE, data collected for this locus may potentially bias subsequent analyses of population differentiation that assume conformation to HWE. After FDR correction, however, most p-values were found to be not statistically significant, indicating that on a population by population basis all populations are conforming to HWE, as shown in the Table 4.2.

To avoid possible bias in subsequent results, due to the deviation from HWE observed at EST MR8, population differentiation analyses were performed including and excluding this locus.

Table 4.2: Probability values of HWE for each locus per each studied populations

Pop	EST MR 5	EST MR 8	EST MR 13	EST MR 14	EST MR 37	EST MR 41	EST MR 51	ALL (by pop)
Sg Timun	0.0497*	0.0062**	1.000	0.0267*	0.1292	0.4775	0.1605	χ^2: 32.6667 Df : 14 P : 0.0032
Kedah	0.1438	0.0012**	0.4504	0.0432*	1.000	0.1381	0.0873	χ^2 : 34.1115 Df : 14 P : 0.0020
Sarawak	0.5612	0.1024	0.3306	0.0208*	0.3099	0.9184	0.4484	χ^2 : 19.7861 Df : 14 P : 0.1370
Terengganu	0.0000**	0.0000**	1.0000	0.5936	0.4539	0.7023	0.2071	χ^2 : Infinity Df : 14 P : High sign
ALL (by loci) (Fisher's Method)	χ^2 : Infinity Df : 8 P : High sig.	χ^2 : Infinity Df : 8 P : High sig.	χ^2 : 3.8087 Df : 8 P : 0.8740	χ^2 : 22.3180 Df: 8 P : 0.0044	χ^2 : 8.0154 Df : 8 P : 0.4320	χ^2 : 6.3143 Df : 8 P : 0.6121	χ^2 : 13.6368 Df: 8 P : 0.0917	χ^2 : Infinity Df : 52 P : High sign

(*p value< 0.05; ** p value significant after FDR correction where α=0.05)

Exact tests for genotypic linkage disequilibrium for all seven different loci did not detect any significant evidence of linkage disequilibrium in any loci ($p > 0.05$), suggesting that no physical linkage is present for any pair of loci, and that no sample represented a mix of divergent non-randomly mating individuals (See Appendix D).

4.4.3 CHARACTERIZATION OF EST-SSR LOCI ISOLATED FROM *M. rosenbergii*

4.4.3.1 Polymorphic Information Content (PIC) and Genetic Variability

The PIC values, the number of alleles detected per locus, and the allelic size range are shown in Table 4.3 below. Those selected seven (Table 4.3) that showed initial high value of PIC, demonstrated considerable variation when tested on a full panel of 120 individuals from the four locations.

Table 4.3: Polymorphism assessment on microsatellites loci using four wild populations

LOCUS	T_A (^OC)	MOTIF	PIC*	R	A
Di-					
EST MR5	55	$(TC)_6$	0.6094 (0.398)	161-179	6
EST MR8	55	$(TC)_6$	0.516 (0.231)	182-202	4
EST MR13	61.4	$(TC)_7$	0.6351 (0.450)	147-161	5
EST MR14	61.4	$(CT)_7$	0.739 (0.649)	153-167	8
Tri-					
EST MR37	64.5	$(GCT)_6$	0.631 (0.429)	147-162	6
EST MR41	63.3	$(TCA)_6$	0.5798 (0.501)	176-197	8
EST MR51	63.3	$(CTT)_7$	0.6355 (0.431)	133-169	10
		TOTAL			47

Annealing temperature (T_A)Number of alleles per locus (A), Allelic size range in bp (R), PIC (Polymorphic Informative Content)
*PIC values calculated based on polymorphism screening on 32 individuals, values in parentheses re-calculated using full panel of 120 samples

All seven primers showed initial high level of polymorphism in the 32 examined samples as the PIC value for the SSR markers was calculated in the range of 0.516 (EST MR8) to

0.739 (EST MR14) with an average of 0.6208 across all loci. Yet, when full data set was screened (120 individuals), PIC values dropped presumably influenced by the sampling size. All di-repeat loci scored values below 0.5 except for EST MR14 (0.649). In addition, tri-repeat loci also demonstrated slight fall in PIC value calculation. For instance, though EST MR51 harbored 10 alleles, its PIC estimate (0.431) was somewhat lower than EST MR41 (0.501).

Among the seven loci assessed, more alleles were detected in tri-nucleotide repeat (mean = 8 per locus) compared to in di-repeat regions, where an average of 5.75 was detected for locus per site. The highest number of alleles per locus was recorded for EST MR51 (10 alleles), whilst the lowest number of 4 alleles scored by EST MR8.

4.4.3.2 Heterozygosity

Based on Table 4.4, calculation of expected heterozygosity and observed heterozygosity were illustrated for all loci, within and among all populations.

Generally, by looking at H_E and H_O values among loci, EST MR14 demonstrated the highest values for both parameter regardless of sampling sites in comparison to other loci, with value ranging from 0.6288 to 0.7009 for H_E, whilst H_O obtained ranges from 0.5333 to 0.8333. Polymorphism assessments on all populations achieved using this primer showed that H_O scored were generally higher than H_E.

Table 4.4: Summary of observed and expected heterozygosity (H_O and H_E) for each locus across four population

Expected heterozygosity (H_E), Observed heterozygosity (H_O), p-value retrieved from Chi-square test conforming HWE (P)

Type of repeat/ locus	SG TIMUN			KEDAH			SARAWAK			TERENGGANU		
	H_E	H_O	P	H_E	H_O	P	H_E	H_O	P	H_E	H_O	P
Di- repeat												
EST MR5	0.5435	0.7667	0.0497*	0.3814	0.5000	0.1438	0.4028	0.5000	0.5612	0.4699	0.6333	0.000**
EST MR8	0.2960	0.2333	0.0062**	0.1576	0.0333	0.0012**	0.1881	0.1333	0.1024	0.2470	0.1083	0.000**
EST MR13	0.4169	0.4333	1.0000	0.5185	0.5862	0.4504	0.5711	0.4483	0.3306	0.5069	0.4615	1.0000
EST MR14	0.6288	0.8333	0.0267*	0.6599	0.8333	0.0432*	0.6955	0.5333	0.0208*	0.7009	0.7417	0.5936
Tri- repeat												
EST MR37	0.4362	0.4333	0.1292	0.2706	0.3000	1.0000	0.3972	0.3333	0.3099	0.4657	0.4250	0.4539
EST MR41	0.5158	0.6667	0.4775	0.6435	0.6667	0.1381	0.4768	0.5333	0.9184	0.5551	0.6250	0.7023
EST MR51	0.3944	0.4000	0.1605	0.5153	0.4667	0.0873	0.3836	0.4000	0.4484	0.4552	0.4583	0.2071

Conversely, the lowest values recorded were shown to be associated with EST MR8, whereby all H_E and H_O scores were very low (less than 0.3). This pattern likely contributed to the significant p-value calculated from statistical test, thus resulting in EST MR8 to depart from HWE.

4.4.4 EST-SSR LOCI FOR CHARACTERIZING POPULATIONS GENETICS OF FOUR WILD LOCATIONS

4.4.4.1 Genetic Diversity

The level of genetic diversity estimated based on seven microsatellite loci are summarized in Table 4.5. All assayed primers detected polymorphism in each of the studied populations, with a minimum of two alleles and a maximum of eight alleles present at each locus in each population. The number of alleles per locus (A_t) observed for Sg Timun ranged from 3 to 6, from 2 to 7 for Kedah, 3 to 8 for Sarawak, and 3 to 7 for Terengganu.

The value of A_t very likely contributed to the higher value of allelic richness (R_s) for Sarawak population in comparisons with others. Kedah scored the lowest for this parameter at 4.542 whilst Sg Timun and Terengganu were in between of those two with 4.809 and 4.824 respectively. Generally, however, average allelic richness was very similar among all sample sites (ranging from 4.542 - 4.961).

On the other hand, the mean effective number of alleles (A_e) was observed to be highest in Terengganu with 2.163 (Std. dev = 0.539), whilst the lowest was achieved by Sg Timun (mean = 1.8982, Std. dev = 0.402).

Table 4.5 : Summary of genetic diversity measures based on seven microsatellite loci in four populations of Sg Timun (Negeri Sembilan), Kedah, Sarawak, and Terengganu: total number of alleles (A_t), effective number of alleles per locus (A_e), allelic richness (R_s) and effective sample size (N)

Type of repeat/ locus	Sg Timun				Kedah				Sarawak				Terengganu			
	A_t	A_e	R_s	N	A_t	A_e	R_s	N	A_t	A_e	R_s	N	A_t	A_e	R_s	N
Di-repeats																
EST MR5	5	2.148	4.933	30	2	1.600	2.000	30	4	1.6559	3.966	30	4	2.1004	3.966	30
EST MR8	3	1.4107	3.000	30	3	1.1834	2.967	30	3	1.2270	3.000	30	3	1.4839	2.967	30
EST MR13	5	1.6949	4.933	30	3	2.0388	3.000	29	5	2.2791	5.000	29	3	1.5502	3.000	29
EST MR14	6	2.6201	5.933	30	7	2.8481	6.933	30	6	3.1634	5.966	30	7	2.7778	6.966	30
Tri-repeats																
EST MR37	4	1.7510	3.999	30	4	1.3626	3.966	30	4	1.6408	3.999	30	5	2.8662	4.967	30
EST MR41	5	2.0293	4.966	30	6	2.7231	5.998	30	5	1.8828	4.933	30	6	2.3018	5.933	30
EST MR51	6	1.6334	5.899	30	7	2.0270	6.993	30	8	1.6057	7.866	30	6	2.0619	5.966	30
TOTAL	**34**	**-**	**-**	**120**	**32**	**-**	**-**	**119**	**35**	**-**	**-**	**119**	**34**	**-**	**-**	**119**
Average (St dev.)	**4.8571 (1.069)**	**1.8982 (0.402)**	**4.809**	**30**	**4.5714 (2.070)**	**1.969 (0.642)**	**4.542**	**29.86**	**5.000 (1.633)**	**1.922 (0.633)**	**4.961**	**29.86**	**4.8571 (1.574)**	**2.163 (0.539)**	**4.824**	**29.86**

Across all loci at all sites a total of 47 alleles were observed (Table 4.3). However, the maximum number observed in any one site was 35 alleles (found in Sarawak population, see Table 4.5), followed by Sg Timun and Terengganu (34), and the lowest recorded in Kedah population (32). Ten private alleles were observed in low frequencies (0.0167-0.0667, Appendix E) across all populations, however generally the most common alleles at each locus were observed at all sites..

Further reference on genotypic and allelic frequencies, refer to Appendix E.

4.4.4.2 Heterozygosity and Inbreeding

Based on Table 4.6, calculation of expected heterozygosity and observed heterozygosity were illustrated for all loci, within and among all populations. H_E estimates ranged from 0.5381 (Sg Timun) to 0.4450 (Sarawak), whereas the highest recorded for H_O was 0.4933 (Terengganu) and the lowest was 0.4117 (Sarawak). This implies that there did not appear to be large differences in genetic variability among the four populations.

The studied populations of *M. rosenbergii* can be ranked based on the H_O (mean value) as follows: Terengganu> Kedah > Sg Timun > Sarawak. The H_O value displayed in the table was recorded to be higher compared to H_E for Kedah and Terengganu. Meanwhile, the value of H_O for Sg Timun and Sarawak are showed to be lower than H_E.

As displayed in the same Table of 4.6, all populations except Sarawak (0.06958) exhibited negative F_{IS} values (Sg Timun: -0.1689; Kedah: -0.08321; and Terengganu: -0.04313). The negative values of F_{IS} are associated with the excess of observed heterozygosity over the expected heterozygosity, and were treated as zero, indicating no inbreeding. Based on the table, F_{IS} value indicating high heterozygosity was only observed in Sarawak, while for other populations, the study probably suggests no inbreeding occurs in the studied population as denoted by the negative value of the coefficient.

Table 4.6 Summary of observed and expected heterozygosity (H_O and H_E) for each locus across four population

POPULATION (MARKERS)	H_E	H_O	F_{IS}
Sg Timun			
EST MR5	0.5435	0.7667	
EST MR8	0.2960	0.2333	-0.16890
EST MR13	0.4169	0.4333	
EST MR14	0.6288	0.8333	
EST MR37	0.4362	0.4333	
EST MR41	0.5158	0.6667	
EST MR51	0.3944	0.4000	
MEAN	0.5381	0.4617	
Kedah			
EST MR5	0.3814	0.5000	-0.08321
EST MR8	0.1576	0.0333	
EST MR13	0.5185	0.5862	
EST MR14	0.6599	0.8333	
EST MR37	0.2706	0.3000	
EST MR41	0.6435	0.6667	
EST MR51	0.5153	0.4667	
MEAN	0.4495	0.4837	
Sarawak			
EST MR5	0.4028	0.5000	0.06958
EST MR8	0.1881	0.1333	
EST MR13	0.5711	0.4483	
EST MR14	0.6955	0.5333	
EST MR37	0.3972	0.3333	
EST MR41	0.4768	0.5333	
EST MR51	0.3836	0.4000	
MEAN	0.4450	0.4117	
Terengganu			
EST MR5	0.4699	0.6333**	-0.04313
EST MR8	0.2470	0.1083	
EST MR13	0.5069	0.4615	
EST MR14	0.7009	0.7417	
EST MR37	0.4657	0.4250	
EST MR41	0.5551	0.6250	
EST MR51	0.4552	0.4583	
MEAN	0.4858	0.4933	

FIS, Wright Inbreeding Coefficient

4.4.4.3 Genetic Differentiation

Pairwise comparisons for among all populations (four locations, six pairs) were performed based on three parameters of fixation indices, namely F_{ST}, R_{ST}, and D. The result of the pairwise F_{ST} test is given in Table 4.7. The F_{ST} values were generally low, ranged from 0.00888 (Sg Timun-Sarawak) to 0.10644 (Kedah-Terengganu). Significant F_{ST} values were only observed between Sg Timun and Kedah, and for comparisons between Terengganu and all other sampling sites ($p < 0.005$), indicating Terengganu was genetically differentiated from all other sample sites. Though the differences were shown to be statistically significant, the low value of F_{ST}s indicate that there is only a very low degree of differentiation present among some of the studied populations.

Table 4.7 Pairwise F_{ST} values between four populations

POPULATION	SG TIMUN	KEDAH	SARAWAK	TERENGGANU
SG TIMUN	-	0.03406	0.00888	0.03142
		0.03506	0.00975	0.03280
KEDAH	0.0000**	-	0.01121	0.10644
	0.0000**		0.01227	0.11137
SARAWAK	0.07207	0.07207	-	0.06797
	0.06306	0.02703**		0.07045
TERENGGANU	0.0000**	0.0000**	0.0000**	-
	0.0000**	0.0000**	0.0000**	

Above dashed line is FST value, and below is p value; *p value< 0.05;
** p value significant after FDR correction where α=0.05;
Number in green color representing value of test performed with omission of MR 8

When pairwise F_{ST} values were calculated without the inclusion of EST MR8, the Sarawak and Kedah appear to be statistically differentiated, suggesting that EST MR8 was influencing results. Again, low values of F_{ST}s indicated low level of genetic differentiation among studied populations, though the differences were again found out to be significant. Nevertheless, overall results from the calculation excluding EST MR8 data were similar to F_{ST} estimation based on all seven loci.

Next, genetic differentiation among all populations was also assessed through estimation of pairwise R_{ST} values (four locations, six pairs). Pairwise tests were again performed for two data sets; with and without inclusion of EST MR8. Table 4.8 shows the R_{ST} values based on pairwise comparison tests. In congruence with the F_{ST} results, low genetic differentiation was exhibited by combinations of these four populations as denoted by generally low R_{ST} value (less than 0.1). Again as earlier observed, each pair-wise estimate involving Terengganu was observed to be statistically significant ($p < 0.05$), thus supporting the results of the F_{ST} analysis.

Table 4.8 Pairwise R_{ST} values between four populations

POPULATION	SG TIMUN	KEDAH	SARAWAK	TERENGGANU
SG TIMUN	-	0.00702	0.0000	0.02689
		0.00634	0.0000	0.02798
KEDAH	0.22523	-	0.0000	0.06670
	0.21622		0.00118	0.07252
SARAWAK	0.61261	0.39640	-	0.04609
	0.62162	0.38739		0.04280
TERENGGANU	0.0000**	0.0000**	0.0000**	-
	0.0000**	0.0000**	0.0000**	

Above dashed line is RST value, and below is p value; *p value< 0.05;
** p value significant after FDR correction where α=0.05;
Number in green color representing value of test performed with omission of MR 8

When the R_{ST} test was performed with the omission of EST MR8, results showed the same pattern as the R_{ST} pairwise test that included data from all seven loci. R_{ST} and F_{ST} tests both generally indicate that Sg Timun, Kedah, and Sarawak are not genetically different, whilst Terengganu was slightly differentiated from all three.

In addition to F_{ST} and R_{ST}, D_{est} was also applied to all four populations to infer the level of their genetic relatedness. D_{est} fixation index for pairwise comparison test is presented in

Table 4.8 below. All values were recorded to be less than 0.1, ranging from 0.00046 (Sg Timun-Sarawak combination) to 0.04322 (Kedah-Terengganu combination). Generally, Terengganu was more different from other sites (D_{est} values ranging from 0.0107 – 0.04322) than the others were to each other, except for the differentiation observed between Kedah and Sg Timun (0.0129) which reflects the significant result observed in the F_{ST} analysis. Overall, however, the small values of D_{est}, show that little if any differentiation exists among all populations was limited.

Table 4.9 Pairwise D_{est} values between four populations

POPULATION	SG TIMUN	KEDAH	SARAWAK	TERENGGANU
SG TIMUN	--	0.012901668	0.000464274	0.010691274
KEDAH	--	--	0.002555735	0.043222995
SARAWAK	--	--	--	0.041001325
TERENGGANU	--	--	--	--

To quantify the similarity between differentiation indices the correlation between the measures was assessed by Mantel tests, see Table 4.9.

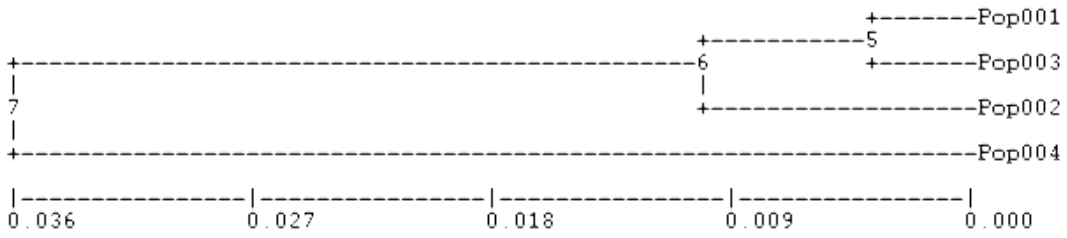
Table 4.10 Values of correlation coefficient between matrices based on Mantel test

Differentiation parameters	Correlation coefficient value
$\mathbf{F_{ST}}$ **-** $\mathbf{R_{ST}}$	0.965995 (0.0490*)
$\mathbf{R_{ST}}$ **-** $\boldsymbol{D_{ST}}$	0.931942 (0.0440*)
$\mathbf{F_{ST}}$ **-** $\boldsymbol{D_{ST}}$	0.956277 (0.0420*)

(Number in brackets represents the p-value, *p-value < 0.05)

As revealed by Mantel tests, the relationships between all three differentiation parameters were strongly correlated. This can be seen in the high correlation coefficient values (more than 90% correlation for each pairwise test) which were all statistically significant ($p<0.05$). Therefore, all pairwise comparisons that measured population differentiation

are in strong agreement with each other, suggesting that conclusions about population differentiation that are drawn based on these fixation indices, D_{est}, F_{ST}, and R_{ST} are robust.



*Pop001: Sg Timun, Pop002: Kedah, Pop003: Sarawak, Pop004: Terengganu

Figure 4.4: UPGMA Phenogram

Next, population tree phenogram in Figure 4.4 above was constructed to further illustrate the relationships among all four populations. Result obtained in the GDA provides further evidence, supporting pairwise comparison analysis. Corresponding to F_{ST} pairwise differentiation based on 7 loci, Terengganu again was observed to be least similar compared to the other three populations. Sg Timun and Sarawak was found to be less genetically differentiated, in contrast to Sg Timun-Kedah pairwise comparison. This is correlated to the earlier observation in which calculated pairwise F_{ST} values for Sg Timun-Kedah was found to be statistically significant, though the magnitude of differentiation is very low. Generally, pairwise analyses revealed that there was little to no differentiation between all sampled populations.

4.4.4.4 Population structure

Bayesian clustering of genotypes revealed that the optimum number of populations represented in the four sample sites was one, indicating that there is no sub-division between all four wild populations. The highest log probability value (Ln P(D)) (closest to zero) was recovered when k=1 (all individuals assigned to the same population). The estimated Ln P(D)) value was less when k=2, and also for all other higher values of k indicating that all studied *M. rosenbergii* populations are not genetically structured. All populations are basically the same in terms of their genetic variation and can be treated as a single panmictic population.

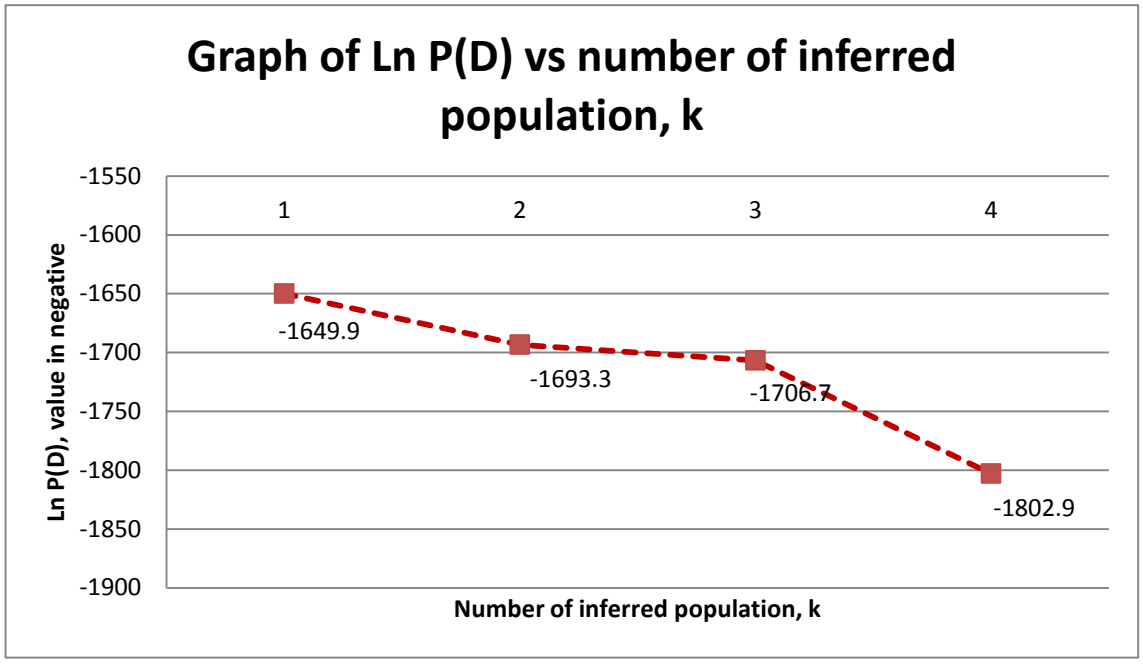


Figure 4.5 : Graph showing value of Ln P(D) (likelihood probability of X| K) estimated for all number of inferred populations using Structure Ver 2.2.

CHAPTER V

Discussion

5.1 MICROSATELLITE LOCI AND PRELIMINARY POLYMORPHISM TESTING

The success ratio of EST-SSR amplification was 50%, with 30 EST-derived primers able to be amplified out of the initial 60 that were tested. Those that failed either were unable to amplify product at all, or showed unspecific amplification. Failure of amplification can be attributed to a variety of reasons, such as location of primer spanning across introns and/or mutations and indels (insertions or deletions) at the primer annealing sites. These issues are especially important when considering EST derived loci, as there are potentially fundamental problems when applying EST loci to genomic DNA regions.

The primers used in this study were designed to amplify microsatellite regions that were identified in sequences that had been assembled from large scale short read length transcriptome sequencing. The transcriptome is the transcribed portion of the genome, and during conversion from genomic DNA to transcribed RNA many alterations occur to the original genomic DNA sequence, for example introns may be excised through splicing mechanisms and exons may be combined into different orders. This means that there is the potential that primers designed based on EST sequence may not amplify genomic DNA, for example if the priming sites are actually located a very long way apart on the chromosome. As another example, priming sites may span the ends of two exons (an intron-exon junction). This would mean that primers would fail to bind to genomic DNA, as their priming site would be bisected by an intron. Large fluctuation of PCR

product size can be caused by insertion of an intron in the region close to or within the microsatellite loci, and also contributing to the unsuccessful amplification (Ellis & Burke, 2007). Alternatively, it is possible that amplification of multiple bands or unspecific amplification can arise if multiple genomic regions are transcribed into homologous EST sequences (Nicot et al., 2004; Aibin-Zhan et al., 2009).

Furthermore, generation of large scale transciptome data set requires accurate assembly of millions of short sequence reads. There is the potential that mis-assembly of raw sequencing data may lead to inaccurate sequence itself, and hence that primers may be synthesized based on incorrect consensus sequences, ultimately resulting in failed amplification.

Clustering step of large amounts of EST data usually involve prior to the assembly process of consensus sequence, ultimately aiming to create a non-redundant representation of a possible putative gene for an organisms transcriptome. During clustering, high quality ESTs are grouped into “clusters” based on sequence similarity. When it comes to cluster large amounts of EST data, usually some large clusters (>500 sequences) are created. Those can lead to iterative contig builds, consumation of lots of computing time and improbable exon alignments, which is unfavourable(Rensing et al., 2003).

In addition, these clusters sometimes contain transcripts for more than one gene, which is not desired as it can lead to mis-assembly. Such large clusters come into existence due to factors such as: (1) large numbers of identical ESTs / high transcript levels; (2) large gene families with highly similar members; (3) false clustering due to contaminant in raw sequence data, such as a) unremoved vector or rRNA sequences, b) undetected cloning artifacts or c) repetitive elements in untranslated regions (UTRs) (Rensing et al., 2003). The poor assembly from raw data possibly causes the manufacturing of incorrect EST sequences contigs, or making the ESTs completely

unrecognizable. In the end, the primer generated from these EST data sets may no longer able to bind the template, thus fail to amplify the original target regions in genome.

In this study, the number of triplets SSRs found in successful amplified products is higher compared to di-nucleotide repeats. Of the 30 di-repeat loci tested, about 40% of the di-repeat motifs were successfully amplified from genomic DNA, while 60% of tri-repeats were able to be amplified. The prevalence of EST-SSR motif type obtained in this current study is incongruent with other studies on aquaculture species which showed the abundance of di-motif type in EST-SSR (e.g: Chinese shrimp-Wang et al., 2005; Pacific oyster-Yu & Li, 2008; Pacific abalone-Li et al., 2010; catfish-Serapion et al., 2004). However, this study showed similar patterns observed in plants, where tri-nucleotide repeats are more common (Varshney et al., 2005, Morgante et al., 2002 as cited in Li et al., 2002). The high distribution of triplet microsatellites in ESTs is perhaps attributable to suppression that limits expansion of nontrimeric SSRs in coding regions, possibly caused by frameshift mutations (Metzgar et al., 2000).

The polymorphism of EST-SSR loci was firstly assayed by calculating the PIC values. The PIC values reported herein can serve as a guide to selecting loci that are most likely to be informative (Chapman et al., 2009, Ye et al., 2010) for *M. rosenbergii*. It is also a helpful guideline to scale genetic variance of *M .rosenbergii* populations. When PIC of a locus is higher than 0.5, the locus is a high polymorphic locus; 0.5> PIC> 0.25, the locus is a mediate polymorphic locus; and below than 0.25, the locus is a low polymorphic locus (Vaiman et al., 1994 as cited in Wang et al., 2009).

In the present study, the PIC values of seven microsatellites loci were found to be higher than 0.5, indicating that these microsatellites are potentially informative and useful. However, estimation of PIC values depends on total number of alleles detected and their allelic frequencies in a population, and so the calculation of PIC is influenced by sample size. In the current study the PIC values declined when re-estimation was performed based on a larger sample, indicating initial sampling bias. Sampling bias is the error resulting from taking a non-representative sample from a larger group, and results based on biased samples may produce misleading conclusions. Comparison of PIC values for di versus tri-loci demonstrates that the value is not totally different and shows an average level of polymorphism.

As shown in the result (Figure 4.3), image taken from GeneMapper displays a characteristic feature namely "stutter bands"- that is, minor products that differ in size from the main product by multiples of the length of the repeat unit (Hauge et al., 1993 & Murray et al. 1993 as cited in Ellegren, 2004). This results from replication slippage of Taq polymerase that occurs during PCR amplification of microsatellite sequences in vitro (Ellegren 2004). The stutter peaks are observed as multiple artifact peaks preceding the true allele peak, and this could lead to incorrect scoring and hinder the genotyping of individuals that can complicate data interpretation. The presence of stutter artifacts are more commonly found in di-nucleotide loci compared to other SSR type motifs (Ellegren, 2004; Bakker et al. 2005). Despite the fact that some stuttering was observed in the present study, as the stutter bands were able to be distinguished from the real alleles their presence did not compromise the data collection process.

5.2 CONFORMITY TO NEUTRAL EXPECTATIONS

Linkage Disequilibrium tests revealed that none of the loci showed significant linkage disequilibrium for all pairs of loci. This indicates that there is probably no physical linkage between any of the loci used in this study, and also that there appears to be no pleiotropic effects resulting in linkage disequilibrium due to association between multilocus microsatellite genotypes and survival. Furthermore, it means that the individuals that were collected at each site are not likely to constitute a mix of two or more breeding populations with very different gene frequencies, as pooling divergent populations in a single sample should, in theory, produce a pattern of linkage disequilibrium.

However, significant departure from HWE was observed in locus EST MR8 suggesting that this locus is not assorting randomly. One possible explanation for violation of HWE is the presence of null alleles. Null alleles can occur as a result of priming site mutations or large or small allele dropout (DeWoody et al., 2006), thus leading to the failure of PCR amplification of one or both microsatellite alleles, resulting in a lack of visible amplicons for one or both alleles in a diallelic genotype (Rodriguez et al., 2009). If null alleles are caused by a mutation in the priming site then individuals that are heterozygous and possess one null allele and one non-null will appear to be homozygous for the visible allele, whereas there are no visible alleles at all in the case of null-null homozygotes. Likewise, long or short allele dropout will result in scoring only one allele per genotype. Hence, falsely recording homozygote genotypes where individuals are true heterozygotes. The presence of null alleles in a population will bias allele frequencies and inflate the number of homozygous genotypes, thus reducing the observed heterozygosity

(DeWoody et al., 2006). Nevertheless, the presence of null alleles is unlikely in this study, as output from Microchecker found no evidence for null alleles.

Another possible reason that causes the departure from HWE is the Wahlund effect. The Wahlund effect occurs when an assumed large randomly mating population actually comprises a number of subdivided populations that differ in allelic frequencies. If pooled samples are taken from subpopulations that have diverged sufficiently in allele frequencies, an overall deficiency of heterozygotes can be observed, thus in cases of Wahlund effect there will be an apparent excess of homozygote genotypes. F-statistics can be applied to measure the reduction in heterozygosity. However, the likelihood of a Wahlund effect being present in the current data set is unlikely, as discussed in the next section of this chapter.

Significant deviations from HWE could also result from selection pressure that may be occurring at the molecular or the phenotypic level. For example, at the molecular level EST MR8 is a di-nucleotide repeat SSR and possibly mutations in repeat number at this locus may interrupt the reading of downstream coding regions (frame-shift mutation), therefore having a deleterious effect on the organism, leading to selection pressure to maintain repeat numbers that do not interrupt the reading frame (Li et al., 2004). Alternatively, this locus has been characterized as immune related (Table 4.1), and the presence of specific allelic variants may have a positive effect on the phenotype of individuals, rendering them with higher immunity, and therefore the locus may be under selection due to its proximity to a gene that is potentially important in the prawns' immune system.

The significant violation from HWE at this locus, however, could simply be caused by the limited sample size of individuals per population used in this study (Wang et al., 2005), or as a result of undetected nulls.

5.3 CHARACTERIZATION OF EST-SSR LOCI ISOLATED FROM *M. rosenbergii*

Among the seven EST-derived SSR loci examined in this study, there were only 4 to 10 alleles per locus. In contrast, previous studies on genomic SSR loci for *M. rosenbergii* studies have detected 12 to 18 alleles per locus (eastern form *M. rosenbergii*: Chand et al., 2005); and 5 to 17 alleles (western form *M. rosenbergii*: Charoentawee et al., 2006). This slightly lower allelic number for EST derived microsatellites is expected, as the evolution of EST loci is potentially constrained by the fact that their accurate transcription may be directly or indirectly linked to important functional processes.

In general, functional constrains in transcribed region of genome appear to be related to lower polymorphism of EST-SSRs since these markers are located within genes, and thus more conserved across species (Ellis & Burke, 2004; Kim et al., 2008). Other factors influence the level of polymorphism of EST-SSR: type of unit repeats (di-,tri-), number of unit repeats, and the region of the gene where they occur (Metzgar et al., 2000, Li et al., 2002, Coulibaly et al., 2005, Varshney et al., 2005, Kim et al., 2008). For examples, Metzgar et al. (2000) reported that all types of SSR repeats (from mono- to hexanucleotides) can be found in excess in noncoding regions of model species across seven eukaryotic clades including *S. cerevisiae, C. elegans*, *Drosophila*, plants, and primates. Besides that, a study by Morgante et al. in 2002 (as reviewed by Li et al., 2002) documented that the occurrence of repeat numbers other than triplets and hexanucleotides were significantly less frequent in protein-coding sequences relative to noncoding fraction in six plant species.

However, as cited in Ellis & Burke (2004), Hamrick & Godt (1996) reported that the level of genetic diversity revealed by these markers are still considerably higher than those revealed by most alternative marker types, such as allozyme, therefore supporting the value of developing this types of markers for understanding genetic variability.

In addition to the relatively low allelic diversity observed for EST markers here, heterozygosity was also low in contrast to heterozygosity in previous studies, and PIC values were also lower than those previously observed in other studies, even where the total number of alleles per locus is similar (e.g. Divu et al., 2008).

5.4 POPULATION GENETIC STRUCTURE AMONG POPULATIONS FROM FOUR WILD LOCATIONS

The level of population genetic structuring among individuals collected from wild populations at four sampling locations was found to be low. Generally, the low degree of genetic differentiation found among sampled wild populations indicated that all *M. rosenbergii* from those locations are genetically similar. Moreover, population structure analysis did not yield any significant genetic differences among the studied populations from Sg Timun, Kedah, Sarawak, thus adding weight to suggestion that these three populations represent a single, large panmictic population.

Whilst population structure analysis revealed no significant heterogeneity between all four sampling sites, the population differentiation estimated by F_{ST} pairwise comparison indices showed the isolation of one population, namely Terengganu. Unlike the other 3 sites, samples from Terengganu were shown to be genetically different from the others. This can be observed both in the pairwise differentiation tests, and it is also reflected in the prevalence of private alleles found in this population. Out of the 10 alleles that were private to single sites, four were private to Terengganu.

Genetic differentiation and structuring of populations can be influenced by the combination of factors; gene flow, natural selection and genetic drift (Freeland, 2010). In general, population differentiation is inversely correlated with gene flow and directly correlated with genetic drift (Frankham et al., 2002; Freeland, 2005 as cited in Freeland, 2010).

High gene flow may have resulted in a lack of genetic structure among sample sites. Populations from all sample sites were very similar, with not only many alleles shared among populations, but also, very similar allelic frequencies among populations. When gene flow is low or being restricted between populations, allele frequencies are expected to diverge as the forces of genetic drift and selection act independently in each population (Slatkin, 1987). The fact that the allelic frequencies were so similar among Sg Timun, Kedah, and Sarawak clearly indicated that either there is ongoing gene flow between them, or that populations have only been isolated in very recent evolutionary history.

Dispersal of *M. rosenbergii* that could lead to the gene exchange between those populations may potentially have resulted from translocations by human. Translocation by human is commonly achieved for economically important freshwater taxa for food commercial purposes (Adamson, 2010), and *M. rosenbergii* is one example of species that owe its distribution pattern to this practice. In less than 10 years, it was reported that *M. rosenbergii* lineages from Sarawak drainage system was brought to the Pulau Sayak Hatcheries Centre in Kedah, to be crossed with Kedah lineages, and the F_1 generation from this cross was then re-stocked in Sg Muda for two purposes: (1) stock enhancement program; and (2) to increase livelihood of local people who heavily relied on prawn farming activity during that time (Dr. Subha Bhassu, personal communication, 7 July, 2011).

On an earlier occasion (2002), recent communication with Dept. of Fisheries revealed that mixing of individuals from Kedah and Sg Timun also occurred as mediated by human translocations for similar purpose, i.e stock enhancement program (Dr. Subha Bhassu, personal communication, 7 July, 2011). There was no evidence that any samples from Terengganu were involved in this program. This might explain the similarity observed between those three lineages, namely Sg Timun, Kedah, and Sarawak; and why Terengganu was genetically differentiated from others. However, this practice has been stopped in 2007, as translocations among and within mainland South-East Asian drainages should be avoided whenever possible. This is crucial in maintaining natural patterns of genetic diversity, since the mixing of divergent lineages is associated with risk that could undermine natural patterns of genetic structure in wild populations from a genetics perspective (Adamson, 2010).

Apart from that, it is quite likely that gene flow is ongoing among these three populations, as the life history of *M. rosenbergii* includes a marine phase. For obligate freshwater species, populations are often significantly structured as a result of reproductive long term isolation at the river drainage level (Malecha et al., 2010). For *M. rosenbergii*, the marine or brackish water phase appears to be the most likely pathway to gene exchange enabling localized dispersal among neighbouring drainages to occur (Malecha et al., 2010). During the marine phase, adult females of *M. rosenbergii* migrate from freshwater rivers into estuarine water to spawn, and larvae can survive in saltwater for up to 22-35 days, potentially enabling them to move to other estuaries and hence colonize non-natal rivers. This could easily result in gene flow among rivers, as juveniles may settle in different rivers from the ones in which their parents occurred.

Alternatively, the lack of significant differentiation observed could result from a very recent barrier to dispersal. Drainage systems in Peninsula Malaysia and the Borneo Archipelago were connected during the last glacial era (Voris, 2000). It is possible that the populations sampled actually represent the remnants of one ancestral large population that inhabited an ancient extended drainage network. As sea levels rose over the past 11,000 years (Voris, 2000), this ancestral population may have been fragmented into the smaller river systems that remained above the higher sea level. In this case the lack of differentiation may be explained by this recent common origin from a single population.

Other than not being touched by human intervention as stated earlier, Terengganu population may also be genetically differentiated because it is geographically isolated from other populations, and therefore high gene flow is infrequent among Terengganu and the other sites. As cited in his review, Woodruff (2003) pointed out that the spine of Thai-Malay Peninsula dominated by north-south ranges of hills has existed for over 100 Myr. The presence of this mountain range acts as a natural barrier which partially isolated Terengganu river from other drainage systems in Peninsula Malaysia, thus restricts the dispersal of Terengganu strain throughout the rivers. Nevertheless, though the differentiation is shown to be statistically significant, the magnitude is very small, implying that all populations are very similar.

As discussed above, it appears that Malaysian *M. rosenbergii* populations are not genetically structured, with the exception of those from Terengganu where very limited differences are present. This indicates that all populations (Sg Timun, Kedah, Sarawak, and Terengganu) can be considered one large panmictic population for management purposes. This finding is important because it serves as preliminary information for future studies. For future studies, more molecular markers could be employed to obtain precise data in characterizing genetic diversity of natural populations as well as cultured lines. Further assessments of wild broodstock will be essential in order to get a better comprehension of overall framework in conservation strategies together with genetic management of local natural resources.

CHAPTER VI

Conclusion

Macrobrachium rosenbergii is one of the important freshwater species for commercial aquaculture, not only for domestic industries, but also all over the world. In an effort to ensure sustainable aquaculture industries, the continuous supply of wild stock *M. rosenbergii* seeds must be ensured. Apart from that, it is also important to conserve their natural reservoir in the entire biological ecosystem to avoid loss of genetic diversity. However, wild stocks of *M. rosenbergii* especially in Malaysia currently face the threat of extinction due to over-exploitation and also from rapid development and human urbanization that contributes to habitat destruction, as well as pollution and many others. All these issues lead to the rise of an alarming call on the active management and conservation strategies of the species population.

Advances in molecular methods have paved the way for the development of state-of-the-art technology to achieve the above goals. The newly developed EST-SSR markers represent a potentially valuable source of gene-based markers for various purposes in genetic studies. One application of EST-SSR is for population genetic analyses. EST-SSR offer benefits such as rapid and inexpensive development, and considerable degree of polymorphism which holds promise for application in genetic diversity assessment of particular species. Therefore, the present study was conducted to demonstrate the utility of novel EST-SSR marker sets developed from *M. rosenbergii* transcriptome for characterizing the genetic variation and genetic structuring of individuals sampled from four locations in Malaysia, namely Sg. Timun, Kedah, Sarawak, and Terengganu as part of a management strategy of this species.

Results from this study showed that PCR amplification using EST-SSR primers successfully produced the desired size of amplicons, with higher prevalence of tri- repeats compared to di-repeats. The occurrence of stutter peaks was also found to be lower than genomic-SSR, indicating less occurrence of unspecific or artifact products. Locus EST MR8 might be under selection as revealed by the deviation from HWE, probably due to its di-repeat motif that is unfavorable in gene sequences to maintain the triplet codon of reading frame, or the selection pressure could be associated with the importance of the locus as an immune-related gene.

Besides that, polymorphism level of EST-SSR such as number of alleles per locus and heterozygosity, though found to be lower than anonymous-SSR were still informative and useful to portray the level of variation that occurs in the selected populations. This implies that EST-SSR markers are suitable to be employed in assessment of genetic diversity of *M. rosenbergii* populations as preliminary screening to compliment further actions that need to be done in upcoming studies.

The findings of this study can be concluded that intra-specific diversity that occurs between all studied populations were not extremely high, as very low variation was detected in pairwise comparisons and genetic structuring analyses. No significant difference was detected between samples involving Sg. Timun, Kedah, and Sarawak. The ongoing gene flow (either naturally or via translocations by humans) or recent isolation from a common ancestors; are possible reasons that accounted for the low magnitude of genetic differentiation. Terengganu population was found to be genetically isolated in pairwise analysis. This was likely due to maintained natural isolation or by barrier restriction. Hence, very limited variations exist. The

overall results suggest that all populations were composed of a single, large, panmictic population.

Last but not least, it is vital to mention here that these findings serve as a baseline for future studies to ensure the prolonged existence of *M. rosenbergii*. While more studies should be carried out in attempt to conserve *M. rosenbergii* populations, other efforts must also be put into the overall framework to explore vast and endless applications of EST-SSR as molecular marker not only in population genetic analyses, but in other related fields as well, such as construction of high-density linkage map, or gene hunting for commercial purposes.