UNIVERSITI SAINS MALAYSIA

UNIVERSITI SAINS MALAYSIA

HVR-I Sequence Analysis of Malay Population of Peninsular Malaysia

Dissertation submitted in partial fulfillment for the Degree of Bachelor Science in Forensic Science

Ayu Suzailiana Binti Muhamad

School of Health sciences
Universiti Sains Malaysia
Health Campus
16150, Kubang Kerian, Kelantan
Malaysia

2007

CERTIFICATE

This is to certify that the dissertation entitled

"HVR-I Sequence Analysis of Malay Population of Peninsular Malaysia"

is the bona fide record of research work done by

Ms. Ayu Suzailiana Binti Muhamad

during the period of December 2006 to May 2007 under my supervision.

Signature of Supervisor : ----------

Name and address of supervisor : Dr Zafarina Zainuddin
School of Health Science,
Health Campus,
Universiti Sains Malaysia,
16150 Kubang Kerian, Kelantan,
Malaysia.

Date :

ACKNOWLEDGEMENT

In the name of Allah, the most gracious, the most merciful. Shalawat and remembrance for the Holy Prophet, Muhammad S.A.W. My gratitude to God Almighty Allah S.W.T. for the blessing in giving me strength, health, wealth, and motivation to accomplish this final year project.

During this project work, I encountered many kind individual that have generously their help and support to finish this research project. I would like to dedicate my special thanks to:

- My beloved supervisor, Dr Zafarina Zainuddin for her guidance, thoughts, ideas, time, and caring. It is luck for me to have such a great supervisor. I will not able to finish this project without her kindness and loving support.
- My parents, Mr Muhamad bin Mustafa and Mrs Suriyani binti Mat Junus for their full moral support and pray.
- My partner, Lee Loong Chuen for her help and support whenever I feel give up and tired.
- All the laboratory staffs of Forensic Science Laboratory and Molecular Biology Laboratory PPSK and all the members of Unit Kemudahan Makmal PPSK for the cooperation.
- All my friends who are generously contribute their time, idea, and support to finish this project, especially to Tuan Khalid bin Tuan Mat, Edi Nur Hisyam bin Atan and Nur Hasanah binti Ruslan.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

bp	:	base pairs
DNA	:	Deoxyribonucleic Acid
dNTP	:	Deoxynucleoside triphosphate
EDTA	:	Ethylenediamine tetraacetic acid
EtBr	:	Ethidium bromide
HVR	:	Hypervariable region
kb	:	kilo base pairs
MgCl	:	Magnesium chloride
mtDNA	:	Mitochondrial DNA
nDNA	:	Nuclear DNA
ng	:	nanogram
PCR	:	Polymerase Chain Reaction
RFLP	:	Restriction Fragment Length Polymorphism
RNA	:	Ribonucleic Acid
TBE	:	Tris Boric EDTA
ul	:	microlitre
UV	:	Ultra violet

ABSTRACT

Sequence of hypervariable 1 (HVS-1) region was analyzed in 37 unrelated Malay population of Peninsular Malaysia. Buccal cell samples were collected from Universiti Sains Malaysia students that came from various parts of Peninsular Malaysia.

A 380 bp fragment of HVS-1 region was amplified using a set of primer, HVA1 and HVB1. Agarose gel electrophoresis was carried to examine the Polymerase Chain Reaction (PCR) product and the product was then purified prior to sequencing using QIAquick PCR Purification kit (QIAGEN). From the total of 37 samples sent for sequencing, only 25 samples were successfully sequenced.

The sequence for each sample was compared to the Anderson reference sequence which act as a reference sequence to report for polymorphism; including total of polymorphic site counted, haplotypes and haplogroup. From result obtained, 52% of the samples were classified under haplogroup M, 24% were classified under haplogroup F, 16% were classified under haplogroup B and another 8% were classified under other haplogroups. This result support the finding reported by other researchers (Schurr and Wallace, 2002; Ballinger *et al.*, 1992).

1. INTRODUCTION

1.1 DNA Fingerprinting

Human identification is the most important task to be done in forensic field. These include determination of the identity of the victim in mass disaster, identifying the missing person, and identifying the accused. There are many ways of individual identification but the most reliable method available is deoxyribonucleic acid (DNA) fingerprinting or also called DNA profiling.

DNA fingerprinting was first submitted in the court as evidence in 1986 due to effort done by Sir Alec Jeffreys in introducing this method in 1985 (Jeffreys *et al.*, 1985). Nowadays, DNA fingerprinting becomes important in anthropological and evolutionary research (Cann *et al.*, 1987 and Vigilant *et al.*, 1991), as well as in forensic studies (Sullivan *et al.*, 1992, Holland *et al.*, 1993, Boles *et al.*, 1995, and Ginther *et al.*, 1992) because it can be obtained from any source of biological materials containing nucleated cells with genomic DNA (Schneider, 1997). In a given individual, every single somatic cell contains the entire genetic information thus enabling the same DNA type to be obtained from any tissue such as hair shaft, nail, skin, bones and semen. Many criminal and civil cases were solved through this method and it becomes admissible as evidence in the court of law.

1.2 Human Mitochondrial DNA

Anderson *et al.* had reported the sequence of human mitochondrial DNA (mtDNA) in 1981. It is known as Anderson sequence and serves as the reference to which subsequent analyses are commonly compared. Recently, this sequence had been revised and named Cambridge reference sequence (Andrews et al., 1999).

Mitochondrial DNA is a circular genome of 16 569 base pairs (bp) (Butler and Levin, 1998). It is inherited exclusively from the mother, thus all the female members in the family will have the same mtDNA information. Its mutation rate is at least 5 to 10 times more that the mutation rate of nuclear DNA (nDNA). mtDNA consist of 2 regions; coding region and non-coding region (Butler and Levin, 1998) and it is comprised of a purine-rich strand and a pyrimidine-rich strand which are commonly called the heavy and light strands respectively.

The coding region contains the sequence for 2 ribosomal ribonucleic acid (RNAs), 22 transfer RNAs and 13 proteins. The non-coding region is also called control region or displacement loop (D-loop) with approximately 1,100 bp in length. This D-loop region consist of hypervariable 1 (HVR 1), ranges from position 16,024 position to 16,365 and hypervariable 2 (HVR 2) regions, positioned from 73 to 340 (Wilson *et al.*, 1993). Recently, the hypervariable 3 (HVR 3) region, situated between region 438 to 574 (Lutz *et al.*, 1997) has also been identified.

Most variability between individuals is found in hypervariable regions. That is why most of forensic applications involve the examination of these regions (Butler and Levin, 1998). This study also used the sequence of hypervariable (specifically, HVR 1) region to classify the haplogroup of Malay population of Peninsular Malaysia. Between unrelated individuals, this control region had been estimated to vary about 1-3% (Piercy *et al.*, 1993). According to Finnila *et al.*, (2001) and Wallace *et al.*, (1995), there are some nucleotide position at this region mutate more often than others. These nucleotide positions have been called as hotspots for base pair substitution.

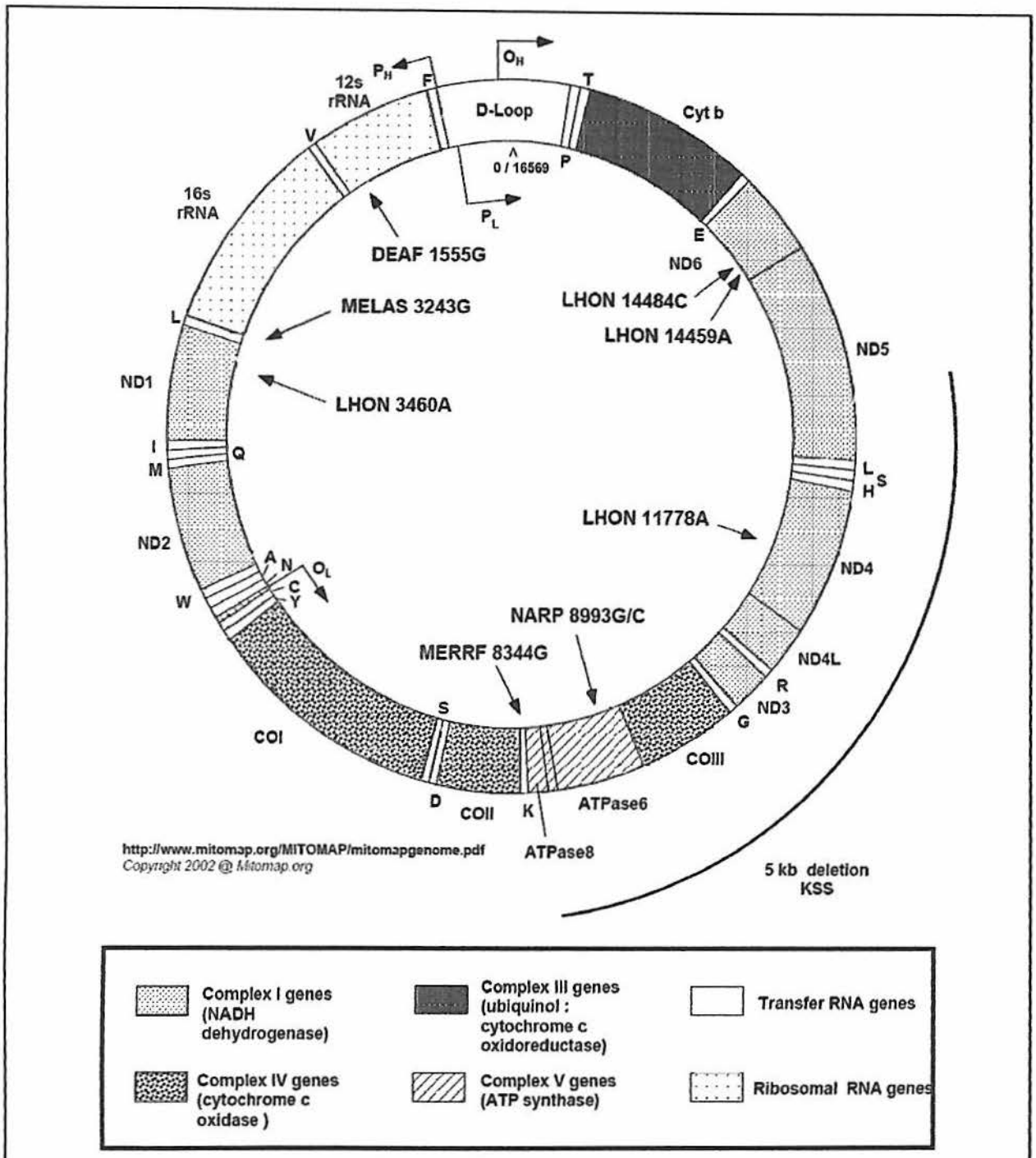


Figure 1: Schematic map of human mitochondrial DNA (www.mitomap.org)

1.3 Advantages of Mitochondrial DNA in Forensic DNA Profiling

There are two types of DNA in human body, nuclear DNA (nDNA) and mitochondrial DNA (mtDNA). Recently, the used of mtDNA had become more popular since it has many advantages. In forensic field, the mtDNA value arises from the sequence variability of the control region between individuals, its efficient polymerase chain reaction (PCR) amplification with limited biological material (Sullivan *et al.*, 1992 and 1991), and its resistance to extreme environmental conditions (Sullivan *et al.*, 1992 and Boles *et al.*, 1995).

Apart from these, we are able to extract the mtDNA from regions or body parts where we usually failed to extract the nDNA. These regions or body parts include teeth (Boles *et al.*, 1995 and Ginther *et al.*, 1992), hair shafts (Hopgood *et al.*, 1992 and Wilson *et al.*, 1995), bone fragments (Sullivan *et al.*, 1992 and Kurosaki *et al.*, 1993) and human feces (Hopwood *et al.*, 1996). These body parts are the most parts that the forensic scientist always deal with in skeletal remain cases. The ability of recovery of mtDNA in such limited sample is due to its high copies per cell. It can exist up to 1000 copies per cell compared to nDNA which is only 2 copies per cell (Butler and Levin, 1998).

1.4 Polymerase Chain Reaction (PCR)

PCR was first introduced by Kerry Mullis in 1989 (Mullis *et al.*, 1986). In Forensic identification using mtDNA, the first task to be carried is Polymerase Chain Reaction (Wilson *et al.*, 1995; Vigilant *et al.*, 1989). PCR is used to amplify a fragment of DNA by using a pair of oligonucleotide primers complementary to one end of the target sequence.

Most forensic cases deal with very limited DNA samples and sometimes the samples are highly degraded. However, with the PCR technology, these problems can be overcome. It can copy the samples into much more quantities so that enables the further analysis to be done. PCR is a cycle of three steps; denaturing, primer annealing, and polymerization.

In the first cycle, the target DNA is separated into two strands by heating at 94°C. The temperature is then reduced before the melting temperature to allow the primers to anneal. The annealing temperature depends on the primer lengths and sequences. After annealing, the temperature is increased to 72°C for optimal polymerization. If PCR is 100% efficient, one target molecule would become 2^n after n cycles. In practice, 20-40 cycles are commonly used (Turner *et al.*, 1997).

1.5 Mitochondria DNA Analysis in Population Studies

mtDNA has become a powerful tool for population study (Stoneking *et al.*, 1993). Most studies used mtDNA analysis to determine the origin of that population due to many advantages such as the sequence variability of its control region between individuals, its efficient polymerase chain reaction (PCR) amplification with limited biological material (Sullivan *et al.*, 1992, 1991), and its resistance to extreme environmental conditions (Sullivan *et al.*, 1992; Boles *et al.*, 1995). It also can be used to trace the maternal ancestry and to understand the human migration and evolution. It has been reported that mtDNA distributions is significantly different among individuals come from different geographic and ethnic origin (Brown *et al.*, 1980).

Among both coding region and control region of mtDNA, control region become more popular to be study. This is because most variability among individuals was found here (Butler and Levin, 1998). For example, a study by Schurr and Wallace in 2002 had analyzed HVR 1 region sequence to determine the mtDNA haplogroup distribution in Southeast Asia. In other study, the HVR 1 and HVR 2 regions also had been sequenced in order to demonstrate the sequence diversity of mtDNA in the Taiwanese Han population (Tsai *et al.*, 2002).

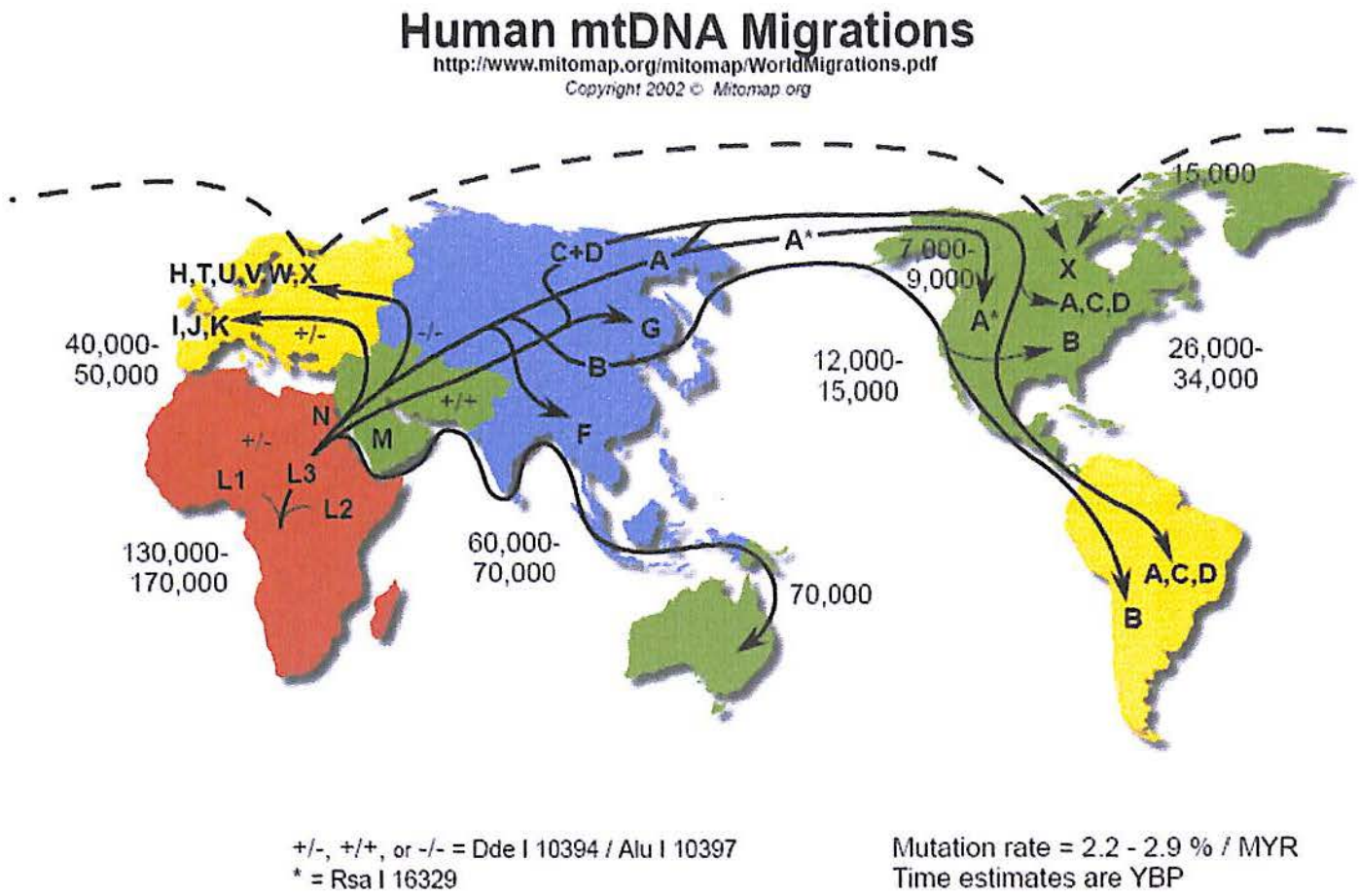


Figure 2: Human mtDNA migrations and its geographic distributions (www.mitomap.org)

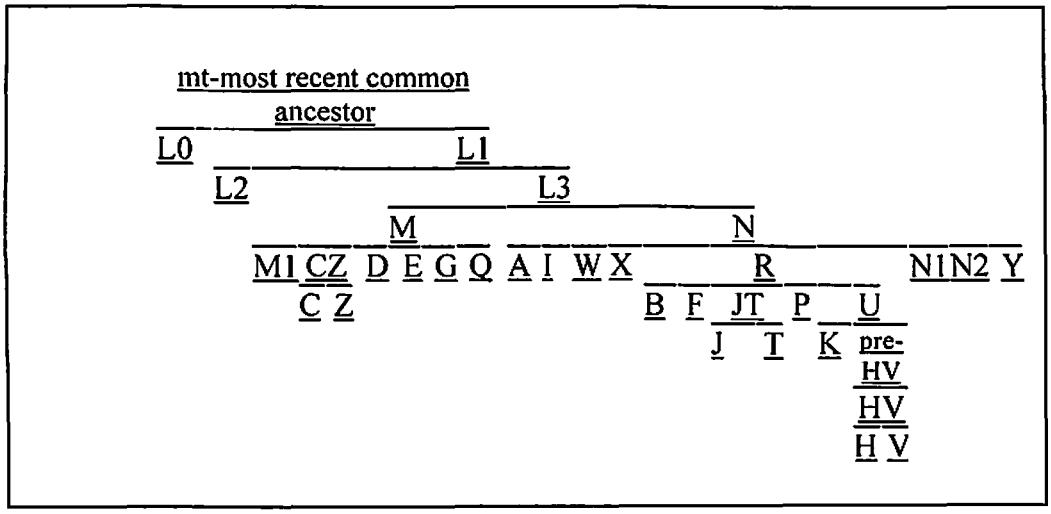
1.6 Mitochondria DNA Haplogroup

Series of alleles at specific locations on a chromosome are called as haplotypes while a collection of haplotypes is called haplogroup. Haplogroup can be defined by differences in human mtDNA. Since mtDNA is passed solely on the matrilineal line, it can be used to define genetic population and to trace the matrilineal inheritance of modern humans back to human origins in Africa and the subsequent spread across the globe.

The mitochondrial haplogroups are divided into 3 main groups, which are designated by the 3 sequential letters L, M, N. Humanity first split between the L group and the rest, and then the M and N groups split (Figure 3). Haplogroups are often geographically oriented and it can be used to define genetic populations. For example, haplogroups L, L1, L2, and L3 are common in Sub-Saharan. Haplogroups H, T, U, V, X, K, I, J, and W which derived from macro-haplogroup N are most common in West Eurasian. In East Eurasian, haplogroup A, B, C, D, E, F, and G are most common. Haplogroup C, D, E, and G are derived from macro-haplogroup M. (en.wikipedia.org /wiki/Haplogroup)

Figure 3: The Human Mitochondrial DNA Haplogroups

(en.wikipedia.org/wiki/Haplogroup)



1.7 Sample Study

Geographically, Malaysia consists of two lands separated by South Chinese Sea. These two lands are named as West and East Malaysia. West Malaysia is also known as Peninsular Malaysia which covers about 132, 000 square kilometers. The samples were randomly taken from the Malay population of Peninsular Malaysia. They are considered as Malay when their previous three generations are Malay without any mixed marriage. Malays are traditionally classified as a member of Mongoloid race. Federal constitution of Malaysia stated that a Malay is a person who born locally, speaks Malay, practice Malay custom and professes Islam. All of the samples collected were the students of Health Campus of Universiti Sains Malaysia, who originate from different parts of Peninsular Malaysia.

2. REVIEW OF LITERATURE

Eurasian mtDNAs belong to two superhaplogroups ("trunks") M (Chen *et al.*, 1995) and N (Quitana-Murci *et al.*, 1999; Alves-Silva *et al.,* 2000). The trunk M encompass the known Asian-specific haplogroups C, D, E, G, and Z while the trunk N encompass the known Asian-specific haplogroups A, B, F, and Y. Restriction fragment length polymorphism (RFLP) analysis has been the primary method of detecting sequence variation in the coding regions of the mtDNA genome whereas sequencing studies have primarily targeted the non-coding control region to detect nucleotide diversity among individuals (Lum *et al.,* 1994; Sykes *et al.,* 1995; Horai *et al.,* 1996; Kolman *et al.,* 1996; Red and Stoneking, 1999).

Mongolia which lies in the heart of east Central Asia, is the first region inhabited by modern humans 20,000 to 25,000 years ago (Fiedel, 1992). In 1993, Nei and Roychoudhury proposed that Mongolia represent the origin of colonizing migrations to the north and south. Mongolians may also represent the first colonists of the New World (Turner, 1984; Neel *et al.,* 1994). According to archeological data, Mongolians represent a group central to the understanding of the evolutionary and population genetic relationships of East Asia and New World populations.

Southeast Asian populations were genetically similar to each other. This suggested that they had a common origin (Schurr and Wallace, 2002). Majority of all mtDNAs in Southeast Asia constitute haplogroups B, F, and M. Haplotypes from a major founding haplogroup in Papua New Guinea were present in Malaysia. The Vietnamese and Malaysian aborigines had high frequencies of haplogroup F, which also seen in most other Southeast Asian population and haplogroup B was present throughout the region. In addition, the Malaysian and Sabah aborigine populations exhibited a number of unique mtDNA clusters that were not observed in other populations (Schurr and Wallace, 2002).

Table 1: Haplogroup designations and region(s) in which it primarily occurs

Haplogroup Designations	*Region(s) in which Haplogroup Primarily Occurs*	*Reference*
A	Asia, Americas	Torroni et al. 1992
B	Asia, Americas	Torroni et al. 1992
C	Asia, Americas	Torroni et al. 1992
D	Asia, Americas	Torroni et al. 1992
E	Asia	Torroni et al. 1994a
F	Asia	Torroni et al. 1994a
G	Asia	Torroni et al. 1994a
H	Europe, Middle East, Asia	Torroni et al. 1994d
I	Europe	Torroni et al. 1994d
J	Europe	Torroni et al. 1994d
K	Europe	Torroni et al. 1994a
L	Africa	Torroni et al. 1992
M	Africa, Asia	Chen et al. 1995
N	Eurasia	Forster et al. 2001
O	?	–
P	PNG	Forster et al. 2001
Q	PNG	Forster et al. 2001
R	Eurasia	Macaulay et al .1999
S	?	–
T	Europe, Middle East	Torroni et al. 1996
U	Europe, Middle East, Asia, Africa	Torroni et al. 1996
V	Europe, Middle East	Torroni et al. 1996
W	Europe	Torroni et al. 1996
X	Europe, Middle East, Asia	Torroni et al. 1996
Y	Asia	Schurr et al. 1999
Z	Asia	Schurr et al. 1999

Table 1 shows the haplogroup designations and its primarily occurs region(s). From the table, we can see that haplogroups A, B, C, D, E, F, G, H, M, U, X, Y, and Z belong to Asian. In fact, haplogroup M (16223 C to T transition) occurs in all Southeast Asian populations at varying frequencies (25%-45%), with the highest frequencies occurring in the Malays and Sabah Aborigines (~60%) (Schurr *et al.*, 2000).

Haplogroup F is also fairly widespread throughout Southeast Asia. Haplogroup F lack the 16223 C to T transition and posses the 16304 T to C transition. Haplogroup F appears in a number of Asian populations (Cann *et al.*, 1987) including Filipinos and Aboriginal Taiwanese (Melton *et al.*, 1995 and 1998) and occurs in highest frequencies in Southeast Asia, specifically in the Vietnamese population (Ballinger *et al.*, 1997). Relatively, its high frequency also can be observed in the Orang Asli of Malaysia, to whom the Vietnamese show linguistic ties (Bellwood, 1979). Based on RFLP composition of its haplotypes, haplogroup F does not appear to be as diverse as haplogroup M (Horai *et al.*, 1984; Harihara *et al.*, 1988; Ballinger *et al.*, 1992).

Haplogroup B defined by nucleotide transition at 16189, 16217 and 16519 (Passarino *et al.,* 1993; Schurr *et al.,* 2000) is found in Vietnamese, Malaysian and Bornean populations (Ballinger *et al.,* 1992). Haplogroup B mtDNAs had a much wider distribution in Asia and the Pacific. In Aboriginal Taiwanese and Taiwanese Han populations also haplogroup B had been observed (Ballinger *et. al., 1992;* Melton *et al.,* 1995, 1998).

Haplogroup M was believed to be the first haplogroups to enter Southeast Asia (Schurr and Wallace, 2002). Haplogroups F and B not as genetically diverse as haplogroup M although they also present throughout the Southeast Asia. Studies conducted have shown that about 57.1% of the Malay population samples are belongs to haplogroup M (Schurr and Wallace, 2002).

3. OBJECTIVES OF THE STUDY

Basically, aim or objective of this study is to analyze the sequence of the hypervariable 1 (HVS-1) region of the samples in order to determine their haplogroup. From the result obtain, we can calculate the percentage of having different haplogroups in the samples of Malay population in Peninsular Malaysia.

Apart from the above two main objectives, other objectives of this study also include:

1. To obtain mitochondrial DNA profile of Malay population in Peninsular Malaysia for forensic database
2. To study the lineage of Malay population in Peninsular Malaysia

4. MATERIALS AND METHODS

4.1 Materials

4.1.1 Chemicals

TBE Buffer

10X TBE buffer was prepared by dissolving 53.9 g of Tris-base and 3.72 g of EDTA in dH2O. Boric Acid was added to adjust the pH at 8.3. TBE buffer was sent for autoclave for sterilization.

Ethidium Bromide (EtBr)

Ethidium bromide ($C_{21}H_{20}BrN_3$) is commonly used as a nucleic acid staining in molecular biology laboratory. It is an intercalating agent and used in agarose gel electrophoresis. 10 drops of EtBr was added into approximately 200 ml of water to stain the agarose gel. When exposed to ultraviolet light, it will fluoresce with a red-orange color, intensifying almost 20-fold after binding to DNA

DNA Ladder

1 kb DNA ladder was used to determine the size of the PCR product and in agarose gel electrophoresis. A total of 2ul of 1 kb DNA ladder was loaded into the gel. It was purchased from Promega, USA.

4.1.2 Apparatus

Bottles, beakers, and measurement cylinder used in this lab work were manufactured by Schott Duran, Germany. All apparatus were sterilized before used to avoid contamination.

4.1.3 ReddyMix™ PCR Master Mix

ReddyMix™ PCR Master Mix contained all components required in PCR; Taq DNA Polymerase, deoxynucleoside triphosphate (dNTPs), reaction buffer, magnesium chloride (MgCl), red dye and precipitant. It was purchased from ABgene, UK. Each vial contains 1.8 ml of 1.1 x working concentration PCR master mix.

4.1.4 Thermal Cycler

PCR process was done by this PTC-0200 DNA Engine™ Thermal Cycler. Its program can be set up and it can hold up to 96 tubes (0.2ml tubes) at one time. The thermal range of this machine is 0°C to 100°C.

4.1.5 PCR Purification Kit

QIAquick PCR Purification Kit (QIAGEN) was used for purification of PCR product. This kit consist of QIAquick spin column, PBI buffer, PE buffer, EB buffer, collection tube (2ml), and loading dye.

4.1.6 PCR Primers

Both forward and reverse primers used were synthesized by invitrogen. The sequence of each primer is listed in table 2.

Table 2: Sequence of the primers

Primers	Sequence	Site of action
HVA1 – forward primer	5'-GAG GAT GGT GGT CAA GGG AC-3'	Hypervariable 1 region (16000 – 16430)
HVB1 – reverse primer	5'-CAC CAT TAG CAC CCA AAG CT-3'	Hypervariable 1 region (16000 – 16430)

4.2 Methods

4.2.1 Sample collection

Samples were collected by previous final year students. A total of 45 samples were collected from Malay population of Peninsular Malaysia with informed consent. They were briefly interviewed in order to determine their background. Buccal swab was collected from individuals that had no mixed marriage background for at least 3 generations.

4.2.2 Sterilization

All glassware, deionised water, pipette tips, PCR tubes, eppendorf tubes and plastic ware were sterilized by autoclaving for 15 minutes at 20 psi.

4.2.3 Separate working areas

Separate working areas were dedicated for PCR preparation and handling of post-PCR products. PCR was prepared in a flow hood. The flow hood was cleaned prior and after all works with 70% ethanol. Each working area was allocated with different set of pipettes and other consumable to avoid cross contamination.

4.2.4 DNA extraction

DNA extraction was successfully done by previous final year students by using the phenol/chloroform method.

4.2.5 Polymerase Chain Reaction (PCR)

A total of 2ul samples (DNA templates), 2ul of each forward and reverse primers (10pm/ul) and 24ul PCR Reddymix™ Master Mix were added into 0.2ml tubes for PCR. The programmed that had been set up in the Thermal Cycler had been used and it took about 2 hours to finish 30 cycles.

At first, the temperature of 94°C was maintained for 5 minutes for initial incubation. This temperature then was retained for 1 minute for each cycle of 30 cycles for denaturing step. After that, for annealing step, the temperature was reduced to 48°C for 30 seconds for each cycle. For another 45 seconds for each cycle, the temperature was raised up to 72°C for amplification. Finally, for final extension, the temperature of 72°C was maintained for 7 minutes. The PCR products were kept at 4°C in a separate fridge from the pre-PCR components to avoid contamination.

4.2.6 Agarose Gel Electrophoresis

Agarose gel was prepared by adding 15g of Agarose LE powder into 100ml of 0.5X TBE buffer. The mixture was heated on the hot plate to dissolve all agarose powder in the buffer. The gel was left to solidify in the casting tray for about 1 hour. After the gel had completely set, it is transferred into the electrophoresis tank and submerged by adding approximately 300 ml of 0.5X TBE buffer. A total of 5 ul of the PCR product and 2 ul of DNA ladder were loaded into the gel and electrophorosed at 80 V for 30 minutes.