UNIVERSITI SAINS MALAYSIA

UNIVERSITI SAINS MALAYSIA

Classification of Haplogroup F in Chinese Population of Peninsular Malaysia

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

Siti Balkiah Binti Ismail

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

2006

CERTIFICATE

This is to certify that the dissertation entitled

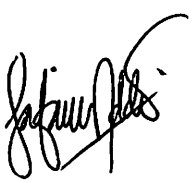
"Classification of Haplogroup F in Chinese Population of Peninsular Malaysia"

is the bonafide record of research work done by

Ms. Siti Balkiah bt. Ismail

During the period December 2005 to April 2006

under my supervision.

Signature of Supervisor: 

Name and address of Supervisor: **Dr. Zafarina bt. Zainuddin**
School of Health Sciences
Universiti Sains Malaysia
Health campus
16150 Kubang Kerian, Kelantan
Malaysia

Date: 14/05/2006

ACKNOWLEDGEMENT

Thanks to Allah the most merciful and the most gracious for His blessing. Thank you for giving me the courage and health to complete my research project.

I would like to dedicate my appreciation and million of thanks to my parents Ismail Yeen and Aminah A.Rahman, my sister and brothers for their continuous moral and financial support. I am very grateful and owe great debt to my supervisor Dr. Zafarina Zainudin, lecturer of School of Health Sciences, who not only contributed by lending journals but also monitoring and maintaining the progress of my research.

I also thanks to Mrs. Rosniah Yusoff, Mrs. Hafizah Harun and Ms. Nur Arnida and all laboratrory staff for their willingness to help me to carry out the laboratory work. My special thanks to all my course mate and friends especially Najiah, Nur Hafiza, Nor Alfarizan, Rosaniza and Khairul Anuar, my beloved brother Zainul Adreen Izzaid and Zainal Abidin for their support and guidance that accompanying me from start till the end of the project.

Finally, thanks to all sample donors for their cooperation and those who participate directly or indirectly in helping me to accomplish this research. The success of my project is owed in part because of them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

bp	:	base pair
DNA	:	Deoxyribonucleic acid
dNTPs	:	Deoxynucleotide triphosphate
etBr	:	Ethium Bromide
HCl	:	Hydrochloric acid
HVS-1	:	Hypervariable region 1
HVS-2	:	Hypervariable region 2
kb	:	kilo base pair
MLP	:	Multi Locus Probe
mtDNA	:	Mitochondrial DNA
NaCl	:	Sodium Chloride
PCR	:	Polymerase Chain Reaction
RE	:	Restriction enzyme
RFLP	:	Restriction Fragment Length Polymorphism
RNA	:	Ribonucleic acid
rRNA	:	Ribosomal ribonucleic acid
SDS	:	Sodium dodecyl sulfate
SLP	:	Single Locus Probe
tRNA	:	Transfer ribonucleic acid

ABSTRACT

mtDNA is passed from one generation to the next, essentially unchanged, solely through the maternal line of a family. A total of 44 samples were randomly collected from USM Kampus Kesihatan, Kubang Kerian, Kelantan. mtDNA was extracted from the buccal swabs. All samples were subjected to PCR by using a set of primer, 12233-F and 12630-R to amplify a 397 bp fragment of mtDNA coding region. This fragment was then digested using *Hinc* II enzymes for determination of the Hinc II site loss or site gain at nucleotide position 12406 for all samples. Samples which did not cut by the Hinc II enzymes were classified into the F haplogroup. Only 2 samples which are 4.5 % from 22 samples analysed were found to belong to haplogroup F.

INTRODUCTION

FORENSIC DNA ANALYSIS.

Over the past decade, a rapid development of DNA technology had increased the ability to identify individuals in forensic science. The first DNA technology used was DNA fingerprinting which had been introduced by Sir Alec Jeffry in 1985 (Schneider, 1997). This technique was first used in 1986 in England in the case of Colin Pitchfork, who had convicted sexual assault and murder of two teenage girls.

Before the use of DNA in forensic science, identification depends on human fingerprints and a set blood group markers (Martin *et.al.*, 2001; Holland *et.al.*, 1999). Both methods had a limitation since the conventional analysis of blood group depends on the availability of blood or related body fluids and the human fingerprints depend on the available of the source material. Meanwhile the DNA profiling does not rely on the nature of the material because each single cell of an individual will carry the entire genetic information (Schneider, 1997).

As the time progressed, forensic DNA analysis has been widely accepted and has become enormously powerful tools for various kinds of investigations. It was quite effective due to the DNA sequence that varies among human. They are now so discriminating and so sensitive and were applicable to wide variety of

biological evidence such as saliva, blood, semen, hair, bone, skin and organs (Andreasson *et.al.*, 2006).

RESTRICTION FRAGMENT LENGTH POLYMORPHISM (RFLP)

Analysis of variable number of tandem repeats of DNA was based on restriction fragment length polymorphism (Carey *et.al.,* 2002). This technique was brought to the attention by Sir Alec Jeffrey in the middle 1980's. The RFLP determination is done by DNA digestion with restriction endonucleases followed by Southern blotting and hybridization to specific probes. The variable length of DNA resulted from the digestion process by the restriction enzyme. The restriction enzyme used to cut the DNA at specific 4 to 6 base pair (bp) recognition sites.

The presence or absence of the recognition site will produce different length of DNA fragment which then separated using the agarose gel. The estimation of each fragment size is done by comparing to the molecular size standard. Differences result from base substitutions, additions, deletions or sequence rearrangements within restriction enzymes recognition sequences.

In RFLP technique, Jeffrey had proposed the use of two probes multi locus probe (MLP) and single locus probe (SLP) (Martin *et.al.*, 2001). In MLP method, the resulting autoradiograph is like a bar code and it was proved to be a problem in conversion to a numerical number. Besides that, it required a large

amount of high molecular weight DNA to be extracted. This method was found to be demanding and not suited to the formation of computer databases.

Meanwhile in SLP method, probe will identify repeat sequences in individual minisatellite to produce maximum two bands per probe (Martin *et.al.*, 2001). The result can be recorded in a numerical data. The introduction of this latest technique had simplified the interpretation in RFLP analysis. Comparison is easier in this method compared to MLP method.

RFLPs can be used in many different ways to accomplish different objectives. It is one of the original applications method used in forensic DNA analysis.

POLYMERASE CHAIN REACTION (PCR)

The polymerase chain reaction (PCR) has strongly enhanced the usefulness of DNA profiling techniques in forensic science. PCR was invented by Kerry Mullis in December 1983. His idea was to artificially multiply the DNA template through repeated cycle driven by an enzyme called DNA polymerase. In forensic cases, usually the DNA sample is often in small quantity and even highly degraded so PCR is found to be very useful in order to make the analysis possible.

Components of PCR includes DNA template, a sets of primers, which flank the region to be amplified, Taq polymerase, dNTPs provide complementary bases for building new DNA strands and buffer which act as suitable chemical environment for the process.

Primers are a short DNA strands usually about 18 to 25 nucleotides. Shorter primer is usually avoided since it can results in non specific annealing while longer primer will increase the melting temperature. Both forward and reverse primer used in PCR should have similar annealing and melting temperature for good amplification result.

PCR is a very sensitive method and cross contamination is mainly the source to be avoided. In order to avoid contamination from post PCR product, the pre-PCR and post-PCR process must be carried out in separate areas. This commonly happened in higher organism and can alter large proportion in final products.

PCR works best for amplification of fragment between 250 to 300 bp. Longer fragments may not be amplified efficiently since DNA polymerase activity starts to fall off. Typical heating cycle may not provide enough time to complete the polymerisations. PCR optimization can be done by manipulating the annealing temperature and concentration of DNA template, primers and magnesium ion.

HUMAN MITOCHONDRIAL DNA

Human cells contains two different types of genome, the nuclear genome and mitochondrial genome. Mitochondria is a semiautonomous organelles possesing their own genome and the machinery for its replication, transcription and for protein synthesis (Saccone. C *et al.*, 2000). Mitochondrial DNA (mtDNA) encodes for essential mitochondria protein as well as tRNAs and rRNAs. Human mtDNA exist as a double stranded, circular, covalent closed molecule of 16,569 base pairs in length (Eshleman *et.al.*, 2003; Castro *et.al.*, 1998).

mtDNA is divided into two main regions, the noncoding region and the coding region. The coding region codes for 2 ribosomal RNAs, 22 transfer RNAs, and 13 proteins (Wilson *et.al.*, 1993). It produces all of the biological molecules responsible for maintaining the mitochondria as well as providing the cell with energy. The mtDNA noncoding region is approximately 1.1 kb in length and it is also known as displacement loop or D-loop (Wilson *et.al.*, 1993). It has two important segments termed the hypervariable region 1 (HVS-1) and hypervariable region 2 (HVS-2) (Figure 1). HVS-1 spans the approximate region of 16,024 to 16,365 and HV2 is located on the side of the origin of replication, approximately encompassing positions 73 to 340 (Wilson *et.al.,* 1993). The hypervariable regions are the segments of greatest importance in forensic DNA analysis because they are highly polymorphic with the mutation rate of 5 to 10 times higher hat the coding region (Wilson *et.al.,* 1993).

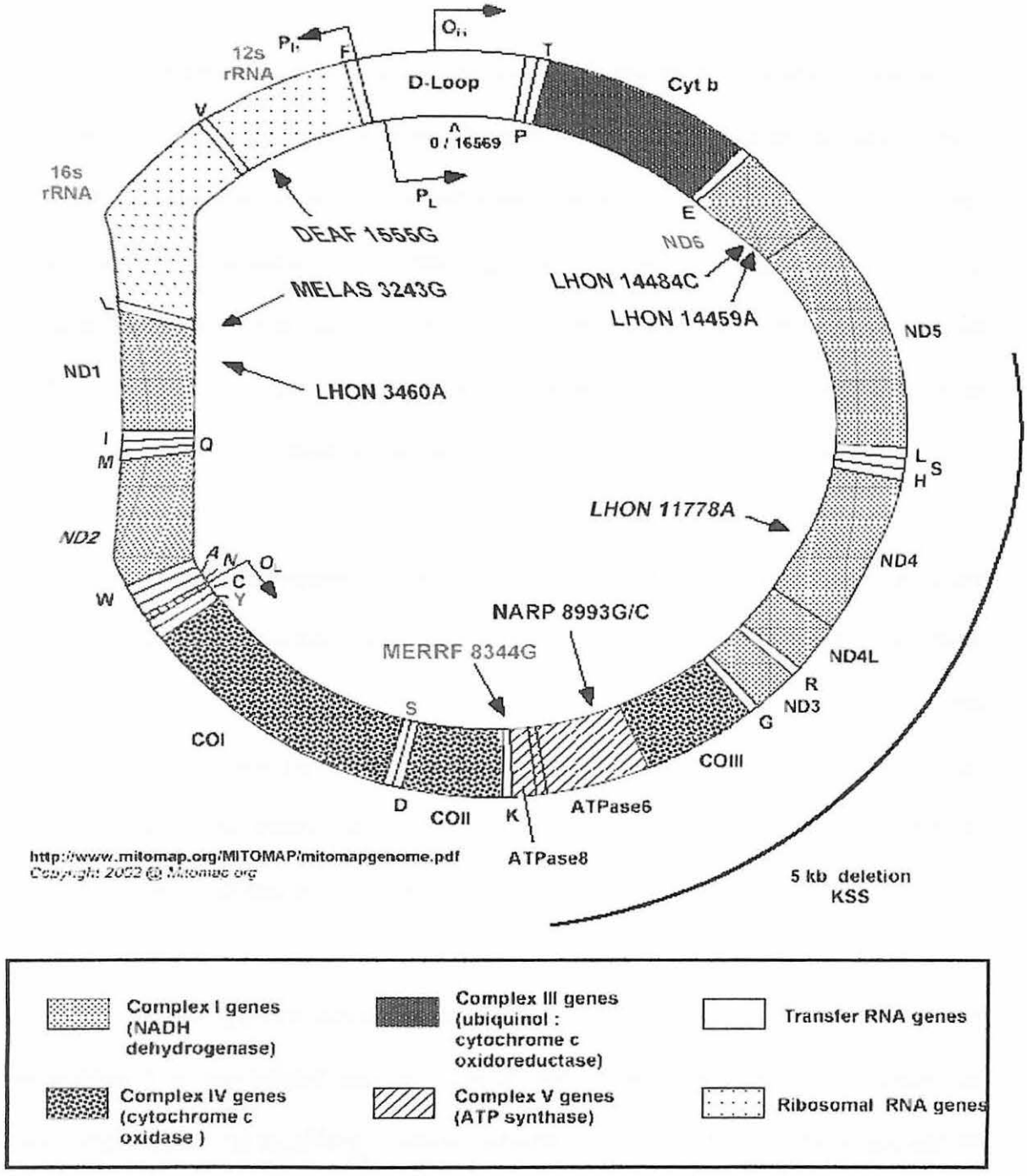


FIGURE 1: Human mtDNA map

ADVANTAGES OF mtDNA

mtDNA has proven to be a powerful exclusionary tool in forensic casework (Fisher *et.al.*,). Some of the noticable feature that differ between nuclear DNA and mtDNA include the mode of inheritance. mtDNA is inherited from mother lineage and it is passed down through the generation. Therefore, excluding mutation, all maternal relatives will have same mtDNA sequences. Eventhough mtDNA cannot be used to identify a specific individual but it still can be serve as reference source in identify the human remains and in the population studies.

The main advantages of mtDNA in forensic DNA analysis is its high copy number. mtDNA molecules are much easier to be found and extracted than nuclear DNA since most cells contains about 500 to 1000 copies of mtDNA. This feature increased the possibility for it to be recovered from biological evidence found at the crime scene which are sometimes degraded and not sufficent in quantities (Andreasson *et.al.,* 2006).

Another important advantages of mtDNA is high number of polymorphism encountered in the control region. This is due to the rapid rate of evolution of their origin (Lima *et.al.*, 2004). Some regions of the mtDNA genome appear to have evolutionary rate between 5 to 10 times faster than the nuclear genome (Wilson *et.al.*, 1993; Castro *et.al.,* 1998). This high evolutionary rate is mainly due to lack of enzymes to repair the replication error and damages in mtDNA (Castro *et.al.,* 1998).

mtDNA ANALYSIS IN POPULATION STUDIES

One of the principal aims of population genetics is to quantify the amount of inheritable variants present in nature (Konskien, 2002). Primary forces that lead to variation are natural selection, random genetic drift, mutations and gene flow (Fisher 1930; Wright 1931; Simpson 1944; Dobzhansky 1951; Simpson 1953). Molecular tools had been used extensively in the evolutionary studies. A key to successful application of the molecular tools is to recognize that different DNA markers vary in their inherent properties. It was also included the consideration of the molecular tools sensitivity and technical features which is commonly determined by their mutation and mode of inheritance (Konskien, 2002).

mtDNA analysis is known as one of the molecular tool which useful for studying prehistory (Konskien, 2002). The first human population studies that based on mtDNA were performed by RFLPs analyisis and it has revealed differences between the four great etchnic groups, Caucasian, Amerindian, African and Asian.

Mitochondria genome offer a very different perspective on human evolution and it is found as a good phylogenetic marker as it is inherited as haploid from maternal lineage (Jorde *et.al.*, 1998; Castro *et.al.*, 1998; Saccone *et.al.*, 2000). It is treated as single locus due to lack of recombination (Fisher et.al, ; Jorde *et.al.,* 1998). The differences between two mitochondrial sequences

represents only the mutation that have taken place since each sequences derived from a common ancestor (Jorde *et.al.*, 1998).

The great amount of variability accumulated in the control region represents a record of the human evolution history. In this sense, the analysis of mtDNA polymorphisms is particularly useful in human evolutionary studies and phylogenetic analysis between populations (Lima *et.al.*, 2004). The amount of mutation is roughly proportional to the time that has passed. The study of mtDNA restriction polymorphism has become a source to detect genetic variation between major human ethnic group (Wilson *et.al.*, 1993; Ivanova *et.al.*, 1999).

mtDNA HAPLOGROUPS

Haplogroups can be used to define genetic populations and are often geographically oriented. Human mtDNA practically haploid and as time passes, mutation accumulate sequentially along less and less related molecule that constitute independent lineage known as haplotypes (Meyer *et.al.* 2001). Basal mutation shared for clusters of lineage known as haplogroups (Bonatto *et.al.*,1997; Meyer *et.al.*, 2001).

Major haplogroups are continentally and ethnically specific. Majority of the Native American mtDNAs could be classified into four distinct clusters which are haplogroups A, B, C and D (Meyer *et.al.* 2001). Sub Saharan African possess

three groups L1, L2, L3 meanwhile the other nine H, I, J, K, T, U, V, W and X encompasses almost mtDNA from European, North African and Western Asian Caucasian (Bonatto *et.al.* 1997). Whereas people in Asia are mostly belong to haplogroup A, B, C, D, G, E and M.

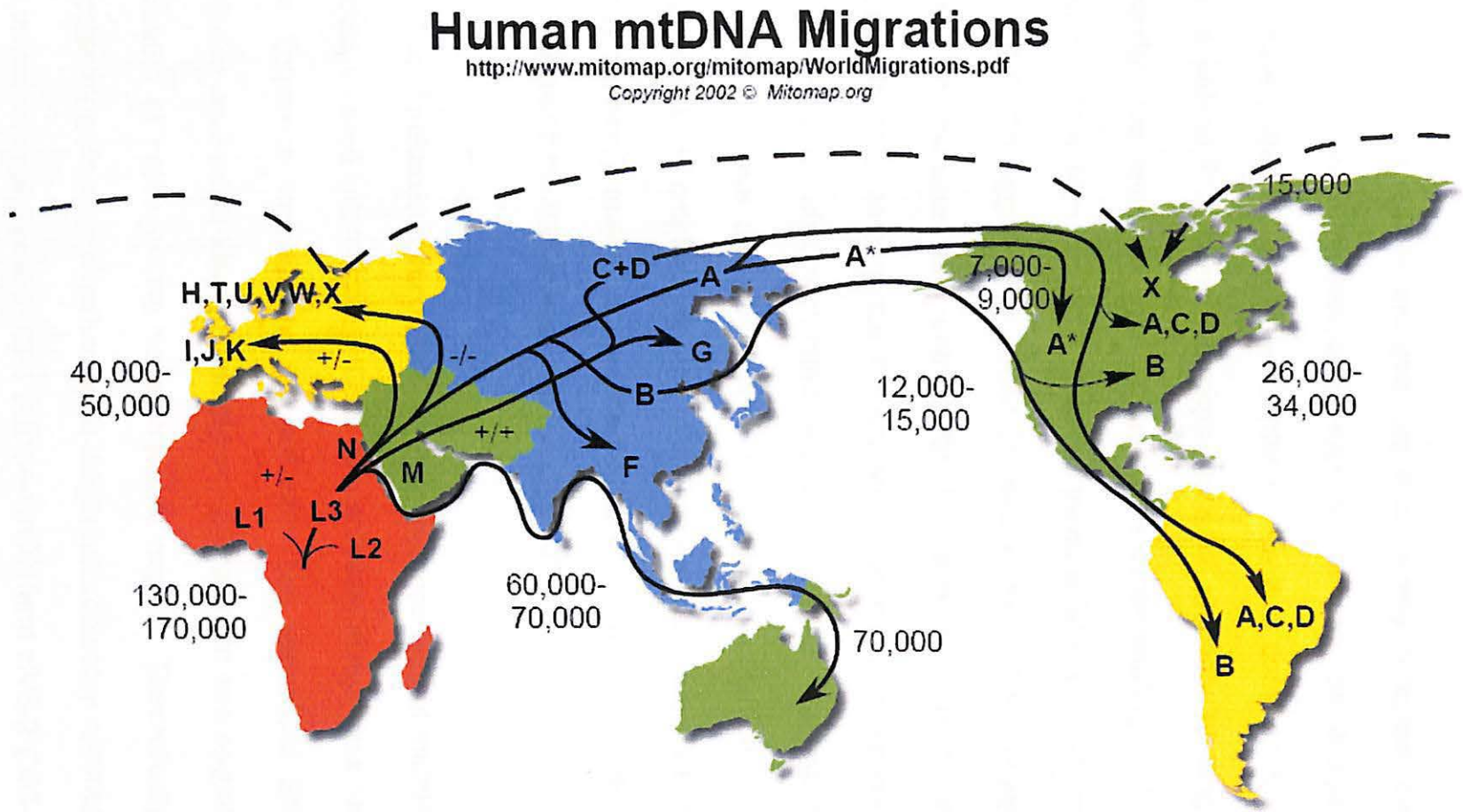


Figure 2: The world mtDNA human migration patterns and each continent specific haplogroup (www.mitomap.org)

ISSUE IN mtDNA ANALYSIS

Eventhough mtDNA analysis has been widely accepted, there are still many issues have been raised by experts involving this analysis including contamination and heteroplasmy. Contamination is a common issue and it is concerns with all PCR based technique since the presence of exogenous DNA will interfere the result. In extreme cases, the contaminating DNA can greatly exceed the DNA from donor and thereby yields a false positive result (Wilson *et.al.,* 1993). 30 amplification cycles is used in the mtDNA analysis. As the amplification increase the opportunity for detectable contamination also increases. In order to minimize the contamination many measures have been taken. Individuals who performed the analysis wear protective gear, all equipment, reagents and also bench is sterilized, extraction is exposed to ultraviolet light and amplification and extraction is done in separate rooms. In the previous studies, it stated that the interpretation of the resulting sequences will not be affected if the contamination is less that 10%.

The simultaneous occurrence of more than one type of mtDNA in a single individual is called heteroplasmy. It can be manifested in two ways sequence and length. Sequence heteroplasmy resulted when two sequences differ by base substitution meanwhile length heteroplasmy arises when two sequences differs in numbers of bases in one homopolymeric regions. Theoretically sequence heteroplasmy can occur anywhere and length heteroplasmy commonly occur at homopolymeric stretches in HVS-1 (16184-16193) and HVS-2 (303-310). Many

studies have shown that level of heteroplasmy may vary within individual as well as between generations.

CHINESE POPULATION IN PENINSULAR MALAYSIA

Malaysia is divided into two regions, known as West Malaysia and East Malaysia. West Malaysia or known as Peninsular Malaysia consists of the southern portion of the Malay Peninsula and nearby islands. North of Peninsular Malaysia is boarded by Thailand and Singapore lies off the southern coastal tip. There are different ethnic groups resides in Malaysia. The major ethnic groups in Peninsular Malaysia are Malays, which comprise 59 % of the total population, followed by Chinese (32 %) and Indians (9 %).

Chinese Malaysian is an overseas Chinese who arrived between 17th to 19th centuries. The Chinese in Malaysia belong to several Chinese dialect groups. The six major dialect groups include the Hakka, Cantonese, Hokkien, Teohchew, Hainanese and Hokchew which is also known as Foochow. A majority of the Chinese Malaysia claim to be Buddhist or Taoist. Chinese Malaysians have traditionally dominated the Malaysian economy.

REVIEW OF LITERATURE

Rapid progress of the mtDNA analysis had brought a set up of two major databases MITOMAP: Human mitochondrial Genome Database and GiiB-JST mtSNP database: Human mitochondrial Genome Polymorphism Database (Umetsu *et.al.,* 2005). Complete sequences of human mtDNA can be obtained through these databases. Both databases provides an information for reconstructing a mtDNA phylogeny and for searching the phylogenetic status of the specific haplogroups (Kong *et.al.,* 2003).

It was believed that modern humans originated 100,000 to 200,000 years ago in Africa (Kivilsid *et.al.,* 1999). The previous extensive global population studies have shown that there are differences in the nature of mtDNAs found in different geographic regions (Mishmar *et.al.,* 2002). The major fraction of mtDNA pool is made up of haplogroups A, B, F and M (Kivilsid *et.al.,* 1999). mtDNA which is present in Southeast Asia is haplogroups B, F and M. It was supported by Toroni (1994) that F was the primary haplogroups in Asia. It has widespread throughout the Southeast Asia.

Vietnamese population studied by Ivanova (1999) was characterized by significant differences in the morph frequencies for *Hinc* II. This morph was also reported in the Taiwanese and Chinese populations (Blanc *et,al.,* 1983). Previous study by Ballinger (1992) stated that haplogroup F is known as haplogroup A. It

was defined by site loss of –*Hinc* II 12406 and –*Hpa* 12406. Ballinger had found there is higher frequency of haplogroup F within Vietnamese and Malay Aboriginal population. These findings suggest that these two populations had a common mtDNA stocks.

There was also recent study of mtDNA polymorphism of haplogroup F done by Tolk. From the analysis, it found that there was considerable frequency of haplogroup F in the population of Hvar Island. Tolk also supported the definition of haplogroup F by Ballinger. It was derived from internal node R of human mtDNA phylogenetic tree and sister group to an eastern Eurasian haplogroup B, as well as to haplogroups H, J, K, T and U. As mentioned earlier haplogroup F is more frequent in Southeast Asia and it has not been reported to be presence in Europe since it was not reported in largest recent collections of mtDNA varieties. So the presence of this haplogroup in Hvar Island indicates that this maternal lineage had been added to the Croatian mtDNA pool. There were two hypothesis being argued in this study whether the haplogroup F was introduced by unidentified source population by sea or from nearby island.

In the sequencing of HVS-1 and HVS-2 of mtDNA of six province of China which involved 263 unrelated Han individuals by Yao (2002), showed there were presence of new subhaplogroup F1 (F1c). Haplogroup F1 which is originally introduced by Torroni (1994) is characterized by 10609. Meanwhile for this new subhaplogroup, it is characterized by site of 10454. Allard *et.al* (2004) supported

Yao *et.al* (2002) by defining haplogroup F1c in the presence of SNPs 16111T, 16129A, 16304C, 152C and a deletion at position 249.

Haplogroup F also appears in a number of Asian populations, including Filipinos (Cann et.al., 1987) and Aboriginal Taiwanese (Melton et.al., 1995, 1998). Study done by Zainuddin and Goodwin (2004) also revealed haplogroup F (F1a) in Malay population but not in Orang Asli.