

**UNIVERSITI SAINS MALAYSIA**



**UNIVERSITI SAINS MALAYSIA**

**Classification of Superhaplogroup M in Chinese  
Population of Peninsular Malaysia**

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

**Nor Alfarizan Binti Mokhtaruddin**

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

**2006**

## **CERTIFICATE**

This is to certify that the dissertation entitled

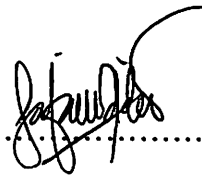
**“Classification of Superhaplogroup M in Chinese Population  
of Peninsular Malaysia”**

is the bonafide record of research work done by

**Ms Nor Alfarizan Binti Mokhtaruddin**

during the period **December 2005 to April 2006** under my supervision.

Signature of supervisor:



Name and address of supervisor:

**DR ZAFARINA ZAINUDDIN**  
School of Health Sciences,  
Health Campus,  
Universiti Sains Malaysia,  
16150 Kubang Kerian Kelantan,  
Malaysia.

Date:

## **ACKNOWLEDGEMENT**

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

**I would like to thank Allah the Almighty for giving me the strength, guide and spirit to finish this research project successfully. Without His blessings, this project may not accomplish and cannot be submitted in time.**

**Sincerest thanks to Dr. Zafarina Zainuddin for her valuable help, support, information and opinions, which guided me in completing this research project. I would also like to express my appreciation to all the donors for their participation, which made this research possible.**

**Special thanks to my parents Mokhtaruddin Md Jelas and Puteh Laili Hashim, my beloved sisters and brothers for their continuous blessing and support. To all my friends and course mates of USM Forensic Science 02/03 especially Nur Hafiza, Najiah Ismail, Siti Balkiah, Rosaniza and those who participate direct or indirectly in helping me to accomplish this research project successfully.**

# TABLE OF CONTENTS

<b>DIVISION</b>	<b>PAGE</b>
<b>ABSTRACT</b>	<b>1</b>
<b>INTRODUCTION</b>	<b>2</b>
<b>REVIEW OF LITERATURE</b>	<b>13</b>
<b>OBJECTIVES OF THE STUDY</b>	<b>18</b>
<b>MATERIALS AND METHODS</b>	<b>19</b>
<b>RESULTS AND DISCUSSION</b>	<b>29</b>
<b>CONCLUSION</b>	<b>42</b>
<b>REFERENCES</b>	<b>43</b>
<b>APPENDIX</b>	

## LIST OF FIGURES

FIGURES		PAGE
<b>Figure 1</b>	Map of human mitochondrial genome and the control region.	7
<b>Figure 2</b>	Phylogenetic position of the eastern-African M1 clade within the human mtDNA phylogeny.	11
<b>Figure 3</b>	Agarose gel electrophoresis showing the products of the extraction process.	31
<b>Figure 4</b>	Agarose gel electrophoresis showing 310 bp PCR product for mtDNA coding region fragment amplified using primer 10291-F and 10556-R.	33
<b>Figure 5</b>	Map of <i>Alu</i> I digestion for mitochondrial DNA coding region fragment amplified using primer 10 291-F and 10 556-R.	35
<b>Figure 6-10</b>	Agarose gel electrophoresis showing digested product of <i>Alu</i> I digestion for mtDNA coding region fragment amplified using primer 10291-F and 10556-R.	36-40

## LIST OF TABLES

<b>TABLES</b>		<b>PAGE</b>
<b>Table 1</b>	Sequence of each primer used in RFLP analysis.	22
<b>Table 2</b>	Master Mix for Restriction Digestion Analysis.	27
<b>Table 3</b>	Results of restriction digestion process of mtDNA coding region fragment and its classification.	41

## LIST OF ABBREVIATIONS

<b>DNA</b>	deoxyribonucleic acid
<b>mtDNA</b>	mitochondria DNA
<b>RFLP</b>	restriction fragment length polymorphism
<b>MLP</b>	multi-locus probe
<b>SLP</b>	single-locus probe
<b>PCR</b>	polymerase chain reaction
<b>ATP</b>	adenosine triphosphate
<b>RNA</b>	ribonucleic acid
<b>tRNA</b>	transfer RNA
<b>rRNA</b>	ribosomal RNA
<b>CR</b>	coding region
<b>HVS-I</b>	hypervariable segment I
<b>HVS-II</b>	hypervariable segment II
<b>HVS-III</b>	hypervariable segment III
<b>bp</b>	base pair
<b>kb</b>	kilobase
<b>np</b>	nucleotide position
<b>STR</b>	short tandem repeat

## ABSTRACT

The evolution of the human mitochondrial genome is characterized by the emergence of ethnically distinct lineages or haplogroups. Mitochondrial DNA (mtDNA) haplogroups were identified on the basis of the presence or absence of a restriction enzyme polymorphism in the coding region along with the sequence polymorphisms in the control region. This study was undertaken to analyze the mtDNA polymorphism in random population of Chinese ethnic of Peninsular Malaysia. A total number of 44 buccal cells samples were collected and subjected to extraction, followed by amplification by PCR and digestion process through restriction fragment length polymorphism (RFLP) analysis. *Alu* I restriction enzyme was used to detect polymorphism at nucleotide position 10397 in the mtDNA coding regions. Result shows that out of 44 samples, 21 of them were identified to be having an *Alu* I site gain at nucleotide position 10397, which classified them under the superhaplogroup M.



# INTRODUCTION

The invention of DNA fingerprinting by Sir Alec Jeffrey in 1985 (Schneider, 1997) has had a tremendous impact in forensic science. It leads a pathway in the investigation of criminal cases. Ever since, the technology and development in DNA analysis expand and become very robust and significant in forensic field.

## **Forensic DNA Analysis**

Earlier, the individualization of biological evidences of human origin was done using the blood typing system. However, this conventional analysis depends upon the nature of the biological material and its availability. In contrast, DNA profiling does not rely on the nature of the material, as the entire genetic information is contained in every single somatic cell of an individual (Schneider, 1997). Thus, DNA can be extracted from various biological specimens such as blood, semen, hair, teeth and saliva although within a minute amount. The degree of accuracy of DNA profiling is as high as 99.9%. As for that reason, results from DNA profiling are highly reliable and has been widely accepted in court of law as evidence.

Besides its ultimate contribution in solving criminal cases, DNA profiling also plays an important role in human identification, for instance, in missing person investigation and identification of mass disaster victims such as in Tsunami tragedy. Implementation of computerized system called DNA Databases that contain DNA profiles make solving of criminal cases and human identification possible and efficient.

## **Restriction Fragment Length Polymorphism (RFLP)**

The first DNA technique adapted in forensic analysis was restriction fragment length polymorphism (RFLP), which was discovered by Sir Alec Jeffrey in 1984. RFLP is used for detecting variation at the DNA sequence level, which occur both in the coding and non-coding regions of DNA. The term polymorphism refers to the physical basis of the differences lies in the nucleotide sequences in the DNA molecule, which reveal the polymorphic site of the particular sequence.

The principle behind RFLP detection relies on the possibility of comparing band profiles generated after restriction enzyme digestion in DNA molecules of different individuals. Some of the most common restriction enzymes used in the RFLP analysis are *Alu I*, *Dde I*, *Bam H I*, *Hin C II*, *Hae II* and *Hae III*. Diverse mutations that occurred affect DNA molecules in different ways, producing fragments of variable lengths. These differences in fragment lengths can be seen after gel electrophoresis, hybridization process and visualisation.

RFLP technique offers a high power of discrimination when using multi-locus probes (MLP). However, this method requires relatively large amount of high molecular weight DNA, with an average fragment size of about 20 kb to 25 kb. Since most of the sample collected from the crime scene were very minute in amount and sometimes are severely degraded, MLP often failed to produce reliable results and also difficult to interpret. To overcome this problem, single locus probes (SLP) based DNA analysis was introduced, which simplified the interpretation in RFLP analysis.

## **Polymerase Chain Reaction**

The polymerase chain reaction or PCR is an innovative technique to increase the amount of a specific sequence of DNA in a particular sample. This amplification technique, which enhanced the usefulness of DNA profiling, was invented by Kary Mullis in 1986. The principle is based on the process of DNA duplication during cells division.

Amplification of the desired target region of DNA sample is done enzymatically. The enzyme, *Taq* polymerase produces millions to billions copies of the selected DNA region within a few hours in a thermal cycler. The components of PCR comprise of DNA template, a set of primers that bind to the flanking region of the target region, *Taq* polymerase, nucleotides (dNTPs) and buffer to provide suitable environment for DNA polymerase. Three steps are involved in one cycle of PCR; denaturation, annealing and extension, which is usually repeated between 25-30 times.

PCR-based DNA typing system have made it possible to analyze DNA obtained from only a few cells as well as from highly degraded human remains. This has been demonstrated in cases such as the identification of the remains of Josef Mengele and the Romanov family (Schneider, 1997).

## **Human Mitochondrial DNA**

Human DNA consists of molecules that carry genetic information, which establish each person as separate and distinct from each other. DNA can be divided into nuclear DNA and mitochondria DNA (mtDNA). Nuclear DNA, which is found in the nucleus, is the product of DNA from both father and mother, while mtDNA that is located in the mitochondria is inherited only from the mother.

Mitochondria are subcellular, semi-autonomously functioning organelles that contain an extrachromosomal genome that is separate and distinct from the nuclear genome (Carracedo *et al.*, 2000). It contains a resident genome that undergoes replication, translation and transcription of their own. The primary functional role of mitochondria is to provide cells with the large bulk of adenosine triphosphate (ATP), synthesized through a process known as oxidative phosphorylation. ATP is used as an energy source to drive cellular reactions.

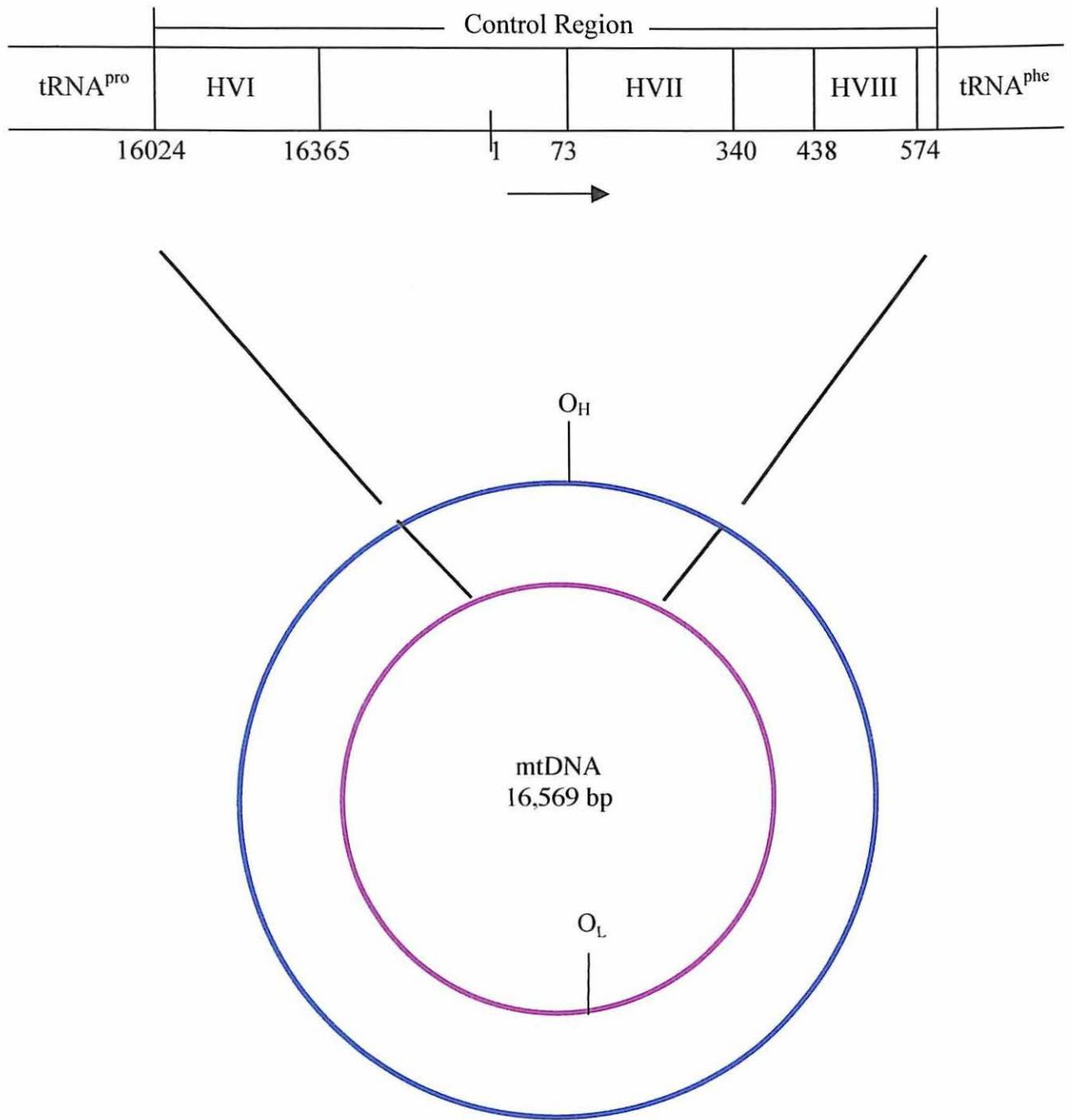
As mentioned earlier, mtDNA is maternally inherited. Shortly after fertilization, all mtDNA molecules in sperm degraded leaving only mtDNA from the mother in all cells of the offspring. Therefore, excluding mutation, mtDNA sequence in maternally related individuals should be identical since the transmission of mtDNA is consistent across many generations. The mtDNA genome has been completely sequenced by Anderson and his co-researchers in 1981 (Anderson *et al.*, 1981).

The genome of human mitochondrion is a closed, double stranded circular DNA molecule. It consists of the pyrimidine-rich strand and purine-rich strand, also

known as light and heavy strands respectively. Human mtDNA is approximately 16,569 base pairs (bp) in length and comprises of the coding and non-coding region.

The mtDNA coding region carries 37 genes, in which 22 of the genes encode transfer RNA's (tRNA), 2 genes encode for ribosomal RNA (rRNA) and 13 genes encode for protein enzymes involved in electron transport chain of oxidative phosphorylation for ATP synthesis. The non-coding region or control region (CR) is approximately 1,100 bp in length and situated between the mitochondrial tRNA<sup>pro</sup> and tRNA<sup>pnc</sup> genes (Figure 1). This region is sometimes called as the displacement loop (D-loop), where sequence variation between individuals is found (Wilson *et al.* 1993).

Nucleotide positions in the mtDNA genome are numbered according to the Anderson Reference Sequence (Butler and Levin, 1998). The numerical designation begins at the origin of replication of the heavy strand and continues around the circle (Wilson *et al.*, 1993). There are three hypervariable segments in CR. Hypervariable segment I (HVS-I) region ranges from position 16,024 to 16,365, hypervariable segment II (HVS-II) region extends from position 73 to 340 and hypervariable segment III (HVS-III) is situated between positions of 438 to 574. These boundaries are not rigidly defined and vary among particular studies or laboratories (Holland and Parsons, 1999). HVS-I and HVS-II regions are found to have the highest degree of variation in mtDNA.



**Figure 1:** Map of human mitochondrial genome and the control region.

- : Pyrimidine-rich strand (light strand)
- : Purine-rich strand (heavy strand)
- O<sub>H</sub> : Origin of replication of heavy strand
- O<sub>L</sub> : Origin of replication of light strand
- : Direction of nucleotide numbering in mtDNA

## **Advantages of mtDNA in Forensic DNA Analysis**

The great advantage of mtDNA in forensic profiling is its high copy number per cell. Compared to nuclear DNA that has only a single copy per cell, mtDNA present in hundreds or thousands of copies per cell (Wilson *et. al*, 1993; Holland and Parsons, 1999). This feature increases the reproducibility of mtDNA from very limited biological samples or even from highly degraded samples. Thus, mtDNA from bone fragments, teeth, hair shafts or even human faeces (Butler and Levin, 1998) could be successfully extracted, in addition to blood, saliva and semen samples that are normally being used.

Another advantage of mtDNA as a powerful forensic tool is its maternal inheritance. Due to its unique mode of inheritance, mtDNA does not undergo recombination. Thus, mtDNA sequence for siblings and all their maternal relatives should be identical, unless if mutation occurred. One case that proved the utility, strength and reliability of mtDNA analysis was the identification of the Romanov family. In addition to STR analysis, which matches the Tsarina to her children, an exact mtDNA sequence match was obtained between Tsarina with Prince Philip of United Kingdom, whose maternal grandmother was the Tsarina's sister (Butler and Levin, 1998).

In general, human mitochondrial DNA has become a meaningful tool in forensic investigation. The maternal inheritance feature of mtDNA and its high copy number are two important characteristics that enable tracking of families, population, and human identification as well in criminal investigation.

## **Application Of mtDNA In Population Studies**

The analysis of mtDNA in relation to its unique properties; the high copy number, the maternal inheritance, the lack of recombination and the high mutation rate, has been a potent method in studying the human population history and evolution. For instance, the mtDNA uniparental mode of inheritance enables researchers to trace related lineages back through time, highlighting the maternal ancestry of a population, without the confounding effects of biparental inheritance and recombination inherent in nuclear DNA (Pakendorf and Stoneking, 2005). The classification of mtDNA haplogroups is based on information gained from RFLP analysis of the coding region and the sequencing of HVS-I and HVS-II region.

According to Cann *et al.*, (1987), mtDNA data in population studies add more knowledge to the history of human gene pool. mtDNA gives a magnified view of the diversity present in human gene pool because mutations accumulate in this DNA several times faster than in the nucleus. Since mtDNA is inherited maternally and does not recombine, it is an important tool for relating individuals to one another and finally. There are about 1016 mtDNA molecules within a typical human, which are usually identical to one another.

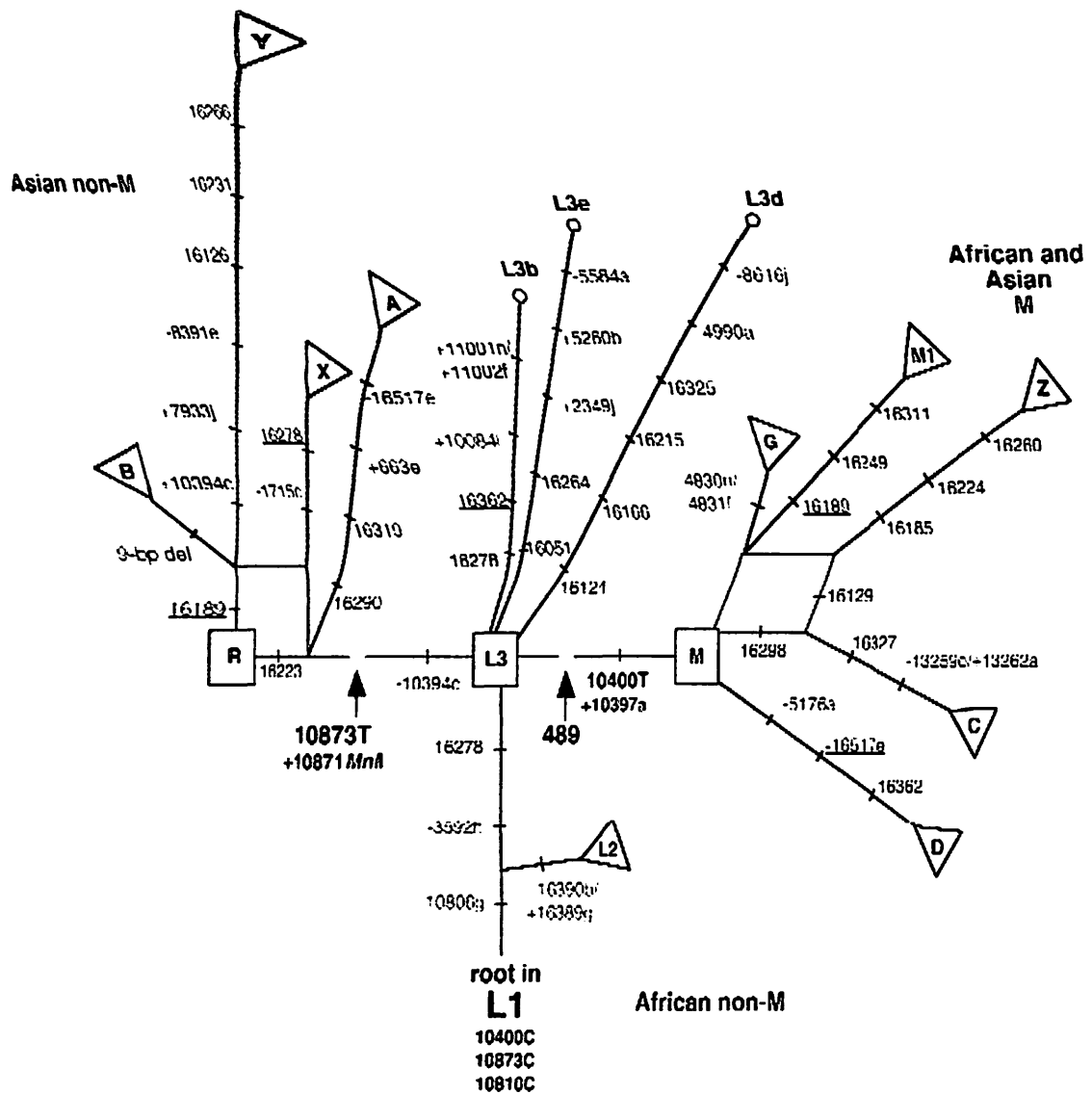
There are two basic approaches to the use of mtDNA in studies of human evolution: the lineage-based approach and the population-based approach (Pakendorf and Stoneking, 2005). The lineage-based approach attempts to unravel the history of mtDNA lineages, called haplogroups, while the population-based approach attempts to study the prehistory of individual populations of geographical regions or of



population migrations by using human population groups as the unit of study and applying population genetic methods to the data.

Based on the molecular analyses, most of human mtDNA sequence variations had accumulated sequentially along radiating maternal lineages from sets of mtDNA founders, during and after the process of human colonization of different geographical regions of the world (Torroni, 2000). For this reason, mtDNA haplogroups are often found to be geographically or ethnically specific.

Based on previous research and study, there are three African, nine European and seven Asian mtDNA haplogroups have been identified. Almost all African mtDNA's fall into haplogroup L, which can be further divided into L1, L2 and L3 (Figure 3). The haplogroups that can be found in European populations are H, I, J, K, T, U, V, W and X, whereas people in Asia are mostly belong to haplogroup A, B, C, D, E, G and M, which is the major haplogroup in Asia.



**Figure 2:** Phylogenetic position of the eastern-African M1 clade within the human mtDNA phylogeny. The network, rooted in the African haplogroup L1, is composed of most parsimonious trees of L3 based on ancestral combined RFLP and HVS-I types (disregarding the additional positions 489 and 10,873; of (i) the nested super-haplogroups R, M, L3 (indicated by squares); (ii) the clade M1; (iii) the East-Asian/Native-American haplogroups A, B, C, D, G, X, Y, Z; and (iv) the African haplogroup L2 (all indicated by triangles) (Quintana-Murci *et al.*, 1999).

## **Chinese Population In Peninsular Malaysia**

Malaysia comprises of Peninsular Malaysia (West Malaysia) and the states of Sabah and Sarawak (East Malaysia). Peninsular Malaysia is separated from the East Malaysia by the South China Sea and neighbouring with Thailand (north) and Singapore (south). Current population of Malaysia is estimated at around 26 million, of whom 83 percent live in the Peninsular Malaysia. The Chinese who are the second larger ethnic in Malaysia form about 30 percent of the population. The Chinese are the descendents of Chinese immigrants from China who arrived between the 15<sup>th</sup> and 19<sup>th</sup> century. Besides speaking in Malay and English, they also speak multiple Chinese dialects. There are three sub-linguistic groups who speak a different dialect of the Chinese language. The Hokkien-speaking group lives predominantly in the island of Penang, the Cantonese in Kuala Lumpur and the Mandarin-speaking group lives predominantly in Johor. Most of the Chinese practice Buddhism, some practice Taoism and also Christian. The Chinese are known for their diligence and keen business sense.

## REVIEW OF LITERATURE

According to Pakendorf and Stoneking (2005), haplogroups represent related groups of sequences that are defined by shared mutations, which tend to show regional specificity. It is known that mtDNA sequence variation has accumulated sequentially along radiating maternal lineages from sets of mtDNA founders during and after the process of human colonization. Numbers of studies had been done to determine the relationship of superhaplogroup M with human population of different geographical regions of the world.

Tanaka *et al.*, (2004) in their effort to construct the East Asia mtDNA phylogeny, had stated that the superhaplogroups M and N were the two early branches that radiated extensively out of Africa. Both superhaplogroups were originated from the L3 African trunk, which was also agreed by Pakendorf and Stoneking (2005) as well as other researchers. They also stated that all the mtDNA lineages detected in Old World populations belong to one of two M and N superhaplogroups with only secondary representatives in Africa. Apart from that, Tanaka and his colleagues concluded that superhaplogroup M was the older exit by which the radiation age was around 30,000 to 58,000 years ago.

Study of Asian and Papuan mtDNA evolution by Forster and his co-researchers in 2001 has showed that the Africa's ancestral mtDNA migration age is about 54,000 years. They proposed that superhaplogroup M migration splits into proto-Papuan and proto-Eurasian. The proto-Papuan M expanded demographically and geographically along the southern route until reaching Papua New Guinea,

allowing the Papuans to retain their overall genetic similarity to the Africans. Meanwhile, the proto-Eurasian mtDNA took 20 or more millenia to be drifted genetically.

Schurr and Wallace (2002) further suggested that the ancient lineage originated from the East Africa had dispersed into East Asia by way of Indian subcontinent with a diverse array of haplotypes evolving in Southeast Asia. Due to this dispersal pattern, superhaplogroup M has been part of the initial expansions of modern human groups into Southeast Asia. As a matter of fact, all Southeast Asian populations belong to superhaplogroup M with varying frequencies range between 25-45% with the highest frequency occurring in the Malays and Sabah aborigines (~60%). Schurr and Wallace believed that the Malaysian and Sabah aborigine populations displayed some unique mtDNA clusters, which were absent in other populations. These populations generally inhabiting the interior areas of the peninsula and islands of the region.

In addition, the highland populations of Melanesia show predominantly M mtDNAs with the coastal populations of these areas having mostly non-M haplotypes (Schurr and Wallace, 2002). Based on this distribution, it was suggested that the earliest population who brought haplogroup M mtDNAs to various parts of Southeast Asia and Melanesia initially settled the interior portion of these areas. The latter arrival of the Austronesian populations, who possessed the non-M haplotypes settled the coastal areas and simultaneously pushed the existing populations further into the interior.

Kivilsid and his co-researchers had performed an Indian mitochondrial DNA variants study in 1999. In their study, the modern Indian populations were found to share common clusters that were specific for both eastern and western Eurasian population. The haplogroups M and U constitute 75% of the variation and the spread is uniform all over India. The sub-branching of haplogroup M in India is profoundly different from that of other Asian populations. There were five novel sub-clusters that form half of the Indian haplogroup M lineages, namely M1, M2, M3, M4 and M5. These five major Indian-specific sub-clusters are not represented to any significant extent elsewhere in Asia. Moreover, the spread of haplogroup M variants revealed some characteristics differences among different population of the Indian peninsula. This has been proved by the presence of sub-cluster M2 as the predominant sub-cluster among the people living in the southern part of the peninsula.

Apart from that, mtDNA RFLP polymorphisms in India have shown that haplogroup M divergence predates the separation of proto-Indians from proto-Eastern Asians. As for this reason, Kivilsid *et al.*, (1999) had concluded that during the past 50,000 years there has been a very limited admixture of Indian populations with the Mongoloid populations living east and north of India.

Kivilsid and his colleagues also stated that haplogroup M can be further divided into discrete sub-clusters according to the accumulation of synapomorphic mutations along the Asian maternal lineages. Sub-clusters C, D, E and G defined by certain RFLP and HVI sequence polymorphisms are spread over vast territories all over Mainland Asia. The split of haplogroup M into these sub-clusters may therefore

reflect a secondary expansion event, which led ultimately to the extension of modern Asian populations to northern and central Asia and to the Americas.

In Ballinger *et al.*, (1992), haplogroup M was defined as having only the *Dde* I site gain at nucleotide position (np) 10394, whereas both haplogroups E and F had the *Dde* I and *Alu* I site gain at np 10394 and 10397 respectively, with these latter two differing by the presence or absence of the *Hae* III 16517 site. This definition was similar to the definition given by Chen *et al.*, (1995) by which the haplogroup M represents the mtDNA possessing the +*Dde*I/+*Alu*I sites.

Upon studying several previous analyses of mtDNA haplotypes by Kivilsid *et al.*, (1999) and Quintana-Murci *et al.*, (1999), Schurr and Wallace (2002) came to the conclusion that haplogroup M is actually a macrohaplogroup. This means that M represents the founding or stem haplogroup from which all subsequent haplogroups bearing the *Dde* I 10394 and *Alu* I 10397 sites evolved. Therefore, any mtDNA with the +*Dde* I/+*Alu* I sites can be said to belong to this macrohaplogroup. Hence, haplogroups C, D, E, G and Z can be considered as smaller branches of haplogroup M.

It was also stated by Ballinger *et al.*, (1992) that the combined *Alu* I at np 10397 and *Dde* I at np 10394 sites essentially split all haplotypes within the Asian populations into two major clusters. Several haplotypes had only the *Dde* I site, which creates a semisite for *Alu* I at np 10397. Thus the *Dde* I site is not only necessary but precedes the creation of the *Alu* I site at np 10397. Apart from that, the overlapping *Alu* I and *Dde* I sites at np 10397 and 10394 appeared to be ancient mutations. The

*Dde* I site has been found in the mtDNAs from every racial group and is present in the most divergent African haplotypes.

The general overview of the pattern of mtDNA diversity in Southeast Asia allows several inferences about the population history. The distribution of superhaplogroup M and other haplogroups as well as the pattern of haplotype sharing by the population that inhabit this region, reflects the process by which it was settled. The antiquity is evident by its estimated age and the many unique clusters that are present in Southeast Asian, Papua New Guinea and Melanesian groups, particularly in ethnic groups representing some of the first populations to enter Asia.