DEVELOPMENT OF MITOCHONDRIAL DNA TYPING METHOD FOR HUMAN IDENTIFICATION

SERI MIRIANTI BINTI ISHAR

UNIVERSITI SAINS MALAYSIA

2018

DEVELOPMENT OF MITOCHONDRIAL DNA TYPING METHOD FOR HUMAN IDENTIFICATION

by

SERI MIRIANTI BINTI ISHAR

Thesis submitted in the fulfillment of the requirement

for the degree of

Doctor of Philosophy

May 2018

ACKNOWLEDGEMENT

In the name of Allah, the most gracious and most merciful. First and foremost, my highest gratitude belongs to the Almighty Creator for making this thesis a reality. I would like to express my deepest appreciation to main supervisor Assoc Prof Dr. Zafarina Zainuddin and co-supervisor Prof Dr. Norazmi Mohd Nor for the opportunity, guidance, support, patience and understanding during years of my research progressions. Thanks also to the Ministry of Higher Education and The National University of Malaysia (UKM) for sponsoring my studies under SLAB program. Thank you Universiti Sains Malaysia for funding this research under short term grant (Grant No: 304/PPSK/61313150). My gratitude also goes to IPharm for allowing me to use laboratory to complete labworks. To people in IPHARM, Dr. Adel, Dr. Wai Kuan, Hadi, Yana, Myra, Nan, Faiz, Zul, Izan, Muzed and others, I appreciate your companionship and support. Thanks also to all Forensic labmates in USMKK especially K.zie, Nany, Waliy and Alia and who have contributed directly or indirectly towards my studies.

My sincere gratitude also goes to my dear husband Talib, Mak and Ayah for their continuous love, support, encouragement and understanding, as well as my siblings and other family members. Last but not least to my little sweethearts, Nadim and Layla, I appreciate your patience and all the love you cherish me with.

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AD	*Anno Domino* (Year of the Lord)
ASPs	Allele specific primers
asPCR	Allele specific PCR
bp	Base pair
BP	Before present
DNA	Deoxyribonucleic acid
ddH_2O	Double distilled water
dNTP	Deoxyribonucleotide triphosphate
EDTA	Ethylenediaminetetraacetic acid
FTA	Flinders Technology Associates
g	Gram
H_2O	Water
HCl	Hydrochloric acid
HV1	Hypervariable I
HV2	Hypervariable II
km	Kilometer
mg	Milligram
ml	Milliliter
mtDNA	Mitochondrial DNA
ng	Nano gram
PCR	Polymerase Chain Reaction
pH	Potential of hydrogen
SNP	Single Nucleotide Polymorphisms

TAE	Tris – Acetate EDTA
Taq	Thermus aquaticus
TBE	Tris – Borate EDTA
TE	Tris - EDTA
T_m	Melting temperature
µl	Microliter
µM	Micromolar
U	Unit
vtASP	Variant type allele specific primer
wtASP	Wild type allele specific primer

PEMBANGUNAN METODOLOGI PENJENISAN DNA MITOKONDRIA UNTUK IDENTIFIKASI MANUSIA

ABSTRAK

Pengesanan mutasi dalam DNA mitokondria (mtDNA) amat berguna dalam laporan perubatan, analisa genetik molekular, variasi *haplogroup*, carian forensik dan pelbagai aplikasi lain. Dalam projek ini, Single Nucleotide Polymorphisms (SNPs) dalam kawasan pengekodan dan kawalan pada mtDNA telah dipilih untuk membangunkan kit DNA. Objektif utama penghasilan kit ini adalah untuk menyediakan kit yang ringkas, kos efektif serta tahan lasak bagi tujuan identifikasi manusia. Tiga puluh SNPs yang spesifik kepada *haplogroup* Asia Tenggara seperti M, B, F, E dan N telah dipilih. Polymerase Chain Reaction (PCR) pusingan pertama dijalankan dengan menggunakan 25 primer yang telah direkabentuk. Produk amplifikasi daripada PCR pusingan pertama yang telah ditulenkan digunakan dalam PCR pusingan kedua atau PCR alel spesifik (asPCR) dan juga penjujukan DNA. Dalam asPCR, 2 jenis primer alel spesifik (ASP) telah direkabentuk iaitu *wild type ASP* (wtASP) dan *variant type ASP* (vtASP) untuk identifikasi variasi yang terdapat dalam fragmen sasaran. Dua puluh ASP telah direkabentuk dan berjaya mengamplifikasi SNPs sasaran. Kawalan positif dalaman PCR (IPC) telah direkabentuk bertujuan untuk kawalan kualiti dan mengelakkan keputusan negatif palsu atau positif palsu. Kajian validasi telah dijalankan dengan menggunakan sampel manusia (Melayu, Cina dan India), mamalia, serangga dan beberapa jenis bakteria. Kajian sensitiviti dijalankan dengan menggunakan sampel DNA daripada tulang purba

dan sampel DNA manusia yang telah dicairkan secara bersiri. Tiada produk amplifikasi diperoleh apabila ASP digunakan bersama template DNA daripada spesis lain dan ini menunjukkan ia adalah spesifik kepada DNA manusia sahaja. Hasil daripada ujian sensitiviti yang telah dijalankan, sebanyak 1ng/µl templat DNA genomik diperlukan bagi amplifikasi optimum PCR pusingan pertama. Berdasarkan keupayaan primer yang direka bentuk untuk amplifikasi kedua-dua pusingan PCR dan keputusan kajian validasi, kit penjenisan yang ringkas serta efektif untuk identifikasi manusia telah berjaya dihasilkan dalam kajian ini. Kit ini amat berguna untuk mengecilkan skop pengkelasan individu terutama dalam kes-kes seperti bencana alam, jenayah forensik serta nahas kapal terbang.

DEVELOPMENT OF MITOCHONDRIAL DNA TYPING METHOD FOR HUMAN IDENTIFICATION

ABSTRACT

Detection of mutations in human mtDNA is useful for the purpose of medical reports, molecular genetic analysis, haplogroup variations, forensic search and many other applications. In this project, Single Nucleotide Polymorphims (SNPs) from coding and control regions of mtDNA were selected for development of mtDNA typing kit. The main objective of this kit is to provide a simple, cost effective yet robust identification of human remains. A total of 30 SNPs that are specific for Southeast Asian haplogroups such as M, B, F, E and N were selected. The first round Polymerase Chain Reaction (PCR) was performed using 25 sets of in house designed primers. The purified amplified products were used in second round PCR known as allele specific PCR (asPCR) and also sent for sequencing to confirm the sequence of the amplified products. In asPCR, two types of allele specific primers (ASPs) for wild type ASP (wtASP) and variant type ASP (vtASP) were designed to identify the polymorphisms within the target fragments. A total of 20 ASPs were successfully designed and able to amplify the targeted SNPs. For quality control purpose and also to avoid false positive or false negative results in asPCR, an internal positive PCR control (IPC) was developed and incorporated together in the same tube during asPCR. A validation study for this mtDNA typing kit was carried out using human DNA samples (Malay, Chinese and Indian), mammals, insects and several types of bacteria. Sensitivity study was carried out using DNA extracted from ancient bones and serial dilution of fresh

human DNA sample. The ASPs are only specific for human DNA as no amplification was observed when using DNA template from other species. Based on the sensitivity validation study, the minimum amount of genomic DNA template required for optimum amplification of first round PCR is 1ng/µl. Based on the successful amplifications of both round PCRs using in house designed primers and the validation studies results, a simple yet effective mtDNA typing kit for human identification has been successfully developed in this study. This kit would be very much useful to narrow down individual classification especially in cases of mass disaster, forensic crime and plane crash.

CHAPTER 1

INTRODUCTION

1.1 Mitochondrial DNA and Single Nucleotide Polymorphisms

Determination of DNA variations contributes to classification and sometimes up to unique identification in human. The Short Tandem Repeats (STRs) and Single Nucleotide Polymorphisms (SNPs) are well known in the studies of human genetic variations. Presently STRs are preferable and provide large contribution in many forensic cases because of their variations at a unique physical location. STRs are a microsatellites, consist of a unit of two to thirteen nucleotides repeated hundreds of times in a row on the DNA strand while SNPs are a variations of a single nucleotide that occur at specific positions throughout the genome. Generally, the SNP which in majority are biallelic markers, are the most abundant forms of DNA polymorphisms observed in human genome and estimated to occur at 1 in every 1,000 bases in human genome (Glover *et al*., 2010).

Although the discrimination power of each SNP is lower to that of STR but the combination and collaboration of the mutation could realize a high discriminating power (Inagaki *et al*., 2004). The SNPs are in great interest currently as they could be used as markers to identify genes that predispose individuals to complex disorders by using linkage disequilibrium. The SNPs can also be applied in forensic genetics and therefore SNPs typing has been focused in this field (Asari *et al.*, 2009). The SNPs have several characteristics that are suitable for several studies especially forensics.

SNPs can be analyzed from short amplicons and this is desirable as almost all forensic samples are degraded (Köhnemann *et al*., 2008). Its low mutation is very precious for paternity testing and it is also suitable for analysis using high-throughput technologies, which have become important for the successful implementation of large population DNA databases and make it easier to perform large population studies (Sobrino *et al*., 2005).

Mitochondrial DNA (mtDNA) becomes interest of this study compared to nuclear DNA as there are 1,000-10,000 copies of mtDNA per cell. Because of this valuable characteristic, it is useful especially in cases of inadequate samples (Hoong & Lek, 2005). In the early years of forensic DNA typing, studies on human DNA are based on restriction fragment length polymorphisms (RFLP) that refers to the differences among samples using different locations of restriction enzymes sites in either genomic DNA or mtDNA. In addition, the circular shape of mtDNA provides stability to survive in extreme conditions which is useful in forensic investigation especially that involves archeological samples. Although mtDNA has little information compared to nuclear DNA as it is solely inherited from mother, the criteria still make it precious in matrilineal study.

In this study, the focus is on mitochondrial DNA SNPs (mtSNPs) due to their abundance throughout the entire mitochondrial genome. It has been shown that mtDNA replacement mutations lead to regional varieties in the mtDNA coding region (Liao & Lee, 2010). In addition, mtDNA coding regions' SNPs are involved in a broad spectrum of diseases. Polymorphisms of mtDNA genome have been reported to be useful for identity testing and analysis of degraded material or samples containing little

or no genomic DNA (Boonyarit *et al*., 2014). Using polymorphisms that are observed in mtDNA genome in relation to the reference sequence, individuals are categorized into monophyletic clades or haplogroups that represent related groups of sequences defined by shared mutation (Anderson *et al*.,1981)

Populations in the same haplogroup normally share their ancestor due to existence of similar haplotypes in the haplogroups. The combination of these alleles at different mtDNA regions caused them to closely link and tend to be inherited together. Due to this fact, it is usually possible to predict haplogroup from these kinds of haplotypes. The collaboration of mtSNPs able to produce specific identification for each population. Thus, selection of these SNPs may indicate population of individual hence giving group determination of individual even at early stage of investigation.

Selections of mtDNA SNPs for this study are based on haplogroups reported specifically for Southeast Asian populations. Simple haplogroup discrimination will not lead to sufficient results as majority of Asian populations belong to superhaplogroup M and N (Ricaut *et al*., 2009). Thus a subtyping of superhaplogroup M and N with further discriminating SNPs is inevitably necessary. The unspecific SNPs like hot-spot SNPs that mutate in different haplogroup background are not preferred as they cannot discriminate haplogroup specifically (Köhnemann *et al*., 2008). Haplogroup E is reported specifically belong to Southeast Asian populations especially Island Southeast Asia and therefore several SNPs in haplogroup E were selected to be incorporated in this mitochondrial DNA typing kit. Other SNPs in subhaplogroups such as M9, M9a, M9a'b, N21, R21, and F1a1a were also added to further enhance the discrimination power of this typing kit.

1.2 Problem statement

Screening of mtDNA SNPs is important nowadays since specific SNPs indicate specific haplogroups. In cases of mass disaster, identification of missing/detached limb/organs using autosomal STRs method can be very expensive. For this kind of cases, this kit may help in grouping the detached limbs or small bones and provides preliminary result in the proses of human genotyping. In addition, autosomal STRs cannot be used to narrow down the identification based on population especially when dealing with mass disaster victims and the technique also requires longer time for identification. (Alonso *et al*., 2005; Glover *et al*., 2010). On the other hand, mtDNA offers greater possibility of getting DNA profile from degraded or very limited DNA samples, which is not possible when using autosomal STRs (Boonyarit *et al*., 2014). The mtDNA typing kit developed in this study will offer a better option for human identification, which is lower in cost, not depending upon expensive high technology instruments, easy yet robust.

Nowadays there are several tools, kits and also techniques that can be used to detect human DNA for human identification. Most of these techniques used STRs as markers for individual identification. In 2004, Japan was hit by massive tsunami (Dawson *et al*., 2004) that have caused thousands of people died. As it attacked coastal area, not only local people were dead but tourists were also among the casualties (Løvholt *et al*., 2014). Corpse identification was a difficult process as many of the bodies were totally smashed during the catastrophic incident. Using normal procedures, classification of those victims was a long and tedious process as search and rescue team has to send DNA samples to laboratories for identification purpose

(Alonso *et al*., 2005). Identification is rather straight forward if family members are able to identify the body or willing to provide their DNA samples as reference. Problem will arise when no family member comes forward to identify the body.

Another example that required massive procedure for victim's identification was MH17 crash in Ukraine in 2014. The ill-fated aircraft with a total of 298 passengers on board was shot down, killing all 283 passengers and 15 crews on board (Samarasekera, 2014). As the plane fell down from 10,000 feet above sea level, the bodies of the victims were severely mutilated. The investigation team reported that they found part of human bodies scattered around within the marked 10 km square feet crime scene. The remains were collected and DNA analysis was carried out for identification purpose. In the beginning, grouping of the intact bodies or limbs or organs according to their population couldn't be performed. Overall, normal procedure of DNA typing in such case is sampling (collection of the intact bodies or fragments), obtain DNA samples from both victims and family members for reference purpose and the last process is DNA analysis by comparing the sample's DNA profile and reference profile. Not all the body parts collected from the crime scene were in good condition. Referring to this case, dealing with internal politic issues has delayed the collection process. The research and rescue teams were only allowed to access the crime scene after several days and therefore when dealing with degraded and small quantity of DNA, mtDNA is considered as the best solution. Based on these two cases discussed, the mtDNA typing kit developed in this study may offer a better option in narrowing down the detection process as it can be used to classify the bodies at the very first stage using a much faster and simple technique that can be brought to the crime scene. Furthermore, the cost is much lower as compared to other methods. Current DNA

typing techniques mostly cost around one thousand ringgit per sample, which may include the initial primer design up to sequencing process for human identification

1.3 Research objectives

The main objective of this project is to develop mitochondrial DNA typing method for the identification of Southeast Asian populations in forensics investigation.

The specific objectives of this project are:

i) To design and optimize primers for amplification of selected SNPs within the mitochondrial DNA control and coding regions.

ii) To design and optimize allele specific primers (ASP) against the amplified fragments.

iii) To design and optimize synthetic oligonucleotide as an internal PCR control in allele specific PCR (asPCR).

iv) To perform specificity validation of the developed kit across different animal species that includes mammals (primates and cat), insects, and bacteria.

v) To determine the sensitivity of the developed kit using degraded DNA samples (DNA extracted from ancient bones) and diluted DNA samples

vi) To validate the developed kit on Malaysian populations (Malay, Chinese and Indian).

1.4 Scope and limitation

Selection of SNPs sites that are specific for Southeast Asian populations is not quite straight forward as these polymorphisms may also represent Asian populations in general. Furthermore, some of the selected SNPs are shared by different haplogroups, which decrease the discrimination process. Increasing the number of SNPs in a way may help to minimize this limitation. Among all of the selected haplogroups, haplogroup E is specifically belongs to Island Southeast Asian and therefore combination of these polymorphisms with other SNPs may discriminate and classify the Southeast Asian populations better. On the hand, the amplified fragments are quite small for the purpose of forensic application. Due to this, the sequencing process may not be able to amplify the required SNPs because of noises production during sequencing for instance due to SNP that is closely located at the primer binding site.

1.5 Importance of study

The main objective of this research is to develop a robust, simple and cost effective mtDNA typing kit for application in forensic cases. Currently the available methods involve laborious techniques and high end instruments, which are not suitable when dealing with huge number of biological samples and most of the analysis cannot be carried out at the crime scene. The most popular DNA typing technique proposed and used nowadays is real time PCR, also known as quantitative PCR (qCPR). This technique could be specific and non-specific fluorescent dyes that intercalate with the DNA sequences. Although it is reported as an accurate typing technique, a specific

software is required to perform the analysis. Prior to qPCR, it involves similar DNA extraction method as any other DNA typing technique hence all the steps are conducted in laboratory and not practical to be carried out in the crime scene.

In order to shorter the analysis period, several new DNA typing techniques have been proposed that can be brought to the investigation field yet the robustness is maintained. The desire to take DNA testing capabilities out of laboratory to a crime scene has leaded the development of portable DNA testing devices. In certain cases, the developed rapid DNA device might aid elimination of innocent suspects early in an investigation. According to (Butler, 2012), much of work so far has been focused on miniaturizing the DNA separation steps and the focus is on STR typing.

A microchip based laboratory, or so called 'labs on a chip' is an application of microfabrication method into miniature to miniaturize the sample preparation and analysis steps in forensic DNA typing. It permits investigation of biological evidence at the crime scene or more rapid and less expensive DNA analysis in a conventional laboratory setting (Paegel *et al*., 2003). The miniature of capillary electrophoresis (CE) in this technique is shorter channels or known as capillaries which will lead to faster DNA separations. According to the report by (P. Liu *et al*., 2008), this CE device has been used to perform investigation onsite in a mobile van and the results managed to be produced within in less than six hours.

Robustness are measured by the reproducibility of the results and capability of this typing kit to be conducted on site. The optimized PCR programs and the designed primers that are specific to human DNA, must be able to produce consistent results

regardless being used in different laboratories and by different operators. With the availability of the portable PCR machine in the market such as Ahram Biosystems, AAS Inc, Biomeme and Biometra that are mostly battery operated, DNA identification process is able to be carried out on site. This will significantly reduce the turnaround time for analysis and the hassle of transporting the samples to laboratories.

However, all of the techniques previously proposed used STR instead of SNPs. Hence, the development of this mtDNA typing kit is important as it can be applied in both individual and population studies.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

The use of mitochondrial DNA instead of nuclear DNA in Forensic DNA analysis is due to several reasons such as quantity and conditions of the sample itself. Samples may be found from low-copy number materials such as shed hair or highly degraded bones due to severe exposure towards environmental conditions. The best way to deal with the problems is by using mtDNA that presents in high copy number within cells (Kline *et al*., 2005). The main purpose of this project is to develop a kit for identification of human remains using mtDNA single nucleotide polymorphisms (mtSNPs) specifically for Southeast Asian (SEA) populations. The selected mtDNA SNPs were the ones classified as the defining SNPs for the particular haplogroups predominantly found in SEA.

2.2 Mitochondrial DNA typing kit

The development of mtDNA SNPs typing kit is not new but the approaches, combination of methods in this kit and the target populations are different. In this kit, a total of 30 mtDNA SNPs within the control and coding regions were selected to be included in this kit. The methods include two rounds of PCR that amplify the templates in the first round PCR and identification of SNPs using allele specific PCR (asPCR).

A group of researchers in National Institute of Standards and Technology Gaithersburg had developed a typing kit using control and coding regions in mtDNA for screening purpose but they covered only eleven loci and specifically designed for Caucasian group. This typing kit was designed to detect mutations using probes attached to the allele specific primer extension (Kline *et al*., 2005).

Another typing kit using mtDNA SNP for screening purpose was developed with 16 targeted SNPs in hypervariable regions for European population but the approach was totally different. This research used mini sequencing method that relies upon a single base extension of a primer immediately adjacent to the SNP using fluorescently labelled ddNTPs prior to sequencing (Brandstätter *et al*., 2004). In 2008, the European DNA Profiling (EDNAP) Group developed the mtDNA typing kit that focused on 16 nucleotide positions in coding region. This typing kit has allowed the discrimination of major West Eurasian mtDNA haplogroups by using multiplex PCR and Snapshot miniseqencing (Parson *et al*., 2008).

A mitochondrial DNA Typing kit that has been developed by Salas *et al.* (2005) was almost similar to the developed kit in this study. The kit was designed to discriminate the nucleotide polymorphisms in haplogroup H, a common European haplogroup for 40-50% of the European populations. In this particular kit, the typing technique allowed to allocate common West Eurasian mtDNA haplotypes into their corresponding branches of mtDNA skeleton. The target mtDNA locations include HV1, HV2 and also coding region. The methods approached in this typing kit were multiplex PCR and also SNaPshot reaction with ddNTPs labeled with dye. From all mtDNA typing kits developed before, the trend of using real time PCR was preferred

and the analysis was based on the fluorescent color that has been used to label ddNTPs or primers. The asPCR method especially has not been used in mtDNA typing kits developed before.

2.3 Disaster victim identification (DVI)

Mass disasters in the form of natural or man-made, can involve loss of life for many victims of the tragedy. Efforts to identify these victims are referred to as disaster victim identification (DVI). Nowadays, DNA test has become routine and expected in DVI in the event of plane crash, large fire or terrorist attack. Normally, mass disasters will leave human remains in pieces or beyond recognition. Sometimes, body parts can be separated from one another and the remains degraded making identification without DNA techniques is quite impossible. The use of fingerprints and dental records still plays an important role in victim identification but the approaches require a finger or an intact skull or jawbone along with archived fingerprint and dental records that can be made available for comparison purposes.

DNA test has a major advantage that it can be used to identify each and every portion of the remains recovered from site provided two conditions that there is sufficient intact DNA present to obtain a DNA type and reference sample is available for comparison purpose. The comparison involves post-mortem (PM) and ante-mortem (AM) data. PM data are generated from the recovered human remains, which may be highly fragmented depending on the type of disaster. AM data come from either direct reference samples or kinship comparisons to biological relatives.

So far, DNA testing has been used to help in identifying victims of numerous airline crashes, the victims of terrorist attacks, recovered remains from mass graves and in natural disasters such as SEA tsunami that occurred in December 2004 and Huricane Katrina that struck New Orleans in August 2005. On 11 September 2001, the terrorist attacks against the World Trade Center (WTC) and Pentagon in Washington DC. More than 19917 pieces of human remains were collected from the tragedy. According to the report, the initial removal and sorting of human remains took place between September 2001 and May 2002. However, the primary DNA identification only went for more than 3 years. One of the largest challenges from this investigation was the process of reviewing the massive amounts of data produced by laboratories (Macqueen, 2009).

Other than that, DNA testing is also applied in individual identification from mass graves. One of the major challenges for performing DNA identification from mass grave is dealing with degraded samples. In this case, mtDNA often is the only source of successful DNA recovery from bones that have been in the ground for many years (Huffine *et al*., 2001). Again, dealing mass DNA samples will consume longer time and increase the cost. Based on this two cases, the developed kit is strongly relevant to be applied in this type of cases especially that involved massive number of victims. DNA identification is directly performed on site with less time consuming yet cost effective (Butler, 2012).

2.4 Selection of haplogroups

Knowledge on mtDNA haplotypes becomes crucial and important in application of mtDNA markers especially in forensic studies. The mtDNA haplotypes strongly correlate to geographic origin (Maruyama *et al*.,, 2010). Even it is seldom preferred in individual unique identification, screening of several mtDNA regions might narrow down the scope of investigation in cases that involved unclassified populations.

A total of 30 mtDNA SNPs from both coding and control regions have been chosen in this study. All the selected mtDNA SNPs represented major haplogroups of SEA based on previous reports (Kong *et al*., 2006; Macaulay *et al*., 2005; M. S. Peng *et al*., 2010). In this study, the selected mtDNA SNPs from coding regions were SNP 13,626, 3552, 1709, 1719, 9080, 3394, 13362, 4491, 3027, 3705, 9512, 7684, 1872 and 8440 while the mtDNA SNPs within the control region were SNP at nucleotide position 16390, 16261, 16266, 16355, 16335, 16291, 16093, 16108, 16274, 153, 146, 16148, 16309, 16287, 195 and 479. The major haplogroups involved were macrogroup M and N together with their branches and subclades of E, R, B, and F (Kong *et al*., 2006).

The nucleotide positions in mitochondrial genome were numbered from 1 to 16569 according to the revised Cambridge Reference Sequence (rCRS) (Bandelt *et al*., 2014). All single nucleotide polymorphisms (SNPs) including insertions or deletions (indels) were scored as differences (mutations) to rCRS. As the previous research were mainly focused on the HV1, HV2 and HV3 regions of mtDNA, this study will focus

on both mtDNA regions. Although most of the previous studies have focused on control regions, there was a study that investigated the complete control region and some specific coding region polymorphisms. This research team has conducted the study to obtain more reliable indicators of mtDNA haplogroups and compare with other datasets on SEA, East Asian and neighboring populations (Maruyama *et al*., 2010).

The non-coding region (also known as control region) is effectively shorter than the coding region with 1.1 kb that covers bases from 16024 to 576 (Figure 2.1). The mutation rate that for these two regions is not equal across the entire mtDNA. It was reported that the overall mutation rate highly occured in the control region as compared to coding region. The control region is a non-coding region that is important for mtDNA replication thus accumulates more mutations. Several types of local differences in the mutation rate among all these nucleotide positions can be observed such as hot spot SNPs or known as SNPs that are prone to mutation and site-specific (hyper) mutability. Among these nucleotide positions, several sites act as mutational hot spots in the control region such as positions 146, 150, 152, 195, 16189, 16311,16362 and 16519 whereas others appear rather stable such as position 477, 493, 16108 and 16129.

Previous reports recorded that non coding region consists of higher overall mutation rate but the mechanism that boosted up this event remains to be elucidated. Due to higher mutation rate, the sequence variation can be easily observed in control region and preferred by the researchers for various applications. But still, the determination of some haplogroups cannot be depended on control regions only even

though there were haplogroups assigned solely based on control region data such as J1b1 and K19a. However, it's more reliable if the data from control regions and coding region are combined together (Oven & Kayser, 2008).

As in control region, the coding region also contains of nucleotide positions that are prone to mutate which will lead to homoplasy or also known as recurrence in phylogeny. The reported nucleotide positions include 709, 1719, 3010, 5460, 10398, 11914, 13105, 13708 and 15884.

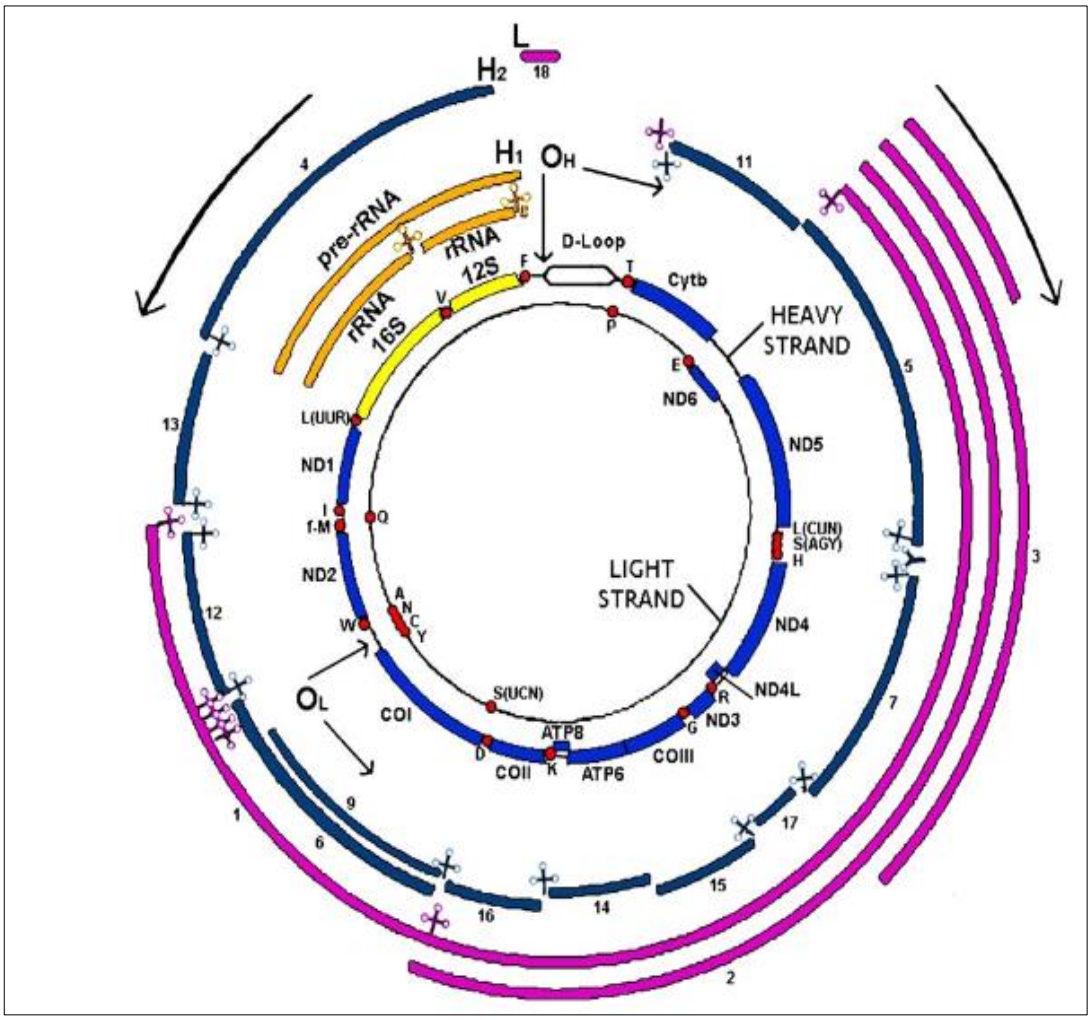

Figure 2.1: Map of human mitochondrial DNA (Montoya *et al*., 2006)

Regarding haplogroups classification, a group of researchers has suggested that climate was the most important factor that influence human mtDNA evolution (Ruiz-Pesini *et al*., 2004) but the issue has been challenged by other investigators stated that climate was not the only factor that determine human mtDNA evolution. Other than climate, population immigration and genetic mutation were claimed as factors that contributed to human mtDNA evolution (Sun *et al*., 2006). Asian populations were mainly belonging to M, N and R macrohaplogroups. Originally, the M and R macrohaplogroups derived from macrohaplogroup L which was reported as root of human mtDNA phylogenetic trees. As such, macrohaplogroup L represented the most ancestral mitochondrial lineage of all currently living modern humans and it was reported as African origin of modern humans. The major subclades listed in macrohaplogroup L were L0, L1, L2, L3, L4, L5, L6 and L7. The non-Africans were all reported exclusively descended from haplogroup L3 that expanded to superhaplogroup M and N. The C, E, G, Q and Z were the haplogroups under M macrohaplogroup. The expansion from N macrohaplogroup were A, I, O, R, S, W, X and Y. The R macrogroups later expands to B, F, J, P, T, HV, H, V and K haplogroups (van Oven & Kayser, 2009). Figure 2.2 previews map of world human mtDNA migrations and each continent specific haplogroups.

The L3 group radiated out of Africa in the form of macrohaplogroup M and N around 60kya (thousand years ago) and reported to enter South Asia and India sub-continent and later drifted to SEA and Australia (Figure 2.2). But actually there was an issue and controversy surrounding the origin of this haplogroup. Some of previous researchers believed that M haplogroup was originated in Asia and represented a backflow to Africa while other researchers believed from their researches that M

haplogroup was truly originated from Africa and then migrated to Asia (Winters, 2010). The report of M haplogroup was originated from Africa has been supported by several recent studies which strengthen the statement (Maji *et al*., 2009; Malhi *et al*., 2007; M. S. Peng *et al*., 2010; Ricaut *et al*., 2009; Winters, 2010). Other reports suggested that East Asian mtDNA pool was locally region-specific and mostly covered by superhaplogroup M and N (Kivisild *et al*., 2002).

Amongst all branches of macrohaplogroup M, group M9 is well distributed in Asian origins. Most basal branches of M9a except M9a1 was reported following the distribution in Southern China and SEA. This pattern suggested that M9a might have a southern origin (M.-S. Peng *et al*., 2011). Apart from haplogroup M9, group M7 also has been found in East and SEA (S. D. Lee *et al*., 1997). According to (Maruyama *et al*., 2010), the M12, M21b and M22 lineage has been found in SEA populations with high percentage in Aboriginal Malay that brought the suggestion of having those haplogroups as Malay markers. Haplogroup E that was rooted out from haplogroup M9 was reported as exclusively belong to Island SEA. The branches of haplogroup E were determined using control and coding region mutations as discrimination between groups in haplogroup E was not possible in some cases using control regions alone (Soares *et al*., 2008).

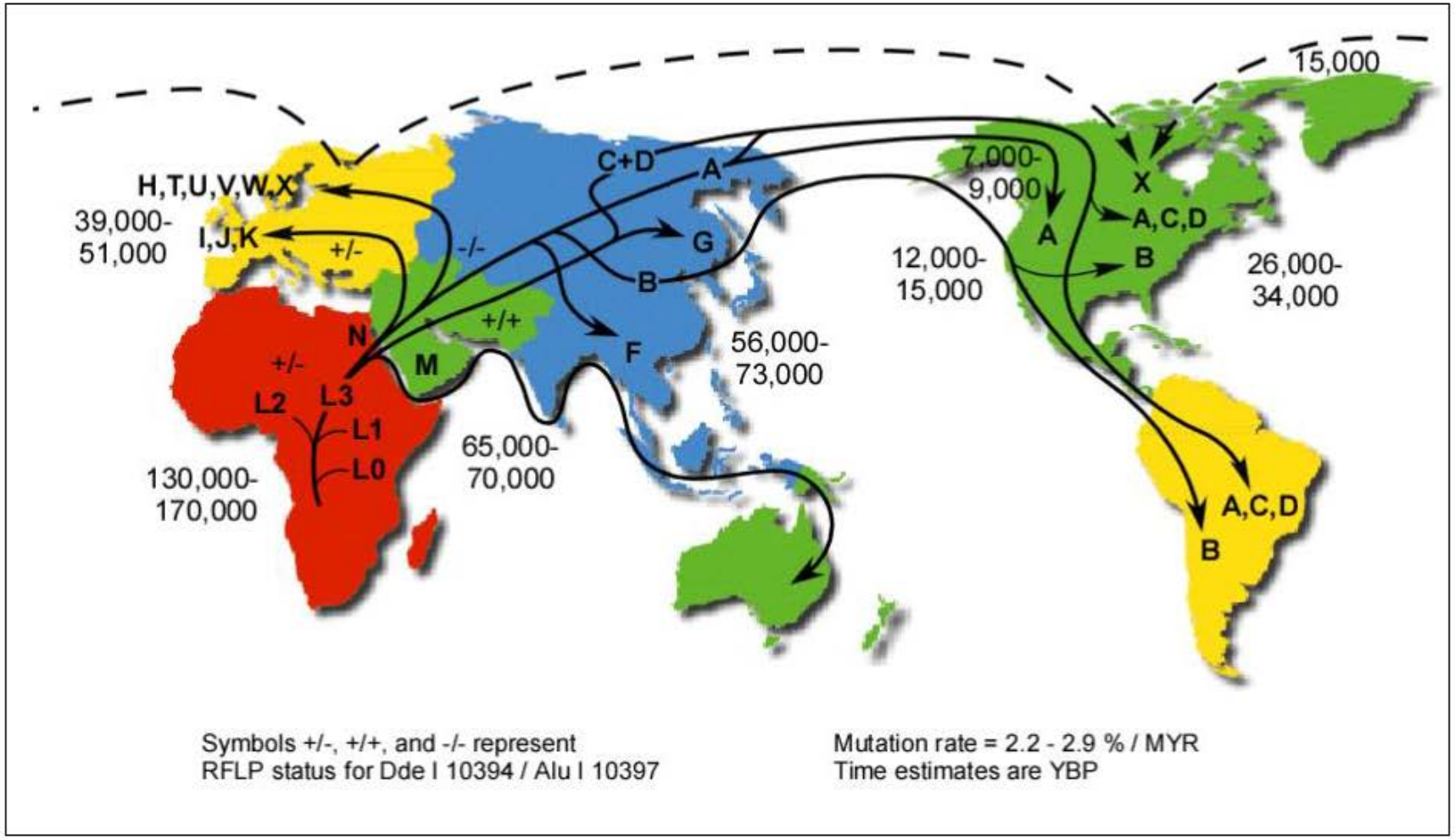

Figure 2.2: Map of world human mtDNA migrations and continent specific haplogroups (http://.mitomap.org/MITOMAP)

Beside M, macrohaplogroup N was another group that rooted out from L3 and radiated to haplogroup R and diverged rapidly within haplogroup N about 60kya (Macaulay *et al*., 2005). From this N macrohaplogroup, N21 and N22 have been found in Southeast Asia and more common in the aboriginal populations in Malay Peninsula (Hill *et al*., 2007; Maruyama *et al*., 2010). The haplogroup R branches out further into haplogroup B and haplogroup F with several subclades each. The B and F were reported as major haplogroups and have been found in Continental East Asia and also Island SEA (Maruyama *et al*., 2010). In previous study, it was mentioned that one of the most common haplogroups in ISEA was haplogroup B, which fell into two main clades, B4 and B5 (Hill *et al*., 2007). In this group, haplogroups B1c1b and B4c1b3 were found more frequently in Island SEA although almost all B lineage were SEA common haplogroups and widely distributed in SEA populations (Maruyama *et al*., 2010).

Other major haplogroups in R was haplogroup F that branched out from R9. Among all kinds of F subhaplogroups, F1a has been found frequently in all East Asia, SEA and ISEA with F1a1a relatively restricted to mainland SEA and Island SEA (Kivisild *et al*., 2002). The F1a2 and F1a3 groups seems to be rare lineage in East Asia and SEA and some recurrence mutations have been found in F subclades for example SNP 16355 that was shared by F3a and M9a (Hill *et al*., 2007). Figure 2.3 shows the summary of human mtDNA lineage.

Various populations have been found in different locations for both Mainland SEA and Island SEA. Above all, the findings of SEA populations were mainly based

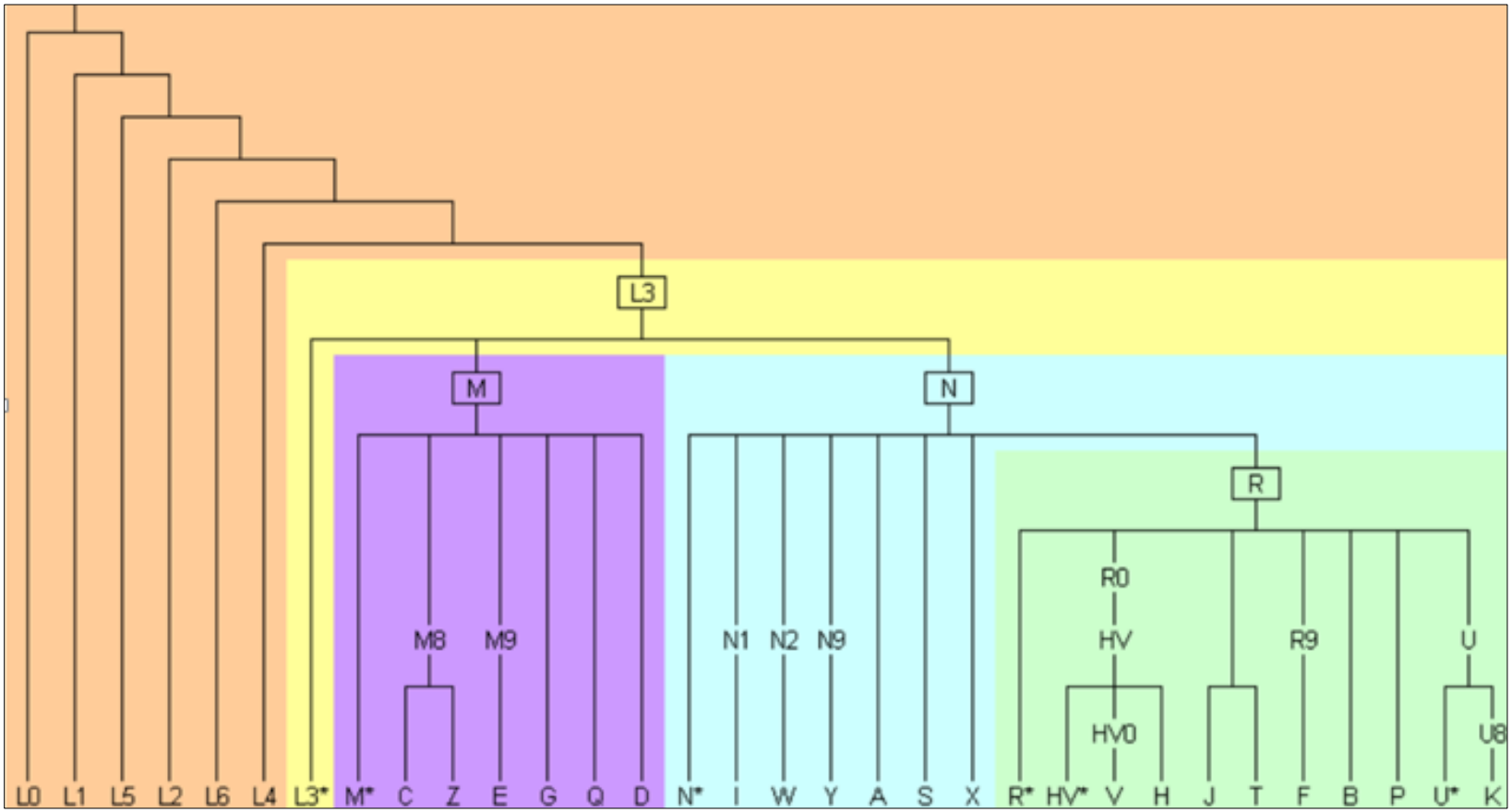

Figure 2.3: Simplified human mtDNA lineages (van Oven & Kayser, 2009)

on the languages used by those populations and of course geographic was the main factor that influenced the languages and their lifestyles. The language used by SEA populations recorded for both Mainland SEA and Island SEA have gave us general information on the migrations of SEA populations. From previous studies that mainly focused on the languages used, several SEA populations were recorded and grouped based on their geographical locations. The recorded languages managed to provide general information on populations found in SEA although it was believed that the reported data still did not cover the whole populations that belong to SEA.

2.5 SNPs of the selected haplogroups

Overall, the selection of mtDNA SNPs in this study were strictly based on SEA haplogroups classification. The SNP variants occurred in many ways such as deletion, insertion, transition and also transversion. From the updated mtDNA haplogroup tree that was accessed from MITOMAP (www.mitomap.org), all recorded mutations were transitions type except indicated other than that. Transition can be explained as nucleotide changes in the same group for example purine to purine and pyrimidine to pyrimidine while transversion can be described as nucleotide changes from purine to pyrimidine or vice versa. The purine group consists of nucleotide A and G while nucleotide T and C is grouped in pyrimidine (Greco & Tor, 2005). The nucleotide variants in each of selected SNPs are listed in Table 2.1. Several SNPs are also shared by other macrohaplogroups.

In previous reports, several SNPs shared similar variants with other haplogroups, even across their macrohaplogroups. Out of the 30 selected SNPs, only 6 of them were recorded as exclusively belong to specific haplogroup. The SNPs were the ones at nucleotide positon 13626 (haplogroup E), 3552 (haplogroup N21), 9080 (haplogroup E2), 13362 (haplogroup M9), 3027 (haplogroup E) and 7684 (haplogroup R9c) and 1872 (R21). Recurrent mutations have been found for several SNPs and they were shared by multiple haplogroups for instance polymorphisms at nucleotide position 16390, 16261, 16291, 16093, 146 and 195. The defining SNPs for each haplogroup are listed in Table 2.2.