PHYLOGEOGRAPHY STUDY OF THE INDIGENOUS PEOPLE OF SABAH: KADAZAN, DUSUN, RUNGUS AND BAJAU

GAN YEE MIN

UNIVERSITI SAINS MALAYSIA

2020

PHYLOGEOGRAPHY STUDY OF THE INDIGENOUS PEOPLE OF SABAH: KADAZAN, DUSUN, RUNGUS AND BAJAU

by

GAN YEE MIN

Thesis submitted in fulfilment of the requirements
for the degree of
Doctor of Philosophy

July 2020

ACKNOWLEDGEMENT

All praise and glory to the God Almighty that this thesis is completed! My first and biggest thanks would be to Him who has been ever so faithful for being there with me every step of the way, for gifting me with the wisdom, perseverance and patience to see this PhD journey through from day one till the very end.

Next, I would like to express my utmost gratitude to my main supervisor, Dr. Eng Ken Khong and my co-supervisor, Dr. Velat Bujeng for their guidance and advice on this thesis. I am especially indebted to Dr. Eng, without whom I would not have the opportunity to work on this project to begin with. It is thanks to his relentless encouragement, support and patience in teaching and guiding me throughout these 3-4 years that kept me motivated and to strive to be better.

This research would also not have been possible without the financial support from several bodies. Firstly, I would like to thank Universiti Sains Malaysia (USM) for funding my studies through the USM Fellowship Scheme (2016-2019). Research work is funded by the Short Term Research Grant, USM (304/PARKEO/6313211) awarded to Dr. Eng, as well as the Long Term Research Grant Scheme, Ministry of Higher Education (304/PPSK/6150115/U132) awarded to Prof. Dr. Zafarina Zainuddin of the Analytical Biochemistry Research Centre (ABrC), USM.

I am also grateful to everyone from Lab 414, School of Biological Sciences, USM, especially Prof. Dr. Nazalan Najimudin and Prof. Dr. Razip Samian for allowing me to use the lab. A special mention goes to Mardani, Haida, Suhaimi and Faizal, who have all been so patient and understanding when guiding me through the various aspects of the lab work, as well as Prof. Dr. Phua Kia Kien and his student Carlos from the Institute for Research in Molecular Medicine (INFORMM), USM who helped me

with imaging of the gel electrophoresis. Further thanks go out to Prof. Mokhtar Saidin and everyone from the Centre for Global Archaeological Research (CGAR) for their financial and technical support.

I would also like to give my biggest appreciation to my family and friends, who have undoubtedly had to tolerate with me during my most stressful moments. To my friends and colleagues, thank you being part of my PhD journey in one way or another. A special mention goes to Deejay Daxter and Siaw Chan who have been so helpful and kind in helping me with the formatting and submission process. To my family, my greatest supporters, I would not have been able to pursue my dreams in archaeology if it were not for your trust, understanding and unconditional love. Thank you all for believing in me, and I hope that I have made each and everyone of you proud. Thank you also to my sister, Yee Wei, who has always responded to my requests for drawing images and maps with nothing but patience and helpfulness.

Finally, I would like to dedicate this thesis to all my grandparents, three of whom have passed away throughout the course of my PhD research. This research has taught me even more of the importance of knowing one’s ancestry, and I will be sure to pass that down to the generations of my family to come.

TABLE OF CONTENTS

LIST OF PUBLICATIONS

LIST OF CONFERENCE PRESENTATIONS

LIST OF TABLES

Page

LIST OF FIGURES

LIST OF ABBREVIATIONS

aDNA	ancient DNA
AMH	anatomically modern humans
ATP	adenosine triphosphate
BC	before Christ
bp	base pair
CRS	Cambridge Reference Sequence
dATP	deoxyadenosine triphosphate
dCTP	deoxycytidine triphosphate
ddATP	dideoxyadenosine triphosphate
ddCTP	dideoxycytidine triphosphate
ddGTP	dideoxyguanosine triphosphate
ddNTP	dideoxynucleoside triphosphate
ddTTP	dideoxythymidine triphosphate
dGTP	deoxyguanosine triphosphate
DNA	deoxyribonucleic acid
dNTP	deoxynucleoside triphosphate
dTTP	dethymidine triphosphate
EA	East Asia
EDTA	ethylenediaminetetraacetic acid
EtBr	ethidium bromide
HLA	human leukocyte antigen
HVS-I	hypervariable segment I
HVS-II	hypervariable segment II
HVS-III	hypervariable segment III
ISEA	Island Southeast Asia
ka	thousand years
kya	thousand years ago
LGM	Last Glacial Maximum
LGP	Last Glacial Period
$MgCl_2$	magnesium chloride

$MgCl_2 \cdot 6H_2O$	magnesium chloride hexahydrate
MJ	median-joining
ML	maximum likelihood
MP	maximum parsimony
MRCA	most recent common ancestor
MSEA	Mainland Southeast Asia
mtDNA	mitochondrial DNA
mya	million years ago
NaCl	sodium chloride
nDNA	nuclear DNA
NG	New Guinea
NMTCN	Nusantao Maritime Trading and Communication Network
np	nucleotide position
OOA	Out of Africa
PAML	Phylogenetic Analysis by Maximum Likelihood
PCR	polymerase chain reaction
RBC	red blood cell
rCRS	revised Cambridge Reference Sequence
RFLP	restriction fragment-length polymorphism
RM	reduced median
rRNA	ribosomal RNA
RSRS	Reconstructed Sapiens Reference Sequence
SA	South Asia
SDS	sodium dodecyl sulphate
SEA	Southeast Asia
SNP	single nucleotide polymorphism
TAE	Tris base, acetic acid and EDTA
Tris-HCl	Tris hydrochloride
tRNA	transfer RNA
UV	ultraviolet
ya	years ago
YTT	Youngest Toba Tuff

LIST OF APPENDICES

KAJIAN FILOGEOGRAFI KAUM-KAUM ETNIK DI SABAH: KADAZAN, DUSUN, RUNGUS DAN BAJAU

ABSTRAK

Kajian filogeografi merupakan kajian yang mengaplikasi penyelidikan filogenetik dalam bidang arkeologi untuk mengkaji corak migrasi dan penempatan manusia pada masa lampau. Di Asia Tenggara, kajian filogeografi telah digunakan secara meluas untuk mengkaji diaspora Austronesia, satu pergerakan kumpulan etnolinguistik yang luas tersebar. Bukti arkeologi dan linguistik telah menunjukkan bahawa petani padi yang bertutur proto-Austronesia berhijrah dari China Selatan ke Taiwan sekitar 5,500 tahun lalu. Bahasa-bahasa Austronesia kemudiannya berkembang di Taiwan dan mula menyebar ke Asia Tenggara, Oceania dan Polinesia sekitar 4,000 tahun lalu. Migrasi ini dikenali sebagai penyebaran "*Out of Taiwan*". Walau bagaimanapun, kajian-kajian genetik di sekitar kawasan ini menunjukkan bahawa keadaan sebenarnya jauh lebih rumit daripada yang dijangka. Malah, sebilangan penyelidik mencadangkan bahawa migrasi telah banyak kali berlaku pada zaman lampau dan bukannya satu kali sahaja, termasuklah migrasi yang bergerak ke arah Taiwan. Di Sabah (Borneo, Malaysia), semua kaum etnik bertutur dalam bahasa-bahasa Austronesia, contohnya bahasa Dusunik dan bahasa Sama-Bajau. Hal ini menunjukkan bahawa pencerobohan, akulturasi atau asimilasi penutur bahasa Austronesia telah berlaku di rantau ini. Namun, kajian filogenetik yang komprehensif belum lagi dijalankan di Sabah berbanding dengan kawasan-kawasan jiran seperti Filipina dan Indonesia. Oleh itu, kajian ini bertujuan mengisi jurang penyelidikan melalui analisis filogenetik terhadap empat kaum etnik utama di Sabah iaitu Kadazan, Dusun, Bajau dan Rungus, dengan menggunakan DNA mitokondria (mtDNA) yang diwarisi dari sebelah ibu. Kajian ini mempunyai dua objektif utama iaitu: (i) untuk mencirikan variasi "*control region*" (atau "*D-loop*") DNA mitokondria kumpulan-kumpulan etnik di Sabah; dan (ii) untuk mengenal pasti sumber populasi bumiputera Sabah samada di Asia Selatan, Asia Timur, daratan Asia Tenggara dan kepulauan Asia Tenggara serta mengkaji corak penempatan dan migrasi mereka menggunakan analisis filogenetik. Seramai 177 individu etnik di Sabah telah dikaji, dan 10 sampel telah dipilih untuk menjalani penurutan genom lengkap mtDNA untuk kajian selanjutnya seperti pembinaan pokok filogenetik dan anggaran masa migrasi serta penempatan. Hasil kajian didapati bahawa *haplogroup* mtDNA 177 individu ini boleh dibahagikan kepada lima episod penyebaran dan penempatan, iaitu (i) penempatan pertama oleh manusia anatomi moden (*anatomically modern humans*), (ii) migrasi awal kumpulan pemburu-pengumpul sebelum ketenggelaman pentas Sunda, (iii) penyebaran manusia selepas tempoh glasial pada Awal Holosen, (iv) penyebaran manusia dari Taiwan semasa Pertengahan Holosen dan (v) penyebaran manusia yang bukan dari Taiwan (tetapi rantau lain seperti Indonesia dan Oceania) semasa Pertengahan Holosen. Secara keseluruhan, hasil kajian ini bukan sahaja menolak model penggantian oleh penutur bahasa Austronesia malah ia turut menunjukkan bahawa kesan Austronesia mungkin berlaku dari aspek budaya dan bahasa dan bukannya dari aspek genetik. Akhirnya, kajian ini merupakan kajian pertama yang menumpu kepada kaum-kaum etnik di Sabah dan diharapkan penyelidikan ini akan turut menjadi rujukan dalam kajian filogenetik di Asia Tenggara pada masa depan.

PHYLOGEOGRAPHY STUDY OF THE INDIGENOUS PEOPLE OF SABAH: KADAZAN, DUSUN, RUNGUS AND BAJAU

ABSTRACT

Phylogeography is a field of study that applies phylogenetic research in the field of archaeology to study past human migration and settlement patterns. In Southeast Asia (SEA), it has been extensively used to study the Austronesian diaspora, possibly the most widespread movement of a single ethnolinguistic group. Through archaeological and linguistic evidence, it has been shown that rice farmers who were proto-Austronesian speakers migrated from South China into Taiwan c. 5,500 years ago. Austronesian languages then developed in Taiwan and Austronesian speakers subsequently spread into SEA, Oceania and Polynesia c. 4,000 years ago. This movement has been coined as the “Out of Taiwan” dispersal. However, recent genetic studies show that the situation is much more complex than this. Some scholars suggested multiple dispersals rather than a one-off migration, as well as a back or reverse migration into Taiwan instead of one that is mono-directional. In Sabah (Borneo, Malaysia), all the indigenous ethnic groups speak Austronesian languages such as the Dusunic and Sama-Bajau languages; this marks an invasion, acculturation or assimilation of Austronesian speakers in the region. However, Sabah has not been subjected to much phylogenetic analysis as compared to neighbouring regions such as the Philippines and Indonesia in spite of its strategic location. Hence, this study aims to fill this research gap by conducting a phylogenetic analysis using the maternally inherited mitochondrial DNA (mtDNA) on four major Sabah ethnic groups, namely the Kadazan, Dusun, Bajau and Rungus. The main objectives of this study are two-fold: (i) To characterise the mtDNA control region (or D-loop) variations of the ethnic groups of Sabah; and (ii) To identify potential source populations in South Asia, East Asia, MSEA and ISEA of the indigenous people of Sabah, as well as study their settlement and migration patterns using phylogenetic analysis. This study involved 177 ethnic individuals from Sabah, out of which 10 samples were selected and subjected to whole mtDNA genome phylogenetic analysis, which further includes drawing of the phylogenetic trees and coalescence time estimation. The resulted mtDNA haplogroups identified from the phylogenetic analysis in this study has yielded the following finding, that the peopling of Sabah can be categorised into five categories. Namely, the peopling of Sabah can be categorised into (i) The first settlement of anatomically modern humans (AMH) in Sabah; (ii) Settlements and/or migrations of early hunter-gatherer populations and the rise of ancient lineages prior to the flooding of the Sunda shelf; (iii) Early Holocene postglacial expansions; (iv) Mid-Holocene dispersal from Taiwan; and (v) Mid-Holocene dispersals that do not originate from Taiwan but are rooted in other regions such as Indonesia and Near Oceania. Overall, not only do the results from this study reject a total replacement model of Austronesian speakers, they also seem to suggest that the "Austronesian effect" could have been in cultural and linguistic terms rather than genetic. Finally, this is the first study focusing solely on the ethnic groups in Sabah, and it is hoped that this research will serve as a reference for future phylogenetic studies in SEA.

CHAPTER 1

INTRODUCTION

Past human settlement and migrations in Southeast Asia (SEA) has remained a highly debated and vigorously discussed subject matter in archaeology over the years. Newly unearthed archaeological evidence and material culture have constantly called for a revision of current knowledge, and the advent of scientific analysis in archaeology has also greatly improved (and sometimes disprove) any prevailing theories and models.

In Island Southeast Asia (ISEA), a geographical region covering modern-day East Malaysia, Brunei, Indonesia, Timor-Leste, Singapore and the Philippines, the "Out of Taiwan" theory is the most predominant model used to describe the spread of Austronesian speakers in the region. It argues for the rise of agriculture as a motivating factor behind the dispersal of Austronesian speakers who subsequently replaced any pre-existing hunter-gatherer populations that they came into contact with. This demic diffusion model has been strongly advocated by scholars such as Bellwood (1992, 2004, 2007, 2017), Bellwood & Dizon (2005, 2008) and Diamond (1988), who based their arguments mostly on archaeological evidence, as well as linguistic analysis largely conducted by Blust (1976, 1984-1985, 1995, 2013). Alternatively, some scholars (e.g. Solheim 1975; Meacham 1984-1985; Solheim 1984-1985; Oppenheimer & Richards 2002) argue that Holocene human dispersal was a response towards climate change and the rise in sea levels. More information on both opposing views are provided in the next chapter.

Within the past two to three decades, DNA studies, especially those pertaining to phylogeography and population genetic analysis, have greatly contributed to our understanding on human migrations and settlement during the Neolithic. Whilst there is now a plethora of DNA studies that has been conducted to shed more light on this subject matter, the currently available literature is seemingly focused on only certain regions within ISEA. Notably, much research has been conducted on the Philippines (e.g. Hill *et al*. 2006, 2007; Tabbada *et al*. 2010; Scholes *et al*. 2011; Delfin *et al*. 2014 etc.) as an immediate "Out of Taiwan" recipient, as well as Indonesia (e.g. Richards *et al*. 1998; Mona *et al*. 2009; Gunnarsdóttir et al. 2011b; Brandão *et al*. 2016; Kusuma *et al*. 2017 etc.) due to its proximate location to Near Oceania, thus acting as the final ISEA step for Austronesian speakers before advancing further into Remote Oceania and Polynesia (Figure 1.1). In spite of its strategic location between both regions as well as its close proximity to Peninsular Malaysia and Mainland Southeast Asia (MSEA), the northern part of Borneo, comprising of Sabah and Sarawak of East Malaysia, is often overlooked in phylogenetic studies. Hence, this region presents a research gap in the phylogenetic studies of ISEA populations.

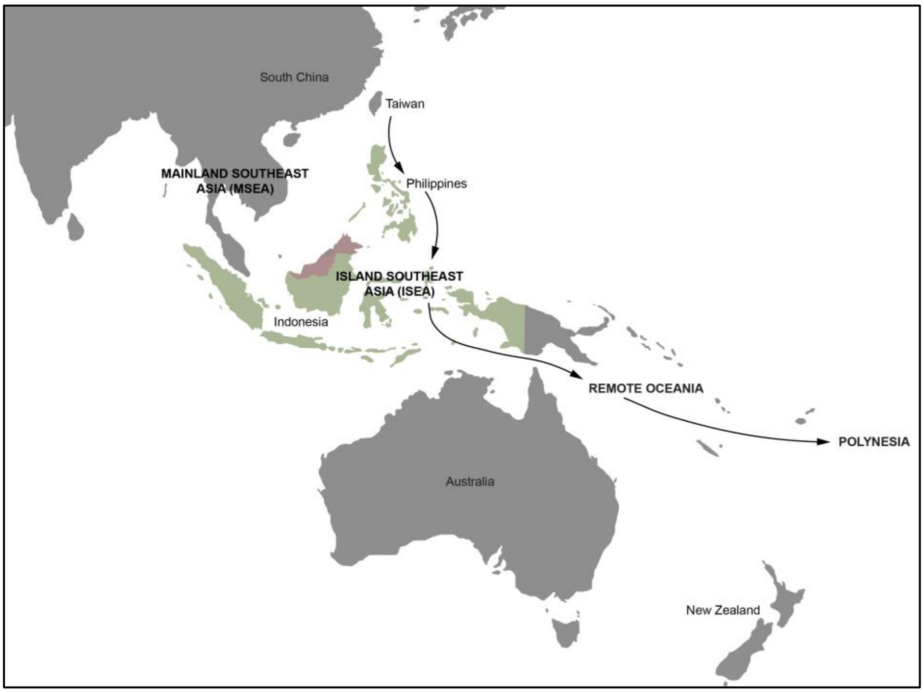

Figure 1.1 Map showing the approximate route for the "Out of Taiwan" migration of Austronesian speakers, originating from Taiwan and spreading into ISEA via the Philippines, subsequently moving into Remote Oceania and Polynesia (Bellwood (1992, 2004, 2007, 2017), Bellwood & Dizon (2005, 2008) and Diamond (1988)). Note that a separate movement of the Austronesian speakers into MSEA/Peninsular Malaysia and further east to Madagascar is not shown here. Highly researched areas in ISEA (the Philippines and Indonesia) have been highlighted in green. By contrast, Sabah and Sarawak (highlighted in red) remains largely under-studied in spite of its strategic location within ISEA and its proximity to MSEA.

1.1 Aims and motivation

For the reason mentioned above, this study hopes to fill in the research gap by subjecting the ethnic groups of Sabah to phylogenetic analysis. The main objectives of this study are two-fold:

1. To characterise the mitochondrial DNA (mtDNA) control region variations of the ethnic groups of Sabah; and
2. To identify potential source populations in South Asia, East Asia, MSEA and ISEA of the indigenous people of Sabah, as well as study their settlement and migration patterns using phylogenetic analysis.

Due to time and financial constraints, only the major ethnic groups in Sabah will be studied. Nevertheless, as the mtDNA control region variations of the Sabah ethnic groups are previously uncharacterised, this will greatly benefit researchers seeking to identify and quantify human settlement and migrations in the region using mtDNA. We would expect to observe some differences in the control region of the mtDNA within the different ethnic groups, and this would shed light on questions regarding demography and migration history of the different ethnic groups. Furthermore, the results from this study will set the ground for further mtDNA characterisation of other ethnic groups, as well as future research utilising other types of DNA, such as the paternally inherited non-recombining Y-chromosome or autosomal DNA.

On a broader scale, the results from this study can be used to test if the ideas put forward for other parts of ISEA holds true for this part of the region. Firstly, we would be able to observe how the genetic data from this study relate to the "Out of Taiwan" theory that is primarily based on archaeological and linguistic evidence; estimated coalescence ages from this study and the published literature will allow us to deduce if the genetic data corroborates or opposes the dates proposed by the "Out of Taiwan" theory. This will subsequently allow us to test population settlements and movements during the Holocene period, as well as study the effects of climate change on human settlements and dispersals since the Last Glacial Maximum (LGM).

1.2 Scope and limitations

As mentioned earlier, due to time and financial constraints, the mtDNA of the major ethnic groups in Sabah, namely the Kadazan, Dusun, Rungus and Bajau, will be the focus of this study. The materials and methodology that have been undertaken to carry out this study is summarised in Figure 1.2. Specific details of the materials and methodology will be provided in Chapter 3.

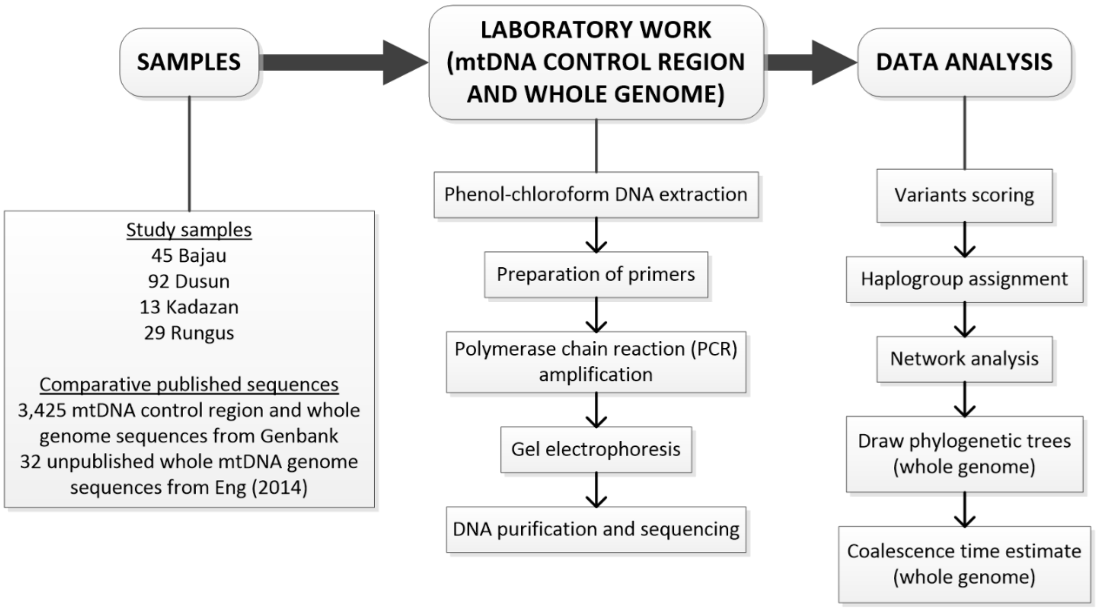
SAMPLES
LABORATORY WORK (mtDNA CONTROL REGION AND WHOLE GENOME)
DATA ANALYSIS
Study samples
45 Bajau
92 Dusun
13 Kadazan
29 Rungus
Comparative published sequences
3,425 mtDNA control region and whole genome sequences from Genbank
32 unpublished whole mtDNA genome sequences from Eng (2014)
Phenol-chloroform DNA extraction
Preparation of primers
Polymerase chain reaction (PCR) amplification
Gel electrophoresis
DNA purification and sequencing
Variants scoring
Haplogroup assignment
Network analysis
Draw phylogenetic trees (whole genome)
Coalescence time estimate (whole genome)

Figure 1.2 Flowchart summarising the materials and methods that are involved in this study.

The Kadazan, Dusun, Rungus and Bajau ethnic groups are chosen, as they represent the largest ethnic groups in Sabah by population (Sabah State Government 2019). On the other hand, the mtDNA is chosen, as it does not reshuffle or recombine, is more easily detectable in the human cell, and has a higher mutation rate, which makes changes and diversity in the mtDNA over a very long period of time more easily detectable. The specifics of the advantages of subjecting mtDNA to phylogenetic analysis is discussed in more details in Section 2.1.1.

However, limitations arise from focusing only on the mtDNA of the major ethnic groups of Sabah. Firstly, Sabah comprises of 33 indigenous groups (Sabah State Government 2019); thus, the findings and outcomes from this study might not reflect the overall phylogenetic signature of the entirety of Sabah. Nonetheless, the findings and outcomes would still provide a snapshot into the mtDNA phylogenetic signature of Sabah, which was previously uncharacterised. Secondly, the mtDNA is maternally inherited; thus, the findings and outcomes from this study would be biased towards the maternal lineage and might not reflect a similar scenario as with the paternal lineage, which could be obtained through studying another type of DNA, the Y-chromosome. Regardless, the findings and outcomes from this study would still benefit current phylogenetic studies in SEA that have focused on the mtDNA by filling in the research gap, as well as serving as a reference study for future phylogenetic studies in Sabah, Malaysia or SEA involving other forms of DNA. It is also important to note that the assignment of the haplogroups in this study is limited to only polymorphisms from the mtDNA control region, or more specifically the HVS-I. Additionally, further assignment into subhaplogroups for some samples will only possible with variants that occur in other parts of the mtDNA, such as the coding region. Regardless, this does not hinder our aim to provide a characterisation of the mtDNA control region and subsequently the genetic signature and diversity of the ethnic groups of Sabah.

Furthermore, the haplogroup diversity that we will observe in the ethnic groups in this study may be partial due to the limiting sample size, especially for the Kadazan individuals who are one of the major ethnic groups in Sabah in terms of population size but are only represented by 13 individuals in this study Figure 1.2). Nonetheless, these limitations will, by no means, impede in shedding new light on the genetic signature and distribution of the ethnic groups from Sabah, a region which has previously been subjected to minimal phylogenetic analysis, as mentioned earlier. As a pilot study, this research will characterise the mtDNA control region haplogroup and genetic diversity of the major ethnic groups in Sabah, and this would unquestionably act as a crucial stepping stone for future mtDNA studies in Sabah and the wider region of Borneo or ISEA.

1.3 Thesis outline

This thesis will have the following chapters:

i. Firstly, Chapter 2 will provide a literature review on topics and any background information that is relevant to the study. This includes, but is not limited to, an introduction to DNA, basic principles and prerequisite knowledge on phylogenetic analysis of human mtDNA, a discussion on modern human origins and dispersals and human migrations in ISEA during the Neolithic, as well as the wider context of SEA and Sabah.

ii. Chapter 3 will detail the materials and methodology used to conduct the research in this study.

iii. Chapter 4 will provide an overview on the mtDNA control region haplogroup profile of each Sabah ethnic group. This is followed by a preliminary discussion on the genetic signature and diversity of each ethnic group.

iv. Chapters 5, 6 and 7 will examine macrohaplogroups M, N and R respectively. These three chapters will present and discuss the haplogroups nested under these macrohaplogroups that have been identified in this study based on the mtDNA control region. These three macrohaplogroups are presented and discussed in separate chapters, as they have distinctive characteristics in terms of their defining mutations, coalescence dates, geographical region spanned and phylogenetic history that make them differ from one another.

v. Chapter 8 will present the results and discussion for the selected haplogroups that have been subjected to whole mtDNA genome analysis.

vi. Lastly, Chapter 9 will bring together all of the data obtained in this study and provide a more comprehensive discussion of the results and their implications. This is followed by a conclusion of the study, as well as some suggestions for further research.

CHAPTER 2

LITERATURE REVIEW

The information provided in this literature review chapter sets out to provide the foundation and in-depth understanding on the subject matters that are relevant to understanding this research study.

2.1 An introduction to human deoxyribonucleic acid (DNA)

Deoxyribonucleic acid (DNA) is the genetic material of all living organisms, except for a few kinds of viruses. DNA is a polymer, formed by a combination of monomeric subunits known as nucleotides. The basic structure of a nucleotide consists of a nitrogenous base, pentose (a type of sugar composed of five carbon atoms), and a triphosphate group (Figure 2.1). There are four nucleotides, which are distinguished by the chemically distinct structure of the bases. The four different bases are adenine (A), cytosine (C), guanine (G) and thymine (T), and they respectively form the four nucleotides known as deoxyadenosine triphosphate (dATP), deoxycytidine triphosphate (dCTP), deoxyguanosine triphosphate (dGTP), and deoxythymidine triphosphate (dTTP).

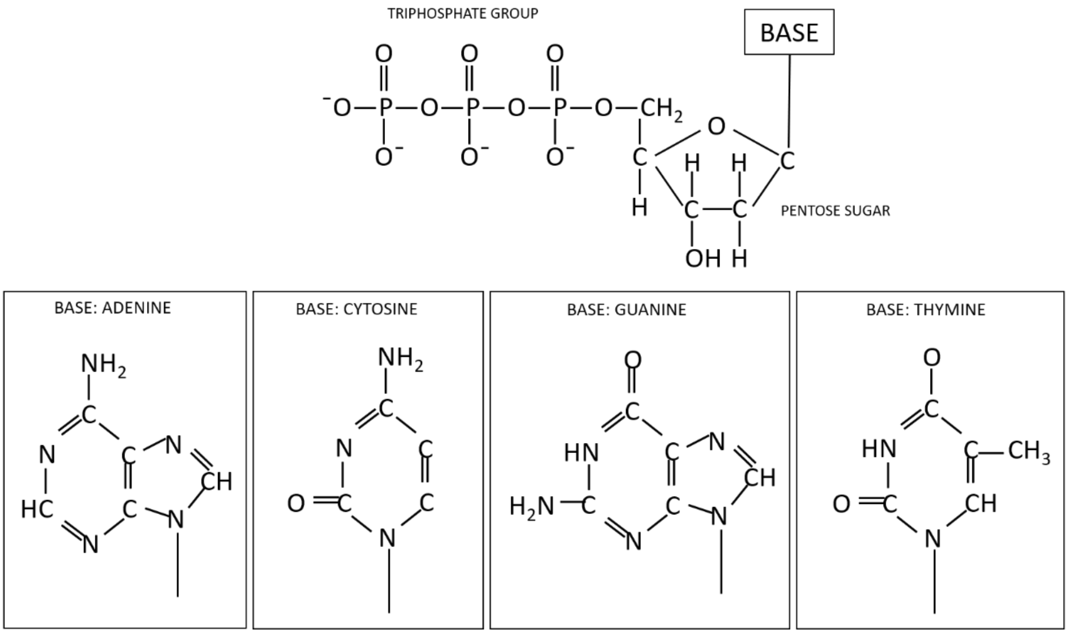

Figure 2.1 Basic chemical structure of a DNA molecule (top) and the four different bases of a DNA nucleotide (bottom). (After Jobling *et al*., 2014).

A DNA molecule consists of two polynucleotide chains that is held together by the hydrogen bonds between the nucleotides, forming the well-known double helix structure (Watson & Crick 1953) of a DNA (Figure 2.2). The bases of the two polynucleotide chains are always complementary to another; in other words, an adenine on one polynucleotide chain is always paired with a thymine on the other polynucleotide chain and vice versa, whereas a cytosine is always paired with a guanine and vice versa. For this reason, the unit length of the DNA molecule is known as a base pair (bp). The two polynucleotide chains are anti-parallel, and the sequences of each polynucleotide chain are read in the direction of 3′ to 5′.

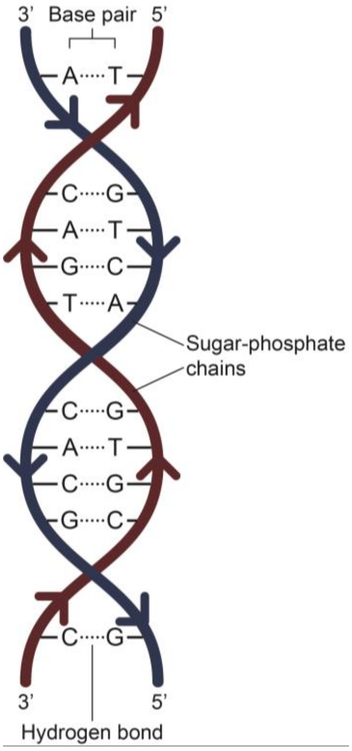

Figure 2.2 Double helix structure of a DNA molecule. (After Brown & Brown, 2011).

As with all multicellular eukaryotic organisms, there are, generally speaking, two different types of genomes in humans – the nuclear genome and mitochondrial genome. The nuclear genome (or nuclear DNA, nDNA) is the type of DNA that is found within the nucleus of a cell. nDNA can be arranged into 46 (i.e. 23 pairs of) linear molecules known as chromosomes. The only exception for this is the reproductive cells (or gametes), which only contain 23 chromosomes. Each member of a chromosome pair is individually inherited from the father and the mother thus nDNA undergoes reshuffling and recombination when it is passed on from the parent to the offspring. In humans, nDNA is ~3,200,000,000 bp long per haploid genome (i.e. reproductive cells) and twice the length for somatic cells (i.e. any cell other than the reproductive cells). There are also only two copies of nDNA per somatic cell.

2.1.1 Human mitochondrial DNA (mtDNA)

By contrast, mitochondrial DNA (mtDNA) is the type of DNA that is found within organelles known as mitochondria that are present in the cytoplasm of a cell. Mitochondria are organelles that converts food into chemical energy in the form of adenosine triphosphate (ATP) to be used by cells. Hence, they are often referred to as the "powerhouses" of cells. Initially a controversial theory by Lynn Sagan, née Margulis (Sagan 1967), it is now widely accepted that mitochondria originated as prokaryotic bacteria, which was incorporated into eukaryotic cells ~1.5 billion years ago through a process known as symbiogenesis; this is also known as the endosymbiotic theory.

The human mtDNA is a circular molecule that is 16,568 bp long in humans (Anderson *et al*. 1981; Andrews *et al*. 1999; Jobling *et al*. 2014) (Figure 2.3). It is double-stranded, with one strand being the heavy strand (H) and the other being the light strand (L). Both strands are characterised by the high content of guanine and cytosine respectively. The mtDNA encodes 37 genes, of which, 13 are for proteins, 22 are for transfer RNA (tRNA) and two are for ribosomal RNA (rRNA). The mtDNA also consists of a non-coding region of approximately 1,100 bp in length known as the control region (or D-loop). This region can be separated into three segments, known as the hypervariable segments I, II and II (HVS-I, II and III), and is often targeted for mtDNA analysis. The control region is a good candidate for phylogeographic studies, as it is within this region of the mtDNA where most mutation occur.

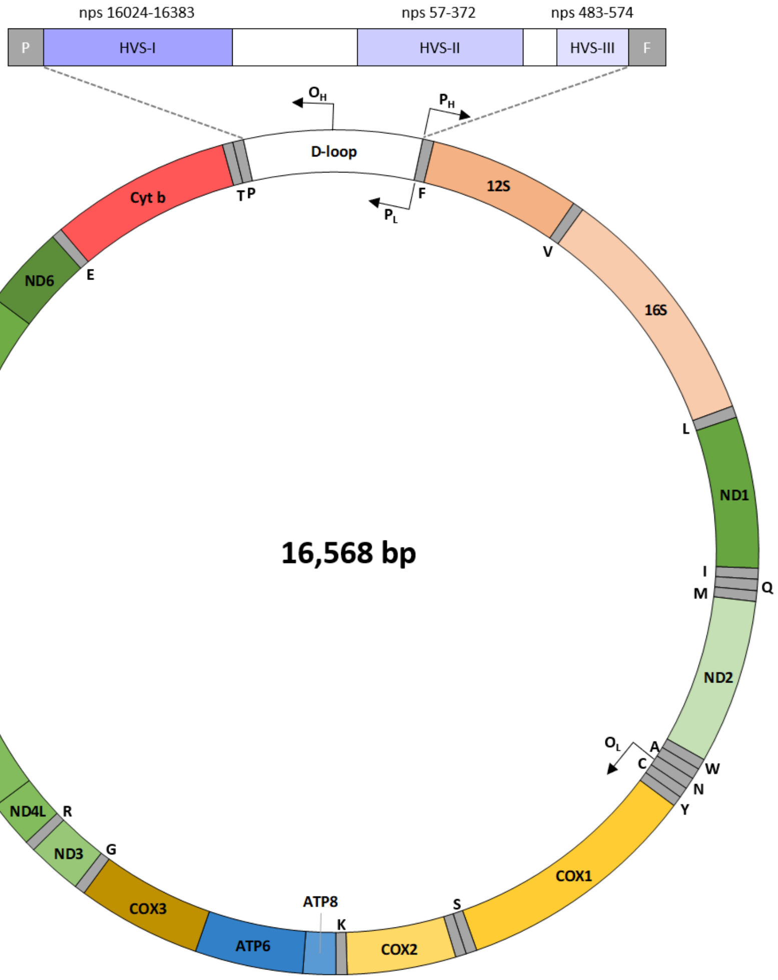

Figure 2.3 Schematic diagram of the human mtDNA. The mtDNA encodes 37 genes, of which, 22 are for tRNAs (indicated by single letter abbreviations e.g. F, V, L etc.), two are for rRNAs (12S and 16S) and the remaining 13 are for proteins (e.g. ND1, COX1 etc.). The D-loop (or control region) consists of HVS-I, II and III; nucleotide positions (nps) of the HVS are according to Andrews *et al*. (1999). Overlapping segments within the mtDNA (e.g. ATP6 and ATP8) are not depicted in this diagram. O_L and O_H indicate the origins of the replication of the light and heavy strands respectively, whereas P_L and P_H indicate the promoters for the transcription of these two strands. (After Jobling *et al*., 2014).

Unlike the nDNA, the mtDNA does not recombine and reshuffle when it is inherited from the parent to the offspring. This is beneficial over nDNA since reshuffling and recombining causes a change over each generation. Therefore, genomic variations are easier to infer through mtDNA, as the mtDNA would retain its genomic signature over generations and any change in the mitochondrial genome would have been purely mutational. In recent years, studies (e.g. Ruiz-Pesini *et al*. 2004; Pereira *et* *al*. 2011; Soares *et al*. 2013; Cavadas *et al*. 2015) have increasingly shown that mammalian and human mtDNA mutations are not random but are instead affected by natural selection (commonly referred to as purifying selection in mtDNA studies). Purifying selection is a process in which deleterious mutations (i.e. mutations that are harmful to the organism) are selectively removed. In animals, selection has been shown to take place during formation of the female germ cells (Jenuth *et al*. 1996; Cree *et al*. 2008; Wai *et al*. 2008), but this has only been recently been demonstrated in humans (Wei *et al*. 2019). Regardless, researchers have been increasingly observing and acknowledging the effect of purifying selection on the mutation rate and molecular clock of the mtDNA, and this will be discussed in more details in Section 2.2.5 later.

The mtDNA is solely maternally inherited, therefore, analysing the mitochondrial genome would allow us to study the maternal lineage of an individual or population. There have been studies (e.g. Schwartz & Vissing 2002; Luo *et al*. 2018) that suggest the paternal input in the human mtDNA; nevertheless, this remains rare and to be conclusively established. Additionally, sperm cells only contain about 50-75 copies of mitochondria, as opposed to oocytes (female gametes), which contain ~100,000 mitochondria, thus the contribution of paternal mtDNA (if present) is very small and essentially negligible (Jobling *et al*. 2014).

Moreover, compared to the nDNA, there are also more copies of mtDNA in a cell, which makes analysing the mtDNA comparatively easier to detect especially in the case of ancient DNA (aDNA). Each mitochondrion contains 2-10 copies of mtDNA, and the number of mitochondria in a cell can vary from hundreds to thousands depending of the amount of energy required by the cell. Therefore, a single cell may contain up to 10,000 copies of mtDNA. Furthermore, the mutation rate of mtDNA is also higher than that of the nDNA (approximately 10 times higher) (Brown *et al*. 1979).

This makes changes and diversity in the mtDNA over a very long period of time more easily detectable. It is precisely these collective characteristics of the mtDNA that make it very suitable for evolutionary studies, especially to investigate the maternal lineage.

2.2 Studying and analysing the human mtDNA

Archaeogenetics is a subdiscipline of biomolecular archaeology that is particularly interested in analysing ancient and modern DNA to study the human past. Examples of the applications of archaeogenetics include sex identification, palaeopathological studies, kinship analysis and human origins studies. Of particular interest here is a field of study that is known as phylogeography. First used by Avise *et al*. (1987), this field of study combines population genetics data (e.g. the mtDNA genealogy) with modern population geographical distribution in an attempt to address how past historical processes (such as climate change and dispersal events) could have potentially shaped the current geographical distribution of said population. Phylogeographic studies require a unique molecular marker where changes can be observed and traced over a significant period of time, and the characteristics of the mtDNA described earlier makes it a very suitable sample for this form of study. The following subsections will provide prerequisite knowledge for conducting a phylogeographic study.

2.2.1 The human mtDNA reference sequence

Early human genetic studies suffered from the lack of knowing exactly the structure of the human genome. This was completely changed when Anderson *et al*. (1981) published the first complete human mtDNA sequence known as the Cambridge Reference Sequence (CRS). It was later corrected and revised by Andrews *et al*. (1999)

and the resulting sequence is known as the revised Cambridge Reference Sequence (rCRS). The rCRS has completely revolutionalised human mtDNA studies, as it has since been used as a reference sequence for variants scoring and building the human mtDNA phylogenetic tree. Note that the rCRS does not represent the oldest human mtDNA sequence (i.e. “Mitochondrial Eve”), but instead belongs to a European individual who fall under haplogroup H (more information on haplogroups will be provided in Section 2.2.4(a) later).

In recent years, another reference sequence known as the Reconstructed Sapiens Reference Sequence (RSRS) has been proposed as an alternative for the rCRS (Behar *et al*. 2012). Unlike the rCRS, the RSRS was constructed by incorporating all available complete mtDNA sequences of *Homo neanderthalensis* into the root of the human mtDNA phylogenetic tree. Because the RSRS is reconstructed from the root of *Homo sapiens* mtDNA phylogeny, it would offer a more accurate account of ancestral versus derived mutations when used as a reference sequence for scoring mutations. Thus, the new RSRS would have the potential to resolve problems pertaining to imprecise estimated coalescence ages that are calculated based on the mutations.

Nonetheless, the rCRS is still chosen as a reference sequence for this study, as not only is it well-established, it would also prevent any errors or ambiguities from arising by adopting the new RSRS system. Moreover, many studies still opt for the rCRS as a reference sequence, which makes comparing the mutations observed in this study to those reported in the published literature more straightforward. Regardless, the reconstruction of the first complete human mtDNA sequence has been an extraordinary advancement in the field of human genetics, and none of this would have been possible without the pioneering development of any of the processes that will be discussed next.

2.2.2 Polymerase chain reaction (PCR)

The development of polymerase chain reaction (PCR) has greatly benefitted molecular biology research because in order to perform any form of DNA study or analysis, large amounts of DNA molecules are required. This is especially the case for archaeogenetic studies dealing with aDNA, as aDNA is often not preserved in huge quantities. Therefore, the DNA samples (regardless of modern or ancient) needs to be amplified in order to be studied. In other words, the amount of DNA in a sample needs to be significantly increased and this can be achieved through a thermo-cycling process known as PCR (Mullis *et al*. 1987). It is a quick and cost-effective way to exponentially amplify targeted segments of the DNA.

To put it simply, PCR synthesise replicated copies of DNA polynucleotides by unwinding the double helix strand and using each original polynucleotide chain as a template to form new complimentary copies. There are, broadly speaking, three stages in a PCR cycle (Figure 2.4):

1. Denaturation of the DNA double helix structure. This is achieved through increasing the temperature to ~95 °C. At this temperature, the hydrogen bonds between the polynucleotides will break thus separating the double-stranded DNA into two single-stranded polynucleotides.
2. Attachment or annealing of primers on the newly formed single-stranded polynucleotides. At this stage, the temperature is decreased to 50-60 °C and the specific temperature would be dependent on the annealing temperature of the primers used. Primers are short strands of nucleotides (i.e. oligonucleotides) of roughly 15-20 bp in length that bind onto the DNA template. Its function is to determine the starting point of the DNA synthesis.

3. Synthesis of new double helix structures. The temperature is raised to 72 °C, the optimum temperature for the *Taq* DNA polymerase, a thermostable enzyme that synthesises new DNA strands. The DNA polymerase synthesises new DNA strands by using the original polynucleotide chain as a template and attaching nucleotides that have been added into the PCR mixture onto the template. Formation of new sequences are always in the 5′ to 3′ direction and are governed by the base-pairing rules.

Whilst the starting point (5′ ends) is determined by the primers, synthesis of the new sequences terminates by chance and may sometimes go beyond the 3′ ends of the targeted sequencing region. However, after the third cycle (Figure 2.4), shorter strands of DNA, which lengths are determined by the primers, would have been formed and used as templates in the subsequent cycles. This would result in the continuous replication of the shorter targeted region, giving rise to more and more DNA strands that are of the targeted sequencing region, assuming that synthesis of the new strands to the desired length is successful at each cycle. Therefore, repeating the three stages in numerous cycles (usually 30-35 cycles) would result in an exponential increment of the amount of DNA, thus allowing us to obtain high amounts of DNA molecules with the targeted length in a relatively short period of time. The fact that PCR allows us to specifically target our desired segment of the DNA for amplification has been greatly advantageous to genetic studies, as this would prevent amplification of contaminants especially with the case of aDNA samples.

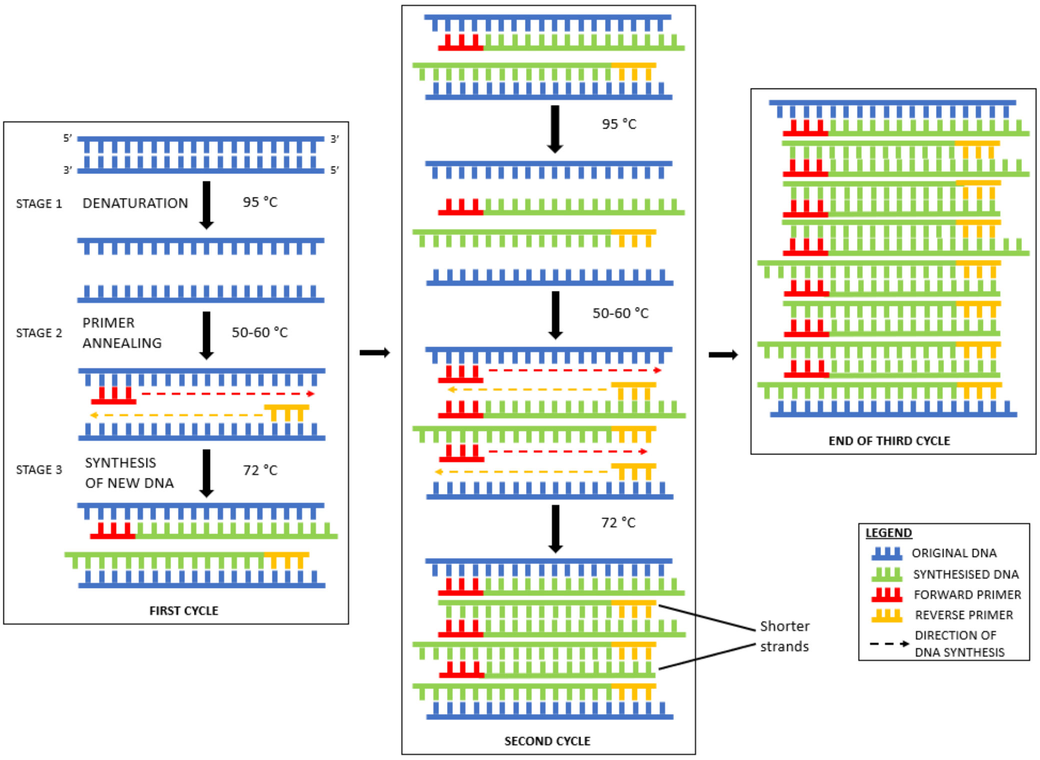

Figure 2.4 First three cycles of the polymerase chain reaction (PCR). At the end of the third cycle, more shorter strands of the targeted sequencing region would have been produced, which would result in an exponential increase of these DNA strands the subsequent cycles. (After Brown & Brown, 2011).

2.2.3 Sanger sequencing

As mentioned earlier, the discovery of the CRS was a major breakthrough for human genetic studies, and this could not have been achieved without first the invention of the Sanger sequencing method. Sanger sequencing is a chain termination method that was developed by Frederick Sanger in the 1970s (Sanger *et al*. 1977). The method is similar to PCR in that it is also a thermo-cycling reaction but has two distinct differences. Firstly, only one primer is required hence replication of the DNA template is linear rather than exponential. Secondly, apart from the standard nucleotides, dideoxynucleotides (ddNTP) are also added into the sequencing mixture. ddNTP differ from standard nucleotides in that they bear a hydrogen (–H) atom rather than the standard hydroxyl (–OH) group at the 3′ carbon (Figure 2.5). This modification prevents further strand synthesis of new DNA sequences because it is the hydroxyl group that the next nucleotide would normally be attached to. Thus, an incorporation of a ddNTP would result in a chain termination. Moreover, each ddNTPs (i.e. ddATP, ddCTP, ddGTP and ddTTP) is labelled with a different fluorescent dye, which allows for easy detection when reading the DNA sequence using a fluorescence detector.

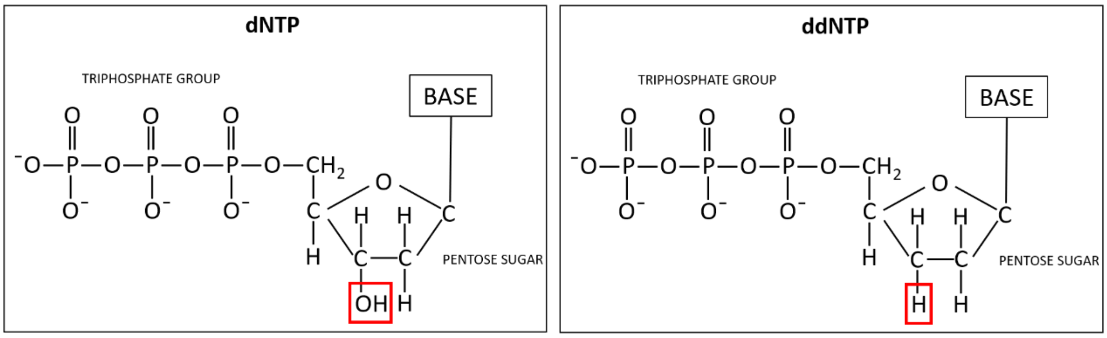

Figure 2.5 Chemical structures of a dNTP and a ddNTP. The latter differ from the former in that it has a hydrogen atom (as opposed to a hydroxyl group) attached to the 3′ carbon. (After Jobling *et al*., 2014).

When synthesising new DNA sequences, the *Taq* polymerase does not discriminate between dNTPs and ddNTPs. Even though either nucleotide could be chosen for synthesis, dNTPs are present in excess as compared to ddNTPs, thus chain termination is less likely to occur so soon. The fact that selection of nucleotides is by random would result in replicated DNA strands of different lengths. These DNA sequences are then loaded into a capillary gel system, which would separate the DNA strands according to their lengths. Once separation is completed, a fluorescent detector is used to detect and identify the ddNTP that is attached to the individual DNA strands, which then provides us with the entire DNA sequence.

2.2.4 Phylogenetic trees and networks

Phylogenetic trees are typically constructed using distance- or character-based methods. The four main methods are neighbour-joining (NJ), maximum parsimony (MP), maximum likelihood (ML) and Bayesian inference methods. NJ is a distance-based method that clusters every sequence at each stage of the tree in order to obtain an overall tree with the shortest branch lengths (Saitou & Nei 1987). By contrast, character-based methods such as MP, ML and Bayesian inference methods reconstruct phylogenetic trees based on the changes (i.e. mutations) in the sequences. MP trees are reconstructed by following the evolutionary pathway with the smallest number of character changes (Fitch 1971). ML trees similarly considers the evolutionary pathway with the least number of changes, but also considers all possible alternatives and produces the tree which has the highest likelihood of explaining the evolutionary data (Schmidt & von Haeseler 2009). And whilst ML considers the tree (prerequisite knowledge) to provide the evolutionary pathway (data) with the highest probability, Bayesian inference methods consider the evolutionary data to provide a tree with the highest probability (Drummond & Rambaut 2009; Ronquist *et al*. 2009). Bayesian inference methods can be carried out using softwares such as BEAST (Drummond & Rambaut 2007) or MRBAYES (Huelsenbeck & Ronquist 2001).

However, the resulting tree may not necessarily represent actual evolutionary processes due to parallel or reverse mutations (Bandelt *et al*. 1995). These mutations subsequently result in homoplasy (diverged haplotypes sharing similar mutations despite following independent evolutionary pathways), recombination (the merging of previously divergent haplotypes), sequence error or superimposed sequences (Bandelt *et al*. 1995; Jobling *et al*. 2014). These processes form cycles or reticulations within a phylogeny, or what is known as a network (Jobling *et al*. 2014).

Like phylogenetic trees, there are various distance- or character-based approaches in constructing network, and an overview of the various methods is provided in Huson & Scornavacca (2010). Here, focus is placed on a type of network, known as median network. This form of network is especially useful for dealing with mtDNA control region sequences that are prone to homoplasy. It takes into consideration all possible evolutionary pathways, which often lead to complex and intangible networks. However, the complexity can be resolved by reducing the reticulations through methods known as reduced median (RM) (Bandelt *et al*. 1999) or median-joining (MJ) (Bandelt *et al*. 1995).

The RM method considers all possible evolutionary pathways before eliminating least probable ones, whereas the MJ method is built based on minimum spanning networks where only more probable evolutionary pathways are considered at each stage (Bandelt *et al*. 1995, 1999; Jobling *et al*. 2014). The least probable evolutionary pathways (or reticulations) in the RM method are resolved by taking into consideration the weight of the characters and the frequency of which the haplotype occurs (Bandelt *et al*. 1995).

2.2.4(a) Human mtDNA phylogenetic tree

Nomenclature of the present mtDNA phylogenetic tree follows a simple cladistic order, which represents the relationship of the more ancestral haplogroups and their nested descendant subclades. A haplogroup is a group of individuals with similar haplotypes that shares a common ancestor; a haplotype is a combination of a set of polymorphisms (i.e. mutations) within a DNA molecule. Labelling of the mtDNA phylogenetic tree follows an alphamerical system, where lowercase alphabets (except for the major clade) and numbers alternate. For example, a major haplogroup would be identified by a single uppercase alphabet, e.g. haplogroup M. Subsequent lineages will be labelled numerically, e.g. M1, M2, M3 etc. If M1 diverges into three further subclades, its subclades would be labelled as M1a, M1b and M1c, and the alternating alphamerical labelling method continues until all subclades have been defined.

Nomenclature of the present mtDNA phylogenetic tree was first introduced using the four main Native American haplogroups A, B, C and D by Torroni *et al*. (1993). Early studies were focused on Native American and Eurasian populations, thus by the time African lineages were systematically studied, few alphabets were left to define the African lineages, which is why deep-rooting human mtDNA clades are called L. In 2009, a global human mtDNA phylogenetic tree was built using the haplogroup nomenclature and published complete mtDNA genome variations (van Oven & Kayser 2009). Available online (www.phylotree.org), the tree has since been used by many studies as a reference and is constantly updated to include newly available complete mtDNA sequences. The most recent version is Build 17 (Figure 2.6), which now contains nearly 5,500 haplogroups built by over 30,000 published sequences (van Oven 2015).

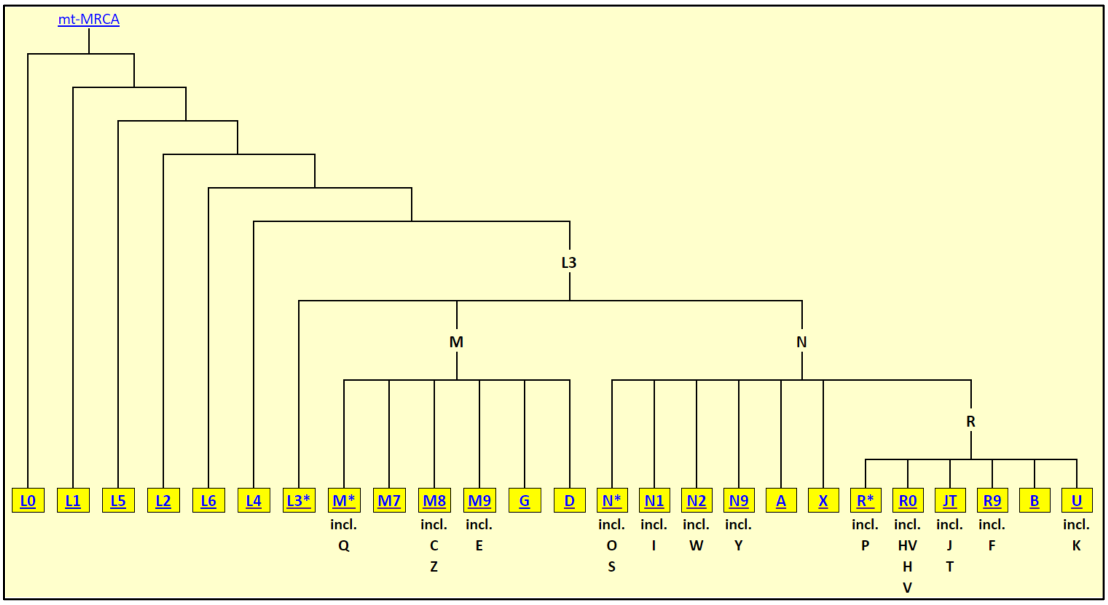

Figure 2.6 Simplified view of the global human mtDNA phylogenetic tree (Build 17, 18 February 2016). mt-MRCA (top left corner) represents the most recent common matrilineal ancestor of humans. Alphabets (e.g. M, N, R etc.) and alphanumerics (e.g. L3, M7, N1, etc.) represent haplogroups. Haplogroups followed by an asterisk (*) (e.g. M*, N* etc.) represent all other descendant lineages of a particular clade except for the ones shown. For example, N* consists of haplogroups N3, N5, N7, N8 etc. (After www.phylotree.org).

2.2.5 Mutation rates

As mentioned earlier, the mtDNA has a mutation rate that is approximately 10 times higher than that of the nDNA. Over the years, researchers have proposed various substitution and mutation rates for the mtDNA (Table 2.1). Mutation and substitution rates differ in that mutation rate is the rate at which mutations occur over time, whereas substitution rate is the rate at which these mutations persists or are fixed in a population (Ho & Larson 2006; Barrick & Lenski 2013). Because mutations tend to be eliminated due to purifying selection or drift, few changes are observed over time, resulting in a slower substitution rate compared to the mutation rate (Ho & Larson 2006).