

https://dspace.org/

Reconstructing the prehistory of the malagasy

Patel, Pareen

2016-10-17
http://hdl.handle.net/10539/21225

Downloaded from a DSpace repository. https://dspace.org/

Reconstructing the prehistory of the Malagasy



Pareen Patel

A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment for the requirements for the degree of Master of Science in Medicine.

Johannesburg, 2016

Declaration

I declare that this dissertation is my own, unaided work. It is being submitted for the degree of Master of Science in Human Genetics at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examinations at any other university. I declare that this work has been approved by the Ethics Committee of the University of the Witwatersrand for Research on Human Subjects, and the certificate numbers are MO90576 and MO90629

- - - - - - - - - - -
Pareen Patel

- - - - - day of February 2016

Abstract

The origins of the Malagasy have been reasonably well documented using historical, cultural and linguistic data. Genetic studies, albeit few, have made use of different types of markers [autosomal DNA, Y chromosome DNA and mitochondrial DNA (mtDNA)], to shed light on the peopling of Madagascar. In this study mtDNA was used to trace the maternal ancestry of 981 Malagasy from different regions in Madagascar. All individuals were screened for the intergenic COII/$tRNA^{Lys}$ 9-bp deletion followed by hypervariable region DNA sequencing in conjunction with phylogenetically informative single nucleotide polymorphism (SNP) typing. The 9-bp deletion occurred at a frequency of 22.73% (223/981) of which 17.04% (38/223) was traced to an African origin, 82.96% (185/223) to Asian origins. Individuals with the Asian form of the deletion who harboured the substitution that defined the "Polynesian motif" were also screened for two additional substitutions at positions 1473(C→T) and 3423(T→A) that define the "Malagasy motif". The Malagasy motif was found in 61.43% (137/223) of individuals with the deletion. Overall, SNP data in conjunction with hypervariable region sequence data delineated 295 unique mtDNA sequences in the sample of which 41.28% (405/981) were traced to African ancestry and 58.71% (576/981) to non-African origins. In addition to the mtDNA haploid marker system, this study has for the first time made use of a 96 SNP ancestry informative marker panel to further shed light on the autosomal contributions to the Malagasy. This study corroborates the findings from other lines of evidence like historical, linguistic and archaeological data concerning the primary/parental origins of the Malagasy. However, these findings have provided a refinement in the data showing that the female gene pool has an appreciably higher contribution from non-African sources of origin compared with African origins. This is in contrast to what is observed using Y chromosome data.

Acknowledgements

I would like to thank:

The subjects who participated in this research project

The National Health Laboratory Service, Genographic Project and the Medical Research Council for funding this work

Professors Trefor Jenkins, Gwyn Campbell, Himla Soodyall, J. Roux, Thomy de Ravel, Doctors Pascal Willem, A. Rakouth and S.Quambec for collecting and processing the samples.

My supervisor, Professor Himla Soodyall, for allowing me this opportunity and for her mentorship and support.

Professor Michele Ramsay for allowing me the use of the BeadXpress equipment.

The members of the Human Genomic Diversity and Disease Research Unit for your friendship and assistance throughout this process.

Carina Schlebusch, for your encouragement and invaluable help.

My husband, mother, father and brother for their love, encouragement and support.

For Zara and Mila

“Every new beginning comes from some other beginning’s end”

-Semisonic

TABLE OF CONTENTS

DECLARATION II
ABSTRACT III
ACKNOWLEDGEMENTS IV
DEDICATION V
TABLE OF CONTENTS VI
TABLE OF CONTENTS (CONT.) VII
TABLE OF CONTENTS (CONT.) VIII
LIST OF FIGURES IX
LIST OF TABLES X
LIST OF ABBREVIATIONS XI
LIST OF ABBREVIATIONS (CONT.) XII
INTRODUCTION
1.1. Geography of Madagascar 1
1.2. The peoples of Madagascar 3
1.2.1. Present-day Malagasy population 3
1.2.1.1. Highland and Lowland Groups 4
1.3. Ancestry of the Malagasy 7
1.3.1. First settlements and possible migratory routes 7
1.3.2. Oral History of the Malagasy people 11
1.3.3. Archaeological Evidence 11
1.3.4. Linguistic Evidence 12
1.3.5. Physical Appearance 13
1.3.6. The culture of the Malagasy 15
1.4. Historical Evidence 18
1.4.1. Written History 18
1.4.2. Slave Trade 18

TABLE OF CONTENTS (CONT.)

1.5. Reconstructing human history using a molecular genetic approach 19

1.6. Aims and Objectives 26

MATERIALS AND METHODS

2.1. Sample collection and study subjects 28

2.2. Molecular Methods 31

2.2.1. Nucleic Acid Extraction and Quantification 31

2.2.2. MtDNA methods 31

2.2.2.1. 1kb mtDNA Hypervariable Region Polymerase Chain Reaction and Sequencing 32

2.2.2.2. Mitochondrial Single Base Extension (SBE) Assay 36

2.2.2.2.1 Multiplex PCR amplification 36

2.2.2.2.2 SBE reaction 37

2.2.2.2.3.TaqMan® SNP Genotyping Assay 38

2.2.2.3. PCR of the Intergenic COII/tRNA^{Lys} 9-bp deletion 39

2.2.2.3.1. Sequencing for the “Malagasy Motif” 41

2.2.2.3.1.1. Primer design 41

2.2.2.3.1.2. 2kb PCR Optimisation 42

2.2.2.4. Autosomal DNA genotyping by use of BeadXpress technology 42

2.2.2.5. mtDNA Data Analysis 44

RESULTS

3.1. Mitochondrial DNA Results 49

3.1.1. Intergenic COII/tRNA^{Lys} 9-bp deletion 49

TABLE OF CONTENTS (CONT.)

3.1.2. Analyses of variation in mtDNA hypervariable regions.....55
3.1.2.1. Haplogroup assignment and structure.....55
3.1.3. Variation within twenty one Malagasy ethnic groups……..……..…64
3.1.4. Genetic affinity of the Malagasy in relation to global population……..…75
3.2. Autosomal DNA results…….....89
3.2.2. SNP Validation……...90
3.2.3. STRUCTURE Analysis……..90
4. DISCUSSION……...102
5. CONCLUSION……112
6. REFERENCES……113
7. APPENDICES…….....129

LIST OF FIGURES

Figure 1: Location of Madagascar.. 2
Figure 2: Ethnic group distribution on Madagascar 4
Figure 3: Computer generated simulations of voyaging displaying directed sailing from the Sunda straights. 10
Figure 4: Picture of a farmer using Zebu cattle to plough rice paddies 17
Figure 5:Picture of an out-rigger canoe 17
Figure 6: Global distribution pattern (excluding North and South America) of the intergenic COII/tRNALys 9-bp deletion 522
Figure 7: Distribution patterns of various forms of the intergenic COII/tRNALys 9-bp deletion within different regions of Madagascar. 53
Figure 8: Phylogeny representing mtDNA haplogroups and sub-haplogroups and their relative frequencies in Madagascar 56
Figure 8 (cont): Phylogeny representing mtDNA haplogroups and sub-haplogroups and their relative frequencies in Madagascar 56
Figure 9: Pie charts of Asian, Eurasian and African contributions to the maternal gene pool of the Malagasy....58
Figure 10: Haplogroup frequency maps of African haplogroups within Madagascar 60
Figure 11: Haplogroup frequency maps of Asian haplogroups within Madagascar 61
Figures 12a and b: Neighbour joining tree representing 295 unique sequences without (Figure 12a) and with (Figure 12b) mtDNA hypervariable region SNPs. The phylogenies are each rooted with four Neanderthal mtDNA sequences. 63
Figure 13: Bar chart representing the mtDNA composition of 21 Malagasy ethnicities 65
Figure 14: Neighbour joining phylogeny of 21 Malagasy ethnicities. 66
Figure 15: Mismatch distribution patterns for some African and Asian mtDNA haplogroups 68
Figure 16: Mismatch distribution patterns for each of the 21 ethnic groups in Madagascar 73
Figure 17a: Median joining network representing mtDNA haplogroup L0a2*. 79
Figure 17b: Median joining network representing mtDNA haplogroup L3b* 80
Figure 17c: Median joining network representing mtDNA haplogroup L2a1* 81
Figure 17d: Median joining network representing mtDNA haplogroup L3e 82
Figure 18a: Median joining network representing mtDNA haplogroup B4a* 83
Figure 18b: Median joining network representing mtDNA haplogroup M* 84
Figure 18c: Median joining network representing mtDNA haplogroup F* 85
Figure 19: MDS plot of Madagascar (MAD_PRESENT) in relation to global populations....87
Figure 20: STRUCTURE results for K=2 to K=5 and various solutions at levels of K.. 94
Figure 21 : STRUCTURE results for K=6 to K=8 and various solutions at levels of K. 95
Figure 22 : STRUCTURE results for K=9 to K=10 and various solutions at levels of K.. 95
Figure 23 : Supervised STRUCTURE results for K=6 95
Figure 24 : Neighbour joining tree created using 96 AIM autosomal SNP data. 99
Figure 25 : Principal component analysis of 96 AIMs 100

LIST OF TABLES

Table 1: Tabulating the breakdown of ethnic groups within various regions of Madagascar. 30
Table 2: Primer sequences used for amplification of the mtDNA 1.1kb hypervariable region and for the subsequent cycle sequencing reactions. 32
Table 3: The general PCR and thermal cycling protocol used to set up the amplification system for the 1.1kb hypervariable region 33
Table 4: Primer sequences used for amplification of the 9-bp deletion in the mtDNA COII/tRNALys intergenic region 40
Table 5: Primer sequences used for amplification and sequencing of the Malagasy motif 42
Table 6: Details of the intergenic COII/tRNALys 9-bp deletion within Madagascar 54
Table 7: Mismatch distribution statistics for various mtDNA haplogroups within Madagascar....68
Table 8: Diversity statistics for various mtDNA haplogroups within Madagascar 70
Table 9: Mismatch distribution statistics for various ethnic groups within Madagascar 72
Table 10: Diversity statistics for various ethnic groups within Madagascar 74
Table 11: Results from mtDNA AMOVA using different groupings 88

LIST OF ABBREVIATIONS

A	Adenine
ATP	Adenosine Triphosphate
AD	Anno Domini
BC	Before Christ
BSA	Bovine Serum Albumin
COII	Cytochrome Oxidase Subunit II
µl	Microlitre
ACD	Acid citrate dextrose
bp	Base pair
^0C	Degrees centigrade
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
ddNTP	Dideoxyribonucleotide triphosphate
D-loop	Displacement loop
dH_20	Distilled water
ddH_20	Distilled deionised water
EDTA	Ethylene-diamine-tetra-acetic acid
EtBr	Ethidium bromide
g	Gram
G6PD	Glucose-6-phosphate Dehydrogenase
H	Heavy strand
HVR1	Hypervariable region 1
HVR2	Hypervariable region 2
HVR3	Hypervariable region 3
NHLS	National Health Laboratory Service
NJ	Neighbour joining
MEGA	Molecular Evolutionary Genetic Analysis

LIST OF ABBREVIATIONS (CONT.)

mtDNA	Mitochondrial DNA
PCR	Polymerase chain reaction
PCA	Principal component analysis
L	Light strand
rpm	Revolutions per minute
TBE	Tris Borate EDTA
TE	Tris EDTA
UV	Ultraviolet
V	Volts
R	Reverse
RE	Restriction enzyme
RFLP	Restriction fragment length polymorphism
s	Seconds
Seq	Sequence
SNP	Single nucleotide polymorphism
Taq	Thermus aquaticus
TBE	Tris borate-EDTA
U	Units

Chapter 1

Introduction

Genetic variation through mutation is passed on to each individual, through time, by our ancestors. Most individuals display genetic variation which is sometimes reflected at a population phenotypic level. The advancement of molecular technology has provided powerful tools to examine inter-individual genetic variation and in reconstructing human history. In this study, aspects of the genetic history of the Malagasy will be uncovered using molecular tools and contribute to the storybook of their genetic history.

The origins of the Malagasy have been reasonably well documented using historical, cultural and linguistic data (see sections 1.3 and 1.4).Genetic studies, albeit few, have made use of different types of markers (see section 1.5) to shed light on the peopling of Madagascar. These studies have suggested that Southeast Asian and African populations have contributed predominantly to the gene pool of the islanders. In the present study, mtDNA will be used to evaluate the role that females have played in shaping the gene pool of the Malagasy. Furthermore, a panel of 96 ancestry informative markers (AIMS) will be used to determine the autosomal contributions to the Malagasy. To better understand the genetic data, some contextualization of Madagascar, its people and history is necessary.

1.1. Geography of Madagascar

The present-day island of Madagascar was initially a part of the Gondwana supercontinent. The western coast of Madagascar formed 160 million years ago when Africa separated from the Gondwana supercontinent (Karanth, 2006). Madagascar lies at a latitude of 20° South and a longitude of 47° East within the Indian Ocean (Figure 1) and is separated from the south-eastern coast of Africa by the 400 kilometre wide Mozambique Channel and lies 5600km from

the Indonesian archipelago (Fitzpatrick and Callaghan, 2008). Expanding over sixteen hundred kilometres long and having a land surface approximately the size of France; Madagascar is classified as the world's fourth largest island, after Greenland, New Guinea and Borneo (Dewar, 1995).

The terrain in Madagascar is marked by a spine of mountains that runs from north to south of the island. These mountain ranges divide the island into various climatic and geographic regions. The North is isolated from the rest of the island by the Tsaratanana mountain range. In the east, a narrow strip of low-elevation forests, referred to as the lowlands or coastal region and in the west a wider section referred to as the high plateau. The interior mountainous region is referred to as the central highlands (Dewar and Wright, 1993). These natural boundaries have created different geographic regions and have assisted in developing differing lifestyles between groups that inhabit these regions (Brown, 1978; Mack, 1986).



Figure 1: *Location of Madagascar*: Map indicating the location of Madagascar within the Indian Ocean. Madagascar lies 400km off the coast of Africa and 5600km from Indonesia Image taken from: Fitzpatrick and Callaghan, 2008.

1.2. The peoples of Madagascar

1.2.1. Present-day Malagasy population

In 2010, the population of Madagascar was estimated at 20 million people. Following the integration of Bantu-speakers from Africa and Austronesian-speakers from 21 loosely defined ethnic groups were formed based on their ethnic, tribal, geographic and political affiliations (Bradt, 2011).

The present-day Malagasy population is quite admixed, displaying cultural and phenotypic traits of Austronesian and Bantu speakers. Malagasy ethnic diversity is broadly classified into Highland and Lowland groups (Blench, 2007). The Highland group is comprised the Merina, Betsileo, Sihanaka, Tanala and Bezanozano (Figure 2). These peoples are deemed to be the more “Asian” of the two groups based on their phenotypic straight black hair and lighter skin and cultural similarities to those in the East. The second group is the Lowland group, which is comprised Antanosy, Antandroy, Antaisaka, Antaifasy, Antaimoro, Antambahoaka, Antankarana, Bara, Betsimisaraka, Makoa, Mahafaly, Vezo, Mikea, Sakalava, Tsimihety and Zafisoro (Figure 2). These people are described to be “African” phenotypically (Tofanelli et al., 2009).

The population numbers for each ethnic group varies throughout the island (Vérin et al., 1969). The Merina, Betsimisaraka and Tsimihety represent the largest frequency of the ethnic groups: 26.10%, 14.90% and 7.00% respectively. The remaining ethnic groups make up less than 10.00% each of the total population.

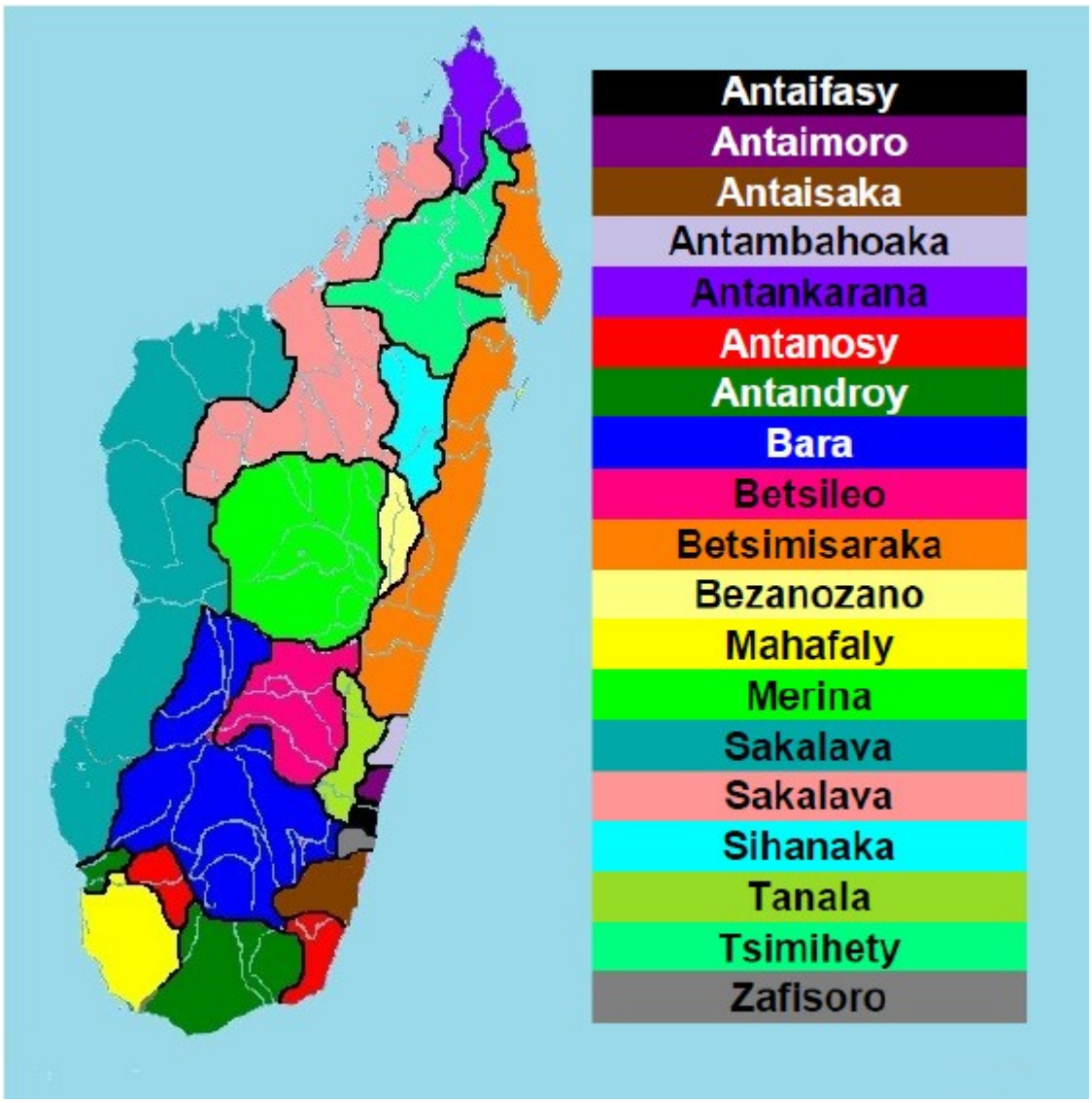


Figure 2: *Ethnic group distribution*. Map of Madagascar showing the distribution of 18 ethnic groups on the island. The Highland group is comprised the Merina, Betsileo, Sihanaka, Tanala and Bezanozano. . The Lowland group is comprised Antanosy, Antandroy, Antaisaka, Antaifasy, Antaimoro, Antambahoaka, Antankarana, Bara, Betsimisaraka, Makoa, Mahafaly, Vezo, Mikea, Sakalava, Tsimihety and Zafisoro
* Map modified from that found on
http://en.wikipedia.org/wiki/File:Ethnic_groups_of_Madagascar_Map.png

1.2.1.1. Highland and Lowland Groups

- Betsileo: These people inhabit the southern parts of the high central plateau. They are phenotypically said to have both African and Malay features (Bradt, 2011).

- Bezanozano are believed to be one of the earliest Malagasy ethnic groups to establish themselves in Madagascar. Their name means "those of many small plaits" in reference to the traditional style in which they wear their hair (Bradt, 2011).

- Merina: These people inhabit the central highlands of Madagascar. They are thought to be descendents of legendary (through oral history) dwarf ancestors called the Vazimba, Malays and Javanese peoples (Serva et al., 2012). The Merina are the largest of all Malagasy ethnic groups and are further subdivided into three groups, (a) Andriana nobles (b) Hova free men or commoners and (c) Mainty, Mpanompo or Andevo slaves. During the nineteenth century the Merina ruled over a third of Madagascar (Brown, 1978; Campbell, 1989).

- Antaimoro: This group of people live on the south-eastern coast of Madagascar. They have a notable Arabic cultural influence, which is demonstrated by their dress choice of Arab robes, turbans and the fez (Bradt, 2011).

- Antaisaka: These people occupy the southern parts of Madagascar and are thought to have high levels of African and Malay admixture (Bradt, 2011).

- Antankarana: This group of people are found in the northern regions of Madagascar and have cultural similarities with that of Arabic culture (Bradt, 2011).

- Betsimisaraka: These people inhabit a narrow coastal strip on the eastern part of the island. They are the second largest group and are thought to be more similar to the Merina than other lowland ethnicities (Campbell, 1996).

- Sihanaka are concentrated around Lake Alaotra and their means the "people of the swamps" in reference to the marshlands that they inhabit (Bradt, 2011).

- Tsimihety: These people are located on the northwest coat of the island and it is suggested that they could well represent the proto-Malagasy people that initially settled in the north (Brown, 1978).

- Antandroy: These people are a semi-nomadic group that inhabit the southern parts of Madagascar (Bradt, 2011).

- Mikea: These people occupy the southwestern regions of Madagascar. They have a mixture of Indonesian and African features. This group can be further separated into two groups; the western group speak the Vezo dialect and the eastern group the Masikoro dialect (Bradt, 2011).

- Mahafaly: This group of people are inhabitants of the southwest coast of Madagascar. They have features and cultural practices that are African influences (Bradt, 2011).

- Bara: The Bara people live in the southern part of Madagascar and are estimated to account for 3% of the Malagasy population (Bradt, 2011).

- Sakalava: This group of people are inhabitants of the west coast. Their customs, dialect and physical appearance shows a resemblance to African customs. The Sakalava are also divided into two groups: the coastal Vezo fisherman and the agro-pastoral Masikoro who inhabit the interior regions of the west coast (Bradt, 2011).

- Vezo: These people are a semi-nomadic, coastal population that inhabits the southern parts of Madagascar. Their name stems from “people who fish” or 'to struggle with the sea' (Bradt, 2011).

- Makoa: These people live in western Madagascar and are thought to be descendents of African slaves who were brought to Madagascar (Bradt, 2011).

- Antanosy: These people mainly live in the Anosy region of south-eastern Madagascar and are among the smallest in both population and geographical range.

- Antaifasy: These people inhabit the southeast coastal region and their name means "people of the sand" (Bradt, 2011).

- Antambahoaka: These people form part of the smallest ethnic group in Madagascar. They inhabit a small region along the south-eastern coast of Madagascar (Bradt, 2011).

- Zafisoro: These people inhabiting a portion of the south-eastern coast of Madagascar.

- Tanala inhabit a forested region in the south-eastern parts of Madagascar. Their name means "people of the forest" (Bradt, 2011).

1.3. Ancestry of the Malagasy

1.3.1. First settlements and possible migratory routes

Since the sixteenth century A.D. Europeans have attempted to elucidate the ancestry of the Malagasy; however this process has been hampered by the lack of written records. Literature on the origins of the Malagasy dates back to the sixteenth century and various theories exist on who were the first people of the island. The first groups to inhabit Madagascar are thought to be the Vazimba, Behosy, Kimosy and Mikea. Other inhabitants are thought to be of either African descent, brought to Madagascar by Indonesian and Swahili traders (Ferrand, 1908; Vérin, 1986) or of proto-Malagasy immigrants of Austronesian descent (Grandidier and Grandidier, 1908; Brown, 1978).

Four theories regarding the origins of the Malagasy have been proposed:

a) The first theory suggests that the people of Madagascar were an autochthonous (native) population (Dewar and Wright, 1993).

b) The second theory suggests that Madagascar was originally settled by people of African origin who were later conquered by Indonesian migrants (Ferrand, 1908). This theory is more feasible than the first given the proximity of Africa to Madagascar; however it lacks conclusive evidence of long distance maritime voyages away from the East African coast (Brown, 1978).

c) The third theory suggests that migrations to Madagascar initially went via Africa. The proto-Malagasy group were therefore comprised genetically mixed groups of Indonesians and Africans. (Ferrand, 1908; Grandidier, 1908).

d) The fourth theory postulates that the initial settlers of Madagascar were migrants from the Malayo-Indonesian-Polynesian region with a lesser contribution of Indian and Semitic people and a much later minuscule contribution of people from Africa. The presence of people from the Malayo-Indonesian-Polynesian region in East Africa during the first millennium A.D. has been taken as evidence against this hypothesis (Murdoch, 1959).

Of these, the third and fourth hypotheses seem more plausible. With these theories two migratory paths have been proposed as the routes which the proto-Malagasy could have used to reach Madagascar (Ferrand, 1908; Vérin et al., 1969). The first one suggests that the proto-Malagasy sailed directly across the Indian Ocean from Java to the eastern coast of Madagascar

in outrigger canoes and the second suggests an indirect route via the Indian sub-continent, Arabian Peninsula along the East African coast then on to Madagascar.

For the first theory to seem plausible, many individuals would have had to migrate over a distance spanning 6000km across the Indian Ocean, which would have been fifteen times further than from East Africa (Fitzpatrick and Callaghan, 2008). Heyerdahl (1971) demonstrated that it is possible to sail on a log raft and travelled 8000km across the Pacific Ocean from South America to the Tuamotu islands in 101 days.

There is a second theory whereby Indonesian mariners would have visited the East African coast before definitely settling in Madagascar (Blench, 1996; Adelaar, 1989). These mariners were thought to follow trade routes along the northern and western coasts of the Indian Ocean, where they would have set up posts and possibly would have inter-married with local East African populations. The initial wave of migrations into Madagascar would have been caused by Indonesian groups in East Africa being removed by the Bantu expansion. A second wave of migration onto the island would have occurred during the ninth to fourteenth centuries whereby Muslim Arab and Swahili traders would have forced settlers from Somalia through to Mozambique onto the Comorian islands and Madagascar (Stiles, 1991). In a recent study conducted by Fitzpatrick and Callaghan, using computer simulations of voyaging, these authors were able to determine that vessels sailing from the Sunda straights during the month of March would have the highest likelihood of passing by the northern coast of Madagascar or landing on the African coast (Figure 3). The authors suggested that the average duration of the voyage would have been 152 days (Fitzpatrick and Callaghan, 2008).

The initial settlers on Madagascar were thought to have a greater Indonesian heritage, whereas later settlers were more African-like. In addition, it has been suggested that the primary

colonists on the island were forced to move further inland by the successive waves of migration on Madagascar. This would have resulted in the Indonesian-like settlers to inhabit the high plateau and the later more African-like settlers to occupy the lowland regions (Campbell, 1996). The second hypothesis seems more plausible due to the shorter trips along the coasts of the Indian Ocean and it is also known that these Austronesian mariners were active traders between the first and fifth centuries A.D. Furthermore Arab and Chinese records indicate that Indonesians had been trading and living in East Africa and Madagascar during this time (Campbell, 1996).

Evidence of recent genetic influences from pirates, traders, slaves, captives and the colonists (African, Indian, Arabian, Portuguese, French, Dutch and British) has been historically documented for both Highland and Lowland groups (Kent, 1962, 1970; Dewar and Wright, 1993). The present-day Malagasy population is described as the "progeny of the Indian Ocean" (Dewar and Wright, 1993). Colonization of the island took place in multiple waves of migration (Campbell, 1988). The import and export of slaves from Madagascar in addition to the movements of traders and migrants would have substantially influenced the gene pool of the Malagasy people (Campbell, 1988).

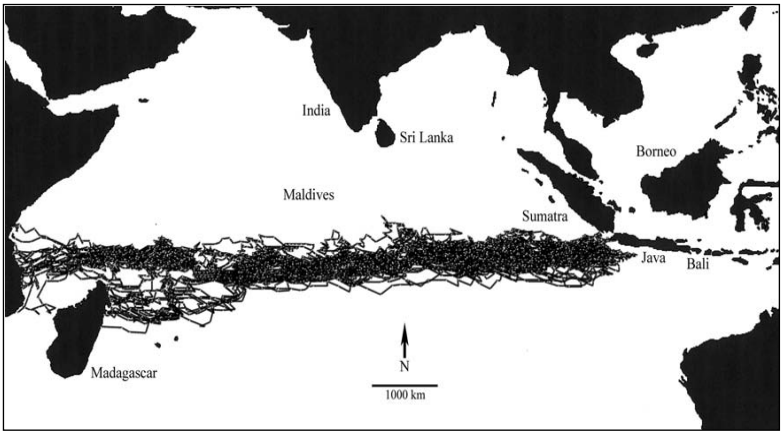


Figure 3: *Hypotheses of several possible routes taken to reach Madagascar*. Computer generated simulations of voyaging displaying directed sailing to Madagascar from the Sunda straights. These simulations took into account: 1) Wind patterns, 2) vessel type, 3) current and 4) propulsion (image taken from: Fitzpatrick and Callaghan, 2008).

1.3.2. Oral History of the Malagasy people

Various sources of information concerning Malagasy origins include oral traditions that were passed down by Merina rulers. A tradition amongst the Antaimoro says that they descend from a group of thirty men who sailed to Madagascar directly from Mecca (Campbell, 1996). However, there is no archaeological evidence or any other form of evidence to suggest that there were direct migrations from Arabia or India to Madagascar. Oral traditions also mention an early group of dwarfs, however there is no evidence on the island to support this theory (Campbell, 1996).

1.3.3. Archaeological Evidence

Historical traditions form an important part of all Malagasy groups. Material testimonies including tombs and former villages serve as evidence for these traditions. The first recorded archaeological findings were made by a Merina historian, Raombana in 1835 and the first systematic archaeology began at the University of Madagascar in 1966 (Dewar and Wright, 1993).

Contrary to some reports, no Stone Age archaeological sites have been discovered on the island of Madagascar, Dewar suggests that even the earliest settlers were fully reliant on iron tools. The only stone tools found were gunflints, weight and net sinkers (Dewar, 1995).

Recent palynological and paleoecological studies have documented the history of vegetation on the island and has yet to yield evidence of human impact on Madagascar that is more than 1900 years old (Burney 1987; Burney et al., 2004). Initial human activity appeared in the first millennium A.D. at two coastal marshes along the south-western coast. This theory is based on extinct pygmy hippopotamus femoral bones that show marks of being cut by metal tool. The

bones were radiocarbon date to the first and fourth centauries A.D. (MacPhee and Burney, 1991).

In 1986 a team of archaeologists discovered a number of rock shelters and small caves, indicative of permanent human occupation on the island. Since this time, sites have been found containing Portuguese jars, Persian glassware, Middle Eastern ceramics, bead necklaces, bronze mirrors and Chinese ceramics, among various other items. Archaeological sites have yielded trade ware that ranges from silver and gold, ceramic vessels to wooden items indicative of a vast trade network (Dewar and Wright, 1993).

Cattle-herding settlements and agricultural villages dated to the eleventh century have been found in the south and south eastern regions of Madagascar. Following the 350 years after 1000 A.D., there is strong evidence that the Malagasy partook in growing rice, herding cattle, fishing, smelting iron and were even active in long-distance trade (Dewar and Wright, 1993).

1.3.4. Linguistic Evidence

Malagasy is the official language of Madagascar and is spoken throughout the island. It is the only language in the African region that belongs to the Austronesian language family. This group includes from 600 to 1000 languages that are spoken in Malaysia, Indonesia, Melanesia, Micronesia, Polynesia, Cambodia, southern Vietnam and Taiwan (Serva et al., 2012).

Since 1603, similarities have been noted between Malagasy and languages spoken in Indonesia (Vérin et al., 1969). The surviving languages (Maanyaan and Ngauju) that are most closely associated with that spoken in Madagascar are found in the Barito Valley in Borneo (Dahl 1951, 1977). There are also elements of Bantu languages found within the morphology, phonology and vocabulary of Malagasy (Dahl, 1988).

There are various regional dialects of Malagasy, each of which have numerous foreign words from English, French, Arabic and Swahili within their vocabulary (Beaujard, 2003). Vérin et al. (1969) proposed that various Malagasy dialects have been diverging for approximately 2000 years and that after the inhabitation of Madagascar; the ancestral group was divided into three groups. One group gave rise to the Antankarana, the next to Tsimihety and the third, the remainder of dialects. The first two groups remained in the north, whereas the third group split into two, one giving rise to dialects in the east and central regions and the second to dialects in the west and south. The theory proposed by Vérin et al. (1969) has been challenged by Noël Gueunier (1988), who showed that the speech patterns on the island reflect clines that may not follow the geographical ones and that the similarity of neighbouring groups may be a result of common ancestry or due to groups of diverse ancestry that have converged over time.

Luis Mariano described two languages in Madagascar; a language that resembled Malay in the interior regions of Madagascar and along the southern, eastern and western coasts and an African language along the north-western coast (Mariano, 1904). There are also non-Malagasy speaking communities on the western coast of Madagascar. Present-day linguistic diversity is far more reduced and there are only a few bilingual communities on the western coast where the Bantu language of Makoa is spoken by recent immigrants and in the northwest, communities that speak both Malagasy and Swahili (Dewar, 1995). It is not yet known as to how or when these influences took place (Adelaar, 1989).

1.3.5. Physical Appearance

In contrast to their language unity, the Malagasy are phenotypically significantly diverse. From as early as the seventeenth century, European observers described “racial” differences amongst the people of Madagascar. Certain populations were described as being “Malay” with

regard to skin colour, hair texture and stature, other Malagasy groups were said to be darker, taller and having curly hair texture. In 1550 and 1661, respectively, both Nagnort and Flacourt reported on physical differences among different groups on the island, labelling them as either black or “les Noirs” and whites or “les Blancs” (Nagnort, 1550; Flacourt, 1661). Grandidier suggested that the physical features, especially that of skin colour was due to the advent of lighter-skin Malays and darker-skinned Melanesians and later arrival of some Africans (Grandidier, 1904). In contrast, Ferrand suggested the darker-skinned peoples to be primarily of African origins (Ferrand, 1908).

Singer et al. (1957) suggested that the present-day Malagasy population has a heterogenous physical appearance with varying degrees of admixture between Asian and African features. Coastal populations of the west and south are said to have African-features with darker skins and short, curly hair whereas highland populations have lighter skin and straight hair. The Merina population, however are said to have a mixture of both African and Asian physical traits (Singer et al., 1957). Buettner-Janusch and Buettner-Janusch (1964) suggested that the highland groups had a different ancestral population from that of the coastal peoples. This hypothesis implied that there was more than one migration to the island. In addition the authors suggested that the original settlers inhabited the highland regions and the groups that formed here forced neighbouring groups to the coastal regions of Madagascar where interactions with peoples from Africa occurred. Interesting as the differences and similarities of Malagasy physical features may be, phenotypic data is not a dependable tool for determining variations between populations.

1.3. 6. The culture of the Malagasy

Malagasy culture is complex and varies greatly regionally. This is consistent with the linguistic, archaeological and physical aspects with influences from both African and Asian sources but also European and Indian contributions (Middleton, 1999).

The influence of Indonesian culture is seen by the Malagasy terminology and techniques used for slash and burn agriculture, the cultivation of rice, taro and bananas. The tube zither, double tuyere bellows for iron smelting and the presence of certain taboos are all indications of Indonesian influences. Dwelling types found in Madagascar are rectangular, wooden huts, which are built on silts. This architectural style is commonly found throughout parts of Asia (Vérin and Wright, 1999). Malagasy implements for pounding rice and some techniques in rice preparation are linked to Indonesian customs (Vérin and Wright, 1999). The cuisine of Madagascar represents the culmination of cultures with Indian, French and Chinese influences.

The introduction of Zebu cattle provides evidence of African contribution to Madagascar. Zebu cattle are known to have descended from the secondary cattle domestication in the 'Fertile Crescent' about 5000 BP (Payne and Wilson, 1999). Cattle herding forms part of the main set of activities of the Sakalava, Bara and Antandroy ethnic groups (Figure 4). Fiber weaving and some burial customs are also linked with African sources (Bradt, 2011).

Malagasy settlers have adopted domestic and agricultural techniques from the entire range that are available. They have shown that material culture do not always map cultural markers. This is clearly indicated by the use of outrigger canoes (Figure 5), which are undoubtedly of Indonesian origin; however are found mostly on the western coast, which is conversely the region that shows strongest African influence (Dewar and Wright, 1993). This may be an

indication that Malagasy living in various regions of the island have adapted to their differing environments by selecting from a panel both Asian and African cultural traits.

Although many features of Malagasy culture, crops and domesticated animals have been linked to Southeast Asian, African, the Near East, European and Indian origins the diversity of these origins reminds us that the present-day Malagasy culture is a product of interaction throughout the Indian Ocean, and not only a combination of Southeast Asian and African origins (Dewar and Wright, 1993).



Figure 4: *Example of African influence on Malagasy culture*. Picture of a farmer using Zebu cattle to plough rice paddies
. Image taken from: http://mongabay.s3.amazonaws.com/madagascar/600/madagascar_4773.jpg



Figure 5: *Example of Asian influence on Malagasy culture*. The out-rigger canoe displayed above is typical of those found in Indonesia
Picture taken from: http://www.peacecorps.gov/wws/educators/enrichment/africa/countries/madagascar/MG0617.jpg

1.4. Historical Evidence

1.4.1. Written History

Due to the lack of written records, scholars attempting to elucidate the ancestry of the Malagasy have been hampered. One of the main sources of historical information would have come from oral accounts passed on by the Malagasy from their rulers since 1000 A.D. (Mourant et al., 1976).

Thus far, there are three sources of written historic records. The first being accounts written by traders and missionaries. Some researchers favoured a theory which suggested that light-skinned Malagasy were descendents of Indonesian settlers, whereas darker-skinned peoples were descendents of Melanesians (Grandidier and Grandidier, 1908). Others favoured a mainly African ancestry, which had contributions from Swahili traders (Ferrand, 1908; Vérin, 1986).

Arab colonies founded in the northwest of the island in the thirteenth century (Brown, 1978) include the second type of written record, which includes ancient Malagasy texts called the Sorabe written in Arabic script. These scripts describe a potential Arabian ancestor. Further documents complied by Père Callet (a Jesuit priest) and a Malagasy historian, Raombana provide the third source (Mack, 1986).

1.4.2. Slave trade

During the nineteenth century Madagascar featured in the slave trade in the Western Indian Ocean. At this time, slaves made up the main item for foreign-trade. Slaves were being exported from Madagascar to Europe, India, the Americas, the Mascarene islands, South Africa and the Middle East. Slavery on the island was so widespread that slaves were progressively replacing the free Merina population of the central highlands. Slaves were not only brought to

Madagascar but were also imported. To supply the demand for the vast slave-trade network, Arab traders took people from the eastern coastlines of Africa to Madagascar as slaves. With an increasing demand for slaves in Europe, Portuguese and Indian traders came to the island in search of supplying this demand. The prohibition of slave exports was brought about in 1820 when the British-Merina treaty was established and the ruler of Madagascar at that time, Radama I, pledged an end to the exportation of slaves. In 1896, after the effective military occupation of Madagascar by the French, the Malagasy slave trade was brought to a halt during the early years of the twentieth century (Campbell, 1981).

1.5. Reconstructing human history using a molecular genetic approach

Advances in human population genetics have allowed us to assess ways in which the gene pool of present-day people has been shaped by past historical and demographic events. The genetic approach is to examine the genetic patterns in living people and then using computational analyses it is possible to reconstruct the history of changes/mutations from the contemporary divergent lineages. Furthermore, we are able to establish relationships between lineages date these lineages and relate all lineages to the most recent common ancestor. DNA markers are powerful tools in distinguishing between individuals (Shriver et al., 1997). This study takes advantage of the discrimination power of the haploid marker, mtDNA as well as autosomal DNA markers to shed light on the possible origins of the Malagasy.

Unlike historical, linguistic and archaeological data, the use of genetic data in uncovering the origins of the Malagasy is limited to a few studies. Different types of data based on autosomal DNA, mtDNA and Y chromosome DNA markers have indicated varying proportions of African, Asian or Pacific contributions to the gene pool of the islanders (Hurles et al., 2005).

Early studies made use of blood group markers and the sickle cell trait to infer ancestry. In one such study Singer et al. (1957) found that two thirds (62-67%) of the Malagasy had a genetic heritage of African origins and the remaining third (33-38%) had a genetic heritage of Indonesian origins. In addition, this group found the sickle cell trait to occur at a frequency of 4.90% within this population and was found to occur at higher frequencies within coastal groups (Buettner-Janusch and Buettner-Janusch, 1964). Since these studies, more refined methods have estimated the haemoglobin S frequency to be to 10.70% and 8.20% (Fourquet et al., 1974). Singer et al. (1957) proposed that the sickle cell trait had first been introduced to Madagascar through African sources at the time when slaves and zebu cattle were brought to the island. Hewitt et al. (1996) showed conclusive proof of African admixture into the Malagasy population using the β^A and β^S sickle cell trait mutation. In addition β^A haplotypes are also suggestive of admixture from Asian and Oceanic populations. Using glucose-6-phosphate dehydrogenase enzyme deficiency, Dangerfield (1997) found that the A^{--} variant, which is almost exclusively African (but also found in the Mediterranean region), was found at a high frequency in the Malagasy. Morar (2000) used eight autosomal and five Y chromosome DNA loci in eleven Malagasy ethnic groups to show that 50.00% of the autosomal data and 68.00% of Y chromosome data had African roots. In addition, the study showed a higher African contribution to the coastal regions of Madagascar.

Cann et al. (1987) highlighted the value of mtDNA, as a tool to reconstruct human origins. Since then mtDNA studies have improved our understanding on historical human migration routes and assessing population affinities. A mitochondrion is an extra-nuclear organelle that is involved in energy production in all eukaryotic cells. Mitochondrial DNA is located within the mitochondrion and is a circular molecule of double-stranded DNA. It contains 16 569 base pairs and has thirty seven genes coding for 22 transfer RNAs, two ribosomal RNAs and 13 proteins (Anderson et al., 1981; Taanman, 1999). The mtDNA non-coding DNA is housed within a 1.122kb region

known as the D-loop or hypervariable region. This region has a high mutation rate and is divided into two sections (HVR I and HVR II). HVR I spans from position 16024-16400 and HVR II from 57-372, however the positions analyzed vary between studies.

MtDNA serves as a useful tool for phylogenetic analysis. Firstly, there is an absence of recombination (simplifying genetic studies) and secondly a high copy number (yielding more DNA for extraction purposes) and fast mutation rates. There are however also limitations to using mtDNA, these being that the inheritance of mtDNA is maternal thus using this marker, one is able to capture a single line of evidence. Given these properties, mtDNA has been subjected to various methods used in ancestry inference and population genetics. These methods include restriction fragment length polymorphisms, hypervariable region sequencing and whole genome sequencing (Taanman, 1999).

Hurles et al. (2005) using a sample size of 37 individuals showed that only 37.80% of the mtDNA contributions were from African sources whereas the remaining 62.20% of the mtDNA lineages were traced to Asian sources. Haplogroup M was the commonest type within the Asian group and was found at a frequency of 73.90% within Madagascar, while haplogroup R represented 26.10%. In contrast, the same study showed that 51.40% of the Y chromosome lineages studied in these individuals had an African derived source. Of the remaining 48.60% non-African lineages, 34.00% were of Southeast Asian or Oceanic origins.

Dubut et al. (2009) investigated whether a connection between India and Madagascar may exist due to the presence of lineages M23 and M46 (also found in India) within Madagascar. They fully sequenced three mtDNA genomes of Malagasy individuals whose mtDNA haplogroups belonged to M23 and M46. The authors proposed that due to the presence of haplogroups M23 and M46 in Madagascar a possible link between India and Madagascar may exist and India

may have served as a stepping stone in migration from Indonesia to Madagascar. The authors note that separate migrations from India to Madagascar may have also occurred and these would have contributed to the Malagasy gene pool. This study was able to further highlight the genetic complexity of the Malagasy population.

More recently Sergio Tofanelli's group used a sample size of 133 males to type 14 unique event polymorphisms (UEPs) and 17 short tandem repeats (STRs) on the Y chromosome. On the mtDNA, 19 SNPs and HVR I sequencing were tested. Their study reported that 61.00% of Highland populations have Asian derived ancestry, whereas 39.00% belonged to African sources. Similar results were observed for the Côtiers (Coastal populations) where 62.00% are represented by Asian haplogroups and 38.00% by African ancestry. They also reported that there were contributions from Eurasia, the Indian subcontinent and the horn of Africa-Arabia, which are represented by approximately 11.00% in the highland group and 4.00% in the coastal groups (Tofanelli et al., 2009).

A recent study of the Malagasy suggests that Madagascar was first settled 1200 years ago by a group of thirty Indonesian women (Cox et al., 2012). This study made use of HVR I sequence from 266 Malagasy individuals as well as RFLP screening for the Malagasy motif (described below). This study reiterates the findings of other studies in saying that the Malagasy people have resulted from interactions of both African and Asian sources.

A further mtDNA marker that proves to be a useful tool in attempting to reconstruct human history is the intergenic COII/tRNA^{LYS} region of mtDNA. This region usually contains two tandemly arranged copies of a 9-bp sequence (Anderson et al., 1981). However, in some cases, one copy of the repeat is deleted most likely due to DNA slippage (Cann et al., 1984). Using RFLP analysis and sequence analysis, Cann et al. were able to show that primary length

variation occurred due to loss of one copy of the 9-bp repeat sequence, CCCCCTCTA (Cann et al., 1984). For the purposes of this study, these features of the intergenic COII/tRNA^{LYS} region will be exploited.

The 9-bp deletion has been found in varying frequencies throughout Asia, Polynesia and the New World and was therefore referred to as an 'Asian-specific" marker (Redd et al., 1995). The deletion has since been found at varying frequencies throughout central and southern Africa is rare in east-central Africa and is absent in western Africa (Vigilant et al., 1991; Soodyall et al., 1996). The deletion is found intermittently in Europe (Torroni et al., 1996; Alves-Silva et al., 1999) and in Australia (Hertzberg et al., 1989).

Subsequent phylogenetic analysis of hypervariable region sequence data have indicated that Asian and African mtDNAs that habour the 9-bp deletion, are not closely related. This would be suggestive of independent origins of the African and Asian forms of the deletion (Vigilant, 1991). In addition, it has been demonstrated that the deletion arose at least three times in Africa. The majority of mtDNA types that are associated with the African form of the deletion have a similar haplotypic profile and is referred to as the major African form of this deletion (Soodyall et al., 1996).

The Asian form of the deletion was then further dissected into the **"**Polynesian motif" which was noted to occur at high frequencies among Polynesians. This form of the deletion is characterised by subhaplogroup B4a1a1a and substitutions at positions 14022 (A→G) a new mutation (found outside the hypervariable region) that is associated with the "Polynesian motif" and three substitutions: 16217 (T→C), 16247 (A→G) and 16261 (C→T) that occur within the hypervariable region (Razafindrazaka et al., 2009).

This lineage may have evolved in eastern island Southeast Asia or Near Oceania around 6,200-10,900 years ago (Pierson et al., 2006). This haplogroup's immediate precursor is deficient of the 16247 (A→G) substitution and has been found in Taiwanese aboriginal groups. The haplogroup is estimated to be 13,200 years of age (Trejaut et al., 2005). This series of dates is consistent with a model whereby Austronesian-speaking populations expanded out of Taiwan during the mid- to late-Holocene. The Austronesian expansion lead to the spread of the immediate ancestor of the Polynesian motif over a vast area covering Taiwan in the north, New Zealand in the south, remote Polynesia in the east and Madagascar in the far west (Razafindrazaka et al.,2009). The "Polynesian motif" reaches fixation (100%) in some Polynesian populations. It is also common in Micronesia and regions of Near Oceania where it is not always restricted to Austronesian-speaking populations but is found in some Papuan speaking groups (Razafindrazaka et al., 2009).

Soodyall et al. (1996) conducted mtDNA studies based on the intergenic COII/tRNA^{Lys} 9-bp deletion and revealed that 26.80% of the Malagasy individuals screened as positive for the deletion, in conjunction with sequence data, displayed the Asian form of the deletion other forms being African and Polynesian and more recently, the Malagasy motif.

In a recent study by Razafindrazaka, a further type of the Polynesian form of the 9-bp deletion was detected in a group of Malagasy samples. The "Malagasy motif" was dated to 6,000 years before present and is found exclusively in the Malagasy. This motif is characterised by the changes that define the Polynesian motif and addition coding region mutations 1473(C→T) and 3423(T→A). Due to the absence of the Malagasy motif in any Melanesian and Polynesian individuals, it is thought that the motif expanded in a short time span after appearing in Madagascar (Razafindrazaka et al., 2009).

The construction of a global mtDNA phylogeny has played a pivotal role in positioning the most recent common ancestor (TMRCA) to Africa and more specifically sub-Saharan Africa. The phylogeny has also assisted in understanding the movements of humans within Africa and ultimately the dispersal and movements outside of Africa (Behar et al., 2008).

1.6. Aims and Objectives

Using a genetic approach this study aims to examine the molecular structure of the Malagasy to shed more light on the evolutionary history of the peoples of Madagascar. Using the matrilineally transmitted haploid mtDNA marker, it would be possible to examine how females have contributed to shaping the Malagasy gene pool. In addition, autosomal DNA markers (which harbour information derived from both males and females), especially those which inform on ancestry (ancestry informative markers or AIMS) will be screened for.

Thus, by using these two marker systems, it would be possible to

a) Evaluate the origin of the 9-bp intergenic COII/tRNA^{LYS} deletion in the Malagasy.
b) Asses the mtDNA diversity in the different ethnic groups across different geographic regions.
c) Examine how males and females have contributed in shaping the gene pool of the Malagasy.

Objectives (a) and (b) will be achieved by:

i) Screening for various forms of the 9-bp intergenic COII/tRNA^{LYS} deletion
ii) Using phylogenetically informative SNPs in conjunction with hypervariable region sequences to classify each individual in accordance with the global mtDNA phylogeny.
iii) Comparing mtDNA data from the Malagasy with other populations from around the Indian Ocean rim to trace the possible geographic region(s) of origins.

Objective (c) will be achieved by:

i) Compiling a panel of 96 AIMs from published sources
ii) Setting up a genotyping system which will be achieved through the use of Illumina's BeadXpress platform.

Finally both mtDNA and autosomal DNA data from this study in conjunction with other genetic data as well as data from other disciplines (historical, archaeological, cultural and linguistic) will be used to refine theories concerning the peopling of Madagascar.

Chapter 2

Materials and Methods

2.1. Sample collection and study subjects

Initially, a total of 507 male samples were chosen from sample pool of 2500 individuals from various regions of Madagascar. The samples chosen for this project were previously typed for Y-chromosome SNPs and STRs. In order to gain a complete understanding of both the paternal and maternal ancestry, mtDNA studies, conducted as part of this study would shed light on the maternal contributions to the peopling of Madagascar. Furthermore, recombining autosomal DNA, studied through the use of 96 ancestry informative markers (AIMS) would be used to elucidate the origin of parental populations. In addition to the 507 samples a further 474 samples were typed by members of the Human Genomic Diversity and Disease Research Unit as part of the Genographic Project, which is a global initiative of genetic anthropology that aims to map human migrations through DNA. These results were pooled with those of the current study to give a total sample size of 981 samples.

This study group consists of individuals from twenty-one ethnic groups in Madagascar. Table 1 outlines the breakdown of ethnic groups and geographic location in Madagascar of people from this sample. Further approval for work on these samples was obtained from the Human Ethics (Medical) Research Committee of the University of the Witwatersrand protocol number: MO90629. All extracted DNA is stored in Professor Himla Soodyall's laboratory under ethics approval by the Human Ethics (Medical) Research Committee (HERC) (Protocol number M980553, renewed in September 2004, renewed in May 2009) (Appendix A). The recipes for reagents used in the present study are described in Appendix B, except in cases where the reagent was commercially obtained.

Once informed consent and information relating to the subject's home language, place of birth, blood group, spoken language and self-classified ethnicity had been obtained by Professor Jenkins' group, 30 – 50ml of venous blood was collected from each individual. Comparative data was obtained from the literature and other sources and is referenced appropriately.

Table 1: Tabulating the breakdown of ethnic groups within various regions of Madagascar.

	REGION/ETHNIC GROUP	NUMBER SCREENED
	CENTRAL HIGHLAND	267
1	Betsileo	109
2	Bezanozano	20
3	Merina	138
	EASTERN CENTRAL	333
4	Antaimoro	81
5	Antaisaka	99
6	Antankarana	15
7	Betsimisaraka	90
8	Sihanaka	48
	NORTHERN	86
9	Tsimihety	86
	SOUTH WESTERN	126
10	Antandroy	45
11	Mikea	34
12	Mahafaly	29
13	Bara	18
	WESTERN	108
14	Sakalava	70
15	Vezo	25
16	Makoa	13
	SOUTH EASTERN	61
17	Antanosy	35
18	Antaifasy	11
19	Antambahoaka	5
20	Zafisoro	6
21	Tanala	4
	TOTAL	**981**

2.2. Molecular Methods

2.2.1 Nucleic Acid Extraction and Quantification

Genomic DNA from EDTA-venous blood was previously extracted from stored buffy coats using the salting out method as described by Miller, 1988 (Miller, Dykes and Polesky, 1988). For the purposes of this study, extracted DNA samples were hydrated and quantified using the NanoDrop ND-1000 Spectrophotometer (Coleman Technologies Inc., LabVIEW®) and appropriate working solutions of 100ng/ul and 50ng/ul were prepared (using distilled deionised water - ddH_20). Samples and working solutions were stored at -20°C.

2.2.2. MtDNA methods

Two approaches were used in order to assign individuals to mitochondrial haplogroups. Initially, mtDNA hypervariable region (CR) I and II were sequenced and compared to the Cambridge Reference Sequence (CRS) to determine the haplogroup of the sample using Phylotree (van Oven M and Kayser, 2009). Secondly, the mitochondrial DNA single base extension (SBE) protocol described by Schlebusch, Naidoo and Soodyall, 2009 was used to allocate the samples into 10 major macro-haplogroups found throughout the world (Schlebusch et al., 2009). In doing so, the SBE method was able to validate the haplogroups assigned to the samples through sequencing. As an alternative approach to the SBE assay, the TaqMan® SNP Genotyping Assay was carried out on a selection of samples to validate the findings of both the sequencing and SBE reactions.

2.2.2.1. 1kb mtDNA Hypervariable region Polymerase Chain Reaction and Sequencing

Polymerase Chain Reaction (PCR) was first described by Mullis et al., 1986. It involves the amplification of a particular region of DNA, where the amplification is achieved by successive rounds of heating and cooling in order for denaturation of double-stranded DNA, annealing of primers and extension of the sequence of interest to take place.

For each of the DNA samples used as part of this project, the 1.1kb non-coding hypervariable region (CR I and CR II) was amplified using the protocol highlighted in Table 3 with the relevant primers (Table 2). CR I and CR II extended over nucleotide positions 15997-16569 and 31 – 407 respectively.

Table 2: Primer sequences used for amplification of the mtDNA 1.1kb hypervariable region and for the subsequent cycle sequencing reactions.

1.1kb Hypervariable region PCR Primers			
Name of Primer	**Region**	**Sequence (5' –3')**	**Reference**
15876F	CR	TCAAATGGGCCTGTCCTTGTAG	Behar et al., 2007
639R	CR	GGGTGATGTGAGCCCGTCTA	Behar et al., 2007
Cycle Sequencing Primers			
15946F	CR I	CAAGGACAAATCAGAGAAAA	Behar et al., 2007
132R	CR I	GACAGATACTGCGACATAGG	Behar et al., 2007
L29	CR II	GGTCTATCACCCTATTAACCAC	Vigilant et al., 1989
H408	CR II	CTGTTAAAAGTGCATACCGCCA	Vigilant et al., 1989

Table 3: The general PCR and thermal cycling protocol used to set up the amplification system for the 1.1kb hypervariable region

PCR Components	Initial Concentration	Final Concentration	Volume (µl)
DNA	N/A	5ng/µl	3
dNTPs (Bioline)	2.5mM	0.1mM	2
10X reaction buffer with added $MgCl_2$	10X	1 X	5
Primer forward (15876)	10mM	0.4µM	2
Primer reverse (639)	10mM	0.4µM	2
FastStart Taq Polymerase	5 U/µl	1 U	0.2
ddH_20	N/A	N/A	35.8
TOTAL			**50**
Thermal cycling conditions			
Cycle step	**Temperature (°C)**	**Time**	**No.of cycles**
Denaturation (Initial)	95	5 minutes	1
Denaturation	95	30 seconds	35
Annealing	55	30 seconds	
Extension	72	2 minutes	
Final Extension	72	10 minutes	1
Hold	4	∞	

A no template control (NTC- sample that contained all reagents but no DNA) was used in all PCRs as a quality assurance method and to check for contamination. Primer sequences for 15876F and 639R are shown in Table 2. All PCRs were performed on a 9700 GeneAmp® PCR System (Perkin-Elmer, Applied Biosystems). All thermal cycling conditions are recorded in Table

3. Negative controls (PCR reactions with no added DNA) were routinely used to serve as a control for contamination. All buffers and Taq Polymerase were supplied by Roche Applied Science.

In order to check for amplification a 5µl of PCR product and 5µl of Bromophenol Blue dye, which contains glycerol or other density increasing chemicals, were combined and the total volume was loaded onto a 3% Agarose gel with ethidium bromide staining. A 1kb plus molecular weight marker (Invitrogen) was used to size the fragments (Appendix C). The products were subjected to gel electrophoresis at 180V in a 1 X TBE buffer for 70 minutes. The products were then visualized under UV light. If amplification was successful, a 1.1kb band was visible (Appendix D, Figure 1).

Post PCR purification was carried out using an enzymatic purification method. Shrimp Alkaline Phosphatase (SAP), which catalyses the release of the phosphate group at the 5' end of DNA, in addition to the enzyme, Exonuclease I [(Exo I) involved in the breakdown of excess primer)] and ddH_20 were combined into a total volume of 4µl per sample, using 1 U (1.40µl) of SAP, 2 U (0.20µl) Exo I (New England Biolabs) and 2.4µl ddH_20 respectively. For each sample, 10µl of PCR product was combined with 4µl of the enzymatic mixture. The entire volume was then incubated at 37° C for one hour, after which the Exo I was inactivated at 75°C for 15 minutes.

The purified PCR product for all samples was then cycle sequenced using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems). This allows for the synthesis of new DNA strands (which are complementary to the single-stranded template) and for the incorporation of fluorescently labelled dideoxynucleotide triphosphates (ddNTPs), which will then terminate the growth of the newly synthesized strand. The primer is chosen according to which strand is being sequenced and one primer is used per reaction. For the L (light) strand of

mtDNA, primers L15876 and L29 were used to amplify CR I and II respectively, in a 5'-3' direction. For the H (heavy) strand of mtDNA, primers H132 and H408 were used to amplify CR I and II respectively, in a 3'-5' direction. CR I spanned across nucleotide positions 15997-16569 and CR II spanned across nucleotide positions 31- 407. Primer sequences are shown in Table 2.

For each primer, the reaction consisted of 1µl of BigDye®, 1µl of primer (3.3mM) and 7µl ddH_20. Samples were cycle-sequenced in the 9700 GeneAmp® PCR System (Perkin-Elmer, Applied Biosystems) and the conditions followed were as follows: denaturation at 96°C for 30 seconds, annealing at 96°C for 30 seconds and extension at 60°C for 4 minutes. All of the steps above were repeated for 25 cycles.

Cycle sequenced products were then purified using the Montage Multiscreen SEQ96 Sequencing Reaction Cleanup Plates (Millipore). This clean up system uses a vacuum based filtration system to elute salts and unincorporated ddNTPs and primer. Cycle sequenced products were combined with 20µl Injection Solution (available as part of the Montage kit) and the total volume was transferred into the bottom of the Multiscreen SEQ_{96} clean up plates. The vacuum was set to 23-25 in Hg and was applied to the plate for 4 minutes until wells appeared "glassy". The plate was then blotted down onto paper towel, removing excess liquid. A further 20µl of injection solution was pipetted into the plate and the vacuum was applied for 3 minutes after which, the plate was blotted on paper towel. The final step involved re-suspending the cycle sequenced products by aliquoting 20µl of injection solution into the clean up plate, covering the plate with adhesive film and placing the plate on a bench shaker to shake at 3500rpm for 10 minutes. The samples were then transferred into an an MicroAmp® Optical 96-Well Reaction Plate (Applied Biosystems) and put into the sequencer for analysis.

Electrophoresis was carried out using the Applied Biosystems, 3130xl Genetic Analyser and sequences were analysed using Sequencing Analysis Software v5.2 (Applied Biosystems).
All sequences were aligned using the Clustal W algorithm employed in the freely available alignment programme Bioedit (Hall, 1999). The revised Cambridge reference sequence (CRS) was used as the reference sequence (Andrews et al., 1999) and variant sites were identified from this. MtDNA haplogroups were defined using nomenclature as per Behar et al., 2008 and Global human mtDNA phylogenetic tree – Build 15 (30 September 2012) [(available at http://www.phylotree.org/)].

2.2.2.2. Mitochondrial Single Base Extension (SBE) Assay

The mitochondrial Single Base Extension (SBE) Assay is a multiplex based system which tests 14 SNP variations that in turn, allows for the categorization of samples into haplogroups L0, L1, L2, L3, L4, L5, L6, M, N and R (Schlebusch et al., 2009).

2.2.2.2.1. Multiplex PCR amplification

The initial multiplex PCR amplified six fragments of differing lengths and the protocol involved creating a reaction mixture of 25µl comprised 1µl of premixed 25 primer mix, 2 U FastStart Taq (Roche Applied Science), 1µl FastStart Taq buffer (without additional $MgCl_2$), 3.5µl MgCl, 3µl dNTPs and ddH_20 to make up the total volume. The protocol that was followed and all primer sequences are listed in Schlebusch et al., 2009.

All PCRs were performed on a 9700 GeneAmp® PCR System Perkin-Elmer, (Applied Biosystems) and the conditions followed were as follows: 95^oC for 6 minutes; 35 cycles of denaturation at 95^oC for 1 minute and 30 seconds, annealing at 60^oC for 1 minute and 30

seconds and elongation at 72°C for 2 minutes. A final elongation step was carried out at 72°C for 10 minutes. Amplicons were checked on a 2% agarose gel with ethidium bromide staining (Appendix D, Figure 2).

Post PCR purification was carried out using an enzymatic cleanup method. Shrimp Alkaline Phosphatase (SAP) [USB Corporation], which catalyses the release of the phosphate group at the 5' end of DNA, in addition to the enzyme, Exonuclease I (New England Biolabs), which is involved in the breakdown of excess primer and ddH_20 were combined into a total volume of 2µl per sample, using 1µl of SAP, 0.1µl Exo I and 0.9µl ddH_20 respectively. For each sample, 5µl of PCR product was combined with 2µl of the enzymatic mixture. The entire volume was then incubated at 37° C for one hour, after which the enzymes were inactivated at 75°C for 15 minutes.

2.2.2.2.2. SBE reaction

The next step involved the single base extension where primers of differing lengths are fluorescently labelled and this in turn allows for improved separation and detection of the extension products. All allele and marker information is listed in Schlebusch et al., 2009. The SBE reaction consisted of 1.5µl of purified PCR product, 1µl of SNaPShot® Multiplex Ready Reaction Mix (Applied Biosystems) and 1µl of the premixed single base extension primer mix made up to a volume of 5µl by ddH_20. The conditions used were as follows: 35 cycles of denaturation at 96°C for 10 seconds, annealing at 50°C for 5 seconds extension at 60°C for 30 seconds.

After extension, the post-extension treatment was carried out in a total volume of 7µl, which contained 5µl of the SBE product and 2µl of the clean-up reagents comprised 0.5µl of SAP,

0.7µl 10X SAP buffer and 0.8µl of water. The entire volume was then incubated at 37° C for one hour, after which the SAP was inactivated at 75°C for 15 minutes.

Following the post-extension clean up step, 7.5µl of Hi-Di Formamide (Applied Biosystems) was combined with 0.5µl of GeneScan-LIZ 120 internal size standard (Applied Biosystems) and 2µl of purified SBE product. All of the reagents and products above were plated into a MicroAmp® Optical 96-Well Reaction Plate (Applied Biosystems) and denatured at 95° C for two minutes then cooled to 4° C. The fragments were separated on the 3130xl Genetic Analyser (Applied Biosystems) in accordance to the instructions in the SNaPShot® Multiplex Kit (Applied Biosystems). All samples were analysed using GeneMapperID v.3.2 software (Applied Biosystems). One of two colours would be indicative of a positive or negative result for a particular SNP, thereby indicating whether the sample was derived or ancestral for the haplogroup in question.

2.2.2.2.3. TaqMan® SNP Genotyping Assay

The TaqMan® SNP Genotyping Assay (Applied Biosystems) was used to confirm the haplogroup status of previously classified (by the hypervariable region sequence) samples of haplogroup R. This system was also used to define a haplogroup if this was not possible using the hypervariable region sequence. This was used as an alternate method to the Single Base Extension assay.

The TaqMan® SNP Genotyping Assay is a real-time PCR based method that includes the use of the 5' exonuclease activity of AmpliTaq Gold®, two locus-specific PCR primers that flank the SNP being investigated, two allele-specific oligonucleotide TaqMan® probes. These allele-specific probes have a fluorescent reporter dye at the 5' end and a non-florescent quencher with

a minor groove binder at the 3' end. A probe that is intact will not fluoresce when excited due to the close proximity of the 5' flourophore to the 3' quencher molecule. Fluorescence will be emitted when an intact probe, which is hybridised to a target allele, is cleaved by the 5' activity of the AmpliTaq Gold®. In this way, using two probes with different fluorophores, allows for the detection of both alleles in a single reaction (De la Vega, Lazaruk, Rhodes and Wenz, 2005).

This system makes use of 1µl DNA diluted to 5ng/µl, 0.25µl Universal Master Mix (Applied Biosystems), 1µl primer mix and 1µl of ddh_20 in a total volume of 5µl. Primer sequences are the property of Applied Biosystems. All PCRs were plated onto an MicroAmp® Optical 384-Well Reaction Plates and PCR was performed on a 9700 GeneAmp® PCR System Perkin-Elmer, (Applied Biosystems) and the conditions followed were as follows: 95°C for 10 minutes; 40 cycles of annealing at 95°C for 15 seconds and elongation at 60°C for 1 minute.

Once the PCR was complete, the plate was placed into the 7900ht-realtime PCR machine (Applied Biosystems) and each sample was labelled and linked appropriately. A laser then detects the intensity of each dye and reports on the fluorescence of each dye accordingly. Genotypes are called using SDS Software v2.3 and AutoCaller Software (Applied Biosystems).

2.2.2.3. PCR of the Intergenic COII/tRNA ^{Lys} 9-bp deletion

Length changes by means of insertions and deletions also serve as useful markers for the inference of population histories. The intergenic COII/tRNA^{Lys} 9-bp deletion region, situated between the COII and Lysine tRNA genes on the mitochondrial genome, is one such length change. In order to detect the presence or absence of the 9-bp deletion in the COII/tRNA^{Lys} intergenic region, each sample was screened with a PCR using primers A and B, described by Wrischnik et al., 1987. Amplifications were performed in a 25µl final volume that contained

100ng DNA template, 10µM of each primer (Metabion), 2.5mM of each deoxynucleotide triphosphate (Bioline), 10mg/ml bovine serum albumin (Roche Applied Science), 5 units of Taq Polymerase (Roche Applied Science), Faststart Taq Reaction buffer (with added $MgCl_2$) supplied by the manufacturer (Roche Applied Science) and double distilled water (ddH_20).

Primer sequences for L8215 (A) and H8297 (B) are shown in Table 4. All PCRs were performed on a 9700 GeneAmp® PCR System (Perkin-Elmer, Applied Biosystems) and the conditions followed were as follows: 94°C for 6 minutes; 30 cycles of denaturation at 94°C for 1 minute, annealing at 56°C for 1 minute and elongation at 72°C for 1 minute. No template controls (PCR reactions with no added DNA) were routinely used to serve as a control for contamination. In addition to the negative controls, samples that were shown, as part of previous studies, with or without the 9-bp deletion, were used as positive controls for this study. For each sample, the entire volume of PCR product and 5µl of Bromophenol Blue dye were combined and the total volume was loaded onto a 3% Metaphor agarose gel with ethidium bromide staining. A 1kb molecular weight marker (Invitrogen) was used to size the fragments (Appendix C). The products were subjected to gel electrophoresis at 180V in a 1 X TBE buffer for 70 minutes. The products were then visualized under UV light. If a 121bp fragment was obtained, the sample did not have the 9-bp deletion (i.e. both copies of the 9-bp repeat were present), however if a 112bp fragment was obtained, only a single copy of the 9-bp repeat was present therefore the deletion was present (Appendix D, Figure 3).

Table 4: Primer sequences used for amplification of the 9-bp deletion in the mtDNA COII/tRNALys intergenic region (Wrischnik et al., 1987)

Intergenic COII/tRNA^{Lys} 9-bp deletion		
Primer Name	**Position of 3' end**	**Sequence (5' –3')**
L8215 (A)	8215	ACAGTTTCATGCCCATCGTC
H8297 (B)	8297	ATGCTAAGTTAGCTTTACAG

2.2.2.3.1. Sequencing for the "Malagasy Motif"

All samples that tested positive for the "Polynesian motif" (described in section 1.5) were sequenced for new diagnostic mutations described by Razafindrazaka et al., 2009.

2.2.2.3.1.1. Primer design

The primer sequences used to amplify to separate 2kb mtDNA coding region fragments encompassing SNPs 1473 (C-T) and 3423 (T-A) were taken from Behar et al., 2007 and are listed in Table 5. The cycle-sequencing primers were designed using Primer3 (http://primer3.wi.mit.edu/), an online computational programme that allows the user to design primers for a specified region. Segments of the mtDNA rCRS were copied into the programme and input parameters were adjusted. Possible primer pairs and annealing temperatures relating to these primers were provided. A suitable primer was chosen for each of the mtDNA coding region SNPs and was modified by ensuring that there was a cytosine residue at both 5' and 3' ends. The primers are listed in Table 5.

2.2.2.3.1.1.2. 2kb PCR Optimisation

A 2kb PCR was carried out in two separate reactions for each of the new mutations described. PCRs were optimised by adjusting annealing temperatures and times. Final amplifications were performed in a 50µl final volume that contained 1µl (~50-100ng) DNA template, 10µM of each primer (Metabion), 2.5mM of each deoxynucleotide triphosphate Bioline), 1 U of Taq Polymerase (Roche Applied Science), Faststart Taq Reaction buffer [with added $MgCl_2$) (Roche Applied Science)] and deionised distilled water ddH_20. All PCRs were performed on a 9700 GeneAmp® PCR System (Perkin-Elmer, Applied Biosystems) and the conditions followed were

as follows: 94°C for 6 minutes; 30 cycles of denaturation at 94°C for 1 minute, annealing at 58°C for 30 seconds and elongation at 72°C for 1.5 minutes. A final elongation step at 72°C for 8 minutes was carried out after the 30 cycle series. Amplicons were checked on a 3% agarose gel (Appendix D, Figure 4). Samples were sequenced employing the same techniques as those of the hypervariable region using primers listed in Table 5 and results were visualised as illustrated in Appendix D, Figure 5.

Table 5: Primer sequences used for amplification and sequencing of the Malagasy motif

Malagasy Motif PCR primers			**Reference**
Primer name	**Position**	**Sequence (5'-3')**	
SC-A (f)	1473 T-C	ATCCTACCCAGCACACAC	Behar et al., 2007
SC-A (r)	1473 T-C	GATTTGCCGAGTTCCTTTTACT	Behar et al., 2007
SC-B (f)	3423 T-A	CTGACAATTAACAGCCCAATATC	Behar et al., 2007
SC-B (r)	3423 T-A	GAATGCTGGAGATTGTAATGGG	Behar et al., 2007
Malagasy Motif Cycle Sequencing Primers			
9-bp SNP_1473F	1473 T-C	CACCACCTCTTGCTCAGCC	Present study
9-bp SNP_3423F	3423 T-A	CCTCCCTGTACGAAAGGAC	Present study

2.2.2.4. Autosomal DNA genotyping by use of BeadXpress technology

In 2007, Illumina Inc released the BeadXpress system, which uses patented VeraCode digital microbead technology for targeted genotyping in multiplexed-based assays (Lin et al., 2009). The BeadXpress system begins with the biotinylation of the input genomic DNA, which is then followed by the immobilization of the biotinylated DNA to streptavidin-coated particles. Annealing of the assay oligonucleotides takes place concurrently. The ligated oligonucleotide carries an address sequence that is able to identify a particular SNP in the assay. In addition,

the oligonucleotide has a universal priming sequence at the 3' region. Unbound oligonucleotides are subsequently washed away and allele-specific extension and ligation of the hybridized oligonucleotide is carried out. The hybridized oligonucleotide is then amplified using fluorescently-labelled universal primers. The PCR product is then hybridized to the VeraCode beads (which have inscribed holographic barcodes) bearing complementary address sequences. A two-colour BeadXpress reader (laser) then detects fluorescence signals to the assign a particular SNP a specific allele. A clustering algorithm assigns the samples to particular genotype clusters (Lin et al., 2009).

As part of this study, we have made use of this genotyping technology to screen for up to 96 AIMs found in African, South Asian, East Asian, European, Middle Eastern and Oceanic populations to enable us to assess the genetic substructure of the Malagasy. A list of 96 AIMs was compiled from literature and the final list is shown in Appendix E. Comparative genotype data was collated for each of the SNPs from publicly available International Hapmap project and the Human Genome Diversity Project. In addition, allele frequency data was collected from ALFRED (Osier et al., 2002). Using these data, Principal component plots were generated to test whether each of the groups would separate from each other. The plots generated from this data showed continental separation between Asia, Oceania, Africa, the Middle East and Europe. Several rounds of validation took place by selecting SNPs that were both informative and functional on Illumina's BeadXpress platform.

In order to type the SNPs, samples had to be normalized to 50ng/ul. Each of the 170 male samples (chosen in equal proportions from each of the ethnic groups) was quantified using the Tecan's Infinite® 200 NanoQuant technology after which they were normalized using the Tecan's Liquid handling system. The procedure to genotype the SNPs on the BeadXpress technology (Illumina) was exactly that recommended by the manufacturer and is described in

detail by Lin et al., 2009. The 24 page protocol, provided by the manufacturer (Illumina) can be downloaded from http://www.illumina.com/support/documentation.ilmn by clicking on GoldenGate Genotyping Assay for VeraCode Manual Protocol Experienced User Card.

2.2.2.5. Data Analysis

CR I and CR II sequences were edited and aligned against the revised Cambridge Reference Sequencing (Andrews et al., 1999), using the Clustal W algorithm (Thompson et al., 1994) in BioEdit v.7.0.5.3 (Hall, 1999). A variant site list was then generated using S-compare (Nelson, 2006). All sequences were manually checked to confirm variant sites. Haplogroups were assigned to samples based on both CR I and CR II variant sites in conjunction with a dataset of over 6700 complete mitochondrial DNA sequences compiled by van Oven and Kayser (http://www.phylotree.org). In addition, all samples that were unable to be defined by their CR I and II sequences, were screened for mtDNA phylogenetically informative SNPs using methods described by Schlebusch et al., 2009. This method assisted in definitively assigning a haplogroup to each sample. Subsequently, both CR I and CR II sequences were combined into a single sequence for further analyses. A unique haplotype list (supplementary data available on request) was generated using DnaSP v.4.10 (Rozas et al., 2003). A haplogroup frequency list based on Malagasy regions was also created (Appendix F).

A unique sequence list was created through DnaSP (Rozas et al., 2003). Two unique datasets of 295 sequences were created, the first containing coding region SNP information along with hypervariable region data and the second which contained only hypervariable region data. Subsequent phylogenetic tree analyses of these sequences were carried out through neighbour-joining analysis using MEGA 5 software (Tamura et al., 2011) in order to determine whether SNP data altered the patterns seen within the phylogenies.

Population pairwise differences were calculated with Arlequin v3.11 (Excoffier, Laval and Schneider, 2005) by using Fst distances (Reynolds, Weir and Cockerham, 1983). Nucleotide diversity (Nei, 1987) for each group was calculated in DnaSP v4.10 (Rozas et al., 2003). Principal component and multidimensional scaling (MDS) plots were produced using haplogroup frequencies and Fst distance matrix (generated through Arlequin v.3.11) through the use of Palaeontological Statistics (PAST), v.1.54 (Hammer, Harper and Ryan, 2001). Furthermore, neighbour joining phylogenies were created through MEGA 5 (Tamura et al., 2011) to determine relationships between various ethnic groups of Madagascar with Neanderthal samples to root the phylogeny. Further samples from two African [(Kenya and Mozambique) (Pereira et al., 2001; Salas, 2002; Brandstätter, 2004)] and two Asian [(Vietnam and China) (Kong et al., 2011; Lum and Cann, 2000; Horai et al., 1996; Irwin et al., 2008) sources were added for comparative purposes.

Networks of sequences were created using the Median Joining algorithm (Bandelt et al., 1999) in Network v.4.5.0.0 (Fluxus-engineering, 2008). Subsequently, the Steiner maximum parsimony algorithm (Polzin and Daneshmand, 2003) within Network 4.5.0.0 was applied. For network analysis transversions were weighted at three times the weight of transitions. External sequence included in phylogenetic and network analyses was the extended hypervariable region reference sequence (Andrews et al., 1999). Additional sequences were taken from published literature for the relevant types of analyses. Based on historical and linguistic data, in addition to the halpogroup data from this study, comparative data was chosen from several East and South Asian (Qin et al., 2010; Hill et al., 2006; Hill et al., 2007; Trejeut et al., 2005; Horai et al., 1996; Kong et al., 2011; Li et al., 2007; Tabbada et al., 2010; Metspalu et al., 2004; Cox, 2005; Eaaswarkhanth et al., 2009, Fucharoen, Fucharoen and Horai, 2001; Gunnarsdóttir et al., 2011; Irwin et al., 2008; Oota et al., 2001; Peng et al., 2011, Kivisild et al., 2002, Kong et al., 2003, Kolman et al., 1996, Mona et al., 2009), African (Quintana-Murci et al., 2008; Cerný et al.,

2004; Graven et al., 1995; Kivisild et al., 2004; Salas et al., 2002; Pereira et al., 2001; Brandstätter et al., 2004; Watson et al., 1996; Brakez et al., 2001; Destro-Bisol et al., 2004, Knight et al., 2003), Oceanic (Deguilloux et al., 2011; Lum and Cann, 2000; Pierson et al., 2006; Whyte, Marshall and Chambers, 2005; Vilar et al., 2008; Sykes et al., 1995), Indian Ocean island populations (Msaidie et al., 2010; Berniell-Lee et al., 2008) and Middle Eastern (Metspalu et al., 2004; Abu-Amero et al., 2008, Terreros et al., 2011) populations.

Mismatch distributions of ethnic groups and haplogroups were calculated in Arlequin v.3.11 (Excoffier et al., 2005). From these results the validity of demographic expansions and the date of these expansions were estimated. To test whether there was actually an expansion at a particular period, a simulation is created of a population going through an expansion and testing whether the given data is significantly different from the simulated expansion values. A non-significant Sum of Squared deviation (SSD) p-value is indicative of an expansion. This expansion can then be dated using the calculation explained in Chapter 3, Results. In addition to dating the time of expansion, an effective population size (N_e) was estimated using the Watterson estimator θ per sequence (Tajima, 1996). This along with selective neutrality tests of Tajima's D (Tajima, 1989, Fu's Fs statistic (Fu, 1997) and the R2 (Ramos-Onsins and Rozas, 2002) statistic were calculated in DnaSP v4.10.

Haplogroup frequency surface maps were created using Surfer v.10 Demo (Golden Software, 2006) and the ordinary Kriging global model (Oliver and Webster, 1990). Mitochondrial contour maps were based on the frequencies of various haplogroups. All total sample sizes in each of the groups was adjusted to the same value.

Inter-population genetic distances were used in Analyses of Molecular Variance (AMOVA), was carried out in Arlequin v.3.11 (Excoffier et al., 2005). The distribution of variance among was

tested at four levels was tested in order to assess relationships among groups of populations. The first level of analysis was to determine the variation contained between individuals within the same population. The second level contained variation between ethnic groups in different geographic regions. The third level contains Madagascan highland ethnicities (Merina, Betsileo, Sihanaka and Bezanozano) in comparison to the remaining lowland ethnicities and the fourth level was to compare Madagascar to the remaining populations within the manually curated comparative dataset.

Results for the intergenic COII/tRNA^{Lys} 9-bp deletion were analysed visually by comparing control samples (run on the same gel as the study samples) to the study samples. Each sample containing the deletion was labelled as “present” whereas those that did not have the deletion were marked as “absent”. In addition, using variant site and haplogroup information, we were able to further classify the intergenic COII/tRNA^{Lys} 9-bp as having either the African, Asian or Polynesian form of the deletion.

All samples containing either Asian or Polynesian forms of the deletion were subsequently sequenced for the Malagasy motif. Sequences were edited and aligned against the revised Cambridge Reference Sequencing (Andrews et al., 1999), using the Clustal W algorithm (Thompson et al., 1994) in BioEdit v.7.0.5.3 (Hall, 1999). SNPs at positions 1473 and 3423 were identified manually after alignment.

All autosomal genotype data was analysed through the use of Illumina’s BeadStudio v.3.2 software. A panel of 96 ancestry informative markers was accumulated from published data (see Appendix E for list of papers).

Comparative data for all 96 SNPs was curated through the Hapmap project and HGDP for all listed populations available on these databases. The datasets were merged which resulted in a combined total of 52 populations, including current dataset. Data was then ordered according to populations and these populations were grouped according to continent. The genotype data was then converted using Plink (found at: http://pngu.mgh.harvard.edu/~purcell/plink/) into numerical values. This format was used to carry out the analysis of population structure.

Analysis of population structure on the 96 SNP set was done in STRUCTURE v.2.3.3 (Falush et al., 2003; Falush et al., 2007). The STRUCTURE analysis of the 96 SNP was done using the following approach: STRUCTURE runs with 10 iterations at K=1 to K=10 were conducted with a burn-in of 10 000 and repeats of a 15 000 for each set. Allele frequencies were correlated and a model with admixture was assumed for all runs. The 10 iterations at each K for each of the 96 SNP sets were condensed into a single consensus run using CLUMPP v1.1.1. (Jakobsson and Rosenberg, 2007). Results were subsequently visualized using DISTRUCT (Rosenberg, 2004). Further distance based analyses were carried out on the autosomal 96 marker panel. To construct inter-population distance matrices, the genotype data was converted in CONVERT 1.3.1 (Glaubitz, 2004) to a format acceptable by GENEPOP (Rousset, 2008), which was subsequently used to create a consensus Fst distance matrix for each of the 96 SNPs. A neighbour joining tree using the Close-Neighbour-Interchange (CNI) method was created using this matrix through MEGA 5. The same matrix was used in Palaeontological Statistics (PAST), v.1.54 (Hammer et al., 2006) to create an MDS plot.

Chapter 3

Results

Thus far, some studies have explored the genetic origins of the Malagasy (Hurles et al., 2005; Tofanelli et al., 2009; Razafindrazaka et al., 2010; Cox et al., 2012 and Kusuma et al., 2015). These studies have used small sample sizes in order to analyse the maternal and paternal contributions to the peopling of the Madagascar. In this study we utilize a sample size (N=981) that is more than double that of previous studies to analyse mitochondrial and autosomal DNA contributions of parental populations to the Malagasy. In the following sections, the results of this study will be shown, analysed and compared to published datasets and the possible origins (based on genetic evidence) of the Malagasy will be discussed.

Firstly, results from the analysis of the intergenic COII/$tRNA^{Lys}$ 9-bp deletion will be presented in the form of global and regional distribution charts and a table. Secondly, sequence and phylogenetically informative SNP data from hypervariable regions will be presented. Haplogroup assignments, a phylogenetic tree, a unique type list and networks that were created using sequence data will be shown. Maternal contributions to the Malagasy will be evaluated by frequency tables, MDS plots, AMOVA, haplogroup frequency maps and mismatch distributions. Finally, parental contributions to the Malagasy will be evaluated through the use of a 96 SNP ancestry informative marker set. Results will be presented in the form of a phylogenetic tree, MDS plot and admixture analysis through the use of STRUCTURE.

3.1. Mitochondrial DNA Results

3.1.1. Intergenic COII/tRNA^{Lys} 9-bp deletion

Most studies of mtDNA base their analysis around nucleotide substitutions (Redd et al., 1995). Length changes by means of insertions and deletions also serve as useful markers for the inference of population histories. The intergenic COII/tRNA^{Lys} 9-bp deletion region, situated between the COII and Lysine tRNA genes on the mitochondrial genome, is one such length change. Independent origins in Asia and Africa make the intergenic COII/tRNA^{Lys} 9-bp deletion region a useful region to study to determine relative parental contributions to the peopling of Madagascar. Initial studies revealed that the 9-bp deletion was found only in Indonesia, Southeast Asia and Oceanic regions, however further analyses have shown it to occur in some African populations (Soodyall et al., 1996). Figure 6 depicts the relative frequencies of the 9-bp deletion globally (from published data).

Table 6 outlines the details of the intergenic COII/tRNA^{Lys} 9-bp deletion and the frequency of each type found in this study. Previous studies by Soodyall et al., (1996) showed that the deletion was present in 26.80% of the sample. The current dataset shows that the intergenic COII/tRNA^{Lys} 9-bp deletion occurs at a frequency of 22.73% within Madagascar. Hypervariable region sequence data in conjunction with SNP variation traced the origins of the 9-bp deletion to African (17.04%) and Polynesian (21.52%) sources. Further analysis on samples containing the Polynesian motif of two coding region SNPs (1473C-T and 3423T-A) revealed that 61.44% of the Polynesian motif samples harbour a new motif called the “Malagasy motif” described by Razafindrazaka et al., 2009 (Table 6). Irrespective of the type of motifs, majority of samples carrying the 9-bp deletion were found in the Eastern Central region of Madagascar, followed by the Central Highland regions (Figure 7). In addition, hypervariable region sequence data of all

samples containing the 9-bp deletion, delineated sixty one unique haplotypes and a haplotype diversity of 0.6871 with the greatest diversity being observed in the group of samples with Asian forms of the deletion.

The African form of the 9-bp deletion is represented by haplogroup L0a2 and its associated sub-haplogroups (Figure 7). Its distribution on the island is not as frequent as that of the Asian form of the deletion, however it is commonly found in East, Central and south-eastern Africa, one of these regions being the parental source for the African form of the deletion in Madagascar. From the thirty eight sequences associated with the African form of the deletion the Eastern Central region of Madagascar has the highest incidence of this motif. The overall haplotype diversity for the African form of the deletion is 0.9132.

Malagasy motif sequence diversity is low. Based on partial coding region sequence data, all samples that are positive for the Malagasy motif contain a cytosine to thymine transition at position 1473 and a thymine to adenine transversion at mtDNA coding region position 3423. In addition, all samples, whether positive for or lacking the Malagasy motif defining SNPs, contain an adenine to guanine change at position 1438. Figure 6 indicates the relative proportions of the 9-bp deletion in various regions of the world (excluding North and South America: Africa (Soodyall et al., 1996; Razafindrazaka et al., 2009), Asia (Metspalu et al., 2004; Melton et al., 1995; Passarino and Semino, 1993; Ballinger et al., 1991; Shields et al., 1992; Stoneking and Wilson, 1989; Horai and Matsunaga, 1986) and Oceania (Redd et al., 1995; Hertzberg et al., 1989; Stoneking and Wilson, 1989) including data from this study representing Madagascar. Furthermore, Figure 7 shows the distribution of the motifs within different regions of the island based on the current dataset.

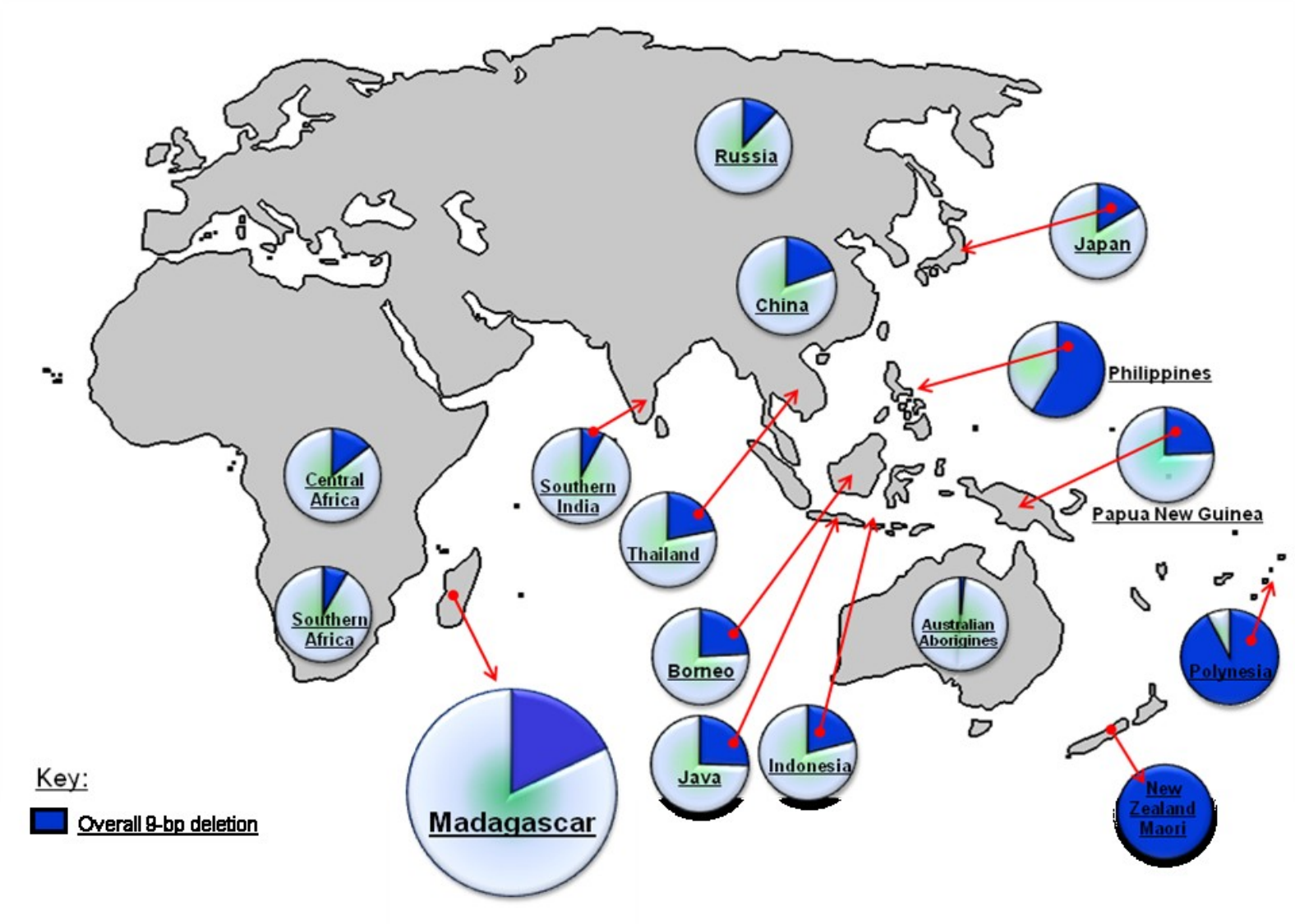


Figure 6: *Global distribution pattern (excluding North and South America) of the intergenic COII/tRNALys 9-bp deletion*. The blue portion represents data from this study. The green shaded area represents comparative data from published sources.

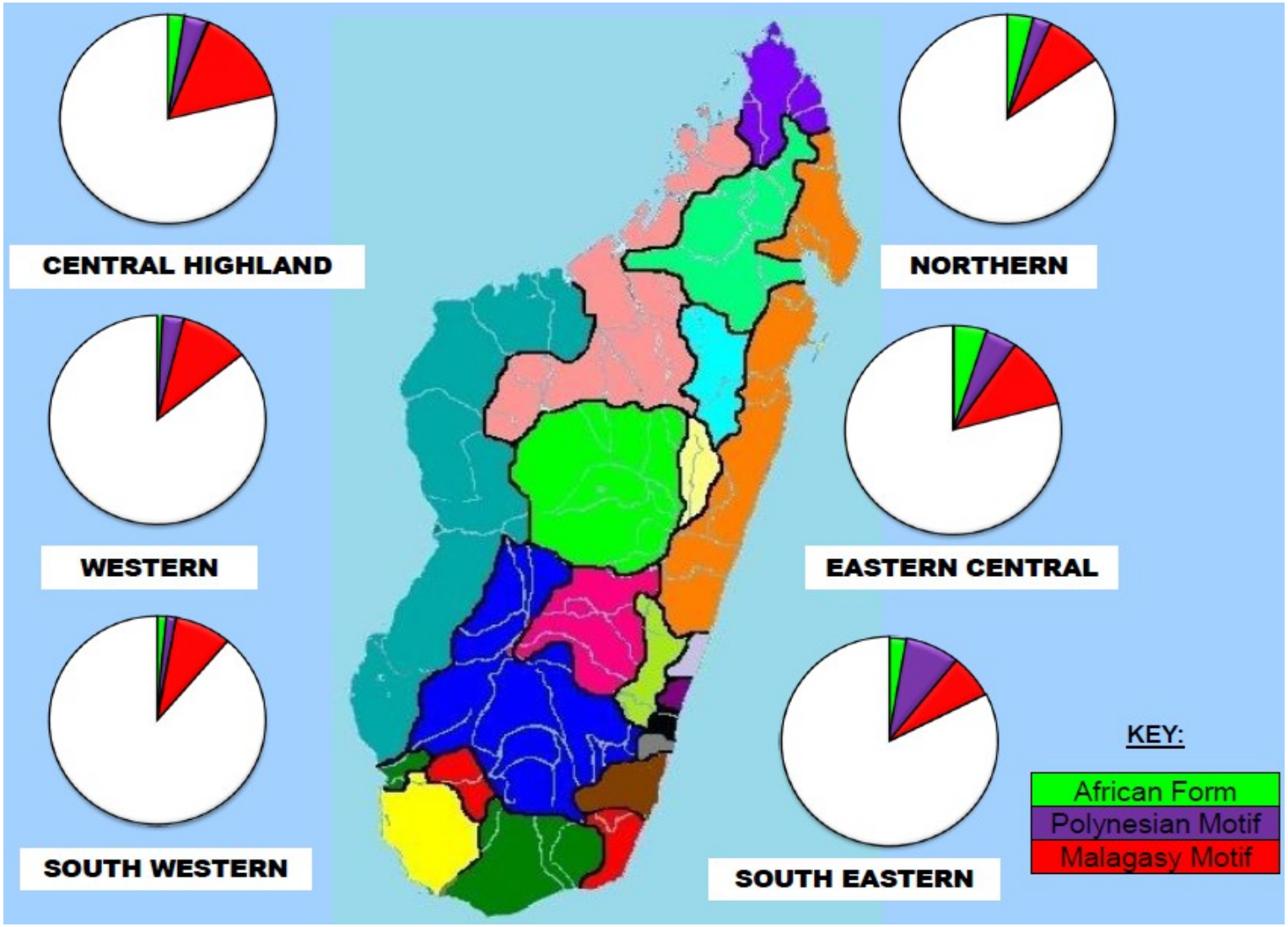


Figure 7: *Distribution patterns of various forms of the intergenic COII/tRNALys 9-bp deletion within different regions of Madagascar*. The shaded regions represent different forms of the intergenic COII/tRNALys 9-bp deletion. The white portion is the number of samples in the present study without the intergenic COII/tRNALys 9-bp deletion.

Table 6: Details of the intergenic COII/tRNALys 9-bp deletion within Madagascar

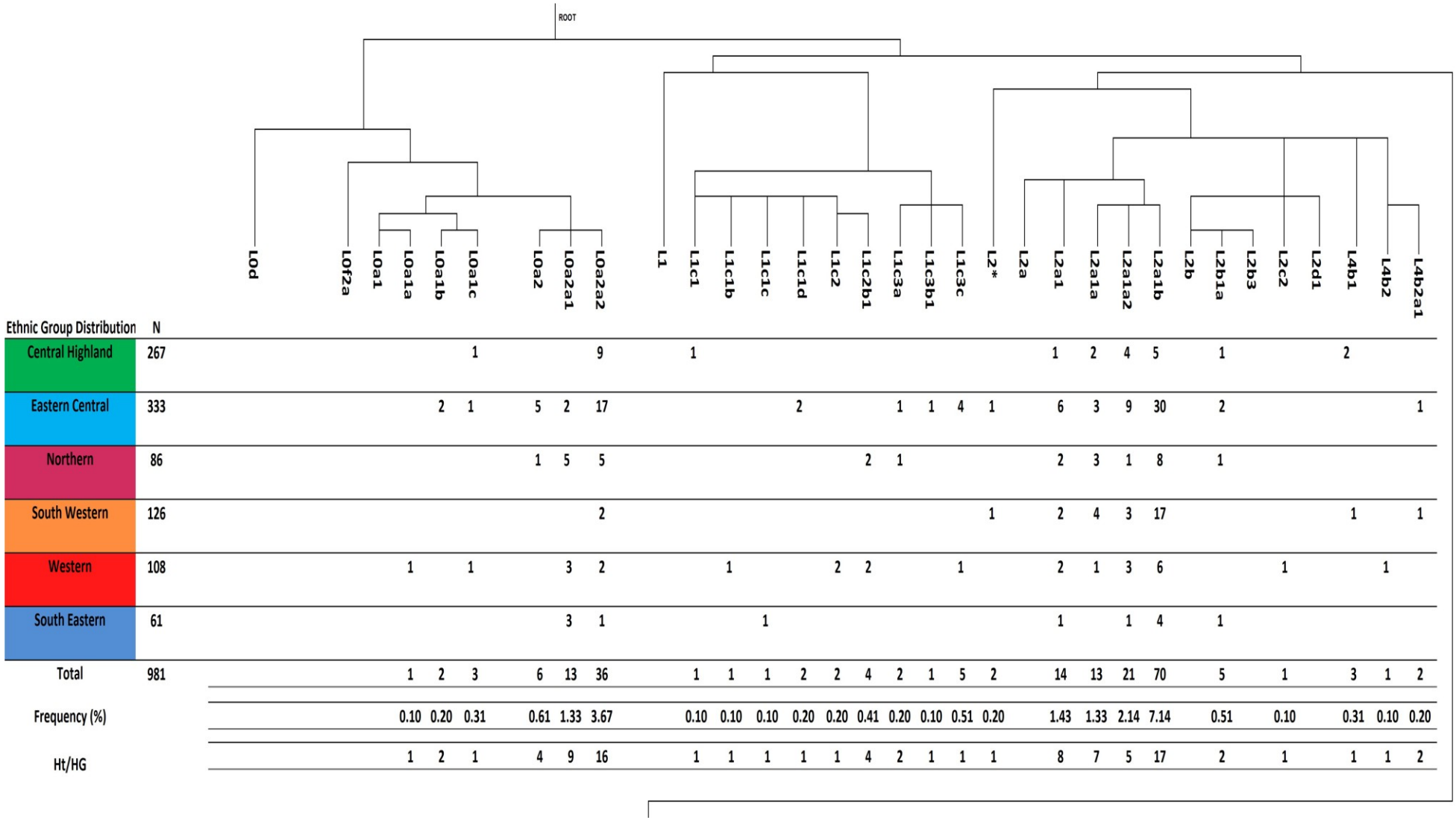
REGION/ETHNIC GROUP	NUMBER SCREENED (N=)	NO. WITH DEL ACCORDING TO MOTIF			
		African	Polynesian	Malagasy	TOTAL
CENTRAL HIGHLAND	267				
Betsileo	109	6	3	19	28
Bezanozano	20		2	1	3
Merina	138	2	8	31	41
EASTERN CENTRAL	333				
Antaimoro	81	6	2	8	16
Antaisaka	99	4	4	16	24
Antankarana	15	2		1	3
Betsimisaraka	90	9	2	15	26
Sihanaka	48		12	7	19
NORTHERN	86				
Tsimihety	86	4	3	9	16
SOUTH WESTERN	126				
Antandroy	45		1	2	3
Mikea	34		1	6	7
Mahafaly	29			3	3
Bara	18	2		1	3
WESTERN	108				
Sakalava	70		2	7	9
Vezo	25	1	1	3	5
Makoa	13		1	3	4
SOUTH EASTERN	61				
Antanosy	35	2	2	4	8
Antaifasy	11				
Antambahoaka	5		1	1	2
Zafisoro	6		1		1
Tanala	4		2		2
TOTAL	981	38	48	137	223

3.1.2. Analyses of variation in mtDNA hypervariable regions

3.1.2.1. Haplogroup assignment and structure

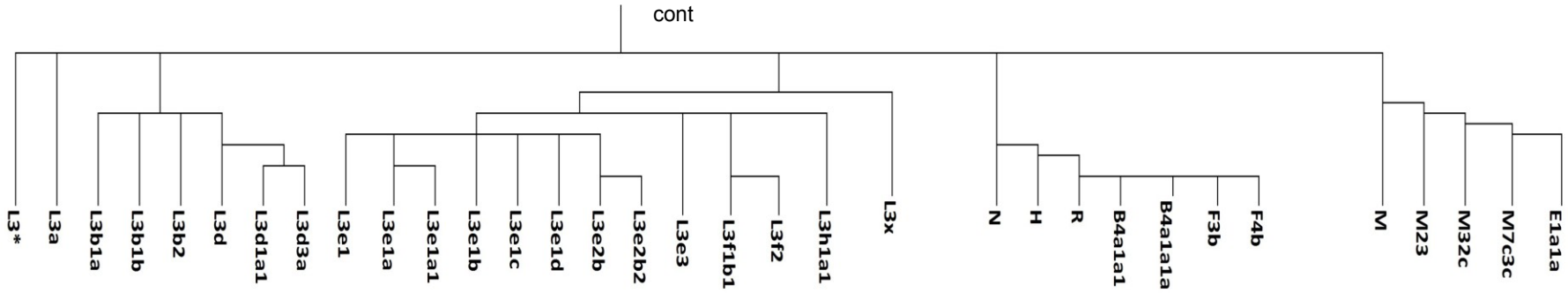
A total of 981 samples were sequenced for the mtDNA hypervariable region (HVR I: and HVR II sequences spanned positions 15997-16569 and 31-407 respectively). A combined sequence from HVR I and II data was used for further analysis. There were 259 variable positions in the combined sequence, where HVR I had 166 variable sites and HVR II had 93. The transversion: transition ratio was 3.47:1. Insertions and deletions were excluded.

Using hypervariable region variable sites and the results from the multiplex single base extension assay (from the coding region), the haplogroup for each of the 981 samples (Figure 6) was called using the placement of the variable site on the mtDNA phylogenetic tree (found at: www.phylotree.org). Furthermore, a breakdown of these haplogroups into Asian, Eurasian and African origins is illustrated in Figure 8, pages 56 and 57, where they are each represented by 58.51%, 0.20% and 41.28% respectively. Of the 981 samples, 57 haplogroups and their sub-haplogroups were identified. Overall, SNP data in conjunction with hypervariable region sequence data defined 295 unique mtDNA sequences. A full haplotype list encompassing both HVR I and II variant sites and gel electrophoresis pictures of the 1kb and single base extension assay PCRs are included in Appendices D and E.



Ethnic Group Distribution	N	L0d	L0f2a	L0a1	L0a1a	L0a1b	L0a1c	L0a2	L0a2a1	L0a2a2	L1	L1c1	L1c1b	L1c1c	L1c1d	L1c2	L1c2b1	L1c3a	L1c3b1	L1c3c	L2*	L2a	L2a1	L2a1a	L2a1a2	L2a1b	L2b	L2b1a	L2b3	L2c2	L2d1	L4b1	L4b2	L4b2a1
Central Highland	267						1			9	1												1	2	4	5	1				2		
Eastern Central	333					2	1	5	2	17					2			1	1	4	1		6	3	9	30	2							1
Northern	86							1	5	5							2	1					2	3	1	8	1							
South Western	126									2											1		2	4	3	17						1		1
Western	108				1		1			3	2		1			2	2			1			2	1	3	6			1				1	
South Eastern	61									3	1			1									1		1	4	1							
Total	981				1	2	3	6	13	36	1	1	1	2	2	4	2	1	5	2			14	13	21	70	5	1				3	1	2
Frequency (%)					0.10	0.20	0.31	0.61	1.33	3.67	0.10	0.10	0.10	0.20	0.20	0.41	0.20	0.10	0.51	0.20			1.43	1.33	2.14	7.14	0.51	0.10				0.31	0.10	0.20
Ht/HG					1	2	1	4	9	16	1	1	1	1	1	4	2	1	1	1			8	7	5	17	2	1				1	1	2

Figure 8: *Phylogeny representing mtDNA haplogroups and sub-haplogroups and their relative frequencies in Madagascar*. The figure continues on the next page. L0d-L4b2a1, represents African haplogroups found throughout the data in the current study. The root of the phylogeny is represented at the top. Frequencies of various haplogroups throughout the six regions in Madagascar are represented at the bottom of the phylogeny



Ethnic Group Distribution	N																														
Central Highland	267		18				5		4	1			2	1		13		4					4	66	13	1		6	21	26	34	22
Eastern Central	333		27			1	11		6			2	1	5	1	15			2				5	67	17	2		4	20	11	35	14
Northern	86		5				1		2		1	1				9							4	10	2	1		1	5	3	9	3
South Western	126	1	9		2		3			2	2	1	1				2		1		1		1	21	7			1	7	6	12	15
Western	108		4	1			3		4	2			1			1	1	2			1			21	3			4	9	3	13	8
South Eastern	61		2				2		2			1	1	3		1							1	11	5			3	2	1	8	6
Total	981	1	65	1	2	1	25		18	5	3	5	6	9	1	39	1	8	2	1	1	1	15	196	47	4		19	64	50	111	68
Frequency (%)		0.10	6.63	0.10	0.20	0.10	2.55		1.83	0.51	0.31	0.51	0.61	0.92	0.10	3.98	0.10	0.82	0.20	0.10	0.10	0.10	1.53	19.98	4.79	0.41		1.94	6.52	5.10	11.31	6.93
Ht/HG		1	15	1	1	1	4		3	3	3	5	5	3	1	15	1	4	1	1	1	1	5	42	10	1		15	14	13	19	15

Figure 8 (cont): Continuation of Figure 8; *Phylogeny representing mtDNA haplogroups and sub-haplogroups* and their relative frequencies in Madagascar. L3*-L3k are representative of African haplogroups. N to E1a1a are of Eurasian origin

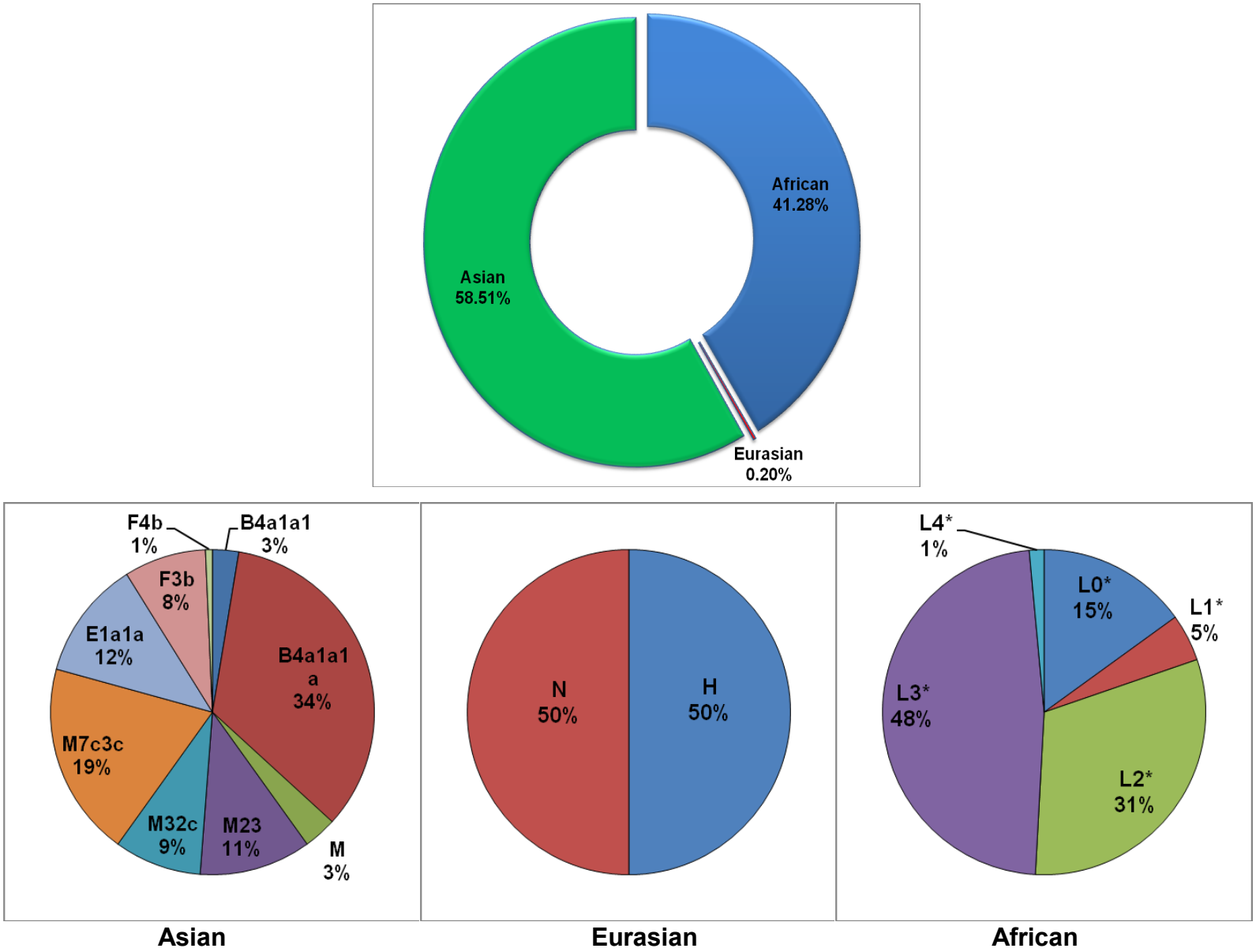


Figure 9: *Pie charts of Asian, Eurasian and African contributions to the maternal gene pool of the Malagasy*. The first pie chart represents all the data in the current study and its contributions per region. The remaining three pie charts represent a haplogroup breakdown at each of the global levels.

In order to further assess the distribution of haplogroups throughout Madagascar, haplogroup frequency maps were created in Surfer v.10 Demo and SAGA (System for Automated Geoscientific Analyses) programs, using haplogroup frequency data based on twenty one ethnic groups within Madagascar. These maps, represented by Figures 10 and 11 were created to show the range distribution of haplogroups B4a1a1a, M7c3c, M23, M32c, L0a2, L2a1a, L2a1b and L3b1a throughout the island, where the darker the colour the more concentrated the haplogroup. All haplogroups except M23 and M32c have a bi or multimodal distribution pattern, which has a single point with the highest frequency and gradually declining frequencies from that point. Haplogroups M23 and M32c show a unimodal distribution pattern.

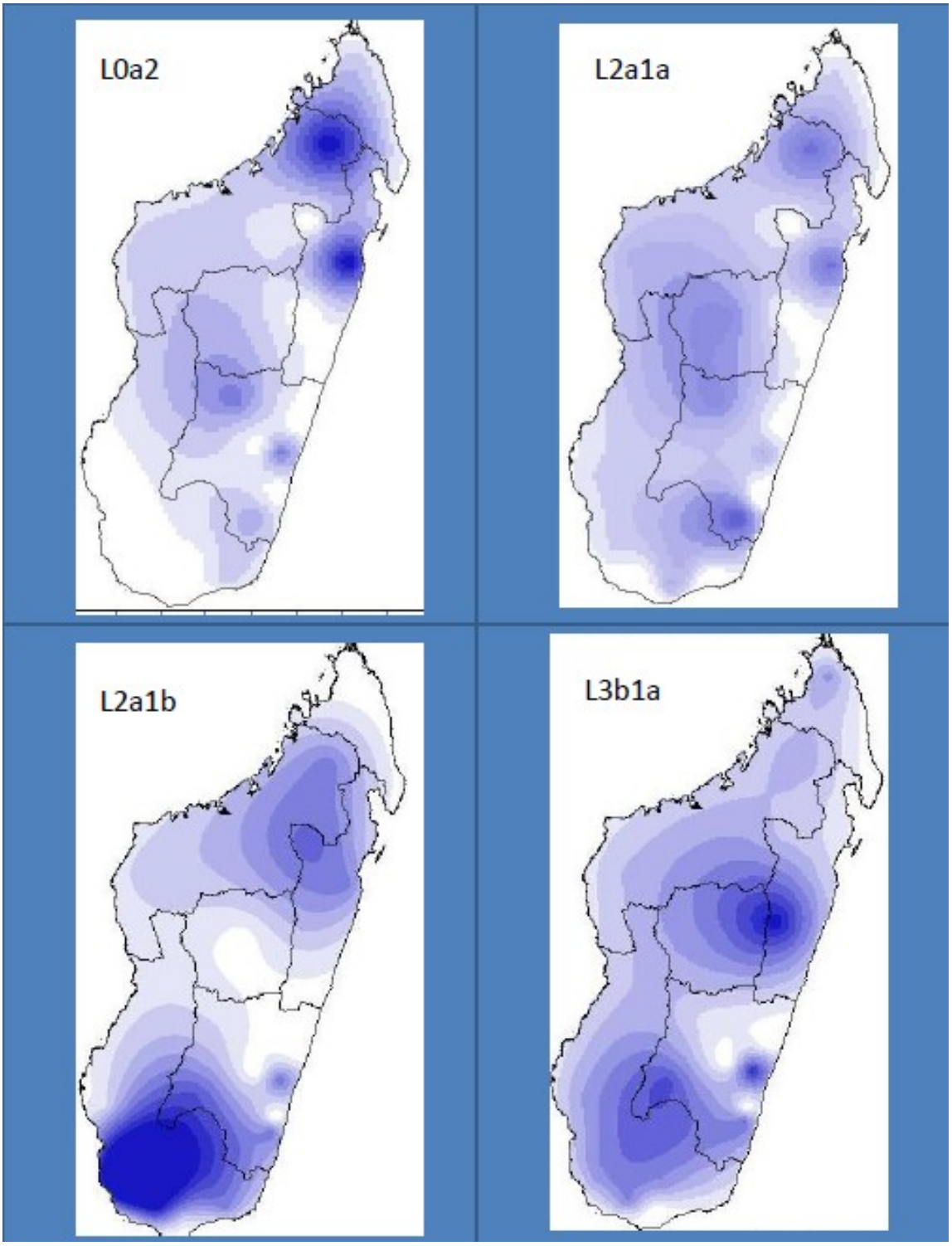


Figure 10: *Haplogroup frequency maps of African haplogroups within Madagascar*. The dark blue colour represents the highest concentration of a particular haplogroup. As the colour becomes lighter, the less frequent the haplogroup in that area. Bi and multi-modal distribution patterns are evident.

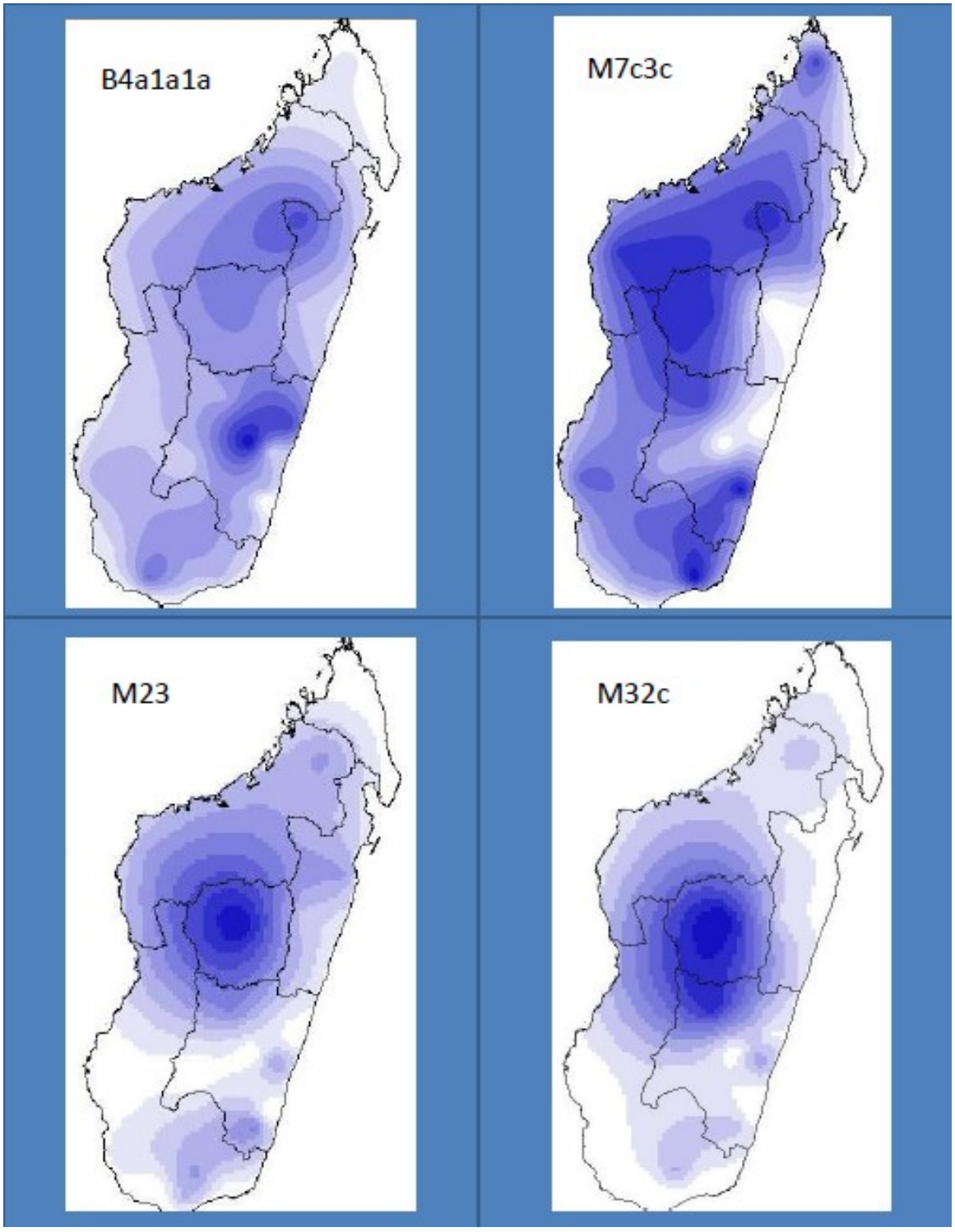


Figure 11: *Haplogroup frequency maps of Asian haplogroups within Madagascar*. The dark blue colour represents the highest concentration of a particular haplogroup. As the colour becomes lighter, the less frequent the haplogroup in that area. Bi and multi-modal distribution patterns are evident.

Phylogenetically informative SNP data obtained through the use of the mitochondrial Single Base Extension assay, in conjunction with hypervariable region sequence data delineated 295 unique mtDNA sequences from a sample pool of 981 sequences. Associations between these unique haplotypes within the macrohaplogroups were assessed using neighbour joining trees (Figures 12a and 12b). The neighbour joining method was used as it is a fast algorithm and is not computationally prohibitive. The phylogenetic trees demonstrate an overall grouping of sub-haplogroups with macro-haplogroups. In Figure 12a, in certain cases, especially with haplogroups L3 and M, the phylogeny lacked structure and haplogroups that were meant to group together, did not. This is somewhat overcome in Figure 12b, where hypervariable region SNP data was added to the sequence data allowing for more structure to haplogroups L3 and M. This is demonstrated further between Figure 12a and 12b by haplogroup L2. Furthermore, when SNP data is included, haplogroups F3b and F4b correctly group together (Figure 12b) where initially without the SNP information, they did not (Figure 12a). The annotations for figure 12a and 12b are representative of 295 unique haplotypes, labelled “HT”. By comparing two phylogenies of the same sequences with (Figure 12a) and without (Figure 12b) phylogenetically informative hypervariable region SNPs, it is clear that these SNPs assist in providing further structure to the tree by correctly grouping sub-haplogroups under relevant macro-haplogroups. The branches of these trees represent each unique haplotype. In the case of this analysis, the colours, which represent each haplogroup are of importance.

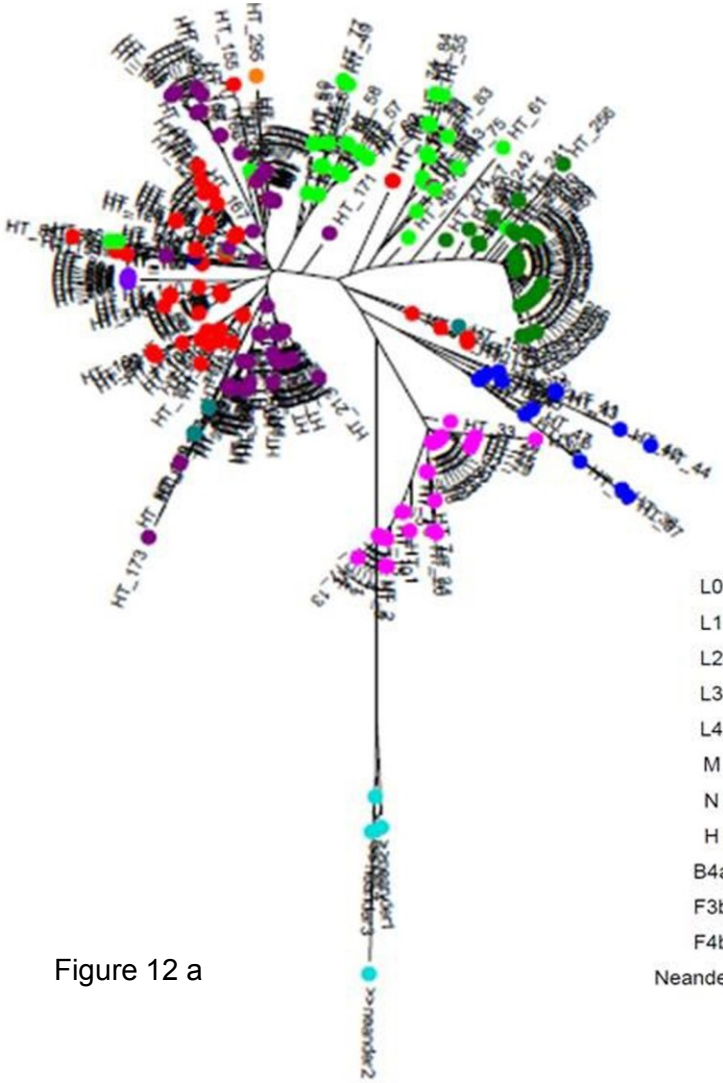




Figure 12 a

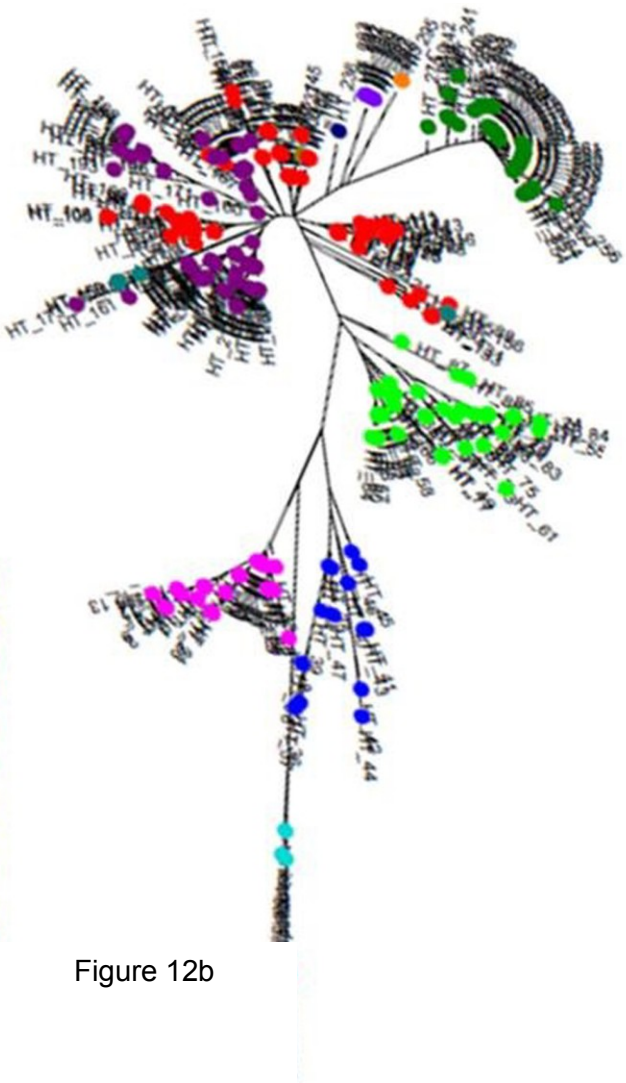


Figure 12b

Figures 12a and b: *Neighbour joining tree representing 295 unique sequences* without (Figure 12a) and with (Figure 12b) mtDNA hypervariable region SNPs. The phylogenies are each rooted with four Neanderthal mtDNA sequences.
These phylogenies show that adding SNP data to sequence data can assist in assigning samples to correct haplogroups.

3.1.3. Variation within twenty one Malagasy ethnic groups

Despite historical records indicating different origins of various Malagasy ethnic groups, in accordance with mtDNA, the present day population is homogenous. The bar chart (Figure 13), using haplogroup distribution displays contributions from African and Asian sources towards the mtDNA. In some cases there were relatively equal contributions from Asian and African sources (Betsileo an Antambohoaka and Tanala). In other's this is not the case. The Sakalava, who are thought to have a higher African origin, according to this data show a greater Asian contribution. This could be explained by sample size bias. A neighbour joining tree created using Fst distances between ethnic groups in Figure 14 shows homogeneity between these groups. The tree also shows some level of separation between Highland (Red) and Lowland (Blue) groups. The tree was rooted using four Neanderthal sequences and Vietnam, China, Mozambique and Kenya were chosen to represent Asian and African populations respectively.

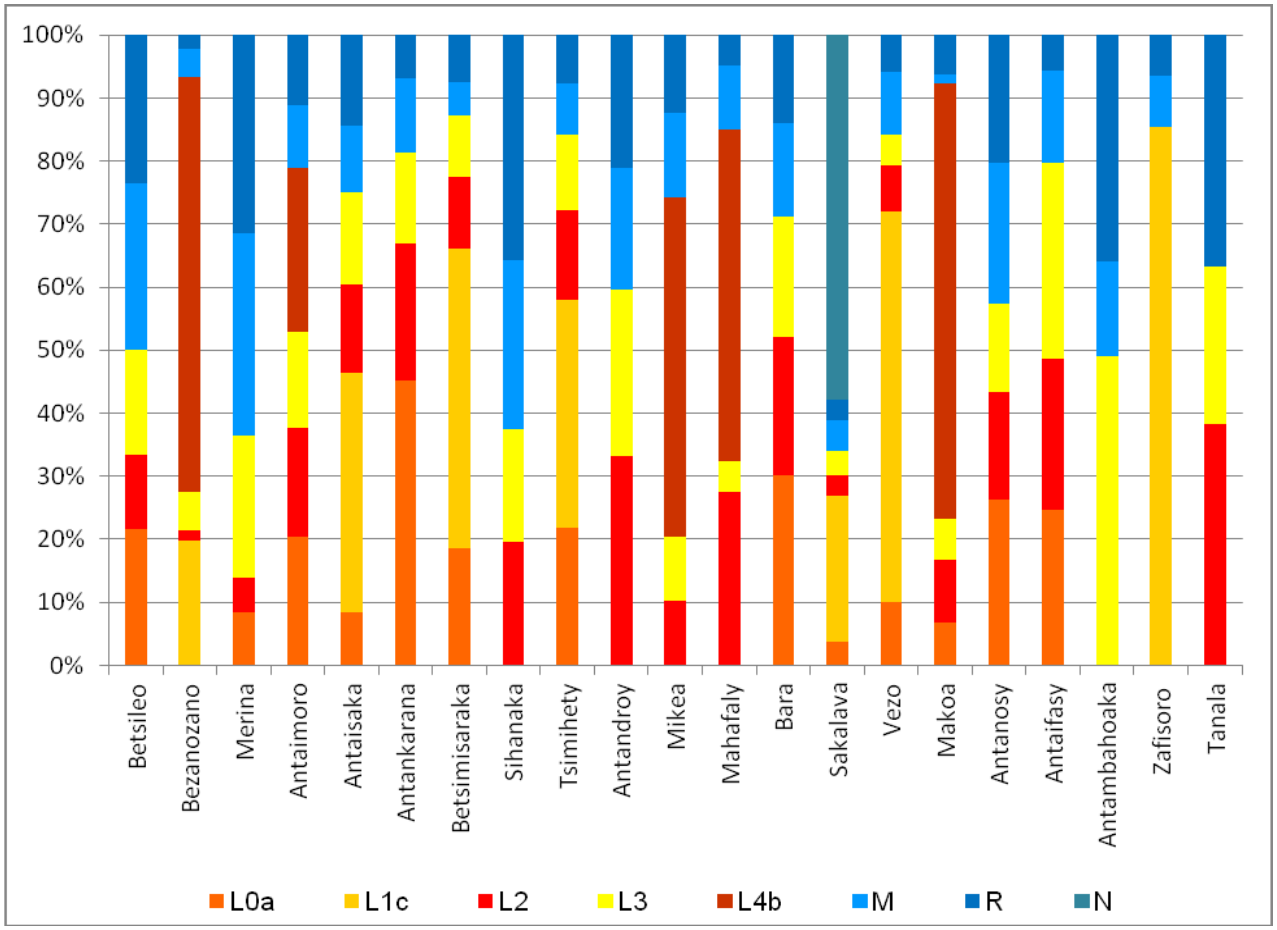


Figure 13: *Bar chart representing the mtDNA composition of 21 Malagasy ethnicities*. All of the ethnic groups have contributions from African (yellow, orange, red and brown colours) and Asian(blue) respectively.

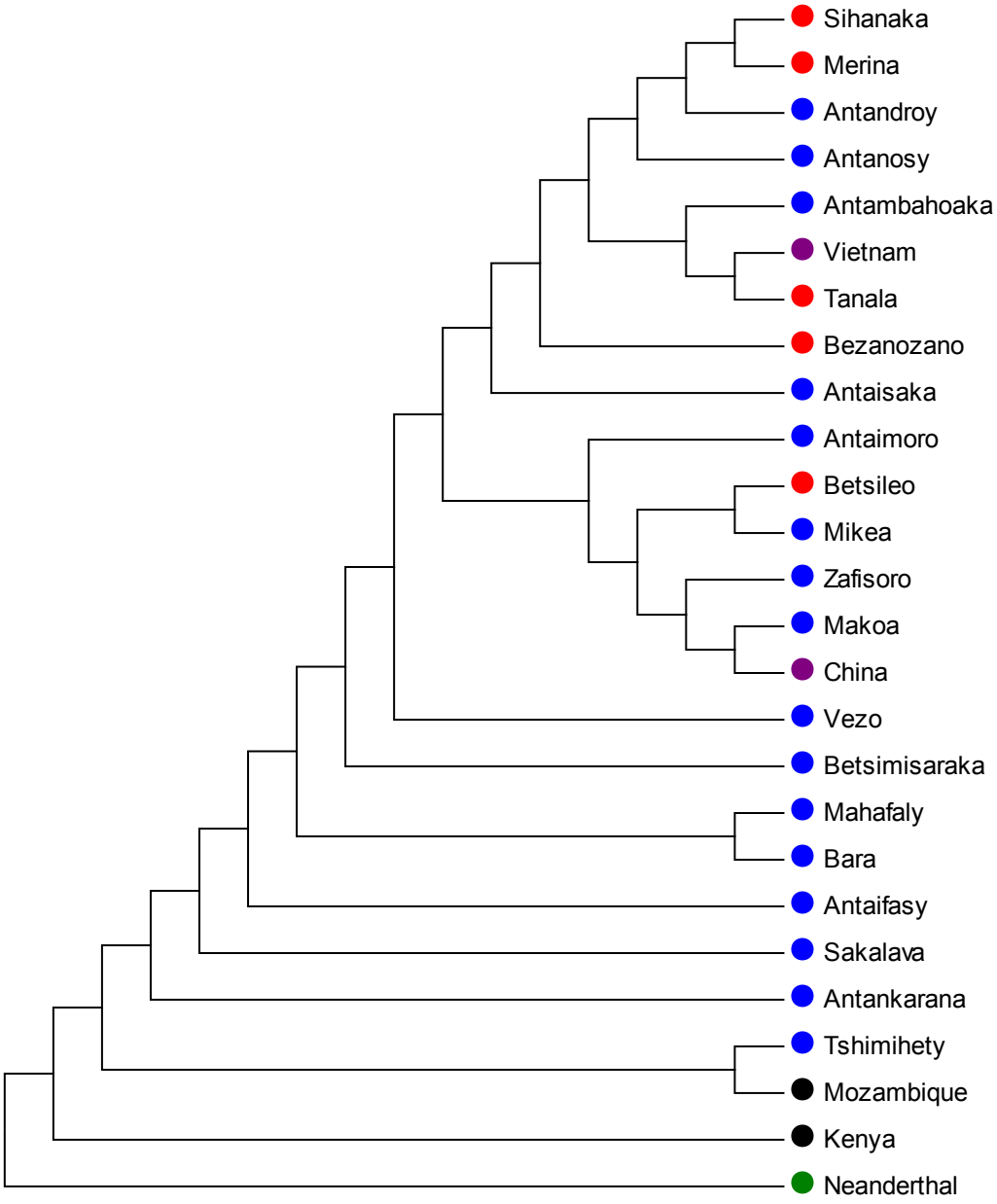


Figure 14: *Neighbour joining phylogeny of 21 Malagasy ethnicities*. Neanderthal sequences were used to root the tree. Red circles represent Highland groups whereas Lowland groups are represented by blue circles. Most of the Lowland groups cluster together, whereas the Highland groups are more evenly distributed. Comparative data from Vietnam and China (purple) group with the historically more Asian, Highland groups, whereas Mozambique and Kenya (black) group with the Highland populations.

In order to further analyse individual haplogroups and the constitution of each Malagasy ethnic group, mismatch distribution analysis was carried out on sequence data to see whether each of these haplogroups and ethnic groups underwent expansions and if so, to date these expansions. The validity of these expansions was also tested. Results are listed in Table 7 and Figure 15 shows the mismatch distribution patterns. The coalescence analysis model which was employed in mismatch distributions analysis assumes a single exponentially growing population (Excoffier and Schneider, 1999).

All haplogroups except L2a1b showed an expansion. Haplogroups that showed expansions with the highest significance included African haplogroup L0a2a2 and Asian haplogroup F3b. Both groups have smooth unimodal curves with low raggedness values. Tau (τ) values differed between African and non-African groups, where African groups showed earlier expansions than the more recent expansions represented by Asian haplogroups. All time values in Table 7 were rounded to the closest 1000 years. Expansions of haplogroups M23 and E1a1a seemed to have occurred around the same time, 2626.05 years before present. Haplogroup L0a2a2, the oldest of the haplogroups within this analysis, showed the oldest date of expansion, 24.7 years before present (Table 7).

Table 7: Mismatch distribution statistics for various mtDNA haplogroups within Madagascar

HG	Raggedness Index	τ	τ (Lower Bound)	τ (Upper Bound)	T=τ/2μ	T=τ/2μ (Lower Bound)	T=τ/2μ (Upper Bound)	Model (SSD) p-value
B4a1a1	0.0677	0.9	0.6	1.3	4.5	3.2	6.8	0.3800
E1a1a	0.1490	0.5	0.2	1.0	2.6	1.2	5.3	0.3140
L0a2a2	0.0262	4.7	0.3	12.4	24.7	1.5	65.3	0.7280
L2a1b	0.0480	0.5	0.3	1.0	2.7	1.3	5.3	0.0030
L3b1a	0.0483	1.1	0.5	2.3	5.5	2.8	12.1	0.5100
L3d1a1	0.3034	3.0	0.3	3.5	15.8	1.8	18.4	0.3850
L3e3	0.0452	2.2	1.3	3.1	11.5	7.0	16.1	0.3280
M23	0.1169	0.5	0.2	1.0	2.6	1.3	5.3	0.2760
F3b	0.0354	1.1	0.3	2.1	5.9	1.3	11.3	0.5980

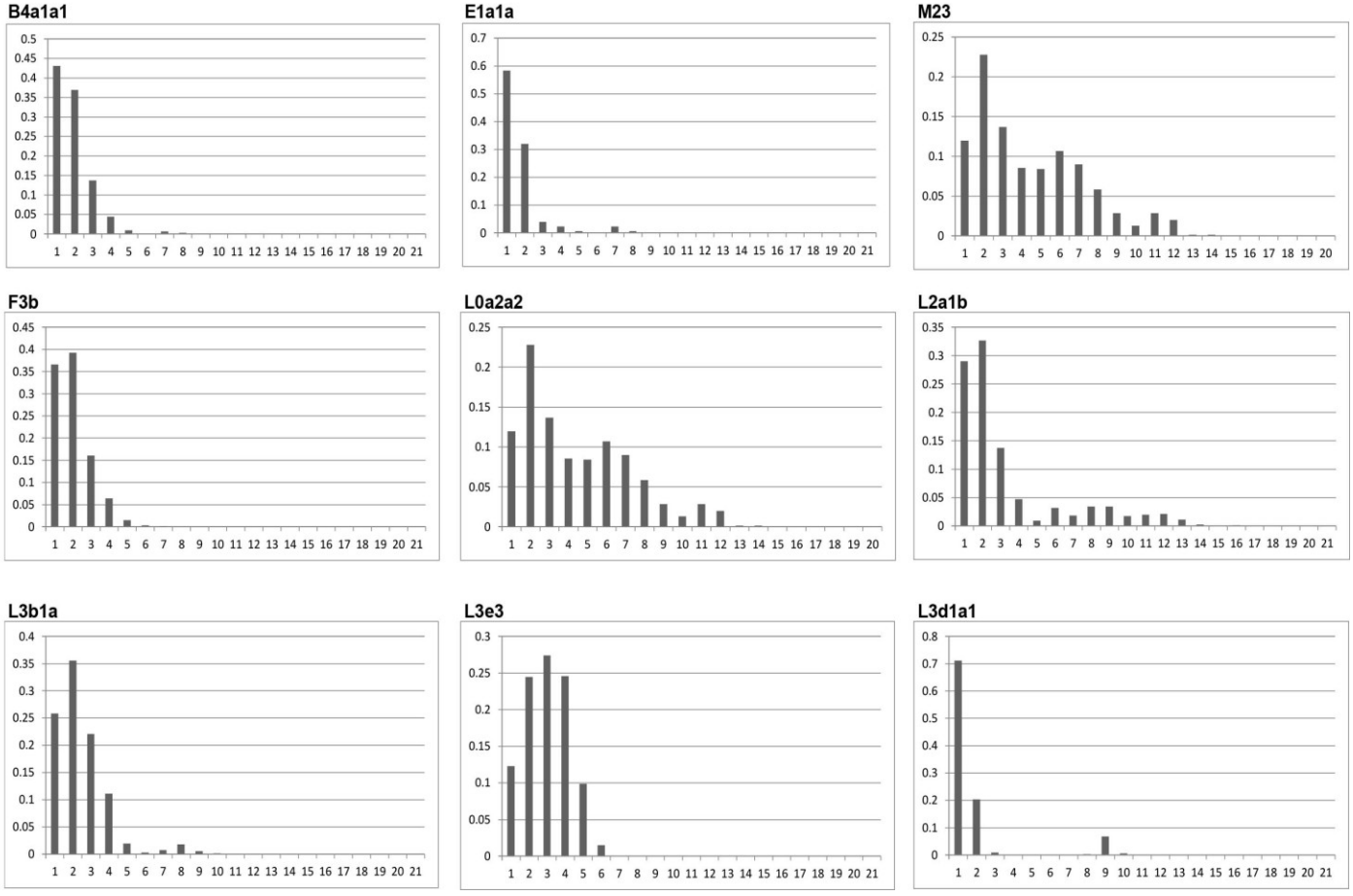


Figure 15: *Mismatch distribution patterns for some African and Asian mtDNA haplogroups*. The highest point in the curve will represent the data at which the haplogroup underwent an expansion.

In addition to mismatch distribution analysis, which has a lesser ability to infer population expansions due to the assumption of a single exponentially growing population, neutrality test summary statistics, which have been reported to show enhanced ability to detect expansions, were carried out (Ramos-Onsins and Rozas, 2002; Pilkington et al., 2008). These include Tajima's D (Tajima, 1989), Fu's Fs (Fu, 1997) and the R2 statistic of Ramos-Onsins and Rozas, 2002. All statistics on haplogroups with more than 10 sequences are represented in Table 7. All test statistics were calculated using DnaSP v4.10. A generation time of 25 years and mutation rate of 2.5*10^-6 per nucleotide per generation (Ward et al., 1991) was used to convert Tau was (τ) to time when expansion occurred (T). This was done using the equation T= τ/2μ, where μ is the mutation rate per gene per generation. As an example, the expansion for haplogroup B4a1a1 was calculated as:

(2.5*10-6)*950bp (total sequence length) = 0.00238 per gene per generation

T= 0.8652/ 2*(0.00238)

= 181.76 generations

= 4544 years (generations x 25)

Furthermore, the effective population size of females (N_e) was estimated from Watterson's estimator θ per sequence. This was calculated as W (θ) / 2μ. In order to test for deviations from the assumptions of neutrality and constant population size, neutrality tests were conducted. A significantly negative value for Tajima's D and Fu's Fs values would suggest population growth and/or positive selection. The same would apply for a significantly positive R2 value. Fu's Fs value is based on the probability of drawing a number of haplotypes that is equal to or greater than the observed number of samples from a population. This method is useful and yields a better outcome to detect population expansions when samples sizes are greater than 50. The R2 value is based on the difference between the average number of nucleotide differences and

the number of unique mutations. This method is suited for detecting population expansions with small sample sizes (Ramos-Onsins and Rozas, 2002; Pilkington et al., 2008).

Of all haplogroups, Asian lineage B4a1a1 had the largest N_e. This was followed by African haplogroups L0a2a2 and L2a1b. The smallest N_e was represented by haplogroup L3d1a1. With Taijma's D and Fu's Fs values, a significantly negative value indicates possible growth or selection. With the R2 statistic, a significantly positive value is indicative of positive selection or population expansion (Jobling, Hurles and Tyler-Smith, 2004). In Table 8 the Taijma's D value for all haplogroups was negative and all expect haplogroup F3b were significant, indicative of population growth or selection. All haplogroups had negative Fs and positive R2 values, however all except L3d1a1 showed significance, which could be explained by sample size.

Table 8: Diversity statistics for various mtDNA haplogroups within Madagascar

HG	Hd	π	W (θ)	Ne	Tajima's D	Tajima's D p-value	Fs	Fs p-value	R2	R2 p-value
B4a1a1	0.5940	0.00095	8.09921	1705	-2.61994	0.0000	-29.3656	0.0000	0.0115	0.0069
E1a1a	0.4166	0.00070	3.75834	791	-2.46853	0.0000	-16.6123	0.0000	0.0334	0.0000
L0a2a2	0.8805	0.00364	7.14016	1503	-1.79169	0.0123	-9.31385	0.0007	0.0505	0.0000
L2a1b	0.7101	0.00244	6.84853	1442	-2.11854	0.0007	-6.74845	0.0080	0.0383	0.0082
L3b1a	0.7419	0.00153	3.40696	717	-1.71861	0.0165	-7.84816	0.0010	0.0506	0.0466
L3d1a1	0.2892	0.00088	2.62057	552	-2.22358	0.0018	-1.24623	0.1443	0.1359	0.4890
L3e3	0.8772	0.00210	3.54786	747	-1.39952	0.0491	-8.40763	0.0001	0.0622	0.0365
M23	0.4385	0.00059	3.17241	668	-2.42966	0.0002	-16.6815	0.0000	0.0320	0.0281
F3b	0.6605	0.00104	2.03773	429	-1.4665	0.0544	-4.64549	0.0093	0.0637	0.0000

In order to further analyse ethnic groups of the island, mismatch distribution analysis was carried out on sequence data per ethnic group to see whether any of these groups underwent expansions and if so, the time at which these expansions took place. Results are listed in Table 9 and Figure 16.

To assess the expansion dynamics of each of the ethnicities, mismatch distribution analysis was carried out. All time values in Table 9 were rounded to the closest 1000 years. All ethnic groups except Betsimisaraka, Mikea and Sinhanaka showed an expansion. Ethnic groups that showed expansions with the highest significance included Bara and Makoa. All ethnicities have smooth unimodal curves with low raggedness values, indicative of an expanding population, Antankarana showed the most recent date of expansion of 39 974 years before present, whereas the Antaifasy ethnic group displayed the oldest expansion which would have occurred 83 403 years before present. The mismatch distribution model assumes a single expansion and dates the older, more pronounced one within the data. These results are therefore indicative of the out of Africa expansion. However, the expansion that can be seen closer to the "0" value could be a more recent expansion, suggesting this as the time at which Madagascar was populated. Many ethnic groups including Vezo, Mikea, Bezanozano, Antambahoaka, Tanala, Antandroy, Merina, Sihanka and Makoa expanded within approximately 6000 years of each other. Betsimisaraka, Antaimoro, Antaisaka, Sakalava and Betsileo also demonstrated expansions within the same time frames, all happening within 4000 years of each other. Once all ethnic groups were combined, the mean date of expansion was represented by the "Madagascar Combined" date of 57 773 years before present.

Table 9: Mismatch distribution statistics for various ethnic groups within Madagascar

Ethnicity	Raggedness Index	τ	τ (Lower Bound)	τ (Upper Bound)	T=τ/2μ	T=τ/2μ (Lower Bound)	T=τ/2μ (Upper Bound)	Model (SSD) p-value
Antaifasy	0.040	15.9	8.9	20.7	83.4	47.0	108.8	0.39
Antaimoro	0.006	10.4	6.6	24.9	54.6	34.5	130.9	0.33
Antaisaka	0.005	10.6	7.0	18.2	55.9	36.9	96.0	0.19
Antambahoaka	0.240	12.4	6.5	18.2	65.0	34.0	95.9	0.19
Antandroy	0.009	12.8	9.6	15.4	67.4	50.3	81.2	0.40
Antankarana	0.015	7.6	3.8	24.6	40.0	20.1	129.7	0.65
Antanosy	0.008	11.2	5.1	23.9	58.6	26.7	125.8	0.21
Bara	0.013	9.3	4.4	22.5	48.6	22.9	118.5	0.73
Betsileo	0.005	10.9	6.9	17.8	57.1	36.4	93.9	0.33
Betsimisaraka	0.007	10.2	6.8	21.2	53.6	35.8	111.8	0.08
Bezanozano	0.014	12.2	7.6	14.9	63.9	40.1	78.3	0.62
Mahafaly	0.027	15.4	10.4	21.0	81.1	54.9	110.6	0.23
Makoa	0.016	13.1	7.4	16.8	68.6	38.9	88.3	0.72
Merina	0.010	12.9	8.1	15.8	67.8	42.6	83.1	0.11
Mikea	0.018	12.0	8.9	14.1	63.0	47.1	74.0	0.09
Sakalava	0.004	10.7	6.7	20.3	56.2	35.0	107.0	0.19
Sihanaka	0.029	12.9	6.9	16.8	68.0	36.1	88.5	0.04
Tanala	0.278	12.7	5.3	16.5	66.8	27.7	86.9	0.42
Tsimihety	0.003	9.0	5.1	25.5	47.1	26.8	134.1	0.23
Vezo	0.008	11.9	7.9	14.6	62.7	41.7	77.0	0.46
Zafisoro	0.236	8.4	2.2	27.5	44.0	11.7	144.9	0.21
Madagascar Combined	0.004	11.0	6.8	20.3	57.8	35.6	107.0	0.20

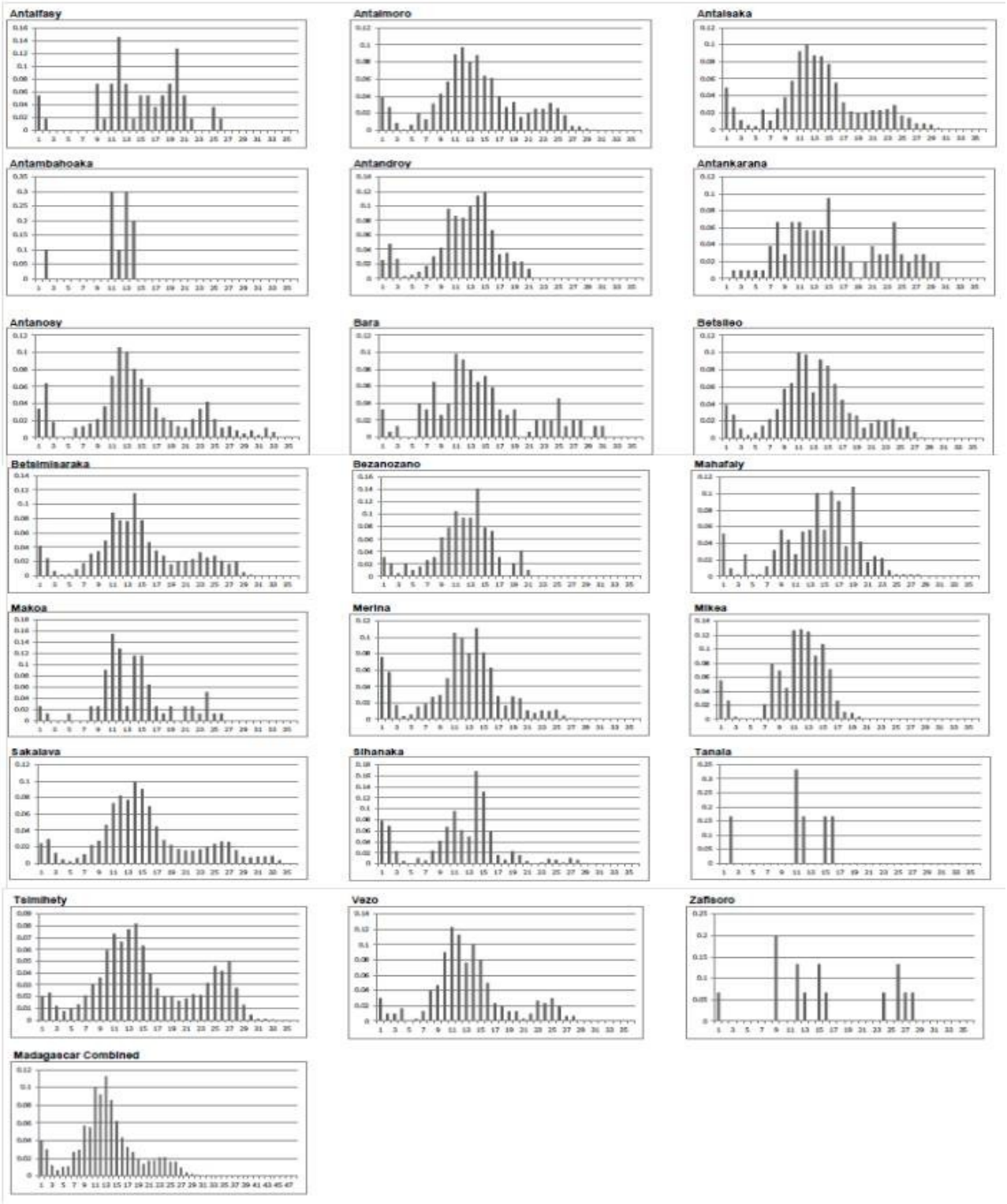


Figure 16: *Mismatch distribution patterns for each of the 21 ethnic groups in Madagascar*. The highest point in the curve will represent the data at which the ethnic group underwent an out of Africa expansion, however the highest point on the smaller curve, closest to the "0" value will represent recent expansion of that ethnic group in Madagascar.

Table 10 shows summary statistics of Tajima's D (Tajima, 1989), Fu's Fs (Fu, 1997) and the R2 statistic (Ramos-Onsins and Rozas, 2002) for the different Malagasy ethnic groups.

Of all ethnic groups the Sakalava had the largest N_e. This was followed by the Vezo and Antanosy respectively. The smallest N_e was represented by the Tanala people. In Table 10 Taijma's D value was negative but not significant for any one ethnic group. However, when all samples were combined into Madagascar Combined, Tajima's D was significantly negative indicative of population growth. All ethnic groups had negative Fs and positive R2 values, however not one of the ethnic groups showed significance. Following the combination of all ethnic groups, Fs and R2 values displayed significance.

Table 10: Diversity statistics for various ethnic groups within Madagascar

Ethnicity	Hd	π	W (θ)	Ne	Tajima's D	Tajima's D p-value	Fs	Fs p-value	R2	R2 p-value
Antaifasy	0.9455	0.01455	16.38802	3450	-0.73967	0.2302	-0.13301	0.4254	0.1072	0.0563
Antaimoro	0.9617	0.01351	17.72236	3731	-0.92417	0.1764	-7.05922	0.0630	0.0697	0.1745
Antaisaka	0.9532	0.01325	17.03025	3585	-0.85654	0.2101	-11.79268	0.0154	0.0691	0.2144
Antambahoaka	1.0000	0.01095	10.56000	2223	-0.11235	0.5405	-0.28641	0.2604	0.1402	0.0711
Antandroy	0.9147	0.01167	14.1788	2985	-0.77555	0.238	-10.06662	0.0079	0.0829	0.2124
Antankarana	1.0000	0.01566	17.83758	3755	-0.71826	0.2439	-5.47703	0.0102	0.1035	0.1082
Antanosy	0.9664	0.01454	17.97072	3783	-0.85563	0.2049	-6.08468	0.0365	0.0852	0.1732
Bara	0.9673	0.01403	19.18854	4040	-1.27004	0.0937	-1.56272	0.2315	0.08770	0.0436
Betsileo	0.9609	0.01279	16.52745	3479	-0.86000	0.2094	-12.60421	0.0133	0.0676	0.2029
Betsimisaraka	0.9578	0.01414	19.32383	4068	-1.01407	0.1498	-17.13706	0.0023	0.0655	0.1410
Bezanozano	0.9684	0.01157	10.99291	2314	0.00066	0.5551	-1.43286	0.2673	0.134	0.5806
Mahafaly	0.9483	0.01377	19.60699	4128	-1.27010	0.0853	-3.67761	0.0987	0.0701	0.0256
Makoa	0.9872	0.01351	17.07909	3596	-1.11887	0.1273	-2.52358	0.0898	0.0997	0.0501
Merina	0.9244	0.01143	14.36144	3023	-0.77052	0.2503	-22.36942	0.0007	0.0669	0.2524
Mikea	0.9447	0.01099	12.22853	2574	-0.53457	0.3297	-2.54712	0.1849	0.0971	0.3035
Sakalava	0.9756	0.01503	20.75313	4369	-1.06526	0.1342	-16.13657	0.0011	0.0679	0.1341
Sihanaka	0.922	0.01126	13.51971	2846	-0.73118	0.2561	-1.56123	0.3346	0.0891	0.3090
Tanala	1.0000	0.0107	9.81818	2067	0.36286	0.7332	0.35340	0.3439	0.1975	0.2274
Tsimihety	0.9814	0.01532	19.6986	4147	-0.87226	0.2047	-23.15552	0.0003	0.0704	0.2027
Vezo	0.9700	0.01304	18.53834	3903	-1.29347	0.0855	-3.66300	0.082	0.074	0.0296
Zafisoro	0.9333	0.01594	17.08029	3596	-0.72682	0.2973	1.54292	0.7013	0.1712	0.2957
Madagascar Combined	0.9619	0.01334	28.66626	6035	-1.60435	0.0122	-23.46122	0.0155	0.0291	0.0

3.1.4. Genetic affinity of the Malagasy in relation to global populations

Further, the relationships between haplotypes from this and comparative datasets were assessed through network analyses of various African (Figures 17a-d) and Asian (Figures 18a-c) haplogroups.

Haplogroup L0a represents one of two branches from the maternal most recent common ancestor (MRCA) and is dated to ~100 000 years old. This haplogroup is commonly found in central, eastern and southeastern Africa yet is almost absent in north and west Africans (Salas, 2002). A phylogenetic analysis was carried out on haplogroup L0a2 samples through a median joining network algorithm in Network 4.5 software. This network contained 535 sequences, including those from the present study and comparative sequences from South Africa, Mozambique, Ethiopia, Senegal, Kenya, Tanzania, Central Africa, Gabon, Cameroon, North Africa (undefined), India and the Middle East. The network represented in Figure 17a shows that haplogroup L0a2 is widespread across the African continent and expanded rapidly therefore allowing no time to show differentiation across geographic regions. Sequences from Madagascar show a great deal of type sharing between sequences from Mozambique, Tanzania and South Africa. This is consistent with haplogroup L0a being prevalent in Southeast Africa; found at 25% in Mozambique (Rosa, Brehm, Kivisild, Metspalu and Villems, 2004).

While the distribution of haplogroup L2a1 and its sub-haplogroups are widespread across the island, they are more commonly found in the Northern and Southwestern regions (Figure 10) Network analysis (Figure 17c) shows that haplogroup L2a1 also demonstrates the regional specificity of haplogroups. In this median joining network of 396 sequences, there is a large amount of type sharing, which suggests that haplotypes found in Madagascar are also

commonly found in South Africa and Mozambique, and to a lesser extent in the Middle East and certain parts of North Africa.

The Malagasy sample also shows appreciable frequencies of L3 and its sub-haplogroups. L3b and L3e (and sub-haplogroups) found at frequencies of 6.93% and 8.76% respectively. Haplogroup L3 and its sub-haplogroups are found throughout Africa with L3d and L3e being the most common. Within the Malagasy L3b1a and L3e3 are the most frequent. Both of these sub-haplogroups have origins in West African Bantu-speakers but are distributed throughout Eastern and Southern Africa (Salas et al., 2002). Median joining networks were created for both L3b and L3e samples (Figures 17b and 17d). The L3b network included 200 sequences whereas the L3e network contained 469 sequences. Both these networks show no clear structure and cannot definitively pinpoint any one population as being the source population of this haplogroup within the Malagasy, however, there is a strong affinity of the Malagasy samples towards samples from Gabon within haplogroups L3e2 and L3e3 and within L3e1 the Malagasy samples show an affinity for samples from Gabon and Mozambique. The L3b network also provides no definitive structure with only a few haplotypes being shared between Madagascar and Benin. This is suggestive of a rapid expansion of these haplogroups, allowing no time for differentiation.

Other African haplogroups and their sub-haplogroups that are represented in this Malagasy sample include L1c (1.93%) and L4 (0.61%) (Figure 8). With regard to haplogroup L1, only L1c and its sub-haplogroups are represented in the Malagasy sample. L1c is the largest and most diverse sub-haplogroup. It is frequently found among central African Bantu speakers and is thought to have originated in western equatorial Africa and is almost absent in eastern and southern Africa (Salas et al., 2002; Vigilant et al., 1991; Destro-Bisol et al., 2004). By matching variant sites from Malagasy L1c2b1 samples to those in a comparative database, 4 matches were found in samples from Gabon, 1 from Mozambique and 1 from Cameroon. Matches for

haplogroup L1c1 revealed exact haplotypic equivalents in 1 sample from Benin, 10 samples from Gabon and 2 samples from Mozambique. L4 and its sub-haplogroups are also found on the island at a frequency of 0.61%. Haplogroup L4 is common in East Africa with its highest frequencies within Hadza (60-83%) and Sandawe (48%) groups of Tanzania (Tishkoff et al., 2007). The closest match between Malagasy L4b samples showed closest matches that differed by 3 positions to samples from Tanzania.

The Asian/Southeast Asian contribution was inferred from the prevalence of haplogroups B (21.50%), M (24.87%), E (6.93%) and F (5.19%). These haplogroups together accounted for 58.51% of the non-African mtDNA lineages in the Malagasy. The remaining 0.20% haplogroups (H and N) was of Eurasian ancestral contribution.

A total of 567 sequences were used to create a median joining network for haplogroup B4a. This network (Figure 18a) contains large nodes indicating a large degree of type sharing between samples from Madagascar, Vietnam, Papua New Guinea, French Polynesia and New Zealand Maoris. Secondary nodes indicate strong type sharing between Madagascar, Taiwan and the Philippines, which in turn show affinity to Southeast China.

Haplogroup M is almost restricted to the Asian continent and is believed to have come to East Asia from migrations through the Indian subcontinent. Haplogroup M and its sub-haplogroups (M23, M32c, M7c3c and E1a1a) are found at frequencies of 1.93%, 6.52%, 5.09%, 11.31% and 6.93% respectively within Madagascar and each of these clades are found at variable frequencies throughout Southeast Asia (Tabbada et al., 2010, Trejeut et al., 2005, Kivisild et al., 2002 and Hill et al., 2007). In this median joining network (Figure 18b), 567 haplogroup and sub-haplogroup M samples were used. Malagasy samples show a strong affinity to those samples from India and a separation is noticeable between samples from India and those from Southeast

China. There is also a certain degree of type sharing between Malagasy, Reunion island and Filipino samples.

Haplogroup F and its sub-haplogroups are geographically localized to Asian populations, including Chinese, Taiwanese, Filipino and at high frequencies in Vietnam. It has also been found to occur in Siberia and Borneo (Hill et al., 2007, Tabbada et al., 2010). Trejeut et al. have reported that F3b has been observed in the Philippines and Sabah from Borneo (Trejeut et al., 2005). Although infrequent, it has also been found to occur in southern and western regions of China (Kivisild et al., 2002, Yao, Kong, Bandelt, Kivisild & Zhang, 2002; Kong et al., 2003). The inverse is true within three southernmost populations of Taiwan where F3b occurs at a high frequency (Trejeut et al., 2005). In an observation by Trejeut et al., it was noted that the variant sites in HVR I differed between Taiwanese and Chinese individuals, whereby Taiwanese F3b sequences were characterized by an A – C transversion at nucleotide position 16220 and lack a transition at nucleotide position 16335 (Trejeut et al., 2005). The same was observed with all of the F3b samples in this current dataset. Furthermore another sub-haplogroup of F was observed in the Malagasy, that of F4b, which was found in over one quarter of Taiwanese individuals in the study by Trejeut et al., 2005, however has been observed at low frequencies <1%) in China Kivisild et al., 2002 and (Kolman et al., 1996) and was completely absent among 519 Thai samples in which the overall haplogroup F frequency was one of the highest in Asia 22%) (Oota et al., 2001; Fucharoen et al., 2001). This was also true for the Malagasy sample set, where F4b was found to occur in four of the 981 samples. A median joining network was created using 262 sequences that were representative of haplogroups F*, F3, F3a, F3b and F4b (Figure 18c). From the network there is minimal type sharing between Taiwanese sequences and Malagasy sequences, however there is a clear separation between sequences from Southeast China and Taiwan, with the Malagasy showing a higher affinity to the Taiwanese.

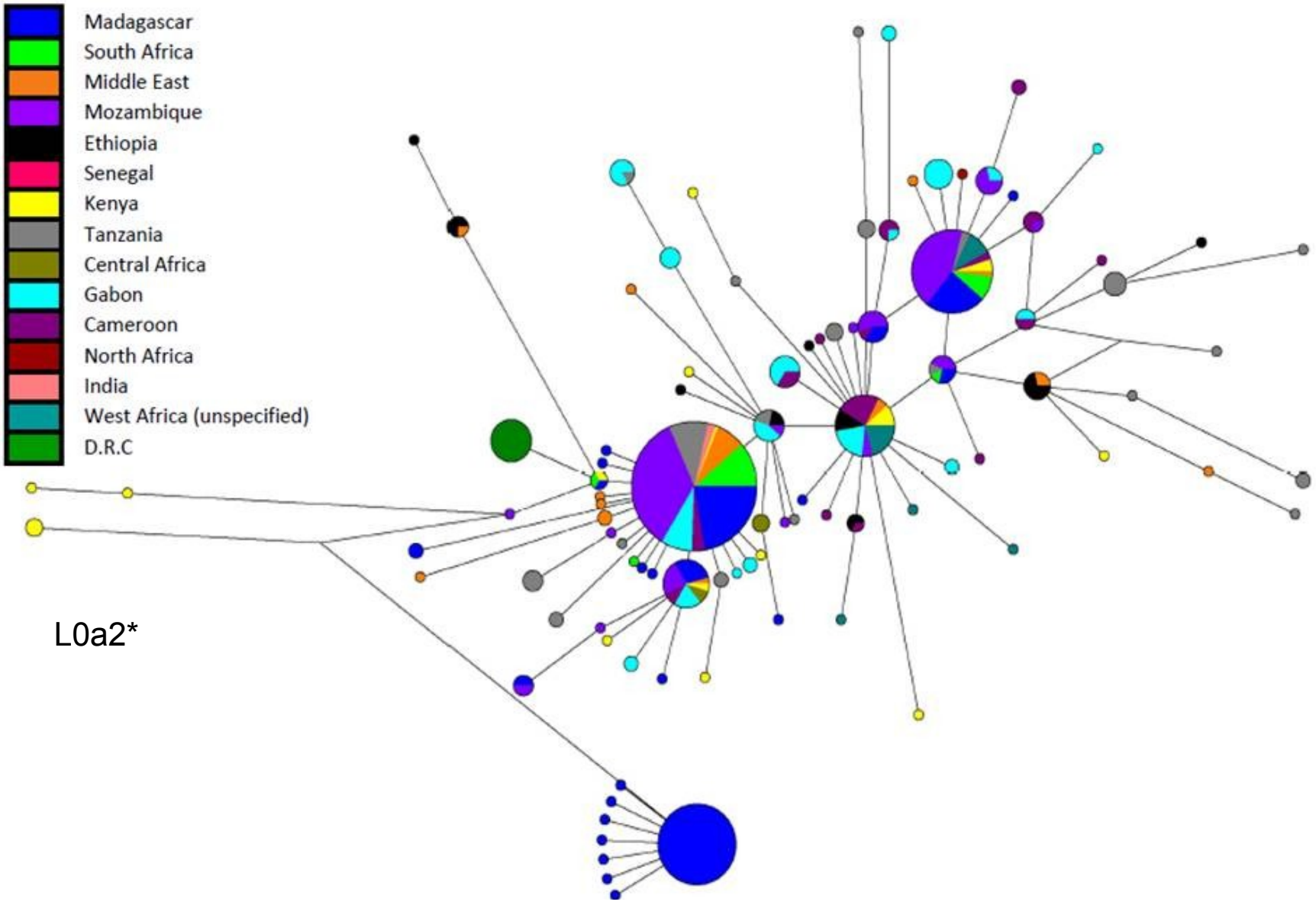


Figure 17a: *Median joining network representing mtDNA haplogroup L0a2**.L0a2 is widespread across the African continent and expanded rapidly therefore allowing no time to show differentiation across geographic regions. Sequences from Madagascar (blue) show type sharing between sequences from Mozambique, Tanzania and South Africa

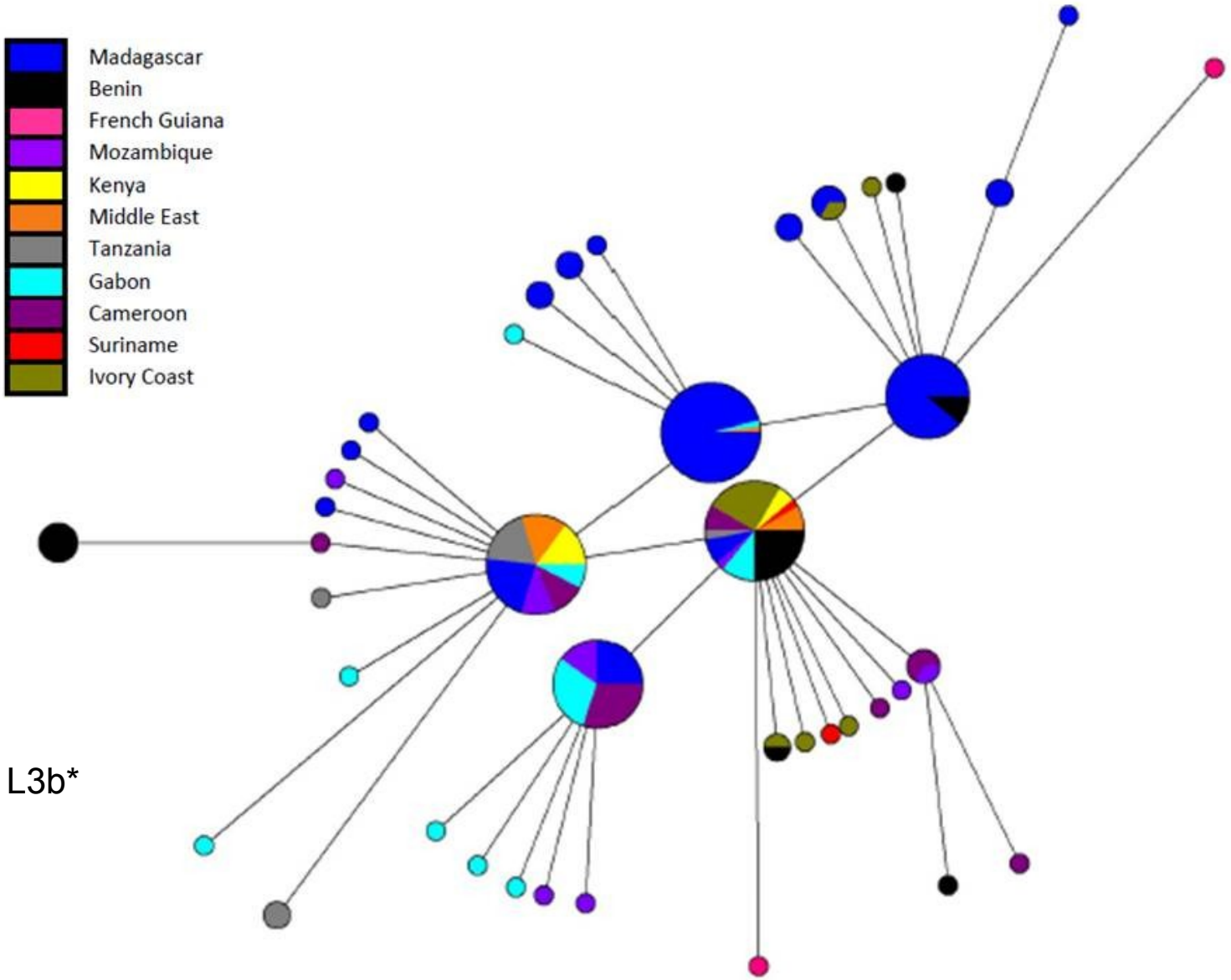


Figure 18: *Median joining network representing mtDNA haplogroup L3b**.The L3b network included 200 sequences. This networks shows no clear structure and cannot definitively pinpoint any one population as being the source population of this haplogroup within the Malagasy

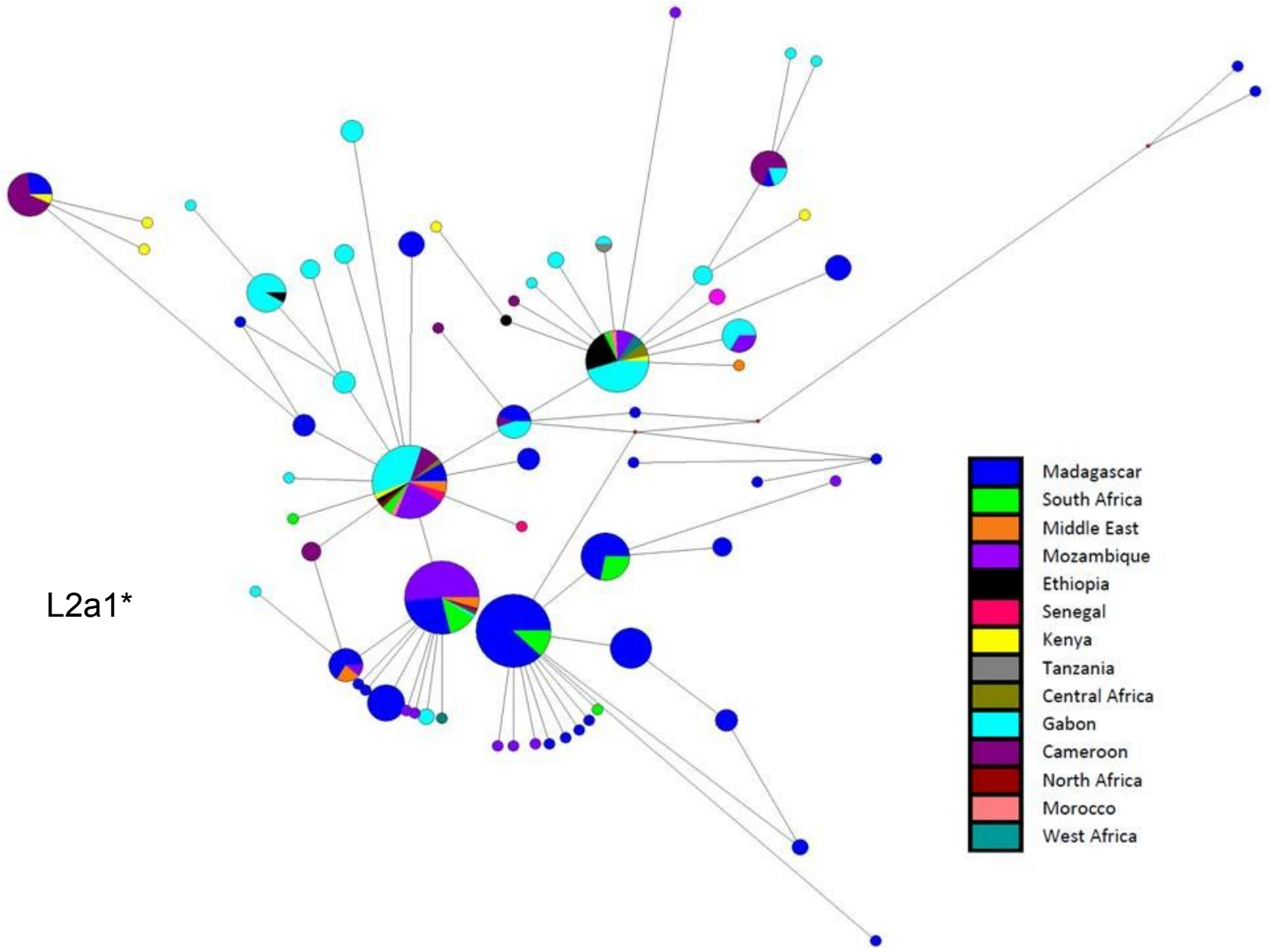


Figure 19: *Median joining network representing mtDNA haplogroup L2a1*.* In this median joining network of 396 sequences, there is a large amount of type sharing, which suggests that haplotypes found in Madagascar are also commonly found in South Africa and Mozambique

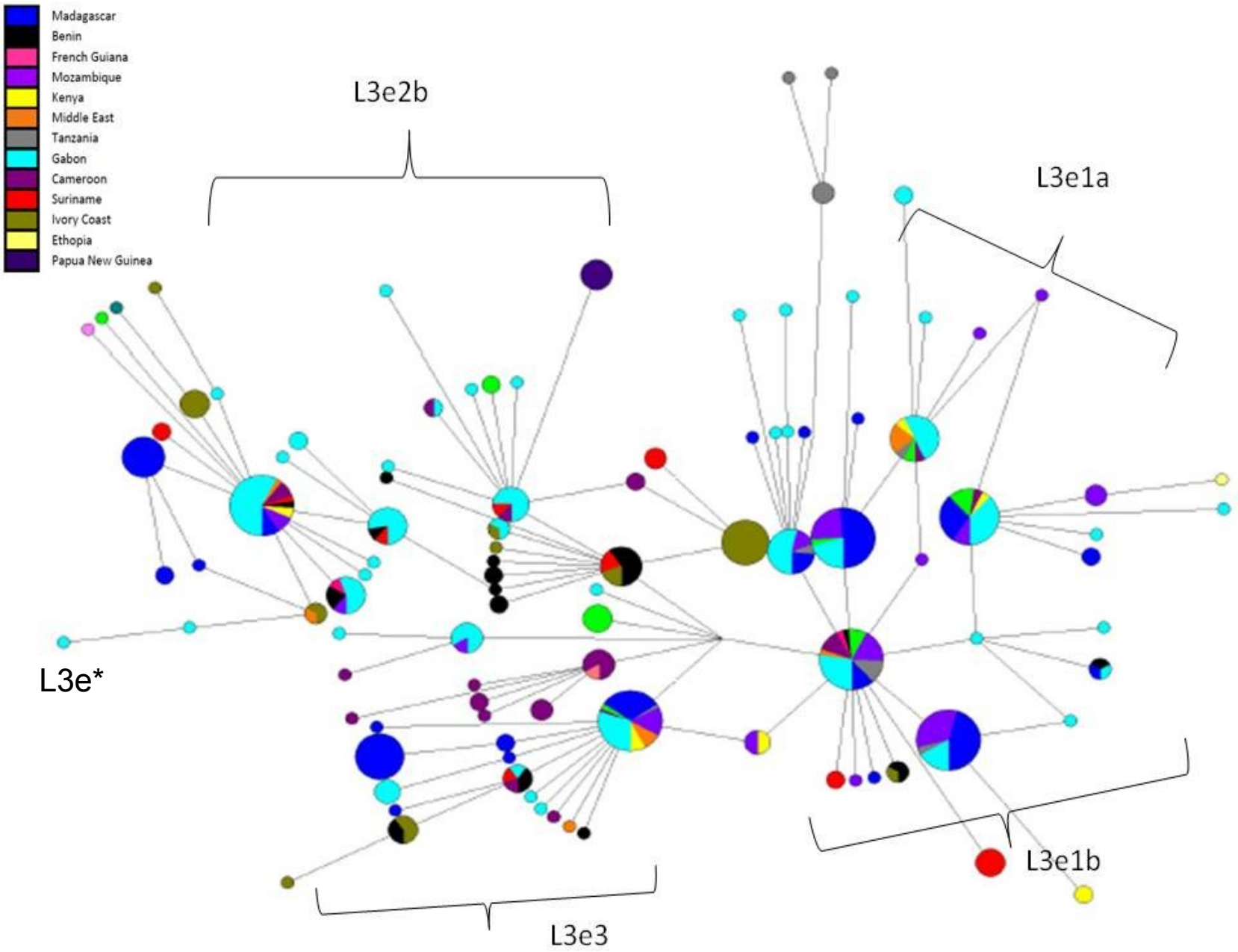


Figure 17d: *Median joining network representing mtDNA haplogroup L3e*. The L3e network contained 469 sequences. This network shows no clear structure and cannot definitively pinpoint any one population as being the source population of this haplogroup within the Malagasy

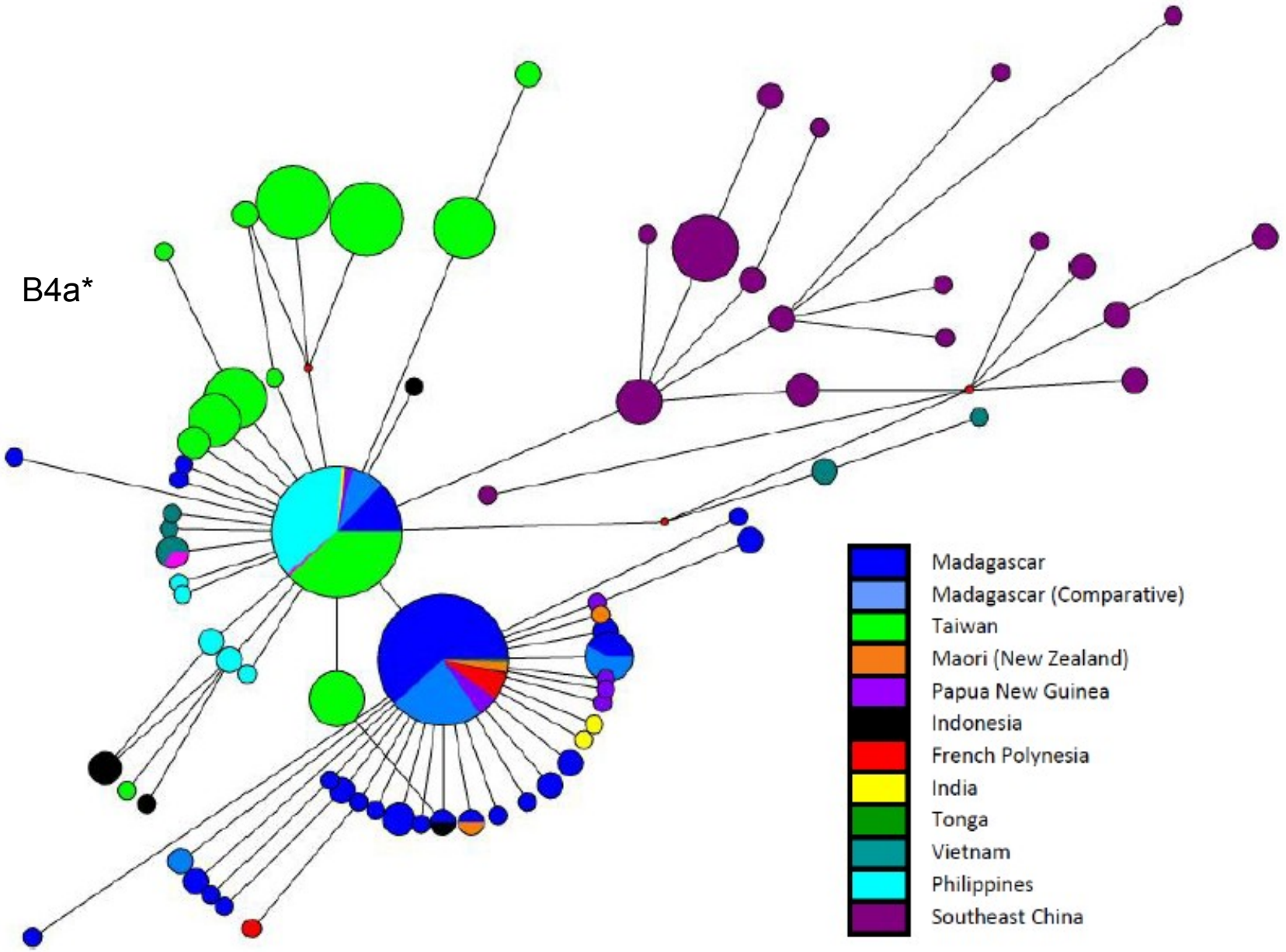


Figure 18a: *Median joining network representing mtDNA haplogroup B4a*.* There is a large degree of type sharing between samples from Madagascar, Vietnam, Papua New Guinea, French Polynesia and New Zealand Maoris.

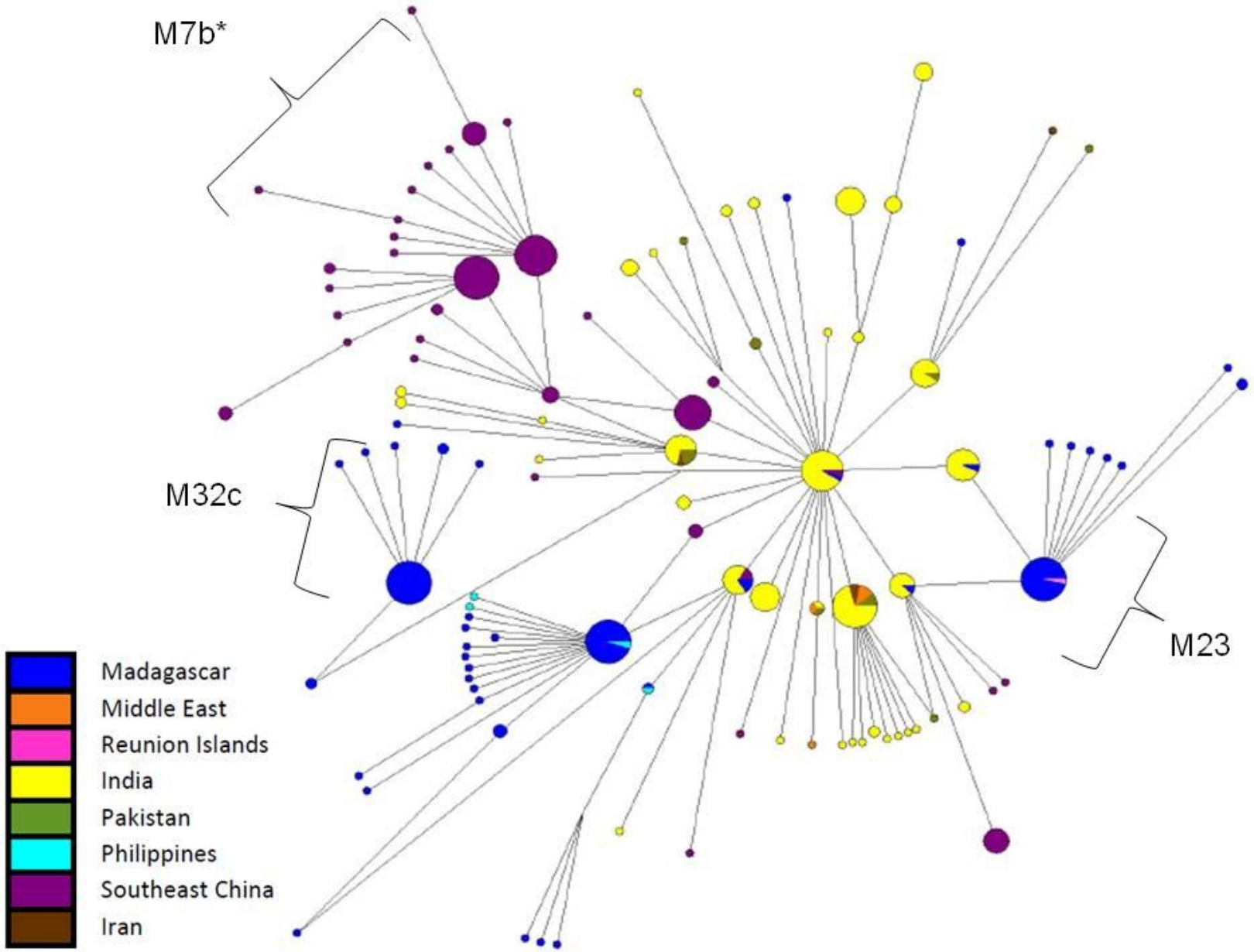


Figure 18b: *Median joining network representing mtDNA haplogroup M**. Malagasy samples show a strong affinity to those samples from India and a separation is noticeable between samples from India and those from Southeast China

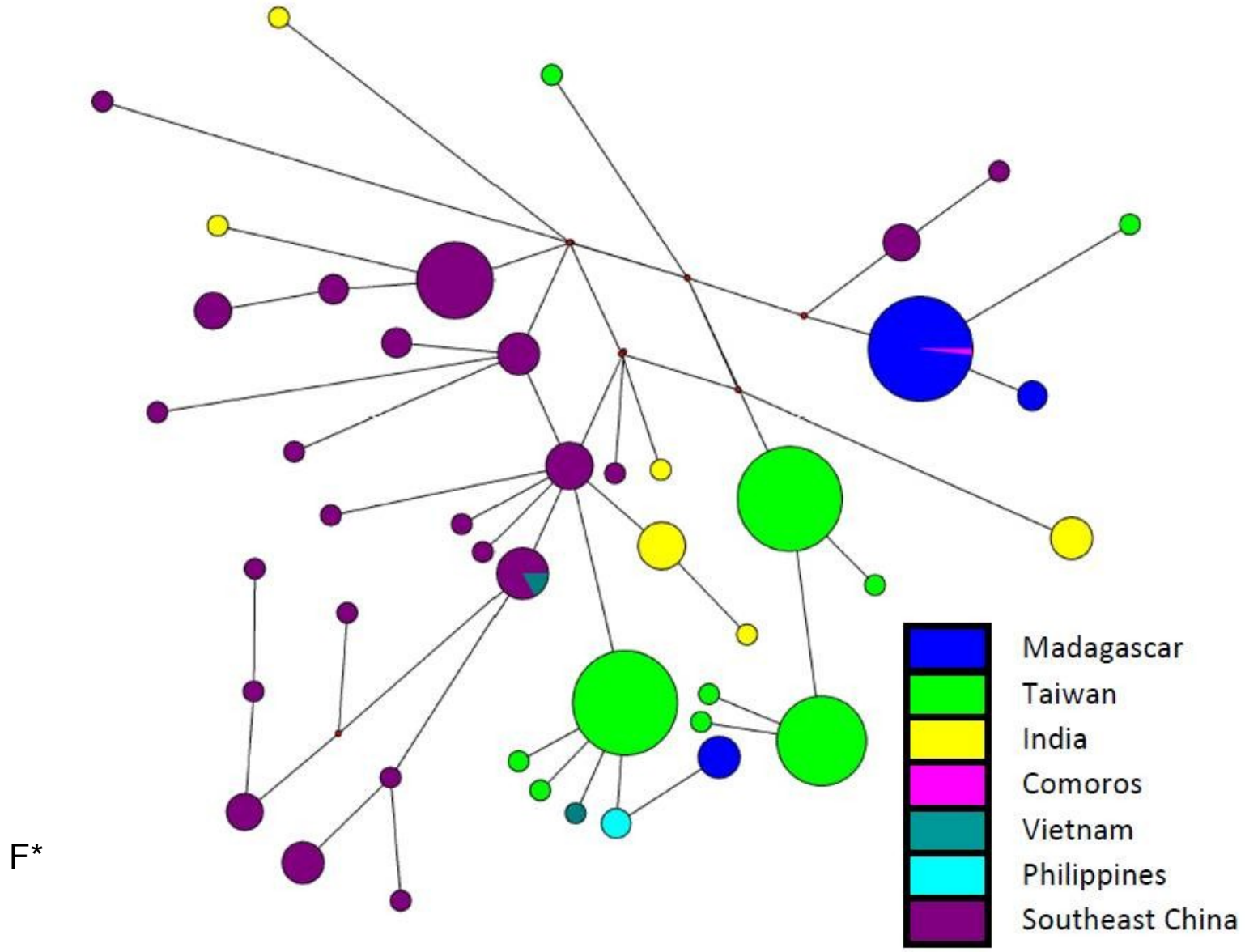


Figure 18c: *Median joining network representing mtDNA haplogroup F**. There is minimal type sharing between Taiwanese sequences and Malagasy sequences, however there is a clear separation between sequences from Southeast China and Taiwan, with the Malagasy showing a higher affinity to the Taiwanese

In order to determine relations between various population groups, mtDNA haplogroup frequency data from African, Asian, Middle Eastern and South Asian populations was used to create Multidimensional Scaling plots (MDS) (Figure 19). The analysis shows a strong affinity between samples from Papua New Guinea, Hawaii, Samoa and New Zealand Maoris. East African countries Mozambique, Kenya, Ethiopia and Tanzania show a higher affinity towards Madagascar (MAD_PRESENT) than other African countries within this study. From the Asian groups used for this MDS plot, Madagascar shows the greatest affinity to Sulawesi and the Philippines. This being said, Madagascar (MAD_PRESENT) is situated between Asian, African and Oceanic region countries, indicating a contribution from each to the peopling of Madagascar. African, Oceanic and Asian countries are also clearly separated, whereas Indian Ocean island populations are indicated to share influences from Africa and Asia to varying degrees. Of the three other studies carried out on Malagasy populations, results from Razafindrazaka et al. are most closely correlated to that of the present study. This may be due to sampling size where the Razafindrazaka study had 266 samples, the Hurles study had 37 samples and the Tofanelli study worked with 133 samples. The Comoros has an interesting position between various Madagascar studies. The Reunion island samples clearly show a stronger affinity towards Asian populations.

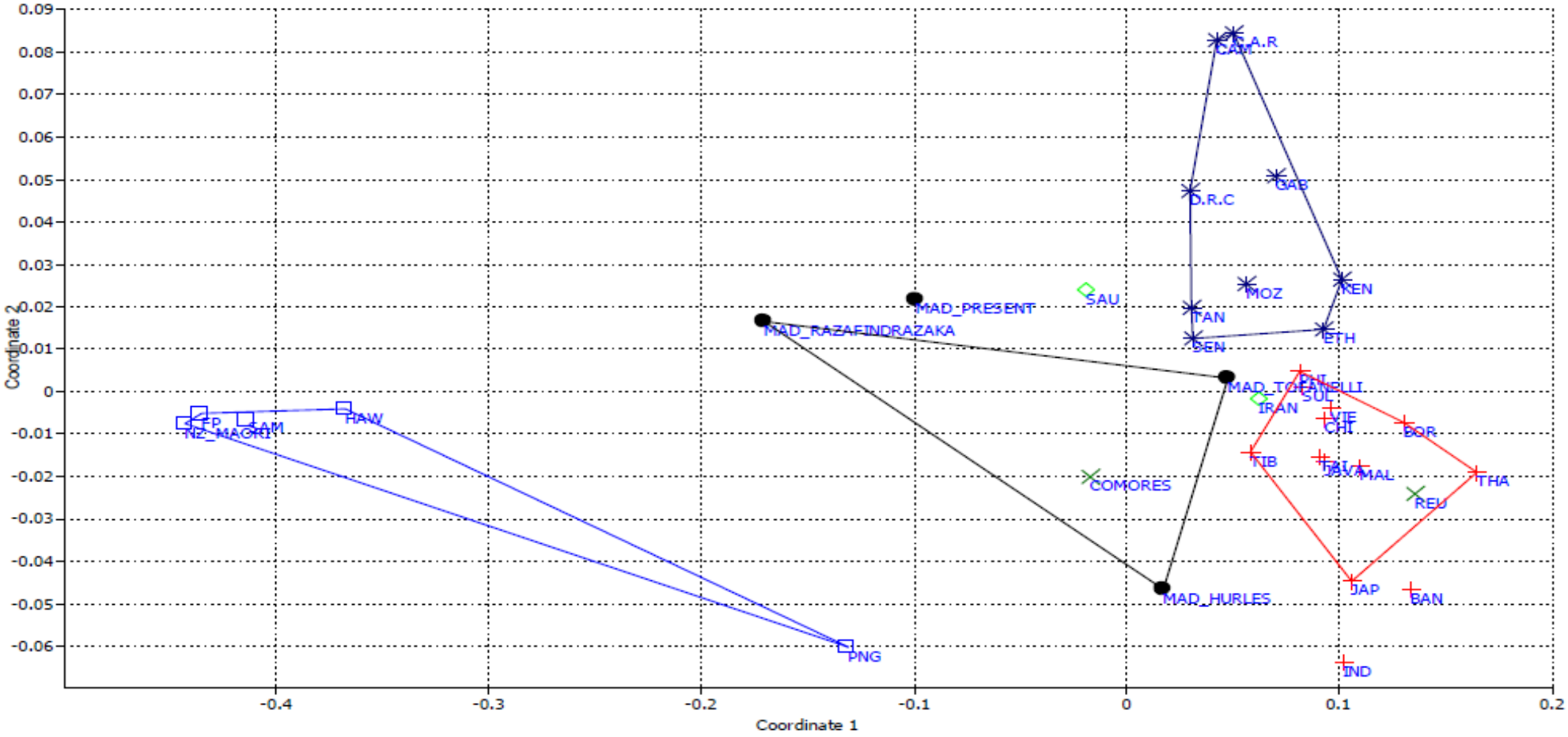


Figure 19: *MDS plot of Madagascar (MAD_PRESENT) in relation to global populations*. East African countries Mozambique, Kenya, Ethiopia and Tanzania show a higher affinity towards Madagascar (MAD_PRESENT) than other African countries within this study. From the Asian groups used for this MDS plot, Madagascar shows the greatest affinity to Sulawesi and the Philippines

Table 11: Results from mtDNA AMOVA using different groupings

Test	[Grouping]	Among groups	Among Populations within groups	Within Groups
A	[ATM ATS ATD ATO ATF ATB ATK BAR BET BEZ BES MER MIK MAH MAK SIH SAK TAN TSI VEZ ZAF]	N/A	0.94	99.06
B	[MER BET BEZ SIH] [ATM ATS ATD ATO ATF ATB ATK BAR BES MIK MAH MAK SAK TAN TSI VEZ ZAF]	0.68	0.61	98.71
C	[MADAGASCAR_PRESENT STUDY] [TAN, KEN, MOZ, ETH, SEN, CAM, GAB, C.A.R, D.R.C] [PNG, NZ_MAORI, SAM, HAW, FP TIB, JAV, BOR, SUL, PHI, MAL, JAP, CHI, VIE, THA, IND, BAN COM, REU IRA, SAU]	1.79	9.26	88.95
D	[MADAGASCAR_PRESENT STUDY] [TAN, KEN, MOZ, ETH, SEN, CAM, GAB, C.A.R, D.R.C] [PNG, NZ_MAORI, SAM, HAW, FP] [TIB, JAV, BOR, SUL, PHI, MAL, JAP, CHI, VIE, THA, IND, BAN] [COM, REU] [IRA, SAU]	4.03	7.57	88.4
E	[MADAGASCAR_PRESENT STUDY, PNG, NZ_MAORI, SAM, HAW, FP] [TAN, KEN, MOZ, ETH, SEN, CAM, GAB, C.A.R, D.R.C] [TIB, JAV, BOR, SUL, PHI, MAL, JAP, CHI, VIE, THA, IND, BAN] [COM, REU] [IRA, SAU]	3.2	8.11	88.69
F	[MADAGASCAR_PRESENT STUDY, TAN, KEN, MOZ, ETH, SEN, CAM, GAB, C.A.R, D.R.C] [TIB, JAV, BOR, SUL, PHI, MAL, JAP, CHI, VIE, THA, IND, BAN] [COM, REU] [IRA, SAU] [PNG, NZ_MAORI, SAM, HAW, FP]	3.88	7.75	88.37
G	[MADAGASCAR_PRESENT STUDY, TIB, JAV, BOR, SUL, PHI, MAL, JAP, CHI, VIE, THA, IND, BAN] [TAN, KEN, MOZ, ETH, SEN, CAM, GAB, C.A.R, D.R.C] [PNG, NZ_MAORI, SAM, HAW, FP] [COM, REU] [IRA, SAU]	4.37	7.71	87.91
H	[MADAGASCAR_PRESENT STUDY, COM, REU] [TAN, KEN, MOZ, ETH, SEN, CAM, GAB, C.A.R, D.R.C] [PNG, NZ_MAORI, SAM, HAW, FP] [TIB, JAV, BOR, SUL, PHI, MAL, JAP, CHI, VIE, THA, IND, BAN] [IRA, SAU]	3.39	7.99	88.62
I	[MADAGASCAR_PRESENT STUDY, IRA, SAU] [TAN, KEN, MOZ, ETH, SEN, CAM, GAB, C.A.R, D.R.C] [PNG, NZ_MAORI, SAM, HAW, FP] [TIB, JAV, BOR, SUL, PHI, MAL, JAP, CHI, VIE, THA, IND, BAN] [COM, REU]	3.76	7.77	88.48

Inter-population genetic distances (Fst), generated using Arlequin v.3.1.1 (Excoffier et al., 2005) were used to conduct Analysis of Molecular Variance (AMOVA). Several grouping were tested in order to determine relationships between groups, between various populations and within the populations. The results of AMOVA analysis with different groupings of the highest-level group are displayed in Table 10. The first level of analysis was to determine the variation contained between individuals of all 21 Malagasy ethnic groups (Group A in Table 10). The second level contained variation between ethnic groups in different geographic regions, namely separating the Highland (Merina, Betsileo, Sihanaka and Bezanozano) and Lowland populations into two groups (Group B in Table 10). For this analysis there was 0.68% variation among groups, 0.61% variation among populations within groups and 98.71% of variation occurred within groups between individuals. The third level was to compare Madagascar to African and Non-African comparative data sources (Group C in Table 10). This showed the greatest level of variation among populations within groups (9.26%). The fourth level of analysis separately compares Madagascar to Africa, Asia, Oceania, the Middle East and Indian Ocean island populations (Group D in Table 10). The subsequent groupings each have Madagascar grouped with either African, Asian, Oceanic, Middle Eastern or Indian Ocean island populations (Groups E-I in Table 10). The highest level of variation among groups is seen in Group G, where Madagascar is paired with samples from Asia (4.37%). Within group results are consistent with what is reported by several studies including that of Li et al. (2008) whereby the highest variation is seen between individuals of the same population rather than between individuals from different populations.

3.2. Autosomal DNA results

Clustering of genotype data is an important way of understanding similarities and differences between populations. Based on genetic similarity, a clustering approach allows for grouping together of similar individuals. A summary of populations through this approach allows for

inferences about stratification and evolutionary history of the populations to be made (Shringarpure et al., 2011).

3.2.2. SNP Validation

A list of 96 AIMs was compiled from various published sources (listed in Appendix E). Comparative allele frequency data was then collated from the ALFRED database (available online at http://alfred.med.yale.edu/alfred/index.asp) to validate the definition of these SNPs into different global regions (i.e. Southeast Asia, Africa, Middle East, Europe, Oceania and South Asia). Subsequently, comparative data was collated from HapMap and Human Genome Diversity Project (HGDP) databases. Data from HapMap and HGDP datasets was grouped according to populations, which resulted in 52 populations represented in Figures 20-23. These populations represented Asia, Africa, the Americas, Europe, Middle East and Oceania.

3.2.3. STRUCTURE Analysis

The average results of the STRUCTURE runs for the 96 SNP set are shown in Figures 20-23. The iterations were done from K=2 to K=10. Ten iterations for K=2 to K=10 are shown. Certain clusters (as with K=2, K=6, K=8, K= 9 and K=10) were split, resulting in two (or three) equally likely solutions. These are represented in Figures 20-23. The first block represents K at a population level, whereas the second block represents individuals within populations at a particular value of K. Each vertical line represents an individual. Population labels from left to right read: Madagascar, Masai, Luhya, African ancestry (USA), Yoruba, Mandenka, Bantu-speakers, San, Mbuti/Biaka Pygmies, Mozabite, Japanese and Chinese, Cambodian, Yakut, Tujia, Tizu, Miaozu, Oroqen, Daur, Mongola, Hezhen, Xibo, Dai, Lahu, She, Naxi, Tu, Uygur, Hazara, Brahui, Makrani, Sindhi, Pathan, Balochi, Burusho, Kalash, Gujarati, Druze, Bedouin, Palestinian, Adeygei, Russia, European, Italian, French, Orcadian, Colombian, Surui, Maya, Karitiana, Pima, Papuan, Melanesian.

At various K values, two or more solutions are given. These solutions are all possible as Structure works by blindly assigning variation to each individual. In the case of this data, the sample size for Madagascar is large and therefore the programme assigns Madagascar to its own cluster. At K=2, solution 1 (5 iterations) clustering separated African (blue) from non-African (yellow) populations (Figure 20). At this level, some African populations (Masai, Luhya and African Americans) showed a small non-African contribution. Individuals from Madagascar show a significantly higher African than non-African contribution. Solution 2 (5 iterations) at K=2, represents a split between African, Oceanic and East Asian populations in comparison to European and South Asian populations. Madagascan samples at solution 2 of K=2, show a higher contribution of African, Oceanic and East Asian populations.

At K=3, additional clusters were resolved in both African and non-African populations. A third cluster (red) separated African populations from the remaining populations. Within Madagascar, majority of contribution is African, with the second highest contributor being Asian populations and the least contributions coming from European and Middle Eastern populations.

At K=4, American populations and Papuan samples separate into its own cluster (green). Interestingly, Melanesian samples do not separate out with the Papuan samples, yet remains clustered with East Asian populations. At this level, Madagascar still shows greatest contributions from Africa, followed by contributions from East Asian populations, then American populations and lastly, least amount of contributions from South Asia, Europe and the Middle East (yellow)

K=5 represents a split between South Asian and Middle Eastern populations by a new cluster (purple). Madagascan samples are now shown to have contributions from mostly African (red) sources followed by East Asian (blue), American and Papuan (green), South Asian (yellow) then European and Middle Eastern populations (purple)

From solution 1 (4 iterations) at K=6, Middle Eastern populations (light green) separate from European populations (purple). At this level, Madagascar is comprised majority contribution from Africa, followed by East Asian populations, Americas and Papuan contributions, South Asian, European and Middle Eastern populations. Solution 2 (4 iterations) at K=6 shows the appearance of a large light green cluster. This suggests that Madagascar is now starting to form its own cluster. This is represented further at K=7, where the light green cluster is replaced by a lilac cluster, which indicates that Madagascar has now separated. The lilac component is represented in other African populations, suggesting Africa as its source of origin, however, due to the sample size of Madagascar in comparison to that of the other African populations, the programme STRUCTURE, assigns the largest component to Madagascar.

From K=8 to K=10 (2 iterations each), this pattern continues, where Madagascar is further divided into various contributions from parental populations. At K=8, one of the Oceanic populations, Papuan separates from other American populations into its own cluster (brown). Interestingly, the remaining Oceanic population comprised samples from Melanesia bears a strong affinity to East Asian populations. Various solutions at K=9 offer equally likely scenarios to the parental population contributions of Madagascar. At each level from K=8 to K=10, Madagascan samples show the most similarity to the Masai people of East Africa.

To overcome the randomness in which STRUCTURE assigns clusters, a supervised STRUCTURE run was carried out, wherein the programme is told to fix ancestral clusters for Africa, Asia, Middle East, Oceania, Europe and the Americas (populations from North, Central and South America) and then to assign each of these fixed clusters to Madagascar with relevant proportions. Figure 23 represents a supervised STRUCTURE run with the top box representing the results thereof at a population level (the average of all individuals). The second box represents the results of each individual. Clusters were fixed at six ancestral populations: Africa (red). Asia (blue), Middle East (purple), Oceania (yellow), Europe (green)

and the Americas (light green). The highest contributor to Madagascar is Africa (67.5%), followed by Asia (11.8%), the Americas (7.1%), Oceania (6.2%), the Middle East (3.9%) and the smallest contribution from European parental populations (3.5%) (Figure 23). It is clear from the individual level analysis that the vast majority of people from Madagascar within this sample have a significant African contribution; however there are many individuals that have a significantly higher Asian contribution.

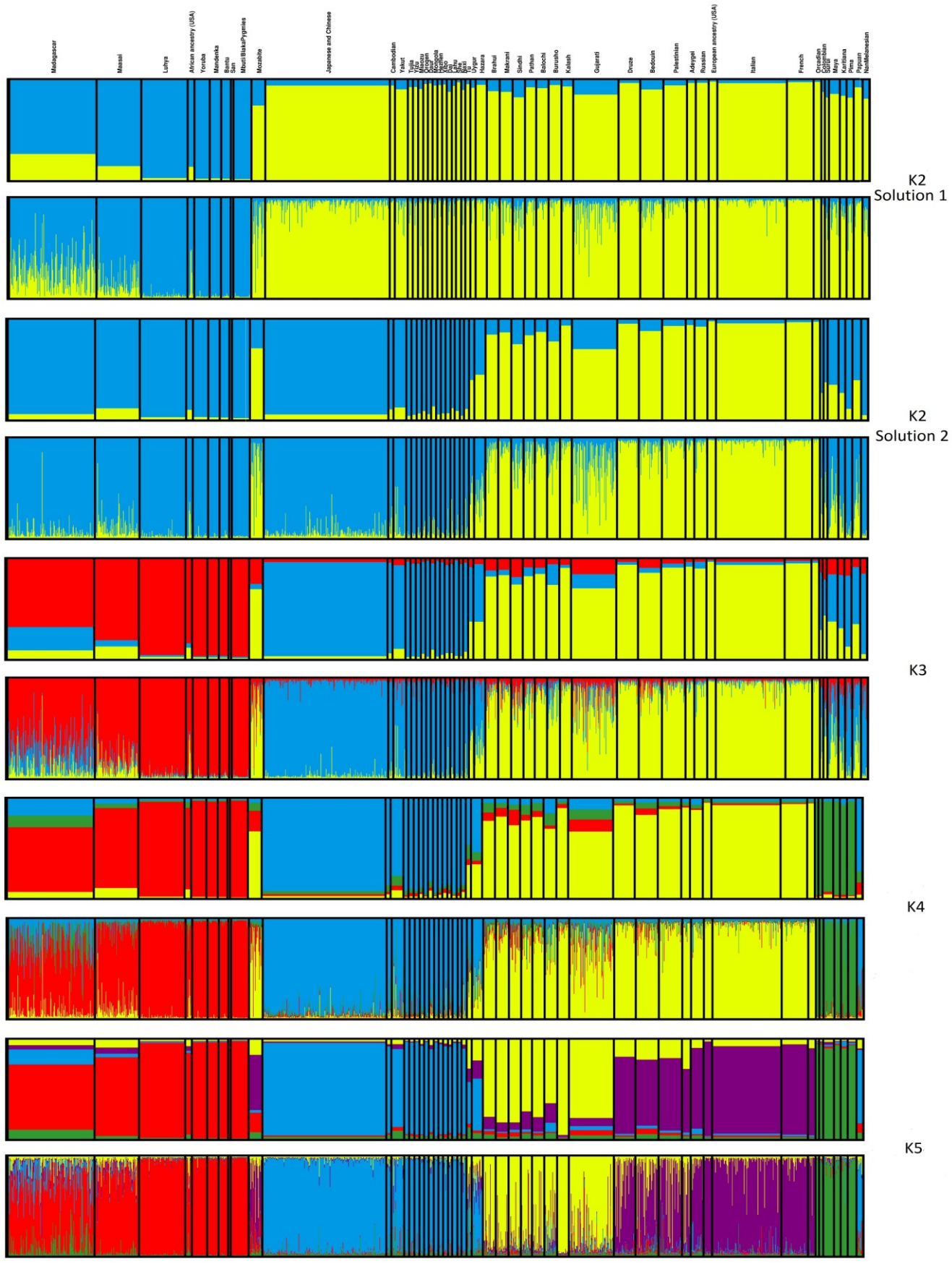


Figure 20: STRUCTURE results for K=2 to K=5 and various solutions at levels of K. The top box represents results at a population level for a particular K value, whereas the bottom box represents individuals within those populations at a particular K value. Each vertical line represents a person. Colours are assigned through the program CLUMPP.

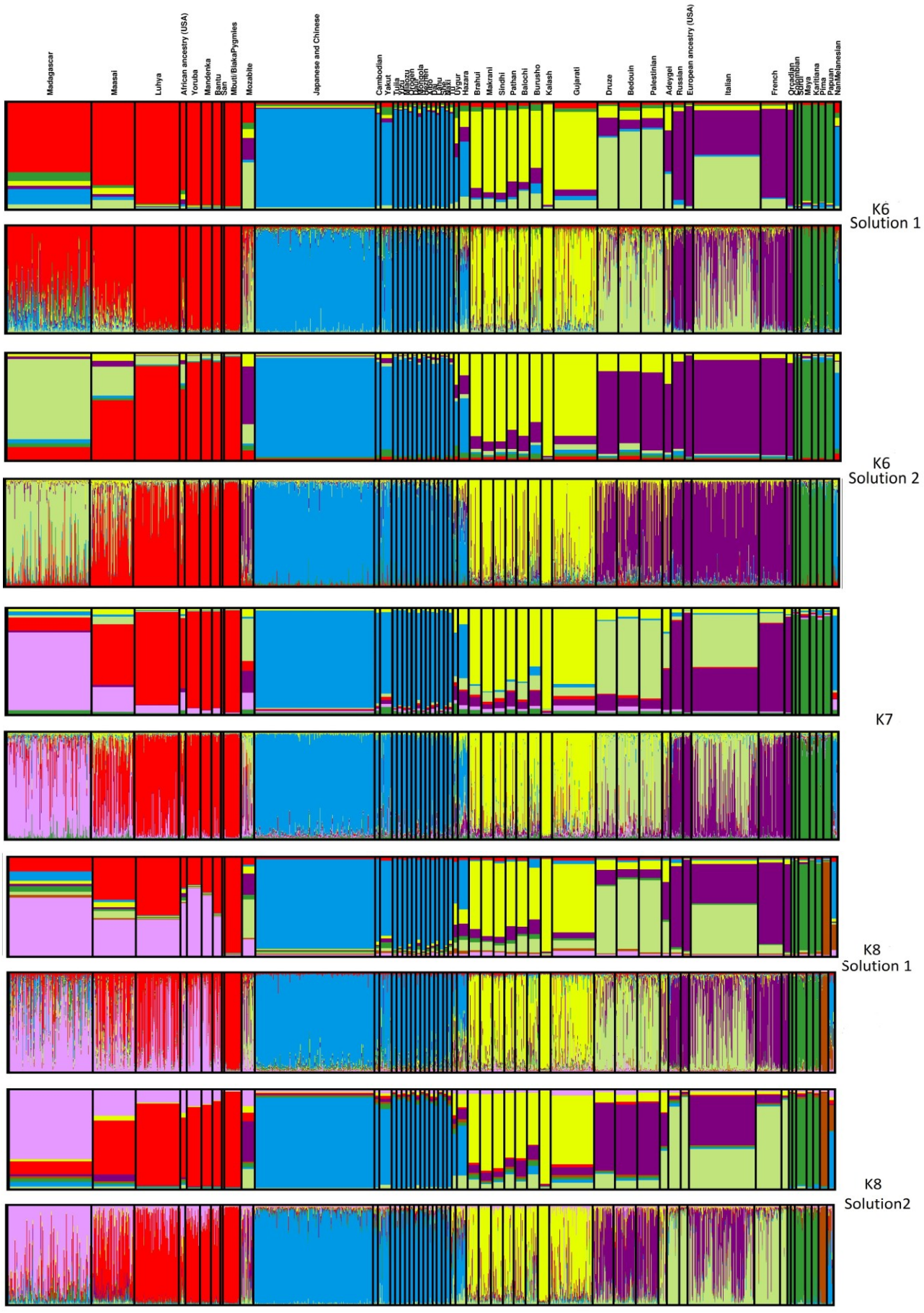


Figure 21 : *STRUCTURE results for K=6 to K=8 and various solutions at levels of K*. The top box represents results at a population level for a particular K value, whereas the bottom box represents individuals within those populations at a particular K value. Each vertical line represents a person. Colours are assigned through the program CLUMPP.

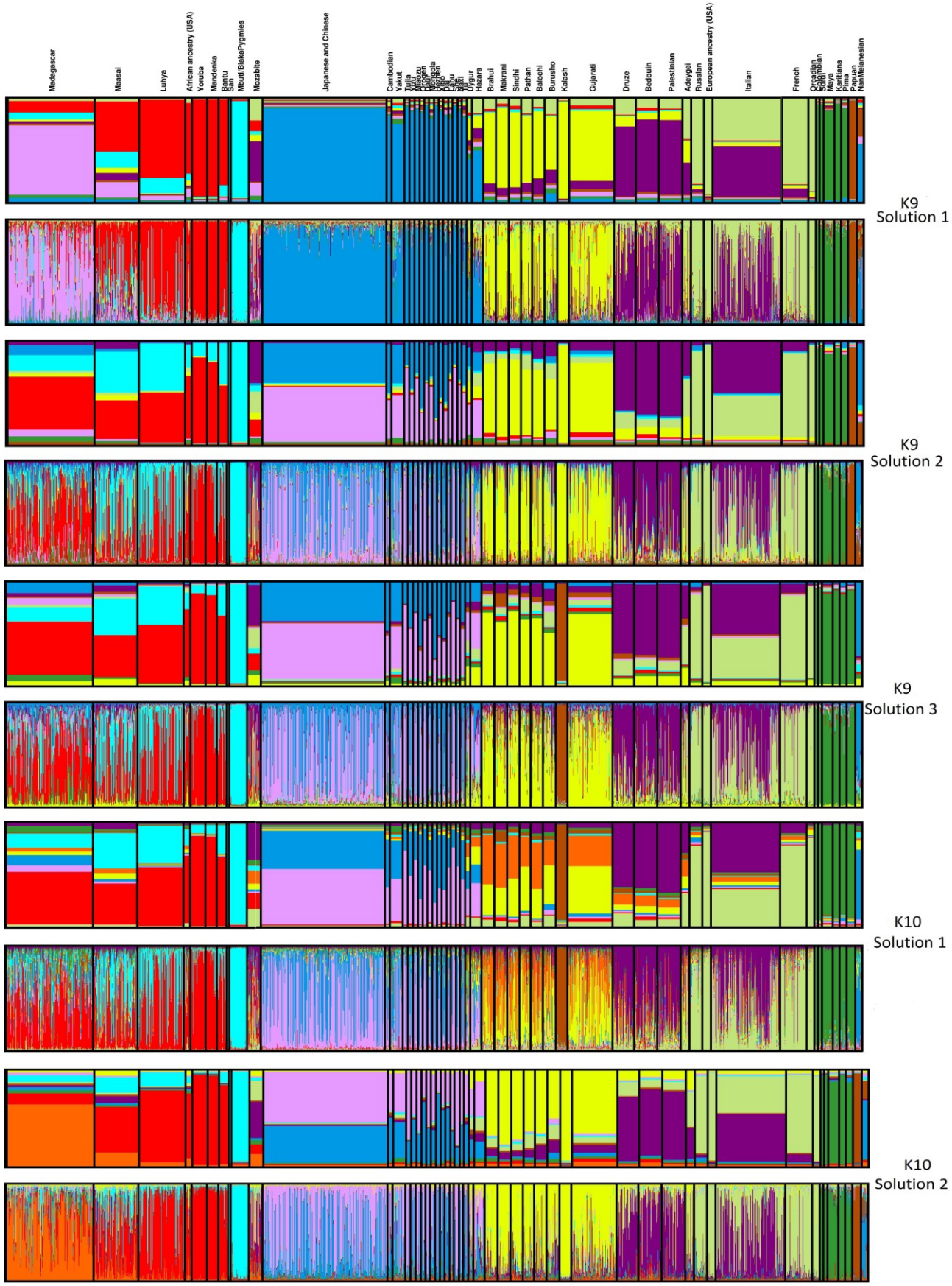


Figure 22: *STRUCTURE results for K=9 to K=10 and various solutions at levels of K*. The top box represents results at a population level for a particular K value, whereas the bottom box represents individuals within those populations at a particular K value. Each vertical line represents a person. Colours are assigned through the program CLUMPP.

.

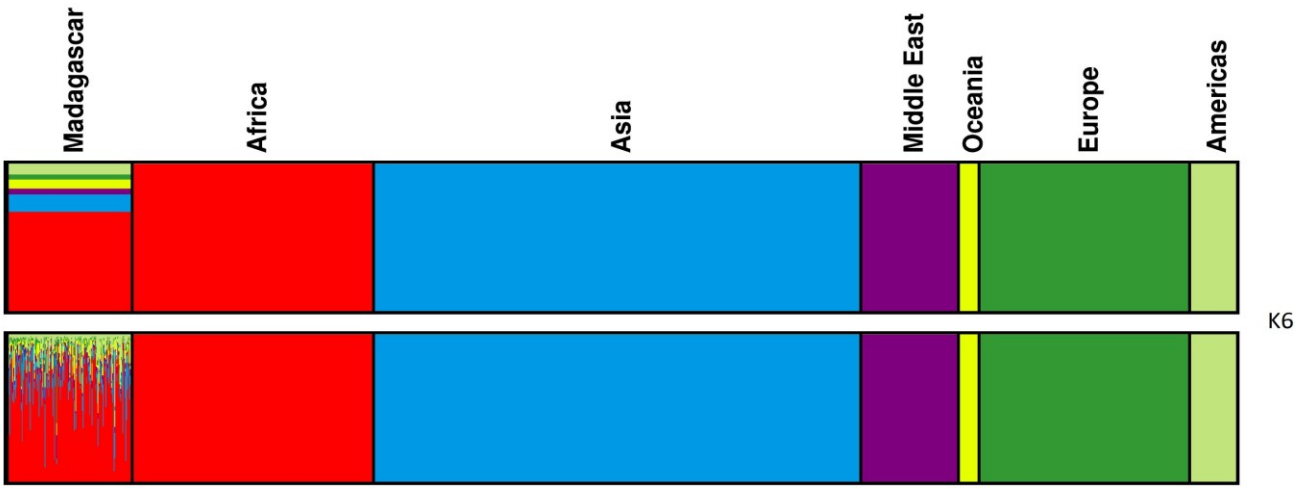


Figure 23: *Supervised STRUCTURE results for K=6*. Each population, except Madagascar is fixed into an ancestral cluster. The top box represents results at a population level for a particular K value, whereas the bottom box represents individuals within those populations at a particular K value. Each vertical line represents a person. Colours are assigned through the program CLUMPP.

The same 96 SNP dataset used in the STRUCTURE analysis was also used in distance based analysis. A distance matrix generated through the GENEPOP software was used to create an unrooted, neighbour joining tree. Figure 24 shows a clear separation of African and non-African populations. The influence of the African component in the Malagasy placed Madagascar closer in position to the African groups. As with the STRUCTURE results (Figures 21-23), Madagascar grouped closely with the Masai sample. Each of the other populations, except Papuan and Melanesian individuals, grouped together with populations from the same continent. This reiterates the results from K=8 to K=10 shown in Figures 21 and 22.

Phylogenetic trees based on distance matrices are subject to influence of admixture within groups. Data represented on these trees, from distance matrices, are optimised in a single dimension. To overcome this, techniques such as principal component analysis (PCA) is able to optimally use the distance matrix in more than a single dimension. The GENEPOP distance matrix shows that the Malagasy population is substantially closer to African populations than Asian and Austronesian populations. Results of the PCA are shown in Figure 25. The PCA reiterates the results from both STRUCTURE and phylogenetic tree analyses, where all populations from the same continent group together. In addition, Papuan and Melanisan samples separate. Although it shares an affinity with Asian populations, Madagascar shows a higher affinity with African populations. Of the African populations affiliated with Madagascar, the Masai, Luhya, Bantu-speakers, Yoruba, Africa-American and Mandenka populations group closely with Madagascar. The San, Mbuti/Biaka Pygmies and the Mozabites are not as strongly affiliated with the Malagasy (Figure 25).

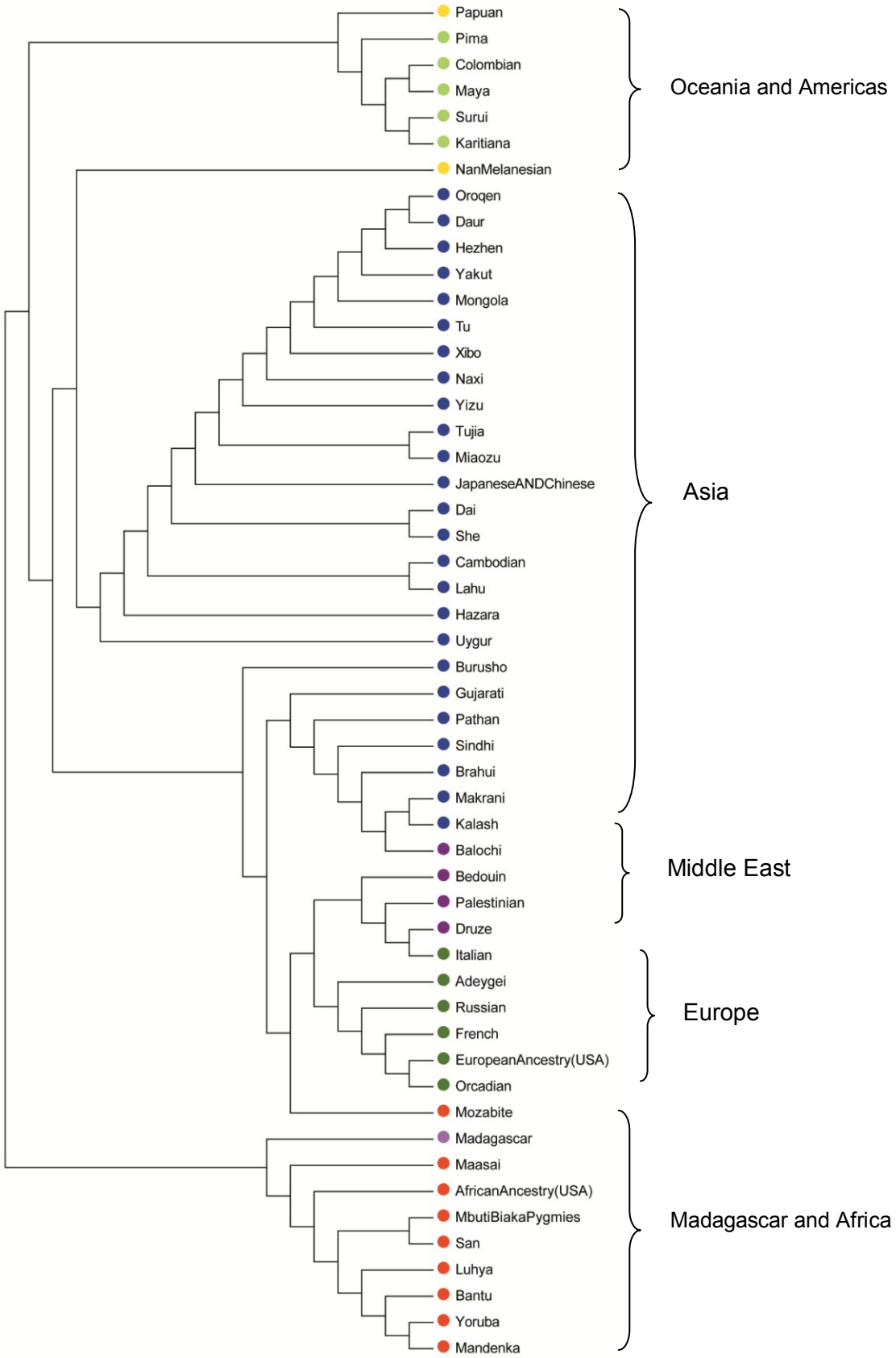


Figure 24: *Neighbour joining tree created using 96 AIM autosomal SNP data*. There is a clear separation of African and non-African populations with Madagascar grouping closely with African populations.



Figure 25: *Principal component analysis of 96 AIMs*. Although it shares an affinity with Asian populations, Madagascar shows a higher affinity with African populations

STRUCTURE results revealed various contributions from parental populations to the Malagasy group in question. Results supported high levels of contribution from African populations and lower contributions from non-African sources. At most K values, the Malagasy sample set closely resembled the Masai sample. The short-comings of the programme STRUCTURE were overcome by doing a supervised structure run, where each parental population showed their relative contributions to the Malagasy. As was consistent with Y chromosome data from previous studies (Morar, 2000; Hurles et al., 2005; Tofanelli et al., 2009; Cox et al., 2012), majority of Malagasy samples showed an African contribution when analysing AIMs. Malagasy mtDNA data however shows a higher Asian than African contribution (Hurles et al., 2005; Dubut et al., 2009; Tofanelli et al., 2009; Cox et al., 2012). Furthermore, results from this study emphasise the heterogeneity of the Malagasy and confirm results of similar studies (Poetsch et al., 2013). The Malagasy example would be key in highlighting the importance of sex-specific processes in human migrations.

Previous studies have shown that geographical distance correlates with genetic distances between populations (Novembre et al., 2008; Ramachandran et al., 2005). This will imply that the amount of admixture and sharing of ancestral population components between the population in question will depend on geographical distance. The analysis of this pattern of sharing between populations has provided important insights into human colonization history including multiple migration waves.

Chapter 4

Discussion

Over the past few years, due to the undertakings of the 1000 Genome Project and other such endeavours, there has been a dramatic increase in the discovery of loci that are polymorphic within both coding and non-coding regions of the human genome. As a result there has been a steady increase in new methods that have been developed to study these loci each of which has its own advantages and disadvantages. To date, various markers have been used to successfully examine the relationships between populations. This study makes use of the properties of some of these markers to uncover the prehistory of the Malagasy.

As part of this study some of the genetic markers mentioned above, were exploited to gain further insight into the peopling of Madagascar, specifically to evaluate the origin of the intergenic COII/tRNA^{LYS} 9-bp deletion in the Malagasy, assess the mtDNA diversity in the different ethnic groups across different geographic regions and to examine how males and females have contributed in shaping the gene pool of the Malagasy. This was achieved by comparing haploid and diploid data from African and non-African sources to evaluate their relative contributions. Most potential parent populations were chosen because they were sampled from geographic areas of Asia and Africa that were most likely to have contributed to the gene pool of the Malagasy. In other instances, some populations were used because the information on these populations was publically available. There have been several studies in the field of archaeology, linguistics, anthropology, history and even genetics that have addressed questions relating to the origins of the Malagasy. To date, many studies later, there remain many unanswered questions. It is fair to say that as yet, no consensus has been reached about the order, time and the amount of people by which the island was first colonized.

There are four hypotheses regarding the settlement of the island: 1) Madagascar already had an indigenous population, 2) Madagascar was initially settled on by Africans 3) Madagascar was colonized by an admixed Indonesian and African population. The admixture was thought to have occurred between Asian traders and natives off the East coast of Africa and 4) Madagascar was populated independently by Africans and Asians.

Only hypotheses 3 and 4 are widely debated with 1 and 2 having been discarded. With respect to hypothesis 3, there are many cultural aspects of modern-day Madagascar that have been derived from both African and Asian cultures. Historical evidence suggests trade activities between Indonesians and Africans between the first and fifth centuries A.D. (Bellwood, 2007). The incorporation of dry rice cultivation, pigs, coconuts and banana crops could have been as a result of trade and not necessarily that the Indonesians themselves settled in Africa. In this instance, Y chromosome markers would prove to be a useful tool in determining the Indonesian contributions to the African gene pool as it would have been most likely that the Asian traders would have taken on African female partners. This would then suggest that the proto-Malagasy would have accumulated African genetic signatures, however since present-day Malagasy languages and way of life are reflective of Asian influences, it suggests that the time spent by Indonesians in Africa did not have a substantial effect.

If as hypothesis 4 suggests, that Madagascar was colonized independently, admixture and the close genetic affinity of ethnic groups on the island could be attributed to interactions between groups and migration of these groups to different regions on the island. Furthermore, traditional Malagasy narratives, known as *ranto*, tell of single or multiple waves of migration from a single parental population (Cox et al., 2012). There is, to date no conclusive evidence that Madagascar was settled in multiple waves of migration from Indonesia, however the mention of *ranto* in Malagasy narratives suggest a possible hypothesis to further investigate.

In addition to the debate surrounding the origins of the parental populations of Madagascar, the route taken by the original settlers is poorly understood and complexities arise when asking whether these migrations were a direct route from across the Indian Ocean to the eastern coast of Madagascar (perhaps as an unintended voyage across the Indian Ocean) or whether it was an indirect route from across the Indian Ocean to the east coast of Africa with a final voyage to Madagascar. The direct transoceanic route has been reviewed based on simulations carried out using monsoon weather patterns, vessel type and ocean currents. From these simulations, the authors of those studies found that the Maldives played an important role in allowing vessels from India, Sri Lanka, Sumatra and Indonesia to voyage across the ocean to Madagascar. Sailors would have resupplied their vessels and waited for favourable weather conditions on the Maldivian archipelago before continuing on. The authors also suggest December through to March as the most favourable times to travel. The simulations were also able to successfully show the interconnectedness of sea routes in the Indian Ocean (Fitzpatrick and Callaghan, 2008). Further evidence of transoceanic voyage from Asia to have reached Madagascar include: wreckage from ships that were destroyed in Java and Sumatra being washed up onto Madagascan shores and sixteenth century shipwrecked Portuguese sailors who reached the shores of Madagascar and whose descendents were found 100 years later by the French in Madagascar, indistinguishable from the other inhabitants except for certain physical traits (Brown, 1978).

The main focus of this study was to determine various contributions using maternally inherited mtDNA and autosomal DNA SNPs to ultimately trace the origins of the parental populations of Madagascar. Historic, cultural and linguistic data are in favour of a bi-parental origin where Asians arrived first followed by an African migration to Madagascar. Hewitt et al., 1994 have suggested that the African contribution would have been from Bantu-speakers from east and central Africa, whereas Campbell (1989) has suggested Mozambique and Malawi.

Assuming hypothesis 4), where Asian population(s) were thought to have first colonized Madagascar followed by Africans, the African mtDNA genetic signatures could have been introduced in one of two possible ways. The first is by means of voluntary migrations of peoples from the eastern shores of Africa to Madagascar give the relatively close proximity of the two land masses. Secondly, Madagascar played a role in the exportation of approximately 72000 slaves to the Mascarenes between 1610 to 1810 (Campbell, 1988). Approximately 5000 slaves were also exported to French plantations from Madagascar. In addition, slaves were being imported to the west coast of Madagascar from Africa to serve as labourers and servants (Campbell, 1989). Many slaves were bought by the ruling dynasty of the Merina who were involved in commercial rice agriculture. By the end of the slave trade, slaves that were freed became incorporated into the Malagasy population. They adopted the ways and culture of the Malagasy. It is estimated that 1 million foreign slaves and 300 000 Malagasy were enslaved during the 19th century.

In accordance with Campbell 1996, it was not uncommon for women to accompany South Sea islander men on transoceanic voyages. In a recent study by Cox et al., 2012, the authors suggest that Madagascar was initially settled on 1200 years ago by a group of thirty women. The study further suggests that ninety three percent of these women were of Indonesian descent. Cox suggests albeit the founding number being small, it has been shown to occur in New Zealand, where the islands were initially settled on by a group of seventy women (Penny, Murray-McIntosh and Harrison, 2002).

Through the judicious use of mtDNA markers such as the intergenic COII/tRNA^{Lys} 9-bp deletion, we are able to trace the likely origins of the marker thereby inferring the possible origins of parental populations.

Sampling of 981 individuals across all ethnic groups of Madagascar has revealed: a) that the intergenic COII/tRNA^{Lys} 9-bp deletion occurs at a frequency of 22.73% on the island; b)

further analysis on samples containing the Polynesian motif of two coding region SNPs (1473C-T and 3423T-A) revealed a new motif called the “Malagasy motif”; c) SNP variation traced the origins of the 9-bp deletion to African (17.04%) and Polynesian (21.52%) and Malagasy (61.44%) sources and d) the ‘Malagasy motif ’ motif is distributed across all Malagasy ethnic groups.

Apart from Madagascar, the Polynesian motif occurs in remote Oceania with low incidence in Melanesia and eastern Indonesia and only sporadic occurrences in Bali and Borneo. It is not found in Indonesians in the Barito River area (Soodyall et al., 1996) and is suggestive of a lack in concordance between genetic and linguistic data regarding the parental origins of the Malagasy. There are three possible explanations for the presence of Polynesian motif in Madagascar; a) It arose independently, b) it was introduced into the Malagasy directly from Polynesia and c) it was introduced through the founding populations of Polynesia and Madagascar. The Polynesian motif is seen at its highest frequencies from Taiwan to the Philippines and Indonesia (Melton et. al., 1995). This corroborates a Taiwanese origin for the proto-Polynesian expansion which spread throughout Oceania by way of Indonesia (Melton et. al., 1995). Furthermore, the absence of the Polynesian motif in India suggests a direct route from the South Pacific as opposed to an indirect route via India. The presence of the Polynesian motif firmly establishes a Southeast Asian connection to Madagascar.

With respect to the Malagasy motif, it cannot be assigned to having originated within a certain region of Madagascar. However, due to the absence of the motif outside of Madagascar, it can be assumed that it originated within Madagascar after the arrival of the Polynesian motif. This hypothesis is however also unlikely for three reasons a) two coding region mutations (nucleotides 1473 and 3423) would have had to appear within 1500-2000 years, b) the motif would have had to spread throughout the island to all ethnic groups and c) the Polynesian motif, which is the immediate precursor to the Malagasy motif would have had to vanish from the island. It is a possibility that the Malagasy motif can still be found in a

few small Indonesian communities (on the relatively understudied islands of Borneo and Sulawesi). Wider sampling within these regions will eventually answer this question. As an alternative, the Malagasy motif may not be present in Indonesia at all. It may have arisen among the earliest Indonesian colonists to Madagascar and subsequently reached high frequencies (due to genetic drift) either in Madagascar or somewhere en route to Madagascar (stopping points along the east African coast) Both these scenarios suggest a substantial Indonesian founder event during the foremost settlement period of Madagascar.

From the above data it is not possible to conclusively establish how the different forms of the motif reached or originated in Madagascar.

Analysis of maternal lineages resulted in haplogroups found in previous studies on Madagascar. A breakdown of these mtDNA haplogroups revealed a 58.51% Asian, 0.20% Eurasian and 41.28% African contribution respectively (Figure 7). Of the 981 samples, 57 haplogroups and their sub-haplogroups were identified.

Historical records indicate different origins of various Malagasy ethnic groups (Buettner-Janusch and Buettner-Janusch, 1904; Hewitt et al., 1998). When considering haplogroup distribution by region, each of the regions, except the Northern region, had relatively equal Asian and African haplogroup contributions. The Northern region had a markedly higher African haplogroup input. Figure 8 comprehensively illustrates these contributions. Regional differences are further illustrated by haplogroup frequency maps (Figures 10 and 11). L0a2 is found at significant levels within the northern regions of Madagascar, where the Tsimihety, Sakalava, Antankarana and Betsimisaraka groups reside (Figure 2). These groups form part of the Lowland group that have a noted African ancestry. L2a1a shows high concentrations in the Bara, Sakalava, Tshimihety and Betsimisaraka ethnicities. L2a1b shows the highest concentration amongst the Lowland Mahafaly ethnic group. L3b1a shows the highest concentration amongst the Lowland Bara and Betsimisaraka groups. For the Asian haplogroups, B4a1a1a has the highest concentration amongst the Betsileo, whereas M7c3c is widely distributed amongst Merina, Sakalava, Antaisaka and Antanosy. Haplogroups M23

and M32c show similar patterning, where both haplogroups are highly concentrated within the central part of Madagascar most notably inhabited by the Merina people (Figure 2). Of the identified haplogroups, M23, which is deemed to be a rare subhaplogroup, was found in 14 of the 981 (6.52%) subjects. These results corroborate findings of a study showing the presence of M23 at a frequency of 1.88% (5/266) (Ricaut et al., 2009). Thus far, M23 has not been detected outside of Madagascar. The authors suggest South East Asian origin for this haplogroup, based on the current distribution of macrohaplogroup M. The current study has an increased sample size and a sample representative of a more diverse Malagasy ethinic group, therefore a representation of 6.52% may be more accurate than that of 1.88% presented by Ricaut et al., 2009.

The bar chart (Figure 13) displays a relatively equal contribution from African and Asian sources towards the mtDNA gene pool in all ethnic groups except the Zafisoro, Antambahoaka and Sakalava groups. This could be due to sample size bias. All ethnic groups have a representation of these African lineages in varying degrees. A higher African contribution is seen within the Mahafaly, Antanosy, Antandroy, Antaisaka, Antaifasy, Antaimoro, Antambahoaka, Antankarana, Bara, Betsimisaraka, Mahafaly, Vezo, Mikea, Sakalava and Tsimihety. This finding supports other data types from previous studies. Reports from Buettner-Janusch and Buettner-Janusch (1904) in addition to a 1998 study by Hewitt et.al, a higher proportion of African derived genes was found in Lowland rather than Highland ethnic groups. Migration, endogamy and assortative pairing would have all played a role in shaping the genetic structure of various ethnic groups**.** Many of the Lowland groups (Vezo to Tshimihety) that are believed to have African origins group closely with the comparative data from Kenya and Mozambique (Figure 14). The remaining Lowland groups that are also thought to have African origins show less affinity to Kenya and Mozambique. The Highland Groups, Sihanaka, Merina, Tanala and Bezanozano show a close affinity to

Vietnam as do the Betsileo with China. This is consistent with phenotypic factors suggesting that these groups are more Asian in appearance.

Figure 19 represents Madagascar in relation to global populations. Madagascar samples from the present study showed greatest affinity to samples from Papua New Guinea, Hawaii, Samoa and New Zealand Maoris. East African countries which include Mozambique, Kenya, Ethiopia and Tanzania show a higher correlation to Madagascar (Present study) than other African countries, whereas populations from the islands of Sulawesi and Philippines come up as the closest Asian relative to Madagascar (Present study) (Figure 19). The Comoros has an interesting position between various Madagascar studies. This could be explained by the small sample size and due to the investigators use of only hypervariable region I sequence data. Comoros also has a very closely shared history to that of Madagascar, and this may explain its close affinity to samples from other studies relating to Madagascar. The Reunion island samples clearly show a stronger affinity towards Asian populations. This may be explained by differing settlement histories between Madagascar and the Reunion island. Through the use of MDS plots, it is clear that the Malagasy people are the product of a culmination of Asian, Oceanic and African origins.

For the AMOVA analyses, several groupings were tested in order to determine relationships between groups, between various populations and within the populations. To test for variation between Highland and Lowland groups, ethnic groups from these regions were separated (Group B in Table 10). For this analysis there was 0.68% variation among groups, 0.61% variation among populations within groups and 98.71% of variation occurred within groups between individuals. The 0.68% shows that there is almost no difference between Highland and Lowland groups based on mtDNA data. The 0.61% also shows that there is little variation within the Highland and Lowland ethnicities. These results are consistent with other forms of analyses carried out in this study, based on mtDNA data. When Madagascar is compared to other groups of populations (African, Asian, Oceanic, Indian Ocean island

populations and Middle Eastern) as in Group D (Table 10), the amount of variation can be seen as the "baseline" where all further groupings can be compared to that of grouping D. From here, Madagascar was paired up with populations from each of the different regions: Africa (Group F), Asia (Group G), Indian Ocean island populations (Group H) and Middle East (Group I). This was done to determine which of the populations, when associated with Madagascar would show the greatest similarity to Madagascar in order to suggest possible parental populations of the Malagasy. When Group D is compared to Groups E-I, the least amount of variation is displayed in Group E, where Madagascar is paired with Oceanic populations (3.20%). The highest variation is displayed when Madagascar is grouped with Asian populations (4.37%) and it is interesting to note that Madagascar does not show the highest variation when grouped with Indian island populations (Group H), this may be due to a shared genetic history between the island populations (Msaidie et al., 2010). Groupings such as in Group G are optimal in comparison to Madagascar being in its own group as in Groups A-D. In this way, Madagascar shows most variation when grouped with Asian populations, 3.88% when grouped with African populations, 3.76% when grouped with Middle Eastern populations, 3.39% when grouped with Indian island populations and 3.2% when grouped with Oceanic populations.

With regard to AIMs, the current study provides additional validation that a SNP set of 96 autosomal markers is able to distinguish a wide variety of continental groups from each other. This is valuable information in separating a single population under investigation into potential contributing ancestral groups. Figures 21 and 22 were able to demonstrate the amount of African and non-African admixture with parental contributions from Africa, Asia, Middle East Europe, the Americas and Oceania. In addition, these analyses also showed that the Malagasy are a heterogeneous population, suggesting of any African-Asian admixture was likely to have taken place at the beginning of Malagasy history.

At K=2, K=6, K=8, K=9 and K=10, there were various solutions, indicating two or three both equally likely, possible outcomes at those K values. In order to visualise the relative proportion of each parental population's contribution, a supervised STRUCTURE run was initiated (Figure 23). In doing this, it was assumed that populations from six regions (Africa, Asia, Middle East, Oceania, Europe and the Americas) contributed to the make-up of modern-day Malagasy. The results supported low levels of contribution from the Middle East and Europe, with increasing contribution from the Americas, Oceania, Asia and the highest contribution being from Africa. These results are consistent with mtDNA, Y chromosome and historical data. Furthermore, studies have shown that geographical distance correlates with genetic distances between populations (Novembre et al., 2008; Ramachandran et al., 2005). This will imply that the amount of admixture and sharing of ancestral population components between the population in question will depend on geographical distance. In this case, Madagascar is situated 400km off the Mozambican shore of Africa. It is therefore likely that Y chromosome lineages were brought from Africa to Madagascar due to its close proximity and the role that Madagascar played as a passage of slave trade (where most slaves would have been male). The Asian and European contributions to the paternal Malagasy gene pool can be attributed to the movement of traders, migrants and colonialists (Hurles et al., 2005; Kusuma et al., 2015). Asian maternal lineages are assumed to have been transmitted to Madagascar by a small cohort of approximately 30 women from Indonesia (Cox et al., 2012). The results demonstrated in this study, allowed for refining the genetic landscape of the Malagasy. In addition, an effective panel of 96 ancestry informative markers was established.

Another technique employed to determine the contributing parental populations to the Malagasy is PCA. In this study, a PCA was generated using comparative data collated from the Hapmap and HGDP (Figure 26). The results indicate that Madagascar lies between African and Asian populations but appears to have a stronger affinity to African populations.

In summary, both mtDNA and autosomal DNA data from the present study confirm that the origins of the Malagasy are complex and lie both in Africa and Asia.

CONCLUSION

It is challenging to establish the level of variation that existed within the proto-Malagasy. A large number of factors have contributed to the make-up of the present day Malagasy. The similarity of gene pools between the twenty one ethnic groups from different regions may be due to recent common ancestry. The data from this present study is unable to resolve whether the proto-Malagasy were an admixed population prior to colonization or whether this admixture occurred after colonization. It is important to recognise that this study has made use of samples sizes far larger than any other study of the Malagasy. Furthermore, these samples represent twenty one ethnic groups from both Lowland and Highland groups. The data from the current study is consistent with that of previous studies, where there is a clear contribution from both African and Asian parental sources; however it is important to note that it would be erroneous to look for only two parental sources of the Malagasy, given the recent colonization of the island. It is feasible to say that the Malagasy are a culmination of both the eastern and western gene pools that have been influenced by no less than 60 000 years of autonomous evolution. The combined contribution from archaeological, historical, linguistic and anthropological evidence will further shed light on the settlement and practices that will determine the prehistory of the Malagasy.

REFERENCES

Abu-Amero, K.K., Larruga, J.M., Cabrera, V.M., and Gonzalez, A.M. (2008). Mitochondrial DNA structure in the Arabian Peninsula. *BMC Evolutionary Biology* 8, 45.

Adelaar, K. A. (1989). Malay influence on Malagasy: linguistic and culture-historical implications. *Oceanic Linguistics*, 28(1), 1-46.

Alves-Silva, J., Guimarães, P. E. M., Rocha, J., Pena, S. D. J., and Prado, V. F. (1999). Identification in Portugal and Brazil of a mtDNA lineage containing a 9-bp Triplication of the intergenic COII/tRNA^{Lys} region. *Human Heredity,* 49(1), 56-58.

Anderson, S., Bankier, A. T., Barrell, B. G., De Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., and Sanger, F. (1981). Sequence and organization of the human mitochondrial genome. *Nature,* 290(5806), 457-465.

Andrews, R.M., Kubacka, I., Chinnery, P.F., Lightowlers, R.N., Turnbull, D.M., and Howell, N. (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics,* 23 (2), 147.

Ballinger, S. W., Schurr, T. G., Torroni, A., Gan, Y. Y., Hodge, J. A., Hassan, K., Chen, K. H., and Wallace, D. C. (1992). Southeast Asian mitochondrial DNA analysis reveals genetic continuity of ancient Mongoloid migrations. *Genetics,* 130(1), 139-152.

Bandelt, H.J., Forster, P., and Röhl, A. (1999). Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution,* 16(1), 37-48.

Beaujard, P. (2003). Les arrivées austronésiennes à Madagascar: vagues ou continuum? (Austronesian arrivals in Madagascar : vague or continuum) *Etudes Océan Indien, 35-36,* 59-147.

Behar, D. M., Villems, R., Soodyall, H., Blue-Smith, J., Pereira, L., Metspalu, E., Scozzari, R., Makkan, H., Tzur, S., and Comas, D. (2008). The dawn of human matrilineal diversity. *American Journal of Human Genetics,* 82(5), 1130-1140.

Berniell-Lee, G., Plaza, S., Bosch, E., Calafell, F., Jourdan, E., Césari, M., Lefranc, G., and Comas, D. (2008). Admixture and sexual bias in the population settlement of La Reunion Island (Indian Ocean). *American Journal of Physical Anthropology,* 136(1), 100-107.

Blench, R. (1996). The ethnographic evidence for long-distance contacts between Oceania and East Africa. In: J. Reade (Ed.).*The Indian Ocean in Antiquity,* (pp. 461-470). London/New York: Kegan Paul/British Museum.

Blench, R. (2007). New palaeo-zoogeographical evidence for the settlement of Madagascar. *Azania XLII: Journal of the British Institute in Eastern Africa,* 42, 69-82.

Bradt, H. (2011). *Madagascar, 10th: The Bradt Travel Guide*: Bradt Publications

Brakez, Z., Bosch, E., Izaabel, H., Akhayat, O., Comas, D., Bertranpetit, J., and Calafell, F. (2001). Human mitochondrial DNA sequence variation in the Moroccan population of the Souss area. *Annals of Human Biology,* 28 (3), 295-307.

Brandstätter, A., Peterson, C.T., Irwin, J.A., Mpoke, S., Koech, D.K, Parson, W., and Parsons, T.J. (2004). Mitochondrial DNA hypervariable region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. *International Journal of Legal Medicine,* 118 (5), 294-306.

Brown, M. (1978). *Madagascar rediscovered: A history from early times to independence*: London: Damien Tunnacliffe.

Buettner Janusch, J., and Buettner Janusch, V. (1964). Hemoglobins, haptoglobins, and transferrins in the peoples of Madagascar. *American Journal of Physical Anthropology,* 22(2), 163-169.

Burney, D. A. (1987). Pre-settlement vegetation changes at Lake Tritrivakely, Madagascar in JA Coetzee (Ed). *Palaeoecology of Africa,* (pp.357-381). Brookfield: A.A. Balkema

Burney, D. A., Burney, L. P., Godfrey, L. R., Jungers, W. L., Goodman, S. M., Wright, H. T., and Jull, A. J. (2004). A chronology for late prehistoric Madagascar. *Journal of Human Evolution,* 47(1-2), 25-63.

Campbell, G. (1981). Madagascar and the Slave Trade, 1810–1895. *The Journal of African History,* 22(2), 203-227.

Campbell, G. (1988). Madagascar and Mozambique in the slave trade of the western Indian ocean 1800-1861. *Slavery and Abolition,* 9(3), 166-193.

Campbell, G. (1989). The East African slave trade, 1861-1895: The" southern" complex. *The International Journal of African Historical Studies,* 22(1), 1-26.

Campbell, G. (1996). *Theories on the origins of the Malagasy: the nineteenth century British school*. Australes – Études historiques aixoises sur l'Afrique austral et l'océan Indien occidental. Paris, L'harmattan.

Cann, R. L., Brown, W. M., and Wilson, A. C. (1984). Polymorphic sites and the mechanism of evolution in human mitochondrial DNA. *Genetics,* 106(3), 479.

Cann, R. L., Stoneking, M., and Wilson, A. (1987). Mitochondrial DNA and human evolution. *Nature,* 325 (6099), 31-36

Cerný, V., Hajek, M., Cmejla, R., Bruzek, J., and Brdicka, R. (2004). mtDNA sequences of Chadic-speaking populations from northern Cameroon suggest their affinities with eastern Africa. *Annals of Human Biology,* 31(5), 554-569.

Cox, M. P. (2005). Indonesian mitochondrial DNA and its opposition to a Pleistocene era origin of proto-Polynesians in Island Southeast Asia. *Human biology*, 77(2), 179-188.

Cox, M. P., Nelson, M. G., Tumonggor, M. K., Ricaut, F. X., and Sudoyo, H. (2012). A small cohort of Island Southeast Asian women founded Madagascar. *Proceedings of the Royal Society B: Biological Sciences,* 279(1739), 2761-2768.

Dahl, O. C. (1951). *Malgache et Maanjan: une comparaison linguistique* (Madagascar and Maanjan : a linguistic comparison) : Oslo: Avhandlinger utgitt av Egede-Instituttet.

Dahl, O. C. (1977). *La subdivision de la famille Barito et la place du Malgache* (The subdivision of the Barito family and the place of the Malagasy) : Copenhagen: Acta orientalia,

Dahl, O. C. (1988). *Bantu Substratum in Malagasy* (special issue: Linguistique de Madagascar et des Comores), Etudes Océan Indien, (9), 91-132.

Dangerfield, B. (1997). *The characterization of Glucose-6-Phosphate Dehydrogenase (G6PD) variation in the Malagasy*. MSc Dissertation. Johannesburg: University of the Witwatersrand.

Deguilloux, M. F., Pemonge, M. H., Dubut, V., Hughes, S., Hänni, C., Chollet, L., Conte, E., and Murail, P. (2011). Human ancient and extant mtDNA from the Gambier Islands (French polynesia): Evidence for an early Melanesian maternal contribution and new perspectives into the settlement of easternmost Polynesia. *American Journal of Physical Anthropology,* 144(2), 248-257.

De la Vega, F.M., Lazaruk, K.D., Rhodes, M.D., and Wenz, M.H. (2005). Assessment of two flexible and compatible SNP genotyping platforms: TaqMan SNP genotyping assays and the SNPlex genotyping system. *Mutation Research,* 573(1-2), 111-135

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglia, A., Tofanelli, S., Spedini, G., and Capelli, C. (2004). Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Molecular Biology and Evolution,* 21(9), 1673-1682.

Dewar, R. E., and Wright, H. T. (1993). The culture history of Madagascar. *Journal of World Prehistory,* 7(4), 417-466.

Dewar, R. E. (1995). Of nets and trees: untangling the reticulate and dendritic in Madagascar's prehistory. *World Archaeology*, 26 (3), 301-318.

Dhandapany, P. S., Sadayappan, S., Xue, Y., Powell, G. T., Rani, D. S., Nallari, P., Rai, T. S., Khullar, M., Soares, P., and Bahl, A. (2009). A common MYBPC3 (cardiac myosin binding protein C) variant associated with cardiomyopathies in South Asia. *Nature Genetics,* 41(2), 187-191.

Dubut, V., Cartault, F., Murail, P., Payet, C., and Thionville, M. D., (2009). Complete Mitochondrial sequences for haplogroups M23 and M46: Insights into the Asian ancestry of the Malagasy population. *Human Biology,* 81(4), 495-500.

Eaaswarkhanth, M., Haque, I., Ravesh, Z., Romero, I. G., Meganathan, P. R., Dubey, B., Khan, F. A., Chaubey, G., Kivisild, T., and Tyler-Smith, C. (2009). Traces of sub-Saharan and Middle Eastern lineages in Indian Muslim populations. *European Journal of Human Genetics,* 18(3), 354-363.

Excoffier, L., Laval, G., and Schneider, S. (2005). Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evolutionary Bioinformatics online,* 1, 47.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics,* 164(4), 1567-1587.

Falush, D., Stephens, M., and Pritchard, J. K. (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular Ecology Notes,* 7(4), 574-578.

Ferrand, G. (1908). L'origine africaine des Malgaches. *Journal Asiatique, 15*, 353-500.

Fitzpatrick, S. M., and Callaghan, R. (2008). Seafaring simulations and the origin of prehistoric settlers to Madagascar. In G. Clark, F. Leach and S. O'Connor (Eds.), *Islands of Inquiry: Colonisation, seafaring and the archaeology of maritime landscapes,* (Vol. 29, pp. 47-58). Sydney: Terra Australis.

Flacourt, E. (1661). *Histoire de la Grande Isle Madagascar*. Paris: Union Coloniale.

Fluxus-engineering (2008). Fluxus Technology Ltd. 2008-2009.

Fourquet, R., Sarthou, J., Roux, J., and Aori, K. (1974). Hemoglobine S et origines du peuplement de Madagascar: Nouvelle hypothese sur son introduction en Afrique (Hemoglobin S and origins for the settlement of Madagascar: new hypothesis on its introduction to Africa). *Archives Institut Pasteur Madagascar,* 43, 185-220.

Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics,* 147(2), 915-925.

Fucharoen, G., Fucharoen, S., and Horai, S. (2001). Mitochondrial DNA polymorphisms in Thailand. *Journal of Human Genetics,* 46(3), 115-125.

Glaubitz, J. C. (2004). convert: A user-friendly program to reformat diploid genotypic data for commonly used population genetic software packages. *Molecular Ecology Notes,* 4(2), 309-310.

GoldenSoftware I (2006). Surfer Demo, Golden Software Inc. 2007-2009.

Grandidier, A. (1904). *Collection des ouvrages anciens concernant Madagascar* (Vol. 2): Comité de Madagascar.

Grandidier, A. and Grandidier, G. (1908). *Ethnographie de Madagascar* (Vol. 1): Imprimerie nationale.

Graven, L., Passarino, G., Semino, O., Boursot, P., Santachiara-Benerecetti, S., Langaney, A., and Excoffier, L. (1995). Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Molecular Biology and Evolution,* 12(2), 334-345.

Gueunier, N. J. (1988). *Dialectologie et lexicostatistique: cas du dialecte malgache de Mayotte (Comores).* Etudes Océan Indien, (9), 143-169.

Gunnarsdóttir, E. D., Li, M., Bauchet, M., Finstermeier, K., and Stoneking, M. (2011). High-throughput sequencing of complete human mtDNA genomes from the Philippines. *Genome Research,* 21(1), 1-11.

Hall, T. A. (1999). *BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT.* Paper presented at the Nucleic acids symposium series.

Hammer, Ø., Harper, D. A. T., and Ryan, P. D. (2001). PAST: paleontological statistics software package for education and data analysis. *Palaeontologia Electronica,* 4(1), 9.

Harrison, C. D., Ballard, D. J., Patel, J., Brown, E. M., Phillips, C., and Thacker, C. R. (2008). Differentiating European and South Asian individuals using SNPs and pyrosequencing technology. *Forensic Science International: Genetics Supplement Series,* 1(1), 476-478.

Hertzberg, M., Mickleson, K. N., Serjeantson, S. W., Prior, J. F., and Trent, R. J. (1989). An Asian-specific 9-bp deletion of mitochondrial DNA is frequently found in Polynesians. *American Journal of Human Genetics,* 44(4), 504.

Hewitt, R., Krause, A., Goldman, A., Campbell, G., and Jenkins, T. (1996). Beta-globin haplotype analysis suggests that a major source of Malagasy ancestry is derived from Bantu-speaking Negroids. *American Journal of Human Genetics,* 58(6), 1303.

Heyerdahl, T. (1971). Atlantic Ocean pollution and biota observed by the "Ra"expeditions. *Biological Conservation,* 3(3), 164-167.

Hill, C., Soares, P., Mormina, M., Macaulay, V., Meehan, W., Blackburn, J., Clarke, D., Raja, J. M., Ismail, P., and Bulbeck, D. (2006). Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Molecular Biology and Evolution,* 23(12), 2480-2491.

Hill, C., Soares, P., Mormina, M., Macaulay, V., Clarke, D., Blumbach, P. B., Vizuete-Forster, M., Forster, P., Bulbeck, D., and Oppenheimer, S. (2007). A mitochondrial stratigraphy for island southeast Asia. *The American Journal of Human Genetics,* 80(1), 29-43.

Horai, S., and Matsunaga, E. (1986). Mitochondrial DNA polymorphism in Japanese. *Human Genetics,* 72(2), 105-117.

Horai, S., Murayama, K., Hayasaka, K., Matsubayashi, S., Hattori, Y., Fucharoen, G., Harihara, S., Park, K. S., Omoto, K., and Pan, I. H. (1996). mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. *American Journal of Human Genetics,* 59(3), 579.

Hurles, M. E., Sykes, B. C., Jobling, M. A., and Forster, P. (2005). The dual origin of the Malagasy in Island Southeast Asia and East Africa: evidence from maternal and paternal lineages. *American Journal of Human Genetics,* 76(5), 894-901.

Irwin, J. A., Saunier, J. L., Strouss, K. M., Diegoli, T. M., Sturk, K. A., O'Callaghan, J. E., Paintner, C. D., Hohoff, C., Brinkmann, B., and Parsons, T. J. (2008). Mitochondrial control region sequences from a Vietnamese population sample. *International Journal of Legal Medicine,* 122(3), 257-259.

Jakobsson, M., and Rosenberg, N. A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics,* 23(14), 1801-1806.

Jobling, M. A., Hurles, M., and Tyler-Smith, C. (2004). *Human evolutionary genetics: origins, peoples and disease*. New York, Garland Publishing.

Karanth, K. P. (2006). Out-of-India Gondwanan origin of some tropical Asian biota. *Current Science,* 90(6), 789-792.

Kent, R. K. (1962). *From Madagascar to the Malagasy Republic*: Greenwood Press.

Kent, R. K. (1970). *Early kingdoms in Madagascar, 1500-1700*: Holt, Rinehart and Winston.

Kidd, J. R., Friedlaender, F. R., Speed, W. C., Pakstis, A. J., De La Vega, F. M., and Kidd, K. K. (2011). Analyses of a set of 128 ancestry informative single-nucleotide polymorphisms in a global set of 119 population samples. *Investigative Genetics,* 2(1), 1.

Kivisild, T., Tolk, H. V., Parik, J., Wang, Y., Papiha, S. S., Bandelt, H. J., and Villems, R. (2002). The emerging limbs and twigs of the East Asian mtDNA tree. *Molecular Biology and Evolution,* 19(10), 1737-1751.

Kivisild, T., Reidla, M., Metspalu, E., Rosa, A., Brehm, A., Pennarun, E., Parik, J.Rr., Geberhiwot, T., Usanga, E., and Villems, R. (2004). Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *The American Journal of Human Genetics,* 75(5), 752-770.

Knight, A., Underhill, P. A., Mortensen, H. M., Zhivotovsky, L. A., Lin, A. A., Henn, B. M., Louis, D., Ruhlen, M., and Mountain, J. L. (2003). African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Current Biology,* 13(6), 464-473.

Kolman, C. J., Sambuughin, N., and Bermingham, E. (1996). Mitochondrial DNA analysis of Mongolian populations and implications for the origin of New World founders. *Genetics,* 142(4), 1321-1334.

Kong, Q. P., Yao, Y. G., Liu, M., Shen, S. P., Chen, C., Zhu, C. L., Palanichamy, M. G., and Zhang, Y. P. (2003). Mitochondrial DNA sequence polymorphisms of five ethnic populations from northern China. *Human Genetics,* 113(5), 391-405.

Kong, Q.-P., Sun, C., Wang, H.-W., Zhao, M., Wang, W.-Z., Zhong, L., Hao, X.-D., Pan, H., Wang, S.-Y., and Cheng, Y.-T. (2011). Large-scale mtDNA screening reveals a surprising matrilineal complexity in East Asia and its implications to the peopling of the region. *Molecular Biology and Evolution,* 28(1), 513-522.

Kusuma, P., Cox, M.P., Pierron, D., Razafindrazaka, H., Brucato, N., Tonasso, L., Suryadi, H.L., Letellier, T., Sudoyo, H., Ricaut, F.X. (2015). Mitochondrial DNA and the Y chromosome suggest the settlement of Madagascar by Indonesian sea nomad populations. *BMC Genomics*, 16, 191

Lao, O., Duijn, K., Kersbergen, P., Knijff, P., and Kayser, M. (2006). Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *American Journal of Human Genetics,* 78(4), 680-690.

Li, H., Cai, X., Winogradâ€□Cort, E. R., Wen, B., Cheng, X., Qin, Z., Liu, W., Liu, Y., Pan, S., and Qian, J. (2007). Mitochondrial DNA diversity and population differentiation in southern East Asia. *American Journal of Physical Anthropology,* 134(4), 481-488.

Li, J. Z., Absher, D. M., Tang, H., Southwick, A. M., Casto, A. M., Ramachandran, S., Cann, H. M., Barsh, G. S., Feldman, M., and Cavalli-Sforza, L. L. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science,* 319(5866), 1100.

Lin, C. H., Yeakley, J. M., McDaniel, T. K., and Shen, R. (2009). Medium-to high-throughput SNP genotyping using VeraCode microbeads. In *DNA and RNA Profiling in Human Blood* (pp. 129-142): Springer.

Lum, J. K., and Cann, R. L. (2000). mtDNA lineage analyses: origins and migrations of Micronesians and Polynesians. *American Journal of Physical Anthropology,* 113(2), 151-168.

Mack, J. (1986). *Madagascar, island of the ancestors*: Published for the Trustees of the British Museum by British Museum Publications.

MacPhee, R. D. E., and Burney, D. A. (1991). Dating of modified femora of extinct dwarf Hippopotamus from southern Madagascar: implications for constraining human colonization and vertebrate extinction events. *Journal of Archaeological Science,* 18(6), 695-706.

Mariano, L. (1904). Relation du voyage de découverte fait à l'Ile Saint Laurent dans les années 1613-1614 (Discovery of the voyage to Ile Saint Laurent during the years 1613-1614). *Collection des ouvrages anciens concernant Madagascar, 11*, 1-64.

Melton, T., Peterson, R., Redd, A. J., Saha, N., Sofro, A. S., Martinson, J., and Stoneking, M. (1995). Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. *American Journal of Human Genetics,* 57(2), 403.

Metspalu, M., Kivisild, T., Metspalu, E., Parik, J. R., Hudjashov, G., Kaldma, K., Serk, P., Karmin, M., Behar, D. M., and Gilbert, M. T. P. (2004). Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genetics,* 5(1), 26.

Middleton,K.(Ed.).1999. *Ancestors, Power and History in Madagascar.* The Netherlands: Brill

Miller, S. A., Dykes, D. D., and Polesky, H. F. (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Research,* 16(3), 1215.

Mona S, Grunz K.E., Brauer S., Pakendorf B., Castrì L., Sudoyo H., Marzuki S., Barnes R.H., Schmidtke J., and Stoneking M. (2009). Genetic admixture history of eastern Indonesia as revealed by Y-chromosome and mitochondrial DNA analysis. *Molecular Biology and Evolution,* 26 (8), 1865-1877

Morar, B. (2000). *An exploration of the origins of the Malagasy using genetic polymorphisms.* PhD Dissertation. Johannesburg: University of the Witwatersrand.

Mourant, A., Lopel, A., and Domaniewski-Sobezak, K. (1976). *The distribution of the human blood groups, and other polymorphisms (2nd edit.)*: Oxford University Press.

Msaidie, S., Ducourneau, A., Boetsch, G., Longepied, G., Papa, K., Allibert, C., Yahaya, A. A., Chiaroni, J., and Mitchell, M. J. (2010). Genetic diversity on the Comoros Islands shows early seafaring as major determinant of human biocultural evolution in the Western Indian Ocean. *European Journal of Human Genetics,* 19(1), 89-94.

Mullis, K. B., Faloona, F. A., Scharf, S. J., Saiki, R. K., Horn, G. T., and Erlich, H. (1986). Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Paper presented at the Cold Spring Harbour Symposium Quant Biol., 51(1), 263-73.

Murdoch, G. P. (1959). *Africa: Its people and their cultural history*. USA: McGraw-Hill Bool Company Inc.

Nagnort (1550). Cited in Singer *et al*., (1957). Physical features, sickling and serology of the Malagasy of Madagascar. *American Journal of Physical Anthropology*, 15 (1), 91-124.

Nei, M. (1987). *Molecular evolutionary genetics*: Columbia University Press.

Nelson, R.M. (2006). S-Compare. 2006-2008.

Oliver, M. A., and Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System,* 4(3), 313-332.

Oota, H., Kurosaki, K., Pookajorn, S., Ishida, T., and Ueda, S. (2001). Genetic study of the Paleolithic and Neolithic Southeast Asians. *Human Biology,* 73(2), 225.

Osier, M.V., Cheung, K.H., Kidd, J.R., Pakstis, A.J., Miller, P.L., and Kidd, K.K., (2002). ALFRED: An allele frequency database for anthropology. *American Journal of Physical Anthropology,* 119(1), 77-83.

Paschou, P., Lewis, J., Javed, A., and Drineas, P. (2010). Ancestry informative markers for fine-scale individual assignment to worldwide populations. *Journal of Medical Genetics,* 47(12), 835.

Passarino, G., Semino, O., Modiano, G., and Santachiara-Benerecetti, A. S. (1993). COII/tRNA (Lys) intergenic 9-bp deletion and other mtDNA markers clearly reveal that the Tharus (southern Nepal) have Oriental affinities. *American Journal of Human Genetics,* 53(3), 609.

Payne W. J. A. and R. T Wilson. 1999. An Introduction to Animal Husbandry in the Tropics. Blackwell Science Ltd, Malden, MA, USA, Paris, London.

Peng, M.-S., He, J.-D., Liu, H.-X., and Zhang, Y.-P. (2011). Tracing the legacy of the early Hainan Islanders-a perspective from mitochondrial DNA. *BMC Evolutionary Biology,* 11(1), 46.

Penny, D., Murray-McIntosh, R. and Harrison, G. L. (2002). Estimating the number of females in the founding population of New Zealand: analysis of mtDNA variation. *Journal of Polynesian Society*, 111 (3), 207–221.

Pereira, L., Macaulay, V., Torroni, A., Scozzari, R., Prata, M. J., and Amorim, A. (2001). Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Annals of Human Genetics,* 65(5), 439-458.

Pierson, M. J., Martinez-Arias, R., Holland, B. R., Gemmell, N. J., Hurles, M. E., and Penny, D. (2006). Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. *Molecular Biology and Evolution,* 23(10), 1966.

Pilkington, M. M., Wilder, J. A., Mendez, F. L., Cox, M. P., Woerner, A., Angui, T., Kingan, S., Mobasher, Z., Batini, C., and Destro-Bisol, G. (2008). Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa. *Molecular Biology and Evolution,* 25(3), 517-525.

Polzin, T., and Daneshmand, S. V. (2003). On Steiner trees and minimum spanning trees in hypergraphs. *Operations Research Letters,* 31(1), 12-20.

Qin, Z., Yang, Y., Kang, L., Yan, S., Cho, K., Cai, X., Lu, Y., Zheng, H., Zhu, D., and Fei, D. (2010). A mitochondrial revelation of early human migrations to the Tibetan Plateau before and after the last glacial maximum. *American journal of Physical Anthropology,* 143(4), 555-569.

Quintana-Murci, L., Quach, H. l. N., Harmant, C., Luca, F., Massonnet, B., Patin, E., Sica, L., Mouguiama-Daouda, P., Comas, D., and Tzur, S. (2008). Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proceedings of the National Academy of Sciences,* 105(5), 1596-1601.

Ramos-Onsins, S. E., and Rozas, J. (2002). Statistical properties of new neutrality tests against population growth. *Molecular Biology and Evolution,* 19(12), 2092-2100.

Razafindrazaka, H., Ricaut, F. X., Cox, M. P., Mormina, M., Dugoujon, J. M., Randriamarolaza, L. P., Guitard, E., Tonasso, L., Ludes, B., and Crubézy, E. (2009). Complete mitochondrial DNA sequences provide new insights into the Polynesian motif and the peopling of Madagascar. *European Journal of Human Genetics,* 18(5), 575-581.

Redd, A. J., Takezaki, N., Sherry, S. T., McGarvey, S. T., Sofro, A. S., and Stoneking, M. (1995). Evolutionary history of the COII/tRNALys intergenic 9 base pair deletion in human mitochondrial DNAs from the Pacific. *Molecular Biology and Evolution,* 12(4), 604.

Reynolds, J., Weir, B. S., and Cockerham, C. C. (1983). Estimation of the coancestry coefficient: basis for a short-term genetic distance. *Genetics,* 105(3), 767-779.

Rosa, A., Brehm, A., Kivisild, T., Metspalu, E., and Villems, R. (2004). MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region. *Annals of Human Genetics,* 68(4), 340-352.

Rosenberg, N.A. (2004). Distruct: a program for the graphical display of structure results. http://www.stanford.edu/group/rosenberglab/distruct.html.

Rousset, F. (2008). Genepop'007: a complete reimplementation of the genepop software for Windows and Linux. *Molecular Ecology Resources,* 8(1), 103-106.

Rozas, J., Sánchez-DelBarrio, J. C., Messeguer, X., and Rozas, R. (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics,* 19(18), 2496-2497.

Salas, A., Richards, M., De la Fe, T. S., Lareu, M. A.-V., Sobrino, B., SÃ¡nchez-Diz, P., Macaulay, V., and Carracedo, A. (2002). The making of the African mtDNA landscape. *American Journal of Human Genetics,* 71(5), 1082-1111.

Schlebusch, C. M., Naidoo, T., and Soodyall, H. (2009). SNaPshot minisequencing to resolve mitochondrial macro haplogroups found in Africa. *Electrophoresis,* 30(21), 3657-3664.

Serva, M., Petroni, F., Volchenkov, D., and Wichmann, S. R. (2012). Malagasy dialects and the peopling of Madagascar. *Journal of The Royal Society Interface,* 9(66), 54-67.

Shields, G. F., Hecker, K., Voevoda, M. I., and Reed, J. K. (1992). Absence of the Asian-specific region V mitochondrial marker in Native Beringians. *American Journal of Human Genetics,* 50(4), 758.

Shriver, M. D., Smith, M. W., Jin, L., Marcini, A., Akey, J. M., Deka, R., and Ferrell, R. E. (1997). Ethnic-affiliation estimation by use of population-specific DNA markers. *American Journal of Human Genetics,* 60(4), 957.

Shriver, M.D., Mei, R., Parra, E.J., Sonpar, V., Halder, I., Tishkoff, S.A., Schurr ,T.G., Zhadanov, S.I., Osipova, L.P., Brutsaert, T.D., Friedlaender, J., Jorde, L.B., Watkins, W.S., Bamshad, M.J., Gutierrez, G., Loi, H., Matsuzaki, H., Kittles, R.A., Argyropoulos, G., Fernandez, J.R., Akey ,J.M., and Jones, K.W. (2005). Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation. *Human Genomics,* 2(2), 81–89.

Singer, R., Budtz Olsen, O. E., Brain, P., and Saugrain, J. (1957). Physical features, sickling and serology of the Malagasy of Madagascar. *American Journal of Physical Anthropology,* 15(1), 91-124.

Soodyall, H., Vigilant, L., Hill, A. V., Stoneking, M., and Jenkins, T. (1996). mtDNA control-region sequence variation suggests multiple independent origins of an" Asian-specific" 9-bp deletion in sub-Saharan Africans. *American Journal of Human Genetics,* 58(3), 595.

Stiles, D. (1991). Tubers and tenrecs: the Mikea of southwestern Madagascar. *Ethnology,* 30(3), 251-263.

Stoneking, M., and Wilson, A.C. (1989). Mitochondrial DNA. In Hill, A. and Serjeantson, S. (eds) *The colonization of the Pacific: a genetic trail* (pp.215-245). Oxford: Oxford University Press.

Sykes, B., Leiboff, A., Low-Beer, J., Tetzner, S., and Richards, M. (1995). The origins of the Polynesians: an interpretation from mitochondrial lineage analysis. *American Journal of Human Genetics,* 57(6), 1463.

Taanman, J. W. (1999). The mitochondrial genome: structure, transcription, translation and replication. *Biochimica et Biophysica Acta (BBA)-Bioenergetics,* 1410(2), 103-123.

Tabbada, K. A., Trejaut, J., Loo, J.-H., Chen, Y.M., Lin, M., MirazÃ³n-Lahr, M., Kivisild, T., and De Ungria, M. C. A. (2010). Philippine mitochondrial DNA diversity: a populated viaduct between Taiwan and Indonesia? *Molecular Biology and Evolution,* 27(1), 21-31.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics,* 123(3), 585-595.

Tajima, F. (1996). The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics,* 143(3), 1457-1465.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution,* 28(10), 2731-2739.

Terreros, M. C., Rowold, D. J., Mirabal, S., and Herrera, R. J. (2011). Mitochondrial DNA and Y-chromosomal stratification in Iran: relationship between Iran and the Arabian Peninsula. *Journal of Human Genetics,* 56(3), 235-246.

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research,* 22(22), 4673-4680.

Tishkoff, S. A., Gonder, M. K., Henn, B. M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T. B., and Ramakrishnan, U. (2007). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Molecular Biology and Evolution,* 24(10), 2180-2195.

Tofanelli, S., Bertoncini, S., Castrì, L., Luiselli, D., Calafell, F., Donati, G., and Paoli, G. (2009). On the origins and admixture of Malagasy: new evidence from high-resolution analyses of paternal and maternal lineages. *Molecular Biology and Evolution,* 26(9), 2009.

Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus, M.-L., and Wallace, D. C. (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics,* 144(4), 1835-1850.

Trejaut, J. A., Kivisild, T., Loo, J. H., Lee, C. L., He, C. L., Hsu, C. J., Li, Z. Y., and Lin, M. (2005). Traces of archaic mitochondrial lineages persist in Austronesian-speaking Formosan populations. *Public Library of Science Biology,* 3(8), 247.

Van Oven, M., and Kayser, M. (2009). Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Human Mutation,* 30(2), E386-E394.

Vérin, P., Koyyak, C., and Gorlin, P. (1969). The glottochronology of Malagasy speech communities. *Oceanic Linguistics,* 8(1), 26-83.

Vérin, P. (1986). *The History of Civilization in Northern Madagascar*. Rotterdam: AA Balkema.

Vérin, P., and Wright, H. (1999). Madagascar and Indonesia: new evidence from archaeology and linguistics. *Bulletin of the Indo-Pacific Prehistory Association,* 18(2), 35-42.

Vigilant, L., Pennington, R., Harpending, H., Kocher, T,D., and Wilson, A.C. (1989). Mitochondrial DNA sequences in single hairs from a southern African population. *Proceedings of the National Academy of Science*, 86(23), 9350–9354.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K., and Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. *Science,* 253(5027), 1503-1507

Vilar, M. G., Kaneko, A., Hombhanje, F. W., Tsukahara, T., Hwaihwanje, I., and Lum, J. K. (2008). Reconstructing the origin of the Lapita cultural complex: mtDNA analyses of East Sepik Province, PNG. *Journal of Human Genetics,* 53(8), 698-708.

Ward, R. H., Frazier, B. L., Dew-Jager, K., and Pääbo, S. (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proceedings of the National Academy of Sciences,* 88(19), 8720-8724.

Watson, E., Bauer, K., Aman, R., Weiss, G., von Haeseler, A., and Pääbo, S. (1996). mtDNA sequence diversity in Africa. *American Journal of Human Genetics,* 59(2), 437-444

Whyte, A. L. H., Marshall, S. J., and Chambers, G. K. (2005). Human evolution in Polynesia. *Human Biology,* 77(2), 157-177.

Wrischnik, L. A., Higuchi, R. G., Stoneking, M., Erlich, H. A., Arnheim, N., and Wilson, A. C. (1987). Length mutations in human mitochondrial DNA: direct sequencing of enzymatically amplified DNA. *Nucleic Acids Research,* 15(2), 529-541.

Xing, J., Watkins, W. S., Witherspoon, D. J., Zhang, Y., Guthery, S. L., Thara, R., Mowry, B. J., Bulayeva, K., Weiss, R. B., and Jorde, L. B. (2009). Fine-scaled human genetic structure revealed by SNP microarrays. *Genome Research,* 19(5), 815-825.

Yao, Y. G., Kong, Q. P., Bandelt, H. J., Kivisild, T., and Zhang, Y. P. (2002). Phylogeographic differentiation of mitochondrial DNA in Han Chinese. *American Journal of Human Genetics,* 70(3), 635-651.

Xing J, Watkins WS, Witherspoon DJ, Zhang Y, Guthery SL, Thara R, Mowry BJ, Bulayeva K, Weiss RB, Jorder LB

APPENDICES

Appendix A

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Professor Himla Soodyall

CLEARANCE CERTIFICATE	**M090576**
PROJECT	Human Genetic Diversity and Disease (Previously M980553)
INVESTIGATORS	Professor Himla Soodyall.
DEPARTMENT	Diversity and Disease Research Unit
DATE CONSIDERED	09.05.29
DECISION OF THE COMMITTEE*	Annual Renewal Approved

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable
cc: Supervisor : Professor T Jenkins

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES...

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Miss Pareen Patel

CLEARANCE CERTIFICATE	**MO90629**
PROJECT	Reconstructing the Prehistory of the Malagasy
INVESTIGATORS	Miss Pareen Patel.
DEPARTMENT	Department of Human Genetics
DATE CONSIDERED	09.06.26
DECISION OF THE COMMITTEE*	Approved unconditionally

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 09.06.26 **CHAIRPERSON** .....
(Professor P E Cleaton Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof H Soodyall

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 10th Floor, Senate House, University.
I/We fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee. **I agree to a completion of a yearly progress report.**

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES...

.....

Appendix B

Recipes for reagents and solutions used

Sucrose-Triton X Lysing buffer
10 ml 1 M Tris-HCl pH8
5 ml 1 M $MgCl_2$
10 ml Triton-X 100
Make up to 1 L with dH_2O and autoclave
Add 109.5 g sucrose just before use
Keep chilled at 4°C

1 M Tris-HCl
121.1 g Tris
1 L dH_2O
Autoclave

1 M $MgCl_2$
101.66 g $MgCl_2$
500 ml dH_2O
Autoclave

T20E5
20 ml 1M Tris-HCl
10 ml 0.5M EDTA pH8
Make up to 1 L with dH_2O and autoclave

0.5 M EDTA
93.06 g EDTA
500 ml dH_2O
pH to 8.0 with NaOH and autoclave

10% SDS
10 g SDS
100 ml dH_2O
Autoclave

Proteinase K 10 mg/ml) Roche Diagnostics)
100 mg Proteinase K stock 100 mg/ml)*
10 ml ddH_20

Proteinase-K mix
For 16 extractions:
400 µl 10% SDS
16 µl 0.5 M EDTA
2.8 ml autoclaved dH_2O
Add 800 µl Proteinase K 10 mg/ml stock) just before use

Saturated NaCl
100 ml autoclaved dH_2O
Slowly add 40 g NaCl until absolutely saturated some NaCl will precipitate out)
Before use, agitate and let NaCl precipitate out

1 X TE buffer
10 ml 1 M Tris-HCl pH8
2 ml 0.5 M EDTA
Make up to 1 L with dH_2O and autoclave

10 X TBE buffer
108 g Tris
55 g Boric acid
7.44 g EDTA
Make up to 1 L with dH_2O and autoclave

1 X TBE 1:10 dilution)
40 ml 10 X TBE
Make up to 200ml with ddH_20

Bromophenol blue Ficoll dye
50 ml dH_2O
50 g sucrose
1.86 g EDTA
0.1 g bromophenol blue
10 g Ficoll
Dissolve
Adjust volume to 100 ml with dH_2O, stir overnight
pH to 8.0
Filter through Whatmann filter paper
Store at room temperature

10 mg/ml Ethidium bromide EtBr)
Add 1 g of ethidium bromide to
100 ml of ddH_2O
Stir for several hours until completely dissolved
Store wrapped in aluminum foil at 4°C

1kb + ladder
285 µl 1kb ladder GibcoBRL)
143 µl Ficoll dye
2 400 µl 1 X TE
310

10 mg/ml BSA
1 g BSA
10 ml ddH_2O
Aliquot into 1 ml amounts and store at –20°C

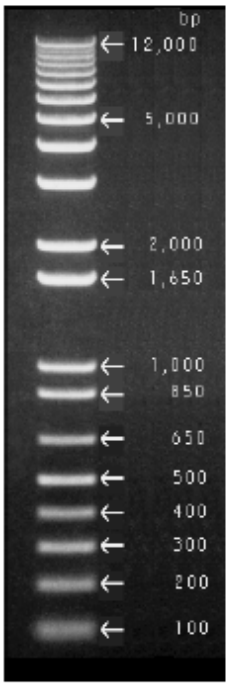
2.5mM dNTPs
Use 100 mM premade stocks of dATP, dGTP, dCTP and dTTP GibcoBRL)
10 µl of each stock dNTP + 360 µl sterile ddH_2O = 400 µl of 2.5 mM dNTPs

Appendix C

Invitrogen
life technologies

1 Kb Plus DNA Ladder

Cat. No. 10787-018 **Size: 250 µg**
Conc.: 1 µg/µl **Store at -20°C.**



1 Kb Plus DNA Ladder
0.7 µg/lane
0.9% agarose gel
stained with ethidium bromide

Structure of Fragments in 1-Kb Increments:

5′-GATCC——— G
G ——— CCTAG-5′
(996 bp)$_n$

Notes:
During 1% agarose gel electrophoresis with Tris-acetate (pH 7.5) as the running buffer, bromophenol blue migrates together with the 500 bp band.

The 1650 bp band is generated from pUC. The bands smaller than 1000 bp are derived from lambda DNA.

Appendix D : Gel Electrophoresis pictures of various PCRs

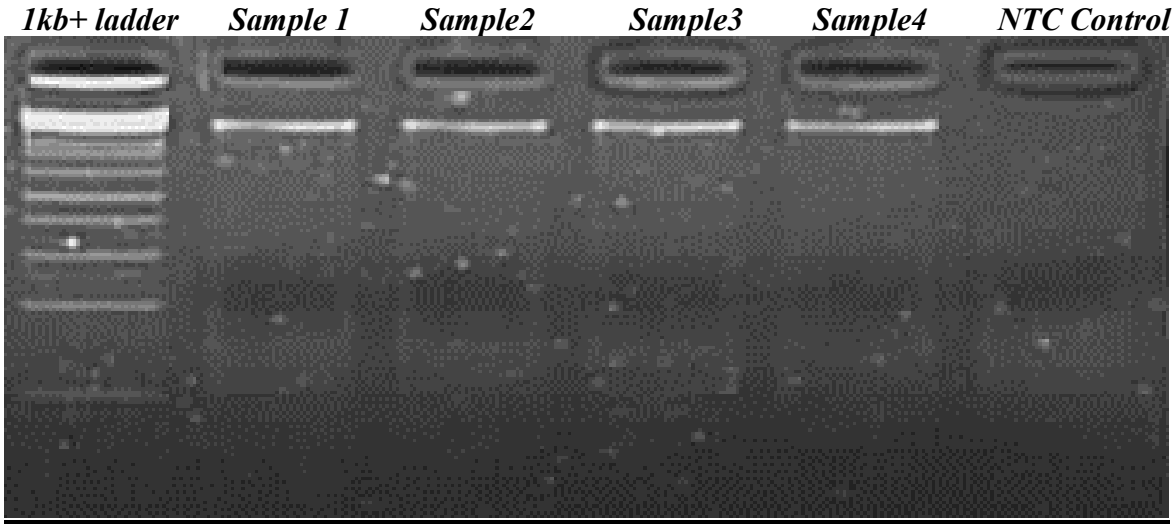


Figure 1: Gel electrophoresis picture of CR I and II 1kb PCR product.

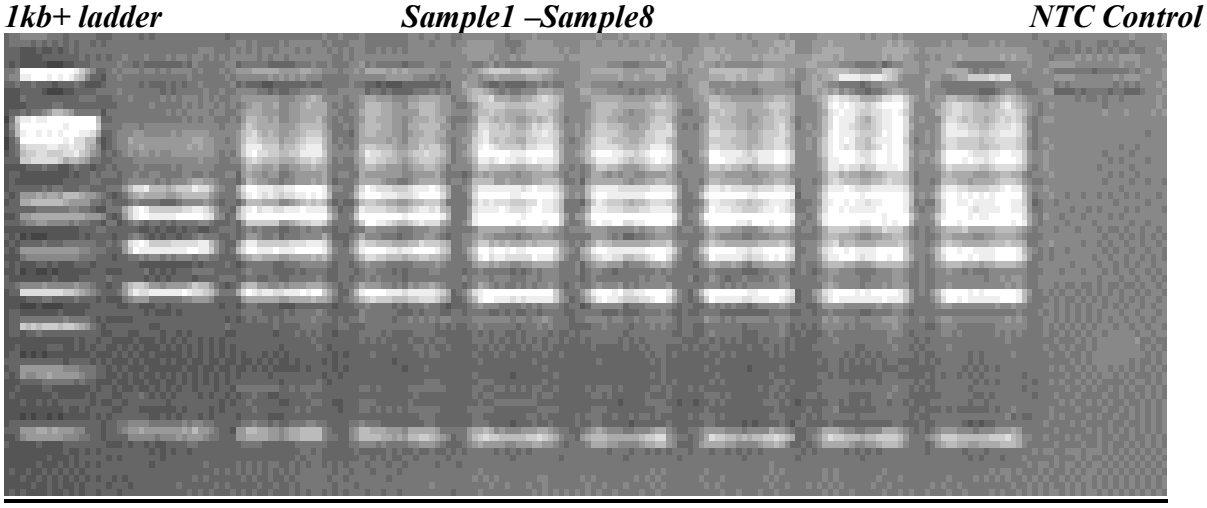


Figure 2: Gel electrophoresis picture of multiplex SBE assay PCR product.

Appendix D : Gel Electrophoresis pictures of various PCRs (CONT.)

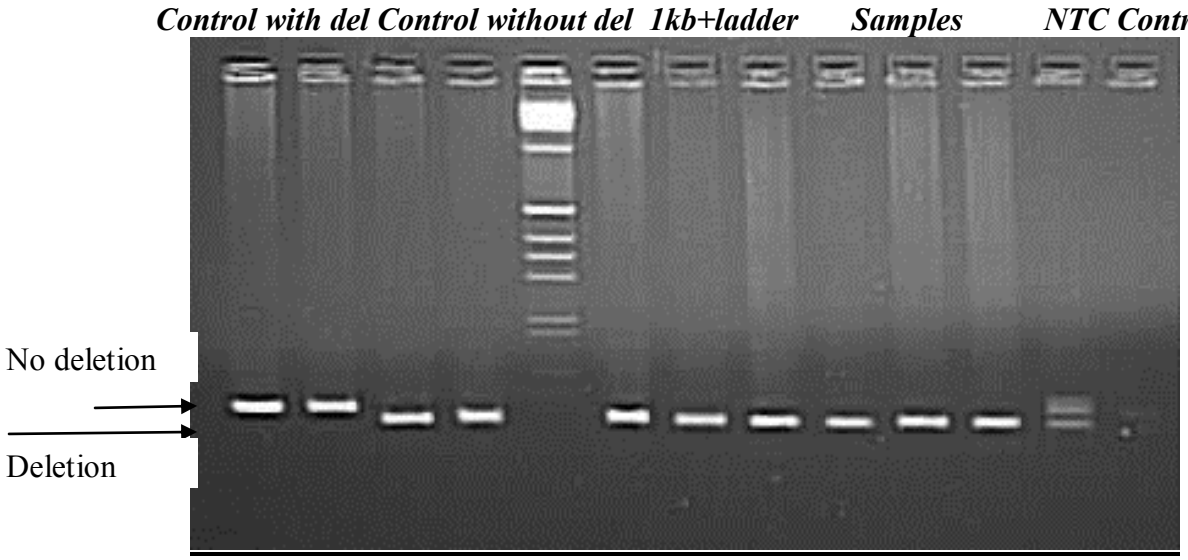


Figure 3: Gel electrophoresis picture of the intergenic COII/tRNALys 9-bp deletion

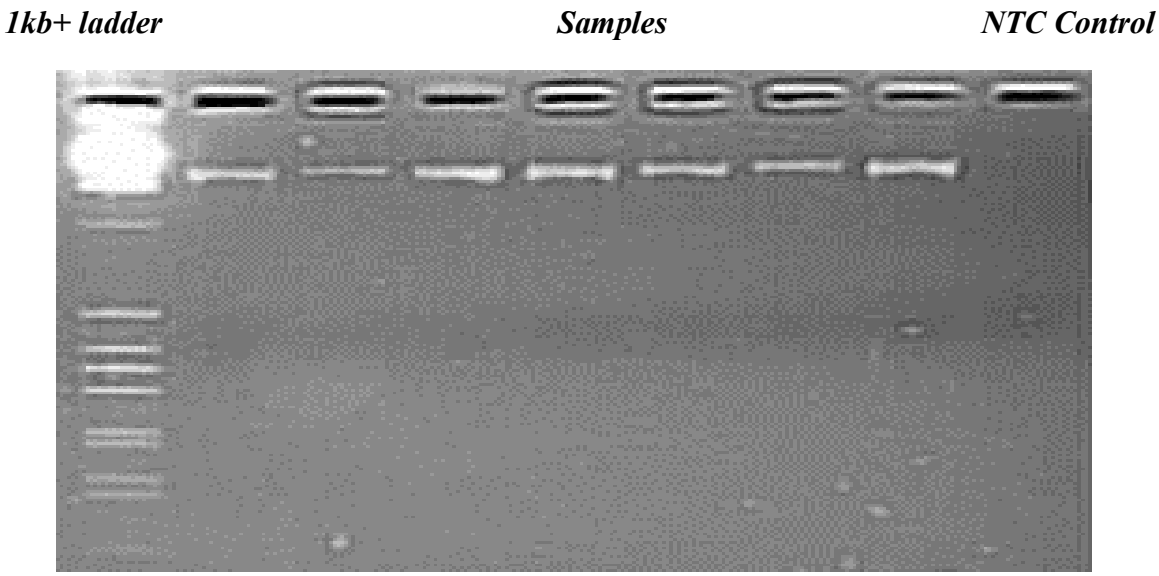


Figure 4: Gel electrophoresis picture of “Malagasy motif” 2kb PCR product.

Appendix D : Gel Electrophoresis pictures of various PCRs (CONT.)

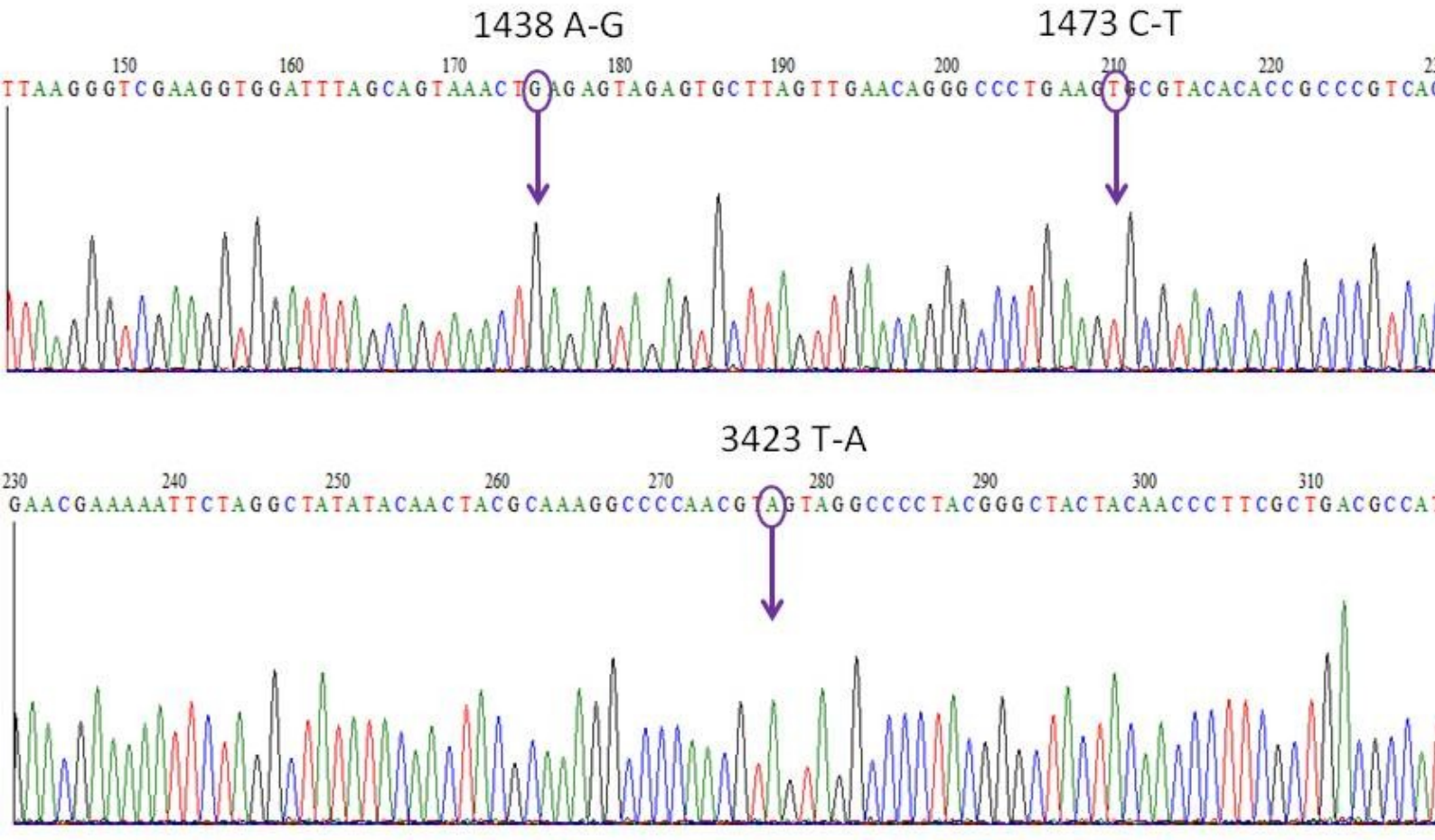


Figure 5: Electropherogram of mtDNA coding region “Malagasy motif” defining SNPs.

Appendix E: List of papers used to compile 96 AIMs panel

Number	Paper	dbSNP	Position on chrom	Country
1	Paschou et al., 2010	rs876901	126006531	Oceania
2	Xing et al., 2009	rs2315523	133085026	Southeast Asia
3	Paschou et al., 2010	rs11096718	17144596	Middle East
4	Paschou et al., 2010	rs9384740	110822900	Middle East
5	Paschou et al., 2010	rs10921877	193861144	Middle East
6	Paschou et al., 2010	rs11119306	207789395	Middle East
7	Paschou et al., 2010	rs1015081	87183612	Europe
8	Paschou et al., 2010	rs12307610	8753428	African
9	Paschou et al., 2010	rs4865414	52453051	African
10	Paschou et al., 2010	rs6060684	33973092	African
11	Paschou et al., 2010	rs6473531	85280011	African
12	Paschou et al., 2010	rs284272	10576215	African
13	Paschou et al., 2010	rs6572814	51613171	African
14	Paschou et al., 2010	rs12356726	91788322	South Asia
15	Paschou et al., 2010	rs3864269	106964864	South Asia
16	Paschou et al., 2010	rs1181804	123969650	South Asia
17	Paschou et al., 2010	rs2348569	86109173	South Asia
18	Paschou et al., 2010	rs4367352	166593062	South Asia
19	Paschou et al., 2010	rs1446770	84006207	South Asia
20	Paschou et al., 2010	rs13088815	59878234	South Asia
21	Paschou et al., 2010	rs2032157	349079	South Asia
22	Harrison et al, 2008	rs1541317	118130009	Europe
23	Harrison et al, 2008	rs1375131	135671267	Europe
24	Shriver, 2005	rs1363448	140763780	East Asia
25	Kidd et al., 2011	rs10007810	41249121	African
26	Kidd et al., 2011	rs1040404	166426514	East Asia
27	Kidd et al., 2011	rs11652805	60417613	African
28	Paschou et al., 2010	rs4886321	60765725	Oceania
29	Paschou et al., 2010	rs7524581	65939087	Oceania
30	Paschou et al., 2010	rs2167597	27667687	Oceania
31	Paschou et al., 2010	rs4941233	53805474	Oceania
32	Paschou et al., 2010	rs1393849	67383123	Oceania
33	Paschou et al., 2010	rs3886608	41742000	Oceania
34	Paschou et al., 2010	rs17712826	75871382	Oceania
35	Paschou et al., 2010	rs10874980	92056487	Oceania
36	Paschou et al., 2010	rs8018542	55440420	Oceania
37	Paschou et al., 2010	rs780881	114688142	Oceania
38	Paschou et al., 2010	rs2738039	10654749	Oceania
39	Paschou et al., 2010	rs1561439	80368206	Oceania
40	Paschou et al., 2010	rs11052381	33060177	Oceania
41	Paschou et al., 2010	rs11153346	112640856	Oceania

42	Paschou et al., 2010	rs2423819	14486864	Middle East
43	Paschou et al., 2010	rs4803910	51416642	Middle East
44	Paschou et al., 2010	rs9822326	158286259	Middle East
45	Paschou et al., 2010	rs3782025	113312817	Middle East
46	Paschou et al., 2010	rs526260	75654161	Middle East
47	Paschou et al., 2010	rs885834	50485518	Middle East
48	Paschou et al., 2010	rs4975217	130241611	Middle East
49	Paschou et al., 2010	rs12427865	88075711	Middle East
50	Paschou et al., 2010	rs4143744	147405354	Middle East
51	Paschou et al., 2010	rs6734107	206840762	Middle East
52	Paschou et al., 2010	rs7873939	23132396	Middle East
53	Paschou et al., 2010	rs3870375	122832789	Middle East
54	Paschou et al., 2010	rs13147558	133677689	Europe
55	Paschou et al., 2010	rs1368684	8914538	Europe
56	Paschou et al., 2010	rs12365527	82483572	Europe
57	Paschou et al., 2010	rs9609759	32043686	Europe
58	Paschou et al., 2010	rs4291226	27240494	Europe
59	Paschou et al., 2010	rs3780610	73518742	Europe
60	Paschou et al., 2010	rs10008492	38442115	Europe
61	Paschou et al., 2010	rs721424	20287310	Europe
62	Paschou et al., 2010	rs1239905	68727252	Europe
63	Paschou et al., 2010	rs12913832	26039213	Europe
64	Paschou et al., 2010	rs4791850	9380286	Europe
65	Paschou et al., 2010	rs1373543	97515211	Europe
66	Paschou et al., 2010	rs779330	193053791	Europe
67	Paschou et al., 2010	rs2420318	118397558	East Asia
68	Paschou et al., 2010	rs103294	59489660	East Asia
69	Paschou et al., 2010	rs12449949	74796899	African
70	Paschou et al., 2010	rs1519590	52541357	African
71	Paschou et al., 2010	rs16841669	135658734	African
72	Paschou et al., 2010	rs10800120	163871934	African
73	Paschou et al., 2010	rs10424191	35749189	African
74	Paschou et al., 2010	rs10072072	118629764	African
75	Paschou et al., 2010	rs2027742	73210278	African
76	Paschou et al., 2010	rs7654542	35300055	African
77	Lao et al. 2006	rs1823718	71934297	Oceania
78	Paschou et al., 2010	rs11583200	50332407	South Asia
79	Paschou et al., 2010	rs7304582	59898929	South Asia
80	Paschou et al., 2010	rs4789396	72503206	South Asia
81	Paschou et al., 2010	rs11841562	21009521	South Asia
82	Paschou et al., 2010	rs10205766	105003863	South Asia
83	Paschou et al., 2010	rs1048048	92825764	South Asia
84	Dhandapany et al., 2010	rs6974649	130463637	South Asia
85	Paschou et al., 2010	rs35607	16069981	South Asia

86	Xing et al., 2009	rs16879442	32444084	Southeast Asian
87	Xing et al., 2009	rs10821148	95323125	Southeast Asian
88	Xing et al., 2009	rs11151371	63326628	Southeast Asian
89	Xing et al., 2009	rs4260664	130990185	Southeast Asian
90	Xing et al., 2009	rs1611350	29806800	Southeast Asian
91	Xing et al., 2009	rs2061740	163719579	Southeast Asian
92	Xing et al., 2009	rs10427222	55660792	Southeast Asian
93	Xing et al., 2009	rs488928	104294431	Southeast Asian
94	Xing et al., 2009	rs1888839	99800921	Southeast Asian
95	Xing et al., 2009	rs2237315	24728375	Southeast Asian
96	Xing et al., 2009	rs4434311	66401913	Southeast Asian

Appendix F: Haplogroup breakdown by region

Key:

Antaimoro	ATM
Antaisaka	ATS
Antandroy	ATD
Antanosy	ATO
Antaifasy	ATF
Antambahoaka	ATB
Antankarana	ATK
Bara	BAR
Betsileo	BET
Bezanozano	BEZ
Betsimisaraka	BES
Merina	MER
Mikea	MIK
Mahafaly	MAH
Makoa	MAK
Sihanaka	SIH
Sakalava	SAK
Tanala	TAN
Tsimihety	TSI
Vezo	VEZ
Zafisoro	ZAF

Appendix F: Haplogroup breakdown by region

	CENTRAL HIGHLAND			EASTERN CENTRAL					NORTHERN	SOUTH WESTERN				WESTERN			SOUTH EASTERN						
mtDNA HG	BET	BEZ	MER	ATM	ATS	ATK	BES	SIH	TSI	ATD	MIK	MAH	BAR	MAK	VEZ	SAK	ATO	ATF	ATB	ZAF	TAN	TOTAL	
Total	109	20	138	81	99	15	90	48	86	45	34	29	18	13	25	70	35	11	5	6	4	981	%
L0a1a																1						1	0.1019
L0a1b				1			1															2	0.2039
L0a1c	1			1											1							3	0.3058
L0a2				1	2	1	1		1													6	0.6116
L0a2a1						1	1		5					1		2	2	1				13	1.3252
L0a2a2	6		3	5	2	1	9		5				2		1	1	1					36	3.6697
L1c1		1			2											1						4	0.4077
L1c1b															1							1	0.1019
L1c1c																				1		1	0.1019
L1c1d					1		1															2	0.2039
L1c2															1	1						2	0.2039
L1c2b1									2							2						4	0.4077
L1c3a							1		1													2	0.2039
L1c3b1							1															1	0.1019
L1c3c							2															2	0.2039
L2				1						1												2	0.2039
L2a											1											1	0.1019
L2a1	1			2	1	2	2		2	1	1	1				2	1					16	1.6310
L2a1a	1		1	1		1	1		3	2	1	1				1						13	1.3252
L2a1a2	2		2	1	4		3		1				1		2	1					1	18	1.8349
L2a1b	3	1	1	8	9		8	5	8	5	1	9	2	2	1	3	3	1				70	7.1356
L2b																						0	0.0000

	CENTRAL HIGHLAND			EASTERN CENTRAL					NORTHERN	SOUTH WESTERN				WESTERN			SOUTH EASTERN						
mtDNA HG	BET	BEZ	MER	ATM	ATS	ATK	BES	SIH	TSI	ATD	MIK	MAH	BAR	MAK	VEZ	SAK	ATO	ATF	ATB	ZAF	TAN	TOTAL	
Total	109	20	138	81	99	15	90	48	86	45	34	29	18	13	25	70	35	11	5	6	4	981	%
L2b1				1																		1	0.1019
L2b3	1						1		1									1				4	0.4077
L2c2														1								1	0.1019
L3a											1											1	0.1019
L3b1a	3	3	12	11	9	1	4	2	5	3	2	2	2	1		3	1	1				65	6.6259
L3b1b																1						1	0.1019
L3b2										1	1											2	0.2039
L3d					1																	1	0.1019
L3d1a1	1	1	3	1	4		5	1	1	3				1		2	1	1				25	2.5484
L3e1a	2	1	1	2	1		2	1	2					1	1	2			2			18	1.8349
L3e1a1			1							1		1				2						5	0.5097
L3e1b									1	2												3	0.3058
L3e1c							1	1	1	1							1					5	0.5097
L3e1d	2			1							1					1		1				6	0.6116
L3e2b			1	1	3		1										2	1				9	0.9174
L3e2b2					1																	1	0.1019
L3e3	5	1	7	2	2	2	7	2	9							1					1	39	3.9755
L3f1b1															1							1	0.1019
L3f2	4										1		1		1	1						8	0.8155
L3h1a1				1	1																	2	0.2039
L3x													1									1	0.1019
L4b1		2										1										3	0.3058
L4b2														1								1	0.1019
L4b2a1				1							1											2	0.2039

	CENTRAL HIGHLAND			EASTERN CENTRAL					NORTHERN	SOUTH WESTERN				WESTERN			SOUTH EASTERN						
mtDNA HG	BET	BEZ	MER	ATM	ATS	ATK	BES	SIH	TSI	ATD	MIK	MAH	BAR	MAK	VEZ	SAK	ATO	ATF	ATB	ZAF	TAN	TOTAL	
Total	109	20	138	81	99	15	90	48	86	45	34	29	18	13	25	70	35	11	5	6	4	981	%
H											1											1	0.1019
N																1						1	0.1019
B4a1a1	2		2	1	2	1		1	4	1							1					15	1.5291
B4a1a1a	23	3	40	10	21	1	17	18	10	9	7	3	2	4	4	13	6		2	1	2	196	19.9796
M	1	1	4	3			1		1	0			1		3	1	1	1		1		19	1.9368
M23	6	3	12	5	6		5	4	5	5		2			4	5	1			1		64	6.5240
M32c	11	2	13	5	3	1	1	1	3	4			2			3			1			50	5.0968
M7c3c	15		19	5	13	2	8	7	9	4	5	2	1	1	2	10	6	1		1		111	11.3150
E1a1a	11	1	10	2	4	1	2	5	3		8	6	1		1	7	5	1				68	6.9317
F3b	7		6	7	7		3		2	2	2	1	2		1	2	3	1		1		47	4.7910
F4b	1		0	1			1		1			0										4	0.4077
																						981	100.0000