AN EXPLORATION OF THE ORIGINS OF THE MALAGASY USING GENETIC POLYMORPHISMS

Bharti Morar

A thesis submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Doctor of Philosophy.

Johannesburg, 2000.

DECLARATION

I, Bharti Morar declare that this thesis is my own, unaided work. It is being submitted for the degree of Doctor of Philosophy in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at this or any other University.

.....B Morar.....

.....7..... day of ...February., 2000.

This work is dedicated to my husband and children, Sameer and Prashant for their understanding and support throughout my research career.

PUBLICATIONS AND PRESENTATIONS

PUBLICATIONS

1. Kidd KK, **Morar B**, Castiglione CM, Zhao H, Pakstis AJ, Bonne-Tamir B, Lu R, Goldman D, Lee C, Nam YS, Grandy DK, Jenkins T, Kidd JR. A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* **103**:211-227 (1998).
2. Soodyall H, **Morar B**, Jenkins T. "Genetic archives". Chapter in book "Refiguring the Archive", organized by the Graduate School for the Humanities and Social Sciences, University of the Witwatersrand (submitted).

PRESENTATIONS AT SCIENTIFIC MEETINGS

1. **Morar B**, Lane AB and Jenkins T. The peoples of Madagascar and their affinities. 9th International Congress of Human Genetics, 18-23 August 1996, Rio de Janeiro, Brazil.
2. **Morar** B, Lane AB, Jenkins T. Origins and affinities of the Malagasy. 9th International Congress of Human Genetics, 18-23 August 1996, Rio de Janeiro, Brazil.
3. **Morar B**, Lane AB and Jenkins T. Genetic memories of population migrations. 7th Biennial Southern African Society of Human Genetics, 18-21 May 1997, Pilanesburg National Park, South Africa.
4. **Morar B**, Lane AB and Jenkins T. The 'Y's of Madagascar. Dual Congress 1998, International Association for the Study of Human Palaeontology and International Association of Human Biologists, 28 June-4 July, Johannesburg and Pretoria.

5. **Morar B**, Lane AB, Soodyall H and Jenkins T. The origins of Y Malagasy chromosomes. 8th Biennial Southern African Society of Human Genetics, 9-12 March 1999, Gordon's Bay, South Africa.
6. Soodyall H, Stermer T, Jo..nou K, Eddy C, Barkhan D, **Morar B** and Jenkins T. Genetic affinities of Sub-Saharan African populations. 8th Biennial Southern African Society of Human Genetics, 9-12 March 1999, Gordon's Bay, South Africa.
7. Soodyall H, **Morar B** and Jenkins T. Genetic diversity in African and Malagasy populations. Wenner Gren Foundation for Anthropological Research, Symposium no. 124, 11-19 June 1999, Teresopolis, Brazil.

ABSTRACT

'*Malgache, qui es-tu?*' This question seeking the ancestry of the Malagasy remains unanswered five centuries after the debate on the proposed hypotheses of Malagasy origins began. Historical, archaeological, linguistic and some genetic data suggest two major sources of ancestry: Africa and Indonesia, with minor contributions from Arabia, India and China, but the evidence for the suggestions is sparse and inconclusive. All Malagasy, irrespective of their ethnic backgrounds speak a common language, Malagasy, but differences in physical appearance and culture suggests that they may have different ancestral histories. The goals of this study are to utilize genetic variation present in the Malagasy in conjunction with available data to reconstruct their prehistory and to provide evidence confirming and/or refining existing theories concerning their prehistory. The genetic profiles of eleven of the eighteen Malagasy ethnic groups, a South African Indian and six African populations were compared at thirteen loci; eight autosomal and five Y chromosomal. The markers include four STR loci each on the autosomes (HUMCSF1PO, HUMTH01, HUMCD4 and an (AC)n repeat in the DRD2 gene) and on the Y chromosome (DYS393, DYS19, DYS390 and DYS391), two *Alu* polymorphisms (CD4-*Alu* and YAP) and three RFLP loci within the DRD2 gene. Some of these loci were used to derive autosomal and Y chromosome haplotypes. Population trees as well as principal component analyses based on the different data sets consistently revealed very close affinities between the eleven Malagasy groups examined. The intermediate clustering of the Malagasy between African and South Asians also reaffirms that these two groups have contributed significantly to the Malagasy gene pool. Admixture estimates made using autosomal data suggest that approximately 50% of the Malagasy gene pool is derived from an African source while Y chromosome data indicates an African contribution of at least 60%. Networks constructed using Y chromosome haplotypes identified a 'Malagasy-specific'

haplotype cluster. A divergence time of 2864 years (95% confidence intervals, 1227 – 7472 years) is estimated for this cluster which is consistent with archaeological data suggesting that the colonization of Madagascar occurred within the past 3000 years. Both autosomal and Y chromosomal data from the present study supports a recent common ancestry for the Malagasy from founders whose gene pool contained contributions from Indonesians and Africans.

ACKNOWLEDGEMENTS

I am indebted to several individuals and organizations that contributed to the completion and success of this research:

My supervisors, Dr Anthony Bruce Lane and Prof. Trefor Jenkins, for the opportunity of carrying out this research project in the department of Human Genetics, SAIMR and for their encouragement and support throughout my studies.

Dr Anthony Bruce Lane, Prof. Trefor Jenkins and Dr Himla Soodyall for reading drafts of this thesis and for their valuable comments and suggestions.

Dr Anthony Bruce Lane for writing and making available a number of DOS-based statistical programmes used in the computational analyses.

Dr Himla Soodyall for her mentoring during the major part of my studies.

I extend my gratitude to all the volunteers who contributed blood samples and to Prof. T Jenkins, Dr T de Ravel and Dr H Soodyall (Dept of Human Genetics, SAIMR), Dr P Vallem (Dept of Haematology, SAIMR), Dr G Campbell (University de Avignon, Avignon), Dr A Rakouth and Dr S Quambek (SALFA, Antananarivo) and Prof. J Roux (Pasteur Institute, Madagascar).

In addition to samples collected by Prof. Jenkins and colleagues in the department, I would also like to acknowledge the contributions of samples from other sources: Prof. KK and Dr J Kidd (Department of Genetics, Yale University School of Medicine, New Haven, USA), Dr JB Clegg and Dr D Rees (MRC Molecular Haematology Unit, John Radcliffe Hospital, Oxford) and Dr DG Woodfield (Department of Transfusion Medicine, Auckland Regional Blood Services, New Zealand).

My colleagues in the department, who, besides assisting with the processing of blood samples for the project, provided a pleasant and friendly work environment.

Dr Alan Redd and Prof. Mark Stoneking for generously making available unpublished Y chromosome data from SE Asian/Oceanic/Australian populations used in the comparative analyses.

I was most fortunate to have personally received financial support from the University of the Witwatersrand and the Stella and Paul Loewenstein Trust which afforded me the opportunity of attending scientific meetings both locally and abroad. In addition, this research project was supported jointly from grants awarded to Prof. Jenkins and Dr Lane from the SAIMR, the University of the Witwatersrand and the National Geographic.

I wish to extend my sincerest gratitude to my parents, Chhotoo and Koosum Madhoo for their unstinting support and encouragement throughout my career.

Lastly, I wish to thank my husband Mahesh, and children Sameer and Prashant who had to endure some stressful times. I am most appreciative of their love and understanding.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

LIST OF TABLES

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

NOMENCLATURE

ACD	acid citrate dextrose
AD	Antandroy
AK	Antankarana
AM	Antaimoro
APS	ammonium persulphate
aq	aqueous
AS	Antaisaka
ASO	allele specific oligonucleotide
Bis	N,N'-methylene-bis-acrylamide
bp	base pair
BS	Betsimisiraka
BSA	bovine serum albumin
BT	Betsileo
°C	degrees celsius
CAR	Central African Republic
Ci	curies
cm	centimetr
DNTP	deoxyribonucleotide triphosphate
DNA	deoxyribonucleic acid
ds	double stranded
EDTA	ethylenediamine tetra-acetic acid
g	centrifugal force
g	gram
HS	human-specific
hr	hour
IAM	infinite allele model
kb	kilobase
KHz	kilohertz
km	kilometre
l	litre
µl	microlitre
µg	microgram
µM	micromolar
M	molar
MF	Mahafaly
MK	Mikea
mg	milligram
$MgCl_2$	magnesium chloride
min	minute
ml	millilitre
mm	millimetre
mM	millimolar
MR	Merina
MW	molecular weight
ng	nanogram
NJ	neighbour-joining
nm	nanometre

nmol	nanomole
%	percentage
PA	polyacrylamide
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
pmol	picomole
PNG	Papua New Guinea
RE	restriction enzyme
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
rpm	revolutions per minute
RT	room temperature
s	second
SA	South African
SDS	sodium dodecyl sulphate
SE	South East
SK	Sakalava
SMM	stepwise mutation model
SNF	supernatant fluid
SNP	single nucleotide polymorphism
STR	short tandem repeat
TE	Tris-EDTA solution
TEMED	N,N,N',N', tetramethylethylene diamine
Tris	tris (hydroxymethyl) amino methane
TS	Tsimihety
U	unit(s)
UV	ultraviolet
V	volt(s)
VNTR	variable number of tandem repeats
vol	volume
vs	versus
w/v	weight per volume
YAP	Y *Alu* polymorphism

CHAPTER 1 - INTRODUCTION

Extensive genetic variation occurs in most populations and this may be reflected at one or more phenotypic levels including variation in physical appearance, chromosome structure (at the gross or molecular level) and protein characteristics. This variation has been passed on to us from our ancestors, modified only by recombination and mutation. With the advent of molecular technology which has made possible the detection and characterization of variation at the DNA level, this record of our history can be read and has proved to be useful in the study of human variation and evolution. Since genetic drift, migration and shared genetic responses to environmental challenges shape gene pools, closely related populations are more likely to have similar frequencies of polymorphic variants than to distantly related ones. The formation of a new population by the amalgamation of groups from two or more different parental populations results in an averaging of allele frequencies. It is possible to obtain information about the progenitors of a population formed in this way by studying its gene frequencies. This has been done in the present study which is aimed at deciphering the ancestry of the Malagasy.

The island of Madagascar separated from Africa about seventy million years ago, before the evolutionary origin of the monkeys and great apes. This long period of separation may explain why so many of its animal and plant species are found nowhere else in the world and why the highest and commonest primates present (other than man) are the lemurs. Because of its long isolation, the island has been described as the laboratory of evolution (Cole, 1992). Colonization by humans occurred relatively recently, and relatively little is known about the island's natural history. Several theories have been put forward to account for the settlement of the island, but a number of questions concerning the nature of the first

settlers (the proto-Malagasy), when and from where they arrived, the routes they used, what parts of the island they initially occupied and the reasons for the colonization still remain unanswered. Historical, archaeological, linguistic and preliminary genetic data suggest two major sources of ancestry: Africa and Indonesia, but the available evidence for this is fairly poor. Many of the early French and Malagasy articles suggest that a stigma was attached to the acknowledgement by a Malagasy of an African ancestry. In most of these articles, the African-Negroid characteristics of the Malagasy are either minimized or described as 'accidental', or associated with Melanesians or Oceanic negritos.

This chapter deals with a number of introductory topics related to the task of elucidating the ancestry of the Malagasy. Amongst these are some summaries of the current knowledge of Madagascar, details of the molecular approaches available for answering questions related to the study and the relative values of different kinds of genetic markers. The chapter ends with a summary of the aims of the study.

1.1 Madagascar - the island

The island of Madagascar is roughly 1600km long and 560km wide and lies in the west Indian Ocean between latitude 12 and 26 degrees south and longitude 43 and 47 degrees east. With an area of almost 600 000 square kilometers, it is the world's fourth largest island and is separated from the southeast African mainland by the 400km wide Mozambique Channel (**Figure 1.1**). Madagascar is thought to have been settled by humans within the past 2000 years, but the exact timing of this settlement and the origins of the first settlers have long been subject to alternative interpretations amongst scholars.

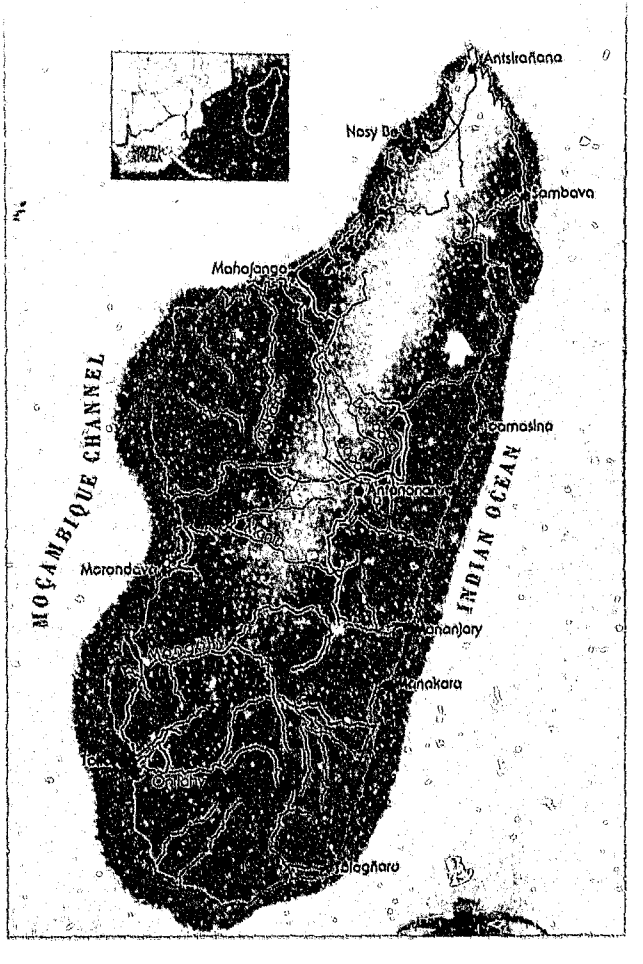


Figure 1.1. The location of Madagascar relative to Africa (Schuurman, 1994).

The island is divided into a number of different geographic regions whose populations, assisted by natural barriers to inter-regional communication, have developed distinct lifestyles (Brown, 1978; Mack, 1986). A longitudinal chain of mountains runs from north to south and divides the island into two parts; a narrow eastern section which slopes steeply, discharging numerous rivers through deep forest gorges into the sea; and a broad western section whose major portion forms the central highlands with an altitude of 2876m. The climate of this latter region is much healthier than the hot and humid fever-ridden eastern coastal plains. Whereas rice cultivation is suited to the highlands, introduced crops such as vanilla, cloves, sugar cane and coffee thrive along the humid coastal belt. The predominant economy in the southwest, which is semi-arid grasslands, is based on cattle herding.

Most of the fauna and flora of Madagascar is indigenous and has been well preserved because of the relatively short period of occupation by man. Much of the prehistory of the island is an enigma but there is no doubt that there are distinct relationships between the fauna of Madagascar and those of Africa and Asia. A single endemic starling species on the island shows ties with Asia rather than Africa (Craig, 1996). Sometime in the last millenium, at least seventeen species of the largest native mammals, birds and reptiles, including large lemurs, cow-sized hippopotami, giant tortoises and elephant birds disappeared from Madagascar (Burney and MacPhee, 1988). Today, no indigenous mammals weighing more than about twelve kilograms are found on Madagascar. Archaeological evidence suggests the co-existence of humans and large native animals in northern Madagascar and it has been suggested that human settlers on the island were partially responsible for the change in the islands flora and fauna.

1.2 The people of Madagascar

Madagascar is one of the world's poorest countries and only 28% of its 14 million people (1997 estimate) are classified as being urban. The people of Madagascar (the Malagasy or the Malgache) have been described as the progeny of the Indian Ocean, because they are thought to have ancestors from all its shores except Australia (Dewar and Wright, 1993). The literature on their origins goes back to the sixteenth century and a number of theories about their ancestry have been put forward. The first inhabitants of the island said to be the Vazimba, Behosy, Kimosy, Mikea and others are thought to be either the descendants of the first proto-Malagasy immigrants (Grandidier and Grandidier, 1908; Verin, 1975; Brown, 1978), or descendants of Africans brought to the island by Swahili and Indonesian traders (Ferrand, 1908; Verin, 1986).

Colonization and settlement of the island is thought to have taken place in multiple waves, probably originating in the SE Asian archipelago in the early part of the first millennium AD. The same people are believed to have introduced food plants like bananas, coconuts, sugar-cane and taro to Africa, which formed the basis of the Bantu agricultural revolution (Deschamps, 1965). Although the Malagasy were initially exporters of slaves, a market for imported African slaves developed in the nineteenth century (Campbell, 1989) and this continuous movement of people to and from the island may have substantially influenced the present-day Malagasy gene pool. Whether or not the island had a social network of its own from the start or was settled by more or less isolated groups, whose main contacts were with people outside Madagascar, remains unclear.

The ethnic, cultural or 'tribal' classification of the people of Madagascar is not based on common descent but is mainly political or based on geographic location. The island has eighteen officially recognized ethnic groups (**Figure 1.2**), each belonging to one of two subgroups based on their geographic location (Brown, 1978):

- **The highlanders** are the groups living on the plateau region and include the Merina (MR), Betsileo (BT), Bezanozano (BZ), Tanala (TN) and Sihanaka (SH).
- **The lowlanders** are the groups living along the coastal and inland low-lying areas and include the Antaisaka (AS), Tsimihety (TS), Sakalava (SK), Mahafaly (MF), Antandroy (AD), Betsimisiraka (BS), Antaimoro (AM), Antankarana (AK), Mikea (MK), Bara (BA), Antaifasy (AF), Antanosy (AN) and the Antambahoaka (AB).

The relative sizes of these groups vary tremendously (Vérin *et al*, 1970). The Merina make up about 27% of the total population, the Betsileo 12%, the Betsimisiraka 15%, the Tsimihety 7%, the Sakalava 6% and the Antaisaka 5%. Tribes living in the lowland areas are collectively called the cotiers; the Betsimisiraka, Tsimihety and Antandroy are the la 'gest of this group. Further details of only the groups included in the present genetic study are discussed below.

1.2.1 The highlanders

a) The **Merina** - This group occupies the central highlands. Popular tradition considers them to be the descendants of the Vazimba (legendary dwarf ancestors), Javanese and Malays. Unusually high proportions of individuals with Indonesian-like features are found in the Merina. This group represents the largest of the Malagasy ethnic groups and in the nineteenth century, they conquered and ruled almost a third of the entire

island (Brown, 1978; Campbell, 1989). The Merina themselves may be divided into three major groups with decreasing hierarchical status (Singer *et al*, 1957): the Andriana or nobles, the Hova or free men (commoners) and the Mainty, Mpanompo or Andevo (the slaves).

b) The **Betsileo** - This group ('the many unconquered') occupies the southern part of the high central plateau and are said to be excellent rice growers. Their physical appearance has been described as being both predominantly Negroid (Singer *et al*, 1957) and intermediate between Negroes and Malays (Brown, 1978, Campbell, 1995).

1.2.2 The lowlanders

a) The **Antaisaka** – Very little has been written about this group who occupy the southern region of the island. They are thought to have high proportions of both Negroid and Malay admixture.

b) The **Tsimihety** - The tribal name means "those that do not use the scissors" (to cut their hair). They occupy the northwest coast and their origins are very obscure (Singer *et al*, 1957). It has been suggested that they represent the original proto-Malagasy group that settled in the northern region of Madagascar (Brown, 1978).

c) The **Sakalava** - The name means "those of the long valley" and includes all the tribal groups of the west coast. They are considered to be mainly of African origin because of their physical appearance, the relatively high proportion of African words in their dialect and the retention of certain customs (tromba/spirit possession and the dady/cult of royal relics) which are clearly of African origin (Brown, 1978). Wooden carvings on the tombs of Sakalava people are also remarkably similar to soapstone carvings of birds found in the ancient African kingdom of Zimbabwe. The Sakalava may be

divided into two main groups: the agro-pastoral Masikoro of the interior and the coastal Vezo who are fishermen, sailors and accomplished swimmers (Stiles, 1988).

d) The **Mahafaly** – These people are found on the southwest coast of the island and are also thought to be more African-like in culture and appearance than the highlanders.

e) The **Antandroy** - This semi-nomad group occupies the southern region of Madagascar and may have a larger African ancestry than the highland people (Brown, 1978).

f) The **Betsimisiraka** - "The inseparables" occupy the long narrow coastal strip on the east. They are the second largest ethnic group on the island. Despite having African admixture, they are thought to be more similar to the Merina than other lowlanders (Campbell, 1995).

g) The **Antaimoro** - The tribal name means "those on the coast". They live along the southeast coast and a strong Arabic/Islamic cultural influence due to a significant Arabian ancestry has been noted amongst the people. The turban and the fez are worn together with versions of Arab robes and mosque-like domes are erected over graves. They are the owners of the sacred manuscripts known as the *Sorabe*.

h) The **Antankarana** – These "people of the rocky country" are dispersed throughout the northern part of the island. Their culture has some similarities with Arabic culture which suggests that they may have some 'Arab' blood.

i) The **Mikea** - This is a group of foraging people in southwestern Madagascar. Marked physical variation exists among them including many combinations of Austronesian and Bantu features. Two divisions of the Sakalava, the Masikoro and the Vezo are the principal neighbors of the Mikea (Stiles, 1988). The Mikea themselves can also be separated into two groups living side by side; the western ones speak the Vezo dialect (Mikea-Vezo) and the eastern ones speak the Masikoro dialect (Mikea-Masikoro). They are said to be a peaceful people who do not welcome outsiders into their country.

Figure 1.2. The approximate geographic locations of the eighteen ethnic groups on Madagascar. The eleven groups examined in the present study are shown in blue (two highland groups) and pink (nine lowland groups).

1.3 Theories of settlement of Madagascar

1.3.1 Origins

Several lines of evidence have led to the formulation of four basic hypotheses regarding the origins of the Malagasy (Campbell, 1995). *Hypothesis 1* suggests that Madagascar possessed an autochthonous (aboriginal or indigenous) population. This hypothesis however was only seriously entertained by the British School (mainly missionaries) until about 1880 when studies on the natural history of the island showed that the earliest evidence of human habitation on Madagascar was in the early part of the first millenium AD (reviewed in Dewar and Wright, 1993).

Hypothesis 2 postulates that the original settlement of Madagascar was by peoples of African origin who may have later been conquered by Indonesian migrants. This hypothesis was presented in 1908 by Ferrand (member of the French School), and though more feasible than *hypothesis 1* because of the proximity of Madagascar to Africa, also lacks support. There is no evidence of the use of any long-distance maritime techniques or any tradition of voyages away from their own coastal waters amongst East African people. Archaeological evidence suggests that Bantu-speakers did not reach the East African coast until after the middle of the first millennium AD, although Negroid peoples could have been present here prior to this. It is therefore highly unlikely that a Bantu migration preceded that of the proto-Malagasy to Madagascar (Murdoch, 1959).

Hypothesis 3 postulates that Madagascar was initially and overwhelmingly settled by migrants from the Malayo-Indonesian-Polynesian region, with smaller migrations of mainly Indian and Semitic people and a minor late contribution from Africa. The

undisputed presence of people from these regions in East Africa during the early part of the first millennium AD and the presence in all Malagasy groups of common Bantu words have been taken as evidence against this hypothesis (Murdoch, 1959).

Hypothesis 4 postulates that the migration/s leading to the initial colonization of Madagascar went via Africa. The proto-Malagasy thus comprised genetically mixed elements, mostly Indonesian and African. This hypothesis is supported by much of the available data as will hopefully become apparent from the rest of this chapter. Historical evidence exists of far-ranging Indonesian voyages into the western Indian Ocean in the early part of the Christian era lending support to the third and fourth theories.

The people of SE Asia have been described as skilled mariners. Austronesian speaking people, after leaving their homeland in Taiwan about 5000 years ago, sailed south into insular SE Asia, where they soon overshadowed pre-existing populations (Bellwood 1985, 1987). From there, they went on sailing east into the Pacific, all the way to Easter Island, leaving few islands unpopulated. It has been suggested that they also went west across the Indian Ocean, partly turning Madagascar culturally and linguistically into an offshoot of SE Asia (Manguin, 1993). The proto-Malagasy brought or soon acquired a mixture of cultural practices, including an east African style of cattle husbandry and Oriental agricultural techniques.

1.3.2 Routes

Hypotheses 3 and *4* (discussed in the previous section) are the only two that remain in contention and two routes have been proposed which the proto-Malagasy may have used to reach Madagascar (Ferrand, 1908; Verin, 1970):

1. A direct route across the southern Indian Ocean from Java to the East Coast of Madagascar (Grandidier and Grandidier, 1908).
2. An indirect route around the northern and western shores of the Indian Ocean via East Africa.

The first seems less likely because of the enormous distance (6400km) and the extremely hazardous sea and weather conditions in those parts of the ocean that would have had to be traversed (**Figure 1.3**). A Kon Tiki style expedition using a suitably styled outrigger canoe tested this possibility and took 49 days to make this crossing (Mack, 1986). This hypothesis is however not seriously considered because it is difficult to imagine a significant colonization of an island in this way. Also, none of the remote intervening islands of the Indian Ocean like Réunion and Mauritius showed evidence of occupation at the time of their discovery by Europeans (Murdoch, 1959) and this argues strongly against a direct crossing from the East to Madagascar. (**Figure 1.3**; also see notes on hypothesis 3 in **section 1.3.1**). The Negroid features among the Malagasy were explained by Grandidier (1908) by a hypothetical ancient direct migration of Negroid featured Melanesians which were later supplemented by the 650 000 slaves from Africa liberated in 1877 and 1896 (Campbell, 1989).

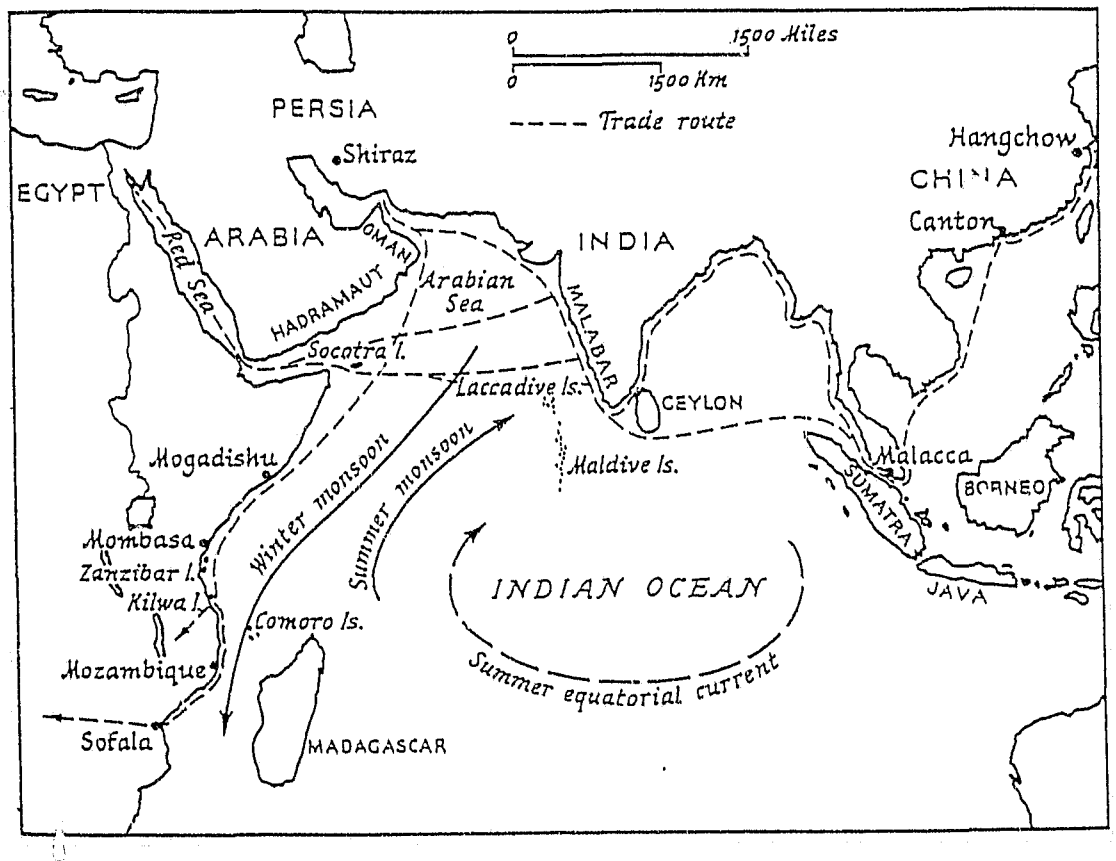


Figure 1.3. Map showing the medieval trade routes and the wind and ocean currents in the Indian Ocean (Oliver and Fagan, 1981).

The second route seems more feasible because it would have been made up of shorter land-hugging trips. Evidence that this route was used at some time is the presence of Asian food crops (bananas, coconuts, sugar-cane and taro), the xylophone, the outrigger canoe and sea-turtle hunting (Deschamps, 1965). The Indonesians taking this route could have inter-married to some extent with the local populations on the way and this would explain the presence of African racial and cultural elements in the basic sub-stratum of the entire Malagasy population. No archaeological evidence of Indonesian settlements have been found on East African shores but the Indonesian/Polynesian outrigger canoe is found in Ceylon, the Maldives, the East African coast and the west coast of Madagascar.

Furthermore Malayo-Polynesian people were known to be active maritime traders between the first and fifth centuries AD (Bellwood 1985, 1987; Stiles, 1988; **Figure 1.4**), which is when the earliest human activity on Madagascar ca. be traced to (Dewar and Wright, 1993). Arab and Chinese records indicate that Indonesians were trading with and living in East Africa and Madagascar (outposts confined to the northwest and west coasts of the island) during this period (**Figure 1.3**). The proposed coastal Indonesian settlements in Africa may have been displaced by the Bantu expansion precipitating the first permanent settlement of Madagascar. Those that remained northern East Africa may have been displaced around the ninth century by Muslim Arabs and Swahili traders dealing in gold, ivory, rhino horn and slaves (Stiles, 1988). This could have resulted in a second migration between the ninth and fourteenth centuries of Indonesians as settlers rather than traders from Somalia through to Mozambique and into the Comoros islands and northern Madagascar. This possible second wave could have completed the occupation of the coasts of Madagascar with settlements on the southern and the more hospitable eastern shores (Brown, 1978).

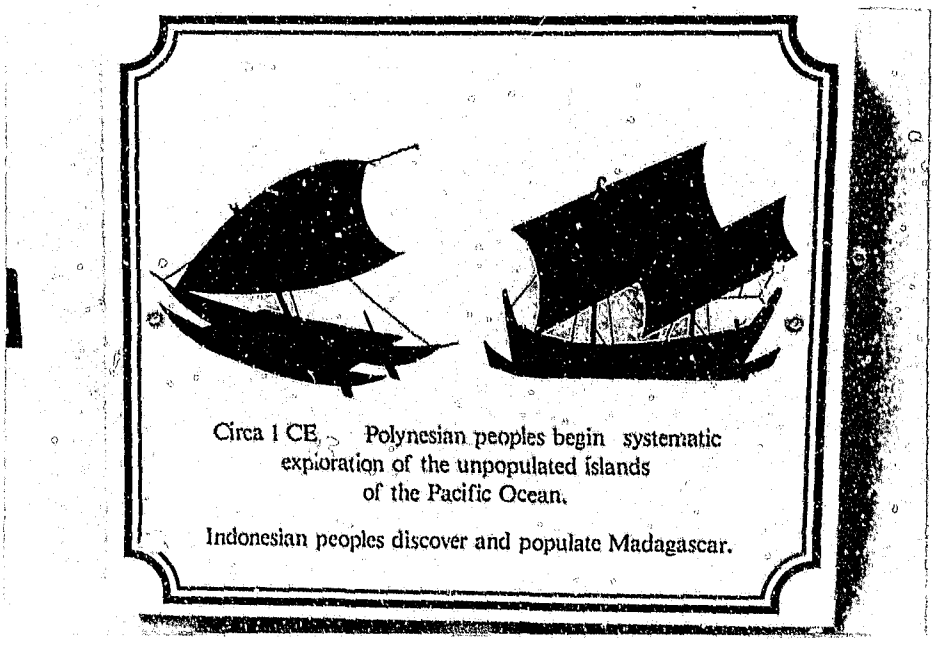


Figure 1.4. Photograph of a plaque outside the maritime museum in Perth. The legend below the pictures of the two sailing vessels clearly supports the colonization of Madagascar by Indonesian people. The outrigger canoe (vessel on the left) is still used in some parts of Madagascar.
(Photograph used with the kind permission of Dr H Soodyall)

The Malagasy could thus be an Indonesian-African mixture in the same way that the Swahili are Arab-African, and both would be products of commerce in the Indian Ocean. By the end of the first millennium, the coasts and some inland areas of Madagascar were probably settled by people who were a racial mixture of Indonesian and African people (Brown, 1978).

1.4 Malagasy ancestry – the evidence

1.4.1 Oral traditions

A number of sources of information about Malagasy history are the verbal traditions passed down by Merina rulers from the eleventh century AD. While no ancient monuments on Madagascar depicting the physical character or mode of living of past people have been recorded, island traditions speak of an early race of dwarfs. A belief held by the Antaimoro states that they descended from 30 men who sailed directly to Madagascar from Mecca during the fourteenth century (Campbell, 1995). However, there is no archaeological evidence or any evidence to suggest direct migrations to Madagascar from either Arabia or India. Arabs most probably migrated to Madagascar via east Africa and the Comoro islands, so that the history of Arab settlements on the latter are probably similar to those on Madagascar.

1.4.2 Cultural evidence

The entire Malagasy culture is essentially Malaysian but crops and domestic animals have been traced to homelands in SE Asia, Africa, the Near East and India. Swidden or slash and burn rice cultivation of dry rice (**Figures 1.5a** and **b**), crops such as taro and bananas,

rectangular wooden huts on stilts (**Figure 1.6**), the outrigger canoe (**Figure 1.7**), the blow gun, some musical instruments, notably the tube zither (**Figure 1.8**), iron smelting using the double tuyere bellows, second burial and the existence of certain taboos are all hallmarks of an Indonesian culture (Brown, 1978). The burial of dead in rectangular tombs decorated with specific designs is also of Indonesian origin (**Figures 1.9a and b**). It is also probable that the ancient SE Asian domesticates, the chicken and the pig, were introduced to Africa some time during the early part of the first millennium and subsequently to Madagascar by Indonesian settlers (Ehret, 1998).

The presence of the hump-backed zebu cattle (**Figures 1.10**) and the calabash-resonated cordophone (Mack, 1986) however, provides evidence of African culture on Madagascar. Cattle farming, which is the chief activity of the people of the vast western and southern plains of Madagascar (Sakalava, Bara and Antandroy), is also of mainland African origin (Brown, 1978). Although the culture of fibre weaving, the presence of musical instruments and certain funeral customs are present throughout the island, they have been modified by the combination of methods from both South-East Asia and Africa. The Arab influence on the East Coast of the island is thought to be pre-Islamic because of the lack of knowledge of the Moslem religion in this region. A somewhat larger and more recent Arab influence in the northwest and northeast of the island is supported by archaeological data (Brown, 1978).

(a)

(b)

Figure 1.5 (a). Bare hillsides on Madagascar's plateau which have been extensively terraced for rice cultivation. The valley bottom usually has flatter paddy fields. (Photograph used with the kind permission of Dr T de Ravel). **Figure 1.5(b)** shows typical terraced fields for rice cultivation in Java, SE Asia (Encarta, 1999).

Figure 1.6. The rectangular huts that exist throughout the island are reminiscent of Oceanic architecture with no obvious African parallel. These huts have a thatched roof and walls of wood, bamboo reeds or mats.

(Photograph used with the kind permission of Dr T de Ravel)

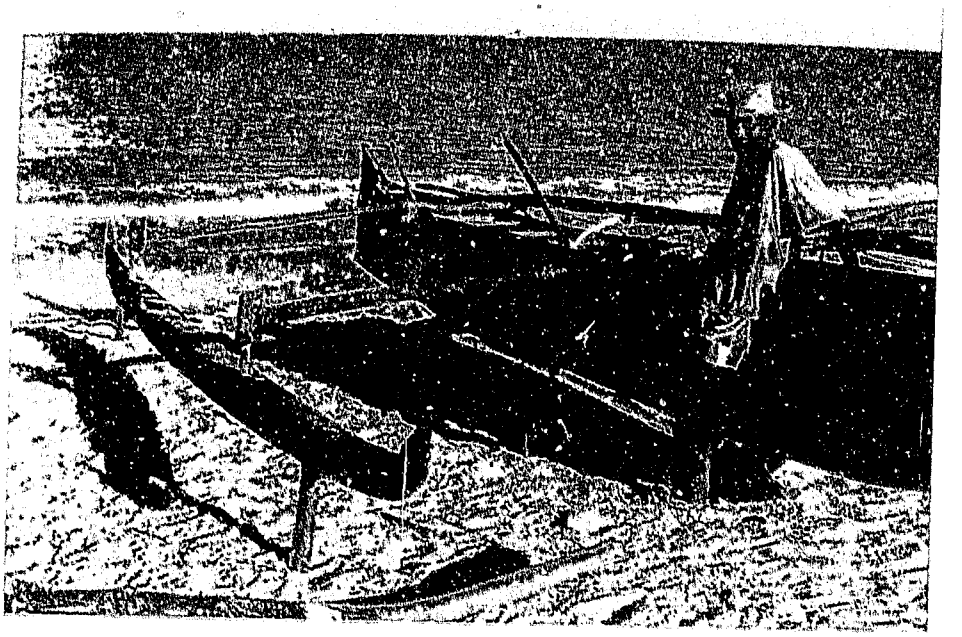
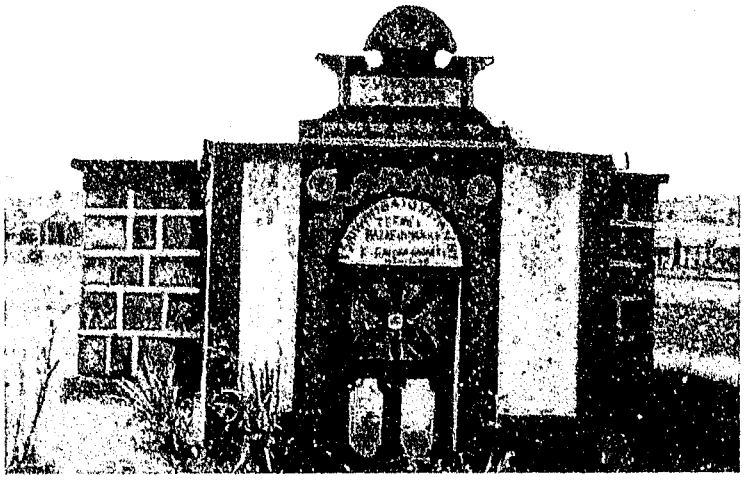


Figure 1.7. The outrigger canoe still used in fishing by the Vezo in southwestern Madagascar is an emblem of SE Asian culture (Schuurman, 1994).

Figure 1.8. The tube zither (valiha) of SE Asian derivation is used during social and religious events and is Madagascar's national instrument. It is often made from a wooden oblong box and bicycle-brake cables (Encarta, 1999).

(a)



(b)

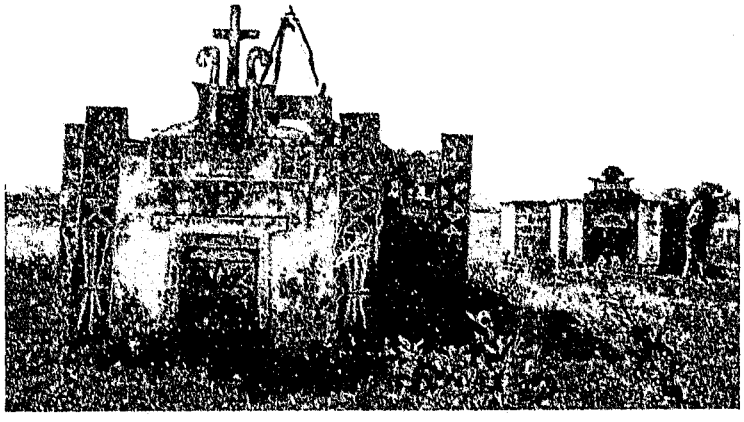


Figure 1.9(a) and (b) show the 'houses' of the dead or tombs; the doors and outside walls of which are frequently decorated with a mixture of floral and/or geometric designs.

(Photographs used with the kind permission of Dr T de Ravel)

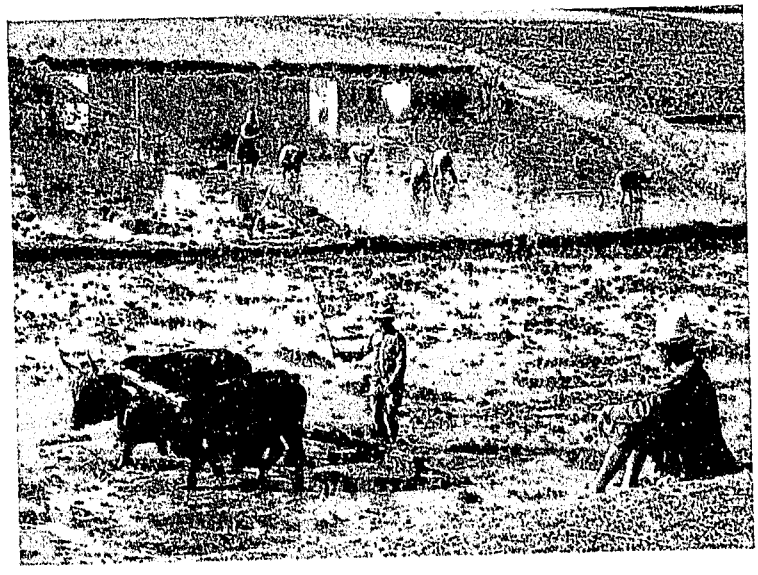


Figure 1.10. Zebu cattle working the paddy fields on Madagascar illustrate the merging of two different cultures.

(Photograph used with the kind permission of Dr T de Ravel)

The distribution of the material culture, however, does not always correlate with that of other cultural markers, e.g. the outrigger canoe which is undoubtedly of SE Asian or Oceanic origin is mostly found along the west coast (used by the Vezo) where the African influence (e.g. cattle-raising and the physical characteristics of the people) is said to be the strongest (**Figure 1.7**). The standard African dugout canoe that predominates elsewhere on the island is not common here. This may indicate that the Malagasy living in different regions of the island have adapted to their respective environments by selecting different available cultural traits. The fertile humid coastal belt of the east favored the development of agriculture, while the vast arid plains of the south and west were more suited to cattle herding (Brown, 1978; Mack, 1986). Another contrast in culture is shown in **Figure 1.10** as the merging of two cultures: rice cultivation by culturally African Malagasy (with the help of Zebu cattle in some areas).

Despite the varied cultural practices present in different Malagasy groups, a single characteristic or tradition draws the different communities together. This is termed fomban-drazana in Malagasy and involves the constant invoking of the ancestors (Mack, 1986).

1.4.3 Historical evidence

There is a paucity of written records documenting the history of the Malagasy, and those available are patchy and cover only very recent time periods. Europeans have attempted to elucidate the ancestry of the Malagasy since the fifteenth century, but have been hindered by this absence of written records. To date, three historic sources have been identified. The first includes accounts and reports written by traders and missionaries. These scripts date to

the middle of the last century when European missionaries taught writing. Grandidier and Grandidier (1908) suggested that the light-skinned Malagasy might be descendants of Indonesian settlers who consisted of the Vazimba, Mikea and a few other minor groups while the darker skinned Malagasy were the descendants of Melanesians. Other European writers (Ferrand, 1908; Verin, 1986) favored a mainly African ancestry from Swahili and Indonesian traders. Historical data also implies the presence of Pygmoid or 'short' people on the island but this hypothesis has been universally rejected based on archaeological data (Murdoch, 1959).

The second set of written records includes Malagasy texts written in Arabic script. Arab colonies were founded in the northwest of the island in the thirteenth century (Brown, 1978) and the ancient documents (the Sorabe) produced by these colonists deal with the history of an alleged Arabian ancestor of a tribe on the southeast coast (Singer *et al*, 1957). Documents compiled by the Jesuit priest Pere Callet and the Malagasy historian Raombana, provide the third set, but these mainly discuss central and SE Madagascar.

From the reports of Arab traders, Portuguese ship captains and early French explorers, it seems that a number of additional elements were added to the Malagasy population during the last millennium. First, the Arabs, settling in the Comoro islands and trading on both sides of the Mozambique Channel contributed to the population and second, the importation of slaves from Africa added an African strain to the population if one did not already exist. There is also historical evidence for the arrival of Indians, Jews, Persians and Chinese to the island during medieval times (Brown, 1978; Dewar, 1992). The culture of the Malagasy however may have been so well established or the disparity between their numbers and those of the immigrants so great, that they were able to assimilate these new

immigrants without any major effects on their language or culture. Because of the lack of detectable cultural and linguistic influences on the Malagasy by these latecomers, they are thought to have had no significant effect on the racial composition of the Malagasy (Brown, 1978).

1.4.4 Physical Appearance

The diversity in the appearance of the Malagasy peoples has been recognized and described for more than 300-400 years (**Figure 1.11**). Nagnort reported both black and white individuals on the island (Singer *et al* 1957) and Flacourt (1661) reported that the inhabitants of the southeast of the island were divided into two sorts- "les Blancs" and "les Noirs". Singer *et al* (1957) hypothesized on the basis of Malagasy physical characteristics – hair, lips, eyes, nose, skin color and prognathism that the Malagasy are about two-thirds African and one-third Indonesian. They suggested that the Merina were clearly different from the other groups, but still having a strong 'mixture' of African physical traits. In the coastal areas, especially in the west and south, the people have darker skins, short crinkly hair and Negroid features. Highland people such as the Merina on the other hand, generally have lighter skins, straight hair and resemble Malayans, Indonesians or Indo-Chinese (Mack, 1986). Physical features are however not always a very reliable measure for determining differences or similarities between populations and this will be evident from the genetic results generated in this study and is discussed in later sections.

Overall, the cultural practices and physical appearance of the different Malagasy groups are very similar and this could be an indication that there has not been enough time for the widely separated groups to develop into separate cultural identities.

Figure 1.11. A group of Malagasy children whose appearance suggests both African and Indonesian ancestry.

(Photograph used with the kind permission of Dr T de Ravel)

1.4.5 Linguistic evidence

Despite the reported variation in the physical characteristics of the Malagasy, all the subgroups speak dialects of the same language - Malagasy. Malagasy is a Malayo-Polynesian language, which belongs to the Austronesian language family. This language family comprises almost 800 languages which are collectively spoken by about 120 million individuals in countries from Madagascar in the west to Easter Island in the east, including the Malay peninsula, southern Vietnam, Indonesia, the Philippines and Taiwan. The Malagasy language thus provides important evidence of SE Asian ancestry. More than 93% of the basic Malagasy vocabulary is Malayo-Polynesian in origin but a small number of Bantu words are also used which is evidence of an African contribution to the Malagasy gene pool (Murdoch, 1959; Vérin *et al*, 1970). Previous genetic studies have also shown that people speaking Malayo-Polynesian languages have a close genetic affinity to each other (Melton *et al*, 1995 and Redd and Stoneking, 1999). This emphasizes the value of molecular markers in identifying the ancestors of the Malagasy.

Two minor exceptions to the uniformity of language spoken throughout the island have been noted: a few bilingual villages are known along the west coast where Makoa, a Bantu language of Mozambique is spoken by the families of recent immigrants. Bilingual Swahili speakers (a language spoken across most of central Africa) can also be found on the island of Nosy-be. A number of Arabic words have been absorbed into the Malagasy language via the Antaimoro people (Brown, 1978), and these include names for the days of the week and words connected with astrology, arithmetic and divining.

Despite the marked dialectical variation in language between the different ethnic groups of Madagascar (Brown, 1978), the dialects are reasonably mutually intelligible which suggests that the original settlers were a relatively unified cultural and linguistic group originating in SE Asia, regardless of where they were living immediately before they

settled Madagascar. Similarities between Malagasy and languages spoken in Indonesia have been recognized since 1603 (Vérin *et al*, 1970), but the actual identity of the Indonesian language from which Malagasy was derived has not yet been established. Studies by Dahl (1951) and Dyen (1953) suggest a close relationship between Merina Malagasy and both Maanyan and Ngaju which are spoken in the Barito River Valley of southeast Borneo. Dahl (1951) has hypothesized that the Proto-Malagasy may have left Borneo around AD 400. His evidence for this is based on the number of Sanskrit loan words in Malagasy and the dating of the beginning of Indian influence in Borneo. As there is no evidence of direct contact between India and Madagascar, the contact between the Proto-Malagasy and Indians must have occurred while the Proto-Malagasy were still in the Indonesian, and probably in the Borneo region.

Using glottochronology, Vérin *et al* (1970) have proposed that the different Malagasy dialects have been diverging for almost 2000 years. The original theory of glottochronology is based on the observation that a certain percentage of words in a language out of a basic word list would be dropped in a given number of years. Applying this principle to sixteen of the dialects spoken in Madagascar, Vérin *et al* (1970) concluded that a single group of people gave rise to the Malagasy in about 100 BC to about AD 225, and almost certainly not later than AD 400. Vérin *et al* (1970) assumed from previous studies that the Proto-Malagasy were a trading population who participated in a vast Indian Ocean trade network which tied Indonesia to regions both east and west of themselves. To the west, the Proto-Malagasy traders would have journeyed through Indian, Arabian and east African regions and eventually arrived on the coasts of Madagascar. Along the east African coast, the presumed predominantly male traders may have taken wives, which would result in cultural, linguistic and genetic admixture. The following generations, with

further contributions from African and Indonesian gene pools would display the tremendous phenotypic diversity resulting from this admixture.

Based on dendograms showing the affinities of the different Malagasy dialects, Vérin *et al* (1970) suggest that early in the history of the Malagasy, the ancestral group split into three populations which eventually gave rise to three linguistic groups of equivalent status; the Antankarana, the Tsimihety and all others. The proto-Antankarana and proto-Tsimihety settled in the north of Madagascar, while members of the third group established trading posts and settlements on the east and west coasts. Around AD 700, a western group began to lose contact with the eastern group and differentiated into a western-southern group as opposed to an eastern-central group. Some inhabitants of the west coast moved to the interior and became cattle herders. In about AD 1300, some eastern individuals may have moved to the highlands and started the irrigated cultivation of wet rice there. Resident pastoralists in this region may have been both absorbed and displaced into the arid zones of the south and west. The Sakalava and Vezo remained on the coast, the former becoming powerful and widespread along the west Coast while the latter remained along the southwest coast as fishermen. Inhabitants of the East Coast remained maritime traders with the Betsimisiraka forming a trade state by the eighteenth century. There is however no corroborating evidence to support specific internal migrations on the island.

The influence of Bantu languages on Malagasy is still unclear. It could have occurred when early Indonesians visited the coast of Africa as traders, or if there were early Bantu-speakers on the island, or it may have happened after African slaves were brought to the island either by Indonesians or Arabs.

1.4.6 Archaeological evidence

Although Madagascar has fascinated archaeologists for a long time, research has only recently begun on the island. No human fossils or prehistoric stone implements have ever been found there and the oldest well-dated remains of human activity are from the second half of the first millennium AD (Vérin, 1986). As soon as humans arrived, however, forest destruction, animal extinction and soil erosion began. Cut marks present on the leg bones of the now extinct pygmy hippopotamus are evidence of butchery by early inhabitants of Madagascar (Burney and MacPhee, 1988).

Unreliable and expensive dating techniques and Malagasy cultural resistance to the opening of tombs in order to examine human remains have hampered archaeological progress on the island. A dating program on the bones of Madagascar's subfossil megafauna suggests a connection between their extinction and the arrival of humans. The island also appears to have been visited several times before permanent residential bases were established (Dewar and Wright, 1993). Archaeologists have failed to find any Stone Age sites and the only stone tools found on the island are net sinkers, weights and gunflints. Archaeological evidence (reviewed in Dewar and Wright, 1993) also indicates human occupation along the southwest, north and eastern coasts in the first millennium AD, but at present, there is no direct evidence of occupation of the interior for this period. A rock shelter in the extreme north (Lakaton'I Anja) discovered in 1986 provides the earliest radiocarbon date of man's presence on the island. Pottery, iron blades, stone weights (probably from nets used to snare birds) and animal bones (including some of marine animals) with cut marks were found and the earliest estimated occupation date is about AD 700.

Trading sites along the East African coast from Mozambique to the Persian Gulf are known to have been in existence since the seventh century AD. By the tenth century, food vessels carved of soapstone-like rock and apparently manufactured on the northeast coast of Madagascar are commonly encountered at these sites. Irodo, a site on the northeast coast of the island not far from the source of the stone food vessels, may have been occupied in the ninth century while agricultural villages and cattle-herding settlements were present in south and southeastern Madagascar by the eleventh century. Pieces of Middle Eastern ceramics and glassware and fragments of late Song Dynasty Chinese ceramics found along the northeast coast of Madagascar can be dated to the eleventh century and suggest the presence there of Arab settlements (Cole, 1992; Dewar and Wright, 1993). The first stone built mosque on the island also dates to around the thirteenth century and provides evidence for the presence of the Islamic faith by this time. From the presence of various tradewares, it seems clear that the island has long been a part of a western Indian Ocean trade network. The first known permanent occupation of the central highlands dates to the fourteenth century (Dewar and Wright, 1993). Studies also show that econ. mic activities were well in place at this stage, e.g. rice growing, bovid herding, fishing, iron smelting and long-distance trade was practiced. Between the fourteenth and sixteenth centuries, a relatively uniform cultural pattern existed on the island

Thus, the first definite traces of human settlement in the seventh century are followed fairly quickly by evidence that the island was widely populated and well integrated into the East African coastal trade network (Dewar and Wright, 1993).

1.4.7 Genetic evidence

Published genetic evidence pertaining to Malagasy ancestry is sparse; only a limited set of genetic markers have been used on relatively small numbers of individuals. In addition, many of the Malagasy ethnic groups have been poorly studied or not studied at all. Early studies by Singer *et al* (1957) and Buettner-Janusch and Buettner-Janusch (1964, 1970) are consistent with a notion that there was a significant African contribution to the founding Malagasy population. Rhesus blood group data clearly shows the presence of a typical African marker (cDe) in the Malagasy which suggests that the African contribution is greater than the Indonesian [Singer *et al* (1957) claimed that the admixture was 2 parts Bantu to 1 part Indonesian]. These studies also showed statistically significant differences in the frequencies of the Hb S (higher in lowland areas) and particular haptoglobin alleles between highland and lowland groups. The frequency of haptoglobin allele 1 in the various coastal and lowland groups is within the range reported for African tribes (Buettner-Janusch and Buettner-Janusch 1964) which suggests strong affinities of these groups with African populations.

At least two explanations are possible for the difference in frequency of the Hb S allele in coastal dwellers and highlanders. The first could be the effect of natural selection in lowland areas because of the coastal areas being highly malarial and the plateau much less so. The second could be that the highland groups, especially the Merina may have maintained strict rules against marriages with the coastal tribes and the descendants of slaves imported from Africa.

The high frequency of the cDe (Ro) Rh haplotype together with the presence of the gene for sickle cell anemia in the Malagasy prompted Singer *et al* (1957) to suggest that the sickle cell gene was brought to Madagascar from Africa and not directly from India. The sickle cell allele is absent . SE Asia. Shorthorn Zebu cattle together with a great number of slaves appear to have been introduced to Madagascar from Africa in about the seventh century AD (Campbell, 198 1989; Mourant *et al*, 1976) and the sickle cell allele could have been introduced at the same time.

Haplotype studies suggest that the sickle cell mutation arose independently in four different regions of Africa and once in Arabia-India (Oner *et al*, 1992; Lapouméroulie *et al*, 1992). The 'Bantu' haplotype, noted only in Bantu-speaking Africa was found on 25 of 28 β^S containing Malagasy chromosomes studied (Hewitt *et al*, 1996). This provides the first definitive evidence that a major component of the island's ancestry is derived specifically from Bantu-speaking Negroids. Since the β^S gene is absent from Indonesia, it would be an appropriate indicator of African gene flow. The apparent absence of the Arabic-Indian β^S haplotype suggests that there was very little or no contribution from this region to the Madagascar.

The mitochondrial genome is maternally inherited and does not undergo recombination. It has been used widely to trace female lineages through evolution. Sequence specific oligonucleotide (SSO) typing in the mtDNA control region (Stoneking *et al*, 1991) has delineated a number of haplotypes for use in evolutionary studies. Soodyall *et al* (1995) showed that a polymorphic 9bp deletion in mitochondria from African peoples occurs on a different haplotypic background to that found among Asians. Using this information, they

have estimated that 70.7% of Malagasy mtDNAs with the 9bp deletion (frequency of 9bp deletion in the Malagasy is 26.8%) are derived from an Asian source, while 29.3% are derived from an African source. One specific haplotypic background of the 9bp deletion is common in Asia and is referred to as the 'Polynesian motif' since it occurs at high frequencies in both ancient and contemporary Polynesians. This motif occurs at very low frequencies in the Indonesian population, and suggests that the original colonists who settled Madagascar may have come directly from Polynesia or that they share a common ancestry with Polynesians, presumably from Indonesia.

Variation at seven RFLP loci in the glucose-6-phosphate dehydrogenase (G6PD) gene has also been exploited to study Malagasy ancestry. An allele at this locus, the A- variant, is associated with partial enzyme deficiency and is found almost exclusively in African populations. Like the sickle cell allele, the A- allele is also thought to protect against malaria. This A- allele is associated with a particular seven-site RFLP haplotype and this A- 'haplotype' has been found at high frequencies in G6PD deficient Malagasy (Dangerfield, 1997). Haplotypes associated with G6PD deficiency in the Mediterranean region have also been detected in the Malagasy, suggesting a Middle Eastern contribution. The haplotype associated with the deficiency in SE Asia was not detected in the study, but this could be due to the limited sample size.

1.4.8 Summary

The following general conclusions regarding Malagasy prehistory can be drawn:

- Permanent settlement of the island occurred prior to the spread of Islam and probably some time after the seventh century AD.
- There is great diversity among the inhabitants within all geographic regions with respect to culture and physical characteristics which may be attributable to varying degrees of Indonesian and African influence. At present, there is very little evidence supporting the presence of Indian, Jewish, Persian or Chinese populations on the island.
- Linguistic evidence indicates a strong Indonesian contribution with minor contributions from African and Arabic sources.
- Genetic data are concordant with major migrations from Indonesia and Africa but not from Melanesia.

1.5 Genetic tools used for exploring population affinities

The human genome has less than five percent of unique coding sequence and more than 95% of non-coding sequences. Seventy five percent of the latter consists of unique or single copy sequences, the function/s of which are not known. Ten to fifteen percent of the remainder consists of tandemly repeated DNA sequences that codes for ribosomal and transfer RNAs, some of which are clustered around the centromeres of certain chromosomes. The other 10-15% of the DNA is made up of two main classes of repetitive DNA sequences that are interspersed throughout the genome; short interspersed sequences (SINES) and long interspersed sequences (LINES). Approximately 300 000 copies of

SINES and 100 000 or so copies of LINES occur in the human genome. *Alu* elements belong to the SINE family. They contain a recognition site for the restriction enzyme *Alu*I and consist of roughly 300 base pairs of sequence. LINES on the other hand, consist of a slightly longer DNA sequence of up to 6 000 base pairs in length.

1.5.1 Polymorphisms

This term can be defined as 'the occurrence together in a population of two or more genetically determined forms (alleles) at a particular locus in such frequencies that the rarest of them could not be maintained by recurrent mutation alone'. Alleles that occur at frequencies of less than one percent are referred to as rare variants. Polymorphisms have been found throughout the human genome, both in coding and non-coding regions are valuable as genetic markers. It is generally presumed that there is no selection for or against polymorphisms in non-coding or 'neutral' DNA. Allele frequencies at many loci in the human genome vary widely among world populations. This variation is considered to be the overall consequence of mutation, genetic drift and natural selection.

DNA polymorphisms have been widely used in the fields of medicine, forensics, gene mapping, population genetics and evolutionary genetics. They have played a role in distinguishing the major human population groups of the world and have helped to resolve uncertainties that result from different interpretations of paleoanthropological data. Many types of polymorphisms, autosomal, mitochondrial as well as those on the X and Y chromosomes are currently being used to answer questions about human ancestry, the colonization of the continents and the 'ancestry' of disease genes. The following sections discuss some examples.

1.5.2 Types of DNA Polymorphisms

Genetic polymorphisms can exist in the form of single nucleotide substitutions or can involve more than one nucleotide in the form of insertions and deletions. Tandemly repeated DNA sequences or variable number of tandem repeats (VNTRs) usually exhibit high degrees of variability with regard to the number and size of repeat copies present. Loci with repeat motifs of a few base pairs (e.g. 2-5 bps) are referred to as microsatellites or simple sequence length polymorphisms (Tautz, 1989), while those with longer motifs are referred to as minisatellites (8-100bp) (Jeffreys *et al*, 1985) or satellites (>100bps).

1.5.2.1 Restriction fragment length polymorphisms

Restriction fragment length polymorphisms (RFLPs) are polymorphisms which can be detected through the use of restriction enzymes. These polymorphisms are detectable as changes in the fragment length resulting from the action of a particular restriction enzyme on a target sequence. In the absence of selection, a newly generated restriction site resulting from a base pair substitution is unlikely to reach polymorphic frequencies more than once in the lifetime of a species. Thus, a specific base pair variant is likely to be identical by descent in whatever population it may be found. Unfortunately, the bi-allelic nature of most RFLPs makes them less informative than multiallelic markers. A number of studies have shown that the majority of known RFLPs are polymorphic in all populations studied to date, albeit at different frequencies (Bowcock *et al*, 1987; Cavalli-Sforza *et al*, 1986). The difference in allele frequencies may be a reflection of the history of the population in terms of drift, migrations and admixture.

1.5.2.2 Short tandem repeat polymorphisms

Short tandem repeat (STR) polymorphisms or microsatellites result from differences in the number of tandemly repeated DNA sequences at different loci in the human genome. The repeat DNA sequences are largely transcriptionally inactive, i.e. occurring in noncoding regions. Mutation and drift both influence the frequencies of alleles in a population with the latter in turn being related to the history of growth and subdivision of the population. Weber and Wong (1993) and Di Rienzo *et al* (1994) examined the pattern of mutability of autosomal STRs during parent-child transmissions and concluded that tetranucleotide repeats mutate at a rate higher than dinucleotide repeats, and that changes in allele sizes of one repeat unit occur more frequently than larger changes. This is consistent with the 'stepwise mutation model' (SMM) first proposed by Ohta and Kimura (1973) to explain the extent of protein polymorphisms. Simulation studies carried out more recently support the SMM for STR loci [Shriver *et al*, 1993; Di Rienzo *et al*, 1994; Slatkin, 1995 and Goldstein, 1995).

Weber and Wong (1993) estimated the average mutation rates at STR loci to be approximately 1.2×10^{-3}. New length alleles are thought to be usually generated by polymerase slippage during replication. Other estimates for the mutation rates at STR loci range from 1.5×10^{-3} to 2×10^{-5} (Edwards *et al*, 1992; Bruford and Wayne, 1993 and Jin *et al*, 1994) suggesting that each STR may have a unique rate of mutation. This rate could be influenced by the size of the repeat unit, the total number of repeats and sequence interruptions within the repeat units.

Dinucleotide repeats occur at intervals of 50-100 kilobases and tri- and tetrameric repeats at intervals of 300-500 kilobases. Because of their high variability (heterozygosities range from approximately 34%-90%) and ease of allele identification using PCR, microsatellites are very popular for mapping genes, for personal identification (Edwards *et al*, 1992; Hammond *et al*, 1994), and for inferring phylogenetic relationships among human populations (Bowcock *et al*, 1994; Di Rienzo *et al*, 1994, Deka *et al*, 1995; Watkins *et al*, 1995; Jorde *et al*, 1995). The small size of STRs also makes it multiplex PCR feasible so that two or more loci can be simultaneously amplified from a single DNA sample (Chamberlain *et al*, 1988). This decreases the cost and time of determining genotypes. Studies by Edwards *et al* (1992) indicate that genotype data from trimeric and tetrameric STR systems conform to Hardy-Weinberg expectation. Furthermore, these markers tend to be highly polymorphic and have mutation rates which are between those reported for protein and VNTR loci.

1.5.2.3 *Alu* polymorphisms

Alu elements belong to a family of SINEs which occur in hundreds of thousands of copies dispersed throughout the primate genome (Weiner *et al*, 1986; Batzer *et al*, 1991). The *Alu* family consists of DNA sequences that are similar to each other and are flanked by direct repeats. *Alu* sequences represent a highly successful class of mobile genetic elements that have retroposed from a few 'master' genes via a RNA intermediate (Rogers, 1983). These elements which are derived from the 7SL RNA gene have multiplied within the human genome during the last 65 million years to a copy number in excess of 500 000. *Alu* sequences are approximately 300 base pairs long and the vast majority are transcriptionally

and transpositionally silent. Once inserted at specific chromosomal locations, most *Alu* elements are not subject to loss or rearrangement, making them stable genetic markers.

Alu sequences may be grouped on the basis of their sequences. One of the most recently formed *Alu* groups is the predicted variant (PV) (Matera *et al*, 1990) or human-specific (HS) group (Batzer *et al*, 1990; 1991), which has between 500 and 2 000 members (Batzer *et al*, 1990; Batzer and Deininger, 1991). Some members of the group have retroposed so recently that they have not become fixed in the human species (Batzer *et al*, 1991). They thus represent an ongoing evolutionary process in the human genome and are a unique source of DNA variation for human population studies. Some useful features of *Alu* elements for reconstructing historical population events are that all loci carrying a particular polymorphic *Alu* insertion are derived from a unique event and hence are identical by descent. Secondly, the ancestral state at each locus where an *Alu* polymorphism exists in man can be reasonably inferred by comparison of the site in other species, i.e. the direction of the mutational change, whether an insertion event took place after the emergence of modern humans or whether an *Alu* sequence was lost since then can be inferred. This is not possible for other types of markers such as RFLPs and VNTR loci.

Alu polymorphisms can be typed by PCR, which facilitates rapid screening of large numbers of samples. Studies on four HS *Alu* polymorphisms by Batzer *et al* (1994) in 34 worldwide populations reveal substantial levels of variation both within and among populations. Stoneking *et al* (1996) have also used *Alu* polymorphisms to study the genetic variation in world populations. They found that African populations exhibit the most between population differentiation and that phylogenetic trees constructed from these data were also rooted in Africa.

1.5.2.4 Haplotype variation

A DNA haplotype can be broadly defined as a particular combination of alleles at syntenic loci. Each combination of alleles is usually transmitted intact through families and can, for convenience, be treated in the same way as an allele at a single locus. The number of possible haplotypes is the product of the numbers of alleles at each of the polymorphic loci. Therefore, the larger the number of alleles at each of the loci constituting the haplotype, the larger is the possible number of combinations or haplotypes. For this reason, the inclusion of STR loci (which usually have many alleles) in haplotypes is advantageous.

Syntenic RFLPs have often been used to define nuclear DNA haplotypes (Hill *et al*, 1989; Ramsay and Jenkins, 1988; Wainscoat *et al*, 1986) and these in turn have been used in phylogenetic studies (Kidd and Kidd, 1996). Recombination, though rare in short stretches of DNA, does occur and may lead to the formation of new haplotypes. Over time, this and mutation can change the diversity and frequencies of haplotypes in populations. The frequencies of these 'newly' formed private and semi-private haplotypes in different populations can provide useful historical and demographic information of a population and may also be used to detect low levels of population admixture (Tishkoff *et al*, 1996; Kidd *et al*, 1998).

The approximate distance between two loci can be measured by the rate of recombination (crossovers) between them and is expressed in centimorgans (cM) or map units (1 map unit ≅ 1cM ≅ 1% recombination ≅ 1Mb). There is very little linkage disequilibrium between genes that are more than 10 map units apart. Alleles at closely linked loci (recombination fraction under 0.5%) however, are often in linkage disequilibrium and are most suitable for

tracing the histories of populations. The age of a particular combination as well as population admixture has a bearing on the extent of the linkage disequilibrium.

1.6 The Y chromosome

The non-recombining portion of the human Y chromosome (NRPY) is transmitted from father to son as a haploid entity without recombination. The variation noted on this chromosome both within and among populations is thus a record of the mutational events which have occurred in male lineages over time. In this context the Y chromosome is the male counterpart of what mtDNA represents for female lineages. These observations make the study of Y chromosome variation highly valuable for describing the evolutionary sequences in a system in which mutation is the sole source of genetic change. In addition, the analyses of worldwide Y chromosome variation may add significantly to the understanding of the radiation of human male lineages.

In the recent past, the use of human Y-chromosome variation was restricted by the paucity of common polymorphisms (Jakubiczka *et al*, 1989; Malaspina *et al*, 1990; Spurdle and Jenkins, 1992a) and the lack of knowledge of the nature of some of these. Various studies (Hammer, 1994; Dorit *et al*, 1995; Whitfield *et al*, 1995 and Underhill *et al*, 1996) show very little or no sequence variation within the sequence of interest in the samples of individuals screened. Three reasons have been given for this reduced diversity (Jobling and Tyler-Smith, 1995):

- A smaller effective population size of the Y chromosome in relation to the size of the autosome population.

- The possibility of only a few males contributing the majority of the Y-chromosomes to the next generation in a population.
- Selective sweeps caused by the selection of advantageous mutations resulting in the rest of the chromosome 'hitchhiking' with the selected allele/mutation. Variants at other loci are lost because of the lack of recombination on this chromosome.

In spite of this apparent low nucleotide diversity, a number of unique mutations have been used to construct haplotypes to trace the radiation and evolution of the Y chromosome itself (Oakey and Tyler-Smith, 1990; Mathias *et al*, 1994; Hammer *et al*, 1997; Hammer *et al*, 1998; Underhill *et al*, 1997). Many bi-allelic Y chromosome markers show large frequency differences between populations from different geographic regions (Underhill *et al*, 1997). The Y *Alu* polymorphism (YAP) (Hammer, 1994), is an example of a very useful insertion polymorphism resulting from a recent insertion of an *Alu* element at Yq11. As described in **section 1.5.2.3**, this is usually a unique event and all chromosomes with this polymorphism stem from the same ancestral chromosome. The frequency of YAP varies in populations worldwide, being highest in sub-Saharan African populations, followed by northern Africans and Europeans. The insertion is rare or absent in Asian and Oceanic populations with the exception of a few east Asian populations viz. the Tibetans and Japanese (Hammer and Horai, 1995).

Recently a number of additional markers have been identified on the Y chromosome, including single base pair variants, insertions, deletions and STR polymorphisms (Jobling and Tyler-Smith, 1995). The individual markers have different mutation rates and depending on the chosen system, different evolutionary time scales can be explored. Haplotypes which include both high and low mutation rate polymorphisms would be most

informative in identifying both individual as well as groups of Y chromosomes (Hammer *et al*, 1997; Scozzari *et al*, 1997 and Skorecki *et al* 1997, Kittles *et al*, 1998, Hammer *et al*, 1998; Karafet *et al*, 1999; Scozzari *et al*, 1999). The YAP polymorphism for instance, together with several other bi-allelic and multi-allelic polymorphisms on the Y chromosome, has been used to study males of geographically diverse origin (Jobling and Tyler Smith, 1995; Hammer *et al*, 1997). These studies indicate that while some haplotypes may be shared between populations, many have a number of unique haplotypes.

A series of highly polymorphic Y-specific microsatellites have been characterized and tested on different population samples (Mathias *et al*, 1994; Roewer *et al*, 1992; Roewer *et al*, 1996 and Cooper *et al*, 1996). These markers show high levels of heterogeneity within and between populations and thus seem to be very useful for population genetic, evolutionary and forensic applications. They are particularly suitable for discriminating between closely related groups (Roewer *et al*, 1996). de Knijff *et al* (1997) have tested seven such markers on a selected number of population samples to assess the reliability of Y STRs for evolutionary and population genetic studies. The distribution of alleles at these loci can be explained in terms of the stepwise mutation model (SMM) which was initially developed for autosomal STRs (Weber and Wong, 1993; Di Rienzo *et al* 1994; Ciminelli *et al* 1995). In the absence of recombination, Y STR allele distribution can be explained by a majority of mutational events involving small changes (increases or decreases of one repeat unit) in allele size and only a few events involving large changes in size. Data collected by de Knijff *et al* (1997) suggest that because of the high mutation rate at Y STR loci (as is the case with their autosomal counterparts), more stable Y-polymorphisms e.g. base-substitutions, or *Alu* insertion/deletion polymorphisms in combination with the Y STRs

would be more useful and reliable for following evolutionary events that occurred more than 50 000 years ago.

Mathias *et al* (1994) observed that the same tandem repeat allele is often found on different Y chromosome backgrounds. This can be attributed to recurrent mutation along different lineages during Y chromosome radiation. Analysis of three STR polymorphisms of the male-specific portion of the Y chromosome by Ciminelli *et al* (1995) shows that partial haplotype randomization is common. Their studies also indicate that the Y chromosome diverges more rapidly than autosomes (Persichetti *et al*, 1992). Locus DYS19 is the most frequently used Y STR at present, with more than 4 500 males screened worldwide (Santos *et al*, 1996; de Knijff *et al*, 1997).

1.7 Anthropological studies

The use of inter-population genetic variation has greatly expanded anthropological research. Gene flow following migration is likely to have been a significant factor in human population diversification. As a consequence of different social patterns around the world, the migration may involve predominantly males, predominantly females or mixed groups. Therefore, an overall picture of the contribution to the gene pool of any population can only be obtained by combining data from autosomal, Y chromosome and mtDNA loci. Analysis of allele frequencies in different populations permits the construction of genetic distance maps (principal components analysis; Cavalli-Sforza *et al*, 1994) which show the genetic relationships between populations. Information about past migrations can be obtained by tracing genes from parental populations to newly formed populations. When following region or population specific markers, this information is more useful.

Unfortunately, applications of this sort are sometimes hampered by the fact that non-African populations most often have a subset of the variation seen in Africans. The widely accepted 'Out of Africa' theory proposed for human evolution may explain this observation. This theory proposes that modern humans evolved in Africa from an anatomically modern H.sapiens ancestor about 100 000 to 200 000 years ago and subsequently spread to the rest of the world replacing all other hominids (Stringer and Andrews, 1988). This migration may have occurred in a single wave or in multiple waves (Lahr and Foley, 1998). The alternate theory proposed is the 'multiregional continuum' theory, which postulates that there was no single origin for modern humans (Wolpoff *et al*, 1989). After the radiation of H.erectus to Europe and Asia 800 000 to 1.8 million years ago (Stringer and Andrews, 1988), modern humans evolved gradually and simultaneously from local forms of H.erectus.

The first model is supported by fossil records (Stringer and Andrews, 1988) which have also shown the existence of modern humans both before and after the existence of the Neanderthals. Phylogenetic trees based on classical blood protein markers (Cavalli-Sforza *et al*, 1988; Nei and Roychoudary, 1993), nuclear DNA RFLP markers (Bowcock *et al*, 1991; Wainscoat *et al*, 1986), *Alu* polymorphisms (Batzer *et al*, 1994), microsatellite loci (Bowcock *et al*, 1994), Y chromosome haplotypes (Hammer *et al*, 1997) and mtDNA (Cann *et al*, 1987; Merriweather *et al*, 1991; Vigilant *et al*, 1991; Horai *et al*, 1995; Penny *et al*, 1995) also support an Africa origin of modern humans. The 'Out of Africa' hypothesis also gains support from the fact that mtDNA sequence diversities (Cann *et al*, 1987; Vigilant *et al*, 1991; Watson *et al*, 1996) as well as allelic diversities at STR loci (Deka *et al*, 1995) tend to be higher in African populations than in non-African populations.

1.8 Aims of this study

Inferences about Malagasy ancestry have been made from historical, archaeological, linguistic and genetic studies, but as can be deduced from previous sections, much remains unresolved. The major aim of the present study was to use additional genetic markers to cast more light on the prehistory of the Malagasy. The following approaches were adopted:

1. Autosomal and Y chromosomal markers were used to estimate the extent of genetic variation in the Malagasy. The same markers were also examined in African and non-African populations for comparison.
2. Autosomal and Y chromosome haplotypes were constructed with the aim of identifying marker haplotypes characteristic of African (especially Bantu and Pygmy populations), Melanesian, Indonesian/SE Asian and Indian populations.
3. Phylogenetic trees were constructed and principal component analyses carried out to gain more information about the genetic relationships between the various population groups.
4. Statistical methods were used to examine the genetic substructure of the Malagasy populations.
5. Networks constructed using Y chromosome haplotypes were used to possibly gain insight on the demographic history of the Malagasy.

CHAPTER 2 - SUBJECTS AND METHODS

2.1 Subjects

To achieve the objectives of this study as outlined in **section 1.8**, the following population groups were examined:

- The Malagasy, represented by eleven of eighteen ethnic groups from different geographic regions in Madagascar,
- Non-Africans represented by an Indian population from South Africa (SA Indian); Austronesians from Indonesia and Polynesia and Melanesians from Papua New Guinea (PNG) and the Solomon Islands,
- Africans represented by samples from Zambia, South Africa [North Sotho (Pedi) and Tsonga] and the Central African Republic [Beya (Gbaya, Mbimou and S.Sangha) and Mbenzele].

Table 2.1 is a list of the populations studied, the corresponding sample sizes and the polymorphic systems typed. Different colors have been used to represent the Malagasy (blue), African (green) and non-African (red) populations in subsequent tables and figures.

Table 2.1. A list of the population groups examined and the sample sizes used to screen for the various polymorphisms.

Population	Autosomal markers			Y linked markers
	STRs	CD4-*Alu*	DRD2	
		2 site haplotype	4 site haplotype	5 site haplotype
Malagasy				
Merina (MR)	37	35	34	49
Betsileo (BT)	36	36	35	49
Antaisaka (AS)	37	36	32	50
Tsimihety (TS)	35	35	35	50
Sakalava(SK)	42	34	37	51
Mahafaly (MF)	34	33	32	27
Antandroy (AD)	35	33	34	44
Betsimisiraka (BS)	38	38	35	51
Antaimoro (AM)	37	36	35	50
Antankarana (AK)	16	16	16	13
Mikea (MK)	50	50	47	33
Non-African				
SA Indian	39	39	39	50
Indonesian (SS)	44	43		
Indonesian (SK)	36	24		
Polynesian	37			
New Guinean	20	20		
Solomom islands	23	23		
African				
Tsonga	35	39	50	50
North Sotho (Pedi)			50	
Zambian	53	50	48	50
Mbenzele	44	44	43	20
Gbaya	35	35	33	30
Mbimou/SSangha	28	28	28	17
N	791	727	663	684

SS=South Sulawesi; SK=South Kalimantan

The approximate geographic locations of the various groups examined are shown in **Figure 2.1**. All samples selected for study are from unrelated individuals collected with their informed consent. Approval for these studies was obtained from the Ethics Committee of the University of the Witwatersrand (ethics clearance certificate numbers 11/6/91, 2/1/95, 25/1/95 and 6/6/95). Whenever possible, at least 70 chromosomes were examined for autosomal markers while at least 50 were selected for Y chromosome studies. For the former, individuals were assumed to belong to a particular group if both their parents belonged to the same group. For Y chromosome studies, only the ethnicity of the father was considered. Haplotype data for Y chromosome analysis from a number of SE Asian/Oceanic populations were kindly provided by Dr Alan Redd of Penn State University (personal communications, 1998). For comparative purposes and where possible, data from the literature have been utilized and is referenced appropriately.

2.1.1 Malagasy populations

Samples representing the Malagasy ethnic groups (see **Figure 1.2** for their location on the island) were collected over a period of five years during field trips to different parts of the island as part of a larger study on the origins of the Malagasy. Each group speaks a different dialect of a common language, Malagasy, which belongs to the Austronesian group of languages. Eleven groups from different regions of the island have been selected for the present study.

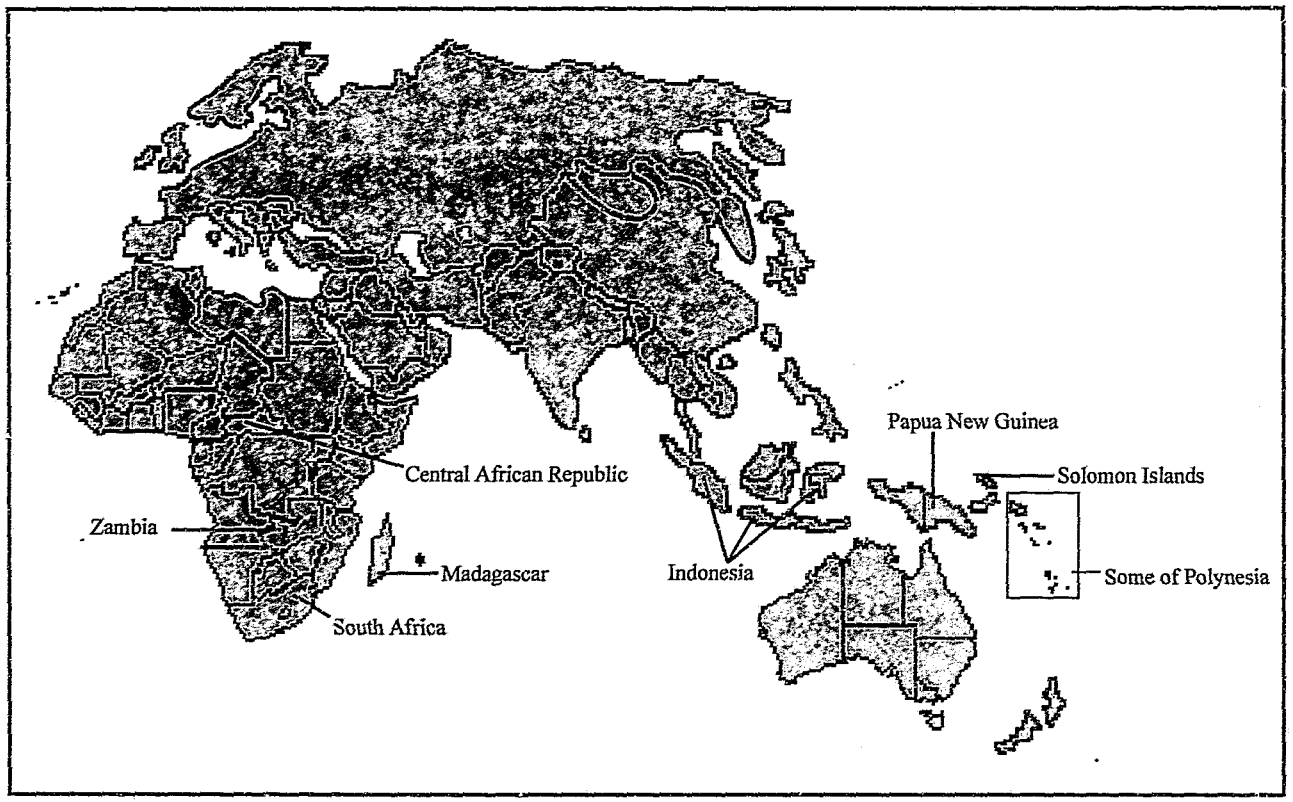


Figure 2.1. Present geographic location of population groups examined in this study.

2.1.2 Non-African populations

2.1.2.1 The South African Indians

The South African Indians (SA Indians) are descendants of immigrants who arrived from India to South Africa between 1860 and 1911 and their origins have been traced to a number of discrete regions in India (Bhana and Brain, 1990). They have been divided into four major religious and/or linguistic subgroups (Grierson, 1927): Moslem Gujerati, Hindu Gujerati, Hindu Hindi and Hindu Tamil. The Tamil subgroup speaks a Dravidian language and the other groups speak an Indo-Aryan language (Mistry, 1965).

2.1.2.2 The Austronesians

(a) Indonesians

Three sets of samples were obtained from different regions in Indonesia: south Sulawesi (Ujung Pandang), Banjamarsin (southern Kalimantan) and a mixed sample from Alor, Medan, Sumba, Bangka, Ambon and Palembang. Austronesian languages predominate in this region of SE Asia.

(b) Polynesians

The Polynesian sample was collected from individuals who speak a language belonging to the Austronesian language family. The sample consists of DNA from Maori, Tongan, Samoan and Rarotongan individuals, mainly from the islands of Tonga and Samoa.

2.1.2.3 The Melanesians

Two samples from Melanesia were included in the study. The first consisted of individuals from Bougainville (Nasioi) in the Solomon Islands and the second was collected from Gimi and Goroka in Papua New Guinea. Both groups speak an Indo-Pacific language.

2.1.3 African populations

Africa is inhabited by three major groups of people; Caucasoids in the north and Negroids and Khoisan in the sub-Saharan region. These three groups have been further sub-divided on the basis of language, culture and history. Greenberg (1963) assigned each of the more than 800 languages spoken in Africa to one of four families; Afroasiatic, Nilo-Saharan, Niger-Kordofanian (including Bantu and Ubangi) and Khoisan ("click"). The four African samples used in this study fall into the Niger-Kordofanian family and are the following:

2.1.3.1 The Tsonga and North Sotho (Pedi)

The Tsonga and North Sotho represent the eastern Bantu of southern Africa but speak different Bantu languages. The Tsonga sample includes individuals who claim to be either Tsonga or Shangaan, but these two names are often used interchangeably.

2.1.3.2 Zambians

The Zambian sample consists of ethnically heterogeneous Bantu-speakers from central Africa and was collected from blood donors and women at an antenatal clinic. The individuals from whom samples were used belong to the Bemba, Chewa, Tonga, Lozi and Ngoni chiefdoms.

2.1.3.3 Mbenzele

The Mbenzele are pygmies from the Central African Republic (CAR) and like the well-studied Biaka or Aka pygmies, they have the characteristic small stature that is believed to be an adaptation to the warm and humid tropical forests. There is no record of the original language spoken by the Pygmies and thus they cannot be classified on the basis of language. The Biaka, or western Pygmies, live in a forest encompassing the area where the CAR, east Cameroon and the northern part of the Republic of Congo meet; they speak a Bantu language belonging to the Niger-Kordofanian family. The Mbenzele have been culturally and geographically isolated from the Biaka and the sample used in this study was collected at a mission station in western CAR. They speak the language of neighboring farming communities, in this case, a Ubangian language.

2.1.3.4 Beya

The three groups known as the Beya people live in a village called Ngola in western CAR. Two of these groups (Gbaya and S.Sangha) speak a distinct language belonging to the

Ubangi subgroup of the Niger-Kordofanian family and take their name from the language that their members speak. The Mbimou speak a Bantu language belonging to the same language family.

2.2 Methods

2.2.1 Sample collection

Venous blood (5-10ml) for DNA extraction was drawn into vacutainer tubes containing the anticoagulant acid citrate dextrose (ACD). Samples collected in rural areas or other countries were transported to the laboratory at ±4°C. Buffy coats were separated from whole blood by centrifugation at 1 000x*g* (±2 300 rpm) for 15 minutes. These were then stored at -20°C until required for DNA extraction.

2.2.2 Preparation of high molecular weight DNA

Total genomic DNA was extracted from stored buffy coats by a slightly modified method of that described by Miller *et al* (1988) (**Appendix A**). This method is simple, safe and relatively inexpensive and relies on the digestion of cellular proteins by proteinase K. Degraded material is salted out by the addition of saturated NaCl and the DNA precipitated by the addition of ethanol. No hazardous organic solvents are required. The recipes for all solutions used are given in **Appendix B**.

2.2.3 Molecular methods

Specific regions of the genome were amplified using the polymerase chain reaction (PCR) which results in the production of multiple copies of the DNA region of interest (Saiki *et al*, 1985). The basic PCR reaction mixture (25µl) consisted of:

- target DNA (50-200ng)
- 10x's *Taq* polymerase buffer (supplied with the *Taq* DNA polymerase)
- 1.5 - 2.5mM $MgCl_2$ (final concentration), (25mM stock supplied with the *Taq* DNA polymerase)
- deoxyribonucleotide triphosphates (each of the four dNTP's at final concentrations ranging from 125-200µM)
- a pair of primers specific for amplifying the target sequence (2-10pmol)
- 0.5U *Taq* DNA polymerase (Advanced Biotechnology or Perkin Elmer)
- spermidine at a final concentration of 2.5mM (optional, only used in some PCR systems)
- radiolabelled deoxyribonucleotide triphosphates [(α-^{32}P)dCTP, (10mCi/ml; 300Ci/mmol, Amersham)] (for selected systems only; ≈0.05µl/25µl PCR)

Sterile tubes (0.5ml volume) were used for all PCR reactions. Any one of the following three thermocyclers was used for PCR: the HYBAID Thermal Reactor, the HYBAID Omnigene or the HYBAID Touchdown thermal cycler. All PCR reactions, except those carried out on the Touchdown cycler, were overlaid with mineral oil to prevent evaporation.

Microsatellite PCR products were labeled for ease of detection during the PCR either by incorporating radiolabelled nucleotides [(α-^{32}P) dCTP] into the growing strands or by using fluorescently labeled primers. For all PCR products that had to undergo polyacrylamide gel electrophoresis (PAGE) or restriction enzyme (RE) digestion, a small volume (5-8µl) of product was electrophoresed on a 2% agarose gel to verify that amplification was successful. PCR products for the detection of the *Alu* repeat polymorphisms (presence or absence of the *Alu*) were electrophoresed in agarose gels of the appropriate concentration to separate alleles by size. The following sections give details of specific PCR reactions and the detection methods used to identify alleles at each of the loci examined.

2.2.4 Amplification and phenotyping

2.2.4.1 Microsatellites

a) Multiplex PCR of autosomal STR loci

The alleles of two tetranucleotide repeat polymorphisms (HUMCSF1PO and HUMTH01) and a pentanucleotide repeat polymorphism (HUMCD4) were amplified in a single multiplex PCR reaction as described by Hammond *et al* (1994). Details on these unlinked loci and the amplification protocols used are given in **Tables 2.2** to **2.4**. Allele designation was according to the number of tandem repeats of the STR sequence (Hammond *et al* 1994).

Table 2.2. Polymorphic loci examined, their chromosomal locations and the primer sequences for their PCR amplification.

SYSTEM	POLYMORPHISM	CHROMOSOME LOCATION	PRIMER	SEQUENCE (5' - 3')	REFERENCE
HUMCSF1PO	STR $[AGAT]_n$	5q33 3-q34	CSF1A	AAC CTG AGT CTG CCA AGG ACT AGC	Hammond *et al* (1994)
			CSF1B	TTC CAC ACA CCA CTG GCC ATC TTC	
HUMTHO1	STR $[AATG]_n$	11p15.5	TOX 1	GTG GGC TGA AAA GCT CCC GAT TAT	Hammond *et al* (1994)
			TOX 2	ATT CAA AGG GTA TCT GGG CTC TGG	
HUMCD4	STR $[TTTTC]_n$	12p12-pter	CD4A	TTG GAG TCG CAA GCT GAA CTA GCG	Hammond *et al* (1994)
			CD4B	CCA GGA AGT TGA GGC TGC AGT GAA	
DRD2 $(CA)_n$	STR $[CA]_n$	11q22-11q23	DRD2CAFOR	TGT CCT TGG TAG GAT GGG A	Castiglione *et al* (1995)
			DRD2CAREV1	GGA GGG CGG TGC GT	
DYS393	STR $(GATA)_n$	Y	DYS393A	GTG GTC TTC TAC TTG TGT CAA TAC	Jobling and Tyler-Smith (1995)
			DYS393B	AAC TCA AGT CCA AAA AAT GAG G	
DYS19	STR $(GATA)_n$	Yp	DYS19A	CTA CTG AGT TTC TGT TAT AGT	Roewer *et al* (1992)
			DYS19B	ATG GCA TGT AGT GAG GAC A	
DYS390	STR $(CTG/AT)_n$	Y	DYS390A	TAT ATT TTA CAC ATT TTT GGG CC	Jobling and Tyler-Smith (1995)
			DYS390B	TGA CAG TAA AAT GAA CAC ATT GC	
DYS391	STR $(CTAT)_n$	Y	DYS391A	CTA TTC ATT CAA TCA TAC ACC CA	Jobling and Tyler-Smith (1995)
			DYS391B	GAT TCT TTG TGG TGG GTC TG	
CD4-*Alu*	*Alu* deletion	12p12-pter	R104	AGG CCT TGT AGG GTT GGT CTG ATA	Tishkoff *et al* (1996)
			R105	TGC AGC TGC TGA GTG AAA GAA CTG	
DRD2-*Taq*I B	RFLP	11q22-11q23	RS3DWST3	GAT GTG TAG GAA TTA GCC AGG	Castiglione *et al* (1995)
			RS3BACK1	GAT ACC CAC TTC AGG AAG TC	
DRD2-*Taq*I D	RFLP	11q22-11q23	TaqDck4	CCT CTG AGG CTT ACT GTC TG	Kidd (personal communications)
			D2EX2UP7	AAA ACT AGG GAG GGT CAG AG	
DRD2-*Taq*I A	RFLP	11q22-11q23	5014MOD	CCT TCC TGA GTG TCA TCA AC	Grandy *et al* (1993)
			TAQACK1	ACG GCT CCT TGC CCT CTA G	
YAP	*Alu* insertion	Yq11	YAP1	CAG GGG AAG ATA AAG AAA TA	Hammer and Horai (1995)
			YAP2	ACT GCT AAA AGG GGA TGG AT	

Table 2.3. Details of PCR conditions for the different polymorphisms, product lengths and electrophoresis details.

System	PRIMER	Final PCR conc	PCR conditions	Product length	Detection
HUMCSF1PO	CSF1A	0.27uM	ID: 95° 2'	295-327bp	4% (39:1) PA,
	CSF1B	0.27uM	95° 45s }		8M urea
HUMTHO1	TCX 1	0.27uM	63° 30s } X's 28	183-207bp	
	TOX 2	0.27uM	72° 30s }		
HUMCD4	CD4A	0.67uM	FE: 72° 10'	125-175bp	
	CD4B	0.67uM			
DRD2 $(CA)_n$	DRD2CAFOR	0.25uM	ID: 94° 5'	114-120bp	6%(19:1) PA,
	DRD2CAREV1	0.25uM	94° 1' }		8M urea
			58° 1' } X's 30		
			72° 1' }		
DYS393	DYS393A	0.16uM	ID: 95° 2'	186-206bp	4.3% (19:1) PA
	DYS393B	0.16uM	95° 30s }		8M urea.
DYS19	DYS19A	0.16uM	54° 30s } X's 28	203-227bp	
	DYS19B	0.16uM	72° 30s }		
DYS390	DYS390A	0.16uM		279-291bp	
	DYS390B	0.16uM			
DYS391	DYS391A	0.16uM		120-132bp	
	DYS391B	0.16uM			
CD4-*Alu*	R104	0.25uM	ID: 94° 5'	1501/1245bp	1% agarose
	R105	0.25uM	94° 1' }		
			58° 1.5' } X's 30		
			72° 1.5' }		
			FE: 72° 10'		
DRD2-*Taq* I B	RS3DWST3	0.20uM	ID: 94° 5'	459bp	2% agarose
	RS3BACK1	0.20uM	94° 1' }		
DRD2-*Taq* I D	TaqDck4	0.20uM	58° 1' } X's 30	300bp	2% agarose
	D2EX2UP7	0.20uM	72° 1' }		
DRD2-*Taq* I A	5014MOD	0.20uM		236bp	2% agarose
	TAQACK1	0.20uM			
YAP	YAP1	0.18uM	ID: 94° 2'	455/150bp	2% agarose
	YAP2	0.18uM	94° 1' }		
			51° 1' } X's 30		
			72° 1' }		

ID - initial denaturation step; FE - final extension step; PA - polyacrylamide

Table 2.4. A summary of the PCR reaction mixtures used to amplify alleles at the different loci using *Taq* DNA polymerase from two sources.

PCR Reagents	PCR Systems (volume in microlitres)		
	Autosomal STRs	CD4-*Alu*	YAP
10x's *Taq* buffer	1.5	2.5	1.5
$MgCl_2$ (25mM)	0.9	1.5	1.5
10x's dNTP's (1,25mM)	1.5	2.5	1.5
spermidine (2.5mM)	1.5		
primers (each)	0.5	0.25	1.5
Taq pol (5U/ul) (Super-Therm DNA pol, Southern-Cross)	0.08	0.1	0.08
water	5.02	17.4	6.92
raiolabelled nucleotides (1:20)	1		
DNA (50-100ng)	0.5	0.5	0.5
	Taq I (B/D/A)	$(CA)_n$	Y STRs
10x's PCR buffer II (Perkin Elmer)	2	2	2.5
$MgCl_2$ (25mM)	1.6	1	1.5
10x's dNTP's (2mM)	2	1	2.5
spermidine (2.5mM)	1.6		2.5
primers (each)	0.26	1	1
Amplitaq pol (5U/ul) (Perkin Elmer)	0.08	0.1	0.1
water	11.7	12.4	7
raiolabelled nucleotides (1:20 dilution of stock)		1	
DNA (50-100ng)	0.5	0.5	0.5

Separation of the radiolabelled PCR products was achieved by electrophoresis through a 4% denaturing PA gel of dimensions 360mm x 160mm x 0.4mm (**Appendix A** and **B**) containing 1XTBE buffer. Two microlitres of each PCR product was mixed with 3µl of formamide dye in the well of a microtitre plate. The plate was heated to 95° on a heating block for two minutes to denature the DNA. Four microlitre aliquots were then loaded into the wells of a vertical PA gel. At least two reference samples containing alleles of known size were included on each gel. Electrophoresis was performed at room temperature at 42V/cm (1 500V) until the xylene cyanol dye had migrated ±22cm from the origin. A fan was used to keep the gel from over-heating. Following electrophoresis, the gel was transferred to a sheet of Whatman 3MM filter paper and dried in a vacuum gel-dryer before being autoradiographed at room temperature using X-ray film.

b) Autosomal $(CA)_n$ repeat

Primers DRD2CAFOR and DRD2CAREV1 (Castiglione *et al*, 1995) (**Table 2.2**) were used to amplify alleles at a $(CA)_n$ dinucleotide locus within the DRD2 gene. The PCR and PAGE conditions stipulated by Castiglione *et al* (1995) were modified and are given in **Tables 2.3** and **2.4**. Electrophoresis was carried out in a 6% denaturing PA gel of dimensions 360mm x 160mm x 0.4mm (**Appendix A** and **B**). Samples were mixed with formamide dye, heated and loaded onto gels as described in **2.2.4.1(a)**. A maximum of three sequential loadings per gel could be made with an interval of at least 50 minutes between loadings. Again, at least two reference DNA samples containing alleles of known sizes were included at least once with every loading.

Electrophoresis was performed in 1 x TBE buffer at room temperature and 42V/cm (1 500V) for three hours after the final loading. A fan was directed at the gel during the run to prevent it from over-heating. Treatment of the gel after the run is as described in **2.2.4.1(a)**.

c) Multiplex PCR of Y chromosome STR loci

Four STRs (DYS393, DYS19, DYS390, and DYS391) located in the non-recombining portion of the Y chromosomes were amplified in a single multiplex PCR reaction. The PCR conditions used were modified from those described by Redd *et al* (1997) and are given in **Table 2.3**. The DYS19 primers used are those described by Roewer *et al* (1992) and those used for the DYS390, DYS391 and DYS393 polymorphisms are those described by Jobling and Tyler-Smith (1995) (**Table 2.2**). Various annealing temperatures, $MgCl_2$ and primer concentrations and the optional use of BSA were tried before a suitable PCR protocol could be established for the successful multiplex amplification of all four STRs. The PCR details are given in **Tables 2.3** and **2.4**. The `A' primer of each primer set (**Table 2.2**) was fluorescently labelled with an ABI dye: DYS19 with fam (blue) and the remaining three with joe (green). This makes it possible to detect and size the PCR products by software written specifically for the automated ABI DNA sequencer (model 377). The `A' primer for three of the systems (DYS393, DYS390 and DYS391) were labelled with the same colour (green) as the sizes of alleles at the three loci do not overlap (**Table 2.3**). The DYS19 `A' primer was labelled with a different colour (joe or blue) because of the overlap in allele sizes at this locus with those of the DYS390 locus.

Electrophoresis was carried out in 4.3% PA gels (0.2mm x 250mm x 120mm). A mixture was made of the dextran-formamide dye (Appendix A and B) and the Genescan-350 Rox (Perkin

Elmer) size standard in the proportion of 2.25µl dye: 0.25µl Rox. One microlitre of PCR product from each sample and 2.5µl of the dye-Rox mix was aliquoted into individual PCR tubes and the samples heated at 95°C for two minutes. Three microlitres of each was then loaded into individual wells (formed by the use of a sharkstooth comb) of the PA gel (one sample per well). Alternate wells were initially loaded and electrophoresis allowed to proceed for 3-5 minutes. The run was then paused while the remaining samples were loaded. This strategy prevented the incorrect scoring of bands due to possible sample spillage between adjacent wells during the loading procedure.

The ABI model 377 coupled with the Genescan software facilitates the automated electrophoresis and analysis of polymorphic markers by the direct detection of fluorescently labelled DNA molecules as they migrate through vertical PA gels. An internal size standard was loaded with every sample to correct for uneven movement of fragments in different lanes of the gel. The analysis programme (ABI Prism™ Genescan™ 2.02) then uses the size standard in each lane to calculate the sizes of the unknown alleles in that lane. The commercially available size standard Genescan –Rox-350 (fragment sizes ranging from 50bp-350bp) was used. The ABI Genotyper® 2.0 software package was also used to facilitate allele assignment.

2.2.4.2 Bi-allelic polymorphisms

The following five bi-allelic polymorphisms were examined:

(a) an *Alu* deletion polymorphism in the CD4 gene,

(b) three *Taq*I RFLPs in the DRD2 gene,

(c) an *Alu* insertion polymorphism on the Y chromosome (YAP).

(a) CD4-*Alu* polymorphism

Genotyping of the CD4-*Alu* deletion polymorphism was achieved using primers R104 and R105 (Tishkoff *et al*, 1996). The PCR conditions were modified slightly and are given in **Tables 2.3** and **2.4**. Amplification products of 1 500bp and 1 250bp in size are obtained for intact and deletion alleles respectively. The PCR products were visualized after electrophoresis in 1% agarose gels containing ethidium bromide and sized by comparison to the fragments of a 1kb molecular weight DNA marker (Gibco BRL) (**Appendix B**) loaded on every gel.

(b) *Taq*I polymorphisms

The primer sets used to amplify the regions in the DRD2 gene in which the three *Taq*I polymorphisms occur are given in **Table 2.3**. The specific order in which the *Taq*I polymorphisms are discussed and used to construct haplotypes (*Taq*I B, *Taq*I D and then *Taq*I A) corresponds to their location in the gene. Annealing temperature as well as $MgCl_2$ concentrations had to be modified from those given in published protocols (**Tables 2.2**). Fortunately, a common PCR protocol could be adapted for all three systems (**Tables 2.3** and **2.4**). The alleles at the three loci were amplified separately for each sample. The PCR product sizes are given in **Table 2.3**.

Following amplification, 20µl of PCR product was digested with 5U of *Taq*I (Boehringer Mannheim, 10U/µl) in a total volume of 25µl. The digests were overlaid with two drops of mineral oil to prevent evaporation and incubated at 65°C for 2hrs. Three µl of Ficoll dye was added to each digest before it was electrophoresed in a 2% agarose gel at 10V/cm for one

hour. A 1kb molecular weight DNA marker (Gibco BRL) (Appendix B) was loaded into one or two slots of every gel for the correct sizing of DNA fragments. Allele assignments were made as follows:

- *Taq*I B locus: *Taq*I site absent - B1 allele (459bp fragment); *Taq*I site present - B2 allele (267bp+192bp fragments)
- *Taq*I D locus: *Taq*I site absent - D1 allele (300bp fragment); *Taq*I site present - D2 allele (212bp+88bp fragments)
- *Taq*I A locus: *Taq*I site absent - A1 allele (236bp fragment); *Taq*I site present - A2 allele (124bp+112bp fragments)

(c) YAP polymorphism

The primers and PCR conditions used to detect the YAP polymorphism are those described by Hammer and Horai (1995) (**Tables 2.2** to **2.4**). A PCR product of 150bp indicates the presence of the ancestral allele (YAP-) while a 455bp fragment indicates the presence of the insertion allele (YAP+). Products were visualised after electrophoresis in a 1% agarose gel and sized with the aid of a 1kb molecular weight DNA marker (Gibco BRL) loaded on every gel.

2.2.4.3 Autosomal haplotypes

(a) Haplotypes at the CD4 locus

Data from samples that had been typed for both the HUMCD4 STR [**2.2.4.1(a)**] and the CD4-*Alu* deletion polymorphism [**2.2.4.2(a)**] were used to construct two-site haplotypes. The two polymorphic loci are located within non-coding regions of the gene on the short arm of chromosome 12. They are 9.8kb apart and are assumed to be selectively neutral (**Figure 2.2**).

(b) Haplotypes at the DRD2 locus

Samples which were successfully typed for the $(CA)_n$ repeat polymorphism [**2.2.4.1(b)**] and the three *Taq*I polymorphisms [**2.2.4.2(b)**] within the DRD2 gene were used for the construction of four-site haplotypes. **Figure 2.3** is a map of part of the DRD2 gene (located on chromosome 11q22-11q23) showing the positions of these polymorphisms.

2.2.4.4 Y chromosome haplotypes

Data from samples that were successfully typed for the four STRs on the Y chromosome [**2.2.4.1(c)**] and the YAP [**2.2.4.2(c)**] were used for the construction of 5-site haplotypes. An effort was made to standardize the nomenclature for alleles at the STR loci and is discussed further in **CHAPTER 5**.

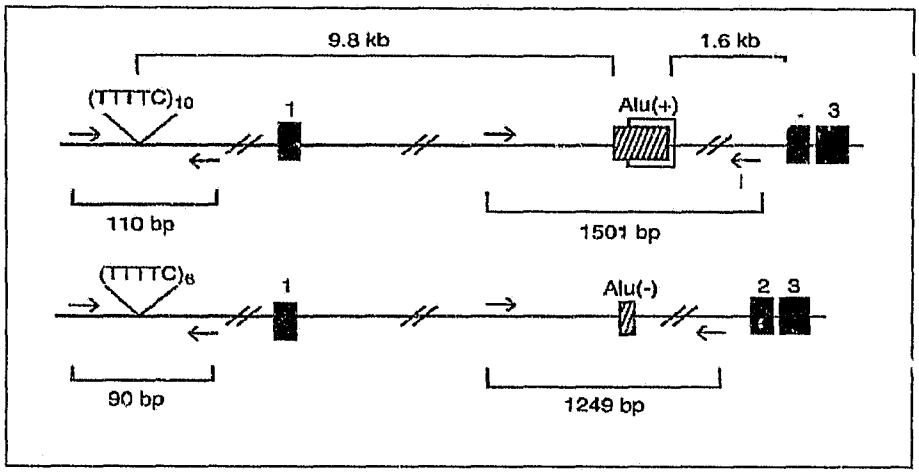


Figure 2.2. Relative positions of the STR $(TTTTC)_n$ and *Alu* polymorphism within the CD4 gene on chromosome 12 [adapted from Tishkoff *et al* (1996)]. The hatched box represents the *Alu* element and the open box the region that is deleted to give the *Alu-* allele. The arrows show the direction of amplification of the primers used to amplify the region of interest. The filled boxes represent exons in the gene.

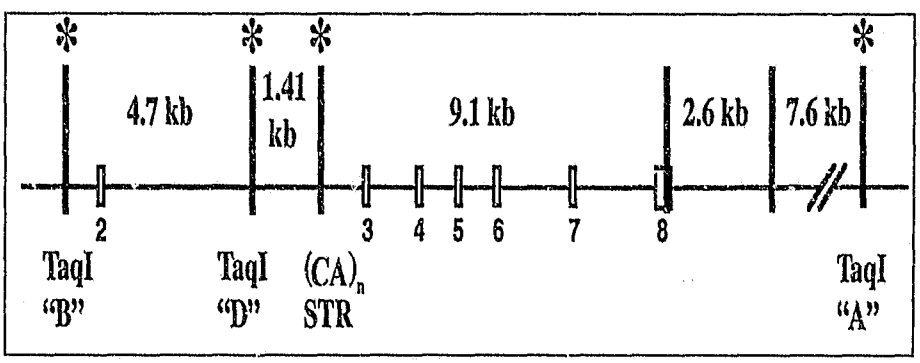


Figure 2.3. Relative positions of the polymorphisms studied (marked by *) within the DRD2 gene on chromosome 11 [adapted from Kidd *et al* (1998)]. Boxes show the positions of exons in the gene.

2.3 Data analysis

a) **Allele frequencies** for each system were estimated directly by gene counting. The standard errors (SE) of the frequency estimates were calculated as described by Nei (1987).

b) **Chi-squared analysis** was applied to all the bi-allelic data to check for deviations from Hardy Weinberg expectation in individual population samples. For the autosomal multi-allelic STR data, a dos based programme, provided by Dr AB Lane was used. This programme is based on the 'exact' method described by Guo and Thompson (1992).

c) **Autosomal haplotype frequencies** were estimated using the computer programmes EM-HAPLO (for the CD4-*Alu* locus) (Hawley and Kidd, 1995) and Arlequin ver 1.1 (for the DRD2 locus) (Schneider *et al*, 1997). Both programs implement an expected maximum (EM) algorithm [Dempster *et al* (1977)] to estimate haplotype frequencies from phenotype data.

d) Genotype data for the Y chromosome was used directly as **Y haplotype data** as each Y chromosome can be regarded as a haplotype. Frequencies of haplotypes were then obtained by direct counting. Haplotype sharing between populations and Reynold's and Slatkin's genetic distances (both based on F_{ST} values) for the Y haplotype data were generated using the computer program Arlequin ver 1.1.

e) Unbiased estimates of **expected heterozygosity** or **haplotype diversity** (and associated standard errors) were calculated according to Nei (1987). The diversity values can theoretically range from 0 to 1, with 0 representing the minimum value (i.e. a single haplotype fixed in a population) and 1 representing the maximum possible genetic variability (i.e. a different haplotype on every chromosome).

f) A variation of Fisher's exact test as described by Weir (1996) was performed to test the significance of **allele frequency differences** between the populations examined.

g) **Genetic distance measurements** are indicators of relatedness among populations and are useful for gaining insights into the historic and phylogenetic relationships among such groups. The distance is a measure of the extent of genetic difference between two populations and is given a numerical quantity. A number of methods exist for genetic distance calculations based on either the infinite allele model (IAM; Kimura and Crow, 1964), or the stepwise mutational model (SMM; Ohta and Kimura, 1973). The genetic distance measure D_{SW} (Shriver *et al*, 1995) was used to analyse microsatellite data while the D_A distances measure (Nei *et al*, 1983) was used to analyse haplotype data. The programme DSW (Ota, 1996) was used to calculate both types of distance measures between each pair of populations.

h) **Phylogenetic trees** are constructed from genetic distance data and represents possible relationships between populations. The most often-used tree drawing methods are the Unweighted Paired Group Method using arithmetic Averages (UPGMA) and the Neighbor Joining (NJ) method of Saitou and Nei (1987). The NJ method was used exclusively for drawing population trees in the present study. The programme DSW (Ota, 1996) was used for cluster analysis as it is able to both generate genetic distances from allele frequency data, as well as draw trees from the generated distance matrix.

i) **Principal components analysis** (PCA) was performed using the computer programme ANTANA (Harpending and Rogers, 1984) to assess the pattern of within and between continent genetic variability as an alternate method to phylogenetic analyses. This is the optimal multi-dimensional representation of the genetic differences among populations. The principal components (PCs) are calculated from genetic distances in a process whereby the genetic distance between populations is apportioned to a few axes. The two

dimensional display of the first two components which retain most of the information of the analysis are used to generate two-dimensional PC maps. The clustering of the populations in the plane provides an indication of their relatedness.

j) Gene flow and population substructure was assessed by the proposed model of regression of heterozygosity and genetic distance (Harpending and Ward, 1982). Assuming an island model of population structure, the genetic distance of a particular island population from the overall mean gene frequency of the island's population (gene frequency centroid) and the relative heterozygosity of that island should be linearly related, if exchange with populations outside the region is the same for each island population. However, if a population loses variability, i.e. becomes more homogeneous, because of a decline in migration, it should show the largest genetic **distance from the centroid**. The expected heterozygosity for each population is calculated from the frequency data and the average heterozgosity is the mean of these individual heterozygosities. The distance from the centroid is calculated for each population and these points are plotted against the values of expected heterozygosity on a Cartesian plane.

k) The least squares method developed by Elston (1971) was used to estimate the percentage of parental population **admixture** in the Malagasy on the assumption that there were only two ancestral parental groups, viz. Africans and Indonesians. A dos-based programme (ADMIX) to carry out these calculations was kindly provided by Dr AB Lane.

l) The programme AMOVA (analysis of molecular variance) is used to examine **population substructure** from estimates of inter-population and intra-population variances. AMOVA calculates an F_{ST} value for the group of sub-populations as a whole and also pairwise F_{ST} values which can be directly or indirectly used as genetic distances. Within limits, the

larger the F_{ST} value, the greater the time since the two populations comprising the group diverged from each other.

m) The age of coalescent, T, for groups of related microsatellite haplotypes can be calculated using a method described by Goldstein (1995) and Slatkin (1995). The theoretical relationship $S=2\mu T\sigma^2$ is applied to the data where S is the expected mean variance in microsatellite repeat length for the loci examined, μ represents the mutation rate and σ^2 is 1 for the SMM (Di Rienzo *et al*, 1994; Slatkin, 1995).

CHAPTER 3 – AUTOSOMAL SHORT TANDEM REPEAT POLYMORPHISMS

3.1 Introduction

Short tandem repeat (STR) polymorphisms have been shown to be extremely valuable in gene mapping by linkage analysis (Todd *et al*, 1991), in the diagnosis of inherited disease (Willems, 1994), in forensic and parentage testing (Edwards *et al*, 1992; Hammond *et al*, 1994)) as well as in population and evolutionary studies (Bowcock *et al*, 1994; Jorde *et al*, 1995). Although several thousand microsatellites have been mapped in the human genome, there is a paucity of data in the frequency and distribution of STR alleles in populations from certain parts of the world. The work described below goes a little way towards addressing this by generating data on Malagasy subgroups and on their purported parental populations.

Although STR loci are abundant in the human genome, not all of them are suitable for use in population genetics: poor PCR product yield, PCR artifacts and sequence microheterogeneity can make the interpretation of results difficult. For instance, four and five base pair repeat polymorphisms tend to amplify more authentically and their alleles are more easily assigned than are alleles at dinucleotide repeat loci (Sprecher *et al*, 1996). The selection of a STR marker would also depend on its heterozygosity and differential distribution in different population groups. Of the four STR loci selected in this study, only one dinucleotide polymorphism was included. This $(CA)_n$ repeat locus is located within the DRD2 gene and recent studies by Kidd *et al* (1998) have shown differences in the distribution of alleles at this locus between African and non-African populations. The same has been shown for the tetranucleotide HUMCD4 locus located within the CD4 gene by

Tishkoff *et al* (1996). Since the Malagasy gene pool is hypothesized to be derived largely from African and Indonesian sources, these two loci seem ideal for this study.

Hammond *et al* (1994) examined a number of STR loci for use in various personal identification settings such as parentage testing and forensics. They developed three standard triplex PCR reactions for amplifying alleles at nine STR loci (including HUMCD4). These loci are generally designated by their GenBank locus names together with the representation of the core repeat motif, e.g. HUMTH01 $[AATG]_n$ refers to the STR polymorphism located in the **hum**an tyrosine **h**ydroxylase gene whose core repeat is [AATG] (Edwards *et al*, 1992). The alleles are numbered according to the numbers of tandem repeats present (originally determined by DNA sequencing). Because of the high heterozygosity rates at some of these loci and the advantages of using multiplex PCR reactions, we selected three systems (HUMCSF1PO, HUMTH01 and HUMCD4) which could be amplified in a single triplex PCR. This proved to be both a quick and cost-effective approach for studying large numbers of samples.

Hammond *et al* (1994) have estimated the mutation rates for the HUMCSF1PO, HUMTHO1 and HUMCD4 loci to be 4.88×10^{-5}, 3.96×10^{-5} and 4.81×10^{-5} respectively. These values are lower than the averages estimated by other studies for STR loci e.g. 1.5×10^{-3} (Jin *et al*, 1994) and 2.1×10^{-3} (Weber and Wong, 1993), which suggests that these loci were well chosen for evolutionary studies. Allele frequency distributions of HUMCSF1PO, HUMTH01 and HUMCD4 have been described previously (Edwards *et al*, 1992; Hammond *et al*, 1994).

The first part of the analyses deals with the distribution of allele frequencies in different populations. The data from each STR locus has initially been analyzed separately. In the second part of the analysis, data from all the loci were pooled for population tree reconstruction and principal components analyses, which summarizes and averages the net genetic distance between the various groups. Caucasoid data from the literature (Edwards *et al*, 1992; Hammond *et al*, 1994) have been included in some analyses to provide some perspective. The distance from the centroid plot was used to estimate the variation and possible differences in gene flow into individual populations compared to that in the entire sample. Elston's least squares method was the used to estimate the proportions of admixture in the highland and lowland Malagasy populations using frequency data from purported parental populations. The details on the analytical methodologies and the software programs used throughout are given in **section 2.3**.

3.2 Allele identification

The radiolabelled alleles at different loci have distinct relative rates of migration during PAGE and can therefore be distinguished from each other. **Figures 3.1** and **3.2** show examples of autoradiographs illustrating the polymorphic alleles seen at the four STR loci studied. Alleles at the three loci which were amplified in the multiplex PCR reaction appear as closely spaced doublets (**Figure 3.1**) while each of the $(CA)_n$ repeat alleles manifests as one main band with three much lighter shadow bands (**Figure 3.2**). The lighter shadow bands sometimes complicate the interpretation of the band pattern. STR loci with more than two nucleotide repeats have been shown to be less prone to repeat slippage during PCR than dinucleotide repeats (Sprecher *et al*, 1996). Hence, they produce fewer shadow or stutter bands on electropherograms.

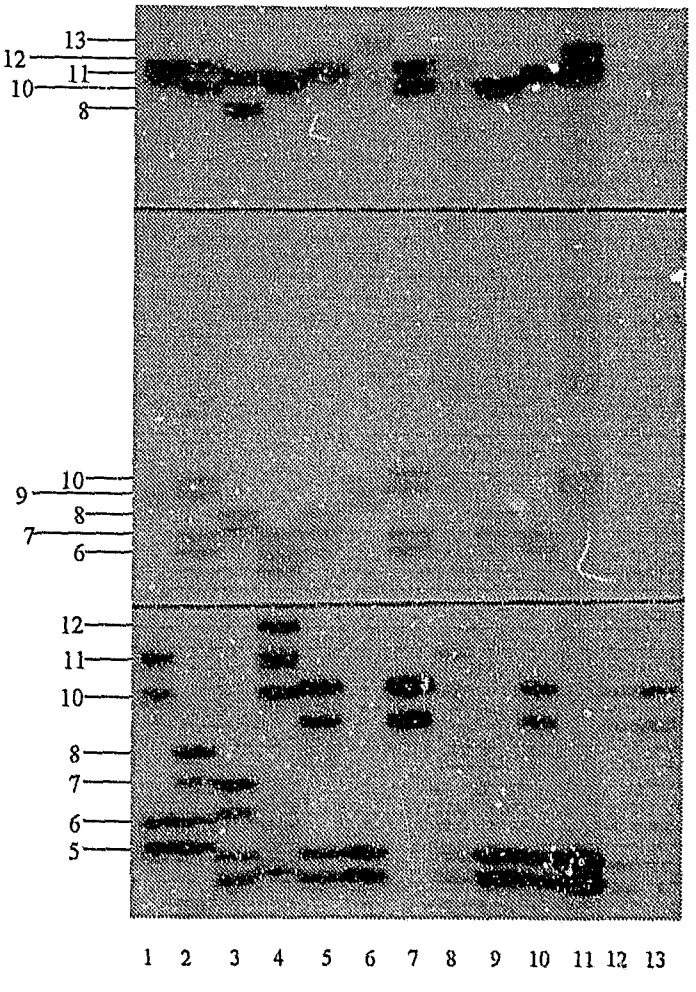


Figure 3.1. Autoradiograph showing separation of alleles amplified in a multiplex PCR reaction. Bands in the top panel represent alleles at HUMCSF1PO, those in the middle panel represent alleles at HUMTHO1 and those in the lowest panel represent alleles at the HUMCD4 locus. Each allele (see labels on the left-hand side of the figure) is represented by a doublet and the top band has been used for allele identification.

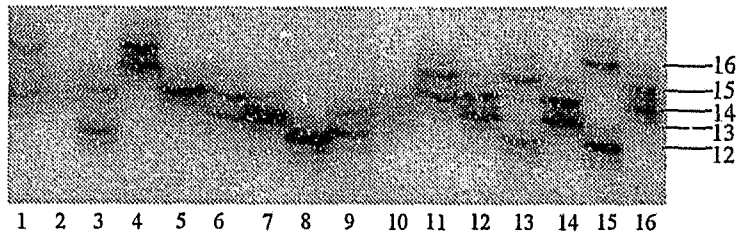


Figure 3.2. Autoradiograph showing the separation of alleles at the $(CA)_n$ repeat locus within the DRD2 gene. Each allele (see labels on the right hand side of the figure) is represented by one major (dark) band shadowed above and below by at least three lighter bands.

3.3 Allele frequency distributions

The observed allele frequency distributions in individual populations are shown in **Table 3.1** through to **Table 3.4**. The frequencies of alleles with the highest frequencies (most frequent or modal alleles) in each population-locus combination appear in bold typeface for easier visualization and comparison. The frequency distribution of STR allele sizes in a population may reflect something of its demographic history. Smooth unimodal distributions suggest large, stable populations that have been expanding for a long time while erratic, multimodal distributions suggest populations that are not growing in size, or have recently passed through a bottleneck or have experienced recent gene flow.

3.3.1 HUMCSF1PO

The DNA from 777 individuals belonging to 22 population groups was typed for alleles at the HUMCSF1PO STR locus. Eight of nine alleles (7-15 repeats) reported in the literature at this locus (Hammond *et al*, 1994) were detected in the present study (**Table 3.1**). The general distribution of allele frequencies appears to be unimodal in non-African populations and bimodal in the African populations. The modal allele in various Malagasy groups varies between alleles 10, 11 or 12. Allele 12 forms the mode in the remaining populations with allele 10 forming a second mode in five of the six African populations. Alleles 7, 8 and 9 occur at low frequencies in almost all the populations screened, except for the Australasian populations from which they are absent. This suggests that all/most of the alleles 7, 8 and 9 in the Malagasy are of African origin. Alleles 10 to 13 have a relatively uniform frequency distribution across most populations while alleles 14 and 15 are rare to absent in the entire study sample.

Table 3.1. Frequencies of alleles at locus HUMCSF1PO in the populations examined. The frequencies of modal alleles are shown in bold typeface.

Population	2N	Allele								
		7	8	9	10	11	12	13	14	15
Malagasy	**772***									
Merina	72	0.028	0.014	0.028	0.208	0.306	**0.347**	0.069	0	0
Betsileo	70	0.029	0	0.057	0.271	**0.300**	0.243	0.100	0	0
Antaisaka	68	0.029	0.044	0.015	**0.309**	0.294	0.279	0.015	0.015	0
Tsimihety	70	0	0.029	0.014	0.271	0.286	**0.314**	0.086	0	0
Sakalava	86	0.047	0.035	0.047	0.279	**0.291**	0.221	0.081	0	0
Mahafaly	64	0.016	0.016	0.031	0.234	0.125	**0.516**	0.063	0	0
Antandroy	72	0.056	0.042	0.028	0.292	**0.333**	0.194	0.056	0	0
Betsimisiraka	76	0.013	0.053	0.053	0.250	0.250	**0.316**	0.053	0.013	0
Antaimoro	76	0.013	0.039	0.053	0.224	**0.316**	0.276	0.079	0	0
Antankarana	32	0.031	0	0.031	0.188	**0.344**	**0.344**	0.063	0	0
Makoa	86	0.023	0	0.023	**0.326**	0.314	0.302	0.012	0	0
Non-African	**390***									
Caucasoid**	346	0	0.003	0.040	0.263	0.266	**0.332**	0.069	0.020	0.006
SA Indian	74	0	0	0.027	0.135	0.311	**0.459**	0.068	0	0
Indonesian	158	0	0	0	**0.266**	0.253	**0.373**	0.108	0	0
Polynesian	72	0	0	0	**0.250**	0.222	**0.375**	0.139	0.014	0
New Guinean	40	0	0.025	0.050	**0.350**	0.150	**0.350**	0.075	0	0
Solomon Islands	46	0	0	0	0.087	0.391	**0.413**	0.109	0	0
African	**392***									
Tsonga	72	0.056	0.083	0	0.194	0.264	**0.375**	0.028	0	0
Zambian	106	0.094	0.057	0.047	**0.217**	0.142	**0.358**	0.085	0	0
Mbenzele	88	0	0.045	0.045	**0.352**	0.261	**0.284**	0.011	0	0
Gbaya	70	0.057	0.114	0.086	**0.271**	0.214	**0.229**	0.029	0	0
Mbimou	22	0	0.045	0	**0.409**	0.273	0.227	0.045	0	0
S.Sangha	34	0.029	0.059	0	**0.324**	0.088	**0.412**	0.088	0	0

* number of chromosomes examined in present study, ** data taken from Hammond *et al*, 1994.

3.3.2 HUMTHO1

Seven hundred and sixty three individuals from 22 populations were typed for alleles at the HUMTH01 STR locus. Seven of the eight alleles (5-12 repeats) reported in the literature (Edwards *et al*, 1992) were detected in the present study. Allele frequency estimates for individual populations are presented in **Table 3.2**. Previous studies have shown alleles 5 and 11 to be rare or absent in most populations (Edwards *et al*, 1992; Perez-Lezaun *et al*, 1997) and the frequencies of alleles 6 to 10 to vary in different populations. The same trends were noted in the populations examined in the present study. Allele 6 occurs at low frequencies in the Tsonga, Zambian, Mbenzele and Mbimou (frequency ranging from 0.030 to 0.057), but is appreciably higher in the Gbaya, S.Sangha, Malagasy and non-African populations (frequencies between 0.065 and 0.348). Allele 7 is common throughout, generally being the modal allele in the African, Malagasy, Indonesian and Polynesian populations. The allele frequency distribution is either unimodal or bimodal with the modes being anywhere from alleles 6 through to allele 10. None of the alleles at this locus seem to be specific to either African or non-African populations. Allele frequency distributions also do not appear to be strikingly different between African and non-African populations.

Table 3.2. Frequencies of alleles at locus HUMTHO1in the populations examined. The frequencies of modal alleles are shown in bold typeface.

Population	2N	Allele						
		5	6	7	8	9	10	11
Malagasy	772*							
Merina	68	0	0.147	**0.368**	0.147	**0.206**	0.132	0
Betsileo	72	0	0.111	**0.361**	0.292	0.139	0.097	0
Antaisaka	72	0	0.111	**0.347**	0.278	0.139	0.125	0
Tsimihety	70	0	0.143	**0.414**	0.186	**0.229**	0.029	0
Sakalava	82	0	0.171	**0.451**	0.110	**0.232**	0.024	0.012
Mahafaly	66	0	0.136	**0.667**	0.091	0.076	0.030	0
Antandroy	70	0	0.100	**0.529**	0.100	**0.200**	0.071	0
Betsimisiraka	76	0	0.184	**0.461**	0.132	**0.145**	0.079	0
Antaimoro	74	0	**0.338**	0.270	0.216	0.135	0.041	0
Antankarana	32	0	0.063	0.188	**0.375**	0.281	0.094	0
Mikea	90	0	0.189	**0.278**	0.222	0.200	0.111	0
Non-African	392*							
Caucasoid**	372	0.005	**0.226**	0.159	0.11	0.142	**0.352**	0.005
SA Indian	74	0	**0.257**	0.149	0.081	**0.338**	0.176	0
Indonesian	158	0	0.114	**0.316**	0.139	**0.323**	0.101	0.006
Polynesian	74	0	0.203	**0.338**	0.081	0.108	**0.257**	0.014
New Guinean	40	0	**0.200**	0.025	**0.500**	0.275	0	0
Solomon Island	46	0	**0.348**	0.174	**0.239**	0.196	0.043	0
African	362*							
Tsonga	66	0	0.030	**0.455**	0.242	0.212	0.061	0
Zambian	88	0	0.034	0.318	**0.330**	0.273	0.045	0
Mbenzele	88	0	0.057	**0.523**	0.193	**0.205**	0.023	0
Gbaya	64	0	0.125	**0.422**	0.219	0.172	0.063	0
Mbimou	22	0	0.045	**0.364**	**0.364**	0.136	0.091	0
S.Sangha	34	0	0.206	**0.294**	0.176	**0.294**	0.029	0

* number of chromosomes examined in present study

** data taken from Edwards *et al*, 1992.

3.3.3 HUMCD4

Very little variation at this locus has been detected in non-human primates (only four alleles of 3 to 6 repeats have been observed). Sequence analysis of the repeats by Tishkoff *et al* (1996) show that alleles smaller than 10 repeats contain identical repeats of '**TTTTC**' throughout whereas allele 10 or larger (found only in humans) have an imperfect '**CTTTC**' as the fourth repeat from the 5' end. From sequence data on human and non-human alleles, they concluded that the ancestral STR allele comprised a small number of '**TTTTC**' repeats. Alleles having ten or more repeats were derived from a single mutant that generated a higher repeat number allele with the imperfect '**CTTTC**' repeat. Tishkoff *et al*'s (1996) nomenclature for alleles at this locus was adopted instead of that used by Hammond *et al* (1994). The allele with five repeat units is termed allele 5 by Tishkoff *et al* (1996) but is considered to be allele 7 by Hammond *et al* (1994) because the latter count the 3' flanking sequence 'CTTCTTTT' as two repeats. Sequencing shows that these imperfect repeats are not variable either in humans or non-human primates and it therefore seems preferable to exclude them from the count of repeats. Tishkoff *et al* (1996) have gathered allele frequency data on this STR for a number of populations from different parts of the world and this has served as a good source of comparative data for the present study.

All together the DNA from 793 individuals from 22 populations was typed for alleles at the HUMCD4 STR locus. Allele frequency estimates for individual populations are given in **Table 3.3**. Of the twelve alleles reported at this locus (4-15 repeats) by Tishkoff *et al* (1996), only three alleles (5, 6 and 10) were detected at appreciable frequencies by them in non-African populations. This large difference in allele distribution worldwide has proven to be very useful in terms of detecting African contribution to the Malagasy. All twelve

alleles were detected in the African sample and ten of these were present in the Malagasy – a range wider than that reported for any other population outside of Africa. More alleles are seen in the lowland Malagasy than in the highland Malagasy e.g. ten alleles were present in the lowland sample but only six were detected in the highland sample. This suggests that the lowlanders have closer affinities with African populations, who show the greatest diversities (**Table 3.6**).

As expected, only alleles 5, 10 and in some instances allele 6 occur at appreciable frequencies in the other non-African populations examined. The distribution of the alleles is complex in most populations, with alleles 5, 8 and 11 being modal in Africans and alleles 5 and 10 being modal in Malagasy and other non-African populations. The frequencies of alleles 5 and 10 in the Malagasy are higher than in the African samples but lower than that in the Indonesians. Alleles 7, 8, 9, 12, 13 and 14 were not detected in the Indonesian sample, but were present in the African sample. Their presence in the Malagasy thus supports an African contribution to the gene pool of the Malagasy.

Allele 4 was detected only twice among the 3 200 chromosomes typed by Tishkoff *et al* (1996) (once in the Kikuyu of Kenya and once in a Biaka pygmy from the CAR). In the present study, allele 4 was detected in two Mbenzele individuals. This is not surprising as the Mbenzele separated from the Biaka less than 200 years ago.

Table 3.3. Frequencies of alleles at locus HUMCD4 in the populations examined. The frequencies of modal alleles are shown in bold typeface.

Population	2N	Allele											
		4	5	6	7	8	9	10	11	12	13	14	15
Malagasy	**788***												
Merina	74	0	**0.514**	0.041	0	0.068	0	**0.338**	0.041	0	0	0	0
Betsileo	72	0	**0.444**	0.125	0	0.125	0	**0.153**	0.139	0.014	0	0	0
Antaisaka	74	0	**0.270**	0.108	0.027	0.095	0.054	**0.243**	0.149	0.027	0	0.027	0
Tsimihety	70	0	**0.386**	0.071	0.057	0.143	0.029	0.129	**0.143**	0.014	0	0.029	0
Sakalava	84	0	**0.405**	0.095	0.060	0.083	0	**0.167**	0.119	0.036	0.012	0.024	0
Mahafaly	68	0	**0.412**	0.029	0.015	0.044	0	**0.279**	0.029	0.147	0.015	0.029	0
Antandroy	72	0	**0.486**	0.056	0.014	0.111	0	**0.153**	0.139	0.028	0	0.014	0
Betsimisiraka	76	0	**0.368**	0.053	0.039	0.053	0.013	**0.303**	0.105	0.039	0.026	0	0
Antaimoro	76	0	**0.408**	0.092	0.026	0.039	0.013	0.145	**0.197**	0.039	0	0.039	0
Antankarana	32	0	**0.344**	0.125	0.094	0.031	0.031	**0.188**	0.094	0.031	0	0.063	0
Mikea	90	0	**0.367**	0.056	0.067	0.089	0.022	**0.189**	0.144	0.033	0.022	0.011	0
Non-African	**398***												
Caucasoid**	382	0	**0.387**	0.306	0	0	0.003	**0.277**	0.021	0.005	0	0	
S.A Indian	78	0	**0.462**	0.064	0	0	0	**0.436**	0.038	0	0	0	0
Indonesian	160	0	**0.606**	0.006	0	0	0	**0.369**	0.019	0	0	0	0
Polynesian	74	0	**0.581**	0.122	0	0	0.081	**0.216**	0	0	0	0	0
New Guinean	40	0	**0.875**	0	0	0	0	**0.125**	0	0	0	0	0
Solomon Island	46	0	**0.870**	0.022	0	0	0.022	**0.065**	0.022	0	0	0	0
African	**400***												
Tsonga	80	0	**0.238**	0.075	0.025	**0.163**	0.038	0.113	**0.175**	0.063	0.038	0.038	0.038
Zambian	106	0	**0.245**	0.123	0.038	**0.132**	0.038	0.085	**0.208**	0.066	0.038	0.028	0
Mbenzele	88	0.023	**0.193**	0.023	0.057	**0.068**	0.011	0.025	**0.307**	0.239	0.034	0.023	0
Gbaya	70	0	**0.157**	0.071	0.057	**0.171**	0	0.086	**0.314**	0.071	0.043	0.029	0
Mbimou	22	0	**0.182**	0.091	0.045	**0.318**	0.045	**0.091**	**0.091**	0.045	0.045	0.045	0
S.Sangha	34	0	**0.206**	0.029	0.088	**0.206**	0.059	0.088	**0.147**	**0.147**	0.029	0	0

* numbers of chromosomes examined in present study.

** data taken from Hammond *et al*, 1994.

3.3.4 $(CA)_n$

Seven hundred and fifty seven individuals were typed for the $(CA)_n$ repeat at the DRD2 locus(**Table 3.4**). Castiglione *et al* (1995) found alleles ranging in size from 6 to 12 repeats in non-human primates. Of the six alleles that occur in human populations (12 to 17 repeats), allele 14 tends to be the commonest. Allele 13 is common in Africa, the Middle East and Europe while allele 15 is common to the latter two regions. Allele 16 appears to be rare in Africa and common in non-African populations. Alleles 12 and 17 are rare worldwide.

Table 3.4 summarizes the allele frequency distributions found in the present study. Most African and the SA Indian popul. 'on show a unimodal allele distribution with allele 14 forming the mode in the former and allele 15 forming the mode in the latter. While all six alleles are present in the African and Malagasy populations, allele 12 appears to be rare or absent in non-Africa populations. The frequency of allele 16 is appreciably higher in the Malagasy and SA Indian samples compared to most of the African samples. Also, only two of the Malagasy groups, the Merina and Betsileo (both highlanders) have higher allele 16 frequencies than those found in Asian (SA Indian and Indonesian) population groups. This would support a larger Asian ancestry for the highland groups compared to the lowland groups. The absence or rarity of allele 12 in non African samples and its presence in all the African as well as all eleven Malagasy groups suggests an African origin of this allele in the Malagasy.

Table 3.4. Frequencies of alleles at the DRD2 $(CA)_n$ locus in the populations examined. The frequencies of modal alleles are shown in bold typeface.

Population	2N			Allele			
		12	13	14	15	16	17
Malagasy	**718***						
Merina	58	0.069	0.207	**0.276**	0.052	**0.397**	0
Betsileo	64	0.063	**0.250**	0.219	0.125	**0.344**	0
Antaisaka	64	0.109	**0.328**	0.313	0.047	**0.172**	0.031
Tsimihety	70	0.057	0.300	**0.400**	0.114	**0.129**	0
Sakalava	72	0.097	0.319	**0.375**	0.083	**0.111**	0.014
Mahafaly	60	0.100	0.167	**0.417**	0.000	**0.150**	0.167
Antandroy	68	0.162	0.235	**0.324**	0.088	**0.162**	0.029
Betsimisiraka	72	0.097	0.222	**0.250**	**0.250**	0.181	0
Antaimoro	68	0.118	0.265	**0.279**	0.147	0.132	0.059
Antankarana	24	0.167	0.167	**0.417**	0.167	0.083	0
Mikea	98	0.204	0.286	**0.296**	0.102	0.102	0.010
Non-African	**96***						
Middle Eastern**	196	0	**0.224**	0.168	**0.486**	0.122	0
European**	80	0.010	0.140	0.170	**0.530**	0.150	0
SA Indian	78	0	0.141	0.167	**0.385**	0.282	0.026
Indonesian	18	0	**0.389**	0.278	0.111	**0.222**	0
African	**700***						
San	196	0.046	0.164	**0.536**	0.209	0.040	0
N Sotho (Pedi)	100	0.210	0.260	**0.360**	0.130	0.040	0
Tsonga	100	0.110	0.330	**0.420**	0.100	0.030	0.010
Zambian	96	0.188	0.240	**0.406**	0.104	0.052	0.010
Mbenzele	86	0.128	0.209	**0.535**	0.047	**0.058**	0.023
Gbaya	66	0.152	0.182	**0.485**	0.121	0.061	0
Mbimou	22	0.136	**0.409**	0.273	0.091	0.091	0
S.Sangha	34	0.029	0.147	**0.618**	0.088	**0.118**	0

* number of chromosomes examined in present study.

** data taken from Kidd *et al*, 1998.

3.4 Statistical analyses

3.4.1 Hardy-Weinberg equilibrium

Population samples were tested for compliance with Hardy-Weinberg expectations using an adaptation of Fisher's exact method (Guo and Thompson, 1992). Results of the tests are given in **Table 3.5**. A *P* value of less than 0.050 was considered to be statistically significant in this study. Seven significant deviations were found among the 88 locus-population combinations tested. These were the Tsonga-HUMCSF1PO, Tsimihety-HUMTHO1, Antaimoro-HUMTHO1, Indonesia (SS)-HUMCD4, Merina-$(CA)_n$, Antandroy-$(CA)_n$ and Zambian-$(CA)_n$ population-locus combinations. None of the seven population-locus samples showed significant deviations from expected homozygosity which implies that if apparent deviations from HWE are not due to chance, they are unlikely to be due to inbreeding. All seven also involve a different population-locus combination which suggests that neither population admixture nor non-random mating is the cause. Using a 95% confidence limit for the test, one would expect one of twenty at-equilibrium populations tested for HWE to show a 'significant' deviation. The number of trials (7 of 88) which had *P* values which were less than 0.050 is slightly more than the expected value of 4.4 but it seems unlikely that this is due to anything other than chance.

Table 3.5. Results of tests aimed at detecting significant deviations from Hardy-Weinberg expectation in the population samples examined.

Population	HUMCSF1PO		HUMTHO1		HUMCD4		$(CA)_n$	
	2N	*P*	2N	*P*	2N	*P*	2N	*P*
Malagasy								
Merina	72	0.940	68	0.340	74	0.073	58	**0.017***
Betsileo	70	0.847	72	0.594	72	0.937	64	0.066
Antaisaka	68	0.523	72	0.066	74	0.708	64	0.291
Tsimihety	70	0.169	70	**<0.001***	70	0.938	70	0.949
Sakalava	86	0.757	82	0.858	84	0.313	72	0.590
Mahafaly	64	0.690	42	0.528	68	0.207	60	0.679
Antandroy	72	0.542	70	0.122	72	0.296	68	**0.037***
Betsimisiraka	76	0.331	76	0.751	76	0.428	72	0.097
Antaimoro	76	0.291	74	**0.019***	74	0.274	68	0.179
Antankarana	32	0.915	32	0.846	32	0.573	24	0.183
Mikea	86	0.528	90	0.627	90	0.438	98	0.101
Non-African								
SA Indian	74	0.913	74	0.743	78	0.528	78	0.266
Indonesian-SS	88	0.663	86	0.473	88	**0.037***		**
Indonesian-SK	70	0.680	72	0.278	72	0.410		**
Polynesian	72	0.169	74	0.249	74	0.171		**
New Guinean	40	0.458	40	0.197	40	0.244		**
Solomon Island	46	0.846	46	0.825	46	0.082		**
AFRICAN								
Tsonga	72	**0.023***	66	0.527	80	0.461	100	0.772
Zambian	106	0.542	88	0.086	106	0.392	96	**0.027***
Mbenzele	88	0.204	88	0.965	88	0.325	86	0.398
Gbaya	70	0.505	64	0.809	70	0.707	66	0.528
Mbimou	22	0.505	22	0.672	22	0.101	22	0.693
S.Sangha	34	0.817	34	0.718	34	0.242	34	0.181
N Sotho (Pedi)							100	0.307

*significant deviations from Hardy-Weinberg expectation (P<0.050).

** data not available.

SS=South Sulawesi; SK=South Kalimantan.

3.4.2 Genetic diversity

The estimated genetic diversity (Nei, 1987) for each population-locus combination examined is given in **Table 3.6** together with the corresponding standard error estimate. Only one non-African population (SA Indian) was screened for variation at the $(CA)_n$ locus. Data for the other populations was not available. The genetic diversity is expected to be higher in a hybrid population than it is in any of the parental populations. This, however, may not be the case if one of the parental populations is African as the latter is thought to be older and has had a larger effective population size. An African population is most probably the parent of all other human populations according to the 'Out-of Africa' hypothesis (Stringer and Andrews, 1988) and the variation found in most non-African populations is often a subset of that seen in Africa.

The HUMCSF1PO locus shows moderate diversity with the average for the locus being 0.756. Diversities in individual population samples varied from 0.679 to 0.827. None of the Malagasy groups have diversity values appreciably different to those obtained for either African or Indonesian groups, but the presence of a few population specific alleles (discussed in **3.3.1**) at this locus increases its value for evolutionary studies.

The average diversity at the HUMTH01 locus is 0.741 and ranges from 0.538 in the Mahafaly to 0.800 in the Mikea and S.Sangha. This locus, like the previous one, shows only a moderate diversity with values mostly between 0.700 and 0.800. Due to the fairly uniform distribution of alleles at this locus in both African and the Indonesian population samples, no obvious conclusion as to Malagasy affinities can be drawn from the results obtained.

The allele frequency variation at the HUMCD4 locus is reflected by differences in diversities in different population samples: 0.820 to 0.914 in Africans, 0.631 to 0.860 in the Malagasy and 0.230 to 0.678 in other non-Africans. The average diversity at this locus is 0.730, which is lower than that noted in the African as well as nine of eleven Malagasy populations. The lower diversities in non-African populations are probably an effect of the limited number of alleles present in them. The two Melanesian populations had the lowest diversity values; 0.230 and 0.249 respectively and are approximately a third of those noted in other populations. These low levels may be partly due to the relatively small sample sizes (20 and 23 individuals respectively) or they may be a reflection of the evolutionary history of these two populations.

The $(CA)_n$ locus also shows a moderate diversity (the average is 0.752) with values ranging from 0.610 in the S.Sangha to 0.820 in the Antaimoro. Whereas the values in eight of eleven Malagasy are greater than the average, only one other population group (the Mbimou) displays a value greater than this. It is interesting that five of the six African groups have lower diversities than the Malagasy. This increased diversity in the Malagasy probably reflects admixture.

Table 3.6. Diversity (H) and standard error (SE) estimates for four autosomal STR loci in the study populations.

Population	HUMCSF1PO		HUMTHO1		HUMCD4		(CA)n	
	H	SE	H	SE	H	SE	H	SE
Malagasy								
Merina	0.757	0.025	0.785	0.025	0.631	0.038	0.742	0.030
Betsileo	0.786	0.092	0.765	0.103	0.749	0.096	0.776	0.108
Antaisaka	0.759	0.093	0.777	0.102	0.843	0.020	0.774	0.108
Tsimihety	0.759	0.091	0.742	0.105	0.807	0.033	0.738	0.030
Sakalava	0.795	0.083	0.718	0.099	0.792	0.088	0.749	0.103
Mahafaly	0.679	0.105	0.538	0.123	0.748	0.036	0.764	0.114
Antandroy	0.779	0.092	0.675	0.111	0.725	0.099	0.803	0.021
Betsimisiraka	0.787	0.089	0.729	0.103	0.772	0.091	0.806	0.089
Antaimoro	0.784	0.088	0.767	0.100	0.780	0.092	0.820	0.093
Antankarana	0.771	0.137	0.781	0.153	0.860	0.138	0.803	0.180
Mikea	0.720	0.082	0.800	0.090	0.809	0.084	0.784	0.077
Non-African								
Caucasoid*	0.747	0.041	0.771	0.045	0.678	0.045	**	
SA Indian	0.687	0.093	0.786	0.100	0.607	0.089	0.743	0.088
Indonesian	0.723	0.061	0.763	0.069	0.677	0.044	**	
Polynesian	0.749	0.091	0.782	0.100	0.611	0.100	**	
New Guinean	0.762	0.124	0.667	0.140	0.230	0.142	**	
Solomom Island	0.687	0.116	0.786	0.127	0.249	0.137	**	
African								
Tsonga	0.762	0.092	0.706	0.109	0.880	0.085	0.706	0.077
Zambian	0.798	0.076	0.729	0.091	0.864	0.074	0.744	0.027
Mbenzele	0.739	0.082	0.659	0.098	0.820	0.083	0.663	0.092
Gbaya	0.827	0.091	0.748	0.110	0.848	0.093	0.712	0.101
Mbimou	0.773	0.166	0.777	0.186	0.914	0.050	0.795	0.168
S.Sangha	0.763	0.137	0.800	0.146	0.902	0.148	0.610	0.153
Average diversity	0.756		0.741		0.730		0.752	

*Hammond *et al*, 1994.

** data not available.

3.4.3 Use of Exact test for population comparisons

A variation of Fisher's exact test (Raymond and Rousset, 1995) was used to test for significant pairwise differences in allele frequency distributions between populations examined in this study. The significance of the differences were assessed by comparisons with the results of 1000 permutations; *P* values of less than or equal to 0.050 were regarded as significant. Two sets of population groups were pooled for all further analyses: the two Indonesian groups were pooled as were the three non-Bantu-speaking Beya groups from the CAR (the Gbaya, Mbimou and S.Sangha). Exact test analysis showed no significant differences between the populations which were pooled.

The results of the pairwise (Exact) tests of differentiation at the HUMCSF1PO locus are given in **Table 3.7**. With one exception (the Mahafaly), none of the Malagasy groups show significant differences from other Malagasy groups, suggesting that they are all relatively homogeneous. The Malagasy groups also show varying degrees of similarities to both the Indonesian and African populations; the overall effect of this is clearer in the phylogeny shown in **section 3.5**.

Of the non-African groups, only the Indonesians show significant differences to other non-African groups (except to the Polynesian group). As would be expected, almost all comparisons between the non-African and African populations revealed significant differences between them. Interestingly, the New Guinean sample showed no significant difference when compared to the Zambian, Mbenzele and Beya populations, but was significantly different to the SA Indian, Indonesian and Solomon island populations. Of the

African populations, only the Mbenzele showed significant differences to other African populations.

Widely varying results of the Exact test were obtained with the HUMTHO1 polymorphism (**Table 3.8**). More within-group differences between the Malagasy ethnic groups were detected at this locus than with the HUMCSF1PO locus. The Sakalava, Mahafaly, Antaimoro and Antankarana differ significantly from at least six of the eleven Malagasy groups. Generally, the Malagasy differ significantly from the Caucasoid, SA Indian, Melanesian, Zambian and Mbenzele populations. They also show varying degrees of affinity with the Indonesian, Tsonga and Beya populations which possibly reflects varying degrees of contributions by Indonesians and Africans to the Malagasy gene pool.

All comparisons within the non-African populations and those between the non-African and African populations revealed significant differences between them. Of the six pairwise comparisons between the African groups, only the Zambian sample showed a significant difference to the Mbenzele and Beya.

Table 3.7. Results of pairwise tests of differentiation at the HUMCSF1PO locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

Population	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	Caucasoid	SA Indian	Indonesian	Polynesian	New Guinean	Solomon Is.	Tsonga	Zambian	Mbenzele	Biaka
MR	0.715	0.431	0.850	0.630	0.192	0.430	0.775	0.881	0.997	0.311	0.166	0.527	0.058	0.285	0.344	0.319	0.292	0.096	0.082	0.116
BT		0.149	0.483	0.843	**0.018**	0.505	0.405	0.742	0.857	0.137	0.162	**0.042**	**0.004**	0.099	0.334	**0.018**	**0.011**	**0.024**	0.053	**0.046**
AS			0.445	0.516	**0.026**	0.735	0.778	0.421	0.555	0.678	**0.006**	**0.011**	**<0.001**	**0.013**	0.292	**0.004**	0.452	**0.032**	0.716	0.404
TS				0.458	0.102	0.322	0.775	0.834	0.638	0.156	0.405	0.159	0.222	0.459	0.561	0.089	0.112	**0.019**	0.228	0.137
SK					**0.006**	0.982	0.656	0.889	0.779	0.113	**0.002**	**0.009**	**0.001**	**0.017**	0.375	**0.006**	0.072	0.112	0.115	0.421
MF						**0.002**	0.232	0.052	0.229	**0.003**	**0.020**	0.087	**0.008**	0.111	0.660	**0.007**	**0.037**	0.217	**0.008**	**0.039**
AD							0.450	0.602	0.652	0.182	**0.001**	**0.001**	**<0.001**	**0.004**	0.158	**0.008**	0.153	**0.030**	0.112	0.301
BS								0.954	0.832	0.148	0.079	0.085	**0.001**	0.062	0.829	**0.037**	0.258	0.141	0.536	0.692
AM									0.930	0.103	0.091	0.125	**0.007**	0.067	0.424	0.102	0.114	0.059	0.183	0.252
AK										0.407	0.377	0.613	0.110	0.284	0.302	0.401	0.404	0.268	0.124	0.227
MK											0.057	**0.008**	**0.007**	**0.006**	0.083	**0.002**	**0.011**	**0.001**	0.343	0.021
Caucasoid												0.203	0.034	0.325	0.360	0.056	**<0.001**	**<0.001**	**0.012**	**<0.001**
SA Indian													**0.045**	0.095	**0.031**	0.633	**0.011**	**0.002**	**0.001**	**<0.001**
Indonesian														0.616	**0.041**	**0.040**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Polynesian															0.216	**0.081**	**0.006**	**0.004**	**0.001**	**<0.001**
New Guinean																**0.004**	**0.048**	0.356	0.332	0.667
Solomon Is.																	**0.011**	**0.001**	**<0.001**	**<0.001**
Tsonga																		0.110	**0.021**	0.170
Zambian																			**0.001**	0.212
Mbenzele																				0.233

Table 3.8. Results of pairwise tests of differentiation at the HUMTHO1 locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

Population	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	Caucasoid	SA Indian	Indonesian	Polynesian	New Guinean	Solomon Is.	Tsonga	Zambian	Mbenzele	Bery
MR	0.289	0.388	0.278	0.132	0.004	0.372	0.599	0.020	0.056	0.640	<0.001	0.009	0.525	0.153	<0.001	0.019	0.054	0.004	0.012	0.384
BT		0.994	0.186	0.009	<0.001	0.031	0.129	0.018	0.237	0.415	<0.001	<0.001	0.017	<0.001	<0.001	0.010	0.227	0.056	0.042	0.609
AS			0.091	0.003	<0.001	0.023	0.136	0.015	0.215	0.493	<0.001	<0.001	0.026	0.008	<0.001	0.011	0.183	0.033	0.016	0.441
TS				0.778	0.011	0.362	0.369	0.039	0.040	0.154	<0.001	<0.001	0.177	0.001	<0.001	0.013	0.162	0.037	0.392	0.816
SK					0.038	0.471	0.359	0.016	<0.001	0.017	<0.001	<0.001	0.053	0.001	<0.001	0.003	0.011	0.001	0.112	0.124
MF						0.173	0.161	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.001	<0.001	<0.001	0.001	<0.001	0.032	<0.001
AD							0.542	<0.001	<0.001	0.019	<0.001	<0.001	0.084	0.002	<0.001	<0.001	0.140	<0.001	0.276	0.121
BS								0.034	0.002	0.171	<0.001	0.001	0.035	0.037	<0.001	0.012	0.021	0.001	0.042	0.259
AM									0.009	0.131	<0.001	<0.001	0.000	0.001	<0.001	0.728	<0.001	<0.001	<0.001	0.023
AK										0.194	<0.001	0.004	0.066	<0.001	0.018	0.044	0.093	0.481	0.005	0.111
MK											<0.001	0.009	0.101	0.005	<0.001	0.206	0.015	0.001	0.001	0.377
Caucasoid												0.010	<0.001	0.042	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
SA Indian													0.050	0.004	<0.001	0.011	<0.001	<0.001	<0.001	0.001
Indonesian														0.002	<0.001	0.002	0.031	0.004	0.003	0.050
Polynesian															<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
New Guinean																0.009	<0.001	<0.001	<0.001	<0.001
Solomon Is.																	<0.001	<0.001	<0.001	0.018
Tsonga																		0.487	0.644	0.218
Zambian																			0.048	0.050
Mbenzele																				0.112

Table 3.9 shows the Exact test results for comparisons at the HUMCD4 locus. With one exception (MR vs. SA Indian), the Malagasy groups show significant differences to the Caucasoid, SA Indian, Indonesian, Polynesian, Mbenzele and Beya populations. This suggests that despite the hypothesis of Africa and Indonesia being the major contributors to the gene pool of the Malagasy, the latter have had sufficient admixture and/or have diverged sufficiently enough (which seems less likely) for their allele frequencies to differ significantly from their purported parental groups. Another explanation would be that some other contributing group has left its mark on the Malagasy gene pool, e.g. the Arabs.

Only the Merina and Mahafaly differ significantly from more than two other Malagasy groups, again suggesting a relative homogeneity in the gene pool of the Malagasy. While the two highland groups show significant differences when compared to the Tsonga and Zambian populations only three of the nine lowland groups differ significantly from these two African populations. This indicates a significant gene contribution to the lowlanders from African populations represented in this study. Almost all the comparisons between non-African groups and those between non-African and African groups yielded significant values. The Mbenzele again, is the only African group that differed significantly from the other African groups.

Table 3.9. Results of pairwise tests of differentiation at the HUMCD4 locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

Population	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	Caucasoid	SA Indian	Indonesian	Polynesian	New Guinean	Solomon Is.	Tsonga	Zambian	Mbenzele	Beya
MR	0.003	0.001	0.002	0.006	0.115	0.025	0.114	0.000	0.003	0.005	<0.001	0.150	0.001	<0.001	0.002	0.001	<0.001	<0.001	<0.001	<0.001
BT		0.102	0.298	0.380	0.002	0.796	0.027	0.197	0.027	0.157	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.014	0.010	<0.001	<0.001
AS			0.287	0.330	0.033	0.129	0.315	0.387	0.692	0.594	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.240	0.157	<0.001	0.002
TS				0.726	0.003	0.559	0.195	0.206	0.297	0.951	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.354	0.332	<0.001	0.039
SK					0.146	0.863	0.517	0.780	0.781	0.955	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.080	0.064	<0.001	0.010
MF						0.080	0.512	0.010	0.103	0.102	<0.001	0.002	<0.001	<0.001	<0.001	<0.001	0.002	<0.001	<0.001	<0.001
AD							0.187	0.584	0.101	0.624	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.042	0.015	<0.001	<0.001
BS								0.147	0.310	0.940	<0.001	0.011	<0.001	<0.001	<0.001	<0.001	0.005	<0.001	<0.001	<0.001
AM									0.649	0.549	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.066	0.103	<0.001	<0.001
AK										0.682	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.225	0.226	<0.001	0.011
MK											<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	0.170	0.153	<0.001	0.006
Caucasoid												<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
SA Indian													0.013	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001	<0.001
Indonesian														<0.001	0.009	<0.001	<0.001	<0.001	<0.001	<0.001
Polynesian															0.004	0.005	<0.001	<0.001	<0.001	<0.001
New Guinean																0.522	<0.001	<0.001	<0.001	<0.001
Solomon Is.																	<0.001	<0.001	<0.001	<0.001
Tsonga																		0.820	0.001	0.498
Zambian																			<0.001	0.663
Mbenzele																				0.001

Exact test results for the $(CA)_n$ locus are given in **Table 3.10**. The two highland groups (Merina and Betsileo) and the Mahafaly differ significantly from most other Malagasy groups examined. The Mahafaly is also the only Malagasy group that shows a significant difference when compared to the Indonesian sample. As noted with the other loci, all comparisons within the non-African populations and those between the non-African and African populations yielded significant differences. None of the African pairwise comparisons yielded significant differences.

Table 3.10. Results of pairwise tests of differentiation at the DRD2-(CA)n locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations for the Malagasy groups are as given in the Nomenclature section.

Population	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	SA Indian	Indonesian	Tsonga	Zambian	Mbenzele	Beya
MR	0.624	0.071	0.009	0.009	<0.001	0.040	<0.001	0.004	0.019	0.001	<0.001	0.310	<0.001	<0.001	<0.001	<0.001
BT		0.066	0.030	0.012	<0.001	0.051	0.129	0.029	0.034	0.002	<0.001	0.669	<0.001	<0.001	<0.001	<0.001
AS			0.294	0.791	0.024	0.775	0.009	0.456	0.205	0.247	<0.001	0.660	0.030	0.040	0.041	0.017
TS				0.887	<0.001	0.205	0.090	0.228	0.371	0.105	<0.001	0.710	0.139	0.076	0.036	0.313
SK					0.002	0.646	0.030	0.483	0.521	0.482	<0.001	0.480	0.437	0.349	0.231	0.333
MF						0.006	<0.001	0.002	0.008	0.000	<0.001	0.010	<0.001	<0.001	0.003	<0.001
AD							0.092	0.768	0.735	0.712	<0.001	0.390	0.027	0.231	0.091	0.076
BS								0.243	0.380	0.031	0.005	0.310	0.001	0.004	0.000	0.002
AM									0.653	0.289	<0.001	0.550	0.028	0.084	0.011	0.010
AK										0.698	<0.001	0.180	0.370	0.844	0.368	0.779
MK											<0.001	0.170	0.076	0.585	0.028	0.063
SA Indian												0.050	<0.001	<0.001	<0.001	<0.001
Indonesian													0.050	0.040	0.040	0.060
Tsonga														0.533	0.206	0.187
Zambian															0.401	0.439
Mbenzele																0.375

3.5 Phylogenetic analyses

3.5.1 Genetic distances

A large number of programs have recently become available for calculating genetic distances from allele frequency data. The choice of the genetic distance to be used depends to some extent on the type of genetic data being analyzed, i.e. whether the infinite allele model or the stepwise mutation model applies [see **section 2.3(g)**]. Other factors that influence the choice of program are the time since the groups being studied are likely to have separated from each other and the sample sizes.

Takezaki and Nei (1996) studied the efficiencies of various genetic distance measures [both traditional (D_C and D_A) and new ($\delta\mu^2$ and D_{SW})] in phylogenetic reconstruction by using computer simulation. They used both the infinite allele model (IAM) and the stepwise mutation model (SMM), for classical markers and microsatellite loci respectively. They found that for both the IAM and SMM, D_C (Cavalli-Sforza and Edward's chord distance) and D_A, [Nei *et al*'s modification of Cavalli-Sforza's distance, (Nei *et al*, 1983)] generally have higher probabilities of obtaining the correct tree topology than other distance measures. Previous computer simulation studies by Nei *et al* (1983) with the IAM showed that D_A is more efficient than D_S (Nei's standard), D_M (Nei's minimum) and D_R (Roger's distance) under the conditions selected.

Both $\delta\mu^2$ (Goldstein *et al*, 1995) and D_{SW} (Shriver *et al*, 1995) increase linearly with time under the SMM unlike D_M and D_S which are nonlinear with increasing time when either the time since divergence or the mutation rate is large. Takezaki and Nei's study (1996) suggests that both $\delta\mu^2$ and D_{SW} are generally less useful than D_C and D_A for making

phylogenetic inferences. They suggest that trees designed with D_A may be sufficient for many purposes unless the evolutionary time span is very long.

Genetic distances between the study populations were calculated using allele frequency data and the computer program DSW (Ota, 1996). This program is able to generate Nei's D_A and D_S distances as well as Shriver's D_{SW} distance from a single input file. Although both the D_A and D_{SW} distances were computed during the present study, D_{SW} distances were used in the final analysis for reasons discussed below. The pairwise genetic distance matrix generated using the four STR loci is given in **Table 3.11**. The smaller the genetic distance between a pair of population samples, the closer the two populations are likely to be to each other in a genetic sense (ignoring sampling error).

The genetic distances between the various Malagasy groups are generally small (0.000 to 0.099) compared to those between them and other population groups (0.001 to 0.249). The apparent distance between the two highland groups, the Merina and Betsileo is zero, as is that between many of the lowland groups. This tentatively supports a recent common ancestry for all eleven Malagasy groups or high levels of gene flow between them.

The genetic distances between the highlanders and the Indonesians are lower than those between the lowlanders and the Indonesians (0.001 and 0.020 for the Merina and Betsileo respectively vs. a range of 0.020 to 0.095 for the lowlanders). On the other hand, the genetic distances between the highlanders and the African groups (e.g. Zambians) are higher than those between lowlanders and the African groups; values of 0.123 and 0.097 are obtained for the Merina and Betsileo respectively compared to a range of 0.006 to 0.120 for lowland groups. This implies a closer association of the highland groups with the

Indonesians and a closer association of the lowland groups with the African populations. The Mbenzele shows the largest distances from the Malagasy. These varying affinities between populations are portrayed more effectively by means of a NJ tree.

3.5.2 Phylogenetic trees

The program DSW is also able to draw phylogenetic trees by the NJ method from the distance matrices it generates. The program can also provide statistical support for the branching in the tree by bootstrapping.

Trees generated from both the D_A and D_{SW} distance matrices gave the expected initial separation of populations. The Malagasy groups fell between the African and non-African branches regardless of whether D_A or D_{SW} was used. Only the order of the Malagasy groups between the African and non-African groups varied between the two trees. It was thought that a possible indicator of the 'better' tree could be obtained by summing the bootstrap values which support their branches. The sums of the bootstrap values for both trees however were found to be almost identical. At this early stage in the study of the ancestry of the Malagasy, there seems to be insufficient data on which to base a choice between the D_A and D_{SW} trees. The D_{SW} distance measure was finally chosen for no reason other than that it has been specifically designed for use with STR loci.

Figure 3.3 depicts the unrooted NJ tree obtained using the D_{SW} distances shown in **Table 3.11**. The SA Indian and Mbenzele groups form the two extremes of the tree, both having very long branch lengths. The two highland groups (Merina and Betsileo) are closest to the Indonesian and SA Indian populations with the Merina clustering with the SA Indians.

Indonesians and a closer association of the lowland groups with the African populations. The Mbenzele shows the largest distances from the Malagasy. These varying affinities between populations are portrayed more effectively by means of a NJ tree.

3.5.2 Phylogenetic trees

The program DSW is also able to draw phylogenetic trees by the NJ method from the distance matrices it generates. The program can also provide statistical support for the branching in the tree by bootstrapping.

Trees generated from both the D_A and D_{SW} distance matrices gave the expected initial separation of populations. The Malagasy groups fell between the African and non-African branches regardless of whether D_A or D_{SW} was used. Only the order of the Malagasy groups between the African and non-African groups varied between the two trees. It was thought that a possible indicator of the 'better' tree could be obtained by summing the bootstrap values which support their branches. The sums of the bootstrap values for both trees however were found to be almost identical. At this early stage in the study of the ancestry of the Malagasy, there seems to be insufficient data on which to base a choice between the D_A and D_{SW} trees. The D_{SW} distance measure was finally chosen for no reason other than that it has been specifically designed for use with STR loci.

Figure 3.3 depicts the unrooted NJ tree obtained using the D_{SW} distances shown in **Table 3.11**. The SA Indian and Mbenzele groups form the two extremes of the tree, both having very long branch lengths. The two highland groups (Merina and Betsileo) are closest to the Indonesian and SA Indian populations with the Merina clustering with the SA Indians.

This supports historical data as to the closeness of the highland groups to Indonesian populations. The three lowland groups that are closest to the African populations are the Mikea, Antankarana and Antaisaka. The remaining Malagasy groups appear to be closer to the two Asian populations. Historical data suggests that the Mikea have the greatest African affinity compared to the other Malagasy groups. Their placement on the tree supports an African affinity but the Antaisaka appear to be closest to the African populations. Very little has been written about the Antaisaka but their apparent closeness to Africans especially to the Zambians was also supported by haplotype studies at the G6PD locus (Dangerfield, 1997). The fact that many of the lowland Malagasy appear closer to the Indonesian rather than the African populations is contrary to the perception that the highlanders are more Indonesian and the lowlanders more African. There is weak statistical support (55%) for the separation of the African populations from the rest and the splits between the SA Indian, Merina and Betsileo from the rest (64%) but less support (<50%) for almost all separations between the other Malagasy groups.

The branch lengths connecting most of the Malagasy are very short suggesting again their recent common ancestry. The Indian, Mahafaly and Mbenzele groups are connected to the tree by long branches which suggests that they diverged relatively long ago or that they have experienced a more rapid rate of evolution.

Table 3.11. Genetic distances between study populations calculated using allele frequency data at four autosomal STR loci. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

Population	MR	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	SA Indian	Indonesian	Tsonga	Zambian	Mbenzele
BT	0.000															
AS	0.052	0.039														
TS	0.026	0.013	0.009													
SK	0.041	0.024	0.011	0.000												
MF	0.056	0.064	0.064	0.041	0.052											
AD	0.025	0.008	0.020	0.000	0.000	0.055										
BS	0.022	0.020	0.005	0.006	0.009	0.022	0.013									
AM	0.037	0.024	0.017	0.003	0.000	0.026	0.009	0.000								
AK	0.035	0.026	0.003	0.000	0.019	0.099	0.028	0.033	0.035							
MK	0.055	0.042	0.000	0.002	0.001	0.081	0.009	0.019	0.012	0.000						
SA Indian	0.028	0.064	0.131	0.110	0.149	0.135	0.140	0.080	0.111	0.085	0.141					
Indonesian	0.001	0.020	0.070	0.027	0.046	0.095	0.040	0.056	0.061	0.020	0.053	0.051				
Tsonga	0.136	0.116	0.008	0.039	0.043	0.099	0.069	0.048	0.059	0.025	0.026	0.220	0.140			
Zambian	0.123	0.097	0.012	0.041	0.043	0.120	0.062	0.056	0.064	0.006	0.024	0.205	0.123	0.000		
Mbenzele	0.249	0.219	0.073	0.132	0.127	0.156	0.156	0.108	0.118	0.131	0.104	0.330	0.273	0.030	0.064	
Bera	0.154	0.126	0.015	0.059	0.057	0.115	0.083	0.048	0.064	0.060	0.044	0.233	0.184	0.001	0.013	0.022

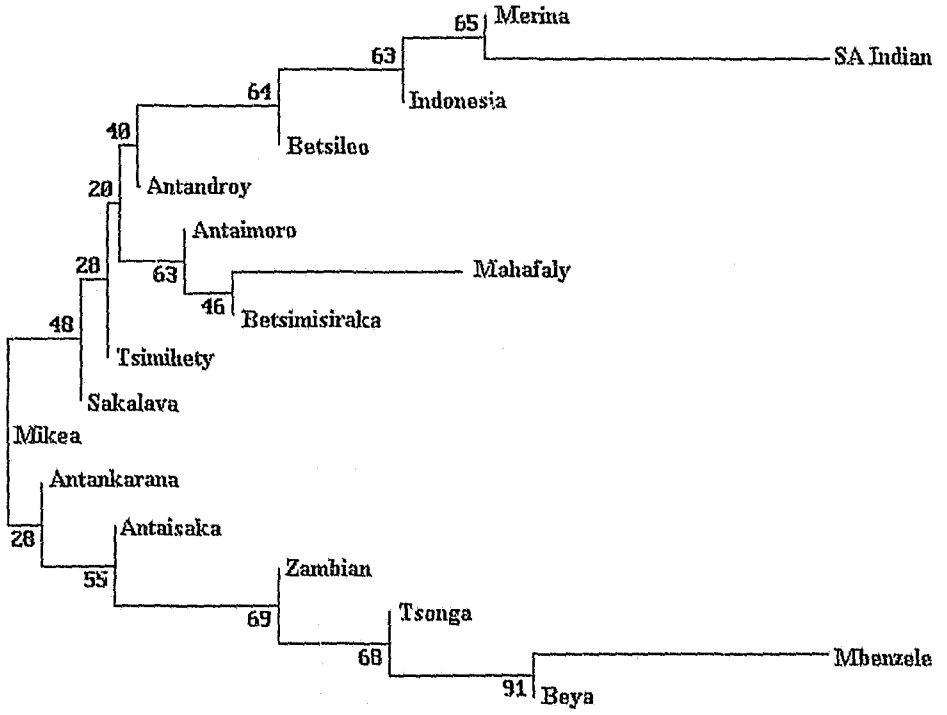


Figure 3.3. Unrooted NJ tree showing possible genetic relationships between several Malagasy, non-African and African populations based on D_{SW} genetic distances generated using allele frequency data from four STR loci [HUMCSF1PO, HUMTH01, HUMCD4 and a $(CA)_n$]. The numbers at each node indicate the percentage of trees (out of 1000 bootstrap replications) with such a node.

When only the results from the multiplexed STR loci were used to generate a NJ tree, it became possible to include published data on the same loci from other populations: namely a Caucasoid population (Edwards *et al*, 1992; Hammond *et al*, 1994) as well as a Polynesian and two Melanesian populations (data on the last three were generated in present study). The resulting unrooted NJ tree is depicted in **Figure 3.4** and allows assessment of the affinities of the Malagasy to Melanesian and Polynesian populations. Two main branches are again observed with the African populations clustering at one end and the non-Africans at the other. The Malagasy cluster closer to the Africans than to the Melanesians who are at the other extreme of the tree, clearly showing little support for Grandidier's theory (which was presented very early on in the search for Malagasy ancestry) of a Melanesian ancestry for the darker skinned Negroid featured Malagasy (Campbell, 1995). The placement of some Malagasy groups are different to that seen in **Figure 3.3** (which includes data from four STR loci and a smaller number of populations in the analysis) and this could be due to the inclusion of data on additional populations or the reduction in the number of loci examined.

Mitochondrial DNA data based on the 9bp deletion in conjunction with control region sequence data (Soodyall *et al*, 1995) suggests that the Malagasy may be closely related to Polynesian or pre-Polynesians. This is borne out by the results depicted in **Figure 3.4**.

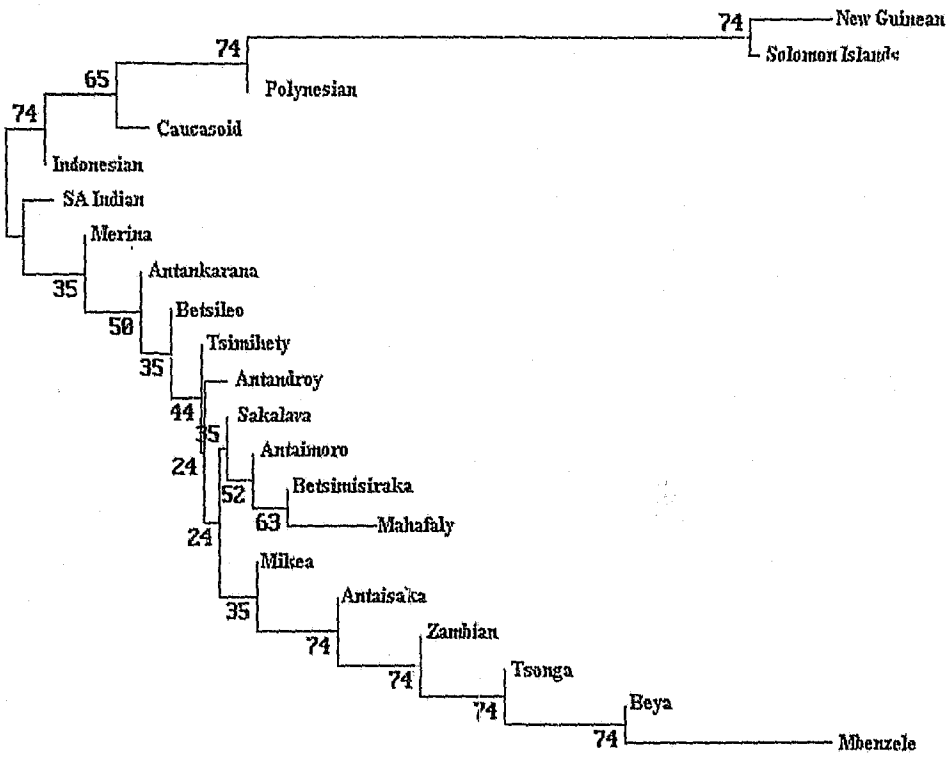


Figure 3.4. Unrooted NJ tree showing possible genetic relationships between several Malagasy, non-African and African populations based on D_{SW} genetic distances generated using allele frequency data from three autosomal STR loci (HUMCSF1PO, HUMTH01 and HUMCD4). The numbers at each node indicate the percentage of trees (out of 1000 bootstrap replications) with such a node.

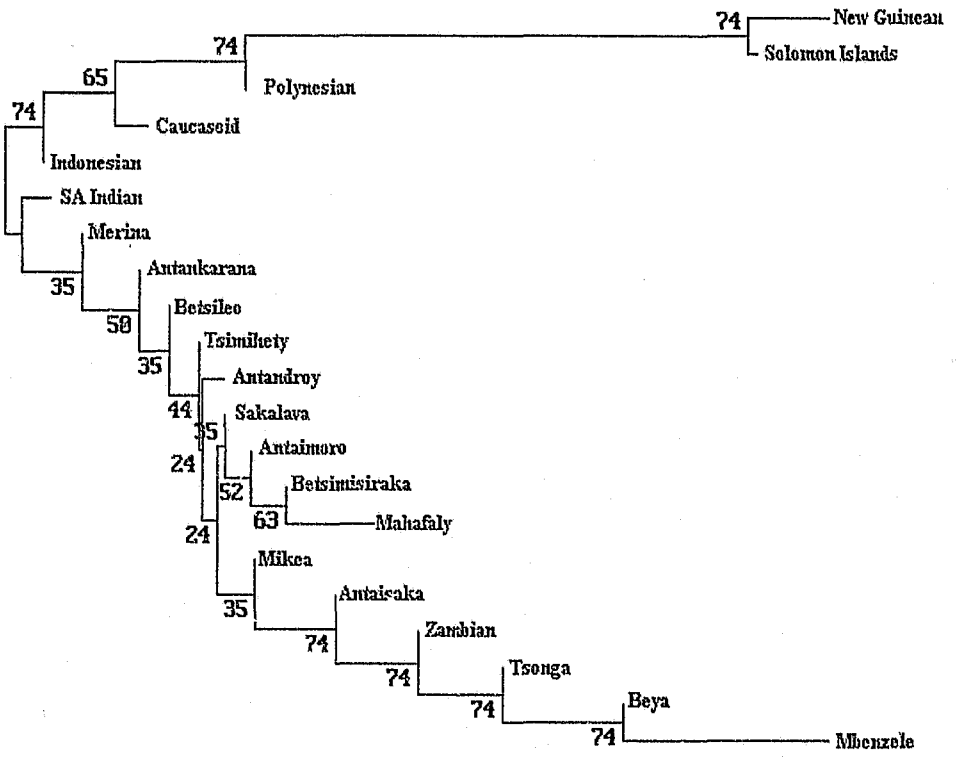


Figure 3.4. Unrooted NJ tree showing possible genetic relationships between several Malagasy, non-African and African populations based on D_{SW} genetic distances generated using allele frequency data from three autosomal STR loci (HUMCSF1PO, HUMTH01 and HUMCD4). The numbers at each node indicate the percentage of trees (out of 1000 bootstrap replications) with such a node.

3.6 Principal Components Analysis

Figure 3.5 displays a two dimensional plot that results from the principal components analysis (PCA) based on allele frequency data from the HUMCSF1PO, HUMTH01 and HUMCD4 loci. A number of populations examined for these three loci lacked data for the $(CA)_n$ locus and this locus was thus excluded from the PCA.

Of the various components generated by the analytical program (20 for the present data set), the first axis/component is usually much more significant than the rest. The population affinities seen in **Figure 3.5** are based on 33.2% of the variation present within the entire sample. The remaining variation is accounted for by 18 other components (each one accounting for <9% of the total variation). The X-axis separates the Africans from the non-Africans but the Y-axis shows that there is additional variation among the non-African groups. This variation however, does not seem to have any concordance with geography. The Malagasy are placed in a tight cluster intermediate between the Africans and non-Africans on both axes. The Merina are placed closest to the non-African cluster showing a closer affinity with the New Guineans rather than with the Indonesians. The Y-axis also highlights differences between the Merina and Betsileo, the two highland groups. Of the lowland groups examined, the Antaisaka rather than the Mikea appear closer to the African populations. The same relationships between the populations that are evident here were seen in **Figure 3.3**, and both support other types of data that suggests that the Malagasy gene pool received contributions from both African and Asian populations. Surprisingly, the Malagasy do not cluster closest to the Indonesian populations examined, and at this stage the reason for this is unclear.

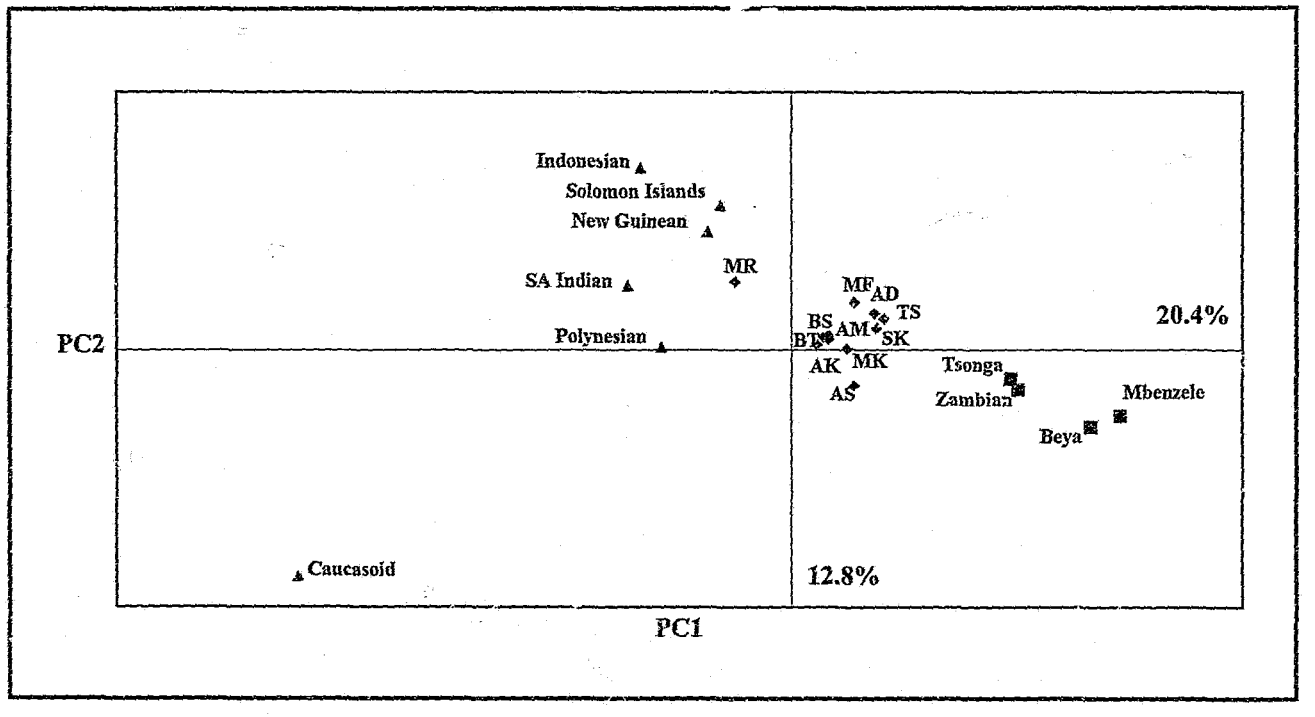


Figure 3.5. Two dimensional principal component plot showing population affinities based on the combined allele frequency data from three STR loci (HUMCSF1PO, HUMTH01 and HUMCD4). Blue diamonds represent Malagasy groups, red triangles represent non-Africans and green squares represent Africans. Abbreviations for the Malagasy are as defined in the Nomenclature section.

3.7 Distance from the Centroid

Previous studies have demonstrated that under the island model, a simple linear relationship is expected between the mean heterozygosity of a population and the genetic distance of that population from the overall mean gene frequency (centroid) of the population [Harpending and Ward, 1982; see **section 2.3 (j)**]. Empirical studies (Harpending and Ward, 1982) have shown this relationship to hold. This model was applied to the data obtained in the present study to assess the relative amounts of gene flow experienced by the different Malagasy groups. A plot of heterozygosity vs distance from the centroid for the populations in this study is shown in **Figure 3.6**.

The predicted 'theoretical regression' line is shown in pink. Except for three Malagasy groups (Merina, Antandroy and Mahafaly), eight of the eleven and the Caucasoid population cluster around the theoretical regression showing a good fit between the observed relationship and that predicted by the model. The close association of the Malagasy groups, and the fact that they are the shortest distance away from the centroid, supports their recent origin from the same proto-Malagasy ancestors who presumably carried in their genomes contributions from African and Indonesian parental populations. Historical data (Dewar and Wright, 1993) as well as genetic studies by Hewitt (1998) suggests that the southwestern groups of Madagascar are descendants in varying proportions of several major African and Asian populations whereas the highland groups are more homogeneous. From **Figure 3.6** it appears that the two southwestern Malagasy groups examined (Antandroy and the Mahafaly) are equally or even more homogeneous than the two highland groups (Merina and Betsileo) examined.

All the African populations appear above the theoretical regression line in the figure. Two factors may have contributed to this observation: gene admixture and a large effective population size. The latter has been demonstrated in Africans for a number of genetic systems (Tishkoff *et al*, 1996; Stoneking *et al*, 1997). All the non-Africans appear below the theoretical regression line, and are thought to be the least genetically diverse because of minimum gene flow effects or due to genetic isolation. It is well known that non-Africans have less genetic variation than Africans and this finding is not surprising.

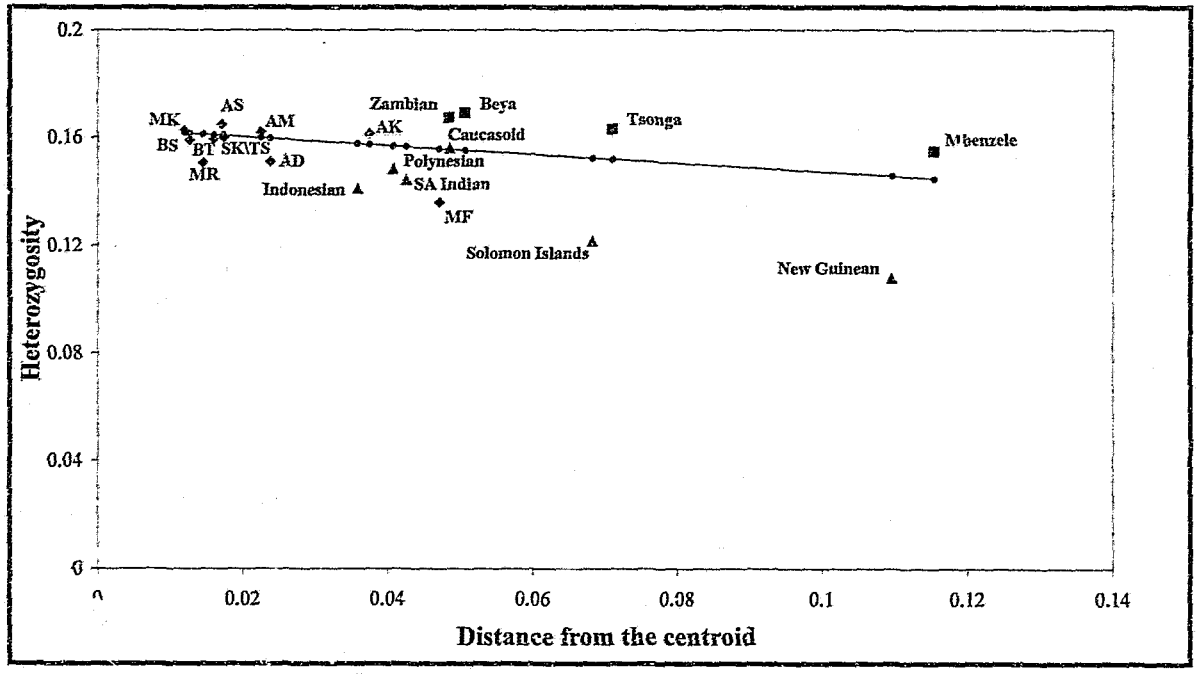


Figure 3.6. Plot of heterozygosity vs. distance from the centroid based on the combined allele frequency data from three STR loci (HUMCSF1PO, HUMTH01 and HUMCD4). Blue diamonds represent Malagasy ethnic groups, red triangles represent non-Africans and green squares represent Africans. Abbreviations for the Malagasy groups are as defined in the Nomenclature section.

3.8 Admixture calculations

On the assumption that the progenitors of the Malagasy consisted of only Africans and Indonesians, the least squares method of Elston (1971) was used to estimate the proportional contributions of these two groups. This approach allows the estimation of admixture proportions using allele frequency data on one or more loci. Two further assumptions were made: firstly, that admixture affects all loci equally and simultaneously, whether gene flow occurs in large amounts over a few generations or in small amounts over many generations, and secondly, that the allele frequencies in the parental populations and the 'hybrid' population have not drifted since the time when admixture occurred. None of these assumptions are strictly valid; there has been gene flow from groups other than the two mentioned and sampling effects are inevitable.

Inferences made using genetic data should also be viewed in the light of social, cultural, ethnohistorical and other available evidence. An example displaying the importance of combining these types of data with genetic data to infer gene flow into a population is discussed in **CHAPTER 5** using the Y *Alu* polymorphism.

Calculations were performed using African (pooled data from the Tsonga and Zambian populations) and Indonesian populations as the two parental populations. Only two of the four African populations were used in the analyses because of their geographic proximity to the East Coast of Africa. Historical evidence also suggests that the slaves taken to Madagascar from Africa included the regions from which the Tsonga and Zambian samples were collected (Campbell, 1995). Ideally, Arab populations should also have been included, but at present this was not possible because of the lack of both genetic material

and comparative data in the literature. The highland and lowland populations represented the hybrid populations in the analysis.

Using allele frequency data from all four loci (33 alleles), the African contribution to the highland population of Madagascar is estimated to be 32.45% and 58.80% to the lowland group. The African contribution in the lowland groups appear to be almost twice that of the highland populations and this disproportionate contribution is supported by historical and other genetic data (Hewitt *et al*, 1996).

3.9 Summary

The allele frequency distributions at four STR loci [(HUMCSF1PO, HUMTH01, HUMCD4 and a (CA_n)] were determined in eleven Malagasy, six African, two Austronesian (Indonesia and Polynesia) and two Melanesian samples (Papua New Guinean and Solomon Island). Because of the differential distributions of alleles at three of the four loci [HUMCSF1PO, HUMCD4 and a $(CA)_n$], a contribution from African as well as non-African populations in the Malagasy could be detected. However, because of the wide distribution of most alleles at STR loci, specific populations cannot be identified as having contributed to the Malagasy gene pool. The variation of modal allele/s noted in most of the Malagasy ethnic groups between those that are modal in Africa and/or Indonesia strongly suggests a contribution by both to the present Malagasy gene pool.

Exact tests used to test for significant pairwise differences in allele frequency data show very few significant differences between Malagasy groups. Varying degrees of similarities are detected between Malagasy and both African and non-African populations, with many

significant differences even with the Indonesians. Possible explanations for this could be that the gene pool of the present day Malagasy has received genetic contributions from populations other than those examined in this study, or, that they have diverged significantly from their purported African and Indonesian parental populations.

The combined allele frequency data from the four loci were used to infer genetic relationships between the populations examined using NJ trees (**Figures 3.3** and **3.4**) and PCA (**Figures 3.5**). NJ trees were generated using both D_{SW} and D_A distance measures. Both suggest that there is a closer genetic affinity between the Indonesians and the highland groups (Merina and Betsileo). The PCA however, places only the Merina group closer to the non-African cluster. The two highland groups are separated based on the variation represented on the Y-axis. The lowland groups appear to have a closer affinity with African populations, especially with the Zambians. The consistent tight clustering of the Malagasy on the PC plots and their short branch lengths on the NJ trees suggests that they have a recent common ancestor. Although there is clear evidence of a major Asian contribution from both the NJ tree and the PC plot, Indonesia does not appear to be the closest parental contributor as previously suggested from historical data. In this respect, it is interesting to note that the SA Indian population seems to have closer genetic affinities with the Malagasy than the Indonesian population studied.

The placement of the different Malagasy ethnic groups on the NJ tree and the PC plot suggests that geographically closer ethnic groups are not necessarily more genetically similar to each other than to groups located further away. This is especially obvious for the cattle-owning Mahafaly and the Antandroy who are neighbors in southern Madagascar and are both thought to be more African-like in terms of physical appearance and lifestyle. The

average genetic distance between these two groups (0.055) however, is larger than that between the other pairs of Malagasy groups (values between Antandroy and other Malagasy range from 0.000 to 0.028) (**Table 3.11**) and their placement in the phylogenetic trees in **Figures 3.3** and **3.4** clearly shows this. This reaffirms the notion that the genetic history of a population can be very different from both it's cultural and geographic history.

An analysis of the distance from the centroid displays the close clustering of the Malagasy to the theoretical regression line. This indicates that there have been no differential or unusually large amounts of gene flow recently into any of the Malagasy groups. In fact, two southwestern Malagasy groups who are thought to be more heterogeneous than highland populations appear below the theoretical regression line in the plot of heterozygosity vs. distance from the centroid. Their position in the plot supports a low variability in the two groups.

Admixture calculations using Elston's least squares method supports historical data of a larger African gene contribution to the lowlanders (58.80%) than to the highlanders (32.45%). Almost all of the analyses carried out using just four STR loci suggest a recent common ancestry for the Malagasy, supporting hypothesis 4 as proposed in **section 1.3.1.** The Malagasy show an affinity to both Bantu speaking populations from southeast Africa as well as to populations from Asia. Despite the wide distribution of alleles at the four STR loci examined, the data supports other types of data (historic, linguistic and genetic) for the ancestry of the Malagasy. Data from other loci and possibly other contributing populations e.g. Arabs, and additional east African populations may refine and give a more accurate picture of the genetic relationships between Malagasy subgroups and their likely ancestral populations.

CHAPTER 4 – AUTOSOMAL HAPLOTYPES

4.1 Introduction

Haplotypes consisting of alleles at multiple genetic markers have proved to be powerful tools for examining the evolutionary histories of populations (Tishkoff *et al*, 1996, 1998; Kidd *et al*, 1998). Depending on the types of polymorphic markers included in the haplotype, different time scales of history can be examined.

Identical restriction fragment length polymorphisms and insertion/deletion polymorphisms are unlikely to arise and reach polymorphic frequencies more than once in the history of humans (Kidd and Kidd, 1995). Alleles at these types of loci are most likely to be identical by descent in whatever population they may be found and frequency differences are most likely to be due to random genetic drift. Alleles at almost all RFLP loci are shared by most populations and therefore by themselves may not be very useful for following recent evolutionary/migratory events. Short tandem repeat loci on the other hand are very useful because of their higher mutation rates and diversity which increases the probability of observing large allele frequency differences between populations due to the effects of drift during population divergence.

Haplotypes consisting of both STRs and RFLPs or insertion/deletion polymorphisms combine the benefits of both types of markers; the high mutation rates of STR loci results in increased variation and the formation of new alleles and the stable flanking markers allow some certainty in tracing the lineage of each haplotype. The increased variability of these haplotypes also increases the likelihood of the occurrence of 'private' haplotypes in

particular populations. The detection of such 'private' haplotypes are ideal for following gene flow and tracing the origins of recently formed populations, e.g. the Malagasy.

Haplotypes consisting of alleles at loci in two autosomal regions have been selected for use in the search for the ancestors of the Malagasy. The first consists of two polymorphic loci within the CD4 gene on the short arm of chromosome 12 (Tishkoff *et al*, 1996; **Figure 2.2**). The second consists of four polymorphisms in and around the DRD2 receptor gene on chromosome 11 (Kidd *et al*, 1998; **Figure 2.3**). Previous studies from both sets of markers have indicated their variability in populations worldwide and emphasize their potential usefulness for evolutionary studies (Tishkoff *et al*, 1996; Kidd *et al*, 1998). This chapter examines the haplotypes at both regions in Malagasy, non-African and African populations and discusses the implications of frequency differences and/or similarities found between the populations.

4.1.1 Haplotypes at the CD4 locus

Two-site haplotypes were defined in terms of the alleles present at a STR locus (HUMCD4) and an *Alu* deletion polymorphism in the CD4 gene [**section 2.2.4.3(a)** and **Figure 2.2**). The two polymorphisms are 9.8kb apart and are assumed to be selectively neutral because of their location in non-coding regions of the gene. The STR locus has already been discussed in detail in **section 3.3.3**, but a few pertinent points will be re-emphasized. Twelve alleles (4-15 repeats) have been detected at this locus in human populations, the majority of which occur in African populations. Tishkoff *et al* (1996) screened individuals from 42 geographically dispersed populations (N=1600) and found

only three alleles (5, 6 and 10) at frequencies greater than 5% outside of Africa making this locus an ideal marker for tracing gene flow from Africa.

The *Alu* polymorphism results from the deletion of 256bp of a 285bp *Alu* element (Edwards and Gibbs, 1992). Studies on chimpanzees, gorillas, orangutans and gibbons indicate that the full-length *Alu* (+ allele) is the ancestral state at this locus with the deletion event having occurred after the divergence of humans and great apes about four to six million years ago. Sequencing results from several African and non-African individuals show that exactly the same sequence is deleted in all individuals (Tishkoff *et al*, 1996). Common ancestry can thus be assumed for all deletion chromosomes. Tishkoff *et al* (1996) have also shown that the deletion (- allele) is rare or absent in Asian, New World and Pacific island populations. In contrast, it occurs at frequencies of about 30% in European and Middle Eastern populations and at variable frequencies (7-28%) in African populations.

Tishkoff *et al* (1996) found that when haplotypes were constructed with alleles at the STR and *Alu* loci, only three haplotypes occurred at appreciable frequencies in non-African populations; the +5 (full length *Alu* sequence and allele 5 at the STR locus) and +10 haplotypes occurred at frequencies greater than 10% while haplotype –6 occurred at frequencies of between 20% and 34% in Europe and the Middle East. In their sample of more than 2200 unrelated non-Africans, they found that STR allele 6 was almost completely associated with the *Alu*- allele. Linkage disequilibrium values between these two alleles suggest that the *Alu* deletion originally occurred on a chromosome with allele 6 at the STR locus. Since this event, there has not been sufficient time for complete equilibrium to be attained (either by crossover events or mutation) in these populations.

Among sub-Saharan populations, the *Alu*- allele is found associated with several STR alleles.

Lower heterozygosity levels were also noted in the non-Africans (0.28-0.72 compared to 0.81-0.90 in Africa) with sub-Saharan African populations showing considerably more variation (more haplotypes) than non-African populations. This difference especially between African and Asian populations has been exploited for examining Malagasy ancestry.

4.1.2 Haplotypes at the DRD2 locus

The dopamine D2 receptor locus (DRD2) is one of five different dopamine receptor genes that have been identified in humans. Several polymorphisms have been identified in the DNA encompassing the coding sequence of this gene which is located on chromosome 11q22-11q23 (Gejman *et al*, 1994). This genetic variation is of immense interest because of the reported association of neutral variants in the non-coding region of the gene with different disorders (primarily alcoholism and substance abuse) (Gejman *et al*, 1994). Four of these polymorphisms have been used by Kidd *et al* (1998) to examine the evolution of haplotypes at the DRD2 locus. The four loci span approximately 25kb [**section 2.2.4.3(b)** and **Figure 2.3**)] and includes a CA repeat polymorphism and three *Taq*I RFLPs.

Considerable variation has been noted at the CA repeat locus in the DRD2 gene (Castiglione *et al*, 1995) and has already been discussed in **section 3.3.4.** Data for the *Taq*I B, D and A loci in the great apes indicate that the sites are not polymorphic in them. The B1, D1 and A1 alleles lack the *Taq*I site while the B2, D2 and A2 alleles have the site.

Alleles B2, D2 and A1 are the only ones seen in chimpanzees, gorillas and orangutans (Iyengar *et al*, 1998). These three alleles are identical in sequence in humans and the great apes and thus have been hypothesized to be ancestral alleles. All three systems are polymorphic in almost all human populations studied (Castiglione *et al*, 1995; Kidd *et al*, 1998) with allele frequencies at each of the loci being similar in popu' ons belonging to the same geographic region.

Kidd *et al* (1998) examined a large number of individuals from 28 distinct populations around the world (seven African, five European, eight East Asian, one Australomelanesian, four North American and three South American) for the four loci. They found that African populations had more haplotypes per population and more haplotypes at frequencies of 5% or more than the other populations examined. Of the 37 haplotypes they detected, four accounted for more than 70% of chromosomes in almost all non-African populations (discussed in detail later, shown in yellow in **Table 4.4**). Two of these accounted for more than 70% of all chromosomes in most East Asian and Amerindian populations, one of which occurs at a much lower frequency in African populations. This is most relevant to the present study. Haplotype analysis by Kidd *et al* (1998) also showed highest average heterozygosities in Africa followed by Europe. Lowest values were obtained for East Asia and the Americas

4.2 Allele identification

4.2.1 HUMCD4 STR alleles

The identification of alleles at this locus has been discussed in **section 3.2.**

4.2.2 CD4-*Alu* alleles

Figure 4.1 is a photograph of a 1% agarose gel in which the PCR products from the CD4-*Alu* locus have been separated. A 1501bp fragment represents the ancestral allele (*Alu*+) while a 1245bp fragment represents the deletion allele (*Alu*-). Heterozygotes are easily identified (lane 2) as are homozygotes for either the *Alu*+ (lanes 3, 4, 6, 7 and 8) or *Alu*– (lane 1) alleles.

4.2.3 $(CA)_n$ alleles

The identification of alleles at this locus has been described in **section 3.2.**

4.2.4 *Taq*I alleles

Figures 4.2, 4.3 and **4.4** are photographs of 2% agarose gels in which the digested PCR products from the *Taq*I B, D and A loci have been separated. Lanes 2 and 5 in **Figure 4.2** (*Taq*I B locus) show B2 homozygotes, lanes 3 and 4 show B1/B2 heterozygotes and lane 6 shows a B1 homozygote. Both digest products (267bp and 192bp) representing the B2 allele are clearly visible on the gel and are distinct from the 459bp B1 (uncut) allele. The

fragments of a 1kb molecular weight marker can be seen in lanes 1, 1 and 7 in **Figures 4.2, 4.3** and **4.4** respectively.

In **Figure 4.3** (*Taq*I D locus), lanes 2-5, 9 and 10 show D2 homozygotes and lanes 6 and 8 D1/D2 heterozygotes. The D1 allele which lacks the *Taq*I site is clearly separated from the D2 allele which is represented by the 212bp fragment. The second digest product of 88bp is difficult to see on the gel.

Lanes 1, 4 and 5 in **Figure 4.4** (*Taq*I A locus) show A1/A2 heterozygotes, lanes 2 and 3 show A2 homozygotes and lane 6 shows an A1 homozygote. The digest products of 124bp and 112bp are not resolved and appear as a single band which is clearly separated from the uncut 236bp fragment.

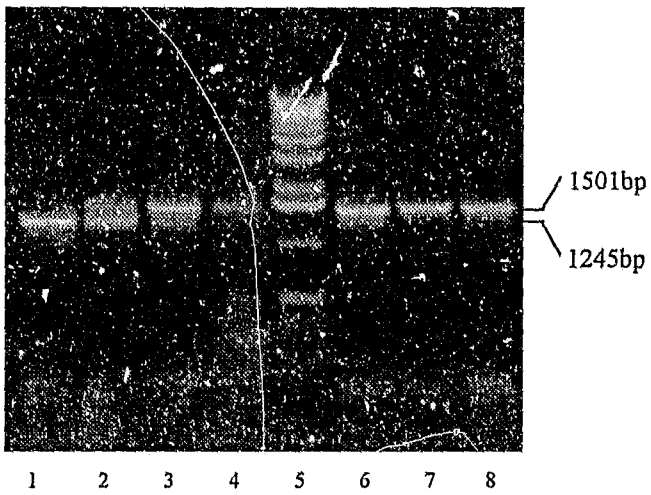


Figure 4.1. Photograph of a 1% agarose gel showing the separation of *Alu*+ and *Alu*- alleles at the CD4 locus. Lane 1 shows a -/- homozygote, lane 2 shows a +/- heterozygote and lanes 3, 4, 6, 7 and 8 show +/+ homozygotes. Lanes 5 shows fragments of the 1kb DNA molecular weight marker (Gibco BRL).

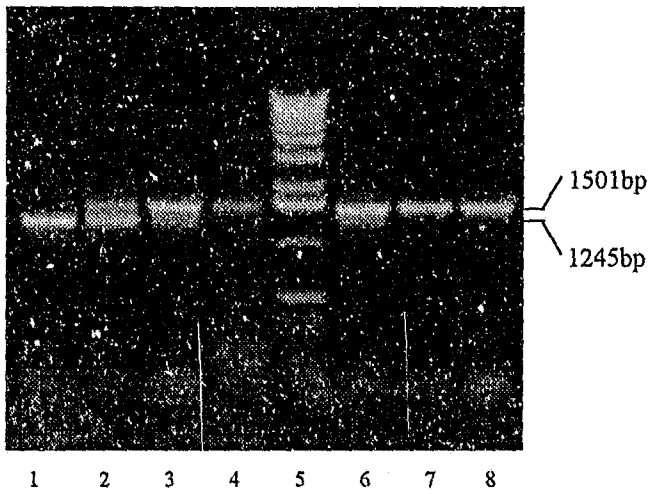


Figure 4.1. Photograph of a 1% agarose gel showing the separation of *Alu*+ and *Alu*- alleles at the CD4 locus. Lane 1 shows a -/- homozygote, lane 2 shows a +/- heterozygote and lanes 3, 4, 6, 7 and 8 show +/+ homozygotes. Lanes 5 shows fragments of the 1kb DNA molecular weight marker (Gibco BRL).

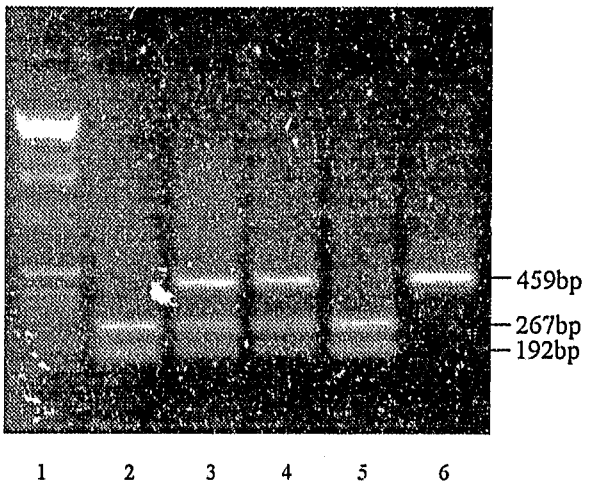


Figure 4.2. Photograph of a 2% agarose gel showing the separation of alleles at the *TaqI* B locus in the DRD2 gene. Lanes 2 and 5 show B2 homozygotes, lanes 3 and 4 show B1/B2 heterozygotes and lane 6 shows a B1 homozygote. Lane 1 show fragments of the 1kb DNA molecular weight marker (Gibco BRL).

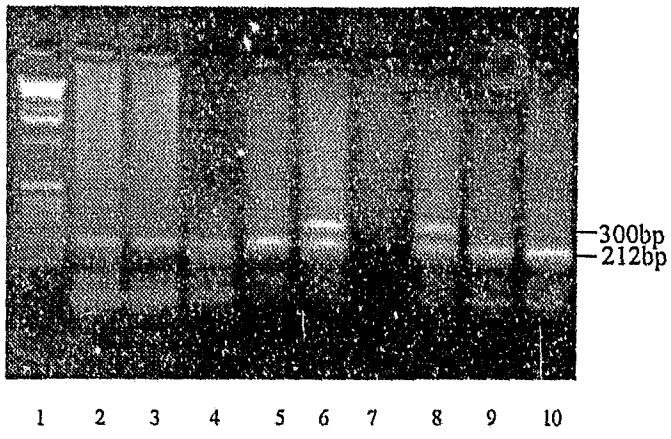


Figure 4.3. Photograph of a 2% agarose gel showing the separation of alleles at the *Taq*I D locus in the DRD2 gene. Lanes 2, 3, 4, 5, 9 and 10 show D2 homozygotes and lanes 6 and 8 show D1/D2 heterozygotes. Lane 1 shows fragments of the 1kb DNA molecular weight marker (Gibco BRL).

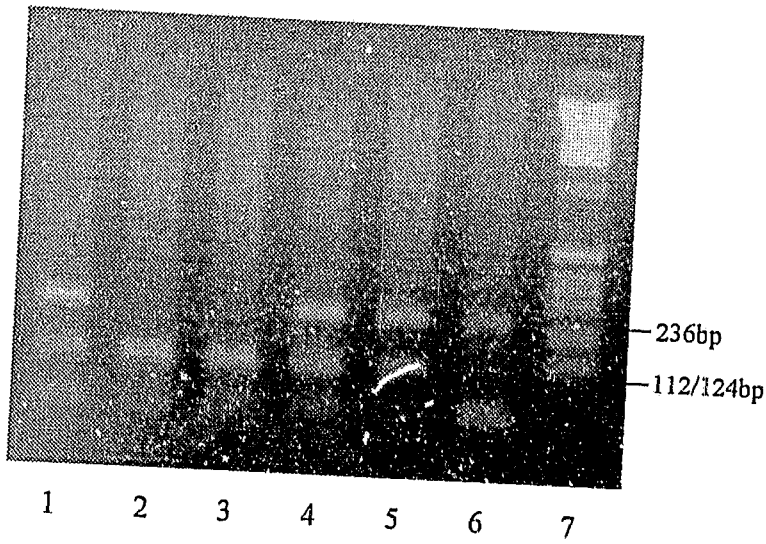


Figure 4.4. Photograph of a 2% agarose gel showing the separation of alleles at the *TaqI* A locus in the DRD2 gene. Lanes 1, 4 and 5 show A1/A2 heterozygotes, lanes 2 and 3 show A2 homozygotes and lane 6 shows an A1 homozygote. Lane 7 shows fragments of the 1kb DNA molecular weight marker (Gibco BRL).

4.3 Allele frequency distributions

4.3.1 HUMCD4 STR polymorphisms

This polymorphism has been discussed in **section 3.3.3** and estimates of the frequencies of alleles at this locus are given in **Table 3.3**.

4.3.2 CD4-*Alu* deletion polymorphism

Estimated *Alu* + and – frequencies and heterozygosity levels for the different population groups are given in **Table 4.1**. Comparative data from the literature are also given in the table. Deletion frequencies in the African populations vary between 5.7% (Mbenzele) and 37% (Zambians) and compares well with published ranges (Tishkoff *et al*, 1996; see **section 4.1.1**). The deletion is also present at a low frequency (6%) in the SA Indians but is absent from the Indonesian sample as would be expected from previous studies (Tishkoff *et al*, 1996). The frequencies in the Malagasy ethnic groups vary from a low of 2.5% in the Mahafaly to 28% in the Antankarana. Because of the absence of the deletion in SE Asia populations, especially Indonesians, the deletion chromosomes in the Malagasy supports the theory that considerable gene flow from African and possibly Middle Eastern populations has occurred into the Malagasy.

The range of heterozygosity levels seen in the Malagasy overlap appreciably with those noted in the African populations. This is probably due to the presence of the *Alu*- allele in the Malagasy at frequencies similar to those seen in Africa. None of the 17 tests carried out to examine samples for significant deviations from the Hardy-Weinberg expectation yielded P values which were less than or equal to 0.050.

Table 4.1. Frequencies of CD4-Alu alleles and heterozygosities for several Malagasy, non-African and African populations.

Population	2N	Frequencies		Heterozygosity
		Alu(+)	Alu(-)	(H)
Malagasy	**684***			
Merina	70	0.914	0.086	0.161
Betsileo	72	0.847	0.153	0.266
Antaisaka	72	0.819	0.181	0.304
Tsimihety	70	0.814	0.186	0.311
Sakalava	68	0.790	0.127	0.337
Mahafaly	66	0.975	0.025	0.061
Antandroy	68	0.882	0.118	0.220
Betsimisiraka	76	0.803	0.197	0.325
Antaimoro	74	0.784	0.216	0.348
Antankarana	32	0.720	0.280	0.431
Mikea	96	0.823	0.177	0.298
Non-Africans	**210***			
SA Indian	78	0.940	0.060	0.123
Indonesian	132	1.000	0.000	0.000
Middle Eastern**	200	0.675	0.325	0.459
Melanesian (PNG and Boug)**	84	1.000	0.000	0.000
Malaysian**	24	0.960	0.040	0.080
Cambodian**	48	0.940	0.060	0.080
Chinese**	88	1.000	0.000	0.000
Japanese**	86	1.000	0.000	0.000
Africans	**292***			
Tsonga	78	0.810	0.190	0.319
Zambian	100	0.630	0.370	0.476
Mbenzele	88	0.943	0.057	0.110
Gbaya	70	0.814	0.186	0.311
Mbimou/S.Sangha	56	0.840	0.160	0.280
Pygmies (Biaka and Mbuti)**	178	0.905	0.095	0.173
Khoisan**	170	0.820	0.180	0.297
Nigerian**	36	0.710	0.290	0.424

*number of chromosomes examined in present study.
**data taken from Tishkoff *et al*, 1996.

4.3.3 $(CA)_n$ repeat polymorphism

This polymorphism has been discussed in **section 3.3.4** and estimates of the frequencies of alleles at this locus are given in **Table 3.4**.

4.3.4 *Taq*I polymorphisms

Allele frequencies at the DRD2 *Taq*I B, D and A loci are shown in **Table 4.2** together with the heterozygosity levels for the different populations. All three loci are polymorphic in the populations examined. The published distribution of the B2 allele at the *Taq*I B locus is as follows (Kidd *et al*, 1998): it occurs at frequencies greater than 74% in African and European populations and less than 65% in most other populations examined. The A2 allele at the *Taq*I A locus follows this distribution very closely in most populations except for slightly lower frequencies in Africa (range: 52% to 73%). At the *Taq*I D locus, the D2 allele occurs at frequencies ranging from 60-90% in African populations, 38-52% in European populations and attains frequencies of more than 90% in most Asian, Pacific and New World populations. The distributions of alleles in the African and non-African populations in the present study follow these results very closely. The distributions in the Malagasy are somewhere between those seen in African and non-African populations

The average heterozygosity at the *Taq*I B, D and A loci are 0.356, 0.247 and 0.496 respectively. Of the Malagasy groups studied, eight have higher heterozygosities at the *Taq*I B locus and ten at the *Taq*I A locus than the average at those two loci respectively. At the *Taq*I D locus on the other hand, only three Malagasy groups have higher values than the average. If a higher heterozygosity is expected in an admixed population, the results at

the *Taq*I B and *Taq*I A loci are understandable, but the reason for the lower than average values at the *Taq*I D loci is not clear.

The genotype frequency data at each of the three loci was tested for agreement with Hardy-Weinberg expectation. Significant departures were observed in 3 of 60 comparisons (20 populations x 3 loci). Because this is exactly the number of comparisons expected to be significant at the 5% level by chance alone and because none of the significant departures cluster by population or locus, these are considered to represent normal statistical fluctuations.

Table 2. Frequencies of alleles and heterozygosity levels at three TaqI RFLP loci in the DRD2 gene in several Malagasy, non-African and African populations.

Population			*TaqI* B				*TaqI* D					*TaqI* A	
	2N	B1	B2	H	2N	D1	D2	H	2N	A1	A2	H
Malagasy	756*				752*				754*			
Merina	74	0.446	0.554	0.508	70	0.143	0.857	0.252	68	0.515	0.485	0.515
Betsileo	70	0.371	0.629	0.481	70	0.200	0.800	0.329	70	0.529	0.471	0.513
Antaisaka	64	0.266	0.734	0.403	64	0.109	0.891	0.201	66	0.470	0.530	0.514
Tsimihety	70	0.129	0.871	0.231	70	0.114	0.886	0.208	70	0.300	0.700	0.432
Sakalava	74	0.203	0.797	0.332	74	0.081	0.919	0.153	74	0.432	0.568	0.505
Mahafaly	66	0.318	0.682	0.447	66	0.015	0.985	0.031	68	0.544	0.456	0.511
Antandroy	68	0.279	0.721	0.415	68	0.132	0.868	0.237	68	0.515	0.485	0.515
Betsimisiraka	70	0.229	0.771	0.363	70	0.229	0.771	0.363	70	0.457	0.543	0.511
Antaimoro	68	0.250	0.750	0.386	68	0.103	0.897	0.190	68	0.471	0.529	0.513
Antankarana	32	0.188	0.813	0.325	32	0.125	0.875	0.233	32	0.406	0.594	0.515
Mikea	100	0.270	0.730	0.402	100	0.120	0.880	0.216	100	0.470	0.530	0.508
Non-Africans	162*				178*				180*			
SA Indian	80	0.288	0.713	0.420	78	0.449	0.551	0.508	80	0.325	0.675	0.450
Indonesian	82	0.451	0.549	0.508	100	0.330	0.670	0.451	100	0.140	0.860	0.246
Middle Eastern**	210	0.080	0.920	0.115	210	0.500	0.500	0.500	210	0.100	0.900	0.185
Cambodian**	50	0.420	0.580	0.497	50	0.040	0.960	0.078	50	0.420	0.580	0.497
Finns**	66	0.120	0.880	0.380	66	0.660	0.440	0.590	66	0.150	0.850	0.330
Africans	406*				404*				402*			
Zambian	96	0.229	0.771	0.361	96	0.094	0.906	0.174	96	0.385	0.615	0.484
Tsonga	100	0.100	0.900	0.184	100	0.140	0.860	0.246	98	0.388	0.612	0.485
Gbaya	66	0.212	0.788	0.345	66	0.167	0.833	0.286	66	0.500	0.500	0.516
Mbimou/S.Sangha	56	0.250	0.750	0.389	56	0.071	0.929	0.138	56	0.357	0.643	0.476
Mbenzele	88	0.125	0.875	0.224	86	0.209	0.791	0.339	86	0.407	0.593	0.494

* number of chromosomes examined in present study.

** data taken from Tishkoff *et al*, 1996.

4.4 Haplotypes at the CD4 locus

Data from both sites at the CD4 locus were available for 786 unrelated individuals belonging to 20 population groups. A large number of individuals were heterozygous for alleles at both sites making it difficult to determine the phase of the alleles on haplotypes without family data. Haplotype frequencies were thus estimated using the computer program EM-HAPLO [see **section 2.3 (c)**] and are given in **Table 4.3** together with heterozygosity estimates for individual populations. Comparative data from the literature (Tishkoff *et al*, 1996) has been included in the table.

Twenty one of the 24 possible haplotypes were detected in this study: all 12 haplotypes possible on the *Alu*+ background and nine of the 12 possible on the *Alu*- background. The greatest number of haplotypes were seen in the African populations followed closely by the Malagasy: 20 of the 21 haplotypes are present in Africans, 19 of the 21 are present in the Malagasy and fewer than six are found in the remaining non-African groups examined. In the non-Africans, allele 6 was detected only in the SA Indian population where it is associated exclusively with the *Alu*- allele, an association already noted by Tishkoff *et al* (1996). This association however, is not as strong in the African or Malagasy populations where the *Alu*- allele is found together with a wide range of alleles. This finding is exclusive to Africa, especially sub-Saharan Africa (present study and Tishkoff *et al*, 1996).

As mentioned in **section 4.1.1**, Tishkoff *et al* (1996) have shown that the pattern of haplotype frequencies in sub-Saharan African populations differs from that in Northeast African and non-African populations. The data from this study on African populations fits in well with the sub-Saharan pattern of haplotype frequencies. The pattern seen in the

Malagasy supports an African ancestry for them based both on the large number of 'African-specific' haplotypes detected and the association of a large number of STR alleles with the *Alu-* allele. The evidence of an Asian-Indonesian contribution to the gene pool of the Malagasy is strengthened by the increased +5 and +10 haplotype frequencies in the Malagasy (average of **0.357** and **0.212** respectively). These values are appreciably higher than the corresponding average frequencies of these haplotypes in most African populations examined (**0.161** and **0.076** respectively). The published +5 and +10 haplotype frequency means for 11 sub-Saharan populations and the Indonesian population studied here are **0.187** and **0.172** (Tishkoff *et al*, 1996) and **0.591** and **0.386** respectively.

As would also be expected from the above data, the apparent diversities in the African populations examined in the present study are greater than those observed for the non-Africans examined (>0.818 in Africans and <0.599 in the non-Africans). The corresponding values for the Malagasy ethnic groups (0.664 to 0.855) exceed those for twelve Asian, Pacific, and Australo-Melanesian populations (range approximately 0.40 to 0.52; Tishkoff *et al*, 1996) and are close to the mean for ten sub-Saharan African populations (0.85; Tishkoff *et al*, 1996). The high values in the Malagasy would be characteristic of an admixed popu' 'on and can be explained both by the differences in the number and frequencies o. haplotypes present in African and non-African populations.

Table 4.3. CD4-Alu haplotype frequency estimates for several Malagasy, non-African and African populations.

		Alu+ STR haplotypes												Alu- STR haplotypes										Diversity
Population	2N	+4	+5	+6	+7	+8	+9	+10	+11	+12	+13	+14	+15	-5	-6	-7	-8	-9	-10	-11	-12	-14	-15	
Malagasy	754*																							
Merina	70		0.457			0.071		0.357	0.029					0.029	0.043					0.014				0.664
Betsileo	72		0.427	0.050		0.125		0.153	0.098	0.014				0.017	0.095					0.041				0.768
Antaisaka	72		0.230	0.014		0.074	0.056	0.250	0.139	0.028		0.028		0.033	0.097	0.028	0.023							0.855
Tsimihety	70		0.333	0.035		0.143	0.029	0.106	0.138	0.014		0.017		0.052	0.017	0.057			0.022	0.005		0.012		0.840
Sakalava	68		0.384		0.015	0.057		0.176	0.104	0.029	0.015	0.029		0.015	0.088	0.059	0.001			0.028				0.805
Mahafaly	64		0.375		0.016	0.047		0.297		0.156	0.016	0.031		0.031	0.031									0.753
Antandroy	66		0.486	0.015		0.088		0.162	0.102	0.029					0.044	0.015	0.015			0.029		0.015		0.726
Betsimisiraka	76		0.267		0.013	0.053	0.013	0.303	0.088	0.039	0.026			0.101	0.053	0.026				0.018				0.821
Antaimoro	74		0.363	0.014		0.027		0.149	0.149	0.041		0.041		0.041	0.08	0.027	0.014	0.013		0.041				0.820
Antankarana	32		0.313			0.031		0.187	0.094			0.063		0.031	0.125	0.094		0.031			0.031			0.853
Mikea	90		0.293	0.014		0.089	0.022	0.189	0.116	0.033	0.022	0.011		0.074	0.042	0.067				0.028				0.852
Non Africans	428*																							
SA Indian	78		0.462					0.436	0.038						0.064									0.599
Indonesian	132		0.591	0.008				0.386	0.015															0.505
Middle Eastern**	198		0.335	0.015				0.240	0.045	0.02	0.025			0.005	0.305		0.005							0.738
Asian**	600		0.605	0.011	0.003	0.003	0.004	0.210	0.023	0.002				0.001	0.037									0.520
New Guinean**	40		0.875					0.125																0.224
Solomon Island**	46		0.869	0.022			0.022	0.065	0.022															0.244
Africans	390*																							
Northeast Africa**	232		0.273	0.010			0.029	0.280	0.113	0.063	0.006	0.027		0.020	0.160	0.010	0.004	0.013						0.780
Sub-Saharan Africa**	806	0.002	0.187	0.023	0.030	0.074	0.026	0.172	0.156	0.107	0.024	0.021	0.005	0.046	0.042	0.009	0.007	0.003		0.054	0.007	0.005	0.001	0.850
Tsonga	76		0.223			0.158	0.026	0.105	0.121	0.066	0.039	0.039	0.039	0.015	0.079	0.026		0.013		0.051				0.892
Zambian	100		0.122		0.020	0.160	0.040	0.080	0.100	0.070	0.040	0.015		0.099	0.110	0.040				0.106		0.015		0.914
Mbenzele	88	0.023	0.193	0.011	0.023	0.068	0.011	0.023	0.295	0.239	0.034	0.023			0.011	0.034				0.012				0.818
Gbaya	70		0.123		0.029	0.144		0.085	0.214	0.048	0.043	0.029		0.034	0.071	0.029	0.028				0.023			0.856
Mbimou'S Sangha	56		0.143		0.018	0.250	0.054	0.089	0.125	0.107	0.036	0.018		0.054	0.053	0.053								0.884

* number of chromosomes examined in present study.

** data taken from Tishkoff et al, 1996. The Middle Eastern sample consists of Israeli Druze and Yemenite Jews.

4.5 Haplotypes at the DRD2 locus

Data for samples which were successfully typed for the $(CA)_n$ repeat polymorphism (**section 3.3.4**) and the three *Taq*I polymorphisms (**section 4.3.4**) were used for the construction of four site haplotypes. Again because of the lack of knowledge of the phase of alleles for haplotype construction in multiply heterozygous individuals, a computer program [Arlequin (ver 1.1), **section 2.3 (c)**] was used to infer haplotype frequencies from the allele frequency data. The availability of this PC-based program at this stage of the analysis favored its use over the program (EM-HAPLO) used for the CD4 haplotype analysis which only runs off a mainframe computer. Estimates of haplotype frequencies are given in **Table 4.4** which extends over two pages. Of the 48 possible haplotypes (the product of the number of alleles at each of the four loci), 36 were estimated to be present in the study sample used; 26 in the Malagasy, 14 in the SA Indians and 20 in the African populations.

Figure 4.5 summarizes the evolutionary relationships proposed for the DRD2 haplotypes defined by the three *Taq*I RFLPs (Kidd *et al*, 1998). The data from studies on worldwide populations suggests that the most likely ancient human haplotypes are **B2-D2-A1-13** and **B2-D2-A1-14**. Both these as well as haplotype **B2-D2-A1-12** are rare outside of Africa but are present in nine of eleven Malagasy groups. Allele 12 has been detected only in Africa, an exception being a single Mayan individual.

Table 4.4. Frequency estimates of haplotypes at the DRD2 locus, gene diversity levels and additional haplotype information on several Malagasy, non-African and African populations.

		HAPLOTYPE																					
		B2	B2	B2	B2	B2	B2	B2	B2	B2	B2	B2	B1	B1	B1	B1	B1	B1	B2	B2	B2	B2
		D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D2	D1	D1	D1	D1
		A1	A1	A1	A1	A1	A1	A2	A2	A2	A2	A2	A1	A1	A1	A1	A1	A1	A2	A2	A2	A2
		12	13	14	15	16	17	13	14	15	16	17	12	13	14	15	16	17	13	14	15	16
Population	2N																					
Malagasy	728*																					
Merina	62	0.016	0.079	0.058		0	0	0.106	0.232	0	0	0	0.065	0	0	0	0.210	0	0.032	0	0	0
Betsileo	64	0	0.067	0	0.016	0.016	0	0.122	0.200	0	0	0	0.063	0	0	0	0.278	0	0	0.018	0.078	0
Antaisaka	62	0.025	0.034	0.032	0	0	0.007	0.226	0.291	0.032	0	0	0.070	0	0	0	0.145	0.025	0	0	0.016	0
Tsimihety	70	0.057	0.050	0.032	0	0	0	0.218	0.353	0.028	0.015	0	0	0	0.015	0	0.114	0	0	0	0.086	0
Sakalava	72	0.042	0.022	0.106	0	0	0	0.253	0.269	0.044	0	0	0.056	0	0	0.014	0.094	0	0	0	0	0
Mahafaly	62	0.048	0.071	0.05		0	0	0.107	0.346		0	0	0.048	0	0	0	0.145	0.161	0	0		0
Antandroy	58	0.038	0.088	0	0	0	0	0.103	0.295	0.029	0	0	0.074	0	0	0	0.162	0.029	0	0	0.059	0
Betsimisiraka	70	0.043	0.094	0.053	0.032	0.014	0	0.086	0.184	0.058	0	0	0.057	0	0	0	0.094	0	0	0.021	0.153	0
Antaimoro	68	0.059	0.097	0.052	0.015	0	0	0.135	0.248	0.062	0	0	0.059	0	0	0	0.117	0.059	0	0	0.070	0
Antankarana	30	0.033	0.050	0.067	0	0	0	0.100	0.334	0.083	0	0	0.100	0	0	0.033	0.067	0	0	0	0.083	0
Mikea	100	0.060	0.088	0.025	0.010	0	0	0.162	0.265	0.010	0	0	0.140	0	0	0	0.097	0.010	0.010	0	0.070	0
Non-Africans	78*																					
SA Indian	78	0	0.013	0	0.003	0	0	0.128	0.128	0.049	0	0	0	0	0.012	0	0.217	0.013	0	0	0.271	0.051
Middle Eastern**	210	0	0.005	0	0.012	0.002	0	0.221	0.154	0.046	0.003	0	0	0.003	0	0	0.049	0	0.014	0.024	0.378	0.064
Finns**	66	0	0	0	0	0	0	0.144	0.137	0	0	0	0	0	0	0.020	0.161	0	0.028	0.081	0.305	0.018
Cambodia**	100	0	0	0	0	0	0	0.140	0.357	0.043	0	0	0	0	0	0	0.420	0	0	0.023	0.017	0
Nasioi**	44	0	0	0	0	0.017	0	0.227	0.269	0	0	0	0	0	0	0	0.227	0	0	0.049	0.159	0.023
African	404*																					
Zambian	96	0.031	0.083	0	0	0	0	0.136	0.395	0.032	0	0	0.136	0.011	0	0	0.052	0.031	0	0	0.063	0
Tsonga	100	0.060	0.153	0.036	0	0	0	0.126	0.384	0.011	0	0	0.040	0	0	0.010	0.030	0.010	0	0	0.079	0
Mbenzele	86	0	0.055	0.050	0	0	0	0.038	0.197	0.023	0	0.023	0.023	0.023	0	0	0.058	0	0	0	0.012	0
Gbaya	66	0.015	0.070	0.112	0	0	0	0.051	0.373	0	0	0	0.121	0.015	0	0.015	0.061	0	0	0	0.076	0
Mbimou/S.Sangha	56	0	0.104	0.057	0	0	0	0.146	0.372	0.018	0	0	0.071	0	0.018	0	0.107	0	0	0	0.053	0

* number of chromosomes examined in present study.

** data taken from Kidd *et al.*, 1998.

Continued overleaf.....

Table 4.4 continued.

		HAPLOTYPE																Apparent gene diversity	No. of haplotypes	No. of haplotypes occuring at frequency >5%
		B2 D1 A1	B2 D1 A1	B2 D1 A1	B2 D1 A1	B2 D1 A1	B2 D1 A1	B1 D2 A2	B1 D2 A2	B1 D2 A2	B1 D1 A2	B1 D1 A2	B1 D1 A2	B1 D1 A1	B1 D1 A1	B1 D1 A1	B1 D1 A1			
Population	2N	12	13	14	15	16	17	13	14	16	12	14	15	12	15	16	17			
Malagasy	728°																			
Merina	62	0	0.016	0	0	0.065	0	0	0	0.145	0	0	0	0	0	0.016	0	0.892	12	7 (58%)
Betsileo	64	0	0.061	0	0	0.031	0	0	0	0.050	0	0	0	0	0	0	0	0.877	12	8 (67%)
Antaisaka	62	0	0.079	0	0	0	0	0	0	0	0	0	0	0.018	0	0	0	0.882	13	5 (38%)
Tsimihety	70	0	0.029	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.838	11	6 (55%)
Sakalava	72	0	0.030	0	0	0.025	0	0	0	0.017	0	0	0	0	0	0.014	0.014	0.860	14	5 (36%)
Mahafaly	62	0.016	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.860	9	6 (67%)
Antandroy	68	0	0.044	0.029	0	0	0	0	0	0	0	0	0	0	0	0	0	0.906	11	7 (64%)
Betsimisiraka	70	0	0.034	0	0	0	0	0	0	0 056	0	0	0	0	0	0.021	0	0.921	15	9 (60%)
Antaimoro	68	0	0 033	0	0	0	0	0	0	0.014	0	0	0	0	0	0	0	0.910	13	9 (69%)
Antankarana	30	0	0.050	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.917	11	9 (82%)
M'kea	190	0	0.030	0	0	0	0	0	0	0.013	0	0	0	0	0.010	0	0	0.914	15	8 (53%)
Non-African	78*																			
SA Indian	78	0	0	0	0	0.062	0.913	0	0	0.014	0	0.026	0	0	0	0	0	0.893	14	6 (43%)
Middle Eastern**	210	0	0.006	0	0	0.016	0	0	0	0.003	0	0	0	0	0	0	0	0.765	>12	5 (42%)
Finns**	66	0	0	0.031	0	0	0	0	0.020	0	0	0.020	0.035	0	0	0	0	0.830	12	5 (42%)
Cambodia**	100	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.670	6	3 (50%)
Nasioi**	44	0	0	0	0.006	0	0	0.023	0	0	0	0	0	0	0	0	0	0.800	8	4 (50%)
African	404*																			
Zambian	96	0	0.010	0.011	0	0.009	0	0	0	0	0	0	0	0	0	0	0	0.873	13	6 (46%)
Tsonga	100	0	0.051	0	0	0	0	0	0	0	0.010	0	0	0	0	0	0	0.800	13	6 (46%)
Mbenzele	86	0.105	0.093	0	0	0		0	0	0	0	0	0	0	0	0	0	0.846	12	6 (50%)
Gbaya	66	0.015	0.046	0	0	0.030	0	0	0	0	0	0	0	0	0	0	0	0.894	13	6 (46%)
Mbimou/S.Sangha	56	0	0	0	0	0	0	0	0.036	0	0	0	0.018	0	0	0	0	0.846	11	6 (55%)

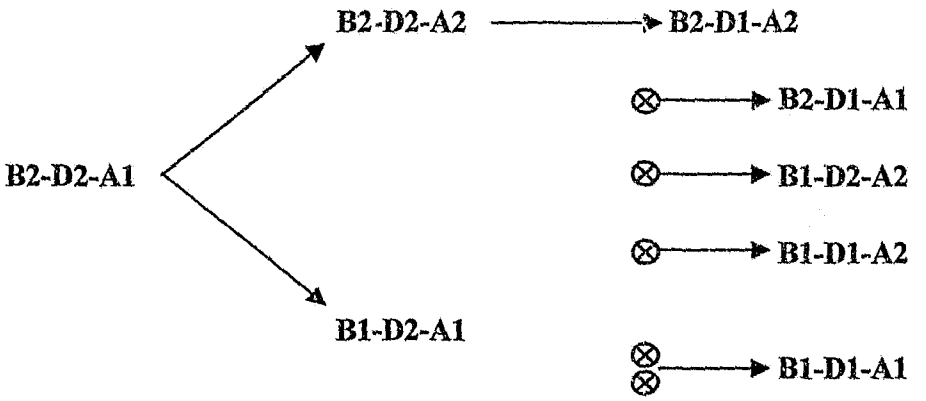


Figure 4.5. Inferred evolutionary relationships of DRD2 haplotypes as defined by three *Taq I* restriction site polymorphisms. The above sequence of events is most compatible with data from a global sample of 28 distinct populations. Adapted from Kidd *et al*, 1998. A single ⊗ symbol represents a single crossover event while two symbols represent a double crossover event.

Haplotype **B2-D2-A2-14** is the most frequent haplotype in most of the populations examined, the exceptions being the Betsileo and the SA Indians in whom the **B1-D2-A1-16** is the most frequent. The latter was also the only haplotype that was shared by all 28 populations examined by Kidd *et al* (1998). The four haplotypes **B1-D2-A1-16, B2-D1-A2-15, B2-D2-A2-13** and **B2-D2-A2-14** (shown in yellow in **Table 4.4**) were found by them to account for at least half of the chromosomes in every population they examined. This also holds true for the African and non-African populations examined in this study.

Alleles 13 and 14 occur primarily on the B2-D2-A2 haplotype, allele 15 on haplotype B2-D1-A2 and allele 16 on the B1-D2-A1 haplotype. A similar trend has been reported by Castiglione *et al* (1995) who suggested a low mutation frequency for the CA repeat locus based on these results. They infer that haplotype frequency variation noted in worldwide populations would thus be primarily the result of genetic drift.

A number of haplotypes are rare or absent from the African populations but are present both in the SA Indian population as well as either Middle Eastern or Cambodian populations (Kidd *et al*, 1998) while other haplotypes are present in the African populations but are absent or rare in non-African populations. These 'population-specific' haplotypes are given in **Table 4.5** together with their incidence in the three major groups of population examined in this study. Many of these 'population–specific' haplotypes are present in the Malagasy supporting both an African as well as an Asian ancestry for them. For example, haplotype **B1-D2-A1-16** is present at lower frequencies in African populations (0.030 – 0.107) than those noted in Asian populations (0.217 – 0.420). This haplotype is present at higher frequencies in most Malagasy groups compared to the

African groups (range is 0.067 – 0.278), especially in the two highland groups (0.210 in the Merina and 0.278 in the Betsileo), supporting a larger Asian ancestry for them.

Overall, the Malagasy populations share 16 haplotypes with Africans, six haplotypes with the Cambodian population (only SE Asian population for which data were available) and 11 with the SA Indian population (**Table 4.4**). The African population shares eight haplotypes with the SA Indian population and only five with the Cambodian population. Four of these shared haplotypes are common to global populations (discussed above) and are thus not informative for admixture studies. Taking this into account, the Malagasy share the most haplotypes with African and the SA Indian population rather than with the SE Asian (Cambodian) population.

A possible explanation for this could be the following: Urbanization is thought to have been a relatively late development in SE Asia. Evidence for the appearance of towns and cities in Thailand, Burma, Cambodia and Java date to between the sixth and eighth centuries AD. These population centres were under Indian influence around the fourth and fifth centuries (Cavalli-Sforza *et al*, 1994). This may explain why the monumental architecture that started flourishing after the seventh century in many different parts of SE Asia is reminiscent of Indian art. It may be possible that a close genetic affinity between Indian and some other (not Cambodian) SE Asian populations does exist and that these SE Asian contributors may also be responsible for some gene flow into the Malagasy. This may also explain the close affinity of the Malagasy to the SA Indian sample as suggested by the autosomal STR data in **CHAPTER 3**. A larger SE Asian sample from different geographic regions needs to be examined to shed light on this. There are historical records of Indians on Madagascar, but because of their endogamous culture (Brown, 1978),

common haplotypes shared by the Malagasy and the SA Indians may be of SE Asian origin rather than recent gene flow directly from Indians on the island.

The apparent diversity values are high in the populations examined (**Table 4.4**), ranging from 0.800 in the Tsonga to 0.921 in the Betsimisiraka. Nineteen haplotypes occur with frequency of at least 5% in at least one population. Except for the Malagasy population groups, the number of haplotypes which occur at a frequency greater than 5% in a population is approximately half the number of haplotypes in that population. In the Malagasy, more than half the haplotypes found in any one group are present at frequencies greater than 5%. This again supports a larger heterogeneity in that population and would be expected of an admixed population. The expected heterozygosity in the SA Indian population (0.893) is almost identical to the average for the Malagasy (0.889) closely followed by the average in the African populations (0.852). This high degree of variation in all the populations examined indicates the value of this locus for population and evolutionary studies.

Table 4.5. Frequencies of some population specific DRD2 haplotypes in Malagasy, Asian and African populations.

Haplotype	Malagasy	Asian	African
B2-D2-A1-12	0.000 - 0.088 (9/11)*	not detected	0.000 - 0.060
B2-D2-A1-13	0.022 - 0.097 (11/11)*	0.000 - 0.013	0.055 - 0.153
B2-D2-A1-14	0.000 - 0.106 (9/11)*	not detected	0.000 - 0.111
B1-D2-A1-12	0.000 - 0.140 (10/11)*	not detected	0.040 - 0.135
B1-D2-A1-16	0.067 - 0.278 (11/11)*	0.217 - 0.420	0.030 - 0.107
B2-D1-A2-15	0.000 - 0.153 (8/11)*	0.017 - 0.271	0.012 - 0.079
B2-D1-A1-13	0.000 - 0.079 (10/11)*	not detected	0.000 - 0.093

* The numbers in parenthesis indicates the proportion of Malagasy groups in which that haplotype was detected.

4.6 Phylogenetic analyses

A NJ tree based on data combined from the CD4-*Alu* and DRD2 loci was generated using the computer program DSW. This combined analysis was limited to 19 populations on whom data were available for both haplotypes. The phylogeny depicting population relationships is given in **Figure 4.6** and shows concordance with both historical and geographic data. All the African populations cluster together, as do the three non-African populations. Nine of the eleven Malagasy groups cluster on the branch leading to the African populations while the two highland groups, the Merina and the Betsileo cluster on the branch leading to the non-African populations. This strongly supports a larger non-African genetic ancestry for the highlanders compared to the lowlanders. None of the analyses indicate a close affinity of the Malagasy to either the Melanesians or Mbenzele. The close clustering of the Malagasy supports their recent common ancestry.

Bootstrap re-sampling was used to assess the strength of support of the data for the branching of the tree. A total of 1000 bootstraps were performed, and except for the branchpoints leading to the Malagasy groups, most other bifurcations have support of more than 73%.

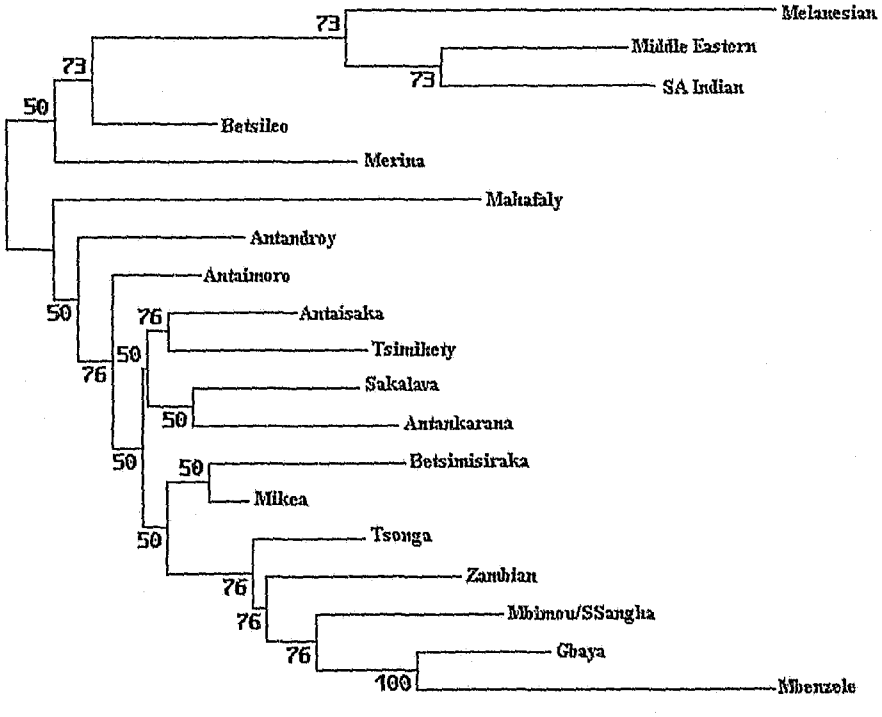


Figure 4.6. Unrooted NJ tree showing the genetic relationships between several Malagasy, non-African and African populations based on D_A genetic distances generated using the combined haplotype data from the CD4 and DRD2 loci. The numbers at each node indicate the percentage of trees (out of 1000 bootstrap replications) with such a node.

4.7 Principal components analysis

A principal components analysis (PCA) was performed on the combined data (both haplotype sets) used in **sections 4.6** and the resulting PC plot is shown in **Figure 4.7**. The first two components, which provide the most information for a two-dimensional depiction of population relationship accounts for 24% of the variance in the data. The PCA plot depicts similar relationships to those that were evident from the NJ tree. Based on the first axis, a cluster of African populations vs the Malagasy vs the two non-African populations can be seen (except for the Mbenzele who come up as a distinct isolate). The Malagasy groups are scattered between the African and non-African populations, with the two highland groups and the Betsimisiraka being closer to the non-African populations. This placement of the Malagasy supports large amounts of gene flow from both African and non-African sources to their gene pool. Like the Mbenzele, the Middle Eastern sample appears distinct on both axes showing very little affinity with other populations in the sample. On the second axis, the Antankarana are closest to the African cluster with the remaining Malagasy (except for the Merina) showing very little separation from the SA Indian and Melanesian populations.

In a plot of the first two PCs generated using only the CD4 haplotype data (and many more population groups) (**Figure 4.8**), the Middle Eastern and European populations trail the African cluster, followed by the SA Indian, Indonesian and Melanesian populations. This trail reminisces the trail of the out of Africa hypothesis. Here again, the Malagasy are scattered between the Asian and African populations and again, closest to the SA Indian sample and not the Indonesian sample as would be expected from historical and linguistic data. As already discussed in **section 4.5**, this apparent affinity between the Malagasy and

the SA Indians may be due to shared ancestry by the Asian branch of Malagasy ancestors (a SE Asian population not yet identified) and the SA Indian population.

The potential value of using haplotype frequencies rather then the frequencies of the individual genetic components which make up the haplotype to examine admixed populations is that the haplotypes retain the identity of the parental population whereas allele frequencies will be an average of that seen in the parental populations. In other words, haplotype data is more specific than data composed of individual locus data, and therefore potentially more valuable. The PC plots generated using either the STR frequency data (**Figure 3.5**) or the haplotype data (**Figures 4.7** and **4.8**) show very similar patterns of population relationships. This is probably due to the very short evolutionary time depths being examined in the study, in which case frequency data would be as useful as haplotype data in resolving population affinities

4.8 Distance from the centroid

A plot of heterozygosity vs distance from the centroid (also from the combined haplotype data set) is given in **Figure 4.9**. Almost all the Malagasy and African populations show a good fit between the observed relationship and that predicted by the model. Possible reasons for deviations from the centroid have already been discussed in **section 3.7** and whereas a slightly higher rate of gene flow than that predicted by the model may explain the placement of some Malagasy groups above the line, the alternative explanation, namely a larger effective population size may be the reason for African populations showing a larger heterozygosity. Many studies have shown greater genetic diversity in African populations (Cann *et al*, 1987; Armour *et al*, 1996; Tishkoff *et al*, 1996; Hammer *et al*, 1998) and the present study supports these. The non-African groups show lower heterozygosity levels than those predicted by the mode.

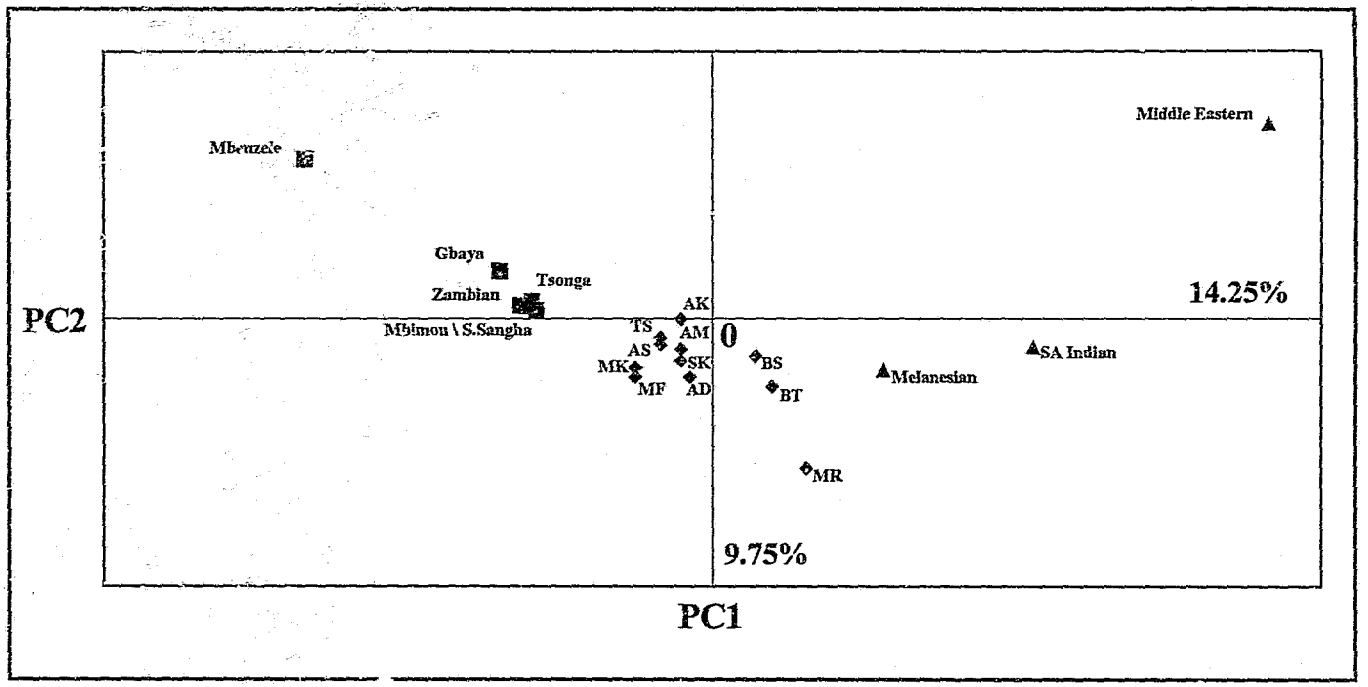


Figure 4.7. Two dimensional principal components plot based on the combined haplotype frequency data from the CD4 and DRD2 loci. Blue diamonds represents Malagasy groups, red triangles represent non-Africans and green squares represents Africans. Abbreviations for the Malagasy are as defined in the Nomenclature section.

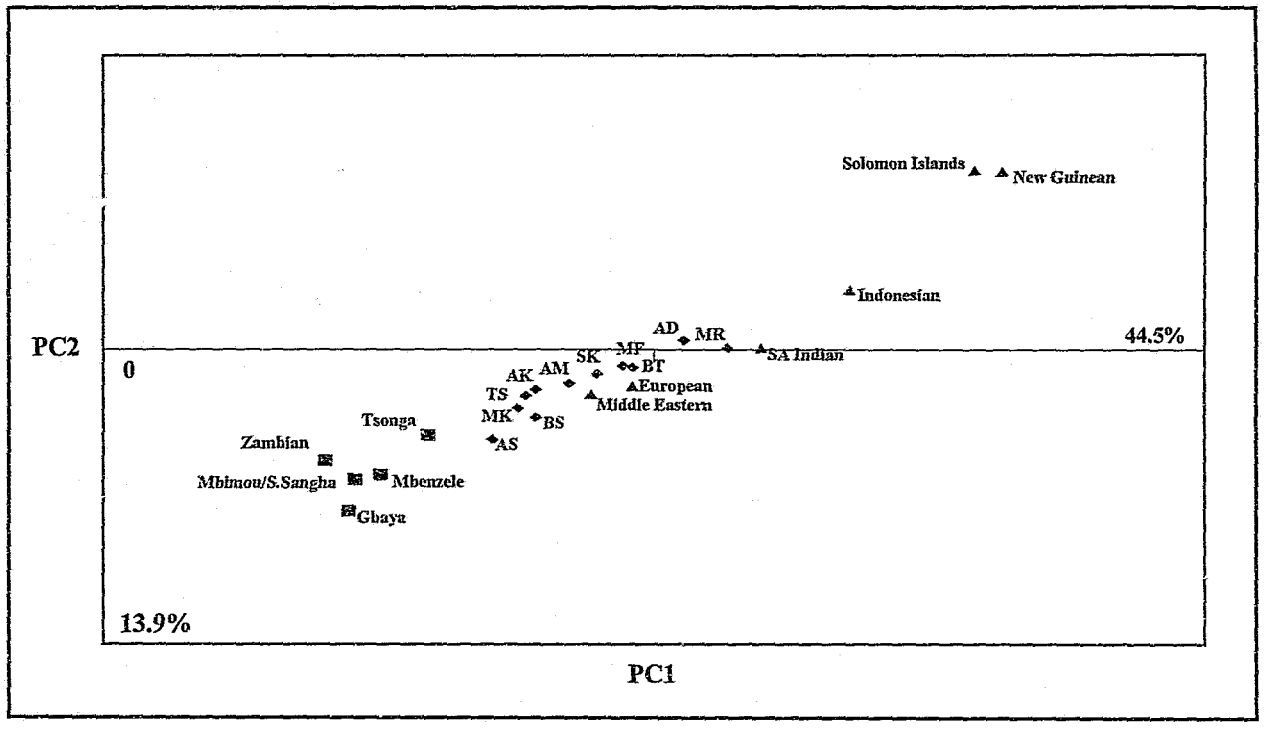


Figure 4.8. Two-dimensional principal components plot based on haplotype data from the CD4 locus. Blue diamonds represent the Malagasy, red triangles represent non-Africans and green squares represent Africans. Abbreviations used for the Malagasy are as defined in the Nomenclature section.

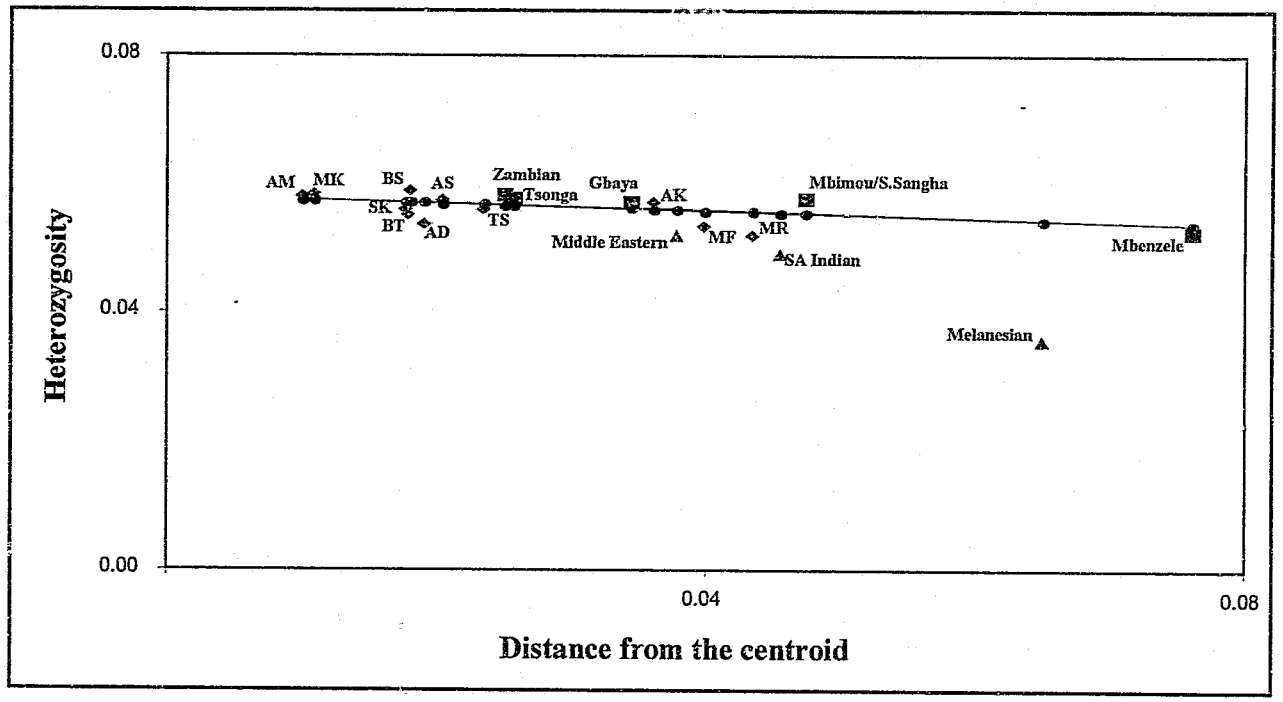


Figure 4.9. Plot of heterozygosity vs. distance from the centroid based on combined haplotype frequency data from the CD4 and DRD2 gene loci. Blue diamonds represent Malagasy ethnic groups, red triangles represent non-Africans and blue squares represent Africans. Abbreviations for the Malagasy groups are as defined in the Nomenclature section.

4.9 Admixture

On the assumption that the progenitors of the Malagasy consisted of only Africans and Indonesians/SE Asians (discussed in **section 3.8**), the least squares method of Elston (1971) can be used to estimate the proportional contributions of these two groups to the Malagasy. Comparative data was not available from Indonesian populations for all the markers examined and thus data from closely related populations had to be used for admixture analysis. The African populations (the Tsonga and Zambians) were used as one of the two parental populations while an Indonesian, Cambodian or SA Indian population sample was used as the second contributing population. The results of these admixture calculations are shown in **Table 4.6**. Data from an Indonesian population was only available for the CD4 haplotypes whereas data for the DRD2 haplotypes were available for both the SA Indian and Cambodian population. The highland and lowland populations represented the hybrid populations. Ideally, Arab populations should have been included, but this was not possible because of the lack of both genetic material and comparative data in the literature from the relevant populations.

From haplotype frequency data at the CD4 locus, the estimated African contribution to the highland population of Madagascar is estimated at **40.72%** and that to the lowland group **60.53%**. The Indonesian contribution to the highlanders is **59.28%** and to the lowlanders is **39.47%**. Data on the DRD2 locus indicates an African contribution to the highland groups as being **47.72%** and that to the lowland groups as **73.70%**. The Cambodian contribution to the highlanders is estimated to be **54.73%** and that to the lowlanders is **22.69%**. Very similar values were obtained when the SA Indian population was used as one of the parental populations instead of the Cambodian sample.

Admixture estimates at both loci suggests that there has been a larger African contribution to the lowland populations (ranging from 60.53% to 77.31%) than to the highland populations (ranging from 40.72% to 47.72%), a finding which is supported by historical data as well as results from STR loci **(CHAPTER 3)**.

Table 4.6. Estimates of proportional contributions to the Malagasy by progenitors including African and Asian populations. The least squares method of Elston (1971) was used on haplotype data from the CD4 and DRD2 loci.

		Percent contribution	
Locus	Population	Highlanders	Lowlanders
CD4-*Alu*	Indonesian	59.28%	39.47%
	African*	40.72%	60.53%
DRD2	SA Indian	52.28%	26.30%
	African*	47.72%	73.70%
DRD2	Cambodian	54.73%	22.69%
	African*	45.27%	77.31%

* The Afrian data consisted of the Zambian and Tsonga data sets.

4.10 Summary

The distributions of two sets of haplotypes were examined in eleven Malagasy and a number of African and non-African populations. The frequencies of both sets of haplotypes, together with comparative data from the literature were used to assess the affinities of the Malagasy to African and Asian/SE Asian populations. These data were also used to generate phylogenetic trees and PC plots and to estimate genetic admixture in the Malagasy.

All six loci examined (two in the CD4 gene and four in the DRD2 gene) were polymorphic in all the populations studied. Higher heterozygosity levels in the Malagasy, comparable to levels noted in Africa supports their admixed ancestry and a large African contribution to their gene pool. The pattern of haplotype variation at the CD4 locus specifically supports this. Many haplotypes from both data sets appeared to be more or less population specific. The presence of such 'African-specific' or 'Asian-specific' haplotypes in different proportions in the Malagasy ethnic groups supports historic data of a larger African genetic contribution to the lowland populations and a larger Asian one to the highlanders. Despite these differences in haplotype frequency distributions, the gene pools of the Malagasy ethnic groups are fairly homogeneous as shown by the phylogenetic analyses and this supports the theory that they share a recent common ancestry.

The phylogenetic analyses as well as the PC plots show the same trend: the highlanders cluster more closely to the Asian population groups whereas the lowlanders cluster more closely to the African groups. An interesting observation made when analyzing data from the CD4 haplotypes is that the Malagasy cluster is closer to the SA Indian sample than it is

to an Indonesian sample (from the Barito River region). Linguistic data suggests that this region could have been the home of the proto-Malagasy but the present genetic data do not support this hypothesis. A possible explanation for the apparent closeness is that the SA Indian sample and the proto-Malagasy share a common ancestor.

The close clustering of the Malagasy groups to each other on the NJ tree and on the PC plots shows that there is very little differentiation between them. Since the formation of the individual Malagasy groups, there appears to have been very little time for factors such as mutation, genetic drift and natural selection to mould the individual ethnic groups into distinct populations. The clo .e genetic affinity displayed by all eleven groups examined supports a common founding ancestral population for them. Any genetic contributions made later by the diverse groups of people who came to the island of Madagascar did not make a significant change to the original gene pool of the founding Malagasy population. A possible exception to this could be the hypothesized significant addition of Asian genes to the highlanders during the fourteenth century. This is the time frame proposed by archaeologists and historians for the settlement of the highlands and also for the arrival on the east coast of the island of a new wave of Asian immigrants. This may account for the closer affinity of highland groups to Asian populations.

In conclusion, the present data supports 'hypothesis 4' proposed for the origins of the Malagasy (**section 1.3.1**), i.e. that the founders of the Malagasy were probably a mixed people who had both African and Asian ancestry. This is apparent in the uniformity of the gene pools of the 11 Malagasy groups examined and by the fact that many of the African or Asian population specific haplotypes are found throughout the Malagasy sample. Many more African and Asian groups need to be examined to identify a possible genetic trail

from SE Asia via Africa to Madagascar. The use of many more polymorphic markers may also help identify the Asian and African identities of the 'proto-Malagasy'.

CHAPTER 5 - Y CHROMOSOME POLYMORPHISMS

5.1 Introduction

The potential of Y chromosome studies for evolutionary or individual profile studies has only recently been realized due to the previous paucity of known polymorphisms on the non-recombining portion of the Y (NRPY). An increasing number of highly polymorphic loci on the Y have however been reported in the last few years (discussed in detail in **section 1.6**). Many of these loci, especially STR loci can be typed by PCR and are currently in use to investigate the extent of genetic diversity in world populations.

Studies from Y STR loci suggest that these loci have mutation rates (≈0.20%/locus/generation; Heyer *et al*, 1997, Kayser *et al*, 1997) comparable to those on autosomes (≈0.21%; Weber and Wong, 1993). Y chromosome STRs, like their autosomal counterparts, also conform to the stepwise mutation model (SMM) and display considerable heterogeneity between even closely related populations (Cooper *et al*, 1996; Roewer *et al*, 1996, de Knijff *et al*, 1997). In an evolutionary context, this high mutation rate and the convergent nature of their mutation processes (Shriver *et al*, 1993) make STRs less efficient than SNPs for tracing long term evolutionary history. Their heterogeneity however offers new avenues for the investigation of recent historic and migratory events.

In a recent global survey, de Knijff *et al* (1997) typed a large number of males from different populations from Africa, Europe, Asia, Oceania and America for a set of seven Y chromosome STRs. They noted markedly different region-specific allele frequency distributions, even when closely related populations were compared. Population trees

based on these Y STR frequencies differed appreciably from those based on autosomal microsatellite frequencies (Bowcock *et al*, 1994) and mtDNA data (Vigilant *et al*, 1991). One of the most likely reasons for this, they suggest, is recurrent mutation. As a result, populations that appear to be related because of shared alleles at STR loci may appear so because of recurrent mutations and the effects of genetic drift and not because of common ancestry (as has recently been found by Deka *et al*, 1996)].

Using Goldstein's genetic distance measure (Goldstein, 1995) which is specific for STR loci, de Knijff *et al* (1997) estimated that Y-STR loci are not suitable for making evolutionary inferences dating back more than 40 000 to 50 000 years ago. For a reliable phylogenetic tree connecting present-day living human populations, Y-STRs would have to be combined with stable polymorphisms such as SNPs or *Alu* insertion/deletion polymorphisms. Although the Malagasy have only existed for about 2 000 years, well short of the 50 000 year maximum, the Y *Alu* polymorphism (YAP) has been included in the present study because of the large allele frequency differences noted between African and non-African populations. Also, the Y *Alu* insertion is thought to have occurred only once in the history of humans and therefore all chromosomes with this polymorphism share identity by descent (Hammer, 1994).

Only a small number of systematic population genetic studies on Y chromosome microsatellites have been reported in the literature (Gomolka *et al*, 1994; Deka *et al*, 1996; Roewer *et al*, 1996; Santos *et al*, 1996; de Knijff *et al*, 1997; Kayser *et al*, 1997) and very few data are available on the purported parental populations of the Malagasy. Alan Redd (personal communications, 1998) kindly provided data on the same four STR loci examined in the present study from SE Asian/Oceanic populations.

The allele distributions at five human Y chromosome loci (YAP, DYS393, DYS19, DYS390 and DYS391) were examined in eleven Malagasy, a SA Indian and five African populations. Haplotypes generated from the genotype data were used to estimate the extent of genetic contributions made to the Malagasy by historically implicated parental populations. Unrooted NJ trees and principal component analyses based on haplotype frequencies and haplotype networks based on unique haplotypes were then used to investigate possible relationships between African, non-African (especially Indonesian) and Malagasy populations. Attempts were made to estimate the length of time over which the observed levels of microsatellite diversity have been accumulating or since rapid population growth (Kittles *et al*, 1998) on Madagascar. The Fst statistic was used to evaluate the degree of divergence among the populations examined. As was done in previous chapters, the results are discussed simultaneously with their presentation in order to avoid repetition.

5.2 Allele identification

5.2.1 Y *Alu* polymorphism

Figure 5.1 is a photograph of a 2% agarose gel in which the PCR products from the YAP locus were separated. A 450bp fragment represents the YAP+ allele (presence of the insertion; lanes 1,2, 4-6, 10 and 12) while a 150bp fragment represents the YAP- allele (absence of the insertion; lanes 3, 7, 9 and 11). As expected for a haploid locus, there are no heterozygotes.

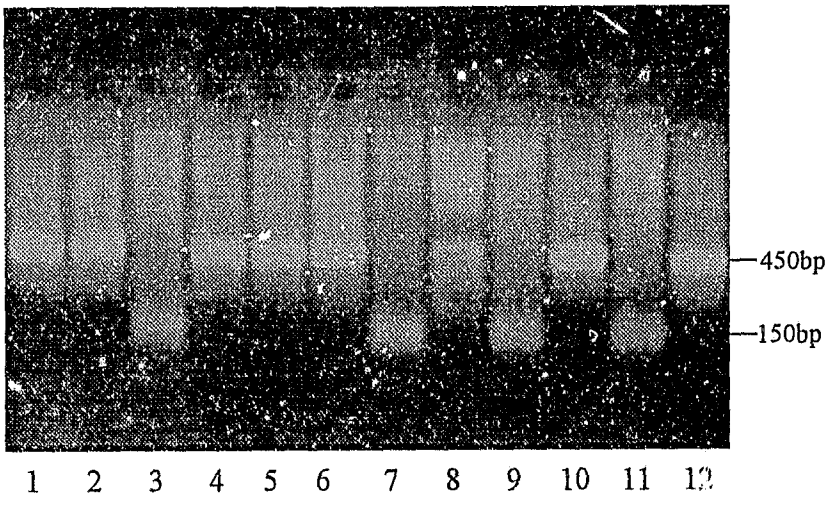


Figure 5.1. Photograph of a 2% agarose gel depicting YAP+ and YAP- alleles. Lanes 1, 2, 4, 5, 6, 10 and 12 represent individuals with the YAP+ allele and lanes 3, 7, 9 and 11 those with the YAP- alleles. Lane 8 represents fragments of a 1 kb DNA molecular weight marker (Gibco BRL).

5.2.2 STR allele identification

The nomenclature for the STR loci examined are as specified in the Genome Database (GDB). The number of times a repeat motif occurs at a locus is used as a label for the allele present (de Knijff *et al*, 1997) according to the recommendations of the International Society of Forensic Haemogenetics (DNA Commission of the ISFH, 1994). For general allele distribution studies this approach works well, but when it comes to representing haplotypes for network analyses, the numerical nomenclature is cumbersome. For this reason an alphabetical allele nomenclature available in the literature for three of the four STR loci (DYS19, DYS390 and DYS391) was adopted (**Table 5.1**). A similar system was adopted for the fourth locus, DYS393, where the letter C and not A was designated to the smallest allele detected to date (11 repeats), leaving a margin for designation of smaller alleles which may be found in future.

To verify the typing and allele designation at the four STR loci, control samples whose alleles have been characterized via sequencing by Kayser (personal communications, 1998) were amplified and typed using the protocols set up in the present study. Total concordance of allele calling was recorded.

The computer-generated image of a polyacrylamide gel through which the products of Y chromosome multiplex PCR were electrophoresed is shown in **Figure 5.2**. The three green bands in each lane represent alleles at the DYS393, DYS390, DYS391 loci respectively (reading from the bottom up) while the single blue bands represent alleles at the DYS19 locus. The red bands in each lane represents internal size standards.

Table 5.1. Criteria for translating apparent amplicon lengths (bp) into corresponding alphabetical codes and numbers of repeats.

	Allele sizes on ABI 377 (bp)	Published allele sizes (bp)	Alphabetical nomenclature	Number of repeats
DYS393	115	116	C	11
	119	120 [a]	D	12 [a]
	123	124 [a]	E	13 [a]
	127	128 [a]	F	14 [a]
	131	132 [a]	G	15 [a]
	135	136 [a]	H	16 [a]
	139	140 [a]	I	17 [a]
DYS19	180	178 [b,e]	Y	11
(DYS394)	184	182 [c,e]	Z [c,e]	12 [a]
	188	186 [d]	A [d]	13 [a]
	192	190 [d]	B [d]	14 [a]
	196	194 [d]	C [d]	15 [a]
	200	198 [d]	D [d]	16 [a]
	204	202 [e]	E [e]	17 [a]
	208	204 [c,f]	F [c,f]	18
DYS390	189	187 [g]	A [g]	17
	193	191	B	18
	197	195 [g]	C [g]	19
	201	199 [g]	D [g]	20 [a]
	205	203 [g]	E [g]	21 [a]
	209	207 [g]	F [g]	22 [a]
	214	211 [g]	G [g]	23 [a]
	218	215 [g]	H [g]	24 [a]
	222	219 [g]	I [g]	25 [a]
	226	223 [g]	J [g]	26 [a]
	230	227 [g]	K [g]	27 [a]
	234	231 [a]	L [a]	28 [a]
DYS391	277	275 [g]	A [g]	8
	281	279 [g]	B [g]	9 [a]
	285	283 [g]	C [g]	10 [a]
	289	287 [g]	D [g]	11 [a]
	293	291 [g]	E [g]	12 [a]
	297	295 [g]	F [g]	13 [a]

[a]-Redd *et al*, 1997; [b]-Ciminelli *et al*, 1995; [c]-Santos *et al*, 1996; [d]-Roewer *et al*, 1992; [e]-Zago *et al*; [f]-Hammer and Horai, 1995; [g]-Deka *et al*, 1996; The remaining nomenclature was added from the present study.

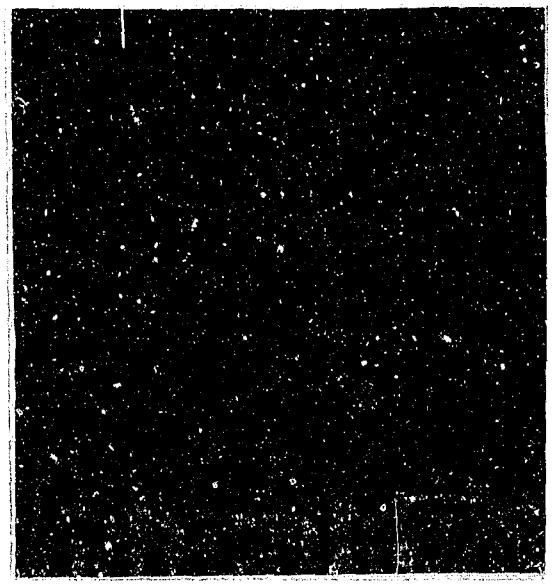


Figure 5.2. Computer generated image of a polyacrylamide gel run in ABI 377 automated DNA sequencer. The red bands in each lane represents fragments of the size standard while the green and blue bands represents alleles at the four loci studied.

5.3 Allele frequency di. ributions

5.3.1 YAP

The results obtained on this locus are summarized in **Table 5.2**. The DNA from 766 males was screened for the insertion at the YAP locus. The insertion was present in all the Malagasy and African groups but was absent from the SA Indians. Interestingly, the frequency of the YAP+ allele in the two highland groups (Merina and Betsileo) is similar to and in some cases higher than it is in lowland Malagasy groups.

The insertion has been shown to be absent from approximately 1 000 SE Asian/Oceanic individuals typed (Hammer *et al*, 1996) but is present at low frequencies in European and Middle Eastern populations. In most African populations, its frequency varies between 0.700 and 1.000 although it is lower in the Khoisan (0.440; Spurdle *et al*, 1994a). Despite its presence at appreciable frequencies in Japanese, Tibetan as well as other East Asian populations (Hammer and Horai, 1996), the insertion can be assumed to be predominantly of African origin in Madagascar since there is no evidence for Japanese or East Asian male contributions to the founding Malagasy population. A small percentage of YAP+ chromosomes in the Malagasy could also be Arabic in origin. The high frequencies of YAP+ alleles in the Malagasy (0.423 to 0.778) suggests a large African ancestry.

Table 5.2. Frequencies of alleles at the YAP locus.

Population	N	YAP +	YAP -
Malagasy	**532***		
Merina	60	0.567	0.433
Betsileo	62	0.581	0.419
Antaisaka	71	0.423	0.577
Tsimihety	63	0.476	0.524
Sakalava	52	0.558	0.442
Mahafaly	27	0.630	0.370
Antandroy	46	0.739	0.261
Betsimisiraka	55	0.655	0.345
Antaimoro	56	0.554	0.446
Antankarana	13	0.692	0.308
Mikea	27	0.778	0.222
Non-African	**50***		
SA Indian	50	0	1.000
SA Jewish[1]	36	0.140	0.860
Polynesian[2]	60	0	1.000
SE Asian/Oceanic[3]	>1000	0	1.000
SW Siberian Altai[3]		0.032	0.968
Mongolian[3]	41	0.026	0.984
Tibetans[3]		0.533	0.467
Japanese[4]	132	0.420	0.580
Taiwanese[4]	21	0	1.000
Europeans[5]	192	0.070	0.930
Saudi Arabia[5]	21	0.100	0.900
African	**184***		
Tsonga	52	0.750	0.250
Zambian	63	0.921	0.079
Gbaya	30	0.767	0.233
S.Sangha	17	0.941	0.059
Mbenzele	22	0.682	0.318

* number of chromosomes examined in present study.
[1]-Spurdle *et al*, 1994a; [2]-Spurdle *et al*, 1994b; [3]-Hammer *et al*, 1996; [4]-Hammer and Horai, 1995; [5]-Hammer, 1994.

5.3.2 DYS393

Approximately 700 males were examined for variation at this locus and the allele frequency distributions in the populations examined are given in **Table 5.3**. Seven alleles have been reported at this locus (11 to 17 repeats or alleles C to I) (Kayser *et al*, 1997; Redd *et al*, 1997), but only six (alleles C to H) were detected in this study. Modal or most common alleles in the population groups are shown in bold typeface in all allele frequency tables that follow.

The trends in allele frequency distributions noted are similar to those in published data on worldwide populations (Kayser *et al*, 1997). Allele C is rare or absent from most populations while allele H was detected in three Malagasy as well as three African populations. Allele H has only been detected in African-American individuals (Redd *et al*, 1997) and its absence in the SA Indian (this study), Chinese and SE Asian/Oceanic populations (Kayser *et al*, 1997; Redd, unpublished data) suggests that its presence in the Malagasy is probably due to African ancestry. Allele D is rare to absent in the African populations but present at high frequencies in Asian populations (0.273 in SE Asian populations; Redd, unpublished data) supporting an Asian origin for this allele in the Malagasy. The widely varying frequencies of alleles E, F and G in all the populations examined do not cast any more light on Malagasy ancestry.

All the populations except the Sakalava and the Tsonga show a unimodal allele frequency distribution with the modal allele in the African and Malagasy populations varying between alleles E and G while the modal allele in the non-African populations is either allele D or E.

Table 5.3. Frequencies of alleles at the DYS393 locus. The frequencies of modal alleles are shown in bold typeface.

Population	N	Alleles					
		C	D	E	F	G	H
Malagasy	**486***						
Merina	52	0	0.058	**0.654**	0.250	0.038	0
Betsileo	50	0	0.020	**0.720**	0.220	0.040	0
Antaisaka	50	0	0.080	**0.680**	0.140	0.080	0.020
Tsimihety	55	0	0.055	**0.564**	0.236	0.145	0
Sakalava	52	0	0.077	**0.616**	0.115	**0.154**	0.038
Mahafaly	27	0	0.037	**0.704**	0.148	0.111	0
Antandroy	45	0.022	0.089	0.244	**0.578**	0.067	0
Betsimisiraka	53	0.038	0.038	**0.566**	0.264	0.094	0
Antaimoro	55	0	0.091	**0.691**	0.127	0.091	0
Antankarana	13	0	0.077	**0.538**	0.308	0.077	0
Mikea	34	0	0.029	0.382	**0.472**	0.088	0.029
Non-African	**50***						
SA Indian	50	0.080	0.320	**0.340**	0.240	0.020	0
Chinese**	36	0	**0.530**	0.310	0.140	0.030	0
Mongolians**	40	0	0.220	**0.620**	0.150	0	0
S-Borneo**	13	0.080	**0.380**	0.230	0.230	0.008	0
African	**173***						
Tsonga	52	0	0.019	**0.578**	0.096	**0.269**	0.038
Zambian	54	0	0	0.352	**0.425**	0.204	0.019
Gbaya	30	0.033	0	**0.533**	0.267	0.167	0
S.Sangha	17	0	0	0	0.471	**0.529**	0
Mbenzele	20	0	0	0.200	0.350	**0.400**	0.050

*number of samples typed in present study.
**Kayser *et al*, 1997.

5.3.3 DYS19

The results obtained for this locus are summarized in **Table 5.4**. At present, this is the most widely used STR and a large amount of comparative data on worldwide populations are available for it (de Knijff *et al*, 1997). Only eight (11 to 18 repeats, alleles Y to F) of the ten alleles reported in the literature (10 to 19 repeats) were detected in the 684 males screened.

This locus is highly polymorphic and shows some very useful geographic variation that has been described in a number of previous studies (Roewer *et al*, 1992; Santos *et al*, 1996). For example, the modal or most common allele in most European populations is allele B while in most African, Asian and Oceanic populations it is either allele C or D. Alleles Y, Z and F are rare to absent in most populations examined while the remaining alleles occur at varying frequencies globally.

The frequency of allele B is higher in the Malagasy and non-African populations compared to that in the African samples. Allele C is modal in the Malagasy, the SA Indian as well as in three of the five African populations. A number of Malagasy groups show a bimodal allele distribution and this would support the theory that the Malagasy were formed from at least two populations, each with appreciably different allele frequency distributions. Because both of the purported parental groups of the Malagasy have the same modal allele as well as similar frequencies of the remaining alleles, this locus on its own is not very useful in elucidating the origins of the Malagasy.

Table 5.4. Frequencies of alleles at the DYS19 locus. The frequencies of modal alleles are shown in bold typeface.

Population	N	Alleles							
		Y	Z	A	B	C	D	E	F
Malagasy	**467***								
Merina	50	0	0	0.060	0.160	**0.360**	0.320	0.100	0
Betsileo	49	0	0	0.041	0.143	**0.449**	0.143	**0.224**	0
Antaisaka	50	0	0	0.040	0.100	**0.440**	0.160	**0.260**	0
Tsimihety	50	0	0	0.020	0.140	**0.580**	0.200	0.060	0
Sakalava	51	0	0	0.078	0.137	**0.511**	0.137	0.137	0
Mahafaly	27	0	0	0.111	0.037	**0.593**	0.074	**0.185**	0
Antandroy	44	0	0	0.023	0.091	**0.523**	0.227	0.136	0
Betsimisiraka	50	0	0	0.080	0.120	**0.500**	0.200	0.100	0
Antaimoro	50	**0.020**	0	0	0.100	**0.700**	0.100	0.080	0
Antankarana	13	0	0	0	0.308	**0.385**	0.230	0.077	0
Mikea	33	0	0	0.000	0.061	**0.636**	0.212	0.091	0
Non-African	**50***								
SA Indian	50	0	0	0.060	0.340	**0.480**	0.120	0	0
NO-Thai**	42	0	0	0	0.070	**0.690**	0.140	0.100	0
SO-Thai**	12	0	0	0	0.420	**0.500**	0.140	0.100	0
Kampuchea**	11	0	0	0	0.180	**0.360**	0.450	0	0
S-Borneo**	13	0	0	0.080	0.150	**0.540**	0.230	0	0
African	**167***								
Tsonga	51	0	0	0.020	0.059	**0.627**	0.196	0.098	0
Zambian	50	0	0	0	0.060	0.360	**0.420**	0.140	0.020
Gbaya	29	**0.069**	0.034	0	0	**0.552**	0.242	0.103	0
S.Sangha	17	0	0	0	0.059	0.176	**0.647**	0.118	0
Mbenzele	20	0	0	0	0.050	**0.650**	0.150	0.150	0

*number of chromosomes examined in present study.
**Kayser *et al*, 1997.

5.3.4 DYS390

Allele frequency results at this locus for individual population groups are given in **Table 5.5**. Of twelve alleles reported in the literature (17 to 28 repeats or allele A to L) (Deka *et al*, 1996; Kayser *et al*, 1997), only eight alleles (C to J) were detected in the 690 males screened in the present study. Alleles A to D and alleles J and K are reported to be rare to absent from most populations examined and alleles E to I show geographic-specific distributions (Deka *et al*, 1996; Kayser *et al*, 1997). Allele E is modal in four of the five African population samples examined, occurring at much lower frequencies in SE Asian populations and being rare to absent from other non-African populations. The high frequency of allele E in the Malagasy thus supports a large genetic input from African populations. The frequencies of alleles F and H vary tremendously between populations and are thus not very informative. Allele G however, occurs at very low frequencies in the African populations examined (0-0.048) compared to frequencies noted in the Malagasy (0.022 to 0.320) and other Asian populations [as high as 0.682 in the Taiwanese (Redd, unpublished data)]. This finding therefore supports a SE Asian origin for allele G in the Malagasy.

Table 5.5. Frequencies of alleles at the DYS390 locus. The frequencies of modal alleles are shown in bold typeface.

Population	N	Alleles							
		C	D	E	F	G	H	I	J
Malagasy	**471***								
Merina	50	0	0.020	**0.420**	0	0.240	**0.260**	0.060	0
Betsileo	50	**0.020**	0	**0.440**	0.080	**0.280**	0.180	0	0
Antaisaka	50	0	0	**0.360**	0.040	**0.320**	0.180	0.080	0.020
Tsimihety	50	0	0	**0.520**	0.020	**0.180**	0.160	0.120	0
Sakalava	51	0	0	**0.510**	0.039	0.118	**0.235**	0.098	0
Mahafaly	27	0	0	**0.593**	0	**0.259**	0.111	0.037	0
Antandroy	45	0	0	**0.668**	0.022	0.022	**0.244**	0.022	0.022
Betsimisiraka	52	**0.019**	0	**0.539**	0.096	**0.135**	0.115	0.096	0
Antaimoro	50	0	0	**0.460**	0.060	**0.320**	0.040	**0.160**	0.020
Antankarana	13	0	0	**0.538**	0.077	0.077	**0.308**	0	0
Mikea	33	0	0.030	**0.636**	0.061	0.061	**0.091**	**0.091**	0.030
Non-African	**50***								
SA Indian	50	0	0	0.040	**0.280**	0.240	0.200	0.180	0.060
Chinese**	36	0	0	0.030	0.050	**0.530**	0.250	0.130	0
Mongolian**	40	0	0	0	0.070	0.250	**0.370**	0.270	0
Borneo***	41	0	0	0.070	0	0.200	**0.370**	0.240	0.120
African	**169***								
Tsonga	50	0	0	**0.700**	0.040	0.040	**0.220**	0	0
Zambian	52	0	0	**0.828**	0.019	0.038	**0.115**	0	0
Gbaya	29	0	0	**0.621**	0.138	0	**0.241**	0	0
S.Sangha	17	0	0.059	0.353	**0.588**	0	0	0	0
Mbenzele	21	0	0.095	**0.524**	0.000	0.048	**0.190**	0.143	0

*number of chromosomes typed in present study.
**Kayser *at al*, 1997.
***Redd, unpublished data.

5.3.5 DYS391

Six alleles have been reported at this locus (8 to 13 repeats or alleles A to F) (Kayser *et al*, 1997) and the frequency distributions of the four alleles (B to E) detected in 685 males in the present study are summarized in **Table 5.6**. Allele C is the modal allele in all the populations screened, with the distributions of the other alleles not varying appreciably between populations. An exception is the S.Sangha in whom allele C appears to have reached fixation; this could however be attributed to the small sample size. These results indicate that by itself, this locus is not particularly useful for elucidating Malagasy ancestry.

Table 5.6. Frequencies of alleles at the DYS391 locus. The frequencies of modal alleles are shown in bold typeface.

Population	N		Alleles		
		B	C	D	E
Malagasy	**470***				
Merina	51	0	**0.725**	0.275	0
Betsileo	48	0	**0.625**	0.333	0.042
Antaisaka	51	0.020	**0.804**	0.176	0
Tsimihety	50	0	**0.660**	0.340	0
Sakalava	51	0	**0.804**	0.157	0.039
Mahafaly	27	0	**0.852**	0.148	0
Antandroy	45	0.044	**0.689**	0.178	0.089
Betsimisiraka	51	0.020	**0.725**	0.255	0
Antaimoro	50	0	**0.720**	0.280	0
Antankarana	13	0.077	**0.769**	0.154	0
Mikea	33	0.030	**0.788**	0.182	0
Non-African	**50***				
SA Indian	50	0.020	**0.82**0	0.160	0
Chinese**	36	0	**0.780**	0.220	0
Mongolian**	40	0.150	**0.670**	0.170	0
S-Borneo**	13	0.150	**0.540**	0.310	0
African	**165***				
Tsonga	50	0	**0.880**	0.120	0
Zambian	50	0.020	**0.880**	0.100	0
Gbaya	28	0	**0.750**	0.250	0
S.Sangha	17	0	**1.000**	0	0
Mbenzele	20	0	**0.650**	0.300	0.050

*number of chromosomes typed in present study.
**Kayser *et al*, 1997.

5.4 Exact test results

Comparisons of allele frequency distributions between populations were performed by an adaptation of Fisher's Exact method (Weir, 1996). The P values for individual comparisons at the five loci examined are shown in **Tables 5.7** to **5.10** and significant differences ($p<0.050$) are shown in bold typeface.

Very few significant differences were noted for most comparisons at locus DYS393, DYS19 and DYS390 between individual Malagasy groups, except when the Antandroy and Mikea were involved. This suggests that their Y chromosome gene pool may be somewhat different to the other nine Malagasy groups. The SA Indian, Zambian and S.Sangha consistently showed significant differences to all other populations while comparisons between each of the Malagasy groups and between the African groups gave varying results.

Results at locus DYS390 show many differences between the highland and lowland groups with the overall pattern suggesting that the lowland groups are closer to the African populations than are the highland groups. Very few significant values were obtained for comparisons at locus DYS391 supporting the idea that this locus is not very useful by itself for addressing questions of population history.

Table 5.7. Results of pairwise tests of differentiation at the DYS393 locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	SA Indian	Tsonga	Zambian	Mbenzele	Gbaya	S.Sangha
MR	0.864	0.392	0.303	0.075	0.512	**<0.001**	0.537	0.306	0.687	**0.042**	**<0.001**	**0.001**	**<0.001**	**<0.001**	0.114	**<0.001**
BT		0.255	0.184	0.059	0.509	**<0.001**	0.440	0.207	0.312	**0.012**	**<0.001**	**0.001**	**0.001**	**<0.001**	0.107	**<0.001**
AS			0.285	0.842	0.873	**<0.001**	0.212	0.806	0.641	**0.009**	**<0.001**	0.068	**0.000**	**0.001**	0.118	**<0.001**
TS				0.350	0.746	**<0.001**	0.709	0.306	0.876	0.091	**<0.001**	0.083	**0.026**	**0.002**	0.570	**<0.001**
SK					0.852	**<0.001**	0.127	0.584	0.546	**0.004**	**<0.001**	0.504	**0.000**	**0.003**	0.173	**<0.001**
MF						**<0.001**	0.638	0.901	0.537	**0.026**	**<0.001**	0.393	**0.006**	**0.003**	0.418	**<0.001**
AD							**0.006**	**<0.001**	0.239	0.432	**<0.001**	**<0.001**	**0.017**	**0.011**	**0.005**	**<0.001**
BS								0.157	0.948	0.184	**0.001**	**0.009**	**0.019**	**0.001**	0.848	**<0.001**
AM									0.451	**0.001**	**<0.001**	**0.033**	**<0.001**	**<0.001**	0.061	**<0.001**
AK										0.769	0.183	0.156	0.187	0.057	0.600	**<0.001**
MK											**<0.001**	**<0.001**	0.462	**0.039**	0.188	**<0.001**
SA Indian												**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Tsonga													**0.003**	**0.019**	0.148	**<0.001**
Zambian														0.221	0.206	**0.003**
Mbenzele															0.052	0.195
Gbaya																**<0.001**

Table 5.8. Results of pairwise tests of differentiation at the DYS19 locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	SA Indian	Tsonga	Zambian	Mbenzele	Gbaya	S.Sangha
MR	0.168	0.105	0.233	0.229	**0.024**	0.434	0.573	**0.001**	0.806	0.111	**0.009**	0.054	0.208	0.164	**0.022**	0.172
BT		0.974	0.150	0.764	0.356	0.588	0.455	0.067	0.450	0.190	**0.001**	0.144	**0.017**	0.646	**0.024**	**0.005**
AS			0.063	0.534	0.401	0.578	0.317	**0.030**	0.249	0.184	**<0.001**	0.132	**0.023**	0.621	**0.040**	**0.002**
TS				0.436	0.073	0.737	0.650	0.421	0.529	0.774	**0.036**	0.727	**0.018**	0.596	0.079	**0.004**
SK					0.584	0.603	0.930	0.147	0.464	0.320	**0.014**	0.310	**0.008**	0.683	**0.037**	**0.004**
MF						0.249	0.370	0.115	0.056	0.153	**0.000**	0.199	**0.003**	0.602	0.079	**0.000**
AD							0.800	0.237	0.399	0.882	**0.002**	0.872	0.211	0.858	0.251	**0.019**
BS								0.128	0.530	0.462	**0.017**	0.491	**0.039**	0.593	0.068	**0.010**
AM									0.107	0.656	**0.001**	0.507	**<0.001**	0.769	0.075	**0.000**
AK										0.147	0.316	0.113	0.114	0.180	**0.043**	0.055
MK											**0.001**	1.000	0.121	0.892	0.466	**0.009**
SA Indian												**<0.001**	**<0.001**	**0.003**	**<0.001**	**<0.001**
Tsonga													**0.041**	0.951	0.262	**0.001**
Zambian														0.128	0.067	0.497
Mbenzele															0.593	**0.006**
Gbaya																**0.016**

Table 5.9. Results of pairwise tests of differentiation at the DYS390 locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	SA Indian	Tsonga	Zambian	Mbenzele	Gbaya	S.Sangha
MR	0.107	0.513	0.385	0.327	0.508	**0.004**	**0.036**	**0.008**	0.356	**0.012**	**<0.001**	**0.002**	**0.001**	0.116	**<0.001**	**<0.001**
BT		0.287	**0.050**	**0.045**	0.353	**<0.001**	0.097	**0.046**	0.492	**0.011**	**<0.001**	**0.003**	**<0.001**	**0.004**	**0.008**	**<0.001**
AS			0.394	0.163	0.512	**0.000**	0.101	0.340	0.298	**0.018**	**<0.001**	**<0.001**	**<0.001**	**0.036**	**<0.001**	**<0.001**
TS				0.790	0.668	**0.019**	0.557	0.163	0.325	0.233	**<0.001**	0.086	**0.003**	0.344	**0.045**	**<0.001**
SK					0.281	0.180	0.478	**0.018**	0.790	0.300	**<0.001**	0.072	0.082	0.690	0.140	**<0.001**
MF						**0.022**	0.402	0.454	0.195	0.219	**<0.001**	**0.017**	**0.009**	0.101	**0.004**	**<0.001**
AD							**0.020**	**<0.001**	0.527	0.229	**<0.001**	0.878	0.286	0.095	0.375	**<0.001**
BS								0.181	0.613	0.701	**<0.001**	**0.021**	**0.008**	0.190	**0.061**	**0.001**
AM									**0.037**	0.063	**<0.001**	**<0.001**	**<0.001**	**0.006**	**<0.001**	**<0.001**
AK										0.561	**<0.001**	0.439	0.076	0.476	0.570	**0.001**
MK											**<0.001**	0.072	0.082	0.690	0.140	**<0.001**
SA Indian												**<0.001**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
Tsonga													0.465	**0.015**	0.364	**<0.001**
Zambian														**0.008**	**0.038**	**<0.001**
Mbenzele															**0.023**	**<0.001**
Gbaya																**0.001**

Table 5.10. Results of pairwise tests of differentiation at the DYS391 locus. Significant differences (p<0.050) are shown in bold typeface. Abbreviations used for the Malagasy groups are as given in the Nomenclature section.

	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	SA Indian	Tsonga	Zambian	Mbenzele	Gbaya	S.Sangha
MR	0.336	0.357	0.530	0.153	0.243	0.037	1.000	1.000	0.204	0.284	0.249	0.080	**0.042**	0.323	1.000	**0.018**
BT		0.057	0.496	0.116	0.092	0.143	0.236	0.346	0.217	0.115	**0.030**	**0.007**	**0.002**	1.000	0.441	**0.006**
AS			0.089	0.644	1.000	0.132	0.724	0.311	0.497	1.000	1.000	0.511	0.709	0.234	0.719	0.113
TS				**0.038**	0.117	**0.016**	0.436	0.652	0.131	0.137	0.062	**0.013**	**0.006**	0.412	0.471	**0.003**
SK					0.882	0.336	0.185	0.150	0.335	0.486	0.570	0.341	0.254	0.411	0.478	0.190
MF						0.295	0.611	0.279	0.453	0.851	1.000	0.713	0.821	0.205	0.492	0.131
AD							0.118	**0.034**	0.846	0.411	0.121	**0.028**	0.058	0.604	0.293	0.084
BS								1.000	0.457	0.803	0.676	0.103	0.084	0.519	1.000	**0.037**
AM									0.210	0.281	0.227	0.079	**0.048**	0.333	1.000	**0.008**
AK										0.832	0.496	0.181	0.354	0.430	0.408	0.074
MK											0.881	0.365	0.507	0.369	0.871	0.107
SA Indian												0.579	0.782	0.180	0.607	0.164
Tsonga													1.000	**0.049**	0.220	0.289
Zambian														**0.041**	0.147	0.472
Mbenzele															0.602	**0.008**
Gbaya																**0.032**

5.5 Haplotype data

The relative positions on the Y chromosome of three of the five loci studied are not known (DYS393, DYS390 and DYS391) and haplotypes were constructed for each individual on the basis of their genotypes in the following order: YAP, DYS393, DYS19, DYS390 and DYS391. The SE Asian/Oceanic haplotype data based on the same four STR loci were provided by Dr Alan Redd (unpublished data) and have been included in some of the comparative analyses that follow. Since the YAP+ allele has not been found in about 1000 Y chromosomes from SE Asia and Oceania (Hammer *et al*, 1996), it has been assumed that the samples used by Alan Redd are all YAP-.

The statistical package ARLEQUIN 1.1 (Schneider *et al*, 1997) was used to calculate haplotype frequencies, haplotype sharing and genetic distances between populations. The hierarchical distribution of Y chromosome diversity, measured as the variance components among individuals, populations and geographic groups was computed by use of the analysis of molecular variance (AMOVA) component of the software.

5.5.1 Haplotype distribution and frequency estimates

The 11 Malagasy groups examined were divided into highland (Merina and Betsileo) and lowland groups (the remaining nine Malagasy groups) for this part of the analysis. The haplotype distributions in these two were then compared to those in three other groups: the SA Indian, SE Asian/Oceanic and African. The haplotype designations and their frequencies are given in **Table 5.11 and 5.12**. To reduce the complexity of the tables, only data for the haplotypes present in the Malagasy and those shared by them with the other

populations are included in the tables. Haplotypes absent from the Malagasy groups were thought to have no relevance to the present discussion. The STR haplotypes have also been divided into two groups or 'frames' based on their YAP status; frame 1 comprises all YAP+ chromosomes (**Table 5.11**) and frame 2 comprises all YAP- chromosomes (**Table 5.12**).

Altogether, 221 unique haplotypes were observed among 1 097 males; 683 males in the present study and 414 in Alan Redd's sample. Fifty-seven of the STR haplotypes occurred on YAP+ chromosomes (h1-h57) and 164 on YAP- chromosomes (h58-h221). The numbers of YAP+ and YAP- haplotypes found in the population groups examined are as follows: 41 YAP+ and 55 YAP- haplotypes in the Malagasy (N=467), 34 YAP- haplotypes in the SA Indians (N=50), 36 YAP- haplotypes in the SE Asian/Oceanic sample (N=414) and 35 YAP+ and 15 YAP- haplotypes in the Africans (N=166). A large number of the haplotypes (h1 to h57) appear to be restricted to African and Malagasy populations while haplotypes h58 to h221 were mainly found in the SA Indian or SE Asian/Oceanic populations. This distribution is directly related to the allele status at the YAP locus. The Malagasy have haplotypes in common with all three groups supporting historical and linguistic data for ancestry from these geographic regions. Forty-seven of the haplotypes detected in the Malagasy also occurred in one or more of the other populations used in the comparative analysis; 27 were also seen in the African populations and 44 in the Asian populations (**Table 5.11**).

Table 5.11. Frequencies of YAP+ haplotypes constructed in the order YAP-DYS393-DYS19-DYS390-DYS391 in Malagasy, non-African and African populations. Haplotyes shown in bold typeface showed homoplasy in the present study.

Number	YAP+	Highland		Lowland		SA Indian		SEA/Oceanic		African		Average
	haplotypes	98		369		50		414		166		frequency
h1	**+DCEC**			1	0.003					1	0.006	0.005
h2	**+EYGC**			1	0.003							0.002
h4	**+EACC**			1	0.003							0.002
h5	**+EAHC**			6	0.016							0.014
h6	**+EAIC**	1	0.010	1	0.003							0.005
h8	**+EBED**			2	0.005							0.005
h9	**+EBGD**	1	0.010									0.002
h10	**+EBHC**	1	0.010	2	0.005					2	0.012	0.012
h11	**+EBHD**	7	0.071							1	0.006	0.019
h12	**+EBIC**			1	0.003							0.002
h13	**+EBID**	1	0.010	9	0.024					1	0.006	0.027
h14	**+EBJC**			3	0.008							0.007
h15	**+ECDC**	1	0.010									0.002
h17	**+ECEC**	10	0.102	56	0.152					28	0.169	0.227
h18	**+ECED**	6	0.061	20	0.054					3	0.018	0.070
h19	**+ECEE**	1	0.010									0.002
h20	**+ECFC**	1	0.010	3	0.008					3	0.018	0.017
h22	**+ECJC**			1	0.003							0.002
h23	**+EDEC**	1	0.010	4	0.011					3	0.018	0.019
h24	**+EDED**	2	0.020									0.005
h25	**+EEEC**	1	0.010	3	0.008					1	0.006	0.012
h26	**+FAHC**			1	0.003							0.002
h28	**+FBEC**			1	0.003							0.002
h30	**+FCDD**			1	0.003							0.002
h31	**+FCEC**	7	0.071	29	0.079					7	0.042	0.104
h32	**+FCED**	2	0.020	6	0.016					2	0.012	0.024
h33	**+FCEE**			3	0.008							0.007
h35	**+FCID**			1	0.003							0.002
h36	**+FDEC**	5	0.051	15	0.041					15	0.090	0.084
h38	**+FDFC**			1	0.003					1	0.006	0.005
h40	**+FEEC**	2	0.020	15	0.041					5	0.030	0.053
h41	**+FEFC**	1	0.010	2	0.005					2	0.012	0.012
h44	**+GCEC**			4	0.011					12	0.072	0.039
h45	**+GCED**	1	0.010	3	0.008							0.010
h46	**+GCEE**			3	0.008							0.007
h47	**+GCFC**			1	0.003							0.002
h49	**+GDEC**	1	0.010	14	0.038					12	0.072	0.065
h52	**+GDGD**			1	0.003							0.002
h53	**+GEEC**	2	0.020	5	0.014					7	0.042	0.034
h54	**+HCEC**			2	0.005							0.005
h55	**+HDEC**			2	0.005					2	0.012	0.010

Table 5.12. Frequencies of YAP- haplotypes constructed in the order YAP-DYS393-DYS19-DYS390-DYS391 in Malagasy, non-African and African populations. Haplotypes shown in bold typeface displayed homoplasy in the present study.

Number	YAP-	Highland		Lowland		SA Indian		SEA/Oceanic		African		Average
	haplotypes	98		369		50		414		166		frequency
h58	-CAFC			1	0.003			1	0.002			0.003
h59	-CAFD			1	0.003							0.001
h60	-CBFB			1	0.003							0.001
h64	-DAGC	1	0.010									0.001
h65	-DAHC			1	0.003							0.001
h67	-DBED			3	0.008			1	0.002			0.006
h68	-DBFC			2	0.005	1	0.020	1	0.002			0.006
h70	-DBGC	1	0.010	6	0.016	2	0.040					0.013
h71	-DBGD	1	0.010	1	0.003			1	0.002			0.004
h73	-DBHD			3	0.008			1	0.002			0.006
h79	-DCGC			3	0.008	1	0.02	4	0.010			0.012
h81	-DCHC			1	0.003			3	0.007			0.006
h82	-DCHD			1	0.003							0.001
h90	-DDGB			1	0.003			1	0.002			0.003
h91	-DDGC	1	0.010			1	0.02	1	0.002			0.004
h92	-DDHB			2	0.00542							0.003
h107	-EAGC	3	0.031	2	0.005			1	0.002			0.009
h108	-EAIC			1	0.003	1	0.020					0.003
h118	-EBGC			1	0.003	1	0.020	6	0.014			0.012
h119	-EBGD			1	0.003			1	0.002			0.003
h121	-EBHD	1	0.010	2	0.005			7	0.017	1	0.006	0.016
h131	-ECEC			2	0.005			2	0.005			0.006
h133	-ECFC	1	0.010					12	0.029			0.019
h136	-ECGC	3	0.031	12	0.033			38	0.092			0.078
h137	-ECGD			4	0.011			6	0.014			0.015
h139	-ECHC	2	0.020	18	0.049	3	0.060	22	0.053	9	0.054	0.079
h140	-ECHD			5	0.014	1	0.020	3	0.007	4	0.024	0.019
h141	-ECIC			2	0.005	2	0.040	7	0.017			0.016
h142	-ECID			3	0.008	3	0.060					0.009
h151	-EDFC			1	0.003			2	0.005			0.004
h152	-EDFD	1	0.010	1	0.003			2	0.005			0.006
h153	-EDGC	4	0.041	9	0.024			11	0.027	1	0.006	0.037
h154	-EDGD	1	0.010	1	0.003			2	0.005			0.006
h155	-EDHC	4	0.041	5	0.014			4	0.010	2	0.012	0.022
h156	-EDHD	1	0.010									0.001
h157	-EDIC	1	0.010	4	0.011	1	0.020	4	0.010			0.015
h162	-EEFC			1	0.003			1	0.002			0.003
h163	-EEGC	6	0.061	16	0.043			8	0.019			0.044
h164	-EEGD	4	0.041	3	0.008			1	0.002			0.012
h165	-EEHC			1	0.003			2	0.005	2	0.012	0.007
h172	-FBCA	1	0.010									0.001
h174	-FBFb			1	0.003							0.001
h177	-FBGC			1	0.003	2	0.040	2	0.005			0.007
h179	-FBHC			1	0.003	1	0.020	1	0.002			0.004
h190	-FCGD			1	0.003			1	0.002			0.003
h191	-FCHC	5	0.051	5	0.014	1	0.020	8	0.019			0.028
h192	-FCHD			2	0.005	1	0.020	10	0.024	2	0.012	0.022
h193	-FCIC			1	0.003	1	0.020	6	0.014			0.012
h194	-FCID			4	0.011			4	0.010	1	0.006	0.013
h199	-FDEC	1	0.010									0.001
h204	-FDHC			1	0.003			4	0.010			0.007
h206	-FEGC			1	0.003			2	0.005			0.004
h214	-GCHC			1	0.003			2	0.005			0.004
h216	-GCID			3	0.008			2	0.005			0.007
h218	-GDEC			1	0.003							0.001

A substantial difference was noted between the numbers of unique haplotypes present on YAP+ and YAP- backgrounds. Fifty-seven unique haplotypes were detected on 415 YAP+ chromosomes whereas 164 haplotypes were detected on 682 YAP- chromosomes. These numbers indicate a greater diversity of YAP- chromosomes. This is not surprising as the insertion of the *Alu* element on the Y chromosome is thought to have occurred relatively recently (55 000 ± 19 000 years ago, Hammer *et al*, 1998) and the YAP- chromosome would thus have been evolving and gathering variation for a much longer time than the descendants of the new YAP+ chromosome.

5.5.1.1 YAP+ haplotypes (frame 1)

All calculations in this section are based on the combined Malagasy and African YAP+ data set. Thirty seven of the YAP+ haplotypes which were exclusive to the African and Malagasy had frequencies of less than 1%, i.e. they were detected less than four times in the sample (**Table 5.11**). Twenty-four of these 37 haplotypes were singletons. Interestingly, nine of the remaining haplotypes that occurred only twice or thrice in the sample were population-specific in that they were detected either only in the African or only in the Malagasy sample. This may support their recent origin in that particular population. Twenty haplotypes were detected at frequencies greater than 1%; 10 occurred at frequencies ranging from 1 to 2%, four at frequencies ranging from 2 to 4% and another four at frequencies ranging from 5 to 9%. Only two haplotypes exceeded a frequency of 10%: h31 at a frequency of 10.4% and h17 at a frequency of 22.7%. Together, these 20 haplotypes account for 87.1% of the YAP+ chromosomes in the present data. The significance of this frequency distribution is discussed further in **section 5.6.1**.

As may be expected from a population formed by the amalgamation of two or more groups, the Malagasy highland and lowland groups together have more unique haplotypes (41) than representatives of one of their progenitors, the African populations, who had only 35. This could however be related to the difference in the sample sizes (almost threefold difference) between the Malagasy and African populations studied; N=467 for the Malagasy and N=166 for the Africans. Most haplotypes that are detected are rare and therefore the number of haplotypes observed in any population would be highly dependent on the population sample size.

5.5.1.2 YAP- haplotypes (frame 2)

Fifty-five of the 164 YAP- haplotypes (h58-h221) detected in the Malagasy, African and non-African populations are shown in **Table 5.12**. One hundred and forty of these haplotypes were present at frequencies of less than 1% (expressed as a proportion of total YAP- chromosomes and thus detected less than seven times in the YAP- sample of 682 chromosomes). Of the remaining 24 haplotypes, 16 occurred at frequencies ranging from 1 to 2%, three at frequencies ranging from 2 to 3% and another three at frequencies of between 3 and 5%. Only two haplotypes exceeded a frequency of 7%: h136 was detected at a frequency of 7.8% and h139 at a frequency of 7.9%. Together, these 24 haplotypes account for 57% of the YAP- haplotypes in the present data set and the significance of these haplotype frequency differences is discussed further in **section 5.6.1**.

5.5.1.3 Haplotype diversity

Haplotype diversities were calculated as described by Nei (1987) and are presented in **Table 5.13**. For all populations except the S.Sangha, the haplotype diversity was greater than 0.85 (0.860 to 0.983) in support of the observation that a large number of the 221 haplotypes observed had a frequency of less than 1%. The decreased diversity in the S.Sangha could be a reflection of the small sample size (N=17), polygamy, founder effect and genetic drift or a combination of these. Haplotype diversities calculated from data on the same four STR loci in African-American, European American and Hispanic population groups ranged from 0.920 to 0.969 (Redd *et al*, 1997). The high diversity obtained using just five Y chromosome polymorphisms indicates their usefulness for tracing recent historic and migratory events.

Table 5.13. Y chromosome haplotype diversities in Malagasy, non-African and African populations.

Population	N	Number of haplotypes	D	SE
Malagasy	**467***			
Merina	49	28	0.967	0.011
Betsileo	49	26	0.960	0.013
Antaisaka	50	26	0.940	0.019
Tsimihety	50	24	0.948	0.017
Sakalava	51	29	0.969	0.010
Mahafaly	27	12	0.860	0.051
Antandroy	44	21	0.940	0.018
Betsimisiraka	51	27	0.961	0.012
Antaimoro	50	24	0.940	0.018
Antankarana	13	10	0.962	0.041
Mikea	33	20	0.934	0.028
Non-African	**50***			
SA Indian	50	34	0.983	0.007
Australian**	107	46	0.927	0.018
New Guinean (highland)**	33	20	0.962	0.016
New Guinean (coastal)**	26	23	0.991	0.013
Moluccans**	36	16	0.910	0.028
Tengarrans**	35	25	0.971	0.016
Borneo**	41	31	0.985	0.009
Javanese**	29	23	0.983	0.014
Malay**	22	20	0.991	0.017
Filippino**	41	18	0.883	0.034
Taiwanese**	44	15	0.895	0.030
African	166*			
Tsonga	50	20	0.889	0.030
Zambian	50	23	0.933	0.020
Gbaya	29	15	0.941	0.025
S.Sangha	17	5	0.772	0.070
Mbenzele	20	15	0.958	0.033

*number of chromosomes typed in present study.
**Redd, unpublished data.

5.5.2 Haplotype sharing

The haplotype sharing data between the 27 population samples in the data set are summarized in **Table 5.14**. The number of haplotypes detected in each population is also shown for perspective, i.e. to give an idea of the number of shared haplotypes versus the number of haplotypes detected in individual populations. As discussed in **section 5.5.1.1**, many haplotypes occur infrequently (singletons or low frequencies) and very few are shared by all the population groups. Populations sharing a large number of haplotypes are likely to be more closely related to each other than to those sharing only a few.

As would be expected, individual Malagasy groups share a large number of haplotypes with each other (up to 14 haplotypes) as do four of the five African groups. Individual Malagasy groups also share many haplotypes (almost as many as the number shared between themselves) with the African populations, especially the Tsonga, Zambian and Gbaya populations. This suggests that these particular African populations or others who are closely related to them have contributed to the Malagasy gene pool. Overall, the Malagasy share almost 50% of their haplotypes (19 of their 41 haplotypes) with the African sample. This large amount of sharing is unlikely to be due to chance or recurrent mutation alone as other studies have shown that even closely related populations share very few haplotypes (Roewer *et al*, 1996). These data suggest that Africans have contributed significantly to the Malagasy gene pool. Thirteen of the 22 Malagasy haplotypes not seen in the African data set presented in this study have been detected in other southern African populations (N=700; Young, 1999 and unpublished data). The origins in the Malagasy of the remaining nine YAP+ haplotypes remains unresolved. A number of these haplotypes could have been missed in the African data set due to sampling effects, or they could have

been inherited from an as yet unsampled African population, or some of them may have arisen on Madagascar. A small proportion could also be of Arabic origin but again, due to the lack of data from Arabic sources, this hypothesis cannot be tested. David Goldstein kindly provided data on the same five loci used in the present study for 49 Yemeni males (unpublished data; 5 YAP+ and 20 YAP- haplotypes). The two YAP+ haplotypes that this data set had in common with the Malagasy have already been detected in African populations and thus the Yemeni data do not help resolve the issue of Arabic contributions to the Malagasy.

Interestingly, the Malagasy share 15 haplotypes with the SA Indians and 39 with the SE Asian/Oceanic sample (**Table 5.11**). Twelve of the fifteen haplotypes shared by the SA Indians and the Malagasy however are also present in the SE Asian/Oceanic sample. This increased degree of haplotype sharing between the SA Indian and the SE Asian populations suggests that they have common male ancestors (see **section 4.5** for a possible explanation of the high haplotype sharing between these two populations). The origins of these 12 haplotypes in the Malagasy can therefore not be resolved. The large extent of haplotype sharing between the SE Asian/Oceanic sample and the Malagasy however strongly supports a genetic contribution from populations from this region to the Malagasy gene pool. The Malagasy also shared seven YAP- haplotypes with the Yemeni sample, but as with the YAP+ haplotypes, all seven were present in one or more of the other non-African population groups examined in this study. Interestingly, h70 (=DBGC) is absent from African and SE Asian/Oceanic populations but is found at a frequency of 0.377 in the Yemeni, 0.040 in the SA Indians and 0.015 in the Malagasy. The ancestry in the Malagasy of most of these shared haplotypes thus remains unresolved.

None of the ten SE Asian/Oceanic groups, not even the sample from the Barito River valley region in Borneo can be singled out as having significantly more haplotypes in common with the Malagasy than the others. The Malagasy and the Borneo sample share only nine haplotypes. This could be due to a number of reasons, one of which may be that the present-day Borneo population may have adopted the language of the true Asian ancestors of the Malagasy, i.e. they may have been linguistically steamrolled (Diamond, 1997). The Bornean, Tengarran, Javanese and Malay groups share many more haplotypes with each other than they do with the other SE Asian groups, which supports a common ancestry for them (Bellwood, 1987).

The non-African and African groups share very few haplotypes (the maximum being three), largely due to huge differences in allele frequencies at the YAP locus. In the African group, the Tsonga and Zambians share the largest number of haplotypes (N=10). This is not surprising since both groups are Bantu-speakers who share a recent common ancestry. The S.Sangha have very few or no haplotypes in common with any other population in the data set, not even with the Gbaya who belong to the same language group and amongst whom they live. This could be due to the effects of a small sample, a different ancestry or a diffe . . emographic history from the other African groups examined.

Table 3.14. Y chromosome haplotype sharing between populations as calculated using Arlequin (ver1.1). The number of chromosomes sampled (N) and the number of unique haplotypes (hts) found in each population are also given.

	N	Unique hts	BT	AS	TS	SK	MF	AD	BS	AM	AK	MK	SAI	AUS	PNGc	PNGh	MOL	TEN	BOR	JAV	MAL	FIL	TAI	TSO	ZAM	MBE	GBA	SSA
MR	49	28	13	12	14	11	7	7	12	10	5	8	5	2	5	4	5	6	4	4	2	4	6	8	10	4	5	2
BT	49	26		11	12	10	6	7	9	12	5	10	2	3	6	3	4	3	4	3	1	4	6	8	11	4	8	1
AS	50	26			11	12	5	8	11	11	5	9	5	2	4	3	3	4	4	5	3	5	6	7	9	6	8	2
TS	50	24				9	6	9	9	12	4	8	4	3	6	7	3	5	6	5	4	5	6	8	7	6	6	1
SK	51	29					7	9	13	9	6	10	9	3	5	2	2	3	7	4	2	2	9	11	13	5	8	2
MF	27	12						3	7	6	3	4	1	0	2	1	1	1	1	1	1	3	3	4	5	4	4	0
AD	44	21							7	8	4	9	3	2	2	4	1	1	3	4	3	2	2	7	8	5	7	1
BS	51	27								10	5	8	5	4	3	3	5	6	5	5	3	6	5	10	11	4	8	2
AM	50	24									5	10	4	1	3	3	2	3	3	4	2	5	6	7	7	4	6	0
AK	13	10										5	0	0	0	1	1	1	1	0	1	0	1	5	6	4	4	0
MK	33	20											3	1	2	1	1	2	3	2	0	1	4	6	10	3	6	1
SAI	50	34												3	8	7	3	6	11	8	4	6	5	1	2	3	2	0
AUS	107	46													6	1	4	4	5	3	0	4	2	3	1	0	0	0
PNGc	26	23														5	4	5	5	6	0	5	4	1	1	1	1	0
PNGh	33	20															5	6	7	8	5	3	4	2	1	3	1	0
MOL	36	16																8	4	2	1	5	3	2	2	1	1	0
TEN	35	22																	9	4	3	7	5	2	2	2	2	0
BOR	41	31																		9	8	5	5	1	3	2	2	0
JAV	29	23																			4	4	5	1	1	3	2	0
MAL	22	20																				4	2	2	2	2	1	0
FIL	41	18																					6	3	2	2	1	0
TAI	44	15																						1	2	2	1	0
SIN	25	12																						0	0	2	0	0
TSO	50	20																							10	5	6	2
ZAM	50	23																								6	8	2
MBE	20	15																								-	4	0
GBA	29	15																									-	2
SSA	17	5																										-

SAI-SA indian, AUS-Australian, PNG-Papua new Guinea, c-coastal, h-highland, MOL-Moluccans, TEN-Tengarans, BOR-Borneo, JAV-Java, MAL-Malaysian, FIL- Fillipino, TAI- Taiwan,TSO-Tsonga, ZAM-Zambians, MBE-Mbenzele, GBA-Gbaya, SSA-S.Sanghs

5.5.3 Genetic distances and phylogenetic trees

The genetic affinities of the Malagasy with African and non-African populations were inferred from NJ trees constructed using a number of different data sets and genetic distance measures. The first of these was fr allele frequency data at the five Y chromosome loci and D_{SW} distances (Shriver *et al*, 1995) computed with the programme DSW (Ota, 1996). The second and third sets of distances used were computed with the programme ARLEQUIN (ver 1.1) (Schneider *et al*, 1997) using haplotype frequency data and either Reynold's distance method (Reynolds *et al*, 1983) or Slatkin's R_{ST} distance method (Slatkin, 1995). Both methods are based on F_{ST} values that have been modified to linearize the distances with time. Neighbour joining trees based on the second and third sets of distances were generated by the computer programme MEGA (Kumar *et al*, 1993) or TREEVIEW, which is part of the DSW package (Tamura, 1993).

The unrooted NJ tree generated using Reynold's genetic distance is shown in **Figure 5.3**. The trees generated from the D_{SW} and R_{ST} genetic distances are almost identical, with minor differences in placement of a few Malagasy and non-African groups in their respective clusters and are therefore not shown. The non-African populations are separated from the African populations with the populations from each of the two geographic regions clustering closely to each other. Eight of the eleven Malagasy ethnic groups also cluster closely together and are placed between the non-African and African clusters. The placement of the Antandroy compared to the other Malagasy was already apparent from results of the Fisher's exact test used to explore population differentiation on the basis of allele frequency distributions (**Tables 5.7, 5.9** and **section 5.4**). The placement of the Malagasy groups relative to the Barito sample in the NJ tree does not support the

hypothesis put forward by linguists that the Malagasy ancestors originated in the Barito River region of Borneo (Dahl, 1951). The close clustering of the SE Asian populations already shown by the haplotype sharing comparisons suggests a recent common ancestry for them and/or continuous gene flow between this group of populations.

The NJ tree also suggests a close association between some of the Malagasy groups and the Mbenzele, contrasting with the haplotype sharing information, but this association probably results from the similarity in YAP+ frequencies between the two groups. The Malagasy appear to be genetically closest to Negroid populations from central and southern Africa.

A very significant clustering of populations by region/continent is portrayed in the NJ tree constructed using Y chromosome haplotype data. The apparent population relationships are largely concordant with those obtained using autosomal STR and RFLP markers (Chapters 3 and 4) as well as other types of markers (Nei and Roychoudary, 1993; Cavalli-Sforza *et al*, 1994; Ruiz Linarez *et al*, 1996).

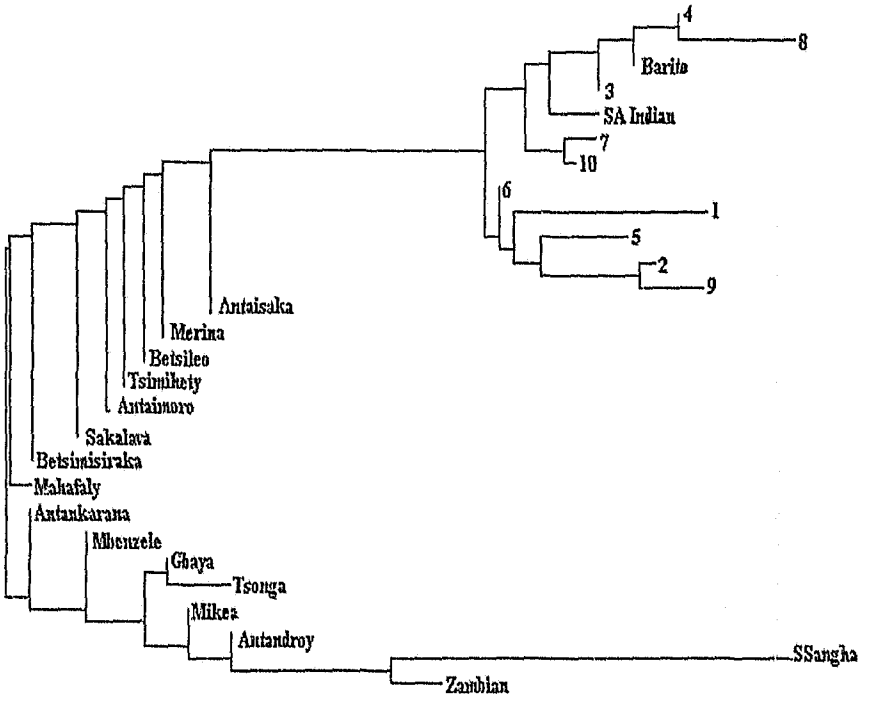


Figure 5.3. Unrooted NJ tree showing the genetic relationships between several Malagasy, African and non-African populations based on Y chromosome haplotype frequencies. Numbers 1 to 10 represent SE Asian/Oceanic populations studied by Alan Redd (personal communications).

5.6 Network analyses

A number of studies (Cooper *et al*, 1996; Roewer *et al*, 1996) have attempted to use parsimony analysis to examine the evolutionary relationships of Y chromosome haplotypes and to see if meaningful population relationships would emerge from the clustering patterns of the haplotypes. A major problem with this analysis is that there are a large number of equally parsimonious trees which can be derived, each with different interpretations of the data. The lack of success with the parsimony approach prompted the use of networks based on the stepwise mutation model (SMM) to examine the relationships and distribution of Y chromosome haplotypes. This approach is similar to that used successfully to analyze mitochondrial sequence data (Bandelt *et al*, 1995) and makes provisions for more than one pathway to describe convergent mutational events that may generate haplotypes which are identical in state (homoplasy). A network of haplotypes is constructed based on the concept of parsimonious networks (Cooper *et al*, 1996), by beginning with the commonest haplotype and sequentially linking adjacent haplotypes on different levels that differ by one repeat only. Lines connect all haplotypes related in this way and haplotypes on each level differ by one repeat from haplotypes to which they are connected in the previous level. This approach has been used in a number of other studies to examine population relationships (Linares *et al*, 1996; Hurles *et al*, 1998; Malaspina *et al*, 1998) and was adopted in this study.

If the haplotypes from a population fit well into an SMM network, one could infer that the population has been in existence long enough to generate the linked haplotypes around the founding haplotypes; or, that the effective population size was large enough to maintain a good representation of the linked haplotypes detected. If the haplotypes from a population

do not form an SMM network, it may be assumed that the population has recently passed through a bottleneck, or that the sample was not large enough, or that there has been a recent influx of individuals from one or more genetically 'different' populations.

In the present study, haplotypes were first divided on the basis of YAP status (frames 1 and 2). Only haplotypes from the African, Malagasy and SA Indian populations are shown on the resulting frame 1 (YAP+) and frame 2 (YAP-) single step networks which are given in **Figures 5.4** and **5.5**. The significance of the colors, shading and different sizes of the boxes used in the network are explained in the **KEYS** and captions in the figures. The color-coding shows the unique or shared haplotypes present in the different population groups. The sizes of the box around each haplotype is proportional to the frequency of the haplotype.

The central cores of both networks appear the most complex with the haplotypes connected to many other adjacent haplotypes, suggesting that different paths may have been followed during the formation of that haplotype. Most of these central haplotypes also occur at frequencies that are greater than 1% (**Table 5.11 and 5.12**) and account for a majority of the Y chromosomes. The haplotypes near the upper part and central core of the network also tend to be shared more often than those at the lower (terminal) ends of the network. This could be an indication of a more recent generation of these terminal haplotypes and there not having been sufficient time for them to spread to other populations or mutating to form links with other haplotypes.

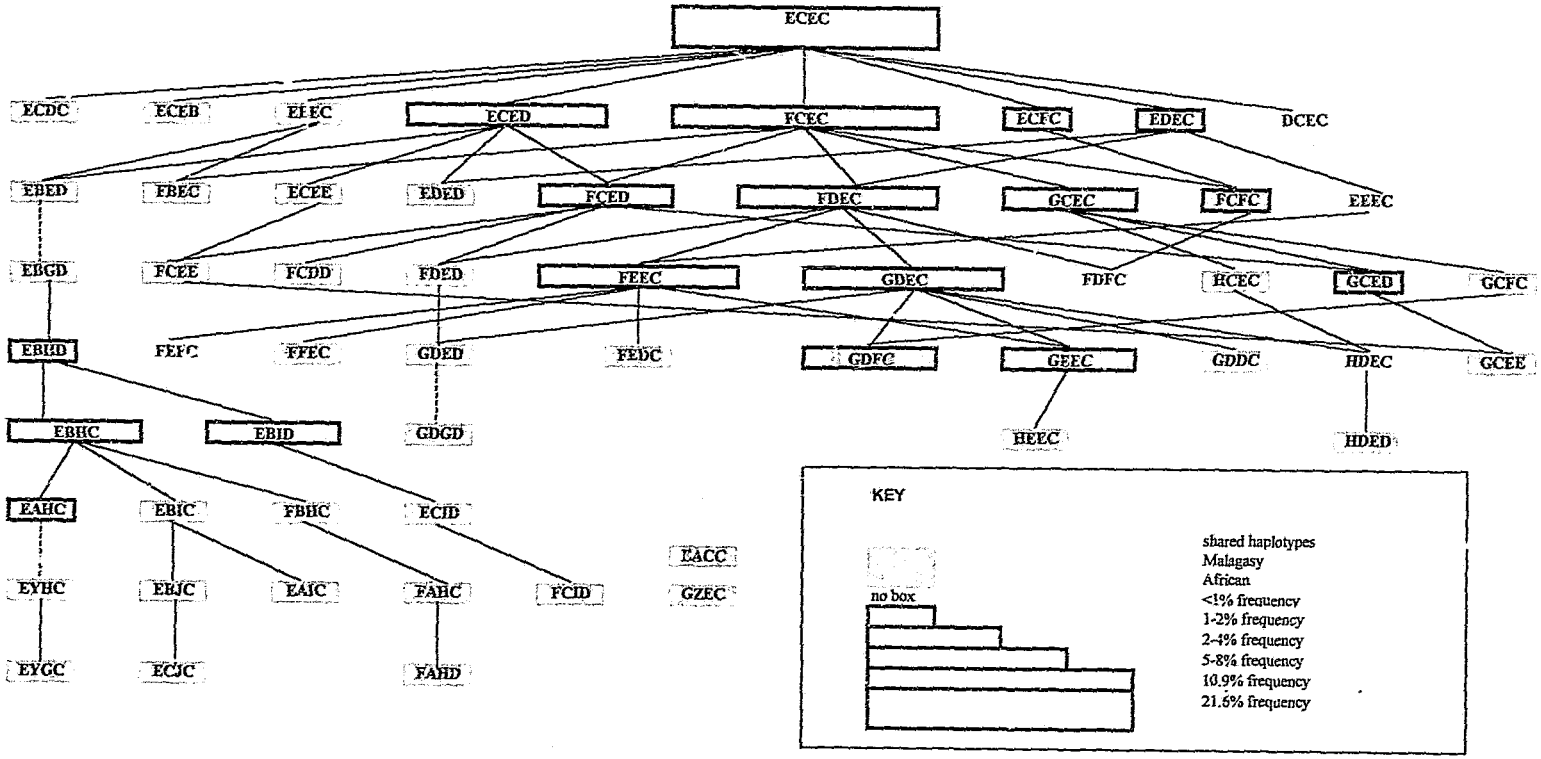


Figure 5.4. Frame 1 network constructed using YAP+ haplotypes present in the Malagasy and African populations. The key describes the usage of colours and different sized boxes. The dashed lines represent connecting haplotypes not detected in the present study.

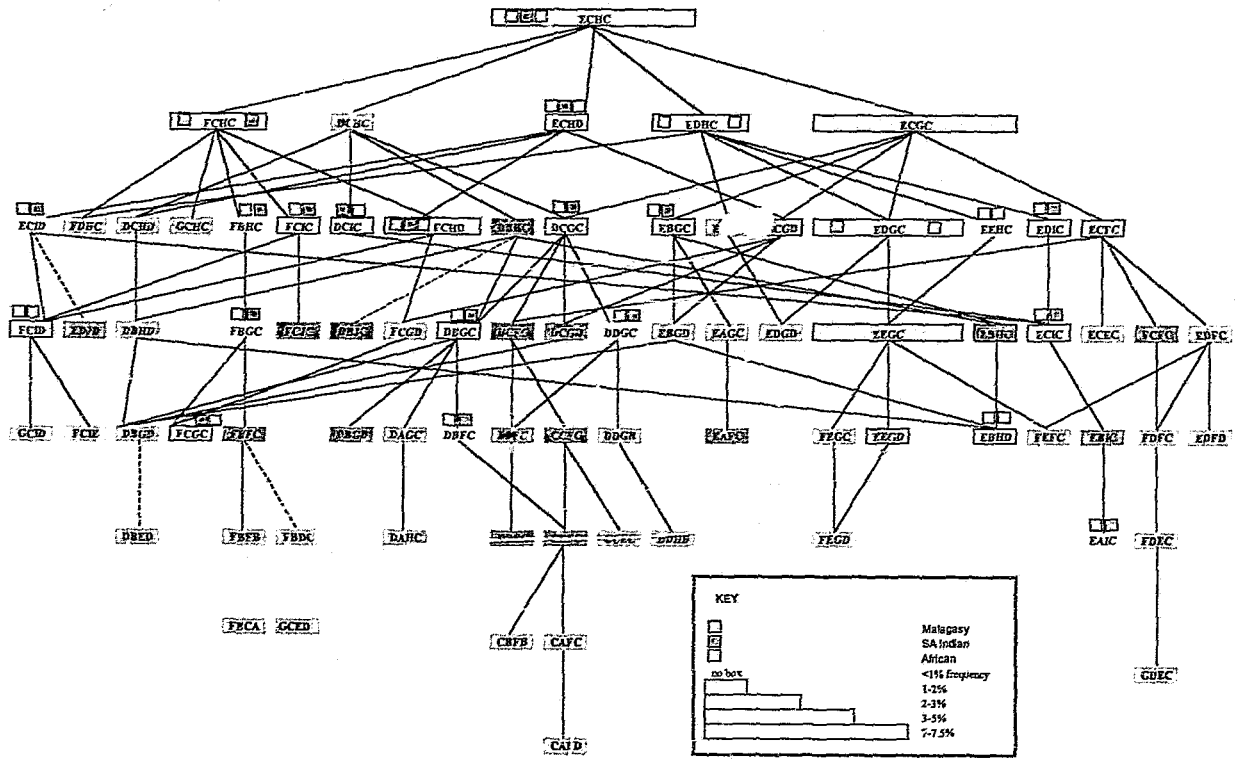


Figure 5.5. Frame 2 network constructed using YAP- haplotypes present in the Malagasy, SA Indian and African populations. The key describes the usage of colours and different sized boxes. The dashed lines represent connecting haplotypes not detected in the present study.

5.6.1 Y chromosome ancestral haplotype

As indicated by the haplotype frequency data discussed in **sections 5.5.1.1** and **5.5.1.2**, most of the haplotypes detected in the present sample have very low frequencies. The commonest haplotype on frame 1 is h17 (+ECEC). It accounts for 22.7% of the YAP+ haplotypes and is found in 15 of the 17 population samples in which the YAP+ was detected. It is also connected to the largest number of haplotypes in the frame 1 network showing the potential for generating all the adjacent haplotypes. Generally, recently arisen haplotypes are unlikely to have had the opportunity to drift to high frequencies or to be present in widespread populations. Thus, common or most frequent haplotypes are most likely to be the oldest under either an infinite allele model (Watterson and Guess, 1977) or a stepwise mutation model (Chakraborty, 1977). This strongly suggests that haplotype ECEC is the ancestral Y haplotype on which the *Alu* insertion occurred. Confirmation of this might be achieved by examining the relationships of YAP+ haplotypes in Japanese and Tibetan populations in which this insertion has also been detected. Unfortunately, these data are not available at this stage for the comparative analysis. Studies by Altheide and Hammer (1997) and Hammer *et al* (1998) suggest that the insertion event probably took place in Asia about 55 000 years ago with a migration back into Africa of these YAP+ chromosomes.

Haplotype 139 (-ECHC) on frame 2 was found at an overall frequency of 7.9%. This value is extremely close to the second highest value of 7.8% noted for h136 (-ECGC). Because h139 is the most widely occurring haplotype in the data set (shared by 20 of 27 populations across racial and geographic boundaries), it is probably the ancestral YAP- haplotype rather than h136 (shared by 13 of 27 populations). Haplotype –ECHC, like haplotype

+ECEC in frame 1 can also be connected by a single mutation step to many other common haplotypes near the top of the network. It is however absent from three Malagasy (Antaimoro, Antankarana and Mikea), the Australian, the New Guinean (coastal) and the S.Sangha population samples. The sample size of four of these populations was less than 35 and this may explain its absence. The apparent absence of –ECHC from the Antaimoro (N=50) and the Australian groups (N=107) is more difficult to explain but could be attributed to genetic drift.

5.6.2 YAP+ network

The network constructed using frame 1 haplotypes is presented in **Figure 5.4**. The *Alu* insertion was only present in African and Malagasy populations in this study and thus this network only represents these two groups. By inferring three intermediate haplotypes (EBFD, GDFD and EZHC represented by small black circles in **Figure 5.4**) not found in our sample, it was possible to connect 55 of the 57 haplotypes on the network. The two singleton haplotypes EACC and GZEC detected in African and Malagasy populations respectively, could not be related by fewer than two repeat changes to any other haplotype in the frame. They are probably in their respective populations due to recent gene flow from populations not examined.

Various lineages or paths can be followed from hypothesized ancestral haplotypes during the evolution of new haplotypes due to mutations at various loci. Haplotypes in different populations may cluster along different lineages giving clues to their demographic histories. Very informative clustering patterns of this sort have been identified in a study of the Lemba of southern Africa (unpublished data) who claim Jewish ancestry and also in a

study examining the affinities of Khoisan-speaking populations of southern Africa (unpublished data). The presence of such unique lineages or clustering of haplotypes may thus provide useful information for understanding the genetic history of a population.

Two clusters on the frame 1 network have been examined in an attempt to extract information regarding Malagasy demographic history. The first and bigger cluster begins with haplotype ECEC and includes all haplotypes connected to it while the second cluster can be seen on the lower left-hand side of the network and begins at haplotype EBGD. This second cluster would be connected to the first if a missing haplotype leading from EBGD to EBED could be found. Each of these clusters have a central core of haplotypes that are both common and shared by the Afri : and Malagasy groups (**Figure 5.4**). The intermediate haplotypes which would connect these two clusters are neither common or shared. These two clusters may thus represent two distinct African populations whose representatives were introduced to the island early on in its colonization or they may represent waves of colonization at different times by population with differing gene pools.

All except one of the 41 YAP+ Malagasy haplotypes (EACC) observed on the 279 YAP+ chromosomes examined fit onto the frame 1 network (**Figure 5.4**). This suggests that the founding male population/s of the Malagasy was relatively large and fairly diverse. All haplotypes found exclusively in the Malagasy (green boxes) can be derived from preceding haplotypes in the network found among African populations (blue and yellow boxes). Most of the haplotypes detected in individual Malagasy groups tend to cluster around the nine common haplotypes at the center of the network (individual networks for each of the Malagasy groups not shown). This reaffirms that Madagascar was settled relatively recently since there has not been sufficient time for regional variation to accumulate. A

number of haplotypes in individual Malagasy groups tend to be placed randomly throughout the network. This could be suggestive of a number of things; it may reflect more recent, erratic gene flow events that occurred in the different groups throughout the island; it may be an indication of population fissions which occurred early on in the colonization of the island; or it may indicate that the original male founders from Africa were not a homogeneous group, and could have arrived on the island in successive waves from different regions, resulting in the complex pattern of variation found throughout the island.

A number of haplotypes at the terminal ends of the network seem to be unique to the Malagasy. This indicates that either a population who has recently contributed to the gene pool of the Malagasy has not yet been sampled or these haplotypes have been generated on the island, i.e. they are 'Malagasy-specific'. Assuming that the latter is the case, a selected cluster consisting of haplotypes found only in the Malagasy has been used (see **section 2.3(m)** and **section 5.6.5**) to estimate the divergence time of these haplotypes on Madagascar.

Thirty-four of the 35 haplotypes observed in Africans fit into the frame 1 network (**Figure 5.4**). The haplotypes detected in the Tsonga, Zambian and Gbaya samples do not form separate clusters but are found in the central common core of the network which to a large extent supports the hypothesis that they too have recently diverged from a common ancestral population. None of the three can be identified as 'clustering' more with Malagasy haplotypes than the other. Thus, the network analyses of YAP+ haplotypes does not cast more light on the specific identity of the African founders of the Malagasy (if they were in fact a single group).

5.6.3 YAP- network

YAP- haplotypes from African, Malagasy and the SA Indian populations were used to construct this SMM YAP- network (**Figure 5.5**). Again, haplotypes that cannot be connected to the network by single mutational events have been linked by indicating possible intermediate connections. All but six haplotypes fit onto the network and the ones that do not fit are separated from the closest neighbour by two steps (shown by a dashed line). The significance of the colors, shading and different sizes of the boxes used in the network is explained in the **KEY** to **Figure 5.5**.

Almost all of the haplotypes found in the Malagasy and SA Indian populations could be linked via single steps on the various lineages on the network. Haplotypes found in the African sample on the other hand, do not follow any clustering pattern but are scattered throughout the network. Again, a number of 'lineages' or branches representing the evolutionary pathways followed by the hypothesized ancestral haplotype ECHC over time can be seen on the network. Certain haplotypes show population specificity and cluster along particular lineages e.g. ECHD-DCHD-DBHD-DBGD appears to be Malagasy-specific and DCFC-DDFC-DDEC SA Indian-specific. Since YAP- haplotypes are uncommon in Africans, their haplotypes are found throughout the network without any specific pattern.

Despite the high levels of diversity and the fact that many unique haplotypes from individual populations may not have been sampled, single steps connect the majority of haplotypes in the groups examined. This suggests that most of the haplotypes in these populations are not likely to be there because of recent gene flow, but may have been

present for many generations. Network analyses may thus be quite useful, not only for detecting gene flow, but also for inferring the demographic history of populations. Core haplotypes may also represent founder or ancestral haplotypes, because not only do they connect to other common haplotypes, thus having the potential to generate them, but they may generally occur in the majority of populations at frequencies higher than those noted for other haplotypes, as is the case for h140 (see **section 5.5.1.2**).

5.6.4 Homoplasy

Homoplasy with regard to haplotypes can be defined as the formation of identical haplotypes by different routes. Since the YAP insertion is hypothesized to have taken place only once in human history, haplotypes in frames 1 and 2 represent two lineages of Y chromosome evolution. All chromosomes with the insertion stem from a single YAP+ ancestor, and all chromosomes without the insertion stem from a single YAP- ancestor. The full range of published alleles at the four STR loci were found to be associated with both the YAP+ and YAP- alleles in the present study. The high mutation rates at STR loci allows conversion of alleles between pre-existing allelic states, and thus identical STR haplotypes in different populations may exist because of mutation rather than due to common descent.

In the present study, 27 four-site STR haplotypes were common to both the YAP+ and YAP- lineages and the ones detected in the Malagasy (24 haplotypes) are shown in bold typeface in **Tables 5.11 and 5.12**. The remaining three haplotypes (ECEB, FAHD and FCFC) detected only in African populations are not shown in the table. These 27 haplotypes account for 47.4% of unique YAP+ haplotypes and 12.2% of unique YAP-

haplotypes in the entire sample. All 27 haplotypes can be localized to the central cores in frame 1 and 2 discussed in detail in **section 5.6** and can be linked to each other via single mutational steps. This clearly demonstrates both the high mutation rate at STR loci and its occurrence in a single step-wise fashion. This mutational process is ongoing on all Y chromosome lineages including YAP- and YAP+ lineages and may explain the confounding phylogenetic relationships observed between populations from around the world when haplotypes based solely on STR data are used (de Knijff *et al* 1997) in comparison to phylogenies based on autosomal and mtDNA data.

5.6.5 Inferences of Y chromosome network antiquity

Another objective of this research was to evaluate Y chromosome data to determine if it could be used to date the colonization of Madagascar. A number of methods are now available to do this using STR data (Goldstein *et al*, 1996) and have been used in a number of studies (Kittles *et al*, 1998; Cooper *et al*, 1999). The protocol followed by Kittles *et al* (1998) has been applied to the present data set [see **section 3.3 (m)**].

A single unambiguous internal cluster in frame 1 (EBHC leading to EBIC/EAHC-EBJC/EAIC-ECJC; see **Figure 5.4**) was identified as possibly being 'Malagasy-specific' and has been used to estimate the time of divergence for this cluster of haplotypes. The theoretical relationship $S=2\mu T\sigma^2$ was applied to the data using the published average mutation rate of 0.0012 (Bianchi *et al*, 1998)]. This value has been put forward based on reports by Heyer *et al* (1997) and Kayser *et al* (1997) with 95% Poisson confidence interval limits of 0.00046- 0.0028. A mutational variance (σ^2) of 1 was used for the one-step mutation model (DiRiezo *et al*, 1994; Slatkin 1995) and a generation time of 25 years

was assumed. The observed variance S was calculated from the haplotype data as the mean number of steps between the root (EBHC in this case) and each of the haplotypes examined in the network (Kittles *et al*, 1998). Using an average mutation rate of 0.0012, the age of the Malagasy-specific cluster is placed at 2864 years (95% confidence intervals, 1227 – 7472 years). The uncertainty of the precise mutation rate for each locus and the fact that the mutation variance for most STR markers have not been formally established, together with other issues such as the size, level of variation and rate of growth of the founding population increases the 95% confidence intervals for the estimate of time depth. The result obtained is compatible with time depths estimated from archaeological data.

5.7 Principal components analysis (PCA)

A PCA was performed using the Y chromosome haplotype frequencies and a plot of the first two components is shown in **Figure 5.6**. These two components account for about 30% of the variance in the data. The X-axis depicts relationships between the populations which were also evident on a NJ tree based on Reynold's genetic distances. The Asian/Oceanic population cluster (red triangles) is clearly separated from the cluster made up of African and Malagasy population groups. This overlap supports a strong affinity of the Malagasy with African populations.

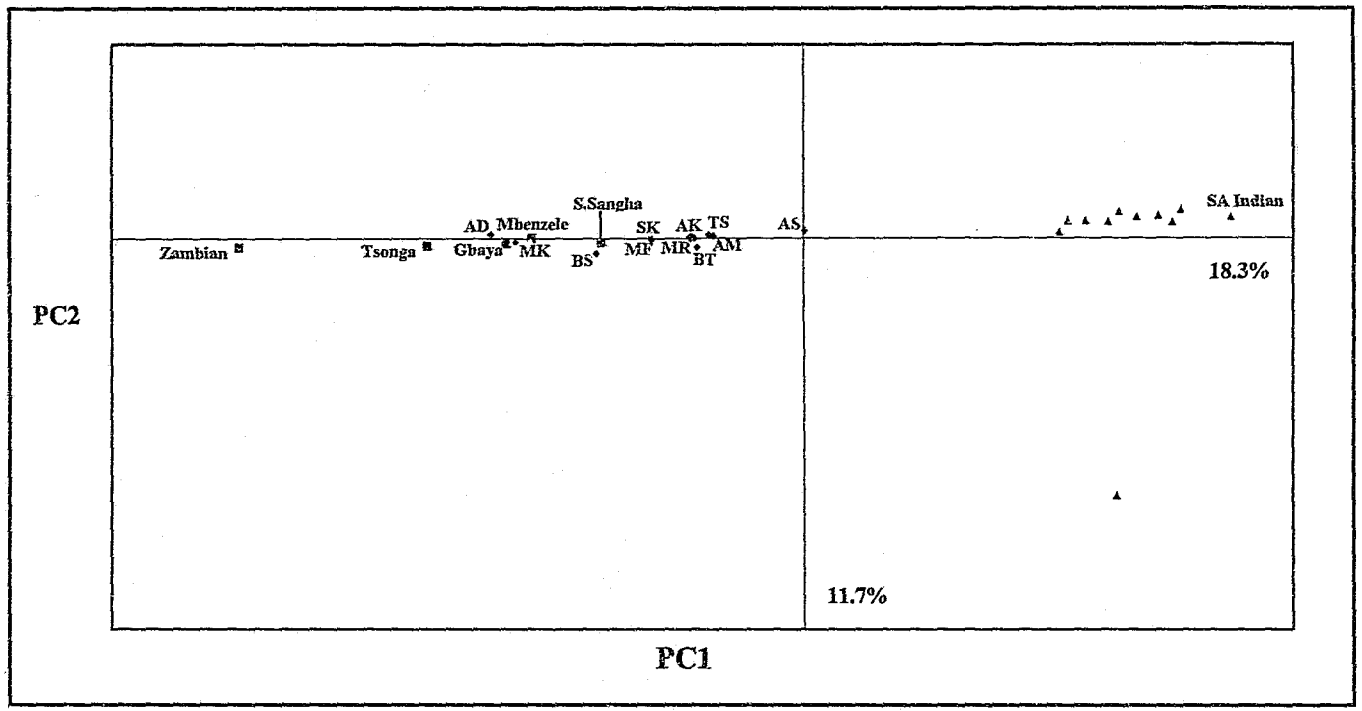


Figure 5.6. Two dimensional principal components plot showing population affinities based on Y chromosome haplotype data. Blue diamonds represent Malagasy groups, red triangles represent non-Africans and green squares represent Africans. The unlabelled red triangles represent SE Asian/Oceanic populations screened by Alan Redd.

5.8 Analysis of Molecular Variance (AMOVA)

AMOVA (Analysis of MOlecular VAriance) was used to examine the hierarchical levels of population structure in the data set. Analysis results showed that when the 11 Malagasy populations were treated as a single group, 97.97% ($p<0.0001$) of the genetic variance was found within the populations whereas only 2.03% ($p<0.0001$) of the genetic variance was attributable to differences among the populations. When all 27 population samples studied were treated as a single group, 84.02% of the genetic variance was found within populations and a more substantial 15.79% of the genetic variance was attributable to differences among populations. The contrast between the results obtained for the Malagasy and the 27 population samples provides some perspective of the similarity of the Malagasy groups to each other. This low degree of genetic differentiation among the Malagasy groups tested supports a recent common ancestry for them. Recent gene flow between groups would however also result in a low percentage.

Studies using classical genetic markers have revealed that most human genetic variation is between individuals within populations rather than between populations (Lewontin, 1992; Nei and Roychoudary, 1972). Autosomal DNA markers have confirmed this, giving nearly the same proportions of genetic substructure as classical markers: 85% among individuals within populations and 15% among populations (Barbujani et al, 1997). When the populations examined in this study were grouped according to their geographic location namely Africa, Madagascar and SE Asia/Oceania, 79.28% of the genetic variance was still found within populations; 15.79% was due to differences among the groups and 4.94% was due to differences among populations within the groups. This supports the findings of Ruiz-Linares et al (1996) and Underhill et al (1997) that the haplotypes of males from the

same geographic region are more similar than the haplotypes of males from different geographic regions.

5.9 Summary

The variation at five polymorphic loci on the nonrecombining portion of the Y chromosome was studied to gain insight on the ancestry of the Malagasy. The YAP locus has two alleles (YAP- and YAP+) while the four STR loci (DYS393, DYS19, DYS390 and DYS391) have 7, 8, 12 and 6 alleles, respectively. Data on the five loci were generated from more than 700 males belonging to 17 populations (467 Malagasy, 166 African and 50 SA Indians). These were used with data from another source (10 SE Asian/Oceanic populations) for comparative purposes.

The YAP+ allele was used to track African male lineages and its presence at high frequencies (greater than 0.500) in the Malagasy ethnic groups confirms that there was a large African male input into the Malagasy gene pool. Three of the four multi-allelic STR loci (DYS393, DYS19 and DYS391) show predominantly unimodal allele frequency distributions with one frequent allele and less frequent neighboring alleles (**Tables 5.3**, **5.4** and **5.6**). Each allele at all four loci differs from its neighbor by one repeat unit and no internal unsampled allele length classes were detected. This distribution is compatible with the SMM whereby new alleles are derived from existing ones by becoming longer or shorter by one repeat unit at a time. Locus DYS390 (**Table 5.5**) on the other hand clearly shows a bimodal or complex distribution in most of the Malagasy and African populations. Bimodal or complex distributions suggest a recent bottlenecks or admixture or the operation of two or more mutation mechanisms. Admixture may explain the bimodal

distribution in the Malagasy but the reason for the bimodal distribution in the African populations is less clear. The data from the other three STR loci examined do not support a bottleneck effect for the African populations. The presence in the Malagasy of particular STR alleles, especially at locus DYS390, is also indicative of either African or SE Asian contributions to them.

Haplotypes constructed in the order: YAP status followed by alleles at locus DYS393, DYS19, DYS390 and DYS391 revealed 221 haplotypes on two frames defined by YAP status; frame 1 defined by YAP+ chromosomes and frame 2 defined by YAP- chromosomes. Fifty-seven of the unique haplotypes were placed on frame 1 and 164 on frame 2. High levels of haplotype diversity (>85% for almost all the populations examined) were noted in all the population samples examined. More than 90% of haplotypes could be placed on SMM networks based on parsimony analysis. Remaining haplotypes could be added by invoking a small number of intermediate states which have either not been sampled, or were bridged by mutational events of more than one step. The clustering of haplotypes from individual populations along different lineages in the network indicates that this type of analysis may prove to be very useful for tracing demographic events in the history of population groups.

Ancestral haplotypes were hypothesized based on them being shared by multiple populations across racial and geographic boundaries; these are haplotype ECEC on a YAP+ background and haplotype ECHC on a YAP- background. The clustering of common shared haplotypes around the hypothesized ancestral haplotype indicates a recent common ancestry and/or extensive male migration during human evolutionary history. The sizes of the population samples used for constructing the networks range from 13 to 51

Although a very small sample size may lead to incorrect conclusions, this range appears to have given results that are compatible with historic events. Network analyses are difficult to carry out and time consuming, but they give an indication of what could be 'old' and 'new' haplotypes based on their position on the network (central or terminal) in relation to the ancestral haplotype. The age of a haplotype can also roughly be determined by examining the number of differences between the haplotypes. The same cannot be done just by examining frequency data.

Y chromosome haplotype sharing data suggests that the different Malagasy groups have a very recent common ancestor. The eleven groups share a large number of haplotypes and cluster very closely to each other on phylogenetic trees generated using at least three different genetic distance measures. Principal component plots derived from these data show the same thing. Malagasy groups also share many haplotypes with the two African Bantu-speaking groups examined, suggesting either a large genetic contribution from them or from groups with a similar genetic make-up. Studies on globin haplotypes (Hewitt et al, 1996) and G6PD polymorphisms (Dangerfield, 1997) also indicate that Bantu-speakers made a large contribution to the Malagasy gene pool. No striking differences in terms of haplotype distribution or clustering on the networks were observed between highland and lowland Malagasy groups. This re-inforces the notion that the ethnic classification on Madagascar is more of a political classification than a biological one. Both the haplotype sharing data and results from AMOVA show a low degree of genetic differentiation among the Malagasy groups tested and this lends further support for their recent origin from a common founding population. The present data do not point to Borneo specifically as the geographic region from which one of the Malagasy ancestral populations emanated as is suggested by linguistic data. The present data supports a large African ancestry (>50%

based on just the YAP data) for the present day Malagasy, not supporting a greater Indonesian genetic influence for highland populations compared to lowlanders.

A large number of former slaves on Madagascar have been incorporated into the different ethnic groups of the island and the high YAP+ frequency may be attributed to recent events rather than an indication of the gene pools of the founding populations. This would however mean that the proportion of slave incorporation was almost uniform and very high throughout the island as indicated by the YAP+ frequency data. Hypothesis 4 given for the origins of the Malagasy as outlined in **section 1.3.1.** which suggests that Madagascar was colonized by genetically mixed populations who had contributions from both Africa and Asia is supported by the Y chromosome data from the present study. This would easily explain the homogeneous picture presented by the eleven Malagasy ethnic groups. The small differences shown by them could be attributed to recent gene flow.

The age of coalescence for a group of related Malagasy-specific microsatellite haplotypes was estimate to be 2864 years (95% confidence intervals, 1227 – 7472 years). This result is compatible with time depths estimated from archaeological data. It is difficult to determine the level of variation that existed within the founding Malagasy population/s at the time/s of settlement, but additional Y chromosome markers generating extended haplotypes may be useful in refining the identification of both African and SE Asian parental populations of the Malagasy.

CHAPTER 6 – DISCUSSION

The recent explosion in the number of known polymorphic loci in the coding and non-coding regions of the human genome has vastly changed the methodology used to study population genetics. Restriction fragment length polymorphisms (RFLPs), short tandem repeat (STR) polymorphisms and *Alu* insertion/deletion (ins/del) polymorphisms have already been used widely and successfully to examine the relationships of human populations. The high variability of STRs and their differential distribution worldwide make them particularly useful markers for reconstructing recent evolutionary events because of the increased possibility of detecting allele frequency differences between populations due to the effects of genetic drift and gene admixture. Their high mutation rates however sometimes confound population relationships. RFLPs and ins/del polymorphisms on the other hand, despite being shared by most populations worldwide and often at similar frequencies, have a much lower mutation rate than STRs, and provide a way of determining identity by descent for the chromosomes being examined. Haplotypes incorporating both types of markers (STRs and RFLPs or ins/del polymorphisms) that differ in their mutation rates increases the usefulness of genetic markers in tracing recent historic events.

This study exploits the properties of some genetic markers to shed light on the prehistory of the Malagasy. Several studies based on historical, anthropological, archaeological and linguistic data have addressed questions pertaining to the origins of the Malagasy, but there are still many questions that remain unanswered. Having reviewed the evidence, Campbell (1995) has justifiably concluded that the colonization of Madagascar has always posed a

problem, there being no academic consensus as to which peoples colonized the island, in what order, when, from where, why, or how.

Of the four hypotheses formulated regarding Malagasy ancestry, only two remain widely debated. The two that have already been discarded are that (1) Madagascar possessed an autochthonous population, and (2) the original settlement of Madagascar was by peoples of African origin. The two theories that are still contended are firstly, was Madagascar colonized by an admixed African/Indonesian population and secondly, was Madagascar colonized independently by Africans and Indonesians. Adding to the complexity of the debate is the postulated route of migration of the proto-Malagasy (people that first colonized Madagascar) to Madagascar: was it a direct route from Indonesian to the east coast of Madagascar? Or was it an indirect route around the northern and western shores of the Indian Ocean via East Africa?

Historical evidence indicates that Indonesians were trading on the east coast of Africa between the first and fifth centuries AD (Bellwood 1985, 1987). They left traces of their contact with Africa in the form of the dry rice cultivation technology and the introduction of chickens, pigs, bananas, coconut, sugar cane and taro (Deschamps, 1965). This cultural contribution could, however, be a result of transient events during episodes of trading with Africa and does not provide evidence that Indonesians settled in Africa or that there was any genetic exchange between Indonesians and Africans. Cultural traits can, of course, also be transmitted indirectly by 'genetically unrelated' traders. If on the other hand, Indonesians did settle in Africa and there was gene flow between these two groups, there should be traces of Indonesian genes in some African populations. The more likely unions would be between Indonesian males and African females, so Y chromosome markers ought

to be useful in assessing the Indonesian contribution to the African gene pool. If gene admixture between Indonesians and Africans could be inferred from genetic studies, then it could suggest that the founding population of Madagascar was already admixed at the time of colonization, resulting in a close genetic affinity between all the ethnic groups of Madagascar. However, if Madagascar was colonized independently by Africans and Indonesians, and admixture occurred on the island, then population fissions and migration to different regions of the island could also explain the close genetic affinity among the ethnic groups.

A major limitation of using genetic markers to resolve the issue of population admixture is that most markers are only capable of resolving historical events that may have occurred between 50-200 000 years ago, making it difficult to discriminate subtle differences which could have occurred in the Malagasy in the 2000 or so years since colonization. The inclusion of faster evolving STR loci in this study has overcome this problem to some extent and has assisted in detecting more recent population movements to the island.

A major goal of this research was to compare the genetic data from eleven of the eighteen ethnic groups present on Madagascar with non-African (Indonesians, Melanesians, Polynesians and Indians) and African (Zambians, Tsonga, Gbaya, Mbimou, S.Sangha and Mbenzele) populations in an attempt to trace the ancestry of the founders of the Malagasy. Most of these 'putative' parental populations were preferentially selected because they were sampled from geographic areas in Africa and Asia most likely to have contributed to the gene pool of the present day Malagasy. Linguistic and historic data favour the hypothesis that the proto-Malagasy originated in SE Asia/Indonesia and Africans arrived later. The role of Polynesians and Indians in the peopling of Madagascar is still debatable.

Historical data suggest that Mozambique and Malawi could be the geographic region of origin of the putative ancestral African populations of Madagascar (Campbell, 1989). Earlier genetic studies on blood groups and a few other sero-genetic markers have shown support for both Asian and African contributions to the Malagasy but these studies have been limited to small sample sizes. More recently, genetic data from systems like the sickle cell trait (Hewitt, 1994) and haplotypes around the glucose-6-phosphate dehydrogenase locus (Dangerfield, 1997) have provided strong support for contributions from Bantu-speakers from central and east Africa to the Malagasy gene pool.

A number of polymorphic loci both on the autosomes and the Y chromosome have been used in this study to examine the genetic structure of present-day Malagasy in order to reconstruct their history. The four autosomal STR loci examined [HUMCSF1PO, HUMTHO1, HUMCD4, and an $(AC)_n$ repeat] have previously been shown to differ in their allele frequency distributions between African and non-African populations (Hammond *et al*, 1994 and Kidd *et al*, 1998). Using frequency differentials it was estimated that Africans contributed 32.4% to the gene pool of the highland groups and 58.8% to the lowland groups. The HUMCD4 locus proved to be the most informative in this regard (**CHAPTER 3, section 3.3**). Studies by Tishkoff *et al* (1996) showed that only three alleles (5, 6 and 10) of the twelve alleles reported at this locus occurred at appreciable frequencies outside of Africa. The frequencies of these three alleles in the Malagasy were greater than those seen in all the African populations examined in this study (**Table 3.3**). The remaining nine alleles detected in the Malagasy were also found in African populations strongly suggesting an African origin for them.

To increase the robustness of the analysis, haplotypes were constructed using alleles at the HUMCD4 STR locus and an *Alu* deletion polymorphism located 9.8kb downstream from it. The deletion has been shown to be absent from Asian populations (Tishkoff *et al*, 1996) but was found in the Malagasy at frequencies of 0.025 to 0.280, comparable to frequencies seen in Africa (**Table 4.1**). Haplotypic data using these two markers provided strong support for a substantial African ancestry for the Malagasy (40.7% for the highland groups and 60.53% for the lowland groups; **CHAPTER 4, section 4.4, Tables 4.3** and **4.6**).

Kidd *et al* (1998) examined the distribution of haplotypes constructed using alleles at four polymorphic loci [three RFLP loci and an $(AC)_n$] around the DRD2 gene in worldwide populations and identified a number of 'population-specific' haplotypes. Again, this haplotype system clearly demonstrates that the Malagasy possess haplotypes that are derived from both Asian and African populations. This is further supported by admixture estimates, which show that Africans have contributed 47.7% to the gene pool of the highland groups and 73.7% to the gene pool of the lowland groups (**CHAPTER 4, section 4.5, Table 4.5** and **4.6**).

Y chromosome data from four STR loci (DYS393, DYS19, DYS390 and DYS391) also proved to be informative in distinguishing contributions made by Africans and SE Asians to the Malagasy (**CHAPTER 5, section 5.3**). In addition, the '+' allele (insertion of an *Alu* element) at the YAP locus, which has been shown to be absent from SE Asian and Oceanic populations (Hammer and Horai, 1995) but found at frequencies of between 0.70 to 1.00 in Negroid populations, was particularly useful in tracing African derived Y chromosomes in the Malagasy. The YAP+ allele was detected on more than 50% of Malagasy Y chromosomes examined, clearly indicating that at least 50% of the Y chromosomes in

Madagascar are derived from an African source (**CHAPTER 5, section 5.3.1**). In a separate study, Soodyall and colleagues (personal communications, 1999) have used M9, a polymorphism which can be used to distinguish between African and non-African YAP- chromosomes, to identify Malagasy YAP- chromosomes that are not of African origin. The M9 polymorphism describes a C to G transversion in non-Africans (Underhill *et al*, 1997). Of 156 YAP- chromosomes examined, 57% had the G variant indicating non-African ancestry. The origin of the remaining 43% of YAP- chromosomes which had the C variant remain unresolved as this variant occurs worldwide.

In summary, both the autosomal and Y chromosome data from the present study confirm that the extant Malagasy show both an African as well as a SE Asian ancestry. Population trees and principal components analysis based on autosomal and Y chromosome allele and haplotype frequency data show the different Malagasy groups as a cluster intermediate between African and Asian clusters. The similarity of the gene pools of the eleven ethnic groups examined from different regions of the island may be due to their recent common ancestry. Unfortunately, the data from the present study cannot resolve whether the proto-Malagasy were an admixed African/Indonesian population prior to colonization or whether admixture between these groups occurred following colonization.

If contact between Indonesian traders and Bantu-speakers in east Africa resulted in gene admixture between the two groups, the proto-Malagasy would have assimilated African genes into their gene pool. The fact that the present-day Malagasy way of life and language still reflects a dominance of Indonesian culture indicates that these two traits were not affected substantially by the time spent by the Indonesians in Africa. The coherence and strength of a common language and culture throughout the island also supports the idea

that new immigrants to the island from different cultures (Jews, Persians, Indians and even Chinese) had minor effects on the lifestyle of the settlers. Only minor isolated cultural and vocabulary links are seen except for techniques relating to cattle herding which were acquired from Africa (Brown, 1978; Mack, 1986).

It is also possible that admixture took place on the island following colonization. If it is assumed that Madagascar was initially colonized by Indonesians, the strong African ancestry demonstrated by the genetic markers used in the present study could be a result of large numbers of African slaves brought to Madagascar during the nineteenth century as well as well as voluntary migration to the island from Mozambique. Madagascar featured strongly in the exportation of slaves and it is estimated that approximately 72 000 slaves were exported to the Mascarenes between 1610 and 1810 (Campbell, 1988). By the early nineteenth century, slaves were being imported to the west coast of Madagascar from Africa by Arab traders as servants and agricultural labourers (Campbell, 1989). Many slaves were sold to the Merina (2000 to 3000 slaves a year by 1824) who were at this stage a ruling dynasty on the island and deeply involved in commercial rice agriculture. The slave population of Antananarivo quadrupled between 1828 and 1861, and an annual influx of possibly 10 000 East African slaves reached the Merina until the mid-1880s (Campbell, 1988). A large number, approximately 5000 slaves per annum during the mid-nineteenth century (Campbell, 1988) were also re-exported to the French plantations. These figures are known to have fluctuated tremendously during the nineteenth century but estimates for Madagascar indicate overall figures of 1 million imported slaves and approximately 300 000 enslaved Malagasy in the course of the nineteenth century.

There is little doubt that the slave trade influenced the gene pool of the Malagasy populations throughout the island because after the slaves were freed at the end of the nineteenth century, they were assimilated into the general Malagasy settlements where they adopted the Malagasy culture. There is evidence, however, to indicate that the slave communities of Madagascar experienced low birth and high mortality rates (Campbell, 1991). These fluctuations were caused by forced harsh physical labour, food shortages and outbreaks of diseases such as malaria, syphilis, smallpox, tuberculosis, cholera and typhoid. Traditional warfare was thought to be a minor contributor to the death rates but these would affect only males whereas the previously mentioned factors would generally affect both sexes. All these factors have caused the gene pool of the Malagasy to vary tremendously during the last century and will make it extremely difficult to establish the size or composition of the population that initially colonized Madagascar

From the close relationships of the Malagasy groups shown in the present study, it would appear that a large and widespread assimilation of African slaves occurred throughout the island. Autosomal data from the present study however do not support an equal African-Indonesian contribution into the various Malagasy ethnic groups. Admixture calculations using Elston's least squares method at the autosomal STR loci [HUMCSF1PO, HUMTHO1, HUMCD4, and an $(AC)_n$ repeat] and the CD4 and DRD2 haplotype loci suggest that there has been a larger African contribution to the nine lowland populations examined than to the two highland populations (Merina and Betsileo), and a larger Indonesian contribution to the highland populations than to the lowland populations (discussed above), a finding which supports other types of data discussed below.

Observations made by historians as well as a study based on anthropometric measures by Singer *et al* (1957) suggest that there were several differences in physical features which distinguished populations living in the highland from those living in the lowland regions of the island. Using physical characteristics like hair form, facial features, and skin colour on a small number of Malagasy individuals, Singer *et al* (1957) concluded that it was difficult to associate phenotypic characteristics with ethnic classification. According to them, the Merina (a highland group) appear to be mainly of Malayo-Indonesian origin but there is no doubt that they have a high proportion of African admixture. They found that though all Malagasy share an essentially Indonesia language and a Indonesian/African culture, the highland groups possess a physiognomy that is clearly more Malay than lowlanders whose physiognomy is more Negroid. Singer *et al* (1957) concluded that "Madagascar provides ample evidence for the belief that there is no intrinsic or necessary relationship between language, culture and physical type. The Malgache language is singularly uniform throughout the island; the cultures, though less uniform, have many basic elements in common; and, the physical types form a sliding scale of varying mixtures of Caucasoid, Mongoloid and Negroid affinities".

Additional evidence for the genetic differences among the highland and lowland populations were drawn from the studies conducted by Buettner-Janusch and Buettner-Janusch (1964) using haptoglobin data and Hewitt (1998) using molecular data at the globin loci. These studies also found a higher proportion of African-derived genes in lowland populations compared with highland populations. Autosomal genetic data therefore appears to be consistent with the phenotypic differences noted by Singer *et al* (1957) between highland and lowland populations. Y chromosome data, however, show that the African male contribution to the two highland groups is equal to and sometimes

even higher than that seen in lowland Malagasy groups (**CHAPTER 5, section 5.3.1**). Genes undoubtedly play a role in morphological/physical appearance and that is why we tend to look like our parents. How then can the strong phenotypic differences noted between the two regions be explained in the context of the Y data?

A number of factors could contribute to the differences noted. Historical data indicates that the highlands were occupied by the Vazimba from around the thirteenth century AD prior to the migration of 'eastern' light-skinned people, currently called the Merina, to the highlands during the sixteenth century (Brown, 1978; Dewar and Wright, 1993). The establishment of the Merina kingdom resulted in the division of the Merina people into three major groups with decreasing hierarchical status (Singer *et al*, 1957): the Andriana or nobles, the Hova or free men (commoners) and the Mainty (the slaves). Under these social circumstances, unions between males from the upper classes with women from the lower class may have been more prevalent than the reverse. Y chromosome data in conjunction with mtDNA data may shed more light on these practices. A large proportion of the slaves liberated in 1896 lived amongst the Merina in the capital city of Madagascar where they formed about 25% of the population (Campbell, 1991). One possibility that may explain the Y chromosome data from this study may be that the Merina sample used in this study was collected from a region densely populated by the Mainty. Differences between highland and lowland groups could also have been maintained through assortative mating and endogamy.

Culture and language evolve much faster than genes because the former are affected by man's adaptation to his physical environment. Physical characteristics like skin colour, height and hair are multi-allelic characteristics and their genetics are poorly understood.

The effects of different combinations of genes controlling these traits would result in an apparently slower effect in an admixed population. An example of this has been described by Brown (1978): shipwrecked Portuguese sailors (sixteenth century) found one hundred years later by the French on Madagascar were indistinguishable from other inhabitants except for certain physical traits. Small groups of 'newcomers' may thus be quickly absorbed into local populations in terms of language and culture, but their genes will retain their origins for much longer as is shown here by the genetic control of physical characteristics and neutral markers like the YAP.

Not much has been written about the number of Indonesian women brought to the island. It was common for South Sea islanders to be accompanied by women on their maritime expeditions (Campbell, 1995). In support of this, a mtDNA marker common in Asia and which has been referred to as the 'Polynesian motif', since it occurs at high frequencies in both ancient and contemporary Polynesians, has been detected in the Malagasy at a frequency of approximately 9% (Soodyall *et al*, 1995). Further studies using mtDNA, together with autosomal and Y chromosomal markers may give a clearer picture of the relative contributions of males and females to the Malagasy and could conceivably localize the origins of male and female parental founders. Whether or not the differences noted between highland and lowland populations could be related to contributions made by differing numbers of males and females from different population groups is not clear at this stage. Polygyny has been observed in the Malagasy (Murdoch, 1959) and details on the extent of such cultural practices will prove to be useful in the interpretation of data from different parts of the genome.

A problem that may be biasing the estimates of contributions by various populations to the Malagasy is that people can and do chose their own ethnic identity especially when they are not living near their place of birth. A large number of the slaves on Madagascar imported during the nineteenth century have been incorporated into various ethnic groups of the island. Because of the stigma attached to the slave caste, slave clans regard themselves as members of the same ethnic groups as their neighbours and do not refer to their families' actual origins [Dewar, personal communications (1992)]. The high frequency of the YAP+ chromosome in the Malagasy could thus easily be attributed to these more recent events like the assimilation of slaves rather than an indication of the gene pools of the earlier founding populations of Madagascar. This would, however, mean that the proportion of slaves incorporated was almost uniform and very high throughout the island as indicated by the high YAP+ frequency in all eleven ethnic groups examined.

It is difficult to determine the level of variation that existed within the founding Malagasy population/s at the time/s of settlement. The types of markers that were used in the present study as well as the fact that a very shallow time depth in evolution is being studied makes it very difficult to distinguish between the independent genetic contributions made to the Malagasy at each wave of migration, either from Africa or SE Asia.There are a also a large number of contributors to the modern Malagasy population and distinguishing between them has been and is going to be difficult. There is no doubt from the present study however, that the Malagasy groups share many autosomal and Y chromosome haplotypes with the two African Bantu-speaking groups examined (Tsonga and Zambian), suggesting either a large genetic contribution from them or from groups with a similar genetic background. Both Bantu-speaking groups show a closer affinity with the Malagasy than do the populations from the Central African Republic. A close affinity of the Malagasy has

also been confirmed to SE Asian populations, but not specifically to an Indonesian population. Although the data from the autosomal markers suggest a link between the Malagasy and the Indian population examined, Y chromosome data indicate that this affinity may be an artifact due to the shared ancestry of Indian and Indonesian populations (Cavalli-Sforza, 1994). It would be incorrect to look for one or two parental populations for the Malagasy, and the inclusion of additional markers in the typing would further resolve the haplotypes detected, making it possible to better distinguish between African and SE Asian contributions to the extant Malagasy gene pool.

Contrary to historical biases, this study shows quite convincingly that there has been considerably more admixture between the African and SE Asian people contributing to the gene pool of the Malagasy. Given the recent colonization of the island and the limited power of resolution of the markers used to study this time period, it is difficult to discriminate between the two contentious hypotheses. Some of the differences detected between Y chromosome data and autosomal data may also be attributed to their mode of inheritance and the influence of their effective population sizes rather than to population differences. As markers from different parts of the human genome have their own biases and shortfalls, ideally, a combined analysis of autosomal, mtDNA and Y chromosomal markers on the same set of individuals may be expected to give a clearer picture of Malagasy ancestry/prehistory. Together with ongoing historical and archaeological studies, more extensive genetic studies may further clarify the prehistory of the Malagasy, which is now beginning to unfold.

APPENDIX A

All solutions were made up using distilled deionised water and stored at room temperature unless otherwise stated.

STOCK SOLUTIONS

40% acrylamide [39:1 (w/v) acrylamide:bisacrylamide]
Stir the following in 30ml water for 30min at room temperature in the dark:

- 19.5g acrylamide
- 0.5g bisacylamide
- 2.55g amberlite

Filter through Whatman no.1 filter paper and make up to 50ml. Store in the dark at 4°C for up to two weeks.

10% ammonium persulphate (APS)
Dissolve 1g of APS in 10ml water. Store at 4°C in the dark.

0.5M ethylenediamine tetra-acetic acid (EDTA)
Dissolve 9.3g Na_2EDTA in 40ml water while adjusting pH to 8.0 with 10N NaOH. When totally dissolved, make up volume to 50ml and autoclave.

Ethidium bromide (10mg/ml)
Dissolve 100mg ethidium bromide in 10ml water by shaking overnight. Filter and store at 4°C in the dark.

Formamide
Stir together 0.5g amberlite MB-6 resin and 100ml formamide at RT for 30min. Filter twice through Whatman no 1 filter paper. Store at -20°C.

1M $MgCl_2$
Dissolve 20.34g $MgCl_2$ in 90ml water. Make up to 100ml, dispense into 25ml aliquots and autoclave.

Proteinase K (10mg/ml)
Dissolve 100mg Proteinase K in 10ml water. Dispense into 1ml aliquots and store at -20°C.

10% Sodium dodecyl sulphate (SDS)
Dissolve 2.5g SDS in 20ml water by stirring at low heat. Make up to 25ml and dispense into 5ml aliquots.

0.4M Spermidine
Dissolve 1g spermidine in 10ml water. Dispense into aliquots and store at -20°C. Dilute 1:40 and store at -20°C.

10Xs TBE buffer
Dissolve 10g hydroxymethylaminomethane (Tris base) and 55g Boric acid in 900ml water. Add 40ml 0.5M EDTA (pH 8.0). Adjust pH to 8.0 with HCl and make up to 1litre.

1M Tris-HCl
Dissolve 121g hydroxymethylaminomethane (Tris base) in 900ml water. Adjust pH to 8.0 with HCl. Make up to one litre and autoclave.

WORKING SOLUTIONS

4% denaturing polyacrylamide (PA) gel
For 30ml gel solution, mix together:
- 3ml 40% PA [39:1 (w/v) acrylamide:bisacrylamide]
- 3ml 10XTBE
- 14.4g urea
- 13ml water

Add 30µl TEMED (Biorad) and 90µl 10% APS to this just before pouring the gel.

4.3% denaturing polyacrylamide (PA) gel (for use in the ABI 377)
For a 100ml solution, mix together:
- 10.6ml de-ionised 40% PA [19:1 (w/v) acrylamide:bisacrylamide]
- 10ml 10XTBE
- 36g urea
- ≈50ml double distilled H_2O

Stir until the urea has dissolved. Filter to remove any solid particles that may fluoresce. Store at 4°C in the dark. To 10ml of the above (sufficient for a single 12cm gel), add 50µl 10% APS and 7µl TEMED just before pouring.

6% denaturing polyacrylamide (PA) gel
For 30ml gel solution, mix together:
- 4.5ml 40% PA [19:1 (w/v) acrylamide:bisacrylamide]
- 3ml 10XTBE
- 14.4g urea
- 11ml water

Add 30µl TEMED (Biorad) and 90µl 10% APS to this just before pouring the gel.

DNA molecular weight marker (1kb ladder)
Mix together:
- 10.9µl 1kb DNA ladder (Gibco BRL; 250µg/242.7µl)
- 5.0µl Ficoll dye, and
- 84µl 1XTE

Store at 4°C. Use 8µl/well in agarose gels.

Dextran-formamide dye
Dissolve 100mg of Blue Dextran in 5ml formamide. Aliquot and store at 4°C.

Ficoll dye
(50% sucrose, 50mM EDTA, 0.1% bromophenol blue, 10% Ficoll).
Dissolve:
- 50g sucrose
- 0.1g bromophenol blue
- 10g Ficoll

in 90ml water. Add 10ml 0.5M EDTA and allow to dissolve overnight. Filter through Whatman 1MM filter paper and make up to 100ml.

Formamide dye
(95% deionised formamide, 20mM EDTA, 0.05% bromophenol blue, 0.05% xylene cyanol)
Mix together:
- 10ml 95% formamide
- 400µl 0.5M EDTA pH 8.0 (20mM)
- 5mg bromophenol blue (0.05%)
- 5mg xylene cyanol FF (0.05%)

Dispense into aliquots and store at -20°C.

Proteinase K working solution
Make up sufficient just before use (0,25ml/tube).
Mix together:
- 1.4ml water
- 0.4ml Proteinase K stock (10mg/ml)
- 0.2ml 10% SDS
- 8µl 0.5M EDTA (pH8.0)

Saturated NaCl
40g NaCl was slowly added to 100ml water until the soln was totally saturated, i.e. some salt remain undissolved.

Sucrose/Triton X-100 lysing buffer
Mix together:
- 10ml 1M Tris-HCl (pH 8.0)
- 5ml 1M $MgCl_2$
- 10ml Triton X-100

Make up to one litre and autoclave. Add 109.5g sucrose just before use. Store at 4°C.

TE buffer
(10mM Tris-HCl, 1mM Na_2EDTA)
Dissolve 121.1mg Tris and 37.2mg Na_2EDTA in 90ml water. Adjust the pH to 8.0 and make up to 100ml. Autoclave.

T20E5 buffer
Make up sufficient just before use (1.5ml/tube).
Mix together 2ml 1M Tris-HCl (pH 8.0) and 1ml 0.5M EDTA (pH8.0) and make up volume to 100ml with water.

APPENDIX B

DNA extraction method

Thawed buffy coats were transferred to 50ml NUNC tubes and the volume increased to 30ml by the addition of ice-cold Sucrose/Triton X-100 lysing buffer. The tubes were inverted to thoroughly mix and then centrifuged at 1 000x*g* for 10min at 4°C. The supernatant fluid (SNF) containing the red blood cells was removed and 20ml of the Sucrose/Triton X-100 lysing buffer added to the pellets. This mixture was cooled on ice for 5min, re-centrifuged at 1 000x*g* for 5min and the SNF discarded. Each pellet was combined with 1.5ml of T20E5 and after thorough mixing, 100µl of a 10% SDS solution was added. Finally, 250µl of a proteinase K working solution was added to each tube and the preparations left at 42°C overnight.

On the following morning, the preparations were vortexed to break up any remaining clumps of cell debri and 500µl of a saturated NaCl solution was added to each tube. The salt causes the dehydration and precipitation of degraded proteins. The samples were vigorously agitated for 15s by inversion and then centrifuged at 1 100x*g* for 15min. A white pellet consisting of salt precipitated protein was visible. The SNF containing the DNA was transferred to a fresh NUNC tube and two volumes of absolute ethanol were added to it. The resulting mixtures were swirled to facilitate mixing and the precipitation of the DNA. The precipitated DNA was 'fished' out of the tube using a sterile automatic pipette tip, washed in 70% ice cold ethanol, transferred to a sterile 1.5ml Eppendorf tube and air dried. Between 0.5ml and 1ml of TE buffer was added to the DNA which was allowed to dissolve for at least 24 hours before use. DNA quality was assessed after agarose gel electrophoresis and visualized by ethidium bromide staining. Aliquots of the DNA samples were kept at 4°C while the remainder was frozen at -20°C.

Pouring agarose gels

An appropriate amount of agarose powder (depending on desired gel concentration) was added to 100ml of 1XTBE buffer. This mixture was boiled until the powder had dissolved completely, giving the appearance of a translucent solution. The gel mixture was cooled to approximately 70°C before the addition of ethidium bromide [3µl Et Br (10mg/ml)/100ml gel)]. The gel was poured to a depth of ≈1cm into a mould with a gel comb in place to form loading wells and allowed to set at room temperature.

Pouring polyacrylamide gels

Glass plates (18cmX40cm) between which PA gels were formed were thoroughly cleaned, first with water and then with 70% ethanol. One of the plates was treated with a siliconising solution. Spacers were placed along the longer edge of the plate and the second plate (usually notched) was placed above the spacers to make a mould. The plates were held together during the entire pouring and gel setting procedure by 'bulldog' clips. The mould was filled with gel solution using a 50ml plastic disposable syringe. A well comb was inserted between the plates at the top end of the gel and polymerization was allowed to occur for at least 45min before the gel was used.

REFERENCES

Altheide TK and Hammer MF (1997). Evidence for a possible Asian origin of YAP Y-chromosomes. *Am J Hum Genet* **61**:462-466.

Armour JA, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR *et al* (1996). Minisatellite diversity supports a recent African origin for modern humans. *Nature Genet* **13**:154-160.

Bandelt HJ, Forster P, Sykes BC and Richards MB (1995). Mitochondrial portraits of human populations using median networks. *Genetics* **141**:743-753.

Barbujani G, Magagni A, Minch E and Cavalli-Sforza LL, (1997). An apportionment of human DNA diversity. *Proc Natl Acad Sci USA* **94**:4516-4519.

Batzer MA and Deininger PL (1991). A human-specific subfamily of *Alu* sequences. *Genomics* **9**:481-487.

Batzer MA, Kilroy GE, Richard PE, Shaikh TH, Desselle TD, Hoppens CL and Deininger PL (1990). Structure and variability of recently inserted Alu family members. *Nucleic Acids Res* **18**:6793-6798.

Batzer MA, Gudi VA, Mena JC, Foltz DW, Herrera RJ and Deininger PL (1991). Amplification dynamics of human-specific (HS) alu family members. *Nucleic Acids Res* **19**:3619-3623.

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, *et al* (1994). African origin of human-specific polymorphic *Alu* insertions. *Proc Natl Acad Sci USA* **91**:12288-12292.

Batzer M, Arcot S, Phinney J, Alegria-Hartman M, Kass D, Milligan S, *et al* (1996). Genetic variation of recent *Alu* insertions in human populations. *J Mol Evol* **42**:22-29.

Bellwood PS (1985). *Prehistory of the Indo-Malaysian archipelago*. Academic Press, Sydney.

Bellwood PS (1987). *The Polynesians*. Thames & Hudson, London.

Bhana S and Brain J (1990). *Setting down roots: Indian migrants in South Africa 1860-1911*. Witwatersrand University Press, Johannesburg.

Bianchi NO, Catanesi CI, Bailliet G, Martinez-Marignac VL, Bravi CM, Vidal-Rioja LB, Herrera RJ and Lopez-Camelo JS (1998). Characterization of ancetral and derived Y-chromosome haplotypes of new world native population. *Am J Hum Genet* **63**:1862-1871.

Bowcock AM, Bucci C, Herbert JM, Kidd KK, Friedlander JS and Cavalli-Sforza LL (1987). Study of 47 DNA markers in five populations from four continents. *Gene Geog* **1**:47-64.

Bowcock A, Kidd J, Mountain J, Herbert J, Carotenuto L, Kidd K and Cavalli-Sforza L (1991). Drift, admixture and selection in human evolution: a study with DNA polymorphisms. *Proc Natl Acad Sci USA* **88**:839-843.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**:455-457.

Brown M (1978). *Madagascar Rediscovered.* Damien Tunnacliffe, London.

Bruford MW and Wayne RK (1993). Microsatellites and their application to population studies. *Curr Opinion Genet Dev* **3**:939-943.

Buettner-Janusch J and Buettner-Janusch V (1964). Haemoglobins, haptoglobins and transferrins in the peoples of Madagascar. *Am J Phys Anthrop* **22**:163-169.

Buettner-Janusch J and Buettner-Janusch V (1970). Hemoglobin S in the Malagasy. *Arch Inst Pasteur Madagascar.* **39**:211-214.

Burney DA and MacPhee RDE (1988). Mysterious Island. What killed Madagascar's large native animals? *Natural History* 7:47-54.

Campbell G (1988). Madagascar and Mocambique in the slave trade of the Western Indian Ocean 1800-1811. *Slavery and Abolition. A journal of comparative studies* 9:1-193.

Campbell G (1989). The East African Slave Trade, 1861- 1895. The "southern" complex. *Int J African Historical Studies*, **22**:1-26.

Campbell G (1991). The state and pre-colonial demographic history: the case of nineteenth century Madagascar. *J Afr History* **32**:415-445.

Campbell G (1995). Theories of the origins of the Malagasy: the nineteenth century British school. *Australes – Etudes historiques aixoises sur l'Afrique australe et l'ocean Indien occidental*, L'Harmattan, Paris, Ed Paillard, pp 127-153.

Cann RL, Stoneking M and Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* **325**:31-36.

Castiglione CM, Deinard AS, Speed WC, Sirugo G, Zhang Y, Grandy DK *et al* (1995). Evolution of haplotypes at the DRD2 locus. *Am J Hum Genet* 57:1445-56.

Cavalli-Sforza LL, Kidd JR, Kidd KK, Bucci C, Bowcock AM, Hewlett BS and Friedlander JS (1986). DNA markers and genetic variation in the human species. *Cold Spring Harbor Symposia on Quantitative Biology*, Cold Spring Harbor, NY, Cold Spring Harbor Laboratory **51**:411-417.

Cavalli-Sforza LL, Piazza A, Menozzi P, Mountain J (1988). Reconstruction of human evolution: Bringing together genetic, archaeological and linguistic data. *Proc Natl Acad Sci USA* **85**:6002-6006.

Cavalli-Sforza LL, Menozzi P and Piazza (1994). *The history and geography of human genes*. Princeton University Press, Princeton NJ.

Chakraborty R (1977). Simulation results with stepwise mutation model and their interpretations. *J Mol Evol* 9:313-22.

Chamberlain JS, Gibbs RA, Ranier JE, Nguyen PN and Caskey CT (1988). Deletion screening of the Duchenne muscular dystrophy locus via multiplex DNA amplification. *Nucleic Acids Res* **16**:11141-11156.

Ciminelli BM, Pompei F, Malaspina P, Hammer M, Persichetti F, Pignatti PF, Palena A *et al* (1995). Recurrent simple tandem repeat mutations during human Y-chromosome radiation in Caucasian subpopulations. *J Mol Evol* **41**:966-976.

Cole M (1992). Return to Madagascar. *New Scientist* **22**:40-43.

Cooper G, Amos W, Hoffman D, Rubinsztein D (1996). Network analysis of human Y microsatellite haplotypes. *Hum Mol Genet* **5**:1759-1766.

Cooper G, William A, Bellamy R, Siddiqui MR, Frodsham A, Hill AVS and Rubinsztein DC (1999). An empirical exploration of the $(\delta\mu)^2$ genetic distance for 213 human microsatellite markers. *Am J Hum Genet* **65**:1125-1133.

Craig A (1996). The successful starlings of Africa. *Birds Africa* **Jan/Feb**:4-8.

DNA Commission of the ISFH (1994). Report concerning further recommendations of the DNA Commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeats). *Int J Legal Med* **107**:159-160.

Dahl OC (1951). *Malgache et Maanyan: Une comparison linguistique*, Egede Institutett, Oslo.
Cited in Dewar and Wright, 1993.

Dangerfield B (1997). The characterisation of Glucose-6-Phosphate Dehydogenase (G6PD) variation in the Malagasy. A dissertation submitted to the Faculty of Medicine, University of the Witwatersrand, Johannesburg, for the degree of Master of Science.

de Knijff P, Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G et al (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. *Int J Legal Med* **110**:134-140

Deka R, Jin L, Shriver MD, Yu LM, DeCroo S *et al* (1995). Population genetics of dinucleotide $(dC\text{-}dA)_n\text{-}(dG\text{-}dT)_n$ polymorphisms in world populations. *Am J Hum Genet* **56**:461-474.

Deka R, Jin L, Shriver MD, Mei Yu L, Saha N, Barrantes R, Chakraborty R et al (1996) Dispersion of human Y-chromosome haplotypes based on five microsatellites in global populations. *Genome Res* **6**:1177-1184

Dempster A, Laird N and Rubin D (1977). Maximum likelihood from incomlete data via the EM algorithm. *J Roy Sta Soc B* **39**:1-38.

Deschamps H (1965). *Histoirede Madagascar*, Editions Berger-Levrault, Paris. Cited in Dewar and Wright, 1993.

Dewar R (1992). Personal communication to J Yellen, Department of Anthropology, University of Connecticut.

Dewar R and Wright H (1993). The culture history of Madagascar. *J World Prehistory*. 7:417-466.

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

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M and Freimer NB (1994). Mutational processes of simple-sequence repeat loci in human populations. *Proc Natl Acad Sci USA* **91**:3166-3170

Diamond JM (1997). The language steamrollers. *Nature* **389**:544-546.

Dorit RL, Akashi H, Gilbert W (1995). Absence of polymorphism at the ZFY locus on the human Y-chromosome. *Science* **268**:1183-1185.

Dyen I (1953). Review of Otto Dahl, Malgache et Maanyan. *Language* 29:577-590. Cited in Verin *et al*, 1970.

Edwards MC and Gibbs RA (1992). A human dimorphism resulting from loss of an Alu. *Genomics* **14**:590-597.

Edwards A, Hammond HA, Jin L, Caskey CT and Chakraborty R (1992). Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. *Genomics* **12**:241-253.

Ehret C (1998). *An African Classical Age: eastern and southern Africa in World History, 1000 BC to AD 400*. The University Press of Virginia, Virginia.

Elston R (1971). The estimation of admixture in racial hybrids. *Ann Hum Genet* **35**:9-17.

Encarta (1999). *Encarta Multimedia Encyclopaedia*. Microsoft Corporation, USA.

Ferrand G (1908). L'origine africaine des Malgaches. *J Asiatique* **15**:353-500. Cited in Dewar, 1995.

Flacourt (1661). *Histoire de la grande ile Madagascar*, Union Coloniale, Paris. Cited in Dewar and Wright, 1993.

Gejman P, Ram A, Gelernter J, Friedman E, Cao Q, Pickar D, *et al* (1994). No structural mutation in the dopamine D_2 receptor gene in alcoholism or schizophrenia. *JAMA* **271**:204-208.

Goldstein DB (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci USA* **92**:6723-6727.

Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL, Feldman MW (1995). Genetic absolute dating based on microsatellites and the origins of modern humans. *Proc Natl Acad Sci USA* **92**:6723-6727.

Goldstein DB, Zhivotovsky LA, Nayar K, Linares AR, Cavalli-Sforza LL, Feldman MW (1996). Statistical properties of the variation at linked microsatellite loci: implications for the history of human Y-chromosomes. *Mol Biol Evol* **13**:1213-1218.

Gomolka M, Hundrieser J, Nurnberg P, Roewer L, Epplen JT and Epplen C (1994). Selected di- and tetranucleotide microsatellites from chromosomes 7, 12, 14, and Y in various Eurasian populations. *Hum Genet 93*:592-6.

Grandidier A and Grandidier G (1908). *Ethnographie de Madagascar*. Imprimerie Nationale, Paris.
Cited in Singer *et al*, 1957.

Grandy DK, Zhang Y and Civelli O (1993). PCR detection of the *Taq*I A RFLP at the DRD2 locus. *Hum Mol Genet* **2**:2197.

Greenberg JH (1963). *The languages of Africa*. Indiana University, Bloomington.

Grierson GA (1927). *Linguistic survey of India*. Motilal Banarsidas, Delhi.

Guo S-W and Thompson EA (1992). Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**:361-372.

Hammer MF (1994). A recent insertion of an *Alu* element on the Y chromosome is a useful marker for human population studies. *Mol Biol Evol* **11**:749-761.

Hammer MF and Horai S (1995) Y-chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet* **56**:951-962.

Hammer MF, Karafet T, Harihara S, Horai S and Stoneking M (1996). Tracking male-mediated migration in Asia: the origin of the Jomonese (abstract). *Am J Hum Genet* to vol **59** suppl:148.

Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, Mitchell RJ, Horai S, Jenkins T, Zegura SL (1997) The geographic distribution of human Y-chromosome variation. *Genetics* **145**:787-805.

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Atheide TK, Jenkins T, Griffiths RC, Templeton AR, Zegura SL (1998) Out of Africa and back again: Nested cladistic analysis of human Y-chromosome variation. *Mol Biol Evol* **15**:427-441.

Hammond HA, Jin L, Zhong Y, Caskey CT and Chakraborti R (1994). Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am J Hum Genet* **55**:175-189.

Harpending H and Rogers A (1984). *Antana: A package for Multivariate Data Analysis.* Harpending & Rogers, Bosque Farms.

Harpending H and Ward R (1982). Chemical systematics and human population. In: Nitecki M (ed) *Biochemical aspects of Evolutionary Biology.* University of Chicago Press, Chicago, pp 213-256.

Hawley ME and Kidd KK (1995). HAPLO: a program using the EM algorithm to estimate frequencies of multi-site haplotypes. *J Hered* **86**:409-411.

Hewitt R (1994). Investigation of the ancestry of the Malagasy population using β-globin haplotypes associated with the sickle cell mutation. A dissertation submitted to the Faculty of Science, University of the Witwatersrand, Johannesburg, for the degree of Bachelor of Science with Honours.

Hewitt R (1998). An investigation into the ancestry of the Malagasy population using variation in the α- and β-globin gene clusters. A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, for the degree of Master of Science.

Hewitt R, Krause A, Goldman A, Campbell G and Jenkins T (1996). β-globin haplotype analysis suggests that a major source of Malagasy ancestry is derived from Bantu-speaking Negroids. *Am J Hum Genet* **58**:1303-1308.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Kniff P (1997). Estimating Y-chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet* **6**:799-803.

Hill A, O'Shaughnessy D and Clegg J (1989). Haemoglobin and globin gene variants in the Pacific. In: Hill A and Serjeantson (eds) *The Colonisation of the Pacific: a genetic trail.* Oxford University Press, Oxford pp 246-285.

Horai S, Hayasaka K, Kondo R, Tsugane K and Takahata N (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci USA* **92**:532-536.

Hurles ME, Irven C, Nicholson J, Taylor PG, Santos FR, Loughlin J, Jobling MA, Sykes BC (1998). European Y-chromosomal lineages in Polynesians: A contrast to the population structure revealed by mtDNA. *Am J Hum Genet* **63**:1793-1806.

Iyengar S, Seaman M, Deinard AS, Rosenbaum HC, Sirugo G, Castiglione CM *et al* (1998). Analyses of cross-species polymerase chain reaction products to infer the ancestral state of human polymorphisms. *DNA Sequence*

Jakubiczka S, Arnemann J, Cooke HJ, Krawczak M, Schmidtke J (1989). A search for restriction fragment length polymorphism on human Y chromosome. *Hum Genet* **84**:86-88.

Jeffreys AJ, Wilson V and Thein SL (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature* **314**:67-73.

Jin L, Zhong Y and Chakraborty R (1994). The exact numbers of possible microsatellite motifs. *Am J Hum Genet* **55**:582-3.

Jobling MA, Tyler-Smith C (1995) Fathers and sons: the Y-chromosome and human evolution. *Trends Genet* **11**:449-456.

Jorde L, Bamshad WS, Zenger R, Fraley AE, Krakowiak PA, Carpenter KD, *et al* (1995). Origins and affinities of modern humans: A comparison of mitochondrial and nuclear genetic data. *Am J Hum Genet* **57**:523-538.

Karafet TM, Zegura SL, Posukh O, Osipova L, Bergen A, Long J, Goldman D, Klitz W, Harihara S, de Knijff P, Wiebe V, Griffiths RC, Templeton AR, Hammer MF (1999). Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am J Hum Genet* **64**:817-831.

Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, herzog B, Hidding M et al (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Leg Med* **110**:125-133.

Kidd KK and Kidd JR (1996). A nuclear perspective on human evolution. In: AJ Boyce, CGN Mascie-Taylor (eds) *Molecular biology and human diversity*. Cambridge University Press, Cambridge, pp 242-264.

Kidd KK, Morar B, Castiglione CM, Zhao H, Pakstis AJ, Speed WC *et al* (1998). A global survey of haplotype frequencies and linkage disequilibrium at the DRD2 locus. *Hum Genet* **103**:211-227.

Kimura M and Crow JF (1964). The number of alleles that can be maintained in a finite population. *Genetics* **49**:725-38.

Kittles RA, Perola M, Peltonen L, Bergen AW, Aragon RA, Virkkunen M, Linnoila M, Goldman D, Long JC (1998). Dual origins of Finns revealed by Y-chromosomes haplotype variation. *Am J Hum Genet* **62**:1171-1179.

Kumar S, Tamura K and Nei M (1993). MEGA: molecular evolutionary genetics analysis, version 1.01. Institute of Molecular Evolutionary Genetics, Pennsylvania State University, University Park.

Lahr MM and Foley RA (1998). Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *Am J Phys Anthropol* Suppl **27**:137-76.

Lapouméroulie C, Dunda O, Duncrocq R, Trabuchet G, Mony-Lobe M, Bodo J *et al* (1992). A novel sickle cell mutation of yet another origin in Africa. The Cameroon type. *Hum Genet* **89**:333-337.

Lewontin RC (1992). The apportionment of human diversity. *Evol Biol* **6**:381-398

Linares RA, Nayar K, Goldstein DB, Hebert JM, Seielstad MT, Underhill PA, Lin AA, Feldman MW, Cavalli-Sforza LL (1996). Geographic clustering of human Y-chromosome haplotypes. *Ann Hum Genet* **60**:401-408.

Mack J (1986). *Madagascar, Island of the Ancestors*. British Museum Publ. Ltd., London.

Malaspina P, Persichetti F, Novelletto A, Iodice C, Terrenato L, Wolfe J, Ferraro M, Prantera G (1990). The human Y chromosome shows a low level of DNA polymorphisms. *Ann Hum Genet* **54**:297-305.

Malaspina P, Cruciani F, Ciminelli BM, Terrenato L, Santolamazza P, Alonso A, Banyko J, Brdicka R et al (1998). Network analysis of Y-chromosomal types in Europe, Northern Africa, and Western Asia reveal specific patterns of geographic distribution. Am J Hum Genet 63:847-860.

Manguin P (1993). Pre-modern Southeast Asian shipping in the Indian Ocean: the Maldives connection. New directions in Maritime history conference, Fremantle, December 1993.

Matera AG, Hellmann U and Schmid CW (1990). A transpositionally and transcriptionally competent Alu subfamily. *Mol Cell Biol* **10**:5424-5432.

Mathias N, Bayes M and Tyler-Smith C (1994). Highly informative compound haplotypes for the human Y-chromosome. *Hum Mol Genet* **3**:115-123.

Merriweather DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T *et al* (1991). The structure of human mitochondrial DNA variation *J Mol Evol* **33**:543-555.

Miller SA, Dyles DD and Polesky HF (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**:1215.

Mistry SD (1965). Ethnic groups of Indians in South Africa. *S Afr Med J* **39**:691-4.

Mourant A, Kopec A and Domaniewska-Sobczak K (1976). *The distribution of the human blood groups and other polymorphisms*. Oxford University Press, London.

Murdoch GP (1959). Africa. *Its people and their culture history*. McGraw-Hill Book Company, Inc, USA.

Nei M and Roychoudary AK (1972). Gene differences between Caucasian, Negro, and Japanese populations. *Science* **177**:434-6.

Nei M, Tajima F and Tateno Y (1983). Accuracy of estimated phylogenetic trees from molecular data. II. Gene frequency data. *J Mol Evol* **19**:153-70.

Nei M (1987) *Molecular evolutionary genetics*. Columbia University press, New York.

Nei M and Roychoudary A (1993). Evolutionary relationships of human populations on a global scale. *Mol Biol Evol* **10**:927-943.

Oakey R, Tyler-Smith C (1990). Y chromosome DNA haplotyping suggests that most European and Asian men are descended from one of two males. *Genomics* 7:325-330.

Ohta T and Kimura K (1973). The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. *Genet Res* **22**:201-204.

Oner C, Dimovski A, Olivieri N, Schiliro G, Codrington J *et al* (1992). β^S haplotypes in various world populations. *Hum Genet* **89**:99-104.

Oliver R and Fagan B (1975). *Africa in the Iron Age*. Cambridge University Press, Cambridge.

Penny D, Steel M, Waddell PJ and Hendy MD (1995). Improved analyses of human mitochondrial DNA sequences support a recent African origin for *Homo sapiens*. *Mol Biol Evol* **12**:863-882.

Perez-Lezaun A, Calafell F, Comas D, Mateu E, Bosch E, Martinez-Arias R, Clarimon J, Fiori G, Luiselli D, Facchini F, Pettener D, Bertranpetit J (1999). Sex-specific migration in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet* **65**:208-219.

Perez-Lezaun A, Calafell F, Mateu E, Comas D, Ruiz-Pacheco R and Bertranpetit J (1997). Microsatellite variation and the differentiation of modern humans. *Hum Genet* **99**:1-7.

Persichetti F, Blasi P, Hammer M, Malaspina P, Jodice C, Terrento L, Novelletto A (1992). Disequilibrium of multiple DNA markers on the human Y chromosome. *Ann Hum Genet* **56**:303-310.

Ramsay M and Jenkins T (1988). Alpha-globin gene cluster haplotypes in the Kalahari San and southern African Bantu-speaking Blacks. *Am J Hum Genet* **43**:527-533.

Raymond M and Rousset F (1995). An exact test for population differentiation. *Evolution* **49**:1280-1283.

Redd AJ, Clifford SL, Stoneking M (1997). Multiplex DNA typing of short tandem repeat loci on the Y-chromosome. *Biol Chem* **378**:923-927.

Reynolds J, Weir BS, Cockerham CC (1983). Estimation of the coancestry coefficient: basis for short term genetic distance. *Genetics* **105**:767-779.

Roewer L, Arnemann J, Spurr NK, Grzeschik K-H, Epplen JT (1992). Simple repeat sequences on the human Y-chromosome are equally polymorphic as their autosomal counterparts. *Hum Genet* **89**:389-394.

Roewer L, Kayser M, Dieltjies P, Nagy M, Bakker E, Krawczak M, De Knijff P (1996). Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. *Hum Mol Genet* **5**:1029-1033.

Rogers J (1983). Retroposons defined. *Nature* **301**:460.

Ruiz-Linares A, Nayar K, Goldstein DB, Hebert JM, Seielstad MT, Underhill PA, Lin AA et al (1996). Geographic clustering of human Y-chromosome haplotypes. *Am Hum Genet* **60**:401-408.

Saiki R, Scharf S, Faloona F, Mullis K, Horn G, Erlich H and Arnheim N. (1985). Enzymatic amplification of β-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anaemia. *Science* **230**:1350-1354.

Saitou N and Nei M (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4:406-425.

Santos F R, Gerelsaikhan T, Munkhtuja B, Oyunsuren T, Epplen JT, Pena SDJ (1996). Geographic differences in the allele frequencies of the human Y-linked tetranucleotide polymorphism DYS19. *Hum Genet* **97**:309-313.

Schneider S, Kueffer JM, Roessli D, Excoffier L (1997). Arlequin version1.1: a software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Geneva.

Schuurman D (1994). Two perspectives of Madagascar. *African Panorama* **39**:6-19.

Scozzari R, Cruciani F Santolamazza P Sellitto D, Cole DEC, Rubin LA *et al* (1997). mtDNA and Y chromosome-specific polymorphisms in modern Ojibwa: implications about the origin of their gene pool. *Am J Hum Genet* **60**:241-244.

Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D *et al* (1999). Combined use of bi-allelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet* **65**:829-846.

Shriver MD, Jin L, Chakraborty R and Boerwinkle E (1993). VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. *Genetics* **134**:983-993.

Shriver MD, Jin L, Boerwinkle E, Deka R, Ferrell R, Chakraborty R, (1995). A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol Biol Evol* **12**:914-920.

Singer R, Budtz-Olsen OE and Brain P (1957). Physical features, sickling and serology of the Malagasy of Madagascar. *Am J Phys Anthropol* **15**:91-123.

Skorecki K, Selig S, Blazer S, Bradman R, Bradman N, Waburton PJ, Ismajlowicz M and Hammer MF (1997). Y chromosomes of Jewish priests (letter). *Nature* **385**:32.

Slatkin M (1995). A measure of population subdivision based on microsatellite frequencies. *Genetics* **139**:457-562.

Soodyall H, Stoneking M and Jenkins T (1995). Polynesian mtDNA in the Malagasy. *Nat Genet* **10**:377-378.

Sprecher CJ, Puers C, Lins AM and Schumm JW (1996). General approach to analysis of polymorphic short tandem repeat loci. *Biotechniques* **20**:266-276.

Spurdle A, Jenkins T (1992). The Y chromosome as a tool for studying human evolution. *Curr Op Genet Dev* **2**:487-491.

Spurdle AB, Hammer MF, Jenkins T (1994a). The Y *Alu* polymorphism in Southern African populations and its relationship to other Y-specific polymorphisms. *Am J Hum Genet* **54**:319-330.

Spurdle AB, Woodfield DG, Hammer MF and Jenkins T (1994b). The genetic affinity of Polynesians: evidence from Y-chromosome polymorphisms. *Ann Hum Genet* **58**:251-263.

Stiles D (1988). Tubers and Tenrecs: the Mikea of southwestern Madagascar. *Ethnology* pp 251-263.

Stoneking M, Hedgecock D, Higuchi RG, Vigilant L and Erlich HA (1991). Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am J Hum Genet* **48**:370-382.

Stoneking M, Fontius J, Clifford S, Soodyall H, Arcot S, Saha N *et al* (1997). *Alu* insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. *Genome Research* 7:1061-1071.

Stringer CB and Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. *Science* **239**:1263-8.

Takezaki N and Nei M (1996). Genetic distances and reconstruction of phylogenetic trees from microsatellite DNA. *Genetics* **144**:389-399.

Tautz D (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. *Nucleic Acids Res* **17**:6463-6471.

Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K *et al* (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science* **271**:1380-1397.

Tishkoff SA, Goldman A, Calafell F, Speed WC, Deinard AS, Bonne-Tamir B *et al* (1998). A global haplotype analysis of the myotonic dystrophy locus: implications for the evolution of modern humans and for the origin of myotonic dystrophy mutations. *Am J Hum Genet* **62**:1389-1402.

Todd JA, Aitman TJ, Cornall RJ, Ghosh S, Hall JRS *et al* (1991). Genetic analysis of autoimmune type 1 diabetes mellitus in mice. *Nature* **351**:542-547.

Underhill PA, Jin L, Zemans R, Oefner PJ, Cavalli-Sforza LL (1996). A pre-Columbian Y chromosome-specific transition and its implications for human evolutionary history. *Proc Natl Acad Sci USA* **93**:196-200.

Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ (1997) Detection of numerous Y-chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* 7:996-1005.

Verin P, Kottak CP and Gorlin P (1970). The glottochronology of Malagasy speech communities. *Oceanic Linguistics* **8**:26-81.

Verin (1975). *Les echelles anciennes du commerce sur les cotes nord de Madagascar*, Service de Reproduction des theses, Lille.
Cited in Dewar and Wright, 1993.

Verin P (1986). *The history of civilization in northern Madagascar*. Balkema, Rotterdam and Boston.
Cited in Dewar and Wright, 1993.

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

Wainscoat JS, Hill AVS, Boyce AL, Flint J, Hernandez M *et al* (1986). Evolutionary relationships of human populations from an analysis of DNA polymorphisms. *Nature* **319**:491-493.

Watterson GA and Guess HA (1977). Is the most frequent allele the oldest? *Theor Popul Biol* **11**:141-60.

Watkins WS, Bamshad M and Jorde LB (1995). Population genetics of trinucleotide repeat polymorphisms. *Hum Mol Genet* **4**:1485-91.

Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, Paabo S (1996). mtDNA sequence diversity in Africa. *Am J Hum Genet* **59**:437-444.

Weber JL, Wong C (1993). Mutation of human short tandem repeats. *Hum Mol Genet* **2**:1123-1128.

Weiner AM, Deininger PL and Efstratiadis A (1986). Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Ann Rev Biochem* **55**:631-661.

Weir BS (1996). *Genetic analysis II: Methods for discrete population genetic data*. Sinauer Associates Inc: Sunderland, Massachusetts, USA.

Whitfield LS, Sulston JE, Goodfellow PN (1995) Sequence variation of the human Y-chromosome. *Nature* **378**:379-380.

Willems PJ (1994). Dynamic mutations hit double figures. *Nat Genet* **8**:213-5.

Wolpoff MH, Spuhler JN, Smith FH, Radovcic J, Pope G, Frayer DW, Eckhardt R and Clark G (1989). Modern human origins (letter) *Science* **241**:772-4.

Zago MA, Silva WA, Tavella MH, Santos SEB, Guerreiro JF and Figueiredo MS (1996). Interpopulational and intrapopulational genetic diversity of Amerindians as revealed by six variable number of tandem repeats. *Hum Hered* 46:274-289.