

https://dspace.org/

Molecular Genetic Variation in Mozambican and Zimbabwean populations

Eddy, Carey-Anne

2001

https://hdl.handle.net/10539/32801

Eddy, C. Molecular Genetic Variation in Mozambican and Zimbabwean populations. 2001. https://hdl.handle.net/10539/32801

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

Molecular Genetic Variation in Mozambican and Zimbabwean populations

Carey-Anne Eddy

2001

A dissertation submitted to the Faculty of Medicine, University of the Witwatersrand, Johannesburg, for the degree of Master of Science.

ABSTRACT

Genetic data, used in conjunction with linguistic, archaeological and anthropological data, is extremely valuable in understanding the evolutionary and demographic history of sub-Saharan African populations. Present-day southern African populations are represented by Khoisan and Bantu-speaking groups. The Bantu languages have their origins in the Benue-Niger area between present-day Nigeria and Cameroon, and has spread to central and southern Africa over the past 2-3000 years. The spread of Bantu-speaking peoples, referred to as the "Bantu expansion", ought to have contributed to the spectrum of genetic variation found in sub-Saharan African populations. In this study we propose to examine the genetic structure of two southeastern Bantu-speaking population groups: the Remba (a Lemba group) from Zimbabwe and a Mozambican group sampled near Inhambane.

Although the Lemba from South Africa live among the Venda and the Remba among the Shona, their intriguing oral history and unique Semitic-like cultural practices have set them apart from their Bantu-speaking neighbours. They claim descent from Jews who came to Africa from "Sena". A previous study using Y chromosome markers suggested both a Bantu and Semitic contribution to the Lemba gene pool. More recently, the presence of a particular Y chromosome type termed the "Cohen Modal Haplotype" (CMH) in the Lemba from South Africa, led others to conclude that the Lemba have a specifically Jewish ancestry. These studies, however, have not resolved conclusively the origins of the Lemba.

The Mozambican gene pool might reflect contributions from other populations particularly due to the crucial geographic position it occupied, linking landlocked countries to coastal

ports, during the days of slavery and trade with the East. There have only been limited genetic studies in populations from Mozambique and even the archaeological record from present-day Mozambique is especially sparse.

We used the non-recombining portion of the Y chromosome and mtDNA data to examine the contribution of males and females, respectively, to the gene pools of the Remba and the Mozambicans and to investigate the prehistory of these two population groups.

The Remba from Zimbabwe and the Lemba from South Africa shared common Y chromosome haplotypes that were connected via single mutational steps, suggesting that they have a common recent ancestry. The CMH was found in only one Lemba individual from South Africa and was absent in the Remba. Furthermore, there were no one-step derivatives of the CMH in the Lemba or the Remba. It is possible that the CMH was introduced into the Lemba fairly recently due to gene flow from South African Jews. Although the Remba and Lemba do have a common Semitic contribution to their Y chromosme gene pool, it is premature to use the CMH as evidence of Jewish ancestry in the Lemba.

There was no evidence from the mtDNA data to suggest any non-African contribution to the mtDNA gene pool of the Remba and Lemba or the Mozambicans. There is, however strong evidence suggesting that the Remba, Lemba, Mozambicans, Venda, southeastern Bantu speaking population groups from South Africa and several other sub-Saharan African

populations share a recent common maternal ancestry that most likely has its roots in the Benue-Niger region prior to the Bantu expansion.

Despite the role the Mozambicans may have played during the days of slavery and trading, their Y chromosome profile shows no reflection of any contribution from non-African populations.

DECLARATION

I declare that this dissertation is my own unaided work. It is being submitted for the degree of Master of Science in Human Genetics at the University of the Witwatersrand, Johannesburg. It has not been submitted for any other degree or examination at any other university.



Carey-Anne Eddy

24th September 2001

This is dedicated to my parents, Meggie and Gert

ACKNOWLEDGEMENTS

There are a number of people and organisations I would like to thank for assisting, in some way or other, with this project:

My supervisor, Dr Himla Soodyall, for all her help, time and patience

Bharti Morar, for always giving so generously of her time, irrespective of how busy she was herself

Debbie Barkhan, Cindy Dorf and Sharmila Manga, not only for checking numerous sequences, but for always being available with time, help and support when they were needed

Professor Jenkins and all the members of the Human Genetics Department

All the individuals who donated blood and everyone who participated in the collection of blood samples

Professor Leopoldo da Costa, Dean of Research, Univeridade Eduardo Mondlane for facilitating the collection of the Mozambican samples

The SAIMR, the University of the Witwatersrand and the Medical Research Council for financial support.

TABLE OF CONTENTS

ABSTRACT.....ii
DECLARATION.....vi
ACKNOWLEDGEMENTS.....viii
TABLE OF CONTENTS.....ix
LIST OF TABLES.....xiii
LIST OF FIGURES.....xiv
ABBREVIATIONS.....xvi

1. INTRODUCTION.....1

1.1 Linguistic studies.....2
1.2 Archaeological and anthropological data.....8
1.3 Reconstructing the prehistory of the Lemba.....10
1.3.1 Historical and cultural evidence.....10
1.3.2 Theories comcerning Lemba origins.....12
1.3.3 Genetic evidence.....14
1.4 The peopling of Mozambique.....19
1.4.1 Genetic studies in populations from Mozambique.....20
1.5 Project objectives.....21
1.6 Genetic markers used in this study.....23
1.6.1 MtDNA studies.....23
1.6.2 The 9-bp deletion.....25
1.6.3 Y chromosome studies.....26
1.6.3.1 The Y *Alu* polymorphism (YAP).....27
1.6.3.2 Y-STR loci.....28

2. SUBJECTS AND METHODS.....30

2.1 Samples.....31
2.2 DNA extraction.....33
2.3 Molecular methods.....33
2.3.1 MtDNA studies.....33
2.3.1.1 Screening for the 9-bp deletion.....33
2.3.1.2 MtDNA sequencing.....34
2.3.2 Y chromosome markers.....37
2.3.2.1 Screening for the YAP.....37
2.3.2.2 Typing of STR loci.....37
2.4 Data analysis.....41
2.4.1 MtDNA data analysis.....41
2.4.2 Y chromosome data analysis.....45
2.4.3 Analysis of molecular variance (AMOVA) and population affinities.....45
2.4.4 Genetic diversities.....46

3. RESULTS.....47

3.1 MtDNA studies.....48
3.1.1 Frequencies of the 9-bp deletion.....48
3.1.2 Phylogenetic analysis of mtDNA sequences.....50
3.1.2.1 The Remba, Lemba and Venda population groups.....50
3.1.2.2 The Mozambican population group.....51
3.1.2.3 A combined phylogeny of the Remba, Lemba, Venda and Mozambicans.....52
3.1.3 MtDNA type diversity.....56
3.1.4 Haplogroup distribution in the Remba, Lemba, Venda and Mozambicans.....62
3.1.4.1 Estimating the age of the MRCA of the L subhaplogroups.....65

3.1.5 MtDNA types associated with the 9-bp deletion.....68
3.1.6 Mismatch distributions.....72
3.1.7 Affinities of the Remba, Lemba and Venda with the Mozambicans and other sub-Saharan African populations.....75
3.2 Y chromosome studies.....78
3.2.1 Variation at individual loci.....78
3.2.2 Haplotype analysis.....81
3.2.3 Y chromosome networks.....85
3.2.4 Frequency and distribution of the Cohen Model Haplotype.....88
3.2.5 Y chromosome affinities of the Lemba, Remba, Venda and Mozambicans.....91
3.3 Genetic structure of the Remba, Lemba, Venda and Mozambicans.....93

4. DISCUSSION.....95

4.1 The prehistory of the Lemba.....96
4.1.1 Origins of Y chromosomes in the Lemba and Remba.....96
4.1.2 The CMH - evidence of a Judaic origin for the Lemba?.....101
4.1.3 Maternal ancestry of the Lemba and Remba.....102
4.1.3.1 The 9-bp deletion.....102
4.1.3.2 Phylogeny of mtDNA types in the Remba and Lemba.....103
4.1.3.3 RFLP haplogroup distribution.....103
4.1.3.4 Demographic history of the Remba and Lemba.....104
4.2 The peopling of Mozambique.....105
4.2.1 Origins of Y chromosomes in the Mozambicans.....106
4.2.2 Maternal ancestry of the Mozambicans.....107
4.2.2.1 The 9-bp deletion.....108
4.2.2.2 Phylogeny of mtDNA types in the Mozambicans.....109
4.2.2.3 RFLP haplogroup distribution.....110
4.2.2.4 Demographic history of the Mozambicans.....110

4.3 Genetic affinities of the Remba, Lemba, Mozambicans and Venda with other sub-Saharan African populations.....111

4.4 Genetic structure of the Remba, Lemba, Mozambican and Venda population groups.....113

4.5 Genetic diversity in the Mozambicans and the Remba.....114

5. CONCLUSIONS.....116

6. REFERENCES.....119

7. APPENDIX.....129

7.1 Recipes and solutions.....130

7.2 Size standards.....133

7.2.1 One kb ladder.....133

7.2.2 ROX-350 (ABI) size standard.....134

7.3 Ethics clearance.....134

7.4 Table of mtDNA types derived from CR1 sequence data.....137

LIST OF TABLES

1. Primers used for mtDNA PCR and sequencing.....36
2. Primers used for YAP and Y-STR PCR.....39
3. Y-STR reaction mixes.....40
4. Frequencies of the 9-bp deletion in SEB speaking populations.....49
5. Diversity parameters for the Remba, Lemba, Venda and Mozambicans.....58
6. MtDNA types found in the Remba, Mozambicans, Lemba and Venda.....59
7. The frequencies and distribution of the haplogroups that corresponded to mtDNA CR sequence types found in the Remba, Mozambicans, Lemba and Venda.....67
8. Summary statistics for the mismatch distributions shown in Figure 9.....74
9. Allele frequencies of YAP and six Y chromosome STRs.....79
10. YAP+ haplotypes found in the Mozambicans, Venda, Lemba and Remba based on STR loci DYS393, DYS19, DYS390, DYS391, DYS388 AND DYS392.....82
11. YAP- haplotypes found in the Mozambicans, Venda, Lemba and Remba based on STR loci DYS393, DYS19, DYS390, DYS391, DYS388 AND DYS392.....83
12. Haplotype diversity and haplotype sharing among YAP+ and YAP- chromosomes in the Mozambicans, Venda, Lemba and Remba.....84
13. YAP- haplotypes associated with the CMH in the Lemba, Remba, South African Jews, Ashkenazi and Sephardic Jews, Yemeni and Sena.....89
14. Percentage of variance at different hierarchical levels using AMOVA for mtDNA sequence data and Y chromosome haplotype data.....94

LIST OF FIGURES

1. The distribution of the major language families of Africa (after Greenburg, 1963)....3
2. A painting of the Hebrew congregation's Krogh Street synagogue which was built in 1946-7.....18
3. Map showing the approximate geographic locations of the population groups examined in this study.....32
4. Unrooted NJ tree showing the phylogenetic relationship of unique mtDNA types in the Remba, Lemba and Venda populations.....53
5. Unrooted NJ tree showing the phylogenetic relationship of mtDNA types found in individuals from Mozambique.....54
6. A NJ tree showing the phylogeny of mtDNA types in the Remba, Lemba, Venda and Mozambicans.....55
7. NJ tree showing two distinct clusters of mtDNA types associated with the 9-bp deletion.....70
8. Median network showing the relationship of mtDNA that correspond to subhaplogroup L1a in individuals with and without the 9-bp deletion.....71
9. Mismatch distributions of sequences within the Remba, Mozambicans, Lemba and Venda.....73
10. NJ population tree showing the mtDNA genetic affinities of the Remba, Mozambican, Lemba and Venda populations with other sub-Saharan African populations.....76

11. PCO-plot showing the genetic affinities of the Remba, Mozambicans, Lemba and Venda with other sub-Saharan African populations using mtDNA CR1 sequence data.....77

12. Median joining network showing the relationship of YAP+ haplotypes found in the Lemba, Remba, Mozambicans and Venda.....86

13. Median joining network showing the relationship of YAP- haplotypes found in the Lemba, Remba, Mozambicans and Venda.....87

14. Network of YAP- haplotypes associated closely with the CMH in the Lemba, Remba, Yemeni, Sena, Ashkenazi and Sephardic Jewish populations, and a South African Ashkenazi jewish population.....90

15. PCO-plot showing the genetic affinities of the Lemba, Remba, Venda and Mozambicans using Y chromosome haplotype data.....92

ABBREVIATIONS

AMOVA	Analysis of molecular variance
APS	ammonium persulphate
bp	base pair
BSA	bovine serum albumin
CMH	Cohen Model Haplotype
CR	control region
ddH_2O	double-distilled water
dNTPs	deoxynucleotide triphosphates
EDTA	ethylenediamine tetra-acetic acid, disodium salt
EtBr	ethidium bromide
kb	kilobase
MEBOOT	Minimum Evolution Bootstrapping
MEGA	Molecular Evolutionary Genetics Analysis
MRCA	most recent common ancestor
mtDNA	mitochondrial DNA
NJ	neighbour-joining
np	nucleotide posotion
PCR	polymerase chain reaction
RFLP	restriction fragment length polymorphism
SEB	southeastern Bantu

STR	short tandem repeats
SWB	southwestern Bantu
TAE	tris-acetate, EDTA
TBE	tris-borate, boric acid, EDTA
TEMED	N, N, N1, N1, -tetramethylethylenediamine
YAP	Y *Alu* polymorphism

1. INTRODUCTION

Linguistic, archaeological, historical and anthropological data has been used extensively to study the prehistory of sub-Saharan African populations. Genetic data used in conjunction with data from these disciplines would lead to a more complete understanding of the prehistory of these populations.

1.1 Linguistic studies

Language has been extremely useful in reconstructing the prehistory of African populations. There is reasonable consensus among linguists concerning the classification of language families in Africa. According to Greenberg's (1963) widely accepted classification, the approximate 1400 languages spoken currently in Africa belong to four major families: Afroasiatic, Nilo-Saharan, Niger-Congo and Khoisan. The distribution of people speaking languages belonging to these families is shown in Figure 1 (simplified from Greenberg, 1963). Languages spoken predominantly in the northern and northeastern regions of the continent belong to the Afroasiaitic language family. The main divisions within this family include the Berber languages of North Africa, the Cushitic languages spoken mainly in Ethiopia and Somalia, and the widespread Semitic family including Arabic, Amharic and Hebrew. Most present-day languages spoken further south; in the central and southern Sahara, southern Sudan and parts of the adjacent savanna extending into parts of east Africa, are classed as Nilo-Saharan, with Nilotic and Sudanic as the main subdivisions. Most of the modern languages of West Africa belong to the Niger-Congo family, which can include Kordofanian, spoken in western Sudan. The distribution of one of the subdivisions of Niger-

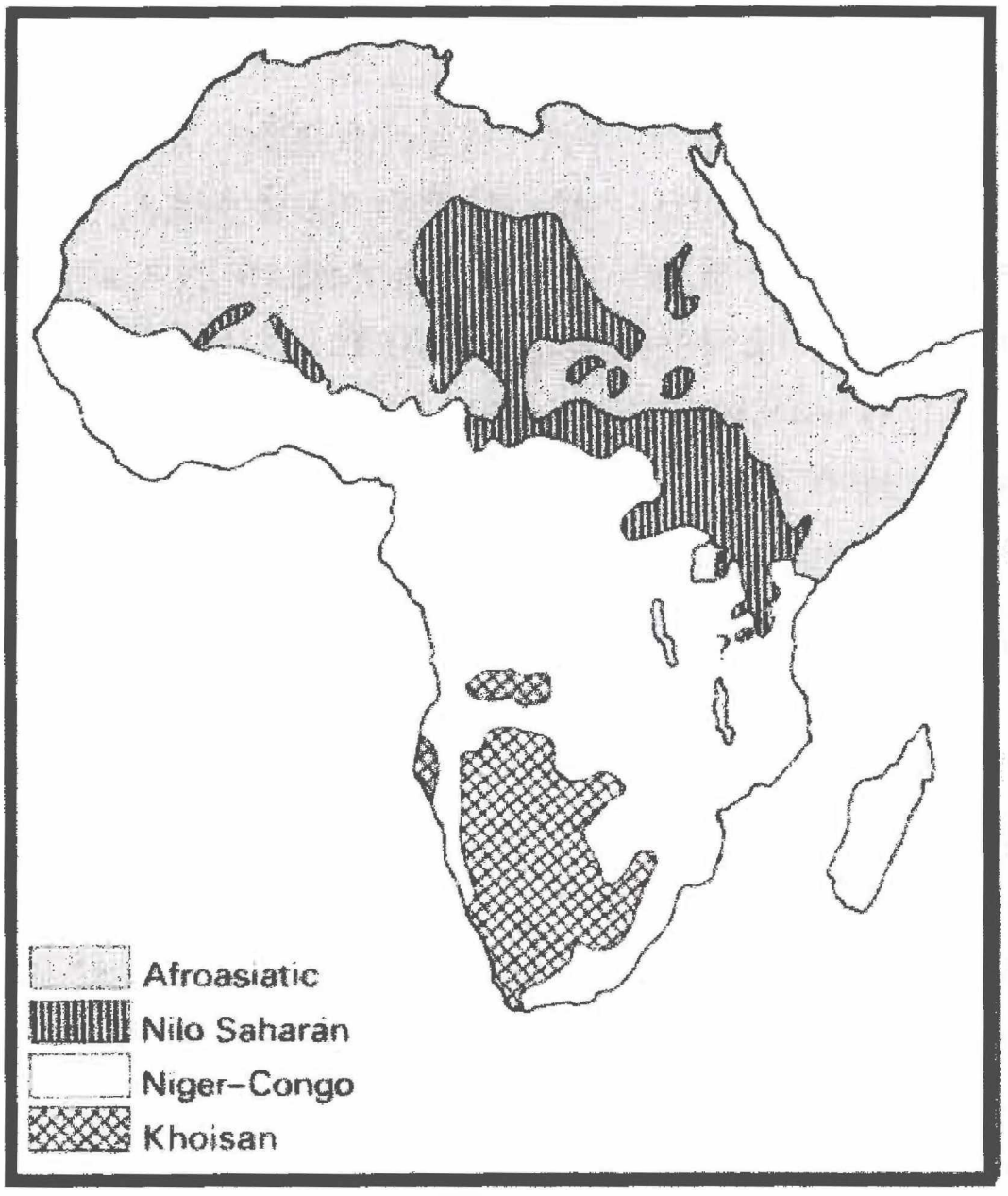


Figure 1. The distribution of the major language families of Africa (after Greenberg, 1963).

Congo, the Bantu languages, extends over most of central and southern Africa, excluding the extreme southwest where Khoisan languages were spoken. Southern Africa was occupied exclusively by Khoisan-speaking peoples before the appearance of Bantu-speakers approximately 2000 years ago (Ehret, 1972). Their distribution has since become restricted to some parts of the central Kalahari Desert in Botswana, and in Namibia.

By incorporating various linguisitic criteria, linguistics can trace the origins of languages as well suggest the areas in which the ancestral language originated (Bleek, 1927; Doke, 1954; Guthrie, 1962; Westphal, 1963; Ehret, 1972). For example, common vocabulary can be used to relate different languages as well as provide clues about the life-styles of the people who spoke them (Greenberg, 1963; Guthrie, 1962). As groups came into contact with each other, words were often borrowed from the one group and assimilated into the other group. Linguistic data, therefore, provides a valuable measure of population distribution and movement in periods without recorded history. Elucidating the origins and expansions of people speaking Bantu languages is a good example of the value of linguistic studies in reconstructing population history.

Most of the Negroid populations occupying present-day sub-Saharan Africa belong to one of over 2000 Bantu-speaking language groups in eastern, southern and central Africa covering an area of nearly nine million square kilometres (Phillipson, 1993). There is clear agreement among linguists that the many Bantu languages are very closely related, implying that Bantu-speakers must have expanded very rapidly to have achieved such a wide geographical

dispersion along with such a small degree of linguistic divergence (Oliver, 1966). Bleek recognised their relationship as early as 1862, and coined the name Bantu for these languages because of their common use of the word "Bantu" for people (Phillipson, 1993). It has been suggested that the Bantu languages originated in the Benue-Niger area, between present day Nigeria and Cameroon, and then spread as a series of waves of migration to other parts of Africa over the past 2-3000 years (Greenberg, 1963; Guthrie, 1962). It is generally accepted that Bantu languages spread as a result of two major routes of migration: a western migration, that permeated through the equatorial forest, and an eastern one that led towards the Great Lakes region to form what are referred to by linguists as the western and eastern Bantu languages, respectively (Phillipson, 1993; Vansina, 1984).

In southern Africa, both southeastern Bantu (SEB) and southwestern Bantu (SWB) languages have diverged from the parental eastern and western Bantu languages, respectively (Westphal, 1963). Many cultural practises of the proto-western Bantu-speakers can be deduced from their vocabulary which suggests that these people were fisherman, farmers of yams, oil palms and minor domestic plants, hunters and gatherers of wild produce and potters (Vansina, 1984). These people did not farm cereal crops nor were they iron or metal workers. These small-scale farmers speaking western Bantu languages gradually occupied much of central Africa where they divided, with one group developing a more aquatic culture associated with fishing and the other proceeding southwards, ultimately giving rise to the SWB speakers found in most parts of Namibia, the Zambezi Valley, the Luangwa Valley, around northern Lake Malawi and Lake Tanganyika. Their expansion was probably arrested

in the southeast when they met farmers speaking eastern Bantu languages (Vansina, 1984). Population groups who speak SWB languages include the Herero, the Himba, the Ambo or Ovambo and the Kavango of Namibia.

Although the eastern Bantu had farmed cereal crops since they left the Cameroon region, it was in the region of the Great Lakes that this group of Bantu-speakers mastered techniques for iron smelting which they eventually carried, along with cereal agriculture and cattle-keeping, throughout eastern and southern Africa in the first centuries of the first millennium AD (Huffman, 1989). Since the Zimbabwean and Mozambican individuals sampled in this study belong to SEB language groups, more attention is given to this group of Bantu speakers than the SWB.

The SEB speakers are comprised of three principal groups: the Nguni, the Sotho/Tswana and the Tsonga (Doke, 1954). Although the Nguni share some common cultural traits, their classification is based on a broad linguistic uniformity. At present the Nguni peoples occupy an area ranging from what used to be the Ciskei to Swaziland. They include the Zulu, Swazi, Xhosa and Ndebele groups (Doke, 1954). The Sotho/Tswana term describes a large group of people who display linguistic and some cultural similarities. They occupy the interior plateau of South Africa and are subdivided into three main chiefdoms (Doke, 1954): the western Sotho or Tswana group, the northern Sotho or Pedi group and the southern sotho (Basotho) who occupy present day Lesotho and adjacent areas (Nurse *et al.*, 1985). The Tsonga and the Shona occupy the southern parts of Mozambique and Zimbabwe, respectively. The Shona

were the initial occupants of present-day Zimbabwe and it was only in the early 19th century that the Ndebele (belonging to the Nguni branch of the SEB language family) from further south (present-day South Africa) settled in the area (Kuper, 1955).

Linguistically, the Shona reveal a particularly close relationship with the Maravi people who migrated into Malawi, and the Tsonga occupy an intermediate linguistic position between the Shona and the Nguni and Sotho to the south (Nurse *et al.*, 1985). The Tsonga have recently received substantial Nguni increments probably initiated by a group of Nguni leaving "Zululand" in 1820-1 under the leadership of Shoshangane (Murdoch, 1959). Although this group was not very numerous, they were strong enough to conquer the lands of the Tsonga in Mozambique and introduce their own language and customs to such an extent that the Tsonga have also become known as the Shaangan or Shangana (Murdoch, 1959). The Tsonga have also been referred to as Thonga – the Zulu pronunciation (Junod, 1927). Despite the Nguni invasion, there are still profound cultural and linguistic differences between the Nguni and the Tsonga with the Tsonga making more closer contacts with the Sotho and Venda of South Africa (van Warmelo, 1974). The Venda are closely related to the Shona and, according to local traditions, it seems quite possible that the Venda were, in fact, Shona immigrants who established themselves over the original non-Venda inhabitants of what used to be the North-Eastern Transvaal (Parfitt, 1997).

Since no precise information appears to have been gleaned concerning the varying speeds at which language changes evolve, the absence of written records documenting the evolution of

language makes it difficult to reconstruct the prehistory of populations belonging to different language groups, for example, the Bantu-speakers described above. Linguistic studies can be used in conjunction with archaeological and genetic data in the attempt to reconstruct African prehistory.

1.2 Archaelogical and anthropological data

Iron technology and ceramic data has been used together with linguistic data to reconstruct the history of Negroid peoples in sub-Saharan Africa (Phillipson, 1993). There is, for example, a strong degree of similarity between archaelogical and linguistic evidence for the spread and development of the Bantu languages and their speakers. The widespread distribution of Early Iron Age (EIA) artifacts throughout sub-Saharan Africa and their different ceramic patterns established that EIA technology was introduced by Negroid peoples (Huffman, 1982). By incorporating cultural data with the archaeological record, like the attitudes about the political role of men, the spiritual role of ancestors, and the importance of cattle, Huffman (1982) found that the SEB could be distinguished from the SWB. He described the SEB settlement pattern as the Central Cattle Pattern and the SWB settlement pattern as the Forest Pattern with the major difference between the two being their spatial organization. The Central Cattle Pattern is characterized by a central cattle byre, containing grain storage pits and elite burials, around which a men's court and an outer arc of rondavel-type houses are arranged according to seniority. The Forest Pattern is identified, among other things, by mass cemeteries facilitating extended burial and rectangular houses usually

arranged in a rough rectangle or in parallel rows on opposite sides of a "street" (Lamen, 1953).

Archaeological data has contributed greatly in understanding the prehistory of the peopling of Zimbabwe. Although the Zimbabwe culture was not typical of the Central Cattle Pattern, several lines of evidence suggest that the Zimbabwe culture shared several traits with the Central Cattle Pattern (Huffman, 1982). It would appear that early Zimbabweans were active traders, and participated in ivory and gold trade with foreign traders along the East Coast, in exchange for glass beads. The accumulation of wealth and power soon replaced the importance of cattle in their culture and these changes can be observed from the archaeological record dating to about AD 1000 (Huffman, 1982). Such changes were first observed in Mapungubwe. By at least AD 1175, Mapungubwe was the capital of a state with several district centers. For unknown reasons, the town was abandoned about AD 1220 and the center of power shifted to Great Zimbabwe, 200 kilometers away (Huffman, 1982). The differences in spatial organization of the settlement pattern were retained at Great Zimbabwe and defined the Zimbabwe Culture (Huffman, 1982).

1.3 Reconstructing the prehistory of the Lemba

1.3.1 Historical and cultural evidence

The Lemba have a unique history in Africa. According to their oral history they claim "we came from Israelites, we came from Sena, we crossed the sea..." (Parfitt, 1997; p9). Their numbers were reduced considerably by the time they reached the coasts of Africa, where they rebuilt their city of Sena. From here the Lemba moved to a third "Sena", in present-day Mozambique (van Warmelo, 1974), where, again, the group suffered many losses. Although the Lemba, nowadays, do not know with any certainty the location of the original Sena, Parfitt (1997) believes it to be Sena in the eastern Hadramaut in the Yemen.

It is possible that some of the Lemba remained in Mozambique, as some Lemba-like groups are still thought to exist here (Margaret Nabarro, personal communicatin, in Parfitt, 1997), before they made their way to Zimbabwe. The Lemba claim to have been instrumental in developing and building the Great Zimbabwe Empire and it was here where the Lemba first encountered the Venda, a group closely related to the Shona. It is relevant to note at this point that in the Shona-speaking parts of Zimbabwe, the Lemba are referred to as Remba, as there is no 'l' sound in Shona. In the late 17th century, after a dispute over the succession of the King, the Lemba nation split into two main groups (Parfitt, 1997). The one group remained around Mberengwe, presumably the Remba, and the other moved south, leading

the Venda, to the Nzhelele valley in what is now Northern Province, South Africa. It was here that the Lemba claim to have helped build the impressive stone construction of Dzata.

The Lemba claim to have spoken an ancestral language "Hiberu" (Parfitt, 1997). The Lemba did not retain this original language and few, nowadays, can even remember any concrete examples of it. Instead they began to adopt the languages of the people among whom they lived, initially that of the Mbire-Shoko people at the "third Sena" in present-day Mozambique (van Warmelo, 1974). Today the Lemba of the Northern Province speak Venda and in Zimbabwe, the Remba speak dialects of Shona. It is estimated that there are currently about 40 000 Lemba living in South Africa, now, and about 15 000 distributed throughout other southern and central African countries (Mathiva, personnal communication, in Parfitt, 1997).

Although the Lemba have lived closely among the Shona and the Venda, they display certain peculiar characteristics that set them apart from their linguistic neighbours (Mourant *et al.*, 1978) and show similarities to those of the Semitic peoples (Mathiva, 1992). They practice a ritualistic religion. Marriage laws are strict and strongly encourage endogamy – outsiders have to undergo ritual purification before admission into the family. Their food laws appear to be essentially Jewish: they do not eat any meat which has not been killed in a 'kosher' manner or prepared by a Lemba, the meat of all ruminants with non-cloven hoofs is avoided, milk and meat are eaten separately. Some observers claim to recognize some Caucasoid features, prominent noses and fairer skins, in the Lemba (Mourant *et al.*, 1978).

According to van Warmelo (1974), the Lemba ancestors were highly skilled in metalwork, pottery, textiles and shipbuilding. They left their hometown of "Sena" and came to southern Africa for the main purpose of trading their goods, especially for gold. They established several trading posts along the eastern African coast and further inland. Although they became well known by the locals throughout their travels, the Lemba remained an elite group, considering themselves superior. It was only when they heard that they could never return home, as their town had been overpowered by the enemy, that they began taking local wives, and the different trading posts marked the establishment of the clans identified today.

The Lemba are composed of over 12 clans, some of which appear to correlate with place names in the Hadramaut (Parfitt, 1997) and several of which have names that are Arabic in origin like Bakali, Haji, Hamisi, Hasane, Sadiki and Salifo (van Warmelo, 1974, p82). The Buba clan is recognized as being the senior clan, the oldest, and for some ritual purposes, the most important (Thomas *et al.*, 2000). The Tovakari clan members were supposedly the main builders of Great Zimbabwe (Parfitt, 1997).

1.3.2 Theories concerning Lemba origins

There appears to be general consensus among anthropologists, historians and geneticists that the Lemba do indeed have Semitic elements in their ancestry (Schapera, 1946; van Warmelo, 1974; Spurdle and Jenkins, 1996; Parfitt, 1997; Thomas *et al.*, 2000). However, differences of opinion haven arisen as to whether they are Arabic or Jewish. There are three predominant

schools of thought concerning the specific nature of the Semitic culture involved in the origin of the Lemba (Connoway, 1976): (i) a pre-Islamic-Judaic-Arab origin, (ii) an Islamic-Arabic origin and (iii) a Falasha-Abyssinic origin.

The pre-Islamic-Judaic-Arabic theory of origin is firmly believed by Gayre of Gayre (1972), who maintains that the Lemba's history can specifically be traced back to the pre-Islamic Saba (or Sheba) in Yemen. According to Gayre of Gayre (1972), Lemba tradition confirms conclusively that the only time Judaised Arabs could have settled in east Africa and present-day Zimbabwe was around the sixth century AD, before Arabs were Christian, pagan, or even Muslim. In addition, he maintains that the Jewish-like cultural practises of the Lemba, for example circumcision and their kosher dietary laws, would be an extremely unlikely coincidence. This hypothesis is further supported by the absence of any trace of "Mohammedanism" in the Lemba's religious ceremonies (Jaques, 1931, pp248-250). However, since this theory appears to be upheld largely by Lemba tradition: their oral history and cultural practises, it has received some criticism. Although the pre-Islamic-Judaic-Arabic theory of Lemba origin has not been entirely refuted, many believe that it is supported by insufficient evidence (Connoway, 1978; Parfitt, 1997; Hendrickx, 1991).

The second school of thought claims that the Lemba have an Islamic-Arabic origin and are convinced that they are descended from Arab traders who took wives from local population groups with whom they traded (Stayt, 1931). Much of this theory is supported by linguistic terms and clan names used by the Lemba that appear to be derived from Arabic place names,

surnames (Bullock, 1927) or the Islamic faith (Mandivenga, 1989). It has been argued, however, that such terminology is likely to emphasise the Yemenite, as opposed to Arabic, origin of the Lemba (Gayre of Gayre, 1972).

A third possible option is that of a Falasha-Abyssinic origin of the Lemba. Thompson (1942) observed many ritual similarities between the Lemba and the Falasha, the "Black Jews" of Abyssinia (Ethiopia) and, based on the style of architecture used by the Lemba, believe that they could not be Arabic in descent. It has also been suggested that the Lemba descended from three originally distinct population groups: the numerically stronger Zambezi Sena, who had been influenced by ethnic ties with the Arabs, a much smaller group of Falashas, who carried Old Testament traditions and rituals with them to the southern interior, and, later, a third group "from the east" (Von Sicard, 1952,p149) arriving at the Sabi River before the end of the 16th century (Von Sicard, 1962). This proposal, in effect, implies a combined Arabic/Jewish origin of the Lemba.

There is clearly no consensus concerning Lemba origins based on historical and traditional evidence. Genetic studies have attempted to refine some of these theories.

1.3.3 Genetic evidence

The Lemba's remarkable story brought them to the attention of a wide audience, particularly geneticists. Y chromosome studies on a group of Lemba from Venda revealed that about

50% of their Y chromosomes were of Semitic origin (Spurdle and Jenkins, 1996). These studies could not distinguish between Jewish and / or Arabic Y chromosomes, and consequently could not infer whether Jewish or Arabic populations contributed to the Y chromosomes of the Lemba. On the other hand, serogenetic (Mourant *et al.*, 1978) and mitochondrial DNA (mtDNA) restriction fragment length polymorphism (RFLP) studies (Soodyall *et al.*, 1996) revealed that the Lemba were indistinguishable from other SEB speaking populations. Preliminary Y chromosome studies on the Zimbabwean Remba indicated that about 70% of their Y chromosomes are derived from Semitic origins (Soodyall, unpublished).

The proposed relationship between the Ethiopian Jews (the Falasha) and the Lemba has not been supported by genetic studies. Data from various classical, nuclear DNA and mtDNA genetic studies (Mourant *et al.*, 1978; Zoossmann-Diskin *et al.*, 1991) indicated that the Falasha are very similar to other Ethiopian populations, exhibiting a genetic profile that represents admixture between African and Caucosoid (probably Mediterranean) groups. However, Y chromosome studies on a small sample of Ethiopian Jews (Ritte *et al.*, 1993) revealed that the Falasha possess specific haplotypes which are rare or absent in both Negroids *and* Caucosoids (Torroni *et al.*,1990; Spurdle and Jenkins, 1992), and one of which is found at notable frequencies only in Khoisan populations (Spurdle and Jenkins, 1992). It is possible that the study on the Ethiopian Jews was subject to sampling error due to the small sample size or that extreme genetic drift has occurred in this largely endogamous group.

Even so, the available Y chromosome genetic data do not support a close genetic relationship between the Falasha and the Lemba.

Recently, Thomas *et al.* (2000) found a particular Y chromosome haplotype, referred to as the Cohen Model Haplotype (CMH), at a frequency of 8.8% in the Lemba from South Africa. Furthermore, this haplotype was most common in one particular Lemba clan, namely the Buba (seven of the 13 Lemba with the CMH were from this clan). The CMH, dated to be approximately 2000 – 3200 years old, has been suggested by Thomas *et al.* (1998) to be a potential signature haplotype of Judaic origin. The CMH is thought to be characteristic of the paternally inherited Jewish priesthood since it has been found at higher frequencies in Ashkenazi and Sephardic Cohanim (Jewish priests) compared to non-Cohanim Jews (Thomas *et al.*, 1998).

More recently, Hammer *et al.* (2000) and Nebel *et al.* (2000) investigated the genetic affinities of Jewish and Middle Eastern non-Jewish populations. The CMH was the most frequent haplotype in Ashkenazi and Sephardic Jews but was found to occur only rarely in Middle Eastern non-Jewish populations (Hammer *et al.*, 2000; Nebel *et al.*, 2000). Even though the two most frequent haplotypes in Israeli and Palestinian Arab population groups were closely related to the CMH, they were found to occur very rarely in Ashkenazi and Sephardic Jews (Nebel *et al.*, 2000). Although each group used different Y chromosome markers for haplotype construction, their results were similar showing that despite some significant differences between the Jewish and Arab populations sampled in each study, their

Y chromosome haplotypes were almost indistinguishable (Hammer *et al.*, 2000; Nebel *et al.*, 2000). These findings support previous studies that suggested a common origin for Jewish and non-Jewish populations living in the Middle East (Santachiara-Benerecetti *et al.*, 1993; Peretz *et al.*, 1997).

Although Thomas *et al.* (2000) concluded that the Lemba have some Jewish ancestry, their origins are still not unequivocally resolved. Considering the common origin of the Jewish and Middle Eastern Arabic populations and their close genetic affinity, it may be more difficult, than previously thought, to determine accurately if the Lemba have specifically Jewish origins. If the presence of the CMH in the Lemba is in fact due to ancient links with Jewish ancestors, we would expect (i) the CMH to be found in the Lemba from Zimbabwe (the Remba) as well as all other Lemba communities in southern Africa and (ii) several other haplotypes derived from the CMH to be present in the Lemba. It is possible that the CMH was introduced into the Lemba only over the past one hundred years. It is known that Ashkenazi Jews have lived in Louis Trichardt, the main town in the area of the South African Lemba, since it was founded in 1898, and appear to have been part of a flourishing community held together by a synagogue in the town's centre (Figure 2). It is feasible that the presence of the CMH in the South African Lemba, sampled by Thomas *et al.* (2000), is a result of a more recent Jewish genetic contribution.

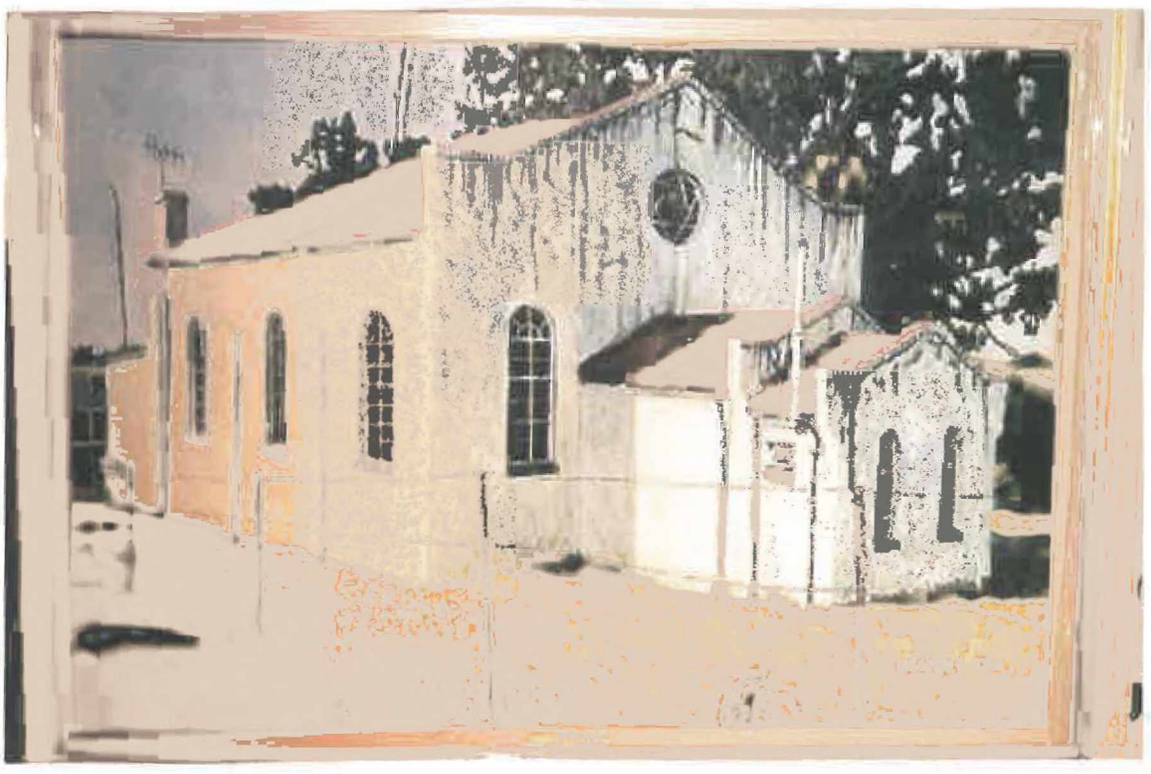


Figure 2. A painting of the Hebrew congregation's Krogh Street synagogue which was built in 1946-7. The building was demolished at the start of 1998 by Old Mutual who had bought the site for the purpose of constructing office buildings. As a gesture Old Mutual commissioned the painting and presented it to the Historical Society of Louis Trichardt.

The first part of this study makes use of mtDNA sequence data and Y chromosome haplotypes to examine the origin of females and male genes, respectively, in the Lemba. These data used in conjunction with historical data ought to shed more light on the prehistory of the Lemba in southern Africa.

1.4 The peopling of Mozambique

Before the arrival of Portuguese explorers in the late 15th century, present-day Mozambique was occupied by Bantu-speaking groups who, like those in Zimbabwe, also belonged to the SEB language family. Four major linguistic groups occupied the area from north to south: the Yao, Maravi, Shona and Shangana-Thonga. No significant intermixing has occurred between the groups mainly because two major rivers, the Zambezi and the Save, served as natural barriers between them (Reys *et al.*, 1970). In the recent past, the Shona and Shangana-Thonga groups have admixed considerably with the Nguni. The Nguni invaded and established a conquest state in Shona and Tsonga territory during the early 19th century (Murdoch, 1959; van Warmelo, 1974). Portuguese and Arabic influences are also found, particularly north of the Zambezi, in coastal tribes. These influences are much smaller in the Shona and Shangana-Thonga groups (Reys *et al.*, 1970). The Mozambican individuals examined herein are from a rural area near Inhambane. They belong to the Shangana-Thonga linguistic group that has subdivisions: Chopi (Tshopi), Hlengwe, Ronga (Baronga), Thonga (Bathonga, Tonga) and Shangana.

1.4.1 Genetic studies in populations from Mozambique

The Mozambican gene pool might reflect contributions from other populations particularly due to its geographical position, linking landlocked countries to coastal ports during the era of extensive slave trade. There have been limited genetic studies in populations from Mozambique. Reys *et al.* (1970) examined variants of glucose-6-phosphate dehydrogenase (G6PD) in the Ronga, Tsonga or Thonga, Shangana and Chopi from southern Mozambique. Although all of these groups belong to the Shangana-Thonga linguistic sub-family, the Chopi are thought, by some, to be more affiliated with the Shona group (Velez-Grilo, 1960). Since the distribution and frequencies of G6PD variants, alleles Gd^{A} and Gd^{A-} were largely unknown in other African populations, Reys *et al.* (1970) found their data difficult to interpret. They did, however, find that the frequencies of the Gd^{A} and Gd^{A-} alleles were higher in the Chopi compared with the Ronga, Thonga and Shangana. Using plasma cholinesterase as a marker, Whittaker and Reys (1975) found that the frequency of the variant E_1^{f} was twice as high in the Chopi compared to the Ronga, Thonga and Shangana. In addition, they found the frequencies of the E_1^{a} and E_1^{f} alleles in their Portuguese sample were very different from those in the SEB groups from Mozambique suggesting that it is unlikely that the Mozambican gene pool received much contribution from the Portuguese during the days of trading and slavery.

An understanding of the genetic affiliation of Mozambican people with their neighbouring countries, including other sub-Saharan African populations would contribute in achieving the

objectives of other projects underway in the Department of Human Genetics. Given the proximity of Mozambique to Madagascar, it is possible that some of the ancestors of the Malagasy originated in Mozambique. By examining the mtDNA and Y chromosome gene pools of Mozambican populations, it would be possible to determine to what extent males and females from Mozambique could have contributed to the gene pool of the Malagasy. Also, it would be possible to examine if the proposed association of the Lemba in Mozambique, at the Lemba's "third Sena" (van Warmelo, 1974), contributed in any way to the gene pool of other populations from Mozambique. An understanding of the genetic structure of Mozambican populations would also shed more light on migrations associated with the Bantu expansion in southern Africa.

1.5 Project objectives

The main objective of this study is to make use of mtDNA and the non-recombining portion of the Y chromosome to examine the contribution of females and males, respectively, to the gene pool of present-day Mozambican populations and a Lemba group from Zimbabwe (the Remba).

There are still many questions that are unresolved concerning Lemba origins. Thus far, genetic studies have been conducted only on the Lemba from South Africa. In the present study we examine a Lemba group from Zimbabwe (the Remba). By comparing Y

chromosome data from the Lemba and the Remba, we hope to address the following questions:

1. What are the origins of Y chromosomes in the Lemba and Remba?
2. What is the frequency of the CMH in the two groups?
3. Is there evidence to support (or refute) the claims made by Thomas *et al.* (2000) of a Jewish ancestry of some Y chromosome found in the Lemba?

MtDNA studies in the Lemba and Remba would elucidate the origins of female ancestors of the Lemba. Using both mtDNA and Y chromosome data, it would be possible to test theories concerning Lemba prehistory and origins outlined in sections 1.3.1 and 1.3.2.

Some of the questions we wish to address from genetic studies on Mozambican populations are:

1. What are the origins of the female and male contributions to the gene pool of Mozambique populations?
2. What are the genetic relationships among the Mozambican population, their neighbouring countries and other sub-Saharan African populations?
3. Is there any genetic evidence to corroborate the claims of the Lemba based on their oral history, discussed in section 1.3.1, of a transient phase in Mozambique?
4. Is there a genetic trail from Mozambique to Madagascar?
5. To what extent does the gene pool of the Mozambican population reflect historical events?

1.6 Genetic markers used in this study

1.6.1 MtDNA studies

MtDNA has been used widely for studies on human evolution and in reconstructing population histories. The mtDNA molecule has several unique features that make it popular for such studies. First, mtDNA is transmitted exclusively by females (Giles *et al.*, 1980), and does not recombine (Gyllensten *et al.*, 1991). Since differences among mtDNA types are due to mutations only, mtDNA lineages can be traced back to a common origin. Second, mtDNA is present in high copy number in human cells. The average somatic cell has just two copies of any given nuclear gene, but hundreds to thousands of copies of mtDNA (Robin and Wong, 1988). This property of mtDNA, along with its extra-nuclear cytoplasmic location, makes it easier to obtain for analysis, and also makes it the molecule of choice for analyses of ancient DNA. Third, mtDNA has been completely sequenced and is well characterized in many species, including humans (Anderson *et al.*, 1981). The final property of mtDNA of evolutionary interest is that it has a high sequence evolution rate, five to 15 times faster than the average segment of nuclear DNA (Brown *et al.*, 1979). This allows useful comparisons to be made among closely related species or populations of the same species, as a sufficient number of mutations will have accumulated over even a short evolutionary time-span.

Certain regions of the mtDNA molecule have been used extensively for population and evolutionary studies. One such region is the 1122 bp control region (Anderson *et al.*, 1981).

Unlike the remainder of the mitochondrial genome which is tightly packed with genes encoding 13 polypeptides, 22 transfer RNAs and two ribosomal RNAs, the control region is non-coding. While no gene function has been assigned to this region, it does contain the origins of transcription and replication, promotors, transcription factor binding sites as well as conserved sequence blocks associated with the start of DNA synthesis (Strachan and Read, 1996).

Although the best resolution for mtDNA comparative population studies would probably be achieved by sequencing the entire mtDNA genome, this would not be time- or cost-efficient. Instead, restriction fragment length polymorphisms (RFLPs) have been used to examine variation and now, commonly, sequencing of the control region. Variation within the control region is concentrated in two regions that have been referred to as hypervariable region I and II (HVRI and II) or CR1 and CR2, consisting of 405 bp and 378 bp, respectively (Anderson *et al.*, 1981).

Analysis based on high (Chen *et al.*, 1995) and low (Graven *et al.*, 1995; Soodyall and Jenkins, 1992, 1993) resolution RFLP mapping have shown that the *Hpa*I site at np 3592, which defines haplogroup L (Chen *et al.*, 1995), is found almost exclusively in Africa. Additional polymorphisms defined by *Hinf*I site gains at np 10806 and 16389 resolved haplogroup L into subhaplogroups L1 and L2 (Chen *et al.*, 1995). More recently, Watson *et al.* (1997) extended the haplogroup designation derived from RFLP studies to incorporate mutations detected by control region DNA sequencing. They resolved haplogroup L further

into subhaplogroups, L1i, L1a, L1b, L2, L3a, L3b and L3c, which have varying frequencies and distribution patterns in different African populations (Chen *et al.*, 1995; Watson *et al.*, 1997). Therefore, mtDNA control region sequences in conjunction with the RFLP haplogroups ought to be useful in examining the structure and affinities of the mtDNA gene pools in the Remba and the Mozambicans.

1.6.2 The 9-bp deletion

Another part of the mtDNA molecule that is useful in population genetics is the intergenic region between the cytochrome oxidase II and transfer RNA for lysine genes. Most humans have two copies of a 9-bp tandem repeat (Anderson *et al.*, 1981). A deletion of one copy of the 9-bp repeat sequence was found at varying frequencies in Asian populations and initially thought to be an Asian-specific marker (Hertzberg *et al.*, 1989). The deletion is not associated with any pathological condition and has subsequently been found at varying frequencies in sub-Saharan African populations, more commonly in central and SEB speakers (Soodyall *et al.*, 1996). The deletion is associated with specific mtDNA types, defined by mutations in the hypervariable control region sequence, which can be resolved by DNA sequencing (Soodyall *et al.*, 1996). Phylogenetic and mismatch distribution analysis of these mtDNA sequences suggest that the 9-bp deletion arose independently in Asia and Africa, with two to three separate origins in sub-Saharan Africa (Soodyall *et al.*, 1996). Thus, the distribution of the 9-bp deletion combined with the corresponding mtDNA control region sequences is informative for tracing population history.

Since mtDNA is a haploid marker and represents a single genetic locus it is useful for interpreting the evolutionary history of genomes contributed by females only, and on its own, it is not very useful in establishing population relationships. Y chromosome data would supplement mtDNA studies and shed more light on the contribution made by males.

1.6.3 Y chromosome studies

Polymorphisms on the non-recombining portion of the Y chromosome can be used to study population history from the perspective of male lineages. Except for the pseudoautosomal region, it is inherited paternally and does not recombine (Ellis, 1991). The non-recombining portion of the Y chromosome, therefore, preserves a unique record of mutational events that occurred exclusively along male lineages from a single ancestral point.

The recent increase in knowledge of the number of highly informative polymorphisms on the NRPY includes variation at short tandem repeat (STR) loci (Roewer *et al.*, 1992; Santos *et al.*, 1993, 1996; de Knijff *et al.*, 1997), an *Alu* insertion polymorphism (Spurdle *et al.*, 1994; Hammer, 1994), single nucleotide polymorphisms (Seielstad *et al.*, 1994; Underhill *et al.*, 1996, 1997; Hammer *et al.*, 1997) and alphoid repeat polymorphisms (Santos *et al.*, 1995). Many of these have been used together to construct compound haplotypes, some of which are useful in tracing population migrations and histories (Hammer and Horai, 1995; Cooper *et al.*, 1996; Roewer *et al.*, 1996; Santos *et al.*, 1996b; Spurdle and Jenkins 1996; Underhill *et al.*, 1996; Hammer *et al.*, 1997, 1998; Scozzari *et al.*, 1997). The use of Y chromosome

microsatellite haplotypes to infer maximum parsimony evolutionary connections (Cooper *et al.*, 1996; Roewer *et al.*, 1996) was ineffective because of the extremely high number of equally probable tree permutations. Consequently, the use of more stable Y chromosome polymorphisms like the *Alu* insertion/deletion polymorphisms in combination with Y-STRs have been recommended for population and evolutionary studies (Deka *et al.*, 1996; de Knijff *et al.*, 1997; Jobling *et al.*, 1997).

1.6.3.1 The Y *Alu* polymorphism (YAP)

The YAP results from the recent insertion of an *Alu* element on the long arm of the Y chromosome at Yq11. *Alu* elements, which are not present in all individuals, are thought to be stable genetic markers. Each insertion event at a specific locus is said to be unique and likely to have occurred only once within the human population (Batzer *et al.*, 1991). This appears to be true for the YAP insertion, as well, which most likely occurred after the divergence of humans and modern apes yet sometime before the divergence of the major African groups (Hammer and Horai, 1995). Hammer *et al.* (1997) raised the possibility that the YAP originated in Asia and was only subsequently carried to Africa. This hypothesis was supported by the finding of high frequencies of the most ancestral YAP+ haplotype in some Asian populations and by the observation of higher levels of diversity associated with Asian versus African chromosomes with this ancestral haplotype (Hammer *et al.*, 1997). Using Templeton's (1993) nested cladistic analysis, Hammer *et al.* (1998) yielded results that,

again, implied an Asian origin for the ancestral YAP+ haplotype. They estimated the age of the YAP insertional event to be 55 000 ± 19 000 years (Hammer *et al.*, 1998).

The YAP is a useful marker in population and evolutionary genetics as its frequency varies considerably among human populations (Hammer and Horai, 1995). The frequency of the YAP in sub-Saharan Africa is about tenfold higher than in Europe, Oceania or most regions of the Far East and reaches its highest frequency in Bantu-speakers and west Africans (78-86%). The frequency is moderately high in Pygmies (66%) and is found at varying frequencies in different Khoisan groups (Hammer, 1994; Spurdle *et al.*, 1994).

1.6.3.2 Y-STR loci

Short tandem repeat loci (including several tetranucleotide repeat microsatellites) exhibit considerable variation and are useful genetic markers in reconstructing Y chromosome evolution. Y chromosome STRs evolve at higher mutation rates than nucleotide changes, and they mutate in a stepwise fashion (Weber and Wong, 1993). Although they are not especially informative on their own, allele frequencies at specific Y-STR loci do vary among human populations (Santos *et al.*, 1996; de Knijff *et al.*, 1997). The YAP in conjunction with Y-STRs can be used to derive Y chromosome haplotypes. Such haplotypes have proven to be especially valuable in revealing genetic differences between closely related populations whose time of divergence has been relatively short (Roewer *et al.*, 1996).

Being haploid systems, both mtDNA and Y chromosome markers are limiting on their own in that the data derived from them tends to be more reflective of the history of the molecule or the chromosome than that of the whole population. It has been shown, however, that mtDNA data in conjunction with Y chromosome data are extremely valuable in understanding how females and males have contributed to the gene pools of populations and hence, in reconstructing population history (Perez-Lezaun *et al.*, 1999).

2. SUBJECTS AND METHODS

2.1 Samples

Blood samples were collected from healthy, unrelated adult volunteers with their informed consent and University Ethics approval for research on human subjects (Himla Soodyall, Protocol Number M980553). Copies of the ethics clearance certificates for the work carried out in this study are given in the Appendix (section 7.3). The age, gender, place of birth and ethnicity of each subject was noted, as well as relationship (if any) to any other individuals in the sample groups. The Mozambican samples were collected at the Chicuque Clinic in Inhambane with the facilitation of Professor Leopoldo da Costa, Dean of Research, Univeridade Eduardo Mondlane. The individuals sampled classified themselves as mainly Chopi, Ronga, Thonga (or Bathonga) and Shangane. It is feasible for the Mozambican sample to comprise these different groups as each group is a subdivision only of one broader linguistic group of SEB speakers, the Shangana-Thonga. The Zimbabwean samples were collected from individuals who all identified themselves as Remba. Figure 3 shows a map illustrating the approximate geographic location of the major population groups studied.

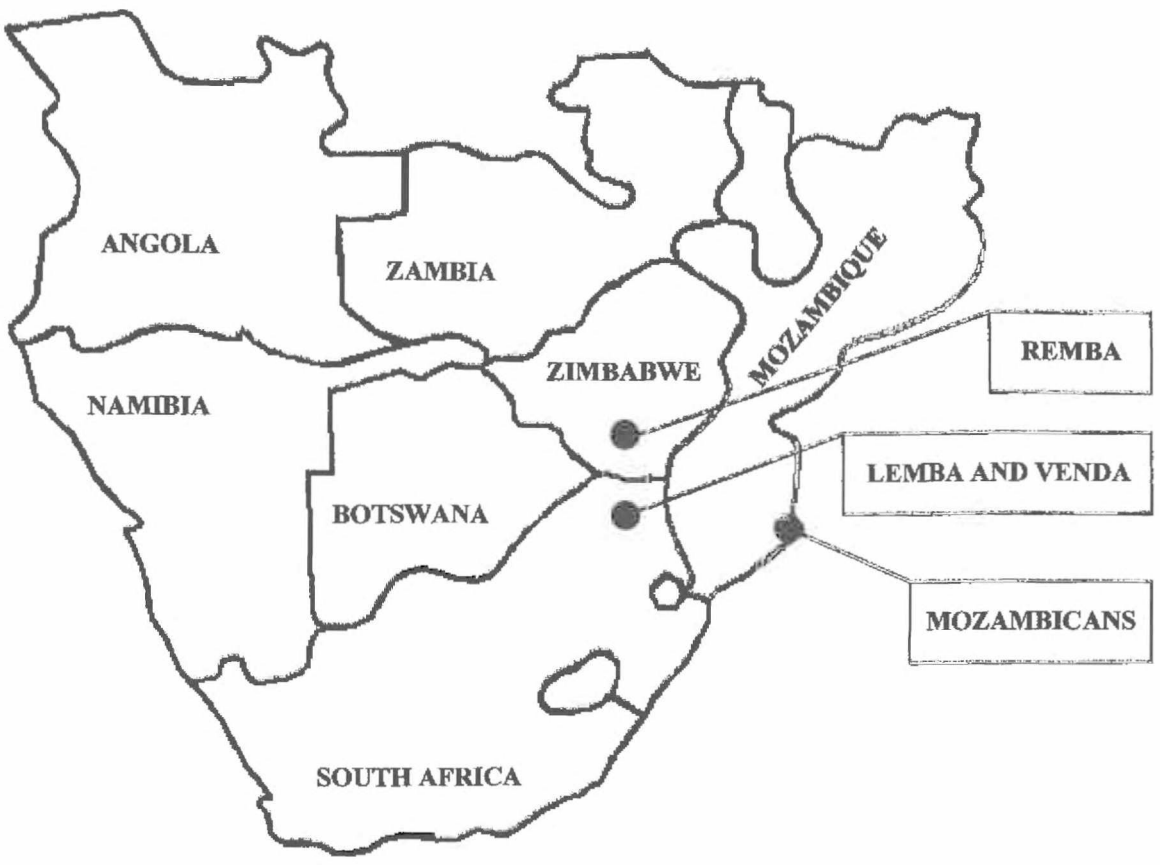


Figure 3. Map showing the approximate geographic locations of the population groups examined in this study.

2.2 DNA extraction

DNA had already been extracted from most of the samples. The salting out method (Miller *et al.*, 1988) was used to extract DNA from stored buffy coats from additional samples.

2.3 Molecular methods

2.3.1 MtDNA studies

2.3.1.1 Screening for the 9-bp deletion

The COII/$tRNA^{Lys}$ intergenic region was amplified using the polymerase chain reaction (PCR) with primers L8215 and H8297 (Table 1). The final reaction mix contained 1X *Taq* DNA polymerase buffer (Boehringer Mannheim), 1M BSA, 0.1M dNTPs, 0.4mM of each primer, 1U *Taq* DNA polymerase and approximately 10 ng of DNA, made up to 50 µl with ddH_2O. Each PCR cycle consisted of denaturing at 94°C for 1 minute, annealing at 56°C for 1 minute and extension at 74°C for 1 minute for 30 cycles (Hertzberg *et al.*, 1989). The products were resolved by gel electrophoresis using 2% Metaphor agarose gels (~10V/cm for 2 hours in 1XTBE buffer) and sized with a 1 kb ladder (Gibco BRL). Positive and negative controls were included in each gel run. Fragment sizes of 112 bp and 121 bp correspond to samples with and without the deletion, respectively. Altogether, 71 Mozambican individuals and 94 Zimbabwean individuals were screened for the 9-bp deletion.

2.3.1.2 MtDNA sequencing

Primers L15996 and H408 (Table 1) were used for PCR amplification of the 1.1 kb mtDNA non-coding control region using the method initially described by Vigilant (1990). The PCR reaction mix and conditions were the same as those used for the amplification of the COII/tRNA^{Lys} intergenic region, described above. The amplified PCR products were resolved by electrophoresis on 2% NuSieve agarose gels in 1XTAE buffer (~100V for 1.5 hours), excised, and purified using Wizard columns (Promega, Whitehead Scientific) according to the manufacturer's recommendation.

Both hypervariable regions of the control region (CR1 = 405 bp; CR2 = 378 bp) were sequenced using the AmpliTaq FS Cycle Sequencing kit (Perkin Elmer). Depending on the strength of the DNA template, between 3 - 8 µl of DNA template and 3.3 µM of primer was used for cycle sequencing, carried out for 25 cycles of 96°C for 30 seconds, 15°C for 15 seconds and 60°C for 4 minutes, each. The L-strand was sequenced using primer L15996 for CR1 and L29 for CR2, while the H-strand was sequenced using H16401 for CR1 and H408 for CR2 (Table 1).

Cycle sequenced PCR products were purified using Centricep columns (Separations) according to the supplier's recommendation and the products were dried in a (Sorval) vacuum dryer for approximately one hour. Purified PCR products were resuspended in 2.5 µl dextran/formamide dye (recipe given in section 7.1), denatured by heating to 90°C for 2

minutes, cooled immediately on ice, and then loaded on a 4.3% polyacrylamide gel. Electrophoresis was carried out in an automated DNA sequencer (Model ABI 377) for 7 hours with 36cm gel plates (3.00 kV, 60mA, 52°C). The mtDNA sequences were analysed using the ABI 377 sequencing software.

Table 1. Primers used for mtDNA PCR and sequencing

Primer	Sequence	Reference
L8215	5'ACA GTT TCA TGC CCA TCG TC3'	Vigilant 1990
H8297	5'ATG CTA AGT TAG CTT TAC AG3'	Vigilant 1990
L15996	5'CTC CAC CAT TAG CAC CCA AGC3'	Hertzberg *et al.*, 1989
H408	5'CTG TTA AAA GTG CAT ACC GCC A3'	Hertzberg *et al.*, 1989
L29	5'GGT CTA TCA CCC TCT TAA CCA C3'	Hertzberg *et al.*, 1989
H16401	5'TGA TTT CAC GGA GGA TGG TG3'	Hertzberg *et al.*, 1989

2.3.2 Y chromosome markers

2.3.2.1 Screening for the YAP

The YAP locus, *DYS287*, was amplified by PCR using primers YAP 1 and YAP 2 (Table 2). The reaction mix contained 1X *Taq* DNA polymerase buffer (Boehringer Mannheim), 1M BSA, 0.1M dNTPs, 0.4mM of each primer, 1U *Taq* DNA polymerase and approximately 10 ng of DNA, made up to 50 µl with ddH_2O. PCR conditions were 94°C for 2 minutes, followed by 30 cycles of: 94°C for 1 minute, 51°C for 1 minute and 72°C for 1 minute (Hammer and Horai, 1995). The PCR products, including positive and negative controls, were resolved on 2% agarose gel electrophoresis in 1XTBE buffer (30-40 min at ~10V/cm). Bands were sized with a 1 kb ladder (Gibco BRL). Fragment sizes of 455 bp correspond to the presence of the *Alu* insertion (designated "YAP+") while fragments of 150 bp in size correspond to the absence of the insertion (referred to as "YAP-"). Forty males from Mozambique were screened for the presence of the YAP element. The Remba sample used in the comparative analysis was previously screened by B. Morar (personal communication).

2.3.2.2 Typing of Y chromosome short tandem repeat (STR) loci

Six Y chromosome STR markers were used in conjunction with the YAP element to examine variation on the Y chromosome. Loci *DYS393, DYS19, DYS390* and *DYS391*, and *DYS388* and *DYS392* were amplified, using fluorescently labeled primers (Table 2), in two multiplex

PCR reactions, respectively. The components of both multiplex reaction mixes are given in Table 3.

PCR conditions for the first multiplex reaction were 95°C for 2 minutes and 28 cycles of 95°C for 30 seconds, 54°C for 30 seconds and 72°C for 30 seconds (Redd *et al.*, 1997). PCR conditions for the second multiplex reaction were modified from Kayser *et al.*(1997) to the same as those for the first multiplex reaction, except for a change in annealing temperature from 54°C to 56°C.

PCR products were resolved on a 4.3% denaturing polyacrylamide gel in an ABI 377 DNA sequencer using 12 cm gel plates (3.00kV, 60mA, 52°C). One µl PCR product plus 0.3µl ABI ROX-350 size standard (see Appendix, section 7.2.2) were loaded in each lane. The software Genescan and Genotyper were used to analyze the data and assign fragment sizes.

Table 2. Primers used for YAP and Y-STR PCR

Locus	Primer	Flourescent Marker	Sequence	Reference
DYS287	YAP 1 YAP 2		5'CAG GGG AAG ATA AAG AAA TA3' 5'ACT GCT AAA AGG GGA TGG AT3'	Hammer and Horai, 1995
DYS393	DYS393-A DYS393-B	ABI-HEX (green)	5'GTG GTC TTC TAC TTG TGT CAA TAC3' 5'AAC TCA AGT CCA AAA AAT GAG G3'	Jobling and Tyler-Smith, 1995
DYS19	DYS19-A DYS19-B	ABI-FAM (blue)	5'CTA CTG AGT TTC TGT TAT AGT3' 5'ATG GCA TGT AGT GAG GAC A3'	Roewer *et al.*,1992
DYS390	DYS390-A DYS390-B	ABI-HEX (green)	5'TAT ATT TTA CAC ATT TTT GGG CC3' 5'TGA CAG TAA AAT GAA CAC ATT GC3'	Jobling and Tyler-Smith, 1995
DYS391	DYS391-A DYS391-B	ABI-HEX (green)	5'CTA TTC ATT CAA TCA TAC ACC CA3' 5'GAT TCT TTG TGG TGG GTC TG3'	Jobling and Tyler-Smith, 1995
DYS388	DYS388-A DYS388-B	ABI-FAM (blue)	5'GTG AGT TAG CCG TTT AGC GA3' 5'CAG ATC GCA ACC ACT GCG3'	Kayser *et al.*, 1997
DYS392	DYS392-A DYS392-B	ABI-Tamra (yellow)	5'TCA TTA ATC TAG CTT TTA AAA ACA A3' 5'AGA CCC AGT TGA TGC AAT GT3'	Kayser *et al.*, 1997

Table 3. Y-STR reaction mixes

Ingredient	Y-STR multiplex 1		Y-STR multiplex 2	
	Volume used (µl)	**Final concentration**	**Volume used (µl)**	**Final concentration**
10X buffer[1]	2.5	1X	2.5	1X
dNTPs (2.5mM)	2.5	0.25mM	2.5	0.25mM
Primers (10µM)[2]	1 each (8)	0.4µM	0.5 each (4)	0.2µM
Spermadine (2.5mM)	2.5	0.25mM	2.5	0.25mM
$MgCl_2$ (25mM)	1.5	1.5mM	1	1mM
Taq (5U/µl)[1]	0.1	0.2U/µl	0.1	0.2U/µl
ddH_20	6.9		13.4	
DNA	1		1	
Total	**25**		**25**	

[1]PE and BM *Taq* DNA polymerase and buffer were used for multiplex 1 and 2, respectively.
[2]The eight primers used for multiplex 1 are: DYS393-A and –B, DYS19-A and –B, DYS390-A and –B and DYS391-A and –B. The four primers used for multiplex 2 are: DYS388-A and –B and DYS392-A and –B.

2.4 Data analysis

2.4.1 MtDNA data analysis

MtDNA sequence data were analysed in two ways. In one analysis, both CR1 and CR2 sequences were analysed together and in the other analysis only CR1 data were used to facilitate comparative analysis with published data for which only CR1 information is available. CR1 and CR2 sequence data were obtained from 67 Remba and 61 Mozambican individuals while CR1 sequences only were obtained from 79 Remba and 62 individuals from Mozambique. These were analyzed in conjunction with the South African Lemba (n = 44) and the Venda (n = 39) (H. Soodyall, unpublished). Individual sequences were compared to the published reference sequence (Anderson *et al.*, 1981) using the GELIN programme developed by Steve Sherry (1991) and aligned manually. After the removal of insertions and deletions and missing data, subsequent analysis was performed using 770 bp of sequence data (nucleotide position (np) 15997 to 16389 for CR1 and 31 to 407 for CR2).

Sequence data were analysed following pairwise comparisons of sequences facilitated by use of the options in MEGA (S. Kumar, personal communication; http//www.megasoftware.net) and MEBOOT (S. Kumar, personal communication). An assessment of the number of mtDNA types found in each group was obtained after the removal of shared mtDNA types. The neighbor-joining (NJ) method (Saitou and Nei, 1987) was used to generate gene trees using the p-distance which measures the proportion of nucleotide sites at which any two

sequences being compared are different. Some gene trees were rooted using the Neanderthal CR1 (Krings *et al.*, 1997) and CR2 (Krings *et al.*, 1999) sequences to give some indication of the evolutionary relationship of mtDNA types. The bootstrap test based on 1000 replications was applied to establish confidence on the topology of the NJ trees (Felsenstein, 1985). The variant sites (i.e. nucleotide substitutions relative to the published reference sequence, Anderson *et al.*, 1981) were compared to those associated with RFLP haplogroups defined by Watson *et al.* (1997). Unique mtDNA types derived from the present study were numbered and grouped according to the nomenclature adopted by Watson *et al.* (1997).

The time of coalescence in years, t, of mtDNA types that corresponded to different RFLP subhaplogroups was estimated by use of the equation $d = 2\lambda t$ (Nei, 1987). The mean pairwise sequence divergence within each subhaplogroup, d, was calculated with the SENDBS programme (N. Takezaki, personal communication; http//www.cib.nig.ac.jp/dda/ntakezak.html#sendbs) using the Gamma distance (Tamura and Nei, 1993) with a distribution value of 0.12 (Soodyall *et al.*, 1996). The Gamma distance corrects for nuceotide substitution rate variation among different sites. By assuming that humans and chimpanzees diverged 4-6 million years ago, Soodyall *et al.* (1996) estimated the rate of substitution per lineage, λ, (with a 95% confidence interval) to be 2.89×10^{-8} (upper limit) and 1.64×10^{-7} (lower limit). The modal rate, calculated under the assumption that humans and chimpanzees diverged five million years ago, was then estimated to be 8.31×10^{-8}, or 16.6%/million years (95% confidence interval 5.78% - 32.88%) (Soodyall *et al.*, 1996).

A median joining network of mtDNA types that corresponded to the L1i haplogroup was constructed by use of the programme NETWORK (developed by Röhl; http//www.fluxus-engineering.com) to further refine the phylogenetic relationships of mtDNA types associated with the 9-bp deletion. Median networks, based on the median algorithm of Bandelt *et al.* (1995), contain all the most parsimonious trees and display graphically the full information content of the sequence data. Unlike in phylogenetic trees where parallel mutation events may obscure the true evolutionary pathways, median networks display the principal character relationships present in the data and resolve likely parallel events while retaining character conflicts in the form of reticulations when ambiguity remains.

Mismatch distributions which reflect episodes of population growth and decline and population separations (Rogers and Harpending, 1992; Harpending *et al.*, 1993) were used to examine the demographic history of the population groups in this study. The distributions were constructed using the IWAVE programme (S. Sherry, personal communication) with a substitution rate of 16.6%/million years (Soodyall *et al.*, 1996). IWAVE computes the pairwise-difference distribution for nucleotide data and fits a theoretical curve to this empirical distribution defined by the parameters tau, τ, and theta, θ, based on the model of sudden expansion (Rogers and Harpending, 1992). The theory suggests that the crest and slope of the leading face of the wave can provide estimates of population expansion times, τ, and pre-expansion equilibrium effective population size, θ. Estimates of past population expansions and separations are measured as $\tau = 2\mu t$ units of mutational time, in which μ is the total mutation rate over all sites in the sample and t is time in generations (Rogers, 1995).

The mean expansion time in years is computed as (τ/effective sequence length) ($1^6/2\mu$) / 2, where the parameter 1^6 counteracts for the mutation rate which is expressed as a percentage per 1^6 years and is used to estimate τ.

Harpending's (1994) raggedness statistic, r, was calculated with the ARLEQUIN package (Schneider *et al.*, 1997). This value measures the "smoothness" of wave profiles generated by pairwise-difference distributions and can be used to differentiate between expanding and constant population size (Tajima, 1989; Harpending, 1994; Aris-Brosou and Excoffier, 1996). Raggedness values between 0.05 and 0.75 suggest constant population size over time, while values below 0.05 imply that the population under investigation has expanded in size (Harpending *et al.*, 1993).

A population tree showing the relationships between various sub-Saharan African populations (Watson *et al.*, 1996; Graven *et al.*, 1995; Vigilant, 1990; Soodyall, 1991; Soodyall, unpublished) and the Remba and Mozambican population groups was constructed, from CR1 sequence data (np 16060 to 16356), using the d_A distance (Nei and Jin, 1989) and NJ method with the SENDBS programme (N. Takezaki, personal communication; http//www.cib.nig.ac.jp/dda/ntakezak.html#sendbs). SENDBS was also used to compute Felsenstein's (1985) bootstrapping probabilities of the NJ tree, using 1000 replicates. Population affinities were examined among the same sub-Saharan African population groups using principal component (PCO) analysis determined with the NTSYS-pc programme

(developed by J. Rohlf). ARLEQUIN was used to calculate a matrix of Slatkin's linearized F_{st} values which was used as the input data for NTSYS-pc (Slatkin, 1995).

2.4.2 Y chromosome data analysis

Allele frequencies were determined at each locus. Haplotypes were constructed using the alleles from each locus in the order YAP, DYS393, DYS19, DYS390 and DYS391. Alleles at loci DYS388 and DYS392 were included to further resolve haplotypes. Median joining networks based on the stepwise model of mutation outlined by Shriver *et al.* (1993) were constructed from haplotypes using the NETWORK programme (developed by Röhl; http//www.fluxus-engineering.com). The networks were used to examine the relationships and distribution of different haplotypes in Mozambican, Remba, Lemba, Venda, Yemeni and Jewish populations.

2.4.3 Analysis of molecular variance (AMOVA) and population affinities

AMOVA was used to test population structuring at different hierarchical levels: linguistic (all four populations grouped together as SEB speakers), geographical (the Lemba and Venda grouped together) and historical (the Remba and Lemba grouped together), using the options contained in the ARLEQUIN package (Schneider *et al.*, 1997). Population affinities among Remba, Lemba, Mozambican and Venda individuals were examined using Y chromosome

haplotype data. PCO plots were constructed in the same way as that described for mtDNA data (section 2.4.1).

2.4.4 Genetic diversities

Y chromosome haplotype diversity and mtDNA sequence diversity were calculated using options within the ARLEQUIN package. Both are measured as gene diversity (Nei, 1987, p180), which is equivalent to the expected heterozygosity for diploid data. It is defined as the probability that two randomly chosen haplotypes or mtDNA types are different in the sample and is estimated as

$$h = n/n - 1 \left(1 - \sum_{i=1}^{k} p_i^2\right)$$

where n is the sample size, k is the number of haplotypes or mtDNA types, and p_i is the sample frequency of the i-th haplotype (mtDNA type) (Nei, 1987, p180).

Nucleotide diversity was calculated for mtDNA sequence data using ARLEQUIN. This is the probability that two randomly chosen homologous nucleotides are different. It is equivalent to the gene diversity at the nucleotide level. Nucleotide diversity was estimated according to the formula of Nei (1987, p257).

3. RESULTS

3.1 MtDNA studies

3.1.1 Frequency of the 9-bp deletion

The frequencies of the 9-bp deletion in the Remba, individuals from Mozambique and other SEB speaking populations are given in Table 4.

The 9-bp deletion frequencies in the Remba and Mozambican populations fell within the range found in other SEB population groups (5.7% - 26.7%) (Soodyall *et al.*, 1996). Although there were no significant differences in the frequency of the 9-bp deletion in the various groups (Table 4), the frequency of the 9-bp deletion in the Mozambican population group was similar to that of the Thonga (Tsonga). This is not surprising since the Shangana-Thonga occupy the most southern region of Mozambique, south of the Save River (Reys *et al.*, 1970). The Mozambican samples in this study were collected in Inhambane which falls within the Shangana-Thonga territory.

Table 4. Frequencies of the 9-bp deletion in SEB speaking populations

Ethnic group	No. of individuals	No. with 9-bp deletion	Frequency of 9-bp deletion (%)	Reference
Remba[1,2,3]	94	17	18.1	Present study
Mozambicans[1,2,3]	71	7	9.9	Present study
Lemba[2,3]	26	7	26.9	Soodyall *et al.*, 1996
Venda[3]	41	9	22.0	Soodyall *et al.*, 1996
Tsonga[3]	35	2	5.7	Soodyall *et al.*, 1996
Nguni[3]	101	13	12.9	Soodyall *et al.*, 1996
Sotho/Tswana[3]	54	11	20.4	Soodyall *et al.*, 1996
Malawian[3]	45	12	26.7	Soodyall *et al.*, 1996

Significant differences in the frequency of the 9-bp deletion:
[1]P = 0.1378
[2]P = 0.0662
[3]P = 0.0692

3.1.2 Phylogenetic analysis of mtDNA sequences

The results of the phylogenetic analyses of sequence data from Remba, Lemba and Venda individuals and those from Mozambican individuals are presented separately first to address questions posed in the Introduction (section 1.5) concerning the origins of the female ancestors of the Lemba and Mozambique populations. The results are then combined to examine the relationships of mtDNA types found in all four population groups. Throughout this section i.e. 3.1.2, 3.1.3, 3.1.4 and 3.1.5, mtDNA types have been named and grouped according to the RFLP haplogroup to which they correspond based on the nomenclature of Watson *et al.* (1997). A description of how the comparison is derived is given in section 3.1.4. Since the nomenclature is based on CR1 data only, "a", "b", "c" etc have been used to designate mtDNA types where different mutations in CR2 differentiate unique CR1 types. This is not technically correct since both CR1 and CR2, and in fact the remainder of the mitochondrial genome, represent one mitochondrial lineage, but we used this nomenclature to assist us in the comparative analysis.

3.1.2.1 The Remba, Lemba and Venda population groups

Both hypervariable regions (CR1 and CR2) of the mitochondrial control region were sequenced in 67 Remba individuals and examined in conjunction with sequence data from 44 Lemba and 39 Venda individuals (Soodyall, unpublished). From pairwise comparisons of mtDNA sequences using the p-distance and the NJ method, we derived 101 unique mtDNA

types in the three population groups. Figure 4 shows an unrooted NJ tree illustrating the phylogenetic relationship of these mtDNA types.

Only 19 mtDNA types were found in more than one individual, with 10 of these shared by more than one population group. Three sequences were common to the Remba, Lemba and Venda: mtDNA types 1a, 93a and 102l (Figure 4). Two mtDNA types were shared by the Remba and the Lemba, four by the Remba and Venda and two by the Lemba and Venda. No trends were observed in the clustering of mtDNA types in any one particular population group. However, mtDNA types associated with the 9-bp deletion, in all three groups formed a distinct cluster (Figure 4).

3.1.2.2 The Mozambican population group

An unrooted NJ tree showing the phylogenetic relationship of mtDNA types found in 61 individuals from Mozambique is shown in Figure 5. Forty-seven unique mtDNA types were derived after pairwise comparisons of mtDNA sequences using the p-distance and NJ method in the Mozambican sample group. Only eight of these types were shared by more than one individual. Again, mtDNA types associated with the 9-bp deletion clustered together.

3.1.2.3 A combined phylogeny of the Remba, Lemba, Venda and Mozambicans

A NJ tree showing the phylogenetic relationships of mtDNA types found in all four population groups is shown in Figure 6. The tree was rooted with a Neanderthal sequence (Krings *et al.*, 1997; Krings *et al.*, 1999) to give some indication of the evolutionary relationship of clusters. From the 211 sequences examined in the Remba, Lemba, Venda and Mozambicans, we derived 141 unique mtDNA types. Only 15 mtDNA types were shared by more than one group. The three types common to the Remba, Lemba and Venda (mtDNA types 1a, 93a and 102l) were also found in individuals from Mozambique. MtDNA types from each of the four groups were dispersed randomly throughout the gene tree. Similarly, no trends in type-sharing were observed between sequences found in Mozambican individuals and those found in the Remba, Lemba or Venda. Since mtDNA types associated with the 9-bp deletion are derived from a unique lineage (Soodyall *et al.*, 1996), it is not surprising that they have clustered together in both separate analyses as well as the combined phylogeny.



Figure 4. Unrooted NJ tree showing the phylogenetic relationship of unique mtDNA types in the Remba (RE), Lemba (LE) and Venda (VE) populations. The bracketed numbers refer to the number of individuals within the population group with that specific mtDNA type. Sequences were compared to the human mtDNA reference sequence (AND) (Anderson *et al.*, 1981).

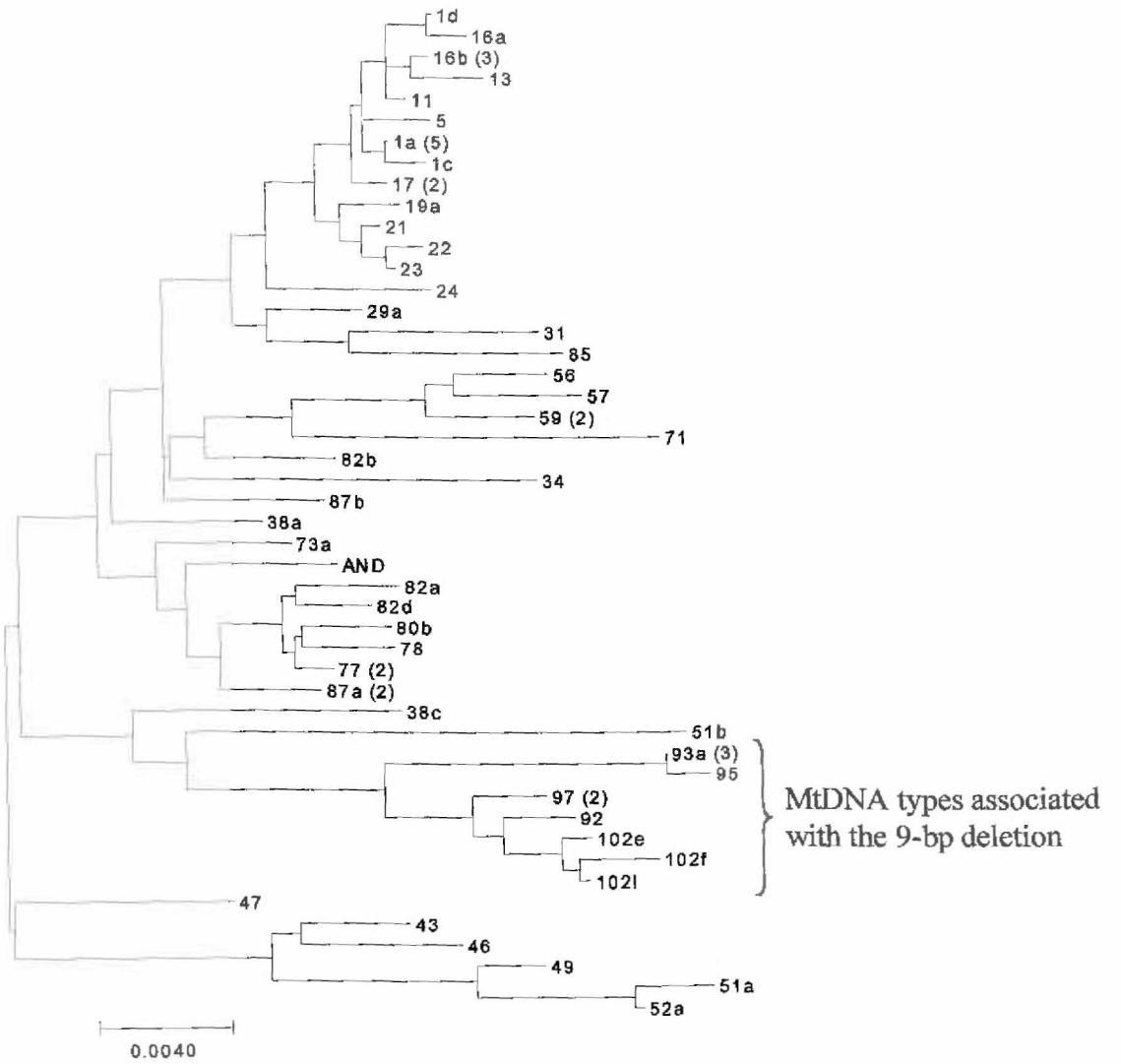


Figure 5. Unrooted NJ tree showing the phylogenetic relationship of mtDNA types found in individuals from Mozambique. The numbers in brackets after some mtDNA types refer to the number of individuals in which those specific types were present. The reference sequence for the human mtDNA control region is included (AND).

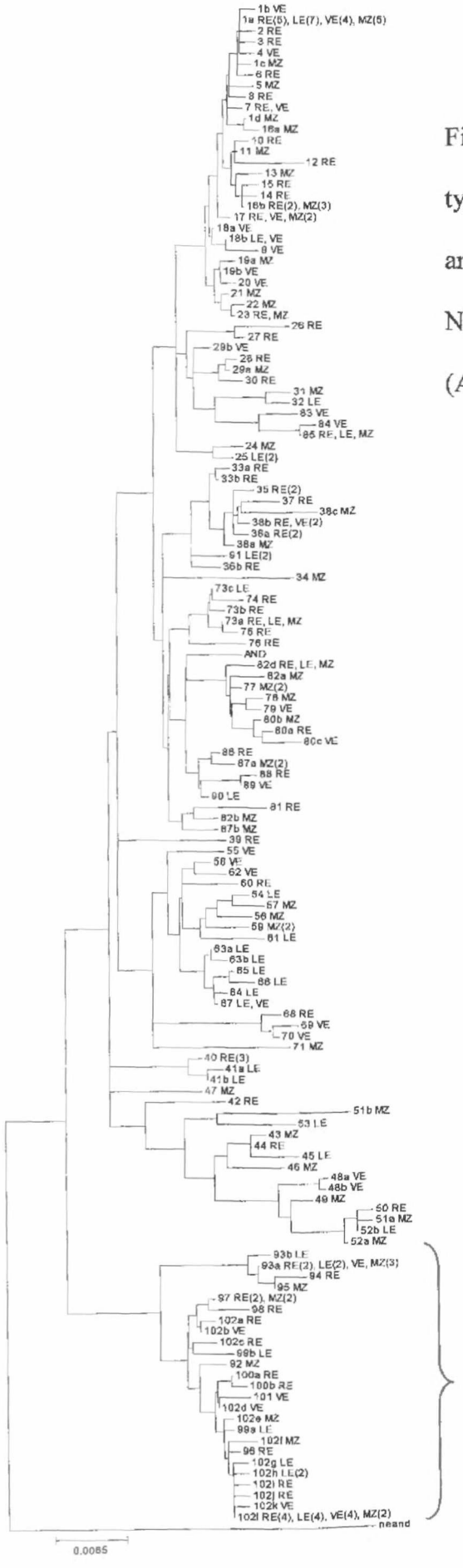


Figure 6. A NJ tree showing the phylogeny of mtDNA types found in the Remba (RE), Lemba (LE), Venda (VE) and Mozambicans (MZ). The tree is rooted with the Neanderthal sequence and the human reference sequence (AND) is included.

MtDNA types associated with the 9-bp deletion

3.1.3 MtDNA type diversity

The 141 unique mtDNA types, described above, were defined by 127 variant sites (82 in CR1 and 45 in CR2) (Table 6). Multiple mutations, involving both transitions and transversions, occurred at np 16286, 16289 and 16293 in CR1, and 185, 189 and 294 in CR2. Transversion mutations occurred at np 16064, 16166, 16188, 16265 and 16291 (CR1), and 95, 186, 209 and 281 (CR2). The remaining mutations were all transitions. The most common mutation occurred at np 16223 and was found in all but four mtDNA types (96.1%). Other common substitutions in CR1 were found at np 16187 (36.3%), 16189 (69.6%), 16278 (61.8%), 16294 (43.1%) and 16311 (51.0%). The most common CR2 variants at np 73, 152 and 263 were found in 80.9%, 80.9% and 85.1% of mtDNA types, respectively (Table 6).

Genetic diversity is defined as the probability that two randomly chosen haplotypes or mtDNA types are different in a population and is, therefore, correlated with the number of unique mtDNA types in the population. Since very little type-sharing was observed within each of the four population groups, it is not surprising that the diversity values calculated for each group were very high ranging from 0.9672 to 1.0000 (Table 5). The genetic diversity values obtained for the Remba, Lemba, Venda and Mozambicans are generally greater than those for Khoisan-speaking populations (average 0.916) (Soodyall, unpublished), Europeans (0.936 – 0.984) and Eastern Asians (0.947 – 0.993) (Comas *et al.*, 1998). Nucleotide diversities, which measure the probability that two randomly chosen homologous nucleotides are different, ranged between 0.021 – 0.023 in the Remba, Lemba, Venda and Mozambican

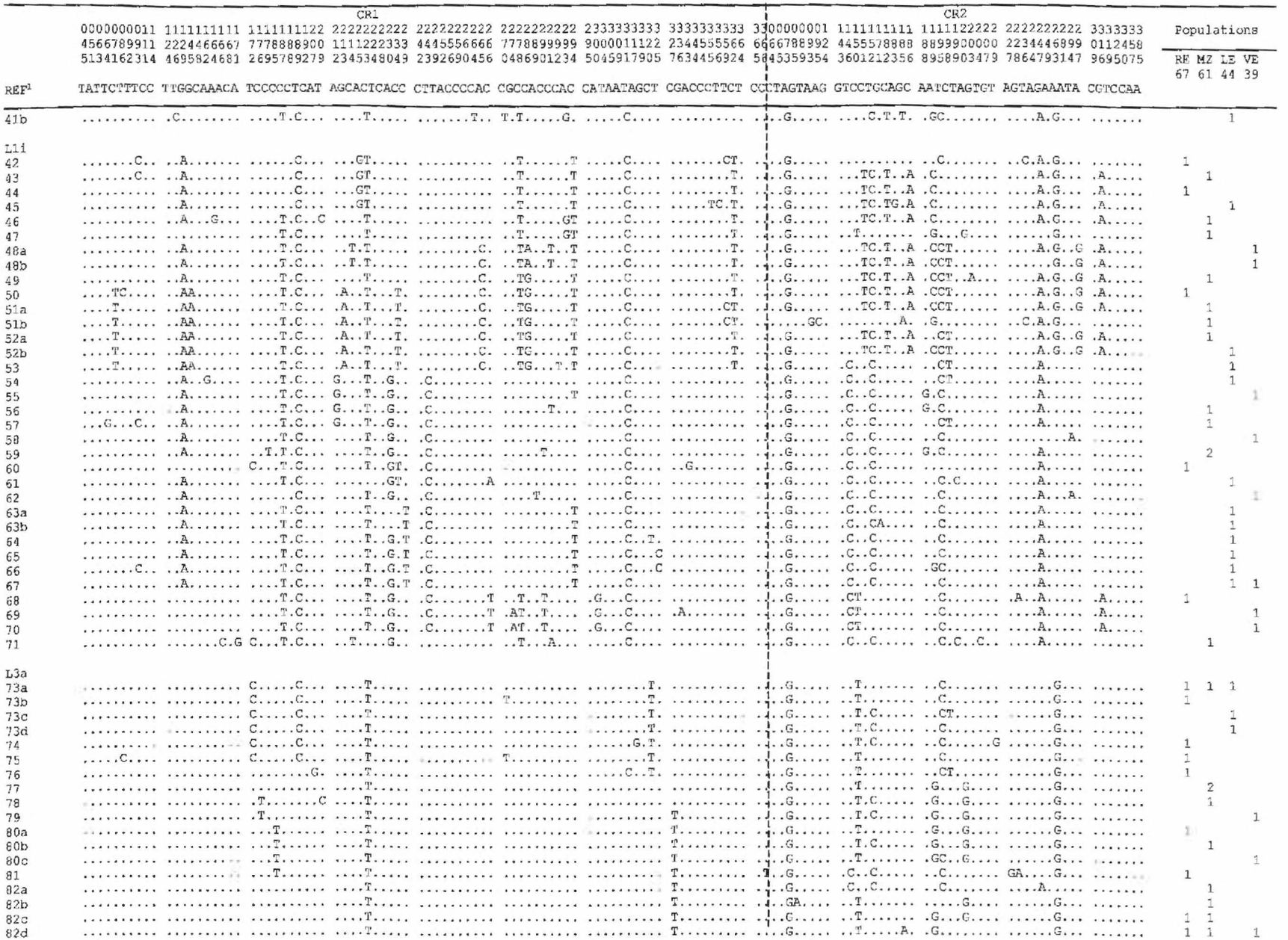
population groups (Table 5). These values are also much higher than those in Asia (0.016 – 0.020) and Europe (0.008 – 0.016) (Comas *et al.*, 1998, and references therein), and the !Kung (0.010) and Sekele (0.019) Khoisan-speaking populations (Soodyall, unpublished).

Table 5. Diversity parameters for the Remba, Lemba, Venda and Mozambicans

	Remba	Lemba	Venda	Mozambicans
Sample size	67	44	39	61
Unique mtDNA types	42	22	23	37
Genetic diversity	0.994 ± 0.004	0.967 ± 0.017	1.000 ± 0.006	0.987 ± 0.006
Nucleotide diversity	0.021 ± 0.010	0.021 ± 0.104	0.021 ± 0.011	0.022 ± 0.011
Mean number of pairwise differences	15.26 ± 6.90	15.83 ± 7.20	15.30 ± 6.99	15.71 ± 7.11

Table 6. MtDNA types (CR1 and CR2) found in the Remba (RE), Mozambicans (MZ), Lemba (LE) and Venda (VE)

	CR1										CR2										Populations			
	0000000011	1111111111	1111111122	2222222222	2222222222	2222222222	2333333333	3333333333	3300000001	1111111111	1111122222	2222222222	3333333											
	4566789911	2224466667	7778888900	1111222333	4445556666	7778899999	9000011122	2344555566	6666788992	4455578888	8899900000	2234446899	0112456											
	5134162314	4695824681	2695789279	2345348049	2392690456	0486901234	5045917905	7634456924	5845359354	3601212356	8958903479	7864793147	9695075								RE	MZ	LE	VE
REF[1]	TATTCTTTCC	TIGGCAAACA	TCCCCCTCAT	AGCACTCACC	CTTACCCCAC	CGCCACCCAC	CATAATAGCT	CGACCCTTCT	CCCTAGTAAG	GTCCTGCAGC	AATCTAGTGT	AGTAGAAATA	CGTCCAA								67	61	44	39
MtDNA type[2]																								
L2																								
1a	.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								5	5	7	4
1b	.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	.....	.CG.....	.C..C.....	..C.....	.....G...	.....											1
1c	.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....A.	.....G...	.....									1		
1d	.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....GG...	.....									1		
2	.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	.....	T.G.....	.C..C.....	..C.....	.....G...	.....								1			
3	.....	.....	.....C...	....T.....	...G.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								1			
4	C.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....											1
5	.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	C	..G.....	.C..C.....	.....	.....G...	.....									1		
6	.....	.....	.....C...	....T.....	.....	..T..T...T	T...G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								1			
7a	.....	.....	.....C...	....T.....	.....	..T..T...T	....G.....	.....T.	..G.....	.C..C.....	..C.....	.....G...	.....											1
7b	.....	.....	.....C...	....T.....	.....	T.T..T...T	....G.....	.....T.	..G.....	.C..C.....	..C.....	.....G...	.....								1			
8	.....	.....	.....C...	....T.T...	.....	..T..T...T	....C.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								1			
9	.....	.....	.....C...	....T.T...	T.....	..T.....T	....G.....	.....	..G.....	.C..C.....	..CT.....	.....G...	.....											1
10	.....	.....	.....CT..	....T.....	.....	..T..T...T	.....	.....T.	..G.....	.C..C.....	..C.....	.....G...	.....								1			
11	.....	.....	.....CT..	....T.....	.....	..T..T...T	.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....									1		
12	.....	.....	.....CT..	....T.....	.....T....	..T..T...T	....G.....	.....T.	..G.....	.C..C.....	..C.....	GA....G...	.....								1			
13	.....C....	.....	....T.CT..	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....									1		
14	.....	.....	.....TCT..	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								1			
15	.....	.....	.....CT..	G...T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								1			
16a	.....	.....	.....CT..	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....GG...	.....									1		
16b	.....	.....	.....CT..	....T.....	.....	..T..T...T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								2	3		
17	.....	.....	.....CT..	....T.....	.....	..T.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								1	2		1
18a	.....	.....	.....C...	....T.....	.....	..T.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....											1
18b	.....	.....	.....C...	....T.....	.....	..T.....T	....G.....	.....	..G.....	.C..C.....	..CT.....	.....G...	.....										1	1
19a	.....	.....	.....	....T.....	.....	..T.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.A....G...	.....									1		
19b	.....	.....	.....	....T.....	.....	..T.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....											1
20	.....C...	.....	.....	....T.....	.....	..T.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....											1
21	.....	.....	.....C...	....T.....	.....	..TT.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....									1		
22	.....	.....	C.....	....T.....	.....	..TT.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....									1		
23	.....	.....	.....	....T.....	.....	..TT.....T	....G.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....									1	1	
24	.....	.....	.....C...	....TC....	....T.....	..T.....	....G.....	.....	..G.....	.C..C.T...	.....	.....G...	.....									1		
25	.....	.....	.....C...	....TC....	.....	..T.....	.....	.....	..G.....	.C..C.TG..	.....	.....G...	.....										2	
26	.....	.....	.....	....T...T.	..C.....	.....	.....T T.	.....	..G.....	AC..C.T...	..C.....	.....G.C.	.....								1			
27	.....	.....	.....	....T...T.	..C.....	..T.....T	T.....	.....	..G.....	.C..C.....	..C.....	.....G...	.....								1			
28	.....	..A.....	.....	....T.....	.....	..T.....T	.....	.....	..G.....	AC..C.T...	..C.....	.....G...	.....								1			
29a	.....	.....	.....	....T.....	.....	..T.....T	.....	.....	..G.....	AC..C.T...	..C.....	.....G...	.....									1		
29b	.....	.....	.....	....T.....	.....	..T.....T	.....	.....	..G.....	.C..C.T...	..C.....	.....G...	.....											1
30	.....	.....	.....	....T.....	.....T...	..T.....	.....	.....	..G.....	AC..C.T...	..C.....	.....G...	.....								1			
31	.....	.....	.....	....T.....	.....T..	..T.....	.....	.....	..G..G..	.CT.C.T...	..CT.....	.....G...	..CT...									1		
32	.....	.....	.....	....T.....	.....T..	..T.....	.....C....	.....	..G..G..	.CT.C.TG..	..CT.....	.....G...	...T...										1	
33a	.....	.....	.....	....T.....	.....	..T.....	.....C....	C	..G.....	.....	.....	.....G...	.....								1			
33b	.....	.....	.....	....T.....	.....	..T.....	.....C....	C	..G.....	....C.....	.....	.....G...	.....								1			
34	.G.....T	.....	.....CT..	....T.....	.....	.....T.	.....C....	C	..G.....	.C..C.....	..C.....	...G..G...	....T..									1		
35	.....	C.....	.....	....T.....	.....	..T.....	....G.....	C	..G.....	.....	.....	.....C...	.....								2			
36a	.....	C.....	.....	....T.....	.....	..T.....	.....C....	C	..G.....	.....	.....	.....G...	.....								2			
36b	.....	C.....	.....	....T.....	.....	..T.....	.....C....	C	..G.....	.C..C.....	..C.....	.....G...	.....								1			
37	.....C..	C.....	.....	....T.....	.....	..T.....	.....	C	..G.....	.....	.....G	.....G...	.....								1			
38a	.....	C.....	.....	....T.....	.....	..T.....	.....	C	.....G..	.....A.	.G.....G	..C.A.G...	.....									1		
38b	.....	C.....	.....	....T.....	.....	..T.....	.....	C	..G.....	.....	.....	.....G...	.....								1			2
38c	.....	C.....	.....	....T.....	.....	..T.....	.....	C	..G.....	....C.....	.....	.....G...	.....									1		
39	.....	..A.....	.....C...	.....	.....	..T.....	.G...C...	....T.....	..G.....	.CT.....	..C.....G	.....G...	.....								1			
40	.....	.C.....	....T.C...	....T.....	.....T..	T.T.....	.....C....	.....	..G.....	....C.T.T.	..C.....	....A.G...	.....G.								3			
41a	.....	.C.....	....T.C...	....T.....	.....T..	T.T.....G.	.....C....	.....	....G.....	....C.T.T.	.GC.....	....A.G...	.....G.										1	


CR1
CR2
Populations
RE MZ LE VE
67 61 44 39
REF[1]
TATTCTTTCC TTGGCAAACA TCCCCCTCAT AGCACTCACC CTTACCCCAC CGCCACCCAC CATAATAGCT CGACCCTTCT CCCTAGTAAG GTCCTGCAGC AATCTAGTGT AGTAGAAATA CGTCCAA
41b
L1i
42
43
44
45
46
47
48a
48b
49
50
51a
51b
52a
52b
53
54
55
56
57
58
59
60
61
62
63a
63b
64
65
66
67
68
69
70
71
L3a
73a
73b
73c
73d
74
75
76
77
78
79
80a
80b
80c
81
82a
82b
82c
82d

	CR1										CR2										Populations			
	0000000011	1111111111	1111111122	2222222222	2222222222	2222222222	2333333333	3333333333	3300000001	1111111111	1111122222	2222222222	3333333											
	4566789911	2224466667	7778888900	1111222333	4445556666	7778899999	9000011122	2344555566	6666788992	4455578888	8899900000	2234446899	0112458											
	5134162314	4695824681	2695789279	2345348049	2392690456	0486901234	5045917905	7634456924	5845359354	3601212356	8958903479	7864793147	9695075								RE	MZ	LE	VE
REF[1]	TATTCTTTCC	TTGGCAAACA	TCCCCCTCAT	AGCACTCACC	CTTACCCCAC	CGCCACCCAC	CATAATAGCT	CGACCCTTCT	CCCTAGTAAG	GTCCTGCAGC	AATCTAGTGT	AGTAGAAATA	CGTCCAA								67	61	44	39
83	.....	..A.....	.....	.A..T.....	.....	..T.....	.....	.....T...C	..G.....	..T.C.T...	..CT...C..	......G...	.....											1
84	.....A	..A.....	.....	.A.....	.....	..T.....	.....	....T.....	..G.....	.CT.C.T...	..CT...C..	......G...	.....											1
85	.....A	..A.....	.....	.A..T.....	.....	..T.....	.....	....T.....	..G.....	.CT.C.T...	..CT...C..	......G...	.....								1	1	1	
L3b																								
86	.....	C.....	.....	....T.....	.....	.....	.....	.....	..G.....	.CT.C.....	.....	......G...	.....								1			
87a	.....	C.....	.....	....T.....	.....	.....	.....A..	.....	..G.....	..T.C.....	.....	......G...	.....									2		
87b	.....	C.....	.....	....T.....	.....	.....	.....A..	.....	..G.....	.C..C.....	..C.....	......G...	.....									1		
88	.....T.	C.....	.....	....T.....	.....	.....	.....	.....T....	..G.....	....C.....	....C.....	......G...	.....								1			
89	.....T.	C.....	.....	....T.....	.....	.....	.....	.....	..G.....	....C.....	....C.....	......G...	.....											1
90	.....	C.....	.....	....T.....	.....	.....	.....C....	.....	..G.....	....C.....	.....	......G...	.....										1	
91	.....	C.....	.....C...	....T.....	.....	..T.....	..C..C....	.....	..G.....	....C.....	.....	......G...	.....										2	
L1a																								
92	.....	.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G.C	....C.....	.....CA.	..C.A.G...	.....									1		
93a	.....	..A.T...T.	C...TGC...	....T..G..	.....	..T.....G.	.....C..T.	.....	.....GC.	.....A.	.G.....	..C.A.G...	.....								2	3	2	1
93b	.....	..A.T...T.	C...TGC...	....T..G..	.....	..T.....G.	.....C..T.	.....	...A.GC.	.....A.	.GC.....	..C.A.G...	.....										1	
94	.....	..A.T...T.	C...TGC...	....T..G..	.....	..T.....G.	...G.C..T.	...T.....	.....GC.	.....A.	.G.....	.AC.A.G...	.....								1			
95	.....	..A.T...T.	C...TGC...	....T..G..	.....	..T.....G.	...G.C..T.	.....	.....GC.	.....A.	.G.....	..C.A.G...	.....									1		
96	.....	..A.T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....CA.	..C.A.G...	.....								1			
97	.....C..	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....	..C.A.G...	.....								2	2		
98	.....C..	....T.....	C...TAC...	..T.T..G..	.....	....G.....	.....C..T.	.....	T....G..	....C.....	.G.....G	..C.A.G...	.....								1			
99a	.....	....T.....	C...TAC...	....T..G..	.....	.....	.....C..T.	.....	T...CGC.	.C..C.....	.G.....	..C.A.G...	.....										1	
99b	.....	....T.....	C...TAC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....CA.	..C.A.G...	.....										1	
100a	.....	....T.....	C...TGC...	....T..GT.	.....	.....	.....C..T.	.....	.....G..	....C.....	.G.....CA.	..C.A.G...	.....								1			
100b	.....	....T.....	C...TGC...	....T..GT.	.....	.....	.....C..T.	.....	.....G..	....C.....	.G.....CAG	..C.A.G...	.....								1			
101	C.C.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	.....G..	....C.....	.G.....CA.	..C.A.G...	.....											1
102a	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....G	..C.A.G...	.....								1			
102b	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....	..C.A.G...	.....											1
102c	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	.C..C.....	.G.....	..C.A.G...	.....G								1			
102d	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	.....G..	....C.....	.G.....CA.	..C.A.G...	.....											1
102e	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	G....G..	....C.....	.G.....CA.	..C.A.G...	.....									1		
102f	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	..T.C.....	.G..C..CA.	..C.A.G...	.....									1		
102g	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....GC.	....C.....	.G.....CA.	..C.A.G...	.....										1	
102h	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T..A.G..	....C.....	.G.....CA.	..C.A.G...	.....										2	
102i	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....CA.	..C.A.G...	T.....								1			
102j	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....CA.	..C.A.GT..	.....								1			
102k	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	TC...G..	....C.....	.G.....CA.	..C.A.G...	.....											1
102l	.....	....T.....	C...TGC...	....T..G..	.....	.....	.....C..T.	.....	T....G..	....C.....	.G.....CA.	..C.A.G...	.....								4	2	4	4

[1]Anderson *et al.*, 1981. The nucleotides in CR1 are given without the prefixed "16".

[2]Unique mtDNA types for CR1 and CR2 (135) correspond to those in Figures 4, 5, 6 and 7. "a", "b", "c" etc have been used to designate types where the CR2 variant sites differentiate unique CR1 types. These types are grouped according to the RFLP haplogroup to which they correspond (based on CR1 types) according to the nomenclature of Watson *et al.* (1997).

3.1.4 Haplogroup distribution in the Remba, Lemba, Venda and Mozambicans

The mtDNA types identified in the four population groups examined corresponded to subhaplogroups L1a, L1i, L2, L3a and L3b defined by Watson *et al.* 1997. Each subhaplogroup was found in all four groups. MtDNA types associated with haplogroup L are found exclusively in Africa (Chen *et al.*, 2000). Therefore, there is no evidence of non-African mtDNA profiles in the Remba, Lemba, Venda and Mozambicans, suggesting that the female gene contribution to these populations is exclusively African. The numbers of individuals whose mtDNA sequences correspond to each subhaplogroup, the numbers of mtDNA types in each subhaplogroup and the distribution of the subhaplogroups among the four populations are summarized in Table 7. The similar distribution of haplogroups among the four populations suggests that they have experienced quite similar female contributions to their gene pools.

Subhaplogroup L2, comprising mtDNA types 1a to 41b (table 6), was not only the most common among the groups, it was also the most heterogenous group made up of 54 different mtDNA types (Table 7). Although L2 was originally defined by the RFLPs +*Hinf*I 16389/ -*Ava*II 16390 (Chen *et al.*, 1995), in this study mtDNA types were assigned to L2 on the basis of common variant sites in CR1 that corresponded to one of the following L2 sequence types defined by Watson *et al.* (1997):

(i) 16189, 16192, 16223, 16278, 16294, 16309;

(ii) 16223, 16278, 16294 or

(iii) 16223, 16278.

L2 was found in several African populations sampled by Watson *et al.* (1997) comprising Afro-Asiatic, Nilo-Saharan and Niger-Kordofanian language speakers but absent in Biaka Pygmies and the !Kung. However, the Vasikela !Kung (or Sekele), sampled by Chen *et al.* (2000) possessed mtDNA types that belonged to a subhaplogroup of L2, namely L2b.

Although Watson *et al.* (1997) found only a small subset of subhaplogroup L1a defined by the 9-bp deletion, in the present study all but four mtDNA types that corresponded to this subhaplogroup were those associated with the deletion (mtDNA types 92 and 97 to 102l, Table 6). The RFLP polymorphism defining L1a is an *Alu*I site gain at np 4310 (Chen *et al.*, 1995). Common mutations defining L1a, here, corresponded to one of two L1a sequence types identified by Watson *et al.* (1997):

(i) 16148, 16172, 16187, 16188C→G/A, 16189, 16223, 16230, 16311, 16320 (mtDNA types associated with the 9-bp deletion) and

(ii) 16129, 16148, 16168, 16172, 16187, 16188C→G, 16189, 16223, 16230, 16278, 16293, 16311, 16320 (mtDNA types found in individuals without the 9-bp deletion).

L1a was the second most common subhaplogroup in each of the four populations except the Mozambicans (Table 7), who were found to have a relatively lower frequency of the 9-bp deletion.

Subhaplogroups L3a and L3b were originally defined by RFLPs –*Hpa*I 3592, +*Mbo*I 2349 and -*Hpa*I 3592, -*Mbo*I 8616, respectively (Chen *et al.*, 1995). MtDNA types in the Remba, Lemba, Venda and Mozambicans were assigned to either L3a or L3b according to the following L3 sequence types identified by Watson *et al.* (1997):

(i) 16129, 16223 (L3a)

(ii) 16172, 16189, 16223, 16320 (L3a)

(iii) 16185, 16223, 16311, 16327 (L3a)

(iv) 16223, 16327 (L3a)

(v) 16124, 16189, 16223, 16278, 16311 (L3b)

(vi) 16124, 16223 (L3b) and

(vii) 16124, 16223, 16319 (L3b)

Fewer mtDNA types in the Remba, Lemba, Venda and Mozambicans corresponded to L3 than those in the combined African populations sampled by Watson *et al.* (1997) where L2 and L3 were the most common subhaplogroups. The subset of L1a defined by the 9-bp deletion may have reached greater proportions than L3 in the four Bantu-speaking population groups studied here as a result of the Bantu expansion. It has also been noted that although L3 mtDNA types are widely dispersed in eastern Africa, they are less prevalent in western and southern Africa (Chen *et al.*, 1995, 2000; Watson *et al.*, 1997).

The remainder of mtDNA types (42 to 71, Table 6) found in the four populations corresponded to the L1i subhaplogroup (Table 7). A median network of the L1i sequences identified by Watson *et al.* (1997) converged on a most common recent ancestor (MCRA)

having CR1 transition mutations at np 16129, 16187, 16223, 16230, 16278 and 16311. MtDNA types identified here with corresponding mutations were assigned to L1i. Sequences corresponding to this subhaplogroup have been termed "isolated lineages" since they were generally found to be population-specific (Watson *et al.*, 1997). It has been suggested that these lineages are the relics of a less dramatic and more ancient expansion event across Africa compared to the L2/L3 cluster (Watson *et al.*, 1997).

3.1.4.1 Estimating the age of the MRCA of the L subhaplogroups

Using a substitution rate of 16.6%/million years and a gamma value of 0.12 (Soodyall *et al.*,1996), the time of coalescence of each subhaplogroup was estimated from the mean sequence diversity within each subhaplogroup (Table 7). Our estimates of the age of the MRCA for the subhaplogroups, based on CR1 and CR2 sequence data, are comparable to those made by other investigators who used both CR1 only and CR1 and CR2 data sets. The coalescent time for the MRCA of mtDNA types corresponding to L1i was 155 000 years before the present (YBP) (95% confidence interval (CI) 78 000 – 445 000) by our analysis, 125 500 – 165 000 according to Chen *et al.* (2000) (CR1 and CR2 data), and 111 000 $\pm$ 5 700 according to Watson *et al.* (1997) (CR1 data). By contrast, the time of coalescence of L1a was much younger at 51 000 YBP (95% CI 26 000 – 147 000 years). Watson *et al.* (1997) arrived at a similar value of 52 000 $\pm$ 4 000 YBP. The estimated age of L2 was 63 000 YBP (95% CI 31 000 – 180 000 years). Similar estimates were made by Chen *et al.* (2000) (59 000 – 77 700 YBP), Watson *et al.* (1997) (56 000 $\pm$ 3 000) and Graven *et al.*

(1995) (70 333 ± 25 710) (CR1 data). Finally, the time of coalescence for L3 was 84 000 YBP (95% CI 42 000 – 240 000 years) by our analysis, 78 300 – 103 200 YBP according to Chen *et al.* (2000), 60 000 ± 2 400 YBP according to Watson *et al.*(1997) and 66 321 ± 24 965 YBP according to Graven *et al.* (1995). Thus, there appears to be consistent trends in the diversity and ages of the major clusters of mtDNAs in African populations.

Table 7. The frequencies and distribution of the haplogroups that corresponded to mtDNA CR sequence types found in the Remba (RE), Mozambicans (MZ), Lemba (LE) and Venda (VE)

RFLP haplogroup	Time to MRCA (YBP) (95% CI)	No. of mtDNA types (CR1) (%) n = 141	No. of individuals (%) n = 211	Distribution of haplogroups [No. (%)]			
				RE 67	MZ 61	LE 44	VE 39
L1i	155 000 (78 000 – 445 000)	34 (24.1)	36 (17.1)	5 (7.5)	12 (19.7)	11 (24.4)	8 (20.5)
L1a	51 000 (26 000 – 147 000)	25 (17.7)	49 (23.2)	17 (25.4)	11 (18.0)	12 (27.3)	9 (23.1)
L2	63 000 (31 000 – 180 000)	54 (38.3)	88 (41.7)	33 (49.3)	25 (41.0)	14 (31.8)	16 (41.0)
L3a	84 000 (42 000 – 240 000)	21 (14.8)	29 (13.7)	10 (14.9)	10 (16.4)	4 (9.0)	5 (12.8)
L3b	39 000 (20 000 – 111 000)	7 (5.0)	9 (4.3)	2 (3.0)	3 (4.9)	3 (6.8)	1 (2.6)

3.1.5 MtDNA types associated with the 9-bp deletion

MtDNA types 92 and 96 – 102l are found in individuals who have the 9-bp deletion. They all correspond to the subhaplogroup L1a according to the nomenclature of Watson *et al.* (1997) (Table 7). The substitutions almost exclusive to these types are found at np 16148, 16172, 16188G/A, 16230 and 16320 in CR1 and 93, 189G, 236 in CR2 (Table 6). Each of these mtDNA types, therefore, correspond to those identified by Soodyall *et al.* (1996) which represent the major form of the 9-bp deletion in Africa. Two types of mutations occur at np 16188: a G was found in 19 mtDNA types and an A in 3 types. Since mtDNA types in this study (99b and 102l) and that of Soodyall *et al.*, (1996) differ only at np 16188, it is likely that first a transversion (either C → G or C → A) occurred at this position (relative to the reference sequence) followed by a transition (either G → A or A → G, depending on which type of transversion occurred first). Other variant sites common to types associated with the 9-bp deletion are at np 16187, 16189, 16223, 16311, 16320 in CR1 and 152, 247, 263 in CR2 (Table 6).

There are four mtDNA types that share all the substitutions exclusive to those types associated with the deletion but are found in individuals without the 9-bp deletion. These are mtDNA types 93a, 93b, 94 and 95 (Table 6). They differ from the others by having additional substitutions at np 16129, 16168, 16278, 16293 in CR1 and 185 in CR2 and by lacking the substitution at np152 in CR2. All mtDNA types associated with the 9-bp deletion, including these four types, form their own distinct cluster in a NJ tree comprised of

all mtDNA types found in the Remba, Lemba, Mozambicans and Venda (Figure 6). However, mtDNA types associated with the deletion and those types without the deletion which harbour a similar sequence profile to the major form of the deletion are separated into two distinct clusters (Figure 7). In addition, a median network of all mtDNA types corresponding to the L1a subhaplogroup separates the cluster of mtDNA types found in individuals without the 9-bp deletion from the cluster of types with the deletion by eight nucleotide changes (Figure 8). It is, therefore, possible that mtDNA types with the deletion as well as the four unique types are either derived from a common mtDNA lineage or a duplication event has restored the second 9-bp copy in the four types without the deletion.

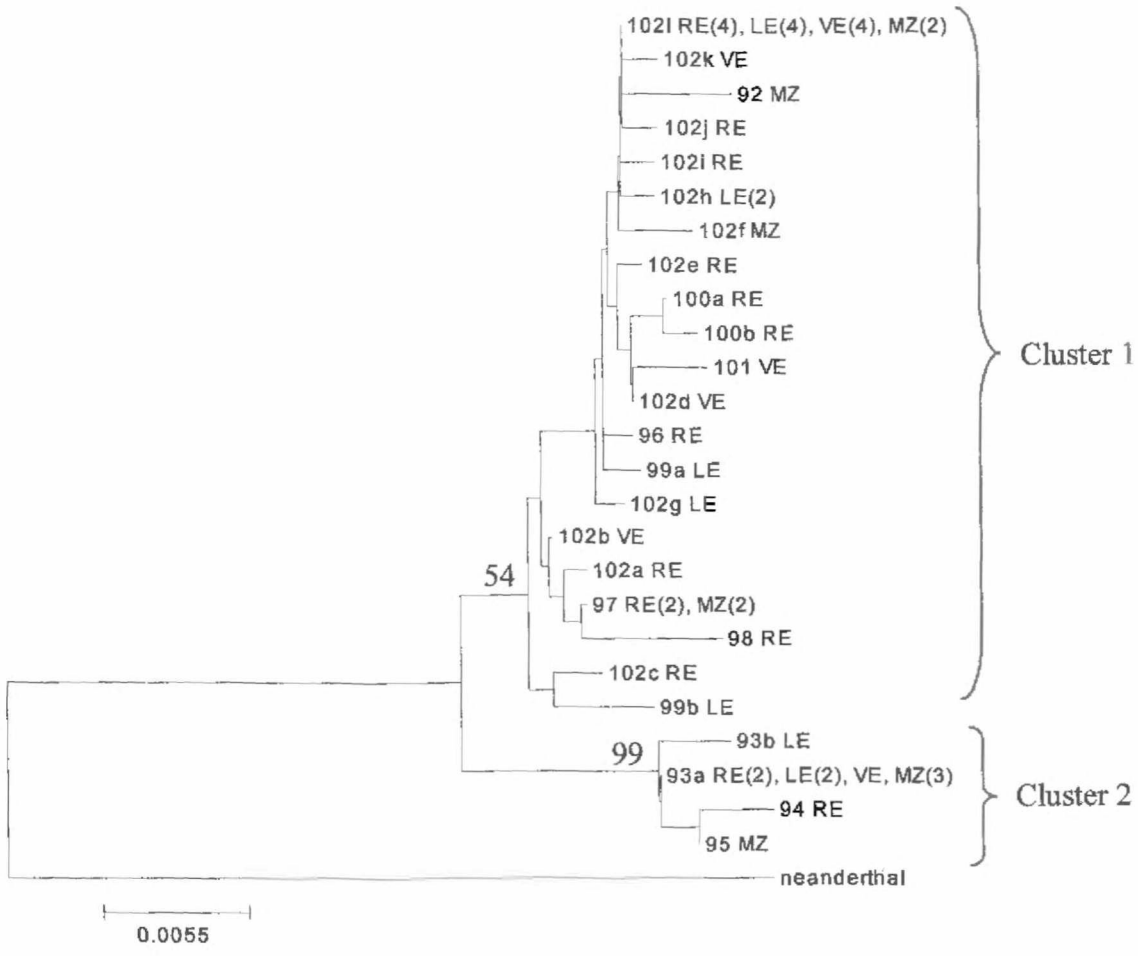


Figure 7. NJ tree showing two distinct clusters of mtDNA types associated with the 9-bp deletion. Cluster 1 comprises mtDNA types found in individuals with the deletion. Cluster 2 contains mtDNA types that have a very similar sequence motif to that found in individuals with the major form of the 9-bp deletion but were found in individuals who lack the deletion. The bootstrap support for the two clusters is given as a percentage.

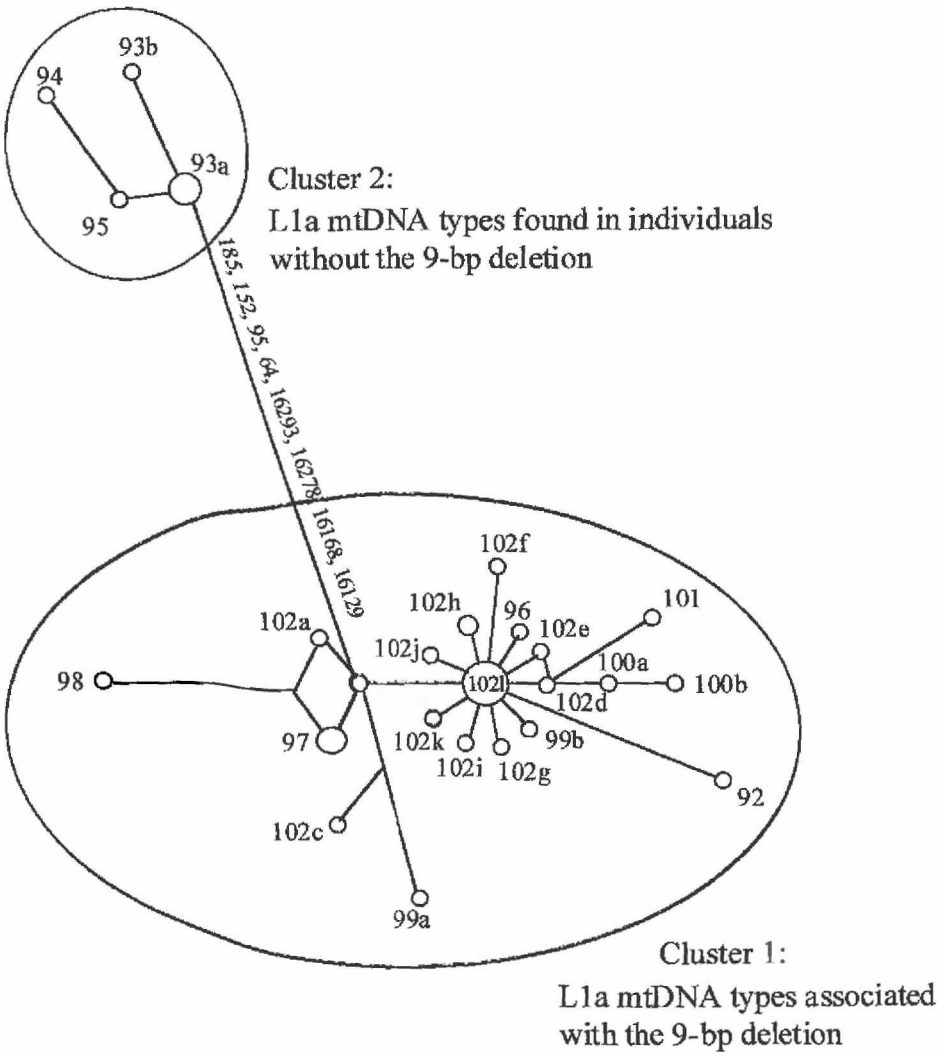


Figure 8. Median network showing the relationship of mtDNA that correspond to subhaplogroup L1a in individuals with (cluster 1) and without (cluster 2) the 9-bp deletion. The node marked in red (mtDNA type 102b) represents the consensus sequence, which deviates from the reference sequence (Anderson *et al.*, 1981) at sites 16148, 16172, 16187, 16188, 16189, 16223, 16230, 16311 and 16320 (CR1), and 65, 93, 152, 189, 236, 247 and 263 (CR2).

3.1.6 Mismatch distributions

Mismatch distributions derived from CR1 sequences only and both CR1 and CR2 sequences in the Remba, Mozambicans, Lemba and Venda are shown in Figure 9. Their summary statistics are given in Table 8. The wave profiles of the distributions derived from CR1 sequences only are similar and, with the exception of a small first peak in each distribution, generally unimodal with raggedness statistics ranging from 0.005 to 0.023. These values fall below the range, 0.05 – 0.75, that suggests constant population size (Harpending *et al.*, 1993) and, therefore, imply that each of the four populations have expanded in size. Similarly, the wave profiles derived from both CR1 and CR2 sequences for each of the four populations are associated with raggedness values (0.002 – 0.006) in keeping with population expansion events. When mtDNA types associated with the 9-bp deletion are excluded from the analyses, the first peaks in both CR1 and CR1 and CR2 distributions for each population group are removed (graphs not shown). This suggests that the mtDNA types associated with the deletion are derived from a different mtDNA lineage. Since a very similar profile has also been seen in South African SEB speakers (Soodyall, personal communication), it is likely that the mtDNA gene pools of the Mozambicans, Remba, Lemba and Venda as well as South African SEB speakers are derived from a common source. The peaks at ~12 and ~22 nucleotide differences in the distributions derived from CR1 and CR2 sequences appear to be due to sequence types that correspond to the L1i subhaplogroup. When these sequences are excluded from the analyses, the wave profiles revert to more normal distributions (graphs not shown).

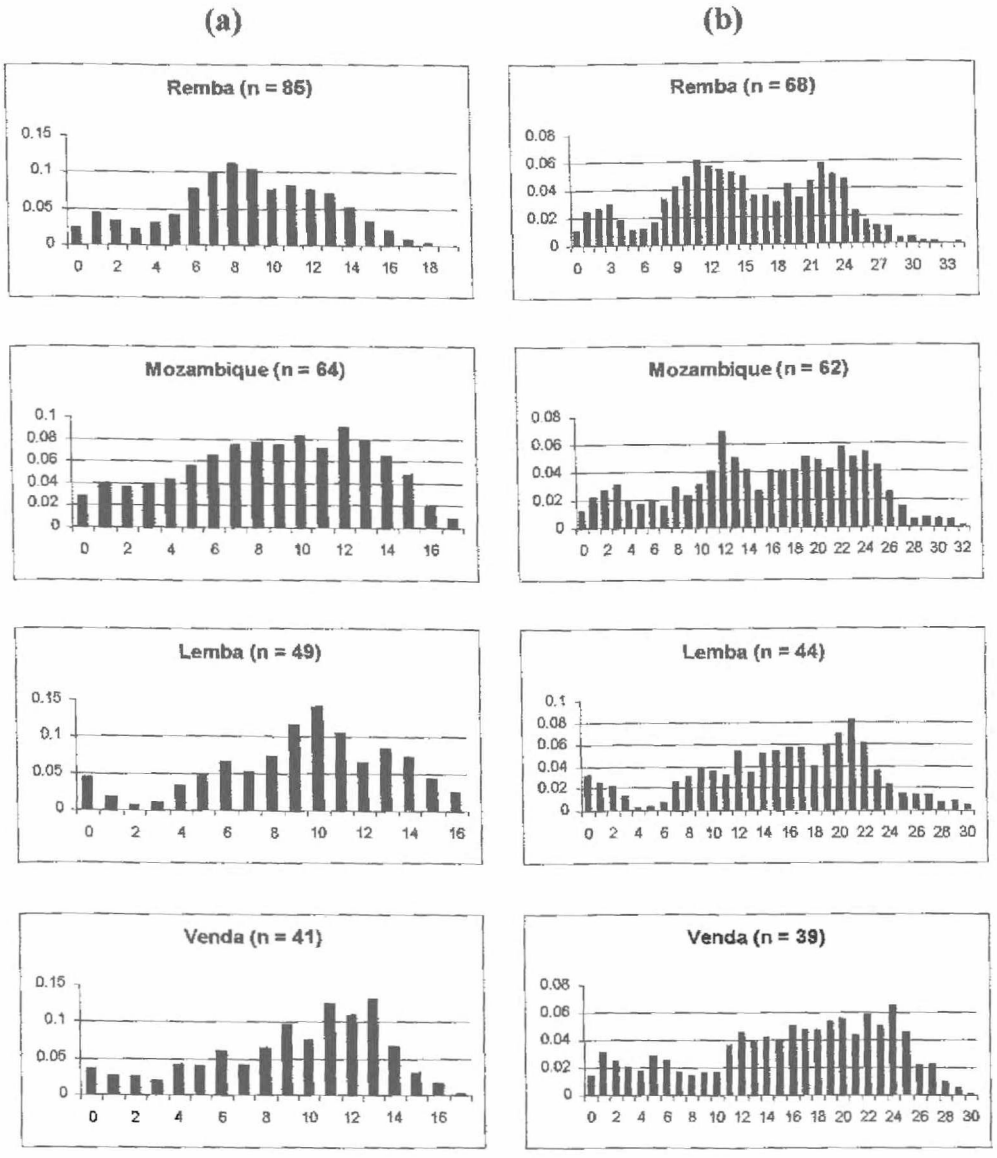


Figure 9. Mismatch distributions of sequences within the Remba, Mozambicans, Lemba and Venda. The distributions in column (a) are derived from CR1 sequences only while the distributions in column (b) are derived from both CR1 and CR2 sequences. The numbers of nucleotide differences between all pairs of sequences are along the x-axis, and the frequency of pairs is along the y-axis.

Table 8. Summary statistics for the mismatch distributions shown in Figure 9.

Population group	Effective sequence length (bp)		Mean		Variance		r[1]		τ^2		Expansion time
	CR1	CR1&2	CR1	CR1&2	CR1	CR1&2	CR1	CR1&2	CR1	CR1&2	CR1&2
Venda	403	780	9.35	15.89	16.05	55.35	0.012	0.004	6.76	9.60	73977
Lemba	404	781	9.29	15.36	15.03	47.32	0.023	0.006	6.90	9.71	74807
Mozambique	404	781	8.71	15.51	18.12	55.53	0.005	0.003	5.64	9.18	70754
Remba	404	778	8.62	14.88	16.05	51.61	0.006	0.002	5.90	8.82	67916

[1]Harpending's (1994) raggedness statistic

[2]$\tau = 2\mu t$ (in which μ = the mutation rate) Rogers (1995)

3.1.7 Affinities of the Remba, Lemba and Venda with the Mozambicans and other sub-Saharan African populations

The data set used for this analysis was comprised of 102 unique mtDNA types derived from 226 CR1 sequences in the Remba, Lemba, Venda and Mozambican population groups (Table 15, in the Appendix, section 7.4). These sequences were analyzed in conjunction with published sequence data from various sub-Saharan African populations. Figure 10 shows a NJ tree for populations, rooted with a Neanderthal sequence. The Remba, Lemba and Venda cluster among the Mozambicans and South African SEB speakers. The very short branch lengths indicate very little differences among these population groups and most of the other sub-Saharan African populations. However, the Khoisan (Sekele, !Kung and Nama) and SWB (Herero and Himba) speakers form their own distinct cluster. The PCO-plot (Figure 11) displays similar patterns of population grouping to the phylogenetic tree with respect to the Remba, Lemba, Venda, Mozambicans and SEB speakers (Figure 10).

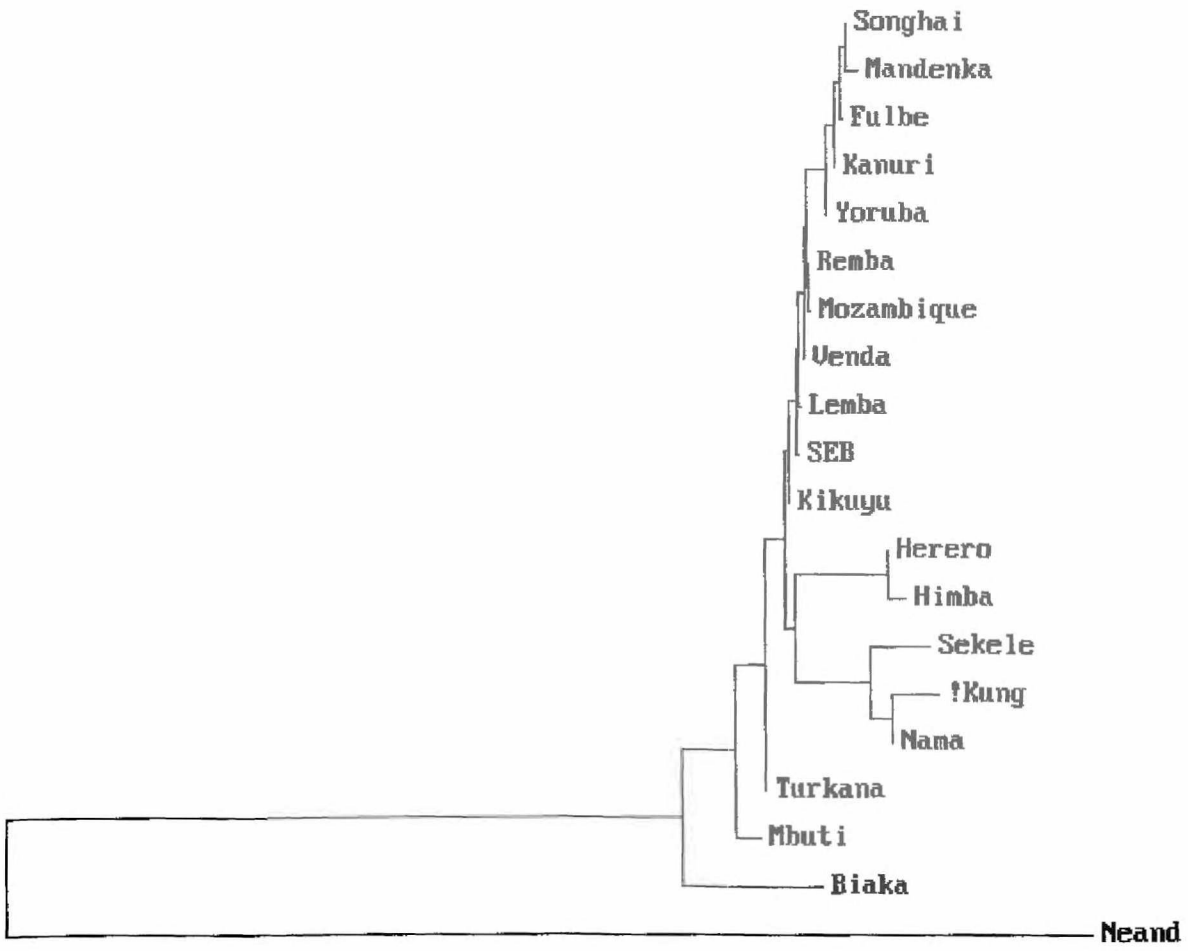
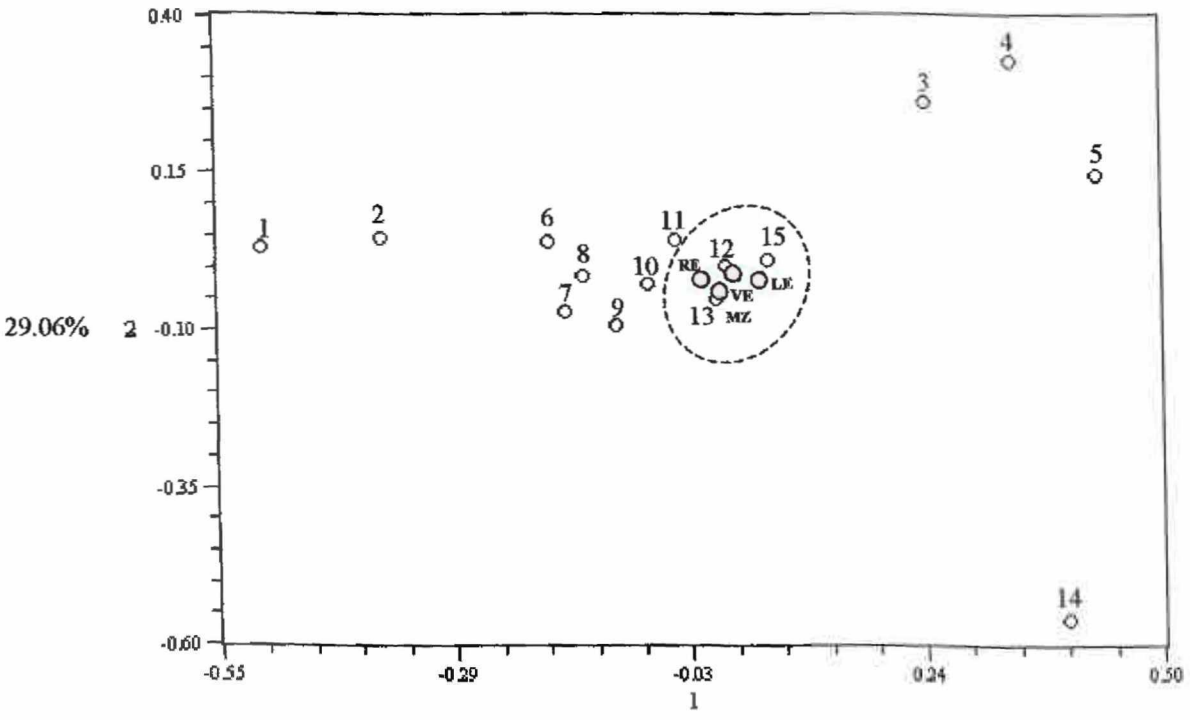


Figure 10. NJ population tree showing the mtDNA genetic affinities of the Remba, Mozambican, Lemba and Venda populations with other sub-Saharan African populations. The tree was rooted with a Neanderthal sequence and constructed using d_a distances from a matrix of the proportion of nucleotide differences.



KEY

1. Himba	5. !Kung	9. Songhai	13. Mbuti Pygmies
2. Herero	6. Kanuri	10. Yoruba	14. Biaka Pygmies
3. Nama	7. Mandenka	11. Kikuyu	15. SEB speakers
4. Sekele	8. Fulbe	12. Turkana	

Figure 11. PCO-plot showing the genetic affinities of the Remba (RE), Mozambicans (MZ), Lemba (LE) and Venda (VE) with other sub-Saharan African populations using mtDNA CR1 sequence data. The x-axis represents the 1st principle component and the y-axis, the 2nd principle component. The scales for each axis are Eigen values which are proportional to the percentage of variance corresponding to the principle components.

3.2 Y chromosome studies

3.2.1 Variation at individual loci

The frequencies of the YAP and the alleles derived for the six STR markers, DYS393, DYS19, DYS390, DYS391, DYS388 and DYS392 in the Mozambican population group are given in Table 9. Data for the other populations (the Lemba, Remba and Venda) were generated by Dr. B. Morar and D. Barkhan from the PEG laboratory and are included here for comparison. YAP+ chromosome were found at high frequencies in the Mozambique and Venda groups, but at low frequencies in the Lemba and Remba (Table 9).

Although individual STR loci are not very reliable on their own, there are clear trends seen in the frequencies and distribution of the alleles among the four populations. At loci DYS393, DYS19 and DYS390, the Mozambicans and the Venda share the most frequent allele as do the Remba and Lemba. At locus DYS391, allele C is most common in all four groups. At loci DYS388 and DYS392, the Mozambicans and Venda share common alleles (12 and 11, respectively) with the Lemba and Remba. Allele 15, however, at both of these loci, is absent in the Mozambicans and the Venda but common in the Remba and Lemba.

Table 9. Allele frequencies of YAP and six Y chromosome STRs (continued over the page)

Locus	STR repeat size[1]	Population			
		Mozambique	Venda	Lemba	Remba
YAP					
N		36	46	47	56
+		0.861	0.783	0.298	0.179
-		0.139	0.217	0.702	0.821
DYS393					
N		36	46	42	57
D	12	0.000	0.043	0.463	0.807
E	13	0.667	0.696	0.293	0.070
F	14	0.250	0.152	0.195	0.123
G	15	0.028	0.087	0.024	0.000
H	16	0.056	0.022	0.024	0.000
DYS19					
N		36	46	42	57
A	13	0.000	0.000	0.048	0.000
B	14	0.250	0.043	0.571	0.456
C	15	0.611	0.630	0.262	0.281
D	16	0.000	0.217	0.119	0.140
E	17	0.139	0.109	0.000	0.123
DYS390					
N		36	46	41	57
D	20	0.000	0.000	0.049	0.000
E	21	0.833	0.717	0.171	0.193
F	22	0.056	0.043	0.146	0.228
G	23	0.000	0.022	0.122	0.088
H	24	0.306	0.196	0.488	0.491
I	25	0.056	0.022	0.024	0.000
DYS391					
N		36	46	41	57
B	9	0.000	0.022	0.000	0.000
C	10	0.750	0.533	0.756	0.877
D	11	0.222	0.043	0.244	0.123
E	12	0.028	0.000	0.000	0.000

Table 9 (continued). Allele frequencies of YAP and six Y chromosome STRs

Locus	STR repeat size[1]	Population			
		Mozambique	Venda	Lemba	Remba
DYS388					
N		33	49	26	44
10	10	0.242	0.163	0.077	0.000
11	11	0.000	0.000	0.000	0.000
12	12	0.667	0.714	0.423	0.386
13	13	0.061	0.081	0.000	0.000
14	14	0.030	0.000	0.000	0.227
15	15	0.000	0.000	0.346	0.386
16	16	0.000	0.041	0.154	0.000
DYS392					
N		35	49	26	44
10	10	0.143	0.204	0.115	0.000
11	11	0.800	0.939	0.308	0.659
12	12	0.057	0.000	0.000	0.000
13	13	0.000	0.000	0.192	0.045
14	14	0.000	0.000	0.000	0.045
15	15	0.000	0.000	0.385	0.250

[1]Redd *et al.*, (1997)

Blue colours highlight common alleles at each locus

3.2.2 Haplotype analysis

Y chromosome haplotypes were constructed using the YAP in conjunction with alleles detected at six STR loci in the order YAP, DYS393, DYS19, DYS390, DYS391, DYS388 and DYS392. Seventy-three unique haplotypes were derived, 37 of which were found on YAP+ chromosomes (Table 10) and 36 on YAP- chromosomes (Table 11). Of the 37 YAP+ haplotypes, 30 were unique to one population, seven found in more than one group and only one of these seven (haplotype 16, Table 10) common to all four population groups. Haplotype 16 was also the most frequently occurring haplotype, suggesting that it might represent the ancestral YAP+ chromosome among these population groups. The most YAP+ haplotype sharing occurred between the Mozambicans and the Venda who shared five common haplotypes (Table 12). Of the 36 unique YAP- chromosomes, four were found in more than one group, three of which were shared by the Remba and Lemba and one by the Mozambicans and the Venda (Table 12). None of the YAP- haplotypes found in the Lemba or Remba were present in individuals from Mozambique or the Venda. This suggests that the YAP- chromosomes in the Lemba and Remba might have a different origin to those in the Mozambicans and the Venda. The haplotype diversity values for the four population groups ranged from 0.8277 to 0.9664 (Table 12).

Table 10. YAP+ haplotypes found in the Mozambicans (MOZ), Venda (VE), Lemba (LE) and Remba (RE) based on STR loci DYS393, DYS19, DYS390, DYS391, DYS388 and DYS392.

Haplotype	Repeat sizes						MOZ	VE	LE	RE
	393	19	390	391	388	392	(n = 35)	(n = 44)	(n = 40)	(n = 54)
1	12	14	22	10	12	11				2
2	12	14	23	10	12	11				1
3	12	15	21	10	12	11		1		
4	12	15	21	10	16	11		1		
5	12	15	24	11	12	11				1
6	13	13	24	10	12	11			1	
7	13	14	21	11	12	11		1		
8	13	14	24	10	12	11	1			
9	13	14	24	10	14	12	1			
10	13	14	24	11	10	10	1			
11	13	14	24	11	12	11	2			
12	13	14	25	10	12	11	1		1	
13	13	15	20	11	12	12			2	
14	13	15	21	10	10	11	1			
15	13	15	21	10	12	10	1			
16	13	15	21	10	12	11	5	18	2	2
17	13	15	21	11	12	11	4		2	
18	13	16	21	10	12	11			1	
19	13	16	23	9	13	11		1		
20	13	16	24	10	13	11			1	
21	13	17	21	10	12	11	2			
22	13	17	24	10	13	11	1	2		
23	14	13	24	10	12	11			1	
24	14	14	25	11	12	11	1			
25	14	15	21	10	10	11	1			
26	14	15	21	10	12	11	2	4	1	
27	14	15	21	12	12	11	1			
28	14	15	22	10	12	11	1			
29	14	16	21	10	12	11		1		
30	14	17	21	10	12	11	1	1		4
31	14	17	24	10	13	11		1		
32	15	15	21	10	12	11		2		
33	15	17	21	10	12	11	1	1		
34	16	14	21	10	12	11	1			
35	16	15	21	10	12	11	1			
36	16	16	21	10	12	11		1		
37	16	16	22	10	12	11			1	
Total YAP+ haplotypes							**30**	**35**	**13**	**10**

Table 11. YAP- haplotypes found in the Mozambicans (MOZ), Venda (VE), Lemba (LE) and Remba (RE) based on STR loci DYS393, DYS19, DYS390, DYS391, DYS388 and DYS392.

Haplotype	Repeat sizes						MOZ	VE	LE	RE
	393	19	390	391	388	392	(n = 35)	(n = 44)	(n = 40)	(n = 54)
1	12	14	22	10	12	11				1
2	12	14	22	10	12	14				2
3	12	14	22	10	12	15			4	8
4	12	14	23	10	14	11				4
5	12	14	23	10	16	11			1	
6	12	14	24	10	12	13				2
7	12	14	24	10	15	11			6	5
8	12	14	24	10	15	13			1	
9	12	14	24	10	16	13			4	
10	12	15	24	10	15	11				5
11	12	15	24	11	12	15				1
12	12	15	24	11	14	11				1
13	12	15	24	11	15	11			2	4
14	12	16	24	10	12	15				1
15	12	16	24	10	14	11				4
16	12	16	24	10	15	11				1
17	13	14	24	10	13	11	1			
18	13	15	21	10	12	15				1
19	13	15	22	10	12	10		1		
20	13	15	24	10	10	10	1			
21	13	15	24	10	10	11	2	1		
22	13	15	24	10	12	11	1			
23	13	16	21	10	12	11		1		
24	13	16	21	10	15	11				1
25	13	16	22	10	12	10		2		
26	13	16	24	10	10	10			2	
27	13	16	24	10	10	11		2		
28	13	16	24	10	16	11		1		
29	13	17	24	10	10	11		1		
30	14	14	23	10	12	15			2	
31	14	14	23	11	12	15			3	
32	14	14	24	11	12	15			1	
33	14	15	22	10	12	10			1	
34	14	17	21	10	12	11				1
35	14	17	21	10	14	11				1
36	14	17	21	10	15	11				1
Total YAP- haplotypes							**5**	**9**	**27**	**44**

Table 12. Haplotype diversity and haplotype sharing among YAP+ (red) and YAP- (blue) chromosomes in the Mozambicans,Venda, Lemba and Remba.

	n (YAP+)	Mozambique	Venda	Lemba	Remba
n (YAP-)					
Mozambique		0.9664±0.0173			
Venda			0.8277±0.0558		
Lemba				0.9538±0.0161	
Remba					0.9469±0.0132

3.2.3 Y chromosome networks

Median joining networks were constructed to show the evolutionary relationships of YAP+ and YAP- chromosomes (Figures 12 and 13, respectively). The 37 YAP+ haplotypes are separated into two major clusters joined by two haplotypes (haplotypes 1 and 2) found only in the Remba. Both clusters, however, contain haplotypes found in each of the four population groups. Seventeen of the 21 haplotypes in cluster 1 are connected to each other via single mutational steps and the remaining four via two mutational steps suggesting a recent common ancestry for these YAP+ chromosomes. Haplotype 16 is not only found in each of the four population groups as well as in the greatest number of individuals, it also has the most single-step derivatives which provides additional support for haplotype 16 representing the ancestral YAP+ chromosome within this cluster.

Of the 36 YAP- haplotypes, those found in Lemba and Remba individuals are grouped closely together albeit in mainly two separate clusters. Although there are fewer YAP- chromosomes in the Mozambicans and the Venda, these haplotypes are also generally more closely associated with each other than to those found in the Lemba or Remba (Figure 13). Since there is largely distinct grouping of Lemba with Remba, and Mozambican with Venda haplotypes, respectively, and no YAP- haplotype common to the four population groups, it is likely that the YAP- haplotypes are not derived from a common ancestor but rather from two or more different lineages.

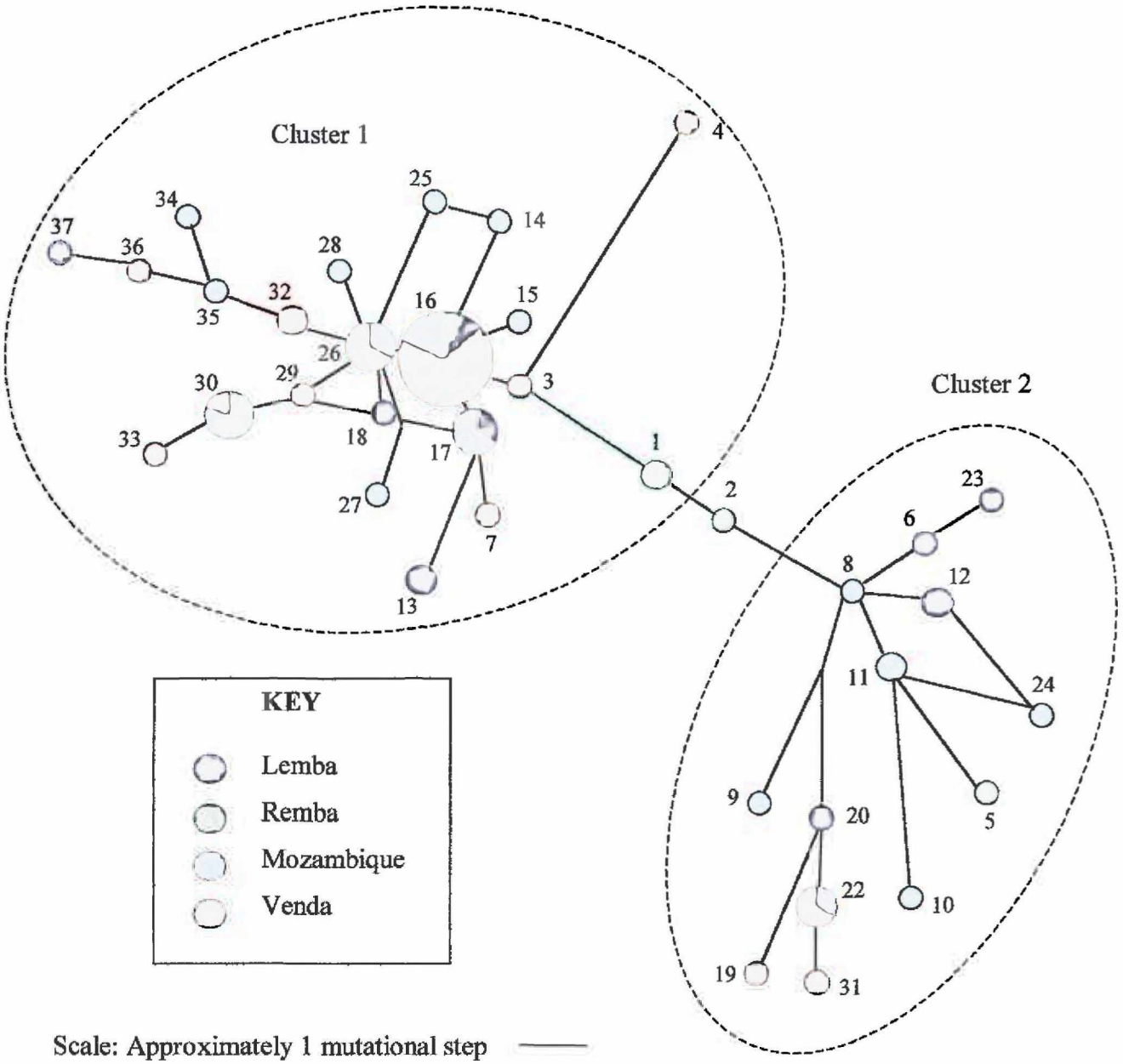
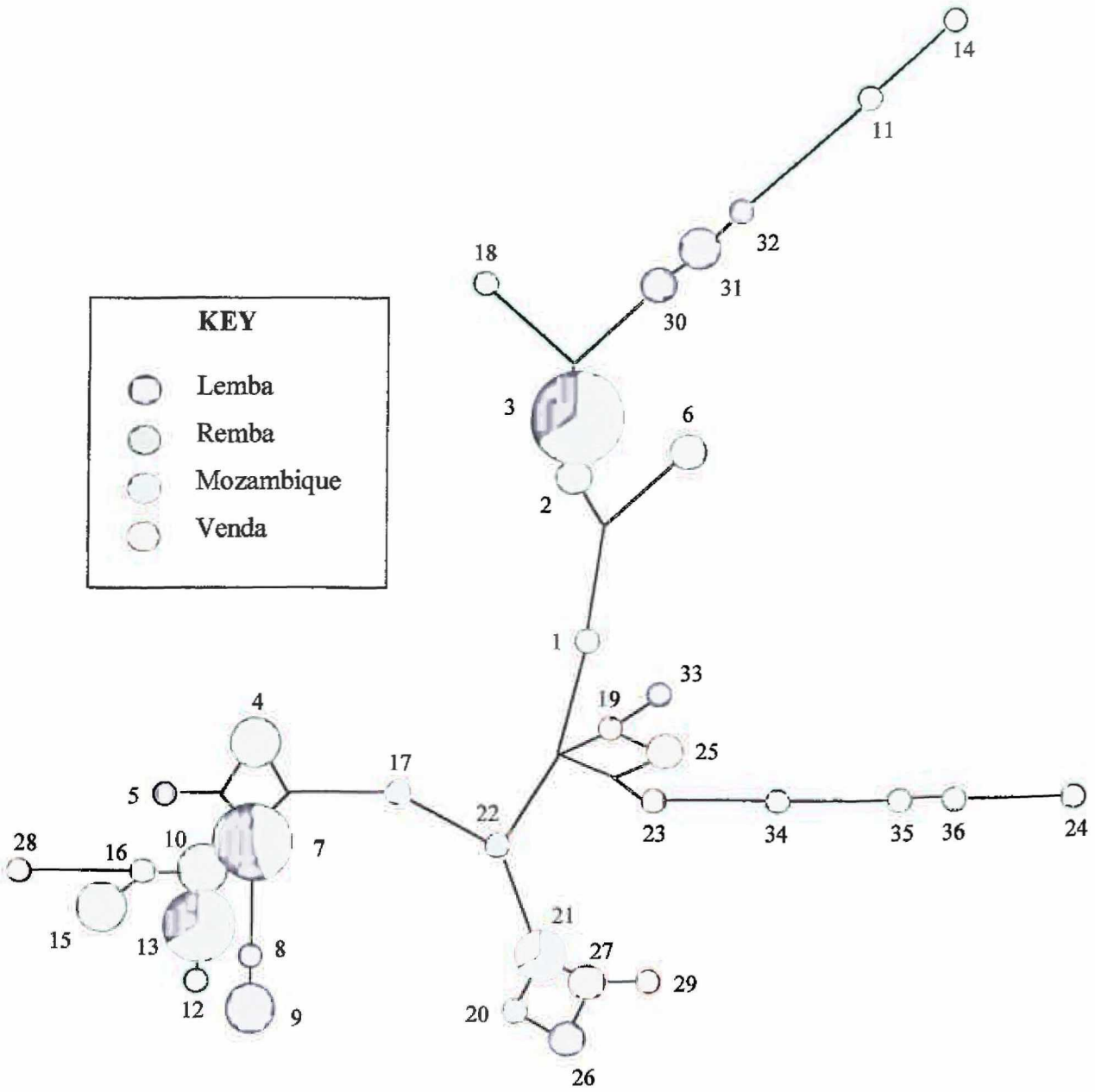


Figure 12. Median joining network showing the relationship of YAP+ haplotypes found in the Lemba, Remba, Mozambicans and Venda. The haplotype numbers correspond to those in Table 10. Haplotypes in Cluster 1 were found in 68 of the 88 YAP+ individuals (77.27%) and haplotypes in Cluster 2 were found in 17 individuals (19.32%).



Scale: Approximately 1 mutational step ——

Figure 13. Median network showing the relationship of YAP- haplotypes in the Lemba, Remba, Mozambicans and Venda. The numbered nodes correspond to the YAP- haplotypes in Table 11. Haplotype 5 is the CMH.

3.2.4 Frequency and distribution of the Cohen Model Haplotype

The frequency and distribution of the CMH, 12 14 23 10 16 11, (haplotype 5), derived from alleles at loci DYS393, DYS19, DYS390, DYS393, DYS388 and DYS392, respectively, was examined in the Remba, Lemba, Mozambicans and a South African Ashkenazi Jewish population. The CMH was absent in the Mozambicans and Remba but was found at frequencies of 0.02 (1/41) and 0.23 (7/31) in the Lemba and South African Jewish population, respectively (B. Morar, unpublished). A network was constructed connecting haplotypes closely associated with the CMH that were found in the Lemba, Remba, Yemeni, Sena and Ashkenazi and Sephardic Jewish populations by including data from Thomas *et al.* (2000) (Table13, Figure 14). Single-step derivatives of the CMH were found in Ashkenazi and Sephardic Jews and individuals from Yemen and Sena but neither in the Lemba nor the Remba. It, therefore, appears that the CMH was introduced into the Lemba more recently and has had insufficient time to gain additional mutations.

Table 13. YAP- haplotypes associated with the CMH in the Lemba, Remba and South African Jews(SAJ)[1] and Ashkenazi Jews (AJ), Sephardic Jews (SJ), Yemeni and Sena[2]

Haplotype	Repeat sizes						Population group
	393	19	390	391	388	392	
1(4)	12	14	23	10	14	11	Remba
2	12	14	23	10	15	11	Yemeni, AJ
3	12	14	23	10	16	10	SJ
4 (CMH) (5)	12	14	23	10	16	11	Lemba, Yemeni, AJ, SJ, SAJ
5	12	14	23	11	16	11	Sena
6 (7)	12	14	24	10	15	11	Lemba, Remba
7	12	14	24	10	16	11	Sena
8 (9)	12	14	24	10	16	13	Lemba
9	12	14	24	10	17	11	SAJ
10	13	14	23	10	16	11	Sena

[1]B. Morar, unpublished
[2]Thomas *et al.* (2000)
The numbers in brackets correspond to the haplotype numbering in Table 11.

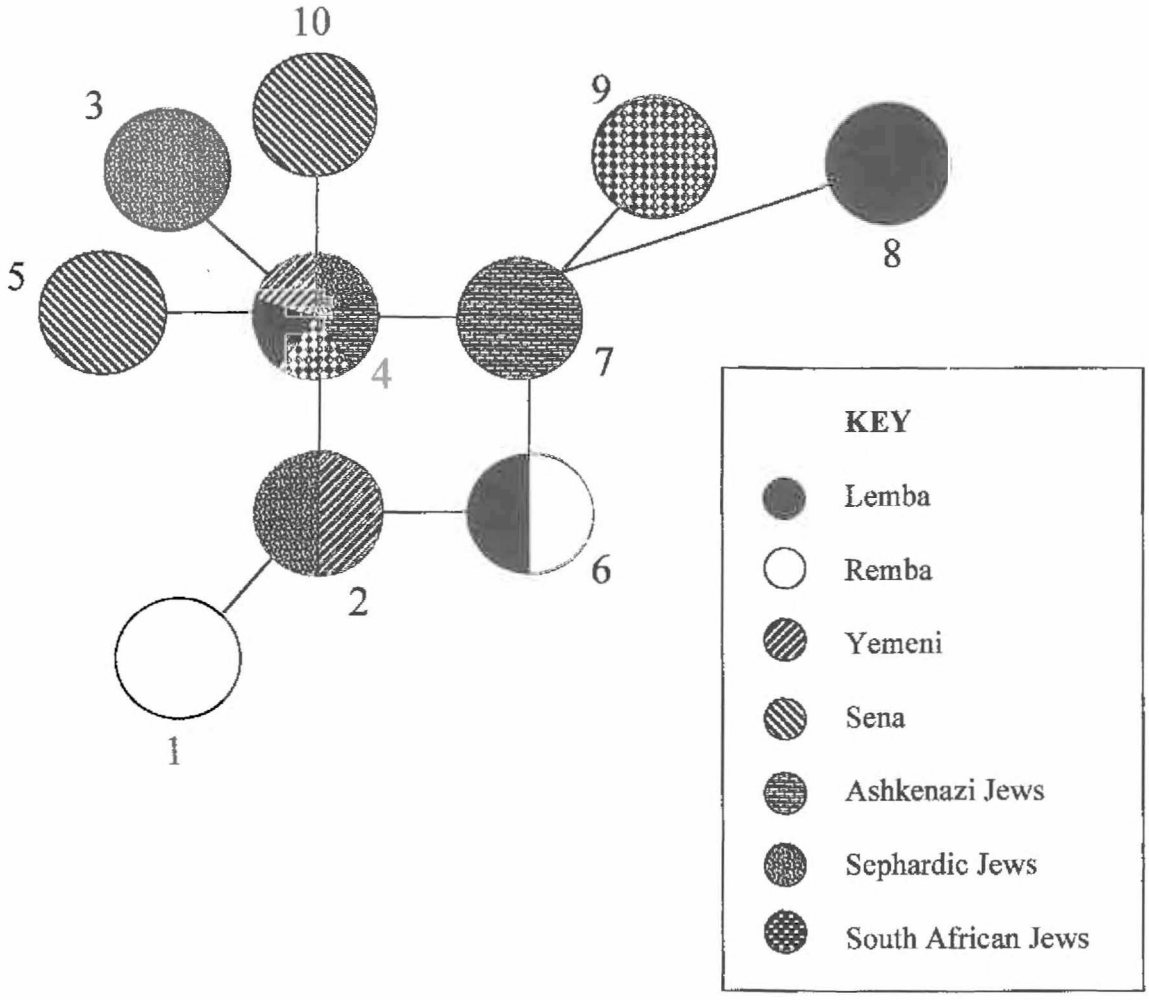


Figure 14. Network of YAP- haplotypes closely associated with the CMH in the Lemba, Remba, Yemeni, Sena, Ashkenazi and Sephardic Jewish populations, and a South African Ashkenazi Jewish population. The haplotypes are numbered according to those in Table 13. Haplotype 4 is the CMH.

3.2.5 Y Chromosome affinities of the Lemba, Remba, Venda and Mozambicans

The relationships of the four populations based on Y chromosome haplotype data was examined using principle co-ordinates analysis (Figure 15). The Mozambicans and the Venda were closely associated using the 1st principle component. The Lemba and Remba were also more closely related to each other than either is to the Mozambicans or the Venda.

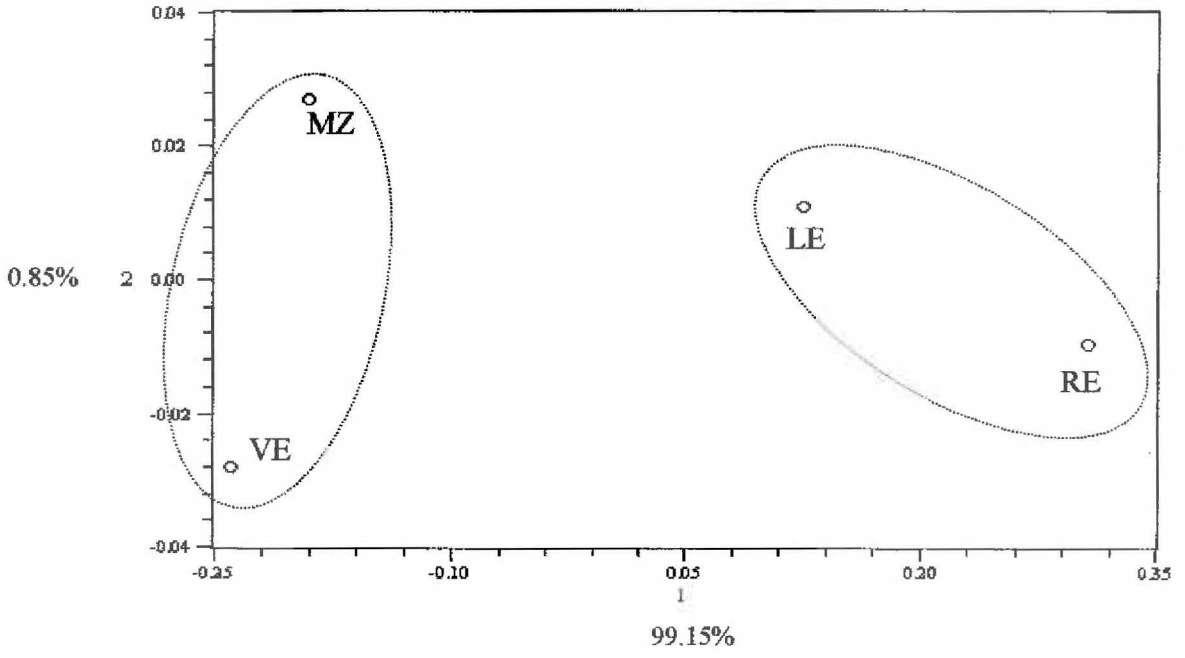


Figure 15. PCO-plot showing the genetic affinities of the Lemba (LE), Remba (RE), Venda (VE) and Mozambicans (MZ) using Y chromosome haplotype data.

3.3 Genetic structure of the Remba, Lemba, Venda and Mozambicans.

AMOVA was used to test population structuring at different hierarchical levels: linguistic (all four populations grouped together as SEB speakers), geographic (the Lemba and Venda grouped together as living among each other), and historical (the Remba and Lemba grouped together as sharing a history of being united as one population group). The results obtained for both mtDNA and Y chromosome variation at each level of testing are given in Table 14. In the instances where negative values were derived for the percentages of variance, the variance among groups is, in essence, negligible. In all levels of grouping for mtDNA, almost all variation was found within populations. Although most Y chromosome variation still occurred within populations, more variation, compared to mtDNA, was attributed to "among groups" and "among populations within groups". Furthermore, the only level of grouping supported by AMOVA was that of Y chromosome data for the Remba and Lemba (grouped together) versus the Mozambicans and Venda (grouped separately). These results suggests that while the Remba, Lemba, Mozambicans and Venda are likely to share a common ancestral mtDNA gene pool, Y chromosomes in the Remba and Lemba appear to have a common origin that is different from that of the other African populations.

Table 14. Percentage of variance at different hierarchical levels using AMOVA for mtDNA sequence data (CR1 and CR2) and Y chromosome haplotype data in the Remba, Lemba, Venda and Mozambicans.

Grouping	Among groups	Among populations within groups	Within populations
MtDNA			
All four populations	0.24 (P=0.290)	-	99.76
Lemba and Venda, Mozambicans, Remba	0.01 (P=0.313)	0 .24 (P=0.884)	99.75 (P=0.254)
Lemba and Remba, Mozambicans, Venda	-1.49 (P=0.681)	1.47 (P=0.510)	100.02 (P=0.281)
Y chromosome			
All four populations	22.77 (P<0.001)	-	77.23
Lemba and Venda, Mozambicans, Remba	-6.22 (P=0.296)	28.24 (P=0.054)	77.98 (P<0.001)
Lemba and Remba, Mozambicans, Venda	23.59 (P=0.188)	2.73 (P=0.024)	73.69 (P<0.001)

'P' is the probability of observing more extreme percentages of variance after 1000 randomizations than those generated in the analysis.

4. DISCUSSION

There were two major objectives to this study. The first was to use genetic markers to shed more light on the ancestry of the Lemba. The second objective was to investigate the genetic structure of the people of Mozambique and to gain some insight into their prehistory. To address these issues, Y chromosome and mtDNA data were used to determine the male and female genetic ancestries, respectively, in these two population groups.

4.1 The prehistoryof the Lemba

Although several genetic studies (Mourant *et al.*, 1978; Spurdle and Jenkins, 1992, 1996; Soodyall, 1993; Thomas *et al.*, 2000) have addressed Lemba origins, the results remain somewhat inconclusive. These studies have only been carried out on a Lemba group from South Africa and their mtDNA gene pool has been investigated using only RFLP data. Here we include a Lemba group from Zimbabwe, the Remba, who, according to historical evidence (Parfitt, 1997), were living with the Lemba as a united group until the late 17th century. In addition, mtDNA sequence data has been used to investigate the maternal history of the Lemba with greater resolution.

4.1.1 Origins of Y chromosomes in the Lemba and Remba

Y chromosome STR allele frequencies and haplotype data were used to assess the similarity of the Y chromosome gene pools of the Lemba and Remba and to test the historical evidence claiming that they share a common, non-African male ancestry. Obvious trends were

observed even in the frequencies and distribution of alleles at individual Y chromosome STR. Firstly the Lemba and Remba differed from the Mozambicans and the Venda by having predominantly YAP- chromosomes (70% in the Lemba and 82% in the Remba). The YAP insertion or YAP+ tends to be largely absent in Caucasoid populations and occurs at high frequencies in Negroids (Spurdle *et al.*, 1994). Secondly, at three Y-STR loci (DYS393, DYS19 and DYS390), the Lemba and Remba shared the most frequent allele, which at each locus was present at only low frequencies in the Mozambicans and the Venda (Table 9). Allele 14 at locus DYS19, occurring most frequently in the Lemba and the Remba, is common in European populations (Santos *et al.*, 1996; de Knijff *et al*, 1997), whereas allele 15 at DYS19 which was the most frequent allele in the Mozambicans and the Venda is predominant in African populations (Santos *et al.*, 1996; de Knijff *et al.*, 1997). At locus DYS391, allele 10 was the most common allele in all four groups. This allele occurs most frequently in several geographical regions, including Australia, New Guinea, Asia, the Arctic, Amerindian, Europe, India and Africa (de Knijff *et al.*, 1997). Allele 15 at both loci DYS388 and DYS392 occurred frequently in the Lemba and Remba but was absent in the Mozambicans and the Venda.

These data alone highlight a similarity in Y chromosome profiles between the Lemba and Remba that differs from the typically African Y chromosomes of the Mozambicans and the Venda. Furthermore the high proportion of YAP- chromosomes in the Lemba and Remba as well as the high frequency of allele 14 at locus DYS19 suggest that non-African populations have contributed to their Y chromosome gene pools.

Haplotype data confirmed these findings. Of the YAP- haplotypes, the Lemba and Remba shared three haplotypes with each other but none with either the Mozambicans or the Venda. Since the YAP+ haplotypes in the Lemba and Remba are likely to represent an African contribution to their Y chromosome gene pools, it is not all that surprising that they do share some YAP+ haplotypes with the Mozambicans and the Venda (Table 12). Even so, PCO analysis of both YAP+ and YAP- haplotypes distinctly separated the Lemba and Remba from the Mozambicans and the Venda (Figure 15).

A network of YAP+ haplotypes found in the Lemba, Remba, Mozambicans and Venda revealed no specific grouping or relationship between the Lemba and Remba (Figure 12). It did, however, separate the haplotypes into two major clusters, both of which contained haplotypes found in each of the four population groups. Most of the haplotypes in the first cluster were connected to each other via single mutational steps with one of these haplotypes (haplotype 16) found in all four population groups as well as in the greatest number of individuals. It also had the most single-step derivatives suggesting that it was the ancestral haplotype in that cluster. This haplotype was found to have the same properties in a network constructed from YAP+ haplotypes found in several South African SEB speaking population groups (B. Morar, unpublished). The presence of this haplotype and some of those closely associated with it in the Lemba and Remba confirms that the Lemba and Remba do share an African, and more specifically SEB, Y chromosome lineage.

YAP- haplotypes in the Lemba and Remba were grouped closely together in two separate clusters in a network constructed from YAP- haplotypes found in all four population groups (Figure 13). No YAP- haplotypes were common to all four groups and there was noticeable grouping of Lemba and Remba, and Mozambican and Venda haplotypes, respectively. This implies that all these YAP- haplotypes are not derived from a common ancestor but rather from two or more different lineages. Haplotypes 7 and 13 were found in more than one individual in both the Lemba and Remba (Figure 13). Haplotype 10 was found in five Remba individuals and four Lemba sampled by Thomas *et al.* (2000). These three haplotypes were connected via single mutational steps and were absent in the Mozambicans and the Venda examined here as well as South African SEB speakers (B. Morar, 2000, unpublished) and rare or absent in the Arabic and Jewish populations sampled by Thomas *et al.* (2000). This lends additional support to the STR frequency data that suggested a common male ancestry from a non-SEB speaking source for the Lemba and Remba.

We know that trade contacts between Bantu-speaking people and Arabs on the East Coast were established in about AD 900 in the area where the modern countries of Botswana, Zimbabwe and South Africa meet. This area was rich in elephant ivory and alluvial gold, and the traditional leaders, already wealthy in cattle, became incredibly wealthy through coastal trade (Huffman 1986). This increase in wealth gradually led to the development of the Zimbabwe culture which was defined on the basis of a settlement pattern first observed at Mapungubwe (AD 1220 to 1290). Although this capital was abandoned, the settlement pattern associated with the Zimbabwe culture was retained at Great Zimbabwe (1290 to

1450) (Huffman 1982). According to Lemba oral tradition, they were instrumental in the building and development of Great Zimbabwe. Archaeological and historical evidence has established that Great Zimbabwe controlled the interior segment of the long-distance trade network involving many countries around the edge of the Indian Ocean. Mainly ivory and gold, along with other trade items from the region, were taken to several trading posts along the east coast where Arabic Swahili traders exchanged the African products for glass beads, cotton and silk cloth, ceramics and other items (Huffman 1986).

According to van Warmelo (1974), the two princes of 'Omân, Sumaimân and Sa'îd were attacked by forces of the Khalif 'Abd al-Makik ibn Marwan of Damascus, and forced to flee to the land of east Africa in the late 7th century. Subsequently a wave of Arab immigrants settled along the coast, the name of their chief apparently Haji Sa'îd. Several Lemba clan names, used nowadays, like Bakali, Haji, Hamisi, Hasane, Sadiki and Salifo are derived from Arabic origins (van Warmelo 1974).

In light of the above, it is quite possible that the non-African contribution to the Y chromosome gene pool of the Lemba, observed in this study, was introduced by Arab immigrants and traders during the Great Zimbabwe era.

4.1.2 The CMH - evidence of a Judaic origin for the Lemba?

The presence of the CMH in the Lemba from South Africa led Thomas *et al.* (2000) to conclude that the Lemba have Jewish ancestry. If the CMH was in fact brought to Africa from a Jewish "Sena" in the north, it should be present in both the Remba and the Lemba, and, over time, single-step derivatives of the CMH should have arisen in both Lemba groups. While the CMH was found in 8.8% of the Lemba sampled by Thomas *et al.* (2000), it was found in only one Lemba individual from the present study (2.5%) and absent in the Remba. It was, however, present in 23% of the South African Jewish population (B. Morar, unpublished). Although single-step derivatives of the CMH were found in the Yemeni, Sena and Jewish populations (data from Thomas *et al.* (2000) included), the haplotypes found in the Remba and Lemba could only be connected to the CMH by at least two mutation events (Figure 14). This suggests that the CMH has been present in the Lemba for insufficient time to accumulate additional mutations.

There is the possibility that the presence of the CMH in the Remba was missed due to sampling. However, the absence of the CMH in the Remba *together* with the lack of single step derivatives of the CMH in both the Remba and the Lemba strongly suggests that the CMH was introduced into the Lemba more recently, presumably after they split from the Remba at least 200 years before the present. The major migration of European Jews to South Africa occurred between 1880 and 1910 when about 40 000 came from Eastern Europe, mainly Lithuania. Before this time individual Jewish men came from the same

region and made a living in South Africa as peddlers, supplying their goods to isolated farms across the country. It is possible that the Lemba acquired the CMH due to gene flow from South African Jews and not from an ancient Jewish population outside of Africa as suggested by Thomas *et al.* (2000).

4.1.3 Maternal ancestry of the Lemba and Remba

We screened for the presence of the 9-bp deletion in 94 Remba individuals and sequenced both hypervariable control regions in 67 Remba individuals to shed more light on their maternal ancestry. The sequence data were analyzed in conjunction with that from 44 Lemba and 39 Venda individuals (Soodyall, unpublished) as well as the Mozambican sequence data, produced in this study, for comparative purposes. Several lines of evidence from the mtDNA data suggest strongly that the Lemba and Remba share a common maternal ancestry with other African, particularly SEB-speaking, population groups.

4.1.3.1 The 9-bp deletion

The deletion was found in 18.1% of the Remba (present study) and 26.9% of the Lemba (Soodyall *et al.*, 1996). These frequencies fall within the range found in SEB-speaking population groups (5.7% - 26.9%) (Soodyall *et al.*, 1996). Furthermore, all the mtDNA sequence types associated with the deletion in the Remba and Lemba had eight particular substitutions (at positions 16148 C→T, 16172 T→C, 16188 C→G/A, 16230 A→G and

16320 C→T in CR1 and 93 A→G, 189A→G and 236 T→C in CR2) in common with each other as well as with a specific cluster of mtDNA types found in SEB-speakers, Pygmies and Malawians (Soodyall *et al.*,1996). Soodyall *et al.*(1996) referred to mtDNA types with this sequence profile as constituting the "major form" of the 9-bp deletion in Africa The Remba and Lemba, therefore, share at least one common mtDNA lineage with other African population groups.

4.1.3.2 Phylogeny of mtDNA types in the Remba and Lemba

We observed no trends in clustering in a NJ gene tree constructed from unique mtDNA types found in the Remba, Lemba and Venda (Figure 4). The Venda were included as they live in close association, geographically, with the Lemba. Unfortunately no Shona population groups, among whom the Remba live, were available for inclusion. Even when mtDNA types found in the Mozambicans were superimposed on the tree (Figure 6), there was no distinct grouping of any one or two population groups.

4.1.3.3 RFLP haplogroup distribution

All the mtDNA types identified in the Remba and Lemba corresponded to subsets of haplogroup L, which is almost exclusive to African populations, i.e. subhaplogroups L1a, L1i, L2, L3a and L3b. These subhaplogroups have varying frequencies and distribution patterns in different African populations (Watson *et al.*, 1997: Chen *et al.*, 2000). The same

subhaplogroups were present in the Venda and the Mozambicans with a very similar distribution pattern. This provides additional evidence suggesting that the Remba and Lemba have African-specific mtDNA sequences that are very closely related to those found in other African populations, including the Venda and Mozambicans. It also emphasises that the Remba and Lemba have not received any noticeable female genetic contribution from a non-African source.

4.1.3.4 Demographic history of the Remba and Lemba

Comparisons of the wave profiles produced in the mismatch distributions partially reveal the demographic histories of a population. The smoothness of the distribution is characterised by Harpending's (1994) raggedness statistic (table 8) and distinguishes distributions from constant-size populations and distributions that reflect expansions. The raggedness values derived from both CR1 data and CR1 and CR2 data (combined) suggest that both the Remba and Lemba have undergone population expansions. It has been suggested that while hunter-gatherer populations were constant in size over a long period of time, producing ragged distributions, agricultural populations increased in population size due to food production, producing smooth distributions (Watson *et al.*, 1996). It is known from archaeological studies, identifying settlement patterns, that SEB-speaking populations were indeed agriculturalists (Huffman, 1982).

The mismatch distributions in the Remba, Lemba and Venda have "wave" profiles similar to those found in other SEB-speakers from South Africa, where the first wave in the distribution was associated with a few mismatches due to sequences related to the 9-bp deletion (Barkhan, 1997). Similarly, when sequences associated with the deletion in the Remba, Lemba and Venda were removed for analysis, the graphs reverted to normal distributions. The population expansion times were similar for each group and ranged from 67 916 to 74 807 years ago. Excoffier and Schneider (1999) found that most populations show signs of expansion during the Pleistocene dating to around 70 000 years before the present. The close similarity of the mismatch distributions, and their summary statistics, derived from these population groups again suggests that their mtDNAs are derived from a common ancestry.

The evidence gleaned from the mtDNA data is in keeping with Lemba oral tradition. They claim that their male ancestors came to this country from a huge town somewhere across the seas to trade their goods. They began leaving some of their men behind with unsold cargo to establish trading stations. They moved further and further inland and only began taking local wives when they heard they could never return home as an enemy people had overpowered their city (van Warmelo, 1974).

4.2 The peopling of Mozambique

The few genetic studies that have been carried out on populations from Mozambique were unable to shed any light on the prehistory of these population groups. Even archaeologists

claim that Mozambique is somewhat of a blank page in their studies (Phillipson, 1993). We used Y chromosome and mtDNA data to investigate the ancestry of present-day Mozambican populations and determine to what extent their gene pool reflects contributions from other populations.

4.2.1 Origins of Y chromosomes in the Mozambicans

Eighty-six percent of the Mozambican males had YAP+ chromosomes. The presence of the YAP insertion occurs at high frequencies in African population groups (Spurdle *et al.*, 1994) which immediately suggests a typically African Y chromosome profile for the Mozambicans. The distribution of alleles at each of the six STR loci examined in the Mozambicans was very similar to that in the Venda. In fact, the two population groups shared the most frequently occurring allele at each locus (Table 9). The most predominant allele at locus DYS19, allele 15, in the Mozambicans and the Venda is most prevalent in the vast majority of African populations (Santos *et al.*, 1996; de Knijff *et al.*, 1997).

Haplotype and network analysis lends further support for a typically African Y chromosome ancestry of the Mozambicans. The Mozambicans shared more haplotypes with the Venda than did either the Lemba or Remba. In the network constructed from YAP+ haplotypes, 23 of the 30 YAP+ haplotypes in the Mozambicans were grouped closely together in Cluster 1 (Figure 12) around haplotype 16, the most likely ancestral haplotype of this cluster (discussed earlier in section 4.1.1). Although these haplotypes are not all connected via single

mutational steps, most of the intermediate haplotypes are found in the Venda (Figure 12) or South African SEB speaking populations (B. Morar, 2000, unpublished). Since haplotype 16 and several single-step derivatives of it are also found in SEB-speaking population groups from South Africa (B. Morar, 2000, unpublished), it appears that most of the Mozambican Y chromosomes share a common lineage with the Venda and these SEB-speaking population groups. In addition, the Mozambicans were most closely associated with the Venda in PCO analysis of Y chromosomes haplotypes (Figure 14).

The five YAP- haplotypes found in the Mozambicans did not cluster with the Lemba and Remba haplotypes which appear to be of Semitic origin. There is, therefore, no evidence that might be suggestive of a non-African contribution to the Y chromosome gene pool of the Mozambicans.

4.2.2 Maternal ancestry of the Mozambicans

As for the Lemba and Remba, the mtDNA data derived from Mozambican individuals implies a maternal African ancestry for the Mozambicans that is shared by the Venda, Lemba and Remba.

4.2.2.1 The 9-bp deletion

The frequency of the 9-bp deletion in individuals from Mozambique was somewhat lower than the frequencies observed in the Remba, Lemba and Venda at 9.9%. Of the SEB-speaking population groups sampled by Soodyall *et al.* (1996), the deletion was found at the lowest frequencies in the Tsonga (5.7%) and the Nguni (12.9%). The Tsonga have settled in South Africa on the escarpment of the Drakensberg (Nurse *et al.*, 1985) and also occupy territory in Mozambique south of the Save River with the Shangana (Reys *et al.*, 1970). In the relatively recent past there has been a considerable amount of admixture between the Tsonga and the Nguni (Bryant, 1924; Murdock, 1959). It has also been suggested that the Shangana represent a drift from the Nguni into the Tsonga (Reys *et al.*, 1970). The Shangana, Tsonga and their closely related ethnic groups (the Bitonga and Ronga) comprise a quarter of the Mozambican individuals sampled here so it is not surprising to find that the frequency of the 9-bp deletion in the Mozambicans lies between that of the Tsonga and the Nguni. The five mtDNA types found in Mozambican individuals with the deletion share the same substitutions common to the Remba, Lemba, Venda and South African SEB speaking population groups that represent the major form of the 9-bp deletion in Africa (Soodyall *et al.*, 1996). This suggests that these mtDNA types, too, are derived from the same ancestral type.

The most likely source of origin of the deletion appears to be central Africa, where the deletion is found in both Negroid and Pygmy populations. Since the deletion is restricted to

central and southern African Negroid and Pygmy populations, it is unlikely that the deletion was found in an ancestor common to both of these populations in which case there would be a wider distribution of the deletion in Africa (Soodyall *et al.*, 1996). It has, therefore, been suggested that the deletion originated in either the Pygmy or Negroid group and was then assimilated into the other group by gene flow (Soodyall *et al.*, 1996). The distribution and frequencies of the 9-bp deletion in the sub-Saharan African populations sampled by Soodyall *et al.* (1996) suggested that its spread in Africa might be due to the recent "Bantu expansion". The frequencies of the deletion in the Mozambicans as well as the Remba, Lemba and Venda and the control-region sequences associated with the deletion suggest that the females carrying the 9-bp deletion that contributed to the mtDNA gene pool of these populations have Bantu-speaking ancestors.

4.2.2.2 Phylogeny of mtDNA types in the Mozambicans

We derived 47 unique mtDNA types from 61 CR1 and CR2 sequences (Figure 5). Only eight types were shared by more than one individual and, as mentioned earlier (section 4.1.3.2), no obvious trends in clustering were seen in a NJ gene tree constructed from mtDNA types found in the Mozambicans, Venda, Lemba and Remba (Figure 6). This suggests that we could, essentially, be seeing the "within population" variation in one hypothetical population ancestral to each of these groups. The fact that so little type sharing was observed among mtDNA types from any of the four population groups emphasises the heterogeneity of their mtDNA gene pools.

4.2.2.3 RFLP haplogroup distribution

The mtDNA types identified in the Mozambicans corresponded to the same subsets of haplogroup L as those in the Remba, Lemba and Venda with a very similar pattern of distribution. The Mozambicans, however, had a lower frequency of mtDNA types corresponding to subhaplogroup L1a which, in this study, represents only those mtDNA types associated with the 9-bp deletion. Given the low frequency of the 9-bp deletion among the Mozambicans, this observation is not that surprising. So, as for the Remba and Lemba, there is no evidence to suggest a non-African contribution to the Mozambican mtDNA gene pool.

4.2.2.4 Demographic history of the Mozambicans

Again, similar results were obtained for the Mozambicans as for the Remba, Lemba and Venda (discussed in section 4.1.3.4). The raggedness value derived from the Mozambican CR1 and 2 sequence data (0.003) suggests that the population has not remained constant in size and reflects a population expansion event that was dated to 70 754 years before the present. This data falls within the range calculated for the Remba, Lemba and Venda (67 916 to 74 807 years ago) and re-emphasises the likelyhood of a common mtDNA ancestry in these populations.

Thus, both Y chromosome and mtDNA data reveal a typically African genetic profile for the Mozambican population group. Despite the crucial position Mozambique occupied during the times of trading and slavery by linking the interior to coastal ports, there is no evidence to suggest any genetic contribution from non-African sources. The Lemba claim to have spent a transient phase in present-day Mozambique, where they built their third "Sena" and apparently some remained in the area. Y chromosome data does not show any link between the Lemba or Remba and the Mozambicans sampled in this study to support this claim. The mtDNA data could not be used to address this issue since the maternally derived gene pools of the Mozambicans, Lemba and Remba are likely to share a common ancestry. However, this does not rule out the possibility that some Lemba groups do still live elsewhere in Mozambique.

4.3 Genetic affinities of the Remba, Lemba, Mozambicans and Venda with other sub-Saharan African populations

MtDNA sequence data was used to examine the genetic affinities of the population groups examined here with other sub-Saharan African populations and hopefully shed more light on migrations associated with the "Bantu expansion" in southern Africa. A NJ population tree indicated very little genetic difference among the Remba, Lemba, Mozambicans, Venda and SEB speakers from South Africa, the Kikuyu, Fulbe, Yoruba and Mandenka, who all also speak languages belonging to the Niger-Kordofanian language group, and the Songhai and Kanuri who speak Nilo-Saharan languages (Watson *et al.*, 1997, and references therein)

4.4 Genetic structure of the Remba, Lemba, Mozambican and Venda population groups

AMOVA was used to test population structuring at different hierarchical levels: linguistic (all four populations grouped together as SEB-speakers), geographic (the Lemba and Venda grouped together as living among each other), and historical (the Remba and Lemba grouped together as sharing a history of being united as one population group). In all levels of grouping for mtDNA, almost all variation was found within populations(99.76%), with only 0.24% of total variance attributed to "among groups" (Table 14). This value for the proportion of variation within populations is extremely high compared to the values obtained from mtDNA studies in other populations. The percentage of variance within Khoisan speaking populations was 92.88% (Soodyall, unpublished) and that derived from a survey of 10 global populations was 80.68% (Excoffier *et al.*, 1992). This lends even more support for a common ancestral mtDNA gene pool in the Remba, Lemba, Mozambicans and Venda.

In contrast, the proportion of Y chromosome variance within populations is a lot lower than similar values obtained in Khoisan speakers (89.93%) (Soodyall, unpublished) and Finnish population groups (90.01%) (Kittles *et al.*, 1998). Furthermore, the only level of grouping supported by AMOVA was that of Y chromosome data for the Remba and Lemba (grouped together) versus the Mozambicans and Venda (grouped separately). This supports previous conclusions that the Y chromosome gene pools of the Remba and Lemba have a common origin that is different from that of other African populations.

4.5 Genetic diversity in the Mozambicans and the Remba

Gene diversities within each of the four populations, although similar to one another for each type of marker, were slightly higher for mtDNA (average, 0.987) than Y chromosome data (average, 0.924). Although a comparison of mtDNA and Y haplotype diversities is not strictly valid, it has been argued that such a comparison is legitimate (Perez-Lezaun *et al.*, 1999). It may be argued that sequence polymorphisms and STRs evolve at very different rates and thus prevent meaningful comparisons. However, Perez-Lezaun *et al.* (1999) showed that mutation rates in mtDNA control regions do not seem to be significantly slower than those of Y chromosome STRs, and, in populations that have split during relatively recent times (as have the four populations studied here), a differential accumulation of genetic variation due to different mutation rates is not likely. Although diversity values depend on the number of loci considered, since neither mtDNA nor the NRPY chromosome undergo recombination, both markers can be regarded as essentially having one locus with multiple haplotypes.

Various reasons could account for the differential behaviour of the Y chromosome haplotype data: (i) intrinsic properties of Y chromosome STRs; (ii) a different effective population size for males versus females; and (iii) a different pattern of male versus female migration. Perez-Lezaun *et al.* (1999) noticed a similar trend in their data and since the differentiation of Y chromosome and autosomal STRs are almost identical (Perez-Lezaun *et al.*, 1997), excluded intrinsic properties as being responsible for the high level of interpopulation diversity shown

by Y chromosome STRs. Instead, they considered a relative reduction of effective population size for males generated by gender-specific migrations.

A traditional structure of patrilocal marriages, where it is customary for the bride to move to the groom's village, would result in a very high female to male ratio of short distance migration. Although the Lemba and Remba favour endogymy, they claim to have taken wives from local populations after they arrived in Africa and still, to some extent, practise patrilocal marriages albeit through rigorous initiation ceremonies (van Warmelo, 1974; Parfitt, 1997 pgs158 and 159). The Venda have had a close relationship with the Shona and then with the Lemba since they travelled south together from Great Zimbabwe, learning skills such as metallurgy and adopting cultural practises, like circumcision, from their apparent leaders. It is, therefore, quite likely that patrilocal marriages occurred among the Venda as well. It is known that both the Shona and Thonga of Mozambique favoured patrilocal residence (Murdoch, 1959). This mechanism could explain the observed difference, in interpopulation differentiation and intrapopulation gene diversity, between Y chromosome and mtDNA markers. It has also been invoked by Salem *et al.* (1996), to explain similar patterns of Y chromosome and mitochondrial diversity in Sinai Bedouins, and by Sielstad *et al.* (1998), in a global survey.

5. CONCLUSIONS

The non-recombing portion of the Y chromosome and mtDNA were used to examine the contribution of males and females, respectively, to the gene pool of the Remba, (a Lemba group from Zimbabwe), and present-day Mozambique populations.

By examining the Remba from Zimbabwe as well as the Lemba from South Africa we found that both groups shared Y chromosome haplotypes that were connected via single mutational steps. This suggests that they do have a common ancestry, presumably that linked with trade along the east African coast associated with Arab traders. The CMH is not as common in the Lemba as reported by Thomas *et al.* (2000) and is, in fact, absent in the Remba. Since the CMH is found at high frequencies among South African Jews, and Jewish traders are known to have been in the Louis Trichardt region since the late 19th century, it is possible that the CMH was introduced into the Lemba, after they split from the Remba, by gene flow from South African Jews. There are no one-step derivatives of the CMH in the Lemba or the Remba which supports this theory of its recent introduction into only the Lemba. It is, therefore, premature to use the CMH as evidence of Jewish ancestry in the Lemba.

No evidence can be gleaned from the mtDNA data to suggest any non-African contribution to the mtDNA gene pool of the Remba and Lemba or the Mozambicans. There is, however strong evidence suggesting that the Remba, Lemba, Mozambicans, Venda, SEB speaking population groups from South Africa and several other sub-Saharan African populations share a recent common maternal ancestry that most likely has its roots in the Benue-Niger region prior to the Bantu expansion.

Despite the role the Mozambicans may have played during the days of slavery and trading, their Y chromosome profile shows no reflection of any contribution from non-African populations.

6. REFERENCES

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R and Young IG (1981). Sequence and organization of the human mitochondrial genome. Nature **290**:457-465.

Aris-Brosou S and Excoffier L (1996). The impact of population expansion and mutation rate heterogeneity on DNA sequence polymorphism. Mol. Biol. Evol. **13**:494-504.

Bandelt HJ (1995). Mitochondrial portraits of human populations using median networks. Genetics **141(2)**:525-535.

Barkhan D (1997). **The evolutionary history of pygmy populations in sub-Saharan Africa**. BSc Honours thesis, University of the Witwatersrand, South Africa.

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.

Bleek D (1927). The distribution of Bushman languages in South Africa. In **Festchrift Meinhof**. Routledge and Kegan Paul, London. pp 419-420.

Brown MD, George M and Wilson AC (1979). Rapid evolution of animal mitochondrial DNA. Proc. Natl. Acad. Sci., USA.**76**:1967-71.

Bullock C (1927). **The Mashona and the Matabele**. Juta, London.

Cavalli-Sforza LL (1986). **African Pygmies**. Academic Press, Inc, New York.

Chen Y-S, Torroni A, Excoffier L, Santachiara-Benerecetti AS and Wallace DC (1995). Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. Am. J. Hum. Genet. **57**:133-149.

Comas D, Calafell F, Mateu E, Perez-Lezaun A, Bosch E, Martinez-Arias R, Clarimon J *et al.* (1998). Trading genes along the silk road: mtDNA sequences and the origin of central Asian populations. Am. J. Hum. Genet. **63**:1824-1838.

Connoway HP (1976). **'n Studie van die Lemba in Venda**. Ongepubliseerde MA-verhanderling, RAU, Johannesburg.

Connoway HP (1978). Herkoms, geskiedenis en genealogie van die Lembasibbes in Venda. SATE **I(2)**:31-44.

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

Ellis NA (1991). The human Y chromosome. Develop. Biol. **2**:231-240.

Doke CM (1954). **The southern Bantu languages**. Oxford University Press for International African Institute, London.

Ehret C (1972). Bantu origins and history: critique and interpretation. Transafr. J. Hist. **11**:1-19.

Excoffier L, Smouse PE and Quatttro JM (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mtichondrial DNA restriction data. Genetics **131**:479-491.

Excoffier L and Schneider S (1999). Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. Proc. Natl. Acad. Sci. USA. **96**:10597-10602.

Felsenstein J (1985). Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39**:783-791.

Gayre of Gayre R (1972). **The Origin of the Zimbabwe Civilisation**. Galaxie, Salisbury.

Giles RE, Blanc H, Cann HM and Wallace DC (1980). Maternal inheritance of human mitochondrial DNA. Proc. Nat. Acad. Sci. USA. 77:6715-6719.

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

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

Guthrie M (1962). Some developments in the prehistory of the Bantu origins. J. Afr. Hist. **3**: 273-282.

Gyllensten UD, Wharton A, Joseffson and Wilson AC (1991). Paternal inheritance of mitochondrial DNA in mice. Nature **352:**255-257.

Hammer MF (1994). A recent insertion of an Alu family member 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 Chromosome DNA variation and the peopling of Japan. Am. J. Hum. Genet. **59**:951-962.

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

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

Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti S, *et al.* (2000). Jewish and Middle Eastern non-Jewish populations share a common pool of Y chromosome biallelic haplotypes. Proc. Natl. Acad. Sci. USA. **97**:6769-6774.

Harpending HC, Sherry ST, Rogers AR and Stoneking M (1993). The genetic structure of ancient human populations. Curr. Anthropol. **34**:483-496.

Harpending H (1994). Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum. Biol. **66**:591-600.

Hendrickx B (1991). The ancient origin of the Lemba (Mwenye). A critical overview of the existing theories. JOAS **3**:172-193.

Hertzberg M, Mickleson KNP, Serjeantson SW, Prior JF, Trent RJ (1989). An Asian-specific 9-bp deletion of mitochondrial DNA is frequently found in Polynesians. Am. J. Hum. Genet. **44**:504-510.

Huffman TN (1982). Archaeology and ethnohistory of the African Iron Age. Ann. Rev. Anthropol. **11**:133-150.

Huffman TN (1986). Iron age settlement patterns and the origins of class distinctions in southern Africa. Advances in World Archaeology **5**:291-338.

Huffman TN (1989). Ceramics, settlements and Late Iron Age migrations. Afr. Archaeol. Rev. 7:155-182.

Jaques AA (1931). Notes on the Lemba tribe of the Northern Transvaal. Ant **26**:245-251.

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

Jobling MA, Pandya A, Tyler-Smith C (1997). The Y chromosome in forensic analysis and paternity testing. Int. J. Legal Med. **110**:118-124.

Junod HA (1927). **The Life of a South African Tribe**. Macmillan, London.

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

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

Krings M, Stone A, Schmitz RW, Krainitzki H, Stoneking M and Pääbo S (1997). Neandertal DNA sequences and the origin of modern humans. Cell **90**:19-30.

Krings M, Geisert H, Schmitz RW, Krainitzki H and Pääbo S (1999). DNA sequence of the mitochondrial hypervariable region II from the Neanderthal type speciman. Proc. Natl. Acad. Sci. USA **96**:5581-5585.

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

Kuper H (1955). **The Shona**. Oxford University Press, London.

Laman K (1953). **The Kongo I**. Studia Ethnographica Upsaliensia 4.

Mandivenga EC (1989). The history and 're-conversion' of the Varemba of Zimbabwe. JRA **19(2)**:98-124.

Mathivha MER (1992). **The Basena/Vamwenya/Balemba.** Morester Printers

Miller S, Dyk DD and Pelesky HF (1988). A simple salting-out procedure for extracting DNA from human nucleated cells. Nucleic Acids Res. **16(3)**:1215.

Mourant AE, Kopec AC and Domaniewska-Sobczak K (1978). **The Genetics of the Jews.** Clarendon Press, Oxford.

Murdoch GP (1959). Shona and Thonga. In **Africa: its peoples and their cultural history.** McGraw-Hill, New York. Pp374-379.

Nebel A, Filon D, Weiss DA, Weale M, Faerman M, Oppenheim A and Thomas MG (2000). High-resolution Y chromosome haplotypes of Israeli and Palestinian Arabs reveal geographic substructure and substantial overlap with hapltoypes of Jews. Hum. Genet. **107(6)**:630-641.

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

Nei M and Jin L (1989). Variances of the average numbers of nucleotide substitutions within and between populations. Mol. Biol. Evol. **6**:290-300.

Nurse GT, Weiner JS and Jenkins T (1985). **The Peoples of Southern Africa and their Affinities**. Clarendon Press, Oxford.

Oliver RO (1966). The problem of the Bantu expansion. J. Afr. Hist. 7:361-376.

Parfitt T (1997). **Journey to the Vanished City**. Phoenix, London.

Peretz H, Mulai A, Usher S, Zivelin A, Segal A, Weisman Z, Mittleman M, *et al.* (1997). The two common mutations causing factor XI deficiency in Jews stem from distant founders: one of ancient Middle Eastern origin and another of more recent European origin. Blood 7:2654-2659.

Perez-Lezaun A, Calafell F, Sielstad M, Mateu E, Comas D, Bosch E and Bertrapetit J (1997). population genetics of Y-chromosome short tandem repeats in humans. J. Mol. Evol. **45**:265-270.

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

Phillipson DW (1993). **African Archaelogy**. Cambridge University Press, Cambridge.

Redd AJ, Takezaki N, Sherry ST, McGarvey T, Sofro ASM and Stoneking M (1995). Evolutionary history of the COII/tRNA^{Lys} intergenic 9 b.p. deletion in human mitochondrial DNAs from the Pacific. Mol. Biol. Evol. **12**:604-615.

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

Reys L, Manso C and Stamatoyannopoulos G (1970). Genetic studies on southeastern Bantu of Mozambique. I. Variants of glucose-6-phosphate dehydrogenase. Am. J. Hum. Genet. **22**:203-215.

Ritte U, Neufield E, Broit M, Shavit D and Motro U (1993). The differences among Jewish communities – maternal and paternal contributions. J Mol Evol **37**:435-440.

Robin ED and Wong R (1988). Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells. J. Cell. Physiol.**136**:507-513.

Roewer L, Arnemann J, Spurr NK, Grzeschik KH and 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, Arnemann J, Spurr NK, Grzeschik K-H and Epplen JT (1996). analysis of molecular variance (AMOVA) of Y chromosome specific microsatellites in two closely related human populations. Hum. Mol. Genet. **5**:1029-1033.

Rogers AR and Harpending H (1992). Population growth makes waves in the distribution of pairwise differences. Mol. Biol. Evol. **4**:406-425.

Rogers AR (1995). Genetic evidence for a pleistocene population explosion. Evolution **49**:608-615.

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

Salem A-H, Badr FM, Gaballah Mf and Paäbo S (1996). The genetics of traditional living: Y-chromosomal and mitochondrial lineages in the Sinai Peninsula. Am. J. Hum. Genet. **59**:741-743.

Santachiara-Benerecetti A, Semino O, Passarino G, Torroni A, Brdicka R, Fellous M and Modiano G (1993). The common Near Eastern origin of Ashkenazi and Sephardic Jews supported by Y-chromosome similarity. Ann Hum Genet **57**:55-64.

Santos FR, Pena SJD, Epplen JT (1993). Genetic and population study of a Y-linked tetranucleotide repeat DNA polymorphism with a simple non-isotopic technique. Hum. Genet. **90**:655-656.

Santos FR, Pena SJD, Tyler-Smith C (1995). PCR haplotypes for the human Y chromosome based on alphoid satellite variants and heteroduplex analysis. Gene **165**:191-198.

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

Schapera I (1946). **The Bantu speaking tribes of South Africa: an ethnographical survey**. Miller, Cape Town.

Schneider S, Keuffer J-M, Roessli D and Excoffier L (1997). Arlequin ver 1.1: a software for population genetic data analysis. Genetics and Biometry Laboratory, Dept of Anthropology and Ecology, University of Geneva.

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

Sherry S (1991). GELIN – an ASCII-based DNA sequence management programme, development version 1.0. PSU.

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.

Sielstad MT, Hebert JM, Lin AA and Underhill PA (1994). Construction of human Y-chromosomal haplotypes using a new polytmorphic A to G transition. Hum. Mol. Genet. **3**:2159-2161.

Sielstad M, Minch E and Cavalli-Sforza LL (1998). Genetic evidence for a higher female migration rate in humans. Nat. Genet. **20**:278-280.

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.

Slatkin M (1995). A measure of population subdivision based on microsatellite allele frequencies. Genetics **139**:457-462.

Soodyall H and Jenkins T (1992). Mitochondrial DNA polymorphisms in Khoisan populations from southern Africa. Ann. Hum. Genet. **56**:315-324.

Soodyal H and Jenkins T (1993). Mitochondrial DNA polymorphisms in Negroid populations from Namibia: new light on the origins of Dama, Herero and Ambo. Ann. Hum. Biol. **20**:477-485

Soodyall H, Vigilant L, Hill AV, Stoneking M and Jenkins T (1996). MtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. Am. J. Hum. Genet. **58**:595-608.

Spurdle A and Jenkins T (1992). Y chromosome probe p49a detects complex *Pvu*II haplotypes and many new *Taq*I haplotypes in southern African populations. Am J Hum genet **54**:107-125.

Spurdle AB, Hammer MF and Jenkins T (1994). 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 and Jenkins T (1996). The origins of the Lemba "Black Jews" of Southern Africa: Evidence from p12F2 and other Y chromosome markers. Am. J. Hum. Genet. **59**:1126-1133.

Stayt HA (1931). Notes on the Balemba (An Arabic-Bantu tribe living among the BaVenda and other Bantu tribes in the Northern Transvaal and Southern Rhodesia). Journal of the Royal Anthropological Institute of Great Britain and Ireland **26**:231-238.

Strachan T and Read AP (1996). Human Molecular genetics. BIOS Scientific Publishers.

Tamura K and Nei M (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. **10**:512-526.

Tajima F and Takezaki N 91994). Estimation of evolutionary distance for reconstucting molecular phylogenetic trees. Mol. Biol. Evol. **11**:278-286.

Thomas MG, Skorecki K, Ben-Ami H, Parfitt T, Bradman N and Goldstein DB (1998). Origins of Old Testament priests. Nature **394**:138-140.

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, Bradman N and Goldstein DB (2000). Y chromosomes travelling south: The Cohen Modal Haplotype and the origins of the Lemba – the "black Jews of southern Africa". Am. J. Hum. Genet. **66**:674-686.

Thompson LC (1942). The Ba-Lemba of Southern Rhodesia. NADA **19**:76-86.

Torroni A, Semino O, Scozzari R, Sirugo G, Spedini G, Abbas N, Fellous M, *et al.* (1990). Y chromosome DNA polymorphism in human populations: differences between Caucosoids and Africans detected by 49a and 49f probes. Ann Hum Genet **54**:287-296.

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

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

van Warmelo NJ (1974). The classification of cultural groups. In: W.D. Hammond-Tooke (ed) **The Bantu-speaking peoples of southern Africa**. Routledge and Kegan, London, pp 56-84.

Vansina J (1984). Western Bantu expansion. J. Afr. Hist. **25**:129-145.

Velez-Grilo VH (1960). Esboco dum guia etnografico de Moçambique. Bol. Inst. Invest. Cient. Moçambique. **1**:126-164.

Vigilant L (1990). **Control region sequences from African populations and the evolution of human mitochondrial DNA**. PhD Thesis. University of California, Berkeley.

Von Sicard H (1952). **Ngoma lungundu. Eine afrikanische Bundeslade**. Studia Ethnographica Upsaliensia V. Almquist & Wiksells Boktrycker, Uppsala.

Von Sicard H (1962). Lemba clans. NADA **39**:68-80.

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

Watson E, Forster P, Richards M and Bandelt H-J (1997). Mitochondrial footprints of human expansion in Africa. Am. J. Hum. Genet. **61**:691-704.

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

Westphal EOJ (1963). The linguistic prehistory of southern Africa: Bush, Kwadi, Hottentot, and Bantu linguistic relationships. Africa **33**:237-264.

Whittaker M and Reys L (1975). Plasma Cholinesterase studies on south-eastern Bantu of Mozambique. Hum. Hered. **25**:296-301.

Zoosman-Diskin A, Ticher, A, Hakim I, Goldwitch Z, Rubinstein A and Bonne-Tamir B (1991). Genetic affinities of Ethiopian Jews. Isr J Med Sci **27**:245-251.

7. APPENDIX

7.1 Recipes and solutions

10% APS

1g APS dissolved in 10ml dH_2O and stored at 4°C for a maximum of 2 weeks.

Bromophenol (Ficoll blue) dye

50g sucrose, 1.86g EDTA, 0.1g bromophenol blue and 10g Ficoll dissolved in ~50ml dH_2O. Made up to 100ml with dH_2O, stirred overnight and filtered.

BSA (10 mg/ml)

1g BSA dissolved in 10ml dH_2O. 1ml aliquots stored at –20°C.

Dextran/formamide dye

20mg Blue Dextran (Sigma) dissolved in 1ml formamide.

EDTA (0.5M) (pH 8.0)

186.1g EDTA mixed with 800ml dH_2O. pH adjusted to 8.0 with 10M NaOH solution while stirring. Made up to 1L once EDTA was completely dissolved.

Ethidium bromide (10mg/ml)

1g EtBr stirred into 100ml dH_2O until completely dissolved. Stored at 4°C, wrapped in aluminium foil.

Ficoll loading dye

50g sucrose, 10ml EDTA pH 7 (50mM), 0.1g Bromophenol blue (0.1%) and 10g Ficoll made up to 100ml with dH_2O

1kb ladder

10.9g ladder, 5µl Ficoll dye and 84µl 1XTE

2% Metaphor gel

2g Metaphor gel powder dissolved, over heat, in 100ml 1X TBE buffer.

2% NuSieve gel

2g NuSieve dissolved, over heat, in 100ml 1X TAE buffer.

4.3% Polyacrylamide gel

36g urea dissolved in 10.6ml de-ionized 40% bis-acrylamide, 10ml 10X TBE and ~50ml ddH_2O. Made up to 100ml with ddH_2O and filtered. Stored covered in foil at 4°C. 30µl TEMED (catalyst) and 250µl 10% APS (facilitates cross-binding) was added to 50ml 4.3% polyacrylamide gel (36cm by 25cm plates – sequencing gels) immediately before the gel was poured. For Y-STR gels (24cm by 25cm plates), 12µl TEMED and 100µl 10% APS was added to 20ml polyacrylamide gel.

50X TAE

121g Tris dissolved in 28.55ml glacial acetic acid and 50ml 0.5M EDTA (pH 8.0). Made up to 500ml with dH_2O.

10X TBE

108g Tris and 55g boric acid dissolved in 40ml 0.5M EDTA (pH 8.0) and 800ml dH_2O. Made up to 1L with dH_2O.

7.2 Size standards

7.2.1 One kb ladder

A 1 kb ladder was used to size PCR fragments on agarose gels.



7.2.2 ROX-350 (ABI) size standard

The peaks in the ABI ROX-350 size standard below were used to size Y-STR alleles, using the ABI genescan and genotyper software programmes. The fragment sizes corresponding to the peaks are, from left to right, 50, 75, 100, 139, 150, 160, 200, 250, 300, 340 and 350 bp.



7.3 Ethics clearance

This study was approved unconditionally by the Committee for Research on Human Subjects (Medical), with the protocol number M990912. Copies of the clearance certificates follow.

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

COMMITTEE FOR RESEARCH ON HUMAN SUBJECTS (MEDICAL)
Ref: R14/49 Eddy

CLEARANCE CERTIFICATE **PROTOCOL NUMBER** M990912

PROJECT Molecular Genetic Variation In Mozambican & Zimbabwean Populations

INVESTIGATORS Ms CA Eddy

DEPARTMENT Dept of Human Genetics, SAIMR

DATE CONSIDERED 991001

DECISION OF THE COMMITTEE *

Approved unconditionally

DATE **CHAIRMAN** .....(Professor P E Cleaton-Jones)

* Guidelines for written "informed consent" attached where applicable.

c c Supervisor: Dr H Soodyall
Dept of Dept of Human Genetics, SAIMR

Works2\Valn0015\HumEth97.wdb\M 990912

==========

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10001, 10th Floor, Senate House, University.

I/we fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee.

DATE 12/10/99 SIGNATURE .....

PROTOCOL NO.: M 990912

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

COMMITTEE FOR RESEARCH ON HUMAN SUBJECTS (MEDICAL)
Ref: R14/49 Eddy

CLEARANCE CERTIFICATE PROTOCOL NUMBER M990912

PROJECT Molecular Genetic Variation In Mozambican & Zimbabwean Populations

INVESTIGATORS Ms CA Eddy

DEPARTMENT Dept of Human Genetics, SAIMR

DATE CONSIDERED 991001

DECISION OF THE COMMITTEE *

Approved unconditionally

DATE CHAIRMAN .....(Professor P E Cleaton-Jones)

* Guidelines for written "informed consent" attached where applicable.

c c Supervisor: Dr H Soodyall
Dept of Dept of Human Genetics, SAIMR

Works2\ain0015\HumEth97.wdb\M 990912

==========

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10001, 10th Floor, Senate House, University.

I/we fully understand the conditions under which I am/we are authorized to carry out the abovementioned research and I/we guarantee to ensure compliance with these conditions. Should any departure to be contemplated from the research procedure as approved I/we undertake to resubmit the protocol to the Committee.

DATE 11/10/99 SIGNATURE .....

PROTOCOL NO.: M 990912

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

7.4 Table of mtDNA types derived from CR1 sequence data in the Remba, Lemba, Mozambicans and Venda

This data set was used to examine the population affinities of the four population groups with other sub-Saharan African populations (section 3.1.6).

Table 15. MtDNA types (CR1) found in the Remba (RE), Mozambicans (MZ), Lemba (LE) and Venda (VE)

	00000000 1111111111 1111111111 2222222222 2222222222 2222222222 2223333333 3333333333 33333	Populations			
	45667899 1122244666 6777788889 0011112223 3344455566 6677788999 9990000011 1222344555 56666				
	51341623 1446958246 8126957892 7923453480 4923926904 5604869012 3450145917 9057634456 94256	RE	MZ	LE	VE
		79	61	45	41
REF[1]	TATTCTTT CCTTGGCAAA CATCCCCCTC ATAGCACTCA CCCTTACCCC ACCGCCACCC ACCACTAATA GCTCGACCCT TCTCC				
MtDNA type[2]					
L2					
1	....C. .....T... ....T..T.. .T.....G..	9	7	7	5
2	....C. .....T... ....T..T.. .T.....G.. ....T.	1			
3	....C. .....T... .....G.... ....T..T.. .T.....G..	1			
4	C..... ....C. .....T... ....T..T.. .T.....G..				1
5	....C. .....T... ....T..T.. .T.....G.. ....C..		1		
6	....C. .....T... ....T..T.. .TT....G..	1			
7	....C. .....T... ....T.T..T.. .T.....G.. ..... .T...	1			1
8	....C. .....T.T. ....T..T.. .T.....G..	1			
9	....C. .....T.T. ..T..... ....T..... .T.....G..				1
10	....CT .....T... ....T..T.. .T..... ..... .T...	1			
11	....CT .....T... ....T..T.. .T.....		1		
12	....CT .....T... .....T.. ....T..T.. .T.....G.. ..... .T...	1			
13	.....C.. ....T.CT .....T... ....T..T.. .T.....G..		1		
14	....TCT .....T... ....T..T.. .T.....G..	1			
15	....CT ..G...T... ....T..T.. .T.....G..	1			
16	....CT .....T... ....T..T.. .T.....G..	2	4		
17	....CT .....T... ....T..... .T.....G..	1	2		1
18	....C. .....T... ....T..... .T.....G..			1	2
19	....T... ....T..... .T.....G..		1		1
20	.....C. ....T... ....T..... .T.....G..				1
21	....C. .....T... ....TT.... .T.....G..		1		
22	....C..... .....T... ....TT.... .T.....G..		1		
23	....T... ....TT.... .T.....G..		1	1	
24	....C. .....TC.. .....T... ....T..... .....G..		1		
25	....C. .....TC.. ....T.....			2	
26	....T... T...C..... ..... .TT.....	1			
27	....T... T...C..... ....T..... .TT.....	1			
28	....A..... ....T... ....T..... .T.....	1			
29	....T... ....T..... .T.....		1		1
30	....T... .....T. ....T.....	1			
31	....T... .....T ....T.....		1		
32	....T... .....T ....T..... .....C.			1	
33	....T... ....T..... .....C. ....C..	2			
34	.G..... .T..... .....CT .....T... .... T.....C. ....C..		1		

	45667899 51341623	1111111111 1122244666 1446958246	1111111111 6777788889 8126957892	2222222222 0011112223 7923453480	2222222222 3344455566 4923926904	2222222222 6677788999 5604869012	2223333333 9990000011 3450145917	3333333333 1222344555 9057634456	33333 56666 94256	Populations RE 79	MZ 61	LE 45	VE 41
REF[1]	TATTCTTT	CCTTGGCAAA	CATCCCCCTC	ATAGCACTCA	CCCTTACCCC	ACCGCCACCC	ACCACTAATA	GCTCGACCCT	TCTCC				
35	.....	..C.....	.....	.....T...	.....	....T.....	.....G..	.....	..C..	2			
36	.....	..C.....	.....	.....T...	.....	....T.....	.....C.	.....	..C..	4			
37	C	..C.....	.....	.....T...	.....	....T.....	.....	.....	..C..	1			
38	.....	..C.....	.....	.....T...	.....	....T.....	.....	.....	..C..	4	2		2
39	.....	....A.....	.....C.	.....	.....	....T.....	...G....C.	.....T..	.....	1			
40	.....	...C.....	.....T.C.	.....T...	.....T	..T.T.....	.....	.....C.	.....	3			
41	.....	...C.....	.....T.C.	.....T...	.....T	..T.T.....	G.....C.	.....	.....			2	
L1i													
42	C	....A.....	.....C.	.....GT...	.....	....T.....	.T.....C.	.....	CT...	1			
43	C	....A.....	.....C.	.....GT...	.....	....T.....	.T.....C.	.....	.T...		1		
44	.....	....A.....	.....C.	.....GT...	.....	....T.....	.T.....C.	.....	.T...	2			
45	.....	....A.....	.....C.	.....GT...	.....	....T.....	.T.....C.	.....TC	.T...			1	
46	.....	....A...G.	.....T.C.	.C....T...	.....	....T.....	GT.....C.	.....	.T...		1		
47	.....	.....	.....T.C.	.....T...	.....	....T.....	GT.....C.	.....	.T...		1		
48	.....	....A.....	.....T.C.	....T.T...	.....	C...TA..T.	.T.....C.	.....	.T...				2
49	.....	....A.....	.....T.C.	.....T...	.....	C...TG....	.T.....C.	.....	.T...		1		
50	....TC..	....AA....	.....T.C.	...A..T...	T.....	C...TG....	.T.....C.	.....	.T...	1			
51	....T...	....AA....	.....T.C.	...A..T...	T.....	C...TG....	.T.....C.	.....	CT...		2		
52	....T...	....AA....	.....T.C.	...A..T...	T.....	C...TG....	.T.....C.	.....	.T...		1	1	
53	....T...	....AA....	.....T.C.	...A..T...	T.....	C...TG...T	.T.....C.	.....	.T...			1	
54	.....	....A..G..	.....T.C.	..G...T..G	...C....	.....	.....	.....C.	.....			1	
55	.....	....A.....	.....T.C.	..G...T..G	...C....	.....	.....	....T...C.	.....				1
56	.....	....A.....	.....T.C.	..G...T..G	...C....	.....	.....T.	.....C.	.....		1		
57	...G...C	....A.....	.....T.C.	..G...T..G	...C....	.....	.....	.....C.	.....		1		
58	.....	....A.....	.....T.C.	.....T..G	...C....	.....	.....	.....C.	.....				1
59	.....	....A.....	....T.T.C.	.....T..G	...C....	.....	.....T..	.....C.	.....		2		
60	.....	.....	..C...T.C.	.....T..G	T..C....	.....	.....	.....C.	.....G....	.....	1		
61	.....	....A.....	.....T.C.	.....G	T..C....	.A.....	.....	.....C.	.....	.....		1	
62	.....	....A.....	.....C.	.....T..G	...C....	.....	.....T..	.....C.	.....	.....			1
63	.....	....A.....	.....T.C.	.....T...	.T.C....	.....	.....	.T.....C.	.....	.....		2	
64	.....	....A.....	.....T.C.	.....T..G	.T.C....	.....	.....	.T.....C.	.T.....	.....		1	
65	.....	....A.....	.....T.C.	.....T..G	.T.C....	.....	.....	.T.....C.	..C.....	.....		1	
66	C	....A.....	.....T.C.	.....T..G	.T.C....	.....	.....	.T.....C.	..C.....	.....		1	
67	.....	....A.....	.....T.C.	.....T..G	.T.C....	.....	.....	.T.....C.	.....	.....		1	1
68	.....	.....	.....T.C.	.....T..G	...C....	.T..T..T..	...G....C.	.....	.....	.....	1		
69	.....	.....	.....T.C.	.....T..G	...C....	.T.AT..T..	...G....C.	....A.....	.....	.....			1
70	.....	.....	.....T.C.	.....T..G	...C....	.T.AT..T..	...G....C.	.....	.....	.....			1
71	.....	C	.GC...T.C.	....T....G	.....	....T...A.	.....	.....C.	.....	.....		1	
72	.....	....A.....	.....T.C.	.....T..G	.....	....T.....	G.....C.	.T.....	.....	.....			1

	1111111111	1111111111	2222222222	2222222222	2222222222	2223333333	3333333333	33333	Populations			
	45667899	1122244666	6777788889	0011112223	3344455566	6677788999	9990000011	1222344555	56666				
	51341623	1446958246	8126957892	7923453480	4923926904	5604869012	3450145917	9057634456	94256	RE	MZ	LE	VE
										79	61	45	41
REF[1]	TATTCTTT	CCTTGGCAAA	CATCCCCCTC	ATAGCACTCA	CCCTTACCCC	ACCGCCACCC	ACCACTAATA	GCTCGACCCT	TCTCC				
L3a													
73	.....	.....	..C.....C.	.....T...	.....	.....	.....	.T.....	.....	2	1	3	
74	.....	.....	..C.....C.	.....T...	.....	.....	.....G	.T.....	.....	1			
75	.....C..	.....	..C.....C.	.....T...	.....	..T.....	.....	.T.....	.....	1			
76	.....	.....	.....	G.....T...	.....	.....	.....C.	.T.....	.....	1			
77	.....	.....	.....	.....T...	.....	.....	.....	.....	.....		2		
78	.....	.....	...T.....	.C....T...	.....	.....	.....	.....	.....		1		
79	.....	.....	...T.....	.....T...	.....	.....	.....	...T.....	.....				1
80	.....	.....	.....T....	.....T...	.....	.....	.....	...T.....	.....	1	1		1
81	.....	.....	.....T....	.....T...	.....	.....	.....	...T.....	....T	1			
82	.....	.....	.....	.....T...	.....	.....	.....	...T.....	.....	2	3		2
83	.....	....A.....	.....	...A..T...	.....	....T.....	.....	.....T.	..C..				1
84	.....	.A..A.....	.....	...A.....	.....	....T.....	.....	.....T..	.....				1
85	.....	.A..A.....	.....	...A..T...	.....	....T.....	.....	.....T..	.....	1	1	1	
L3b													
86	.....	..C.....	.....	.....T...	.....	.....	.....	.....	.....	1			
87	.....	..C.....	.....	.....T...	.....	.....	.....	A.....	.....		3		
88	.....	T.C.....	.....	.....T...	.....	.....	.....	.....T.	.....	1			
89	.....	T.C.....	.....	.....T...	.....	.....	.....	.....	.....				1
90	.....	..C.....	.....	.....T...	.....	.....	.....C.	.....	.....			1	
91	.....	..C.....	.....C.	.....T...	.....	....T.....	.....C..C.	.....	.....			2	
L1a													
92	.....	.....	..C...TGC.	.....T..G	.....	.....	.....C.	.T.....	.....		1		
93	.....	....A.T...	T.C...TGC.	.....T..G	.....	....T.....	G.....C.	.T.....	.....	3	3	4	1
94	.....	....A.T...	T.C...TGC.	.....T..G	.....	....T.....	G.....G.C.	.T....T...	.....	1			
95	.....	....A.T...	T.C...TGC.	.....T..G	.....	....T.....	G.....G.C.	.T.....	.....		1		
96	.....	....A.T...	..C...TGC.	.....T..G	.....	.....	.....C.	.T.....	.....	1			
97	C	.....T...	..C...TGC.	.....T..G	.....	.....	.....C.	.T.....	.....	3	1		
98	C	.....T...	..C...TAC.	....T.T..G	.....	.....G...	.....C.	.T.....	.....	1			
99	.....	.....T...	..C...TAC.	.....T..G	.....	.....	.....C.	.T.....	.....			2	
100	.....	.....T...	..C...TGC.	.....T..G	T.....	.....	.....C.	.T.....	.....	2			
101	C.C.....	.....T...	..C...TGC.	.....T..G	.....	.....	.....C.	.T.....	.....				1
102	.....	.....T...	..C...TGC.	.....T..G	.....	.....	.....C.	.T.....	.....	8	5	7	7

[1]Anderson *et al.*, 1981. The nucleotide positions are given without the prefixed "16".

[2]Unique mtDNA types (assigned numbers 1-102) are grouped according to the RFLP haplogroup to which they correspond.

ACKNOWLEDGEMENTS

There are a number of people and organisations I would like to thank for assisting, in some way or other, with this project:

My supervisor, Dr Himla Soodyall, for all her help, time and patience

Bharti Morar, for always giving so generously of her time, irrespective of how busy she was herself

Debbie Barkhan, Cindy Dorf and Sharmila Manga, not only for checking numerous sequences, but for always being available with time, help and support when they were needed

Professor Jenkins and all the members of the Human Genetics Department

All the individuals who donated blood and everyone who participated in the collection of blood samples

Professor Leopoldo da Costa, Dean of Research, Univeridade Eduardo Mondlane for facilitating the collection of the Mozambican samples

The SAIMR, the University of the Witwatersrand and the Medical Research Council for financial support.