

https://dspace.org/

An investigation into the ancestry of the Malagasy population using variation in the alpha- and beta-globin gene cluster

Hewitt, Rachel

2014-04-07
http://hdl.handle.net10539/14511

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

AN INVESTIGATION INTO THE ANCESTRY OF THE MALAGASY POPULATION USING VARIATION IN THE α- and β-GLOBIN GENE CLUSTERS

Rachel Hewitt

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

1998

AN INVESTIGATION INTO THE ANCESTRY OF THE MALAGASY POPULATION USING VARIATION IN THE α- and β-GLOBIN GENE CLUSTERS

Rachel Hewitt

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

1998

ABSTRACT

The issue of Malagasy ancestry has been controversial, and has still not been completely resolved. The historical, linguistic, archaeological and some genetic evidence points to the fact that modern Malagasy are the descendants of immigrants who arrived on the island over the past 2000 years, from South and Southeast Asia, Africa and the Near East. In more recent centuries, mainly in the twentieth century, there have been significant numbers of Indian, Chinese and French immigrants. In addition, archaeological and historical studies of specific regional populations of Malagasy suggest a complex pattern of internal migration within the island, extending back in time to the first European contacts with the island in the sixteenth century. The 22 Malagasy ethnic groups may be classified as "highland" or "lowland" depending on their geographic distribution on the island. Within the ethnic groups, the founding populations have made different genetic contributions: the highland groups are said to have a greater Indonesian contribution to their ancestry, while the lowlanders have a greater African contribution to their ancestry.

Genetic studies on the Malagasy have been limited by small sample sizes, deficiencies in sampling procedures and in the limited number of polymorphisms studied. In light of the paucity of written records, the Department of Human Genetics, SAIMR, has undertaken a large study in Madagascar to reconstruct the biological history of its people, using genetic variation. This thesis forms a part of this study.

Variation in the α- and β-globin cluster has been extensively studied in many parts of the world, and has been shown to be population specific, with specific variants having distinct geographical distributions. Thus haemoglobin and its related disorders have been the subject of extensive studies for determining the origin(s) of particular populations. In this study, some of the α- and β-globin variation present in the Malagasy was characterised. Seven RFLPs/HVRs in the α-globin gene cluster and seven RFLPs in the β-globin gene cluster were analysed. The common α- and β-globin gene cluster haplotypes differ between African and Asian populations. Frequencies also vary between populations in a specific geographical regions. The aim of this study was to

characterise the haplotypes present in the Malagasy, to provide information on the relative genetic contributions of different populations to the peoples of Madagascar.

DNA samples from randomly selected, haematologically normal individuals were analysed. Individuals were chosen from six Malagasy ethnic groups: two "highland" populations (Merina and Betsileo), two "lowland" populations (Antasaka and Tsimiheti) and two populations from the south-west of the island (Mahafaly and Vezo). The groups chosen cover a broad range of Madagascar and thus provide some representation of the Malagasy population as a whole. The number of individuals studied in each ethnic group are as follows: Merina: 88; Betsileo: 78; Antasaka: 67; Tsimiheti: 67; Mahafaly: 26; Vezo: 25.

The frequencies of the α- and β-globin RFLP sites and α-globin HVRs in the Malagasy were calculated. 5' and 3' β-globin haplotypes were constructed on the basis of homozygosity. A maximum-likelihood algorithm was used to obtain frequencies of 5' β-globin haplotypes that could not be assigned on the basis of homozygosity. These data were then subject to statistical analysis. The frequencies of the 5' β-globin haplotypes (consisting of the five sites *Hin*cII 5' to ε, *Hin*dIII within $^{G}\gamma$ and $^{A}\gamma$, *Hin*cII within ψβ and 3' to it) were the most informative data set for comparing the Malagasy ethnic groups to each other and to other world populations. Unfortunately, the maximum-likelihood estimates of 5' β-globin haplotypes could not be used for comparative analyses due to the lack of similar data in other populations. However the strong correlation between the maximum-likelihood frequencies and the observed frequencies illustrated the ability of the algorithm to determine haplotype frequencies from otherwise uninformative individuals.

5' β-globin haplotypes were assigned unambiguously for 248 Malagasy chromosomes. Ten haplotypes were found; of these, nine have been reported previously in other world populations and one has not been reported and has thus been called "rare" in this study. The frequencies of unambiguous 5' β-globin haplotypes in the Malagasy and the proposed parental populations were initially analysed with χ^2 tests. For a more accurate comparison between these

populations, genetic distances were calculated and used for the construction of phylogenetic trees, principle component analysis was carried out, and a study of heterozygosity versus distance from the centroid was performed. Admixture estimates of two African populations and one Indonesian population to Malagasy ancestry were calculated.

Certain general trends were noted in all the analyses. The results are in agreement with the historical data which provides evidence for both African and Asian contributions to Malagasy ancestry. The highlanders were more closely affiliated to the Indonesian/Polynesian populations, while the south-west groups showed the strongest associations with the African populations. The lowlanders were consistently intermediate in position between the highlanders and the south-west groups, with the Antasaka being slightly more closely related to the African populations than the Tsimiheti. The Malagasy were shown to have high heterozygosities, similar to those of African populations, and this high degree of diversity is probably a reflection of the many sources of ancestry of the Malagasy. The south-west groups were the furthest outliers in the model of heterozygosity versus distance from the centroid, suggesting that these groups are the most genetically admixed of all the Malagasy groups that were studied. Estimates of ancestral population admixture confirmed these trends, with the highlanders having the highest proportional contribution by Indonesians (53%), but the lowest total African contribution (47%), while the south-west groups have the highest Bantu contribution (65%). The Indonesian and African contributions to the lowlanders are intermediate between those to highlanders and south-west groups. Overall the Malagasy subjects included in this study showed a 61% African admixture contribution and a 39% Indonesian admixture contribution.

It is hoped that the results obtained in this study will contribute to the larger project concerning the origins of the Malagasy, and that they may be used to shed further light on the much debated issue of Malagasy ancestry.

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 at any other university.

Rachel Hewitt

----- day of March, 1998

To my family and friends

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to the following people, for their involvement in my Masters project:

My supervisors, Professor T. Jenkins and Dr A. Krause.

The members of the field trips to Madagascar for collecting Malagasy blood samples, especially Prof T. Jenkins, Dr T. de Ravel, Dr P. Willem, Dr G. Campbell, Prof. J. Roux, Dr A. Rakouth and Dr S. Quanbeck.

The Department of Haematology of the SAIMR for all the haematological analyses for the project.

Dr H. Soodyall for all her advice concerning the statistical analysis of the data.

Bharti Morar for her help with the use of the EM programme.

Bruce Dangerfield for his endless help with the computer programmes that were used in this write-up.

Dr T. de Ravel and Dr H. Soodyall for kindly permitting use of the photographs that were recaptured in Chapter 1.

The FRD, SAIMR and University of the Witwatersrand for providing financial support for this project

All the members of the Human Genetics Department; especially those in the Molecular and Serogenetics laboratories for their support, help and technical advice: Michele Ramsay, Tony Lane, Andie Goldman, Prashiela Manga, Gwynn Stevens, Michelle Starfield, Bruce Dangerfield, Fahmida Essop, Robyn Kerr, Nayna Manga, Angela Turner, Latifa Toffie, Premila John, Carolyn Padoa and Anton Bergheim. Thanks also to Gloria Teckie and Tina Wessels and to Barbara and Pat for all their help and support throughout this study.

TABLE OF CONTENTS

ABSTRACT i

DECLARATION iv

DEDICATION v

ACKNOWLEGEMENTS vi

TABLE OF CONTENTS vii

LIST OF TABLES xiv

LIST OF FIGURES xvii

LIST OF ABBREVIATIONS AND SYMBOLS xx

CHAPTER 1

INTRODUCTION 1

1.1 THE POLYMORPHISM CONCEPT 3

1.1.1 PROTEIN POLYMORPHISM 3

1.1.2 DNA POLYMORPHISMS 4

1.1.2.1 RESTRICTION LENGTH FRAGMENT POLYMORPHISMS (RFLPs) 5

1.1.2.1.1 Detection of RFLPs 5

1.1.2.1.2 Applications of RFLPs 6

1.2 HAEMOGLOBIN 7

1.2.1 HAEMOGLOBIN STRUCTURE 7

1.2.2 MOLECULAR GENETICS OF HUMAN HAEMOGLOBIN 9

1.3 THE HUMAN β-LIKE GLOBIN GENES 11

1.3.1 β-LIKE GLOBIN GENE FINE STRUCTURE 11

1.3.2 β-GLOBIN CLUSTER STRUCTURAL VARIANTS 13

1.3.3 DNA SEQUENCE POLYMORPHISMS WITHIN THE β-GLOBIN GENE CLUSTER 13

1.3.3.1 POLYMORPHISM OF THE β-GLOBIN GENE: β-GENE FRAMEWORKS 17

1.3.3.2 β-GLOBIN CLUSTER HAPLOTYPES: RANDOM VERSUS NON-RANDOM ASSOCIATION OF POLYMORPHIC SITES 19

1.4 THE HUMAN α-LIKE GLOBIN GENES 22

1.4.1 α-LIKE GLOBIN GENE FINE STRUCTURE 22

1.4.2 VARIATION IN THE α-GLOBIN CLUSTER 23

1.4.2.1 VARIATION IN THE NUMBER OF α-GLOBIN GENES 23

1.4.2.2 GENE CONVERSION IN THE α-GLOBIN CLUSTER 23

1.4.2.3 RESTRICTION SITE POLYMORPHISMS 24

1.5 MADAGASCAR 27

1.5.1 THE PEOPLES OF MADAGASCAR 28

1.5.1.1 PHYSICAL APPEARANCE AND LINGUISTICS 32

1.5.1.2 CULTURE AND AGRICULTURE 33

1.5.2 THE ANCESTRY OF THE MALAGASY: "LA PLUS BELLE ENIGME DU MONDE" 38

1.5.2.1 THE BRITISH SCHOOL 38

1.5.2.2 THE FRENCH SCHOOL 39

1.5.2.3 THE MODERN SCHOOL 40

1.5.2.3.1 Initial settlement of Madagascar 40

1.5.2.3.2 Later immigrants and settlement 43

1.5.2.3.3 Genetic traits of the Malagasy 45

1.5.2.4 SUMMARY OF MALAGASY ANCESTRY 46

1.6 AIMS OF THE STUDY 48

CHAPTER 2

SUBJECTS AND METHODS 49

2.1 COLLECTION AND STORAGE OF BLOOD 49

2.2 HAEMATOLOGICAL METHODS 49

2.3 SUBJECTS INCLUDED IN THIS STUDY 50

2.4 EXTRACTION OF HIGH MOLECULAR WEIGHT DNA 50

2.5 QUANTITATION OF DNA 50

2.6 POLYMERASE CHAIN REACTION 51

2.6.1 β^S-MUTATION SCREEN 51

2.6.1.1 PCR PRODUCT DIGESTION AND AGAROSE GEL ELECTROPHORESIS 52

2.6.2 β-GLOBIN CLUSTER 52

2.6.2.1 PCR PRODUCT DIGESTION AND AGAROSE GEL ELECTROPHORESIS 57

2.6.3 α-GLOBIN CLUSTER 57

2.7. CALCULATION OF ALLELE FREQUENCIES AND HAPLOTYPE CONSTRUCTION 63

2.8. STATISTICAL ANALYSIS OF THE DATA 64

2.8.1 CHI-SQUARED (χ^2) ANALYSIS 64

2.8.2 GENETIC DISTANCE AND PHYLOGENETIC TREE ANALYSIS 65

2.8.3 PRINCIPLE COMPONENT (PC) ANALYSIS 66

2.8.4 HETEROZYGOSITY AND DISTANCE FROM THE CENTROID 67

2.8.5 ESTIMATION OF GENETIC ADMIXTURE 69

CHAPTER 3

RESULTS AND DISCUSSION 71

3.1 ALLELE FREQUENCIES OF β-GLOBIN RFLP SITES 71

3.1.2 CHI-SQUARE (χ^2) VALUES AND SIGNIFICANCE TESTING OF RFLP FREQUENCIES 74

3.2 UNAMBIGUOUS 5' β-GLOBIN HAPLOTYPES 89

3.2.1 CHI-SQUARED ANALYSIS OF 5' β-GLOBIN HAPLOTYPE FREQUENCIES 95

3.2.2 PHYLOGENETIC ANALYSIS 98

3.2.3 PRINCIPLE COMPONENT ANALYSIS OF 5' β-GLOBIN HAPLOTYPES 110

3.2.3.1 PC ANALYSIS OF THE SIX MALAGASY ETHNIC GROUPS 111

3.2.3.2 PC ANALYSIS OF THE SIX MALAGASY ETHNIC GROUPS AND OTHER WORLD POPULATIONS 113

3.2.3.3 PC ANALYSIS OF THE THREE MALAGASY REGIONAL GROUPS AND OTHER WORLD POPULATIONS 113

3.2.3.4 PC ANALYSIS OF MALAGASY AND OTHER WORLD POPULATIONS 116

3.2.4 AVERAGE HETEROZYGOSITY AND DISTANCE FROM THE CENTROID 118

3.2.4.1 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE SIX MALAGASY ETHNIC GROUPS 119

3.2.4.2 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE SIX MALAGASY ETHNIC GROUPS AND OTHER WORLD POPULATIONS 121

3.2.4.3 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE MALAGASY REGIONAL GROUPS AND OTHER WORLD POPULATIONS 123

3.2.4.4 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE MALAGASY AND OTHER WORLD POPULATIONS 125

3.2.5 ESTIMATION OF ADMIXTURE IN THE MALAGASY 127

3.3 MAXIMUM LIKELIHOOD ESTIMATES OF 5' β-GLOBIN HAPLOTYPES 130

3.4 3' β-GLOBIN HAPLOTYPES 135

3.5 FREQUENCIES OF α-GLOBIN RFLP SITES AND HVRS 138

3.5.1 CHI-SQUARE (χ^2) VALUES AND SIGNIFICANCE TESTING OF RFLP AND HVR FREQUENCIES 140

3.5.2 PHYLOGENETIC ANALYSIS 149

CHAPTER 4

GENERAL CONCLUSIONS 151

REFERENCES 161

APPENDIX 1

SOLUTIONS AND REAGENTS 169

LIST OF TABLES

Table 2.1 PCR primer sequences and concentrations used for the RFLP sites in the β-globin cluster. 53

Table 2.2 PCR conditions for the RFLP sites in the β-globin cluster. 54

Table 2.3 Sizes of the PCR products expected after amplification and restriction enzyme digestion, and the type of agarose gel used to visualise the fragments. 56

Table 2.4 PCR primer sequences and additional component specifications for the α-globin cluster RFLP sites and HVRs. 58

Table 2.5 PCR conditions for the RFLP sites and HVRs in the α-globin cluster. 60

Table 2.6 Sizes of the PCR products and restriction enzyme digestion fragments expected for the RFLP sites and HVRs in the α-globin gene cluster, and the percentage agarose gel used to visualise the fragments. 62

Table 3.1 Frequencies of the presence (+) of the β-globin RFLP sites and their standard errors (S.E.) in the six Malagasy ethnic groups studied. 72

Table 3.2 Frequencies of the presence (+) of the β-globin RFLP sites and their standard errors (S.E.) in the highland, lowland and south-west groups. 73

Table 3.3 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy groups and other world populations at the *Hind*II/ε site. 75

Table 3.4 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy groups and other world populations at the *Hind*III/$^{G}\gamma$ site. 76

Table 3.5 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy groups and other world populations at the *Hind*III/$^{A}\gamma$ site. 77

Table 3.6 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy groups and other world populations at the *Hind*II/5'ψβ site. 78

Table 3.7 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy groups and other world populations at the *Hind*II/3'ψβ site. 79

Table 3.8 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy groups and other world populations at the *Ava*II/β site. 80

Table 3.9 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy groups and other world populations at the *Hinf*I/β site. 81

Table 3.10 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy regional groups and other world populations at each β-globin RFLP site. 83

Table 3.11 5' β-globin haplotype frequencies and standard errors in the six Malagasy groups, Indonesians and Sub-Saharan Africans. 90

Table 3.12 5' β-globin haplotype frequencies and standard errors in the Malagasy regional groups, Indonesians and Sub-Saharan Africans. 93

Table 3.13 Chi-square values and p values for a pairwise comparison of Malagasy ethnic groups and other world populations for 5' β-globin haplotypes. 96

Table 3.14 Chi-square values and p values for a pairwise comparison of Malagasy regional groups, Bantu and Indonesian populations for 5' β-globin haplotypes. 97

Table 3.15 Admixture estimates (m) of the putative parental populations to the hybrid Malagasy regional groups and the total Malagasy population. 128

Table 3.16 5' β-globin haplotypes in the Malagasy: maximum likelihood estimates (EM), observed frequencies (OBS) and their correlations (r) and average heterozygosities (h) of EM estimates. 131

Table 3.17 Frequencies of the 3' β-globin haplotypes in the Malagasy. 136

Table 3.18 Chi-square values and p values (df=1) for a pairwise comparison of 3' β-globin haplotypes of the Malagasy ethnic groups and other world populations. 137

Table 3.19 Frequencies of the presence (+) of the α-globin RFLP sites and HVRs in the Malagasy ethnic groups and other world populations, and their standard errors (S.E.). 139

Table 3.20 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy ethnic groups and other world populations at the *Xba*I site. 142

Table 3.21 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy ethnic groups and other world populations at the *Sac*I site. 143

Table 3.22 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy ethnic groups and other world populations at the *Bgl*I site. 144

Table 3.23 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy ethnic groups and other world populations at the IZ/Z HVR. 145

Table 3.24 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy ethnic groups and other world populations at the PZ/Z HVR. 146

Table 3.25 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy ethnic groups and other world populations at the *Acc*I site. 147

Table 3.26 Chi-square values and p values (df=1) for a pairwise comparison of Malagasy ethnic groups and other world populations at the θ*Pst*I site. 148

LIST OF FIGURES

Figure 1.1 Molecular shape of haemoglobin (the haem group is shown as a disc) 8

Figure 1.2 The human β- and α-globin loci and the developmental expression of the globin genes. 10

Figure 1.3 The position of the exons (shaded areas) and introns (open areas) in the α- and β-globin genes. 12

Figure 1.4 The β-globin gene cluster showing the position of the 7 RFLP sites studied. 16

Figure 1.5 Normal β-globin gene frameworks. 18

Figure 1.6 The α-globin gene cluster showing the positions of the RFLP sites and HVRs studied. 26

Figure 1.7 Map of Madagascar showing the distribution of the Malagasy ethnic groups. 29

Figure 1.8 Typical market scenes in Madagascar, showing some of the cultural diversity present among the Malagasy. 31

Figure 1.9 Rectangular houses and rice paddies occur mainly in the highlands of Madagascar, and are typically Indonesian. 34

Figure 1.10 An Arab-style tomb on the plateau provides further evidence of Malagasy ancestry. 37

Figure 1.11 An outrigger canoe seen off the coast of Madagascar is of Polynesian/Indonesian origin. 42

Figure 1.12 The use of Zebu cattle for rice farming provides an illustration of the fusion of African and Indonesian cultures in Madagascar. 44

Figure 3.1 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5 5' β-globin RFLP sites in the Malagasy ethnic groups, Indonesians and Bantu. 86

Figure 3.2 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5 5' β-globin RFLP sites in the Malagasy regional groups, Indonesians and Bantu. 88

Figure 3.3 Histogram showing the 5' β-globin haplotype frequencies in the Malagasy ethnic groups, Indonesians and Sub-Saharan Africans. 91

Figure 3.4 Histogram showing the 5' β-globin haplotype frequencies in the Malagasy regional groups, Indonesians and Sub-Saharan Africans. 94

Figure 3.5 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5' β-globin haplotypes in the 6 Malagasy ethnic groups. 99

Figure 3.6 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations. 100

Figure 3.7 Standard neighbour joining tree based on the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (reconstructed from PHYLIP, using Nei's genetic distance). 102

Figure 3.8 Standard neighbour joining tree based on the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (reconstructed from PHYLIP, using Cavalli-Sforza's Chord measure). 103

Figure 3.9 Standard neighbour joining tree based on the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (reconstructed from PHYLIP, using Reynold's distance) 105

Figure 3.10 Standard neighbour joining tree based on the frequencies of the 5' β-globin haplotypes in the Malagasy regional groups and other world populations (drawn with DISPAN using Nei's genetic distance). 107

Figure 3.11 Standard neighbour joining tree based on the frequencies of the 5' β-globin haplotypes in the Malagasy and other world populations (drawn with DISPAN using Nei's genetic distance). 109

Figure 3.12 PC map of 5' β-globin haplotype frequencies in the 6 Malagasy ethnic groups. 112

Figure 3.13 PC map of 5' β-globin haplotype frequencies in the Malagasy ethnic groups and other world populations. 144

Figure 3.14 PC map of 5' β-globin haplotype frequencies in the Malagasy regional groups and other world populations. 115

Figure 3.15 PC map of 5' β-globin haplotype frequencies in the Malagasy and other world populations. 117

Figure 3.16 Heterozygosity vs distance from the centroid for the 6 Malagasy ethnic groups. 120

Figure 3.17 Heterozygosity vs distance from the centroid for the Malagasy ethnic groups and other populations. 122

Figure 3.18 Heterozygosity vs distance from the centroid for the regional Malagasy groups and other world populations. 124

Figure 3.19 Heterozygosity vs distance from the centroid for the Malagasy and other world populations. 126

Figure 3.20 Standard neighbour joining tree based on the maximum-likelihood frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (drawn with DISPAN using Nei's genetic distance). 134

Figure 3.21 Standard neighbour joining tree based on the frequencies of the 5 α-globin RFLP sites and 2 HVRs (drawn with DISPAN using Nei's genetic distance). 150

LIST OF ABBREVIATIONS AND SYMBOLS

χ^2	Chi-Square test
ACD	Acid Citrate Dextrose
AS	Antasaka/Antesaka (Malagasy ethnic group)
βLCR	β-globin Locus Control Region
bp	Base pairs
BSA	Bovine serum albumin
BT	Betsileo (Malagasy ethnic group)
°C	Degrees centigrade
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
EDTA	Ethylenediamine Tetra-Acetic Acid
EM	Expectation-maximisation
EtBr	Ethidium Bromide
FBC	Full Blood Count
g	Gram
mg	Microgram
G6PD	Glucose-6-Phosphate Dehydrogenase
HbA	Adult haemoglobin
HbF	Foetal haemoglobin
HbS	Sickle haemoglobin
Hct	Haemocrit
HGT	High Gelling Temperature (agarose)
HVR	Hypervariable Region
IVS	Intervening Sequence
kb	Kilobase
mA	Milliamperes
MCH	Mean Cell Haemoglobin
MCV	Mean Cell Volume
MF	Mahafaly (Malagasy ethnic group)
min	Minute
mg	Milligram
ml	Millilitre
ml	Microlitre
mM	Millimolar
MR	Merina (Malagasy ethnic group)
OD	Optical Density
PC	Principle Component
PCR	Polymerase Chain Reaction
pmol	Picomoles
PNG	Papua New Guinean
RCC	Red Cell Count
rpm	Revolutions per minute
RFLP	Restriction Fragment Length Polymorphism
SA	South African
SAIMR	South African Institute for Medical Research
SE	South East
SDS	Sodium Dodecyl Sulphate

TBE	Tris-Boric Acid-EDTA
TE	Tris-EDTA
Tris	Tris-(hydroxymethyl)amino methylene
TS	Tsimiheti/Tsimihety (Malagasy ethnic group)
UV	Ultraviolet
V	Volt
VNTR	Variable Number of Tandem Repeats
VZ	Vezo (Malagasy ethnic group)

CHAPTER 1

INTRODUCTION

The human haemoglobin gene family has for many years played an important role in our understanding of human genetics. The principles of many hereditary human genetic diseases were first defined on the molecular level in the haemoglobin disorders. The most notable was Pauling's discovery of an abnormal haemoglobin as the cause of sickle cell anaemia which gave rise to the concept of " molecular disease" (Pauling *et al.* 1949). Ingram's subsequent localisation of the abnormality to a single amino acid substitution introduced the principle of "one gene, one polypeptide chain" (Ingram 1958). Additional discoveries of abnormal haemoglobins defined the relationship between the structure and function of haemoglobin. In so doing, they provided models for analysing the control of gene expression and gene regulation, and thus formulated many basic concepts of molecular genetics (Weatherall *et al.* 1984). The globin genes were among the first genes to be cloned; furthermore, since the precise localisation of the globin genes has been determined, they have been extremely useful in linkage studies (Magenis *et al.* 1985, Reeders *et al.* 1985, Breuning *et al.* 1987, Buckle *et al.* 1988). In addition, haemoglobin variants are valuable tools in the study of evolution, and may be used to trace the origin(s) and/or past migrations of populations.

Madagascar's culture is a unique fusion of elements drawn from the western, northern and eastern shores of the Indian Ocean. While the ancestry of the Malagasy has long been discussed, the evidence is incomplete and the patterns complex. Historical, linguistic, genetic and some archaeological evidence point to Africa and Indonesia as the major sources of ancestry of the Malagasy, with additional contributions from Arabia and India (Dewar and Wright 1993).

This study forms part of a large inter-disciplinary research project that is intended to test the various hypotheses concerning the origins of the Malagasy by combining social science techniques with those of human genetics. The project aims to identify the origins and historical movements of the proto-Malagasy groups, and to integrate into

these findings the latest historical and archaeological evidence on the origins of the Malagasy and the first settlement of Madagascar.

In this introduction, the concept of genetic polymorphism, and the structure and molecular genetics of human haemoglobin will be reviewed. Theories that have been proposed concerning the colonisation of Madagascar and the ancestry of the Malagasy will be presented. Finally, the aims of the study will be outlined.

1.1 THE POLYMORPHISM CONCEPT

A large number of different enzymes and proteins are synthesised in the human body and each of their distinct polypeptide chains is coded for by a DNA sequence in the genome. Studies have shown that there are a considerable number of structurally variant forms, which are genetically determined. These inherited variants or alleles are attributed to specific gene mutations which occurred in single individuals in earlier generations. The majority of the alleles are rare, but occasionally occur with an appreciable frequency, and may become common enough to be described as a genetic polymorphism. More formally, polymorphism has been defined as the "occurrence of two or more alleles at a particular locus, where at least two alleles appear with frequencies of more than one percent, and the rarer cannot be maintained by recurrent mutation alone" (Bodmer and Cavalli-Sforza 1976).

The incidence and distribution of polymorphisms or allelic variants which are observed in natural populations are considered to be the overall consequences of three processes:

1. Mutation, which generates new alleles.
2. Natural selection, which tends to cause the spread of alleles which increase "fitness", where fitness is measured by the relative contributions which individuals of different genotypes make to the next generation.
3. Random genetic drift, which is due to the sampling process determining which gametes produced by members of one generation will give rise to individuals of the next generation (Harris 1980).

1.1.1 PROTEIN POLYMORPHISM

The phenomenon of polymorphism has been recognised in genetics since the early 1900's. One of the first well-studied examples of protein polymorphism in man was the ABO blood group polymorphism, detected at the beginning of this century by Karl Landsteiner. Other polymorphisms that were also recognised on the basis of clear-cut physiologic differences between people include red-green colour blindness and the

phenylthiourea taste sensitivity polymorphism, discovered by Fox in 1932 (Bodmer and Cavalli-Sforza 1976).

Genetic variations in humans have classically been detected by the demonstration of differences in the structures or properties of gene products, by serological and biochemical methods. Many protein and enzyme polymorphisms were detected in the 1960's and 1970's. These included the haemoglobin, glucose-6-phosphate dehydrogenase, haptoglobin and serum cholinesterase polymorphisms (Bodmer and Cavalli-Sforza 1976). The existence of polymorphisms at the protein level is a reflection of polymorphic differences in DNA.

1.1.2 DNA POLYMORPHISMS

In the late 1970's, rapid advances in techniques for nucleic acid analysis allowed the detection of polymorphisms at the DNA level. In 1975, Southern introduced a method (Southern Blotting) (Southern 1975), which has been used to study the organisation of many eukaryotic genomes, and in particular, has been used to analyse structural variants of DNA (Mears *et al.* 1978, Orkin *et al.* 1978).

Some variations may occur in the coding regions of DNA that result in an altered protein product, and thus may be detected as protein polymorphisms. Other variations are "silent": they are common in certain regions of DNA and are not apparent on protein electrophoresis. These may be classified according to the sites in the DNA at which they cause alterations, as follows (Kazazian *et al.* 1983):

1. Synonymous mutation or single base change in the DNA, which alters a particular codon, but does not cause a change in the amino acid at that position.
2. Mutation in the 5' or 3' untranslated regions of a gene.
3. Mutation in an intervening sequence.
4. Mutation in non-coding DNA.

1.1.2.1 RESTRICTION FRAGMENT LENGTH POLYMORPHISMS (RFLPs)

DNA restriction fragment length polymorphisms (RFLPs) are manifested as differences that occur in a fragment length of DNA when it is cleaved by a specific restriction endonuclease. An RFLP may be caused by one of the following:

1. A single base change or nucleic acid substitution that leads to the loss or formation of a recognition site.
2. A difference in the number of tandem repeats of short (dinucleotide, trinucleotide or tetranucleotide) DNA sequences, also referred to as microsatellites, or a variable number of tandem repeat (VNTR) DNA sequences between different chromosomes.
3. Any other DNA structural alteration, such as a deletion, inversion, translocation or duplication that occurs between the recognition sites of a restriction enzyme.

1.1.2.1.1 Detection of RFLPs

RFLPs are usually detected by one of two methods, firstly by Southern blotting, using hybridisation probes complementary to the variant DNA fragments, and secondly, by the polymerase chain reaction (PCR) technique for amplification of specific DNA sequences. PCR is a fast, simple and accurate method of detecting RFLPs, where the appropriate DNA sequence information is available for the synthesis of oligonucleotide primers. A region of DNA can be amplified and scored for the presence or absence of restriction sites after digestion with the appropriate enzymes. The first PCR amplification of genomic DNA was a 110bp-globin fragment for the diagnosis of sickle cell anaemia (Saiki *et al.* 1985). The technique is capable of producing a selective enrichment of a specific DNA sequence by a factor of 10^6, greatly facilitating a variety of subsequent analytical manipulations (Saiki *et al.* 1988).

It has been claimed that more RFLPs occur at restriction enzyme recognition sites that contain the dinucleotide CpG. The theoretical basis for this is the fact that the human genome is highly methylated at this dinucleotide. 5-methyl-cytosine is subject to frequent replacement by thymine due to deamination of the methylated base. Such deamination would result in the addition or removal of a restriction enzyme recognition site (Cooper and Schmidtke 1984).

1.1.2.1.2 Applications of RFLPs

RFLPs are useful in linkage analysis, prenatal diagnosis and in particular, anthropological and evolutionary studies. Much of the history of recent human evolution is preserved in the genome. The application of interpopulation genetic variation to the fields of population and evolutionary genetics has greatly expanded anthropological research. The best way to study population affinities is by comparing DNA sequences of individuals from different populations, since DNA analysis is very informative, direct and unambiguous (Wainscoat *et al.* 1986). The large number of RFLPs available, together with the fact that most RFLPs are not associated with any functional differences and are not deleterious, allows for a powerful means to examine allelic variation in a large number of individuals. The analysis of allele frequencies in different populations permits the construction of genetic distance maps between the populations and allows genetic relationships between populations to be determined (Cavalli-Sforza *et al.* 1994).

Studies of gene flow between populations are greatly enhanced by mitochondrial RFLPs, in conjunction with RFLPs on the Y- chromosome. Because of the maternal origin of mitochondrial DNA, mitochondrial DNA polymorphisms will give an indication of the female contribution to the gene pool of a particular population. Likewise the Y-chromosome RFLPs can be used to determine the source of the male contributions to a population's gene pool (Wainscoat *et al.* 1986).

1.2 HAEMOGLOBIN

Haemoglobin is one of the best known proteins and there is a vast amount of literature concerning it. A brief review of the information that is relevant to this study has been included.

1.2.1 HAEMOGLOBIN STRUCTURE

Haemoglobin (Hb) is a tetramer that consists of two α-like and two β-like globin subunits. The subunits are encoded by two clusters of genes, each of which is expressed sequentially during development. Each subunit is associated with a haem molecule and binds oxygen, and the monomers are linked together by electrostatic and hydrophobic interactions (Figure 1.1). Haemoglobin is an allosteric protein, which changes conformation as it binds oxygen, such that oxygen binding at one site favours binding at the three other sites (reviewed in Antonarakis *et al.* 1985).

The alpha (α) and zeta (ζ) chains comprise the alpha-like globins, and contain 141 amino acids each. The epsilon (ε), gamma (γ), delta (δ) and beta (β) chains comprise the beta-like globins, and contain 146 amino acids each. All globin genes are expressed only in erythroid cells. In earliest embryonic life, the two ζ chains and the two ε chains combine to form Hb Gower I ($\zeta_2\varepsilon_2$); two α chains and two ε chains form Hb Gower II ($\alpha_2\varepsilon_2$); and two ζ chains and two γ chains form Hb Portland ($\zeta_2\gamma_2$). In fetal life, these embryonic forms of haemoglobin are replaced by fetal haemoglobin (HbF) (α_2, γ_2). The γ chains, of two types, are nearly identical and are designated $^{G}\gamma$ and $^{A}\gamma$. In postnatal life, HbF is replaced by adult haemoglobin (HbA) ($\alpha_2\beta_2$), and HbA_2 ($\alpha_2\delta_2$), a minor adult component. HbA represents about 98% of all haemoglobin found in an individual's erythrocytes after the age of six months. The remaining 2% consists of HbA_2, and a small amount of HbF is also found in adult peripheral blood (Antonarakis *et al.* 1985).

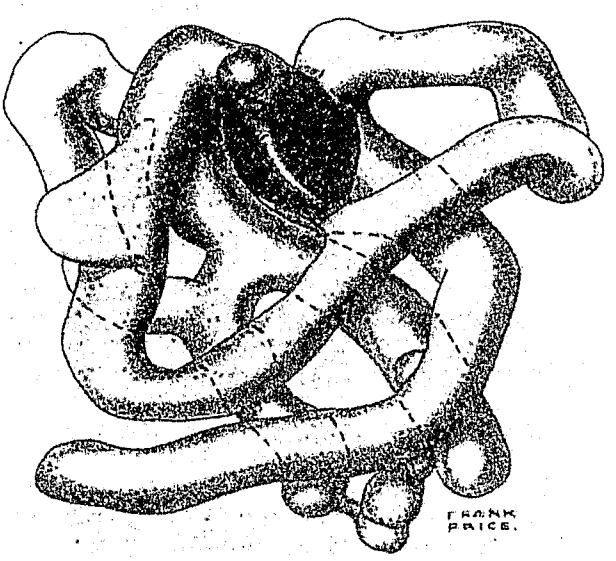


Figure 1.1 Molecular shape of haemoglobin (the haem group is shown as a disc) (Lehmann and Huntsman 1974).

During development, there is a co-ordinated switching of haemoglobin synthesis, affecting both the site of erythropoiesis, as well as the types of polypeptide chains and hence haemoglobins synthesised. This is illustrated in Figure 1.2. These switches are under tight genetic control, the mechanism of which is relatively complicated and is presently under investigation (Grosveld *et al.* 1993).

1.2.2 MOLECULAR GENETICS OF HUMAN HAEMOGLOBIN

Somatic cell hybridisation studies localised the α-globin gene complex to a 25 kilobase (kb) region on chromosome 16, and the β-globin gene complex to a 50 kb region on chromosome 11 (Deisseroth *et al.* 1977, 1978, Maniatis *et al.* 1980). The genes are arranged in the order in which they are expressed during development. Conserved sequences essential for eukaryotic gene expression, including promoters and enhancers for transcription and sequences necessary for mRNA processing are found in and around the globin genes (reviewed in Higgs *et al.* 1989).

It appears that the globin gene family has evolved from a single ancestral gene. The ancestral α-like and β-like genes were probably separated by a translocation or transpositional event about 500 million years ago, so that the gene families are now present on different chromosomes. Each cluster has evolved by gene duplication, modified by sequence divergence, deletion due to unequal crossing over between repeat sequences, gene conversion and retroposition (Efstratiadis *et al.* 1980).

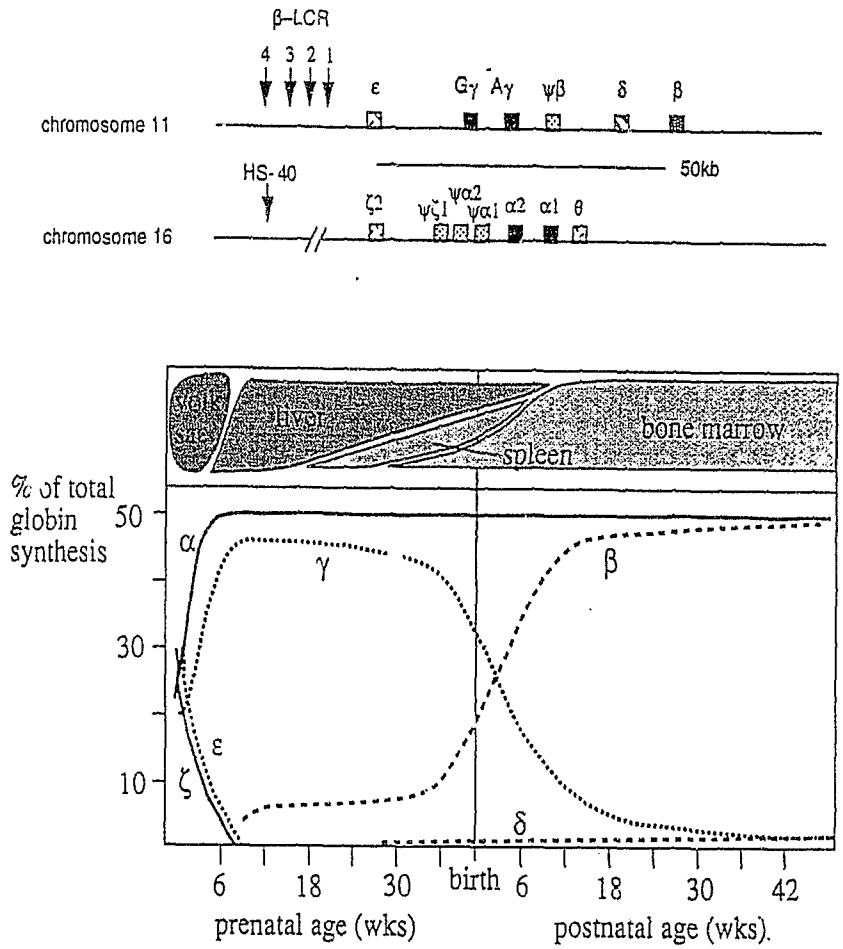


Figure 1.2 The human β- and α-globin loci and the developmental expression of the globin genes (reviewed in Grosveld *et al.* 1993).

1.3 THE HUMAN β-LIKE GLOBIN GENES

1.3.1 β-LIKE GLOBIN GENE FINE STRUCTURE

The complete nucleotide sequences of the five human β-like globin genes and flanking regions were determined in 1980 (reviewed in Maniatis *et al.* 1980). Later, the gene was located to 11p15 by *in situ* hybridisation and utilisation of rearrangements on chromosome 11 (Magenis *et al.* 1985).

The gene cluster is illustrated in Figure 1.2. The ε gene is located 13.3kb 5' to the $^{G}\gamma$ gene. The two γ genes are situated 3.5kb apart, and are thought to have arisen by gene duplication. There is only one nucleotide sequence difference in the coding region between these two genes, in codon 136, where GGA (coding for glycine) is present in $^{G}\gamma$, while GCA (coding for alanine) is present in $^{A}\gamma$. The genes also diverge in the 3' untranslated regions. The $^{A}\gamma$ and δ genes are 15kb apart, and the δ and β 6.5kb apart. The δ- and β-chains differ by 8 out of 146 amino acids (reviewed in Bank *et al.* 1980).

All five β-like globin genes are interrupted by two introns at identical positions: the first is 125-150 base pairs (bp) long and is located between codons 30 and 31; the second, 700-900 bp long, is located between codons 104 and 105 (Lawn *et al.* 1978) (Figure 1.3). A pseudogene, ψβ1, is also included in the gene cluster. This is a gene that has 75% - 80% sequence homology with the β-gene, but contains mutations which prevent its expression. Hence no globin polypeptide has been identified, and the function of ψβ1 is unclear (Maniatis *et al.* 1980).

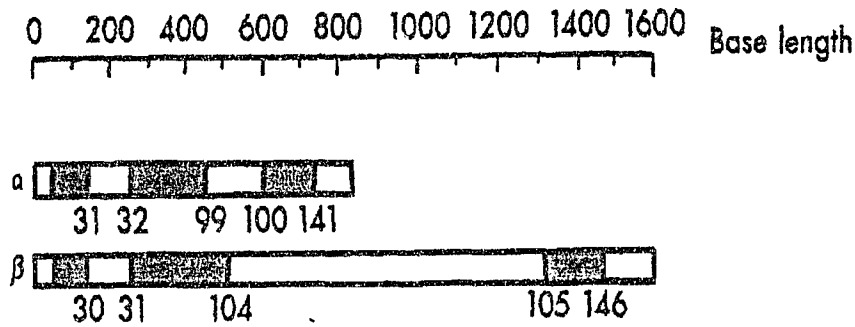


Figure 1.3 The position of the exons (shaded areas) and introns (open areas) in the α- and β-globin genes (Proudfoot 1980).

1.3.2 β-GLOBIN CLUSTER STRUCTURAL VARIANTS

The most common structural variants of the β-globin polypeptide are simple amino acid substitutions resulting from single changes in the nucleotide sequence. The majority of these have no demonstrable effect on globin gene activity or stability. However some amino acids substitutions result in severe clinical effects. The common haemoglobinopathies are β-thalassemia and the three β-globin abnormalities β^S ($\beta^{6\ Glu \rightarrow Val}$), β^E ($\beta^{26\ Glu \rightarrow Lys}$) and β^C ($\beta^{6\ Glu \rightarrow Lys}$). β-Thalassemia may be divided into two types, β^O- and β^+-thalassemia, depending on whether the underlying mutation abolishes or only reduces the production of functional protein. Both β^O- and β^+-thalassemia are due commonly to single nucleotide substitutions or small deletions or insertions; about 120 mutations are known to cause β-thalassemia. Homozygotes for the β^S gene have sickle cell anaemia, while homozygotes for β^E and β^C have mild haemolytic disorders. Significant illness accompanies double heterozygotes involving HbS and HbC (SC disease), and HbE and β-thalassaemia (Nagel and Ranney 1990). The fusion proteins, eg. Hb Lepore, are another class of haemoglobin variant, that result from unequal cross-over and fusion between β-like genes. Variations of other genes in the β-globin gene cluster commonly involve the γ-globin genes, and include $^G\gamma \rightarrow ^A\gamma$ and $^A\gamma \rightarrow ^G\gamma$ replacements; deletions of one of the γ-globin genes; γ-globin gene triplications (γγγ); and γ-globin gene quadruplications (γγγγ) (Huisman 1987).

1.3.3 DNA SEQUENCE POLYMORPHISMS WITHIN THE β-GLOBIN GENE CLUSTER

Several genetic polymorphisms of restriction enzyme cleavage sites have been detected in the β-globin gene cluster. Jeffreys (1979) estimated that one in every 100 nucleotides in the human genome was polymorphic. Later, Kazazian *et al.* (1983) proposed that any two β-globin gene clusters chosen at random would differ by one nucleotide in every 500.

The first polymorphism was detected by Lawn *et al.* (1978). However this is a rare polymorphism and not very useful. The first useful polymorphism in the cluster was an alteration in an *Hpa*I recognition site located 5kb 3' to the β-gene (Kan and Dozy 1978). The two normal fragments obtained on digestion with *Hpa*I occurred at different frequencies in Black, Caucasian and Asian populations, but a third variant was detected in Black individuals only, and the *Hpa*I RFLP was thus described as a "private" polymorphism, found in Negroid peoples only. The frequent association of the absence of this site with the β^S-gene was exploited as a tool for DNA-based prenatal diagnosis (Kan and Dozy 1978, Feldenzer *et al.* 1979).

Subsequently, a number of other polymorphic sites in the β-globin gene cluster were identified, mostly occurring in flanking DNA or introns. There are over 20 polymorphic sites within the β-globin gene cluster, but most analyses are limited to the following: *Hin*cII (equivalent to *Hind*II) 5' to ε, *Xmn*I 5' to $^G\gamma$, *Hin*dIII within $^G\gamma$ and $^A\gamma$, *Hin*cII (equivalent to *Hind*II) within ψβ and 3' to it, *Ava*II within β, and *Hpa*I, *Hin*dIII and *Bam*HI (equivalent to *Hin*fI) 3' to β. The sites analysed in this study are shown in Figure 1.4. All the polymorphisms described until 1985 are single nucleotide substitutions. Most occur in all racial groups, however there are differences in the frequencies of the allele morphs in different populations. A few of the sites are "private" and are only polymorphic in Blacks, such as the *Taq*I site 5' to the ε-globin gene, and the *Hin*fI and *Hpa*I sites 3' to the β-globin gene (Chakravarti *et al.* 1984).

In addition to single nucleotide substitutions, there is also polymorphism for several structural features in the gene cluster. The number of simple sequence repeats within the $(AT)_xN_{12}(AT)_y$ motif varies among individuals. These sequences are located 6kb - 18kb upstream of the ε-globin gene in the β-globin locus control region (βLCR), a region that has been shown to be critical to the *in vivo* regulation of the developmental switching and expression of all the globin genes (Perichon *et al.* 1993). There is also polymorphism for the number of repeats (4, 5 or 6) of the sequence (ATTTT) in a region approximately 1.4kb 5' to the β-globin gene, and for a series of 16-17 (AC) repeats 2.7kb 5' to the β-globin gene (Poncz *et al.* 1983). It is suggested that these insertion-deletion polymorphisms arose by mispairing of chromosome homologues and

unequal crossing over (Orkin and Kazazian 1984). Another tandemly repeated sequence, the $(AT)_xT_y$ motif, is located closer to the β-globin gene. A number of structures of this repeat have been described in different population groups (Chebloune *et al.* 1988).

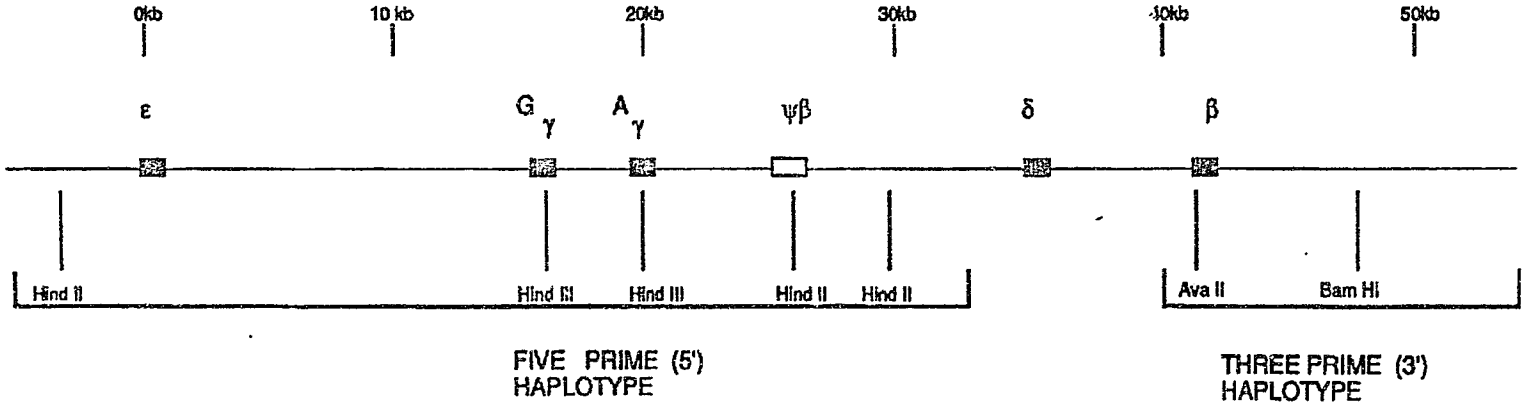


Figure 1.4 The β-globin gene cluster showing the position of the 7 RFLP sites studied (Flint *et al.* 1993a).

1.3.3.1 POLYMORPHISM OF THE β-GLOBIN GENE: β-GENE FRAMEWORKS

The β-globin gene exhibits polymorphism limited almost exclusively to common sequence types that have been designated β-gene frameworks (Orkin *et al.* 1982, Kazazian *et al.* 1983, Antonarakis *et al.* 1984). These frameworks are normal β-gene sequences, differing from one another by single nucleotides. Three frameworks (1, 2 and 3) were initially identified in β-genes of Mediterraneans. The polymorphisms lie in codon 2 and in IVS2, positions 16, 74, 81 and 666. Frameworks 1 and 2 differ at a single nucleotide (at IVS2 position 74), whereas framework 3 has the structure of framework 2 plus the 4 additional substitutions. Further studies on Asian β-genes revealed a further framework that differed from the Mediterranean framework 3 at one site, and this was thus designated "3 Asian" (Figure 1.5). In each racial group, the frequency of the three basic frameworks (1, 2 and 3 Asian) is found to vary. These frameworks can be de'_cted in uncloned DNAs by analysis of both intragenic and extragenic restriction sites, since the frameworks are in linkage disequilibrium with 4 polymorphic sites (AvaII/β, *Hpa*I/β, *Hind*III/β and *Bam*HI/β sites) (Antonarakis *et al.* 1985).

Another framework variant was reported by Carestia *et al.* (1991) that is present in 100% association with the African 5' haplotype [----+]. It appears to be ancient since it is associated with different HpaI/3'β polymorphic fragments and with normal and mutant genes. These fixed sequence variants of the β-gene are termed frameworks because they are the backgrounds upon which mutations leading to haemoglobinopathies have occurred. The existence of these frameworks has been explained by neutral mutation followed by genetic drift (Kazazian *et al.* 1985).

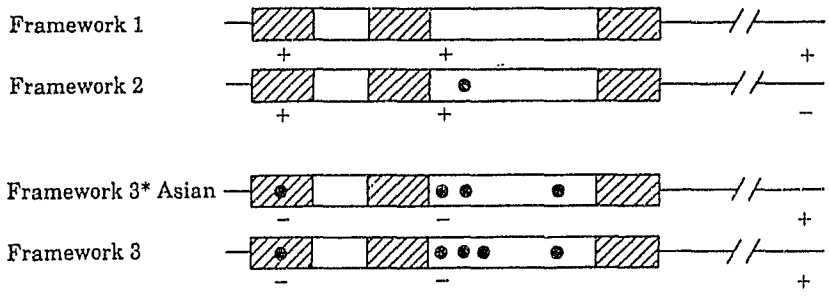


Figure1.5 Normal β-globin gene frameworks. Filled areas are exons in the β gene and open areas are introns. The + and - correspond to the *Ava*II, *Hpa*I and *BamH*I restriction polymorphic sites in the 3' region of the β-globin gene cluster . Black dots in frameworks 2 and 3 indicate the locations of single nucleotide differences from the sequence of framework 1 (Kazazian *et al.* 1983).

1.3.3.2 β-GLOBIN CLUSTER HAPLOTYPES: RANDOM VERSUS NON-RANDOM ASSOCIATION OF POLYMORPHIC SITES

A β-globin cluster haplotype is defined as a group of linked loci that are characterised by the presence (+) or absence (-) or specific polymorphic restriction endonuclease recognition sites on the cluster. Given the possible presence or absence of n sites, one would expect 2^n haplotypes if polymorphisms were randomly associated on a chromosome. However, only a small number of haplotypes have been found for regions of the β-globin cluster, providing evidence that polymorphic sites are not randomly associated (Antonarakis *et al.* 1985, Flint *et al.* 1993a, 1993b).

There are two regions of DNA characterised by non-random association of sequences, known as the 5' and 3' regions of the β-globin cluster:

(i) The 5' region of the β-globin cluster

This comprises a 34.6kb region extending from the ε-globin gene to 5' of the δ-globin gene. Eight polymorphic sites are included in this partial haplotype. Three haplotypes (out of $2^8 = 256$ possible) account for about 94% of β^A chromosomes in Greek, Italian and Asian-Indian populations. The three haplotypes differ from one another by more than one site, and none can be derived from any other by a single point mutation or cross-over event (Antonarakis *et al.* 1985). In American Blacks, a fourth haplotype was found at a much higher frequency than expected, and was rare in non-Black populations (Antonarakis *et al.* 1982). These data suggest a non-random association of polymorphic sites in the 5' region. This non-randomness is termed linkage disequilibrium. Eighteen 5' β-globin haplotypes have now been described in various world populations (Flint *et al.* 1993a).

(ii) The 3' region of the β-globin cluster

Non-random association of polymorphic restriction sites is also observed from the 5' end of the β-globin gene, extending 19.4kb in the 3' direction. Six polymorphic sites are included in this partial haplotype. About 90% of the chromosomes in various populations contain only one of four partial haplotypes in this region. The β-gene itself is contained within this portion of the cluster, and its frameworks are specifically associated with these common 3' patterns (Orkin *et al.* 1982).

Between the 5' and 3' sequence clusters lies a 9.1kb region that displays random association with either segment (Antonarakis *et al.* 1982). That is, the observed frequencies of the combinations of the 5' and 3' clusters of polymorphic sites are nearly identical to those expected if they were randomly associated. Thus, whereas the alleles in the 5' and 3' subclusters show strong linkage disequilibrium, no linkage disequilibrium exists between the two subclusters. Since recombination is the primary evolutionary mechanism for reducing linkage disequilibrium, the implication is that the recombination rate is much higher between the two subclusters than within each. Thus, on an evolutionary timescale, this region has been designated a recombination "hot spot" (Chakravarki *et al.* 1984). It has been suggested that the polymorphic structural features located 5' to the β-globin gene, such as the (ATTTT) and $(AC)_n$ repeats, increase the recombination frequency between the 5' and 3' haplotypes (Semenza *et al.* 1984).

Since the polymorphisms discussed above are largely present in all racial groups, they may be considered ancient markers of β-globin gene clusters. Recently a 3kb region encompassing the β-globin gene has been sequenced in individuals from nine populations from Africa, Asia and Europe to detect allelic polymorphisms that can be used to trace African and Asian lineages in the genetic ancestry of modern humans (Harding *et al.* 1997a, 1997b). In their study, a β-globin gene tree was constructed and the most recent common ancestor was a sequence found only in Africa and estimated to have arisen approximately 800 000 years ago. The patterns of β-globin diversity between Africa and Asia were explained by greater African effective population size rather than by greater time depth. The patterns are also not reconciled with the "Out of

Africa" hypothesis of modern human evolution, which assumes a unidirectional migration out of Africa 100 000 years ago and total replacement of archaic populations in Asia (Harding *et al.* 1997a, 1997b).

1.4 THE HUMAN α-LIKE GLOBIN GENES

1.4.1 α-LIKE GLOBIN GENE FINE STRUCTURE

The α-globin locus occurs at 16p13.1 (Nicholls *et al.* 1987).The cluster includes two α-globin structural genes (α_1 and α_2), an embryonic α-like gene (ζ_2), three pseudogenes ($\psi\alpha_1$, $\psi\alpha_2$ and $\psi\zeta_1$) and a related gene of unknown function θ_1 (Proudfoot *et al.* 1980, Marks *et al.* 1986). These are arranged in the order 5'-ζ_2-$\psi\zeta_1$-,$\psi\alpha_2$-$\psi\alpha_1$-α_2-$\alpha_1\theta_1$-3' (Figure 1.2). The distance between ζ and ψζ varies between 10 and 12 kb and that between ψζ and $\psi\alpha_2$-$\psi\alpha_1$ is 3.5kb, while that between $\psi\alpha_1$ and α_2 and between α_2 and α_1 are 4.2kb and 3.7kb respectively (Lauer *et al.* 1980, Michelson and Orkin 1983). As with the β-globin gene locus, the α-globin genes are divided into three exons by two noncoding intervening sequences (IVS1 and IVS2) of approximately 95bp and 125bp, located between codons 31 and 32 and 99 and 100 respectively (Maniatis *et al.* 1980) (Figure 1.3). The positions of the introns are conserved, probably reflecting their common origin from a single ancestral gene. The α-globin cluster and surrounding DNA also contain many Alu-family repeats, but in contrast to the β-globin cluster, they are GC rich (60%), early replicating and are associated with non-methylated CpG-rich islands. The α_2 and α_1 genes are highly homologous, encode identical proteins and only differ within IVS2 and their 3' noncoding regions (17% divergence) (Michelson and Orkin 1983). Despite this the α_2 gene encodes two- to three-fold more protein than α_1 (Higgs *et al.* 1989).

The α-globin gene complex includes a number of hypervariable regions (HVRs) consisting of tandem repeats of different unit sequences which nevertheless share a core sequence. They are located at the 3' end of the complex (α-globin 3'HVR), between the ζ and ψζ genes (interzeta HVR) and within the introns (IVS1 and IVS2) of the ζ-like genes (ζ-intron HVR) (Goodbourn *et al.* 1983). A further HVR (5'HVR), located 70kb upstream of the α-globin gene cluster is also highly polymorphic and together with the other HVRs at this locus, provides a valuable genetic marker on 16p (Jarman and Higgs 1988). The gene for adult polycystic kidney disease was mapped

using these HVRs (Reeders *et al.* 1985). Each hypervariable region appears to be inherited in a Mendelian fashion, although new alleles may be produced at a relatively high rate, and little is known about the evolution or function of these regions (Jeffreys *et al.* 1985). However, they are of great value as genetic markers, particularly in forensic and anthropological studies, since the HVRs have very significant in-built population specific information in their allelic distributions (Martinson *et al.* 1994).

1.4.2 VARIATION IN THE α-GLOBIN CLUSTER

1.4.2.1 VARIATION IN THE NUMBER OF α-GLOBIN GENES

Phenotypically normal individuals may have 3,4,5 or 6 α-globin genes and 3,4,5 or 6 ζ-like genes, as a result of unequal genetic exchange,. The triplicated ζ-gene (ζζζ) is particularly common in Southeast Asia, Melanesia, Micronesia and Polynesia but elsewhere is uncommon (Winichagoon *et al.* 1982). These variants are thus useful population markers. The α-thalassaemias are classified into 4 types according to the number of α-globin genes involved (1,2,3 or 4). Although the majority of α-thalassaemias are due to gene deletion, some non-deletional types of defects have been reported (Kan *et al.* 1977, Maniatis *et al.* 1980).

1.4.2.2 GENE CONVERSION IN THE α-GLOBIN CLUSTER

Analysis of the downstream ζ-like gene in several populations has shown that it exists in two distinct forms. In one its structure is clearly that of a pseudogene ($\psi\zeta_1$) and in the other (ζ_1) the $\psi\zeta_1$ gene has undergone a gene conversion by the ζ_2 gene such that it becomes more similar to the functional ζ gene although it still appears not to be expressed in vivo. The frequency of the ζ_2-ζ_1 chromosome varies from one population to the other (0.14 to 0.57), and phenotypically normal individuals homozygous for either ζ_2-$\psi\zeta_1$ or ζ_2-ζ_1 chromosomes have been observed (Hill *et al.* 1985).

1.4.2.3 RESTRICTION SITE POLYMORPHISMS

It has been estimated that on average 1 in 80 bp of the α-cluster varies between individuals, some identified as RFLPs. They can be divided into common (the frequency of the rarer allele in most populations being 0.05 to 0.50) and uncommon (<0.05) variants; the latter are frequently race or population specific. Analysis of nine common polymorphisms (illustrated in Figure 1.6), including the pseudozeta/zeta ($\psi\zeta_1/\zeta_1$) polymorphism and the interzeta hypervariable region (IZHVR), showed that linkage disequilibrium occurs between these markers and some common α-globin haplotypes could be derived. Although there is a theoretical total of 768 different haplotypes, only a limited number have been observed, suggesting that linkage disequilibrium exists in the α-globin cluster, as in the β-globin cluster (Higgs *et al.* 1986). The range of haplotypes is also not uniformly distributed.

α-Globin haplotype groups have been defined according to the polymorphisms present at the 5 dimorphic sites located at the 3' end of the haplotype (PZ/Z, *Acc*I, *Rsa*I, α*Pst*1 and θ*Pst*I sites). The group I haplotypes end in (PZ++--), group II in (PZ+---), group III in (Z----), group IV in (PZ+-++) and group VII in (Z+---). Subdivision within a haplotype group is based on the 4 markers (3 dimorphic sites and one triallelic VNTR) at the 5' end of the haplotype. None of the five major haplotypes (Ia, IIa, IIIa, IVa ,VIIa) is related to each other by single mutational events (Higgs *et al.* 1986). In each population only one or two haplotypes are common and several others are present at low frequencies. World-wide, haplotype Ia is the most common, with IIa also being widely distributed. Compared to other populations, only Melanesians have high frequencies of haplotypes IIIa and IVa, but haplotypes Ia and IIa are rare or absent in Melanesians, as they are in the San and West African populations (Hill *et al.* 1989). Haplotype Ia is common throughout Polynesia and IIa is common in South-east Asia (Hill *et al.* 1989, O'Shaughnessy *et al.* 1990). As noted for the β-globin complex (Wainscoat *et al.* 1986), Africans appear to have a wider mixture of haplotypes, including some that are rare in Eurasians (eg. IIIf and Vc) (Flint *et al.* 1993a) . In addition to the usefulness of α-haplotypes that occur on normal chromosomes, α-haplotypes are of value in elucidating the molecular basis and geographical distribution of the common forms of

α-thalassaemia, since particular α-thalassaemia mutations are known to be associated with specific α-haplotypes (Flint *et al.* 1993a).

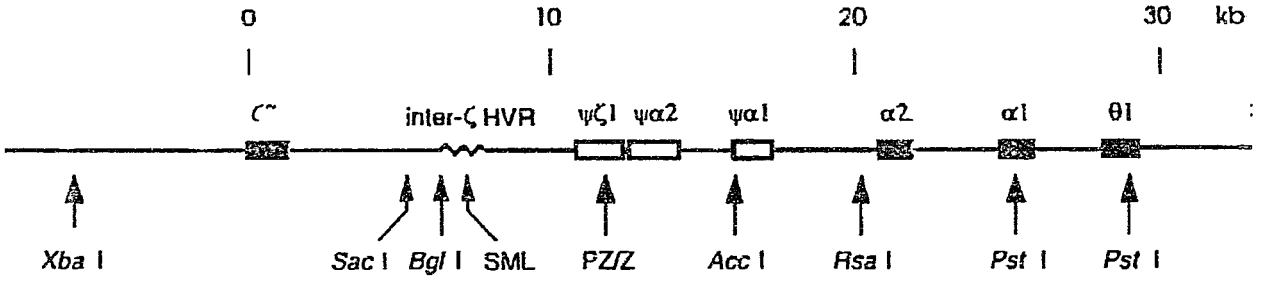


Figure 1.6 The α-globin gene cluster showing the positions of the RFLP sites and HVRs studied (modified from Martinson *et al.* 1996).

1.5 MADAGASCAR

Madagascar is one of the great islands of the world, the fourth largest after Greenland, New Guinea and Borneo. Madagascar is 1600km long and 500km wide, and is situated in the Indian Ocean, 400km from the coast of the African continent. It has been separated from Africa since the end of the Cretaceous period, ie. for about 70 million years, according to evidence from studies of the few remaining biological links with Africa. Most fauna and flora is unique to Madagascar as a result of the island's long separation from other land masses.

Madagascar is environmentally very diverse. The east coast is a narrow strip of land, mainly covered with an impenetrable tropical forest. Introduced crops such as vanilla, cloves, sugar-cane and coffee thrive along this humid stretch of coastline. The central part of the island, the so-called plateau, is cooler and drier and is largely covered by grasslands and savannah. The highlands descend slowly to the west, where the climate becomes tropical again. The south-west is much drier and is semi-desert-like, with the main vegetation a thorny scrub punctuated by baobabs. However, there is a well-watered region along the coastal plains in the south-west, where there are large animal herds and crops of rice, cotton, tobacco and manioc.

1.5.1 THE PEOPLES OF MADAGASCAR

A single term, Malagasy. is used to refer to the island's people and its language. However, there is as much cultural variation in Madagascar as there is diversity of climate, vegetation or wildlife. Madagascar has twenty officially recognised ethnic groups, as well as two "unofficial" groups, the Vakinankaratra (VA) and the Zafisoro (Z), and these are distributed on the island as follows (see Figure 1.7 also):

East-Coastal Madagascar: Betsimisaraka (BS), Antemoro (AM), Antefasy (AF), Antambahoaka (AB), Antesaka (AS), Zafisoro (Z). Tanala (TN), Antanosy (AN).

South-West Madagascar: Antandroy (AD), Bara (BA), Mahafaly (MF), Vezo (VZ), Masikoro (MS).

West Madagascar: Sakalava (SK), Makoa (MK).

North Madagascar: Antankarana (AN), Tsimiheti (TS), Sihanaka (SH).

Central Madagascar: Merina (MR), Betsileo (BT), Bezanozano (BZ), Vakinankaratra (VA) (Brown 1978).

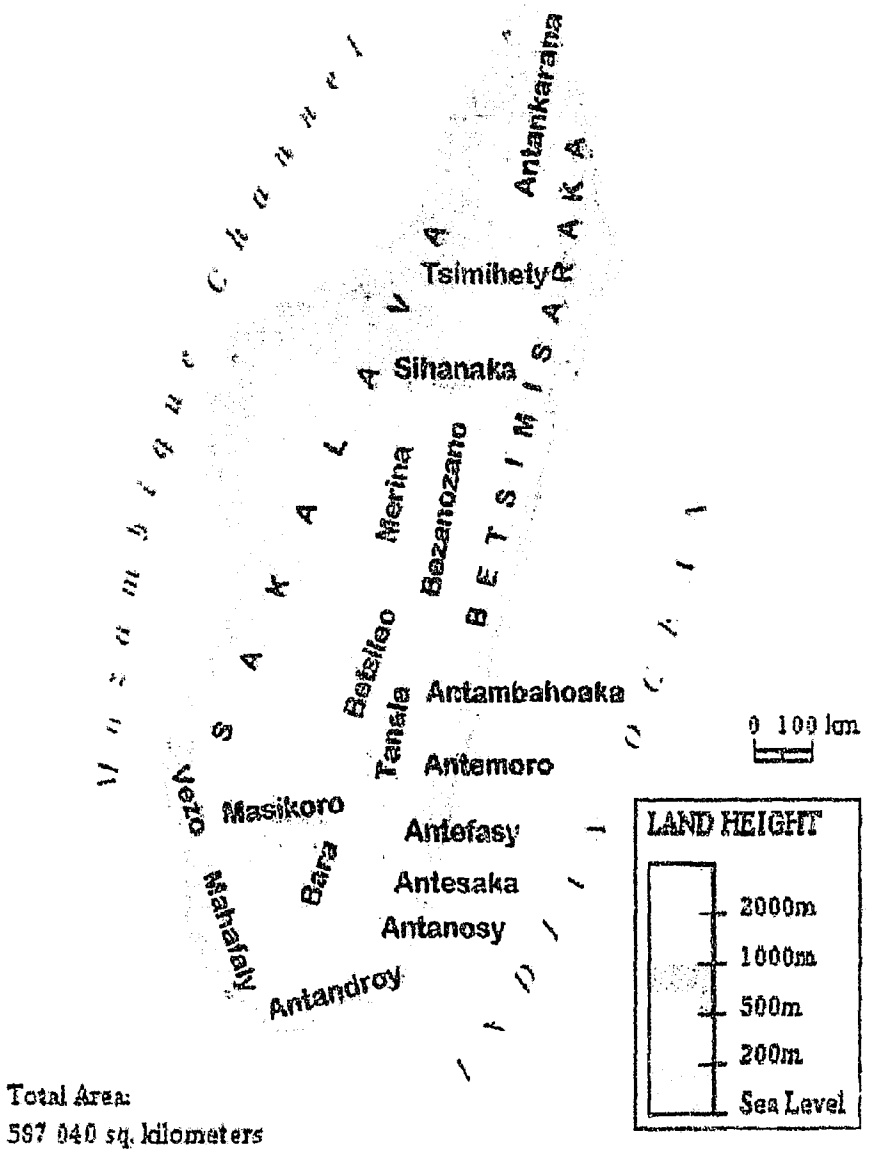


Figure 1.7 Map of Madagascar showing the distribution of the Malagasy ethnic groups.

The largest ethnic groups comprise the Merina and Betsileo. The Merina, Betsileo, Bezanozano and Vakinankaratra (who are of Merina and Betsileo admixture) are usually referred to as the "highland" groups. Tribes living in the peripheral areas are collectively known as *côtiers* (lowlanders), of whom the most numerous are the Betsimisaraka, Tsimihety and Antandroy (Brown 1978). A number of the larger ethnic groups are generally divided into three castes: the nobles (caste I), the commoners or freemen (II), and the former slaves and their descendants (III). Among the Merina, these castes are also called the Andriana, Hova and Mainty respectively. These social distinctions are no longer adhered to very strictly, but they do form the basis of other social and economic distinctions, and are manifest especially on ceremonial occasions. This population structure is seen as a possible problem for sampling the Malagasy population according to ethnic groups, and differentiating racial traits (Dewar 1992).

There is marked diversity of physical appearance, cultural practices and economic systems among the Malagasy. This large scale diversity is a reflection of their many sources of ancestry, of which the major ones are Indonesian, African, Arabic and Indian, and will be discussed in the sections below.





Figure 1.8 Typical market scenes in Madagascar, showing some of the cultural diversity present among the Malagasy

1.5.1.1 PHYSICAL APPEARANCE AND LINGUISTICS

The majority of people are a mixture in varying degree of the extreme African and Asian types. The diversity in the appearance of the Malagasy people has been known for over 400 years - in 1550 Nagnort reported two types of inhabitants, " the ones black with frizzy hair; the other white with long hair" (reviewed in Singer *et al.* 1957). In the central plateau, many people are said to have light skins and straight hair and, in physical features, often resemble Malayans, Indonesians or Indo-Chinese. In the coastal areas however, especially in the west and south, the people are said to be much darker in colour and have many Negroid features, and short, crinkly hair. Within those groups where the caste system operates, there is a complex mixture of physical characteristics. It is generally assumed that the noble caste has the purest Indonesian blood, whereas the slave caste, which formed about one-third of the total population by the mid-nineteenth century, comprised mainly Africans as well as indigenous, local captives (Campbell 1991). In eastern and south-eastern societies, notably the Antaisaka, Antanosy, Antaifasy and Antambahoaka, Arabic (and possibly Indian) blood is considered to be restricted to the noble caste (reviewed in Jenkins 1992).

The principal unifying factor is the common language, which is spoken throughout the island with only relatively minor differences of dialect. Malagasy belongs clearly to the Austronesian family of Indonesian/Polynesian languages, entirely so in its grammar and structure and about 80% in its vocabulary. The most closely related surviving languages to Malagasy are found in the Barito valley of Borneo. A further complication is the influence of Arabic, French, English, and in particular, a Bantu language or languages on Malagasy. This "Bantu substratum" is found in elements of its vocabulary, for example, a possible origin for the name Masikoro, a name applied to a sub-division of the Sakalava, is *mashokora*, a word used in parts of Tanzania for scrub forest of a kind similar to some of the vegetation in south-west Madagascar (Dewar and Wright 1993). Most of the linguistic variation is ordered clinally north to south, with the people of the far north being linguistically different from southern populations, simply due to the fact that the north of the island has received far more migrants over the past century. Thus the linguistic distinctiveness of the Malagasy groups is no guarantee of

their biological and cultural distinctiveness. There are a few villages on the west coast where Makoa, a Bantu language of Mozambique, is spoken by the families of recent immigrants. In the north-west, especially on the island of Nosy-be, there are some speakers of Swahili (Dewar 1992).

1.5.1.2 CULTURE AND AGRICULTURE

"Tradition" in Malagasy is "fomban-drazana", literally, the ways of the ancestors or ancestral practice. It can be said to draw together the diverse cultural practices of the Malagasy, the constant invoking of the ancestors and ancestral practice being the first and most significant point of reference (Mack 1986). However, the culture also shows regional variation, even more so than the language, often because of differing geographical conditions, and sometimes because of the presence of certain ethnic influences in some areas. But in some cases it may be difficult to conclusively identify the sources of the different elements in the culture.

i) African and Indonesian influences

Much of the way of living of the Malagasy is that of the Eastern archipelagos. Rice culture, utensils, weaving of plaited basketry, iron-smelting (using double tuyère bellows), clothing, some musical instruments (notably the tube zither or *valiha*, the conch and the cone drum), totem poles, beliefs and superstitions came from the East (mainly Indonesia) and not from Africa (Singer *et al.* 1957, Mack 1986). The construction of rectangular houses aligned in certain directions, because of the symbolic significance attached to various sides and corners of the house, is of definite Indonesian origin, and occurs mainly in the central highlands. (Figure 1.9) Slash-and-burn rice cultivation, practised in areas of the eastern escarpment is typical of Indonesia and the hill tribes of Indo-China, while knowledge of irrigated rice cultivation, as practised on the plateau, was probably brought at a later stage from India or Ceylon (Figure 1.9).

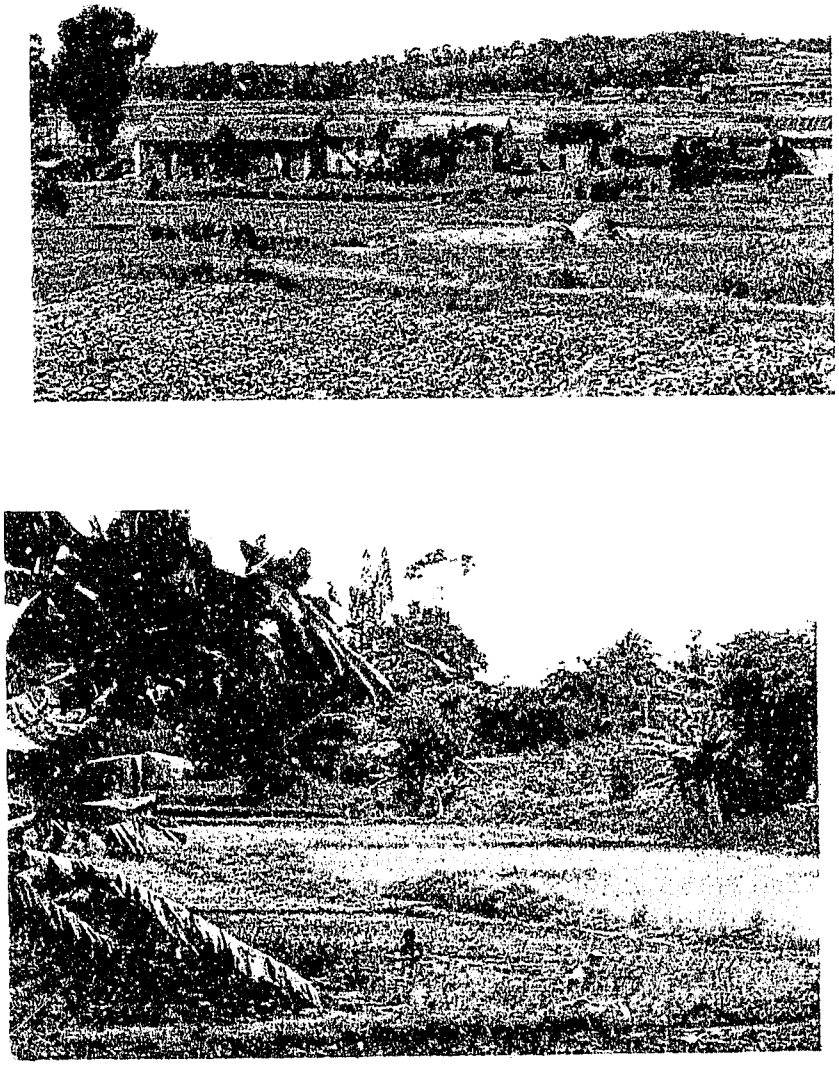


Figure 1.9 Rectangular houses and rice paddies occur mainly in the highlands of Madagascar, and are typically Indonesian.

The building up of herds of Zebu cattle and their association with the ancestors is a well-known feature of some African societies, and is more commonly associated with the land-extensive fishing and swidden economies of the lowland peoples (Campbell 1991). Other fibre weaving practices and musical instruments are of African origin. The Mahafaly, Antandroy and Bara display particularly strong African influence in their cultures (Brown 1978).

Some funeral customs and types of tombs can be traced to Asia and others to Africa: there is a clear African origin for many of the funerary rites among the Sakalava such as the *tromba* (spirit possession) and the *dady* (cult of royal relics), while the sacrificing of cattle and the use of cattle skulls and horns to decorate tombs is a link with Indonesia and Indo-China. Communities in the north east, especially the Betsimisaraka, have various features of social organisation and mortuary ritual (such as second burial and turning of the dead) that are common in South-East Asia (reviewed in Brown 1978).

ii) Arabic and Islamic influences

The evidence of the Arabic infiltration of Madagascar is mainly ethnographic. Traditional histories document that the north west and south east coasts of Madagascar were the areas with closest relations with Islamised traders from nearby islands and the coast of Africa (Oliver and Atmore 1981b). However, Malagasy culture shows evidence of having been very thoroughly infiltrated by broader aspects of Islamic culture, and the Islamic-inspired notions are not limited to the areas of original coastal settlement. In particular, the intricate system of divination or *sikidy*, and the significance attached to different orientations can be related to Muslim inspiration and pervade the island's diverse peoples, regardless of whether or not they can be shown to have some direct contact with the Islamic world. The Malagasy are well-acquainted with the Islamic system of astrology. The Malagasy names for the lunar months and for the days of the week are both direct borrowings from the Arabic. There are numerous other traces of Arab or Islamic cultural influences. The ruins of a few Arab-

style houses and tombs, forts and mosques still remain (Figure 1.10). Some musical instruments (eg. the flute) and textile traditions of Arab origin are still used by the Malagasy today, and in the south east, the turban and fez are worn together with versions of Arab robes (Mack 1986).

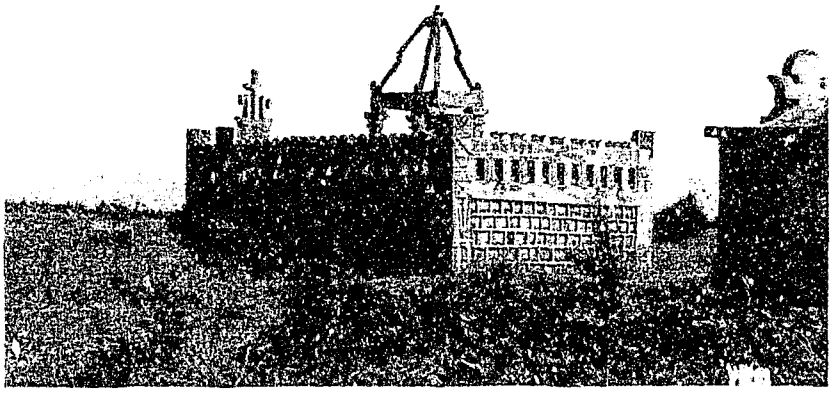


Figure 1.10 An Arab-style tomb on the plateau provides further evidence of Malagasy ancestry.

1.5.2 THE ANCESTRY OF THE MALAGASY: "LA PLUS BELLE ENIGME DU MONDE"

How and why Madagascar was settled remains an open question. To try and address these issues, the three schools of thought (British, French and modern) concerning Malagasy origins, will be discussed in this section (reviewed in Campbell 1995).

1.5.2.1 THE BRITISH SCHOOL

The first attempts at resolving Malagasy ancestry were made in earnest at the beginning of the nineteenth century by Protestant missionaries, who compared the Malagasy to other societies of the Indian Ocean rim in terms of their physiognomy, language and customs. Ellis (in 1859) and later Oliver (in 1868) noted two physiognomic types of Malagasy: the Merina were the most Malay-looking, and the lowland peoples in the south and west of the island were of Negroid ancestry, while the east coast peoples were of mixed ancestry, "more Negroid than the Merina but not fully African". According to Sibree (1880) the East coast peoples were lighter-skinned than the Merina because of the strong genetic input from European traders, pirates and planters. West Madagascar had received a strong genetic input from Africans imported by the Arabs as slaves (reviewed in Campbell 1995). The Imerina, among whom the caste system operated strongly, possessed greater mixture of physical types than other regions of Madagascar due largely to wars in the nineteenth century when women and children from other ethnic groups were brought to the island as slaves. However, the different groups of castes remained distinct because of a rigid caste prohibition against sexual relations between slaves and the "free" population (Campbell 1991).

It was generally accepted that although one language was spoken in Madagascar, its division into different dialects suggested that the people speaking them were of different origins, who moreover arrived on the island at different times (reviewed in Campbell 1995). The year AD 400 was proposed as an estimated date of departure of proto-Malagasy speakers from Indonesia, based on the limited number of Sanskrit loan

words in Malagasy. However this date has since been questioned as it does not agree with that evident from the archaeological findings (Dewar and Wright 1993). Since the oral history reports the Merina to have arrived after the other peoples of the island, it was thought that in order for the "Negroid" people to have arrived before the Merina and yet speak a dialect of the same language, the former were "aboriginal" and the latter late arrivals were conquerors of the first. The Merina and Betsileo were closely related because Betsileo kings were called "Hova", a term also used to refer to the free Merina caste.

Sibree (1880) and Shaw (1885) contended an all Malayo-Polynesian origin to Malagasy ancestry, noting the difficulty of navigation in East African waters, the similarities in the physiognomy of the non-Merina people and the reflection of Melanesian dialects in the Malagasy language. But, they failed to recognise the African influences on Malagasy language, physiognomy and traditions. However Dahle (1883) argued for a dual migration to Madagascar, from the Malayo-Polynesia region and from Africa. He proposed an initial migration from Africa based on its proximity to Madagascar, followed by a deliberate, major immigration from the East by Malayo-Polynesian peoples who conquered and inter-married with the first African settlers on the island (reviewed in Campbell 1995). The general view was that both divisions of Malagasy (African and Malay), whether they voyaged directly across the Indian Ocean or reached Madagascar via Africa, originated from the East.

1.5.2.2 THE FRENCH SCHOOL

The British School laid the basis for the emergence of the French School, the most notable contributions of which were made by the Grandidiers and Ferrand (1880-1970). The Grandidiers drew on Ellis' argument of an Austronesian origin to all Malagasy. They proposed that the Malay-featured Malagasy originated from the Malay-Indonesian region, and the Negroid-featured Malagasy came from Melanesia. By contrast Ferrand argued that most Malagasy originated from Indo-African communities that had formed from the intermarriage between Indonesians and Africans along the East African coast (reviewed in Campbell 1995). However, this argument suffers two weaknesses. Firstly

there is no archaeological evidence of Indonesian settlements on the African coast, and secondly, it is necessary to assume subsequent migrations to account for the regional heterogeneity of the Malagasy (Dewar 1995).

1.5.2.3 THE MODERN SCHOOL

The modern School (1970-present) has given greater clarity to Malagasy origins and the history of the settlement of Madagascar, by revisiting old ground as well as exploiting new techniques especially in linguistics and human genetics.

1.5.2.3.1 Initial settlement of Madagascar

Archaeological evidence from radio carbon datings reveal the first human presence in Madagascar to be certain only from the eighth century AD, and the first permanent human settlements only from the ninth century AD. It is significant that no stone implements, arrow-heads or axe-heads have yet been found on Madagascar, such as have been found in abundance in Africa (Brown 1978, Dewar and Wright 1993). Another sign of recent arrival is the fact that settlement of the entire island is still incomplete. To a lesser degree, the cultural homogeneity and the similarity of racial composition of the Malagasy, are an indication that there has not been much time for the widely separated groups of settlement to form separate cultural and racial identities. Finally, a comparison between the level of culture of the proto-Malagasy and the spread of various cultural features in Indonesia provides some clue as to when the ancestors of the first Malagasy left Indonesia. For example, there are no significant traces of Buddhism in early Malagasy culture, and the proto-Malagasy seem to have been ignorant of iron-working. As both Buddhism and knowledge of iron-working began to appear in Indonesia at the beginning of the Christian era, it is likely that the ancestors of the Malagasy left Indonesia about 2000 years ago (reviewed in Brown 1978).

The Indonesians probably arrived in Madagascar via the northern and western shores of the Indian Ocean, rather than directly across the Indian Ocean, a voyage of about

6400km. The present-day distribution of the Indonesian/Polynesian outrigger canoe in the Indian Ocean (Figure 1.11), and notably in Ceylon, the Maldives, the East African coast and, significantly, the west coast of Madagascar among the Vezo, is support for the northern route. This route also helps to explain the presence of African racial and cultural elements in the basic stratum of the population throughout Madagascar (reviewed in Brown 1978).

The earliest migrants to the uninhabited island would have been fisherman and tropical vegeculturalists, growing bananas and cocao-yams, and wherever there was the possibility of irrigation, rice. They would have begun by colonising the well-watered north-east coastal region of the island. In the drier lowlands of the west and south, they would have been confined to scattered settlements around river mouths. For as long as possible, they would have avoided the central highland region, which was covered with primaeval forest. From an early stage in their settlement, these oceanic pioneers may have been joined by mixed elements from the East African mainland, resulting initially in the establishment of Vezo fishing communities, and, later, in the formation of the pastoral Sakalava society, who would have brought with them the African millet cereals and domesticated animals, especially cattle, the words for which are recognisably of Bantu origin. The presence of the original Bantu communities was thought to be due to slave raiding of Mozambique by the Malagasy. However, there is strong evidence that most of these were not slaves, but deliberate colonists: they lived in independent communities, following African social customs and religious beliefs. Those who live in the southern interior adapted Malagasy speech but retained African clan names (Brown 1978).

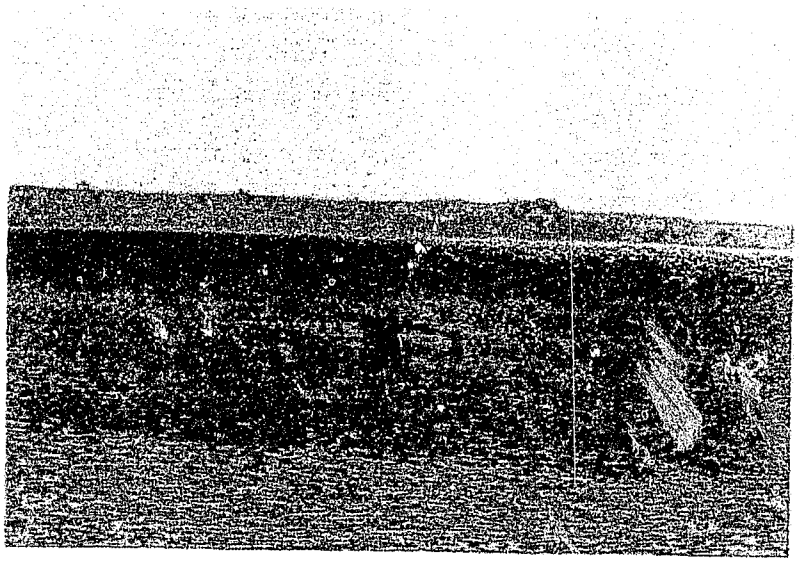


Figure 1.11 An outrigger canoe seen off the coast of Madagascar is of Polynesian/Indonesian origin.

1.5.2.3.2 Later immigrants and settlement

From about the middle of the first millennium, external pressures on the settlers in Africa compelled more and more of them to seek refuge in Madagascar. During the fourteenth and fifteenth centuries, a second and more substantial migration of Indonesians and Africans to Madagascar took place, completing the occupation of the southern and eastern coastal regions. Western Madagascar was extensively colonised by African peoples from Mozambique, who were transported by Arab shipping or else followed by canoe in the wake of trade ships (Oliver and Atmore 1981a). Some original colonists were forced to look for new settlements further inland, where they came to form the people known as the Merina and Betsileo. By the end of the first millennium, the coasts and some inland areas of Madagascar were firmly settled by a group of people, recognised as the ancestors of the present-day Malagasy, racially a mixture of Indonesian and African. These original settlers comprise the Betsimisaraka, Antanosy and Antankarana. Their culture and civilisation was so well established that they were able to absorb new immigrants from different, and sometimes more advanced cultures, with only relatively minor effects on their language and culture (Brown 1978). Archaeological findings have included Jews, Persians, Indians and Chinese among these immigrants, who are said to have arrived on the island in medieval times (Dewar 1992).

However, later immigrants who made a greater impact on the culture and history of Madagascar were Arabs, trading down the east coast of Africa from the tenth century onwards (Oliver and Atmore 1981a). These Arabs would already have had an appreciable African admixture by the time they reached Madagascar, but would also have maintained many Arabic customs. Many peoples of south-east Madagascar, particularly the Antaifasy, Antaimoro and Antambahoaka reveal substantial Arab/Islamic influence (Brown 1978). Once settled on the island, the Arabs were known to have imported African slaves to Madagascar, first to act as their servants and agricultural labourers, and later to sell to the Malagasy (Singer *et al.* 1957). This involvement in the East African slave trade had important influences on the ancestry of certain Malagasy

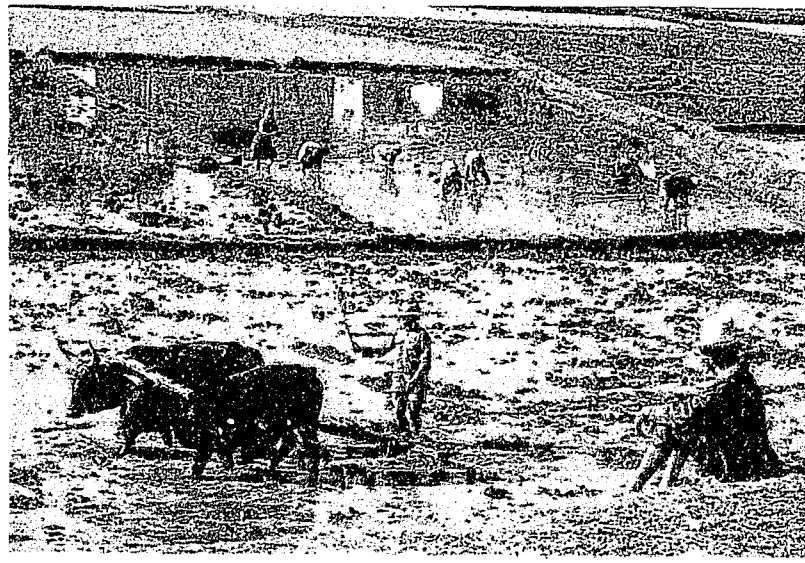


Figure 1.12 *The use of Zebu cattle for rice farming provides an* illustration of the fusion of African and Indonesian cultures in Madagascar.

ethnic groups (Campbell 1995). The Islamised populations, which included the Antemoro and the Antambahoaka, were among the last people to arrive from outside the island before its discovery by the Portuguese in 1500 and the independent documentation which began to build up from that date (Mack 1986). In the north-west, there is a sizeable community of recent Muslim immigrants from the Comoros or Pakistan (Oliver and Atmore 1981b).

1.5.2.3.3 Genetic traits of the Malagasy

"Of the large land masses, the one we know least about the blood factors of the inhabitants is Madagascar", Mourant *et al.* (1976). This statement may be said to still be partially true, since little data have been generated since it was written. Relatively few genetic studies on the Malagasy have been carried out. The first of these studies investigated the incidence of various haemoglobins, haptoglobins, transferrins and other blood factors among the Malagasy, and confirmed the strong African contribution to the ancestry of the Malagasy (Buettner-Janusch and Buettner-Janusch 1964). However, the number and range of the polymorphic systems used was very limited, and relatively few individuals were investigated, providing serious deficiencies in the sampling of the populations. The most recent investigations into Malagasy ancestry using genetic techniques include a mitochondrial DNA study (Soodyall *et al.* 1993), β^S-haplotype analysis (Hewitt *et al.* 1996) and a G6PD study (Dangerfield 1997) of the Malagasy. All of these studies provide evidence for African and Indonesian contributions to Malagasy ancestry, and are discussed in more detail in Chapter 4.

1.5.2.4 SUMMARY OF MALAGASY ANCESTRY

It cannot be stated with certainty which people were the first to arrive on the island of Madagascar. However, the vast amount of anthropological, ethnographic and biological data reflect long-term interactions between diverse populations on and off the island. In particular, the data reveal that the base stock of Malagasy people is a mixture of African and Indonesian immigrants, with the African element predominating today, but speaking an essentially Indonesian language and possessing a culture which contains many elements from both areas of origin.

In addition, several other likely sources of origins of peoples contributing to the modern Malagasy population have been suggested (reviewed by Dewar 1992):

* The coasts of Mozambique and Tanzania - through trans-Mozambique channel slave trade by Muslims of Swahili-Arab origin living in Madagascar in the 1860's (Campbell 1988), and possibly earlier through commerce.

* The Comoros islands - possibly settled a bit earlier and a stop off point for Indonesians, Arabs and other groups travelling down the East African coast. Later, Comorian migrants also settled in north-western Madagascar. From 1785-1820, the Betsimisaraka launched periodic slave raids on the Comoros islands (Campbell 1988).

* The "Swahili Coast" of Tanzania and Kenya - migrants of traders to the north-west and north-east coasts of Madagascar (Oliver and Atmore 1981a).

* The Persian Gulf - many Malagasy claim descent from "Shirazi" ancestors (possessing a strong Negro element); there is some historical evidence of migrations in the 14th and 15th centuries (Oliver and Atmore 1981a).

* "Malay sailors" of the 13th or 14th centuries and perhaps a second "Indonesian wave" - accounts for the distinctively "Asian" appearance of many of the highland groups.

* Europeans - the Portuguese were the first to arrive on the island in 1500, followed by French, and later by British sailors (Oliver and Atmore 1981b).

* Additional associations with Great Zimbabwe, the Rift Valley of East Africa, and South India in the forging of Malagasy society and culture.

1.6 AIMS OF THE STUDY

In light of the unresolved debate concerning the ancestry of the Malagasy and the source(s) of the biological and cultural diversity existing in the population, this study of α- and β-globin genetics was undertaken to try and resolve some of these uncertainties, as part of the larger study of the Department of Human Genetics, SAIMR. In this way, it was hoped to confirm or refute existing theories that have been based on historical and archaeological evidence. There are three main objectives of this study:

1. To analyse the frequencies of the RFLP sites within the β-globin gene cluster statistically to gain information on the population affinities and possible source(s) of origin of the Malagasy.
2. To construct β-haplotypes of Malagasy chromosomes.
3. To compare the Malagasy β-haplotypes to the common haplotypes that have been reported in different populations in the world, specifically those in Africa, Indonesia and Arab-Indian populations, using different statistical methods, to deduce the possible source(s) of ancestry of the Malagasy.
4. To analyse the frequencies of the RFLP sites and HVRs within the α-globin gene cluster statistically to gain additional information on the population affinities and possible source(s) of origin of the Malagasy.

CHAPTER 2

SUBJECTS AND METHODS

2.1 COLLECTION AND STORAGE OF BLOOD

Field trips to Madagascar have been undertaken for two periods each year, beginning in 1992, to collect blood samples from the Malagasy for a large research project entitled "The origins and affinities of the people of Madagascar". Samples from four field trips that were undertaken during the periods May/June 1992, October 1992, June 1993 and October 1993 have been used in this study.

Blood samples were collected in ACD and EDTA tubes from all individuals. Blood samples were flown to Johannesburg for analysis in the South African Institute for Medical Research (SAIMR) laboratories. Two ACD's from each individual were centrifuged in a Beckman benchtop centrifuge for 10-15 minutes at 3000rpm, at 4°C, to separate the blood into its major components. The buffy coat (white blood cells), together with a few red blood cells were removed from the centrifuge tube and stored in 5ml capped test tubes at -20°C until required for DNA extraction.

Throughout the study, the reagents used were those generally available in the laboratory, unless otherwise specified when one particular supplier was used.

2.2 HAEMATOLOGICAL METHODS

All EDTA blood samples were analysed initially in the Department of Haematology, SAIMR. Full blood counts (FBCs) were performed to determine the red cell count (RCC), haematocrit (Hct), mean cell volume (MCV) and mean cell haemoglobin (MCH). Haemoglobin electrophoresis, using cellulose acetate strips was carried out to identify and quantify the haemoglobin type(s) (HbA, HbF, HbS) of each individual. The presence of HbS was confirmed by citrate agar electrophoresis.

2.3 SUBJECTS INCLUDED IN THIS STUDY

Individuals were chosen from six Malagasy ethnic groups: two "highland" populations (Merina and Betsileo), two "lowland" populations (Antesaka and Tsimiheti) and two populations (Mahafaly and Vezo) from the south-west of the island (highlighted in Figure 1.7). The groups chosen cover a broad geographical range of Madagascar and thus provide a reliable representation of the Malagasy population as a whole. Only haematologically normal, specifically non-β^s, non-thalassaemic individuals (MCV >80, MCH >27) were studied. The number of individuals studied in each ethnic group are as follows: Merina: 88; Betsileo: 78; Antasaka: 67; Tsimiheti: 67; Mahafaly: 26. Vezo: 25.

2.4 EXTRACTION OF HIGH MOLECULAR WEIGHT DNA

DNA was extracted from all buffy samples according to the simple salting out method of Miller *et al.* (1988). The DNA was collected by spooling, then dissolved in TE and stored at -70°. Working stock aliquots were stored at 4°C.

2.5 QUANTITATION OF DNA

All the DNA samples were quantified by a rather crude means, but one which was accurate enough to estimate the amount of DNA required for the PCR reactions. An aliquot of DNA (5μl) was electrophoresed on a 0.8% agarose gel, stained with ethidium bromide, together with 10 μl of a 1kb λ ladder, to estimate the approximate DNA concentration. In some cases (α-globin PCR), the exact concentration of DNA was required, and was determined spectrophotometrically. DNA samples were made up to an appropriate dilution and aliquots were placed in quartz cuvettes. OD readings were taken at a wavelength of 260nm in a Varian DMS 70 spectrophotometer. The concentration of DNA was then calculated using the conversion formula : 1 OD ≈ 50μg/ml double-stranded DNA and correcting for the dilution factor (Sambrook *et al.* 1989).

2.6 POLYMERASE CHAIN REACTION

The polymerase chain reaction (PCR) is used to amplify a segment of DNA that lies between two regions of known sequence. In this study, all PCR reactions were carried out in either a Perkin Elmer Cetus® DNA thermal cycler or a Hybaid® thermal reactor. Each reaction mix contained 100 nanograms of genomic DNA; forward and reverse primers of differing concentrations depending on the system; 1/10 volume of 10x deoxynucleotides (adenine, thymine, cytosine and guanine) as defined in Appendix 1; 1 unit of *Taq* DNA polymerase; 1/10 volume of the recommended *Taq* DNA polymerase buffer (10x concentrated, consisting of 100mM Tris-HCl, 15nM $MgCl_2$, 500mM KCi, 1mg/ml gelatine; pH 8.3); and in some cases, spermidine, BSA, Triton X and additional $MgCl_2$, (as specified in Sections 2.6.2 and Table 2.4). For each sample, the reaction mix was overlaid with mineral oil to prevent evaporation of the sample during the reaction.

2.6.1 β^S-MUTATION SCREEN

In order to confirm that the individuals included in this study were non-β^s, all DNA samples were screened for the sickle cell mutation by PCR amplification of the 5' region of the β-globin gene, followed by restriction enzyme digestion with *Dde*I. The primer sequences (supplied by J. Old, personal communication to A. Krause) are as follows:

5' GGCCAATCTACTCCCAGGAG

3' ACATCAAGGGTCCCATAGAC

One hundred nanograms of DNA, 5pmol of each primer, and 1/10 volume of 10x deoxynucleotides were used per PCR reaction; the final volume of the reaction mix was 50µl. The reaction consisted of 30 cycles of denaturation at 94°C for 48 seconds; annealing at 59°C for 48 seconds; and extension at 72°C for 90 seconds. The reaction was completed with a final extension period of 10 minutes at 72°C. The PCR product was checked on a 4% composite gel (3% Nusieve:1% HGT). The size of the expected fragment is 605bp.

2.6.1.1 PCR PRODUCT DIGESTION AND AGAROSE GEL ELECTROPHORESIS

The entire volume of PCR product was digested with 10 units of *Dde*I and 1/10 volume of recommended 10X buffer. Digestion was carried out at 37°C for 2 hours. The products of digestion were checked on 4% composite gels, stained with ethidium bromide. The gels were submerged in troughs containing 1×TBE running buffer, and for each gel, 5µl of a 1kb λ ladder was loaded in the outer well to serve as a molecular weight standard. Electrophoresis was carried out at 100V. The gels were viewed under UV-illumination and recorded with the gel documentation system. Each individual was scored for the presence or absence of the *Dde*I recognition site according to the fragments seen upon UV-illumination of the gel. The absence of the *Dde*I recognition site (on a β^S-chromosome) results in a fragment of length 351bp, but if the site is present (as expected on a β^A-chromosome) fragments of length 150bp and 201bp will be seen. In addition, all individuals will have four constant bands of lengths 89bp, 88bp, 40bp and 37bp. The smaller fragments were not seen since they had run off the gel by the time that the gel had been run far enough in order to separate the fragments of length 201bp and 150bp.

2.6.2 β-GLOBIN CLUSTER

Seven RFLP sites in the β-globin cluster were analysed using PCR, followed by digestion of the PCR product with the appropriate restriction enzyme. Each individual was scored for the presence (+) or absence (-) of each of the seven sites shown in the diagram of the β-globin gene cluster (Figure 1.4).

Each PCR reaction was performed in a final volume of 25µl, and required 2.5µl of a 1/40 dilution of 0.1M spermidine, and 1unit of Taq polymerase (Boehringer Mannheim). Table 2.1 contains the information of the PCR primer sequences, and the amount of primer used per reaction. Table 2.2 shows the PCR conditions used for each RFLP site.

TABLE 2.1

PCR primer sequences and concentrations used for the RFLP sites in the β-globin cluster.

RFLP SITE	PRIMER SEQUENCES[1]	QUANTITY OF PRIMER/PCR REACTION
*Hin*cII/ε	TTAAGAGAGCTAGAACTGGTGAG AAGCCTCATATAAAGGAGCAAATC	5pmol
*Hin*dIII/^{G}γ	AGTGCTGCAAGAAGAACAACTACC CTCTGCATCATGGGCACTGAGCTC	5pmol
*Hin*dIII/^{A}γ	ATGCTGCTAATGCTTCATTAC TCATGTGTGATCTCTCTCAGCAG	5pmol
*Hin*cII/5'ψβ	TCCTATCCATTACTGTTCCTTGAA ATTGTCTTATTCTAGAGACGATTT	5pmol
*Hin*cII/3'ψβ	GTACTCATACTTTAAGTCCTAACT TAAGCAAGATTATTTCTGGTCTCT	2.5pmol
*Ava*II/β	ACTCCCAGGAGCAGGGAGGGCAGG TTCGTCTGTTTCCCATTCTAAACT	5pmol
*Hin*fI/β	TGGATTCTGCCTATTAAAA GGGCCTATGATAGGGTAAT	12.5pmol

[1] Primer sequences are shown from 5' to 3' for the forward and reverse primer respectively. Primer sequences were supplied by Dr J Old in 1990. The *Hin*dIII/^{G}γ, *Hin*dIII/^{A}γ and *Hin*fI/β sequences were previously published (Kulozik et al. 1988, Semenza et al. 1989, Varawalla *et al* 1992).

TABLE 2.2

PCR conditions for the RFLP sites in the β-globin cluster.

	STEP CYCLE			
SYSTEM	DENATURING	ANNEALING	EXTENSION	NO. OF CYCLES
Hinc II/ε, *Hin*dIII/$^{G}\gamma$, *Hin*dIII/$^{A}\gamma$ *Hinc*II/5'ψβ, *Ava*II/β	93°C/ 1min	65°C/ 1min	72°C/1.5min	30
*Hinc*II/3'ψβ	93°C/ 1min	64°C/ 1min	72°C/1.5min	30
*Hin*fI/β	93°C/ 1min	59°C/ 1min	72°C/1.5min	30

The sizes of the expected PCR products as well as the types of gels used to visualise the fragments are specified in Table 2.3.

TABLE 2.3

Sizes of the PCR products expected after amplification and restriction enzyme digestion, and the type of agarose gel used to visualise the fragments.

β-GLOBIN RFLP SITE	PRODUCT SIZE (bp)	AGAROSE GEL	FRAGMENTS AFTER DIGESTION (bp) [1]		
			-	+	C
*Hin*cII/ε	158	3% Nusieve: 1% HGT	158	86/72	
*Hin*dIII/^{G}γ	333	1.5% Nusieve: 1.5% HGT	333	235/98	
*Hin*dIII/^{A}γ	635	1.5% Nusieve: 1.5% HGT	635	327/308	
*Hin*cII/5'ψβ	794	1.5% Nusieve: 1.5% HGT	794	687/107	
*Hin*cII/3'ψβ	914	1.5% Nusieve: 1.5% HGT	914	480/434	
*Ava*II/β	676	1.5% Nusieve: 1.5% HGT	315	214/101	361
*Hin*fI/β	714	3% Nusieve: 1% HGT	341	213/128	244/154

[1] - and + denote absence and presence of the polymorphic site respectively; C denotes constant bands.

2.6.2.1 PCR PRODUCT DIGESTION AND AGAROSE GEL ELECTROPHORESIS

PCR products were digested with 5-10 units of the appropriate restriction enzyme, together with 1/10 volume recommended 10X buffer. Digestion was carried out in the PCR Eppendorf tubes, at 37°C for 2 hours. The products of digestion were electrophoresed in ethidium bromide stained composite gels as specified in Table 2.3. The gels were submerged in troughs containing 1×TBE running buffer, and for each gel, 5μl of a 1kb λ ladder was loaded in the outer well to serve as a molecular weight standard. The gels were viewed under UV-illumination and recorded with the gel documentation system. The expected sizes of fragments following digestion are listed in Table 2.3.

2.6.3 α-GLOBIN CLUSTER

Nine sites were studied (Figure 1.6), including 7 RFLPs and 2HVRs, however results for only seven of these sites could be obtained. The primer sequences and other PCR components are shown in Table 2.4. For completeness, the specifications from the original protocol (Norwich 1997, personal communication to A. Krause) for the two sites *Rsa*I and α*Pst*I for which the template failed to amplify (see below) are also given. Five picomoles of each primer was used in each reaction, together with 1unit of Advanced Biotechnology Taq polymerase, in a total reaction volume of 100μl.

TABLE 2.4 PCR primer sequences and additional component specifications for the α-globin cluster RFLP sites and HVRs.

SITE	PRIMER SEQUENCE[1]	AMOUNT DNA	[$MgCl_2$]	dNTPs	ADDITIONAL COMPONENTS
*Xba*I	CTCAACCAGATGCCAGCCAAACAGAC	500ng	3.0mM	200mM	
	GAACCATTACACCAAGTGAAGGACG				
*Sac*I	AAGATGGAGAAGGAAGGATGACTGC	500ng	2.0mM	200mM	
	GAGAGCAAGCCGATGTGGTAGAATG				
IZ/*Bgl*I[2]	CCAGGACACAGATGGAGGCTAATGT	500ng	1.5mM	200mM	2,5ul of 1/40 dilution of 0.1M Spermidine
	TTGGAGTTGGGTGCATGAGGC				
PZ/Z	GGGTGGGCAGAAGGGGAGAC	500ng	1.0mM	200mM	
	CAGATGCTGGGACTTTAGAGGCTGT				
*Acc*I	TTTTAATAGAGACACAGTTTCACCATG	500ng	3.0mM	100mM	0.1mg BSA, 0.1% Triton-X
	CTTGATTGTTTCTTGATCTGG				2.5ul of 1/40 dilution of 0.1M Spermidine
*Rsa*I	CCTTCTCTCCCCTGTCCTTTCCCTACCC	50ng +	1.0mM	200mM	
	GACCTCCACCCTCCCCCTC				
α-*Pst*I	GGTTGAATGCTCCAGCCGGTTCCAGCTATTGCT	50ng +	2.0mM	200mM	
	GAGTCTGTGGGGACCAGAAGAGT				
θ-*Pst*I	ACGGTGTCTTCGTTATTTCATAGG	500ng	3.0mM	200mM	
	TCAGTGCTTCATTTCTTTTTGTGGC				

[1] Primer sequences are shown from 5' to 3' for the forward and reverse primer respectively. Primer sequences were supplied by Dr J Old (personal communication to Dr A Krause). [2] One PCR system is used to analyse both the IZ HVR and *Bgl*I RFLP.

The PCR conditions for the RFLP sites and HVRs studied in the α-globin cluster are provided in Table 2.5.

TABLE 2.5 PCR conditions for the RFLP sites and HVRs in the α-globin cluster.

SITE	STAGE*	DENATURATION	ANNEALING	EXTENSION
*Xba*I	1	95°C/3.5 min	61°C/2 min	72°C/3.5 min
	2	95°C/1.5 min	62°C/1 min	72°C/3 min
	3	95°C/1.5 min	62°C/1 min	72°C/5 min
*Sac*I	1	98°C/2.5 min	59°C/1.5 min	72°C/1 min
	2	98°C/1.5 min	60°C/1 min	72°C/1 min
	3	98°C/1.5 min	60°C/1 min	72°C/5 min
IZ/*Bgl*I[1]	1	95°C/3.5 min	61°C/2 min	72°C/3 min
	2	95°C/1.5 min	62°C/ 1min	72 °C/3 min
	3	95°C/1.5 min	62°C/1 min	72 °C/5 min
PZ/Z	1	98°C/2.5 min	61°C/1min	72°C/2 min
	2	98°C/1.5 min	62°C/1min	72°C/2 min
	3	98°C/1.5 min	62°C/1min	72 °C/5 min
*Acc*I	1	98°C/2.5 min	54°C/2 min	72°C/3 min
	2	98°C/2 min	55°C/1.5 min	72°C/2 min
	3	98°C/2 min	55°C/1.5 min	72°C/5 min
*Rsa*I	1	98°C/2.5 min	60°C/2 min	72°C/1.5 min
	2	98°C/1.5 min	61°C/1 min	72°C/1.5 min
	3	98°C/1.5 min	61°C/1 min	72°C/5 min
α-*Pst*I	1	95°C/3.5 min	59°C/2 min	72°C/3 min
	2	95°C/1.5 min	60°C/1 min	72°C/3 min
	3	95°C/1.5 min	60°C/1 min	72°C/5 min
θ-*Pst*I	1	95°C/3.5 min	59°C/3 min	72°C/4 min
	2	95°C/1 min	60°C/2 min	72°C/4 min
	3	95°C/1 min	60°C/2 min	72°C/5 min

* Stage 1: 2 cycles, Stage 2 : 27 cycles, Stage 3 : 1 cycle.

[1] One PCR system is used to analyse both the IZ HVR and *Bgl*I RFLP.

The expected sizes of the fragments produced after amplification and restriction enzyme digestion are given in Table 2.6. For the seven α-globin sites for which PCR was successful, the PCR products and the products of restriction enzyme digestion were checked on a 2% HGT agarose gel under the same gel electrophoresis conditions as described in Section 2.6.2.1 for the β-globin RFLP systems.

The template failed to amplify during the PCR reactions for the two RFLP sites *Rsa*I and α*Pst*I, despite much troubleshooting. The trials included using:

- a range of 10X dNTPs (1/20, 1/10, 1/5 volume per PCR reaction)
- a range of primer concentrations (2.5pmol, 5pmol, 10pmol, 15pmol)
- a range of $MgCl_2$ concentrations (0.5 - 3mM at 0.5mM intervals)
- a special α-globin PCR reaction buffer (specified in Appendix 1)
- BSA (100mg/ml)
- Formamide (0.5%)
- Triton X-100 (0.1%)
- Spermidine (2.5μl of 1/40 dilution of 0.1M stock)

The PCR conditions were also altered in order to get a PCR product, including using a range of annealing temperatures, from 54°C upwards; longer extension times; and a range of the number of cycles. For *Rsa*I, a new primer was also synthesised (GCAGGAGACAGCACCATGGT instead of GACCTCCACCCTCCCCCTC) since no oligonucleotide complementary to the given primer sequence could be found from a search of the gene sequence. The sequence of the reverse primer was found in the gene search, at the expected position (818 bp from the end of the forward primer). However either no PCR product or a PCR product of an incorrect length was obtained. (The sequences of the primers for the αPstI site were found on the same gene search. The problem of incorrect primer sequence was then excluded in this case). Thus it was decided to generate data for the α-globin cluster using seven sites only, and unfortunately haplotypes could not be constructed for comparison with other world populations, since the 3' haplotypes determine the major α-globin haplotype groups.

TABLE 2.6

Sizes of the PCR products and restriction enzyme digestion fragments expected for the RFLP sites and HVRs in the α-globin gene cluster, and the percentage agarose gel used to visualise the fragments.

		DIGESTION PRODUCTS (bp)[1]		
SITE	**PCR PRODUCT (bp)**	**+**	**-**	**AGAROSE GEL**
*Xba*I	950	586	950	2% HGT
		364		
*Sac*I	560	366	560	2% HGT
		194		
IZ/*Bgl*I[2]	S - 900	M - 1600	S - 900	2% HGT
	M -1800		M - 1800	
	L -2600		L - 2600	
PZ/Z	PZ - 989		PZ -989	2% HGT
	Z -620		Z - 620	
*Acc*I	826	434	826	2% HGT
		392		
*Rsa*I	818	273	818	2% HGT
		545		
α*Pst*I	937	435	594	2% HGT
		343	343	
		159		
θ*Pst*I	887	715	887	2% HGT
		172		

[1] + and - denote presence and absence of the polymorphic site respectively.

[2] S,M and L denote the small, medium and large variants of the IZ HVR. One PCR system is used to analyse both the IZ HVR and *Bgl*I RFLP.

2.7 CALCULATION OF ALLELE FREQUENCIES AND HAPLOTYPE CONSTRUCTION.

The frequency of the alleles of each β- and α-globin RFLP site and α-globin HVR variant was calculated for each of the 6 ethnic groups studied. Standard errors (SE) were calculated according to the formula

$$SE = \sqrt{\frac{p(1-p)}{n}}$$, n = number of chromosomes.

β-globin haplotypes of each individual's two chromosomes were constructed where possible, on the basis of homozygosity/hetereozygosity at one site only. The frequencies of these unambiguous β-globin haplotypes in the Malagasy were compared to the haplotype frequencies of other populations that are documented in the literature, using a few analytical methods.

The haplotypes present in an individual can only be resolved if all RFLP sites are homozygous, or if only one site is heterozygous. If an individual is heterozygous at more than one site, then the haplotype can only be determined with the assistance of genealogical data, so that the coinheritance of RFLPs can be followed. In the absence of such data, generally, the haplotypes of fewer than half of the individuals sampled can be resolved since multiple-locus heterozygosity masks genotypic categories. The correspondence between genotypic and phenotypic categories may be further obscured by recessive alleles (not applicable in this study since the alleles are co-dominant). Genetic estimation is also complicated by the fact that with high polymorphism, sample sizes tend to be smaller than the number of genotypic categories, so that any particular genotype is likely to be unique or absent in a sample (Long *et al.* 1995).

The approach of ignoring multiply heterozygous samples is wasteful of data and may also introduce an ascertainment bias into the overall haplotypes. Most multiply heterozygous individuals will be heterozygous for at least one of the high-frequency haplotypes, and the majority of rare haplotypes will be present only in heterozygous individuals. Consequently the haplotype spectrum obtained will underestimate both the

frequencies of the common haplotypes and the number of rare haplotypes (Martinson *et al.* 1995). This problem can be addressed by using the statistical approach of maximum-likelihood estimation. The expectation-maximisation (EM) algorithm (written by Dempster *et al.* 1977) is an efficient method for estimating allele and haplotype frequencies in multiple highly polymorphic genetic systems (Long *et al.* 1995). For those individuals who were heterozygous at more than one site in the β-globin gene cluster, a computer programme (EM UTILITY) based on the EM algorithm was used to obtain maximum-likelihood estimates of β-globin haplotype frequencies for each ethnic group. The EM algorithm finds the allele frequencies that maximise the probability of obtaining the observed data. The EM algorithm works by first guessing the allele frequencies, then calculating the expected genotype frequencies using the guessed allele frequencies and assuming Hardy-Weinberg proportions. Using the expected genotype frequencies, maximum likelihood estimates of allele frequencies are obtained, and the cycle is repeated until allele frequencies converge to stable values, which are the maximum likelihood values.

2.8 STATISTICAL ANALYSIS OF THE DATA

2.8.1 CHI-SQUARED (χ^2)ANALYSIS

In order to compare the allele frequencies in the α- and β-globin gene clusters and β-globin haplotype frequencies, and to assess any significant differences between the Malagasy ethnic groups and other populations, the Easistat computer programme was used to generate chi-square values and their corresponding p values. Yates' correction was used to correct for the small sample size. The differences were assessed at the 99% significance level. The χ^2 analysis is a crude measure of distance between different populations: a large χ^2 value, with a correspondingly small p value, indicates a relatively large difference between two populations. A more accurate measure of the degree of difference or relatedness of two populations may be obtained by calculating genetic distances.

2.8.2 GENETIC DISTANCE AND PHYLOGENETIC TREE ANALYSIS.

Genetic diversity within the Malagasy and between the Malagasy ethnic groups and other world populations was estimated by calculating genetic distances, based on 5' β-globin haplotype frequencies. Two programmes were used: the first was DISPAN (Genetic Distance and Phylogenetic Analysis, written by Tatsuya Ota of Pennsylvania State University, 1993), which constructs trees from a matrix of genetic distances computed according to Nei's standard genetic distance formula. The second programme used was PHYLIP (Phylogeny Inference Package, version 3.5 written by Joseph Felsenstein, 1995), which computes three different genetic distances from allele/haplotype frequency data, according to three different formulae: Nei's standard genetic distance, the Cavalli-Sforza chord measure, and the genetic distance of Reynolds *et al.* The three measures have different assumptions. All assume that all differences between populations arise from genetic drift. Nei's distance is formulated for an infinite isoalleles model of mutation, in which there is a rate of neutral mutation. It is assumed that all loci have the same rate of neutral mutation, and that the genetic variability initially in the population is at equilibrium between mutation and genetic drift, with the effective population size of each population remaining constant. The two other distance measures assume that there is no mutation, and that all gene frequency changes are due to genetic drift alone. However they do not assume that population sizes have remained constant and equal in all populations. They cope with changing population size by having expectations that rise linearly not with time, but with the sum over time of 1/N, where N is the effective population size. Thus if population size doubles, genetic drift will be taking place more slowly, and the genetic distance will be expected to be rising only half as fast with respect to time. Both distances are different estimates of the same quantity under the same model (Felsenstein, 1995:PHYLIP package manual).

Genetic distance provides an estimate of a parameter (time or cumulated inverse effective population size) of the model which is thought to have generated the differences seen. A Phylogenetic tree can be viewed as a simplified description of a matrix of distances. It offers a graphic aid for visualising the complex relationships

between populations and a path to infer the possible evolutionary history behind them. Standard Neighbour Joining trees were constructed by both DISPAN and PHYLIP programmes. Neighbour Joining is a distance matrix method producing an unrooted tree without the assumption of a clock. A tree specifies no order to the populations, other than that available at the nodes, and rotation of the nodes involves no loss of information. The lengths of the branches are proportional to the distance between the populations, and thus the relationships between populations may be inferred according to their position and clustering on the tree (Cavalli-Sforza *et al.* 1994). DISPAN uses bootstrap analysis to test for the effects of sampling error on the tree. In this study, 100 repetitions of the bootstrapping procedure were performed.

2.8.3 PRINCIPLE COMPONENT (PC) ANALYSIS

Principle Component Analysis is a procedure for satisfying multivariate data with a minimum loss of information. Principle components offer the simplest mode of analysis of a set of populations-by-gene frequency data. A PC may be described as a weighted average of all the gene frequencies of that population, with weights calculated so as to maximise the amount of information about a population that can be condensed into a single metric value. For p alleles, the first PC is obtained by plotting the frequencies of all p alleles in a p-dimensional space. A straight line is then drawn as close as possible to the points, such that the sum of the distances of the points to the line is a minimum. The points are then projected orthogonally onto the "regression line" thus obtained to produce the (first) principle axis or component. By replacing the original data with their first PC value, some information has been lost, but the information is now contained in a simpler, one-dimensional display instead of a bidimensional display. One can continue to add more axes, each orthogonal to and crossing at the centre of the previous PC, and projecting the points of each population orthogonally onto the line. However, it is usually sufficient to stop at the second and third PC's, and produce a two- or three-dimensional picture, since the highest PC's contain the most information. Further components will recover smaller and smaller amounts of the residual variation. The amount of information recovered by the various components can be calculated from the eigenvalues of the

data matrix, which are obtained by squaring the singular values generated in the computer programme. The two PC's retaining the most information can be plotted against each other as abscissa and ordinate of a Cartesian plane to produce a two-dimensional PC map. In such maps the two axes are scaled to express the relative importance of the PC's: the first is more important than the second etc. Such a plot provides a genetic "map" of similarity among groups, which can be compared to geography, or to cultural or linguistic indicators of proximity. The distances of the populations in the map mimic their genetic differences and the patterning of local differentiation, but relationships may be distorted depending on the amount of information lost. PC analysis is extremely useful in simplifying a data set containing a large number of variables and reducing it to two- or three-dimensions, where the loss of the total genetic information is minimised (Cavalli-Sforza *et al.* 1994).

2.8.4 HETEROZYGOSITY AND DISTANCE FROM THE CENTROID

To assess the relative amount of gene flow experienced by the Malagasy, and the correspondence between social structure and genetic structure, a model of regression of heterozygosity on genetic distance, described by Harpending and Ward (1982) was used. This model evaluates the extent to which local genetic structures result from local social structures, and the extent to which they result from other influences.

Assuming an island model of population structure, the genetic distance of an island from the gene frequency centroid (the overall mean gene frequencies of the population) and the relative homozygosity of that island should be linearly related, if exchange with populations outside the region is the same for each island. If migration declines monotonically with distance, islands on the edge of the collection (ie. furthest away) should show the largest genetic distance from the centroid, and they should lose genetic variability the most rapidly because they are peripheral (and thus be more genetically homogeneous).

$$h = \left(\frac{N}{N-1}\right)\left(1-\sum_{i=1}^{J} x_i^2\right)$$

N = number of chromosomes, x = haplotype frequency, j = number of haplotypes in the population.

In order to calculate the distance from the centroid (r_i) for a population i with k haplotypes, we first calculate the distance from the centroid (r_j) per haplotype, where

$$r_j = (p_j - P)^2/(P)(1 - P)\ ,\ \ j = 1,2, \ldots..,k$$

and p_j and P are the frequencies of the j'th haplotype in population i and in the total population respectively (Batzer *et al.* 1996).

Then r_i is the average of the r_j's, ie.

$$r_i = \frac{\sum_{j=1}^{k} r_j}{k}$$

The heterozygosity and distance from the centroid for each population are plotted as abscissa and ordinate on a Cartesian plane, forming a scatter of points.

Next, the expected heterozygosities are calculated for each population. The theoretical expectation is that there should exist a linear relationship between heterozygosity and distance from the centroid. The expected heterozygosity (h_i) for a population I, is calculated according to the formula:

$$h_i = H(1 - r_i)$$

where H is the mean of the haplotype frequencies of all the populations included , and r_i is the distance from the centroid of population i as calculated above. These values of expected heterozygosity are plotted against the corresponding distance from the centroid for each population, on the same Cartesian plane as above. The points form a straight line, the so-called line of theoretical regression, from which high or low outliers may be identified. In this type of analysis, the outliers are of particular interest, since populations that have experienced more gene flow than average should be more heterozygous, and will thus will fall above the theoretical linear expectation, while isolated populations that have experienced less gene flow than average will fall below the theoretical prediction. In this analysis, and in

particular when used for the Malagasy, one also needs to take into account the sample size of the population and consider the effects of genetic drift.

2.8.5 ESTIMATION OF GENETIC ADMIXTURE

Elston's least-squares method (Elston 1971) was used to estimate the proportional contributions of the proposed parental populations in the formation of the "hybrid" Malagasy population, in a further attempt to clarify the historical affinities between these populations. This model is applicable for an admixture model with an arbitrary number of parental populations, with the requirement that the number of independent alleles/haplotypes does not exceed the number of proportional contributions of the parental populations in the hybrid gene pool.

In the model, there are two crucial assumptions:
1. The exact ethnic composition of the ancestral populations and their allele frequencies must be known.
2. There must have been no systematic change in allele frequencies owing to factors other than gene flow between ancestral and contemporary generations within each parental population and in the hybrid (Chakraborty 1986).

For j parental populations and k haplotypes, the least squares method estimates of admixture proportion (m) is obtained from the equation

$$m = (X'X)^{-1}X'y$$

(provided X'X is a non-singular matrix), where the matrix X and the vector y are alterations of the matrix P(a k x j matrix of k haplotype frequencies in j parental populations), and the vector p_h (a k x 1 vector of the corresponding haplotype frequencies in the hybrid population (Long *et al.* 1991).

If P is represented by its column vectors p_1, p_2, ...p_j, then the ith column of X is obtained as $x_i = p_i - p_j$, and y is obtained as $y = p_h - p_j$. The calculation provides results for the first (j - 1) parental populations and the admixture estimate for the jth parental population is obtained from the equation

$$m_j = 1 - \sum_{i=1}^{j-1} m_i$$

(Long *et al.* 1991).

CHAPTER 3

RESULTS AND DISCUSSION

3.1 ALLELE FREQUENCIES OF β-GLOBIN RFLP SITES.

The frequencies of the presence (+) of the seven RFLP sites studied in the β-globin cluster in each of the six Malagasy ethnic groups studied are shown in Table 3.1. (The difference in the number of chromosomes analysed at each polymorphic site, within each Malagasy ethnic group, reflects the fact that some difficulty was encountered in amplifying or restricting older or contaminated samples. Only a limited number of individuals were available from the Vezo and Mahafaly groups, hence the small sample sizes of these two groups).

The six populations were then grouped as highlanders (MR and BT), lowlanders (AS and TS) or south-westerners (MF and VZ) according to their regional distribution on the island, and the allele frequencies of each of the seven RFLP sites for this classification were calculated (Table 3.2).

Table 3.1 Frequencies of the presence (+) of the β-globin RFLP sites and their standard errors (S.E.) in the six Malagasy ethnic groups studied.

	Merina			Betsileo			Antasaka			Tsimiheti			Mahafaly			Vezo		
	n[1]	freq.	S.E.	n[1]	freq.	S.E.	n[1]	freq.	S.E.	n[1]	freq.	S.E.	n[1]	freq.	S.E.	n[1]	freq.	S.E.
Hincll/ε	176	0.472	0.038	156	0.436	0.039	132	0.394	0.042	134	0.351	0.041	52	0.250	0.059	46	0.174	0.057
*Hind*III/$^{G}\gamma$	176	0.386	0.038	152	0.263	0.035	134	0.269	0.038	134	0.388	0.042	50	0.460	0.071	44	0.318	0.071
*Hind*III/$^{A}\gamma$	168	0.226	0.033	154	0.162	0.030	134	0.907	0.025	134	0.202	0.035	44	0.159	0.055	42	0.095	0.045
*Hinc*II/5'ψβ	172	0.122	0.025	152	0.072	0.021	132	0.099	0.025	134	0.134	0.030	48	0.063	0.035	36	0.056	0.038
*Hinc*II/3'ψβ	166	0.530	0.038	144	0.590	0.041	120	0.558	0.045	118	0.593	0.045	50	0.700	0.065	32	0.438	0.088
*Ava*II/β	172	0.808	0.030	138	0.848	0.031	130	0.823	0.034	126	0.833	0.033	50	0.800	0.057	46	0.891	0.045
*Hinf*I/β	172	0.715	0.034	142	0.634	0.041	132	0.697	0.039	124	0.621	0.044	50	0.300	0.065	34	0.500	0.086

[1]n is the number of chromosomes studied

Table 3.2 Frequencies of the presence (+) of the β-globin RFLP sites and their standard errors (S.E.) in the highland, lowland and south-western groups.

	HIGHLAND[1]			LOWLAND[1]			SOUTH-WESTERN[1]		
	n[2]	freq.	S.E.	n[2]	freq.	S.E.	n[2]	freq.	S.E.
HincII/ε	332	0.460	0.027	266	0.370	0.030	98	0.210	0.041
*Hind*III/$^{G}\gamma$	328	0.330	0.025	268	0.330	0.028	94	0.390	0.049
*Hind*III/$^{A}\gamma$	322	0.200	0.023	268	0.150	0.021	86	0.130	0.037
*Hinc*II/5'ψβ	324	0.100	0.017	266	0.120	0.020	84	0.060	0.025
*Hinc*II/3'ψβ	310	0.560	0.028	238	0.580	0.033	82	0.600	0.054
*Ava*II/β	310	0.830	0.021	256	0.830	0.024	96	0.840	0.037
*Hinf*I/β	314	0.680	0.027	256	0.660	0.030	84	0.380	0.052

[1] HIGHLAND comprise Merina and Betsileo; LOWLAND comprise Antasaka and Tsimiheti; and SOUTH-WESTERN comprise Vezo and Mahafaly.

[2] n is the number of chromosomes studied

3.1.2 CHI-SQUARE (χ^2) VALUES AND SIGNIFICANCE TESTING OF RFLP FREQUENCIES.

The results of the pairwise chi-square tests performed for each of the seven β-globin RFLP sites for the six Malagasy ethnic groups and several of their proposed founding populations are given in Tables 3.3 to 3.9 respectively. Only differences significant at the 99% confidence interval are highlighted. In the χ^2 calculations, the frequencies of the presence of the sites in Melanesia, Polynesia, Micronesia and Indonesia were obtained from Chen *et al.* (1990), and those in the Bantu and San, were from Ramsay and Jenkins (1987). Unfortunately there were no frequencies published for the *Ava*II and *Hinf*I sites in any South-East Asian population.

There are few significant differences between the six Malagasy ethnic groups, with the exception of the Vezo, who differ significantly from the Merina and Betsileo at the *Hinc*II/ε site (Table 3.3) and the Mahafaly, who differ significantly from the Merina at the *Hinc*II/ε site (Table 3.3) and from all the other five Malagasy groups at the *Hinf*I site (Table 3.9). However, when the Malagasy are compared to other world populations, significant differences between the Malagasy and the Melanesians, Polynesians and Micronesians are observed at almost all of the seven RFLP sites. There are unexpected differences between the Malagasy groups and the Indonesians at three of the five 5' β-globin RFLP sites (*Hinc*II/ε, *Hind*III/$^{A}\gamma$ and *Hinc*II/3'ψβ), in particular at the *Hind*III/$^{A}\gamma$ site, where the Merina and Betsileo show significant differences from the Indonesians but the Mahafaly and Vezo do not. This contradicts some of the other evidence (mainly the historical and ethnographic data) which suggests that the highland populations have a greater Indonesian contribution to their ancestry than the lowland populations. The Malagasy groups do not differ significantly from the Bantu at many of the seven sites: the χ^2 results for the *Hinc*II/ε site (Table 3.3) are as expected, where the Merina and Betsileo show the greatest differences from the Bantu, but at the *Hinf*I site (Table 3.9), the differences are unexpected since the Mahafaly and Vezo differ significantly from the Bantu but the Merina and Betsileo do not.

Table 3.3 Chi-square values and p values (df = 1) for a pairwise comparison of Malagasy groups and other world populations at the *HincII*/ε site.

	BT	AS	TS	MF	VZ	MELANESIAN	POLYNESIAN	MICRONESIAN	INDONESIAN	BANTU	SAN
MR[1] χ^2	0.293	1.545	4.08	7.202	12.158	25.64	96.478	102.913	23.702	13.301	26.902
p	0.5882	0.2138	0.0434	0.0073	0.0005	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0003	$< 10^{-4}$
BT[1] χ^2		0.36	1.843	4.913	9.303	31.545	104.45	111.101	28.056	10.259	22.606
p		0.5487	0.1746	0.0266	0.0023	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0014	$< 10^{-4}$
AS[1] χ^2			0.362	2.782	6.439	37.555	112.229	119.119	32.872	7.191	18.09
p			0.5473	0.0953	0.0112	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0073	$< 10^{-4}$
TS[1] χ^2				1.31	4.248	49.327	127.373	134.548	40.697	4.839	14.516
p				0.2525	0.393	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0278	0.0001
MF[1] χ^2					0.448	38.196	113.071	120.946	37.821	0.603	5.789
p					0.5032	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.4375	0.0161
VZ[1] χ^2						46.887	126.173	134.607	45.34	0.033	2.144
p						$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.8563	0.1431

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

Table 3.4 Chi-square values and p values (df = 1) for a pariwise comparison of Malagasy groups and other world populations at the *Hind*III/$^{G}\gamma$ site.

	BT	AS	TS	MF	VZ	MELANESIAN	POLYNESIAN	MICRONESIAN	INDONESIAN	BANTU	SAN
MR[1] χ^2	5.062	4.215	0.008	0.598	0.439	4.819	59.033	64.594	5.506	0.226	12.398
p	0.0245	0.0401	0.9304	0.4392	0.5078	0.0281	$< 10^{-4}$	$< 10^{-4}$	0.0189	0.6345	0.0004
BT[1] χ^2		0.001	4.535	5.907	0.279	0.287	27.541	31.764	0.07	4.441	26.395
p		0.9768	0.0332	0.0151	0.5976	0.5923	$< 10^{-4}$	$< 10^{-4}$	0.7908	0.0351	$< 10^{-4}$
AS[1] χ^2			3.807	5.273	0.194	0.142	27.529	31.768	0.119	3.932	24.424
p			0.051	0.0217	0.6593	0.706	$< 10^{-4}$	$< 10^{-4}$	0.7299	0.0474	$< 10^{-4}$
TS[1] χ^2				0.511	0.426	4.049	54.988	60.511	5.069	0.183	11.322
p				0.4747	0.5139	0.0442	$< 10^{-4}$	$< 10^{-4}$	0.0244	0.6686	0.0008
MF[1] χ^2					1.423	5.251	51.075	57.052	6.511	$< 10^{-4}$	4.08
p					0.233	0.0219	$< 10^{-4}$	$< 10^{-4}$	0.0107	0.9835	0.0434
VZ[1] χ^2						0.044	23.252	27.347	0.601	0.927	10.856
p						0.8336	$< 10^{-4}$	$< 10^{-4}$	0.4382	0.3358	0.001

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS - Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

Table 3.5 Chi-square values and p values (df = 1) for a pairwise comparison of Malagasy groups and other world populations at the *Hind*III/$^{A}\gamma$ site.

	BT	AS	TS	MF	VZ	MELANESIAN	POLYNESIAN	MICRONESIAN	INDONESIAN	BANTU	SAN
MR[1] χ^2	1.696	7.966	0.143	0.58	2.829	34.586	37.221	24.54	15.065	1.015	4.396
p	0.1929	0.0048	0.7055	0.4461	0.0926	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0001	0.3137	0.036
BT[1] χ^2		2.13	0.501	0.033	0.706	14.651	22.135	11.737	7.803	0.002	1.477
p		0.1445	0.4789	0.8567	0.4006	0.001	$<10^{-4}$	0.0006	0.0052	0.9625	0.2242
AS[1] χ^2			4.966	0.733	0.07	1.991	8.589	2.305	1.982	0.434	0.005
p			0.0258	0.3919	0.7908	0.1582	0.0034	0.129	0.1592	0.5099	0.9454
TS[1] χ^2				0.16	1.81	23.628	29.971	18.039	11.628	0.401	3.018
p				0.6893	0.1786	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0006	0.5266	0.0823
MF[1] χ^2					0.317	5.105	13.485	5.356	4.57	0.013	0.755
p					0.5732	0.0239	0.0002	0.0207	0.0325	0.9095	0.385
VZ[1] χ^2						0.419	4.156	0.686	0.818	0.167	0.013
p						0.5174	0.0415	0.4075	0.3659	0.6829	0.91

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS - Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

Table 3.6 Chi-square values and p values (df = 1) for a pairwise comparison of Malagasy groups and other world populations at the *Hinc*II/5'ψβ site.

	BT	AS	TS	MF	VZ	MELANESIAN	POLYNESIAN	MICRONESIAN	INDONESIAN	BANTU	SAN
MR[1] χ^2	1.718	0.215	0.021	0.827	0.749	17.114	7.878	5.794	0.484	0.002	5.802
p	0.19	0.6428	0.8842	0.3632	0.3868	$< 10^{-4}$	0.005	0.0161	0.4865	0.9676	0.016
BT[1] χ^2		0.331	2.359	0.008	0	27.575	1.295	0.519	3.969	0.196	12.638
p		0.5651	0.1246	0.9276	0.9938	$< 10^{-4}$	0.255	0.4711	0.0463	0.658	0.0004
AS[1] χ^2			0.518	0.206	0.222	18.241	3.846	2.425	1.446	0.016	7.717
p			0.4716	0.6498	0.6377	$< 10^{-4}$	0.0499	0.1194	0.2291	0.8983	0.0055
TS[1] χ^2				1.152	1.022	11.651	9.008	6.812	0.132	0.058	4.142
p				0.2831	0.312	0.0006	0.0027	0.0091	0.716	0.81	0.0418
MF[1] χ^2					0.111	9.956	0.101	0.001	1.953	0.16	6.395
p					0.7393	0.0016	0.7506	0.9749	0.1623	0.6887	0.0114
VZ[1] χ^2						7.827	0	0.056	1.687	0.183	5.361
p						0.0051	1	0.8125	0.194	0.6692	0.0206

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

Table 3.7 Chi-square values and p values (df = 1) for a pair-wise comparison of Malagasy groups and other world populations at the *Hinc*II/3'ψβ site.

	BT	AS	TS	MF	VZ	MELANESIAN	POLYNESIAN	MICRONESIAN	INDONESIAN	BANTU	SAN
MR[1] χ^2	0.901	0.124	0.872	3.857	0.588	16.671	106.34	107.401	20.374	16.557	31.614
p	0.3426	0.7246	0.3505	0.0495	0.4432	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$
BT[1] χ^2		0.158	0.006	1.457	1.901	26.203	121.893	123.054	27.897	11.332	24.594
p		0.6907	0.9377	0.2274	0.1679	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0008	$< 10^{-4}$
AS[1] χ^2			0.171	2.391	1.036	17.309	105.385	106.427	21.496	13.233	27.04
p			0.6794	0.1221	0.3087	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0003	$< 10^{-4}$
TS[1] χ^2				1.283	1.886	23.071	115.675	116.788	26.113	10.66	23.592
p				0.2573	0.1697	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0011	$< 10^{-4}$
MF[1] χ2					4.553	22.678	112.315	113.403	27.662	3.287	12.574
p					0.0329	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	$< 10^{-4}$	0.0698	0.0004
VZ[1] χ^2						0.792	41.453	42.024	3.7	15.026	29.283
p						0.3735	$< 10^{-4}$	$< 10^{-4}$	0.0544	0.0001	$< 10^{-4}$

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

Table 3.8 Chi-square values and p values (df = 1) for a pair-wise comparison of Malagasy groups and other world populations at the *Ava* II/β site.

	BT	AS	TS	MF	VZ	BANTU	SAN
MR[1] χ^2	0.585	0.033	0.164	0.006	1.214	3.186	0.298
p	0.4443	0.8562	0.6851	0.9401	0.2705	0.0743	0.5854
BT[1] χ^2		0.146	0.023	0.312	0.239	1.492	1.396
p		0.7027	0.8783	0.5766	0.625	0.222	0.2373
AS[1] χ^2			0.003	0.021	0.727	2.4	0.559
p			0.9587	0.886	0.3938	0.1213	0.4548
TS[1] χ^2				0.093	0.489	1.982	0.828
p				0.761	0.4846	0.1592	0.3628
MF[1] χ^2					0.902	2.525	0.058
p					0.3424	0.112	0.8092
VZ[1] χ^2						0.114	2.005
p						0.7356	0.1568

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.

Table 3.9 Chi-square values and p values (df = 1) for a pair-wise comparison of Malagasy groups and other world populations at the *Hinf*I/β site.

	BT	AS	TS	MF	VZ	BANTU	SAN
MR[1] χ^2	1.999	0.047	2.501	26.644	5.086	1.776	0.753
p	0.1574	0.8278	0.1138	$< 10^{-4}$	0.0241	0.1827	0.3855
BT[1] χ^2		0.957	0.008	15.308	1.537	5.06	3.476
p		0.3279	0.9292	0.001	0.215	0.0245	0.0622
AS[1] χ^2			1.325	21.978	3.82	0.035	1.135
p			0.2497	$< 10^{-4}$	0.0507	0.8524	0.2866
TS[1] χ^2				13.472	1.157	5.547	3.925
p				0.002	0.282	0.0185	0.0476
MF[1] χ^2					2.637	24.729	22.63
p					0.1044	$< 10^{-4}$	$< 10^{-4}$
VZ[1] χ^2						8.218	6.532
p						0.0041	0.0106

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

These general trends are confirmed in Table 3.10, which shows the χ^2 values and corresponding P values for a pair-wise comparison of the Malagasy when classified into the regional groups and proposed parental population(s). The highlanders show no significant differences to the lowlanders at any of the RFLP sites, but the south-westerners differ from both the highlanders and lowlanders at two sites, *Hinc*II/ε and HinfI/β. The Malagasy groups show greater differences to Indonesians than they do when compared to the Bantu, although for a number of sites, the differences are not significant.

Table 3.10 Chi-square values and p values (df = 1) for a pairwise comparison of the Malagasy regional groups and other world populations at each β-globin RFLP site.

*Hinc*II/ε				
	LOWLAND	SOUTHWEST	BANTU	INDONESIAN
HIGHLAND χ^2	3.813	17.252	13.15	31.332
p	0.0509	$<10^{-4}$	0.0003	$<10^{-4}$
LOWLAND χ^2		7.381	6.758	46.783
p		0.0066	0.0093	$<10^{-4}$
SOUTHWEST χ^2			0.209	61.923
p			0.6479	$<10^{-4}$
*Hind*III/$^{G}\gamma$				
	LOWLAND	SOUTHWEST	BANTU	INDONESIAN
HIGHLAND χ^2	0.004	1.071	1.723	2.46
p	0.9489	0.3007	0.1894	0.1168
LOWLAND χ^2		1.038	1.69	2.284
p		0.3082	0.1936	0.1307
SOUTHWEST χ^2			0.1041	4.616
p			0.7469	0.0317
*Hind*III/$^{A}\gamma$				
	LOWLAND	SOUTHWEST	BANTU	INDONESIAN
HIGHLAND χ^2	1.875	1.666	0.39	12.7
p	0.1709	0.1967	0.5322	0.0004
LOWLAND χ^2		0.099	0.025	7.252
p		0.7535	0.8737	0.0071
SOUTHWEST χ^2			0.001	3.706
p			0.978	0.0542
*Hinc*II/5'ψβ				
	LOWLAND	SOUTHWEST	BANTU	INDONESIAN
HIGHLAND χ^2	0.316	0.815	0.01	2.277
p	0.5743	0.3666	0.9223	0.1313
LOWLAND χ^2		1.674	0.002	0.869
p		0.1958	0.9627	0.3513
SOUTHWEST χ^2			0.392	3.619
p			0.5315	0.0571
*Hinc*II/3'ψβ				
	LOWLAND	SOUTHWEST	BANTU	INDONESIAN
HIGHLAND χ^2	0.105	0.267	15.505	29.382
p	0.7457	0.6055	0.0001	$<10^{-4}$
LOWLAND χ^2		0.047	13.498	30.468
p		0.8279	0.0002	$<10^{-4}$
SOUTHWEST χ^2			9.431	22.514
p			0.0021	$<10^{-4}$

Table 3.10 Chi-square values and p values (df = 1) for a pairwise comparison of the Malagasy regional groups and other world populations at each β–globin RFLP site.

*Ava*II/β			
	LOWLAND	SOUTHWEST	BANTU
HIGHLAND χ^2	0.002	0.064	2.62
p	0.9689	0.7999	0.1055
LOWLAND χ^2		0.036	2.466
p		0.8498	0.163
SOUTHWEST χ^2			1.477
p			0.2243
*Hinf*I/β			
	LOWLAND	SOUTHWEST	BANTU
HIGHLAND χ^2	0.137	23.527	3.478
p	0.7115	$<10^{-4}$	0.6622
LOWLAND χ^2		19.261	4.239
p		$<10^{-4}$	0.0395
SOUTHWEST χ^2			21.941
p			$<10^{-4}$

The data of the individual RFLP sites do not provide very informative comparisons when treated separately since not all RFLP sites are necessarily population specific, and differences between populations may be due to chance rather than being a reflection of the degree of relatedness of the populations. Thus to generate more useful comparisons, the frequencies of the five 5' β-globin RFLP sites in each of the six Malagasy ethnic groups, as well as in the Indonesians (Chen *et al.* 1990) and the Bantu (Ramsay and Jenkins 1987) were considered together (each locus acting independently). Pairwise Nei's standard genetic distances were calculated (although not strictly correct since the five sites are not independent of each other) and a standard neighbour-joining tree was constructed using the programme DISPAN to give a visual representation of the data. (Throughout this chapter, the standard neighbour-joining trees constructed using the programme DISPAN were constructed from Nei's standard genetic distance by the neighbour joining method). From the tree (Figure 3.1), it can be seen that the Merina show the greatest affinity to the Indonesians, (as indicated by the branch lengths), while all the other Malagasy groups appear on the "Bantu branch". Along this branch, the Mahafaly show the greatest affinity to the Bantu, while the Betsileo are the least closely affiliated to the Bantu. This suggests that the Betsileo have a greater genetic input from another population (eg. Indonesians) than the Mahafaly. The other three populations (AS, VZ and TS) are placed intermediate on the Bantu branch, suggesting a cline of genetic admixture of Bantu and other parental populations (eg. Indonesians).

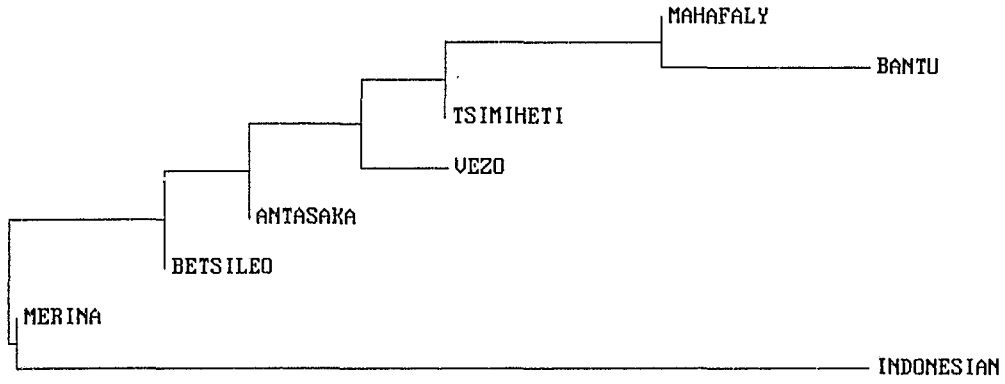


Figure 3.1 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5 5' β-globin RFLP sites in the Malagasy ethnic groups, Indonesians and Bantu.

Similarly, the frequencies of the five 5' β-globin RFLP sites in each of the three Malagasy regional groups, as well as in the Indonesians (Chen *et al.* 1990) and the Bantu (Ramsay and Jenkins 1987) were considered together (each locus acting independently). Pairwise standard genetic distances were calculated and a standard neighbour-joining tree was constructed using the programme DISPAN to give a visual representation of the data. The tree is shown in Figure 3.2. The highland group is the most closely affiliated to the Indonesians, while the south-west group is the most closely associated with the Bantu, with the lowland being the next most closely related Malagasy group to this putative parental population. These results are in precise agreement with all the other evidence of Malagasy ancestry (historical, archaeological and ethnographic) which suggests that the highland populations have the greatest Indonesian input to their ancestry, while the lowlanders are typically more African in origin, and the populations in the south-west of the island show the greatest African influence (reviewed in Dewar and Wright 1993, Campbell 1995).

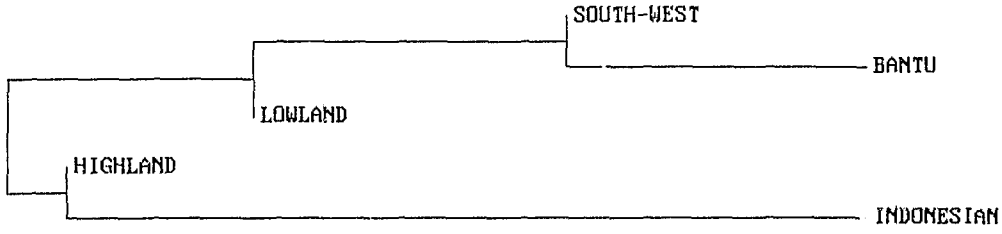


Figure 3.2 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5 5' β-globin RFLP sites in the Malagasy regional groups, Indonesians and Bantu. (HIGHLAND = Merina and Betsileo; LOWLAND = Antasaka and Tsimiheti; SOUTH-WEST = Vezo and Mahafaly).

3.2 UNAMBIGUOUS 5' β-GLOBIN HAPLOTYPES

5' β-globin haplotypes were assigned unambiguously for 248 Malagasy chromosomes, where individuals were homozygous at all five sites or heterozygous at one site only. Ten haplotypes were found and are shown in Table 3.11. Of these, nine have previously been reported in other world populations (haplotypes 1 to 9 in Table 3.11) (Flint *et al.* 1993a), and one (haplotype 10), which is found in both the Tsimiheti and Mahafaly, has not been reported and has thus been called "rare" in this study. From a comparison of the haplotype frequencies of the Malagasy ethnic groups and other world populations (Table 3.11), certain general trends can be seen. (In the discussion that follows, the haplotype numbers refer to those in Table 3.11). The two "African-specific" haplotypes (haplotypes 4 and 5) occur at highest frequencies in the lowland groups, the Antasaka (0.357 and 0.375 respectively) and the Tsimiheti (0.119 and 0.143 respectively), and also at appreciable frequencies in the south-west groups (0.313 in the Vezo, and 0.250 in the Mahafaly). Another haplotype (haplotype 6) which occurs at highest frequencies in Africans (although it is also found in Indonesians), was only found in the Tsimiheti, Mahafaly and Vezo. However, the common Asian haplotype (haplotype 1) occurred at higher frequencies in the Merina and Betsileo (0.515 and 0.540 respectively) than in the lowland and south-west groups. These general trends may be seen more clearly from a graphic representation of the 5' β-globin haplotype frequencies in the histogram (Figure 3.3). The rare haplotype found in this study in two of the Malagasy groups could also simply be the result of new mutation. But it may also suggest that the set of proposed parental population that we are using for comparisons is not complete, or else that the incorrect parent populations have been used.

Table 3.11 5' β–globin haplotype frequencies and standard errors (shown in brackets) in the six Malagasy groups, Indonesians and Sub-Saharan Africans.

NO.	5' HAPLOTYPE[1]	MERINA	BETSILEO	TSIMIHETI	ANTASAKA	VEZO	MAHAFALY	SUB-SAHARAN AFRICAN[2]
		* [66]	* [48]	* [42]	* [56]	* [16]	* [20]	* [60]	* [191]
1	+ - - - -	0.515 (0.061)	0.540 (0.072)	0.381 (0.075)	0.321 (0.062)	0.187 (0.098)	0.100 (0.066)	0.900 (0.038)	0.063 (0.017)
2	- + + - +	0.091 (0.035)	0.042 (0.028)	0.071 (0.040)	0.089 (0.042)	0.000	0.050 (0.048)	0.017 (0.017)	0.073 (0.018)
3	- + - + +	0.015 (0.016)	0.000	0.000	0.000	0.000	0.050 (0.048)	0.033 (0.023)	0.136 (0.025)
4	- - - - +	0.152 (0.044)	0.292 (0.065)	0.357 (0.074)	0.375 (0.065)	0.313 (0.116)	0.250 (0.096)	0.000	0.461 (0.037)
5	- + - - +	0.076 (0.033)	0.021 (0.021)	0.119 (0.049)	0.143 (0.047)	0.063(0.061)	0.250 (0.096)	0.000	0.209 (0.030)
6	- + - - -	0.000	0.000	0.048 (0.033)	0.000	0.063 (0.061)	0.050 (0.048)	0.017 (0.017)	0.021 (0.010)
7	- - - - -	0.060 (0.030)	0.021 (0.021)	0.000	0.000	0.313 (0.116)	0.050 (0.048)	0.000	0.000
8	+ - - - +	0.076 (0.033)	0.083 (0.040)	0.000	0.071 (0.034)	0.063 (0.061)	0.050 (0.048)	0.000	0.021 (0.010)
9	- + + + +	0.015 (0.016)	0.000	0.000	0.000	0.000	0.000	0.000	0.000
10	- - + - +	0.000	0.000	0.024 (0.024)	0.000	0.000	0.150 (0.079)	0.000	0.000

[1]Sites 1 - 5 in Figure 1.4 were used to construct the unambiguous 5' β-globin haplotypes.

[2] Flint *et al.* 1993a.

* Number in brackets is the number of chromosomes studied.

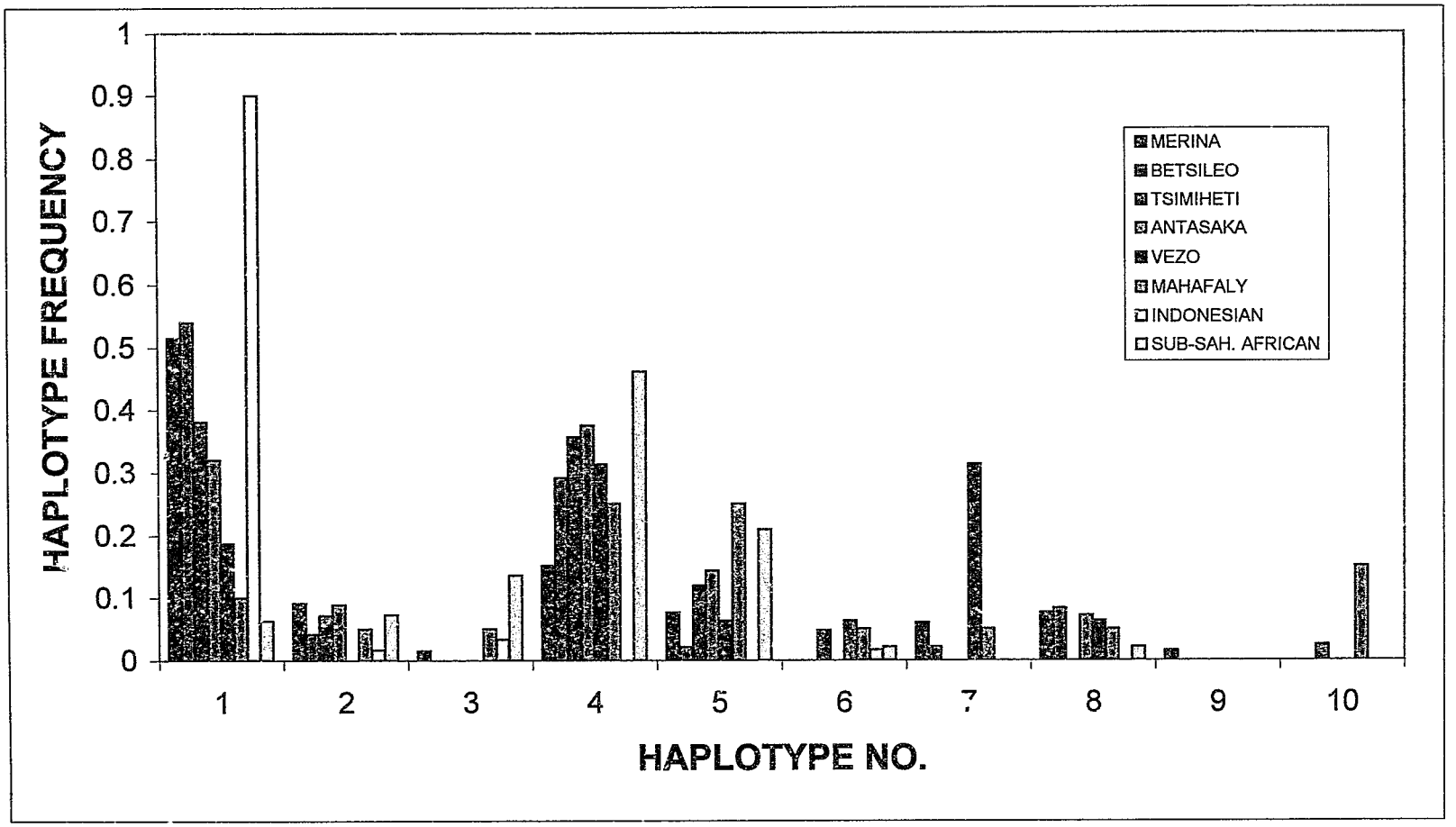


Figure 3.3 Histogram showing the 5' β-globin haplotype frequencies in the Malagasy ethnic groups, Indonesians and Sub-Saharan Africans. (Haplotype no. refers to that in Table 3.11.).

The frequencies of the 5' β-globin haplotypes in the regional Malagasy groups are given in Table 3.12, and are represented graphically in a histogram (Figure 3.4). Once again, it can be seen that the African-specific haplotypes (haplotypes 4 and 5) occur at highest frequencies in the lowland and south-west groups, while the common Asian haplotype (haplotype 1) occurs at the highest frequency in the highlanders. Further, haplotypes 3 and 6, which occur at low but appreciable frequencies in Africa (Flint *et al.* 1993a) occur at the highest frequencies in the south west groups, and haplotype 6 is absent in the highlanders.

The 5' β-globin haplotype frequency data were used in all the further analyses, rather than the individual RFLP site data because of the strength of haplotype analysis in investigating population affinities. Also, a number of different African and Asian populations, all of which are not necessarily putative parental populations of the Malagasy according to previous studies, were included in the analyses that follow. This was done to assess which populations may have contributed to Malagasy ancestry, and to place the Malagasy in a world context.

Table 3.12 5' β-globin haplotype frequencies and standard errors (shown in brackets) in the Malagasy regional groups, Indonesians and Sub-Saharan Africans.

NO.	5' HAPLOTYPE[1]	HIGHLAND	LOWLAND	SOUTH WEST	INDONESIAN[2]	SUB-SAHARAN AFRICAN[2]
		* [114]	*[98]	*[36]	*[60]	*[191]
1	+ - - - -	0.526 (0.047)	0.347 (0.048)	0.139 (0.058)	0.900 (0.038)	0.063 (0.017)
2	- + + - +	0.07 (0.024)	0.082 (0.028)	0.028 (0.027)	0.017 (0.017)	0.073 (0.018)
3	- + - + +	0.009 (0.008)	0.000	0.028 (0.027)	0.033 (0.023)	0.136 (0.025)
4	- - - - +	0.211 (0.038)	0.367 (0.048)	0.278 (0.075)	0.000	0.461 (0.037)
5	- + - - +	0.053 (0.021)	0.133 (0.034)	0.167 (0.062)	0.000	0.209 (0.030)
6	- + - - -	0.000	0.020 (0.014)	0.056 (0.038)	0.017 (0.017)	0.021 (0.010)
7	- - - - -	0.044 (0.020)	0.000	0.167 (0.062)	0.000	0.000
8	+ - - - +	0.079 (0.025)	0.041 (0.020)	0.056 (0.038)	0.000	0.021 (0.010)
9	- + + + +	0.009 (0.008)	0.000	0.000	0.000	0.000
10	- - + - +	0.000	0.010 (0.010)	0.083 (0.047)	0.000	0.000

[1]Sites 1 - 5 in Figure 1.4 were used to construct the unambiguous 5' β–globin haplotypes.

[2] Flint *et al.* 1993a.

* Number in brackets is number of chromosomes studied.

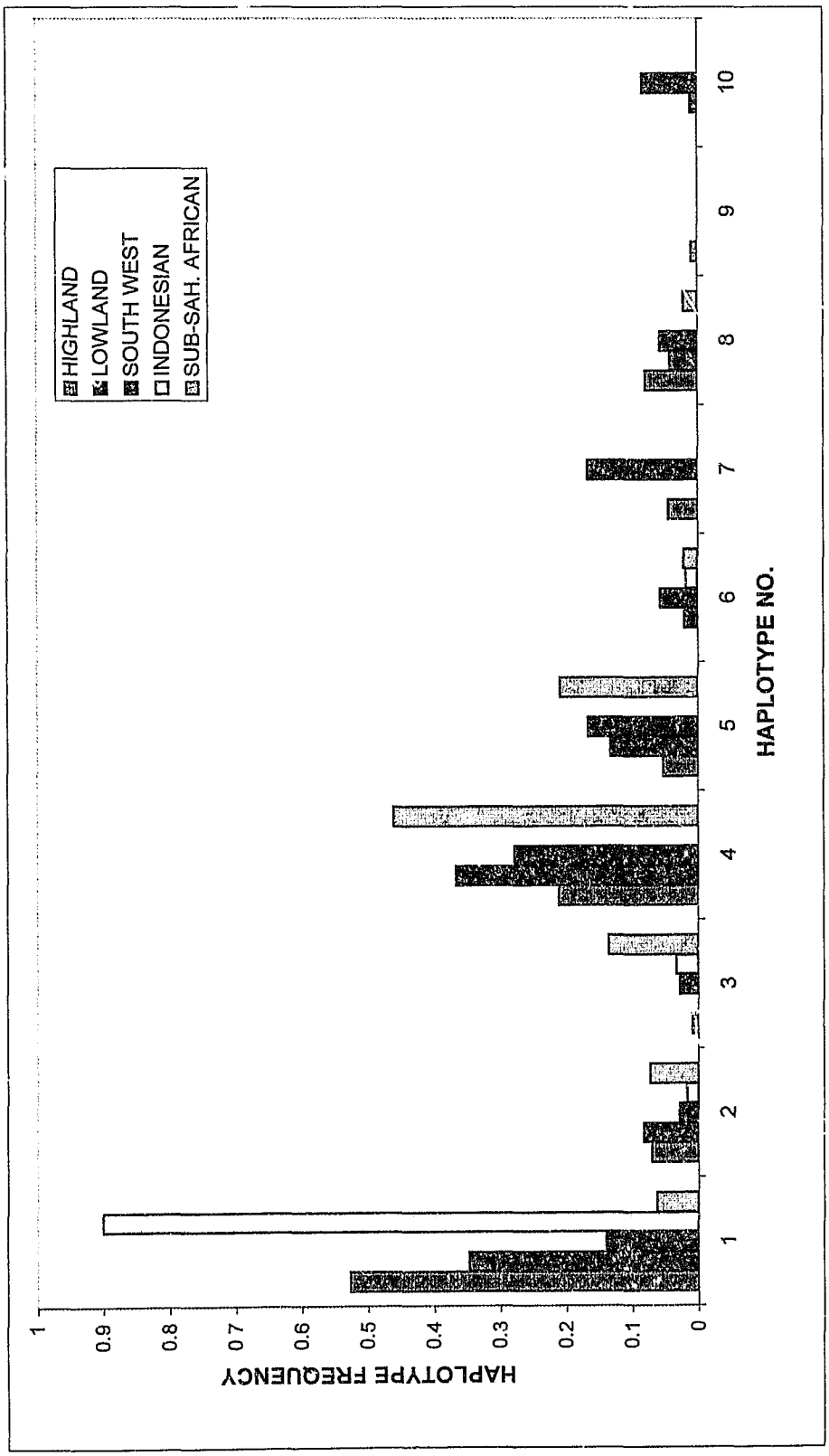


Figure 3.4 Histogram showing the 5' β-globin haplotype frequencies in the Malagasy regional groups, Indonesians and Sub-Saharan Africans. (Haplotype no. refers to that in Table 3.11.).

3.2.1 CHI-SQUARED ANALYSIS OF 5' β-GLOBIN HAPLOTYPE FREQUENCIES.

Chi-square tests were performed using the 5' β-globin haplotype frequencies, to investigate any significant differences within the Malagasy groups, and between the Malagasy and possible parental populations (Table 3.13). The haplotype frequency data for the Indonesian, Nigerian and Algerian populations were obtained from Gueirrerio *et al.* (1994), and those for the Bantu and San, from Ramsay and Jenkins (1987). There are no significant differences between the Malagasy groups. The most noteworthy trends observed when the Malagasy are compared to other populations, are that all six groups differ significantly from Indonesians at the 99% level; only the Merina differ from the Bantu at the 99% level. Both the Merina and Betsileo differ significantly from a Nigerian population. These results are again in accordance with the historical and ethnographic evidence of the highlanders being more Indonesian and less African, and are also in accordance with the β^S (Hewitt *et al.* 1996) and G6PD (Dangerfield 1997) studies, which both suggest that the major African contribution to Malagasy ancestry was made by Bantu-speakers from Sub-Saharan Africa, perhaps in the region of Zambia in particular.

The six groups were then classified into their regions, and chi-square tests were performed (Table 3.14). Under this broader classification, the south-westerners now show a significant difference to the highlanders. Again, all groups differ significantly from the Indonesians, an unexpected finding especially for the highland group. The highland group and Bantu differ significantly from each other, which is in agreement with the historical data which suggests that the highlanders have a greater Indonesian than African contribution to their ancestry than the lowlanders, for whom the opposite is true.

Table 3.13 Chi-square values and p values for a pair-wise comparison of Malagasy and other populations for 5' β-globin haplotypes. Number in brackets is the number of degrees of freedom.

	BT	AS	TS	MF	VZ	INDONESIAN	NIGERIAN	ALGERIAN	BANTU	SAN
MR[1] χ^2	3.47 (7)	10.014 (7)	8.91 (8)	15.421 (9)	11.241 (9)	23.348 (9)	32.089 (8)	19.498 (8)	23.581 (8)	41.372 (7)
p	0.8383	0.1878	0.4456	0.08	0.1884	0.0055	0.0001	0.0124	0.005	$<10^{-4}$
BT[1] χ^2		6.26 (5)	5.505 (7)	16.531 (8)	11.004 (6)	23.614 (7)	23.545 (7)	21.49 (6)	19.972 (6)	42.323 (6)
p		0.2817	0.4809	0.0354	0.0883	0.0027	0.0014	0.0031	0.0104	$<10^{-4}$
AS[1] χ^2			2.478 (6)	8.726 (8)	14.896 (6)	44.426 (6)	13.089 (6)	28.426 (6)	9.595 (5)	25.622 (5)
p			0.8709	0.366	0.0211	$<10^{-4}$	0.0416	0.0001	0.2127	0.0003
TS[1] χ^2				6.22 (8)	11.431 (7)	31.996 (6)	11.497 (5)	23.03 (7)	8.367 (6)	24.501 (6)
p				0.6226	0.1209	$<10^{-4}$	0.0742	0.0017	0.3014	0.0004
MF[1] χ^2					4.826 (8)	32.29 (8)	6.085 (8)	19.849 (8)	3.628 (8)	4.114 (8)
p					0.776	0.0001	0.6377	0.0109	0.8891	0.8467
VZ[1] χ^2						37.309 (6)	10.721 (5)	23.955 (7)	13.097 (6)	18.567 (7)
p						$<10^{-4}$	0.0974	0.0023	0.1086	0.0174

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. $p<0.01$.

Table 3.14 Chi-square values and p values for a comparison of Malagasy regional groups, Bantu and Indonesians for 5' β-globin haplotypes. (No. in brackets is no. of degrees of freedom.)

	LOWLAND	SOUTHWEST	BANTU	INDONESIAN
HIGHLAND[1] χ^2	14.428 (8)	25.396 (9)	30.335 (8)	28.235 (9)
p	0.1079	0.0026	0.0044	0.0009
LOWLAND[1] χ^2		20.055 (8)	13.705 (6)	47.757 (6)
p		0.0101	0.0567	$<10^{-4}$
SOUTHWEST[1] χ^2			9.527 (8)	46.872 (8)
p			0.2998	$<10^{-4}$
BANTU χ^2				59.17 (5)
p				$<10^{-4}$

[1]HIGHLAND = Merina and Betsileo; LOWLAND = Antasaka and Tsimiheti; SOUTHWEST = Vezo and Mahafaly.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. $p<0.01$.

3.2.2 PHYLOGENETIC ANALYSIS

Pairwise genetic distances were calculated from the 5' β-haplotype frequency data and standard neighbour-joining trees were constructed using the programme DISPAN. The data was analysed at four levels:

1. A comparison of the six Malagasy ethnic groups (Figure 3.5).
2. A comparison of the six Malagasy ethnic groups and other world populations (Figures 3.6 - 3.9).
3. A comparison of the three Malagasy regional groups and other world populations (Figure 3.10).
4. A comparison of Madagascar and other world populations (Figure 3.11).

The haplotype frequency data for the "other world populations" was obtained from the review by Hong *et al.* (1997).

From the neighbour joining tree in Figure 3.5, it can be seen that the genetic affinities of the Malagasy ethnic groups are synonymous with the regional grouping: the Merina and Betsileo, two of the highland groups cluster together; two of the south west groups, the Mahafaly and Vezo cluster together; and the Antasaka and Tsimiheti, two of the lowland groups are both positioned intermediate between the first two groups.

In order to determine the affinities of each ethnic group with some of the proposed parental populations, the neighbour joining tree in Figure 3.6 was constructed using DISPAN. The Merina are found closest to the Indonesian/Polynesian cluster of populations and are thus the most closely affiliated to these populations. The Betsileo are found along the same branch, further from this cluster than the Merina, suggesting a stronger genetic input from other populations (eg. African) than that which occurs in the Merina. The Mahafaly are found the furthest from the Indonesian/Polynesian cluster and closest to the Bantu, suggesting that the Mahafaly have the greatest African influence to their ancestry of all the 6 Malagasy groups studied. The Vezo are also closely associated with the Bantu, but less

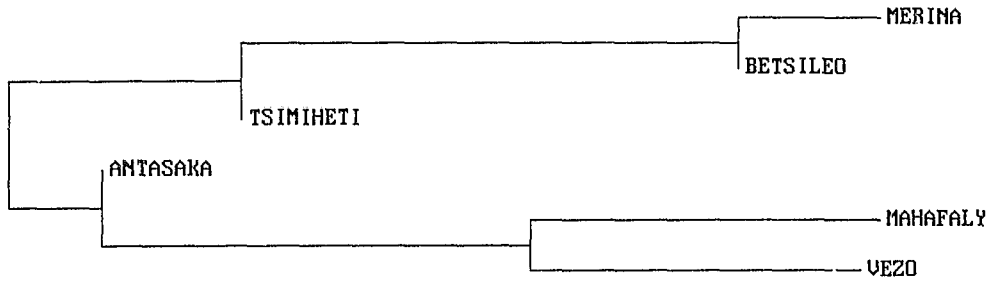


Figure 3.5 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5' β-globin haplotypes in the 6 Malagasy ethnic groups.

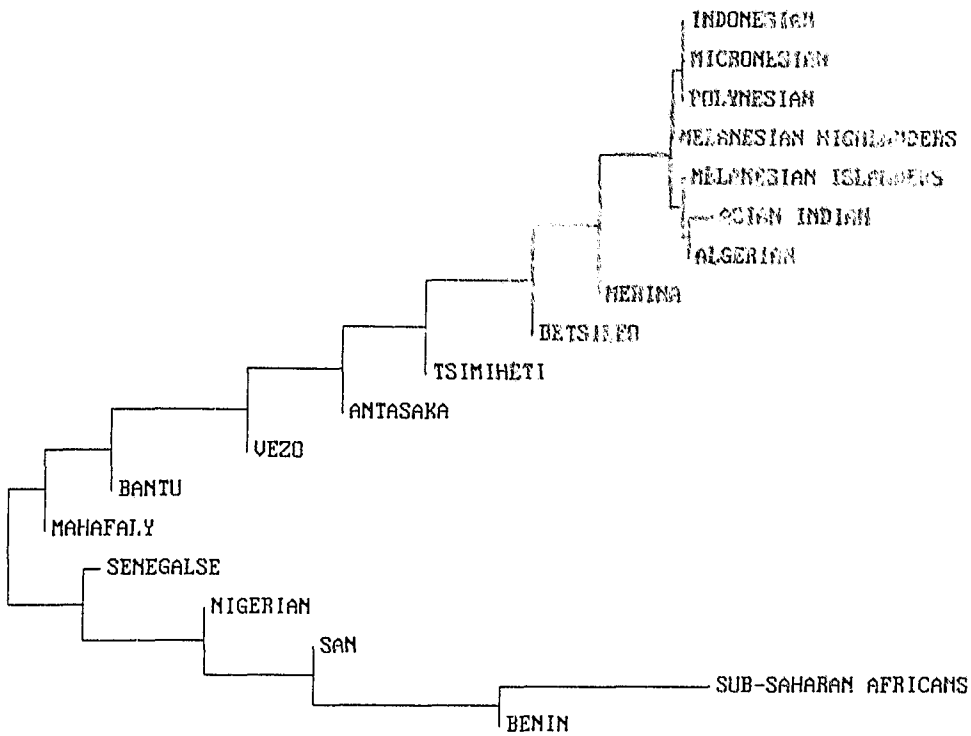


Figure 3.6 Standard neighbour joining tree drawn with DISPAN, based on the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations.

closely so than the Mahafaly, as indicated by the differences in branch lengths between the Mahafaly and Bantu and between the Vezo and Bantu. The Antasaka and Tsimiheti are placed between the Vezo and Betsileo on this same branch, with the Antasaka affiliating more closely with the Bantu and the Tsimiheti affiliating more closely with the Indonesian/Polynesian populations. This tree may be slightly biased since not all of the populations have been analysed at a consistent level: in the larger groups , eg. the Polynesians, there is certainly some sub-structure to the sample (which has not been accounted for in the literature) which may be similar to that of the ethnic grouping in the Malagasy.

In order to assess how different genetic distance formulae affect tree structure, three neighbour joining trees were constructed with the programme PHYLIP, using Nei's standard genetic distance, the Cavalli-Sforza chord measure and Reynold's genetic distance. The trees were redrawn by hand for greater clarity, hence the branch lengths may not be entirely accurate. The neighbour-joining tree obtained using Nei's standard genetic distance (Figure 3.7) is exactly the same as that constructed with DISPAN (Figure 3.5). Figure 3.8 shows the neighbour-joining tree obtained using the Cavalli-Sforza chord measure. In this tree, the Merina are again the most closely affiliated with the Indonesian/Polynesian populations, with the Betsileo occurring next along the branch, a little further from the Indonesian/Polynesian cluster. Also the Mahafaly are again the most closely associated with the African populations, with the Antasaka and Tsimiheti placed intermediate on the same branch. However, the Vezo are found closer to the Indonesian/Polynesian cluster on this tree than on the tree constructed using Nei's genetic distance. This may be due to the relatively small size of the Vezo sample, since the Cavalli-Sforza chord measure is affected by population size. Also, the Bantu occur further towards the end of the African branch than before, and this time the Malagasy groups are more closely affiliated with the Benin population than they are with the Bantu.

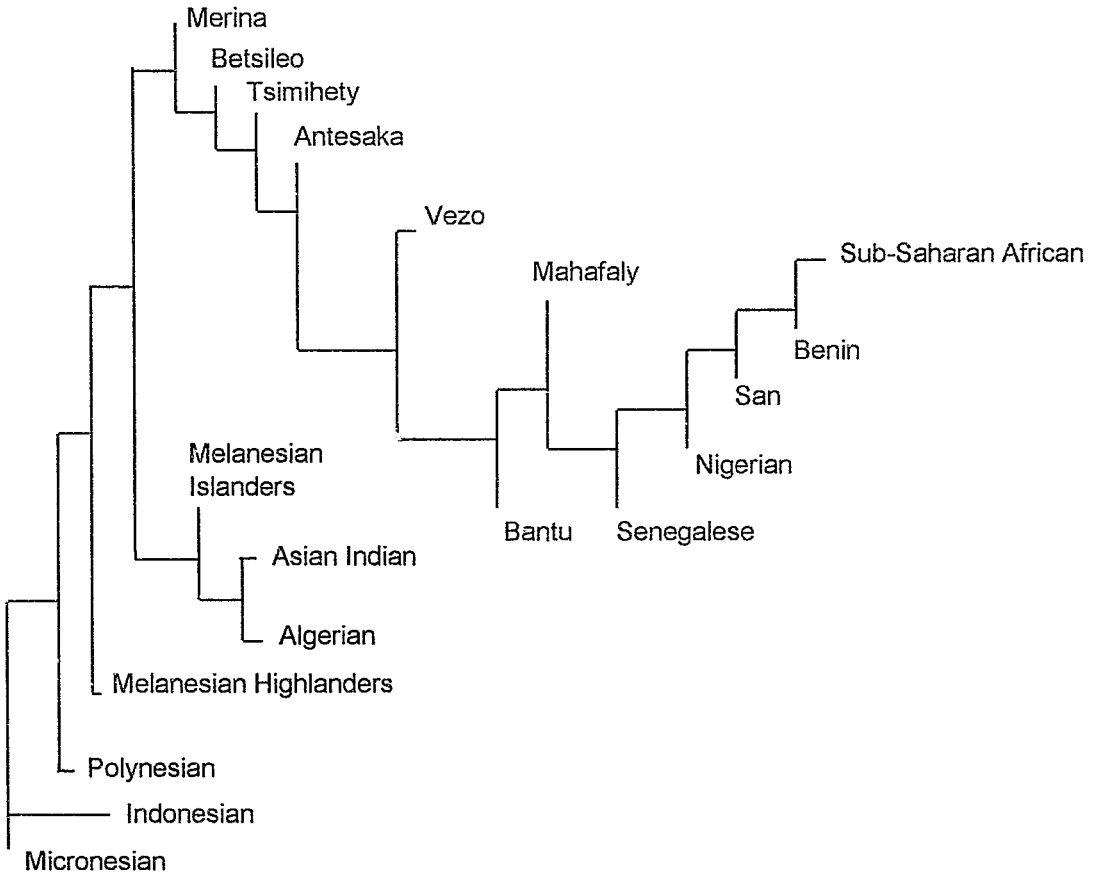


Figure 3.7 Standard neighbour joining tree based the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (reconstructed from PHYLIP, using Nei's standard genetic distance).

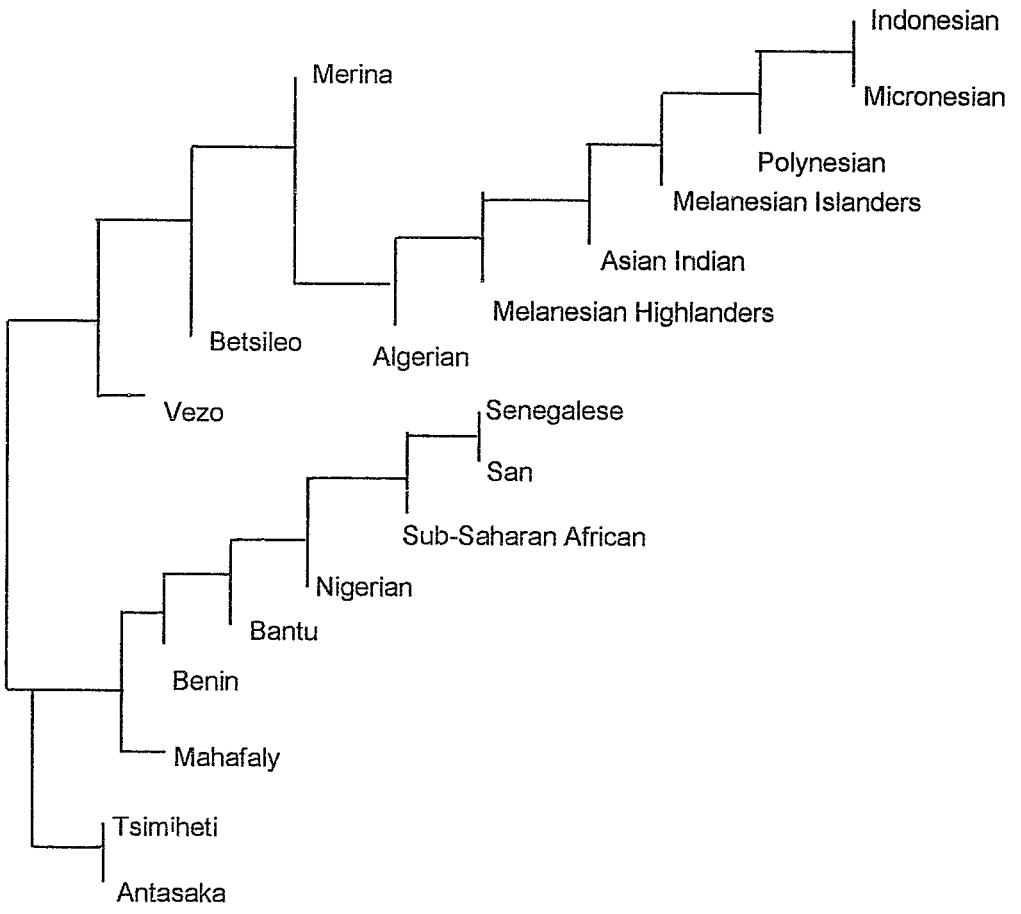


Figure 3.8 Standard neighbour joining tree based the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (reconstructed from PHYLIP, using Cavalli Sforza's Chord measure).

The tree obtained using Reynold's Genetic Distance is shown in Figure 3.9 and is slightly different to the above-mentioned two trees. All the Malagasy groups are found on the main "African" branch of this tree. Again, the Mahafaly are closest to the African populations, with the Vezo next along this branch, followed by the Antasaka, Tsimiheti, Betsileo and Merina respectively. This suggests the same cline of genetic admixture in these groups as before, where the Betsileo and Merina, being the furthest from the African populations but closest to the Asian populations, are assumed to have a greater genetic input from other (ie. Indonesian/Polynesian) populations than the other, more African Malagasy groups. However, this time the Mahafaly are the most closely affiliated with the Senegal population, rather than with the Bantu, and occur on a separate branch of the main African branch to the other Malagasy groups. Once again the differences between this tree and the trees constructed using Nei's Distance and the Cavalli-Sforza measure may be attributed to the effects of sample size.

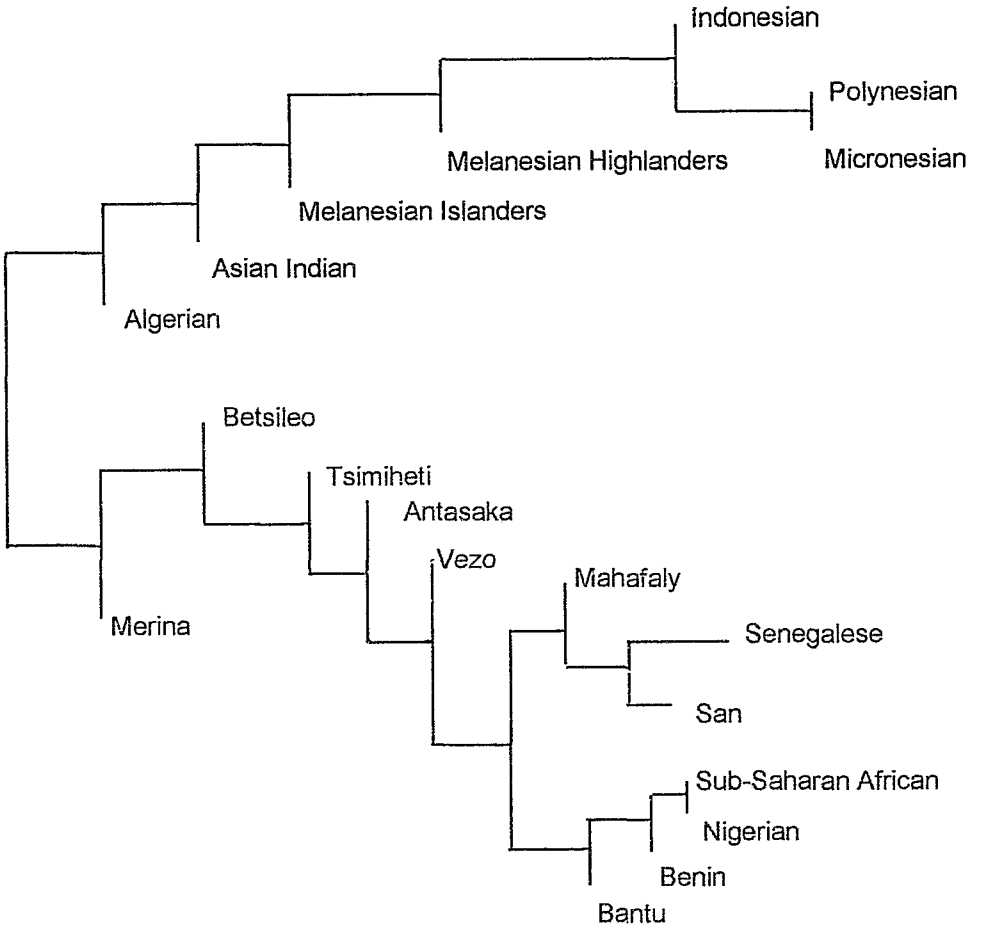


Figure 3.9 Standard neighbour joining tree based the frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (reconstructed from PHYLIP, using Reynold's distance).

For a slightly more consistent comparison of the Malagasy with other world populations, the 5' β-globin haplotype frequencies of the Malagasy regional groups (highland, lowland and south-west) were used to construct a neighbour joining tree, together with the data of the same other world populations as before. Figure 3.10 shows the tree that was obtained for these populations using DISPAN. The clustering and positioning of the populations is in agreement with the historical information of Malagasy ancestry. The highlanders are seen to occur closest to the Polynesian/Indonesian cluster of populations, while the south-west group is the least most closely associated (of the three Malagasy regional groups) with this cluster, but is the most closely affiliated to the Bantu. The lowlanders are placed intermediate between the highland and south-west groups, on the same branch leading to the Polynesian/Indonesian cluster of populations.

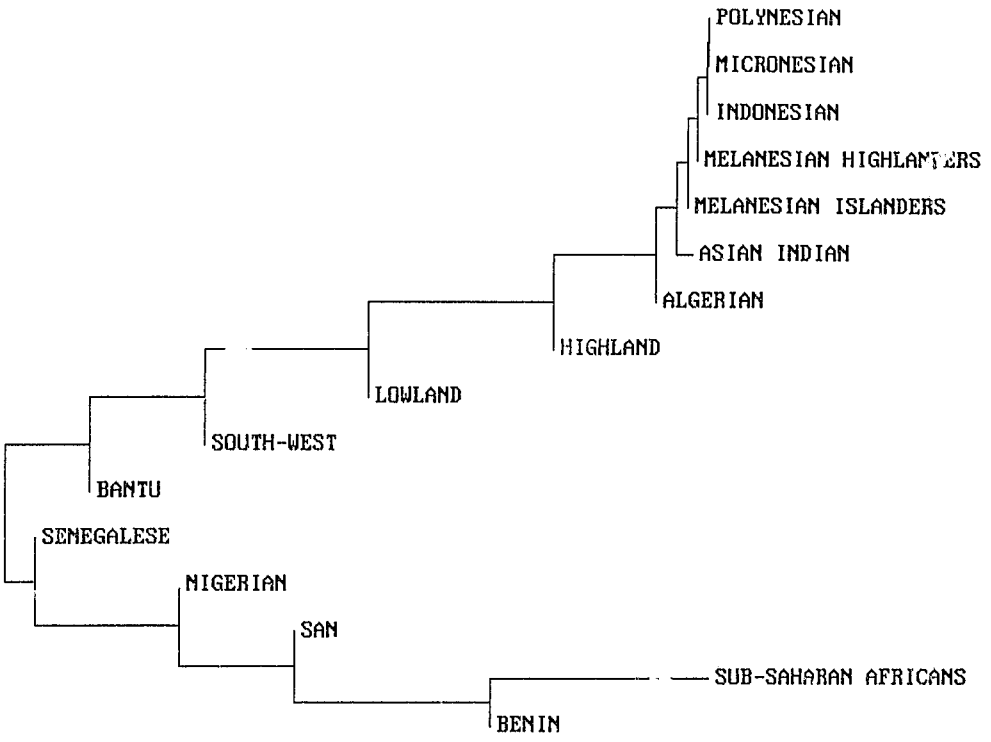


Figure 3.10 Standard neighbour joining tree based on the frequencies of the 5' β-globin haplotypes in the Malagasy regional groups and other world populations (drawn with DISPAN using Nei's genetic distance). (HIGHLAND = Merina and Betsileo; LOWLAND = Antasaka and Tsimiheti; SOUTH-WEST = Vezo and Mahafaly).

At the last level of population sub-structure, the 5' β-globin haplotype frequencies of all six Malagasy ethnic groups were pooled into a single "Madagascar" sample, and this data was used with the other world population data to generate a neighbour joining tree (using DISPAN). This tree is shown in Figure 3.11, and once again the clustering of populations is consistent with all the other evidence of Malagasy ancestry. As expected, the Malagasy occur on a branch intermediate in position between the two putative source populations, the Polynesian/Indonesian cluster of populations and the African populations, in particular, being the most closely affiliated to the Bantu of all the African populations.

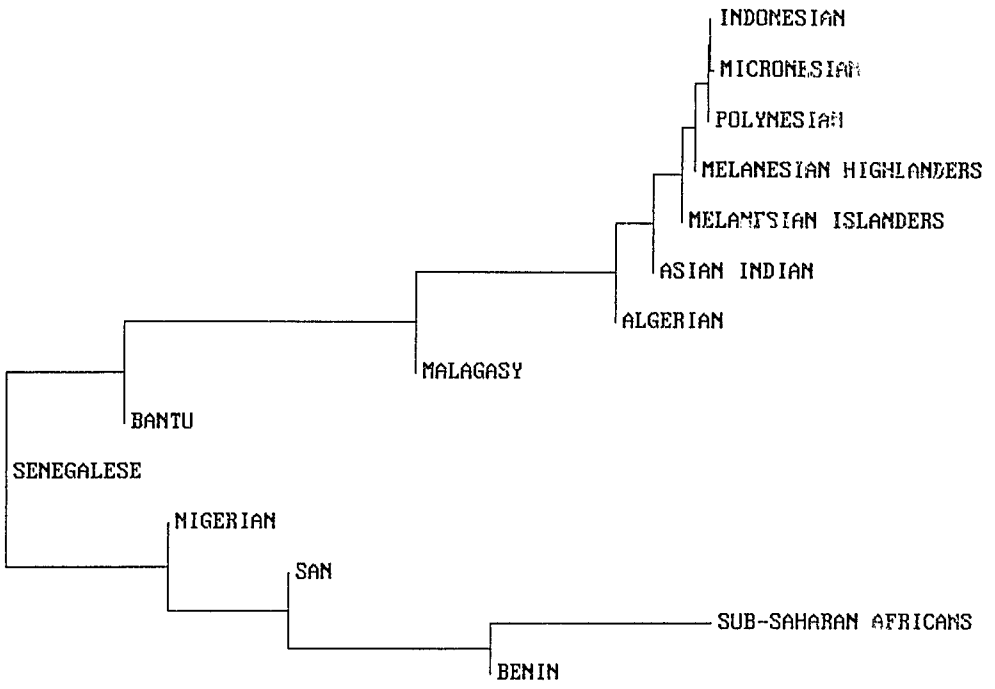


Figure 3.11 Standard neighbour joining tree based on the frequencies of the 5' β-globin haplotypes in the Malagasy and other world populations (drawn with DISPAN using Nei's genetic distance).

3.2.3 PRINCIPLE COMPONENT ANALYSIS OF 5' β-GLOBIN HAPLOTYPES.

In this study, the Principle Component analysis, as described in Section 2.8.3., provides a two-dimensional representation of the set of 5' β-globin haplotype data, and is used to complement the information contained in phylogenetic trees, which provide a one-dimensional interpretation of the data. The analysis was carried out using the programme ANTANA, version 1.5 (written by Harpending and Rogers, Pennsylvania State College, 1985), a programme for Multivariate Data Analysis.

The data was analysed at the same four levels as in Section 3.2.2:

1. A comparison of the six Malagasy ethnic groups
2. A comparison of the six Malagasy ethnic groups and other world populations.
3. A comparison of the three Malagasy regional groups and other world populations.
4. A comparison of Madagascar and other world populations.

The data matrix used to determine the principle components in each case was a matrix of the frequencies of the ten haplotypes (given in Table 3.11) of the populations concerned. For each population, the sum of the haplotype frequencies had to equal 1, and for levels of analysis 2,3 and 4 listed above, this was corrected for in the other world populations since some haplotypes occur in these populations that were not found in the Malagasy. This was done by assigning an 11th "haplotype", which was calculated as: 1-(the sum of the frequencies of the ten haplotypes). Singular values of the principle components were computed, and these were converted to eigenvalues by squaring the singular value. The significance of each principle component (11 in each case, corresponding to the 11 haplotypes of the data matrix) was determined by calculating the ratio of each eigenvalue to the total sum of the eigenvalues. The two principle components that retained the most information were used to generate a two-dimensional PC map. In each case, the first two principle components retained the most information, and the singular values corresponding to only these two PC's are shown in Sections 3.2.3.1. to 3.2.3.4. below. The clustering of the populations in the plane provided an indication of their

affinities. Since the first PC (the x-axis) retained more information than the second PC (the y-axis) in each case, the horizontal displacement between the populations was more significant than the vertical displacement, in identifying the genetic relatedness of the populations.

3.2.3.1 PC ANALYSIS OF THE SIX MALAGASY ETHNIC GROUPS.

The PC map for the six Malagasy ethnic groups studied is shown in Figure 3.12. The first PC expresses 45% of the total information (variance) and the second 31%; the total for the two-dimensional representation is thus 76%.

The trends seen are similar to those identified from the phylogenetic analysis. The Merina and Betsileo cluster relatively tightly on the negative x-axis, while the Tsimiheti and Antasaka form another, separate tight cluster. The Mahafaly and Vezo seem widely scattered, but considering their horizontal displacement from the other groups, the Vezo are relatively closely associated with the two lowland groups, with the Mahafaly less closely so. But, taking the vertical displacement of the groups into account, the Mahafaly are closely related to the lowlanders, more so than they are to the highlanders. Overall, the clustering suggests the same relationships between the Malagasy groups, with the highlanders (Merina and Betsileo) and south-westerners (Vezo and Mahafaly) appearing at the two ends of the scale and the lowlanders (Antasaka and Tsimiheti) placed in between the two.

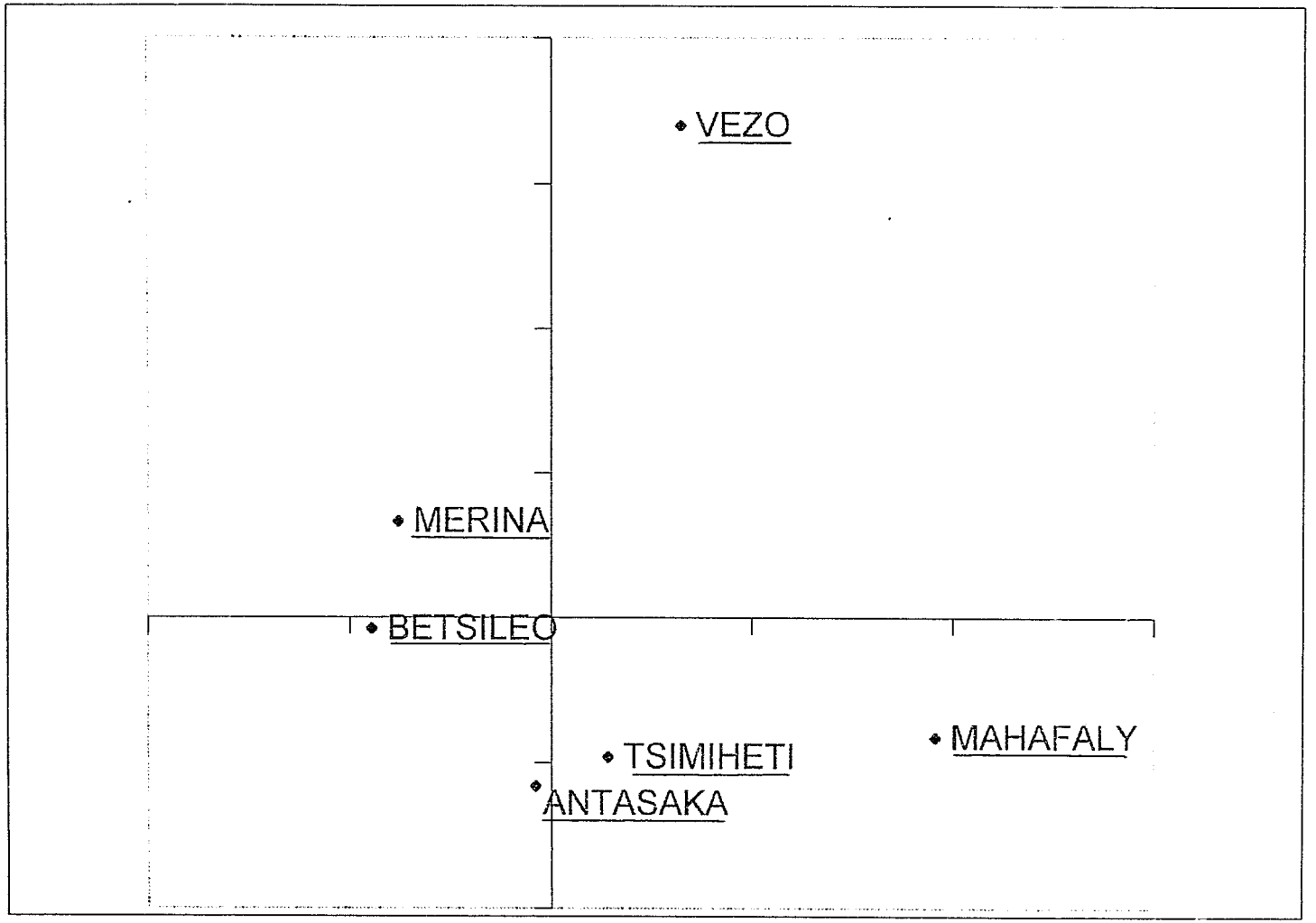


Figure 3.12 PC map of 5' β-globin haplotype frequencies in the 6 Malagasy ethnic groups.

3.2.3.2 PC ANALYSIS OF THE SIX MALAGASY ETHNIC GROUPS AND OTHER WORLD POPULATIONS.

The PC map for the six Malagasy ethnic groups and the other world populations is shown in Figure 3.13. The first PC expresses 55% of the total information (variance) and the second 16%; the total for the two-dimensional representation is thus 71%.

All six Malagasy ethnic groups form a cluster in the bottom left quadrant, and are closely associated with a cluster of African populations. The Polynesians/Indonesians again form a separate cluster in the upper and lower right-hand quadrants. Once again the Malagasy appear intermediate between the African and Polynesian/Indonesian clusters, but positioned closer to the Africans. The Merina and Betsileo are the closest of the Malagasy groups to the Polynesian/Indonesian cluster, while the Mahafaly and Antasaka are the closest to the African populations

3.2.3.3 PC ANALYSIS OF THE THREE MALAGASY REGIONAL GROUPS AND OTHER WORLD POPULATIONS.

The PC map for the regional Malagasy groups and the other world populations is shown in Figure 3.14. The first PC expresses 60% of the total information (variance) and the second 17%; the total for the two-dimensional representation is thus 77%.

Once again, populations from the same continents are seen to cluster, with the Malagasy regional groups also forming a cluster in the bottom left-hand quadrant. They are closer to the African cluster than they are to the Asian cluster. But the highlanders tend more to the Polynesian/Indonesian cluster than the other two regional groups, who tend more towards the African cluster.

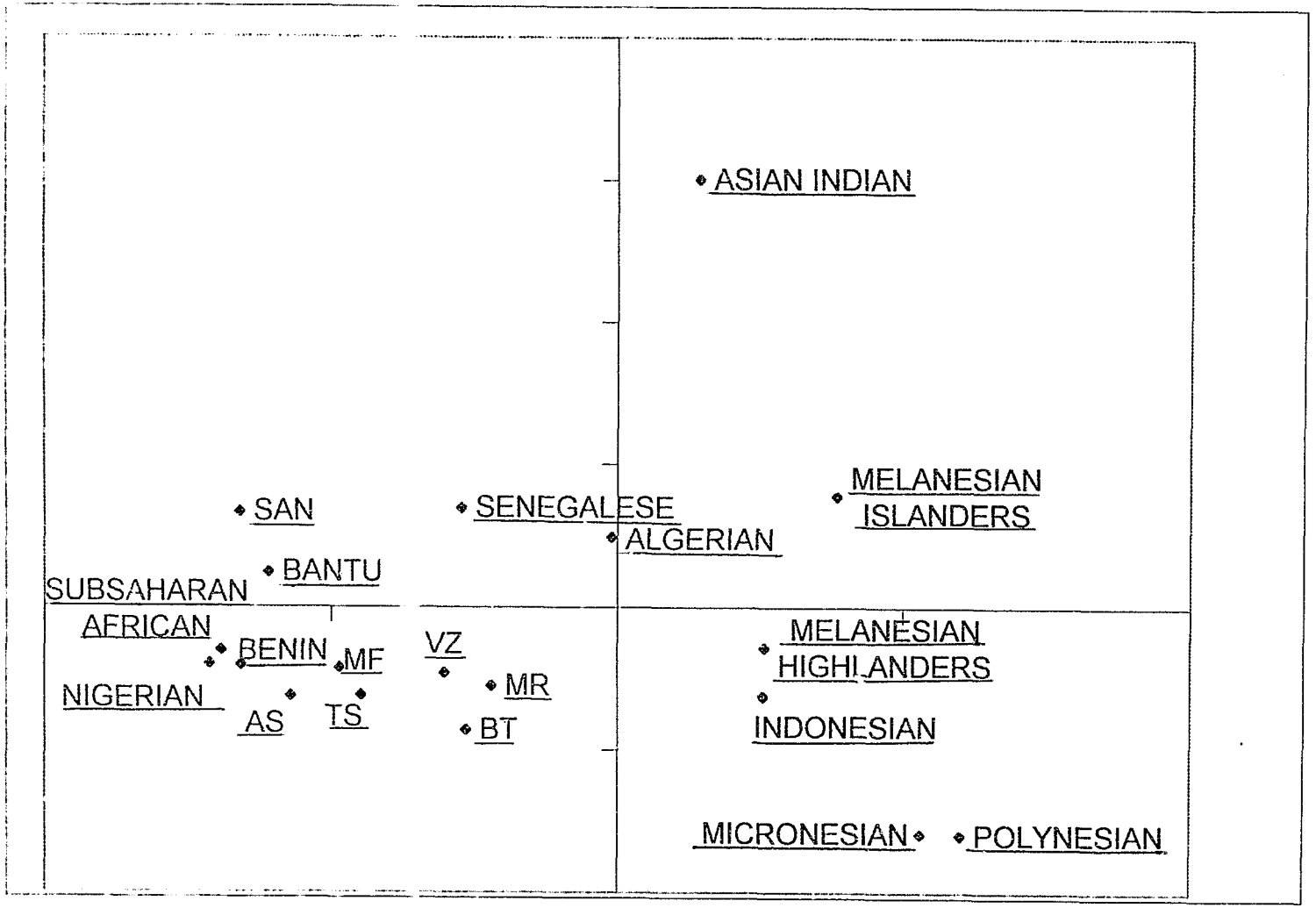


Figure 3.13 PC map of 5' β-globin haplotype frequencies in the Malagasy ethnic groups and other world populations. MR=Merina; BT=Betsileo; AS=Antasaka; TS=Tsimiheti; MF=Mahafaly; VZ=Vezo.

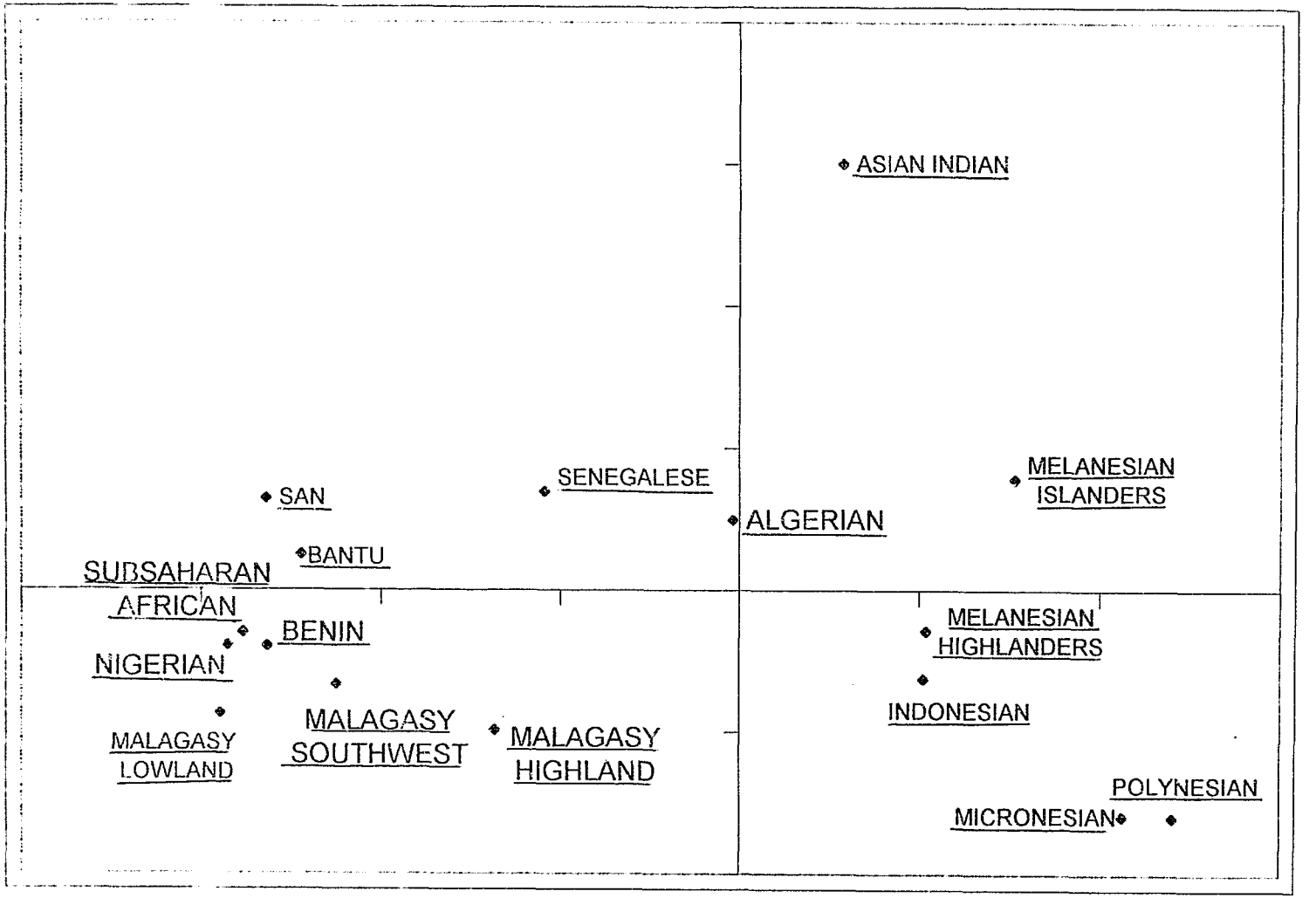


Figure 3.14 PC map of 5' β-globin haplotype frquencies in the Malagasy regional groups and other world populations.

3.2.3.4 PC ANALYSIS OF MALAGASY AND OTHER WORLD POPULATIONS

The PC map for the total Malagasy population and the other world populations is shown in Figure 3.15. The first PC expresses 65% of the total information (variance) and the second 19%; the total for the two-dimensional representation is thus 84%. This map thus provides the most accurate representation of the population affinities of the four PC maps, having the highest value of expression of total variance.

A similar set of trends is seen in this PC map, as in Figures 3.13 and 3.14, with different continents forming separate clusters at different positions on the map. The position of Madagascar on the map is not quite as expected, considering the fact that all of the populations have been analysed at the same level of substructure. One would expect to find the Madagascan population intermediate in position between the two clusters of the putative source populations, at least with respect to the first PC (the horizontal displacement of the populations) even though they are intermediate between the African cluster and Indonesians along the second PC.

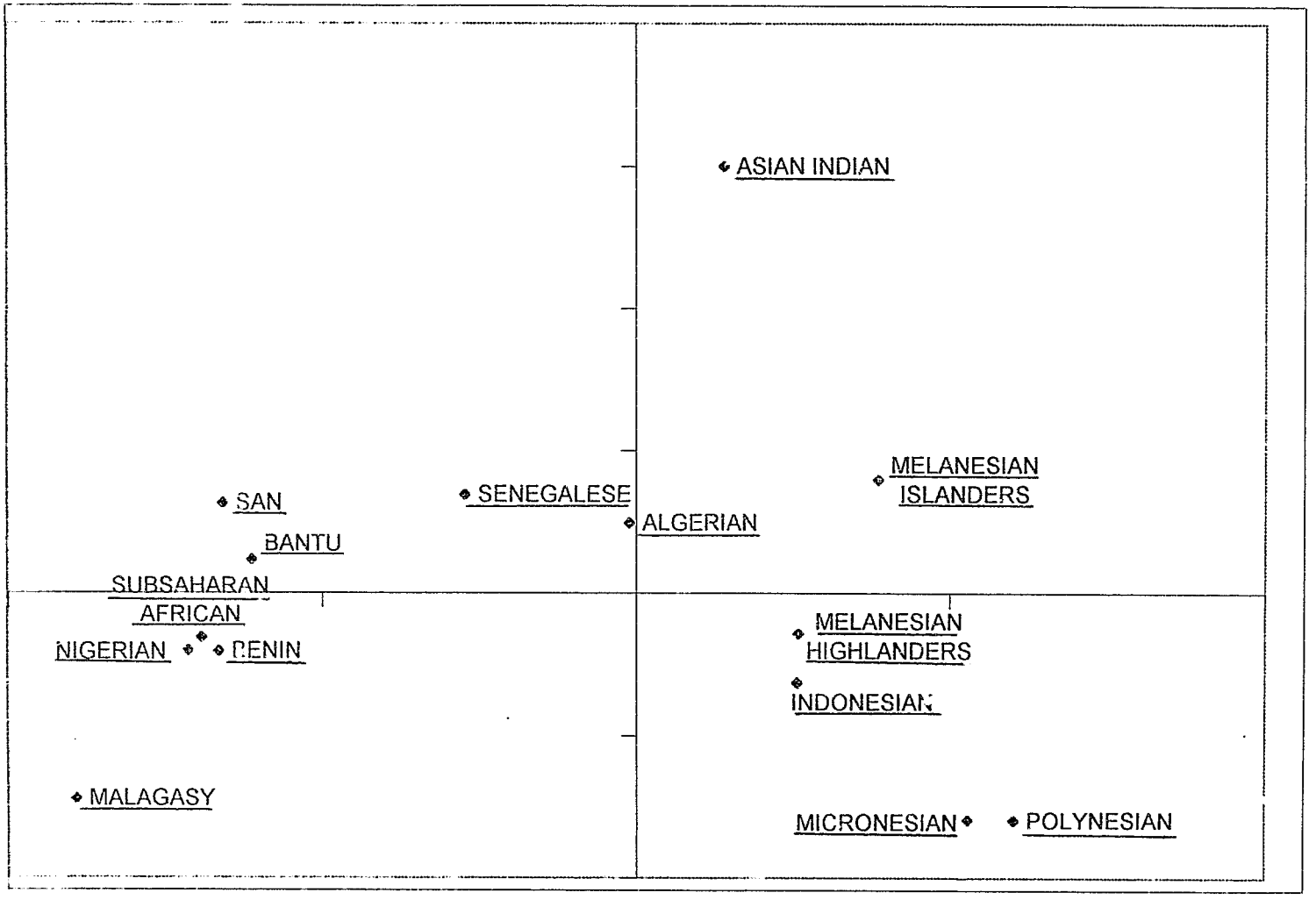


Figure 3.15 PC map of 5' β-globin haplotype frequencies in the Malagasy and other world populations.

3.2.4 AVERAGE HETEROZYGOSITY AND DISTANCE FROM THE CENTROID.

To determine the relative amounts of gene flow experienced by the populations included in this study, the heterozygosity of each group was compared to the genetic distance from the centroid for each group, according to the Harpending and Ward model (1982), described in Section 2.8.4. The 5' β-globin haplotype data was analysed at the same four levels as in previous sections:

1. A comparison of the six Malagasy ethnic groups
2. A comparison of the six Malagasy ethnic groups and other world populations.
3. A comparison of the three Malagasy regional groups and other world populations.
4. A comparison of Madagascar and other world populations.

At each of these levels, the heterozygosity for each population was calculated from their corresponding 5' β-globin haplotype frequencies (according to the formula shown in Section 2.8.4.), Then the average heterozygosity of all the populations involved was calculated (by obtaining the mean of the individual heterozygosities of all the populations). The distance from the centroid was calculated for each population as described in Section 2.8.4. These points were plotted against each other on a Cartesian plane. The values of expected heterozygosity for each population were plotted against the corresponding values of distance from the centroid. In each case, the expected linear relationship was obtained, shown as the "theoretical regression" line on each scattergram (Figures 3.15 to 3.18).

3.2.4.1 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE SIX MALAGASY ETHNIC GROUPS.

Figure 3.16 shows a scattergram of heterozygosity versus distance for the six Malagasy ethnic groups included in this study. The predicted regression line is shown on the figure. There is a clustering around the theoretical regression of the two highland and the two lowland populations, although the Betsileo appear to be a low outlier. This suggests that the Betsileo are the least genetically diverse, with the least amount of gene flow having taken place, and may suggest that they have been relatively genetically isolated. However, the Vezo and Mahafaly are the most heterozygous, and are shown to be high outliers to the model, indicating that they are more admixed than the other Malagasy groups. This is in accordance with the history of the Malagasy (reviewed in Dewar and Wright, 1993), which shows that the populations in the south-west of Madagascar are descendants in varying proportions of several major African and Asian populations, whereas the highland groups are more homogeneous, claiming to be of Indonesian ancestry, with relatively little African admixture.

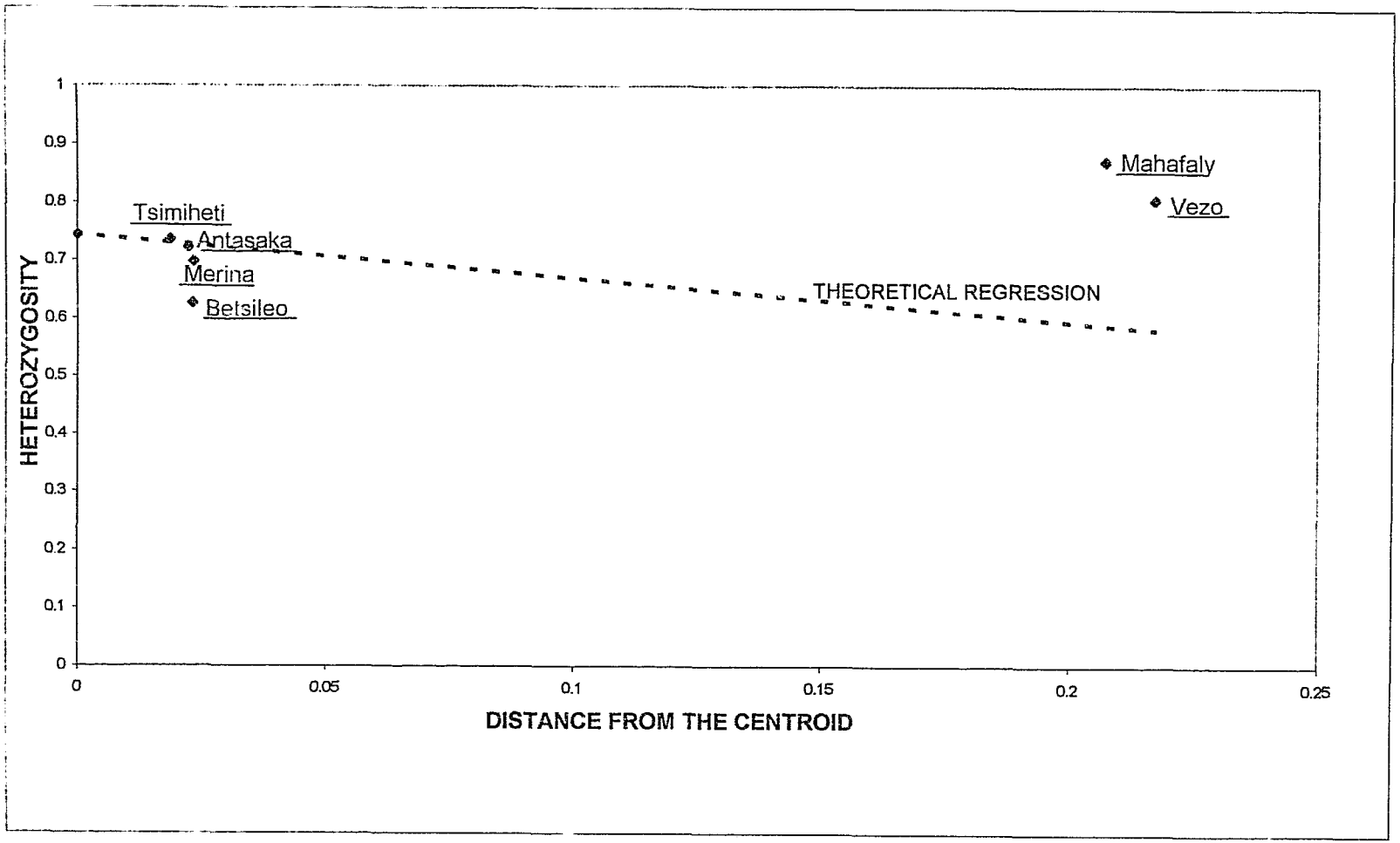


Figure 3.16 Heterozygosity vs distance from the centroid for the 6 Malagasy ethnic groups.

3.2.4.2 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE SIX MALAGASY ETHNIC GROUPS AND OTHER WORLD POPULATIONS

A scattergram of heterozygosity versus distance for the six Malagasy ethnic groups included in this study and other world populations is shown in Figure 3.17. The Malagasy groups have similar heterozygosities to the African populations, which are higher than those of the Asian populations. The highland (Merina and Betsileo) and lowland (Antasaka and Tsimiheti) groups cluster together, with the Antasaka and Tsimiheti being higher outliers, and thus having higher heterozygosities than expected as compared to the Merina and Betsileo. The Vezo and Mahafaly are also high outliers, but lie far from the theoretical regression line and have much higher heterozygosities than expected. This indicates that they are more genetically admixed than the other Malagasy groups, and have experienced a higher rate of gene flow. (This may be slightly exaggerated due to the relatively small sample sizes of the Vezo and Mahafaly). The Asian populations are low outliers to the model, in agreement with what is known about non-African populations having relatively little genetic variation. It is well-known that African populations are the most genetically diverse populations, and this greater diversity has always been thought to reflect a greater antiquity of human populations in Africa. However, most recently an analysis of heterozygosity versus distance from the centroid for a number of African and non-African populations revealed that African populations probably have had a larger effective population size than non-African populations, and this may account for the larger heterozygosity of African populations (Stoneking *et al.* 1997).

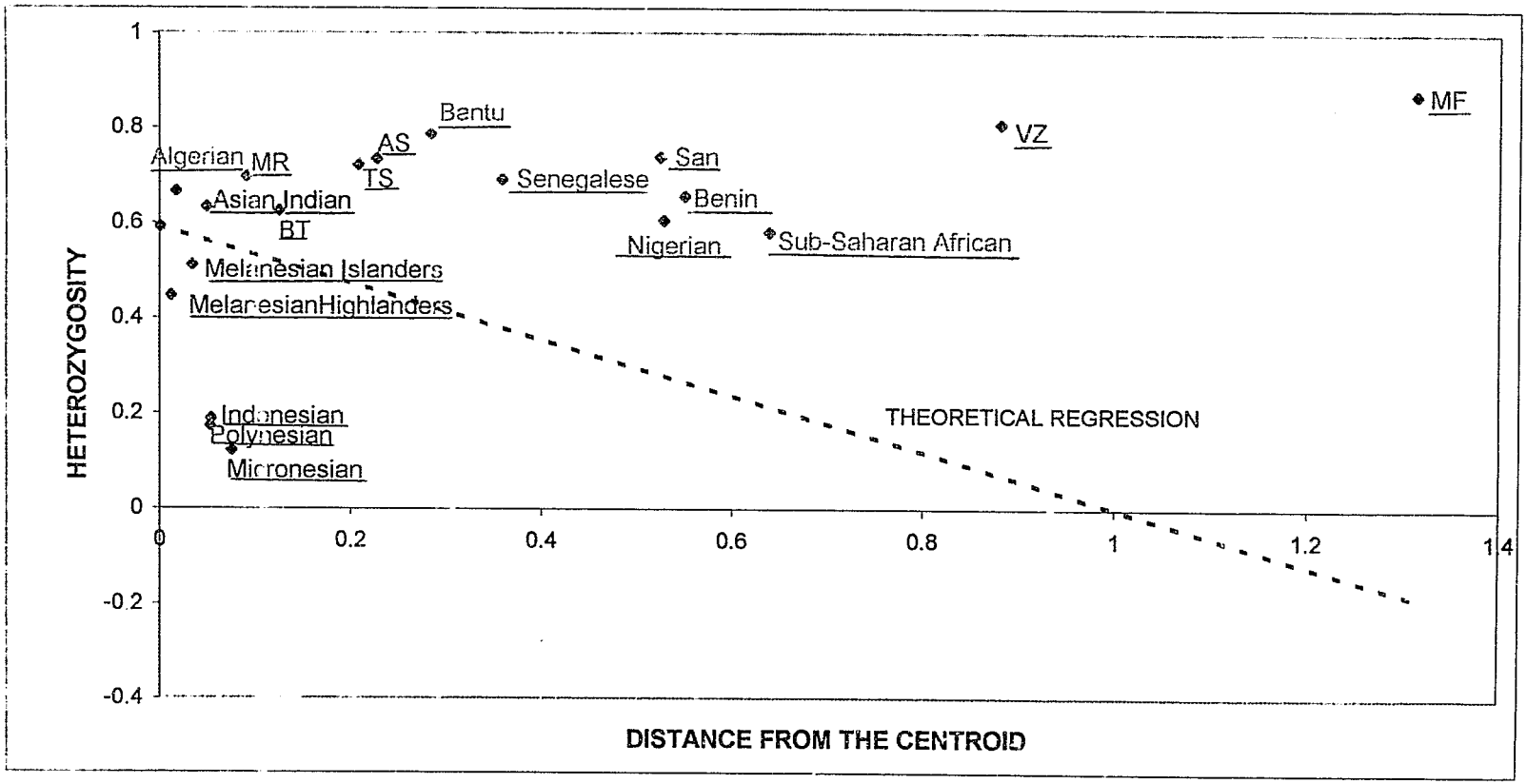


Figure 3.17 Heterozygosity vs distance from the centroid for the Malagasy ethnic groups and other populations. MR=Merina; BT=Betsileo; AS=Antasaka;TS=Tsimiheti; MF=Mahafaly; VZ=Vezo.

3.2.4.3 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE MALAGASY REGIONAL GROUPS AND OTHER WORLD POPULATIONS

In order to correct for the bias in sampling of the individual Malagasy ethnic groups compared to that of the other larger populations, the Malagasy were grouped into the larger regional samples. A scattergram of heterozygosity versus distance from the centroid for the regional Malagasy ethnic groups and other world populations is shown in Figure 3.18. Once again, all the African populations and the Malagasy are high outliers. In particular, the south-west groups are much more heterozygous than expected, with the lowland groups somewhat more heterozygous than expected, and the highland group closest to the theoretical line. This pattern is what would be expected for a hybrid population such as the Malagasy. Moreover, this is also in accordance with the suggestion that the south-west populations have received more gene flow and/or gene flow from more populations than the lowland populations, who are in turn more admixed than the relatively homogeneous highland populations. The Asian populations again fall below the theoretical line, in accordance with their lower levels of heterozygosity, compared to other world (particularly African) populations, as discussed above in Section 3.2 4.2.

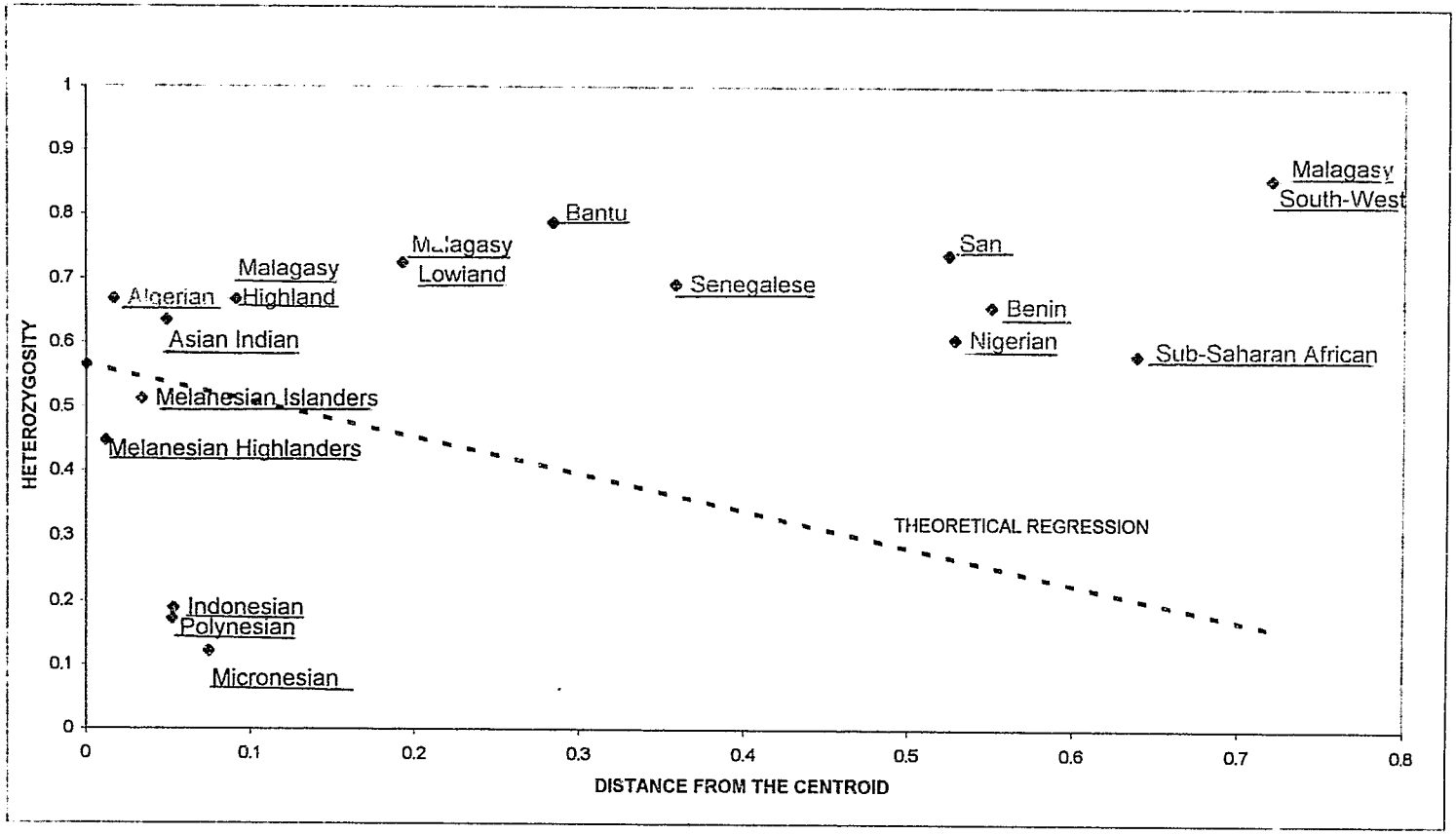


Figure 3.18 Heterozygosity vs distance from the centroid for the regional Malagasy groups and other world populations. HIGHLAND=Merina and Betsileo; LOWLAND=Antasaka and Tsimiheti; SOUTHWEST=Vezo and Mahafaly.

3.2.4.4 HETEROZYGOSITY VERSUS DISTANCE FROM THE CENTROID FOR THE MALAGASY AND OTHER WORLD POPULATIONS

To standardise the sampling even further, the 5' β-globin haplotype data for all 6 Malagasy ethnic groups were pooled into one "Malagasy" sample, and this was compare to the data for the other world populations. A scattergram of heterozygosity versus distance from the centroid for these populations is shown in Figure 3.19. The Malagasy population is a high outlier in this model, as well as the African populations, but it is closer to the theoretical line than the Bantu population is. However the reasons for which each of these two populations show higher heterozygosities than expected may be different. In the case of the Bantu, this may be a reflection of an older and/or larger population. For the Malagasy however, this argument does not hold, but rather, the deviation from the expected may be a reflection of the large amount of admixture from a number of populations, and may be a reflection of the smaller size of the Malagasy population.

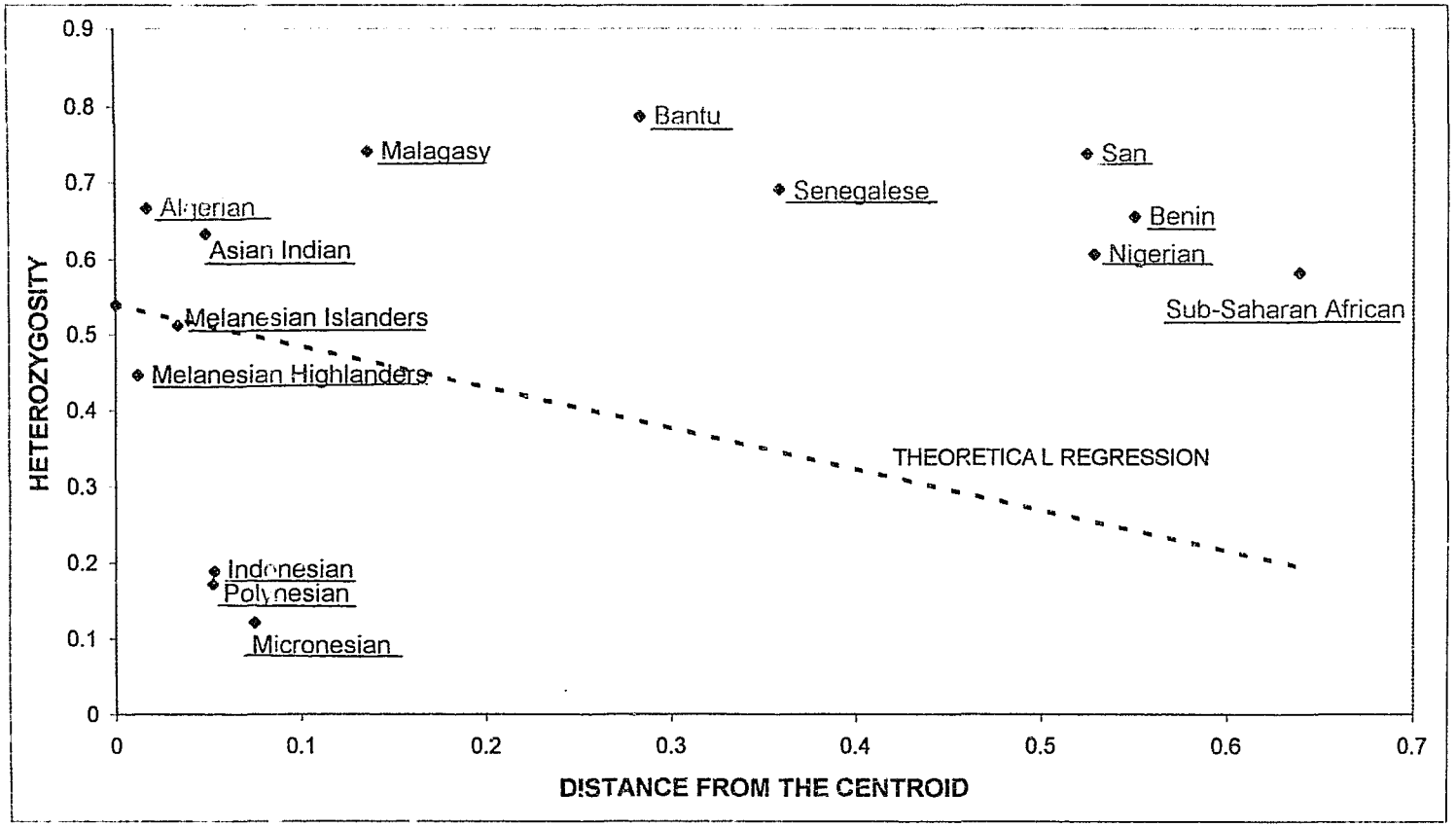


Figure 3.19 Heterozygosity vs distance from the centroid for the Malagasy and other world populations.

3.2.5 ESTIMATION OF ADMIXTURE IN THE MALAGASY

The proportional contributions of three putative parental populations, West Africans, Bantu and Indonesians, to the Malagasy gene pool were calculated according to Elston's least squares method of genetic admixture, described in Section 2.8.5. In the calculations, the haplotype frequency data for the Indonesian, and West African populations were obtained from Flint *et al.* (1993a), and those for the Bantu were from Ramsay and Jenkins (1987). The results are shown in Table 3.15.

In admixture studies, proper identification of ancestral populations is an important issue. In this study, we were limited by the amount of data that has been published in determining the ancestral populations. Thus, although the West African population is not necessarily a true putative parental population of the Malagasy, it is one of the few other African populations for which data has been published (Flint *et al.* 1993a), and it was necessary to include it in this analaysis to obtain some measure of the genetic admixture of the Malagasy. In the admixture estimates calculation, if an m value (proportional contribution) of a parental population is negative, it is possible that a parental population is inappropriate or that the alleles included in the study may be affected by natural selection (Chakraborty 1986). This was the case when the Polynesians were included in this admixture calculation, which may suggest that they are an inappropriate parental population of the Malagasy. When the Polynesian were excluded from the calculation, more reliable estimates were obtained, which are more in keeping with the trends of population affinities that have been generated thus far.

The Indonesian contribution is highest in the highlanders (53%), then next highest in the lowlanders (31%) and lowest in the south west groups (12%). The Bantu contribution is highest in the south-west group (65%), while the highlanders have a greater Bantu contribution (36%) than the lowlanders (26%). This is surprising considering all other evidence of Malagasy ancestry, in particular the β^s study (Hewitt *et al.* 1996), which showed that the major African contribution to Malagasy **Table**

Table 3.15 Admixture estimates (m) of the putative parental populations to the hybrid Malagasy regional groups and the total Malagasy population.

	Highland	Lowland	South-west	Total Malagasy
West African	11.3%	43.2%	22.9%	25.1%
Bantu	35.7%	25.5%	64.5%	36.3%
Indonesian	53.0%	31.2%	12.6%	38.5%

ancestry was from Bantu-speaking Negroids, and that the lowlanders are the most typically African of all the Malagasy groups. This could be due to the fact that the incorrect African parental populations were used in the analysis. However, when the total African contribution to each regional Malagasy group is taken into account, it is shown to be higher in the lowlanders (69%) than in the highlanders (47%). Overall, the African contribution to Malagasy ancestry is 61% and the Indonesian contribution is 39%. These estimates may be slightly biased, since the ethnic groups that were included in this study comprised two highland groups but four lowland groups (the VZ and MF are geographically lowland groups as well).

The admixed nature of the Malagasy was also apparent from the plot of heterozygosity versus distance from the centroid (Figure 3.19) and is again consistent with the existing evidence for both African and Indonesian contributions to Malagasy ancestry. Firstly, populations from the same continent tend to cluster and the admixed population (the Malagasy) tends to be intermediate in position between the two source populations (ie. African and Asian), and secondly, the Malagasy had higher heterozygosities than expected, indicative of the greater level of gene flow experienced by this group.

3.3 MAXIMUM-LIKELIHOOD ESTIMATES OF 5' β-GLOBIN HAPLOTYPES.

The analysis up to this point in the study has been based on unambiguous 5' β-globin haplotypes, and thus has ignored the rare haplotypes and those haplotypes that could not be assigned on the basis of homozygosity/heterozygosity at one RFLP site only. To investigate the possible loss of information from otherwise uninformative individuals, maximum likelihood estimates of 5' β-globin haplotypes in the Malagasy were generated. Using the maximum-likelihood (EM-HAPLO) programme, 26 5' β-globin haplotypes were observed in the Malagasy. The haplotype frequencies estimated by the algorithm are shown in Table 3.16, together with the unambiguous, directly observed values. Eight haplotypes are new to this survey (as indicated in the table) and have not been reported previously in any other world population (Hong *et al.* 1997). The rare haplotype (--+-+) observed in the TS and MF was constructed unambiguously and therefore certainly exists (Section 3.2). The lowest frequency with which a haplotype can occur in a sample of *n* chromosomes is 1/*n*, which for each of the groups is: MR: 0.0064; BT: 0.0075; AS: 0.0083; TS: 0.0085; MF: 0.025; and VZ: 0.0417. Taking this into account, only three of the eight new haplotypes (-+++-, ++-+- and ---+-), for which frequencies were generated by the EM algorithm but had observed frequencies of 0, could possibly be new 5' β-globin haplotypes. The first of these (-+++-) is found in both the AS and TS; the second (++-+-) in both the MR and VZ; and the third (---+-) in the AS only. Thus it is likely that many of these low-frequency haplotypes are randomly defined by the EM algorithm and do not represent haplotypes actually present in the sample. The frequencies of these haplotypes would eventually drop to zero, after further iteration of the EM algorithm (Martinson *et al.* 1995).

A very strong correlation (r) exists between the two sets of frequencies in each ethnic group (shown in Table 3.16), indicating the ability of the algorithm to derive haplotype frequencies from otherwise uninformative individuals. If additional populations are to be included in the large study, it may be useful to obtain haplotype frequencies using the EM algorithm to prevent losing useful data.

Table 3.16 5' β-globin haplotypes in the Malagasy: maximum likelihood estimates (EM) observed frequencies (OBS) and their correlations (r), and average heterozygosities (h) of EM estimates.

	MERINA		BETSILEO		ANTASAKA		SIMIHETI	
	EM (± S.E.) (78)*	OBS	EM (± S.E.) (67)*	OBS	EM (± S.E.) (60)*	OBS	EM (± S.E.) (59)*	OBS
+++++	0	0	0	0	0	0	0	0
++++-	0	0	0	0	0.001417 (0.004216)	0	0	0
+++--	0.000021 (0.000067)	0	0.000004 (0.000111)	0	0	0	0.000006 (0.000077)	0
++---	0.007528 (0.007528)	0	0.013172 (0.015986)	0	0	0	0	0
+----	0.39179 (0.044628)	0.515	0.374183 (0.046140)	0.54	0.369416 (0.046279)	0.321	0.338563 (0.046082)	0.381
-----	0.032558 (0.019316)	0.06	0.007852 (0.007855)	0.021	0	0	0	0
+-+++	0	0	0	0	0	0	0	0
+-++-	0	0	0	0	0	0	0	0
+-+--	0	0	0	0	0	0	0	0
+-+-+	0	0	0	0	0	0	0	0
++-++	0.000007 (0.000032)	0	0.000078 (0.000504)	0	0	0	0.005009 (0.007042)	0
+--++	0	0	0	0	0	0	0	0
++--+	0	0	0	0	0	0	0	0
+---+	0.047081 (0.021841)	0.076	0.052822 (0.019464)	0.083	0.037437 (0.018399)	0.071	0.010185 (0.015985)	0
----+	0.137463 (0.0.9645)	0.152	0.296436 (0.041246)	0.292	0.301478 (0.048673)	0.375	0.261152 (0.050746)	0.357
---++	0	0	0	0	0	0	0	0
--+++	0	0	0	0	0	0	0	0
-++++	0.019658 (0.011218)	0.015	0	0	0.014579 (0.012683)	0	0.000034 (0.000663)	0
-+-++	0.066078 (0.020122)	0.015	0.059623 (0.020105)	0	0.030829 (0.017437)	0	0.089326 (0.027603)	0
-+-+-	0.027349 (0.013661)	0	0	0	0.032955 (0.018023)	0	0.022845 (0.015082)	0
-+--+	0.057085 (0.018409)	0.076	0.016679 (0.018039)	0.021	0.093526 (0.030580)	0.143	0.052497 (0.020496)	0.119
-+---	0	0	0	0	0.017629 (0.019533)	0	0.017033 (0.017038)	0.048
-++-+	0.204679 (0.030967)	0.091	0.156361 (0.030557)	0.042	0.080485 (0.028522)	0.089	0.164104 (0.033545)	0.071
-++--	0	0	0.007774 (0.007774)	0	0.00003 (0.00906)	0	0.009953 (0.010090)	0
-+++-	0.000001 (0.000003)	0	0.007462 (0.007462)	0	0.011823 (0.012241)	0	0.018379 (0.012940)	0
--+--	0	0	0	0	0	0	0	0
++-+-	0.008619 (0.008676)	0	0	0	0.000061 (0.001277)	0	0	0
+--+-	0.000082 (0.000270)	0	0	0	0.000002 (0.000012)	0	0	0
--+-+	0	0	0	0	0	0	0.008744 (0.008746)	0.024
--++-	0	0	0	0	0	0	0	0
---+-	0	0	0.007463 (0.007464)	0	0.008333 (0.008333)	0	0	0
+++-+	0	0	0.00004	0	0	0	0.00217 (0.007634)	0
r	0.944519927		0.944516495		0.956891621		0.936472018	
h	0.7814		0.7481		0.7609		0.7859	

* Number in brackets is number of individuals.

Table 3.16 5' β-globin haplotypes in the Malagasy: maximum likelihood estimates (EM), observed frequencies (OBS) and their correlations (r), and average heterozygosities (h) of EM estimates.

HAPLOTYPE	MAHAFALY		VEZO	
	EM (± S.E.) (20)*	OBS	EM (± S.E.) (12)*	OBS
+++++	0	0	0	0
++++-	0	0	0	0
+++--	0	0	0.000357 (0.040006)	0
++---	0.000001 (0.000026)	0	0.000005 (0.049203)	0
+----	0.221617 (0.078445)	0.1	0.166662 (0.122385)	0.187
-----	0.025942 (0.025985)	0.05	0.308914 (0.122062)	0.313
+-+++	0	0	0	0
+-++-	0	0	0	0
+-+--	0.000042 (0.049572)	0	0	0
+-+-+	0	0	0	0
++-++	0	0	0	0
+--++	0	0	0	0
++--+	0	0	0	0
+---+	0.05334 (0.036662)	0.05	0.041667 (0.041667)	0.063
----+	0.159518 (0.057956)	0.25	0.232757 (0.091054)	0.313
---++	0	0	0	0
--+++	0.013455 (0.023705)	0	0	0
-++++	0.000068 (0.000808)	0	0	0
-+-++	0.036477 (0.041449)	0.05	0	0
-+-+-	0.025 (0.025)	0	0	0
-+--+	0.250708 (0.076534)	0.25	0.059267 (0.069465)	0.063
-+---	0.052398 (0.035903)	0.05	0.107395 (0.119286)	0.063
-++-+	0.085348 (0.092490)	0.05	0	0
-++--	0	0	0	0
-+++-	0	0	0	0
--+--	0	0	0	0
++-+-	0	0	0.041667 (0.041623)	0
+--+-	0	0	0	0
--+-+	0.076086 (0.082012	0.15	0	0
--++-	0	0	0	0
---+-	0	0	0	0
+++-+	0	0	0.041309 (0.054810)	0
r	0.883362919		0.973201095	
h	0.8496		0.8105	

* Number in brackets is number of individuals.

Average heterozygosities were calculated from the EM frequency estimates (Table 3.16), and again, the Mahafaly and Vezo are the most heterozygous groups (0.8496 and 0.8105) while the Betsileo are the least heterozygous (0.7481). This suggests a greater genetic admixture of the populations that founded the south-west groups than the highlanders, who do claim to be of a "purer stock".

Standard genetic distances were calculated using this data set for the Malagasy and the 5' β-globin haplotype frequencies of the Indonesian and Sub-Saharan African populations used in Section 3.2 (Flint et al 1993a). We are not entirely justified in including these two putative parental populations in this phylogenetic analysis, since their haplotype frequencies were calculated on the basis of homozygosity, rather than by the EM algorithm, as for the Malagasy. However, it was necessary to included some parental populations for comparative purposes. A standard neighbour-joining tree was constructed using DISPAN, and is shown in Figure 3.20. The clustering and position of the populations is fairly similar to that for the tree constructed using unambiguous 5' β-globin haplotype (Figure 3.6). Once again, the Merina are the most closely affiliated to the Indonesians, with the Betsileo being the next closest Malagasy group along this branch. The south-west groups are found along the African branch of the tree, but in this case, the Vezo are more closely affiliated to the African group than the Mahafaly. These differences are probably due to the bias of rare haplotypes obtained from the EM programme.

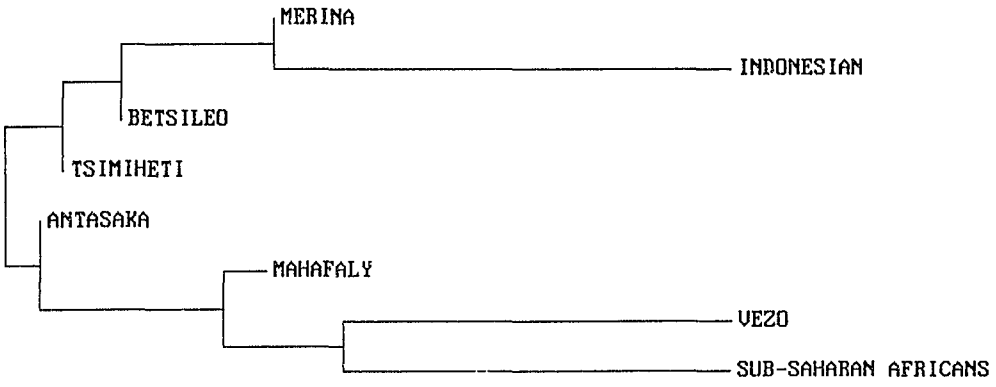


Figure 3.20 Standard neighbour joining tree based on the maximum-likelihood frequencies of the 5' β-globin haplotypes in the Malagasy ethnic groups and other world populations (drawn with DISPAN using Nei's genetic distance).

3.4 3' β-GLOBIN HAPLOTYPES

The frequencies of unambiguous 3' β-globin haplotypes (consisting of the two sites *Ava*II/β and *Hinf*I/β) in each ethnic group are shown in Table 3.17. These results are relatively uninformative and thus for comparative purposes, the frequencies of haplotypes (+ +), (+ -) and (- +) were compared to the frequencies of the frameworks 1, 2 and 3 respectively, consisting of the four 3' RFLP sites, that have been reported in various ethnic groups (Antonarakis *et al.* 1985) (described in Section 1.3.3.1.).

Chi-square tests were performed and the results of the pairwise analysis are shown in Table 3.18. Data for the frequencies of the β-globin gene frameworks in Asiatic Indians and South East Asians were obtained from Antonarakis *et al.* (1985); for the Bantu and San, from Ramsay and Jenkins (1987). All Malagasy ethnic groups differ significantly from South East Asians. Only the Merina show no significant difference to the Asiatic Indians, which may be expected since they are characteristically more Asian than the other Malagasy groups. However, the Mahafaly and Bantu differ significantly from each other, which contradicts the historical data. However, since the Malagasy haplotypes are incomplete frameworks, this finding cannot be regarded as being very important. The Mahafaly are also shown to differ significantly to all the other Malagasy groups (except the Tsimiheti), a finding to which little importance can be attached.

Table 3.17 Frequencies of 3' β-globin haplotypes in the Malagasy.

	Merina (158)	Betsileo (114)	Antasaka (114)	Tsimiheti (104)	Mahafaly (36)	Vezo (28)
+ +	0.582	0.561	0.606	0.519	0.222	0.500
+ -	0.253	0.351	0.263	0.365	0.694	0.464
- +	0.152	0.088	0.105	0.010	0.000	0.000
- -	0.013	0.000	0.026	0.106	0.083	0.036

Number in brackets is number of chromosomes tested.

Table 3.18 Chi-square values and p values (df = 1) for a pair-wise comparison of 3' β-globin haplotypes of the Malagasy ethnic groups and other world populations.

	BT	AS	TS	MF	VZ	ASIATIC INDIAN	SE ASIAN	BANTU	SAN
MR[1] χ^2	3.785	0.927	23.65	30.021	6.337	9.736	45.503	2.063	0.12
p	0.2857	0.3189	$< 10^{-4}$	$< 10^{-4}$	0.0963	0.0209	$< 10^{-4}$	0.5593	0.9894
BT[1] χ^2		2.656	15.284	20.32	2.511	20.291	47.062	3.249	2.455
p		0.4477	0.0016	0.0001	0.4733	$< 10^{-4}$	$< 10^{-4}$	0.197	0.293
AS[1] χ^2			13.075	23.434	4.606	14.736	49.203	1.156	1.262
p			0.0045	$< 10^{-4}$	0.2031	0.0021	$< 10^{-4}$	0.7637	0.7383
TS[1] χ^2				10.714	1.373	44.554	68.373	4.142	16.785
p				0.0134	0.7118	$< 10^{-4}$	$< 10^{-4}$	0.0418	0.0008
MF[1] χ^2					3.881	43.189	27.006	21.687	18.711
p					0.1436	$< 10^{-4}$	$< 10^{-4}$	0.0001	0.0003
VZ[1] χ^2						15.836	20.587	5.212	5.369
p						0.0012	0.0001	0.1569	0.1467

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

3.5 FREQUENCIES OF α-GLOBIN RFLP SITES AND HVRS.

Due to the limited amount of data available, only the most general statistical analysis has been accorded the data of the α-globin RFLPs and HVRs. The frequencies of the presence of the five RFLP sites and two HVRs studied in the α-globin cluster in each of the six Malagasy ethnic groups are given in Table 3.19. (The difference in the number of chromosomes analysed at each polymorphic site, within each Malagasy ethnic group, reflects the fact that some difficulty was encountered in amplifying or restricting older or contaminated samples. Only a limited number of individuals were available from the Vezo and Mahafaly groups, hence the small sample sizes of these two groups).

Table 3.19 Frequencies of the presence (+) of the α-globin RFLP sites and HVRs in the Malagasy ethnic groups and other world populations, and their standard errors (S.E.).

	MR[1]	BT[1]	AS[1]	TS[1]	MF[1]	VZ[1]	SA BLACK[2]	INDIAN[3]	NIGERIAN[3]	PNG[3]	SAUDI ARABIAN[3]	SE ASIAN[3]	ISLAND MELANESIAN[3]
*Xba*I	0.603	0.526	0.481	0.472	0.563	0.674	0.620	0.610	0.500	0.400	0.470	0.410	0.630
S.E.	0.040	0.047	0.049	0.048	0.072	0.069	0.076	0.057	0.083	0.064	0.088	0.086	0.076
n*	156	114	104	108	48	46	40	74	36	60	32	32	40
*Sac*I	0.558	0.533	0.510	0.406	0.340	0.540	0.440	0.780	0.420	0.020	0.470	0.690	0.080
S.E.	0.040	0.045	0.049	0.048	0.066	0.071	0.079	0.048	0.081	0.018	0.081	0.082	0.042
n*	156	122	100	106	50	50	39	74	38	58	38	32	40
*Bgl*I	0.116	0.132	0.087	0.127	0.146	0.227	0.070	0.110	0.030	0.580	0.160	0.190	0.380
S.E.	0.027	0.031	0.030	0.033	0.051	0.064	0.041	0.037	0.028	0.062	0.059	0.069	0.083
n*	146	114	92	102	48	44	40	74	38	62	38	32	34
IZ HVR S	0.425	0.404	0.457	0.441	0.354	0.341	0.260	0.080	0.210	0.360	0.110	0.160	0.500
S.E.	0.041	0.045	0.052	0.049	0.065	0.072	0.071	0.031	0.066	0.064	0.051	0.065	0.079
M	0.548	0.561	0.511	0.510	0.625	0.568	0.740	0.590	0.790	0.640	0.760	0.079	0.500
S.E.	0.041	0.047	0.052	0.049	0.065	0.075	0.071	0.057	0.066	0.064	0.069	0.086	0.079
L	0.027	0.035	0.032	0.049	0.021	0.901	0.000	0.330	0.000	0.000	0.130	0.440	0.000
S.E.	0.014	0.017	0.018	0.021	0.021	0.045	0.000	0.055	0.000	0.000	0.055	0.088	0.000
n*	146	114	92	102	48	44	39	74	38	58	38	32	40
PZ HVR	0.720	0.578	0.460	0.491	0.462	0.625	0.550	0.860	0.550	0.430	0.710	0.790	0.580
S.E.	0.037	0.045	0.049	0.048	0.065	0.065	0.079	0.041	0.081	0.064	0.074	0.075	0.078
n*	150	116	100	110	52	48	40	74	38	58	38	29	40
*Acc*I	0.387	0.357	0.281	0.453	0.435	0.674	0.720	0.850	0.000	0.470	0.670	0.720	0.580
S.E.	0.040	0.045	0.045	0.048	0.051	0.069	0.072	0.041	0.000	0.064	0.078	0.089	0.078
n*	150	112	96	106	46	46	39	74	30	62	36	25	40
θ*Pst*I	0.020	0.043	0.032	0.069	0.080	0.040	0.210	0.030	0.130	0.400	0.060	0.030	0.450
S.E.	0.011	0.018	0.018	0.025	0.038	0.028	0.065	0.023	0.055	0.064	0.057	0.030	0.112
n*	152	116	94	102	50	50	40	60	38	58	18	32	20

[1]MR = Merina; BT = Betsileo; AS = Antasaka; TS =Tsimiheti; MF = Mahafaly; VZ =Vezo.

[2] Ramsay and Jenkins (1988).

[3] Higgs *et al.* (1986).

* n is the number of chromosomes studied.

3.5.1 CHI-SQUARE (χ^2) VALUES AND SIGNIFICANCE TESTING OF RFLP AND HVR FREQUENCIES

The results of the pairwise chi-square tests performed for each of the five α-globin RFLP sites and two HVRs for the six Malagasy ethnic groups and several of their proposed founding populations are given in Tables 3.20 to 3.26 respectively. In the χ^2 calculations, frequency data for the South African Black were obtained from Ramsay and Jenkins (1988) and for all the other populations from Higgs *et al.* (1986). As with the comparisons for the RFLP frequencies in the β-globin gene cluster, there are certain limitation to this analysis with regard to using the correct and/or complete parental populations. Only differences significant at the 99% confidence interval are highlighted.

There are very few differences between the six Malagasy groups. At the PZ/Z site (Table 3.24), the Merina differ from the two lowland groups and the Mahafaly at a level of significant difference; at the *Acc*I site (Table 3.25), the Vezo differ significantly from the two highland groups and the Antasaka. These differences are not unexpected and are consistent with the history of the different groups and the differing degrees to which the parental populations have influenced the different ethnic groups: the groups in the south west are thought to have the greatest African contribution to their ancestry, while the highlanders show the greatest Indonesian influence.

All the Malagasy groups show quite significant differences to the Papua New Guinea (PNG) population and Island Melanesians at the *Sac*I (Table 3.21), *Bgl*I (Table 3.22) and θ*Pst*I (Table 3.26) sites, which is not very surprising since not many studies include these as putative parental populations. The Malagasy also differ significantly from the Indian population at the *Sac*I (Table 3.21), IZ/Z (Table 3.23), PZ/Z (Table 3.24) and *Acc*I (Table 3.25) sites. These findings are also not unexpected since the historical data do not provide evidence for a major genetic contribution to the Malagasy by an Indian population, and may also be supportive evidence of a direct migration across the Indian Ocean of the first Indonesian settlers to Madagascar,

rather than the proposed route via the northern and western shores of the Indian Ocean (Brown 1978).

There are few differences between the Malagasy and other African populations; the only noteworthy differences are seen for the *Acc*I (Table 3.25) and θ*Pst*I (Table 3.26) sites, where large chi-square values are obtained when the highland groups, in particular, are compared to the South African Black and the Nigerian populations. There are a few significant differences between the Malagasy and the South-East Asian population, occurring only at the *Sac*I (Table 3.21), IZ/Z (Table 3.23), PZ/Z (Table 3.24) and *Acc*I (Table 3.25) sites. However these differences are unexpected since they mostly occur between the South-East Asians and the highland groups, who, on the contrary, should show the greatest similarity to this population of all the Malagasy groups. Once again however, although the information provided by the RFLP sites and HVRs is of some use, it is difficult to make assumptions regarding the ancestry of the Malagasy and to determine the relative genetic input of the parental populations from these type of data. As with the β-globin RFLP sites, the α-globin RFLP sites and HVRs are not necessarily population specific and do not hold the same strength of analysis that haplotype analysis does.

Table 3.20 Chi-square values and p values (df = 1) for a pair-wise comparison of Malagasy ethnic groups and other world populations at the *Xba*I site.

	BT	AS	TS	MF	VZ	SA BLACK	INDIAN	NIGERIAN	PNG	SAUDI ARABIAN	SE ASIAN	ISLAND MELANESIAN
MR[1] χ^2	1.267	3.269	3.869	3.423	0.494	0.006	0.004	0.879	6.38	1.441	3.411	0.006
p	0.2603	0.0706	0.0492	0.0643	0.4821	0.938	0.949	0.3484	0.0115	0.023	0.0648	0.938
BT[1] χ^2		0.288	0.451	0.062	2.34	0.801	0.908	0.007	2.031	0.141	1.001	0.801
p		0.5918	0.502	0.8032	0.1261	0.3708	0.3406	0.9329	0.1541	0.7072	0.3172	0.3708
AS[1] χ^2			0	0.581	4.044	1.865	2.329	0	0.703	0.007	0.288	1.865
p			0.9896	0.4459	0.0443	0.1721	0.127	0.9956	0.4019	0.9336	0.5916	0.1721
TS[1] χ^2				0.752	4.492	0.024	2.731	0.009	0.548	0.028	0.208	2.15
p				0.3857	0.034	0.8777	0.0984	0.9233	0.4591	0.8673	0.6484	0.1426
MF[1] χ^2					0.807	0.141	0.097	0.121	2.211	0.353	1.302	0.141
p					0.3689	0.7069	0.7551	0.7283	0.137	0.5524	0.2538	0.7069
VZ[1] χ^2						0.061	0.284	1.868	6.767	2.49	4.464	0.061
p						0.8042	0.5944	0.1717	0.0093	0.1146	0.0346	0.8042

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. $p<0.01$.

Table 3.21 Chi-square values and p values (df=1) for a pair-wise comparison of Malagasy ethnic groups and other world populations at the *Sac*I site.

	BT	AS	TS	MF	VZ	SA BLACK	INDIAN	NIGERIAN	PNG	SAUDI ARABIAN	SE ASIAN	ISLAND MELANESIAN
MR[1] χ^2	0.086	0.382	5.243	6.334	0.003	1.402	10.063	1.775	48.799	0.563	1.342	27.958
p	0.7699	0.5363	0.022	0.0118	0.9557	0.2363	0.0015	0.1828	$<10^{-4}$	0.453	0.2466	$<10^{-4}$
BT[1] χ^2		0.041	3.185	4.539	0.007	0.756	11.365	1.035	42.8	0.203	1.879	24.074
p		0.839	0.0743	0.0331	0.9345	0.3845	0.0007	0.3091	$<10^{-4}$	0.6521	0.1704	$<10^{-4}$
AS[1] χ^2			1.857	3.231	0.03	0.356	12.478	0.552	38.168	0.036	2.414	21.019
p			0.173	0.0722	0.8624	0.5509	0.0004	0.4573	$<10^{-4}$	0.8489	0.1203	$<10^{-4}$
TS[1] χ^2				0.373	1.965	0.832	23.788	0.001	26.865	0.288	6.746	13.221
p				0.5416	0.16098	0.3616	$<10^{-4}$	0.9787	$<10^{-4}$	0.5914	0.0094	0.0003
MF[1] χ^2					3.287	0.496	22.765	0.309	17.883	1.101	8.105	7.561
p					0.0698	0.4814	$<10^{-4}$	0.5784	$<10^{-4}$	0.294	0.0044	0.006
VZ[1] χ^2						0.579	7.133	0.793	35.536	0.161	1.205	19.581
p						0.4467	0.0076	0.3732	$<10^{-4}$	0.6883	0.2723	$<10^{-4}$

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

Table 3.22 Chi-squa.e values and p values (df = 1) for a pair-wise comparison of Malagasy ethnic groups and other world populations at the *Bgl*I site.

	BT	AS	TS	MF	VZ	SA BLACK	INDIAN	NIGERIAN	PNG	SAUDI ARABIAN	SE ASIAN	ISLAND MELANESIAN
MR[1] χ^2	0.032	0.255	0.004	0.81	2.558	0.213	0.002	1.848	46.974	0.171	0.631	12.191
p	0.8583	0.6134	0.9491	0.7765	0.1097	0.6444	0.9674	0.1741	$<10^{-4}$	0.6796	0.427	0.0005
BT[1] χ^2		0.622	0.013	0	1.523	0.452	0.064	2.328	37.199	0.018	0.262	9.164
p		0.4304	0.9103	0.9926	0.2171	0.5014	0.801	0.127	$<10^{-4}$	0.89<	0.609	0.0025
AS[1] χ^2			0.456	0.61	3.954	0.013	0.038	0.738	41.849	0.767	1.498	13.543
p			0.4996	0.4346	0.0467	0.9091	0.8458	0.3903	$<10^{-4}$	0.3812	0.221	0.0002
TS[1] χ^2				0.003	1.617	0.712	0.024	2.123	35.669	0.036	0.313	9.13
p				0.9589	0.2035	0.3988	0.8766	0.1451	$<10^{-4}$	0.8491	0.576	0.0025
MF[1] χ^2					0.542	0.497	0.114	2.314	19.695	0.022	0.034	4.823
p					0.4614	0.4807	0.7356	0.1282	$<10^{-4}$	0.8823	0.8528	0.0281
VZ[1] χ^2						2.641	2.179	5.465	11.684	0.261	0.018	1.535
p						0.1041	0.1399	0.0194	0.0006	0.6093	0.8926	0.2153

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

Table 3.23 Chi-square values and p values (df = 1) for a pair-wise comparison of Malagasy ethnic groups and other world populations at the IZ/Z HVR.

	BT	AS	TS	MF	VZ	SA BLACK	INDIAN	NIGERIAN	PNG	SAUDI ARABIAN	SE ASIAN	ISLAND MELANESIAN
MR[1] χ^2	0.031	0.178	0.402	0.613	2.305	3.685	49.502	5.807	1.199	14.446	44.783	0.503
p	0.9846	0.9148	0.8179	0.7358	0.3159	0.1584	$<10^{-4}$	0.0548	0.549	0.0007	$<10^{-4}$	0.7774
BT[1] χ^2		0.46	0.31	0.243	1.192	2.996	39.058	4.845	1.15	11.407	34.195	0.923
p		0.7946	0.8563	0.8858	0.5511	0.2236	$<10^{-4}$	0.0887	0.5627	0.0033	$<10^{-4}$	0.6303
AS[1] χ^2			0.054	1.132	1.782	4.7	38.799	6.914	1,971	13.746	30.739	0.306
p			0.9733	0.5677	0.4103	0.0954	$<10^{-4}$	0.0315	0.3733	0.001	$<10^{-4}$	0.8583
TS[1] χ^2				1.137	0.959	5.09	36.48	7.283	2.749	12.253	28.168	0.927
p				0.5665	0.6192	0.0785	$<10^{-4}$	0.0262	0.2529	0.0022	$<10^{-4}$	0.6292
MF[1] χ^2					1.037	0.684	21.77	1.669	0.02	7.126	19.014	1.193
p					0.5954	0.7102	$<10^{-4}$	0.4341	0.9902	0.0284	0.0001	0.5509
VZ[1] χ^2						2.886	14.452	3.834	3.311	4.871	10.415	2.974
p						0.2362	0.0007	0.147	0.191	0.0876	0.0055	0.226

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. $p<0.01$.

Table 3.24 Chi-square values and p values (df = 1) for a pair-wise comparison of Malagasy ethnic groups and other world populations at the PZ/Z HVR.

	BT	AS	TS	MF	VZ	SA BLACK	INDIAN	NIGERIAN	PNG	SAUDI ARABIAN	SE ASIAN	ISLAND MELANESIAN
MR[1] χ^2	5.279	16.065	13.223	10.278	1.137	5.05	3.474	8.928	3.705	22.372	25.491	10.935
p	0.0216	0.0001	0.0003	0.0013	0.2864	0.0246	0.0624	0.0028	0.0542	$<10^{-4}$	$<10^{-4}$	0.0009
BT[1] χ^2		2.524	1.375	1.508	0.15	16.099	0.014	1.468	0.003	8.388	11.313	2.206
p		0.1121	0.241	0.2194	0.6984	0.0001	0.9055	0.2257	0.9568	0.0038	0.0008	0.1374
AS[1] χ^2			0.096	0.024	2.905	28.264	0.601	0.003	1	2.637	4.98	0.035
p			0.7569	0.878	0.0883	$<10^{-4}$	0.4381	0.9536	0.2479	0.1044	0.0256	0.8509
TS[1] χ^2				0.033	1.905	25.293	0.207	0.076	0.64	3.871	6.432	0.281
p				0.8565	0.1676	$<10^{-4}$	0.6488	0.7834	0.4235	0.0491	0.0112	0.5961
MF[1] χ^2					2.067	21.71	0.398	0.007	0.874	2.059	4.142	0.019
p					0.1505	$<10^{-4}$	0.5281	0.9355	0.35	0.1513	0.0418	0.8903
VZ[1] χ^2						8.165	0.245	2.031	0.149	8.274	11.07	2.75
p						0.0043	0.6207	0.1541	0.6994	0.004	0.0009	0.0973

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. $p<0.01$.

Table 3.25 Chi-square values and p values (df = 1) for a pair-wise comparison of Malagasy ethnic groups and other world populations at the *Acc*I site.

	BT	AS	TS	MF	VZ	SA BLACK	INDIAN	NIGERIAN	PNG	SAUDI ARABIAN	SE ASIAN	ISLAND MELANESIAN
MR[1] χ^2	0.129	2.429	0.865	0.169	10.588	41.228	12.395	24.413	0.88	8.133	8.382	3.842
p	0.7192	0.1191	0.3525	0.681	0.0011	$<10^{-4}$	0.0004	$<10^{-4}$	0.3482	0.0043	0.0038	0.05
BT[1] χ^2		1.038	1.693	0.537	11.974	42.065	13.79	25.845	1.604	9.411	9.587	4.901
p		0.3083	0.1932	0.4635	0.0005	$<10^{-4}$	0.0002	$<10^{-4}$	0.2053	0.0022	0.002	0.0268
AS[1] χ^2			5.64	2.653	18.253	52.252	20.135	33.251	4.94	14.819	14.521	9.254
p			0.0176	0.1033	$<10^{-3}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0262	0.0001	0.0001	0.0023
TS[1] χ^2				0.001	5.427	27.616	7.007	17.138	0.001	4.098	4.757	1.281
p				0.9776	0.0198	$<10^{-4}$	0.0081	$<10^{-4}$	0.9787	0.0429	0.0292	0.2578
MF[1] χ^2					4.4	21.71	5.781	14.741	12.483	3.484	4.212	1.168
p					0.0359	$<10^{-4}$	0.0162	0.0001	0.0004	0.062	0.0401	0.2797
VZ[1] χ^2						4.269	0.041	3.96	3.749	0.028	0.018	0.523
p						0.0388	0.8393	0.0466	0.0528	0.867	0.8946	0.4697

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. $p<0.01$.

Tab 3.26 Chi-square values and p values (df=1) for a pair-wise comparison of Malagasy ethnic groups and other world populations at the θ*Pst*I site.

	BT	AS	TS	MF	VZ	SA BLACK	INDIAN	NIGERIAN	PNG	SAUDI ARABIAN	SE ASIAN	ISLAND MELANESIAN
MR[1] χ^2	0.565	0.031	2.673	2.482	0.076	0.007	15.861	6.859	51.53	0.016	0.068	44.004
p	0.4524	0.86	0.102	0.1152	0.7831	0.932	0.0001	0.0088	$<10^{-4}$	0.8999	0.7941	$<10^{-4}$
BT[1] χ^2		0.003	0.278	0.348	0.109	0.009	7.641	2.377	33.245	0.14	0.042	26.337
p		0.9532	0.5983	0.5555	0.7417	0.9263	0.0057	0.1231	$<10^{-4}$	0.7078	0.8374	$<10^{-4}$
AS[1] χ^2			0.709	0.758	0.051	0.174	8.408	3.133	31.111	0.039	0.319	26.328
p			0.3998	0.3841	0.8241	0.6761	0.0037	0.0767	$<10^{-4}$	0.843	0.572	$<10^{-4}$
TS[1] χ^2				0.006	0.113	0.35	3.95	0.712	23.991	0.095	0.123	18.128
p				0.9371	0.7362	0.5539	0.0469	0.3988	$<10^{-4}$	0.7585	0.7256	$<10^{-4}$
MF[1] χ^2					0.177	0.425	1.828	0.19	12.712	0.035	0.182	10.602
p					0.6737	0.5147	0.1763	0.663	0.0004	0.8526	0.6695	0.0011
VZ[1] χ^2						0.106	4.254	1.38	17.237	0.155	0.158	15.168
p						0.7448	0.0392	0.24	$<10^{-4}$	0.6938	0.6913	0.0001

[1] MR = Merina; BT = Betsileo; AS = Antasaka; TS = Tsimiheti; MF = Mahafaly; VZ = Vezo.
Highlighted p values indicate differences significant at the 99% confidence interval, ie. p<0.01.

3.5.2 PHYLOGENETIC ANALYSIS

For consistency, a standard neighbour joining tree similar to the tree constructed for the 5' β-globin RFLP sites (Figure 3.1) was constructed (using DISPAN). The frequencies of the five α-globin RFLP sites and two HVRs in the Malagasy and other world populations (Higgs *et al.* 1986) were used to calculate a matrix of genetic distances. The tree is shown in Figure 3.21. It is clear that this analysis is not very accurate, since the clustering of the populations is not entirely in keeping with the trends that have been observed from the more reliable β-globin haplotype data. However, the trends that are consistent, are that the Merina and Betsileo are shown to be closely related, with the Antasaka and Tsimiheti intermediate in position again between the highland groups and the south-west groups. The Vezo are the most closely affiliated of all the Malagasy groups with the African cluster of populations. But the Mahafaly are the furthest from the African cluster along the same branch, and are closest to the node of the tree which leads to the Island Melanesian/PNG populations.

Complete α-globin haplotypes, consisting of the seven RFLPs and two HVRs, could not be constructed due to the problems described in Section 2.6.3. with the PCR for the *Rsa*I and α*Pst*I sites. Further, it was not possible to construct even partial unambiguous haplotypes based on the results obtained for the five RFLP sites and two HVRs in the Malagasy, because of the high degree of heterozygosity. For a complete analysis, it will be necessary to obtain data for all the sites in the α-globin cluster, in particular the five 3' sites that define the major α-globin haplotype groups (discussed in Section 1.4.2.3.), and to generate α-globin haplotypes. Using the frequencies of the α-globin haplotypes thus generated, meaningful comparative analyses can be performed, from which conclusions may be generated.

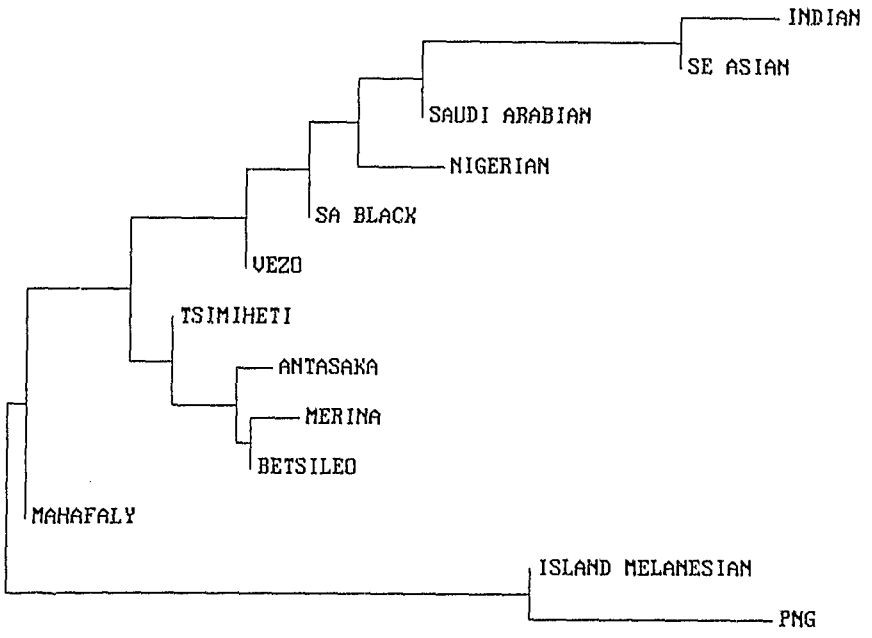


Figure 3.21 Standard neighbour joining tree based on the frequencies of the 5 α-globin RFLP sites and 2 HVRs (drawn with DISPAN using Nei's genetic distance).

CHAPTER 4

GENERAL CONCLUSIONS

In the previous chapter, the results obtained in each section were discussed individually, since a number of different statistical methods of analysis were used. This chapter serves as a general discussion of all the results that were obtained, in an attempt to highlight the most significant trends that were noted in the study, as well as the general conclusions that can be drawn, in conjunction with the history and other genetic studies, concerning the ancestry of the Malagasy.

DNA polymorphisms have provided population geneticists with powerful tools for the study of population affinities, migration patterns and gene flow between populations, as well as with estimates of recombination frequencies and mutation rates (Ramsay and Jenkins 1988). The present study deals with variation in and around the α- and β-globin gene clusters in the Malagasy in an attempt to provide additional information on the origins and affinities of this population.

Individually, the RFLP sites did not provide many useful results upon chi-square analysis of their frequencies in the Malagasy and their parental populations. No significant trends concerning population affinities were obtained. This is because not all the single site RFLPs are specific to particular populations, and any variation in frequencies between populations may be due to chance rather than being a reflection of the degree of relatedness of the populations. Contrary to this, haplotype analysis reveals much more information. A haplotype represents a type of "combined variation" of the information contained in the polymorphic sites that make up the haplotype. The polymorphic sites are said to exist in strong linkage disequilibrium since they are so close together on the chromosome that there is little chance that recombination will separate them and they are inherited as single alleles. Consequently, genetic drift and the relatively slow dispersal of haplotypes across the world means that genetically separated populations will contain unique and characteristic distributions of haplotypes. This is particularly true for the 5' β-globin and α-globin haplotypes, since in each

population there are a few common haplotypes and a larger number of rarer ones, but what is common in one population is not necessarily so in another (Flint *et al.* 1993b). The frequencies of the unambiguously assigned 5' β-globin haplotypes in the Malagasy were an extremely useful data set, providing the basis from which a number of analytical methods could be carried out, at different levels of population sub-structure.

Throughout this study, populations from Africa and Asia (particularly Indonesia) were used for comparative purposes since the historical, linguistic, archaeological and other genetic data suggest that the major sources of Malagasy ancestry are Africa and Indonesia (reviewed in Dewar and Wright, 1993, Campbell 1995). In this study however, we were limited by the data available for the choice of the putative parental populations (for example, no data were available for an Arab population). We were also limited in knowing exactly which populations were appropriate to include as "parental" populations. Thus a number of African and Asian populations were included as far as possible in most analyses to avoid any bias in selection of parental populations and also to try and place the Malagasy in a broad world context, to ascertain the closest affinities with other populations. In such an investigation of population affinities, it would be most appropriate to compare our data with that of African populations from the Mozambique-Malawi region, since this is the most likely region from which the putative ancestral African population of the Malagasy is thought to have come (Campbell 1997). However, for most of the analyses used in this study, only data from Bantu, San and a general "Sub-Saharan African" population were available. Similarly, the most appropriate Asian population to include in this study for comparative purpose would be one from the Barito River region of Indonesia, since the linguistic data suggest that the first Indonesian to settle on Madagascar came from this area (Dewar and Wright 1993), but as no such data were available, only the data for a broad "Indonesian" population was used in most analyses. Therefore the results of the comparative analyses of the different populations included in this study provided only a measure of the populations' affinities, and did not provide conclusive evidence as to who the true ancestral populations of the Malagasy were.

The results consistently showed that the Malagasy are an admixed population of both Indonesian and African (more specifically, Sub-Saharan/Bantu) ancestry. There is some variation between the Malagasy ethnic groups, which becomes masked when the groups are compared to other world populations, suggesting that the Malagasy groups are more closely related to each other than any one group is related to a parental population. Some of this variation may be due to the difference in the number of individuals studied in each ethnic group, with particular reference to the extremely small sample of Vezo and Mahafaly. Any differences within the Malagasy become more apparent when they are classified into the larger regional groups: the highland, lowland and south-west groups clearly show closer affiliations to some parental population(s) rather than another. The highlanders consistently show a greater Indonesian influence compared to the lowlanders and south-west groups, who show a greater African influence to their ancestry, with the south west groups being the most African in character. This is in accordance with the historical and ethnographic data (discussed in Section 1.5), that documents definite regional differences between the Malagasy ethnic groups, for example in physical appearance and cultural practices (Brown 1978, Campbell 1995).

Chi-square analysis of the 5' β-globin haplotypes provided a useful overview of these trends, but a more accurate estimate of the population affinities was obtained from genetic distances. In order to provide more information for inferences on population genealogy (rather than summary statistics), these were represented graphically, in one-dimension as phylogenetic trees, and in two-dimensions, as Principle Component maps. The clustering of the populations in the diagrams (Figures 3.12-3.15) showed that the highlanders are more closely affiliated to the Indonesian/Polynesian populations, while the south-west groups are most closely associated with the African populations. The lowland groups consistently appear intermediate in position between the highland and south-west groups.

Buettner-Janusch and Buettner-Janusch (1964) also found a higher African element to predominate in the lowland populations than in the highlanders from a study of haptoglobins within the Malagasy. They reported the frequencies of the haptoglobin allele Hp1 to be 66% and 54% in coastal and highland populations respectively, with that of the lowlanders being closer to the African frequency and higher than the value of 24% found in Malaysia. In a study of the ABO blood groups in the Malagasy, the distribution of the B gene (which is a South-east Asian rather than an African characteristic) also occurred at varying frequencies among the Malagasy. Among the Merina, the Andriana or nobles had 18% of B; the Hova or commoners were found to have 21% of B genes; and the Mainty or former slave population had 15% (Mourant *et al.* 1976). This is consistent with the origins of the different castes, with a mainly Indonesian origin for the Andriana and Hova and an African one for the Mainty (Campbell 1997). The Mahafaly were found to have only 18% of B genes (Mourant *et al.* 1976), which is also consistent with the results of the current study, that suggest that the Mahafaly have a greater African than Indonesian contribution to their ancestry. Also, recently Morar *et al.* (1997, personal communication) used a number of autosomal polymorphic loci to estimate the proportional contributions of Africans and Indonesians/South East Asians to highland and lowland Malagasy groups. They found that the African contribution to highland groups was 38% compared to a 71% contribution to the lowland groups.

The findings of the present study, together with these, differ somewhat from the results obtained from a study of G6PD variation in the Malagasy (Dangerfield 1997). Phylogenetic and PC analysis of Malagasy G6PD-haplotypes showed the Vezo and Mahafaly (the "south-west" groups) to associate more strongly with Austronesian populations, than even the highland groups. The south-west groups were shown to have the highest proportion of Indonesian admixture (81%) compared to the highland groups (50%) and the lowland groups (15%). In the present study however, the Vezo and Mahafaly show very close associations to the African populations compared to the highland and lowland groups. They are the most genetically diverse Malagasy groups,

and show the widest spread of haplotypes (Table 3.3). Moreover, the south-west groups are the most heterozygous Malagasy groups (based on unambiguous as well as EM-derived 5' β-globin haplotype frequencies) which is more in keeping with a stronger African input to their ancestry, rather than a strong input from an Indonesian population (who show relatively little genetic diversity). From the heterozygosity versus distance analysis, the south-west groups consistently show higher heterozygosities than expected, indicative of higher rates of gene flow experienced by these populations. Again these findings are in agreement with the historical evidence that the south-west of Madagascar received a strong genetic input from African slaves imported mainly from Mozambique, in addition to the deliberate colonists to the island from Africa (Campbell 1995).

The admixture estimates generated in this study also show that the south-west groups have the highest proportion of African admixture (78%) compared to the highland (47%) and lowland (69%) groups. Despite this, however, the present-day Vezo and Mahafaly peoples still claim a greater Indonesian ancestry, since there seems to be a stigma attached to admitting African ancestry (Professor Roux, Pasteur Institute, Madagascar, personal communication). The relatively low β^S gene frequencies of the Mahafaly and Vezo (0 and 0.034 respectively) lend some support to these claims of a greater Indonesian influence to their ancestry (since β^s occurs in Africa but not in Indonesia). However, the frequencies of the β^S gene may be lower in these populations since the south-west of Madagascar is drier and malaria is not present in such endemic proportions as compared to the other coastal regions. Hence there may not be the same strong selective pressure to maintain the β^S gene at a high frequency. The south-west groups also have lower frequencies of G6PD deficiency (0.073) than the highland groups (0.089) (Dangerfield 1997). This supports the selection argument for lower β^S gene frequencies, since G6PD occurs in south-east Asia as well as in Africa, and G6PD deficiency variants are also influenced by selection.

In the present study and that of Dangerfield, the conflicting admixture estimates for the regional Malagasy groups require further explanation. In these two studies, there has been some consistency with respect to selecting the Malagasy groups studied (groups from the highland, lowland and south-west were chosen in both cases). But once again there have been limitations set by the data available for the putative parental populations (ie. The specific African and Indonesian populations) and of knowing with certainty which parental populations to include in the analysis. Long *et al.* (1991) has suggested that heterogeneity of admixture estimates across different loci may be due random drift and natural selection, as well as to errors in estimating the ancestral gene frequencies. In particular, random genetic drift may be a plausible explanation since the effective population sizes are difficult to estimate but are most certainly much smaller for the hybrid Malagasy ethnic groups than for the putative ancestral populations. Also the effective population size may change depending on the locus one is analysing: autosomal, where there are equal contributions made by males and females, or X-linked, where the contribution made by males is only half of that made by females. Natural selection may also bias the estimates, particularly in the G6PD and the present studies, where the loci that have been analysed may be under strong selective pressures, unlike the autosomal markers, which are assumed to be neutral markers. The selective pressures may differ geographically within Madagascar (highland versus lowland groups in accordance with the distribution of malaria), and also between Madagascar and the areas from which the ancestral Malagasy are thought to have come (Indonesia and Africa). Selection may also vary temporally, since with time, the selective pressure could change at differing rates in Madagascar and the areas of the respective ancestors.

Hill (1996) has also addressed this problem in connection with inferring population affinities from genetic distances and the subsequent phylogenetic analyses. In this approach, one assumes that no selection is acting to affect genetic differentiation of the populations studied Again this is especially relevant in the current study since we know of the selective pressures exerted by malaria on α- and β-alleles. It may not be

consistent to include populations from Africa in the same comparison as populations from Indonesia/Polynesia, since there may be different strains of malaria which may confer different selective pressures on the genes involved in the different areas.

A similar explanation for differing results for the same population may hold true when we compare the results generated in this study to the data from a mitochondrial study of the Malagasy (Soodyall *et al.* 1993). Mitochondria represent only the female contributions to the population, and it is possible that the male and female contributions of one or more of the parental populations were disproportionate. Analysis of Malagasy mtDNA has shown that approximately 71% of mtDNA associated with the 9bp deletion in the Malagasy is derived from Asia, whereas 29% is derived from Africa. In the present study, we estimate the African contribution to the overall Malagasy population to be 1%, while the corresponding Indonesian contribution was estimated to be 39%. In addition, the mtDNA "Polynesian" motif was found in 96% of Malagasy individuals who had the "Asian " deletion, suggesting that the Malagasy are more closely affiliated to Polynesians than they are to Indonesians (Soodyall *et al.* 1993), in contrast to the present study, in which the results (particularly those of the chi-square tests) consistently showed the Malagasy to be more closely related to the Indonesians than the Polynesians. There are also discrepancies between the linguistic evidence and genetic data: the mtDNA analysis showed that the "Polynesian" motif was absent in Indians, Africans and Indonesians from the Barito River area, demonstrating that linguistic affiliation between populations is not a good indicator of biological relatedness. However, the mtDNA analysis is only one study and the results may be slightly biased since they only account for the female contribution to the Malagasy gene pool; mtDNA is maternally inherited, unlike the autosomal genomic α- and β-globin loci where equal contributions are made by both parents. But it does lend support to the hypothesis that Madagascar was colonised directly from the South Pacific region, and not by an indirect route via India and Africa, as has been previously postulated (Brown 1978). The β^S-study provides similar support for the direct route, since no "Arab-Indian" β^S-haplotype

was detected among the Malagasy, suggesting that a large Arab-Indian contribution to Malagasy ancestry is unlikely (Hewitt *et al.* 1996).

It is clear that it is difficult to specify the exact putative Indonesian population as the ancestral Malagasy population, especially when one considers a number of loci. Similarly, it is difficult to specify the exact ancestral African population, since there have been waves of migrants form Africa to Madagascar, as deliberate colonists and as slaves, but this may be narrowed down slightly more than for the case of the ancestral Indonesian population. In the β^s study (Hewitt *et al.* 1996), it was suggested that a major source of African ancestry to the Malagasy was derived from Bantu-speaking peoples from central or east Africa but not Africa south of the Zambezi. This was based on the fact that the β^s-gene is found at very low frequencies south of the Zambezi River and is virtually absent in populations living south of the Limpopo, but does occur in Madagascar, at overall frequencies of 0.0156 in the highland populations and 0.0744 in the lowland populations. Also, since none of the other African β^s-haplotypes (Senegal, Benin or Cameroon) was found at significant frequencies in the Malagasy, it is unlikely that the African ancestors came from Atlantic west Africa, central west Africa or the Cameroon. The single Senegal haplotype found in the β^s study has since been attributed to French soldiers who had been based in Senegal, and who presumably carried the Senegal β^s-haplotype to Madagascar where they were subsequently stationed. Support for these β^s gene frequencies is obtained from the distribution of G6PD deficiency variants in Africa. The G6PD deficiency frequency in Madagascar is similar to the frequencies of the deficiency in central and east Africa (notably in Mozambique) which support the suggestion made in the β^s study for the putative parental African population. The most prevalent G6PD deficiency variant found in the Malagasy was the African specific A^- variant, occurring at an overall frequency of 87% (Dangerfield 1997). In the same G6PD study, it was suggested that a Zambian population is a good candidate ancestral population based on admixture calculations, and also since the G6PD deficiency frequency in Madagascar is concordant with the frequency found in a Zambian population. Contribution from a Tsonga population was

also suggested from admixture calculations, and this population is a good candidate based on the proximity of Mozambique to Madagascar. Non-genetic studies have suggested the Mozambique-Malawi stretch to be the area from which most African Malagasy arose (reviewed in Campbell 1995). However these may have been the later immigrants to Madagascar, imported as slaves, and not necessarily the true ancestral population.

The discrepancies encountered from these sets of analyses, emphasise the need to perform a number of studies at different levels (protein variation, mtDNA and nuclear DNA RFLPs) to generate unequivocal data. It has occurred in other studies that some loci have been uninformative of the pattern of affinities among populations (eg. in a study by Roberts-Thomson et al. (1996, who used α-globin cluster rearrangements and haplotypes to investigate affinities between Australian Aborigines and PNG highlanders and island Melanesians, to trace gene flow and colonising events into Australia). In particular, Zago et al (1995) found a homogeneous distribution of β-globin haplotypes but heterozygous distribution of α-globin haplotypes and rearrangements among different tribes of South American Indians, reflecting the action of various genetic mechanisms (genetic drift, founder effect, consanguinity) on small isolated population groups with a complicated history of divergence-fusion events. It is not unlikely that the same could hold true for the Malagasy.

Once again, we note the importance of considering a number of loci (including neutral markers) in a study which aims to reconstruct one population's history through analysing its affinities with other populations. In this study and the other genetic studies of Malagasy ancestry, we note the complexities of studying small groups of populations and the corresponding bias of sample size and its effects of founder effect and genetic drift. Kus and Raharijaona (1992) succinctly referred to the Malagasy as "a culture whose integrity has been forged in place". Kent also remarked that "to "crack" the Malagasy origins has been the stated or secret dream of most Malgachisants" (Dewar 1995). Yet all of the available linguistic, biological and cultural data clearly reflect long-

term interactions between diverse populations on and off the island. It is hoped that once all the data from this large genetic study of the Malagasy being undertaken by the Department of Human Genetics are amalgamated, a clearer and more comprehensive picture of Malagasy origins will emerge.

REFERENCES

Antonarakis S., Boehm C., Giardina P. and Kazazian H. (1982) Nonrandom association of polymorphic restriction sites in the β-globin gene cluster. Proc. Natl. Acad. Sci. 79:137-141.

Antonarakis S., Boehm C., Serjeant G., Theisen C., Dover G. and Kazazian H. (1984) Origin of the β^S-globin gene in Blacks: The contribution of recurrent mutation or gene conversion or both. Proc. Natl. Acad. Sci. 81:853-856.

Antonarakis S., Kazazian H. and Orkin S. (1985) DNA polymorphism and molecular pathology of the human globin gene clusters. Hum. Genet. 69:1-14.

Bank A., Mears J., Ramirez F., Burns A., Spence S., Feldenzer J. and Baird M. (1980) The organisation of the γ-δ-β gene complex in normal and thalassaemia cells. Haemoglobin 4:497-507.

Batzer M., Arcot S., Phinney J., Alegria-Hartman M., Kass D., Milligan S., Kimpton C., Gill P., Hochmeister M., Ioannou P., Herrera R., Boudreau D., Scheer W., Keats B., Deininger P. And Stoneking M. (1996) Genetic variation of recent Alu insertions in human populations. J. Mol. Evol. 42:22-29

Bodmer W. F. and Cavalli-Sforza L.L (1976) Genetics, Evolution and Man. W. H. Freeman, San Francisco.

Breuning M., Madan K., Verjaal M., Wijnen J., Meera Khan P. And Pearson P. (1987) Human α-globin maps to pter-p13.1 in chromosome 16 distal to PGP. Hum Genet 78:287.

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

Buckle V., Higgs D., Wilkie A., Super M. And Weatherall D. (1988) Localisation of human α-globin to 16p13.3-pter. J Med Genet. 25:847.

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

Campbell G. (1988) Madagascar and Mocambique in the slave trade of the Western Indian Ocean 1800-181. Slavery and Abolition. A journal of comparative studies 9:1-193.

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

Campbell G. (1995) Theories of the origins of the Malagasy: The nineteenth century British School. In: Michel M. And Paillard Y. (Eds) Australes pp 127-153.

Campbell G. (1997) The Origins and Demography of slaves in nineteenth century Madagascar: A chapter in the history of the African ancestry of the Malagasy. In Press.

Carestia C., Pagano L., Lacerrra G., Fioretti G., De Angioletti M., de Bonis G. and De Rosa L. (1991) β-globin gene framework variant: an African type? Nucleic Acids Res. 19:6968.

Cavalli-Sforza L. Kidd J., Kidd K., Bucci C., Bowcock A., Hewlett B. and Friedlaender J. (1986) DNA markers and genetic variation in the human species. Cold Spri.g Harbour Symp. Quant. Biol 51:411-417.

Cavalli-Sforza L., Menozzi P. And Piazza A. (1994) The History and Geography of Human Genes. New Jersey: Princeton University Press. Ch 1 pp. 24-42.

Chakraborty R. (1986) Gene admixture in human populations: Models and predictions. Yearbook of Phys Anthrop 29:1-43.

Chakravarti A., Buetow K., Antonarakis S., Waber P., Boehm C. and Kazazian H. (1984) Nonuniform recombination within the β-globin gene cluster. Amer. J. Hum. Genet. 36:1239-1258.

Chebloune Y., Pagnier J., Trabuchet G., Faure C., Verdier G., Labie D. and Nigon V. (1988) Structural analysis of the 5' flanking region of the β-globin gene in African sickle cell anaemia patents: Further evidence for three origins of the sickle mutation in Africa. Proc. Natl. Acad. Sci. 85:4431-4435.

Chen L., Easteal S., Board P. And Kirk R. (1990) Evolution of β-globin haplotypes in human populations. Mol. Biol. Evol. 7(5):423-437.

Cooper D. and Schmidtke J. (1984) DNA restriction fragment length polymorphisms and heterozygosity in the human genome. Hum. Genet. 66:1-16.

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

Deisseroth A., Nienhuis A., Turner P., Velez R., Anderson W., Ruddle F., Lawrence J., Creagan R. And Kucherlapati R. (1977) Localisation of the human α-globin structural gene to chromosome 16 in somatic cell hybrids by molecular hybridisation assay. Cell 12:205-218.

Deisseroth A., Nienhuis A., Lawrence J., Giles R., Turner P. And Ruddle F. (1978) Chromosomal localisation of human β-globin gene on chromosome 11 in somatic cell hybrids. Proc. Natl. Acad. Sci. 75(3):1456-1460.

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

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

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

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

Efstratiadis A., Posakony J., Maniatis T., Lawn R., O'Connell C., Spritz R., DeRiel J., Forget B., Weissman S., Slightom J., Blechl A., Smithies O., Baralle F., Shoulders C. and Proudfoot N. (1980) The structure and evolution of the human β-globin gene family. Cell 21:653-668.

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

Feldenzer J., Mears J., Burns L., Natta C. and Bank A. (1979) Heterogeneity of DNA fragments associated with the sickle-globin gene. J. Clin. Inves. 64:751-755.

Flint J., Harding R., Boyce A. and Clegg J. (1993a) The population genetics of the haemoglobinopathies. Balliere's Clinical Haematology 6:215-262.

Flint J., Harding R., Clegg J. and Boyce A. (1993b) Why are some genetic disease so common? Distinguishing selection from other processes by molecular analysis of globin gene variants. Hum. Genet. 91:91-117.

Goodbourn S., Higgs D., Clegg J. And Weatherall D. (1983) Molecular basis of length polymorphisms in the human ξ-globin gene complex. Proc. Natl. Acad. Sci. 80:5022-5026.

Grosveld F., Dillon N. and Higgs D. (1993) The regulation of human globin gene expression. Bailliere's Clinical Haematology 6:31-55.

Guerreiro J., Figueiredo M. and Zago M. (1994) Beta-globin gene cluster haplotypes of Amerindian populations from the Brazilian Amazon region. Hum Hered. 44:142-149.

Harding R., Fullerton S., Griffiths R. And Clegg J. (1997a) A gene tree for β-globin sequences from Melanesian. J. Mol. Evol. 44: Suppl:5133-5138.

Harding R., Fullerton S., Griffiths R., Bond J., Cox M., Schneider J., Moulin D. And Clegg J. (1997b) Archaic African and Asian lineages in the genetic ancestry of modern humans. Am. J. Hum. Genet. 60 (4):772-789.

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

Harris H. (1980) The principles of human biochemical genetics. Elsevier/North Holland Biomedical Press, Amsterdam.

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

Higgs D., Wainscoat J., Flint J., Hill A., Thein S., Nicholls R., Teal H., Ayyub H., Peto T., Falusi Y., Jarman A., Clegg J. And Weatherall D. (1986) Analysis of the human α-globin gene cluster reveals a highly informative genetic locus. Proc. Natl. Acad. Sci. 83:5165.

Higgs D., Vickers M., Wilkie A., Pretorius I., Jarman A. And Weatherall D. (1989) A review of the molecular genetics of the human α-globin gene cluster. J of American Soc of Hemat. 73(5):1081-1104.

Hill A. (1996) Measuring natural selection by malaria. In: Boyce A. And Mascie-Taylor C. (Eds) Society for the study of Human Biology Symposium 38. Ch7 pp 81-92.

Hill A., Nicholls R., Thein S. And Higgs D. (1985) Recombination within the human embryonic ξ-globin locus: A common ξ-ξ chromosome produced by gene conversion of the ξ-gene. Cell 42:809.

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

Hong S., Goh S., Soodyall H., Stoneking M., Namkoong Y. And Lee C. (1997) β^A-globin gene haplotypes in two Korean populations. Hum Biol. (In Press).

Huisman T. (1987) A short review of human γ-globin gene abnormalities. Acta. haemat. 78:80-84.

Ingram V. (1958) Abnormal human haemoglobins. I. The comparison of normal human and sickle-cell haemoglobins by "fingerprinting". Biochim. Biophys. Acta 28:539-545.

Jarman A. And Higgs D. (1988) A new hypervariable marker for the human α-globin gene cluster. An J Hum Genet 43:249-256.

Jeffreys A. (1979) DNA sequence variants in the ^{G}γ-, ^{A}γ-, δ- and β-globin genes of man. Cell 18:1-10.

Jeffreys A., Wilson V. And Thein S. (1985) Hypervariable "minisatellite" regions in human DNA. Nature 314:67.

Jenkins T. (1992) Origins and Affinities of the people of Madagascar. Research proposal submitted to National Geographic. South African Institute of Medical Research and University of the Witwatersrand.

Kan Y. and Dozy A. (1978) Polymorphism of DNA sequence adjacent to human β-globin structural gene: Relationship to sickle mutation. Proc. Natl. Acad. Sci. 75:5631-5635.

Kan Y., Dozy A., Trecartin R. and Todd D. (1977) Identification of a nondeletional defect in α-thalassemia. N. Engl. J. Med. 297:1081-1084.

Kazazian H., Chakravarti A., Orkin S. and Antonarakis S. (1983) DNA polymorphisms in the human b globin gene cluster. In: Evolution of Genes and Proteins. Nei M. and Koehn R. (Eds), Sinauer Associates Inc., Massachusetts.

Kulozik A., Lyons J., Kohne E., Bartram C. And Kleihauer E. (1988) Rapid and non-radioactive prenatal diagnosis of β-thalassemia and sickle-cell disease: application of the polymerase chain reaction (PCR). Br. J. Haemat. 70:455-458.

Lauer J., Shen C., and Maniatis T. (1980) The chromosomal arrangement of the human α-like globin genes: sequence homology and α-globin gene deletions. Cell 20:119-130.

Lawn R, Fritsch E, Parker R, Blake G and Maniatis T (1978) The isolation and characterisation of linked γ- and β-globin genes from a cloned library of human DNA. Cell 15:1157-1174.

Lawn R., Efstratiadis A., O'Connell C. and Maniatis T. (1980) The nucleotide sequence of the human β-globin gene. Cell 21:647-651.

Lehman H. and Huntsman R. (1974) Man's Haemoglobins. North Holland Publishing Co., Amsterdam. p 89.

Long J., Williams R., McAuley J., Medis R., Partel R., Tregellas W., South S., Rea A., McCormick S. and Iwaniec U. (1991) Genetic variation in Arizona Mexican Americans: estimation and interpretation of admixture proportions. Am. J. Phys. Anthrop. 84(2):141-157.

Long J., Williams R. and Urbanek M. (1995) An EM algorithm and testing strategy for multiple-locus haplotypes. Am. J. Hum. Genet. 56:799-810.

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

Magenis R., Donlon T. and Tomar D. (1985) Localisation of the β-globin gene to 11p15 by *in situ* hybridisation: Utilisation of chromosome 11 rearrangements. Human Genetics 69:300-303.

Maniatis T., Fritsch E., Lauer J. and Lawn R. (1980) The molecular genetics of human haemoglobins. Ann. Rev. Genet. 14:145-178.

Marks J., Shaw J. And Shen C. (1986) Sequence organisation and genomic complexity of the primate θ1 globin gene, a novel α-globin like gene. Nature 321:785.

Martinson J. (1996) Molecular perspectives on the colonisation of the Pacific. In: Boyce A. And Mascie-Taylor C. (Eds) Symposium Proceedings of the Society for the Study of human Biol pp 171-195.

Martinson J., Boyce A., And Clegg J. (1994) VNTR alleles associated with the α-globin locus are haplotype and population related. Am. J. Hum Genet. 55:513-525.

Martinson J., Excoffier L., Swinburn C., Boyce A., Harding R., Langaney A. And Clegg J. (1995) High diversity of α-globin haplotypes in a Senegalese population, including many previously unreported variants. Am. J. Hum Genet. 57:1186-1198.

Mears J., Ramirez F., Leibowitz D., Nakamura F., Bloom A., Konotey-Ahulu F. and Bank A. (1978) Changes in restricted human cellular DNA fragments containing globin gene sequences in thalassaemias and related disorders. Proc. Natl. Acad. Sci. 75:1222-1226.

Michelson A. And Orkin S. (1983) Boundaries of gene conversion within the duplicated human α-globin genes. J. Biol. Chem. 258:15245-15254.

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

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

Nagel R. and Ranney H. (1990) Genetic epidemiology of structural mutations of the β-globin gene. Semin. in Haem. 27:342-359.

Nicholls R., Jonasson J., McGee J., Patel S., Ionasescus V., Weatherall D. and Higgs D. (1987) High resolution gene mapping of the human α-globin locus. J Med Gen 24:39-46.

O'Shaughnessy d., Hill A., Bowden D., Weatherall D. And Clegg J. (1990) Globin genes in Micronesia: Origins and affinities of Pacific Island Peoples. Am. J. Hum. Gen. 46:144-155.

Oliver R. and Atmore A. (1981a) The African Middle Ages 1400 - 1800. Cambridge University Press, New York.

Oliver R. and Atmore A. (1981b) Africa since 1800. Cambridge University Press, New York.

Orkin S., Alter B., Altay C., Mahoney M., Lazarus H., Hobbins J. and Nathan D. (1978) Applications of endonuclease mapping to the analysis and prenatal diagnosis of thalassaemias caused by globin gene deletion. New Engl. J. Med. 299:166-171.

Orkin S., Kazazian H., Antonarakis S., Goff S., Boehm C., Sexton J., Waber P. and Giardina P. (1982) Linkage of b-thalassaemia mutations and β-globin gene

polymorphisms with DNA polymorphisms in human β-globin gene cluster. Nature 296:627-631.

Orkin S. and Kazazian H. (1984) The mutation and polymorphism of the human β-globin gene and its surrounding DNA. Ann. Rev. Genet. 18:131-171

Pauling L., Itano H., Singer S. and Wells I. (1949) Sickle cell anaemia, a molecular disease. Science 110:543-548.

Perichon B, Ragusa A., Lapoumeroulie C., Romand A., Moi P., Ikuta T., Labie D., Elion J. and Krishnamoorthy R. (1993) Inter-ethnic polymorphism of the β-globin gene locus control region (LCR) in sickle cell anaaemia patients. Hum Genet. 91:464-468.

Poncz M., Schwartz E., Ballantine M. and Surrey S. (1983) Nucleotide sequence analysis of the β-globin gene region in humans. J. Biol. Chem. 258:11599-11609.

Proudfoot N., Shander M., Manley J., Gefter M. And Maniatis T. (1980) Structure and in vitro transcription of human globin genes. Science 209:1329-1336.

Ramsay M. and Jenkins T. (1987) Globin-gene associated restriction-fragment-length polymorphisms in southern African peoples. Am. J. Hum. Genet. 41:1132-1144.

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

Reeders S., Breuning M., Davies K., Nicholls R., Jarman A., Higgs D., Pearson P. and Weatherall D. (1985) A highly polymorphic DNA marker linked to adult polycystic kidney disease on chromosome 16. Nature 317:542.

Roberts-Thomson J., Martinson J., Norwich J., Harding R., Clegg J. And Boettcher B. (1996) An ancient common origin of Aboriginal Australians and New Guinea Highlanders is supported by α-globin haplotype analysis. Am. J, Hum Genet. 58:1017-1024.

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

Saiki R., Gelfand D., Stoffel S., Scharf S., Higuchi R., Horn G., Mullis K. and Erlich H. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487-491.

Sambrook J., Fritsch E. and Maniatis T. (1989) Molecular Cloning: A laboratory manual. Cold Spring Harbor Laboratory Press. Appendix E5.

Semenza G., Malladi P., Surrey S., Delgrosso K., Poncz M. and Schwartz E. (1984) Detection of a novel DNA polymorphism in the β-globin gene cluster. J. Biol. Chem. 259:6045-6048.

Semenza G., Dowling C. And Kazazian H. (1989) HinfI polymorphism 3' to the human β-globin bene detected by the polymerase chain reaction (PCR). Nuc. Acids Res. 17(6):2376.

Singer R., Budtz-Olsen O. and Brain P. (1957) Physical features, sickling and Serology of the Malagasy of Madagascar. Am. J Phy. Anthr \p. 15:91-124.

Soodyall H., Jenkins T. And Stoneking M. (1993) 'Polynesian' mtDNA in the Malagasy. Nature Gen. 10:377-378.

Southern E.M. (1975) Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol 98:503-517.

Stoneking M., Fontius J., Clifford S., Soodyall H., Arcot S., Saha N., Jenkins T., Tahir M., Deininger P. And Batzer M. (1997) Alu insertion polymorphisms and human evolution: Evidence for a larger population size in Africa. Genome Res. (In Press).

Varawalla N., Fitches A. And Old J. (1992) Analysis of β-globin gene haplotypes in Asian Indians: origin and spread of β-thalassemia on the Indian subcontinent. Hum. Genet. 90:443-449.

Wainscoat J., Hill A., Boyce A., Flint J., Hernandez M., Thein S., Old J., Lynch J., Falusi A., Weatherall D. and Clegg J. (1986) Evolutionary relationships of human populations from an analysis of nuclear DNA polymorphisms. Nature 319:491-493.

Weatherall D., Higgs D., Wood W. and Clegg J (1984) Genetic disorders of haemoglobin as models for analysing gene regulation. Phil. Trans. R. Soc. Lond. B. 307:247-259.

Winichagoon P., Higgs D., Goodbourn S., Clegg J. And Weatherall D. (1982) Multiple arrangements of the human embryonic ξ-globin genes. Nuc. Acids Res. 10:5853.

Zago M., Santos., Clegg J., Guerreiro J., Martinson J.,Norwich J. And Figueiredo M. (1995) Alpha-globin gene haplotypes in South American Indians. Hum. Biol. 67(4):535-546.

APPENDIX 1

SOLUTIONS AND REAGENTS

10X α-GLOBIN PCR HAPLOTYPE BUFFER

	(Final Concentration/100μl reaction)
1M Tris-HCl (pH 8.8 at 20°C)	66.7mM
1M $(NH_4)SO_4$	16.7mM
1M $MgCl_2$	1.0-3.0mM
0.1 M Na_2EDTA (pH 8.0 at 20°C)	66.7μM
0.2% Gelatin	0.01%
DMSO	10.0%
β-mercaptoethanol	140mM

10X DEOXYNUCLEOTIDES (10X dNTPs)

125μl 10mM dATP
125μl 10mM dCTP
125μl 10mM dGTP
125μl 10mM dTTP
Make up to 1ml with double-distilled water. Use 1/10 volume per PCR reaction.

EDTA

0.5M stock, pH 8.0

ETHIDIUM BROMIDE

10mg/ml stock. Store at 4°C, in a dark container.

FICOLL DYE

50% sucrose
50mM EDTA (pH 7.0)
0.1% bromophenol blue dye
10% Ficoll

1κb λ LADDER

109ml λ ladder (0.25mg/ml stock)
50ml Ficoll dye
840ml 1X TE

SATURATED NaCl

40 mg NaCl
100 ml sterile water

PROTEINASE K MIX (Simple salting out procedure)

1% SDS
2 mM EDTA
2 mg/ml Proteinase K (10mg/ml stock)

PROTEINASE K SOLUTION

10mg/ml stock. Make up with sterile water and store at -20°C.

SUCROSE-TRITON X LYSING BUFFER

10 ml 1M Tris-HCl pH 8
5 ml 1M $MgCl_2$
10 ml Triton X 100

Make up to 1l with distilled water. Autoclave. Keep solution at 4°. Add 109,5 g sucrose just before use. The solution cannot be stored with sucrose.

T20E5

20 mM Tris-HCl
5 mM EDTA

10X TBE BUFFER (pH 8.3)

108g Tris Base
55g Boric acid
7.44g EDTA

Make up to 1 litre with distilled water and adjust pH to 8.3.

10X TE BUFFER (pH 8.0)

10mM Tris-HCl
1mM EDTA

TRIS-HCl
121.1g Tris-HCl
800 ml double distilled water
Adjust to pH 8.0 with Hcl and make up to 1l with distilled water.

Author Hewitt R
Name of thesis An Investigation Into The Ancestry Of The Malagasy Population Using Variation The A- And B-Globin Gene Clusters Hewitt R 1998

PUBLISHER:
University of the Witwatersrand, Johannesburg
©2013

LEGAL NOTICES:

Copyright Notice: All materials on the University of the Witwatersrand, Johannesburg Library website are protected by South African copyright law and may not be distributed, transmitted, displayed, or otherwise published in any format, without the prior written permission of the copyright owner.

Disclaimer and Terms of Use: Provided that you maintain all copyright and other notices contained therein, you may download material (one machine readable copy and one print copy per page) for your personal and/or educational non-commercial use only.

The University of the Witwatersrand, Johannesburg, is not responsible for any errors or omissions and excludes any and all liability for any errors in or omissions from the information on the Library website.