THE CHARACTERISATION OF GLUCOSE-6-PHOSPHATE DEHYDROGENASE (G6PD) VARIATION IN THE MALAGASY

BRUCE THORNLEY FRANK DANGERFIELD

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

1997

ABSTRACT

ABSTRACT

Much of the history of the peoples of Madagascar, the Malagasy, is unknown and uncertain. The human inhabitants of Madagascar show a large amount of biological diversity. Studies of physical appearance, linguistics, ethnology, archaeology and history have implicated the African, Indonesian, Arabian, Indian and Chinese populations in contributing to the present-day diversity of the people of Madagascar. The major contributors to the inhabitants of Madagascar appear to be the African and Indonesian populations. Large amounts of African admixture are said to be present in the lowland ethnic groups and large amounts of Indonesian admixture in the highland groups. Few genetic studies have been performed to characterise the Malagasy and quantitate the proportions of genetic contributions to the diverse gene pool. Studies of genetic contributions could help to elucidate the origins and affinities of the people of Madagascar. A good candidate marker for this type of genetic study is the G6PD locus.

G6PD deficiency is a potentially serious condition in which haemolytic episodes may occur in an affected individual, usually initiated by an environmental trigger, such as ingestion of fava beans or infection. Symptoms vary in severity and include neonatal jaundice and chronic, non-spherocytic haemolytic anaemia. Approximately 400 million individuals world-wide are estimated to have a deficiency gene. G6PD has a distinct geographic distribution, with high frequencies in tropical areas due to a selective advantage conferred against *Plasmodium falciparum* malaria. Over 400 biochemical variants (over 60 characterised at the molecular level) in the X-linked G6PD gene have been characterised, many of them causing deficiency. Several of the variants have been found at polymorphic frequencies in different populations and these have been shown to display geographical specificity. Several neutral polymorphisms have been found within the G6PD gene exclusively in particular population groups (notably sub-Saharan Africans). The current study aimed to characterise the Malagasy people using a combination of deficient and non-deficient variants detected by RFLP analysis.

The frequency of G6PD deficiency in Malagasy males ranged from 0.073 in the Antandroy group to 0.313 in the Makua group. Higher frequencies tended to occur in lowland ethnic groups than in highland ethnic groups. Both the G6PD A^- and G6PD Mediterranean deficiency variants were found in the Malagasy samples studied. Most deficiency was attributed to the G6PD A^- variant (68 of 78 deficient males). The Mediterranean variant occurred in 9 of 78 deficient males and the variant occurring in the remaining one deficient male was unidentified. G6PD A^- is common in Africa, but rare in other parts of the world. The G6PD Mediterranean variant appears to have arisen at least twice and there are thus two associated haplotypes. The haplotype common in the Middle East and found at a low frequency in Indonesia was found in 8/9 $G6PD^{Mediterranean}$ deficient Malagasy males and the haplotype common in India, which is also found at a high frequency in Indonesia, occurred in one of 9.

Samples from several possible "parental" populations were obtained. The African populations studied were the Zulus, Tsonga, Zambians and a group from the Central African Republic. Indonesian and Polynesian populations were also studied. A sample of randomly selected, unrelated male individuals (and female individuals where sample sizes were small) was selected from each of 6 Malagasy ethnic groups from different geographical regions: 63 Merina (MR) individuals and 54 Betsileo (BT) individuals (from the central highland); 60 Antesaka (AS) individuals (a southeastern lowland group); 56 Tsimihety (TS) individuals (a northern lowland group); and 26 Vezo (VZ) and 21 Mahafaly (MF) individuals (lowland groups in the southwest). Seven-site RFLP haplotypes of the deficient and non-deficient variants were constructed from results obtained on these samples. These were used to estimate the degree of association between ethnic groups and between ethnic groups and their ancestral populations. The analytical methods employed were: the calculation of haplotype frequency distributions, χ^2 analysis testing differences between subgroups within the populations and between different populations, calculation of genetic distances, construction of phylogenetic trees, calculation and plotting of the principal components, and admixture estimates.

The results obtained indicated strong affiliations between the highland groups and the Austronesian (Indonesian and Polynesian) populations. Genetic distances between the Austronesian populations were small. A strong affinity was observed between the Antesaka and Tsimihety lowland groups and the African populations, particularly with the Zambian population. The Vezo and Mahafaly groups, however, associated with the Austronesian populations, contrary to expectations for lowland groups. Admixture calculations showed a fairly high (50%) proportion of Indonesian admixture in the highland groups, a low (15%) proportion of admixture in the lowland group, and a very high (81%) proportion in the Vezo and Mahafaly groups. In the lowland group, the contribution from a Zambian-like population is 75% and a Tsonga-like population contributed 11%. In the highland group, the Zambian-like population contributed 27% and the Tsonga-like population contributed 22%. In the Vezo and Mahafaly groups, the Tsonga-like population contributed 16% and the Zambian-like population contribution was small or absent. While it is possible that two or more African groups contributed substantial proportions of genes to the Malagasy, an alternative interpretation concordant with these data is that a single parental population, with characteristics of both the present-day Zambian and Tsonga populations was responsible for the majority of the African contribution to the Malagasy gene pool. A good candidate for this single putative parental population, by virtue of its location, is the population of Mozambique. Different selective pressures acting in the lowland and highland areas, due to different levels of malaria endemicity, could be responsible for the apparently different proportions of the Zambian-like and Tsonga-like haplotypes present in the hybrid Malagasy groups.

This study has provided some data on the different contributions to Malagasy ethnic groups. Together with other loci and non-genetic evidence, it is hoped that it will contribute to a comprehensive picture of the origins of the Malagasy.

DECLARATION

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

.....

Bruce Thornley Frank Dangerfield

17th day of October 1997

For Britta, for reasons that would mean nothing to others but mean everything to me.

ACKNOWLEDGEMENTS

I would sincerely like to thank the following people and organisations:

My supervisors Dr Amanda Krause and Professor Trefor Jenkins for enabling me to carry out this project and for their time, support and patience.

All of the individuals who kindly donated blood.

All of those individuals who assisted in blood sample collection, especially Prof. T Jenkins, Dr T de Ravel, Dr P Willem, Dr G Campbell, Dr A Rakouth, Dr S Quambek and Prof. J Roux.

Dr D Rees and Dr J Clegg of the MRC Molecular Haematology Unit of John Radcliffe Hospital in Oxford, for providing Indonesian DNA samples.

Dr DG Woodfield of the Department of Transfusion Medicine of the Auckland Regional Blood Services in New Zealand, for providing the Polynesian DNA samples.

Dr Gwyn Campbell (University of the Witwatersrand) and Dr J Roux (Institute Pasteur) for providing sociological information on Madagascar.

Dr M Stoneking and Dr H Soodyall for advice on the statistical analysis.

The Haematology Department of the SAIMR, under the direction of Prof. B Mendelow for performing the haematological testing used in this project.

The Serogenetics Laboratory, under the direction of Dr A Lane for performing starch gel electrophoresis on the samples used in this project.

The SAIMR, the University of the Witwatersrand, the Medical Research Council and the National Geographic for financial support and laboratory facilities.

All of the people in the laboratories of the Human Genetics Department who were always available with time and help when I needed it, especially: Michéle Ramsay, Tony Lane, Thomy de Ravel, Prashiela Manga, Bharti Morar, Rachel Hewitt, Michelle Starfield, Andy Goldman, Gwynn Stevens, Fahmida Essop, Robyn Kerr, Premilla John, Anton Bergheim, Carolyn Padoa and Melanie Israelstam.

TABLE OF CONTENTS

Abstract ... ii
Declaration ... v
Dedication ... vi
Acknowledgements ... vii
Table of Contents ... viii
List of Abbreviations and Symbols ... xiii
List of Tables ... xv
List of Figures ... xvii

1. INTRODUCTION ... 1

1.1 MADAGASCAR ... 3
1.1.1 GEOGRAPHY AND NATURAL HISTORY OF MADAGASCAR ... 3
1.1.2 THE PEOPLES OF MADAGASCAR ... 5
1.1.2.1 ETHNIC CLASSIFICATION OF THE MALAGASY ... 5
1.1.2.2 THEORIES OF THE ORIGINS OF THE MALAGASY ... 7
1.1.2.3 EVIDENCE INDICATING THE ORIGINS OF THE MALAGASY ... 9
1.1.2.3.1 THE MALAGASY LANGUAGE ... 9
1.1.2.3.2 ARCHAEOLOGICAL FINDS IN MADAGASCAR ... 10
1.1.2.3.3 CULTURE AND ETHNOLOGY OF THE MALAGASY ... 11
1.1.2.3.4 HISTORICAL EVIDENCE OF MALAGASY ANCESTRY ... 13
1.1.2.3.5 PHYSICAL APPEARANCE OF THE MALAGASY ... 14
1.1.2.3.6 GENETIC STUDIES OF MALAGASY PEOPLES ... 15
1.1.2.3.7 SUMMARY OF MALAGASY ANCESTRY ... 17
1.2 GLUCOSE-6-PHOSPHATE DEHYDROGENASE ... 18
1.2.1 CLASSIFICATION OF G6PD DEFICIENCY ... 19
1.2.2 CLINICAL CONSIDERATIONS OF G6PD DEFICIENCY ... 20
1.2.2.1 HISTORY OF THE DISEASE ... 20
1.2.2.2 SIGNS AND SYMPTOMS OF G6PD DEFICIENCY ... 21
1.2.2.3 PREVENTION, SCREENING AND TREATMENT OF G6PD DEFICIENCY ... 22
1.2.3 GEOGRAPHIC DISTRIBUTION OF G6PD DEFICIENCY ... 23
1.2.3.1 SELECTIVE ADVANTAGE AGAINST MALARIA ... 23
1.2.4 THE G6PD PROTEIN ... 27
1.2.4.1 STRUCTURE OF THE PROTEIN ... 27
1.2.4.2 METABOLIC FUNCTION OF THE PROTEIN ... 28
1.2.5 THE G6PD GENE ... 30
1.2.5.1 LOCATION OF THE G6PD GENE ... 30
1.2.5.2 STRUCTURE OF THE GENE ... 30

1.2.6 VARIANTS WITHIN THE G6PD GENE *32*
1.2.6.1 SPORADIC MUTATIONS AND POLYMORPHISMS 32
1.2.6.2 MOLECULAR ANALYSIS 33
1.2.6.3 NUMBER AND GEOGRAPHIC DISTRIBUTION OF G6PD VARIANTS 35
1.2.6.3.1 AFRICAN VARIANTS *35*
1.2.6.3.1.1 THE A VARIANT 35
1.2.6.3.1.2 THE A⁻ DEFICIENCY VARIANT 36
1.2.6.3.1.3 "SILENT" POLYMORPHISMS 37
1.2.6.3.1.4 POLYMORPHISMS IN INTRONS 37
1.2.6.3.1.5 HAPLOTYPE CONSTRUCTION 37
1.2.6.3.2 MEDITERRANEAN AND EURASIAN VARIANTS *40*
1.2.6.3.2.1 G6PD MEDITERRANEAN 40
1.2.6.3.3 SOUTHEAST ASIAN VARIANTS *41*
1.2.7 G6PD VARIANTS IN POPULATION STUDIES *42*
1.3 AIMS OF THE STUDY 44

2. SUBJECTS AND METHODS 45

2.1 SUBJECTS 46
2.1.1 MALAGASY INDIVIDUALS *46*
2.1.1.1 RANDOMLY SELECTED INDIVIDUALS FOR DNA STUDIES 46
2.1.1.2 G6PD DEFICIENT INDIVIDUALS 47
2.1.2 AFRICAN INDIVIDUALS *48*
2.1.2.1 ZAMBIAN SAMPLES 48
2.1.2.2 CENTRAL AFRICAN REPUBLIC (CAR) SAMPLES 48
2.1.2.3 TSONGA SAMPLES 49
2.1.2.4 ZULU SAMPLES 49
2.1.3 AUSTRONESIAN INDIVIDUALS *50*
2.1.3.1 INDONESIAN SAMPLES 50
2.1.3.2 POLYNESIAN SAMPLES 50
2.2 METHODS 51
2.2.1 PROCESSING OF BLOOD *51*
2.2.1.1 HAEMATOLOGICAL TESTING 51
2.2.1.2 STARCH GEL ELECTROPHORESIS 51
2.2.1.3 DNA EXTRACTION 52
2.2.2 VARIANTS STUDIED *53*
2.2.2.1 ELECTROPHORETIC MOBILITY VARIANTS 53
2.2.2.2 DEFICIENCY VARIANTS 54
2.2.2.3 POLYMORPHISMS 54
2.2.2.4 PCR AMPLIFICATION 57
2.2.2.5 PREPARATION OF GELS 59

2.2.2.6 VISUALISATION OF PCR PRODUCTS 60
2.2.2.7 RESTRICTION ENZYME DIGESTION 60
2.2.2.8 VISUALISATION OF DIGESTED PRODUCTS 60
2.2.3 HAPLOTYPE CONSTRUCTION *64*
2.2.4 STATISTICAL ANALYSIS OF THE DATA *65*
2.2.4.1 FREQUENCY DISTRIBUTIONS 65
2.2.4.1.1 ENZYME DEFICIENCY FREQUENCY *65*
2.2.4.1.2 ALLELE FREQUENCIES *66*
2.2.4.1.2.1 LINKAGE DISEQUILIBRIUM 66
2.2.4.1.3 HAPLOTYPE FREQUENCIES *67*
2.2.4.1.3.1 HAPLOTYPE FREQUENCY HISTOGRAMS 67
2.2.4.2 CHI-SQUARE χ^2 ANALYSIS 68
2.2.4.3 CALCULATION OF GENETIC DISTANCES 69
2.2.4.4 TREE CONSTRUCTION FROM GENETIC DISTANCES 70
2.2.4.5 PRINCIPAL COMPONENT ANALYSIS 71
2.2.4.6 ADMIXTURE CALCULATIONS 73
2.2.4.6.1 THE BERNSTEIN METHOD *73*
2.2.4.6.2 ROBERT AND HIORNS' LEAST SQUARES METHOD *74*
2.2.4.6.3 ELSTON'S LEAST SQUARES METHOD *75*

3. RESULTS AND DISCUSSION 76

3.1 FREQUENCY DISTRIBUTIONS 76
3.1.1 G6PD ENZYME DEFICIENCY FREQUENCY *76*
3.1.1.1 ENZYME DEFICIENCY IN AFRICAN POPULATIONS 77
3.1.1.1.1 CAR POPULATION ENZYME DEFICIENCY FREQUENCY *77*
3.1.1.1.2 ENZYME DEFICIENCY FREQUENCIES IN OTHER AFRICAN POPULATIONS *78*
3.1.1.2 ENZYME DEFICIENCY IN AUSTRONESIAN POPULATIONS 79
3.1.1.3 ENZYME DEFICIENCY IN THE MALAGASY ETHNIC GROUPS 80
3.1.2 ALLELE FREQUENCIES FOR THE 7 RFLP SITES *83*
3.1.2.1 DEFICIENCY VARIANT ALLELE FREQUENCIES 86
3.1.2.1.1 MBO II SITE (MEDITERRANEAN DEFICIENCY VARIANT) FREQUENCY *88*
3.1.2.1.2 NLA III SITE (A⁻ DEFICIENCY VARIANT) FREQUENCY *89*
3.1.2.1.3 COMPARISON OF MOLECULAR AND HAEMATOLOGICAL METHODS *92*
3.1.2.2 MOBILITY VARIANTS 93
3.1.2.2.1 THE FOK I SITE (THE A MOBILITY VARIANT) *93*
3.1.2.3 ALLELE FREQUENCIES OF NEUTRAL POLYMORPHISMS 96
3.1.2.4 LINKAGE DISEQUILIBRIUM 98

3.1.3 HAPLOTYPE FREQUENCIES *102*
3.1.3.1 HAPLOTYPE FREQUENCY HISTOGRAMS 107
3.2 CHI SQUARE (χ^2) ANALYSIS 113
3.2.1 χ^2 COMPARISON OF HAPLOTYPES WITHIN POPULATIONS *114*
3.2.1.1 χ^2 ANALYSIS BETWEEN MALAGASY ETHNIC GROUPS 114
3.2.1.2 χ^2 ANALYSIS BETWEEN CAR GROUPS 116
3.2.1.3 χ^2 ANALYSIS BETWEEN INDONESIAN GROUPS 118
3.2.2 χ^2 COMPARISON OF HAPLOTYPES BETWEEN POPULATIONS. *120*
3.2.2.1 χ^2 COMPARISONS BETWEEN THE AFRICAN POPULATIONS 120
3.2.2.2 χ^2 COMPARISONS BETWEEN NON-AFRICAN POPULATIONS ... 120
3.2.2.3 χ^2 COMPARISON OF MALAGASY POPULATIONS TO OTHER POPULATIONS 122
3.2.2.3.1 GENERAL TRENDS *122*
3.2.2.3.2 χ^2 COMPARISON OF HIGHLAND GROUPS TO OTHER POPULATIONS *123*
3.2.2.3.3 χ^2 COMPARISON OF LOWLAND GROUPS TO OTHER POPULATIONS *124*
3.2.2.3.4 χ^2 COMPARISON OF VEZO AND MAHAFALY GROUPS TO OTHER POPULATIONS *125*
3.3 GENETIC DISTANCES 126
3.3.1 GENETIC DISTANCES BETWEEN AFRICAN POPULATIONS *126*
3.3.2 GENETIC DISTANCES BETWEEN NON-AFRICAN POPULATIONS *126*
3.3.3 GENETIC DISTANCES BETWEEN MALAGASY AND NON-MALAGASY POPULATIONS *128*
3.4 PHYLOGENETIC TREES 130
3.5 PRINCIPAL COMPONENTS (PC) ANALYSIS 136
3.5.1 SINGULAR VALUES OF PC ANALYSIS *136*
3.5.2 CARTESIAN PLOTS OF PC ANALYSIS *137*
3.6 ADMIXTURE CALCULATIONS 141
3.6.1 THE BERNSTEIN METHOD OF ADMIXTURE CALCULATION *142*
3.6.2 ROBERTS AND HIORNS LEAST SQUARES ADMIXTURE ESTIMATE *145*
3.6.3 THE LEAST SQUARES ADMIXTURE ESTIMATION OF ELSTON. *148*
3.6.4 EVALUATION OF THE THREE ADMIXTURE METHODS *150*
3.6.4.1 PROPORTION OF INDONESIAN ADMIXTURE 150
3.6.4.2 PROPORTION OF AFRICAN ADMIXTURE 151

4. GENERAL DISCUSSION 153

4.1 G6PD DEFICIENCY SCREEN 153

4.1.1 ENZYME DEFICIENCY IN THE MALAGASY 153

4.1.2 DEFICIENCY VARIANTS FOUND IN MADAGASCAR 158

4.2 HAPLOTYPE ANALYSIS IN THE RANDOMLY SELECTED MALAGASY 161

4.2.1 CHARACTERISATION OF PUTATIVE ANCESTRAL POPULATIONS 162

4.2.1.1 HAPLOTYPE ANALYSIS IN AFRICAN POPULATIONS 162

4.2.1.2 HAPLOTYPE ANALYSIS IN AUSTRONESIAN POPULATIONS 163

4.2.2 CHARACTERISATION OF THE MALAGASY ETHNIC GROUPS 164

4.2.2.1 HAPLOTYPE ANALYSIS OF THE HIGHLAND GROUPS 164

4.2.2.2 HAPLOTYPE ANALYSIS OF THE LOWLAND GROUPS 165

4.2.2.3 HAPLOTYPE ANALYSIS OF THE SOUTHWESTERN MALAGASY GROUPS 166

4.3 ANCESTRAL POPULATION ADMIXTURE IN THE MALAGASY 167

5. CONCLUSIONS 170

6. REFERENCES 173

APPENDIX A - SOLUTIONS AND REAGENTS 183

APPENDIX B - CALCULATIONS USED IN THE ANALYSIS 186

LIST OF ABBREVIATIONS AND SYMBOLS

λ = Lambda
χ^2 **test** = Chi Square test
°C = Degrees Celsius
µg = Microgram
µl = Microlitre
AB = Antambahoaka (Malagasy ethnic group)
AD = Antandroy (Malagasy ethnic group)
AF = Antefasy (Malagasy ethnic group)
AK = Antankarana (Malagasy ethnic group)
AM = Antemoro (Malagasy ethnic group)
AN = Antanosy (Malagasy ethnic group)
AS = Antesaka (Malagasy ethnic group)
ASOH = Allele Specific Oligonucleotide Hybridisation
BA = Bara (Malagasy ethnic group)
bp = Base pairs
BS = Betsimisaraka (Malagasy ethnic group)
BSA = Bovine serum albumin
BT = Betsileo (Malagasy ethnic group)
BZ = Bezanozano (Malagasy ethnic group)
CAR = Central African Republic
CAR(grp) = CAR(MSG) = Combined Mbindu (M), Sangha Sangha (S) and Gbayo (G) language groups of the CAR
cm = Centimetre
cM = Centimorgan
CV = Colour vision
ddH_2O = Deionised distilled water
dNTP = Deoxyribonucleotide triphosphate
EDTA = Ethylenediamine Tetra-Acetic Acid
EtBr = Ethidium Bromide
FIX = Factor IX
HPRT = Hypoxanthine phosphoribosyltransferase
FRAX = Fragile X
FVIII = Factor VIII
g = Gram
G6PD = Glucose-6-phosphate dehydrogenase
HTF = *Hpa* II tiny fragments
IND(CS) = Indonesian smple from Central Sulawesi
IND(grp) = Grouped Indonesian sample (i.e. IND(SK), IND(CS) and IND(SS))
IND(SK) = Indonesian sample from South Kalimantan
IND(SS) = Indonesian sample from South Sulawesi
kb = Kilobase
kDa = Kilodalton
K_m = Michaelis constant
Mb = Megabases
MF = Mahafaly (Malagasy ethnic group)
mg = Milligram

MK = Makua (Malagasy ethnic group)
ml = Millilitre
mM = Millimolar
MR = Merina (Malagasy ethnic group)
MS = Masikoro (Malagasy ethnic group)
NADP = Nicotinamide-adenine dinucleotide phosphate
NADPH = Reduced Nicotinamide-adenine dinucleotide phosphate
NJ = Neighbour joining
PC = Principal components
PCR = Polymerase chain reaction
POL = Polynesian population
RE = Restriction enzyme
RFLP = Restriction Fragment Length Polymorphism
SDS = Sodium Dodecyl Sulphate
SE = Standard error
SH = Sihanaka (Malagasy ethnic group)
SK = Sakalava (Malagasy ethnic group)
SSCP = Single Stranded Conformational Polymorphism
TBE = Tris- Boric acid- EDTA
TE = Tris-EDTA
TN = Tanala (Malagasy ethnic group)
Tris = Tris-(hydroxymethyl) amino methane
TS = Tsimihety (Malagasy ethnic group)
u = Units
UPGMA = Unweighted Paired-Group method using arithmetic Averages
V = Volt
VZ = Vezo (Malagasy ethnic group)
VZ&MF = **VZ/MF** = Vezo and Mahafaly combined Malagasy ethnic groups
WHO = World Health Organisation

List of Tables

Table 2-1: Mutation locations, PCR primer sequences and the enzymes used to cut the fragments used in analysis of the G6PD locus 55

Table 2-2: Annealing temperatures (T°) used for the Cetus and Omnigene thermocyclers for the 7 RFLP systems studied at the G6PD locus and the expected sizes of fragments before and after digestion with the appropriate restriction enzyme 58

Table 3-1: G6PD deficiency frequencies according to region and ethnic group for Malagasy males determined by haematological screening 81

Table 3-2: Allele frequencies for the 7 G6PD systems in the random Malagasy groups 84

Table 3-3: Allele frequencies for the 7 G6PD systems in the non-Malagasy populations 85

Table 3-4: Specific molecular deficiency variants in G6PD deficient Malagasy males by ethnic and regional grouping 87

Table 3-5: D values for comparison of the linkage disequilibrium figures found in this study (for the CAR group) compared with those found by Vulliamy *et al* (1991b) in an African sample 99

Table 3-6: Chi square and p values for the linkage disequilibrium calculations for the CAR group 100

Table 3-7: G6PD haplotype frequencies for the various Malagasy groups studied 103

Table 3-8: G6PD haplotype frequencies for the African populations studied 104

Table 3-9: G6PD haplotype frequencies for the Austronesian populations studied 105

Table 3-10: Chi square values, p values and degrees of freedom for the comparison of the haplotype distributions in the ethnic and regional groups within the Malagasy population 115

Table 3-11: Chi square values, p values and degrees of freedom for the comparison of the haplotype distributions of the subgroups within the CAR population 117

Table 3-12: Chi square values, p values and degrees of freedom for the comparison of the haplotype distributions of the regional groups within the Indonesian population 119

Table 3-13: Chi square values, degrees of freedom and p values for the comparison of the haplotype distributions between the Malagasy ethnic groups and the non-Malagasy population groups 121

Table 3-14: Nei's standard genetic distances (D) (calculated using the DISPAN program) between all the groups which formed the basis of the study 127

TABLE 3-15: ADMIXTURE (M) VALUES OF THE ZAMBIAN PARENTAL POPULATION IN THE HYBRID POPULATIONS TOTAL MALAGASY, HIGHLANDERS, LOWLANDERS AND COMBINED VEZO & MAHAFALY USING THE METHOD OF BERNSTEIN 144
TABLE 3-16: ADMIXTURE PROPORTION (M) VALUES FOR THE AFRICAN AND INDONESIAN PARENTAL POPULATIONS IN THE HYBRID POPULATIONS TOTAL MALAGASY, HIGHLANDERS, LOWLANDERS AND VEZO & MAHAFALY GROUPS USING THE METHOD OF ROBERTS AND HIORNS 146
TABLE 3-17: ADMIXTURE (M) VALUES OF THE PARENTAL POPULATIONS ZAMBIAN, TSONGA AND INDONESIA FOR THE HYBRID POPULATIONS TOTAL MALAGASY, HIGHLANDERS, LOWLANDERS AND VEZO & MAHAFALY USING THE ELSTON METHOD 149

List of Figures

Figure 1-1 - Relief map of Madagascar 4
Figure 1-2 - Map showing the distribution of 19 Malagasy ethnic groups 6
Figure 1-3 - Map showing medieval trade routes in the Indian Ocean and wind and ocean currents 8
Figure 1-4: Pictures of the typical rectangular houses and tombs found on the central plateau indicating Indonesian influence 12
Figure 1-5 - Map showing geographic distribution and frequencies of G6PD deficiency 24
Figure 1-6 - Geographic distribution of malaria in 1991 25
Figure 1-7 - The function of G6PD in the pentose phosphate pathway 29
Figure 1-8 - Location and structure of the G6PD gene 31
Figure 1-9 - Some variants found in the coding regions of the G6PD gene 34
Figure 1-10 - Haplotypes of the G6PD gene 38
Figure 1-11 - Postulated evolutionary sequence of the mutations in the G6PD gene based on the smallest number of mutational steps required to give the 7 haplotypes 39
Figure 1-12 - Map showing the geographical locations of some of the major deficiency variants and their relative contributions to deficiency in those areas 43
Figure 2-1: The map positions and expected haplotypes for the 7 variants studied in this project 56
Figure 2-2: Digitised photograph of the *Bcl I* digest run on 5% Metaphor gel 61
Figure 2-3: Digitised photograph of a sample of the *Nla III* digest run on 4% composite agarose gel 62
Figure 2-4: Digitised photograph of the *Mbo II* digest run on a 4% composite agarose gel 63
Figure 3-1: Map showing the distribution of the biochemical GD^{A-} deficiency variant in southern Africa 91
Figure 3-2: Map showing the distribution of the mobility variant GD^{A} in southern Africa 94
Figure 3-3: Histogram showing the haplotype frequencies at the G6PD locus of the African populations studied 108
Figure 3-4: Histogram showing the haplotype frequencies at the G6PD locus of the Austronesian populations studied 109
Figure 3-5: Histogram showing the haplotype frequencies at the G6PD locus of the Malagasy groups studied 110
Figure 3-6: Histogram showing the haplotype frequencies at the G6PD locus of all of the populations studied 111

FIGURE 3-7: NEIGHBOUR JOINING TREE DRAWN IN DISPAN, BASED ON NEI'S STANDARD GENETIC DISTANCES FOR THE UNGROUPED DATA 131
FIGURE 3-8: NEIGHBOUR JOINING TREE DRAWN IN DISPAN USING NEI'S STANDARD GENETIC DISTANCES FOR THE GROUPED DATA 133
FIGURE 3-9: NEIGHBOUR JOINING TREE DRAWN IN DISPAN, USING NEI'S STANDARD GENETIC DISTANCES, FOR THE B HAPLOTYPES (NUMBERS I, VII AND II) FOR THE UNGROUPED DATA 134
FIGURE 3-10: NEIGHBOUR JOINING TREE DRAWN WITH DISPAN USING NEI'S STANDARD GENETIC DISTANCES FOR THE B HAPLOTYPES (NUMBERS I, VII AND II) FOR THE GROUPED DATA 135
FIGURE 3-11: PLOT OF THE PRINCIPAL COMPONENTS OF THE G6PD HAPLOTYPE FREQUENCIES FOR ALL OF THE POPULATIONS STUDIED 138
FIGURE 3-12: PLOT OF THE PRINCIPAL COMPONENTS OF THE G6PD HAPLOTYPE FREQUENCIES FOR THE MALAGASY ETHNIC GROUPS STUDIED 140
FIGURE 4-1: PLOT OF SICKLE CELL DISTRIBUTION COMPARED TO G6PD DEFICIENCY DISTRIBUTION FOR THE 5 GEOGRAPHICAL REGIONS WITH A LINEAR REGRESSION LINE 156

CHAPTER ONE

INTRODUCTION

1. INTRODUCTION

The island of Madagascar is biologically diverse, both in its human and non-human inhabitants. Much of the history of the human colonisation of the island is unknown and the origins of the peoples of Madagascar, the Malagasy, are uncertain. This has been a subject of much debate over the years and this study forms a part of a growing body of evidence which is accumulating to uncover the secrets of the island's past. The present study involves molecular genetic analysis of the variants occurring within the G6PD gene, using this gene as a marker to investigate admixture from putative parental populations. The literature review which follows therefore comprises two parts: a review of the current state of knowledge of Madagascar and a review of the current knowledge of G6PD variation.

Madagascar is an island off the east coast of Africa. As G6PD deficiency confers a selective advantage against malaria (Allison 1960), the tropical climate of Madagascar is an important consideration in this study. The climatic conditions in different parts of the island are also important, since different selective pressures may be acting in different regions in the island. The geographic features may also lead to geographic isolation, increasing the amount of endogamy occurring within the ethnic groups.

Observations on the ancestry of the Malagasy are presented and the evidence used in making these observations. This evidence includes the physical appearance of the people, the language spoken, archaeological finds, historical evidence and general ethnology. Genetic studies of the Malagasy to date are limited but those that have been performed are discussed.

G6PD deficiency can be a severe and often life-threatening condition. G6PD deficiency shows geographical specificity and is usually found at high frequencies in tropical regions because G6PD deficient individuals have some natural resistance to malaria infection (Ruwende *et al* 1995). This geographic distribution can be useful in population studies such as this one. These geographically specific variants can be studied using molecular techniques. The gene location and protein structure is briefly described and the geographical specificity of several polymorphic deficiency and non-deficiency variants is discussed in detail.

The major deficiency variants discussed are the common Caucasoid deficiency variant G6PD Mediterranean and the common African deficiency variant A^-. The A^- variant occurs almost exclusively in Africa and is rare outside of Africa, while the Mediterranean variant is rare or absent in sub-Saharan Africa but occurs at high frequencies outside of Africa. Five other neutral polymorphic variants which have some geographic specificity are also discussed in this study. Together, the 7 polymorphisms can be used to construct haplotypes for maximum informativity in characterising populations and population movements.

The major aim of this study was to describe the G6PD variation present in the Malagasy people in order to identify, using the geographical distribution of these variants, the ancestral populations of the people of Madagascar and their relative contributions to the present-day Malagasy peoples.

1.1 MADAGASCAR

1.1.1 GEOGRAPHY AND NATURAL HISTORY OF MADAGASCAR

Madagascar is the fourth largest island in the world after Greenland, New Guinea and Borneo. A relief map of Madagascar is shown in Figure 1-1. Madagascar has a total surface area of 587 041 km^2, it is 1600km long and 500km wide. The island lies approximately 400 km off the coast of Africa in the Indian Ocean. A central plateau, partly volcanic, runs down the centre of the island (north/south) and is 2876m above sea level at its highest point (Cole 1992). To the east, the land drops sharply, often precipitously, down to sea level and is covered mainly with tropical forest. The west side of the island has more gentle slopes of savannah and plains.

Madagascar is tropical and is often hit by tropical cyclones. The coastal areas of the island are generally hot throughout the year, whereas the central plateau is more temperate. The eastern part of the island receives much rain, sometimes exceeding 3050mm p.a. The central plateau receives less moisture and the south-west regions may receive less than 380mm p.a (Encarta 1994). In the wetter lowland areas of Madagascar, *Plasmodium falciparum* malaria has been endemic, although malaria outbreaks in the usually malaria-free highlands have occurred in wetter seasons (Campbell 1991).

Madagascar is thought to have broken away from Africa at least 120 million years ago (Cole 1992). This long isolation has resulted in the development of many unique species of plants and animals. It is thought that the timing of the extinction of the larger indigenous animal species could give some indication of the time of the arrival of humans on the island (Burney and MacPhee 1988).

The people of Madagascar are called the Malagasy. Only 22% of the population is urban. The population is ethnically diverse and the two official languages are Malagasy and French. Although the recent history of Madagascar is well established, much of the ancient history is uncertain and speculative.

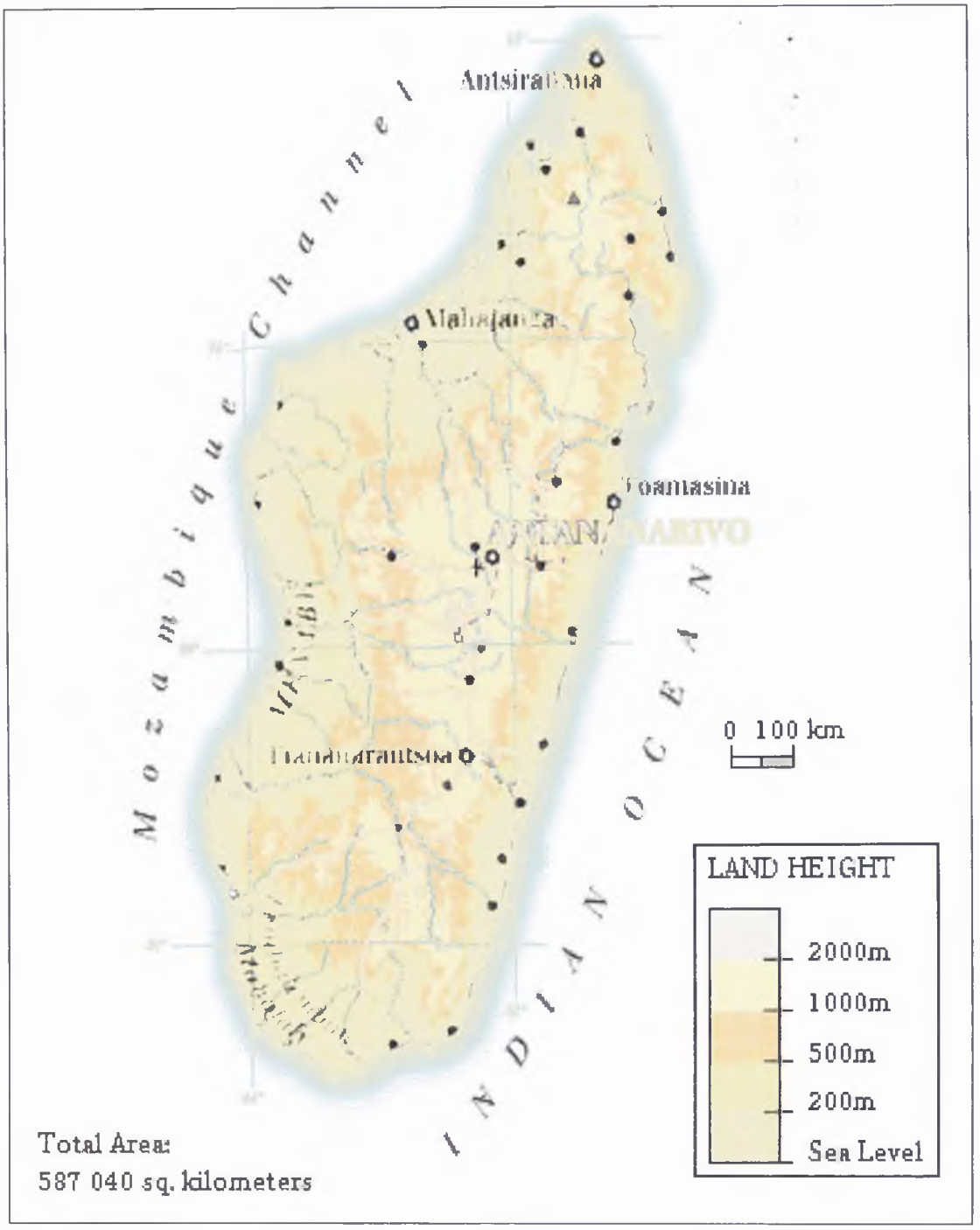


Figure 1-1 - Relief map of Madagascar (after Dorling Kindersley 1995)

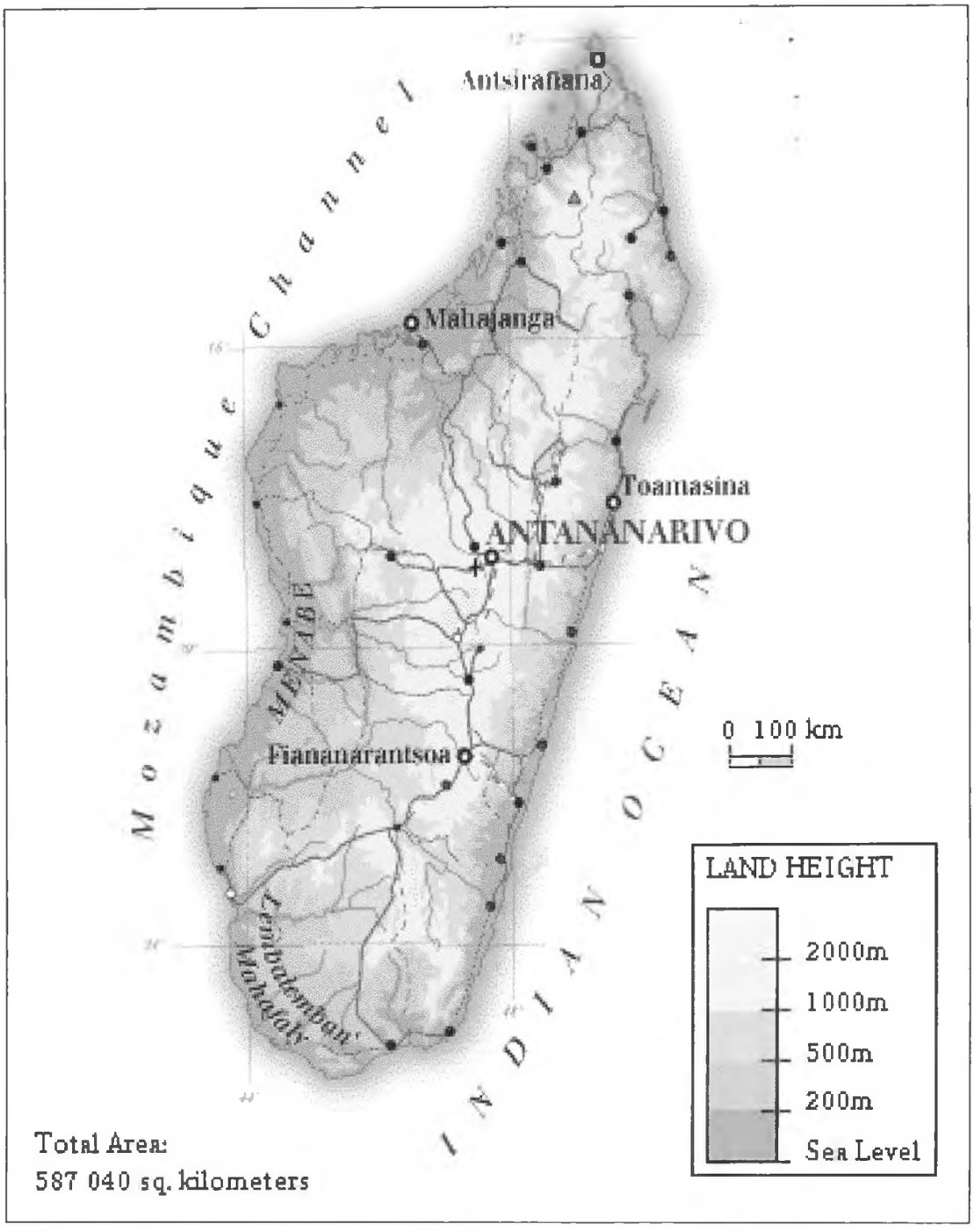


Figure 1-1 - Relief map of Madagascar (after Dorling Kindersley 1995)

1.1.2 THE PEOPLES OF MADAGASCAR

Much of the available information concerning the origins of the people of Madagascar is anecdotal or found in journals which are not easily available, often in French. Comprehensive reviews are available (Brown 1978, Dewar and Wright 1993) and these have been used in the sections that follow.

1.1.2.1 ETHNIC CLASSIFICATION OF THE MALAGASY

The fact that there is a single term for the people of Madagascar, the Malagasy, is misleading since there is much ethnic diversity on the island. Figure 1-2 shows the locations of 19 loosely defined ethnic groups. The ethnic groups may be subdivided into the highland peoples (those living on the plateau, specifically the Merina and Betsileo groups) and the lowland (coastal) peoples (Brown 1978).

The ethnic groups can also be classified geographically as follows:

North: Antankarana (AK), Tsimihety (TS), Sihanaka (SH)

Central: Merina (MR), Betsileo (BT), Bezanozano (BZ)

East-Central: Betsimisaraka (BS), Antemoro (AM), Antefasy (AF), Antambahoaka (AB), Antesaka (AS), Tanala (TN), Antanosy (AN)

Southwest: Antandroy (AD), Bara (BA), Vezo (VZ) Mahafaly (MF), Masikoro (MS)

West: Sakalava (SK)

The Merina can be subdivided into three groups or castes: Andriana ("nobles" or caste I), Hova ("commoners" or caste II) and Mainti (slaves or caste III) (Singer *et al* 1957). The slave and commoner castes are thought to have a greater degree of African admixture than the noble class which is thought to have high proportions of Indonesian admixture. Not mentioned above or shown on the map is the western lowland group Makua (MK). The Makua is not an official ethnic group and is thought to originate mainly from slave peoples arriving on the island in the 19th century (Campbell, The Origins of the Malagasy, manuscript in preparation).

Figure 1-2 - Map showing the distribution of 19 Malagasy ethnic groups (Brown 1978)

1.1.2.2 THEORIES OF THE ORIGINS OF THE MALAGASY

There are two major theories regarding the origins of the Malagasy. The first was proposed by Grandidier after the French occupation in 1895 and suggested that the peoples of Madagascar arose exclusively from people of eastern origin. It was proposed that the "Malay-like" highland peoples arose primarily from Indonesian peoples and that the "Negroid-like" lowland peoples arose primarily from Melanesian peoples. It was thought that these eastern populations crossed the Indian Ocean in outrigger canoes as early as the sixth century BC (reviewed by Campbell 1996). The single migration theory implies a migration of a substantial number of Indonesians across the Indian Ocean. While this is possible given the wind and ocean currents (shown in Figure 1-3), it seems improbable for the time period suggested.

A more likely explanation was proposed by Ferrand, who suggested that the peoples of Madagascar originated from both African and Indonesian populations and that the Indonesian people followed trade routes along the northern periphery of the Indian Ocean (shown in Figure 1-3). It was further suggested that the Indonesians set up trading posts on the east coast of Africa, establishing a proto-Malagasy colony (reviewed by Campbell 1996). Due to increasing pressure from African expansion, this colony was gradually forced on to the island in successive waves (Brown 1978). These waves of people would be expected to have had different proportions of African and Indonesian admixture. The first settlers could be expected to be most Indonesian-like and subsequent waves of colonists would have been more African-like. It is suggested that the initial colonists on Madagascar were forced further and further inland by the population expansion and the result was that the first (Indonesian-like) settlers ended up on the plateau with the more African-like peoples inhabiting the lowland.

Diversity in the island was further compounded by cultural, political and religious influences. The ruling classes (i.e. the Merina) tended to have a rigid endogamy, while the other groups tended to be less discriminating. The introduction of slaves from east Africa added additional African genes to the already diverse gene pool (Campbell 1991).

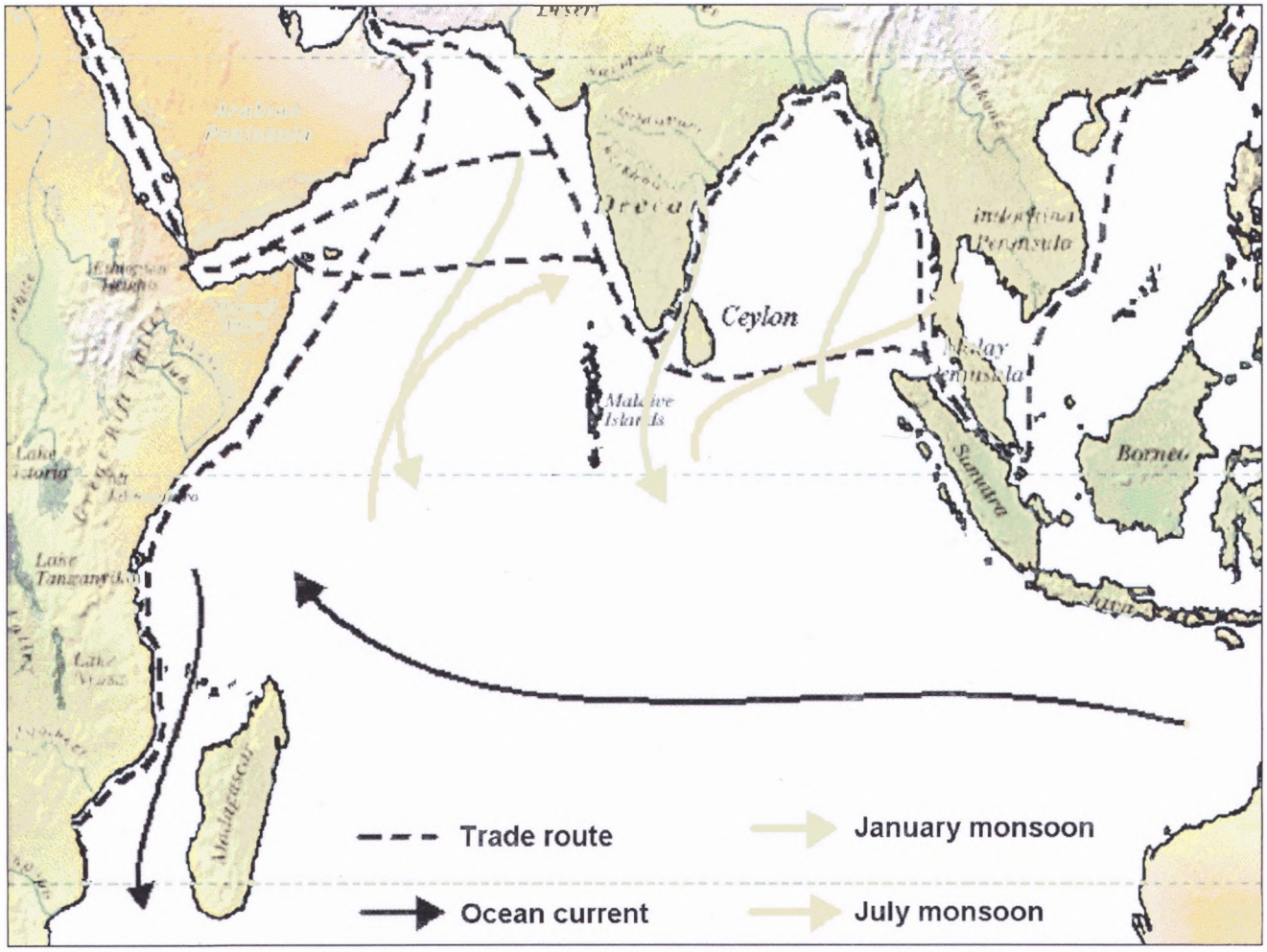


Figure 1-3 - Map showing medieval trade routes in the Indian Ocean and wind and ocean currents (after Oliver and Fagan 1975 and Dorling Kindersley 1995)

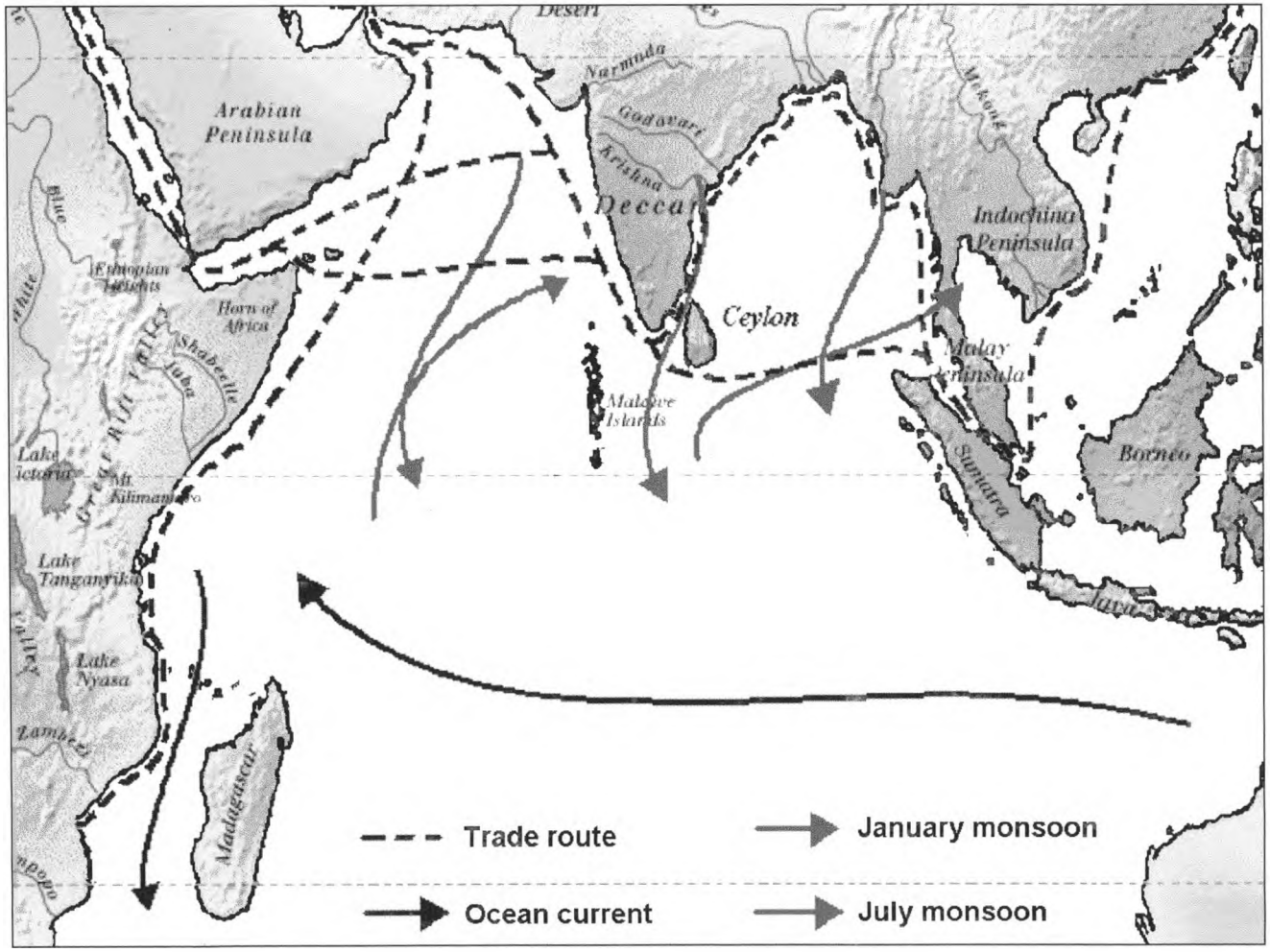


Figure 1-3 - Map showing medieval trade routes in the Indian Ocean and wind and ocean currents (after Oliver and Fagan 1975 and Dorling Kindersley 1995)

1.1.2.3 EVIDENCE INDICATING THE ORIGINS OF THE MALAGASY

1.1.2.3.1 THE MALAGASY LANGUAGE

The language spoken almost universally in Madagascar today, Malagasy, is an Austronesian language, although it does contain some Bantu vocabulary. Dialectical variation within the island has been studied and, using glottochronology, it has been proposed that 2000 years ago is the approximate time of divergence of the dialects (Vérin *et al* 1970). The most closely related language is spoken in the Barito valley in Borneo. Other studies based on the number of Sanskrit loan-words present in Malagasy have, however, argued that the first migrants arrived on the island sometime from the fifth to seventh centuries AD (Dewar and Wright 1993).

The language studies indicate a long contact with Bantu languages and also indicate that the original inhabitants of Madagascar remained in contact with the languages of southeast Asia for some time after settling on the island (Dewar and Wright 1993). This would support the theory of various waves of colonists, rather than a single exodus from Indonesia. If there was ongoing trade, this could explain ongoing contact with the language(s) of southeast Asia.

1.1.2.3.2 ARCHAEOLOGICAL FINDS IN MADAGASCAR

Very few archaeological studies have been carried out in Madagascar until fairly recently. Bones excavated at the turn of the century on the south-western coast appear to bear marks inflicted by metal tools. The bones have been dated to AD 80 (30 BC. - AD 230) and AD 380 (AD 240-440). It has been suggested that these may indicate the first signs of human activity. Other settlement sites in the southwest have been dated to AD 150 and AD 190. It is thought, however, that these are indications of hunting by non-residents (Dewar and Wright 1993).

The first indications of permanent occupation have been found on the northeast coast, dating to the late eighth century. These finds contain tradeware (iron smelting evidence and a necklace of silver, gold, carnelian and glass beads), indicating that the sites formed part of a trade network. A site known as Ihàrana on the northeast coast has revealed signs of a flourishing Moslem community. Tombs have been found containing Chinese porcelain, Persian glassware, bead necklaces, bronze mirrors and silver jewellery. Some of these date to the twelfth century, others are as recent as the 17th century. Other tombs on the west coast are thought to be linked with the Sacumbe people who lived in Mozambique. Both the name of the place, Sakoambé, and the traditional bird-carvings on the tombs, which are similar to soapstone carvings found in Zimbabwe, implicate a link with Africa (Brown 1978). As yet, there is no indication of occupation of the interior of the island in the first millennium. The first record of occupation of the central highlands is dated to the 13^{th} or 14^{th} century. At the same date, there is evidence of occupation of large portions of the coastlines (Dewar and Wright 1993).

1.1.2.3.3 CULTURE AND ETHNOLOGY OF THE MALAGASY

Much of the culture of the Malagasy is derived from Southeast Asia, Africa, the Near East and India. This indicates that the "Malagasy culture is a product of interaction throughout the Indian Ocean and not exclusively a marriage of Southeast Asian and African origins" (Dewar and Wright, 1993).

Indicators of Indonesian ancestry include rectangular houses (shown in Figure 1-4), particularly in the highlands, aligned in particular directions. Other indicators include typical Indonesian tombs (Figure 1-4), risiculture techniques, utensils, basket weaving and the use of the Indonesian outrigger canoe (Brown 1978, Dewar and Wright 1993).

The African influence in Madagascar can be seen in some of the religious beliefs, fibre-weaving techniques, musical instruments and the herding of the African Zebu cattle in the lowlands. African customs, especially among the Sakalava of the west coast include *tromba* (spirit possession) and the *dady* (cult of royal relics). The name Masikoro possibly originates with the Tanzanian word *mashokora*, relating to a type of scrub forest (Brown 1978).

Arabic and Islamic influences include the use of words borrowed from Arabian (such as words for the days of the week and words connected with arithmetic and divining). Islamic astrology is also widespread. Arabic cultural influence is prominent among the southeastern groups, Antemoro, Antanosy and Antambahoaka (Brown 1978).

Cultural markers are not always consistent. The outrigger canoe is found along the western coast (Mack 1986) which, from other cultural markers is the most African part of the island (Dewar and Wright 1993). One must bear in mind, however, that the use of cultural markers depends on the way of life. For example, the highlands are geographically better suited to risiculture than the lowland plains, which were better suited to cattle-herding (Brown 1978). In the same way, the outrigger canoe would only be found along the coast, irrespective of the nature of the people living there.

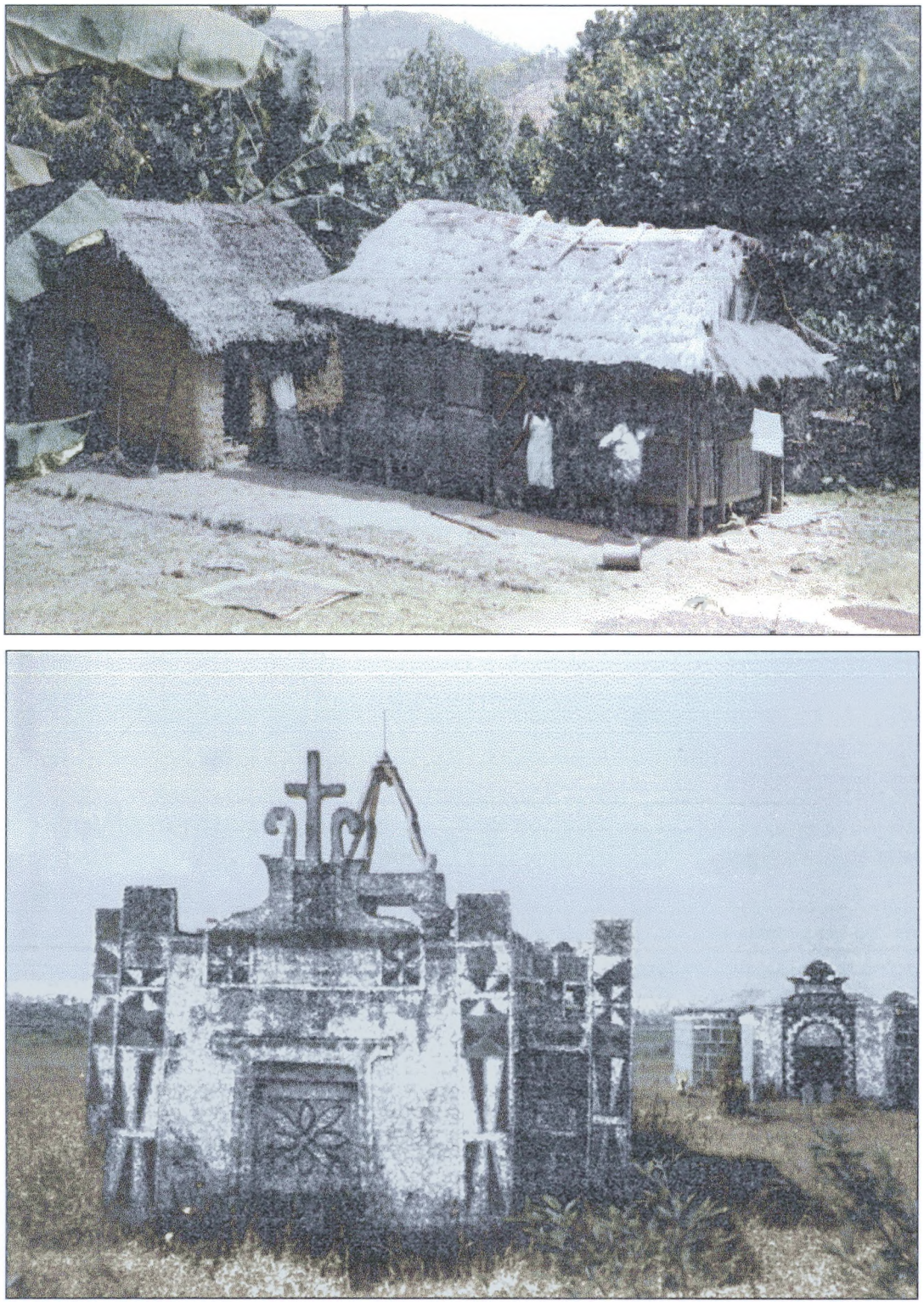


Figure 1-4: Pictures of the typical rectangular houses (top) and tombs (bottom) found on the central plateau indicating Indonesian influence

(Photographs used with the kind permission of Prof. T Jenkins)

1.1.2.3.4 HISTORICAL EVIDENCE OF MALAGASY ANCESTRY

Written Malagasy historical records are sparse, largely because there was little indigenous written language until the arrival of European missionaries in the middle of the last century (Singer *et al* 1957), although there are oral traditions. In the southeast are a group of people who call themselves Zafy-Ibrahim (descendants of Ibrahim or Abraham). It is doubtful that they are Jews, as they claim, but rather of Arabian descent (Brown 1978). Along the east coast there are numerous Arab or Islamic groups who have no knowledge of central aspects of the Moslem religion. This indicates that they arrived either in pre-Islamic times or shortly thereafter.

Another example of verbal history comes from the Sakalava tradition that a group of people landed on the southwest coast with a shipload of gold. These people probably came from Zimbabwe, a major source of gold at the time, whereas gold was not mined in Madagascar until the nineteenth century. There is also archaeological evidence linking Madagascar to Zimbabwe (as discussed in section 1.1.2.3.2).

Oral traditions among the Merina go back to the sixteenth century, saying that a light-skinned people "from the east", the Hova, overcame the indigenous dark-skinned Vazimba, partly in battle and partly by intermarriage. This has been interpreted to mean that these late arrivals came from Malaya or Java, explaining the greater proportion of Asian features in these highlanders (Brown 1978).

There are some examples of written history, mainly among the Antemoro. It is thought that they originally wrote in Arabic, but that this was adapted to the Malagasy language at an early stage. These written articles fall into two categories. The first deals with subjects such as astrology, divination and medicine. The other deals with the history of the Antemoro, although their reliability is uncertain. Some seem to have been written recently or copied, with variable degrees of accuracy, and the oldest two cannot be viewed by outsiders (Brown 1978).

1.1.2.3.5 PHYSICAL APPEARANCE OF THE MALAGASY

The Malagasy people are said to range between the two extremes of the African and Asian physical appearance. Individuals of Indonesian-like appearance tend to be found in the highland regions while people of African-like appearance are found in the lowlands.

An attempt was made to quantify and compare Malagasy physical features in the highland and lowland groups in a study by Singer *et al* (1957). The physical characteristics were traits selected to give indications of ethnic origin, including the hair ("frizzy" or straight), lips, eyes (for example presence or absence of epicanthic folds), the nasal width and the skin colour. The results of these analyses were vague but indicated the presence of African admixture in the gene pool. The comparison between the highland and lowland groups showed that the representative groups were significantly different (Singer *et al* 1957).

1.1.2.3.6 GENETIC STUDIES OF MALAGASY PEOPLES

There have been few genetic studies performed on the Malagasy. Until recently, those that have been performed have been deficient in sampling procedure, sample size and availability of informative markers. Hostility of the locals towards research performed by outsiders has also been a complication (Cole 1992).

In 1957, Singer *et al* presented findings on the physical features (see section 1.1.2.3.5), sickling and various serological traits of the Malagasy. The sickling and serology results confirmed the presence of African admixture and the authors calculated a figure for Malagasy ancestry composition of one third Indonesian and two thirds African (Singer *et al* 1957).

More recent publications by Buettner-Janusch and his colleagues (Buettner-Janusch and Buettner-Janusch 1964, Buettner-Janusch *et al* 1973) described an analysis of transferrins, haemoglobins, haptoglobins and ceruloplasmins in the Malagasy. Although the sample sizes were small, the sickle cell trait was found to occur at a significantly higher frequency in the lowlanders than the highlanders. Although this could be interpreted as indicating a larger proportion of African admixture, one cannot be sure of this because the sickle cell trait provides a selective advantage in the heterozygote against malaria and is therefore subject to selection. It could indicate a greater selective advantage in the lowlands than in the highlands. Results for the other blood proteins also support the indication of greater African contribution to the lowlanders (Buettner-Janusch and Buettner-Janusch 1964). In the 1973 study, two ethnic groups were analysed: Merina (representative of the highland groups) and Sakalava (lowland). Again, a significant difference was found in the red blood cell protein frequencies studied (Buettner-Janusch *et al* 1973).

In a recent analysis of haplotypes associated with the sickle cell mutation (Hewitt *et al* 1996), it was concluded that the major African component of the Malagasy ancestry is derived from Bantu-speaking Negroid people from Central and East Africa. Like the other genetic studies, this study fails to show substantial admixture from other countries. The sickle cell trait is not, for example, found in south-east Asia and no conclusions concerning the proportion of Indonesian admixture can therefore be drawn.

Analysis of Malagasy mtDNA has shown that approximately 70.7% of mtDNA associated with the 9-bp deletion in the Malagasy is derived from Asia, whereas 29.3% is derived from Africa (Soodyall *et al* 1995). Since mtDNA is maternally inherited it only provides information on female contributions. It does show that there were Asian women among the colonists. In 96.2% of the Asian-origin Malagasy mtDNA, the 9bp deletion is associated with a haplotype that is common in Polynesia, but rare or absent in Indonesia. It is conceivable that a single ancestral Indonesian population was responsible for the colonisation of both Polynesia and Madagascar (Soodyall *et al* 1995).

1.1.2.3.7 SUMMARY OF MALAGASY ANCESTRY

In summary, the Malagasy of today are thought to be composed of mainly African and central Indonesian elements. The Indonesian culture and language predominate but there is substantial admixture from other cultures, notably African and Arabian.

The central highland groups appear to be more Indonesian-like. The Merina group seems to have a particularly strong Indonesian influence. The Tsimihety and the Tanala have a more Indonesian/African mix, thought to be similar to the original Indonesian/African mix of the first colonists, while the Betsileo are thought to have slightly more African admixture. The lowland groups, Sakalava and Bara, appear to have particularly strong African influences, while the Mahafaly and Antandroy have slightly more Indonesian admixture. The tribes on the southeast coast, especially the Antemoro, the Antambahoaka and the Antenosy, appear to show a strong Arabic/Islamic affinity (Brown 1978).

1.2 GLUCOSE-6-PHOSPHATE DEHYDROGENASE

The glucose-6-phosphate dehydrogenase (G6PD) protein is a cytoplasmic enzyme found in all cells. Mutations in the G6PD gene cause G6PD deficiency, probably the commonest disease-producing enzyme disorder in human beings. More than 400 G6PD variants have been characterised biochemically and approximately 7.5% of the world population (400 million individuals world-wide) are estimated to have a G6PD deficiency gene (Vulliamy *et al* 1993). This high prevalence world-wide is largely due to the selective advantage conferred to carriers of deficiency variants against *Plasmodium falciparum* malaria (Ruwende *et al* 1995). Although G6PD deficiency is an X-linked condition, approximately 10% of G6PD deficient individuals are female. This is due to the high frequency of the gene(s) causing enzyme deficiency. In addition, about 10% of heterozygous females are deficient due to non-random X inactivation of the normal allele (WHO Working Group 1989).

The G6PD protein has a vital housekeeping function and so, while variants abound, a mutation completely inactivating the protein is thought to be incompatible with life in humans (WHO Working Group 1989). A detailed biochemical discussion of the G6PD enzyme will not be provided in this review as the focus of this study is DNA analysis of the variants. With the cloning of the G6PD gene, diagnosis of deficiency by detection of deficiency variants is now possible using molecular methods in the absence of clinical manifestations. This makes large-scale screening feasible, particularly for the common deficiency variants. Many of these variants are localised to distinct geographical regions and this type of screening is therefore suitable for use in studies of populations and population movements. Few data regarding frequencies are provided in this introduction since they will be presented in the discussion for the purposes of comparison with published reports (Chapters 3 and 4).

1.2.1 CLASSIFICATION OF G6PD DEFICIENCY

G6PD variants may be subdivided into five classes, based on the level of enzyme activity in the red blood cells and on the associated clinical manifestations (WHO Working Group 1989). All variants, both so-called "deficient" and "non-deficient" can be placed into one of these classes.

Class I variants: severely deficient and associated with chronic, non-spherocytic haemolytic anaemia (CNSHA).

Class II variants: severely deficient (residual activity of < 10%) with no associated CNSHA (e.g. G6PD Mediterranean).

Class III variants: moderately deficient (residual activity is 10-60%) (e.g. the common African deficiency, A^-).

Class IV variants: normal enzyme activity (60-150%) and usually electrophoretically different from B (e.g. the A variant).

Class V variants: increased enzyme activity.

Clinical manifestations are associated mainly with variants from classes I, II and III. These are the so-called "deficient" variants. Most of the common pathological variants are found in classes II and III, since class I variants are never polymorphic, presumably due to their severity (Luzzatto and Mehta 1995). Many of the variants, both deficient and non-deficient, occur at polymorphic frequencies in different populations of the world.

1.2.2 CLINICAL CONSIDERATIONS OF G6PD DEFICIENCY

1.2.2.1 HISTORY OF THE DISEASE

The recognition of G6PD deficiency dates back to observations by Pythagoras, who warned against eating the broad bean *Vicia faba* (Luzzatto and Mehta 1995). By the turn of the century, physicians recognised G6PD deficiency as a clinical condition (Beutler 1993). In the 1950's it was discovered that primaquine, an antimalarial drug, caused haemolytic anaemia in certain subjects (Carson *et al* 1956).

The variability of deficiency-causing variants was recognised by the late 1960's and the WHO suggested guidelines for the characterisation of variants to improve standardisation (Betke *et al* 1967). By 1983, almost 400 biochemical variants had been described (Yoshida and Beutler 1983), although it was often difficult to determine whether these were distinct (for example, Fairbanks *et al* 1980).

The G6PD gene was localised to Xq28 in 1980 (Pai *et al* 1980) and in 1986, the G6PD gene was cloned (Persico *et al* 1986, Takizawa *et al* 1986), enabling mutation analysis to be carried out. More recent advances in G6PD research include the discovery of the binding domain for NADP (Hirono *et al* 1989), the determination of the full sequence of the G6PD gene (Chen *et al* 1991) and the postulation of the three dimensional (3D) structure of the G6PD enzyme based on the 3D structure of the G6PD enzyme from *Leuconostoc mesenteroides* (Naylor *et al* 1996).

1.2.2.2 SIGNS AND SYMPTOMS OF G6PD DEFICIENCY

Luzzatto and Mehta (1995) reviewed the spectrum of symptoms in 1995. Since the orientation of this study is not clinical, they will be mentioned here but not discussed in detail.

Most individuals with G6PD deficiency are asymptomatic. When the condition manifests (usually as acute anaemia), the body usually compensates rapidly and the condition remains undetected. The symptoms of G6PD deficiency usually manifest in hemizygous males in response to an environmental trigger. In severe cases, symptoms manifest in the neonate as neonatal jaundice. Haemolysis in G6PD deficient individuals may be induced by infection (viral, rickettsial or bacterial), by oxidant drugs (including primaquine and naphthalene) or by the ingestion of fava beans (i.e. favism). The severity of the clinical manifestations is dependent on the type of variant causing the deficiency. For example, favism is associated with the Mediterranean deficiency variant but not usually with the A^- African deficiency variant (Sodeinde 1992). (At least one individual with A^- deficiency associated with favism has been described (Calabrò *et al* 1990).)

1.2.2.3 PREVENTION, SCREENING AND TREATMENT OF G6PD DEFICIENCY

G6PD deficiency constitutes a significant health burden. Prevention of the symptoms of G6PD deficiency can be accomplished effectively by educating the public and health care workers. (Sodeinde 1992). Education in combination with screening or diagnosis is particularly effective since at-risk individuals can then be targeted specifically.

Neonatal screening has been recommended by the WHO in areas where the G6PD deficiency prevalence is 3-5% or more in males (WHO Working Group 1989). Treatment for neonates includes exchange transfusion and phototherapy (Tan, 1976). In developing countries it is a simple and cost-effective way of reducing the health burden (WHO Working Group 1989).

1.2.3 GEOGRAPHIC DISTRIBUTION OF G6PD DEFICIENCY

G6PD deficiency has a distinct global geographical distribution with a deficiency range of 0% to 26% in most populations and up to 70% in isolated populations as shown in Figure 1-5. Since deficiency frequencies will be discussed in detail in the discussion (section 3.1.1), specific frequencies will not be discussed in detail in this section. The highest reported frequency of G6PD deficiency is approximately 70% in the Kurdish Jewish population (Cohen 1971). Deficiency in Southern Africa ranges from 0% in Sotho individuals (Jenkins *et al* 1968) to 33.3% in a Bemba group in Zambia (Barclay *et al* 1970, reviewed by Jenkins 1972). Deficiency in Asia ranges on average from 0.1% to 7.2% (Sodeinde 1992). In general, the frequency of G6PD deficiency decreases with increasing altitude, apparently because of lower levels of malarial endemicity (Sodeinde 1992).

1.2.3.1 SELECTIVE ADVANTAGE AGAINST MALARIA

Figure 1-6 shows the distribution of malaria in 1991. Comparing Figures 1-5 and 1-6, one can see that the distribution of malaria and the distribution of G6PD deficiency correlate closely. This led to the suggestion that G6PD deficiency provided a selective advantage against *Plasmodium falciparum* malaria (Allison 1960).

Until recently, it was thought that protection from malaria by the A^- ($G6PD^{A-}$) allele was limited to heterozygous females but not hemizygous males and that non deficienct $G6PD^A$ hemizygotes were also protected (Bienzle *et al* 1972). It has recently been shown, however, that there is a substantial protection both to the hemizygous male (58% protection against severe malaria) and to the heterozygous female (46% protection against severe malaria) for individuals carrying the $G6PD^{A-}$ African deficiency variant (Ruwende *et al* 1995). While protection can be seen in both mild and the rarer severe malaria, it has been found that severe malaria provides a more sensitive system than mild malaria in assessing protection. The study showed no difference in protection between the A and B non-deficiency variants (Ruwende *et al* 1995), unlike the study by Bienzle *et al* (1972).

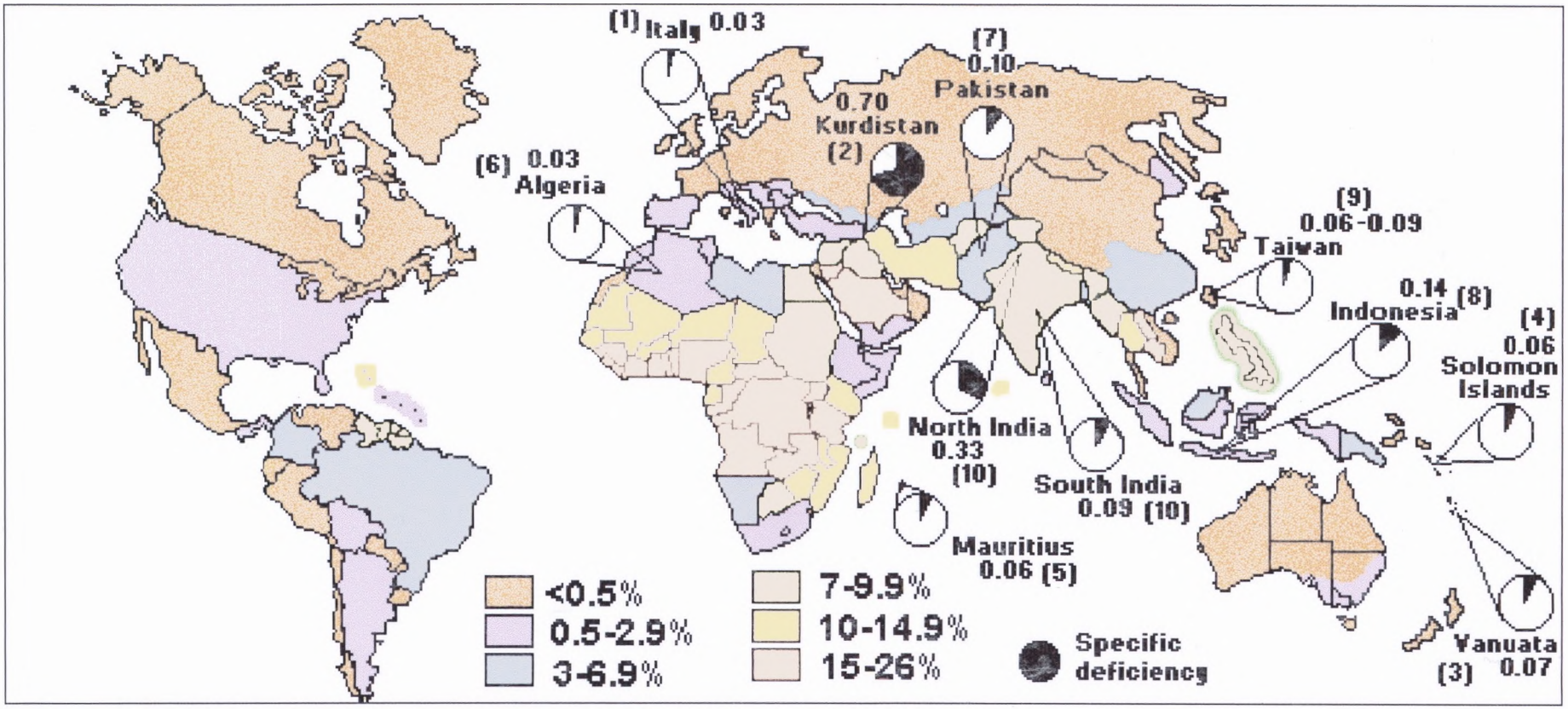


Figure 1-5 - Map showing geographic distribution and frequencies of G6PD deficiency (adapted from WHO Working Group 1989)

Specific deficiency frequencies are displayed in pie charts as reported by:

(1) Calabro *et al* 1990, (2) Cohen 1971, (3) Ganczakowski *et al* 1995, (4) Hirono *et al* 1995, (5) Kaeda *et al* 1996, (6) Nafa *et al* 1994, (7) Saha *et al* 1994, (8) Soemantri *et al* 1995, (9) Tang *et al* 1995, (10) Kaeda *et al* 1995.

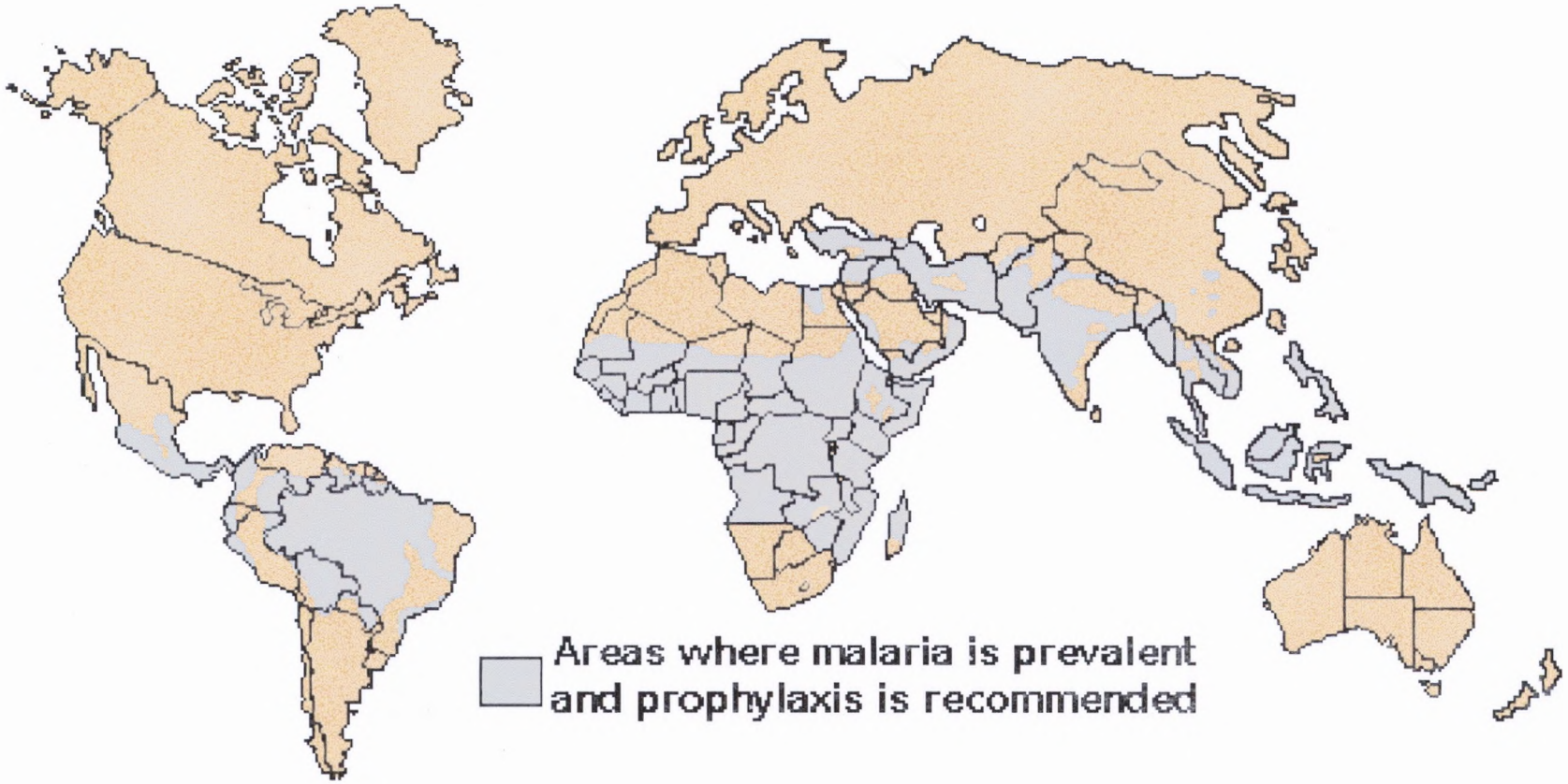


Figure 1-6 - Geographic distribution of malaria in 1991 (after Bartholomew World Environment Atlas 1991)

The mechanism of the malarial protection is not entirely clear. *P. falciparum* growth inhibition has been shown *in vitro* in enzyme deficient erythrocytes (Roth *et al* 1983). It has been shown that *P. falciparum* development is inhibited under conditions of oxidative stress (Friedman 1979) and so it is possible that the mechanism of protection is similar to that of the sickle cell trait, where parasitised cells die quicker than non-parasitised cells. G6PD A^- deficiency affords less protection than the sickle cell trait, but the same or greater protection than that provided by HLA and thalassaemia variants (Ruwende *et al* 1995).

The enzyme deficiency produced by the A^- variant is small compared to deficiency variants in Asia and the Mediterranean. It remains to be determined whether deficiency variants other than the A^- variant produce a greater degree of protection.

1.2.4 THE G6PD PROTEIN

1.2.4.1 STRUCTURE OF THE PROTEIN

The active G6PD enzyme exists either as a dimer or a tetramer, containing tightly bound NADP (Mason 1996). The primary structure is a 515 amino acid polypeptide with a molecular weight of 59 265 daltons. The three dimensional structure of the human G6PD protein is not as yet known. The 3D structure of G6PD of the bacterium *Leuconostoc mesenteroides* has been determined and this has been used to postulate the 3D structure of the human G6PD (Naylor *et al* 1996).

Because the 3D structure of the enzyme is not available, other strategies have been adopted to identify functional domains. The region surrounding Lys 205 (exon 6), is thought to be the binding domain for glucose-6-phosphate (G6P), because of a lowered K_m for G6P associated with mutations in this area. The NADP binding site is thought to be located near exon 10 (amino acid residues 352-429) since mutations in this area often result in CNSHA and there is some reactivation of the enzyme in the presence of Mg^{2+} and NADP (Beutler 1994). G6PD Orissa (44 Ala→Gly) has a K_m for NADP fivefold higher than the normal. This could indicate that the site is a co-enzyme binding domain (Kaeda *et al* 1995). Analysis of conserved residues or tracts of residues has also been used to postulate functional domains (Luzzatto and Mehta 1995).

1.2.4.2 METABOLIC FUNCTION OF THE PROTEIN

G6PD is a "housekeeping" enzyme vital to the life of every cell but most conspicuous in the red cell. The enzyme catalyses the first step in the pentose phosphate pathway (reviewed in Luzzatto and Mehta 1995), converting glucose-6-phosphate (G6P) to 6-phosphogluconolactone and reducing NADP to NADPH. This reduced co-factor is important in protecting cells from oxidative damage. The enzyme is vital in red blood cells (which are particularly prone to oxidative damage) because it is the only enzyme which produces NADPH in these cells. The NADPH reduces glutathione, which in turn reduces the oxidant peroxide (H_2O_2) to water (WHO Working Group 1989). Superoxide, another oxidative radical, can be converted to H_2O_2 by superoxide dismutase as shown in Figure 1-7.

An alternate pathway for peroxide detoxification is mediated by the enzyme catalase. It has been found, however, that catalase has four moles of NADPH per mole of catalase tightly bound to it as a co-enzyme (Kirkman and Gaetani 1984) and G6PD is thus required for this H_2O_2 detoxification pathway too. The catalase pathway may become more important when there are low levels of G6PD i.e. in G6PD deficiency (Gaetani and Ferraris 1988).

The mechanism by which environmental triggers in G6PD deficient individuals cause haemolysis is not clearly understood. It is thought that the oxidation of -SH groups in vital proteins decreases cell deformability (when cross-linking occurs on the surface of the cell) and possibly makes the cell recognisable to macrophages for destruction. This implies that a threshold of NADPH is required to prevent haemolysis, and could explain why class I variants cause spontaneous haemolysis. Class II and III variants presumably produce sufficient NADPH for normal functioning and haemolysis only occurs when extra NADPH is needed (i.e. during oxidant stress) (Luzzatto and Mehta 1995).

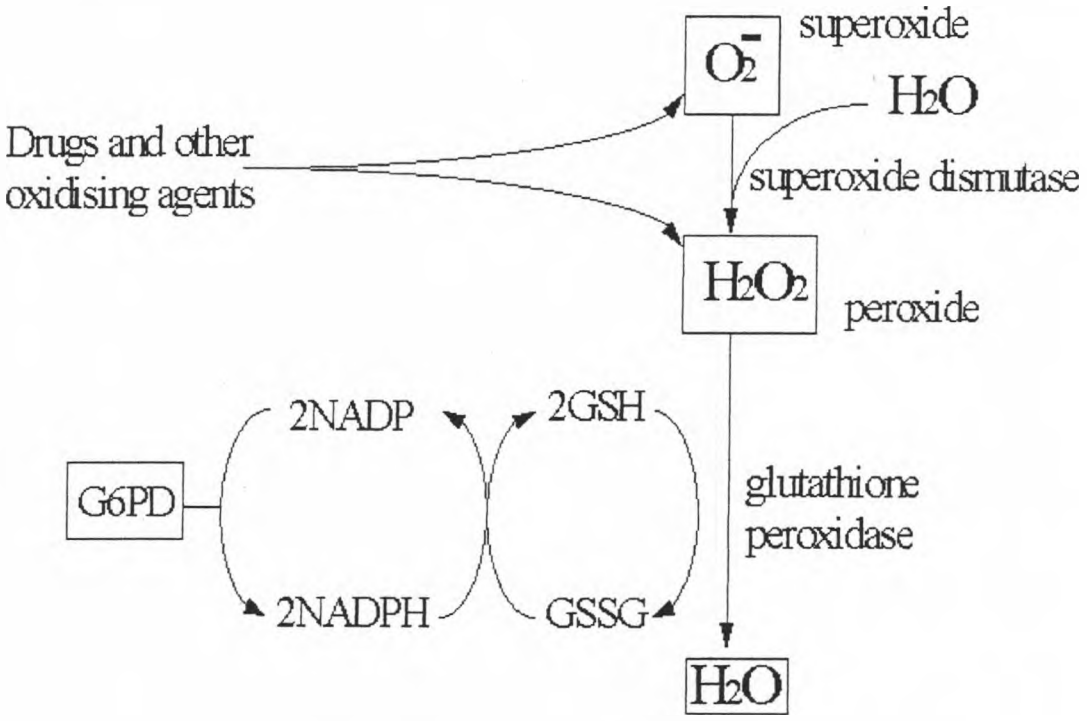


Figure 1-7 - The function of G6PD in the Pentose Phosphate pathway (after WHO Working Group 1989)

1.2.5 THE G6PD GENE

1.2.5.1 LOCATION OF THE G6PD GENE

The G6PD gene was localised to the region Xq28 in 1980 (Pai *et al* 1980), by studying translocation breakpoints in X-autosome translocations. The position of G6PD along with the FVIII and the red-green colour blindness genes is shown in Figure 1-8. The complete sequence was determined by Chen *et al* (1991).

1.2.5.2 STRUCTURE OF THE GENE

The G6PD gene (shown schematically in Figure 1-8) is approximately 18kb long, containing 13 exons. The coding sequence is 1545bp long, beginning in exon 2, coding for a protein sequence 515 amino acids in length. The 5' untranslated sequence is 71 bp long and the 3' untranslated sequence is 608 bp long. The largest intron is intron 2 with a total length of 9.82kb (Chen *et al* 1991).

The promoter region contains typical housekeeping gene elements. These include several potential Sp1-binding sites, no CAAT box, an atypical TATA box and a segment homologous to the SV40 21-bp repeat. The region from -1200bp to within the first exon is GC rich with many *Hpa* II sites, characteristic of an HTF island (reviewed by Luzzatto and Mehta 1995).

Sixteen *Alu* sequences or portions of *Alu* sequences occur at a high concentration (4 to 5 times higher than expected for a random 20kb genomic sequence). Three of the *Alu* sequences are located in the region 5' to the transcription initiation site, one is located at the 3' region of the gene and the remaining 12 cluster within intron 2 (Chen *et al* 1991). A number of open reading frames are also present in intron 2, although a search of GenBank has not revealed homology with other regions (Chen *et al* 1991).

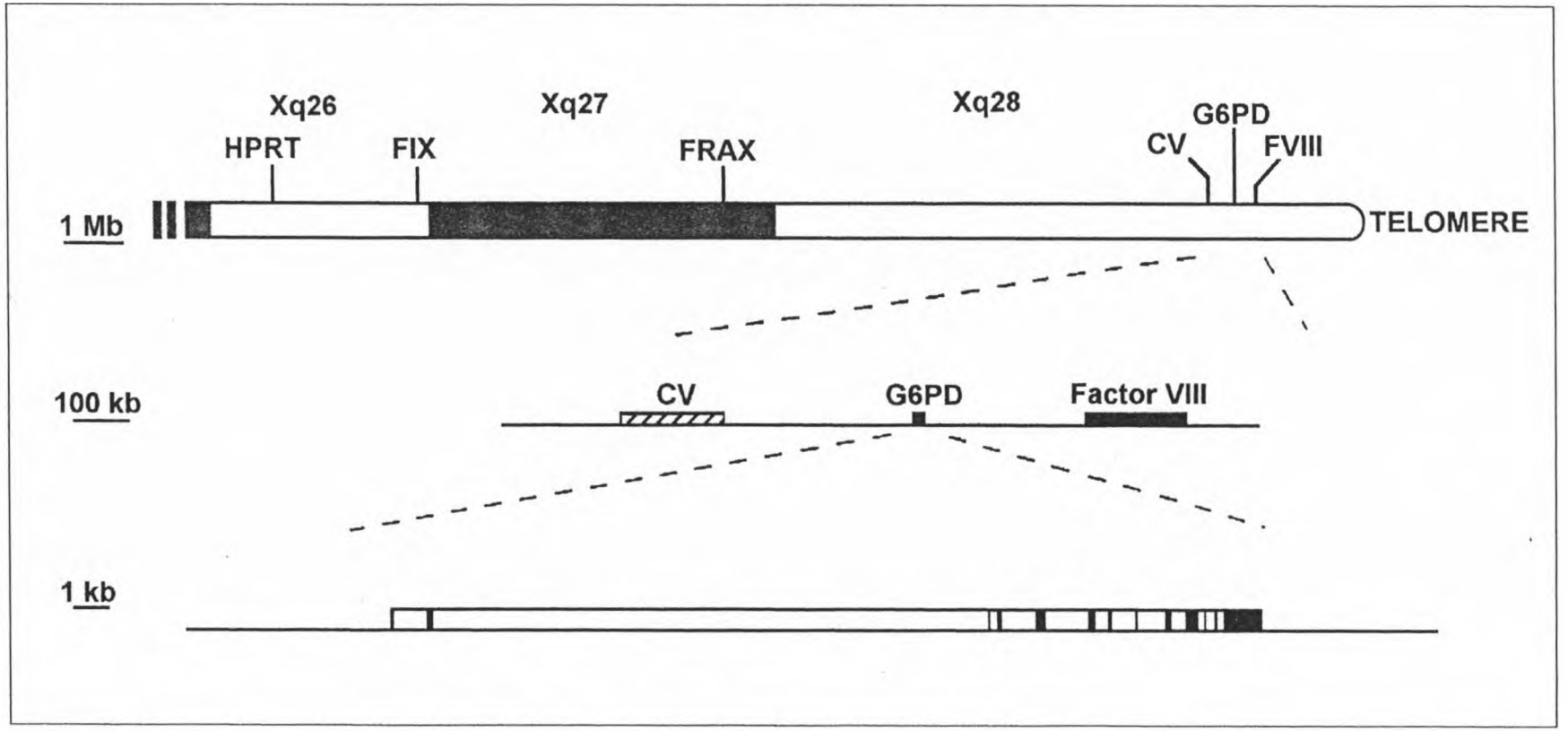


Figure 1-8 - Location and structure of the G6PD gene (after Vulliamy *et al* 1992)

Included on the diagram are genes in the Xq26 region (HPRT = hypoxanthine phosphoribosyltransferase and FIX = Factor IX gene); in the Xq27 region (FRAX = Fragile X A gene) and in the Xq28 region (CV = gene for red-green colour vision, and FVIII = the factor VIII gene).

1.2.6 VARIANTS WITHIN THE G6PD GENE

1.2.6.1 SPORADIC MUTATIONS AND POLYMORPHISMS

Variants may be classified as sporadic (i.e. new mutations) or polymorphic (a frequency >1%) (Sodeinde 1992). Variants from class I (related to CNSHA) are almost never polymorphic, probably because they are too severe to become balanced by malarial selection.

The majority of G6PD variants are sporadic and are unaffected by drift or selection. Most of them have been discovered because of the severe clinical manifestations they produce. A sporadic variant must be severe enough to limit its own survival in red blood cells, but not severe enough to be incompatible with the life of the individual.

To attain polymorphic frequencies, a variant must be neutral (e.g. in introns) or advantageous with respect to selection (Town *et al* 1992). Most polymorphic variants are associated with G6PD deficiency, apparently due to the selective effect of malarial protection. The variant must then be severe enough to cause deficiency, but not severe enough to outweigh the selective advantage of malarial protection. In other words, the variant must cause haemolysis in times of oxidant stress, but not spontaneously. It is therefore not surprising that no polymorphic variants fall into class I (CNSHA-associated) and that nearly all polymorphic variants fall into classes II or III (i.e. deficient variants) (Luzzatto and Mehta 1995). Polymorphic variants are useful in population studies since the variants often have distinct geographical distributions (see section 1.2.7).

1.2.6.2 MOLECULAR ANALYSIS

Characterisation of red cell G6PD deficiency variants has classically been performed biochemically (Betke *et al* 1967). Molecular characterisation of variants, which is more definitive than biochemical characterisation, has revealed that some variants thought to be different based on biochemical analysis are due to the same mutation. Also, variants which were thought to be the same based on biochemical characterisation have been found to be due to different mutations (for example the A- deficiency variant discussed in section 1.2.6.3.1.2).

Molecular techniques used in the characterisation of variants include allele-specific oligonucleotide hybridisation (ASOH) analysis and single-strand conformation polymorphisms (SSCP) analysis (Calabrò *et al* 1993), followed by sequencing. For a known mutation, ASOH analysis can also be useful in rapid screening (Calabrò *et al* 1993). Restriction fragment length polymorphism (RFLP) is useful for rapid detection of a known mutation which destroys or creates a recognition sequence for a restriction enzyme. For large rearrangements or deletions, RFLP analysis may be inappropriate and even a single nucleotide substitution does not always change a site (although primers can often be designed to introduce a restriction site from the mutation).

Most of the deficiency-causing mutations are single amino acid missense mutations, presumably because major rearrangements are incompatible with life. Variants not caused by missense mutations include: single amino acid deletions (G6PD Sunderland, Tsukui and Urayasa); a deletion of 2 amino acids (G6PD Stonybrook); a substitution of 3 amino acids (G6PD Vancouver); a deletion of 8 amino acids (G6PD Nara); and a splice site mutation (G6PD Varnsdorf). The only nonsense mutation, which has been found in a heterozygote, is G6PD Georgia (Huang *et al* 1996, Mason 1996). Some of the variants characterised at the molecular level are shown in Figure 1-9.

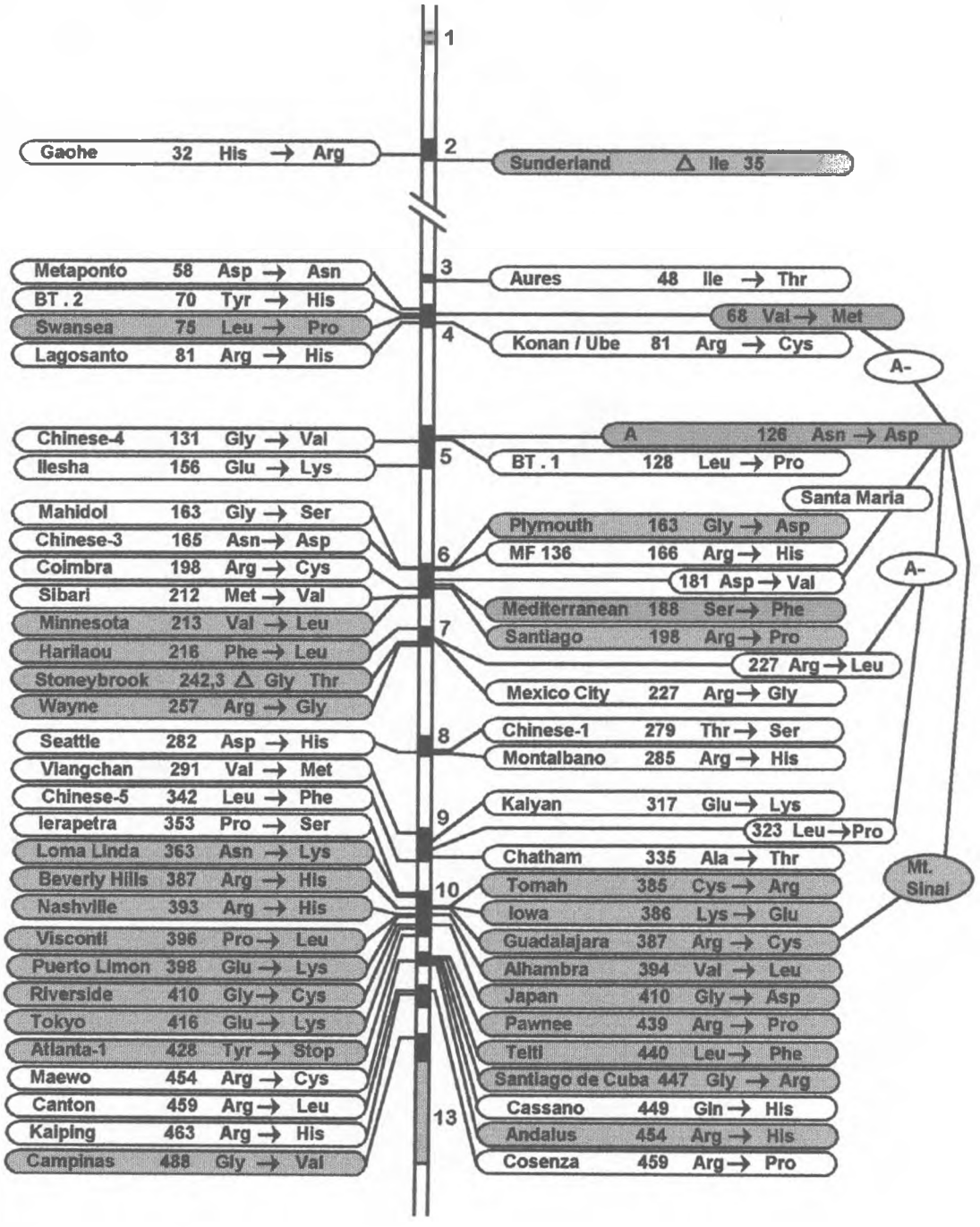


Figure 1-9 - Some variants found in the coding regions of the G6PD gene (after Vulliamy *et al* 1993, Luzzatto and Mehta 1995)

Exons are shown as numbered boxes (black for coding and grey for non-coding exons). Variants shaded in grey are those associated with CNSHA (i.e. class I), and those shaded in red are variants examined in this study.

1.2.6.3 NUMBER AND GEOGRAPHIC DISTRIBUTION OF G6PD VARIANTS

Many of the polymorphisms and deficiency variants are population specific. The major implication for population genetics is that geographically specific variants can be used to trace population movement and admixture. What follows is a description of some of the more common variants and their geographical distributions. Although some frequencies are reported here to convey an impression of the global frequencies, details of the frequencies in pertinent geographical regions are presented for comparison purposes in the discussion (Chapters 3 and 4).

1.2.6.3.1 AFRICAN VARIANTS

1.2.6.3.1.1 THE A VARIANT

The G6PD A variant, 376 A→G (126 Asn → Asp) (Takizawa *et al* 1987), is electrophoretically faster than the more common form (B), with a normal or slightly lower (90±15% of normal) activity (Town *et al* 1992). Although both G6PD A and G6PD B phenotypes have been found in chimpanzees (where G6PD A is caused by the same mutation as in humans) (Beutler *et al* 1989), it is thought that the ancestral form is G6PD B. The mutation creates a *Fok* I restriction enzyme site, making screening by RFLP analysis possible (Takizawa *et al* 1987).

G6PD A has been found mainly in individuals of African descent, although it has been found in some European populations (e.g. Italian and Spanish) at low frequencies (Calabro *et al* 1990, Rovira *et al* 1995). This is probably due to gene flow from Africa. It is found at a frequency of 0.2 in African Americans. In sub-Saharan Africa, the frequency ranges from 0.00 to 0.36 (Reys 1970, Jenkins 1972, Coetzee *et al* 1992).

1.2.6.3.1.2 THE A⁻ DEFICIENCY VARIANT

The most common deficiency variant occurring in the African Negroid population is electrophoretically fast and enzymatically deficient and is therefore called G6PD A⁻. This variant is a composite of at least 2 mutations. The first, 126 Asn→Asp, causes the B→A mobility change (as discussed in section 1.2.6.3.1.1) and the second mutation causes the deficiency itself (Hirono and Beutler 1988). Three mutations have been identified which cause the deficiency found in G6PD A⁻: 202 G→A (68 Val→Met) (which is the most common) (Hirono and Beutler 1988); 680 G→T (227 Arg→Leu) and 968 T→C (323 Leu→Pro) (Beutler *et al* 1989), which are much rarer.

G6PD A⁻ is not found exclusively in Africa, although it is rare outside Africa (Beutler *et al* 1989) and in other regions is probably present due to gene flow from Africa. It has been found in individuals from Spain (Rovira *et al* 1995), Puerto Rico, Mexico (Beutler *et al* 1989) and Italy (Calabro *et al* 1990). The mutations involved include the 68 Val → Met and 227 Arg → Pro mutations on A chromosomes (126 Asn → Asp). A⁻ frequencies in and around Africa include: Mauritania (in Northwest Africa) 0.07; Sahara 0.115; Morocco 0.016; Berber (a population in North Africa) 0.000 (Nafa *et al* 1994). In sub-Saharan Africa, frequencies range from zero in the Khoi (Jenkins 1972) to 0.19 in various ethnic groups in Mozambique (Reys 1970).

The common mutation, 68 Val → Met, creates a restriction site for *Nla* III (Hirono and Beutler 1988). The mutation has not been found on B (126 Asn) chromosomes and it is thought that the mutation originally occurred on an A (126 Asp) chromosome. It has also been shown *in vitro* that the 68 Val → Met mutation does not cause deficiency on its own (Town *et al* 1992). Both the 126 Asn → Asp and the 68 Val → Met mutations by themselves cause a small reduction in enzyme activity (90% and 76% of normal respectively). When both mutations are present, the reduction in enzyme activity is additive (approximately 50% of normal). However, the enzyme yield when both mutations are present is only 4% of normal (Town *et al* 1992).

1.2.6.3.1.3 "SILENT" POLYMORPHISMS

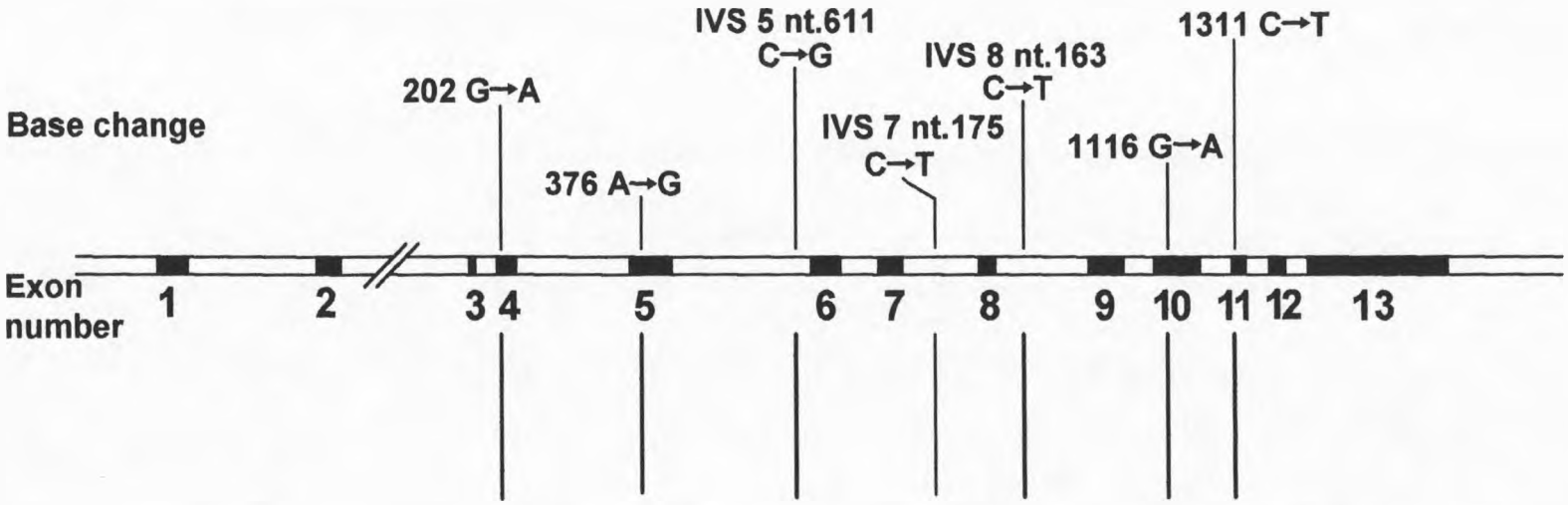
Mutations do not necessarily result in a change in the amino acid sequence. A silent G → A transition, removing a *Pst I* site, occurs at nt 1116 within the coding region (D'Urso *et al* 1988). The 1116 G→A mutation occurs in the last base of the codon CAG and the glutamine is therefore not changed. The mutation is polymorphic in people of African origin, but not in Europeans (D'Urso *et al* 1988). Few frequency data are available for this polymorphism but a frequency of 0.22 has been reported in Nigerians and the site is monomorphic in Europeans (D'Urso *et al* 1988).

1.2.6.3.1.4 POLYMORPHISMS IN INTRONS

Some G6PD polymorphisms are silent because they occur in introns. Four of these polymorphisms are: IVS5 nt611 C → G which creates a *Pvu II* site (Yoshida *et al* 1988); IVS2 nt9722 A → G which does not change a restriction site but mismatched primers can be used to generate a *Sca I* site (Vulliamy *et al* 1991b); IVS8 nt163 C→T which creates a *Bsp HI* site (Vulliamy *et al* 1991b); and IVS11 nt1311 C→T which does not create or destroy a site, although a *Bcl I* site can be introduced by mismatched primers (Kurdi-Haidar *et al* 1990). The *Bcl* I site is polymorphic in non-African and African populations, while the other sites are polymorphic only in African populations (Yoshida *et al* 1988, Fey *et al* 1990, Vulliamy *et al* 1991b).

1.2.6.3.1.5 HAPLOTYPE CONSTRUCTION

Using the combination of the intronic polymorphisms, the silent mutation, the A mobility variant and the common A^- deficiency variant, it is possible to construct haplotypes. Only 7 haplotypes (shown in Figure 1-10) have been described of a possible 128. Using this approach, it was shown that strong linkage disequilibrium exists between the alleles at the RFLP loci (Vulliamy *et al* 1991b). A likely evolutionary sequence of the mutations has been postulated (shown in Figure 1-11), based on the smallest number of mutational steps required to obtain the 7 haplotypes (Vulliamy *et al* 1991b).



Restriction site		*Nla* III	*Fok* I	*Pvu* II	"*Sca*I"	*Bsp* HI	*Pst* I	"*Bcl*I"	G6PD
Haplotype number:	I	-	-	-	-	+	+	-	B
	VII	-	-	-	-	+	+	+	B
	II	-	-	-	-	+	-	-	B
	III	-	+	-	-	+	+	-	A
	IV	-	+	-	-	-	+	-	A
	V	-	+	+	+	-	+	-	A
	VI	+	+	+	+	-	+	-	A

Figure 1-10 - Haplotypes of the G6PD gene (adapted from Vulliamy *et al* 1991b)

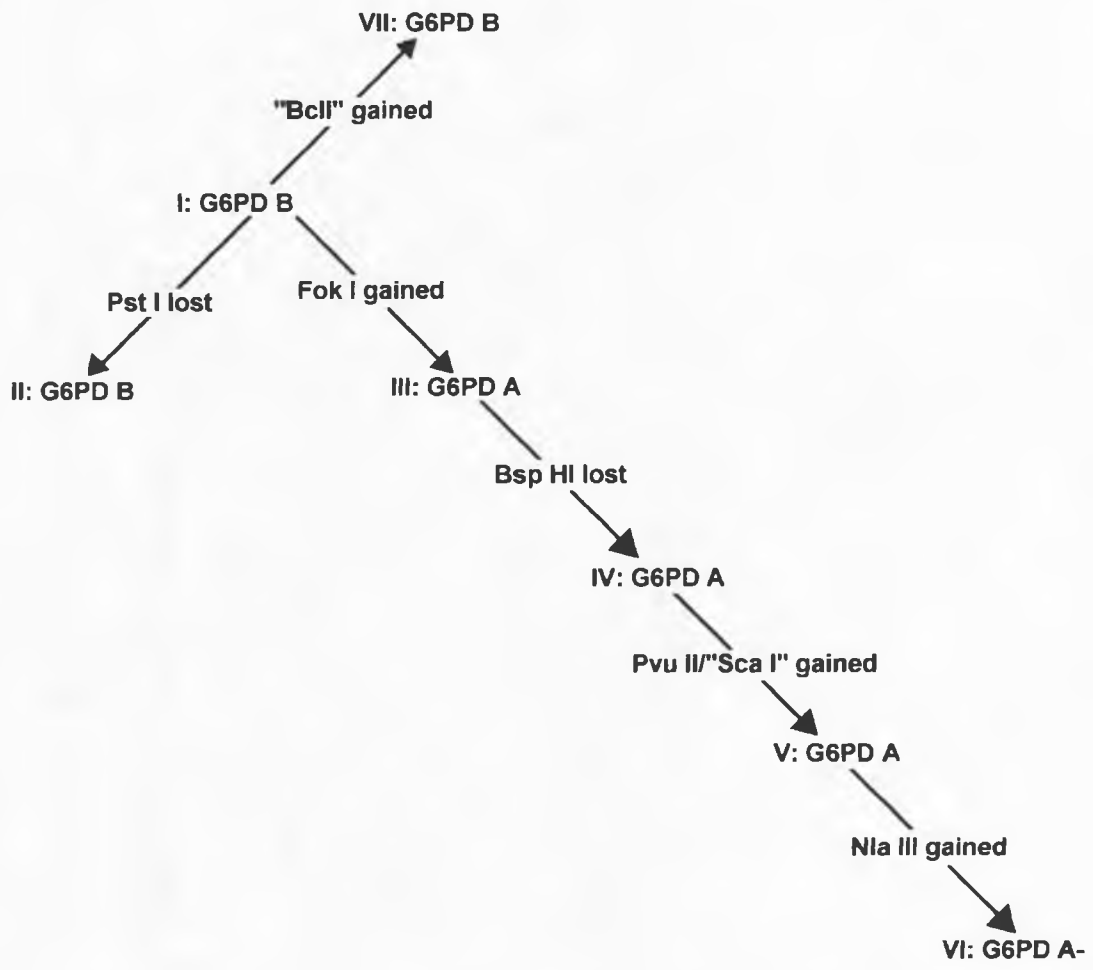


Figure 1-11 - Postulated evolutionary sequence of the mutations in the G6PD gene based on the smallest number of mutational steps required to give the 7 haplotypes (Vulliamy *et al* 1991b)

1.2.6.3.2 MEDITERRANEAN AND EURASIAN VARIANTS

1.2.6.3.2.1 G6PD MEDITERRANEAN

The most common deficiency variant in the Mediterranean region is G6PD Mediterranean (563 C→T) (188 Ser → Phe) (Vulliamy *et al* 1988). This variant has been found in Greece, southern Italy, Spain, Egypt, Israel, Lebanon, Turkey, Bulgaria, Rumania and Pakistan (Beutler and Kuhl 1990b, Kurdi-Haidar *et al* 1990, Oppenheim *et al* 1993). Deficient males in Europe and the Middle East with G6PD Mediterranean tend to have a T at nt 1311 (the "*Bcl* I" site discussed in section 1.2.6.3.1.4) (Kurdi-Haidar *et al* 1990). The G6PD Mediterranean variant has also been found in India. However, these individuals tend to have a C at nt 1311 (Beutler and Kuhl 1990b). This indicates that there are probably at least two origins of the Mediterranean mutation which can be distinguished by the *Bcl* I associated site (Beutler and Kuhl 1990b).

The proportion of G6PD deficient individuals with the G6PD Mediterranean deficiency variant varies widely in Europe. Reports of the G6PD Mediterranean frequency in deficienct individuals (i.e the proportion of total deficient variants) in Italy include 0.415 in Cosenza (Calabrò *et al* 1993) and 0.667% in Ferrara (Ninfali *et al* 1993). The Kurdish Jewish population has the highest reported G6PD deficiency frequency in the world (70% of males). The deficiency is almost entirely attributable to G6PD Mediterranean (0.955 of all deficient variants) (Oppenheim *et al* 1993). Individuals in northwest Pakistan and Punjabi individuals from an Indian community in western London have the Indian associated Mediterranean haplotype (Saha *et al* 1994, Kaeda *et al* 1995).

1.2.6.3.3 SOUTHEAST ASIAN VARIANTS

The most common deficiency variant occurring in south-east Asia is G6PD Canton (1376 G→T), which creates an *Afl* II site (Stevens *et al* 1991). The Mahidol (487 G→A) variant, which creates an *Alu* I site, the G6PD Union variant and the Mediterranean variant are found at high frequencies in some areas (Soemantri *et al* 1995).

In the Javanese in Indonesia, the Mediterranean variant (which comprises 0.313 of the total deficiency variants) is more common than the Mahidol variant (0.120) (n= 16). In 4 of 5 cases of the Mediterranean variant, the silent nt1311("*Bcl* I" site) mutation was absent, indicating the Indian haplotype associated with the variant (Soemantri *et al* 1995). G6PD Canton is also found at a high frequency (0.187 of total deficiency) in this area. The Mahidol variant has not been described in the Vanuata archipelago, although the G6PD Union variant is found at a high frequency (Ganczakowski *et al* 1995).

1.2.7 G6PD VARIANTS IN POPULATION STUDIES

The frequencies and geographical specificities of the deficiency variants discussed above are summarised in Figure 1-12. Genetic analysis is by now a well established method of tracking population migration and admixture. The G6PD gene has a particularly high level of area-specific variation and is therefore well-suited to this type of study.

The close linkage of the neutral polymorphisms in the G6PD gene allows for the construction of informative haplotypes. The fact that G6PD is X-linked makes it particularly useful in haplotype analysis since the phase of the mutations is not a problem if determined in males. There is strong linkage disequilibrium between the polymorphisms and this means that few haplotypes out of the possible combinations occur, making them more specific and easier to track. The major problem is that, while there is a high level of variation in Africa, there tends to less variation in other populations and finding non-African population specific haplotypes is difficult.

The deficiency variants, however, are highly specific to geographic regions. For example, G6PD A^- is rare outside of Africa, whereas G6PD Canton, Mediterranean and Mahidol are rare or absent within Africa. Using a combination of the anlaysis of the deficiency variants and the non-deficient polymorphisms in haplotype analysis, G6PD analysis should provide an informative system for the study of the Malagasy population.

There are few data published on full haplotypes, especially in Africa and the G6PD locus has not been widely exploited for population studies. This is largely because most of the variation observed occurs in African populations. Some studies which have applied sites from these haplotypes to population studies include determination of the origins of the G6PD Mediterranean variant (Beutler and Kuhl 1990b) and the analysis of admixtures and founder effects in Taiwanese aboriginal groups (Tang *et al* 1995).

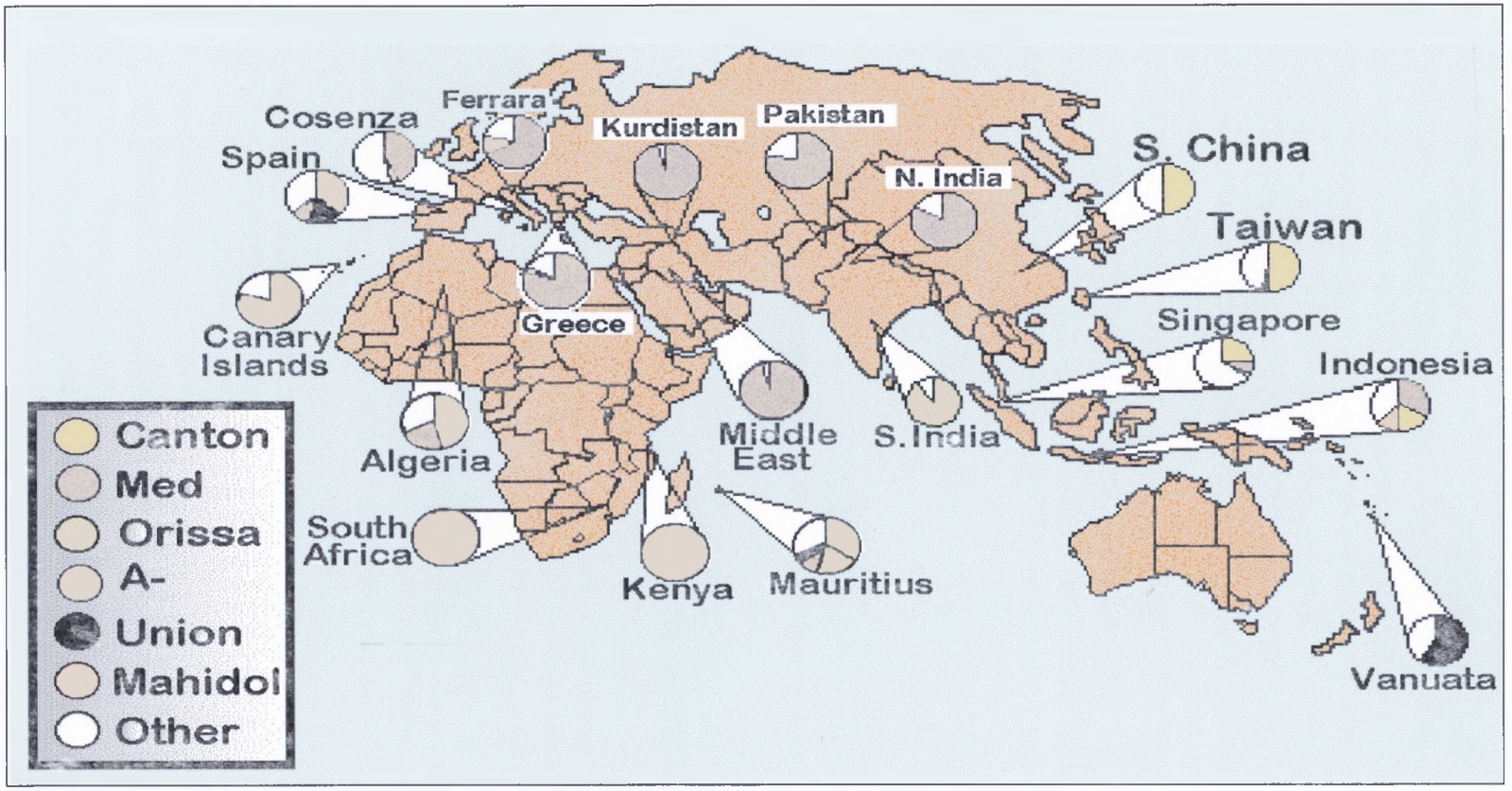


Figure 1-12 - Map showing the geographical locations of some of the major deficiency variants and their relative contributions to deficiency in those areas

The pie charts indicate the frequency of the variants, expressed as a proportion of the total deficiency in each area (see sections 1.2.6.3.2 and 1.2.6.3.3 for specific discussion of the variants).

1.3 AIMS OF THE STUDY

The aims of this study were:

1. To establish G6PD deficiency frequencies in the Malagasy groups.

2. To characterise the G6PD deficiency variants in the Malagasy population in order to obtain information on the origins and affinities of this population.

3. To construct haplotypes of variants and polymorphic sites in the G6PD gene in the Malagasy people, the African populations and the Indonesians to assist in their determination of the origins and affinities of the peoples of Madagascar.

CHAPTER TWO

SUBJECTS AND METHODS

2. SUBJECTS AND METHODS

Three different subsets of populations were sampled. The first was the Malagasy individuals themselves, and the second and third were African and Austronesian (Indonesian and Polynesian) populations. The Malagasy individuals were selected in two ways: (i) individuals were selected randomly from 6 different ethnic groups; and (ii) individuals already determined to be deficient haematologically were selected irrespective of ethnic group. Using the randomly selected samples it was possible to analyse the haplotypes (calculating frequencies in the randomly selected individuals and getting an idea of genetic distances and admixtures). In the G6PD deficient samples, the deficiency variants themselves were characterised and the high degree of geographic specificity of the variants used to provide information on the ancestral population. The African and Austronesian populations were selected in order to establish allele and haplotype frequencies in the putative parental populations.

The method used predominantly in this project is RFLP analysis by the polymerase chain reaction. Details of blood processing, serological testing, gel preparation and visualisation of the digested and pre-digested product will also be presented. The statistical methodology and background are also covered.

2.1 SUBJECTS

2.1.1 MALAGASY INDIVIDUALS

Peripheral blood samples from Malagasy individuals were collected during the course of field-trips conducted by this department (the Department of Human Genetics, South African Institute for Medical Research, Johannesburg, South Africa) from 1992 to 1996. The ethnic groups of the parents of the Malagasy individuals sampled were determined by questioning the individuals. DNA analysis of 347 unrelated Malagasy individuals was performed in this project.

2.1.1.1 RANDOMLY SELECTED INDIVIDUALS FOR DNA STUDIES

Since G6PD is X-linked, the phase of adjacent markers can be unequivocally determined in males. Males were therefore sampled preferentially, although females were also studied when the numbers of available males were small. The maternal ethnic group was considered to be the ethnic group of the males being sampled. For the females sampled, both the maternal and paternal ethnic groups are important in determining the ethnic group. These samples were studied at seven RFLP sites (discussed in section 1.2.6.3) and seven-site haplotypes were constructed.

Irrespective of deficiency status, 280 unrelated individuals were randomly selected from the following ethnic groups (based on the maternal ethnic group for the males and both maternal and paternal ethnic groups for the females):
Merina (a central highland group) - 63 males;
Betsileo (a central highland group) - 54 individuals (38 males and 16 females);
Antesaka (a southeastern lowland group) - 60 males;
Tsimihety (a northern lowland group) - 56 males;
Vezo (a southwestern lowland group) - 26 individuals (12 males and 14 females); and
Mahafaly (a southwestern lowland group) - 21 individuals (9 males and 12 females).

The Vezo and Mahafaly sample numbers are small. They have therefore been grouped together for some of the analysis, based on their geographical proximity.

2.1.1.2 G6PD DEFICIENT INDIVIDUALS

From a total of 1019 individuals tested for G6PD, 93 DNA samples (78 males and 15 females) from G6PD deficient individuals were obtained. Haematological analysis had been performed on all of the Malagasy individuals and these deficient samples were selected based on a chemiluminescence G6PD assay. The deficient samples were studied with the same systems as those in the random screen (i.e. with the seven RFLPs) to determine the type of deficiency alleles prevalent and the haplotypes associated with the alleles.

2.1.2 AFRICAN INDIVIDUALS

2.1.2.1 ZAMBIAN SAMPLES

Peripheral blood samples were collected from Zambian individuals on field-trips undertaken by this department (the Department of Human Genetics, SAIMR, Johannesburg, South Africa) in June 1995. A total of 50 unrelated Zambian individuals were studied in this project. The people of Zambia are ethnically heterogeneous with over 70 different ethnic groups (Dorling Kindersley 1995). The sample studied here is similarly heterogeneous (21 groups) and none of the groups sampled was large enough to study on its own. The Zambian population is a Bantu group which represents a possible Malagasy parental population from tropical southern Africa.

2.1.2.2 CENTRAL AFRICAN REPUBLIC (CAR) SAMPLES

Samples from 62 unrelated individuals from the Central African Republic (CAR) were collected in field-trips undertaken in 1995 by this department. These were divided into three language groups: 11 Mbindu individuals (2 males and 9 females), 17 Sangha Sangha individuals (15 males and 2 females) and 34 Gbaya individuals (29 males and 5 females). The samples were drawn from the CAR village of Ngola. This village contains approximately 200 inhabitants, speaking three distinct non-Bantu languages. Again, the groups were not large enough to analyse separately and were therefore pooled.

2.1.2.3 TSONGA SAMPLES

Fifty unrelated Tsonga samples (31 males and 19 females) were collected from random samples available in the department collected for previous research projects. The Tsonga are a southern African Bantu group located in close proximity to Mozambique.

2.1.2.4 ZULU SAMPLES

From the pool of samples available in the department, 42 unrelated Zulu samples were drawn and analysed. These samples were obtained from blood transfusion samples or from other miscellaneous sources. The samples selected comprised 34 males and 8 females.

2.1.3 AUSTRONESIAN INDIVIDUALS

2.1.3.1 INDONESIAN SAMPLES

Indonesian samples were obtained from three different regions: Ujung Pandang in South Sulawesi (45 males), Banjamarsin in South Kalimantan (Borneo) (47 males) and Central Sulawesi (52 males). These samples were provided by JB Clegg and D Rees of the MRC Molecular Haematology Unit of John Radcliffe Hospital in Oxford in 1994.

2.1.3.2 POLYNESIAN SAMPLES

The Polynesian group studied comprised 59 samples from unrelated Polynesian males. The samples are from the Polynesian islands of Tonga and Samoa and were kindly supplied by Dr DG Woodfield of the Department of Transfusion Medicine in Auckland. The sample is a mixed group of Maori, Samoans, Rarotongan, Tongan and Nuiean from various Polynesian islands. None of these groups was large enough to study independently and they were therefore combined in a single pooled Polynesian sample.

2.2 METHODS

2.2.1 PROCESSING OF BLOOD

Peripheral blood was collected from the subjects in ACD and EDTA tubes. Tubes were centrifuged to separate the red and white blood cells and plasma. The plasma and red blood cells were preserved for future analysis and the white cells (buffy coats) were stored at -20°C for DNA extraction.

2.2.1.1 HAEMATOLOGICAL TESTING

The 5ml EDTA tubes were sent to the Department of Haematology, SAIMR, where they were stored at 4°C and tested within 3-4 days of collection. Among the tests performed by the Department of Haematology of the SAIMR was a screen for G6PD deficiency. The technique used was a chemiluminescent assay, where the presence or absence of enzyme activity is determined. These results were used to select the G6PD deficient samples used in this project. The test detects deficiency status, but cannot identify specific variants or quantitate enzyme activity.

2.2.1.2 STARCH GEL ELECTROPHORESIS

Starch gel electrophoresis was performed in the Serogenetics Laboratory of this department by Dr AB Lane and his assistants, to determine the electrophoretic mobility of the samples. In practice, in Africa, the major distinction determined is the difference between the B and A mobility variants. The test can sometimes also detect some deficiency, simply by observing a lighter product and so it is possible to detect an A variant, an A deficient (e.g. A^-) variant, a B variant and a B deficient (e.g. Mediterranean) variant, but not to distinguish between, for example, two different B deficient variants (such as Mediterranean and Mahidol).

2.2.1.3 DNA EXTRACTION

DNA was extracted from the stored white blood cells using a salting out method (Miller *et al* 1988) or the Quiamp DNA extraction kit. In the salting out method, a detergent solution, sucrose-Triton X, is used to lyse the cells. Cellular protein is degraded using proteinase K (along with a chelating buffer, $T_{20}E_5$, containing EDTA and a detergent, SDS) and precipitated using a saturated NaCl solution. The precipitated protein is pelleted out and discarded and the genomic DNA is precipitated from the supernatant using absolute ethanol. The DNA is then dissolved in TE buffer. The genomic DNA was stored at -70°C or 4°C until required. All solutions are listed in Appendix A.

The Quiamp kit is produced by Qiagen Inc. The kit is a rapid DNA extraction method for extracting DNA from a variety of different tissues. It is often useful where small volumes of the tissue are available. In this project, the Quiamp technique was used primarily as a backup for the salting out technique, often when samples were degraded or when the quantities of blood available were small.

The DNA concentration was estimated crudely by running the high molecular weight DNA on 0.8% agarose gel alongside a sample of known concentration. The DNA was either used undiluted or a 1/10 dilution was used for PCR amplification.

2.2.2 VARIANTS STUDIED

The genomic DNA obtained from the peripheral blood was amplified using the polymerase chain reaction (PCR) and the variant restriction sites investigated using RFLP analysis. All of the samples were analysed at all seven of the RFLP loci. Only one enzyme cuts each site and a particular enzyme cuts only a single site. For the sake of convenience, therefore, the variants will be referred to by the restriction enzymes used to cleave them in each case. Details of the RFLP sites, the enzymes used to detect each mutation and the primer sequences are listed in Table 2-1. The references from which the primer sequences were obtained are also listed. The map location of the sites are shown in Figure 2-1. This figure was adapted from Vulliamy *et al* (1991b) and, since the same 7 haplotypes as those found by Vulliamy *et al* (1991b) were found in this project, the haplotype numbers are the same as those listed by Vulliamy *et al* (1991b). In addition, the *Mbo* II site, which determines the Mediterranean deficiency, is included. The "*Sca I*" was omitted since it is in complete linkage disequilibrium with the *Pvu* II site and therefore does not furnish any further information.

2.2.2.1 ELECTROPHORETIC MOBILITY VARIANTS

Starch gel electrophoresis determined whether a sample had the A or B mobility. The mutation causing this mobility shift can be confirmed using molecular methods by analysing the *Fok* I site. This was done for all samples. The *Fok* I (+) indicates an A mobility chromosome and *Fok* I (-) indicates a B mobility. Not all of the samples were analysed by starch gel electrophoresis and so no comparison between the molecular and biochemical analyses was made.

2.2.2.2 DEFICIENCY VARIANTS

The deficiency or non-deficiency of each sample was determined haematologically. The deficiency variants expected in this project were the common African deficiency, A^- (as discussed in section 1.2.6.3.1.2), and the most common non-African deficiency, G6PD Mediterranean (discussed in section 1.2.6.3.2.1). A^- is found on an A mobility (*Fok* I (+)) chromosome and Mediterranean variant occurs on a B mobility (*Fok* I (-)) chromosome. The most common A^- deficiency-causing mutation is detected using the enzyme *Nla* III, where *Nla* III (+) indicates the most common A^- mutation (202 G→A) and *Nla* III (-) indicates its absence. The rarer A^- mutations (680 G→T and 968 T→C) were not tested for. The Mediterranean variant is detected using *Mbo* II where *Mbo* II (+) indicates the presence of the Mediterranean variant and *Mbo* II (-) its absence.

2.2.2.3 POLYMORPHISMS

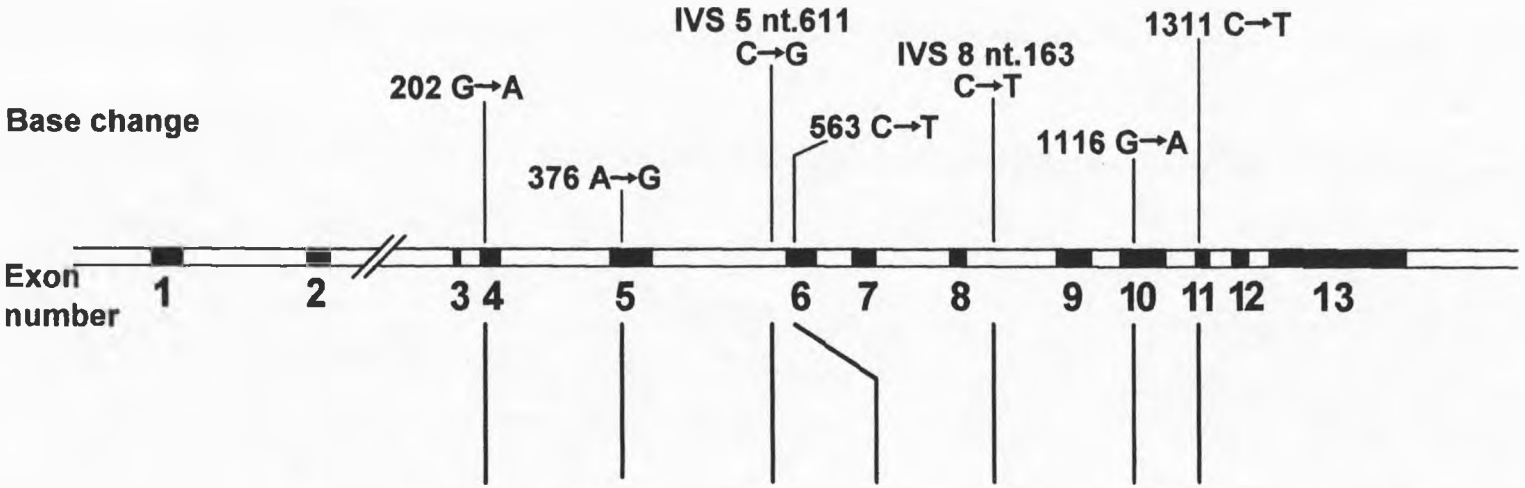
For all the samples, random and deficient, a further 4 RFLP sites (*Pvu* II, *Bsp* HI, *Pst* I and *Bcl* I) were analysed to construct a haplotype. Details of the primers used for each are shown in Table 2-1. These 4 sites, in combination with the 3 variants mentioned above (sections 2.2.2.1 and 2.2.2.2), were used to construct a 7-site haplotype for each sample. These haplotypes are shown in Figure 2-1, where a '+' indicates the presence of the site and a '-' indicates the absence of the site.

The G6PD Mediterranean variant is postulated to have two different origins based on the existence of two distinct associated haplotypes (Kurdi-Haidar *et al*, 1990). These haplotypes may be distinguished by the *Bcl* I polymorphism. The Middle Eastern associated haplotype is distinguished by a T at position nt1311 and is therefore *Bcl* I (+) (Kurdi-Haidar *et al* 1990). The Indian haplotype has a C at nt1311 and is *Bcl* I (-) (Beutler and Kuhl 1990b). This means that in the haplotypes shown in Figure 2-1, haplotype Med VII is equivalent to the haplotype common in the Middle East, and the Med I and Med II haplotypes are those common in India. The middle eastern haplotype is also common in the Mediterranean region (Beutler and Kuhl 1990b).

Table 2-1: Mutation locations, PCR primer sequences and the enzymes used to cut the fragments used in analysis of the G6PD locus

Mutation	Enzyme	Primer pair	Forward Primer Sequence	Reverse Primer Sequence	Ref
nt 563 C→T	*Mbo* II	G6B	ACT CCC GAA GAG GGG TTC AAG G	CCA GCC TCC CAG GAG AGA GGA AG	1
nt 202 G→A	*Nla* III	G6A2 (A-)	CAG CCA CTT CTA ACC ACA CAC CT	CCG AAG CTG GCC ATG CTG G	2
nt 376 A→G	*Fok* I	G6AI	CTG TCT GTG TGT CTG TCT GTC C	GGC CAG CCT GGC AGG CGG GAA GG	2
IVS 5 nt 611 C→G	*Pvu* II	G6PVU	AGG AGG TTC TGG CCT CTA CT	CTC CAC GAT GAT GCG GTT CC	3
IVS 8 nt 163 C→T	*Bsp* HI	G6BSP	GGA GCT AAG GCG AGC TCT GGC	TGC CTT GCT GGG CCT CGA AGG	4
nt 1116 G→A	*Pst* I	G6PST	TCC CTG CAC CCC AAC TCA AC	CAC CAT GTG GAG TCC CCC GG	3
nt 1311 C→T	*Bcl* I	G6BCL	TGT TCT TCA ACC CCG AGG AGT	AAG ACG TCC AGG ATG AGG TGA TC	5

1. Hirono and Beutler 1988
2. Nafa *et al* 1994
3. Beutler and Kuhl 1990a
4. Vulliamy *et al* 1991b
5. Kurdi-Haidar *et al* 1990



Haplotype	*Nla* III	*Fok* I	*Pvu* II	*Mbo* II	*Bsp* HI	*Pst* I	*Bcl* I	G6PD
I	-	-	-	-	+	+	-	B
VII	-	-	-	-	+	+	+	B
II	-	-	-	-	+	-	-	B
Med I[1]	-	-	-	+	+	+	-	B-
Med VII[2]	-	-	-	+	+	+	+	B-
Med II[1]	-	-	-	+	+	-	-	B-
III	-	+	-	-	+	+	-	A
IV	-	+	-	-	-	+	-	A
V	-	+	+	-	-	+	-	A
VI	+	+	+	-	-	+	-	A-

Figure 2-1: The map positions and expected haplotypes for the 7 variants studied in this project (after Vulliamy et al, 1991b)

[1]Med I and Med II indicate the possible Indian associated G6PD^{Mediterranean} haplotypes (i.e. *Mbo* II (+), *Bcl* I (-)) (Kurdi Haidar *et al* 1990).
[2]Med VII indicates G6PD^{Mediterranean} haplotype found in the middle east (*Mbo* II (+), *Bcl* I (+)) (Kurdi-Haidar *et al* 1990).

2.2.2.4 PCR AMPLIFICATION

All samples were amplified in 15µl PCR reaction mixes. These mixes were made up as follows: approximately 100ng DNA; 1mM dNTP; 1x Taq polymerase buffer; 1.2µl of a 1/150 dilution of a 1µg/µl primer stock and 1 unit of Taq polymerase. A 1/40 dilution of 0.1M spermidine was made up and 0.3µl of the dilution was added to some samples. The primer sequences are listed in Table 2-1. PCR amplification was initially performed on a Perkin Elmer Cetus thermocycler.

All seven of the RFLP systems used the same reaction mix listed above. The only variable that needed to be adjusted during PCR optimisation was the annealing temperature of each system. The annealing temperatures for the 7 systems are shown in Table 2-2. The PCR conditions used were as follows: denaturing: 94°C for 30 seconds, annealing: variable (see Table 2-2) for 30 seconds, extension: 72°C for 45 seconds. The cycle was repeated 30 times.

The PCR reactions were subsequently re-optimised on a newer machine in the laboratory, the Hybaid Omnigene thermocycler. In re-optimising for the Omnigene, the length of time taken for each cycle was adjusted. This was required because the Cetus has a cooling unit whereas the Omnigene is air-cooled. The result is that the Cetus has the ability to adjust to different conditions quicker than the Omnigene. All steps were increased to 1 minute each (still with 30 cycles). The annealing temperature had to be readjusted for the Omnigene and these temperatures are also shown in Table 2-2.

Table 2-2: Annealing temperatures (T°) used for the Cetus and Omnigene thermocyclers for the 7 RFLP systems studied at the G6PD locus and the expected sizes of fragments before and after digestion with the appropriate restriction enzyme

	Annealing Temperature		Restriction fragment sizes before and after digestion		
System	**Cetus T°**	**Omnigene T°**	**Uncut (bp)**	**-**	**+**
Mbo II	67°C	63°C	547	377+119+26+25	277+119+100+26+25
Nla III	65°C	61°C	353	338+15	215+123+15
Fok I	65°C	59°C	295	295	154+141
Pvu II	63°C	59°C	140	140	90+50
Bsp HI	67°C	61°C	794	794	293+501
Pst I	67°C	59°C	587	587	374+213
Bcl I	61°C	59°C	203	203	180+23

2.2.2.5 PREPARATION OF GELS

All of the gels used in this study were made up following the same basic protocol. A gel mix of 300 ml (in 1xTBE) was made and dissolved on a hotplate or in a microwave oven. Ethidium Bromide (EtBr) was added to the liquid gel after dissolution and cooling. The gel was then poured and allowed to set. The final EtBr concentration for all gels except the 5% Metaphor was 0.03μg/ml. The EtBr concentration of the 5% Metaphor gel was doubled to 0.06μg/ml. Four types of gels were used in this project.

A 0.8% agarose (HGT-FMC Bioproducts) gel was used to examine high molecular weight DNA to estimate the quantity and assess the quality of the DNA as discussed in section 2.2.1.3. A 2.5% agarose (HGT-FMC Bioproducts) gel was used for visualising PCR products prior to digestion. A 4% composite agarose (2:2 HGT:NuSieve - FMC Bioproducts) gel was used for examining digestion products.

A 5% Metaphor agarose (FMC Bioproducts) gel was used for examining *Bcl* I digestion products. Metaphor gel is a high quality, high resolution gel. The high resolution was required to distinguish digestion products for *Bcl* I which were very similar in size (180bp and 203bp). The gel was prepared on a hot plate since dissolution was slow and the solution tends to boil over in the microwave oven. Because the gel needed to be run for a long time (5-6hrs at 100-120V) the EtBr tended to fade and this is why a higher concentration of EtBr (0.06μg/ml) was used. This was sometimes not sufficient and the gels needed to be re-stained before results could be obtained.

2.2.2.6 VISUALISATION OF PCR PRODUCTS

Five microlitres of each PCR product was loaded with Ficoll loading dye on 2.5% agarose gel made in 1XTBE. Samples were electrophoresed at 100-120V for approximately 1hr, or until distinct bands could be visualised. As these gels contained EtBr, products were visualised under UV light. The PCR product had to be visualised prior to digestion to determine that the genomic DNA had been amplified successfully so as not to waste expensive enzymes. A molecular weight marker, 1kb λ ladder was loaded in an adjacent well to determine the sizes of the bands produced. The sizes of the undigested fragments are shown in Table 2-2.

2.2.2.7 RESTRICTION ENZYME DIGESTION

The 10µl of PCR product remaining after PCR product visualisation was digested according to the manufacturers' instructions. All digests were performed at 37°C except for the *Bcl I* digest which was performed at 61°C. A 30µl digestion reaction was set up using 5 units of enzyme in each sample. The sizes of the resulting bands are shown in Table 2-2.

2.2.2.8 VISUALISATION OF DIGESTED PRODUCTS

The entire 30µl digest for each sample was run on agarose gel for visualisation. For all systems except the *Bcl* I system, the samples were run on 4% composite agarose gel (as detailed in section 2.2.2.5 above). The *Bcl I* fragments to be resolved were so similar in size (203bp and 180bp) that a 5% Metaphor gel was required to resolve them, as shown in Figure 2-2. Figures 2-2, 2-3 and 2-4 are examples of electrophoresed digests. The other systems were based on the same principle but are not shown here.

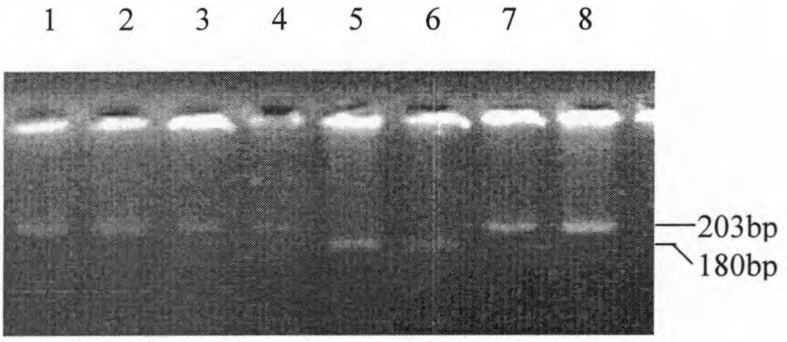


Figure 2-2: Digitised photograph of the *Bcl I* digest run on 5% Metaphor gel

Lanes 1, 2, 3, 4, 7 & 8 show the *Bcl I* - type (203bp). Lanes 5 and 6 show samples which are *Bcl* I + (180bp). The 23bp fragment in the *Bcl I* + samples has run off and therefore cannot be seen, but the resolution allows for clear distinction between the 180bp and 203bp fragments.

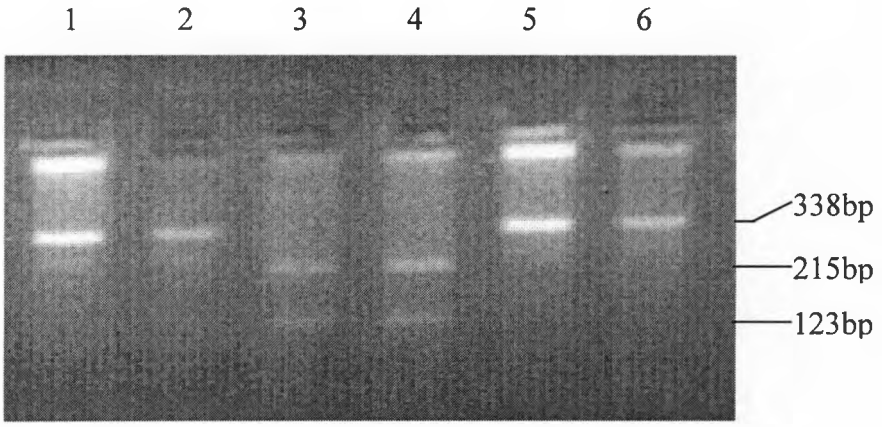


Figure 2-3: Digitised photograph of a sample of the *Nla III* digest run on 4% composite agarose gel

Lanes 1, 2, 5 & 6 contain *Nla III* -ve samples (338bp). Lanes 3 & 4 contain *Nla III* +ve samples (215bp+123bp) indicating the presence of the G6PD A- deficiency variant. The 15bp fragment has run off.

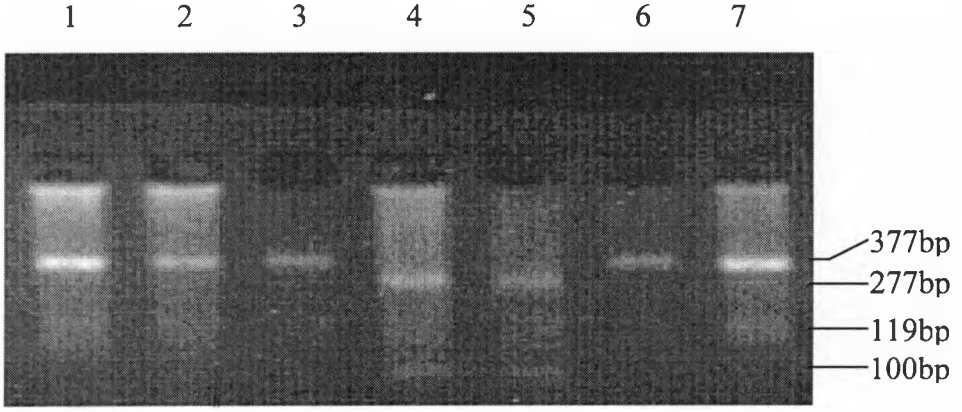


Figure 2-4: Digitised photograph of the *Mbo II* digest run on a 4% composite agarose gel

Lanes 1, 2, 3, 6 & 7 contain *Mbo II* -ve samples (377bp+119bp+26bp+25bp). Lanes 4 & 5 contain *Mbo II* +ve samples (277bp+119bp+100bp+26bp+25bp) i.e. individuals with G6PD Mediterranean deficiency. The smaller fragments in each case (26bp and 25bp) are not visible and the 119bp fragments are very faint. However, the distinction between the 377bp and 277bp fragments is clear.

2.2.3 HAPLOTYPE CONSTRUCTION

The results at the seven RFLP loci were used to generate haplotypes for maximum informativity. The expected haplotypes based on haplotypes previously reported in individuals of predominantly African ancestry (Vulliamy *et al* 1991b) are shown in Figure 2-1. In addition the *Mbo* II site, which determines the Mediterranean deficiency (as discussed in section 1.2.6.3.2.1), has been included and the expected associated haplotypes (middle eastern and Indian as discussed in section 1.2.6.3.2.1) are identified.

Since the G6PD locus is on the X chromosome, haplotype construction is easy in males since one does not have to determine phase. However, where females were used, the phase needed to be determined. It was assumed that the haplotypes would correspond with those most frequently encountered (Figure 2-1), based on published findings (Vulliamy *et al* 1991b). This is a justifiable assumption because the same 7 haplotypes were found in the males studied in this project. The haplotypes constitute a small proportion (7) of the number of possible haplotypes for 7 sites, namely 128, as a result of the strong linkage disequilibrium which exists between the alleles at the 7 loci. Using this assumption, the haplotypes in females were inferred (where heterozygotes were identified) based on the most likely haplotypes observed.

For both the male and female groups, results could sometimes not be produced for all seven sites. For a number of the samples, therefore, full haplotypes could not be generated because a result for a critical site was missing. Data for these samples were not used in the haplotype analysis, but they were used in the allele frequency analysis.

2.2.4 STATISTICAL ANALYSIS OF THE DATA

2.2.4.1 FREQUENCY DISTRIBUTIONS

The 3 types of frequencies calculated were: G6PD enzyme deficiency frequencies, allele frequencies (including deficiency variant allele frequencies) and, after construction of haplotypes, haplotype frequencies.

2.2.4.1.1 ENZYME DEFICIENCY FREQUENCY

For those populations where G6PD deficiency results were available, the enzyme deficiency frequency was determined both for the total sample and for each ethnic group in each geographical region. These results can be useful in comparisons to other published frequencies and can also be useful in comparisons between and within populations. One can compare the observed frequencies between the Malagasy groups and between published reports on the putative ancestral populations (African and Indonesian). Because of the wide range in frequency in Africa, this could possibly also help to pinpoint specific groups in Africa as putative ancestral groups.

The enzyme deficiency frequency can only be calculated for males because of the variability of expression in females. Females will usually only express as deficient if they are homozygous (or compound heterozygotes) for deficiency variants, meaning that the estimate of deficiency would be lower than expected from the male frequency. We could not calculate a gene frequency based on the assumption that females are haematologically deficient only if they carry two deficient variants because approximately 10% of heterozygotes (i.e. carrying only one deficiency variant) can be expected to be haematologically deficient, presumably due to non-random X-inactivation (WHO working group 1989). For simplicity and accuracy therefore, the deficiency frequency was calculated only in males.

2.2.4.1.2 ALLELE FREQUENCIES

The allele frequencies for the seven sites (deficient and non-deficient) were determined and compared to other published reports of these frequencies. Calculating these frequencies in the populations and comparing these to previous publications is useful in the characterisation of the populations. It is necessary to compare populations at the allele level since very few publications have reported on the full seven-site haplotype analysed in this project. Many of the variants are African-specific and studies performed on non-African populations therefore have reported on fewer sites. There are few molecular data available on African populations, limiting comparison potential.

2.2.4.1.2.1 LINKAGE DISEQUILIBRIUM

Once allele frequencies had been determined, they were used to calculate linkage disequilibrium between alleles. This was done by calculating D values. The observed association between two alleles was compared to the values which would have been expected if this association was a chance event (i.e. for independent events, the expected frequency is the product of the individual allele frequencies for that association). The difference between the observed and expected values (i.e. the D value) is a measure of linkage disequilibrium between the alleles. These values were compared to the values reported by Vulliamy *et al* (1991b).

It was determined whether or not these values were significant by χ^2 analysis. Again, these findings were compared to other published reports. A worked example showing the method used is presented in Appendix B (I).

2.2.4.1.3 HAPLOTYPE FREQUENCIES

Haplotypes were constructed (as discussed in section 2.2.3) and the frequencies of the haplotypes in all population samples were determined. These were compared to other studies. A number of alleles analysed together in haplotypes are more powerful in comparisons than are the alleles considered individually.

2.2.4.1.3.1 HAPLOTYPE FREQUENCY HISTOGRAMS

The haplotype frequency data were used to generate histograms. This gives a visual impression of the similarity or dissimilarity of the populations. The histograms show both the comparisons of the haplotypes and, because the deficiency variants are included within the haplotypes, the association of the haplotype with a particular type of deficiency in different populations.

2.2.4.2 CHI-SQUARE χ^2 ANALYSIS

The haplotype frequencies were compared between populations using chi-square (χ^2) analysis to determine whether significant differences existed between the populations. χ^2 analysis was also used to determine whether or not differences existed between groups within the same population. For example, 3 groups were compared within the CAR population to determine whether or not it was justifiable to combine all of the CAR samples into a single group for further analysis.

χ^2 analysis is based on the equation:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

The null hypothesis is that there is no difference between the observed and expected figures. The observed figures in this case are the haplotype numbers observed in the first population. The expected values are generated by normalising the second population with respect to the size of the first, so that differences in population size do not bias the estimate. The χ^2 value is therefore a quantification of the difference between the two populations and is used to generate a p value. A p value of less than 0.05 (i.e. 95% confidence interval) is considered statistically significant in this study. A worked example of the χ^2 analysis is shown in Appendix B (II).

The χ^2 analysis is a crude measure of distance between different populations. A large χ^2 value (small p value) between two populations indicates a large difference (and therefore genetic distance) between them compared to a small χ^2 value. A more accurate measure of the distance between two populations is obtained by calculating the genetic distance.

2.2.4.3 CALCULATION OF GENETIC DISTANCES

Genetic distances between the populations were calculated using Nei's standard distance calculation (Nei 1972, discussed in Chakraborty 1986). Nei's standard genetic distance calculation determines a simple deviation from the mean. It can be calculated without the bias of population size as follows:

$$I = \frac{1 - a - b}{1 - b}$$, and

$$D = -\ln I$$

where D is Nei's standard genetic distance, I is genetic identity (a measure of the probability of drawing identical alleles from two or more populations), a is the measure of between-population heterozygosity and b is a measure of the within-population heterozygosity.

In this project, the genetic distances were calculated using the computer program DISPAN (Genetic Distance and Phylogenetic Analysis), written by Tatsuya Ota of the Pennsylvania State University. Calculation of genetic distances between populations is in itself useful for quantifying the difference between populations. The distance data can also be used to construct phylogenetic trees for visual assessment of the grouping of the populations.

2.2.4.4 TREE CONSTRUCTION FROM GENETIC DISTANCES

There are a number of different possible ways to draw phylogenetic trees (reviewed in Cavalli-Sforza *et al* 1994). The most widely used are the neighbour joining (NJ) tree and the Unweighted Paired-Group Method using arithmetic Averages (UPGMA) tree. The use of the UPGMA tree is appropriate when the evolutionary rate is considered to be constant. Since the "alleles" examined in this section are actually haplotypes, the rate of evolution can not be considered to be constant and the use of this tree is therefore not appropriate. The method utilised exclusively in this project is the NJ tree.

DISPAN calculates Nei's standard genetic distances from population frequency data and draws trees based on these data. It then establishes the statistical accuracy of the fit using bootstrapping. This is done by generating a number of random distance matrices derived from the original matrix and using these to draw trees. The proportion of these trees agreeing with the original tree is then calculated. The bootstrap value is the number of times the new data set is generated. In this project, a bootstrap value of 1000 was used in each case.

The construction of trees is a useful way of graphically representing the genetic distance data. From the tree, the relationships between populations can be viewed as clustering of branches. The lengths of the branches is proportional to the genetic distances between the populations and their relative rates of differentiation. Rotation of the nodes (i.e. swapping branches at a node) of a good tree should result in no change in the interpretation of the data.

2.2.4.5 PRINCIPAL COMPONENT ANALYSIS

Principal component (PC) analysis is a method of simplifying a multivariate dataset using singular value decomposition. It involves simplifying multivariate data for visual representation in such a way that there is a minimum of data loss. The computer program used to perform the PC analysis in this project is ANTANA version 1.5 (written by Henry Harpending and Alan Rogers of Pennsylvania State College).

During PC analysis, the population-by-gene-frequency data are represented in multidimensional space depending on the number of alleles. So for a biallelic system, studied in 7 populations, the populations are plotted on a 2-dimensional plane with the one axis representing the first allele and the second axis representing the second allele. The frequencies of the alleles for the populations are plotted with the first allele frequency as the abscissa and the second allele frequency as the ordinate. The 7 populations therefore produce a cluster of 7 points. A line is then drawn as close to the 7 points as possible, such that the sum of the differences between the points and the line is minimised. This is similar to drawing a line of regression. This "regression line" becomes the first principal component which has an arbitrary scale and the 7 points are plotted on it at intervals determined by their perpendicular projection onto the line. In this way, the data from two dimensions are reduced to 1 dimension. Similarly, p alleles may be represented in p-1 dimensions, with p-1 principal components.

Some loss of data occurs in this reduction process, although it is minimised. For the analysis of many "alleles" (in this project, 7 haplotypes), even this reduction is not sufficient for visualisation purposes (i.e. it is impossible to conceptualise and draw a six-dimensional graph). Usually, however, two or three of the principal components contain the majority of the data. (The different principal components, because they are perpendicular to each other, contain no duplicated information.) The amount of data contained in the two or three principal components used can be determined by analysis of the eigenvalues (the proportion of information contained in each principal component) or the singular values (the square roots of the eigenvalues of the covariance matrix) of the data matrix.

Once it has been determined which of the principal components contain the majority of the data (based on the singular or eigenvalues), the two principal components containing the majority of the data can be plotted against each other and the full dataset can be represented in two dimensions with little loss of data. In this way, clustering of the populations can be observed and interpreted based on the significance of the axes (the principal components) as determined by the ratio of the singular values to each other. The closer the populations cluster, the smaller the differences between them and the greater their affinity.

2.2.4.6 ADMIXTURE CALCULATIONS

Admixture was calculated using three different methods. The reason that more than one method was used, was for comparison of the results produced by different methods and because some of the methods have limitations that the others do not have.

2.2.4.6.1 THE BERNSTEIN METHOD

Bernstein was the first person to estimate population admixture based on allele frequencies. The method that he used was appropriate only for bi-parental populations. It is used in this analysis by comparing a single African population with a single Austronesian population at a time in estimating their proportionate genetic contribution to the Malagasy gene pool. The advantage of the Bernstein method is its simplicity, but its drawback is the limitation of considering only 2 parental populations.

The principle proposed by Bernstein (as discussed in Chakraborty 1986) can be summarised in the following equations. If p_1, p_2,...,p_k are the frequencies of a particular allele in k parental populations, and these populations contribute m_1, m_2,...,m_k fractions to the hybrid population ($0 \leq m_i \leq 1$; $\Sigma m_i = 1$) then the expected frequency of that allele in the hybrid population, P_h, will be given by:

$$P_h = \sum_{i=1}^{k} mi \cdot pi$$

This equation, because of the number of variables, is only appropriate for determining the m values (admixture fractions) for hybrid populations of bi-parental origin. This is calculated for the first parental population as follows:

$$P_h = m_1 \cdot p_1 + (1 - m_1) \cdot p_2$$

$$\therefore m_1 = \frac{(p_h - p_2)}{(p_1 - p_2)}$$

The m value for the second parental population (m_2) is calculated by: $m_2 = 1\text{-}m_1$.

2.2.4.6.2 ROBERT AND HIORNS' LEAST SQUARES METHOD

A worked example of the calculations used here appears in Appendix B (III). The calculations developed by Roberts and Hiorns were based loosely on the Bernstein equations, but were extended to more than two parental populations. The calculations provide a least squares estimate of the admixture proportions. This means that it is the minimum estimate of the sum of the squares of the difference between the hybrid population frequency and the proportional contribution of each parental population.

The method of calculation for the least squares method as proposed by Roberts and Hiorns (as discussed in Elston 1971), is as follows. If there are P (>1) parental populations, let $X=(x_{ij})$ be a matrix of k rows with p columns, x_{ij} representing the i^{th} allele in the j^{th} population (i=1,2,...,k; j=1,2,...p) and y be a vector of dimension k representing the allele frequencies of the hybrid population. Then

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

where X' is the transposed matrix of X, and $m'=(m_1,m_2,...,m_p)$ is the proportional contributions of the p parental populations as in the previous calculation method of Bernstein, provided that $X'X$ is non-singular. However, since the allele frequencies at a locus always sum to 1, if all of the alleles at a locus are used, then $X'X$ would be singular and the calculation would be invalid. Roberts and Hiorns therefore arbitrarily excluded an allele from the matrix X.

2.2.4.6.3 ELSTON'S LEAST SQUARES METHOD

A worked example of the calculations used here appears in Appendix B (IV). Elston (1971) said that the technique of arbitrarily discarding one of the alleles in the original parental population frequency matrix used by Roberts and Hiorns was unnecessary. He also showed that the m values could differ significantly depending on which allele was discarded. He therefore adapted the least squares method so that one of the alleles need not be discarded. This was done by generating a second matrix, X*, from the original matrix X. X* is a (k x p-1) matrix generated as follows. The j^{th} column of X* is obtained as $x^*_j=P_j-P_p$. The vector y* is obtained by $y=P_h-P_p$. The calculation of the m values for the first p-1 parental populations is then

$$m = \left(X^{*\prime} X^*\right)^{-1} X^{*\prime} y^*,$$

again provided that $X^{*\prime}X^*$ is non-singular. The m value of the p^{th} population is given by:

$$m_p = 1 - \sum_{j=1}^{p-1} m_j.$$

Elston added that if any of the m values were negative, the corresponding parental populations should be discarded from the analysis i.e. the m values should be re-computed with the smallest m value set to 0. In this analysis, both the Zulu and Polynesian groups had to be discarded from analysis for this reason.

CHAPTER THREE

RESULTS AND DISCUSSION

3. RESULTS AND DISCUSSION

The following chapter is a presentation of results and discussion of the results, while chapter 4 provides a general discussion of the trends observed. This has been done to avoid repeating results when discussing them. Section 3.1.1 presents results of enzyme deficiency in the total Malagasy group. Section 3.1.2.1 is a discussion of the origins of the population specific deficiency variants found in the samples chosen on the basis of enzyme deficiency. Sections 3.1.2.2 to 3.1.2.4 is a discussion of the remainder of the non-deficiency variants in the randomly selected sample. Sections 3.2 to 3.6 involve analysis of the haplotypes constructed from all of the alleles (including deficiency alleles) for the randomly selected sample.

3.1 FREQUENCY DISTRIBUTIONS

The frequency distributions which are presented and discussed here are the enzyme deficiency, the allele frequencies and the haplotype frequencies. The G6PD enzyme deficiency frequency is obtained from haematological analysis as described in section 2.2.1.1. The allele frequencies presented are those of all seven of the RFLP sites studied, including those identifying the deficiency variants ($G6PD^{A-}$ and $G6PD^{Mediterranean}$) and the mobility variant A. The allele frequency data were used to calculate linkage disequilibrium between alleles at sites within the haplotype. The haplotype frequency data show the distributions of the haplotypes in all populations and are used in all statistical calculations hereafter.

3.1.1 G6PD ENZYME DEFICIENCY FREQUENCY

The frequency of enzyme deficiency is calculated only in males in this project. In this project, the only populations for which haematological deficiency data were available were the Malagasy and the CAR samples.

3.1.1.1 ENZYME DEFICIENCY IN AFRICAN POPULATIONS

3.1.1.1.1 CAR POPULATION ENZYME DEFICIENCY FREQUENCY

Out of the 62 male CAR samples selected for DNA studies for the 7 molecular systems, 2 (0.032) were found to be G6PD deficient. Compared with other published frequencies, this figure is rather low. The WHO reported a frequency of 15-26% for this area (WHO working group 1989) and frequencies in Zaire range from 0.054 in non-Bantu speaking individuals to 0.235 in Bantu-speaking individuals (Motulsky *et al* 1966). However, an unpublished report by Cavalli-Sforza (reviewed in Betke *et al* 1967) indicates a deficiency frequency of 4% for the Babinga pygmies in the Central African Republic, and this correlates more closely with the results in this project. In Zaire, a frequency of 0.061 was found in the non-Bantu-speaking Ngabaka (Motulsky *et al* 1966). Of all of the people tested in this report, the Ngabaka people live closest to the southern CAR border.

Both of the deficient individuals were from the Sangha Sangha language group. However, the chi square comparison between the language groups (section 3.2.1.2, Table 3-11) indicates that the CAR is a homogeneous group and so there is no particular reason to report the deficiency frequency as being higher in the Sangha Sangha group than in the other CAR groups. Both of the samples found to be enzymatically deficient were identified by molecular analysis as having the A^- (*Nla* III) deficiency variant. The molecular results therefore confirm the enzyme activity results and the A^- frequency in this groups is a good measure of total G6PD deficiency frequency. For both samples, the type of deficiency was the A^- type (*Nla* III). No Mediterranean deficiency variants were found. The total deficiency found in this group is therefore accounted for by a single variant (A^-), although it would be inappropriate to make too much of this based on only two deficient samples.

3.1.1.1.2 ENZYME DEFICIENCY FREQUENCIES IN OTHER AFRICAN POPULATIONS

From previously published reports (Beutler *et al* 1989), the largest proportion (if not all) of the deficiency in Africa can be expected to be due to the common African variant A^- studied here. The frequency of the A^- variant in males is therefore probably a good estimate of enzyme deficiency frequency in the African samples tested in this project. In general, there is a decrease in the frequency of the A^- variant from north to south in sub-Saharan Africa (as discussed in section 3.1.2.1.2, Figure 3-1) (Jenkins 1972). In a large study of various ethnic groups, the G6PD deficiency frequency was found to range from 0.000 to 0.333 (Jenkins 1972).

Based on the presence of the *Nla* III site in males, the frequency of deficiency in the Zambian population is higher than in the CAR populations, 4/16=0.250. Comparison with other published data reveals comparable results. In Yaka villages in the southwest of Zaire (close to the Zambian border), the frequency was 0.234 (Motulsky *et al* 1966) and in a large survey of many Zambian ethnic groups, the deficiency frequency ranged from zero to 0.333 (Barclay *et al* 1970).

Based on the presence of the *Nla* III site in males, the frequency of deficiency in the Tsonga sample is 1/31 (0.032). This frequency is comparable to the deficiency frequency of 0.053 in the Tsonga group reported by Hitzeroth and Bender (1980).

3.1.1.2 ENZYME DEFICIENCY IN AUSTRONESIAN POPULATIONS

The Austronesian populations studied in this project (Indonesian and Polynesian) were not screened for G6PD deficiency and no deficient variants were found in the molecular analysis. It is therefore difficult to draw conclusions regarding deficiency frequency or type in this sample. Published frequencies of total deficiency in Indonesia indicate a frequency of 0.136 (Soemantri *et al* 1995). The variants found in Indonesia are mainly the Canton and Mahidol variants, although the Mediterranean variant accounts for approximately 0.313 of the total deficiency. We could therefore expect a frequency of approximately 4% (0.136x0.313) for the Mediterranean variant in the Indonesian sample. The frequency reported in this study is low, possibly due to a sampling bias. Other reports of deficiency frequency in Indonesian regions include 0.02 and 0.11 in two groups in Bali (Brequet *et al* 1982), and the deficiency frequency in Chinese and Malay males in Singapore are 0.035 and 0.202 respectively (Saha and Banerjee 1971). The G6PD deficiency frequency in southeast Asia tends to be lower than that in African regions from which putative Malagasy ancestral populations are thought to have arisen.

3.1.1.3 ENZYME DEFICIENCY IN THE MALAGASY ETHNIC GROUPS

Table 3-1 shows a breakdown of the Malagasy G6PD deficiency frequencies by ethnic group and region. The highest regional frequency (0.182) was found in the west, which is an area thought to be populated by people of characteristically African origin. Comparing this to published reports on African deficiency frequencies (section 3.1.1.1), this high frequency is expected. Low regional frequencies were found in the Indonesian-like (Campbell 1996) populations of central Madagascar (0.089). Again, the low frequencies are not surprising given the low frequencies reported in Indonesia (section 3.1.1.2). Low G6PD deficiency frequencies were also, suprisingly, found in the southwest (0.073), which is thought to have a more African-like population (Campbell 1996).

The highest ethnic group deficiency frequency is found in the Makua, the most African ethnic group (Campbell, The Origins of the Malagasy, manuscript in preparation). High frequencies were also found in other lowland groups such as Antemoro and Antesaka in the east central coastal region. Deficiency frequencies in the highland groups were low (0.074 in the Merina and 0.093 in the Betsileo). Some groups had deficiency frequencies of zero, although the sample number for these groups was small and the absence of deficiency may represent sampling bias.

Out of a total number of 1019 male subjects sampled in the field trips to Madagascar, 134 (0.133) were found to be deficient using the chemiluminescent screening test. This overall frequency of deficiency correlates with other published frequencies. The WHO reports a frequency of 10-14.9% in Madagascar (WHO working group 1989), while Dodin (reviewed in Betke *et al* 1967) indicates a frequency of 14-16%. The deficiency frequencies on the east coast of Africa (Kenya and Tanzania), a likely location for the Malagasy African ancestral population, range from 17% to 23%. The frequency was 28% in Nairobi and 15% near Lake Victoria (Allison 1960). The average deficiency frequency reported in various groups in Mozambique is 19% (Reys *et al* 1970), and this also corresponds with deficiency frequencies found in many of the lowland Malagasy groups.

Table 3-1: G6PD deficiency frequencies according to region and ethnic group for Malagasy males determined by haematological screening

Region and Ethnic Group	No. Tested	Deficient	
		No.	Frequency (±SE)
North:	**106**	**17**	**0.160 (0.036)**
Antankarana	14	0	0
Tsimihety	88	17	0.193 (0.042)
Sihanaka	4	0	0
Central:	**259**	**23**	**0.089 (0.018)**
Merina	108	8	0.074 (0.025)
Betsileo	129	12	0.093 (0.026)
Bezanozano	22	3	0.136 (0.073)
East Central:	**430**	**72**	**0.167 (0.018)**
Betsimisaraka	97	14	0.144 (0.036)
Antemoro	77	17	0.221 (0.047)
Antefasy	24	2	0.083 (0.056)
Antambahoaka	35	4	0.114 (0.054)
Antesaka	143	27	0.189 (0.033)
Tanala	9	0	0
Antanosy	45	8	0.178 (0.057)
Southwest:	**109**	**8**	**0.073 (0.025)**
Antandroy	41	3	0.073 (0.041)
Bara	22	1	0.045 (0.044)
Vezo	21	2	0.095 (0.064)
Mahafaly	20	2	0.100 (0.067)
Masikoro	5	0	0
West:	**77**	**14**	**0.182 (0.044)**
Sakalava	61	9	0.148 (0.045)
Makua[1]	16	5	0.313 (0.116)
Total[2]	**1019**	**136**	**0.133 (0.011)**

[1] The Makua group is not an official ethnic group but is an unofficial grouping, mainly originating from slave peoples (Campbell, The Origins of the Malagasy, manuscript in preparation).

[2] The remaining 38 samples (making up a total of 1019) were individuals for whom ethnic group was not established.

There are a number of conclusions which can be drawn from these results. The high deficiency frequencies in the lowland groups could be due to a high G6PD deficiency frequency in the African ancestral population and the initial colonists of Madagascar. The deficiency frequency in parts of Africa from which the Malagasy are thought to have originated is slightly higher than that of Indonesia according to published reports. However, the deficiency frequencies in the hybrid populations may not be representative of the ancestral populations for a number of reasons. If a small number of founders were responsible for colonising the island, and the frequency in these founders was (by chance of sampling bias) smaller (or greater) than that of the parent populations from which they came, the resulting frequency in Madagascar would also not be representative of the ancestral populations. Furthermore, the G6PD deficiency frequency could be expected to be affected by malarial selection within Madagascar. There is marked climatic variation within the island. The central highland is more temperate, while the east coast regions are tropical with a high rainfall and the southwest lowland region is more temperate and arid. This climatic variation could result in different malarial distribution and therefore different selective pressures. These figures for enzyme deficiency are consistent with higher African contribution in most of the lowland groups as opposed to greater Indonesian admixture, but are also consistent with a greater degree of selection due to increased malaria frequency in the low-lying areas. The specific variants involved could also have different selective advantages.

The time span involved could also contribute to the different frequencies. The Makua group from Mozambique was probably the last group to colonise the island and this group has the highest deficiency frequency. With time, the frequency could decrease due to the selective pressure on the island being less than that in the homeland areas of the respective ancestors.

It is most likely that the distribution observed on the island today is the result of a combination of all or some of the above factors.

3.1.2 ALLELE FREQUENCIES FOR THE 7 RFLP SITES

The frequencies of the alleles for the 7 RFLP sites described in section 2.2.2 were determined for the randomly selected individuals in each of the 13 populations. These allele frequencies for the random individuals from the Malagasy groups are listed in Table 3-2 and the frequencies for the non-Malagasy populations are shown in Table 3-3.

The allele frequencies are reported for the entire sample, male and female. The most obvious features in Tables 3-2 and 3-3 are the large proportion of polymorphisms in the African and Malagasy populations compared to the Austronesian populations. There are also suprisingly few polymorphisms present in the Vezo and Mahafaly Malagasy groups. As discussed in section 1.2.6.3.1, many of the sites which are polymorphic in Africa are not polymorphic outside Africa.

The discussion of the allele frequencies will be divided into 3 sections. Firstly, a discussion of the enzyme deficient samples which were selected where DNA was available and studied using molecular methods (section 3.1.2.1). Studying these data, the specific deficiency variants (*Nla* III and *Mbo* II sites) responsible for the deficiency, and the distribution of these variants across the island could be established. Secondly, a discussion of the randomly selected (i.e. irrespective of deficiency status) individuals for the mobility variant (*Fok* I site) (section 3.1.2.2) and thirdly, a discussion of the randomly selected individuals for the neutral polymorphisms (*Pvu* II, *Bsp* HI, *Pst* I and *Bcl* I sites) (section 3.1.2.3). The deficient individuals were analysed to characterise the deficiency variants and the random individuals were studied to determine haplotype frequencies for comparison with the putative parental populations. Some of the randomly selected individuals were also found to be deficient.

Table 3-2: Allele frequencies[1] for the 7 G6PD systems in the random Malagasy groups

	Antesaka		Betsileo		Merina		Tsimihety		Vezo		Mahafaly	
	n*	Freq. (±SE)[2]	n*	Freq. (±SE)[2]	n*	Freq. (±SE)[2]	n*	Freq. (±SE)[2]	n*	Freq. (±SE)[2]	n*	Freq. (±SE)[2]
***Mbo* II**	59	0.017 (0.017)	61	0	62	0	56	0.054 (0.030)	39	0	33	0
***Nla* III**	56	0.179 (0.051)	53	0.038 (0.026)	60	0.067 (0.032)	56	0.161 (0.049)	31	0	32	0.063 (0.043)
***Fok* I**	57	0.263 (0.058)	68	0.176 (0.046)	62	0.097 (0.038)	54	0.222 (0.057)	40	0.025 (0.025)	33	0.121 (0.057)
***Pvu* II**	59	0.203 (0.052)	66	0.076 (0.033)	63	0.063 (0.031)	54	0.148 (0.048)	37	0	33	0.061 (0.042)
***Bsp* HI**	57	0.719 (0.060)	65	0.846 (0.045)	62	0.903 (0.038)	55	0.764 (0.057)	38	0.974 (0.026)	33	0.879 (0.057)
***Pst* I**	56	0.964 (0.025)	58	0.966 (0.024)	60	0.933 (0.032)	54	0.926 (0.036)	30	1.000 (0)	33	0.879 (0.057)
***Bcl* I**	59	0.339 (0.062)	55	0.327 (0.063)	62	0.339 (0.060)	54	0.333 (0.064)	32	0.250 (0.077)	32	0.156 (0.064)

[1]The frequencies are all reported for the presence of the RFLP site (e.g. *Mbo* II (+)).

[2]Standard error (SE) was calculated by: $SE = \sqrt{\frac{(p)(1-p)}{n}}$, where n is the number of chromosomes.

* n represents the number of chromosomes analysed. For those groups where only males were analysed, the number of chromosomes is the same as the number of individuals, but for those groups where females were included, the number of chromosomes is not equal to the number of individuals. The total number (n) is not constant for each population for different systems because sometimes a result could be obtained for one site but not another for a particular sample.

Table 3-3: Allele frequencies[1] for the 7 G6PD systems in the non-Malagasy populations

	Zambians		CAR(grp)[3]		Tsonga		Zulus		IND(grp)[4]		Polynesians	
	n*	Freq. (±SE)	n*	Freq. (±SE)	n*	Freq. (±SE)	n*	Freq. (±SE)	n*	Freq. (±SE)	n*	Freq. (±SE)
***Mbo* II**	84	0 (0)	78	0 (0)	69	0 (0)	47	0 (0)	131	0 (0)	55	0 (0)
***Nla* III**	81	0.222 (0.046)	74	0.054 (0.026)	68	0.044 (0.025)	44	0 (0)	34	0 (0)	56	0 (0)
***Fok* I**	84	0.357 (0.052)	78	0.256 (0.049)	69	0.188 (0.047)	48	0.250 (0.063)	138	0 (0)	57	0 (0)
***Pvu* II**	84	0.262 (0.048)	78	0.103 (0.034)	69	0.029 (0.020)	46	0.065 (0.036)	90	0 (0)	58	0 (0)
***Bsp* HI**	84	0.631 (0.053)	77	0.714 (0.051)	69	0.797 (0.048)	44	0.727 (0.067)	95	1.000 (0)	53	1.000 (0)
***Pst* I**	81	0.926 (0.029)	78	0.859 (0.039)	66	0.742 (0.054)	43	0.791 (0.062)	93	1.000 (0)	44	1.000 (0)
***Bcl* I**	82	0.317 (0.051)	77	0.364 (0.055)	62	0.355 (0.061)	44	0.295 (0.069)	112	0.179 (0.036)	58	0.172 (0.050)

[1]The frequencies are all reported for the presence of the RFLP site (e.g. *Mbo* II (+)).

[2]Standard error (SE) was calculated by: $SE = \sqrt{\frac{(p)(1-p)}{n}}$, where n is the number of chromosomes.

* n represents the number of chromosomes analysed. For those groups where only males were analysed, the number of chromosomes is the same as the number of individuals, but for those groups where females were included, the number of chromosomes is not equal to the number of individuals. The total number (n) is not constant for each population for different systems because sometimes a result could be obtained for one site but not another for a particular sample.

[3]CAR(grp) is the combination of individuals belonging to the Mbindu, Sangha Sangha and Gbaya CAR groups.

[4]IND(grp) is the combination of all Indonesian individuals (from south and central Sulawesi and Kalimantan).

3.1.2.1 DEFICIENCY VARIANT ALLELE FREQUENCIES

The deficiency variants present in the Malagasy and in some of the African samples have, until now, not been characterised at the molecular level. In the Malagasy population, 93 enzyme deficient individuals (78 males and 15 females) (as described in section 2.1.1.2) were randomly selected where DNA was available and studied using molecular methods. The deficiency variants in the Malagasy sample were identified for 77 of the 78 deficient male subjects and for 11 of the 15 female subjects. The total number of male and female individuals for whom the deficiency allele was unidentified was therefore 5. The regional and ethnic group breakdown for the male subjects showing the deficiency variants found is shown in Table 3-4. All of the males except one had the A^- or Mediterranean variants. The male whose deficiency variant could not be established was found to have an A mobility chromosome (i.e. *Fok* I (+)). The deficiency variant most likely to be present in this individual is therefore one of the rare A^- mutations (as discussed in section 1.2.6.3.1.2).

The 11 female deficient subjects whose deficiency variant could be identified were all homozygous for the common A^- mutation (i.e. *Fok* I (+), *Nla* III (+)). No females were found to have deficiency caused by the Mediterranean variant and no haematologically deficient females were found that were heterozygous for either the A^- or Mediterranean deficiency alleles. Of the 4 samples for which the deficiency variant was unidentified, 1 was homozygous for an A mobility chromosome (*Fok* I (+)) and 3 were homozygous for non-A mobility chromosomes (*Fok* I (-)). The A mobility sample may have one or both of the uncommon A^- deficiency variants (in homozygous or compound heterozygous form). The non-A mobility samples have non-African deficiency variant(s) other than the Mediterranean variant. Candidates would include the Canton and Mahidol deficiency variants which are common in southeast Asia. However, other non-A mobility (and non-B electrophoretic) variants in Africa could be responsible. These variants could be identified in the future using RFLP analysis or sequencing.

Table 3-4: Specific molecular deficiency variants in G6PD deficient Malagasy males by ethnic and regional grouping

REGION AND ETHNIC GROUP	DEFICIENCY VARIANT				Total[5]
	A[-1]	Med I[2]	Med VII[3]	Other A[4]	
North:	**8**	**1**	**2**	**0**	**11**
Antankarana	0	0	0	0	0
Tsimihety	8	1	2	0	11
Sihanaka	0	0	0	0	0
Central:	**11**	**0**	**0**	**0**	**11**
Merina	4	0	0	0	4
Betsileo	6	0	0	0	6
Bezanozano	1	0	0	0	1
East Central:	**29**	**0**	**6**	**1**	**36**
Betsimisaraka	8	0	3	1	12
Antemoro	6	0	1	0	7
Antefasy	1	0	0	0	1
Antambahoaka	0	0	1	0	1
Antesaka	9	0	1	0	10
Tanala	0	0	0	0	0
Antanosy	5	0	0	0	5
Southwest:	**8**	**0**	**0**	**0**	**8**
Antandroy	4	0	0	0	4
Bara	1	0	0	0	1
Vezo	2	0	0	0	2
Mahafaly	1	0	0	0	1
Masikoro	0	0	0	0	0
West:	**10**	**0**	**0**	**0**	**10**
Sakalava	8	0	0	0	8
Makua	2	0	0	0	2
Unknown	2	0	0	0	2
Total	**68**	**1**	**8**	**1**	**78**

[1] A⁻: the common A⁻ deficiency variant *Nla* III (+) and *Fok* I (+) (haplotype VI).

[2] Med I refers to the Mediterranean variant (*Mbo* II site (+)) associated with haplotype I (i.e. *Bcl* I (-), the Indian associated haplotype).

[3] Med VII refers to the Mediterranean variant associated with haplotype VII (*Bcl* (+), the middle eastern associated haplotype).

[4] Other A indicates that the variant has not been identified, but occurs on an A (*Fok* I (+)) chromosome.

[5] Total refers to the total number of enzyme deficient male individuals analysed using the molecular methods.

3.1.2.1.1 *MBO* II SITE (MEDITERRANEAN DEFICIENCY VARIANT) FREQUENCY

The *Mbo* II (+) variant is associated with the G6PD Mediterranean deficiency. In this study, the *Mbo* II site was not found in any of the randomly selected sub-Saharan African samples. In other studies the Mediterranean variant has only been found in the extreme northern parts of Africa such as Algeria (Nafa *et al* 1994). The *Mbo* II (+) allele was not, however, found in any of the Austronesian populations either (the Indonesian and Polynesian samples) in this study. The Mediterranean variant is reported to form a significant proportion (5/16=0.31) of the Indonesian deficiency variants (Soemantri *et al* 1995). For the sample size in this project (144) and the lack of G6PD activity data, this under-representation is not unexpected. Also, the study of Soemantri *et al* (1995) used a small sample size which could lead to sampling bias.

The frequency at which the Mediterranean variant is found in the deficient male Malagasy sample is very small (8/78=0.10). No *Mbo* II (+) female individuals were found. This is not surprising in view of the much lower frequency of the Mediterranean variant in the males and the small sample size (15) in the deficient females. All but one of the Mediterranean deficient samples had simultaneous presence of the *Bcl* I site. This indicates a predominance of the Middle Eastern associated haplotype, rather than the Indian haplotype, unlike the proportion in Indonesia where the Indian haplotype predominates (Soemantri *et al* 1995).

Interestingly, the Mediterranean variant was only found in lowland populations (Table 3-4). This could be because of a low frequency in the original Indonesian ancestral population, a low frequency in the founder population which settled Madagascar, or could simply be a bias from the small sample size. A lower selective advantage in the highlands where the Indonesian gene contribution is expected to be greater or greater selective advantage of the A^- variant may also explain the distribution. It is difficult to draw conclusions without knowing the $G6PD^{Mediterranean}$ frequency in the possibly small group of founders. Whether or not the Mediterranean variant is a less effective selective agent under local conditions than the A^- variant is also unknown.

3.1.2.1.2 *NLA* III SITE (A^- DEFICIENCY VARIANT) FREQUENCY

The combination of *Fok* I (+) and *Nla* III (+) detects the most common of the three A^- deficiencies. As expected, no deficiency variant other than the common A^- deficiency variant was found in African populations in this study. The deficiency in the only African population for which enzyme deficiency data are available, the CAR, is entirely accounted for by the A^- mutation analysed in this population, although this is on the basis of only 2 deficient samples. Although other A^- variants exist, they occur at a low frequency in individuals of African descent (Beutler *et al* 1989).

The *Nla* III (+) variant, A^-, is the predominant deficiency variant in the Malagasy (70/78=0.90). It is spread all over the island (as can be seen from Table 3-4) and there does not seem to be a proportional under-representation of A^- compared to Mediterranean deficiency in the highland groups or an over-representation in the lowland groups. This suggests a strong African contribution to all Malagasy groups. The *Nla* III (+) variant (A^-) is not found in any of the Austronesian samples and is the only deficiency variant found in the African samples.

In the present study, the *Nla* III (+) allele was never found on a *Fok* I (-) chromosome (out of a total number of 468 *Fok* I (-) chromosomes from the African and Malagasy populations). Town *et al* (1992) predicted that the *Fok* I (-), *Nla* III (+) combination could occur at a low frequency due to recombination and that these individuals would not be G6PD deficient. The fact that the combination has not been found in this study is a measure of the strong linkage disequilibrium that exists between these alleles. The *Nla* III (+) allele exists on only a single haplotype (haplotype VI).

As most reports on G6PD deficiency in Africa have studied enzyme deficiency rather than variation at the DNA level, one needs to be cautious in comparing these with the work in this project since three possible mutations cause A^- deficiency. From previous reports (Beutler *et al* 1989), however, we can expect the largest proportion (if not all) of the deficiency to be due to the common African variant A^- studied here. The frequency of the *Nla* III(+) allele is therefore probably a good estimate both of the deficiency frequency and the Gd^{A-} allele frequency obtained biochemically.

In general, there is a clinal decrease in Gd^{A-} allele frequency from north to south in southern Africa (as shown in Figure 3-1). While the Gd^{A-} variant (ascertained biochemically) frequency in Nigerians was 0.22 (Porter *et al* 1964), the frequency in southern African Bantu males was 0.056 and was absent in the Khoi (Jenkins and Nurse 1972) and San (Balinsky and Jenkins 1966). In a larger study of the San people, a low frequency of deficiency (0.015-0.045) was found (Jenkins *et al* 1968). The average frequency of Gd^{A-} found in Mozambique (0.19) (Reys *et al* 1970), a good candidate for the ancestral population of the Malagasy, is similar to that in the lowland Malagasy individuals tested in this project. Other reported frequencies include 0.053 in the Tsonga and 0.059 in the Zulu (Hitzeroth and Bender 1980).

The A^- variant is not entirely specific to Africa and has been found in white male individuals in Spain (Rovira *et al* 1995) and Italy (Calabro *et al* 1990). These A^- variants are thought to represent African admixture since the A^- variants are associated with the same haplotypes as found in African samples (Vulliamy *et al* 1991b). However, the A^- variant is very ancient and so it is difficult to determine how long ago the movement from Africa occurred. Furthermore the dates of the colonisation of Madagascar are uncertain. In the present study, the A^- mutation has been found to predominate. If the people of Madagascar are truly the "Progeny of the Indian Ocean" (Dewar and Wright 1993), it may be inappropriate to assume that all of the A^- variants in Madagascar originated in Africa. In the present study only one variant is unidentified in the male population. If some of the A^- in Madagascar represents movement from other countries, other variants, occurring at higher frequencies in those countries of origin, would be expected. The A^- variant has not been found in Indonesia in this project or in other publications (Soemantri *et al* 1995). It is probably reasonable, therefore, to conclude that most (if not all) of the A^- variants found in Madagascar are of African origin. Other factors which must be taken into account in this type of comparison are the size of the founder population and the extent of modification of the allele frequency due to selective pressure, two factors which are difficult to quantitate.

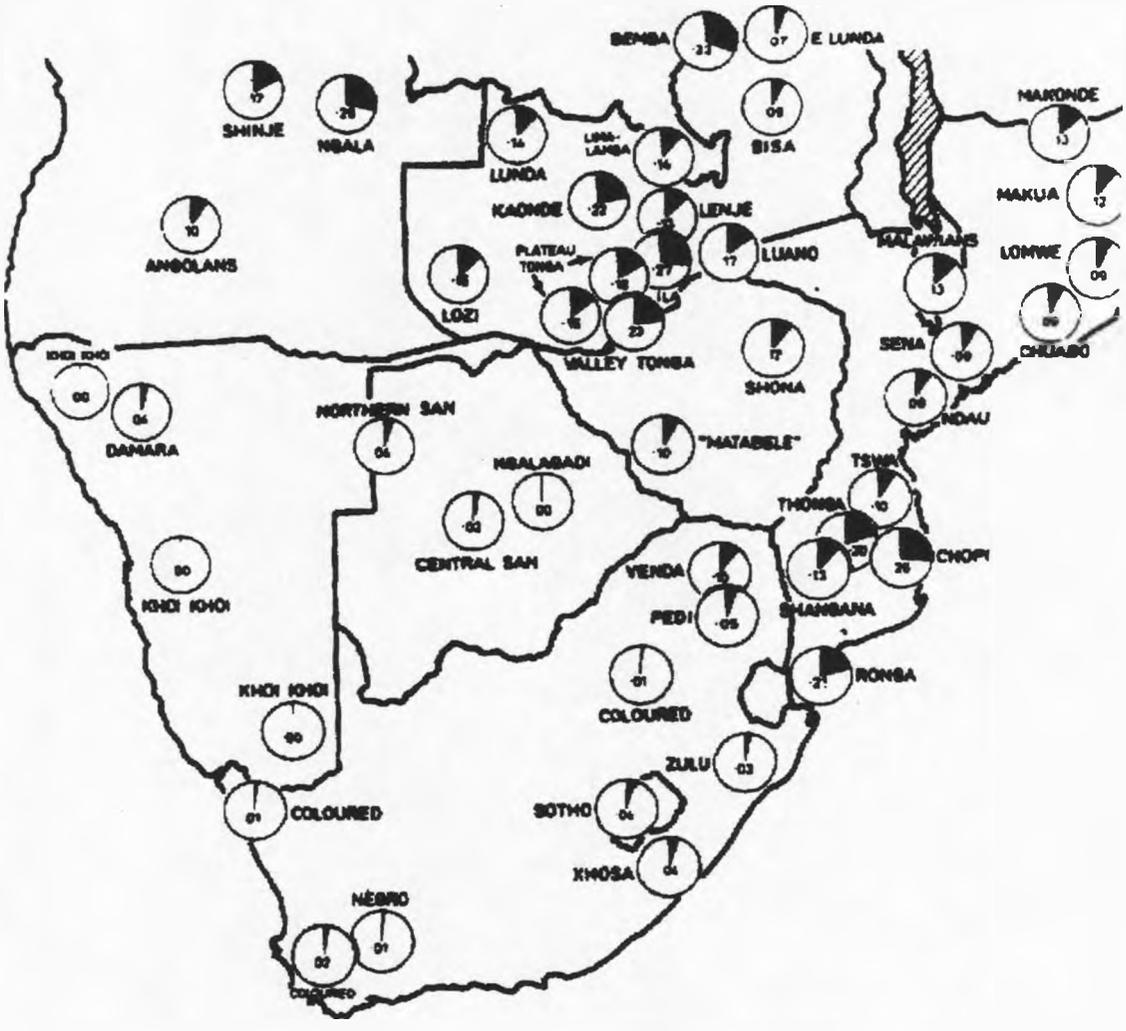


Figure 3-1: Map showing the distribution of the biochemical Gd^{A-} deficiency variant in southern Africa (Jenkins 1972)

3.1.2.1.3 COMPARISON OF MOLECULAR AND HAEMATOLOGICAL METHODS

The Austronesian and most of the African samples analysed in this study were not tested for G6PD deficiency. A comparison of enzyme status with molecular results in these populations is therefore not possible.

For 77 of the 78 Malagasy males (and 88/93 including females) found to be G6PD deficient, the molecular defect could be characterised. The type of deficiency present in the other 5 samples has yet to be characterised. In addition, 4 samples had deficiency variants (i.e. they were *Fok I* (+) and *Nla III* (+)), but were found to be non-deficient by the G6PD screening test. Two of these samples were females, heterozygous for the A^- mutation, and would therefore not be expected to be enzyme deficient. The other two were males, hemizygous for the A^- deficiency variant and would be expected to be G6PD deficient. The enzyme activity screen may thus underestimate G6PD deficiency slightly.

The screening test for G6PD deficiency detected 78/80 (97.5%) of the deficient males while the molecular testing, using 2 deficiency variants only, detected 79/80 (98.75%) of the deficient males. This type of molecular testing is only effective in detecting known, expected variants. By far the most common variant (68/78) in Madagascar is the A^- variant (*Fok* I (+), *Nla* III (+)) and the majority of the remainder (9/78) are the Mediterranean variant (*Mbo* II(+)). Other deficiency variants are present in the Malagasy, indicating that molecular screening would not be 100% effective.

The preponderance of the A^- and Mediterranean variants has important implications for public health. In principle, 79/80 (98.75%) of deficient individuals could be detected using fast and accurate RFLP analysis. Although neither variant is very severe (neither results in CNSHA) they may be responsible for some cases of neonatal jaundice and reactions to anti-malarials (e.g. primaquine). The Mediterranean variant does also predispose to favism, haemolytic disease of the newborn and haemolytic anaemia.

3.1.2.2 MOBILITY VARIANTS

3.1.2.2.1 THE *FOK* I SITE (THE A MOBILITY VARIANT)

The *Fok I* RFLP allows for a molecular method of detecting the variant responsible for the commonest G6PD electrophoretic mobility shift from B (normal) to A (fast). The *Fok* I (+) result corresponds to an A-mobility result and the *Fok* I (-) corresponds to a B-mobility result. The *Fok* I (+) site was found only in the African and Malagasy groups tested in this project as shown in Tables 3-2 and 3-3; it was not found in the Austronesian populations. This supports previous reports (discussed in section 1.2.6.3.1.1) where the A variant was rare or absent in non-African populations.

Like the Gd^{A-} allele, the Gd^{A} allele follows a clinal decrease in frequency from north to south (as shown in Figure 3-2), although it is not as marked as the Gd^{A-} profile (Jenkins 1972), possibly because it offers less or no selective advantage against malaria (Ruwende *et al* 1995). A high frequency (0.22) of the A variant was found in Nigerians (Porter *et al* 1964). A similarly high frequency (0.20) was found in a Negroid population from South West Africa (Balinsky and Jenkins 1966) and at the same time it was determined that there were frequencies of 0.12 in southern African Bantu-speaking males, 0.00 in Khoi individuals, and 0.028 in San males. When the Bantu-speaking group was subdivided into different chiefdoms, a frequency of 0.154 was found in the Zulu, a frequency of 0.178 in the Basuto and 0.195 in the Tswana (Balinsky and Jenkins 1967).

The highest frequency found in this project (Table 3-3) was in the Zambians (0.357), followed by the CAR(grp) (0.256), Zulu (0.250) and Tsonga (0.188). Comparison with published results show a frequency of 0.211 in the Tsonga group (Hitzeroth and Bender 1980), comparable to the value of 0.188 found in this study. The *Fok* I frequency in the Zulu population found in this study is 0.250, slightly higher than the Gd^{A} frequency report of 0.154 by Balinsky and Jenkins (1967), but comparable to the frequency of 0.223 reported by Hitzeroth and Bender (1980). The Zambian *Fok* I frequency (0.357) in this project is one of the highest A frequencies reported in Africa; (a summary of frequencies is reviewed in Nurse *et al* 1985).

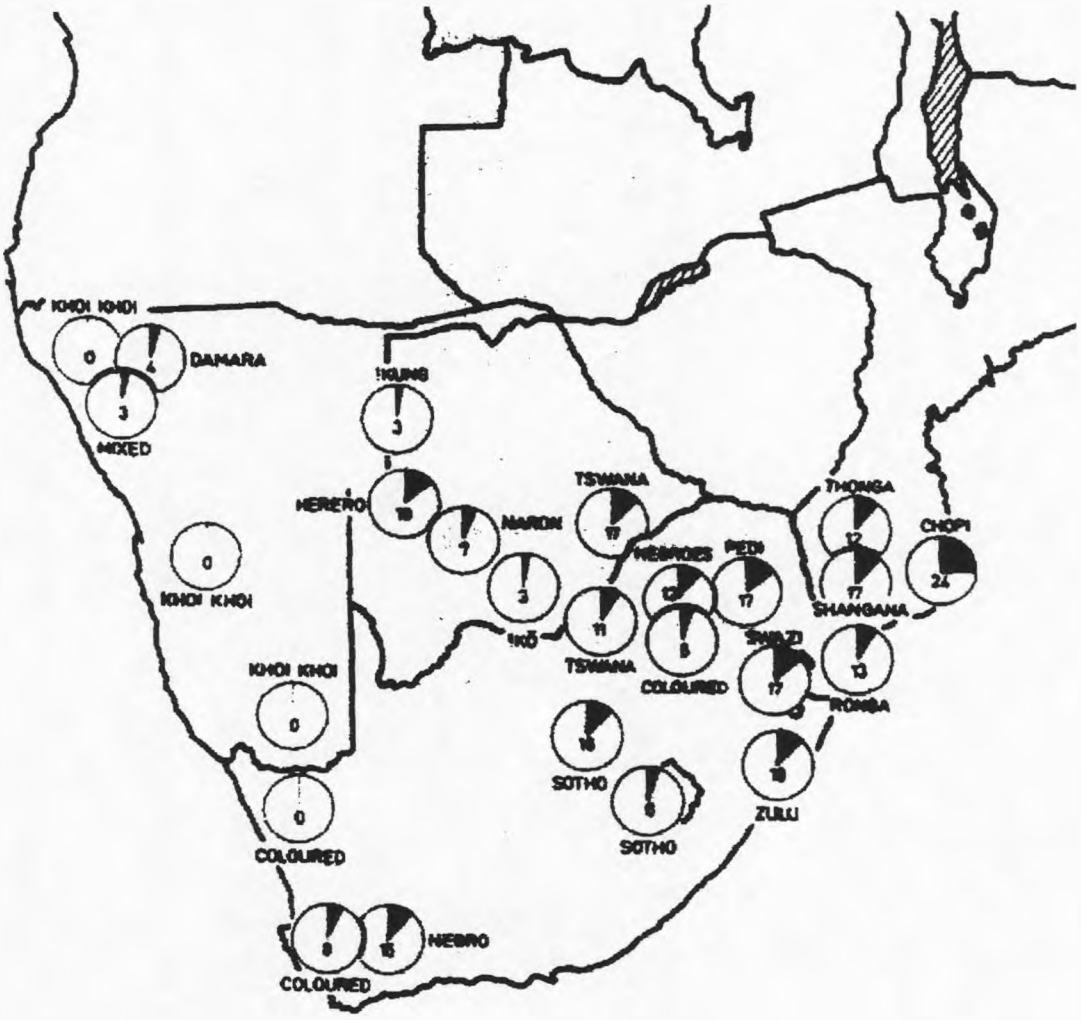


Figure 3-2: Map showing the distribution of the mobility variant *Gd*^{A} in southern Africa (Jenkins 1972)

The Malagasy groups (Table 3-2) have intermediate *Fok* I(+) frequencies between the Austronesian groups (zero) and the highest African frequencies. The lowland groups (Antesaka and Tsimihety) tend to have frequencies closer to the African groups and are slightly lower than the average frequency (0.36) found in Mozambique (Reys *et al* 1970). The Merina group, who are thought to be the most Indonesian-like Malagasy group (as discussed in section 1.1.2.2), have a very low *Fok* I frequency (0.097) (i.e. relatively closer to the Indonesian frequency). The other highland group, Betsileo, is thought to be more cosmopolitan, and less endogamous than the Merina. The *Fok* I frequency (0.176) is higher and closer to the African values.

Interestingly, the southwest (Vezo and Mahafaly) ethnic groups also show a very low *Fok* I (+) frequency. The lack of the *Fok* I (+) allele is not enough to conclude Indonesian ancestry, but it could indicate less African admixture since the A variant is African-specific. Because of the large range of variation of this polymorphism in the African populations (0.091 to 0.357) it does not even necessarily indicate less African ancestry i.e. the B mobility variant is not specific to non-African populations. Low frequencies in the highland groups and in the Vezo and Mahafaly could therefore indicate high levels of admixture from non-African populations (such as Indonesian) or admixture from an African group which has a low proportion of A chromosomes.

3.1.2.3 ALLELE FREQUENCIES OF NEUTRAL POLYMORPHISMS

A major problem in studying population admixture where one of the parental populations is African, is that there are relatively few polymorphisms which are specific to non-African areas. In this project the *Bcl* I polymorphism is the only one which is not African-specific (as discussed in section 1.2.6.3.1.4).

In the populations studied in this project (Tables 3-2 and 3-3), the Austronesian populations have relatively low frequencies of *Bcl I* (+) (0.179 for the Indonesians and 0.172 for the Polynesians). Published results include frequencies of 0.31 in an Indonesian sample (Soemantri, *et al*, 1995), 0.053 in a sample from the Solomon Islands in Melanesia, 0.20 in an Indian sample, and 0.24 for a mixed African sample (Vulliamy *et al* 1991b). The African samples in this project (Table 3-3) tend to have higher *Bcl* I (+) frequencies than in the Austronesian samples (0.295 in the Zulu population and 0.355 in the Tsonga population). The Malagasy groups can be expected to have intermediate values. In this study, the lowland groups Antesaka (0.339) and Tsimihety (0.333) have frequencies approaching the African frequencies. However, the highland values are of similar magnitudes. The Vezo (0.250) and Mahafaly (0.156) groups again have lower frequencies, approaching that of the Indonesian and Polynesian populations.

The trend of the frequencies in the Malagasy samples having intermediate frequencies occurs in some of the other neutral polymorphisms, although where the frequencies in the African and Austronesian samples are very similar, this is not easy to distinguish. In general, the southwest Vezo and Mahafaly groups have frequencies more similar to Austronesian than African population frequencies.

The *Pvu* II site has been found to be polymorphic in African populations only and is only polymorphic on A mobility chromosomes. This is true for this project too: the site was not polymorphic in the Indonesian and Polynesian populations and polymorphic in the African populations, as expected. In North American Negroid individuals, a *Pvu* II(+) frequency of 0.190 and in South African Bantu-speaking Negroids, a frequency of 0.021 (Coetzee *et al* 1992) has been reported. In a sample of black individuals in London, the *Pvu* II(+) frequency was found to be 0.27 (Vulliamy *et al* 1991a). These figures are consistent with the data obtained for the African populations analysed in this study (Table 3-3).

Few data are available for the frequency of the African-specific *Bsp* HI polymorphism. The mutation abolishes a *Bsp* HI site (as discussed in section 1.2.6.3.1.4). Tables 3-2 and 3-3 show the frequencies of the *Bsp* HI (+) allele. The RFLP is not polymorphic in non-African populations, as shown by the non-African populations (Indonesians and Polynesians) in this project where the frequency is 1. Vulliamy *et al* (1991b) found a frequency of *Bsp* HI (+) of 43% in African individuals. By comparison, the frequencies found in the African populations in this study range from 0.631 (in the Zambian population) to 0.797 (in the Tsonga).

The *Pst* I (-) RFLP is found on a B mobility chromosome and has only been found in individuals of African descent in this study and previously (as discussed in section 1.2.6.3.1.3). Frequencies of *Pst* I (+) of 0.11 in African individuals, predominantly from West Africa (Vulliamy, *et al*, 1991b) and 0.12 in black males in London (Vulliamy, *et al*, 1991a) have been previously reported. In this project, frequencies of *Pst* I (+) range from 0.742 in the Tsonga population to 0.926 in the Zambian population (as shown in Table 3-3). The *Pst* I (+) site was not polymorphic in the Indonesian or Polynesian populations, as expected.

3.1.2.4 LINKAGE DISEQUILIBRIUM

Vulliamy *et al* (1991b) calculated linkage disequilibrium between the alleles at six of the seven RFLP loci studied in this project. The method used, described in section 2.2.4.1.2.1, the same as that of Vulliamy and colleaues (1991b), was the calculation of D values, which are the absolute values of the differences between the observed and expected frequencies of the association of two alleles. The population studied by Vulliamy *et al* (1991b) was a mixed group of individuals of African origin (mainly from Nigeria and Kenya). For purposes of comparison, D values in this project were calculated for the combined CAR group (CAR (grp)). This group is the combination of the Mbindu, Sangha Sangha and Gbayo individuals. It is justifiable to group these samples since they are not significantly different (see section 3.2.1.2). Since there are no Mediterranean deficient samples in the CAR(grp) group or that of Vulliamy *et al* (1991b), the *Mbo* II site was excluded from analysis. The results reported by Vulliamy *et al* (1991b) are compared to those calculated in this project in Table 3-5. In both cases, the allele studied was the least common allele.

The D values were analysed with a chi square test to determine the significance of the difference between the observed and expected figures. The chi square and p values for these are shown in Table 3-6. A significant difference indicates that the two alleles in question are in linkage disequilibrium. Because there are seven biallelic sites in the haplotypes, 2^7=128 possible haplotypes are expected. Vulliamy *et al* (1991b) found a total of seven haplotypes out of the possible 128 haplotypes, indicating a strong linkage disequilibrium. In this study (section 3.1.3) the same seven haplotypes were found. At least two additional haplotypes were observed in the present study, which occur at a low frequency (see section 3.1.3).

Table 3-5: D values for comparison of the linkage disequilibrium figures found in this study (for the CAR group) compared with those found by Vulliamy *et al* (1991b) in an African sample

RE Site	*Nla* III (+)		*Fok* I (+)		*Pvu* II (+)		*Bsp* HI (-)		*Pst* I (-)	
	This study	Vulliamy[ϕ]	This study	Vulliamy[ϕ]	This study	Vulliamy[ϕ]	This study	Vulliamy[ϕ]	This study	Vulliamy[ϕ]
***Fok* I (+)**	0.038[†]	0.14[‡]								
***Pvu* II (+)**	0.046[‡]	0.18[‡]	0.091[‡]	0.16[‡]						
***Bsp* III (-)**	0.037[†]	0.15[‡]	0.186[‡]	0.24[‡]	0.088[‡]	0.17[‡]				
***Pst* I (-)**	0.008	0.03	0.036	0.05	0.014	0.03	0.040	0.05		
***Bcl* I (+)**	0.020	0.06	0.093*	0.10*	0.037	0.07[†]	0.051	0.10*	0.051	0.03

*$p < 0.01$; †$p < 0.05$; ‡$p < 0.0001$. The significance values for the Vulliamy study are taken from Vulliamy *et al*, 1991b, while the χ^2 and p values for this study are shown in Table 3-6.

[ϕ]Vulliamy *et al* (1991b).

Table 3-6: Chi square and p values (1 degree of freedom) for the linkage disequilibrium calculations for the CAR group

RE Site		*Nla* III (+)	*Fok* I (+)	*Pvu* II (+)	*Bsp* HI (-)	*Pst* I (-)
Fok I (+)	χ^2	5.602$^{\Uparrow}$				
	p	0.018†				
Pvu II (+)	χ^2	13.942$^{\Uparrow}$	24.674			
	p	1.89E-04‡	6.79E-07‡			
Bsp HI (-)	χ^2	4.782$^{\Uparrow}$	39.439	20.763		
	p	0.029†	3.38E-10‡	5.20E-06‡		
Pst I (-)	χ^2	0.322$^{\Uparrow}$	2.466$^{\Uparrow}$	0.785$^{\Uparrow}$	2.803$^{\Uparrow}$	
	p	0.571	0.116	0.376	0.094	
Bcl I (+)	χ^2	1.168$^{\Uparrow}$	7.456$^{\Uparrow}$	2.558$^{\Uparrow}$	3.720$^{\Uparrow}$	3.720$^{\Uparrow}$
	p	0.280	0.006*	0.110	0.054	0.054

*$p < 0.01$; †$p < 0.05$; ‡$p < 0.0001$.

$^{\Uparrow}$ Yates' correction for continuity was applied where necessary for comparisons with small observed values.

In both the study by Vulliamy *et al* (1991b) and the present study, there is a general trend of decrease in linkage disequilibrium with increasing distance between the loci for which the alleles are being compared (see Figure 2-1 for a map of the locations of the RFLP loci). Although the D values produced are often of a different magnitude for the two studies, the significance values are usually similar. For example, the comparison of the *Nla* III (+) allele and the *Pvu* II (+) allele produce p values which are significant at the 99.99% confidence interval for both studies, even though the D values are different (0.046 and 0.18). There is thus good corroboration between this study and the one by Vulliamy *et al* (1991b). The only comparisons which differ between the two studies are the *Pvu* II (+), *Bcl* I (+) comparison which was significant in the Vulliamy study (at the 95% confidence interval) but not in this study and the *Bsp* HI (-), *Bcl* I (+) comparison which was significant in the Vulliamy study (at the 99% confidence interval) but not in this study.

The values reported here serve to confirm the observation of Vulliamy *et al* (1991b) that there is a strong linkage disequilibrium between most of the sites within the G6PD cluster, possibly in part because of the close proximity of the sites. Only seven haplotypes out of a possible 128 were reported by Vulliamy *et al* (1991b) and they are the same seven found in this study, despite the fact that a substantially larger and more clearly defined set of populations was studied in this project. In addition, at least two novel haplotypes were found in this study (see section 3.1.3). Other studies have also found strong linkage between the alleles at the RFLP loci examined here, but none of them analysed full six- or seven-site haplotypes (Beutler and Kuhl 1990a, Vulliamy *et al* 1991a).

3.1.3 HAPLOTYPE FREQUENCIES

The use of haplotypes in the comparison of the affinities of different populations is more powerful than the use of individual alleles. In the remainder of the project, therefore, haplotypes were used for comparison rather than alleles. The Malagasy haplotype frequencies are shown in Table 3-7, the African frequencies in Table 3-8 and the Austronesian frequencies in Table 3-9. In all cases, the frequencies are calculated for all individuals studied (male and female) and the total, n, is therefore the number of chromosomes for which haplotypes could be assigned. The numbers reported here are slightly lower than those reported for the allele frequencies, because for some individuals, full haplotypes could not be constructed or inferred.

On examination of Tables 3-7 to 3-9, it is apparent that there are fewer haplotypes present in the Austronesian populations than in the African and Malagasy populations. There are only 2 haplotypes present in the Austronesian samples whereas in the African and Malagasy samples, all 7 of the haplotypes reported by Vulliamy *et al* (1991b) are present. In addition, the Mediterranean deficiency haplotypes Med I (the Indian associated haplotype) and Med VII (the Middle Eastern associated haplotype) are present in the Malagasy. No Mediterranean variants (and therefore no Mediterranean haplotypes) were present in the African or Austronesian populations. Haplotype III does not occur in any of the African populations studied although it is present in the Malagasy. Vulliamy *et al* (1991b) showed that haplotype III does, however, occur in Africa. This could indicate that inappropriate African populations are being considered as ancestral Malagasy populations in this project. However, with a haplotype that occurs at such a low frequency (0.02) (Vulliamy *et al* 1991b) (or 0-0.019 in this study) it could simply indicate a sampling error.

Table 3-7: G6PD haplotype frequencies, (with Standard errors in brackets) for the various Malagasy groups studied

Group (n[1]) Haplotype	ANTESAKA (57) Freq. (±SE)	BETSILEO (54) Freq. (±SE)	MERINA (62) Freq. (±SE)	HIGH[2] (116) Freq. (±SE)	TSIMIHETY (54) Freq. (±SE)	LOW[3] (111) Freq. (±SE)	VEZO (30) Freq. (±SE)	MAHAFALY (32) Freq. (±SE)	VZ&MF[4] (62) Freq. (±SE)
I	0.316 (0.062)	0.481 (0.068)	0.516 (0.063)	0.500 (0.046)	0.352 (0.065)	0.333 (0.045)	0.700 (0.084)	0.594 (0.087)	0.645 (0.061)
VII	0.333 (0.062)	0.333 (0.064)	0.323 (0.059)	0.328 (0.044)	0.278 (0.061)	0.306 (0.044)	0.267 (0.081)	0.156 (0.064)	0.210 (0.052)
II	0.035 (0.024)	0.037 (0.026)	0.065 (0.031)	0.052 (0.021)	0.074 (0.036)	0.054 (0.021)	0	0.125 (0.058)	0.065 (0.031)
III	0	0.019 (0.018)	0	0.009 (0.009)	0	0	0	0	0
IV	0.070 (0.034)	0.074 (0.036)	0.016 (0.016)	0.043 (0.019)	0.074 (0.036)	0.072 (0.025)	0.033 (0.033)	0.031 (0.031)	0.032 (0.022)
V	0.018 (0.017)	0.019 (0.018)	0	0.009 (0.009)	0	0.009 (0.009)	0	0.031 (0.031)	0.016 (0.016)
VI	0.193 (0.052)	0.037 (0.026)	0.065 (0.031)	0.052 (0.021)	0.167 (0.051)	0.180 (0.036)	0	0.063 (0.043)	0.032 (0.022)
"Other"	0.018 (0.017)	0	0.016 (0.016)	0.009 (0.009)	0	0.009 (0.009)	0	0	0
Med I	0	0	0	0	0.019 (0.018)	0.009 (0.009)	0	0	0
Med VII	0.018 (0.017)	0	0	0	0.037 (0.026)	0.027 (0.015)	0	0	0

[1] n= number of chromosomes. For groups where only males were sampled, this is the same as the number of individuals.

[2] HIGH= combined highland groups (BT and MR combined). The groups were combined after χ^2 testing (section 3.2.1.1) and showed the differences between the frequencies in the various populations were not significant.

[3] LOW = combined lowland groups (AS and TS). The groups were combined after χ^2 testing (section 3.2.1.1).

[4] VZ&MF = combined Vezo and Mahafaly groups. The groups were combined after χ^2 testing (section 3.2.1.1) and due to geographical proximity.

Table 3-8: G6PD haplotype frequencies, (with standard errors in brackets) for the African populations studied

Population (n) Haplotype	Zambian (82) Freq. (±SE)	CAR (M)[1] (20) Freq. (±SE)	CAR (S)[2] (19) Freq. (±SE)	CAR (G)[3] (38) Freq. (±SE)	CAR (MSG)[4] (77) Freq. (±SE)	Tsonga (62) Freq. (±SE)	Zulu (42) Freq. (±SE)
I	0.268 (0.049)	0.300 (0.102)	0.105 (0.070)	0.237 (0.069)	0.221 (0.047)	0.242 (0.054)	0.262 (0.068)
VII	0.341 (0.052)	0.350 (0.107)	0.474 (0.115)	0.316 (0.075)	0.364 (0.055)	0.339 (0.060)	0.310 (0.071)
II	0.024 (0.017)	0.100 (0.067)	0.053 (0.051)	0.211 (0.066)	0.143 (0.040)	0.226 (0.053)	0.167 (0.058)
III	0	0	0	0	0	0	0
IV	0.098 (0.033)	0.200 (0.089)	0.158 (0.084)	0.105 (0.050)	0.143 (0.040)	0.097 (0.038)	0.167 (0.058)
V	0.049 (0.024)	0	0.105 (0.070)	0.079 (0.044)	0.065 (0.028)	0	0.071 (0.040)
VI	0.207 (0.045)	0.050 (0.049)	0.105 (0.070)	0.026 (0.026)	0.052 (0.025)	0.048 (0.027)	0
"Other"	0.012 (0.012)	0	0	0.026 (0.026)	0.013 (0.013)	0.048 (0.027)	0.024 (0.024)
Med I	0	0	0	0	0	0	0
Med VII	0	0	0	0	0	0	0

[1]CAR(M) = Mbindu language group of the CAR

[2]CAR(S) = Sangha Sangha language group of the CAR

[3]CAR(G) = Gbayo language group of the CAR

[4]CAR(MSG) = combined Mbindu, Sangha Sangha and Gbayo groups of the CAR (combined after the χ^2 comparison in section 3.2.1.2 showed no significant difference).

Table 3-9: G6PD haplotype frequencies (with standard error in brackets) for the Austronesian populations studied

Population (n) Haplotype	IND (SS)[1] (39) Freq. (±SE)	IND (SK)[2] (25) Freq. (±SE)	IND (CS)[3] (17) Freq. (±SE)	IND (grp)[4] (81) Freq. (±SE)	POL[5] (46) Freq. (±SE)
I	0.795 (0.065)	0.833 (0.076)	0.529 (0.121)	0.750 (0.048)	0.783 (0.061)
VII	0.205 (0.065)	0.167 (0.076)	0.471 (0.121)	0.250 (0.048)	0.217 (0.061)
II	0	0	0	0	0
III	0	0	0	0	0
IV	0	0	0	0	0
V	0	0	0	0	0
VI	0	0	0	0	0
"Other"	0	0	0	0	0
Med I	0	0	0	0	0
Med VII	0	0	0	0	0

[1]IND (SS) = Indonesian group from South Sulawesi

[2]IND(SK) = Indonesian group from South Kalimantan

[3]IND(CS) = Indonesian group from Central Sulawesi

[4]IND(grp) = combined IND(SS), IND(SK) and IND(CS) groups (combined after the χ^2 analysis in section 3.2.1.3 showed no significant difference)

[5]POL = Polynesian population.

The haplotype category "Other" includes haplotypes not reported by Vulliamy *et al* (1991b). These occur at low frequencies (a total of eleven haplotypes, seven of them found in females). In the four males, two individuals each carried two different haplotypes. The two novel haplotypes (*Nla* III, *Fok* I, *Pvu* II, *Mbo* II, *Bsp* HI, *Pst* I, *Bcl* I) found were - - - - - + - (one found in a Tsonga male and one found in a Merina male) and - + - - + + + (one found in an Antesaka male and the other in a male from the CAR). It can be seen from Figure 2-1 (a map of the common haplotypes) that these haplotypes could have been produced by recombination between haplotype VII (which is common in Africa and elsewhere) and haplotype IV (which is common in Africa) in an African population. They could have been present in the Malagasy ancestral populations at a very low frequency.

No anomalous haplotypes were found in the Austronesian populations. In females there is the problem of inferring the phase of the linked markers and so it is difficult to conclude the presence of a new haplotype, particularly if there are true novel haplotypes found in the population. A possible solution to the problem of establishing phase would be the use of long range PCR.

3.1.3.1 HAPLOTYPE FREQUENCY HISTOGRAMS

The histograms of haplotype frequencies in Figures 3-3 to 3-6 give a visual representation of the similarities and differences between the populations. The Austronesian samples show a preponderance of haplotypes I and VII (Figure 3-4). In the Indonesian populations from south Sulawesi (SS) and south Kalimantan (SK) and the Polynesian population, the ratio of haplotype I to haplotype VII is approximately 80:20. The Indonesian sample from Central Sulawesi, with a ratio of almost 50:50, has a different appearance from the Polynesian and other Indonesian groups. In general, however, the Indonesians and the Polynesians have similar distributions, as evidenced by chi square testing (section 3.2.2.2) where no significant difference between these two populations was found.

By comparison, the African samples show a very wide range of haplotypes (Figure 3-3), with every haplotype except the Mediterranean (MED) haplotypes and haplotype III being represented. As with the allele frequencies, there is considerably more variation in the African populations than in the non-African populations. In the Austronesian samples, haplotype I has the highest frequency, while in Africa it is haplotype VII. The common African deficiency associated haplotype, haplotype VI, is present in all of the groups except the Zulu. The Zambian population has the highest frequency (0.207) of this haplotype (Figure 3-3, Table 3-8).

All of the 7 haplotypes are present in the Malagasy samples (Figure 3-5). The highland groups (Merina and Betsileo) and the Vezo and Mahafaly groups have high frequencies of haplotype I, the common haplotype in the Austronesian populations, compared to the lowland groups (Antesaka and Tsimihety) which have lower frequencies of this haplotype. The Antesaka and Tsimihety groups, however, have a higher frequency of the common African deficiency haplotype, haplotype VI compared to the highland groups and the Vezo and Mahafaly groups. Since haplotype III is absent in the African populations studied so far but present in the Malagasy, it could be a useful marker haplotype in identifying the true Malagasy ancestral population. However, the low frequency at which it is found means that it might not be identified in the ancestral population.

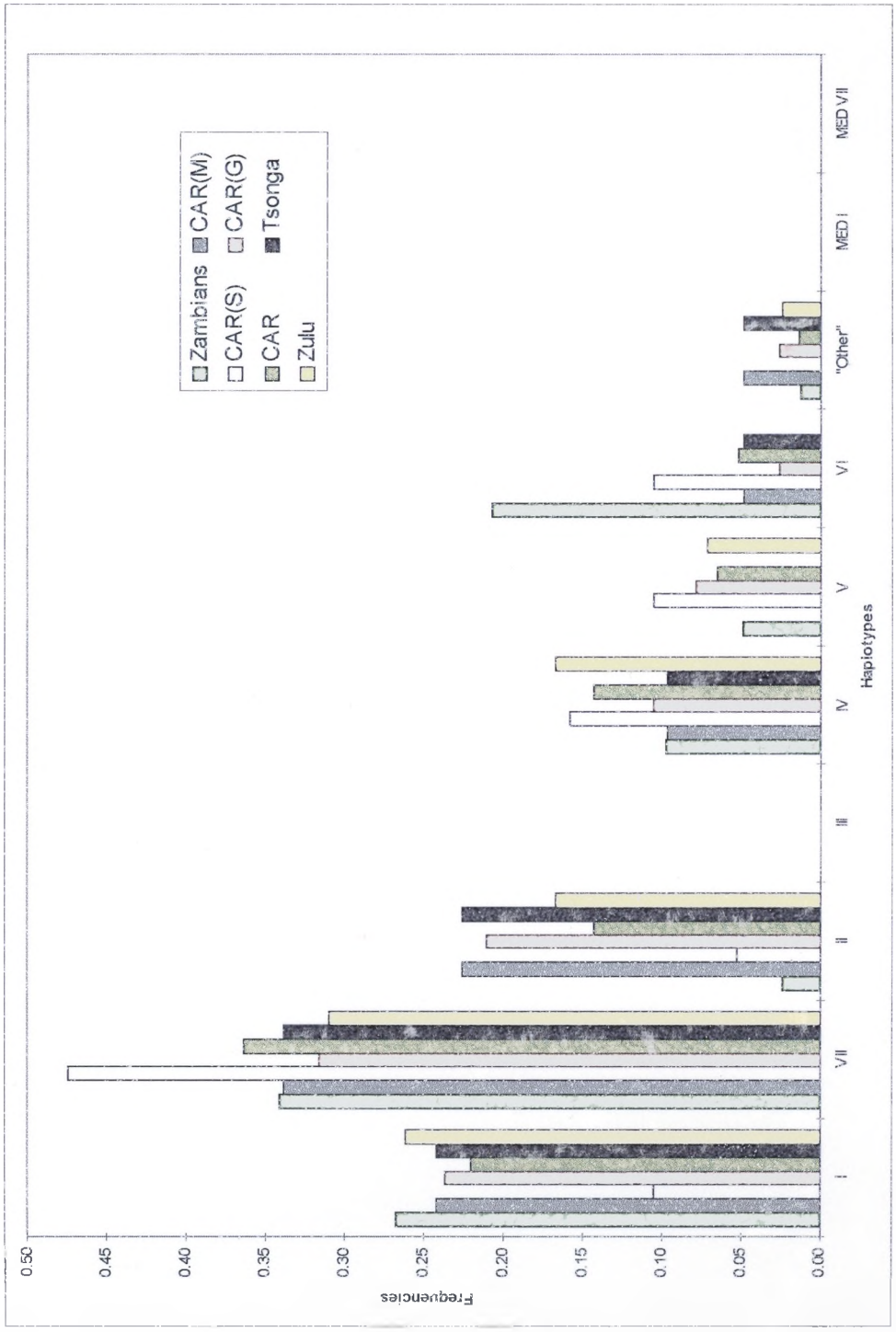


Figure 3-3: Histogram showing the haplotype frequencies at the G6PD locus of the African populations studied

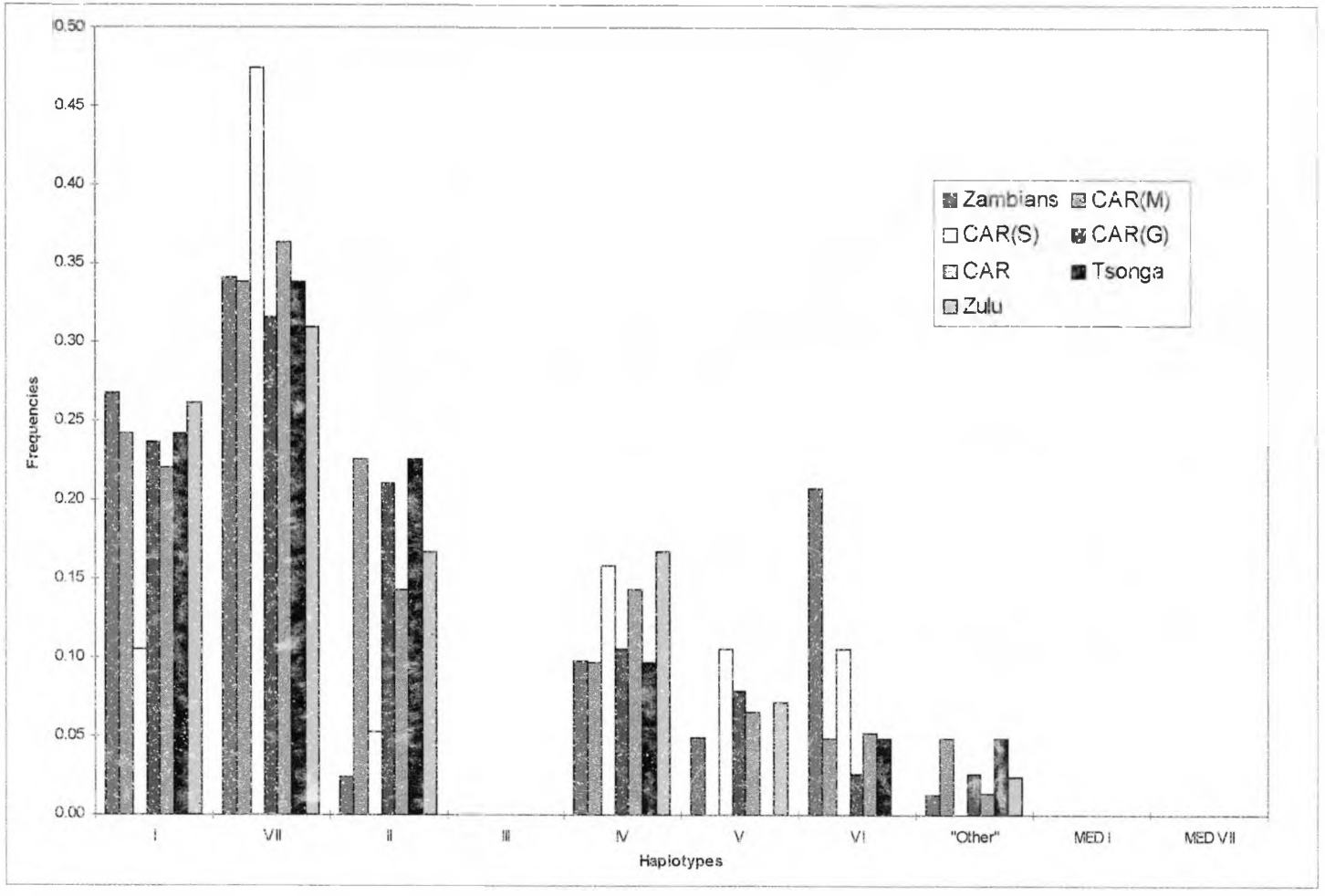


Figure 3-3: Histogram showing the haplotype frequencies at the G6PD locus of the African populations studied

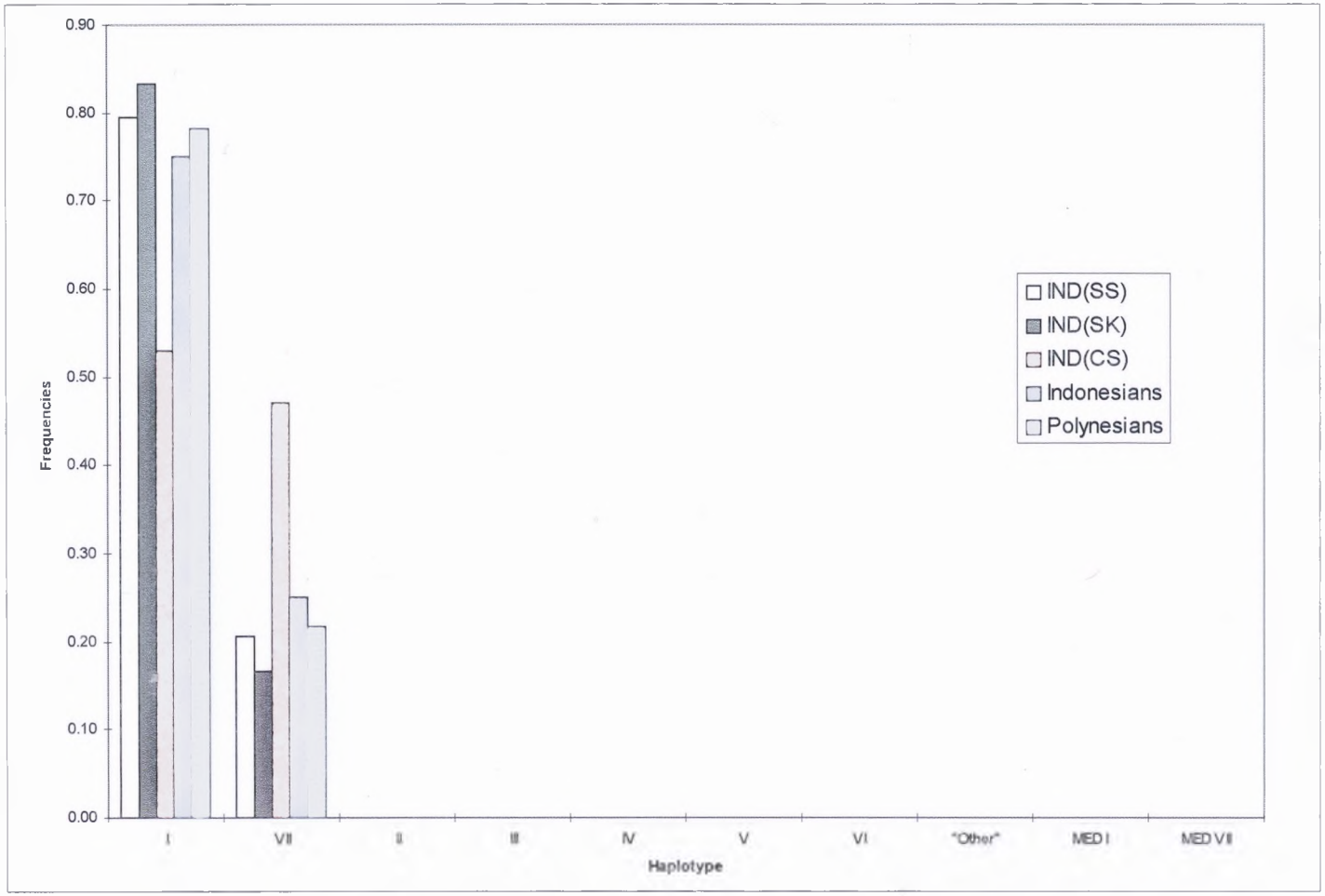


Figure 3-4: Histogram showing the haplotype frequencies at the G6PD locus of the Austronesian populations studied

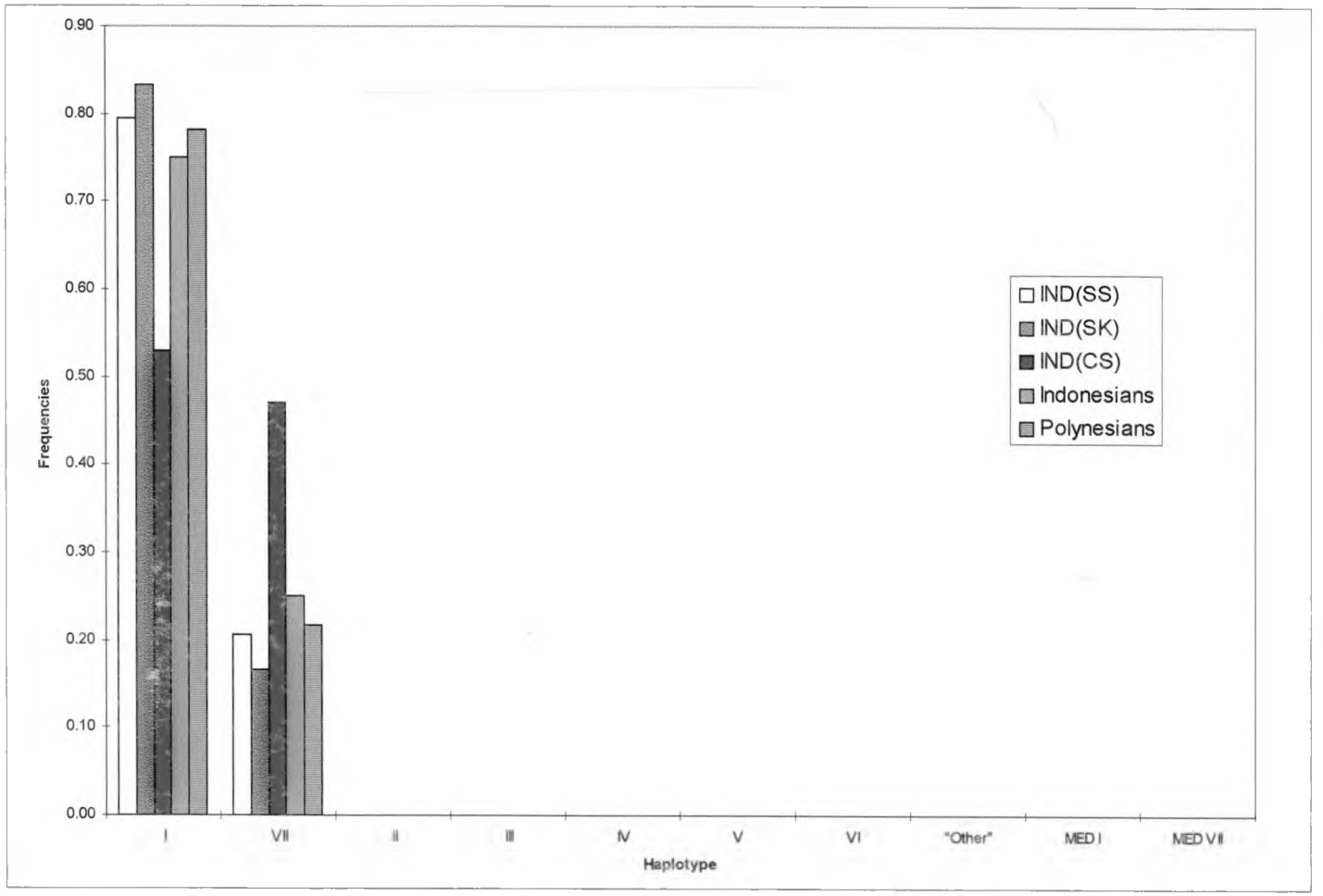


Figure 3-4: Histogram showing the haplotype frequencies at the G6PD locus of the Austronesian populations studied

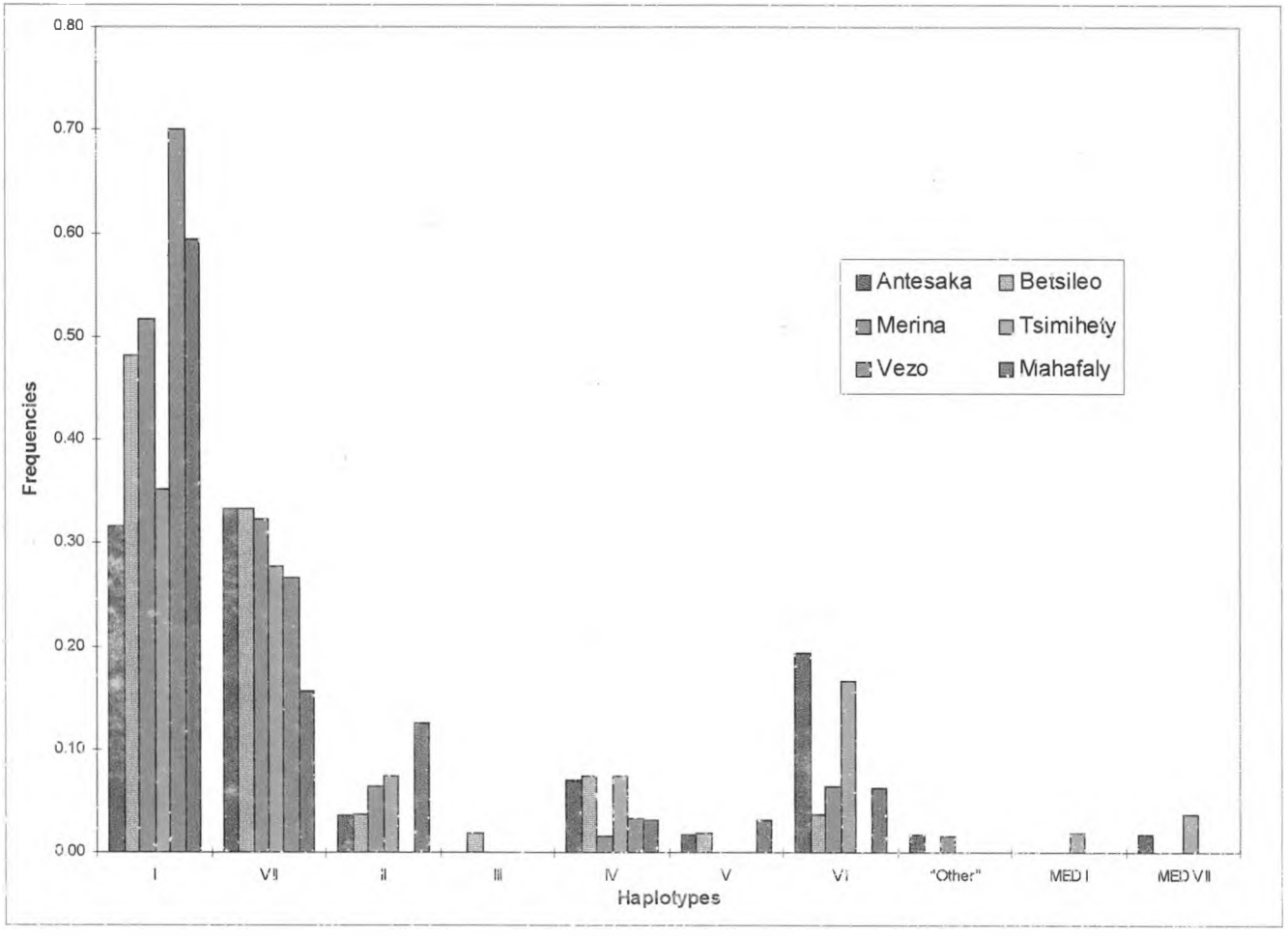


Figure 3-5: Histogram showing the haplotype frequencies at the G6PD locus of the Malagasy groups studied

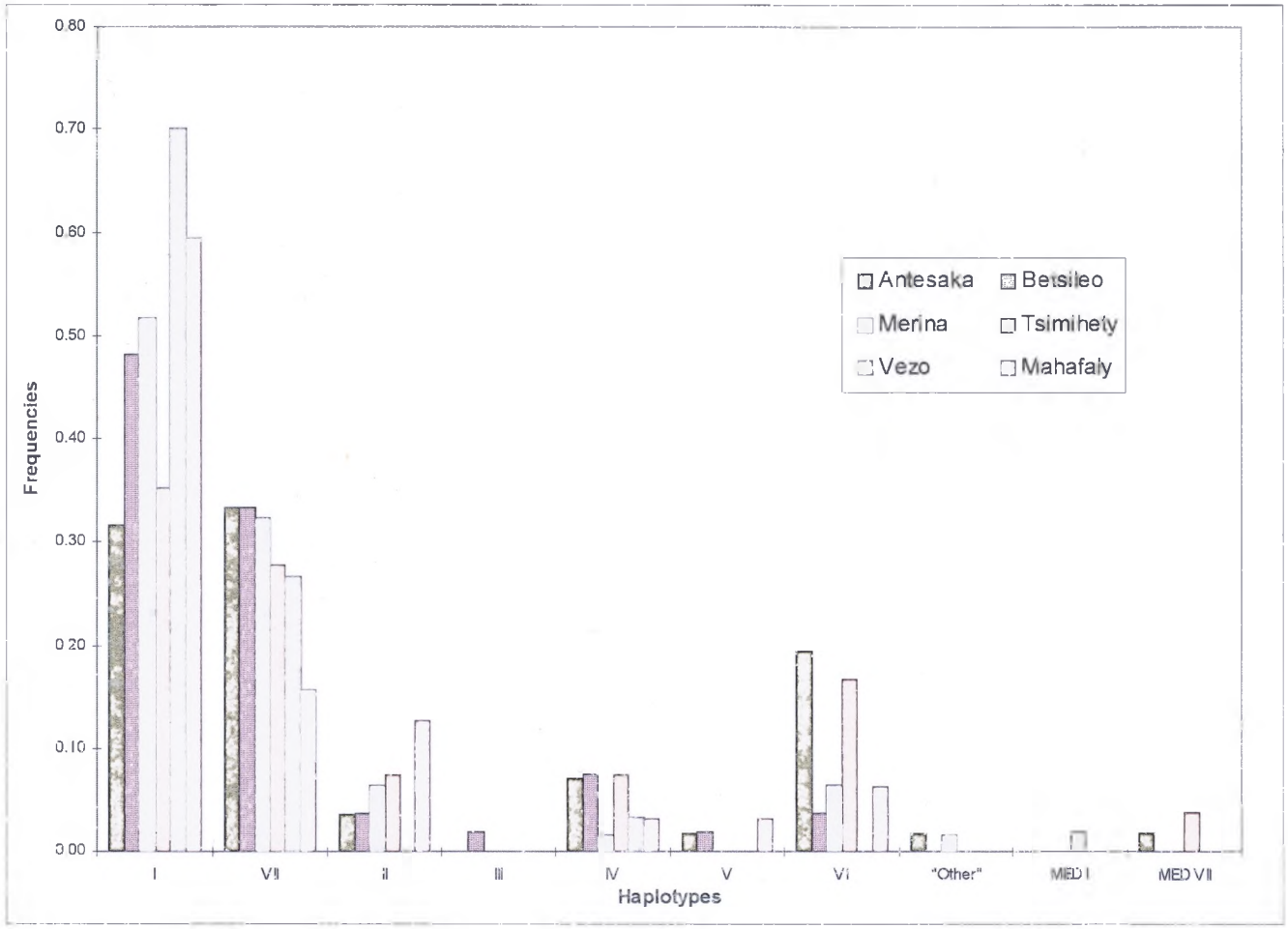


Figure 3-5: Histogram showing the haplotype frequencies at the G6PD locus of the Malagasy groups studied

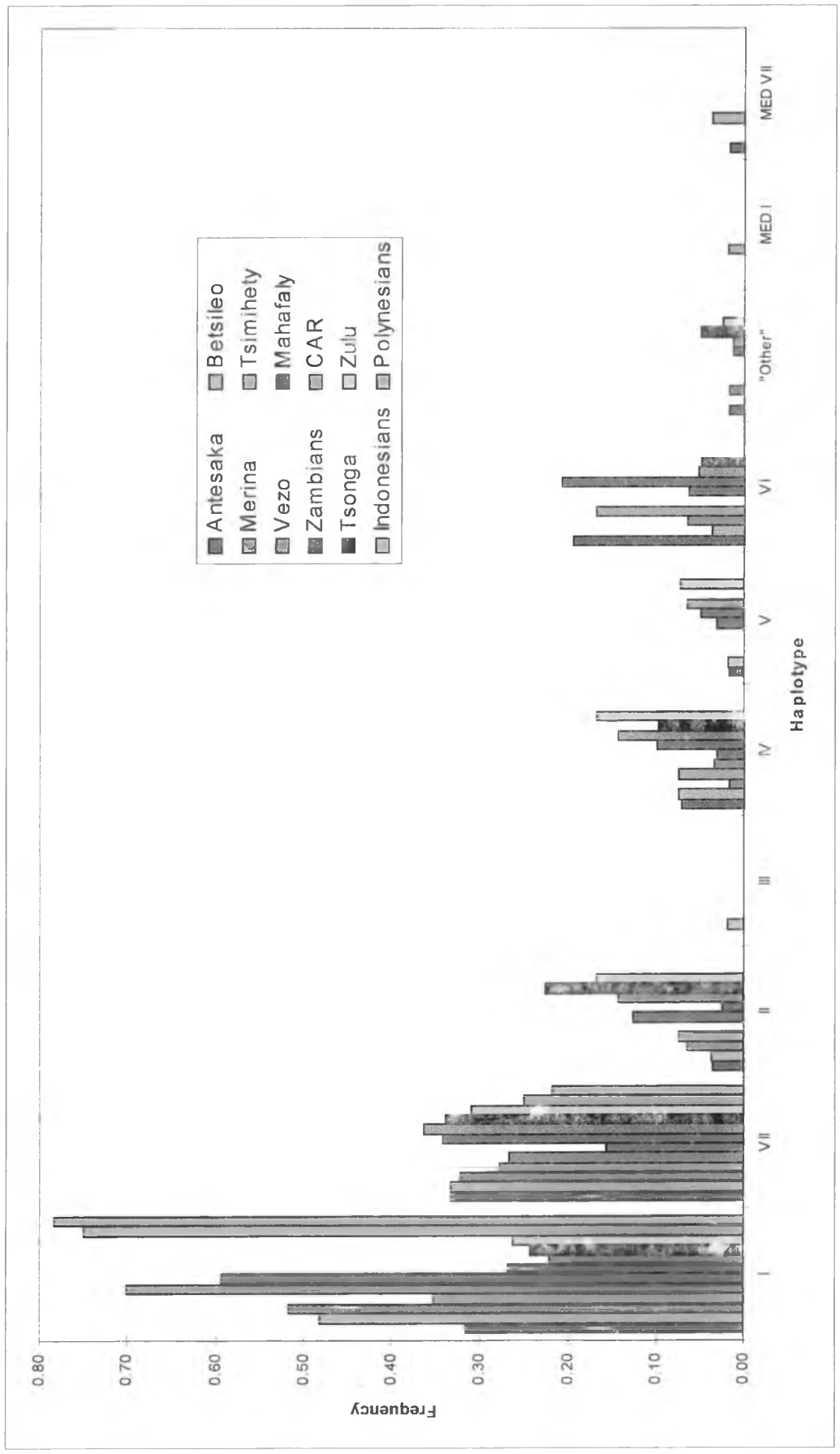


Figure 3-6: Histogram showing the haplotype frequencies at the G6PD locus of all of the populations studied

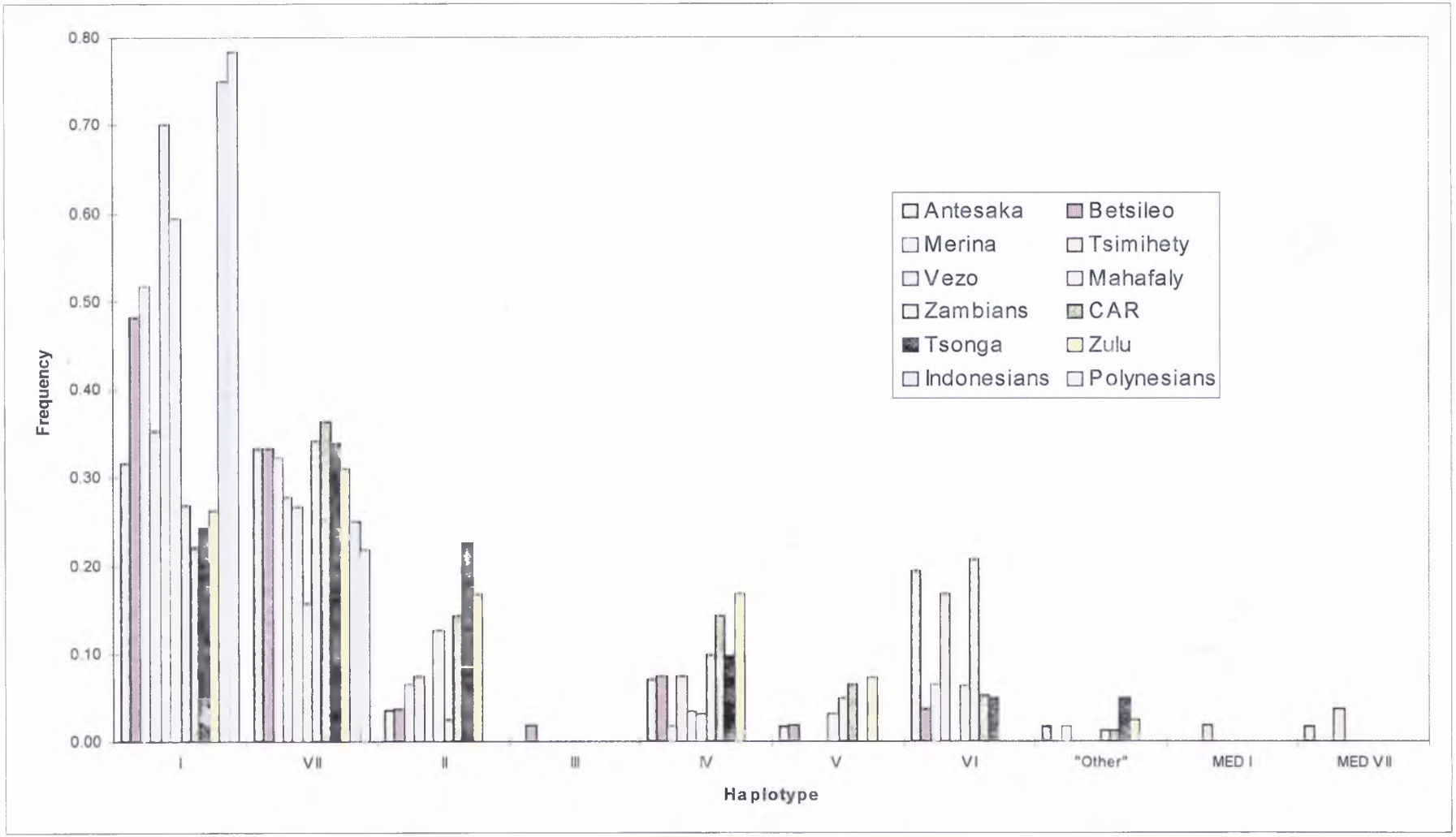


Figure 3-6: Histogram showing the haplotype frequencies at the G6PD locus of all of the populations studied

It is difficult to assess the affinities of the Malagasy groups with individual populations within the Austronesian group because of lack of variation in these samples. The differences between the African populations are also not clear-cut enough to show that one of these is a more likely ancestral population than any other. However, Figure 3-6 highlights the large differences between the African and Austronesian populations.

In comparing all of the samples (Figure 3-6), one can see that the Malagasy samples show frequencies intermediate between the African and Austronesian extremes. The highland and Vezo and Mahafaly groups have frequencies more similar to the Austronesian populations, while the Antesaka and Tsimihety lowland groups have frequencies more similar to the African groups. The Vezo and Mahafaly groups, which are geographically situated in the lowland areas, appear to be more similar to the Austronesian populations than the African populations. However, there is little variation in non-African samples as only 2 haplotypes on B-mobility chromosomes have been found. These similarities therefore do not necessarily indicate affinity with the Austronesian populations but could simply indicate dissimilarity with the African populations or similarity with another non-African population.

3.2 CHI SQUARE (χ^2) ANALYSIS

Chi square analysis was used to determine whether significant differences exist between the populations studied. Since chi square analysis quantitates the difference between groups, it is also a crude measure of distances between populations i.e. if large differences exist between 2 populations, the χ^2 value will be larger (and the p value smaller) than if smaller distances exist between 2 populations.

Comparison was performed using the haplotype frequencies i.e. the ten haplotypes were treated as ten alleles at a single locus. Where the expected number of haplotypes for a particular group was one or absent, that haplotype was grouped with another in the group. The expected figure in each class in the χ^2 analysis was normalised with respect to population size to avoid sample size bias (see Appendix B(II)).

In addition to comparing different populations, different groups within the same population were compared. Where a population was broken down into different subgroups (i.e. the ethnic grouping of the Malagasy, the language grouping of the CAR samples and the geographical grouping of the Indonesians), these subgroups were compared to determine whether the subgroups have a biological basis. If no significant differences existed between subgroups within a population, the groups were combined for further analysis.

3.2.1 χ^2 COMPARISON OF HAPLOTYPES WITHIN POPULATIONS

3.2.1.1 χ^2 ANALYSIS BETWEEN MALAGASY ETHNIC GROUPS

In addition to comparing each Malagasy ethnic group to the other ethnic groups, the highlanders as a group were compared to the lowlanders as a group and the south-west (Vezo and Mahafaly) ethnic groups as a group. The results of the chi-square analysis for the ethnic groups are shown in Table 3-10. When comparing the individual ethnic groups, no significant differences were found, although the p value tends to be lower when comparing a highland to a lowland group.

The χ^2 differences are significant (at the 95%, 99% or 99.9% confidence intervals) in comparing the Antesaka and Tsimihety groups with the Vezo and Mahafaly groups (with the exception of Mahafaly vs. Tsimihety), but not significant in comparing the Vezo and Mahafaly with the Merina and Betsileo groups. The Vezo and Mahafaly groups therefore appear to resemble the highland groups rather than the lowland ones. The Vezo and Mahafaly groups are significantly different to each other at the 95% confidence interval, despite their geographic proximity, possibly because of their strict traditional endogamy (Roux, Institute Pasteur, Madagascar, personal communication), although there could also be sampling bias due to the small sample sizes of the Vezo and the Mahafaly.

There is a significant difference between the combined highland group and the combined lowland group at the 95% confidence interval. The combined lowland group is significantly different from the combined Vezo and Mahafaly group at the 99% confidence interval, but there is no significant difference between the combined southwest group and the combined highland group. This again shows an affinity between the southwest and highland groups. However, this possibly indicates a lower African gene contribution rather than specific admixture from Austronesia or common ancestry. Based on the differences between the southwest groups with other lowland groups and on geographical proximity, they were treated as a group distinct from the other lowland groups in further analysis, despite the difference between them, since the sample sizes are small.

Table 3-10: Chi square values, p values and degrees of freedom for the comparison of the haplotype distributions in the ethnic and regional groups within the Malagasy population

		BT[2]	MR[3]	TS[4]	VZ[5]	MF[6]	LOW[7]	VZ&MF[8]		
AS[1]	χ^2	7.97	10.49	1.28	14.96	11.80			X^2	AS
	df	6	5	5	2	5			df	
	P	0.240	0.063	0.937	$5.63x10^{-4}$‡	0.038*			P	
BT[2]	χ^2		3.66	6.68	5.32	6.00			X^2	BT
	df		5	5	2	5			df	
	P		0.600	0.245	0.070	0.306			P	
MR[3]	χ^2			8.24	4.16	3.79			X^2	MR
	df			5	2	5			df	
	P			0.144	0.125	0.580			P	
TS[4]	χ^2				13.68	7.09			X^2	TS
	df				2	5			df	
	P				0.001†	0.214			P	
VZ[5]	χ^2					6.18			X^2	VZ
	df					2			df	
	P					0.046*			P	
HIGH[6]	χ^2						14.28	4.26	X^2	HIGH
	df						6	5	df	
	P						0.027*	0.513	P	
LOW[7]	χ^2							19.48	X^2	LOW
	df							5	df	
	P							0.002†	P	

*p<0.05; †p<0.01; ‡p<0.001.
[1]AS = Antesaka, [2]BT = Betsileo, [3]MR = Merina, [4]TS = Tsimihety, [5]VZ = Vezo, [6]HIGH = Combined highland groups, [7]LOW = Combined lowland groups, [8]VZ&MF = Combined Vezo and Mahafaly (southwest) groups

3.2.1.2 χ^2 ANALYSIS BETWEEN CAR GROUPS

The 3 language groups of the Central African Republic samples were compared to each other. The results of the chi square analyses are shown in Table 3-11. None of these groups were significantly different from each other. There are some differences, but for practical purposes, the CAR sample may be considered to be a homogeneous group. For this reason, the three groups were grouped together into one group (called CAR(grp)) for comparison to other populations by chi square analysis and in the other analysis methods (e.g. the determination of genetic distances and drawing of phylogenetic trees).

Table 3-11: Chi square values, p values and degrees of freedom for the comparison of the haplotype distributions of the subgroups within the CAR population

		CAR (S)[2]	CAR (G)[3]
CAR (M)[1]	χ^2	4.50	2.87
	df	4	4
	p	0.342	0.579
CAR (S)[2]	χ^2		6.33
	df		5
	p		0.275

[1]CAR(M) = Mbindu group of the Central African Republic

[2]CAR(S) = Sangha Sangha group of the Central African Republic

[3]CAR(G) = Gbayo group of the Central African Republic

3.2.1.3 χ^2 ANALYSIS BETWEEN INDONESIAN GROUPS

The Indonesian samples were taken from three different geographical locations. The chi square analysis of the groups within the Indonesian group are shown in Table 3-12. As suggested by the haplotype distributions, there are subtle differences between the three Indonesian groups. Significant differences do occur between the central Sulawesi (CS) group and both the south Sulawesi group (SS) ($p=0.043$) and the south Kalimantan group (SK) ($p=0.035$) at the 95% confidence interval. Given the small size of the central Sulawesi group ($n=17$), however, it was felt justifiable to combine the three Indonesian populations into one group (IND(grp)). When the Indonesian group was compared to the other populations (in chi-square testing and genetic distance evaluation), this grouped sample was used.

Table 3-12: Chi square values, p values and degrees of freedom for the comparison of the haplotype distributions of the regional groups within the Indonesian population

		IND (SK)	IND (CS)
IND (SS)	χ^2	1.43×10^{-3}	4.09
	df	1	1
	P	0.706	0.043*
IND (SK)	χ^2		4.44
	df		1
	P		0.035*

*$p < 0.05$

3.2.2 χ^2 COMPARISON OF HAPLOTYPES BETWEEN POPULATIONS

Table 3-13 shows the chi square and p values for the haplotype distribution comparison between the populations studied. The grouped sample CAR(grp) includes all of the CAR samples combined on the basis of the absence of significant differences (as discussed in section 3.2.1.2). The (IND(grp)) group includes all of the Indonesian samples, again based on the heterogeneity of the sample (as discussed in section 3.2.1.3). Comparison between different Malagasy groups are not shown in Table 3-13, since they have already been reported in Table 3-10.

3.2.2.1 χ^2 COMPARISONS BETWEEN THE AFRICAN POPULATIONS

Table 3-13 shows that the comparison of different African populations yields significantly different values for most comparisons. This is not surprising given the high degree of variation that has been observed in African populations in previous studies. There is no significant difference between the Tsonga and the Zulu samples at the 95% confidence interval.

3.2.2.2 χ^2 COMPARISONS BETWEEN NON-AFRICAN POPULATIONS

There is no significant difference between the combined Indonesian group and the Polynesian group, the only two Austronesian populations studied in this project. A large amount of variation is not expected in the Austronesian populations compared to the African populations since the African populations are much older, and because of the fact that there are only 3 haplotypes on the B-mobility chromosomes (i.e. little variability outside Africa). The lack of significance in the comparison between the populations is therefore not surprising.

Table 3-13: Chi square values, degrees of freedom (in brackets) and p values for the comparison of the haplotype distributions between the Malagasy ethnic groups and the non-Malagasy population groups

Zambians	CAR(grp)[1]	Tsonga	Zulu	IND(grp)[2]	Polynesians		
2.46 (6)	14.86 (6)	14.16 (5)	15.79 (5)	39.90 (2)	22.24 (1)	χ^2 (df)	**Antesaka**
0.873	0.021*	0.015*	0.007†	2.2×10^{-9}‡	2.4×10^{-6}‡	p	
12.24 (6)	13.26 (6)	12.49 (5)	11.16 (5)	19.23 (2)	9.56 (1)	χ^2 (df)	**Betsileo**
0.057	0.039*	0.029*	0.048*	6.7×10^{-5}‡	0.002†	p	
17.57 (5)	19.71 (5)	16.44 (5)	13.84 (4)	16.51 (2)	8.04 (1)	χ^2 (df)	**Merina**
0.004†	0.001†	0.006†	0.008†	2.6×10^{-4}‡	0.005†	p	
3.49 (5)	9.76 (5)	9.92 (5)	6.51 (4)	38.39 (2)	18.62 (1)	χ^2 (df)	**Tsimihety**
0.624	0.082	0.077	0.164	4.6×10^{-9}‡	1.6×10^{-5}‡	p	
20.54 (2)	24.79 (2)	21.44 (2)	18.03 (2)	2.76 (2)	0.66 (1)	χ^2 (df)	**Vezo**
3.5×10^{-5}‡	4.1×10^{-6}‡	2.2×10^{-5}‡	1.2×10^{-4}‡	0.251	0.416	p	
18.50 (5)	16.15 (5)	12.32 (5)	11.19 (5)	21.69 (2)	3.24 (1)	χ^2 (df)	**Mahafaly**
0.002†	0.006†	0.031*	0.048*	2.0×10^{-5}‡	0.072	p	
22.63 (6)	24.15 (6)	19.97 (5)	21.09 (5)	19.67 (2)	10.80 (1)	χ^2 (df)	**Highlanders**
0.001†	4.9×10^{-4}‡	0.001†	0.001†	5.4×10^{-5}‡	0.001†	p	
6.90 (6)	20.45 (6)	16.91 (5)	19.64 (5)	45.24 (2)	26.39 (1)	χ^2 (df)	**Lowlanders**
0.330	0.002†	0.005†	0.001†	1.5×10^{-10}‡	2.8×10^{-7}‡	p	
27.24 (5)	27.20 (5)	22.00 (5)	18.25 (5)	12.40 (2)	2.39 (1)	χ^2 (df)	**Vezo and Mahafaly**
5.1×10^{-5}‡	5.2×10^{-5}‡	0.001†	0.003†	0.002†	0.122	p	
	0.02 (6)	19.51 (5)	16.13 (5)	50.93 (2)	31.46 (1)	χ^2 (df)	**Zambians**
	0.018*	0.002†	0.006†	8.7×10^{-12}‡	2.0×10^{-8}‡	p	
		2.51 (5)	1.56 (5)	57.31 (2)	37.07 (1)	χ^2 (df)	**CAR(grp)[1]**
		0.775	0.906	3.6×10^{-13}‡	1.1×10^{-9}‡	p	
			1.52 (4)	51.57 (2)	30.97 (1)	χ^2 (df)	**Tsonga**
			0.823	6.3×10^{-12}‡	2.6×10^{-8}‡	p	
				45.92 (2)	23.92 (1)	χ^2 (df)	**Zulu**
				1.1×10^{-10}‡	1.0×10^{-6}‡	p	
					0.17 (1)	χ^2 (df)	**IND(grp)[2]**
					0.679	p	

*p<0.05; †p<0.01; ‡p<0.001.

[1]CAR(grp) = Pooled Mbindu, Sangha Sangha and Gbayo groups of the Central African Republic

[2]IND(grp) = Pooled Indonesian samples from southern and central Sulawesi and Kalimantan.

3.2.2.3 χ^2 COMPARISON OF MALAGASY POPULATIONS TO OTHER POPULATIONS

3.2.2.3.1 GENERAL TRENDS

Given the distance between the African and Austronesian populations and the representation of only 2 of the African haplotypes in the Indonesians, none of the African samples is expected to resemble any of the Austronesian samples. This is confirmed in this study, as can be seen in Table 3-13.

When comparing the African and Austronesian populations to the Malagasy groups, the general trend tends to be greater differences between the Austronesian populations with the Malagasy groups and smaller differences (less significant) for comparison of African groups with Malagasy groups. This could indicate that, in general, there is a greater proportion of African admixture than Austronesian admixture in the Malagasy people. However, it may be misleading to draw too much of a conclusion from this trend. The fact that there is so much more variation in the African groups could again be biasing this assessment. Selection could also influence this. Admixture calculations (section 3.6) may give a more accurate impression.

3.2.2.3.2 χ^2 COMPARISON OF HIGHLAND GROUPS TO OTHER POPULATIONS

In general, the highland Malagasy ethnic groups are significantly different from the Indonesian and Polynesian samples. This is unexpected as the highland groups have generally been thought to be very similar to the Indonesians (as discussed in section 1.1.2.2), although it is not entirely unexpected considering the fact that there is so much more variation in the African samples (and therefore the African component of the Malagasy). A small difference therefore constitutes many more haplotypes and a large increase in the amount of variation. The comparisons of the highland groups with the African populations are generally also significantly different.

Looking at distinct groups within the Malagasy, there is a closer affinity between the highland groups (Merina and Betsileo) and the Austronesian groups (i.e. smaller differences) than between the lowland (and African) groups and the Austronesian groups (i.e. greater differences). The combined highland group is still significantly different from the Indonesian and Polynesian populations, although the trend is a closer affinity between the highland and Austronesian groups than between the highland groups and the African groups.

These results indicate that, while there are significant differences between the Austronesian groups and the highland groups, there is still a closer association between them than between the highland groups and the African groups. This is consistent with other evidence suggesting that the highland groups are closer to the Indonesian ancestors than the lowland groups, who more closely resemble the Africans.

3.2.2.3.3 χ^2 COMPARISON OF LOWLAND GROUPS TO OTHER POPULATIONS

The Antesaka group is significantly different from both of the Austronesian populations. While still significantly different from some of the African populations, the differences appear smaller (for example, ZU: p=0.007, Tsonga: p=0.015), than for the comparisons with the Indonesian ($p=1x10^{-8}$) and Polynesian ($p=2.4x10^{-6}$) groups. The highest p value obtained is for the Zambian population (p=0.873). This is also the only Antesaka comparison yielding a non-significant p value at the 95% confidence interval.

The Tsimihety sample tends to share a closer kinship with the African samples than the Antesaka sample does. There is a significant difference between this group and both of the Austronesian groups. All of the African samples are not different to the Tsimihety group. As with the Antesaka group, the Tsimihety group appears least different from the Zambian population.

The combined lowland group is significantly different from both of the Austronesian populations. The comparisons with the African populations produce p values which are still significantly different at the 99% interval although the values are substantially higher than for the Austronesian comparisons. This indicates that, while the lowlanders as a group are distinct from the African and the Austronesian groups, they share closer affinities with the African than the Austronesian populations. The only exception is the comparison with the Zambian sample, where the p value is not significant. This indicates that the lowland group may be most closely related to a Zambian-like group, but does not indicate that the ancestors of the lowlanders came from Zambia, but perhaps from a related population.

3.2.2.3.4 χ^2 COMPARISON OF VEZO AND MAHAFALY GROUPS TO OTHER POPULATIONS

The Vezo and Mahafaly groups are lowland groups, although they appear to share stronger associations with the Austronesian populations than with the African populations. The comparison of the Vezo group with the Indonesian and Polynesian populations yields no significant differences, but the comparison of the Vezo group with all of the African populations yields significant p values at either the 99.9% or 99% confidence interval. The Mahafaly group associates with the Polynesian group and is significantly different from the Indonesian group. The Mahafaly group compared to the African populations are always significantly different. The comparison of the Vezo and Mahafaly groups with the lowland group Antesaka and Tsimihety (Table 3-10) show significant differences, indicating that the Vezo and Mahafaly groups are not typical lowland groups. When the Vezo and Mahafaly groups are combined, the trends observed are similar to the comparisons of the two groups individually.

The affinity between the Vezo and Mahafaly and the Austronesian populations could simply indicate a non-African association because of low variation in non-African populations, since the haplotypes present in other non-African populations could be the same and the frequencies could be similar (by chance) in the Indonesian people and, for example, an Arabian population. The correlation could, therefore, indicate admixture from another non-African population. To confirm this, other non-African populations would need to be screened for comparison and non-African specific markers would need to be defined.

The Vezo and Mahafaly groups are known to have kept certain strict traditions, apparently since they left Indonesia, in many cases more so than other groups. The Mahafaly group is said to have a large amount of cultural similarity with the peoples of Sulawesi in Southeast Asia, adjacent to Borneo (Dr J Roux, Institute Pasteur, Madagascar, personal communication). It is, therefore, conceivable that the results observed here do indicate Austronesian gene contribution, with low African contribution.

3.3 GENETIC DISTANCES

The results of Nei's unbiased genetic (D) distances (calculated using DISPAN) for the population comparisons are shown in Table 3-14. Genetic distances were calculated for the comparison between the different ethnic groups of the Malagasy, together with the grouped highland, lowland and southwest (Vezo and Mahafaly) populations. The combined CAR group (CAR (grp)) and the combined Indonesian group (IND(grp)) were also used, but not the separate Indonesian and CAR subgroups. The lower the value (i.e. small positive or large negative value) for a comparison, the closer the populations are to each other.

3.3.1 GENETIC DISTANCES BETWEEN AFRICAN POPULATIONS

Comparisons between African populations range from 0.1108 (most dissimilar) for the Tsonga vs. Zambian comparison to -0.0473 (most negative i.e. smallest, therefore most similar) for the CAR(grp) vs. Zulu comparison. The Zulu and the Tsonga group are very close. This was also seen in the chi square analysis (Table 3-13) and is confirmed in the low D value (-0.0321). The CAR(grp) is also close to both the Zulu and Tsonga samples, both in chi square analysis and genetic distance analysis.

3.3.2 GENETIC DISTANCES BETWEEN NON-AFRICAN POPULATIONS

There is a very small D value for the comparison of the two Austronesian populations (-0.0082). This close relationship between the two was also observed in the chi square analysis (Table 3-13). The strong affinity between these two populations is possibly exaggerated because of the low variation in the Austronesian populations compared to that of the African populations. The number of populations studied is also limiting in characterising variation in Austronesia.

Table 3-14: Nei's standard genetic distances (D) (calculated using the DISPAN program) between all the groups which formed the basis of the study

	AS[1]	BT[2]	MR[3]	TS[4]	VZ[5]	MF[6]	High[7]	Low[8]	VZ/MF[9]	Zambia	CAR(grp)[10]	Tsonga	Zulu	IND(grp)[11]
BT[2]	0.0336													
MR[3]	0.0415	-0.0222												
TS[4]	-0.0416	0.0048	0.0056											
VZ[5]	0.1526	0.0062	0.0034	0.0879										
MF[6]	0.1570	0.0311	0.0123	0.0673	-0.0064									
High[7]														
Low[8]							0.0383							
VZ/MF[9]							0.0194	0.1295						
Zambia	-0.0370	0.0839	0.1041	-0.0081	0.2395	0.2463	0.1004	-0.0127	0.2442					
CAR(grp)[10]	0.0566	0.1005	0.1428	0.0667	0.3000	0.2896	0.1289	0.0719	0.2967	0.0532				
Tsonga	0.0871	0.1156	0.1275	0.0699	0.2904	0.2369	0.1276	0.0894	0.2665	0.1108	-0.0160			
Zulu	0.0865	0.0618	0.1045	0.0661	0.2144	0.1871	0.0904	0.0872	0.2031	0.0966	-0.0473	-0.0321		
IND(grp)[11]	0.1904	0.0322	0.0228	0.1187	-0.0155	-0.0011	0.0326	0.1661	-0.0062	0.2866	0.3619	0.3410	0.2684	
POL[12]	0.2172	0.0464	0.0344	0.1365	-0.0140	-0.0029	0.0454	0.1884	-0.0073	0.3200	0.4052	0.3780	0.2992	-0.0082

[1]AS = Antesaka, [2]BT = Betsileo, [3]MR = Merina, [4]TS = Tsimihety, 5VZ = Vezo, [6]MF = Mahafaly, [7]High = combined highland groups (Merina and Betsileo), [8]Low = combined Antesaka and Tsimihety groups, [9]VZ/MF = combined Vezo and Mahafaly groups, [10]CAR(grp) = combined CAR groups, [11]IND(grp) = combined Indonesian groups, [12]POL = Polynesian samples.

3.3.3 GENETIC DISTANCES BETWEEN MALAGASY AND NON-MALAGASY POPULATIONS

In general with the genetic distances (as with the chi square values, Tables 3-10 and 3-13) the African and Austronesian populations form the extremes and the Malagasy groups appear at intermediate positions within the two extremes. So, for example, the D value for the comparison of the Polynesian and CAR(grp) populations is 0.4052, the largest D value generated. The combined lowland group compared to the Polynesian population produces a D value of 0.1884 and comparison of lowland to CAR(grp) population produces a D value of 0.0719, placing the lowland group intermediate between Polynesians and the CAR(grp).

The phylogenetic trees generated from the genetic distance data (section 3.4) are a more effective way of visualising the genetic distance data than by values. A mention of general trends will therefore be made here instead of a detailed discussion of each comparison.

There are low D values between the highland groups and the Austronesian populations (for example, Merina vs. IND (grp) - D=0.0209), but higher values compared to the African populations (for example, Merina vs. Tsonga - D=0.1275). When the Merina and Betsileo groups are combined as the highland group the trend is the same as for the individual groups. Highland vs. Tsonga yields D=0.1276, while highland vs. IND yields D=0.0311.

The lowland groups have high D values (large genetic distances) when compared to the Austronesian populations and low D values (small genetic distances) in comparisons with African populations. There is therefore a stronger association with the African groups than with the Austronesian groups. For both the Antesaka and Tsimihety groups, the association with the Zambians is the strongest of all of the African comparisons with these lowland groups.

The association of the lowland groups with African populations is stronger than the association of highland groups with African populations. Where the comparison of Merina with Tsonga produced D=0.1275, the association of Antesaka with Tsonga yields D=0.0871 and the comparison of Tsimihety with Tsonga yields D=0.0699. In addition, the distance between the highlanders and the lowlanders (0.0383) is greater than the distance between the highlanders and the Indonesians (0.0311) and greater than the difference between the lowlanders and the Zambians (-0.0127). This means that there is possibly more similarity between the groups on the island and the parental populations than similarity between different groups within the island. This could explain some of the variation within groups on the island that has been observed in other markers (for example archaeological - section 1.1.2.3.2 and cultural - section 1.1.2.3.3). This can also be seen in the southwest group frequencies. Interestingly, the Vezo are closer to the Indonesians (-0.0156) than to their geographical neighbour, the Mahafaly (-0.0064). There is a greater distance between the Vezo and Mahafaly (which are close geographically) than between each of these groups and any of the Austronesian groups, again indicating large diversity within the Malagasy, perhaps making pooling of these two unjustified.

The distance between the combined southwest group and the Austronesian groups is smaller than the distance between the highland groups and the Austronesian groups. The D value for the comparison of Vezo and Indonesian populations is -0.0155 while the comparison of Merina and the Indonesian generates a D value of 0.0228. This does not necessarily indicate that the southwest groups are more Indonesian-like than the highland groups, but does indicate again that they are less African-like. Other non-African populations would have to be analysed to determine whether this correlation also indicates an association, particularly since the southwest groups are lowland groups, where a large Indonesian admixture is not expected.

3.4 PHYLOGENETIC TREES

The phylogenetic trees in this section were drawn using the computer program DISPAN (Genetic Distance and Phylogenetic Analysis), using the NJ (neighbour-joining) method as discussed in section 2.2.4.4. Trees were drawn for the individual ethnic groups and for the combined ethnic groups (highland, lowland and southwest groups). Figure 3-7 is a dendrogram of a NJ tree drawn using the distance matrix in Table 3-14, for the ungrouped data (all ethnic groups separately) and Figure 3-8 shows the NJ tree for the distances using the grouped data (highland, lowland and southwest groups).

From Figure 3-7, it can be seen that the populations cluster into two major groups, the African populations on one branch, the Austronesian populations on the other and the Malagasy groups located between these two extremes. The highland groups cluster more tightly with the Austronesian populations than with the African populations and the Antesaka and Tsimihety lowland groups associate more strongly with the African populations than the Austronesian populations. The Vezo and Mahafaly groups are located closer to the Austronesian populations than are the highland groups.

The Austronesian clustering is very close, as would be expected for non-African populations where there is less variation than in African populations. To establish more definitive relationships, one would need to include other non-African groups. The African populations, by comparison, are more spread out, showing greater variation.

The Merina group is closer than the Betsileo group to the Austronesian populations. This is expected for the very Indonesian-looking Merina group. However, the current sample is a mixture of three distinct castes (discussed in section 1.1.2.1). It would be interesting to analyse the 3 castes separately to see if the Andriana (the highest caste) associated more strongly with the Indonesians and if the Mainti (the lowest caste) associate more strongly with the African populations. Small sample sizes at present prevent this type of analysis.

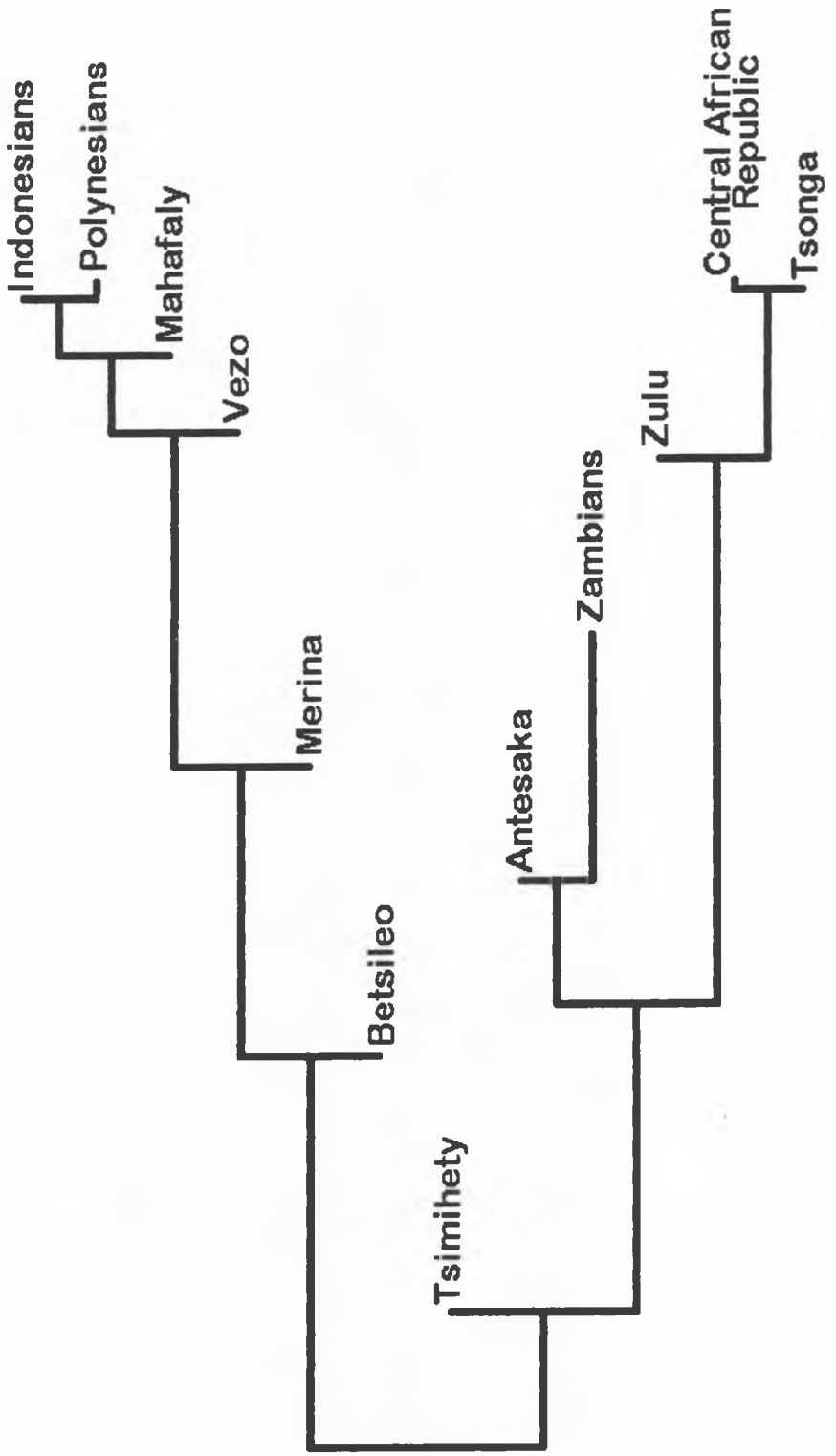


Figure 3-7: Neighbour Joining tree drawn in DISPAN, based on Nei's standard genetic distances for the ungrouped data

Both of the lowland groups are closer to the Zambian group than to any other African group. The association between Antesaka and the Zambians is particularly close, as was noted in the analysis of the genetic distances. However, the other African groups are not so distant that they could be excluded as putative ancestral populations. All of the African populations are found on the same branch.

From the length of the branches, it can be seen that there is more internal variation within Madagascar than between Malagasy groups and putative ancestral populations (e.g. the Merina group is closer to the Indonesian population than to the Antesaka group). This trend can be confirmed quantitatively by genetic distance investigation (Table 3-14).

When the grouped data are used to construct a tree (Figure 3-8), similar trends are observed. The lowlanders still group most closely with the Zambian population and the highlanders still associate more strongly with the Austronesian populations than with the African populations, although the southwest groups associate even more strongly with the Austronesians than do the highland groups.

A and A^- variants may offer a selective advantage to malaria and thus be selected for (as discussed in section 1.2.3.1). Therefore in an effort to reduce the effect of selection and improve the resolution of the trees in the non-African populations, trees were drawn based on genetic distances calculated for B haplotypes only (since only B haplotypes are found in non-African populations). Again, the trees drawn were for the ungrouped (Figure 3-9) and grouped (Figure 3-10) data. Virtually identical relationships noted for the previous trees can be seen in these trees. Since only two non-African parental populations were studied, even this analysis of a subset of data does not clarify the relationships between the groups. More non-African populations need to be studied to establish definitive relationships to non-African populations.

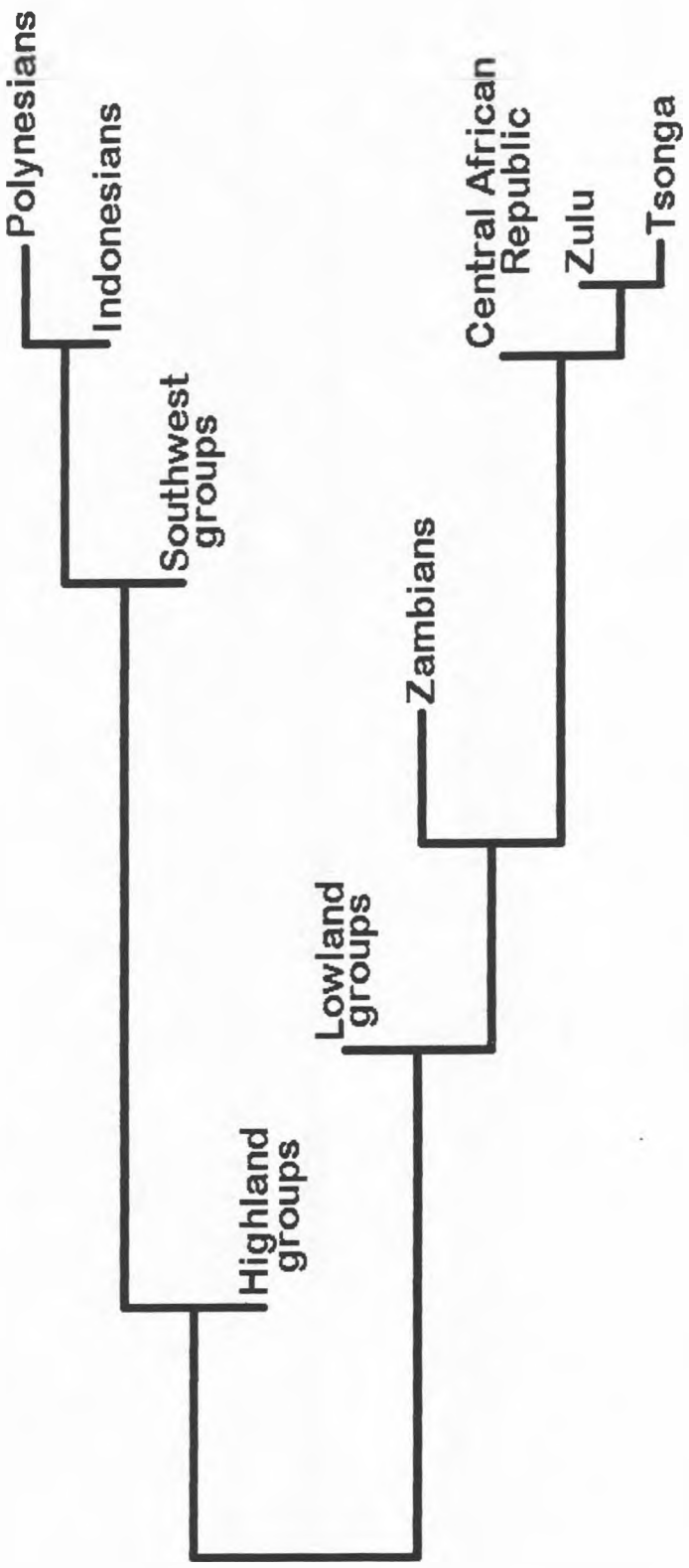


Figure 3-8: Neighbour Joining tree drawn in DISPAN using Nei's standard genetic distances for the grouped data

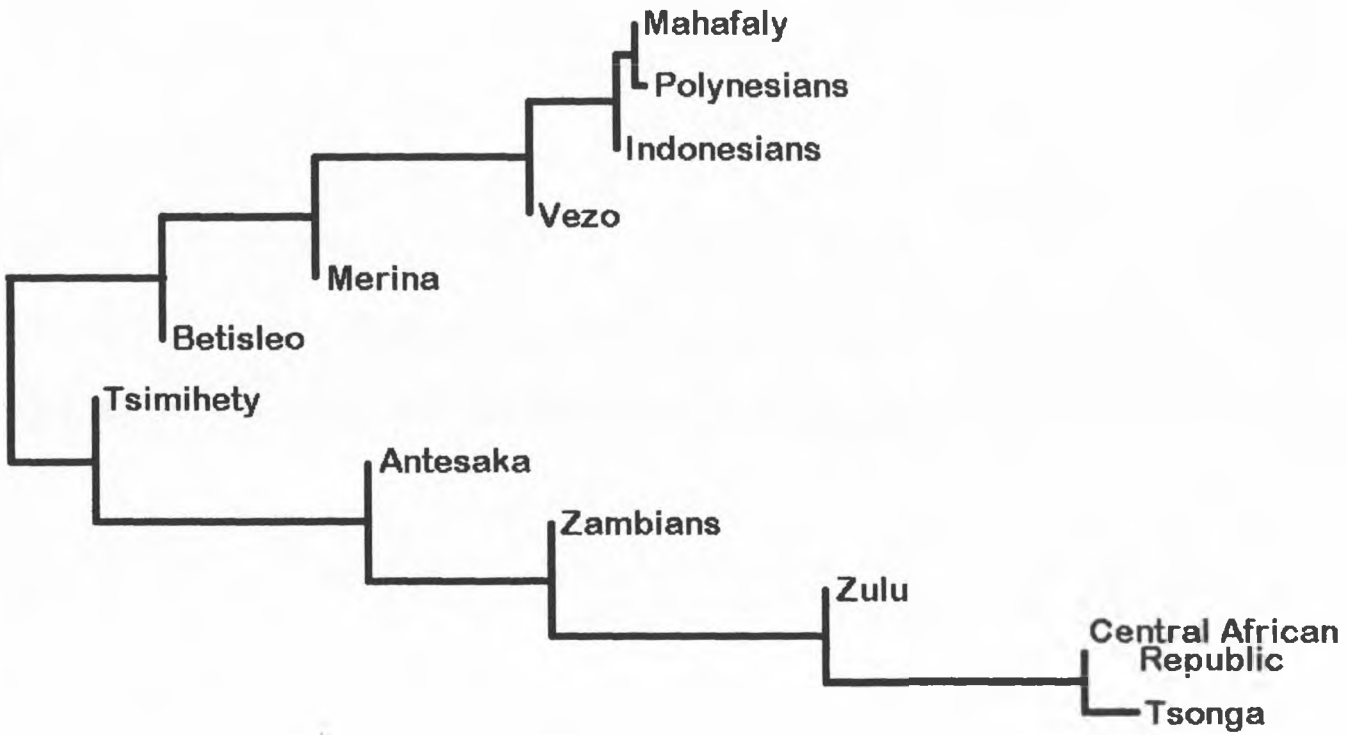


Figure 3-9: Neighbour Joining tree drawn in DISPAN, using Nei's standard genetic distances, for the B haplotypes (numbers I, VII and II) for the ungrouped data

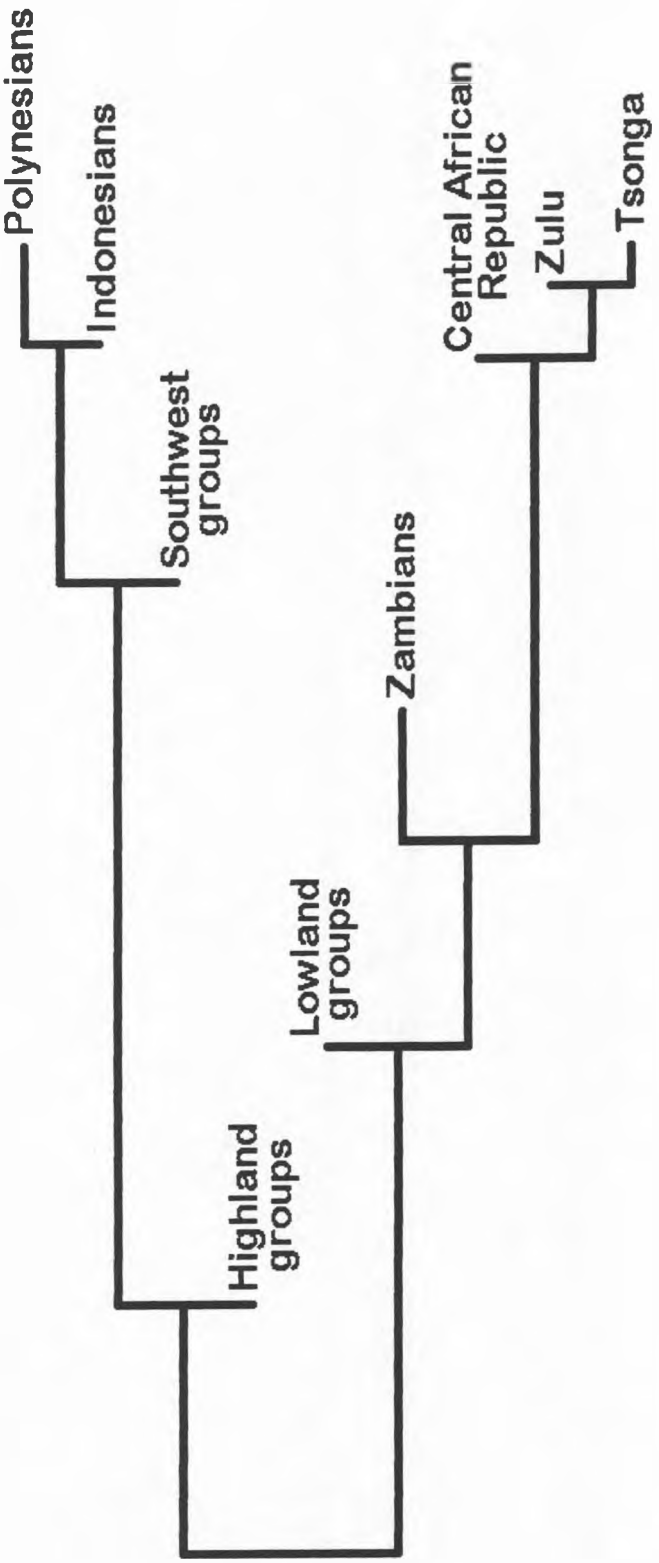


Figure 3-10: Neighbour Joining tree drawn with DISPAN using Nei's standard genetic distances for the B haplotypes (numbers I, VII and II) for the grouped data

3.5 PRINCIPAL COMPONENTS (PC) ANALYSIS

Principal component (PC) analysis provides an alternative method of representing the relationships between the populations. The relationships between the populations can be clearly discerned and both subtle distinctions and overall trends can be evaluated. In this study, PC analysis was performed using the computer program ANTANA, written by Henry Harpending and Alan Rogers of Pennsylvania State College. The data matrix used to determine the principal components was a matrix of haplotype frequencies for the populations. Singular values were first generated to determine the significance of each principal component before plotting the two most significant principal components against each other. PC analysis was performed for all of the populations studied in this project and independently for the Malagasy ethnic groups alone.

3.5.1 SINGULAR VALUES OF PC ANALYSIS

The singular values obtained for the 10 haplotypes when all of the populations were studied are the following:

0.52846; 0.32802; 0.20190; 0.15687; 0.14637; 0.10491; 0.04501; 0.00002; 0.03446; 0.01

The majority of the data from the multivariate decomposition is therefore contained in the first two dimensions. In combination, these two principal components account for 0.52846+0.32802= 0.85648 out of a total of 1.55602 (55.0% of the data). Approximately, 55% of the data can thus be represented by plotting the first two principal components against each other. The amount of data retrieved are often more than this (Cavalli-Sforza *et al* 1994). The remainder of the data is lost in the representation. A further 0.20190/1.55602 (13.0%) of the data could be retrieved in a third dimension, but it was found that the clarity lost in 3D representation was not worth this comparatively trivial gain in data retrieval. When only the Malagasy samples were examined, 59.4% of the data were retrieved in the first two dimensions.

3.5.2 CARTESIAN PLOTS OF PC ANALYSIS

The scatter plot representation of the first two principal components for the 12 populations plotted against each other is presented in Figure 3-11. The numbers represented on the axis are not a measure of genetic distance or allele frequency. They are arbitrary except in determining the scale. The 1st principal component (x-axis) is more significant than the 2nd principal component (y-axis) by a ratio of approximately 5:3, (0.52846:0.32802) and the horizontal displacement is therefore slightly more significant in identifying outlying groups than is the vertical displacement.

The trends shown in Figure 3-11 are similar to those seen in the phylogenetic tree analysis. The Malagasy ethnic groups occur at positions between the extremes formed by the African and Austronesian populations. There is a clustering of lowland groups, along with the Zambian population in the top left quadrant and a clustering of the other African groups in the bottom left quadrant. The Austronesian populations cluster together on the positive x-axis and the highland groups are more closely associated with the Austronesian groups than with the African or lowland groups. The Vezo and Mahafaly associate more strongly with the Austronesian populations than do the highland groups.

The clustering of the Austronesian populations is very tight compared to the loose African structuring, indicating smaller variation between the Austronesian populations than between the African populations. There are larger Cartesian distances between the Malagasy groups than between the African populations, and often closer associations between the Malagasy groups and the putative parental populations than with each other. For example, Merina is closer to the grouped Indonesian sample than it is to either Antesaka or Tsimihety. The close clustering of the Vezo group with the Austronesian populations is particularly striking in this plot. This analysis (Figure 3-11) confirms visually the χ^2 result (Table 3-10) which showed significant differences between the Vezo and Mahafaly groups. In this representation, the Mahafaly associate more closely with the highland groups than with the Vezo group.

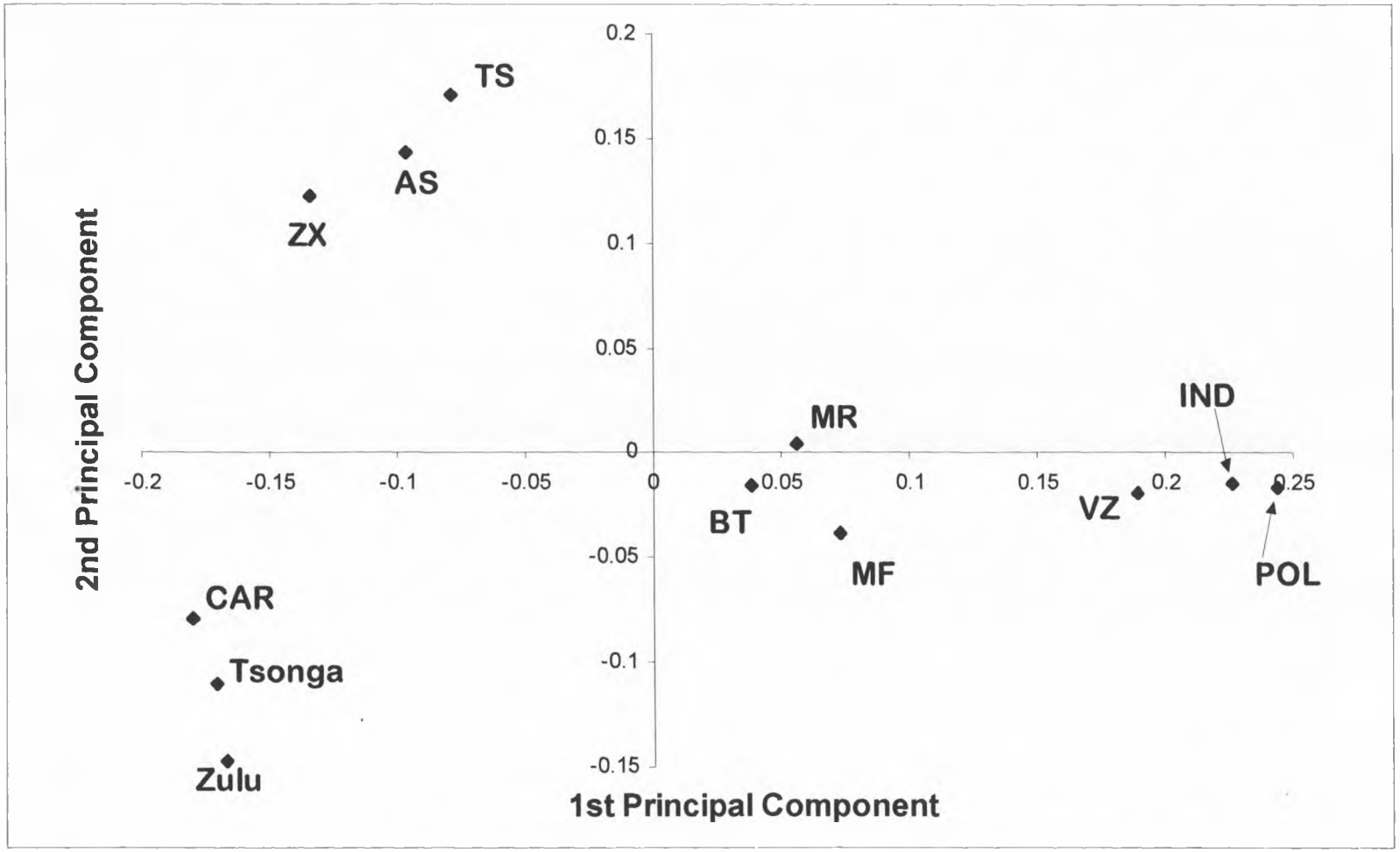


Figure 3-11: Plot of the principal components of the G6PD haplotype frequencies for all of the populations studied

AS = Antesaka, BT = Betsileo, CAR = Central African Republic, IND = Indonesian, MF = Mahafaly, MR = Merina, POL = Polynesian, TS = Tsimihety, VZ = Vezo, ZX = Zambian.

Figure 3-12 shows the scatter plot of the first and second principal components of the singular value decomposition for the Malagasy groups by themselves. The ratio of the first to the second principal components is approximately 2:1 (i.e. 0.39426:0.22180). This indicates that the horizontal axis is more significant than the vertical axis by a factor of 2.

The relationship between the Malagasy groups (Figure 3-12) when considered apart from the other populations is very similar to their relationship when considered together with the other populations (Figure 3-11), although the resolution is much higher. Given that the horizontal axis is more significant than the vertical axis, the Merina, Betsileo, Vezo and Mahafaly are still associating most strongly and the African-like lowland groups Antesaka and Tsimihety are the outliers.

The close relationship between the two highland groups, Merina and Betsileo, can be seen clearly in Figure 3-12. The similarly close relationship between the lowland groups is also apparent as is the larger distance between the geographically neighbouring southwest groups. What is not clear from the plot is whether the relationship between the highland and southwest groups is simply because of distance from the African populations or whether they are truly similar to each other. If the Merina group could be split into different castes and the Andriana class showed clustering with the southwest groups, this could help to answer this question. It should be kept in mind, however, that the sample sizes for the southwest groups are small.

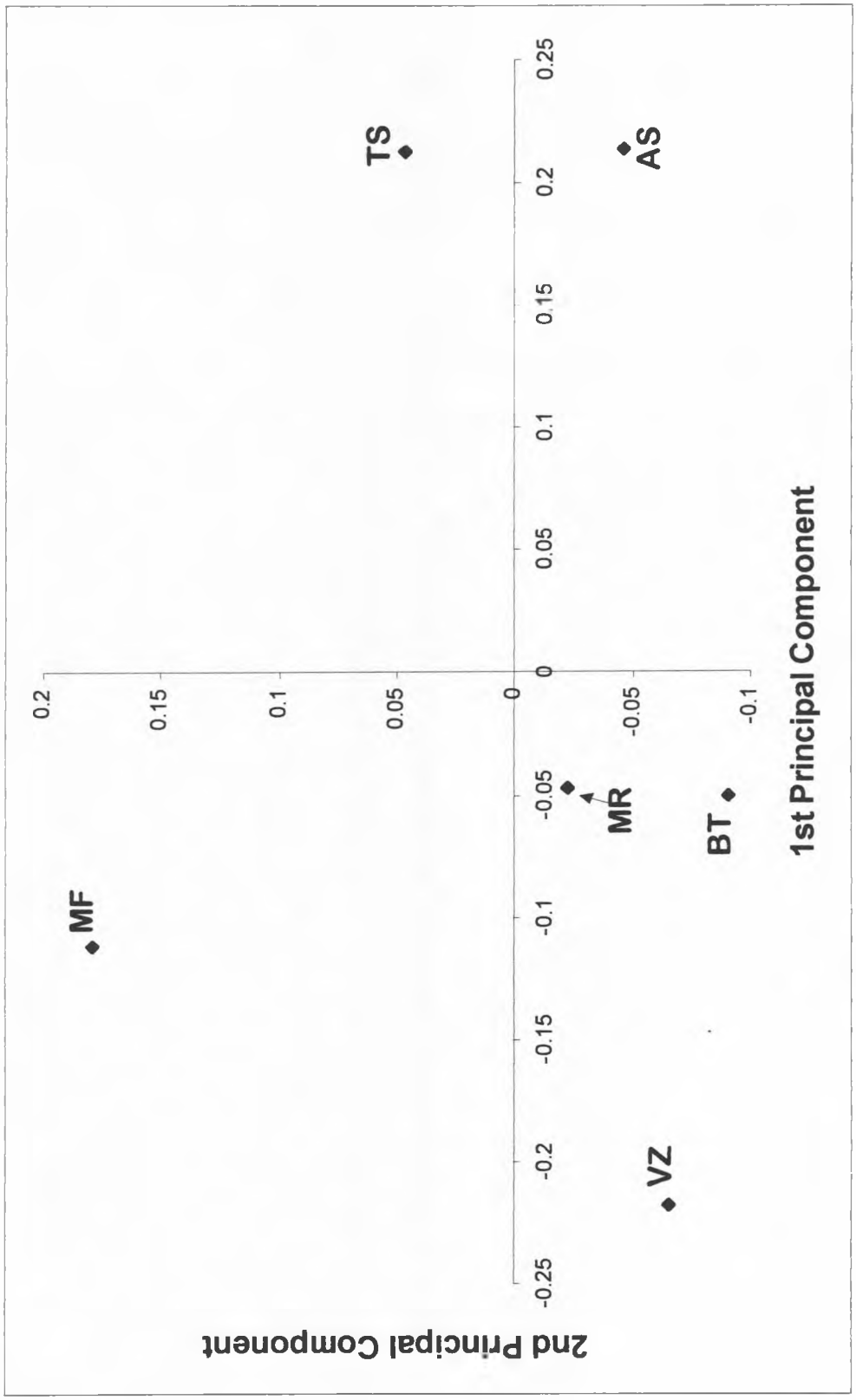


Figure 3-12: Plot of the principal components of the G6PD haplotype frequencies for the Malagasy ethnic groups studied

AS = Antesaka, BT = Betsileo, MF = Mahafaly, MR = Merina, TS = Tsimihety, VZ = Vezo.

3.6 ADMIXTURE CALCULATIONS

The admixture (proportion of parental population contribution) in the hybrid populations was determined using three methods. Bernstein was the first to estimate population admixture using gene frequencies. This is the first method used in this analysis. The second method was the least squares method presented by Roberts and Hiorns. The third method was the least squares method as it was adapted by Elston (as described in section 2.2.4.6.3).

The principle of admixture calculations relies on two major assumptions. The first is that the exact ethnic composition of the ancestral populations is known without error, and the second is that there has been no systematic change in allele frequencies owing to factors other than gene flow between ancestral and contemporary generations within each parental stock and in the hybrid (Chakraborty 1986).

These systematic changes include, amongst others, selection. The G6PD gene frequency may well be affected by the selective advantage conferred against malaria for some alleles. In the polymorphisms analysed, one might expect some amount of selection bias. The extent of this bias is unknown. Further, the admixture calculations may not be entirely valid due to the lack of knowledge of the parental population structure. However, the admixture calculations can give useful information even if the accuracy of this information is compromised by bias, provided the possibility of that bias is recognised. One method to test the accuracy of the admixture calculations, is to perform the calculations on a variety of other non-selected loci. Since this project forms part of a larger study involving many other loci, it should ultimately be possible to get a reasonably accurate admixture value for each putative parental population in each hybrid population group.

3.6.1 THE BERNSTEIN METHOD OF ADMIXTURE CALCULATION

The analysis of m values was performed using the Indonesians as one parental population and the Zambian population combined as the second parental population. The Zambian population was chosen because of the close relationships observed between the Zambians and lowlanders (sections 3.4 and 3.5). The hybrid population frequencies analysed were the entire random sample (all Malagasy ethnic groups) combined into one group, the highlanders as a group, the lowlanders as a group and the Vezo and Mahafaly combined as a group. Although there are differences between the Vezo and Mahafaly groups, this was done for computational expediency and because the admixture results can only be used to indicate general trends. The analysis was performed using each haplotype as an allele. The admixture (m) values for the Zambian parental population using this analysis are shown in Table 3-15. The m values for the Indonesian parental population are (1-m) for each allele.

The major limitation of the Bernstein estimate is that only two parental populations can be investigated at a time. There is a large amount of variation between the African populations, preventing the combining of the populations for one analysis and so the populations need to be investigated one at a time. From the results obtained in previous sections, a Zambian-like population appears to most closely resemble the African parental population and the Zambian data were therefore used for the calculations.

The mean m values show a high proportion of African admixture in the lowlanders (67.87%), a lower proportion of admixture in the highlanders (41.69%) and the lowest proportion of African admixture is observed in the southwest (Vezo and Mahafaly) group (20.38%). For the grouped random Malagasy sample, the Zambian parental population constitutes almost half of the admixture (47.18%). Although there is a large amount of variation, the trend is generally consistent for each haplotype. This is another limitation of the Bernstein method. The m values are calculated for each "allele" (haplotype) separately as it does not allow for the incorporation of many alleles at one locus.

The Bernstein calculation is simply a ratio of haplotype frequencies and the sum of the m values is therefore defined as being 1 (100%). However, in all of the hybrid populations, m values above 100% were obtained for haplotype II. This indicates an invalid underlying assumption concerning the parental population because of a relative overrepresentation of haplotype II in the hybrid population. The frequency of haplotype II in both parental populations (Table 3-8 and 3-9) is lower than that in the hybrid populations (Table 3-7), meaning that the haplotype II frequency in the hybrid populations could not have been contributed to by only the parental populations studied. Similarly, the negative figure for haplotype VII in the southwest hybrid group is caused by proportional under-representation of haplotype VII in the hybrid population. Haplotypes II and VII were therefore excluded from the calculation of the admixture mean.

The presence of the invalid classes (haplotypes II and VI) in Table 3-15 do not invalidate the Bernstein estimate. While they could be accounted for by sampling error, they could also indicate that another parental population was present. The Bernstein calculation, because it is limited to two parental populations, cannot take the possibility of a third parental population into account. The m values were therefore recalculated for the parental population combination of Tsonga vs. Indonesian. This comparison validates haplotype II, but invalidates haplotype VI. It therefore appears that both a Tsonga-like population and a Zambian-like population may have been involved in the ancestry of the Malagasy groups, although it is possible that a different African population, with attributes from both of the putative parental populations examined here, was the true ancestral population. This will be discussed further in section 3.6.3.

Table 3-15: Admixture (m) values of the Zambian parental population in the hybrid populations Total Malagasy, Highlanders, Lowlanders and combined Vezo & Mahafaly using the method of Bernstein

	Total Malagasy	Highlanders	Lowlanders	Vezo & Mahafaly
	m (Zambian)[1]	m (Zambian)[1]	m (Zambian)[1]	m (Zambian)[1]
I	57.91%	50.96%	86.23%	20.23%
VII	49.93%	85.32%	62.82%	-39.38%
II	357.15%	326.94%	346.28%	433.15%
IV	53.20%	44.18%	73.87%	33.06%
V	21.28%	17.67%	18.47%	33.06%
VI	46.73%	24.95%	86.91%	15.56%
"Other"	56.75%	70.69%	73.87%	0.00%
Mean[2]	47.18%	41.69%	67.87%	20.38%

[1] Admixture proportion (m) for the Zambian parental population. The Admixture proportion for the Indonesian parental population is calculated by 1-m.

[2] The mean was calculated only for the valid classes i.e. haplotypes I, IV, V, VI and "Other". See text (section 3.6.1) for detail.

3.6.2 ROBERTS AND HIORNS LEAST SQUARES ADMIXTURE ESTIMATE

A worked example of the calculations used in this section is provided in Appendix B(III).The least square method of Roberts and Hiorns (discussed in Elston 1971) performs admixture analysis for a number of different parental populations (not limited to 2 as in the Bernstein method) and for a number of alleles for a number of loci (each allele is not calculated separately as for the Bernstein method). In this way, the large diversity in the African populations can be taken into account and not assumed to be part of one population. Because of the requirement to exclude one of the "alleles", the haplotype "other" was excluded from the calculation in each example as this haplotype represents a small proportion of the total. The admixture proportions contributed by the parental populations (Zambians, Tsonga, Zulu, Indonesians and Polynesians) were calculated for the same hybrid populations as in the Bernstein method (total Malagasy, highlanders, lowlanders and VZ/MF). These data are shown in Table 3-16(a).

The amount of African admixture shown in Table 3-16(a) in the highland group is lower than in the lowland group. The amount of African admixture is very low in the southwest (VZ/MF) group. As with the Bernstein method, there are several examples of invalid entries (e.g. the Zulu contribution to the total Malagasy, lowland and VZ/MF groups and the Polynesian contribution to the Highlanders, Lowlanders and VZ/MF groups). The sum of the m values for all of the parental classes should equal 1. The presence of invalid results indicates invalid underlying assumptions of the parental populations. This can be useful in eliminating parental populations. The repeatedly invalid Zulu, Indonesian and Polynesian contributions could be indications that some or all of these populations do not represent parental populations. The data were therefore re-analysed with the Zulu and Polynesian populations excluded from the analysis. These results are shown in Table 3-16(b).

Table 3-16: Admixture proportion (m) values for the African and Indonesian parental populations in the hybrid populations Total Malagasy, Highlanders, Lowlanders and Vezo & Mahafaly groups using the method of Roberts and Hiorns

a) Admixture calculated for the parental populations Zambians, Tsonga, Zulu, Indonesians and Polynesians

	Total Malagasy	Highlanders	Lowlanders	Vezo & Mahafaly
	m	m	m	m
Zambian	40.68%	24.57%	76.99%	5.83%
Tsonga	26.88%	4.29%	44.86%	36.97%
Zulu	-9.36%	6.03%	-30.49%	-0.31%
Indonesian	11.15%	233.92%	-102.64%	-201.95%
Polynesian	30.02%	-169.25%	109.70%	260.21%

b) Admixture calculated for the parental populations Zambians, Tsonga and Indonesians

	Total Malagasy	Highlanders	Lowlanders	Vezo & Mahafaly
	m	m	m	m
Zambians	39.57%	28.68%	73.09%	-0.07%
Tsonga	16.38%	24.85%	9.55%	12.76%
Indonesians	43.24%	49.95%	14.88%	81.46%

In Table 3-16(b), the Indonesian contribution in the highlanders (49.95%) is larger than the Indonesian contribution in the lowlanders (14.88%) and the largest Indonesian contribution is found in the southwest (VZ/MF) group (81.46%). For the total Malagasy group, the Indonesian contribution is less than half (43.24%).

In general, there seems to be a larger African contribution from a Zambian-like population than the Tsonga-like population sample in most of the groups. The negative value in the VZ/MF group from the Zambian sample would, however, tend to indicate an error in the underlying assumption of the Zambians as a parental population. This could, however, be an indication of the influence of selection or an indication that a single African population is responsible for the African contribution to the Malagasy gene pool, rather than two separate populations. It has already been shown that there are substantial differences between the Vezo and Mahafaly group and this be part of the reason for this negative result obtained. However, to split the combined VZ/MF group into separate groups is not appropriate given the sample sizes (Vezo:n=30, Mahafaly:n=32) and the potential bias due to selection. An attempt could be made to determine if more selection is occurring in these groups if the sample sizes for deficient individuals was bigger.

The limitation of the Roberts and Hiorns method of admixture estimation is that one of the alleles from each locus must be excluded. The exclusion of one of the alleles, however, biases the sample and reduces the accuracy of the calculation. It has been found that the outcome can vary depending on which allele is excluded (as reviewed in Elston 1971).

3.6.3 THE LEAST SQUARES ADMIXTURE ESTIMATION OF ELSTON

A worked example of the calculations used in this section appears in Appendix B(IV). Table 3-17 shows the m values obtained after analysis using the least squares method of Elston (1971). The approach used was the same as detailed above for the Roberts and Hiorns method. Because of invalid results being generated for some parental groups, the Polynesian and Zulu groups were discarded from the analysis and the admixture calculations were performed using the Zambian, Tsonga and Indonesian populations as parental populations. As before, the hybrid populations analysed were the total Malagasy group, the highlanders, the lowlanders and the combined Vezo and Mahafaly groups.

The same trend is observed (shown in Table 3-17) in this method as in the previous 2 methods. The highlanders have an Indonesian admixture proportion of 50.63%, while the lowlanders have a substantially lower Indonesian component (14.48%) and the Vezo and Mahafaly group have a substantially higher Indonesian component (80.47%). The combined Malagasy hybrid group has an Indonesian admixture component of 43.14%. There are different proportions of admixture for the Zambian and Tsonga populations. This could indicate that the African contribution to the gene pool was made by at least 2 different African populations: one similar to the Zambians and the other similar to the Tsonga, or it could indicate admixture from a single African population with characteristics of both the Zambian-like and Tsonga-like populations. As with the Roberts and Hiorns method, the Zambian-like population seems to contribute a larger proportion of the African genes than the Tsonga-like population in most of the groups. In the VZ/MF group, however, the contribution by the Zambian-like population appears to be smaller than that of the Tsonga-like population.

Table 3-17: Admixture (m) values of the parental populations Zambian, Tsonga and Indonesia for the hybrid populations Total Malagasy, Highlanders, Lowlanders and Vezo & Mahafaly using the Elston method

	Total Malagasy	Highlanders	Lowlanders	Vezo & Mahafaly
	m	m	m	m
Zambians	40.31%	26.96%	74.74%	3.63%
Tsonga	16.55%	22.41%	10.78%	15.90%
Indonesians	43.14%	50.63%	14.48%	80.47%

3.6.4 EVALUATION OF THE THREE ADMIXTURE METHODS

Of the three admixture estimates used in this project, the least squares method of Elston can be expected to be the most accurate since it can incorporate more than two parental populations (unlike the Bernstein method) and one of the classes need not be excluded (unlike the Robert and Hiorns method). However, all three methods have shown some internal validity and it has been shown that various different methods of admixture calculation can produce similar results of admixture proportions (Elston 1971).

3.6.4.1 PROPORTION OF INDONESIAN ADMIXTURE

The method of Bernstein generates an average proportion of Indonesian admixture value for the Vezo and Mahafaly group of 79.62% (calculated from Table 3-15 using 1-m), while the Roberts and Hiorns method give a value of 81.46% (Table 3-16(b)) and the method of Elston (Table 3-17) generates a value of 80.47%. All 3 values are within 2% of each other. The proportion of Indonesian admixture in the other three hybrid groups for the three methods becomes more discrepant in hybrid groups where more African admixture is involved. The average Indonesian contribution to the highland groups shows a range of almost 9% for the three methods (41.69% for the Bernstein method to 50.63% for the Elston method) and in the lowland hybrid group, the discrepancy has a range of almost 18% (32.13% using the Bernstein method to 14.8% for the Roberts and Hiorns method).

The Roberts and Hiorns and the Elston methods, which are more accurate than the Bernstein method, produce comparable results for all three hybrid groups. From all three methods there is a consistent trend of highest proportion of Indonesian admixture in the southwest groups, less Indonesian admixture in the highland groups and the least Indonesian admixture in the lowland groups, confirming the trend observed with the other analyses performed (sections 3.4 and 3.5).

3.6.4.2 PROPORTION OF AFRICAN ADMIXTURE

The proportion of African admixture in the Vezo and Mahafaly groups calculated for each of the methods is very small. The average value calculated for the Bernstein method (from the Zambian population, Table 3-15) is 20.38%, using the Roberts and Hiorns method 12.76% (from the Tsonga population, Table 3-16(b)) and for the Elston method 19.53% (for all African groups Table 3-17). This confirms the observation made in previous sections that there is little affinity between the Vezo and Mahafaly groups and the African populations. The negative value obtained for the Zambian proportion of the VZ/MF hybrid population could indicate that the Zambian population is not a parental population of the Vezo and Mahafaly groups. This is supported by the Elston method analysis (Table 3-17), where a very low proportion of Zambian admixture (3.63%) is generated. This is unexpected since the highlanders and lowlanders seem to derive most of their African contribution from a Zambian-like population.

As expected from previous analyses, but contrary to expectations from linguistic and cultural markers, the highlanders have a higher proportion of African admixture than the Vezo and Mahafaly lowland groups. Using the Bernstein method, there is an African (Zambian) contribution of 41.69% in the highlanders. By comparison, using the Roberts and Hiorns method (Table 3-16(b)), this proportion is 28.68%+24.85%=53.53% and using the Elston method (Table 3-17), the African contribution sums to 26.96%+22.41%=49.37%. The large African admixture present in the highland group could be due in part to the presence of all three Merina castes, and it would be interesting to separate out the three castes if more samples become available in the future.

The proportions of African gene admixture due to the Zambian-like and Tsonga-like populations are consistent between the Roberts and Hiorns and the Elston methods in the highland sample. The Zambian-like population seems to have contributed approximately the same proportion of the genes in the highland as the Tsonga-like parental group, with the Zambian contribution slightly higher than the Tsonga.

The lowlanders, as expected, have a larger amount of African gene contribution than the highlanders. Using the Bernstein calculation the m value for African contribution is 67.87% (Table 3-15), while, using the Roberts and Hiorns calculation, the figure is 82.64%. The estimate calculated using the Elston method is 85.52%. The last two values can be expected to be most accurate and they are within 3% of each other. The proportions of Zambian and Tsonga admixture for the two methods are also comparable.

The proportional contributions from the Zambian population and the Tsonga population are not similar in the highland and lowland groups. Both the Roberts and Hiorns method and the Elston method indicate proportional contributions of approximately 7:1 (Zambian:Tsonga) in the lowland group compared to a 1:1 proportion in the highland group. The reason for this could be that different populations made up the majority of the highlanders compared to the lowlanders. This seems to indicate that two or more distinct populations contributed different proportions of genes to the highland and lowland populations. If only one African population was responsible for the majority of African admixture in the Malagasy, one would expect the ratio of the contributions of the populations studied here to be equivalent, even though the magnitude of the total African contribution would still be larger in the lowland than in the highland groups. One can not be this specific, however, since it is not known which African populations contributed most of the genes to the present-day Malagasy and the admixture calculations used here are possibly not adequate for this degree of subtlety of interpretation, especially given the presence of selection possibly biasing the results.

CHAPTER FOUR

GENERAL DISCUSSION

4. GENERAL DISCUSSION

Many of the sections examined so far give similar indications of the ancestry of the Malagasy. Due to requirements of presenting and discussing the results consecutively, however, the different methods have not been compared. Chapter 3 provided detailed discussion of results generated from each type of analysis. This chapter is designed to provide an overview of the results and general discussion from which trends and conclusions can be drawn.

4.1 G6PD DEFICIENCY SCREEN

The samples selected for analysis in the deficiency screen were selected on the basis of enzyme deficiency. A comparison of the haematological and molecular methods reveals good concordance. The haematological screen is therefore an appropriate method of retrieving deficient samples for the analysis of the deficiency variants involved and the calculation of deficiency frequencies in the populations. Unfortunately, enzyme deficiency data were not always available for all of the populations studied.

4.1.1 ENZYME DEFICIENCY IN THE MALAGASY

Haematological data were available for the Malagasy and CAR individuals. For populations for which haematological data are unavailable, the frequencies can be estimated from published reports. A deficiency frequency of 13.6% is reported in Java, Indonesia (Soemantri *et al* 1995). There is a range of deficiency frequencies reported for different regions of Africa. Some deficiency frequencies reported are very low, e.g. in the Khoi (Jenkins and Nurse 1972) and San (Balinsky and Jenkins 1966). However, these populations are not thought to be prominent contributors to the Malagasy gene pool. The frequency reported from the east African coast is 17%-23% (in coastal Tanzania), and even higher (28%) in Nairobi (Kenya) (Allison 1960). The east coast of Africa, particularly Mozambique, is a good candidate for the region of the ancestral contribution to the Malagasy in terms of geographical proximity.

From the admixture results (section 3.6) the major African contributors to the Malagasy gene pool seem to be a Zambian-like and a Tsonga-like population. The published G6PD deficiency frequency in the Tsonga population is 5.3% (Nurse *et al* 1985) and the Zambian A^- deficiency variant frequency found in this study is 22% but could be higher if deficiency variants other than the common A^- are present.

The frequencies of enzyme deficiency in the candidate African populations are high (up to 28% on the African east coast (Allison 1960)) compared to the frequency in Indonesians (13.6%) (Soemantri *et al* 1995). A higher frequency of deficiency could therefore be expected in the more African-like Malagasy ethnic groups. In this study the highest deficiency frequency is found in the Makua group (0.313, Table 3-1), the most African of the Malagasy ethnic groups (Campbell, The Origins of the Malagasy, manuscript in preparation). High frequencies are also found in other lowland (African-like) ethnic groups (Table 3-1).

Caution should be used in comparing deficiency frequencies. Selective pressures in Madagascar could be different compared to the original locations of the ancestral populations, resulting in lower or higher frequencies of deficiency over time. The period over which selection has been acting should also be considered. The Makua group originated as Mozambican slaves who arrived on the island in the 19th century (Campbell, The Origins of the Malagasy, manuscript in preparation). If selection conditions in Mozambique in the 19th century were different from the selective pressure in Madagascar today, the high frequency in this group (Table 3-1) could be due to the lack of time required for equilibrium to be reached under local conditions.

In addition, the large climatic variation results in different distributions of malaria and therefore different selective pressures between the highland and lowland populations. The tropical lowland areas have traditionally been areas of malaria endemicity, whereas the drier highland areas have been malaria-free (Campbell 1991). A larger selective pressure (and thus higher G6PD deficiency frequency), could therefore be expected in the lowland groups compared to the highland groups, irrespective of the frequencies in the ancestral populations.

A similar distribution is observed for frequencies of sickle cell anaemia ($Hb\beta^s$) (which also confers a selective advantage against malaria) in Madagascar, where a high frequency is found in lowland populations compared to highland groups (Hewitt *et al* 1996). The observed distribution of both traits is probably due mainly to the presence of differential selective pressures in the lowland and highland regions. The sickle cell anaemia frequency is highest on the east lowland coast (Hewitt *et al* 1996), whereas G6PD deficiency frequency is highest on the west coast. The $Hb\beta^s$ distribution reported by Hewitt *et al* (1996) for 5 geographical regions was plotted against the G6PD deficiency distribution for the same regions (as shown in Figure 4.1). The Pearson's product moment correlation coefficient (r) was calculated to be 0.740, indicating a relatively strong correlation.

There has been some debate centred around the G6PD deficiency distribution in African populations. Hitzeroth and Bender (1980) have reported that the G6PD deficiency distribution is consistent with the malaria selection hypothesis. However, Jenkins (1972, 1982) has argued that the distribution and the observed lack of correlation between the G6PD deficiency and $Hb\beta^s$ distributions is best explained by population movement of the Bantu peoples and the postulation that the A^- mutation is older than the $Hb\beta^s$ mutation (Jenkins 1972). Assuming that a single African population contributed most of the African genes in Madagascar, differences in the two distributions in Madagascar should only be due to processes occurring after the colonisation of the island. The strong correlation between the two distributions in Madagascar thus lends support to the hypothesis of Jenkins (1972) that the lack of correlation in Africa is based on population movement rather than on selection.

Since $Hb\beta^s$ is found at very low frequencies south of the Zambezi River, the most likely origin of the Malagasy African parental population is north of the Zambezi River (Hewitt *et al* 1996). The G6PD deficiency frequency in Madagascar is concordant with the frequencies found in this area of Africa, notably in Mozambique (Reys *et al* 1970), which is a good candidate by virtue of geographic location. However, the wide G6PD deficiency range in different Malagasy groups makes it difficult to specify a region as the origin of the putative ancestral population.

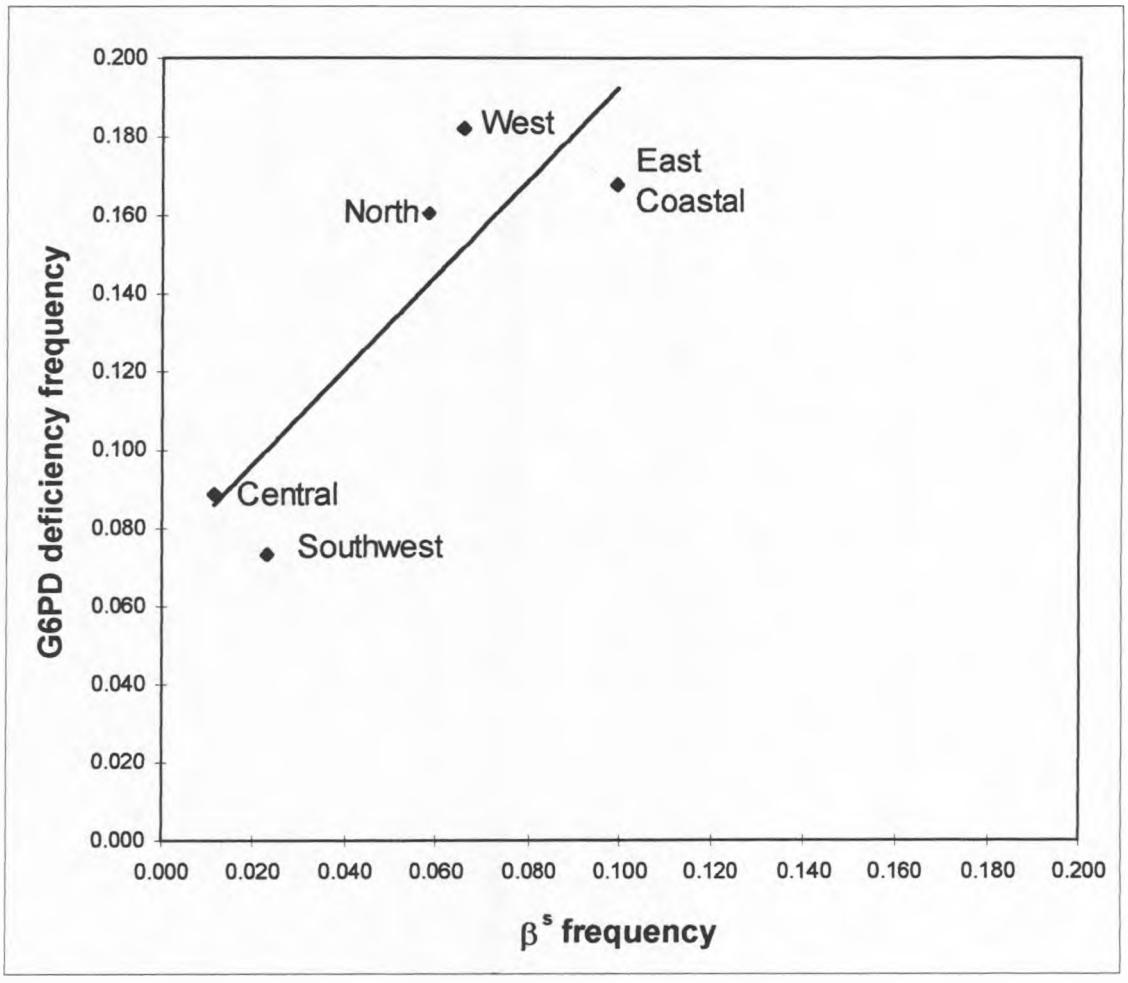


Figure 4-1: Plot of sickle cell distribution compared to G6PD deficiency distribution for the 5 geographical regions with a linear regression line

Geographical regions were compared as opposed to specific ethnic groups because some of the groups sampled were very small, requiring pooling of the data.

The deficiency frequencies in the founders of the ethnic groups should also be taken into account. If a small group of Indonesian settlers contributed to the Malagasy gene pool and the deficiency frequency in this group was (by chance or sampling bias) different to the majority of the Indonesian population from which they came, it could result in deficiency frequencies in the highland populations higher or lower than those expected from analysis of present day Indonesian populations. The possibility of founder effect is also true for the African populations. If there were larger numbers of African settlers, the frequency in the lowlanders could be expected to more closely resemble the African population frequencies because founder effect would be less likely to have operated. However, the numbers of settlers are unknown.

4.1.2 DEFICIENCY VARIANTS FOUND IN MADAGASCAR

By far the most prevalent (87%) deficiency variant represented in the Malagasy is the A^- deficiency variant. The only other deficiency variant shown to be present is the Mediterranean variant (11.5%). The A^- variant has been shown to occur at high frequencies in Africa, with a clinal decrease in frequency from North to South (Jenkins 1972). The fact that it is essentially African-specific is useful in identifying the African contribution to the Malagasy. The variation in frequencies in Africa may be useful (to a lesser extent) in pinpointing particular African populations contributing to the Malagasy gene pool.

In the Zambian population, a good candidate ancestral population by admixture calculations (section 3.6), the frequency of A^- is high (0.22) (Table 3-3). In Mozambique, a good candidate population based on geographical proximity and cultural markers in Madagascar, the frequency is 0.19 (Reys *et al* 1970).

The A^- variant is not present in Indonesia, although the Mediterranean variant is present at frequencies of 31.3% of the total deficiency variants i.e. a gene frequency of approximately 4% (Soemantri *et al* 1995). The frequency of the Mediterranean variant in the Malagasy people is similarly low. However, in the Malagasy sample, 8/9 samples have the Middle Eastern associated haplotype (*Bcl* I (+)). This is not the same haplotype which predominates in Indonesia, although the Middle Eastern haplotype is found in Indonesia at a low frequency (Soemantri *et al* 1995). Also, the regional distribution of the Mediterranean and A^- variants does not entirely correlate with what is known about the ethnic groups. The Mediterranean variant would be expected to occur at higher frequency in highland groups, but in this study it has only been found in lowland groups, perhaps because selective pressure has increased an initial low frequency in this area. This could indicate Indonesian contribution to the lowlands. Similarly, the the A^- variant is an indicator of African contribution in the highlands. The haplotype frequency studies (discussed in section 4.2.2) also support this indication of both African and Indonesian contribution to the peoples of the highland and lowland areas, occurring in different proportions in different groups.

The deficiency variants are subject to selective pressure. As with the overall G6PD deficiency, therefore, many factors may influence the frequency distribution in the Malagasy sample (discussed in section 4.1.1). One possible explanation for the presence of the Mediterranean variant in the lowland areas and not the highland areas is that, because of the absence of malaria in the highland areas, the Mediterranean variant frequency has not increased and therefore has not been detected there because of the low frequency. The greater frequency of the A- variant all over the island allows it to be detected even though its frequency may not have increased over time, especially in the highlands. By analysis of the total deficiency frequency on the island, there does seem to be lower enzyme deficiency frequencies in the highland areas. On the other hand, given that the haplotype associated with the Mediterranean variant is the Middle Eastern rather than the Indian haplotype, a more likely explanation is that the Mediterranean contribution in the Malagasy is derived from Arabic/ Islamic sources rather than from Austronesia. This adds support to the linguistic and cultural evidence of substantial contribution from Arabia in lowland communities, particularly in the southeast regions (Brown 1978) (discussed in section 1.1.2.3.7). In a study of sickle cell anaemia (Hewitt *et al* 1996) no Arab-Indian β^s haplotypes were found. The indication from the low frequency of the Middle East associated $G6PD^{Mediterranean}$ deficiency variant is that the Arabian contribution in the Malagasy is small and since the sample size reported by Hewitt *et al* (1996) is relatively small (35 chromosomes) an Arab-Indian β^s haplotype at low frequencies may not have been detected. Again, the lack of representation of Mediterranean G6PD variants associated with the Indian haplotype of Indonesian origin in the highland areas could be due to a decrease in the frequency over time as a result of lower selective pressure in the highland areas as confirmed by analysis of total deficiency frequency (Table 3-1).

Selective pressures in Madagascar could be different for the two deficiency variants. The Mediterranean deficiency variant is a class II variant, while the A^- deficiency variant is a class III variant. It has been shown that the $G6PD^{Mediterranean}$ variant is associated with favism while the A^- is usually not, and so differential selective pressures for these two variants is not impossible. Based on 9 samples, however, it is difficult to verify these suppositions.

Given the large distances between Indonesia and Madagascar, compared to the distances between Africa and Madagascar, one could expect a smaller number of founding settlers from Indonesia than from Africa. This allows for a higher incidence of founder effect from Indonesia which could affect the observed frequencies. Other deficiency variants are present in the Indonesian population. While the Mediterranean variant is the commonest variant in Indonesia, it accounts for only one third of the deficiency variants present. One would therefore expect to observe other deficiencies (at lower frequencies) in the Malagasy. The 3 as yet unidentified non-A deficiency variants present in deficient Malagasy females could be these "underrepresented" variants.

Dewar and Wright (1993) describe the Malagasy as the "Progeny of the Indian Ocean". In many of the Indian Ocean rim countries, other non-Mediterranean, non-A^- deficiency variants (for example Canton, Union, Orissa and Mahidol) occur at high frequencies. The fact that so few unidentified deficiency variants were found in the Malagasy is an indication of the low level of these contributions. The 3 unidentified Malagasy deficiency variants may be due to these non-southeast Asian contributions to the Malagasy gene pool. Further investigations of these unidentified deficiency variants may give an indication of the origins of the ancestral populations providing these small contributions since they are often geographically specific.

4.2 HAPLOTYPE ANALYSIS IN THE RANDOMLY SELECTED MALAGASY

The selection of samples in the random screen was performed on the basis of ethnic group of the parents. This may not be the most appropriate method of analysis since there is no fixed definition of ethnic groups. A possible alternative is geographical classification of the individuals sampled using a grid system or geographical information system (GIS). However, given the presence of ethnic prejudice and geographic isolation there seems to be a relatively large amount of endogamy and so the ethnic classification system used is probably appropriate in that it defines small groups.

χ^2 analysis was used to identify (and to quantitate) differences between groups and populations. Genetic distance calculations, which were also performed, are a more accurate measure of the affinities between groups. For visualisation purposes, the genetic distances were used to construct phylogenetic trees. The second method of visualisation, which is probably more appropriate for this data set since it indicates relationships between samples without implying a temporal process as indicated by a phylogenetic tree, was the calculation and plotting of the principal components. There is much overlap of methodology in all of these methods. While some methods were more appropriate for working with the dataset, a range of methodology was used to compare the results to establish that a trend was valid using more than one method. All of the methods were used to establish affinities between and within the putative ancestral populations, within the Malagasy population and between the Malagasy population and the putative ancestral populations.

4.2.1 CHARACTERISATION OF PUTATIVE ANCESTRAL POPULATIONS

4.2.1.1 HAPLOTYPE ANALYSIS IN AFRICAN POPULATIONS

The large amount of variation within the African populations compared with the Austronesian populations can be seen in the haplotype frequencies (Tables 3-8 and 3-9) and in the frequency distributions (Figure 3-3 and 3-4). At least 7 (possibly 9) haplotypes have been found in the African populations compared to 2 in the Austronesian populations. Many of the RFLP sites examined here are African-specific and this can cause difficulties in this type of population study. With so much variation in the African populations and comparatively little in the southeast Asian populations, the variation in the southeast Asian populations becomes overwhelmed by the variation in the African populations. This is why the study of the deficiency variants is useful. Deficiency variants are found which are specific to Africa and others which are specific to non-African regions. The limitation of the deficiency variants analysis is that quantification of the proportion of contribution from each ancestral population is difficult because of the influence of selection, frequencies in the original populations and unknown factors such as founder effect as discussed in section 4.1.1.

There is also a large amount of variation between the different African populations. This is shown numerically in large χ^2 values (section 3.2.2.1) and large genetic distances (section 3.3.1) between African populations and, graphically, in the loose clustering of the African populations in principal component analysis (section 3.5) and in the phylogenetic tree analysis (section 3.4). As the history of the Malagasy is so poorly defined, identification of the putative parental populations is difficult.

4.2.1.2 HAPLOTYPE ANALYSIS IN AUSTRONESIAN POPULATIONS

Only two Austronesian populations were studied in this project: the Indonesian and Polynesian populations. The two populations are very similar when compared to African populations. There was no significant difference observed in the χ^2 analysis. The association in genetic distance is very strong as is shown in the phylogenetic analysis, and the clustering in the principal components analysis is very tight.

The implications to the Malagasy of the large diversity in the African populations and the small diversity in the Austronesian populations is that at first glance the Austronesian genetic contribution will be masked by the African contribution. However, if less diversity is observed in an ethnic group, this in itself can be an indication of less African contribution in that group (although there are other explanations, such as population bottlenecks). This study would benefit from the addition of other non-African populations. With only two non-African populations studied and particularly since these two populations are rather similar to each other with respect to G6PD variation, the only correlation which can be made is a non-African correlation. If more non-African populations are studied, it may be possible to establish affinities between the Malagasy and another population on the basis of its own character as opposed to on its non-African nature. However, since most of the polymorphisms are specific to Africa, with only two haplotypes occurring outside Africa (and both of these also occurring within Africa) this type of problem is recurrent. Future analysis could include the detection of non-African specific polymorphism for use in this type of system.

Analysis of mtDNA haplotypes associated with a 9bp deletion has revealed that a haplotype common in Polynesia, but rare or absent in Indonesia, is present at high frequencies in the Malagasy (Soodyall *et al* 1995). Because of the paucity of G6PD polymorphism outside Africa, the Polynesian and Indonesian populations tend to associate strongly with all of the statistical methods used in this project. This lack of resolution in the Austronesian populations prevents the corroboration of the results of Soodyall *et al* (1995).

4.2.2 CHARACTERISATION OF THE MALAGASY ETHNIC GROUPS

4.2.2.1 HAPLOTYPE ANALYSIS OF THE HIGHLAND GROUPS

In general the results obtained in this study confirm expectations concerning the ethnic groups. There is good concordance between the results obtained using different methods of analysis. All of the methods display Africa and Austronesia as the extremes between which the Malagasy ethnic groups are located, according to the nature of the Malagasy ethnic group.

The highland group is more closely related to the Austronesian populations than to the African populations in all analyses. Given the more Indonesian physical appearance (Singer *et al* 1957) and cultural identity (Dewar and Wright 1993), this is not surprising. The Merina group associates more strongly with the Austronesian populations than does the Betsileo group and again this corroborates the physical and ethnological evidence (as discussed in section 1.1.2.3).

All 3 of the Merina castes are contained within the Merina group used in this project and it is therefore somewhat surprising that there is such a strong correlation between the Merina group and the Austronesian populations. The Adriana ("nobles") might be expected to have a substantially higher level of Indonesian admixture than the Mainti ("slaves"). It is not appropriate to separate the castes however, because of small sample sizes and because for some Merina individuals, the caste was unknown. In future studies, if the sample sizes could be increased, the Adriana may be expected to display a strong affinity with the Indonesian population, while the Mainti may be expected to show a stronger affinity with the African populations.

The Betsileo sample, while correlating less strongly with the Austronesian populations than does the Merina, still correlates more strongly with the Austronesian populations than with the African populations with all of the methods of analysis. However, there is still substantial evidence for significant African contribution to both highland groups.

4.2.2.2 HAPLOTYPE ANALYSIS OF THE LOWLAND GROUPS

The lowland groups Antesaka and Tsimihety show closer affinities with the African populations than they do with the Austronesian populations. Both the Antesaka and Tsimihety share the greatest affinity with the Zambian group, but the Antesaka is the closer. There is a stronger association between these lowland groups and the Zambian population than between the lowland groups and the highland groups on all analyses. In fact, there is a larger distance between the Antesaka and the Tsimihety than between the Antesaka and the Zambian population using genetic distance analysis (Table 3-14) and on principal components analysis (Figure 3-11).

This provides genetic substantiation of the observation of large variation between the groups on the island. The results show that the distance between the ethnic groups is often larger than the distance between the ethnic groups and the putative parental populations. This is perhaps surprising given the linguistic homogeneity on the island.

4.2.2.3 HAPLOTYPE ANALYSIS OF THE SOUTHWESTERN MALAGASY GROUPS

The most unexpected finding from this project is that the Vezo and Mahafaly groups, which are technically lowland peoples, have a greater Indonesian character than an African one. This is confirmed in the χ^2 (section 3.2), genetic distance (section 3.3), phylogenetic trees (section 3.4) and principal components (section 3.5) analyses. Both the Vezo and the Mahafaly groups correlate more strongly with the Austronesian populations than the Merina group do. The Vezo correlates with the Austronesian populations more strongly than does the Mahafaly group. Also, there is a significant difference between the Vezo and the Mahafaly groups. This difference, however is probably not particularly significant because the sample sizes of the Vezo and Mahafaly are small and because of the strong affinity that both the Vezo and Mahafaly groups show with the Indonesian and Polynesian populations.

The Vezo and Mahafaly groups are geographically lowland groups. However, much of the evidence in this study points to a more Indonesian-like character of these people. The Vezo and Mahafaly groups are also close geographically. However, there are substantial differences between these two groups. The relationship with the non-African populations reported here does not necessarily indicate an association with the Indonesian population. Some ethnological data do bolster this association. For example, the presence of the outrigger canoe as a cultural marker (Mack 1986) and the fact that some of the Indonesian traditions have been maintained in the Vezo and Mahafaly (Dr J Roux, Institute Pasteur, Madagascar, personal communication). The lower G6PD frequencies reported here and the lower sickle cell anaemia frequency reported by Hewitt *et al* (1996) could implicate a more Austronesian genetic affinity for the southwest Malagasy groups although both of these low frequencies could be explained by lower malaria selective pressure due to the drier climate.

4.3 ANCESTRAL POPULATION ADMIXTURE IN THE MALAGASY

The admixture values reported are what might be expected in the highland and lowland hybrid populations. There is a larger amount of Indonesian admixture in the highland populations (approximately 50%) than in the lowland populations (approximately 15%). The Vezo and Mahafaly populations have proportions of Indonesian or Indonesian-like admixture even larger than the highland population (approximately 81%), making them appear to be more Indonesian-like rather than typically lowland populations. The African contribution to the Malagasy gene pool is the inverse to that of the Indonesian proportion (approximately 50% in the highlanders, approximately 75% in the lowlanders and approximately 19% in the Vezo and Mahafaly).

Each method of admixture calculation has furnished further evidence for the affinities. Invalid results generated using the Bernstein method indicate either that more than two parental populations are involved in the ancestry of the Malagasy, or that the putative parental populations being studied in this project are inappropriate. The Robert and Hiorns and Elston methods indicate the proportions of the population admixture if two or more African populations are involved.

In general, the admixture results (as well as other haplotype analyses) implicate Zambian-like and Tsonga-like putative parental populations in Malagasy African admixture. The Tsonga population is a good candidate population by virtue of its geographic location on the east coast of Africa, close to Mozambique. The Zambian population seems to have the greatest affinity with the lowland groups from analysis of genetic distances, principal components and χ^2 analyses. This does not necessarily indicate that these two populations contributed significantly to the Malagasy gene pool, but rather that the Malagasy ancestral population probably bore a genetic resemblance to both of these two present-day populations.

The results from the Roberts and Hiorns and the Elston methods indicate that if both a Zambian-like and a Tsonga-like population contributed to the hybrid populations, the proportion of Zambian-like and Tsonga-like genes contributed varies between the three hybrid populations. In the lowlanders, there is a 7:1 ratio of Zambian-like to Tsonga-like contribution, while in the highlanders, these two contributions appear to be in an equal ratio. In the Vezo and Mahafaly populations, there seems to be little or no Zambian admixture. This could indicate that two or more waves of African people settled on Madagascar and contributed in different proportions to the Malagasy ethnic groups. If only one African population was involved, the ratio of Zambian-like and Tsonga-like contribution found here would be expected to be the same in the three hybrid populations, even though the total African contribution would still be lower in the highland and southwest groups. Alternatively, another explanation is that a single population contributed and the different ratios observed in the highland and the lowland are due to the selective pressure on particular (deficiency) haplotypes. The difference between the Zambian-like and Tsonga-like ratios in the lowland and southwest groups could then be due to a very low selective advantage of certain alleles because the arid region could result in very low malaria levels.

When the three admixture methods were compared (section 3.6.4), it was found that the discrepancy in admixture values between the three groups is larger when a larger proportion of African admixture is encountered. This could be a reflection of the greater diversity found in the African populations or could indicate that more inaccuracy is encountered from the African populations than from the Indonesian populations. If one population with Zambian-like and Tsonga-like characteristics is involved, this could account for this, although selective pressures could also be important.

It is uncertain how accurate the admixture calculations are for G6PD polymorphisms, since the structural gene is subject to selection. The polymorphisms forming the haplotypes used here are probably not subject to selection, removing most of this inaccuracy in the calculations, but it is inappropriate to take these values as absolute indications of parental population contributions. Also, the values do not indicate that these are the parental populations. For example, a large value for the Zambian population could merely indicate that the parental population resembled the Zambian population in gene frequency more than it resembled the Tsonga population.

Based on sickle cell β^s haplotype frequencies, it has been suggested that a major source of ancestry of the Malagasy is derived from Bantu peoples of central or east Africa, but not from Africa south of the Zambezi, where Hbβ^s occurs at low frequency (Hewitt *et al* 1996). This is in concordance with the data presented in the present project, where the Zambian population shows the greatest affinity with the Malagasy ethnic groups. Future prospects of this work would include characterising samples from Mozambique which is a better candidate as an ancestral population than the Zambian population, based on geographic proximity and ethnological evidence in the Malagasy. Few genetic studies have been performed on the populations of Mozambique and it is therefore difficult to draw definitive conclusions at this point. Hopefully future genetic studies will shed light on the plausibility of a Mozambican population being the source of African genetic admixture in the Malagasy.

CHAPTER FIVE

CONCLUSIONS

5. CONCLUSIONS

The G6PD deficiency frequency in the various Malagasy ethnic groups ranges from 0.073 to 0.313. Deficiency frequencies are higher in the lowland groups (which are traditionally thought to be more African in their makeup) than in the highland groups (which are more Indonesian-like). The deficiency frequencies are intermediate between the high deficiencies found in the putative African ancestral populations and the slightly lower deficiency frequency of the putative ancestral Indonesian population. The sizes of the founding populations and the selective factors present on the island are not known precisely and quantitative comparison is therefore not possible. The presence of differing climatic regions could indicate differing selective pressures within the island, particularly as the lowlands are malaria endemic regions.

The deficiency variants identified are the common A^- variant and the Mediterranean variant. The A^- deficiency variant comprises (87%) of G6PD deficiency and the Mediterranean variant constitutes most of the remainder (11.5%) of the deficiency variants. The A^- variant is virtually African-specific, while the $G6PD^{Mediterranean}$ variant is rare or absent in sub-Saharan Africa and more common in Indonesia. The haplotype most commonly associated with the Mediterranean haplotype in Indonesia is the same one common in India. The most commonly occurring $G6PD^{Mediterranean}$ associated haplotype found in Madagascar is the one common in the Middle East. The presence of this could indicate a founder effect from the Indonesian founding population, but more likely represents gene flow from Arabia. The distribution of the variants do not correlate entirely with the expected location of the ethnic groups (i.e. the Mediterranean variant was only found in lowland peoples). This is probably due to the smaller selective pressure of malaria protection in the highland than in the lowland, but could also be due to differing selective advantages conferred by the two variants and/or biasing of frequencies due to founder effects. The A^- variant is present in highland and lowland groups, indicating African admixture in all groups rather than just in the lowland ('African-like') groups. Two A mobility deficiency variants which were not A^- (on the basis of haplotype) and three non-A-mobility deficiency variants (not Mediterranean on the basis of haplotype) were found but not identified.

In addition to the previously reported 7 haplotypes (Vulliamy *et al* 1991b), two novel haplotypes were found. The new haplotypes can be explained by recombination between haplotypes VII and IV. Strong linkage disequilibrium was found between alleles at the 7 RFLP sites, confirming the findings of other reports (Vulliamy *et al* 1991b, Beutler and Kuhl 1990a). The relationships between the Malagasy ethnic groups and the putative ancestral populations were investigated using the techniques of χ^2 analysis, genetic distance calculations, phylogenetic tree analysis and principal components analysis.

There is a larger degree of diversity among the African populations than among the Austronesian populations. The highland Malagasy groups show stronger affinities with the Austronesian than with the African populations. The Antesaka and Tsimihety lowland Malagasy groups have stronger affinities with the African than with the Austronesian groups. This correlates with cultural and linguistic evidence (Dewar and Wright 1993). The Vezo and Mahafaly groups appear to have stronger affinities with the Austronesian populations than with the African populations and these affinities appear to be even stronger than those between the highland and Austronesian groups. This is not entirely unexpected as there is some evidence of Indonesian culture in these groups, as evidenced by cultural markers such as the Indonesian outrigger canoe (Mack 1986; Dr J Roux, Institute Pasteur, Madagascar, personal communication).

Admixture calculations were performed to determine the proportion of admixture in the different groups. The highlanders seem to have an Indonesian admixture of approximately 50%. The Indonesian admixture calculated for the lowlanders is approximately 15%, and the Indonesian contribution to the Vezo and Mahafaly groups was calculated as approximately 81%.

The African contribution seems to have come from two different population types, one resembling the Zambian population (central Africa) and the other resembling the Tsonga population (eastern Africa). The proportion of the contributions of these two African populations was different in the three Malagasy groups. In the lowlanders, the Zambian-like population predominated at a ratio with the Tsonga of 7 to 1; in the highlanders, the proportions were approximately equal and in the Vezo and Mahafaly, there was little or no contribution by the Zambian-like population. While it is possible that various different African populations contributed to the Malagasy gene pool in this way, a more feasible interpretation of the data is that a population other than the one examined here was involved and that the differing ratios here are due to differing selective pressures in the lowland and highland areas.

It is hoped that the data generated in this study may eventually benefit the peoples of Madagascar by increasing the Malagasy awareness of their history and cultural identity. It could also help to clarify the origins of the Malagasy, confirm the evidence of cultural traditions and place the Malagasy population in context with the rest of the world. The characterisation of the G6PD variants involved in deficiency could also be important in public health and diagnosis.

G6PD is only one of many loci which is being studied in a large, interdisciplinary project involving many loci. If a large number of loci are studied, this will reduce errors due to factors such as selection and drift. Unusual features of the putative parental populations can be exploited and combined from many loci to give a clearer overall impression of the contributions of these populations to the present-day Malagasy. Population-specific markers from different loci can be identified and used to trace the contributions of putative parental populations. In combination with non-genetic markers, the genetic description of the Malagasy peoples is an important factor in determining the origins and affinities of the peoples of Madagascar.

CHAPTER SIX

REFERENCES

6. REFERENCES

Allison AC. (1960) Glucose-6-phosphate dehydrogenase deficiency in red blood cells of East Africans. *Nature* **186**:531-532.

Balinsky D, Jenkins T. (1966) Glucose-6-phosphate dehydrogenase: occurrence of the "A" electrophoretic variant in Hottentots, Bushmen and Bantu. *S. Afr. J. Med. Sci.* **31**:124.

Balinsky D, Jenkins T. (1967) Electrophoretic variants of glucose-6-phosphate dehydrogenase and phosphoglucomutase in Bantu and Coloured patients. *S. Afr. J. Med. Sci.* **32**:96.

Bartholomew World Environment Atlas. (1991) *Bartholomew World Environment Atlas.*:41. Harper Collins Publishers, Edinburgh.

Barclay GPT, Jones HI, Splaine M. (1970) A survey of the incidence of sickle cell trait and glucose-6-phosphate dehydrogenase deficiency in Zambia. *Trans. Roy. Soc. Trop. Med. Hyg.* **64**(1)78-92.

Betke K, Brewer GJ, Kirkman HN, Luzzato L, Motulsky AG, Ramot B, Siniscalco M. (1967) Standardization of procedures for the study of glucose-6-phosphate dehydrogenase: Report of a WHO scientific group. *Wld. Hlth. Org. Techn. Rep. Ser.* **366**:1-53.

Beutler E. (1993) Study of glucose-6-phosphate dehydrogenase: history and molecular biology. *Am. J. Hematol.* **42**:53-58.

Beutler E. (1994) G6PD deficiency. *Blood* **84** (11):3613-3636.

Beutler E, Kuhl W. (1990a) Linkage between a Pvu II restriction fragment length polymorphism and G6PD A-202A/376G: evidence for a single origin of the common G6PD A- mutation. *Hum. Genet.* **85**:9-11.

Beutler E, Kuhl W. (1990b) The nt 1311 polymorphism of G6PD: G6PD Mediterranean mutation may have originated independently in Europe and Asia. *Am. J. Hum. Genet.* **47**:1008-1012.

Beutler E, Kuhl W, Vives-Corrons JL, Prchal JT. (1989) Molecular heterogeneity of glucose-6-phosphate dehydrogenase A-.*Blood* **74** (7):2550-2555.

Bienzle U, Ayeni O, Lucas AO, Luzzatto L. (1972) Glucose-6-phosphate dehydrogenase and malaria - Greater resistance of females heterozygous for enzyme deficiency and of males with non-deficient variant. *Lancet* **1**:107-110.

Brequet G, Ney R, Kirk RL, Blake NM (1982) Genetic survey of an isolated community in Bali, Indonesia. II. Hemoglobin types and red cell enzymes. *Hum. Hered.* **32**: 308-317.

Brown M. (1978) *Madagascar rediscovered.* Chapter 3. Damien Tunnacliffe, London.

Buettner-Janusch J, Buettner-Janusch V. (1964) Hemoglobins, haptoglobins and transferrins in the peoples of Madagascar. A*m. J. Phys. Anthropol.* **22**:163-170.

Buettner-Janusch J, Buettner-Janusch V, Coppenhaver D, Reisman R. (1973) Transferrins, haptoglobins and ceruloplasmins among the Merina and Sakalava. *Arch. Inst. Pasteur Madagascar* **42** (1):193-200.

Burney DA, MacPhee RDE. (1988) Mysterious Island. *Natural History* **7**:47-54.

Calabro V, Giacobbe A, Vallone D, Montanaro V, Cascone A, Filosa S, Battistuzzi G. (1990) Genetic heterogeneity at the glucose-6-phosphate dehydrogenase locus in southern Italy: a study on a population from the Matera district. *Hum. Genet.* **86**:49-53.

Calabrò V, Mason PJ, Filosa S, Civitelli D, Cittadella R, Tagarelli A, Martini G, Brancati C, Luzzatto L. (1993) Genetic heterogeneity of glucose-6-phosphate dehydrogenase deficiency revealed by single-strand conformation and sequence analysis. *Am. J. Hum. Genet.* **52**:527-536.

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

Campbell GR. (1996) *Theories of the origins of the Malagasy: the nineteenth century British school. Australes - Études historiques aixoises sur l'Afrique australe et l'océan Indien occidental,* L'Harmattan, Paris, Ed Paillard, pp 127-153.

Carson PE, Flanagan CL, Ickes CE, Alving AS. (1956) Enzymatic deficiency in primaquine-sensitive erythrocytes. *Science* **124**:484-485.

Cavalli-Sforza LL, Menozzi P, Piazza A. (1994) *The history and geogrpahy of human genes*: 25-42. Princeton University Press Chichester, United Kingdom.

Chakraborty R. (1986) Gene admixture in human populations: models and predictions. *Yearbook of Physical Anthropology* **29**:1-43.

Chen EY, Cheng A, Lee A, Kuang W-J, Hillier LD, Green P, Schlessinger D, Ciccodicola A, D'Urso M. (1991) Sequence of human glucose-6-phosphate dehydrogenase cloned in plasmids and a yeast artificial chromosome. *Genomics* **10**:792-800.

Coetzee MJ, Bartleet SC, Ramsay M, Jenkins T. (1992) Glucose-6-phosphate dehydrogenase (G6PD) electrophoretic variants and the PvuII polymorphism in Southern African populations. *Hum. Genet.* **89**:111-113.

Cohen T. (1971) Genetic markers in migrants to Israel. *Isr. J. Med. Sci.* 7 (12):1509-1514.

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

D'Urso M, Luzzatto L, Perroni L, Ciccodicola A, Gentile G, Peluso I, Persico MG, Pizzella T, Toniolo D, Vulliamy TJ. (1988) An extensive search for RFLP in the human glucose-6-phosphate dehydrogenase locus has revealed a silent mutation in the coding sequence. *Am. J. Hum. Genet.* **42**:735-741.

Dewar RE, Wright HT. (1993) The culture history of Madagascar. *Journal of World Prehistory* 7 (4):417-466.

Dorling Kindersley . (1995) *World Reference Atlas - Multimedia Atlas*. Dorling Kindersley, London.

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

Encarta . (1994) *Encarta Multimedia Encyclopaedia*. Microsoft Corporation, USA.

Fairbanks VF, Nepo AG, Beutler E, Dickson ER, Honig G. (1980) Glucose-6-phosphate dehydrogenase variants: reexamination of G6PD Chicago and Cornell and a new variant (G6PD Pea Ridge) resembling G6PD Chicago. *Blood* **55** (2):216-220.

Fey MF, Wainscoat JS, Mukwala EC, Falusi AG, Vulliamy TJ, Luzzatto L. (1990) A Pvu II restriction fragment length polymorphism of the glucose-6-phosphate dehydrogenase gene is an African-specific marker. *Hum. Genet.* **84**:471-472.

Friedman MJ. (1979) Oxidant damage mediates variant red cell resistance to malaria. *Nature* **280**:245-247.

Gaetani GF, Ferraris AM (1988) Recent developments on Mediterranean G6PD. *Br. J. Haematol.* **68**:1-2.

Ganczakowski M, Town M, Bowden DK, Vulliamy TJ, Kaneko A, Clegg JB, Weatherall DJ, Luzzatto L. (1995) Multiple glucose 6-phosphate dehydrogenase-deficient variants correlate with malaria endemicity in the Vanuatu Archipelago (Southwestern Pacific). *Am. J. Hum. Genet.* **56**:294-301.

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

Hirono A, Beutler E. (1988) Molecular cloning and nucleotide sequence of cDNA for human glucose-6-phosphate dehydrogenase variant A(-). *Proc. Natl. Acad. Sci. USA* **85**:3951-3954.

Hirono A, Kuhl W, Gelbart T, Forman L, Fairbanks VF, Beutler E. (1989) Identification of the binding domain for NADP+ of human glucose-6-phosphate dehydrogenase by sequence analysis of mutants. *Proc. Natl. Acad. Sci. USA* **86**:10015-10017.

Hirono A, Ishii A, Kere N, Fujii H, Hirono K, Miwa S. (1995) Molecular analysis of glucose-6-phosphate dehydrogenase variants in the Solomon Islands. *Am. J. Hum. Genet.* **56**:1243-1245.

Hitzeroth HW, Bender K. (1980) Erythrocyte G-6-PD and 6-PGD genetic polymorphisms in south African Negroes, with a note on G-6-PD and the malaria hypothesis. *Hum. Genet.* **54**:233-242.

Huang C-S, Hung K-L, Huang M-J, Li Y-C, Liu T-H, Tang TK. (1996) Neonatal jaundice and molecular mutations in glucose-6-phosphate dehydrogenase deficient newborn infants. *Am. J. Hematol.* **51**:19-25.

Jenkins T. (1972) Genetic polymorphisms of man in southern Africa. MD Thesis, University of London.

Jenkins T. (1982) Human evolution in southern Africa. *Progress in Clinical and Biological Research, vol **103A**. Human Genetics Part A: The Unfolding Genome*. Ed. B Bonné-Tamir, Alan R Liss, Inc., New York.

Jenkins T, Nurse GT. (1972) Blood group gene frequencies. *S. Afr. Med. J.* **46**:560.

Jenkins T, Blecher SR, Smith AN, Anderson CG. (1968) Some hereditery red-cell traits in Kalahari Bushmen and Bantu: Haemoglobins, glucose-6-phosphate dehydrogenase deficiency, and blood groups. *Am. J. Hum. Genet.* **20**:299-309.

Kaeda JS, Chhotray GP, Ranjit MR, Bautista JM, Reddy PH, Stevens D, Naidu JM, Britt RP, Vulliamy TJ, Luzzatto L, Mason PJ. (1995) A new glucose-6-phosphate dehydrogenase variant, G6PD Orissa (44 Ala->Gly), is the major polymorphic variant in tribal populations in India. *Am. J. Hum. Genet.* **57**:1335-1341.

Kaeda JS, Kotea R, Shun NK, Manraj M, Krishnamoorthy R, Bautista JM, Vulliamy T, Mason PJ, Luzzatto L. (1996) G6PD deficiency on the island of Mauritius. *Br. J. Haem.* **93** (supp 1):326.

Kirkman HN, Gaetani GF. (1984) Catalase: a tetrameric enzyme with four tightly bound molecules of NADPH. *Proc. Natl. Acad. Sci. USA* **81**:4343-4347.

Kopelman AE, Ey JL, Lee H. (1978) Phototherapy in newborn infants with glucose-6-phosphate dehydrogenase deficiency. *J. Pediatr.* **93** (3):497-499.

Kurdi-Haidar B, Mason PJ, Berrebi A, Ankra-Badu G, Al-Ali A, Oppenheim A, Luzzatto L. (1990) Origin and spread of the glucose-6-phosphate dehydrogenase variant (G6PD-Mediterranean) in the Middle East. *Am. J. Hum. Genet.* **47**:1013-1019.

Luzzatto L, Mehta A. (1995) Glucose 6-phosphate dehydrogenase deficiency. *Metabolic and molecular bases of inherited disease* **VIII** 7th Ed., Chapter 111:3367-3398. C.R. Scriver

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

Mason PJ. (1996) New insights into G6PD deficiency. *Br. J. Haematol.* **94**:585-591.

Meloni T, Costa S, Dore A, Cutillo S. (1974) Phototherapy for neonatal hyperbilirubinemia in mature newborn infants with erythrocyte G-6-PD deficiency. *J. Pediatr.* **85** (4):560-562.

Meloni T, Corti R, Naitana AF, Arese P. (1982) Lack of effect of phototherapy on red cell riboflavin status and on glucose-6-phosphate dehydrogenase activity in normal and G-6-PD-deficient subjects with neonatal jaundice. *J. Pediatr.* **100** (6):972-974.

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

Motulsky AG, Vandepitte J, Fraser GR. (1966) Population genetic studies in the Congo: Glucose-6-phosphate dehydrogenase deficiency, hemoglobin S, and Malaria. *Am. J. Hum. Genet.* **18** (6):514-537.

Nafa K, Reghis A, Osmani N, Baghli L, Ait-Abbes H, Benabadji M, Kaplan J-C, Vulliamy T, Luzzatto L. (1994) At least five polymorphic mutants account for the prevalence of glucose-6-phosphate dehydrogenase deficiency in Algeria. *Hum. Genet.* **94**:513-517.

Naylor CE, Rowland P, Basak AK, Gover S, Mason PJ, Bautista JM, Vulliamy TJ, Luzzatto L, Adams MJ. (1996) Glucose 6-phosphate dehydrogenase mutations causing enzyme deficiency in a model of the tertiary structure of the human enzyme. *Blood* **87**(7):2974-2982.

Nei M. (1972) Genetic distance between populations. *The American Naturalist* **106** (949):283-292.

Ninfali P, Baronciani L, Ruzzo A, Fortini C, Amadori E, Dall'ara G, Magnani M, Beutler E. (1993) Molecular analysis of G6PD variants in northern Italy: a study on the population from the Ferrara district. *Hum. Genet.* **92**:139-142.

Nurse GT, Weiner JS, Jenkins T. (1985) *The peoples of Southern Africa and their affinities.*:276-279; 332-333. Clarendon Press, Oxford.

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

Oppenheim A, Jury CL, Rund D, Vulliamy TJ, Luzzatto L. (1993) G6PD Mediterranean accounts for the high prevalence of G6PD deficiency in Kurdish Jews. *Hum. Genet.* **91**:293-294.

Pai GS, Sprenkle JA, Do TT, Mareni CE, Migeon BR. (1980) Localization of loci for hypoxanthine phosphoribosyltransferase and glucose-6-phosphate dehydrogenase and biochemical evidence of nonrandom X chromosome expression from studies of a human X-autosome translocation. *Proc. Natl. Acad. Sci. USA* **77** (5):2810-2813.

Persico MG, Viglietto G, Martino G, Toniolo D, Paonessa G, Moscatelli C, Dono R, Vulliamy T, Luzzatto L, D'Urso M. (1986) Isolation of human glucose-6-phosphate dehydrogenase (G6PD) cDNA clones: primary structure of the protein and unusual 5' non-coding region. *Nucleic Acids Res.* **14**:2511-2522.

Porter IH, Boyer SH, Watson-Williams EJ, Adams A, Szeinberg A, Siniscalco M. (1964) Variation of glucose-6-phosphate dehydrogenase in different populations. *Lancet* **I**:895-899.

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

Roth EF, Raventos-Suarez C, Rinaldi A, Nagel RL. (1983) Glucose-6-phosphate dehydrogenase deficiency inhibits in vitro growth of Plasmodium falciparum. *Proc. Natl. Acad. Sci. USA* **80**:298-299.

Rovira A, Vulliamy T, Pujades MA, Luzzatto L, Vives Corrons JLL. (1995) Molecular genetics of glucose-6-phosphate dehydrogenase (G6PD) deficiency in Spain: identification of two new point mutations in the G6PD gene. *Br. J. Haem.* **91**:66-71.

Ruwende C, Khoo SC, Snow RW, Yates SNR, Kwiatkowski D, Gupta S, Warn P, Allsopp CEM, Gilbert SC, Peschu N, Newbold CI, Greenwood BM, Marsh K, Hill AVS. (1995) Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. *Nature* **376**:246-249.

Saha N, Banerjee B (1971) Erythrocyte G-6-PD deficiency among Chinese and Malays of Singapore. *Trop. Geogr. Med.* **23**:141-144.

Saha N, Ramzan M, Tay JSH, Low PS, Basair JB, Khan FM. (1994) Molecular characterisation of red cell glucose-6-phosphate dehydrogenase deficiency in north-west Pakistan. *Hum. Hered.* **44**:85-89.

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

Sodeinde O. (1992) Glucose-6-phosphate dehydrogenase deficiency. *Baillière's Clin. Haematol.* **5** (2):367-382.

Soemantri AG, Saha S, Saha N, Tay JSH. (1995) Molecular variants of red cell glucose-6-phosphate dehydrogenase deficiency in central Java, Indonesia. *Hum. Hered.* **45**:346-350.

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

Stevens DJ, Wanachiwanawin W, Mason PJ, Vulliamy TJ, Luzzatto L. (1991) G6PD Canton a common deficient variant in southe east Asia caused by a 459 Arg-Leu mutation. *Nucl Acids Res.* **18**:7190.

Takizawa T, Huang IY, Ikuta T, Yoshida A. (1986) Human glucose-6-phosphate dehydrogenase: primary structure and cDNA cloning. *Proc. Natl. Acad. Sci. USA* **83**:4157-4161.

Takizawa T, Yoneyama Y, Miwa S, Yoshida A. (1987) A single nucleotide base transition is the basis of the common human glucose-6-phosphate dehydrogenase variant A(+). *Genomics* **1**:228-231.

Tan KL. (1976) Phototherapy for neonatal jaundice in erythrocyte glucose-6-phosphate dehydrogenase-deficient infants. *Pediatrics* **59** (Neonatology supp. 6, pt 2):1023-1026.

Tang TK, Huang W-Y, Tang C-JC, Hsu M, Cheng T-A, Chen K-H. (1995) Molecular basis of glucose-6-phosphate dehydrogenase (G6PD) deficiency in three Taiwan aboriginal tribes. *Hum. Genet.* **95**:630-632.

Town M, Bautista JM, Mason PJ, Luzzatto L. (1992) Both mutations in G6PD A- are necessary to produce the G6PD deficient phenotype. *Hum. Mol. Genet.* **1** (3):171-174.

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

Vulliamy TJ, D'Urso M, Battistuzi G, Estrada M, Foulkes NS, Martini G, Calabro V, Poggi V, Giordano R, Town M, Luzzatto L, Persico MG. (1988) Diverse point mutations in the human glucose-6-phosphate dehydrogenase gene cause enzyme deficiency and mild or severe hemolytic anemia. *Proc. Natl. Acad. Sci. USA* **85**:5171-5175.

Vulliamy TJ, Othman A, Town M, Nathwani A, Falusi Y, Luzzatto L. (1991a) Linkage disequilibrium of polymorphic sites in the G6PD gene in African populations and the origin of G6PD A-. *Gene Geogr.* **5**:13-22.

Vulliamy TJ, Othman A, Town M, Nathwani A, Falusi AG, Mason PJ, Luzzatto L. (1991b) Polymorphic sites in the African population detected by sequence analysis of the glucose-6-phosphate dehydrogenase gene outline the evolution of the variants A and A-.Proc. *Natl. Acad. Sci. USA* **88**:8568-8571.

Vulliamy T, Mason P, Luzzato L. (1992) The molecular basis of glucose-6-phosphate dehydrogenase deficiency. *Trends Genet.* **8** (4):138-143.

Vulliamy T, Beutler E, Luzzatto L. (1993) Variants of glucose-6-phosphate dehydrogenase are due to missense mutations spread throughout the coding region of the gene. *Hum. Mutat.* **2**:159-167.

WHO working group . (1989) Glucose-6-phosphate dehydrogenase deficiency. *Bull. WORLD Health Organ.* **67** (6):601-611.

Yoshida A, Beutler E. (1983) G-6-PD variants: another up-date. *Ann. Hum. Genet.* **47**:25-38.

Yoshida A, Takizawa T, Prchal JT. (1988) RFLP of the X chromosome-linked glucose-6-phosphate dehydrogenase locus in Blacks. *Am. J. Hum. Genet.* **42**: 872-876.

APPENDIX A

SOLUTIONS AND REAGENTS

APPENDIX A - SOLUTIONS AND REAGENTS

10mM dNTPs

Dissolve each dNTP in dH_2O to an approximate concentration of 10mM.
Adjust pH to 7.0 with 1M tris-base.
Adjust concentration precisely, using the following equation:

$$\text{Molarity} = \text{Absorbance} \times \frac{\text{dilution factor}}{\text{Extinction coefficient}}$$

BASE	WAVELENGTH	EXTINCTION COEFFICIENT
A	259	1.54×10^4
G	253	1.37×10^4
C	271	9.10×10^3
T	260	7.40×10^3

Aliquot into 1.5ml Eppendorf tubes.
Store at -20°C.

DEOXYNUCLEOTIDES

- 125µl 10mM dCTP
- 125µl 10mM dTTP
- 125µl 10mM dATP
- 125µl 10mM dGTP

Make up to 1ml with ddH_2O (i.e. 500µl ddH_2O).

0.5M ETHYLENEDIAMINE TETRA-ACETIC ACID (EDTA) STOCK

- 18.61g $Na_2EDTA.2H_2O$
- 80 ml dH_2O

Adjust pH to 7.5/8.0 with NaOH pellets (±2g).
Autoclave.

10mg/ml ETHIDIUM BROMIDE STOCK

- 0.1g EtBr
- 10ml sterilised dH_2O

Store at 4°C.

FICOLL LOADING DYE
- 50g sucrose (50%)
- 10ml EDTA pH 7 (50mM)
- 0.1g Bromophenol blue (0.1%)
- 10g Ficoll (10%)

Make up to 100ml with dH_2O.

1kb λ LADDER
- 10.9µl λ ladder
- 5µl Ficoll dye
- 84µl 1xTE

Load 5-10µl.

PROTEINASE K SOLUTION
- 0.2ml 10% SDS
- 8µl 0.5M EDTA
- 0.4ml 10mg/ml proteinase K stock
- 1.4ml ddH_2O

SODIUM CHLORIDE (SATURATED)
Autoclave 100ml dH_2O.
Add 40g NaCl slowly, while stirring at room temperature, until saturated (some NaCl remains undissolved.
Agitate and allow to settle before use.

10% SODIUM DODECYL SULPHATE (SDS)
Autoclave 1l of dH_2O.
- 100g SDS
- 900ml sterile dH_2O

Heat to dissolve.
Adjust pH to 7.2 with conc HCl.
Make up to 1l with sterile dH_2O.

0.1M SPERMIDINE

MWT = 254.6
∴25,46g in 1 000ml = 0.1M
For 20ml:

$$\frac{25.46 \times 20}{1000} = 0.5g$$

- 0.5g spermidine
- 20ml ddH_2O

Aliquot into 1ml eppendorfs.
Store at -20°C.

SUCROSE-TRITON X
- 10ml 1M Tris-HCl
- 5ml 1M $MgCl_2$
- 10ml Triton-X 100

Make up to 1l with dH_2O.
Autoclave if the solution is not being used immediately.
Add 109.5g sucrose immediately before use.

10x TBE
- 108g Tris-base
- 55g Boric acid
- 7.44g EDTA

Make up to 1l.
Adjust pH to 8.0.

1x TE
- 1ml 1M EDTA
- 10ml Tris-HCl (pH 8.0)
- 50ml dH_2O

Adjust pH to 8.0.
Make up to 1l with dH_2O.
Autoclave.

$T_{20}E_5$
- 0.6ml 1M tris-HCl (pH 8.0)
- 0.3ml 0.5M EDTA

Make up to 30ml with dH_2O.

TRIS-HCl
- 121.1g Tris-base
- 800ml dH_2O

Adjust pH as required (i.e. to 7.5, 7.9 or 8.0) with conc. HCl.
Make up to 1l with dH_2O.
Autoclave.

APPENDIX B

CALCULATIONS USED IN THE ANALYSIS

APPENDIX B - CALCULATIONS USED IN THE ANALYSIS

B I) LINKAGE DISEQUILIBRIUM CALCULATIONS

Linkage disequilibrium calculations are performed by calculating a D value based on the difference between the observed frequency and the expected frequency of allele associations. The following is the calculation of the D value between the least common allele at the *Fok* I RFLP and the least common allele at the *Pvu* II RFLP for the CAR(grp) population group.

The observed figure is calculated by the observed frequency with which the *Fok* I(+) allele associates with the *Pvu* II(+) allele. The only haplotypes in which the *Fok* I(+) occurs with the *Pvu* II (+) are haplotypes V and VI. The observed frequency of this association is the sum of the frequencies of haplotypes V (0.065) and VI (0.052) in the CAR(grp) population = 0.117. The expected figure is obtained by multiplying the allele frequency of the *Fok* I(+) allele (0.256) by the allele frequency of the *Pvu* II (+) allele (0.103) to give 0.026. The D value (a measure of linkage disequilibrium) is obtained by finding the absolute value of the difference between the observed and expected values i.e. 0.117-0.026 = 0.091 (as shown in Table 3-5).

Whether a D value is significant or not can be determined by chi square analysis. To perform the χ^2 test, numbers as opposed to frequencies are used. To convert the observed and expected frequencies to numbers, they were multiplied by the total number of haplotypes (i.e. chromosomes) in the sample i.e. 77:

0.117x77=9 (observed) (or the sum of number of chromosomes with haplotype V and haplotype VI)

0.026x77=2.025 (expected)

This gives one class for χ^2 analysis. The second class is the remainder of the chromosomes. i.e. 77-9=68 (observed) and 77-2.025=74.975 (expected).

The χ^2 value is given by:

$$\chi^2 = \sum \frac{(observed - expected)^2}{expected}$$
$$= \frac{(9 - 2.025)^2}{2.025} + \frac{(68 - 74.975)^2}{74.975}$$
$$= 24.674$$

Since there are 2 classes, there is 1 degree of freedom. A χ^2 value of 24.674 at 1 degree of freedom corresponds to a p value of 6.70×10^{-7} (Table 3-6), which is significant at the 99.9% confidence interval.

Where the observed numbers were less than 5, Yates' correction for continuity was applied.

B II) CHI SQUARE (χ^2) ANALYSIS CALCULATIONS

The χ^2 analysis to test for significant differences between the groups CAR (M) and CAR (S) is shown below.

	CAR(M)		CAR(S)		Total
Class	**Observed**	**Expected**	**Observed**	**Expected**	**Observed**
III+V+VI+Other+ Med I + Med VII	1	3	4	2	5
I	6	4	2	4	8
VII	7	8	9	8	16
II	2	2	1	1	3
IV	4	4	3	3	7
Total	**20**		**19**		**39**

The observed values are the number of haplotypes in each class. Because some of the haplotype classes were empty, they had to be grouped. The first class therefore consists of haplotypes III, V, VI, "Other", Med I and Med VII. The expected numbers are generated from the observed numbers to take sample sizes into account. The expected figure for the Mbindu group (M) for the first class (i.e. 3) is thus calculated as follows:

Total of observed figures for the Mbindu group divided by the total observed multiplied by total of the first observed class. i.e. 20/39x5=3. The other expected figures are calculated in the same way. e.g. the expected figure for the group S in the 4th class (haplotype II) = 19/39x3=1.

The significance is determined by calculating a χ^2 total using these observed and expected values (=4.50). For 5 classes, there are 4 degrees of freedom and this corresponds to a p value of 0.342 (not significant at the 95% confidence interval).

B III) ADMIXTURE CALCULATIONS - ROBERTS AND HIORNS

The calculation of admixture for the total Malagasy hybrid population for the parental populations Zambian, Tsonga, Zulu, Indonesian and Polynesian, using the Roberts and Hiorns method (Table 3-16) is shown below. The Roberts and Hiorns method calculates m′(=m1,..m2,..mp) (the proportional contributions of the p parental populations), based on the following equation:

$$\hat{m} = \left(X'X\right)^{-1} X'y,$$

where X is the matrix of allele frequencies for the populations, X′ is the transpose of matrix X, and y is a vector of allele frequencies for the hybrid population. The haplotype "Other" has been excluded.

The vector y, the haplotype frequencies in the hybrid ("Total Malagasy") population, is the following:

	y
"Allele"	Total Malagasy
I	0.47
VII	0.29
II	0.06
III	0.00
IV	0.05
V	0.01
VI	0.10
MED I	0.00
MED VII	0.01

The matrix X, the haplotype frequencies in the parental populations, is the following:

X

"Allele"	Zambian	Tsonga	Zulu	Indonesian	Polynesian
I	0.268	0.242	0.262	0.750	0.783
VII	0.341	0.339	0.310	0.250	0.217
II	0.024	0.226	0.167	0	0
III	0	0	0	0	0
IV	0.098	0.097	0.167	0	0
V	0.049	0	0.071	0	0
VI	0.207	0.048	0	0	0
MED I	0	0	0	0	0
MED VII	0	0	0	0	0

The matrix X′ is the transpose of matrix X:

X′

Allele	I	VII	II	III	IV	V	VI	MED I	MED VII
Zambian	0.268	0.341	0.024	0	0.098	0.049	0.207	0	0
Tsonga	0.242	0.339	0.226	0	0.097	0	0.048	0	0
Zulu	0.262	0.310	0.167	0	0.167	0.071	0	0	0
Indonesian	0.750	0.250	0	0	0	0	0	0	0
Polynesian	0.783	0.217	0	0	0	0	0	0	0

The matrix product of X (a 9x5 matrix) and X′ (a 5x9 matrix) is the 5x5 matrix:

X′X

0.244	0.206	0.200	0.283	0.284
0.206	0.236	0.222	0.266	0.263
0.200	0.222	0.225	0.272	0.272
0.283	0.266	0.272	0.610	0.633
0.284	0.263	0.272	0.633	0.660

The matrix product is the matrix where each element is the sum of the products of the elements of the row-vector in X and the corresponding column vector in X′. For example, the element at $(X'X)_{23}$ (0.222) is the vector product of the X′ row vector 2 (i.e. 0.242, 0.339, 0.226, 0, 0.097, 0, 0.048, 0, 0) and the X column vector 3 (i.e. 0.262, 0.310, 0.167, 0, 0.167, 0.071, 0, 0, 0).

The element at $(X'X)_{23}$ is therefore:

(0.242)(0.262) + (0.339)(0.310) + (0.226)(0.167) + (0)(0) + (0.097)(0.167) + (0)(0.071) + (0.048)(0) + (0)(0) + (0)(0) = 0.222.

The product of a non-singular matrix and its inverse matrix is the identity matrix (a matrix where the diagonals are singular and all other values have a value of 0).

The matrix $(X'X)^{-1}$ is the inverse of the matrix X'X.

$(X'X)^{-1}$

19.218	-3.182	-4.726	-62.294	54.746
-3.182	93.991	-64.329	-283.7	262.824
-4.726	-64.329	70.7847	63.944	-62.924
-62.294	-283.7	63.944	2712.31	-2490.5
54.746	262.824	-62.924	-2490.5	2290.13

The matrix product X'y is given by:

X'y

0.253
0.235
0.232
0.419
0.430

The matrix product of the X'y matrix and the inverse (X'X)-1 matrix is the vector:

Population	**$(X'X)^{-1}X'Y$**
Zambian	0.40685
Tsonga	0.26883
Zulu	-0.0936
Indonesian	0.11147
Polynesian	0.30023

These are the m values for the proportional contribution of each parental population (shown in Table 3-16). Many of these matrix manipulation functions can be automated using commercial packages such as spreadsheets. The matrix calculations in this project were performed using MS EXCEL v 5.0.

B IV) ADMIXTURE CALCULATIONS - ELSTON

The following calculations generate the m values in the hybrid population ("Total Malagasy") for the parental populations (Zambian, Tsonga and Indonesian), using the method of Elston. The results are shown in Table 3-17. For the first p-1 parental populations, the m value is calculated by:

$$m = \left(X^{*\prime} X^{*}\right)^{-1} X^{*\prime} y^{*},$$

and the m value for the p^{th} parental population is calculated by:

$$m_p = 1 - \sum_{j=1}^{p-1} m_j,$$

where X* is a modified form of the matrix X in the Roberts and Hiorns method above, X*′ is the transpose of X* and y* is a modified form of the y hybrid population frequency in the Roberts and Hiorns method above.

The matrix X below is a matrix of the p parental population frequencies for all of the alleles (no allele classes are discarded):

	X		
Haplotype	Zambian	Tsonga	Indonesian
I	0.268	0.242	0.750
VII	0.341	0.339	0.250
II	0.024	0.226	0
III	0	0	0
IV	0.098	0.097	0
V	0.049	0	0
VI	0.207	0.048	0
Other	0.012	0.048	0
MED I	0	0	0
MED VII	0	0	0

The vector y is the frequencies in the hybrid population "Total Malagasy":

	y
Haplotype	Total Malagasy
I	0.467
VII	0.294
II	0.055
III	0.003
IV	0.052
V	0.010
VI	0.097
Other	0.007
MED I	0.003
MED VII	0.010

The matrix X* is obtained as follows:

The j^{th} column of X* is obtained by Xj-Xp. So for example for X^*_{11}, Xj=0.268 and Xp=0.750. Thus X^*_{11} is 0.268-0.750 = -0.472.

The matrix X is a 10x3 matrix and the matrix X* is a 10x2 (i.e. p-1) matrix:

X*

-0.472	-0.499
0.095	0.092
0.012	0.213
0	0
0.098	0.097
0.049	0
0.207	0.048
0.012	0.048
0	0
0	0

In a similar way, the vector y* is obtained by y-Xp. The first element of this vector, y^*_{11}, is thus y_{11} (=0.467) - Xp (=0.741); i.e. y^*_{11} = 0.467-0.741 = -0.274.

The vector y* is therefore:

y*

-0.274
0.047
0.043
0.003
0.052
0.010
0.097
0.007
0.003
0.010

The transpose of matrix X* (i.e. X*′) is:

X*'

-0.472	0.095	0.012	0.000	0.098	0.049	0.207	0.012	0.000	0.000
-0.499	0.092	0.213	0.000	0.097	0.000	0.048	0.048	0.000	0.000

The product of this transpose and X* is:

X*'X*

0.287	0.267
0.267	0.317

and the inverse of this matrix is:

$(X^{*\prime}X^*)^{-1}$

16.034	-13.510
-13.510	14.539

The matrix product X*′y* is:

X*'y*

0.160
0.160

And the matrix product $(X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*}$, i.e. the m values for the p-1 parental populations, is:

	$(X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*}$
Zambian	0.4031
Tsonga	0.1655
Total	0.5686

The m value for the pth parental population (i.e. the Indonesians) is calculated by 1- the total of the first p-1 parental populations; i.e. 1-0.5686 = 0.4314. These results are shown in Table 3-17.