

Genetic variation in Khoisan-speaking populations from southern Africa

BY

Carina Maria Schlebusch

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

Johannesburg, 2010

DECLARATION

I declare that this thesis is my own unaided work. It is being submitted for the Degree of Doctor of Philosophy in Human Genetics at the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination at any other university. I declare that this work has been approved by the Ethics Committee of the University of the Witwatersrand for Research on Human Subjects, and the certificate numbers are M050902 and M980553.

Carina M. Schlebusch

Date

ABSTRACT

The San and Khoe people currently represent remnant groups of a much larger and widely distributed population of hunter gatherers and cattle herders, respectively, who had exclusive occupation of southern Africa before the arrival of Bantu-speaking groups in the past 1,200 years and sea-borne immigrants within the last 350 years. This project made use of mitochondrial DNA (mtDNA), Y-chromosome DNA and autosomal DNA markers to examine the population structure of various San and Khoe groups and to reconstruct their prehistory. The groups included in the study consists of six different Khoe-San groups (ǀKhomani, Nama, Khwe, !Xun, /Gui + //Gana + Kgalagari and Ju'hoansi), four different Coloured groups and five other population groups that were included in the comparative analysis.

For the mtDNA study a minisequencing technique was successfully developed which allowed the assignment of mtDNA lineages into the 10 global mtDNA macro-haplogroups. Haplogroups were further resolved using control region sequence data obtained from both hypervariable regions (HVR I and HVR II). Using this approach 538 individuals (both males and females) were screened and their mtDNA types were resolved into 18 haplogroups encompassing 245 unique haplotypes. In addition, 353 males were examined for Y-chromosome DNA variation using 46 bi-allelic Y-chromosome markers and 12 Y-STR markers. The Y-chromosomes in the sample were assigned into 29 haplogroups (using bi-allelic variation) following the nomenclature initially recommended by the Y-chromosome Consortium and resolved into 268 unique haplotypes (Y-STR variation). To assess the level of autosomal variation, 220 genome wide autosomal SNPs were typed in 352 individuals. These SNPs were combined in different datasets and analysed using two different approaches allowing for genotype and haplotype analyses. Data from these three marker systems were analysed using different analytical methods (distance based phylogenetic analysis, network analysis, dating of lineages, principal components analysis, phylogeographic analysis, AMOVA analysis, population structure analysis, and population genetic summary statistics) to assess the ancestral associations and the genetic affinities of the various San, Khoe and Coloured populations.

The most striking observation from this study was the high frequencies of the oldest mtDNA haplogroups (L0d and L0k) and Y-chromosome haplogroups (haplogroups A and B) found in Khoe-San and Coloured groups. The sub-haplogroups were, however, differentially distributed in the different Khoe-San and Coloured groups which suggested different demographic histories.

The current distribution of Khoe-San groups comprises a wide geographic region extending from southern Angola in the north to the Cape Province (South Africa) in the south. Linguistically Khoe-San groups are also divided into northern Khoisan-speaking groups (Ju division) and southern Khoisan-speaking groups (Tuu division) with an additional linguistic group (Khoe) associated with some Khoe-speaking San groups in Botswana and the Khoe herders of South Africa and Namibia (such as the Nama). For all three genetic marker systems, northern groups (Ju speaking - !Xun, Ju'hoansi and Khoe-speaking San - /Gui + //Gana) grouped into one cluster and southern groups (historically Tuu speaking - #Khomani and Coloured groups) grouped into a second cluster with the Khoe group (Nama) clustering with the southern Khoe-San and Coloured groups.

The Khwe genetic profile was very different from the other Khoe-San groups. Although high proportions of Bantu-speaking admixture were identified in the Khwe group, they also contained a unique distribution of other mtDNA and Y-chromosome lineages. A previously published theory suggested that, based on the presence of a specific E-M35 Y-chromosome haplotype, the Khwe might be descendants of an east African pastoralist group that introduced the pastoralist culture to a region located in the present day northern Botswana. This pattern also mirrors what archaeologists have found with respect to the introduction of pastoralism to southern Africa. The theory was further supported and elaborated on in the present thesis. Considering the frequency and distribution of E-M35, the highest frequency (46%) was found in the Khwe with a present-day distribution in northern Botswana and southern Angola while a decrease in frequency is observed towards the south with low frequencies (<10%) in the Karoo Coloured groups. Conversely, none of the mtDNA (female) L0k and L0d lineages observed in the Khwe group were observed in the southern Khoe-San and Coloured groups. From these observations a theory was proposed that after introduction into the region of northern Botswana, the

southwards spread of pastoralism was not a clear-cut demic or cultural diffusion. Rather some male individuals integrated with the southern tribes and took with them the pastoralist practice and likely also their Khoe-language.

Altogether this thesis presented new insights into the multifaceted demographic history that shaped the existing genetic landscape of the Khoe-San and Coloured populations of southern Africa.

To: My parents

ACKNOWLEDGEMENTS

I am grateful to all subjects who participated in this research project and would like to thank them for their contributed blood and saliva samples for DNA extraction and subsequent analyses. I would also like to thank Professor Trefor Jenkins and colleagues in the Division of Human Genetics for assistance with fieldwork and the processing of samples. In addition, I am thankful for the support and mediation provided by the South African San Council and the Working Group of Indigenous Minorities in Southern Africa (WIMSA).

I am particularly appreciative to Prof. Fourie Joubert (Bioinformatics and Computational Biology Unit, University of Pretoria (UP)) for accommodating and assisting me in running analyses requiring intensive computational time on their cluster computer system at UP.

During my studies I was supported by a National Research Foundation Prestigious Doctoral Scholarship. Travel grants from the University of the Witwatersrand and the National Research Foundation allowed me to present some of this work at an international conference. This research was supported by grants awarded to me by the NHLS Research Trust and to Professor Himla Soodyall by the NHLS, University of the Witwatersrand, the NRF and the MRC.

My sincere appreciation goes to Professor Himla Soodyall, my supervisor, for her assistance and guidance throughout this study, and for reading several drafts of this thesis. I also wish to express my gratitude to my colleagues in the HGDDRU (TJ Naidoo, Heeran Makkan, Akashnie Maharaj, Raj Mahabeer, Christoff Erasmus and Pareen Patel) for their help, friendship and motivation.

Finally, I would like to express my deepest gratitude to my parents, family and friends, for their constant support and encouragement. Most of all, my appreciation goes to my husband Ronnie who always have been an incredible source of help, love and encouragement throughout the years.

TABLE OF CONTENTS

DECLARATION	ii
ABSTRACT	iii
ACKNOWLEDGEMENTS.....	vii
TABLE OF CONTENTS	viii
LIST OF FIGURES	xiii
LIST OF TABLES	xx
LIST OF ABBREVIATIONS	xxiii
Note on terminology adopted in thesis	xxvi
1. INTRODUCTION	1
1.1 Khoe-San today	5
1.1.1 Group classification	8
1.1.1.1 The Ju	12
1.1.1.1.1 The !Xun	12
1.1.1.1.2 The Ju\’hoansi	13
1.1.1.1.3 The ‡X’ao//’ãesi	13
1.1.1.2 Khoe-speaking San groups	14
1.1.1.2.1 The Tshua and Shua of eastern Botswana.....	14
1.1.1.2.2 The Khwe of northern Botswana and southern Angola.....	15
1.1.1.2.3 The /Gui and //Gana of the central Kalahari.....	15
1.1.1.2.4 The Naro	17
1.1.1.2.5 The Hai\om	18
1.1.1.3 The Kwadi	18
1.1.1.4 The Khoe.....	18
1.1.1.4.1 The Korana	19
1.1.1.4.2 The Cape Khoe.....	19
1.1.1.4.3 The Nama	19
1.1.1.5 The !Xóó and the ‡Hǃǃ (Tuu division).....	20
1.1.1.6 Remnants and descendants of Khoe and San groups living in South Africa	21
1.1.1.6.1 N//η people (“Mountain Bushmen”).....	22

1.1.1.6.2 The //Xegwi	22
1.1.1.5.3 /Xam descendants	23
The Karretjie people	24
1.1.1.6.4 The ‡Khomani	25
1.1.1.5.5 South African Khoe descendant groups	26
1.1.1.5.6 The !Xun and Khwe of Platfontein	26
1.2 Khoe-San history	29
1.2.1 Linguistics, Archaeology and Ethnography	29
1.2.1.1 Khoisan Linguistic Family	29
1.2.1.2 Khoe-San History according to Linguistics	30
1.2.1.3 Khoe-San History according to Archaeology and Ethnography	32
1.2.1.4 Khoe-San history according to Physical Anthropology	40
1.2.2 Khoe-San history according to molecular genetic studies	42
1.2.2.1 Serological studies	42
1.2.2.1.1 Differences between San and Khoe	42
1.2.2.1.2 Differences between Khoe-San subgroups	43
1.2.2.1.3 Commonalities between Hadza, Sandawe and Khoe-San	48
1.2.2.1.4 Khoe-San admixture into other population groups	49
1.2.2.2 Mitochondrial DNA studies	49
1.2.2.3 Y-chromosome studies	58
Y-chromosome tree structure	58
The age of the Y-chromosome tree	66
Y-chromosome studies in the Khoe-San	66
Y-chromosome and mtDNA comparative studies	68
1.2.2.4 Autosomal DNA studies	70
1.3 Aims	78
2. SUBJECTS AND METHODS	83
2.1 Subjects	83
2.2 Methods	87
2.2.1 DNA extraction	87
2.2.2 MtDNA methods	88
2.2.2.1 MtDNA minisequencing method	88

2.2.2.1.1 PCR-multiplex amplification	91
2.2.2.1.2 Minisequencing reaction	92
2.2.2.2 HVS amplification and sequencing	96
2.2.2.3 MtDNA data analysis	98
2.2.3 Y-chromosome methods	102
2.2.3.1 Y-chromosome RFLP	105
2.2.3.2 Y-chromosome minisequencing	111
2.2.3.3 Y-chromosome STR	113
2.2.3.4 Y-chromosome data analysis	114
2.2.4 Autosomal SNP methods	116
2.2.4.1 Autosomal SNP data analysis (Genotypic)	118
2.2.4.2 Autosomal SNP data analysis (Haplotypic)	121
3. MITOCHONDRIAL-DNA STUDIES	124
3.1 Minisequencing	125
3.2 HVS-I and II variation	130
3.3 Haplogroup assignment and structure	133
3.3.1 Haplogroup L0d/k	136
3.3.2 Khoe-San associated haplogroups L0d and L0k – Further analysis	142
3.3.3 Discussion of analyses of Khoe-San associated haplogroups L0d and L0k	151
L0k	151
L0d	155
L0d3	157
L0d1 and L0d2	161
L0d1	161
L0d1a	162
L0d1b	164
L0d1c	166
L0d2	168
L0d2a	168
L0d2b	171
L0d2d	172
L0d2c	172

L0dx	173
3.3.4 Summary of haplogroup histories	174
3.3.5 Haplogroup contributions from neighboring population groups	175
3.4 Mitochondrial genetic relationships between different Khoe, San, Coloured and neighboring groups	176
3.4.1 Summary: Genetic Affinities between the Khoe-San and Coloured groups as inferred from mtDNA analysis.....	194
4. Y-CHROMOSOME STUDIES.....	197
4.1 Haplogroup allocation and geographic distribution	197
4.2 Haplogroup diversity	200
4.3 African haplogroup analyses and discussion	201
Haplogroup A – Internal structure	201
Haplogroup A - Discussion	206
Haplogroup B – Internal structure	208
Haplogroup B - Discussion	212
Haplogroup E – Internal structure	213
Haplogroup E-M75.....	213
Haplogroup E-M2.....	216
Haplogroup E-M35.....	219
Haplogroup E - Discussion	223
4.4 Eurasian haplogroups	228
Haplogroup R – Internal structure.....	228
Eurasian haplogroups - Discussion	228
4.5 Analyses of Y-chromosome genetic relationships between different Khoe, San, Coloured and neighbouring groups	231
4.5.1 Discussion on the genetic affinities between Khoe-San and Coloured populations from southern Africa	242
5. AUTOSOMAL DNA STUDIES	245
5.1 Results and discussion (Genotypes).....	245
5.1.1 Heterozygosity.....	245
5.1.2 STRUCTURE analyses	248
5.1.3 Variation across STRUCTURE datasets	256

5.1.4 Distance based analysis of unlinked SNP sets.....	259
5.1.5 AMOVA analysis.....	269
5.2 Results and discussion (Haplotypes)	272
5.2.1 Inferred haplotypes.....	272
5.2.2 Distance analysis.....	274
5.3 Summary of autosomal results.....	285
6. GENERAL DISCUSSION	286
7. CONCLUSION.....	293
8. REFERENCES	296
9. APPENDICES	309
Appendix A: Ethics approval	310
Appendix B: Recipes for reagents and solutions used	313
Appendix C: Physical distance matrix (in km) between Khoe-San and Coloured groups	316
Appendix D: Details of SNP used in autosomal analyses	317
Appendix E: Haplotype list of HVR I and HVR II variation.....	322
Appendix F: Graphs – Physical vs. Genetic distance (L0d/k sequences and L0d sequences)	338
Appendix G: Haplotype list of 12 marker Y-STR panel	339
Appendix H: Bar charts showing haplotype frequencies for 44 inferred short haplotypes	347

LIST OF FIGURES

	Page
Figure 1.1	Map indicating the current distribution of Khoe-San groups 7
Figure 1.2	Map representing the historical geographic spread of the Khoe-San according to their language groups 10
Figure 1.3 A	Cluster analysis of distance matrix data from Jenkins (1986) 46
Figure 1.3 B	Principal Component Analysis of distance matrix data from Jenkins (1986) 47
Figure 1.4	Tree showing global mtDNA macro-haplogroups according to the nomenclature of Behar <i>et al.</i> , (2008) 51
Figure 1.5	Haplogroups within the L0 macro-haplogroup according to the nomenclature of Behar <i>et al.</i> , (2008) 53
Figure 1.6	Sub-haplogroups within the L0d haplogroup according to the nomenclature of Behar <i>et al.</i> , (2008) 54
Figure 1.7	Tree showing global Y-chromosome macro-haplogroups according to the nomenclature of Karafet <i>et al.</i> , (2008) 59
Figure 1.8	Sub-haplogroups within haplogroup A according to the nomenclature of Karafet <i>et al.</i> , (2008) 60
Figure 1.9	Sub-haplogroups within haplogroup B according to the nomenclature of Karafet <i>et al.</i> , (2008) 61
Figure 1.10	Sub-haplogroups within haplogroup E according to the nomenclature of Karafet <i>et al.</i> , (2008) 65
Figure 1.11	Distribution of Pygmies according to Cavalli-Sforza (1986) 77
Figure 2.1	Map indicating the place of origin for the Coloured and Khoe-San individuals who participated in the study 85
Figure 2.2	Tree showing the 10 mtDNA macro-haplogroups that are distinguished by typing 14 SNPs 90

		Page
Figure 2.3	The Y-chromosome haplogroup tree with nomenclature according to Karafet <i>et al.</i> , (2008) indicating the branch-defining mutations screened for by using SNaPshot minisequencing panels and RFLP assays in the HGDDRU laboratory	104
Figure 2.4	SNP selection strategy illustrated on a chromosome	116
Figure 2.5	Diagram illustrating how STRUCTURE results for 100 SNP sets were condensed into one consensus run	119
Figure 3.1	A 2% agarose gel showing the six amplified fragments that result from the multiplex PCR	126
Figure 3.2	Electropherogram examples showing peak profiles of haplogroups L0, L1, L3 and M	127
Figure 3.3	Mitochondrial haplogroup tree with nomenclature according to Behar <i>et al.</i> , (2008), listing haplogroup frequencies in the different populations in the study group	131
Figure 3.4	Graphical illustration of percentage mitochondrial haplogroup assignment in the populations used in comparative population analysis	132
Figure 3.5a	Maximum likelihood tree representing the substructure of L1 to L5	134
Figure 3.5b	Maximum likelihood tree showing the relationships of the different mtDNA haplotypes within haplogroup L0	135
Figure 3.6	Median joining network representing L0 substructure in the different populations of the study group	137
Figure 3.7	L0d structure as published in Behar <i>et al.</i> , (2008) with suggested changes	138
Figure 3.8	Graphical illustration of percentage L0d/k sub-haplogroup assignment in the populations used in comparative population analysis	139

		Page
Figure 3.9	Graphic representation of coalescent times and times of divergence of the mtDNA sub-haplogroups of L0d and L0k	141
Figure 3.10	Bar-graph indicating the clinal distribution of the L0d/k subgroups	142
Figure 3.11	Contour plots indicating the frequency distributions of L0d/k subgroups	143
Figure 3.12	Contour plots of L0d1c split into two subgroups, L0d1c1 and the remaining L0d1c sequences (L0d1c-)	144
Figure 3.13	Mismatch distributions of L0d/k sub-haplogroups and comparative groups	146
Figure 3.14	Bayesian Skyline plots of haplogroups showing changes in N_e through time	150
Figure 3.15	L0d3 branch after adding comparative published sequences	159
Figure 3.16	Principal component analysis of Fst values between different populations in the study group	178
Figure 3.17	Cluster analysis tree representing mitochondrial Fst values between different populations in the study group	179
Figure 3.18	Pairwise comparisons between physical geographic distance (X-axis) and mitochondrial Fst genetic distance (Y-axis)	181
Figure 3.19	Principal component analysis of L0d/k Fst values between different populations in the study group	183
Figure 3.20	Cluster analysis tree representing L0d/k Fst values between different populations in the study group	184
Figure 3.21	Principal component analysis of L0d Fst values between different populations in the study group	186
Figure 3.22	Cluster analysis tree representing L0d Fst values between different populations in the study group	187
Figure 3.23	Mismatch distributions of populations in the study group	191

	Page
Figure 4.1	Y-chromosome haplogroup tree with nomenclature according to Karafet <i>et al.</i> , (2008), listing haplogroup frequencies in the different populations in the study group 198
Figure 4.2	Graphical illustration of percentage Y-chromosome haplogroup assignment in the populations used in comparative population analysis 199
Figure 4.3	Contour plots indicating the frequency distributions of Y-chromosome haplogroups in the Khoe-San and Coloured populations 199
Figure 4.4	Neighbour Joining tree representing the substructure of Haplogroup A 203
Figure 4.5	Median joining network representing Haplogroup A substructure in the different populations of the study group 204
Figure 4.6	MDS plot visualizing the $\delta\mu^2$ distance matrix for haplogroup A 205
Figure 4.7	Neighbour Joining tree representing the substructure of Haplogroup B 209
Figure 4.8	Median joining network representing Haplogroup B substructure in the different populations of the study group 210
Figure 4.9	MDS plot visualizing the $\delta\mu^2$ distance matrix for haplogroup B 211
Figure 4.10	Neighbour Joining tree representing the substructure of Haplogroup E-M75 214
Figure 4.11	Median joining network representing Haplogroup E-M75 substructure in the different populations of the study group 215
Figure 4.12	Neighbour Joining tree representing the substructure of Haplogroup E-M2 217
Figure 4.13	Median joining network representing Haplogroup E-M2 substructure in the different populations of the study group 218
Figure 4.14	Neighbour Joining tree representing the substructure of Haplogroup E-M35 220

		Page
Figure 4.15	Median joining network representing Haplogroup E-M35 substructure in the different populations of the study group	221
Figure 4.16	MDS plot visualizing the $\delta\mu^2$ distance matrix for haplogroup E-M35	222
Figure 4.17	Neighbour Joining tree representing the substructure of Haplogroup R	229
Figure 4.18	Median joining network representing Haplogroup R substructure in the different populations of the study group	230
Figure 4.19	Principal Component Analysis of Y-chromosome Fst values between different populations in the study group	233
Figure 4.20	Principal Component Analysis of Y-chromosome Rst values between different populations in the study group	234
Figure 4.21	Cluster analysis tree representing Y-chromosome Fst values between different populations in the study group	235
Figure 4.22	Cluster analysis tree representing Y-chromosome Rst values between different populations in the study group	237
Figure 4.23	Pairwise comparisons between physical geographic distance (X-axis) and Y-chromosome Fst and Rst genetic distance (Y-axis)	238
Figure 4.24	Graphical illustration of percentage Y-chromosome haplotype for Khoe-San associated haplogroups in the Khoe-San and Coloured groups	238
Figure 4.25	Principal component analysis of Y-chromosome Rst values (excluding Eurasian and BS associated haplogroups) between Khoe-San and Coloured groups	239
Figure 4.26	Cluster analysis tree representing Y-chromosome Rst values (excluding Eurasian and BS associated haplogroups) between Khoe-San and Coloured groups	240
Figure 5.1	Scatter plot of heterozygosities in the 14 populations and the total sample set for each of the 100 sample sets	247

		Page
Figure 5.2	Correlation between heterozygosity and the variation observed between the 100 datasets	247
Figure 5.3	Averaged results of the Structure runs of the 100 different SNP sets	253
Figure 5.4	Triangle plot of individual cluster assignment at K=3 with the Khoe-San, non-African and BS associated clusters on the three different corners of the triangle	255
Figure 5.5	Graphical representation of the variation between the population cluster assignments across the 100 runs	258
Figure 5.6a	The Majority Rule consensus tree constructed from a 100 NJ trees	261
Figure 5.6b	The consensus tree constructed from the average of 100 distance matrices	262
Figure 5.7	Principal component analysis of autosomal genotypic distances between different populations in the study group	263
Figure 5.8	Principal component analysis of the average individual distance matrix	266
Figure 5.9	Pairwise comparisons between physical geographic distance (X-axis) and autosomal genotypic distance (Y-axis)	269
Figure 5.10	Bar charts of inferred haplotypes and their frequencies in each of the 14 populations	273
Figure 5.11	Principal Component Analysis of autosomal haplotype distance values between different populations in the study group	275
Figure 5.12	Principal Component Analysis of autosomal haplotype distance values between different individuals in the study group	276
Figure 5.13	Principal Component Analysis of autosomal representative haplotype distance values between different populations in the study group	280

		Page
Figure 5.14	Cluster analysis tree illustrating autosomal representative haplotype distance values between different populations in the study group	281
Figure 5.15	Splits decomposition network showing the different trees that explain the relationships between the representative composite haplotypes of the different populations	282
Figure 5.16	Pairwise comparisons between physical geographic distance (X-axis) and autosomal haplotype genetic distance (Y-axis)	284

LIST OF TABLES

	Page
Table 1.1	Internal classification of southern African Khoisan linguistic group 9
Table 1.2	MtDNA haplogroup frequencies in San populations studied to date 53
Table 1.3	Published mtDNA sub-haplogroup frequencies in San populations as fractions of the total number of L0d/k haplotypes in the sample group 55
Table 1.4	Y-chromosome haplogroup frequencies (%) of Khoe-San populations studied to date 68
Table 2.1	Number of individuals in which mtDNA, Y-Chromosome and autosomal variation were examined, their group and group-code, and place of sampling and origin 84
Table 2.2	Primer sequences, binding sites, amplicon sizes and concentrations for multiplex PCR amplification of 6 fragments 92
Table 2.3	Minisequencing primers used to distinguish haplogroups L0-L6, M, N and R 94
Table 2.4	Chromatogram band profile for identifying haplogroups L0-L6, M, N and R 95
Table 2.5	Sequences of primers used to amplify and sequence HVS-I and II 96
Table 2.6	PCR ingredients and cycling conditions for amplification and sequencing of HVS-I and II. Final concentrations of ingredients are shown 97
Table 2.7	SNPs typed in RFLP assays to determine Y-chromosome haplogroup 106
Table 2.8	Conditions and concentrations used during Y-chromosome RFLP typing 107

		Page
Table 2.9	Information on the seven Y-chromosome minisequencing panels used to resolve haplogroups according to Figure 2.3	112
Table 2.10	Y-STR PCR Thermal Cyclers Conditions	113
Table 3.1	Results of the minisequencing screening and classification of 699 sequences compared to classification based on HVS sequences	129
Table 3.2	TMRCAs calculated for the L0d/k subgroups. Four different mutation rates are applied	140
Table 3.3	Mismatch distribution statistics (haplogroups)	145
Table 3.4	Diversity statistics and neutrality tests of L0d/k subgroups and comparative haplogroups	147
Table 3.5	Mitochondrial population pairwise F_{st} values	177
Table 3.6	Results from mitochondrial AMOVA analysis using different groupings on the first level	188
Table 3.7	Mismatch distribution statistics (Groups)	192
Table 3.8	Diversity statistics and neutrality tests for populations in the study group	193
Table 4.1	Pairwise genetic distances between the 15 study groups calculated from Y-chromosome data	232
Table 4.2	Results from Y-chromosome AMOVA analysis using different groupings on the first level	242
Table 5.1	Average proportion of polymorphic loci, heterozygosities and gene diversities in each population over the 100 different SNP datasets	246
Table 5.2	Averaged population cluster assignments of the STRUCTURE runs from the 100 different SNP sets	254
Table 5.3	Average likelihood and delta-K scores across the 100 runs	255
Table 5.4	Average population distance matrix of autosomal genotypic data	260

		Page
Table 5.5	Results from autosomal genotypic AMOVA analysis using different groupings on the first level	269
Table 5.6	Maximum composite likelihood population distances of individual haplotypes	274
Table 5.7	Maximum composite likelihood population distances of population representative haplotypes	279

LIST OF ABBREVIATIONS

AFE	AFR + EUR
AFR	Afrikaner
aLRT	approximate likelihood-ratio test
AMOVA	analysis of molecular variance
ASD	average square distance
Ave	average
BP	before present
bp	base pairs
BS	Bantu-speakers
BSA	bovine serum albumin
BSP	Bayesian skyline plot
CAC	Cape Coloured
CEPH	Centre d'Etude du Polymorphisme Humain (Center for the Study of Human Polymorphisms)
CI	confidence interval
CKGR	Central Kalahari Game Reserve
CNC	Northern Cape Coloured
COL	Karoo Coloured
CRS	Cambridge reference sequence
ddH ₂ O	deionised distilled water
ddNTP	dideoxyribonucleotide triphosphate
del	deletion
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
DRC	Manyanga
DUM	Duma San
EDTA	ethylene-diamine-tetra-acetic acid
ESA	Earlier Stone Age
EUR	European
F	forward
FNLA	Frente Nacional de Libertação de Angola (National Front for the Liberation of Angola)
FPO	First Peoples of the Kalahari
g	gram
Gd	Gene Diversity
GTR	general time-reversible
GUG	/Gui, //Gana and Kgalagari
Hd	Haplogroup Diversity
HER	Herero
Het	Heterozygosity
HG	haplogroup
HGDDRU	Human Genomic Disease and Diversity Research Unit
HGDP	Human Genome Diversity Project
HPLC	high performance liquid chromatography

Ht	haplotype
HVS-I	hypervariable segments I
HVS-II	hypervariable segments II
IND	Indian
ins	insertion
JOH	Ju\’hoansi
KAR	Karretjie people
kb	kilobase
KHO	±Khomani
km	kilometre
KSC	Khoe-San + Coloured
KWE	Khwe
LGM	Last Glacial Maximum
LSA	Later Stone Age
m	migraton rate
M	molar
MALDI-TOF	matrix-assisted laser desorption/ionisation-time of flight mass spectrometry
Mb	megabase
MCMC	Markov Chain Monte Carlo
MDS	multi dimentional scaling
mg	milligrams
MgCl ₂	magnesium chloride
min	minutes
ml	millilitre
mM	millimolar
MP	maximum parsimony
MRC	Medical Research Council
MRCA	most recent common ancestor
MSA	Middle Stone Age
mtDNA	mitochondrial DNA
n	number
NaCl	sodium chloride
NAM	Nama
NAR	Naro
N _e	effective population size
NEAN	Neanderthal
ng	nanogram
NGO	non-governmental organization
NHLS	National Health Laboratory Service
NJ	neighbour joining
NRF	National Research Foundation
°C	degrees Centigrade
OTH	Other
P	probability
PC	principal component
PCA	principal component analysis
PCR	polymerase chain reaction

qt	quartile
R	reverse
r	correlation co-efficient
RE	restriction enzyme
RFLP	restriction fragment length polymorphism
RNA	ribonucleic acid
s	seconds
SA	South Africa
SADF	South African Defense Force
SASC	South African San Council
SDS	sodium dodecyl sulfate
SEB	south-eastern Bantu-speakers
Seq	sequence
SNP	single nucleotide polymorphism
SOT	Sotho, Tswana
SSD	sum of squared differences
STD	standard deviation
STR	short tandem repeat
subHG	subhaplogroup
SWB	south-western Bantu-speakers
SWZ	Swazi
T	time
Taq	Thermus aquaticus
TBE	tris borate-EDTA
TE	Tris EDTA
TMRCA	time to most recent common ancestor
TOT	total
μ	mutation rate
U	units
μg	microgram
μl	microlitre
μM	micromolar
UV	ultraviolet
v	version
WIMSA	Working Group of Indigenous Minorities in Southern Africa
XEG	//Xegwi
XUN	!Xun
YAP	Y Alu polymorphism
YCC	Y-chromosome Consortium
ZUX	Zulu, Xhosa
π	Nucleotide Diversity
τ	Tau
AFE	AFR + EUR
AFR	Afrikaner

Note on terminology adopted in thesis

The term “Khoisan” was first used by Leonard Schultze in 1928 (Schultze, 1928) and was intended to be used as a biological label. It was further popularised by Isaac Schapera in the 1930s (Schapera, 1930). The term has a collective meaning for two groups of people, the Khoi (old Nama word) or Khoe (modern Nama word), who were traditionally the pastoralist groups, and the San, who were hunter-gatherers. This grouping was introduced by European scholars who used mode of subsistence to distinguish the two groups. More recently this division has been challenged by present-day San and Khoe communities and there still debate as to whether this grouping presents a true reflection of subdivision. The word “Khoi” or “Khoe” means “person” in Nama. Two surviving pastoralist groups, the Nama and Korana, use the word “Khoenkhoen”, meaning “people of the people”. The word “San” is the Khoe word for “foragers” or “bushmen” (Barnard, 1992).

In 2002, at a meeting attended by the Working Group of Indigenous Minorities in Southern Africa (WIMSA) and the South African San Council (SASC), the San people decided that they wanted to be referred to by their individual community names (!Xun, ‡Khomani, etc.) or collectively as San. When collectively referring to the San and the Khoe, the term Khoe-San was suggested (Crawhall, 2006).

In this thesis individual groups will be referred to by their preferred community names. The application and spelling of the community names are in accordance with the usage in the book “Voices of the San” (le Roux and White, 2004); a book compiled by young representatives from San communities. The collective word “San” and “Khoe” will be adopted for the traditional hunter-gatherers and pastoralist groups, respectively, while “Khoe-San” will be adopted for the Khoe and the San populations. When referring to the linguistic grouping, the term Khoisan-speaking will be adopted. The use of the word Khoisan is in no way meant to be derogatory and is used only for the sake of continuity with current linguistic classification. When referring to the sub-grouping of the Khoisan linguistic group the nomenclature suggested by Güldemann (Güldemann, In Press) was followed.

1. INTRODUCTION

Finally, we believe that identifying genetic differences between races and ethnic groups, ..., is scientifically appropriate. What is not scientific is a value system attached to any such findings. Great abuse has occurred in the past with notions of 'genetic superiority' of one particular group over another. The notion of superiority is not scientific, only political, and can only be used for political purpose...

We need to value our diversity rather than fear it. Ignoring our differences, even with the best intentions, will ultimately lead to disservice of those who are in the minority.

*Neil Risch et al. (2002)
Genome Biology 3:1-12*

The study of genetic differences between individuals has a variety of implications and benefits that influence how we see our past and shape our future. We are entering a new era where the field of medicine, involving prevention and treatment, are becoming increasingly more customizable at an individual level. Methods of identifying individual differences in disease susceptibility and individual responses to drug treatment are developing rapidly. Various studies have shown that the human population is not homogenous in terms of disease risk and response to treatment (Jorde *et al.*, 2001; Risch *et al.*, 2002; Bamshad *et al.*, 2004). For the effective planning of prevention and treatment strategies the goal is to characterize risks both at individual as well as population levels.

While the medical field is still developing and our knowledge and technology is not yet capable of individual based risk assessments, we are forced to rely on risk assessments within population groups. A "race-neutral" approach in the biomedical field would not be advantageous to all groups of people. Instead, such an approach may in the end be disadvantageous to minority groups (Cavalli-Sforza *et al.*, 1994; Jorde *et al.*, 2001; Risch *et al.*, 2002; Bamshad *et al.*, 2004; Jobling *et al.*, 2004a). Various studies have shown that there are genetic substructure in the human population and that individuals within a certain group are genetically more similar to each other than to individuals within another group (Cavalli-Sforza *et al.*, 1994; Rosenberg *et al.*, 2002; Jakobsson *et al.*, 2008; Li *et al.*, 2008; Tishkoff *et al.*, 2009).

Sub-structure within the human population has largely resulted as a consequence of genetic drift and migration of sub-groups of humans, which led to isolation. The isolation

between sub-groups caused non-random mating which in turn resulted in genetic divergence. The field of human evolutionary genetics studies these genetic differences in order to unravel the history of humans. By employing different molecular genetic techniques, population subdivision, population expansion dynamics and human migration patterns are investigated.

There is only one true history of humankind and scholars have adopted several methods to reconstruct this past. In addition to molecular evolutionary genetic approaches, various other fields have been, and are still, actively studying human history and evolution. History in the form of recorded text goes back only as far as 4 000 years before present (BP). To study history older than this, other methods of investigation are required. Historical linguistics investigates the history of languages and their relationships to one another. Languages spoken by different groups of people retain evidence of their origin and are related to other languages in a measurable fashion. Language, however, also has a relative shallow time-history and linguists have suggested that languages do not retain evidence of their origin for more than 10 000 years (Jobling *et al.*, 2004a). Archaeology has a greater time depth and studies human history captured in physical remains, such as bones, stone tools, pottery, waste deposits and dwellings left over by past human groups. Palaeontology investigates the very deep ancestors of humans by investigating fossilized remains. The use of molecular genetics is a recent addition to the methods of studying human history (Cavalli-Sforza *et al.*, 1994; Jobling *et al.*, 2004a). Within the present thesis various molecular genetic markers and analyses techniques are utilized to aid in the inference of African history.

The first study that illustrated genetic differentiation between groups was a study on the ABO blood groups at the beginning of the 20th century (Landsteiner, 1901). The magnitude of this genetic variation only became apparent in the 1950s to 1960s when individual differences in proteins could be systematically studied (Cavalli-Sforza *et al.*, 1994). The study of protein variation was merely the beginning. When analysis methods for the hereditary material itself, DNA, became available, genetic variation could be studied directly and the field of evolutionary genetics expanded rapidly (Cavalli-Sforza *et al.*, 1994; Jobling *et al.*, 2004a).

Until recently, most studies that investigated the origin and dispersal of anatomically modern humans concentrated on two haploid compartments of the human genome, namely, the mitochondrial DNA and the Y-chromosome (Jobling and Tyler-Smith, 2000; Jobling and Tyler-Smith, 2003; Forster, 2004; Torroni *et al.*, 2006; Underhill and Kivisild, 2007). A few studies did investigate autosomal variation. These studies, however, were usually on particular genes that were under investigation due to their influence on a specific phenotypic property or disease risk. The variation therefore would have been subject to selection pressures. Recent advances in the human genome project have allowed us access to large amounts of information on neutral genetic variation that would give a more complete insight into human evolutionary history (Cavalli-Sforza, 1998; Przeworski *et al.*, 2000; Garrigan and Hammer, 2006). In this thesis neutral autosomal variation as well as the haploid mitochondrial genome and Y-chromosome were used in a three-pronged approach to study the evolutionary history of selected groups of southern African individuals.

All studies to date provide substantial support for an African origin of modern humans. The greatest genetic variation is present within African populations and variation outside of Africa is a subset of the African diversity (Jobling and Tyler-Smith, 2003; Garrigan and Hammer, 2006; Torroni *et al.*, 2006; Underhill and Kivisild, 2007). Africa has remarkable cultural, linguistic and genetic diversity and more than 2 000 distinct ethnic groups and languages exist on the continent (Gordon, 2005). Despite the pivotal role that Africa has played in the evolution of humankind and main residence of *Homo sapiens* for most of their existence, the history and population dynamics within the continent remain poorly understood. The present thesis try to contribute to the understanding of the history of the African continent by using molecular markers in selected groups of aboriginal human inhabitants of southern Africa.

The majority of sub-Saharan Africans (>200 million people) speak one of ~500 very closely related languages, even though they are distributed over an area of ~500 000 km². These languages are collectively referred to as “Bantu” languages, based on the word meaning

“people” (Bleek, 1862). The current distribution of these groups is largely a consequence of the movement of people (demic diffusion) rather than a diffusion of only language (Ehret and Posnansky, 1982; Huffman, 1982). This expansion is commonly referred to as the “Bantu Expansion” (Greenberg, 1963) and is thought to be due to the development and spread of agriculture and, possibly, the use of iron (Greenberg, 1972; Phillipson, 1993; Newman, 1995). The “Bantu Expansion” began ~3 000 - 5 000 years BP (Ehret, 1982; Vansina, 1990) and originated in the Cross River Valley, in the region of current eastern Nigeria and western Cameroon (Johnston, 1913; Greenberg, 1972; Huffman, 1982; Vogel, 1994).

To a certain extent the expansions of Bantu-speaking groups masked the earlier history of non-Bantu-speaking African populations. Groups that existed all over the African continent before the Bantu-expansions were replaced and/or assimilated by the Bantu-speaking groups. Traces of these pre-Bantu groups might still be found in the genetic variation, language and cultural practices of various Bantu-speaking groups where they have been incorporated/ assimilated. Very few sub-Saharan African ethnic groups have retained a cultural, linguistic and genetic identity that distinguishes them from the Bantu-speaking groups. Examples of such groups of people are the Hadza and Sandawe from East Africa, the Khoe-San populations from southern Africa and the Pygmy populations from central Africa. These populations (excluding the Khoe) did not adopt an agricultural lifestyle but instead kept a hunter-gatherer lifestyle. Their cultural practices, lifestyle and language (for the Khoe-San, Hadza and Sandawe) distinguish them from Bantu-speakers.

This distinction is also visible in the comparative genetic analysis of these populations in relation to the Bantu-speakers. In both Y-chromosome and mitochondrial DNA studies, these populations tend to carry unique and older lineages than the lineages associated with the Bantu-speaking people. In fact the deepest clades known among modern humans for both the Y-chromosome and mitochondria are found commonly and at their highest frequencies in the Khoe-San people (Behar *et al.*, 2008; Karafet *et al.*, 2008). Additionally, in autosomal studies Khoe-San people group in a distinct cluster from that of Bantu-

speakers (Cavalli-Sforza *et al.*, 1994; Rosenberg *et al.*, 2002; Jakobsson *et al.*, 2008; Li *et al.*, 2008; Tishkoff *et al.*, 2009). Thus, these unique relict populations of hunter-gatherers who carry genetic variation belonging to the deepest clades known among modern humans are crucial links to the past. It is important to extensively study their genetic contribution to the human gene pool. This is becoming increasingly difficult as the Khoe-San groups are losing their cultural identities, lifestyles and languages and are integrating into surrounding groups. In the current thesis the genetic variation from various Khoe-San groups are examined and analysed using multiple methodologies. The analyses are used to make inferences about the relatedness of the different Khoe-San groups, their affinities to neighboring groups and their place in African history.

To fully understand and interpret the genetic relatedness between the different Khoe-San groups included in this study it is important to review their present geographical distribution and demographics. Furthermore one must consider their relationship to neighbouring Khoe-San groups and neighbours from other population groups. Another important factor to take into consideration is the classification system used to classify the various Khoe-San groups. The following sections will review and summarise these different aspects.

1.1 Khoe-San today

The Khoe-San people of southern Africa consist of a collection of small diverse groups of people who share common cultural, linguistic and genetic features. Some of the groups are pastoralists, while others are hunter-gatherers or fishermen. Most Khoe-San individuals today, however, work as herdsman or laborers for members of other ethnic groups (Barnard, 1992; Smith *et al.*, 2000; le Roux and White, 2004).

Almost all Khoe-San groups are affected by social ills such as economic dependency, alcoholism, malnutrition, and societal breakdown. Many of these problems are because policies regarding the Khoe-San were developed without their participation and the recognition of their cultural legitimacy. On a continent that was and still is being rapidly colonized for its resources, their egalitarian values have left the San groups especially

vulnerable. With their land and food resources been taken away by surrounding groups and governments, their freedom has been restricted and their cultures and traditions have deteriorated. Only now, have certain groups began to reclaim their culture and basic human rights. This happened after the international community brought attention to the struggle of these marginalized indigenous people. Today, various organisations represent, handle and concentrate on land rights and ownership, political recognition and representation, and cultural rights and development projects involving Khoe and San groups (Broyhill *et al.*, Current).

Different San and Khoe groups are distributed throughout southern Africa where they live among and to some extent are admixed with the various Bantu-speaking populations surrounding them (See Figure 1.1) (Barnard, 1992; Smith *et al.*, 2000; le Roux and White, 2004). Today, the greatest proportion and the largest diversity of Khoe-San people can be found in Botswana followed by Namibia. Small groups of San people are also found in the southern parts of Angola and to a lesser extent southern Zambia and eastern Zimbabwe. The San people of South Africa have to a large extent lost their identities and have integrated into other populations. The Khoe groups still extant today are mainly found in Namibia while descendants of various mixed Khoe and San groups found in South Africa are known as the Coloured population (Barnard, 1992; Smith *et al.*, 2000; le Roux and White, 2004).

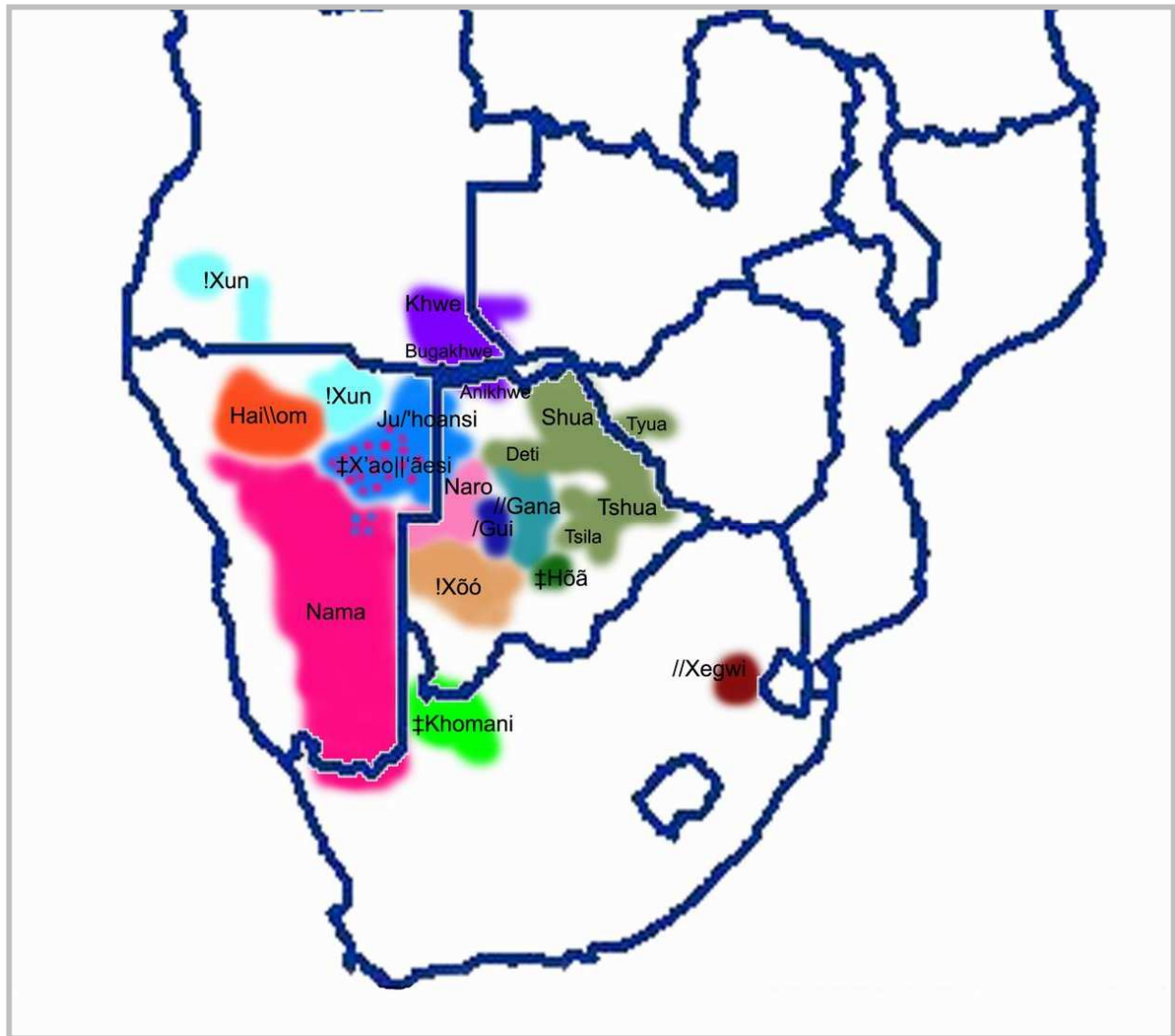


Figure 1.1 Map indicating the current distribution of Kho-San groups

1.1.1 Group classification

To classify Khoe-San groups into their individual ethnic groups is, in many ways, problematic. Different words and spellings have been used to refer to the same groups of people over the years. Linguistic classification is the method most commonly used to identify different groups.

As mentioned previously, historical inference based on language has a shallow time depth, maximum 10 000 years. These are very short times relative to historical inferences made from investigating genetic lineages such as mitochondrial genomes and Y-chromosomes. Relationships between these molecular markers can go back to over 100 000 years. When a hierarchical classification of possibly related groups such as the Khoe-San is made using a linguistic system, it will not necessarily reflect group classifications that can be made based on genetic information. It is one of the aims of this thesis to see if some of the group relationships inferred from linguistics can also be observed in the genetic analysis of Khoe-San groups. It is therefore necessary to first review the linguistic classification of the different Khoe-San groups, investigate how this classification is used to infer the relatedness between the different Khoe-San groups and finally how linguistics are used to infer the history of Khoe-San groups.

Table 1.1 shows the linguistic groupings (Güldemann, In Press) and Figure 1.2 the historical geographic spread of the Khoe-San groups based on languages and dialects.

The main Khoe-San language families include Ju (Northern Khoisan), Khoe-Kwadi (Central Khoisan) and Tuu (Southern Khoisan). The Khoe-Kwadi group includes Kwadi, the extinct language of Angola and the Khoe language branch. The Khoe language branch includes the people, known commonly today, as the Khoe (linguistic branch “KhoeKhoe”) as well as the San groups that speak languages more closely related to Khoe languages than to other San languages (linguistic branch “Kalahari”) (Güldemann, In Press) (Table 1.1 and Figure 1.2).

Table 1.1 Internal classification of southern African Khoisan linguistic group (Güldemann, In preparation)

Lineages and branches	Languages and dialects	Remarks
Ju-ǀHǎa		
ǀHǎa	Single language	Newly affiliated to Ju
Ju (= Northern Khoisan)		
Northwest	!O!Xǃu, !Xǃu	
Southeast	Ju/'hoan, ǀKx'au//e	
Khoe-Kwadi		Possibly related to Sandawe
Kwadi	Single language	Newly affiliated to Khoe
Khoe (= Central Khoisan)		
KhoeKhoe		
North	Eini, Nama-Damara, Hai//om	
South	!Ora, Cape varieties	
Kalahari		
East		
Shua	Cara, Deti, /Xaise, Danisi, Ts'ixa, etc.	
Tshwa	Kua, Cua, Tsua, etc.	
West		
Kxoe	Khwe, //Ani, Buga, G/anda, etc.	
G//ana	G//ana, G/ui, ǀHaba, etc.	
Naro	Naro, etc.	
Tuu (= Southern Khoisan)		
Taa-Lower Nossob		
Taa		
West	N/u//en, West !Xǃo	
East	'N/ohan, N/amani, East !Xǃo, Kakia	
Lower Nossob	/Auni, /Haasi	
!Ui	N//ng; ǀUngkue; /Xam; //Xegwi	

Bold – Independent lineage; Underlined – Earlier classification unit

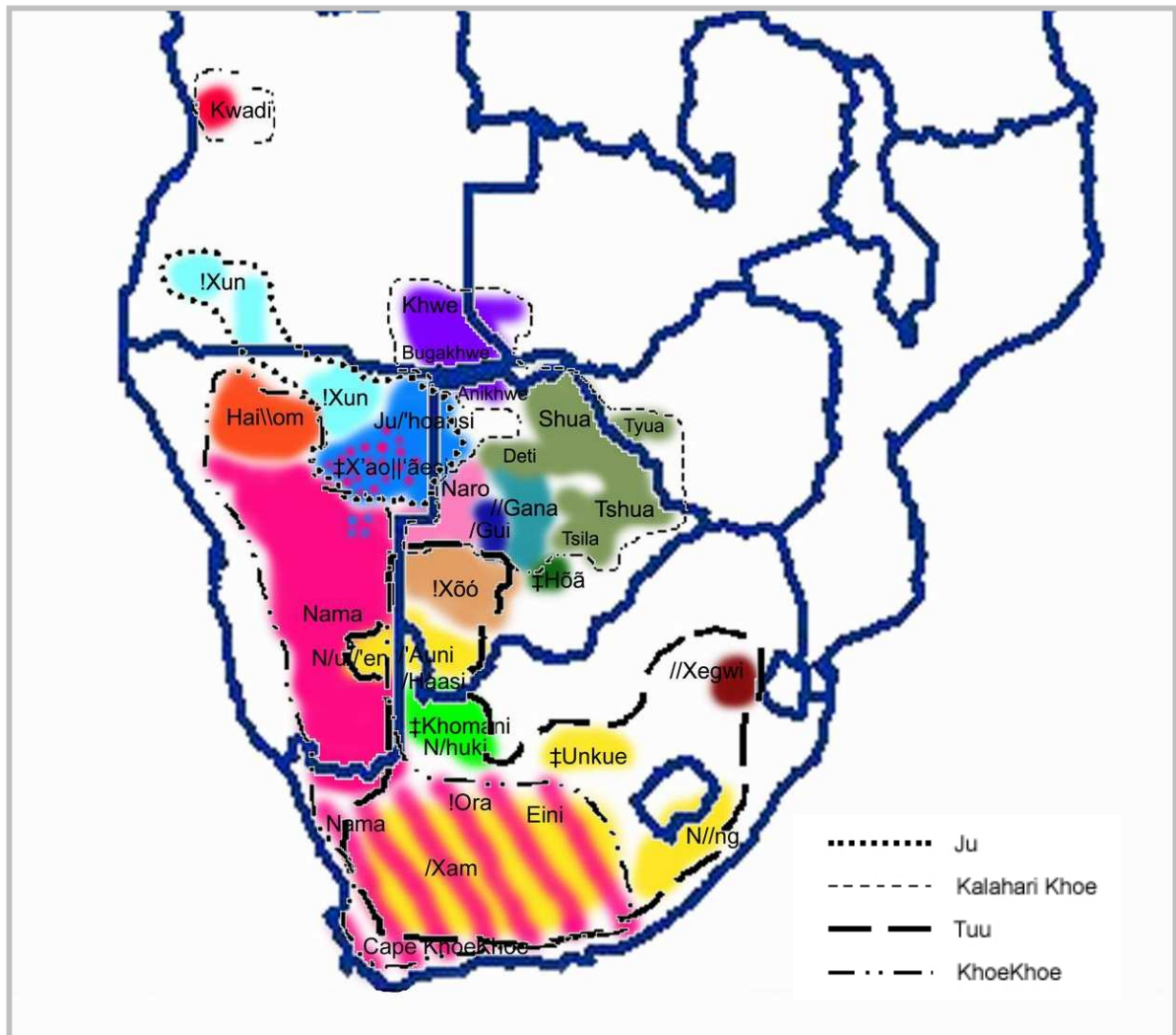


Figure 1.2 Map representing the historical geographic spread of the Khoe-San according to their language groups

Ju groups are linguistically split into a Northwest and Southeast division (Table 1.1). The Northwest groups include the !Xun groups of Angola and Northern Namibia while the southwest groups include the Ju'hoansi of northern Botswana and northeastern Namibia and the ǀX'ao//ǁǁesi (Auen) of western Botswana and northeastern Namibia. The ǀHǁǁǁ, is a south-eastern Botswana San group that may represent a linguistic intermediary of Ju and Tuu speakers (Güldemann, In Press) (Figure 1.2).

The only distinct Khoe group (speaking the KhoeKhoe language grouping) living today is the Nama of Namibia. The Korana (!Ora) and Cape Khoe (Cape KhoeKhoe) of South Africa represent extinct groupings of Khoe language and culture but their descendants live in the Coloured population of South Africa (Figure 1.2). The Hai\om of north Namibia also speak a KhoeKhoe language, however, this group is thought to have originated as result of contact between the Nama and the !Xun of northern Namibia (Barnard, 1992; Smith *et al.*, 2000; le Roux and White, 2004).

The Khoe-speaking San groups (speaking the Kalahari Khoe language grouping) are the most numerous and culturally diverse of the San language groups (Table 1.1). They inhabit the central and northern parts of Botswana, including the central Kalahari Desert and Okavango swamps, the southern parts of Angola and the Caprivi Strip of Namibia (Figure 1.2). Groups included into this language group are the Naro of western Botswana, the /Gui, //Gana and Deti of central Botswana, the “river Bushmen” of northern Botswana and southern Angola (the different Khwe groups), the Tshua and Shua of eastern Botswana and the Tyua of western Zimbabwe (Figure 1.2) (Barnard, 1992; Smith *et al.*, 2000; le Roux and White, 2004).

The Tuu language branch is divided into three groups, namely, the Taa, Lower Nossob and !Ui (Table 1.1). Most of the groups belonging to the Tuu language division have lost their language and cultural identity completely but their descendants are found in other population groups and in many cases classify themselves as “Coloured”. The “Lower Nossob” language group is extinct and the only remaining Taa group is the !Xǃǃ, who live in the south central Kalahari of Botswana. Remaining !Ui (!Wi) speakers consist of a few remaining San groups that are geographically scattered throughout South Africa. What is

presently known about the San peoples of South Africa derived from studies on very few remnant populations that survived into the 1700-1800s. These include the //Xegwi who lived in the Lake Chrissie area of the now Mpumalanga province of South Africa, the ‘mountain Bushman’ or N//η or ‘People of the Eland’ who lived in Lesotho, KwaZulu Natal and the Eastern Cape, the ‡Khomani who lived in the northern part of the Northern Cape province that borders with Botswana and Namibia (roughly where the Kalahari Gemsbok park is located now) and the /Xam who occupied the Karoo area of the Western, Northern and Eastern Cape provinces (Barnard, 1992; Smith *et al.*, 2000; le Roux and White, 2004); (Traill, 1973; Westphal, 1974); (Güldemann, In Press).

In the following sections the individual San and Khoe populations within the linguistic groupings will be described based on their identity, geographic spread, demography and what is known about their history.

1.1.1.1 The Ju

The three main ethno-linguistic groups of the Ju; the !Xun, the Ju\’hoansi (meaning “real people”) and ‡X’ao//’ãesi (also called Auen or linguistic-branch úKx’au//’e) correspond to indigenously defined dialects that also parallel three different cultural units and geographic areas (Figure 1.2 and Table 1.1). The word !Xun (or the different spelling !Kung) has been widely used to describe all three of these groups, however, the only group that uses the term as self-identification are the !Xun groups of Angola and northern Namibia (!xũ is a word indicating “person” in !Xun languages). The three groups together are estimated to comprise 25 000 to 30 000 individuals (Marshall and Ritchie, 1984; Gordon, 1986). The largest group is the central Ju\’hoansi while the northern !Xun is distributed over a larger geographic area (Gordon, 1984; Barnard, 1988).

1.1.1.1.1 The !Xun

The northern !Xun do not live in the Kalahari like the other two groups but rather in the forested areas of southern Angola and northern Namibia. Their self-designation is !o !xũ which means “forest people” (Bleek, 1928). Two groups found in Angola are known locally as Kwankala (Vakwankala) and Sekele (Vasekele) (De Almeida, 1965). In the local Bantu-speaking languages these names have derogatory connotations (meaning poor uncivilized

wanderers) and are not used anymore. The !Xun lived in close association with the local Ambo (Ovambo) population for centuries. It is through this association that the !Xun learned crop cultivation, herding and fishing with nets and spears. In the 1950s very few groups still followed a foraging lifestyle supplemented with assisting Bantu-speakers in the winter harvests in exchange for grain. In 1970-1980 Angola was a battleground between the government and guerrillas. Since then no ethnographic studies have been conducted to assess the extent of damage the war has had on the !Xun way of life (De Almeida, 1965; Barnard, 1992).

1.1.1.1.2 The Ju!hoansi

The central Ju!hoansi groups occupies areas with a large supply of water and plant resources. The area has over a hundred edible plants, the most important among these the Mongogo nut, a nutritious nut that can be gathered virtually the whole year. Bands of people usually camp out near permanent waterholes and Mongogo groves. In the past they only camped out during the dry winter and moved away during the wet season to exploit other territories. In Botswana, however, over the past century, groups have increased their time camping out. Today most groups have settled at the waterholes, and depend on Herero and Tswana residents for their livelihood. Development projects, including schools and handicraft tourist shops, were implemented by the Botswana government and anthropologists. In Namibia a “homeland” reserve for the Ju!hoansi (Bushmanland) was established and a school and administrative camp were built at Tsumkwe. In 1978 the South African Defense Force (SADF) built a military base at Tsumkwe and recruited Ju!hoansi soldiers. Many families lived off the earnings from the military base. Traditional subsistence techniques started vanishing because of this and the fact that the reserve was too small to support the number of people. Anthropologists were partially successful in encouraging them to adopt cattle husbandry in the reserve but met with opposition from wildlife officials (Marshall, 1960; Lee, 1979; Guenther, 1986; Barnard, 1992).

1.1.1.1.3 The ǀX'ao//ǃesi

The ǀX'ao//ǃesi (Auen) occupy a region in Botswana that overlaps with another San group, the Khoe-speaking Naro San group. This land is also shared with Bantu-speaking Tswana and to a lesser extent Hereros. White ranchers, mainly Afrikaners, own most of the

land. The white settlers arrived in 1897 from the Cape colony and occupy what is known as the Ghanzi farm block of western Botswana. Linguistic and anthropological evidence suggest ancient contact between the Khoe-speaking Naro and the ǀX'ao//'ãesi. The direction of borrowing seems to be from the ǀX'ao//'ãesi to the Naro. Today, overpopulation of the area by humans and livestock prevents traditional hunting and gathering practices. Many of the families have settled in towns and on ranches where they are laborers or they act as tourist attractions in exchange for the permission to use the land for gathering practices. They also earn small salaries and tips from tourists (Marshall, 1960; Barnard, 1992).

1.1.1.2 Khoe-speaking San groups

The Khoe speaking San groups speak languages (linguistic grouping Kalahari Khoe) that are more closely related to KhoeKhoe languages than to other San languages. This relationship, however, is distant. The Khoe speaking San groups are distributed over most of Botswana and their regions overlap with some of the other San groups.

1.1.1.2.1 The Tshua and Shua of eastern Botswana

The Tshua (south) and the Shua (north) consists of a number of scattered groups distributed over a large area, from the Kweneng district in the southeast of Botswana to the Ngamiland district in the northeast. They have lived in close association with Bantu-speaking groups including the Tswana, Kgalagari and Kalanga (closely related to the Shona of Zimbabwe) for over a century. The groups have various names for self identification (Tshua, Hietshware, Kua, Shua, Ts'ixa, Danisi, Deti) but all speak the eastern Khoe dialect (Table 1.1) where the word *Tshua* or *Shua* is used to refer to a “person” rather than the word *Khoe*. These eastern Khoe-speaking San are herders and cultivators as well as hunters. They also engage in extensive trade activities and “contract work” with neighbouring Bantu-speakers. This contract work entails an agreement, (known locally as a “mafisa” relationship) between a local Bantu-speaking tribe and a San group. In this mafisa relationship the San group will look after the cattle of the Bantu-speaking group and in return have the right to the milk, meat in case of an accidental death, the right to use them in ploughing and in some cases the right to keep the calves. Due to these relationships

many of the San groups of eastern Botswana settled at cattle posts (Dornan, 1975; Barnard, 1992).

1.1.1.2.2 The Khwe of northern Botswana and southern Angola

The Khoe-speaking San of northern Botswana, southern Angola and western Zimbabwe comprise the various Khwe (linguistic grouping - Kxoe) groups (including the Bugakhwe and //Anikhwe) (Table 1.1). They live in the Okavango swamp area and surrounding regions. This area is infested by tsetse flies; as a result livestock rearing is not viable. They sustain themselves through fishing as well as hunting and gathering. Linguistically, they are closer to the central Khoe speaking San than the eastern groups. Phenotypically, however, they resemble Bantu-speakers and genetic evidence also suggests a genetic makeup similar to the Bantu-speaking populations that surround them (Nurse and Jenkins, 1977; Cashdan, 1986). They share their territory with various Bantu-speakers including the Mbukushu (cultivators), the Yei (fishermen) and to a lesser extent the Tswana, Kgalagari and Herero herders. Each group operates in a different ecological niche. The San groups are concentrated on the banks of the Okavango River and the delta area as their informal name “river Bushmen” implies (Barnard, 1992). It is not clear whether these northern Khoe speaking San groups are Khoe-San groups with extensive Bantu-speaking admixture, Bantu-speakers that lost their cattle, another pastoralist population closely related to Bantu-speakers who occupied the region before the Bantu expansions or maybe a mixture of various refugee groups driven from the grazing grounds into the Okavango swamps (Cashdan, 1986).

1.1.1.2.3 The /Gui and //Gana of the central Kalahari

The /Gui and //Gana groups lived in an area now occupied by the Central Kalahari Game Reserve (CKGR) in central Botswana. /Gui has no specific meaning other than the reference to the group while //Gana is derived from a word that means “people of the well”. The /Gui and //Gana also shared the CKGR territory with the Kgalagari. The Kgalagari are the oldest existing Bantu-speaking tribe in Botswana. //Gana individuals all tend to speak Kgalagari as well as their own language and it is believed by the //Gana themselves that they originated from a intermixing of the /Gui and the Kgalagari. The /Gui occupied the region adjacent to the western CKGR as well as the western part of the CKGR and //Gana

the central and eastern part as well as the region adjacent to the eastern CKGR. The CKGR was established in 1961 and extends over 52 600 square kilometers. Only the southern (wooded zone) and central (bushveld) parts have enough vegetation to support human occupation. The central part is good hunting territory. From the 1960s to the 1980s the population in the CKGR declined from 2 000 to approximately 1 000 individuals. The Ghanzi district commissioner George Silberbauer studied the /Gui and //Gana groups extensively and constructed a borehole in the south central parts of the CKGR near the ‡Xade pan. Subsequently ‡Xade became a settlement with permanent occupation which grew from ~200 in the 1960s to ~700 in the late 1970s. In the late 1970s the people of ‡Xade were taught subsistent farming practices but with little available water this was not a successful strategy. The introduction of farming led to an increased number of livestock such as horses, donkeys and goats, which put further pressure on water supplies. Hunting on horseback and donkeys also ensued which caused a decline in large game and attracted the attention of wildlife park officials (Silberbauer, 1965; Barnard, 1992). A compromise was reached in which the San groups may stay as long as they only used traditional means of hunting.

In 1986 the government decided that the CKGR should strictly be a wildlife reserve and that residents should be relocated. San groups wished to stay in the reserve and proposed to work with park officials to sort out problems. This was declined and the resistance to resettlement was met with threats from the government and discontinuation of services. In 1997 the people of the CKGR were resettled from ‡Xade in the Central Kalahari Game Reserve to New ‡Xade, a large settlement in Ghanzi District, southwest of the reserve, and Kaudwane, a large settlement in Kweneng District not far from Khutse Game Reserve. Promises of large compensation to people that move soon were made. In reality very little compensation was paid-out and people struggled to keep their livelihoods. A San run NGO, First peoples of the Kalahari (FPO), worked with CKGR residents and took the Botswana government to court. In 2005, the government ruled that the CKGR was off limits to people even though some residents still lived there. San people trying to access the CKGR were shot at by government officials with teargas and rubber bullets, some individuals were injured, arrested and detained. In 2006 the final decision of the court was that San groups were unlawfully removed. The government, however, was not required to restore services

because it was not unlawful for them to have stopped these services. At the end of 2006 San groups were allowed to return but without any domestic stock. They are only allowed to live from hunting and gathering practices. Hunting licenses, however, are still not issued and people are living mainly of wild foods from the reserve and food they obtain from outside (Broyhill *et al.*, Current).

1.1.1.2.4 The Naro

The Naro live in the western parts of Botswana with the !Xõó to the south and the ǀX'ao//'ãesi to the north. They are the most numerous of the San groups and are estimated to be one fifth of the total San population. In the 1980s they numbered approximately 9 000 individuals; ~5 000 in Botswana and ~4 000 in Namibia (Barnard, 1992). Since the late 1800s the Naro shared a large part of their eastern territory (Ghanzi block) with white and recently, black ranchers. Southwest of the Ghanzi farm block their territory overlaps with the Xanagas farm block where ranches are mostly owned by individuals of mixed white-black ancestry and also mixed Nama ancestry. Other areas south of the farm blocks and in-between the blocks are shared between the Naro and Tswana, Kgalagari and Herero subsistence herders. The Kgalagari entered the area in the early 19th century while the Tswana and Herero have migrated there since the settlement of the white ranchers. A few small San groups (Ts'aokhoe, Qabekhoe, N/haints'e and ǀHaba) that are not Naro live in the northern parts of the Ghanzi block. They are linguistically grouped with northern and central Khoe speaking San groups with some linguistic similarities to the Naro. There is very little information available on these smaller groups (Guenther, 1986; Barnard, 1992; Guenther, 1996).

The areas occupied by the Naro have a relatively good water supply. Because of the ranches, however, the majority of the Naro have settled permanently at ranch boreholes, cattle-herding posts and towns. They supplement their traditional livelihood with herding, mafisa relationships and wage labor. They also act as tourist attractions on game farms in exchange for permission to use the land for gathering practices. They earn small salaries from ranch owners and/or tips from tourists. Some settled in the outskirts of the towns like Ghanzi and D'Kar and government settlement schemes such as the settlement at Hanahai. Unemployment and alcohol abuse is a big problem. The general mood under the Naro is

one of powerlessness, despair and deprivation. They lost their land and dignity and see themselves as weaker and less intelligent than surrounding groups. Names for themselves include “voiceless people” and “rubbish people”, but they still take pride in their language and traditions such as the trance-dance (Guenther, 1986; Barnard, 1992; Guenther, 1996) (Personal observation).

1.1.1.2.5 The Hai\lom

The Hai\lom live in the northern parts of Namibia in the areas around the Etosha pan. Their name means “tree” or “bush-sleepers”. They speak a language closely related to Nama and have been classified as !Xun who acquired the Nama language. The Hai\lom themselves, however, maintain to be a separate group with a separate group identity (Barnard, 1992).

1.1.1.3 The Kwadi

Very little is known about the Kwadi people of Angola. Their language is now extinct and the people have to a large extent integrated into surrounding groups. Records of their language suggest that they did speak a Khoe-related language (Table 1.1). Their language, however, were very distantly related to the extant Khoe languages as well as to the Khoe-speaking San languages. The closest language to Kwadi is geographically the most distant, namely, one of the eastern Khoe-speaking San group languages spoken by the Hietshware. Not much research has been done on the Kwadi but they seem to have been a large group of people in the past. They were mentioned in accounts of various navigators, historians and adventurers from the 16th to the 19th century. All mentioned the group of San people that lived near the mouth of the Curoca River. In the 1930 they were reported to be dying out and integrating into surrounding groups. In the 1950s only a few families remained. Their current status is unknown (De Almeida, 1965; Estermann, 1976; Barnard, 1992).

1.1.1.4 The Khoe

The Khoe can be divided into three ethnic divisions, namely, the !Ora (or Korana), the Cape Khoe and the Nama (Figure 1.2). Early reports also made mention of a fourth division, the Einiqua (language – “Eini”) that lived along the Orange River to the east of the

Korana, but very little is known about this group (Figure 1.2) (Schapera, 1930; Elphick, 1985; Barnard, 1992; Smith, 1995).

1.1.1.4.1 The Korana

The Korana (!Ora) were pastoralists that occupied much of the Karoo area of the Cape province but their descendants became absorbed / transformed into the Baster, Griqua and Coloured population of the area. Their early raiding activities were, however, recorded and remnants of their cultural practices survived into the 20th century (Schapera, 1930; Engelbrecht, 1936; Barnard, 1992). It is widely assumed that an essential proportion of the !Ora group came from the Cape Khoe people (see below) who fled from European colonization from the 1600's onwards. These fleeing Cape Khoe met and mixed with other people on their way and finally settled at the confluence of the Vaal and Orange Rivers where they also had contact with North KhoeKhoe-speaking pastoralists like the Nama of the Lower Orange and the Eini to the east (Güldemann, 2006b).

1.1.1.4.2 The Cape Khoe

The Cape Khoe was the pastoral population encountered by the first white settlers in 1652 at the Cape of Good Hope. They were spread over the southern parts of the Cape Province and three subdivisions were distinguishable, namely, the Eastern, Central and Western Cape Khoe. Periods of warfare between the Cape Khoe and the white settlers ensued but their final cultural collapse took place shortly after an overwhelming smallpox epidemic in 1713. Today their descendants are found among the "Coloured" population of the Cape province (Elphick, 1985; Barnard, 1992).

1.1.1.4.3 The Nama

The Nama are the best-known Khoe group. Today around 90 000 Nama individuals live in south and central Namibia, and to a lesser extent in the northern Cape (SA) and eastern parts of Botswana. The Nama people most probably came from an area located in the current northern parts of the Cape province (SA) and divided into two large subdivisions of people, the Great and the Little Nama (Westphal, 1963; Hoernle, 1985; Barnard, 1992).

The Great Nama (Gai-Naman) settled in the great Namakwaland area of Namibia prior to European contact. Several tribes existed with certain associated territories. In recorded history the Great Nama were divided into seven tribes (the Gai-//haun or *Rooi Nasie*; the !Gami-ṁnûn or *Bondelswarts*; the //Haboben or *Veldskoendraers*; the !Khara-khoen or *Kopers*; the //Khau-/gôan or *Swartboois*; the //Ô-gain or *Groot Doden*; the ṁAonin or *Topnaars*) (Westphal, 1963; Hoernle, 1985; Barnard, 1992). The Nama presently use mainly the Afrikaans group delineations (*italic*).

The Little Nama (ṁKham-Naman) only migrated into Namibia in the 19th century in separate tribal groups. They were also known collectively as the 'incoming groups' and the 'Oorlams'. The Little Nama tribes were the /Hôa-/aran or *Afrikaners*; the /Khobesin or *Witboois*; the !Aman or *Bethaniers*; the /Hai-khauan or *Bersebaers* and the Gai-/khauan or *Lamberts* or *Amraals*. These Little Nama tribes came from the south in search for better grazing but met with the Great Nama and Herero that were already there and conflicts developed. The Nama, who remained south of the Orange River, became incorporated into the "Coloured" population of South Africa (Westphal, 1963; Hoernle, 1985; Barnard, 1992).

The Nama lived a nomadic life and were pastoralists. With the incursion of Bantu-speakers and Europeans into their territory, their tribal organization shifted from hereditary chiefs to military leaders and chiefs. Early forms of tribal organization and social structure quickly deteriorated with German colonization in 1890. Additional factors include a severe drought and a rinderpest epidemic. The Nama revolt and resultant wars (1904-7) finally broke up traditional tribal structure. Although the tribes are dispersed today there are still some chiefs that maintain control over their traditional locations (Westphal, 1963; Hoernle, 1985; Barnard, 1992).

1.1.1.5 The !Xõó and the ṁHõã (Tuu division)

The !Xõó belong to the Southern Khoe-San language division (Tuu division, Taa branch). They identify themselves by a variety of names, !Xõó is the most widely used. The !Xõó consist of widely scattered groups of people that live in the southern parts of Botswana in one of the poorest environments of the Kalahari. Game and plant foods are sparse and permanent waterholes are few. The people, however, have extensive knowledge of their

environment and are able to identify and utilize over 200 plant species. Today the eastern parts of their territory have ample water supplies due to boreholes associated with the trans-Kalahari highway that runs through the area. A development project at Bere (south of Takathswaane) involved the construction of a borehole, shop, school and projects with guidance in livestock rearing. Many !Xóó people moved there and have settled around this area (Barnard, 1992).

The ǀHõã are distantly related to the !Xóó but they live in close proximity to them. Their language is thought to be an intermediate between Ju and Tuu. Their region is also shared by the Kgalagari herders who have been in the area for centuries and Nama individuals from Namibia (refugees from the time of German occupation).

1.1.1.6 Remnants and descendants of Khoe and San groups living in South Africa

The Khoe-San people of South Africa have to a large extent completely lost their identities and have integrated or transformed into other populations. What we presently know of Khoe and San peoples of South Africa are derived from studies on very few remnant populations that survived into the 1700-1800s.

The South African San groups belonged to the !Ui family of the Tuu (Southern Khoisan) language division. In historical times a large diversity of !Ui languages were spoken throughout all parts of the interior of South Africa. Their geographic range stretched from the Namaqualand in the west through the northern Cape, the Free State and Lesotho to KwaZulu-Natal and the south-eastern parts of Mpumalanga (old Transvaal). The best known of these languages is /Xam, a language mainly spoken in the Karoo, south of the Orange River. There were, however, numerous other !Ui languages more or less related to /Xam throughout South Africa. A few of these languages were recorded and still had a few active speakers in recent history like //Xegwi in the southeastern Transvaal. Of the other !Ui languages very little other than a name is known, like //Kx'au of Kimberly, //Ku //e (ǀUngkue) of Theunissen in the Free State, Seroa (N//ǀ or N//ng) of the Free State and Lesotho and !Gã !ne of the eastern Cape area (Traill, 1996).

Of the South African Khoe culture, language and traditions, very little also remains. In 1652 the Khoe pastoralists of the Cape or the Cape Khoe, spoke either the eastern or the western Khoe dialect. The speakers of these dialects, however, rapidly converted their language to Afrikaans or Xhosa (on the eastern frontier). The western dialect survived until recently in the form of !Ora (Korana) and Xiri (Griqua) among groups of Cape Khoe who migrated from the Cape to the Orange River area. The descendants of the Korana and Griqua adopted Afrikaans as their mother tongue and today South African Khoe languages are virtually extinct outside a few scattered individuals who retained some knowledge of the languages. One such individual lived near Colesberg. He spoke a dialect of !Ora that was largely unintelligible to Nama speakers, illustrating the differences between these two Khoe languages (Traill, 1996).

The next few sections describe the little knowledge we have about the history of these South African Khoe-San groups.

1.1.1.6.1 N//η people (“Mountain Bushmen”)

The N//η people or “People of the Eland” were groups of San people that inhabited the mountainous regions of Lesotho, Natal, Griqualand East and the former Transkei (from there their name “Mountain Bushmen”). Archaeological evidence indicates that the mountain regions were only occupied by San groups, with the influx of Bantu-speaking agriculturists into the regions of the KwaZulu-Natal midlands (Mazel, 1996). They were encountered by travelers and administrators of the 19th century but were already declining in numbers by then. At that stage, the available land was owned by Nguni and Sotho herders, and the San people lived by raiding the livestock of these herders. With the incoming white settlers the few remaining groups finally dwindled in numbers and they either died out and / or were absorbed into the Bantu-speaking groups (Wright, 1971; Vinnicombe, 1976; Barnard, 1992).

1.1.1.6.2 The //Xegwi

The //Xegwi is a group of San people that lived in the eastern Transvaal (now Mpumalanga) near Lake Chrissie. In the 1950s only 66 individuals were left (Potgieter, 1955; Ziervogel, 1955; Barnard, 1992). Today only single individuals who still recognize

their San ancestry remain, however, no one speaks the language or know of the cultural practices anymore (Personal observation). The last //Xegwi speaker who died in 1988 spoke their own language and Southern Sotho (Potgieter, 1955; Ziervogel, 1955; Barnard, 1992). The San of Lake Chrissie are believed to have been a collection of remnants from the original Transvaal San, such as those that inhabited the Honingklip shelter (Korsman and Plug, 1992) and also scattered refugee groups from the Orange Free State (Potgieter, 1955) and the Natal Drakensberg/Lesotho (Prins, Unknown). These groups fled from the in-coming Boer and English settlers and the turmoil that resulted from clashes between settlers and the Bantu-speakers. Various historical documents recorded a large group of San individuals migrating from the central Natal Drakensberg to the southern Transvaal highveld (Prins, Unknown). It is believed that these fleeing Drakensberg San composed a large part of the more recent San groups from Lake Chrissie. This is corroborated by the fact that the //Xegwi language were very similar to the languages of the “Mountain Bushman” and that their second language was Southern Sotho, a language spoken by Sotho people from Lesotho and surrounding areas (Potgieter, 1955; Prins, Unknown).

1.1.1.5.3 /Xam descendants

The /Xam inhabited a region of the Cape Province known as the great Karoo. The great Karoo area of South Africa is an arid scrubland with dispersed hills that stretch over an area of 400 000 sq/km of the Northern, Eastern and Western Cape provinces. This area was inhabited by both San and Khoe groups up until the late 1800's. The San group was the /Xam and the pastoralist Khoe group was part of the Korana group. The /Xam had subgroups (“Ss'wa ka” or “Plain bushmen”, “/nussa” or “Grass bushmen”, “!Kaoken ss'o” or “Mountain bushmen” and “Brinkkop bushmen”) but they all spoke the /Xam language with minor dialect differences (Traill, 1996).

The western world has learnt about the /Xam through the pioneering work of Wilhelm Bleek, a 17th century linguist who moved from Germany to the Cape Province. Bleek, his sister in law, Lucy Lloyd and his daughter, recorded the cultural practices, language and religion of the /Xam people while providing shelter to various /Xam individuals (www.lloydbleekcollection.uct.ac.za) (Deacon, 1996).

There are many reasons for the apparent disappearance of the /Xam; the principal factor probably is the advance of Bantu-speaking herders from the north and white colonists from the south, which led to the occupation and conquest of the great Karoo in the 18th century. Colonist hunters and farmers moved in and occupied all the remaining hunting ground previously used by the /Xam. The occupation of their resources was not the only reason for the disappearance of the /Xam, they were physically hunted by colonists and bounties were placed on their heads. Hunting parties were organized to hunt 'Bushmen'. Males that were not killed by hunters fled into the hilltops or were sent off to prisons. Females and children were relocated to farms to serve as farmhands, the so-called tame-bushmen. In the same way Khoe farmers living in the area were in competition with colonists for grazing ground. The Khoe people, however, claimed right to certain lands and had cattle to trade. They therefore generally received more respect from colonists than the San people (Barnard, 1992; Penn, 1996; Traill, 1996; Bennun, 2004).

The descendants of the /Xam females and children who were relocated to farms, today still live on some of the farms but became admixed with the local Xhosa (Bantu-speaking) population. Older farm owners still call some of their labourers "Bushmen" or recall that parents or grandparents of their workers were "Bushmen". Many farmers, however, tell the tale of "Bushmen" that couldn't settle in one place and had "wanderlust". These people became the "Karretjie" people that had their donkey carts as mobile units and moved from place to place to do different periodic jobs (De Jongh, 2002).

The Karretjie people

The word "Karretjie" is an Afrikaans word for 'donkey cart', alluding to their mobile lifestyle on donkey carts. Throughout the great Karoo there exist small bands of people living this mobile lifestyle but due to recent changes in economical factors, this way of living is quickly disappearing. The Karretjie people phenotypically resemble Khoe and San people. Oral and archaeological records also suggest Khoe and San ancestry but the group completely lost their original language and culture. They identify themselves as "Coloured" and speak Afrikaans. Most of the "Karretjie" people are sheep shearers and fencers. Typically they have a home base or as they call it "uitspan" or outspan where they keep their cart in between jobs. These outspans are usually on a neutral piece of land such as the section of

land between a road and a farm fence. They would stay in this space until their skills in shearing or fencing was required by a farmer. When this happened, they would pack their donkey cart and the whole family and living unit would move to the farm until the work was completed, after which they would move back to the same outspan (De Jongh, 2002).

1.1.1.6.4 The ‡Khomani

The ‡Khomani together with the /'Auni tribe and several other now extinct groups lived in the far northern parts of the northern Cape (north of Upington), the southern part of Botswana and the southern parts of Namibia. Roughly where the Kalahari Gemsbok Park is located today. They all spoke branches and dialects of the Taa-Lower Nossob branch of the Tuu family of Khoisan languages. In 1980 there were only few individuals left who remembered a lifestyle of active hunting and gathering in this area. They self identified as N/amani and !gabani but by then only spoke Nama (only one woman could speak the N/u language, but remembered only words). The individuals said that in the past the San of the Gemsbok park area used to live in small scattered groups in the summer and aggregated in the area of the Nossob River (southern Botswana) in the winter. There they traded goods (ostrich eggshell beads and animal skins) with Tswana groups. Their main food sources were gemsbok and small game as well as tsama melons and other wild food (Steyn, 1984; Barnard, 1992).

The Khoe-San people presently living in this area, spanning the borders of northern South Africa, southeast Namibia and southern Botswana, are from several different tribes that lost their individual tribal identities and speak either Afrikaans or Nama. The southern parts of Namibia, before the Nama colonization, had many San groups from the Taa language family. Today, however, all their descendants speak Nama (Güldemann, 2006a). The South African descendants of these San groups mostly classify themselves as Coloured. The following passage from Steyn illustrates how most of South Africa's Khoe and San have been reclassified as Coloured individuals.

“Regarding their present 'ethnic' status, Regopstaan and his wife, as well as Axerob and G/okos, said that they were classified as 'coloureds' and, with the exception of G/okos who was too young, are all 'coloured' pensioners. Although the others seemed to take some pride in what they apparently saw as an improvement on 'Bushman' status, Regopstaan took exception to this. He told the registering officer that he was no 'coloured', but a Bushman, a category that does not exist in the South African population classification system. In a sense he had his way; although not classified as a 'Bushman', he proudly showed me his identity card on which the bearer's name was registered as R. Boesman! “ (Steyn, 1984)

A group of South African descendants of these scattered southern Kalahari tribes now call themselves collectively ‡Khomani. They have had a recent rediscovery of their identity; they won a land claim and organized themselves into a community governed by a council. Only very few old individuals, from the Northern Cape (SA) and Botswana, however, still speak the N/u language. The term ‡Khomani was not known to the N/u speakers, it was introduced to San descendants of the northern Cape by representatives of the South African San Institute (SASI). Other than N/u, the only other extant Tuu language is !Xóõ, of southern Botswana. Unlike N/u, however, !Xóõ is still an active language and is being taught to children (Crawhall, 2003; Sands *et al.*, 2007).

1.1.1.5.5 South African Khoe descendant groups

The Khoe groups of South Africa included the Cape Khoe of the southern parts of the Cape Province, the Korana who occupied large parts of central South Africa extending over the Northern Cape into the Free State and the Nama of the North Western Cape region in the Richtersveld area extending into Namibia. Although Cape Khoe and Korana do not exist anymore today as specific populations their descendants were incorporated into “mixed culture” groups like the Griqua, Baster and Coloured groups with their associated cultures. Certain aspects of Khoe culture can still be recognized in rural areas where livestock rearing is the prime economic goal. In a way the Khoe culture formed the base of the Griqua, Baster and Coloured cultures that developed (Barnard, 1992).

1.1.1.5.6 The !Xun and Khwe of Platfontein

Although not originally from South Africa, the !Xun and Khwe of Platfontein now made South Africa their permanent home. They originally came from Angola and were employed by the South African Defense Force (SADF) before they were relocated to SA. Five

hundred veterans of the SADF together with 3500 dependants were relocated in 1990 from Namibia to South Africa (Sharp and Douglas, 1996). They currently live in Platfontein, near Kimberly.

The people of Platfontein are two different San groups with separate identities. One third of the people are known as Khwe (also were called Barakwena) and two thirds are !Xun (also were known as Vasekele). They speak different languages and have a different phenotypic appearance. The groups have remained separate and have insisted to be settled in different parts of the camp. The !Xun group retained a much more cohesive nature and cling to their San identity. They have not mixed with outsiders beyond the camp and have retained a much more unified group than their Khwe counterparts. The Khwe have been more ambivalent about their group identity and have established relationships with surrounding South African groups (Sharp and Douglas, 1996).

Although the people of Platfontein have separated themselves into these two groups, members within these groups were not individuals that came from the same area or even knew one another. The !Xun came from a wide region in central Angola around Serpa Pinto (currently Menongue) where many of them lived as stock farmers or cultivators alongside Bantu-speaking groups. !Xun men from different regions were recruited into the Portuguese colonial military in the late 1960s. When the Portuguese moved out the !Xun affiliated with a liberation force, FNLA, in the Serpa Pinto region. FNLA had links with the SADF and when FNLA collapsed the !Xun were recruited by the SADF and brought to the Omega military base in the Caprivi strip of the then South West Africa (Namibia) (Guenther, 1986; Sharp and Douglas, 1996).

The Khwe on the other hand originally came from south-east Angola where they have lived along the river systems as cultivators and cattle keepers. They have also originally come from a widespread region of southeast Angola and were recruited into a different unit by the Portuguese army. When the Portuguese moved out of Angola, the Khwe fled into neighboring countries like southwest Zambia, northwest Botswana and the Caprivi Strip of South West Africa where there were other Khwe people amongst whom many of the Khwe

soldiers had kin. From there they were recruited into the SADF (Guenther, 1986; Sharp and Douglas, 1996).

This difference in recruiting background underlies the differences in the attitude that the two groups had towards the SADF. The !Xun had a favorable opinion of the SADF because the SADF saved them from Angola when FNLA collapsed. Also, there were no resident !Xun population in the Caprivi and they were dependant on the SADF. On the other hand the Khwe were much more skeptic about the army and what the army had to offer them. This is because the Angolan Khwe blended into the local Khwe population and only joined the army at Omega base as a source of employment (Guenther, 1986; Sharp and Douglas, 1996).

Many of the !Xun were later (late 1970s) relocated to the second "Bushman battalion" in Tsumkwe. At Tsumkwe they were meant to join up with the Ju!'hoansi of Nyae Nyae but the Ju!'hoansi saw the !Xun as invaders and they had to be kept in isolated bases in western Bushmanland. Thus, in 1990 a large number of !Xun opted to come to South Africa while many of the Khwe stayed in the Caprivi where they had local contacts (Guenther, 1986; Sharp and Douglas, 1996).

Both these groups were relocated in 1990 to the Schmidtsdrift military base. The South African government was reluctant to allocate land or commit funds to secure the future of the San groups. The SADF saw these two groups as "former mercenaries who have outlived their usefulness" (Guenther, 1986; Sharp and Douglas, 1996). The !Xun and Khwe trust were established in 1993 to look after the interests of the groups. They remained in tented camps near the Schmidtsdrift military base for several years until recently, the new South African government allocated land to them in Platfontein near Kimberley, where they settled (Guenther, 1986; Sharp and Douglas, 1996).

1.2 Khoe-San history

1.2.1 Linguistics, Archaeology and Ethnography

1.2.1.1 Khoisan Linguistic Family

The languages of Africa are divided into four super language families, namely, Afro-Asiatic, Niger-Kordofanian, Nilo-Saharan and Khoisan. It has been believed for a long time that Khoisan is a single linguistic family with a common ancestor giving rise to all Khoisan languages (Greenberg, 1963). Recently, however, linguists studying Khoisan languages argue that all of the Khoisan languages are not necessarily genealogically related (Westphal, 1971; Güldemann, Forthcoming-a) and the similarities between some of the main branches of Khoisan may be due to areal language contact. These main branches might be genealogically related and have very deep roots, but even the best linguistic methods cannot distinguish chance, inheritance, and contact over time depth of over 10 000 years. Thus, current linguistic methods do not have the resolution to prove that all of the main Khoisan branches are genealogically related (Güldemann, 2007; Güldemann, Forthcoming-a).

Current understanding indicates that the Hadza, Sandawe, Khoe-Kwadi, Ju and Tuu language families and possibly the ꞤHǃ language are linguistic independent lineages within the Khoisan language group (see Table 1.1). They represent separate genealogical groups, which have not yet been proved to be linguistically related to each other or to any other language in the world (Güldemann, 2007; Güldemann, Forthcoming-a). While Hadza appear to be totally unrelated to all the other Khoisan languages, a recent study, however, does note a promising relationship between Khoe-Kwadi and Sandawe (Güldemann and Elderkin, Forthcoming).

There are only a few other similarities in the main branches of Khoisan languages than the fact that they use clicks as phononemic speech sounds. Many languages over the world, however, use clicks as paralinguistic speech sounds. Also at least one other language not related or in contact with the Khoisan language family, in aboriginal Australia, use clicks within their language. Instead of genealogical relationships between all the Khoe-San

languages it might be possible that there was an earlier linguistic macro-area that stretched from eastern Africa to the southern Africa with a linguistic-areal connection. The Bantu expansion into eastern and southern Africa erased this connection by causing the extinction of a many local languages, which might have shared clicks as a common phoneme type (Trautmüller, 2003; Güldemann, 2007).

To determine if there is deeper genealogical structure within the main branches of Khoisan, proto languages were inferred predating the current languages. In doing so it was discovered that the Khoe language was related to the now extinct language Kwadi. Khoe and Kwadi would have formed two sister branches deriving from an ancestral language, Proto-Khoe-Kwadi (Güldemann, Forthcoming-b; Güldemann and Elderkin, Forthcoming). Khoe-Kwadi also showed promising links to the east African language Sandawe (the other east African click language, Hadza, however, show no relationship to any other language) (Güldemann and Elderkin, Forthcoming).

1.2.1.2 Khoe-San History according to Linguistics

The Ju and Tuu branches (non-Khoe branches) of the southern African Khoisan family show some linguistic homogeneity but the link is unclear from a historical perspective. It can either be due to a very old common ancestor or due to areal convergence of two distinct lineages over a very long time. This group in cultural-ethnology terms consists of foragers only and shows continuity from very old archaeological records (Güldemann, In Press). The Tuu (southern) branches is thought to have a separation that goes back the furthest in history, based on the degree of linguistic distances of languages within the branches. The languages within the Tuu branch differ widely among themselves suggesting an extended process of divergent development. In the Ju branch languages are closely related to each other but not always mutually intelligible (Vossen, 1998; Miller-Ockhuizen and Sands, 1999).

The Khoe-Kwadi branch is the largest attested Khoisan lineage; it contains considerable internal sub-branching and has a wide geographic spread. All of this suggests divergence and expansion of this family. The population is also diverse in cultural-ethnology terms and consists of both foragers and pastoralists (Vossen, 1998; Güldemann, In Press).

In historic times the Cape region had two groups of people that spoke two genealogical unrelated Khoisan languages. One group spoke a language belonging to the Khoe branch of Khoisan languages and the other group spoke an unrelated language belonging to the Tuu branch of Khoisan languages. Only two languages in these two branches still have active speakers in the Cape today. Nama (Khoe branch) are spoken by a few thousand people in the Richtersveld area in the northwestern corner of South Africa and N//u (Tuu branch) are spoken by fewer than 20 individuals scattered over the Northern Cape region north of the Orange River. A few extinct languages from this area, namely, !Ora (Khoe branch) and ǀUngkue and /Xam (both Tuu branch) have sufficient recordings to be linguistically analysed (Güldemann, 2006b).

The history of the Cape region of southern Africa, inferred from a linguistic perspective can be summarised as follows. The oldest known ethno-historical layer was the foraging society of the San. In the Cape the group involved, correlated with the !Ui linguistic unit. From 2 000-2 500 years BP, a new cultural type with animal husbandry appeared according to archaeological findings. In the Cape this group correlate with the distinct linguistic group, the KhoeKhoe. The archaeological record of the trajectory of pastoral expansion suggests that the KhoeKhoe entered the Cape from the north rather than the east. Corroborating this is the fact that the linguistic groups most closely related to the KhoeKhoe (the Kalahari Khoe) live in Botswana, Namibia and Angola. Due to their mode of life, pastoralists did not inhabit inhospitable areas like the Karoo and Kalahari. In coastal areas, however, and areas around great rivers a co-habitation of the !Ui foragers and the KhoeKhoe pastoralists for around two millennia is assumed. Because of the asymmetric relationship that usually exists between hunter-gatherers and pastoralist it would be expected that there would be an incorporation of hunter-gatherer females into the pastoralist group together with culture and language elements but not the other way around. From this would follow that the !Ui language would have an influence on KhoeKhoe. This can clearly be seen in a linguistic analyses of the KhoeKhoe language compared to the !Ui languages. Compared to the other Tuu languages the !Ui language structure stayed relatively unchanged while KhoeKhoe diverged from other Khoe languages (Kalahari Khoe and Kwadi) and incorporated many linguistic elements from the

!Ui branch of the Tuu languages and leading to a situation where KhoeKhoe have a strong linguistic substrate of the Tuu languages. This scenario would imply that some gene-flow has occurred, probably from the San to the Khoe (through the incorporation of San females by the Khoe). Thus in a genetic sense the geneflow from the southern San, !Ui speakers into the KhoeKhoe would be apparent through studies on mitochondrial DNA but not in Y-chromosome studies, while autosomal markers would give an intermediate picture. The KhoeKhoe of southern Africa later expanded and moved back into Namibia and became the Nama of Namibia, but still retain the evidence of contact with the southern San !Ui in their language and presumably also would in their genetics (Güldemann, 2006b).

1.2.1.3 Khoe-San History according to Archaeology and Ethnography

Archaeology is widely used to study and infer the history of the human population. An advantage that archaeology has is that some of the material used in investigations are very robust and withstand deterioration through time very efficiently. Depending on the material used in investigation (i.e. wood, bone, stone, etc.) the time depth investigated could be very deep. Since early hominid species used stone tools, archaeology can investigate hominid associated culture and demographics up to millions of years before present.

It is generally assumed that the presence of flaked stone artifacts in the archeological record indicate the presence of true humans of the genus *Homo*. The time period in which members of the genus *Homo* had the ability to use and manipulate stone is known as the Stone Age. The Stone Age started ~2.5 million years BP and is divided into three stages, namely, the Earlier, Middle and Later Stone Age. Throughout all the stages of the Stone Age, humans were present in southern Africa. Their signature was left behind in the changes they caused in their environment and are studied by archaeologists.

By studying the archeological record it is possible to identify certain demographic tendencies in the human populations involved. For example, it is possible through looking at the sizes and frequencies of archaeological sites to infer population densities and thereby identify population expansions and contractions. These expansions and contractions can then be linked to certain events in the paleoenvironment. Similarly, by studying the genetic variation present in extant populations one can also identify historical

population expansion and contraction patterns which can be dated to certain times in the past. It is therefore one of the aims of this thesis to identify these genetic signatures of population expansions and contractions and to try and correlate it with information available from the field of archaeology. The next section thus reviews the different stages in the archaeological record, their associated times, paleoenvironment and signatures of human occupation.

The Earlier Stone Age (ESA) occupied the time period from 2.5 million years BP to 250 000 years BP in southern Africa and is characterized by the use of large rudimentary flaked artifacts like handaxes. Throughout this stage there is evidence of the occupation of southern Africa by various hominans (humans and their extinct relatives) (Deacon and Deacon, 1999; Mitchell, 2002; Wadley, 2007).

The Middle Stone Age (MSA) saw the introduction of “cores” (pieces of rocks that are skillfully prepared to produce flakes of regular size and shape) into the archaeological record and stretched from 250 000 years BP to ~30 000 years BP in southern Africa. MSA tools were generally smaller than ESA tools and lack the large handaxes and cleavers. There is no consensus on the definition of the MSA. Some archaeologists believe that it is a time related sequence while, others identify it as a package of technologies. For some archaeologists the MSA in southern Africa is associated with the appearance of anatomically modern people (*Homo sapiens*) (Wadley, 2007; Lombard, 2008). This was confirmed with the discovery of remains of early modern human fossils dated to 90 000 and 110 000 - 120 000 years BP at the Klasies River site in the eastern Cape. Further proof was the discovery of early modern human remains of a similar time period at a site named Border Cave on the KwaZulu-Natal, Swaziland border. Furthermore some form of symbolism can be dated as far back as 77 000 years BP. The shell beads from this period found at Blombos Cave imply individual or group identity and symbolism. While these cognitively and anatomically modern humans were roaming southern Africa, the European landscape was still dominated by Neanderthals (Deacon and Deacon, 1999; Henshilwood *et al.*, 2002; Wadley, 2007; Lombard, 2008). The earliest known set of morphological characteristics associated with modern humans, however, appears in fossil remains from Ethiopia, dated to ~150 000 – 190 000 years BP (White *et al.*, 2003; McDougall *et al.*,

2005). This finding does not exclude the probability that modern morphological traits existed in other regions of Africa (such as southern Africa) during this time. In other regions specimens may have been less well preserved or archaeological and paleontological investigations may not have been conducted as yet (Lahr and Foley, 1998; Reed and Tishkoff, 2006). Presently a multiregional origin model for modern humans within Africa is not as unlikely as it would be for global populations (Lahr and Foley, 1998; Campbell and Tishkoff, 2008).

Regarding the paleoenvironment of the MSA, it was previously believed that the period between 60 000 and 25 000 years BP was marked by very arid conditions in southern Africa, which led to a continuous population decline (Klein, 2000; Klein *et al.*, 2004). This was partly inferred by an impoverished archaeological record for this period. Recently, however, a paper by Mitchell (Mitchell, 2008) summarised paleoenvironmental data that refute the presence of hyperarid conditions in southern Africa during this period. Furthermore, he showed that a substantial archaeological record does exist for this period, albeit not as well studied as the periods that flank this stage (the earlier Stilbaai and Howiespoort cultures and the later LSA period). In addition many of the human foci, which exploited coastal resources during this period, are today submerged since the sea-level was 30-60m below the present level (Mitchell, 2008).

The Later Stone Age (LSA) display technology to produce small specialized tools, such as microlithic tools and saw the introduction of bows and arrows, needles, bored stones, fishing equipment, etc. This period stretched from between 30 000 - 20 000 years BP to 2 000 years BP. The transition from the MSA to LSA is an uncertain concept, while some archaeologists believe the LSA began as early as 40 000 years BP, others insist that in certain regions MSA technology can only be found as recent as 20 000 years BP. It was suggested that the division of the MSA and LSA might be more of an archaeological construct than a real divide. The LSA do, however, have marked technological innovations and a regular occurrence of behaviour that were only rarely found in the MSA (Wadley, 2007). It is almost certain that the LSA sites were occupied by the descendants of the people who practiced MSA technology. Many sites have evidence for both complexes. San art, tools, burials and other remains of San hunter-gatherer lifestyle is associated with the

LSA and can be traced back with confidence as far as 22 000 years BP in the archaeological record. Also San social structure is very evident in archaeological remains for the past 10 000 years. Archaeological deposits from the MSA suggest that the social organization and rules of group behaviour did not change with the transition of the MSA into the LSA and were the same for the last 100 000 years or more. It is most likely that the MSA people that lived in southern Africa were the direct ancestors of the LSA people, namely, the San (Deacon and Deacon, 1999).

Concerning the paleoenvironment of the LSA, the period leading up to the Last Glacial Maximum (LGM) (28 000 to 19 500 years BP) is marked by the occurrence of 'higher energy' human settlement during certain periods and at certain sites. These sites include Lesotho, southern Cape, Caledon valley, southern Namibia and the southern Kalahari. The LGM period (18 000 years BP) was associated by significantly colder conditions and intensified aridity before moister and milder conditions returned after 16 000 years BP. The LGM associated period (19 000 – 15 000 years BP) is marked by a major downturn in population size and distribution and may have caused localized extinctions. The rise in population numbers after the LGM was initially seen only at the few sites that existed though the LGM. The rise was slow until 13 500 years BP, thereafter population growth accelerated and deserted sites were reoccupied and new sites established. Distinct technological traditions for this period are reported for sites from South Africa (Robberg industry) compared to Namibian, northern Botswana and Mashonaland sites. It was suggested that this distinction could reach back to the distinct Tuu and Ju linguistic traditions and possibly also genetic distinctness (Mitchell, 2002).

Relative cool conditions remained throughout the Pleistocene to Holocene conversion (10 000 years BP) and maximum temperatures were only reached 8 000 years BP. The rise in sea level was effectively completed around 9 000 years BP and submerged large areas of previously exposed grassland. Groups became more concentrated and social exchange between groups increased. The later Holocene sites (~4 000 years BP) documents rising populations, expansions into new habitats and elaboration of material culture, especially in the Cape Fold Belt and Thukela basin. Technologies, which were characterized by delayed rather than immediate returns, developed and increased. For instance, the "firestick-

farming” technology developed and practiced in the southern and eastern Cape, which regulates flowering and production times of geophytes, increased food production capabilities of populations dramatically (Mitchell, 2002).

The archaeological record from 2 000 years BP changed radically with the introduction of pastoralism to southern Africa. This transition is marked by the introduction of pottery and sheep remains in the archaeological record followed by the introduction of cattle and domesticated dogs. The herder way of life is associated with the people who spoke the Khoe languages. The general feeling among current researchers is that a sheep herding economy and ceramics were adopted by aboriginal Khoe-speaking hunter-gatherers from Bantu-speaking agro-pastoralists. These agro-pastoralists were spreading south from east Africa and arrived in Zambia/Zimbabwe ~ 2 100 years BP. Current theories suggests that the transfer took place in southeastern Angola, southwestern Zambia or northern Botswana. From the core area of northern Botswana the sheep together with the Khoe-speaking herders migrated southwards and gradually settled in between the hunter-gatherers from South Africa (Smith, 1983; Smith, 1992; Sadr, 1998).

Two migration routes are proposed, the first hypothesizes that stock keepers came west through northern Botswana and Namibia, down the Atlantic coast to the Cape and then further along the south coast and inland Cape areas (Stow, 1905; Cooke, 1965). This theory is based on the occurrence of paintings of sheep and shepherds in Zimbabwe and the ecological improbability of moving through the central Kalahari. This theory is also supported by records of oral traditions (Stow, 1905; Cooke, 1965). The second theory proposed that Khoe groups from northern Botswana acquired livestock from their Bantu-speaking Iron Age neighbors to the north. Subsequently their population and herds grew and the population spread south along the Zimbabwe/Botswana border, east of the Kalahari, towards the confluence of the Orange and Vaal Rivers. From there some groups spread south to the coast following one of the river valleys such as the Seekoe River and from there east and west along the coast. Other groups followed the Orange River to the Atlantic from where they spread north into Namibia and south into Namaqualand (Elphick, 1977). At the moment various archaeological findings lends more support to the Atlantic coastal route through Namibia (Mitchell, 2002). The earliest dates for the arrival of sheep

and ceramics in the Cape is 2 100 years BP in the northern Cape and 1 900 years BP on the southern Cape coast (Sealy and Yates, 1994; Henshilwood, 1996). It is further suggested that some of the hunter-gatherers of the Cape area were recruited and incorporated into Khoe culture. Those who remained hunter-gatherers moved into areas unsuitable for domestic stock or settled into an established working and trading relationship with the herders (Smith, 1983; Smith, 1992; Deacon and Deacon, 1999).

Both theories support population movement from the northern Botswana Khoe groups together with the pastoralism culture to the southern parts of Africa. This is supported by the linguistic, glottochronology findings that the KhoeKhoe languages of the south, diverged from the Kalahari Khoe languages ~2 000 years BP (Ehret and Posnansky, 1982). The homogeneity of the KhoeKhoe dialects further indicates a rapid recent expansion. The KhoeKhoe expansion is, however, only one component of the spread on the Khoe language group. The explosion of sites across Botswana in the last 2 000 years coupled to the oral traditions of Khoe-speaking San groups that they formerly owned livestock might be an indication of the Kalahari Khoe expansion linked to pastoralist groups/culture (Walker, 1995; Mitchell, 2002).

Certain evidence in the archaeological record, however, indicates that a simple demic diffusion model might not be sufficient to explain the spread of pastoralism. The spread of ceramics is thought to be associated with the spread of pastoralism and the two technologies form a package. The rapidity of the spread of ceramics ahead of the pastoralist culture and their occurrence in sites where herders never penetrated raises questions. If ceramics and pastoralism was a package spread by the KhoeKhoe herders they would regularly co-occur, which is not always the case. There would also be a ceramic stylistic chain that link assemblages in the Cape and southern Namibia to those in Botswana from whence they came. The stylistic chain would thus mirror the migration routes of the people. Archaeological sites with sufficient material are not adequate to make definite conclusions. Thus far, however, evidence of radical differences between styles argues against a common origin (Sadr, 1998; Mitchell, 2002).

The theories supporting a demic diffusion, argues that it would be very improbable of hunter-gatherers to adopt the pastoralist culture and therefore a population that spread the pastoralist culture is essential. Two hypotheses exist about the interaction between hunter-gatherers and herders. The first is that there is a great deal of overlap between these social and economic categories, and hunters who obtained stock could easily convert to herding, while herders who lost their stock easily fall back to hunting (Elphick, 1977). This theory thus support that at least some of the first livestock diffused southward from one to another group of hunter-gatherers (Deacon *et al.*, 1978; Deacon, 1984; Klein, 1986; Kinahan, 1995). The other hypothesis supports separate social and economic groups that do not interchange easily (Parkington, 1984; Parkington *et al.*, 1986). These separate groups can be identified archaeologically through different cultural signatures in deposits (Smith *et al.*, 1991; Smith, 1992). Hunter-gatherers are seen as groups living on the fringe of herding society. They utilize wild resources, but occasionally interact as clients, through trading with herders. The hunters will also make forays against the herds of the herders leading to persecution and wars. Hunting and herding is thus quite discrete economic categories with hunter-gatherers occupying niches on the fringes of pastoralist society in a lower class or subservient status. These theories argues that the pastoralism culture requires a fundamental change in how hunter-gatherer social relations are organized and that the conversion of hunter-gatherer culture to pastoralism is very improbable (Parkington, 1984; Parkington *et al.*, 1986; Smith, 1986; Smith *et al.*, 1991; Smith, 1992; Boonzaier *et al.*, 1996).

Archaeology can, however, not conclusively prove whether the spread of pastoralism is associated with a demic diffusion of populations together with the pastoralist culture or a diffusion of the culture on its own. An intermediate model where only few individuals, perhaps only males, spread and transferred the pastoralist tradition and their language to resident hunter-gatherer groups further south is also possible. A genetic approach using male specific and female specific markers would be ideal in this case and would be addressed in one of the aims of this thesis.

The introduction and manipulation of iron and copper tools in southern Africa is known as the Iron Age and is associated with the arrival of the pre-colonial Bantu-speaking farmers

(Deacon and Deacon, 1999). The relationship and interaction between the hunter-gatherers and the in-moving Bantu-speakers is another hotly debated topic in the archaeological community. While some groups support hunter-gatherers as affluent independent communities (Marshall, 1976; Lee, 1979) others support the theory that in-moving Bantu-speakers marginalized, dispossessed and isolated San communities (Wilmsen, 1989; Wilmsen *et al.*, 1990). There is also support for a theory that San-Bantu-speaker relations varied temporally and geographically. In some instances they may have retained their independent hunter-gatherer lifestyles and in some they may have been marginalized and subjected by Bantu-speakers (Campbell, 1990; Sadr, 1997). Furthermore, some communities may in fact have had active and beneficial trade relations with Bantu-speakers and therefore benefiting indirectly from the cultivator/pastoralist culture (Nurse, 1983; Denbow and Wilmsen, 1986; Campbell, 1990; Sadr, 1997). Some resolution to this problem might be found in the analysis of population expansion and bottleneck/contraction signals found in genetic data. If in-moving pastoralists adversely affected San communities a signal of a recent population contraction would be evident. Such a post-Neolithic population bottleneck was indeed proposed recently through analysis of hunter-gatherer genetic data (Excoffier and Schneider, 1999) (see further discussion in section 1.2.2.2). An investigation of genetic evidence for recent population bottlenecks associated with the in-moving herders will form the basis of one of the aims of the present thesis.

The major events evident in the archaeological record in the history of the LSA San hunter-gatherers and the Khoe herders can be summarised as follows. During the MSA to LSA transition (30 000 – 20 000 years BP) there was the introduction of the specialized LSA technology and certain sites showed increases in population sizes but only for truncated periods. The population density only increased noticeably from 13 500 years ago and especially in the last 4 000 years. The hunter-gatherers from northern Botswana adopted a herding economy (and perhaps the Khoe-language) 2 000 years BP and migrated southwards into South Africa. During the same time Bantu-speakers moved southwards from East Africa and settled in the eastern parts of South Africa (Deacon and Deacon, 1999; Mitchell, 2002). There was trade and interaction between San, Khoe and the metal-working Bantu-speaking agriculturists of the Early Iron age. At the time of European

colonization the eastern part of southern Africa had been populated by Iron Age Bantu-speakers for about 1 000 years. Hunter-gatherers had developed working relationships with the Bantu-speakers as well as the Khoe herders who had been settled in the southern and western parts of southern Africa for at least 1500 years. This situation was disrupted by the loss of control over land with the start of the European colonization (Deacon and Deacon, 1999).

1.2.1.4 Khoe-San history according to Physical Anthropology

While the archaeological record attest to a continuous human occupancy of southern Africa from the Earlier Stone Age to present times, it is difficult to directly link specific signatures and fossils in the record to ancestors of present day populations. Aside from ancient DNA studies, which up to now have not been successfully conducted on human fossils from southern Africa, physical anthropology provides a possible solution to the problem (Morris, 2005; Morris, 2008). The field of physical anthropology studies the osteological features of fossils and compares them to current osteological features from different populations. Although this field of research are regarded by some scientists as controversial and/or obsolete, this is still an active area of research that uses state of the art statistical procedures and contributes valuable hypotheses about the history of the Khoe-San in southern Africa (Morris, 2005; Morris, 2008). It should, although, be stressed that genetic studies conducted to date have not been able to show correlation between morphological features and genetic variants. In other words, it is not possible to explain how the different anthropometric traits found in modern humans have come about and which gene(s) are responsible for generating particular traits. The next section briefly outlines how studies based on methods used by physical anthropologists have contributed to reconstructing the early history of southern African populations.

Craniometric studies suggest that the earliest appearance of the morphological traits found in South African Khoe-San could be traced to around the terminal Pleistocene and early Holocene period (around 12 000 BP) (Stynder *et al.*, 2007a; Stynder *et al.*, 2007b). The fossil evidence before 50 000 years BP are difficult to link to any contemporary population (Beaumont, 1980; Grun *et al.*, 1990; Morris, 1992), while late MSA osteological features such as the Hofmeyer cranium (36 000 years BP) falls outside the range of modern Khoe-

San variation (but surprisingly fall within the range of European Upper Paleolithic cranial variation) (Grine *et al.*, 2007). A hypothesis that was put forward was that the aboriginal populations from southern Africa developed the distinct Khoe-San morphological traits after a period of isolation, caused by the arid conditions of the LGM, in which drift and selection acted on the isolated southern African populations (Morris, 2002).

In contrast to the uncertainty surrounding morphological features of fossils before the LGM, it has been shown that there was continuity in the morphological features of fossils from the terminal Pleistocene until present day. Two periods of possible genetic discontinuity was identified around 4 000 years BP, when the population sizes increased dramatically and around 2 000 years BP, when pastoralism was introduced. It was however concluded that the variations of morphology during these two stages was most likely due to *in situ* changes of populations in response to environmental factors (Stynder *et al.*, 2007a). Aside from the evidence of population continuity across the time when pastoralism was introduced to the southern parts of Africa, further craniometric studies failed to find evidence for distinctive features between southern African hunter-gatherers and herders (Stynder, 2009). The study note a small increase in variation during this period but fail to find support for a large-scale immigration of morphologically different herders or the long term co-existence of two different populations. The two hypotheses offered to explain the small increase of variation were (1) a small-scale immigration of morphologically distinct herders or (2) increased morphological variation in response to lifestyle changes due to the adoption of pastoralist practices (Stynder, 2009).

Another area in which physical anthropology contributed valuable information is the question of whether a connection exists between the Khoe-San and east African populations. While evidence of recent common ancestry or contact between these two groups exist in the linguistic and genetic fields, contemporary physical anthropology have not found any support of morphological commonalities (Morris, 2003; Morris and Ribot, 2006). For decades the presence of ancient Khoe-San populations in east Africa has been accepted in anthropological literature, however, a review of the initial studies and evidence failed to find any support of a overlap between east African and Khoe-San morphological variation (Reviewed in Morris, 2003 and Morris, 2008).

1.2.2 Khoe-San history according to molecular genetic studies

During the course of the 20th century molecular biology and genetics started to contribute towards inferences of the histories of different population groups. Studies in physical anthropology that concentrated on morphological trait differences in the different groups were also common (Reviewed in Tobias, 1985). However, because phenotypic traits are not inherited in a straightforward manner and are more sensitive to observational error and environmental influence many reports and articles seems to be contradictory. While some of the studies could find differences between San and Khoe groups and their different subgroups other studies failed to find significant differences (Jenkins, 1986). The first molecular biology study on Khoe-San groups was based on the use of ABO blood groups in 1932 (Pijper, 1932). Since then several other serogenetic markers have been used to examine patterns of genetic affinities of the Khoe-San (Reviewed in Nurse *et al.*, 1985; Jenkins, 1986 and Jenkins, 1988). Section 1.2.2.1 will give an overview of these serological studies and highlight the important findings. More recently work on the hereditary material itself, DNA, were published. These studies are, however, few and involve only few selected Khoe-San groups. Sections 1.2.2.2 - 1.2.2.4 will review the published genetic studies on Khoe-San.

1.2.2.1 Serological studies

1.2.2.1.1 Differences between San and Khoe

When serological studies were conducted on the San and the Khoe the most prominent difference between them were found using the ABO and Rhesus blood group systems as well as the haptoglobins. In the ABO system the B allele has a very low frequency in the San groups (including the Khoe speaking San groups like the /Gui, //Gana and Naro) and occur at frequencies less than 0.04. In the Khoe groups (Sesfontein Topnaars, Tsumaris Nama and Nama from southern Namibia) it was found at frequencies 4-8 times higher than that in the San groups. These frequencies are similar to the Zimbabwean and Zambian Bantu-speakers and marginally lower than the South African Bantu-speakers (Pijper, 1932; Pijper, 1935; Zoutendyk *et al.*, 1955; Jenkins and Nurse, 1972; Jenkins, 1986). In the Rhesus system the allele frequencies of the different alleles differed significantly between San and Khoe groups and the frequencies in Khoe groups correspond more to the

frequencies found in Bantu speakers. Haptoglobin Hp^1 frequencies is also different in San and Khoe groups with a low frequency in San groups and a higher frequency in Khoe groups again corresponding to the higher frequencies in Bantu-speakers (Jenkins and Nurse, 1972; Jenkins, 1986). A possible explanation for the correspondence of the Khoe frequencies to the Bantu-speaking frequencies rather than to the San groups can be due to the high amount of Nama (a subservient group with Bantu-speaking ancestry) admixture into the Nama groups. Since the Nama is the only extant Khoe group the hypothesis cannot be tested by comparing it to the frequencies of other Khoe groups (Jenkins, 1986).

1.2.2.1.2 Differences between Khoe-San subgroups

In 1971 Jenkins *et al.*, combined allele frequencies from several serogenetic studies (blood groups, serum protein and red cell enzyme systems) in multivariate analysis through genetic distances. Data from gene frequencies in 10 loci in 18 southern African populations were compared through genetic distance measures coupled to clustering methods (Jenkins *et al.*, 1971; Jenkins, 1986). Figure 1.3 A and B shows a clustered tree and Principal Component Analysis (PCA) plot adapted from the genetic distance matrix published in Jenkins (1986).

In the cluster analysis (Figure 1.3 A) all the Khoe and Khoe descendant groups cluster together. The grouping of the Hai//om within the Khoe cluster is in contrast to the theories that the Hai//om is a !Xun group that acquired the Nama language. The Hai//om does not cluster with any of the four Ju groups (!Xun-Vasekele, !Xun-Kavango valley, Ju\`hoansi, ǀX'ao//ãesi) but rather with two Nama groups. Interestingly, the Hai//om that inhabit northern Namibia (Figure 1.1), cluster more closely with the two southern Nama groups of Keetmanshoop and !Kuboes (Richtersveld) than the more northern Rehoboth Nama. This, however, might have to do with the various amount of admixture from Caucasoid and Bantu-speaking groups into the different Nama groups. The clustering of the Bastards and Coloured populations within the Khoe cluster confirm the large inputs of Khoe groups into these two groups of hybrid ancestry.

All of the San groups except the !Xun (previously also referred to a Vasekele) form a monophyletic cluster. The clustering within the San cluster does not conform to linguistic

clustering but rather with geographic proximity. The speakers of northern San languages of the Ju linguistic group (Tsumkwe Ju\’hoansi, ǀX’ao//’ãesi, Kavango valley !Xun and the Vasekele !Xun) does not form a uniform cluster and neither do the Khoe-speaking San groups of the central Kalahari (Naro, /Gui and //Gana). Rather the Naro (Khoe linguistic group) and ǀX’ao//’ãesi (Ju linguistic group) who have geographically overlapping territories form the closest cluster; followed by the Ju\’hoansi (Ju) which is also geographically close. The G!ang!ai !Xun then joins the cluster also according to geographic distance. The two central Kalahari San groups /Gui and //Gana, forms a separate branch within the San cluster. This indicates that geographic separation has a greater influence on geneflow than linguistic barriers. The !Xun cluster with two Bantu-Speaking groups suggests a higher amount of Bantu-Speaking admixture.

The Bantu-speaking groups (Kgalagari from Botswana, Ngwato - a Tswana group from Botswana, Herero from Namibia) form two separate clusters that also include the Khwe and Dama who are classified as “Khoisan speaking Negroes” by Jenkins (1986). The Khwe appear to be most closely related to the Herero supporting observations that the Khwe phenotypically resemble Bantu-speakers rather than Khoe-San groups. The Khwe cluster closer with the Western Bantu speaking Herero group than with the Eastern-Bantu speaking Ngwato. The Dama also clearly cluster with the Bantu-speakers confirming historic accounts that the Dama were similar genetically to Bantu-speaking people before they adopted Nama as a language, possibly a consequence of their enslavement by the Nama. This study together with others (Nurse *et al.*, 1976; Nurse and Jenkins, 1977) have thus shown that the Khwe (as well as the Dama) have genetic profiles that are more similar to Bantu-speakers. In a further study it was found that the Khwe most closely resemble their Mbukushu neighbours and an Ambo chiefdom (the Ndonga) (Nurse and Jenkins, 1977).

The first axis on the PCA plot contains more than half of the total variation (56%) and summarises the Bantu-speaking versus Khoe-San variance component. The Bantu-Speaking groups cluster to the one side of this component while the northern San (NS) groups occupy the other extreme. Even though the Dama speak a Khoe language they show very little admixture from Khoe-San groups. The Kgalagari is known to live in close

contact with the /Gui and //Gana and the relative higher Khoe-San contribution into this group is evident in the first component. The Khwe, although grouping with the Bantu-Speaking groups show a higher contribution from the Khoe-San variance component than most of the Bantu-Speaking groups. The Nama, Baster and Coloured groups are located between the San and Bantu-Speaking groups on the first component, indicating higher admixture from Bantu-Speaking groups into these groups. This evidence of geneflow from Bantu-Speaking into the Khoe and Khoe descendant groups are expected, due to enhanced contact between these groups as a consequence of the sharing of the pastoralist culture (versus the absence of pastoralism in the San groups). With exception of the Vasekele !Xun and the //Gana, all the other northern and central San groups contain little Bantu-Speaking admixture. As mentioned before, the Vasekele !Xun lived in close association with the local Ambo (Ovambo) population for centuries, from whom they learned crop cultivation and herding. The higher Bantu-Speaking component in this group is therefore not surprising. Also, the higher Bantu-Speaking component in the //Gana (G//ana) is expected since it is believed by the //Gana themselves that they originated from a intermixing of the /Gui (G/wi) and the Kgalagari.

The second and third component seem to both summarise a component of variance that exist between the Khoe groups and the !Xun groups. The second component (21%) summarises variation between the !Xun of the Kavango valley in northern Namibia and the Khoe groups. The /Gui, //Gana, Ju'hoansi, ǀX'ao//'ãesi and Naro occupy intermediate positions with the /Gui, //Gana located more to the !Xun side and the other three groups more towards the Khoe side. The third component separate the Vasekele !Xun from the Khoe groups with the Ju'hoansi, ǀX'ao//'ãesi and Naro intermediate.

This thesis will investigate the genetic relationships between different Khoe and San groups further to see if the mitochondrial, Y-chromosome and autosomal genetic variation reflect group affinities that were apparent from serological studies. Through analysis of the genetic systems this study will investigate the genetic relatedness of the Khwe to the other Khoe-San populations and to the Bantu-speakers. The amount of admixture from Bantu-speakers into different Khoe-San groups will also be analysed. Furthermore this study will investigate if the genetic systems suggest differences between San and Khoe populations.

The study will also focus on how physical distance between groups influence their genetic relatedness, since cluster analyses of serological studies suggest a strong influence.

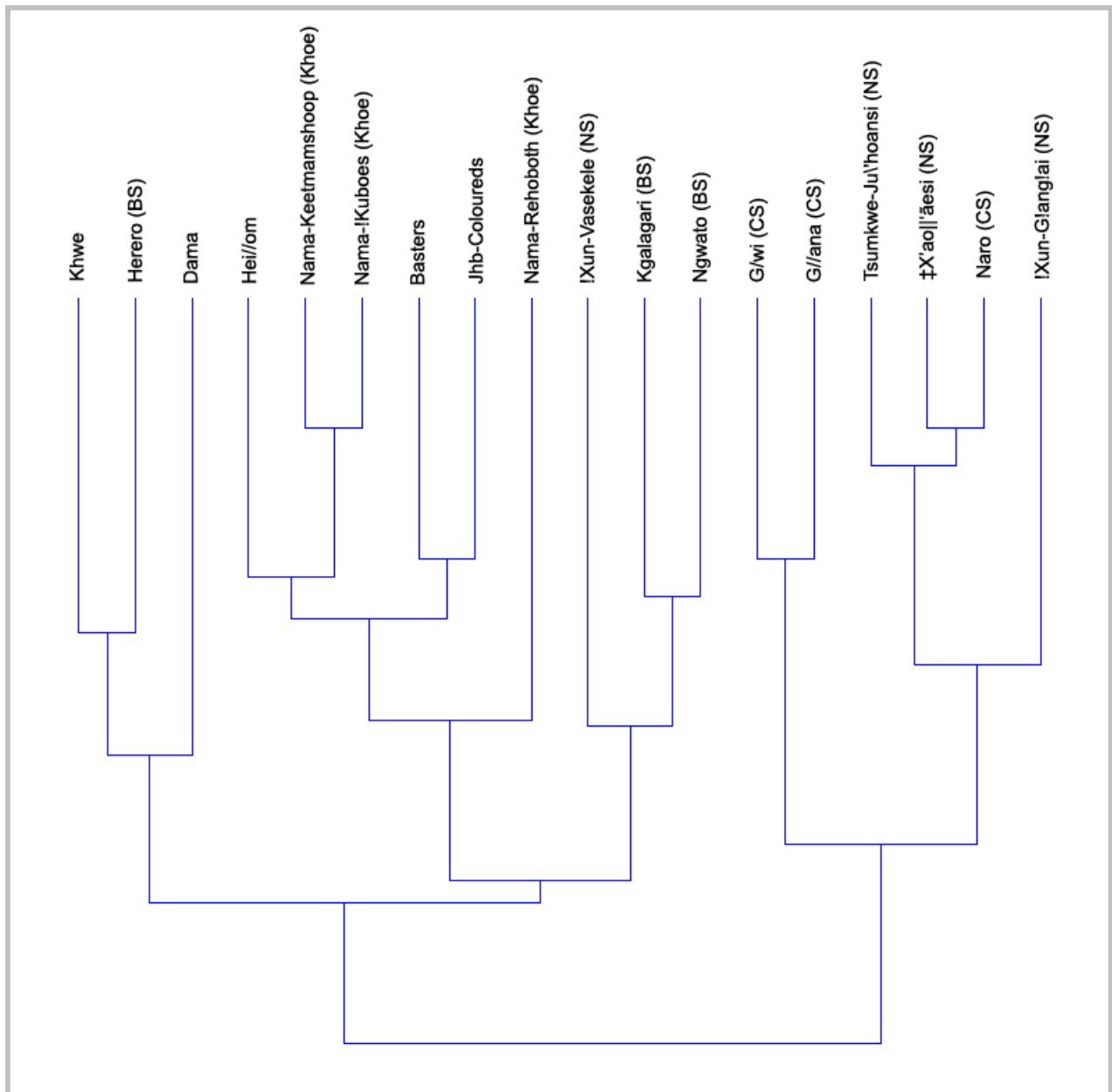


Figure 1.3 A Cluster analysis of distance matrix data from Jenkins (1986).

11 loci in 18 populations. NS – Northern San, CS – Central San, BS – Bantu-speaking

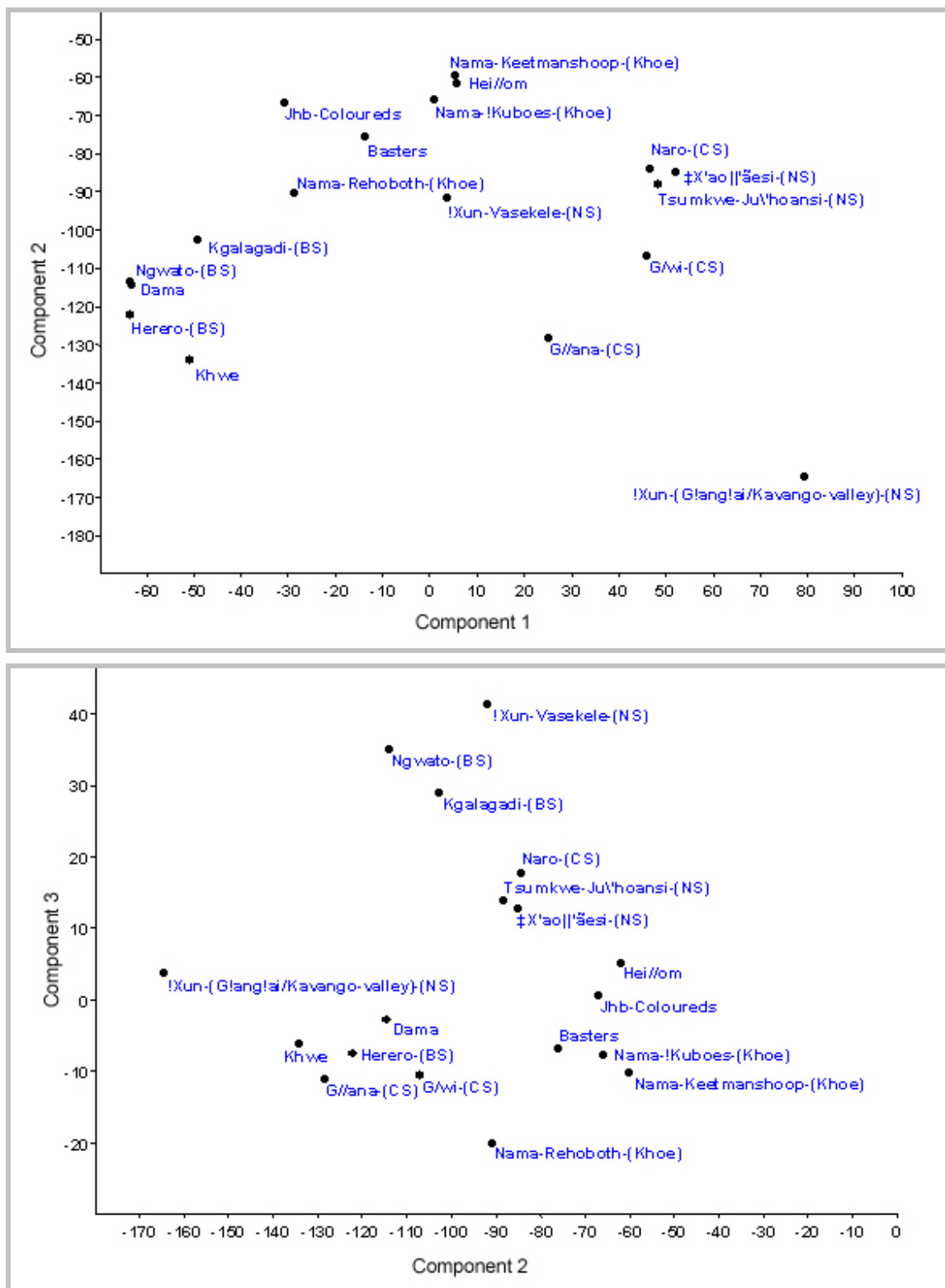


Figure 1.3 B Principal Component Analysis of distance matrix data from Jenkins (1986). 11 loci in 18 populations. (axis 1 = 56.1 % variation, axis 2 = 21.1%, axis 3 = 8.1% variation)

1.2.2.1.3 Commonalities between Hadza, Sandawe and Khoe-San

Jenkins (1982) did correspondence analysis on 23 sub-Saharan populations for 11 serogenetic systems containing 32 alleles (Jenkins, 1982). Correspondence analysis is a multivariate statistical method similar to principal component analysis, except it applies to categorical data rather than continuous data. The analysis revealed that the Hadza group closest to the Babinga Pygmies and the Sandawe closest to east African Bantu-speakers especially the Nyaturu. The Nyaturu are the Bantu-speaking neighbours of the Sandawe with whom they have intermarried frequently. They, however, also show similarities to the Dama on axis 1 of the correspondence plot. On the second axis they are similar to the Keetmanshoop Nama but the presence of malaria protective alleles in the Sandawe separates them on the first axis from the Nama. It might be that there are some genetic similarities between the Khoe and the Sandawe that are masked by the extensive intermarrying by the Nyaturu and the effects of selection on protein coding alleles, which are affected by geography (Jenkins, 1982). It would be interesting to look for genetic similarities between these groups by looking at genetic variation that are not affected by selection through the study of neutral polymorphisms.

As mentioned before the Sandawe have possible linguistic links with the Khoe-San, but not the Hadza. According to Ten Raa (1970) the southern Sandawe groups are also phenotypically similar to the Khoe-San while the central Sandawe groups resemble Bantu-speakers and Nilotes (Ten Raa, 1970). He describes the resemblance to the Khoe-San as follows: “a short stature, a yellow skin, peppercorn hair, the epicanthic fold, excessive wrinkling of the skin at an advanced age, and a typical pentagonal Bushman-like skull: even steatopygia appears to occur in some women”. Phenotypic features are rarely used today but this description contributes to the hypothesis that Khoe-San like hunter gatherer groups existed from mount Kenya to the Cape of Good Hope before the Bantu-expansions (Ten Raa, 1970; Traunmüller, 2003). Current physical anthropological research, however, found no support for morphological similarities between east African and Khoe-San groups (see section 1.2.1.4).

1.2.2.1.4 Khoe-San admixture into other population groups

Immigrant Bantu-speakers had close contact with the indigenous Khoe-San people. This is evident from linguistic borrowings as well as morphological and genetic characteristics. The amount of Khoe-San admixture into various Bantu-speaking groups have been estimated by making use of an immunoglobulin allotype system known as *Gm* (Jenkins *et al.*, 1970). This system contains a specific haplotype that is characteristically and almost exclusively Khoe-San. This specific haplotype was used to determine the amount of Khoe-San admixture into certain Bantu-speaking groups. The group with the highest admixture was the Cape Nguni (Xhosa) population, with frequencies of over 50%. Other southern African Bantu-speaking groups with appreciable frequencies were the Sotho/Tswana people and the other Nguni people. The frequency, however, declines in the more northern groups like the Pedi (14%) and the Tsonga (12%). The Namibia southwestern Bantu-speakers show very low frequencies of admixture. The Kavango group shows no admixture at all while the Herero and Himba show slightly elevated frequencies at about 12%. Studies in other gene marker systems also confirmed these proportions (Jenkins and Corfield, 1972; Jenkins, 1974; Jenkins and Dunn, 1981).

In this thesis, southwestern, southeastern and central African Bantu-speaking groups were included as comparative groups to the Khoe-San populations. Although not the main focus of the thesis, the amount of admixture from Khoe-San groups into these Bantu-speaking groups will also be analysed using the different genetic systems.

1.2.2.2 Mitochondrial DNA studies

Following on the influential paper by Cann *et al.*, concerning the value of mtDNA in reconstructing human origins, mtDNA studies have continued to advance our understanding on historical human migration routes and assessing population affinities (Cann *et al.*, 1987).

Mitochondrial DNA (mtDNA) is located in an extra-nuclear organelle, the mitochondria (a cytoplasmic organelle involved in energy production in eukaryotic cells). The mitochondrial genome is a circular molecule of double-stranded DNA that contains 16 569 basepairs (bp) (Anderson *et al.*, 1981). Each mtDNA contain genes coding for 13 proteins, 22 transfer

RNAs and two ribosomal RNAs (Anderson *et al.*, 1981; Wallace, 1995). Nearly all the non-coding DNA of the mtDNA molecule is contained in a 1.122kb region known as the control region or D-loop (Anderson *et al.*, 1981). This non-coding region has an extremely high mutation rate and is divided into two hypervariable regions, named hypervariable segments I and II (HVS-I and HVS-II). These two regions have been used extensively in phylogenetic studies. Their positions vary between studies but roughly correspond to base pair positions 16024-16400 for HVS-I and 57-372 for HVS-II (Stoneking and Soodyall, 1996; Stoneking, 2000).

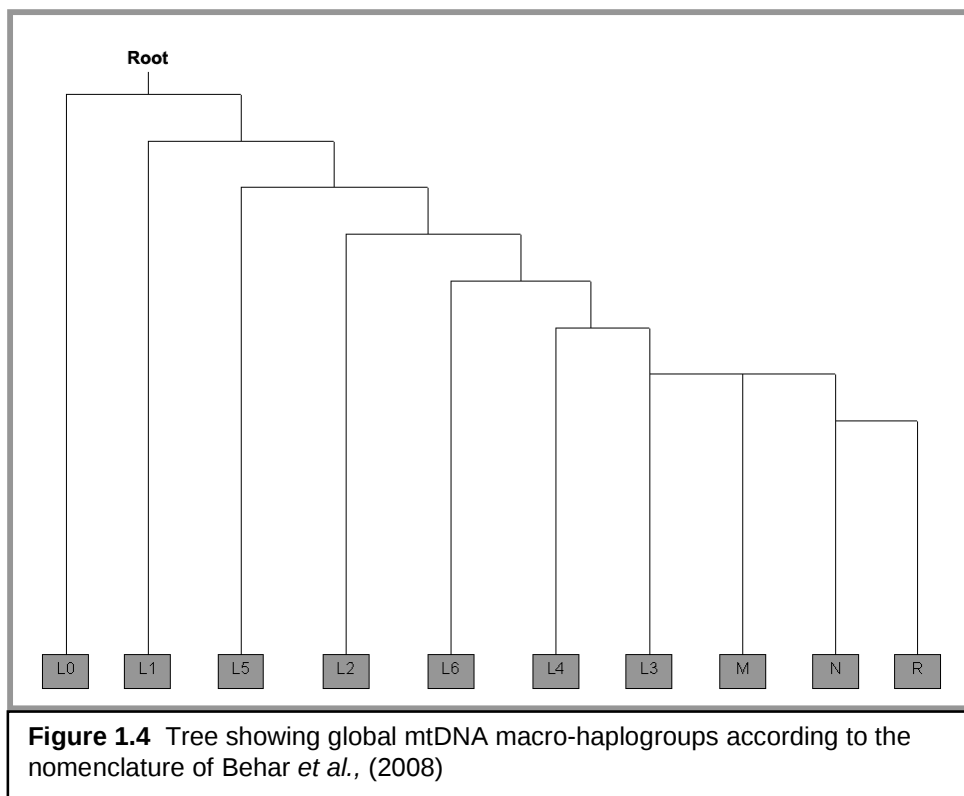
The mtDNA phylogeny has played a central role in locating the human maternal most recent common ancestor (MRCA) to sub-Saharan Africa. It also indicated an initial and modest spread of humans within Africa more than 100 000 years BP, a prominent expansion within Africa 60 000 – 80 000 years BP, leading ultimately single dispersal wave out of Africa that populated the rest of the world (Forster, 2004; Reed and Tishkoff, 2006; Torroni *et al.*, 2006; Behar *et al.*, 2008).

Several factors make mtDNA ideal for phylogenetic analysis over the time scale of modern humans, i.e. the absence of recombination, combined with a high copy number and fast mutation rates. A caveat, however, is that due to the inheritance from mother to child, mtDNA captures the history of the maternal lineage only. Another problem that arises when using only the mtDNA control region is that this part of the mtDNA genome is subject to saturation due to excessive homoplasmy because of the rapid mutation rate. Furthermore the distribution of mutations in the control region is non-random, leading to problematic rate heterogeneity issues when calculating divergence date estimates (Tamura and Nei, 1993; Excoffier and Yang, 1999; Meyer *et al.*, 1999). Furthermore, there is an ongoing discussion on whether human mtDNA evolves neutrally. An assumption behind various population genetic analyses is the selective neutrality of the genetic markers employed. There have been reports on natural selection affecting mtDNA, with temperature being highlighted as a possible selective force (Torroni *et al.*, 2001; Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004). Several other studies, however, concluded that human mtDNA sequence variation has not been significantly influenced by climate (Elson *et al.*, 2004; Kivisild *et al.*, 2006; Amo and Brand, 2007; Ingman and Gyllensten, 2007; Balloux *et al.*, 2009). Despite these

caveats, mtDNA remains by far the most widely used genetic marker in studies of human populations.

Various methods such as high resolution restriction fragment length polymorphisms (Merriwether *et al.*, 1991; Semino *et al.*, 1991; Soodyall and Jenkins, 1992) control region sequencing (Vigilant *et al.*, 1991; Richards *et al.*, 1996; Yao *et al.*, 2003), and a combination of these two methods (Torroni *et al.*, 1996; Torroni *et al.*, 1998; Macaulay *et al.*, 1999) have been used to screen for mtDNA variation. The analysis of whole mtDNA sequence data has reaffirmed the observation deduced from other methods of mtDNA analysis that certain mtDNA polymorphisms show geographical differentiation (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008).

African mtDNA haplogroups are divided into seven macro-haplogroups (L0'1'2'3'4'5'6), while the rest of the worlds' lineages are classified as subgroups of macro-haplogroups M, N and R (Figure 1.4) (Behar *et al.*, 2008).



The first split in the human mtDNA phylogeny is between the two daughter branches, L0 and L1'2'3'4'5'6 (L1-6), located on opposite sides of the root. They split from each other 133 000 – 155 000 years BP (Behar *et al.*, 2008). The archaeological record from this period is too poor to reliably propose hypotheses to for this separation event. Recent studies, however, show that stressful climatic fluctuations known to have occurred throughout the MSA might have caused sporadic settlements of *Homo sapiens* in northwest Africa, the Near East, Chad, and southern Africa (Walter *et al.*, 2000; Henshilwood *et al.*, 2002; Bouzouggar *et al.*, 2007). Today, the L1-6 branch haplogroups are far more widespread while the L0 haplogroups (Figure 1.5) are limited to certain sub-Saharan African population groups. Studies that predate the recognition of L0 as sister to L1-6 suggest that the spread of the haplogroups now labeled as haplogroups within L0 and L1 is the result of an early expansion of modern humans from a location often suggested to be East Africa, to most of the African continent (Maca-Meyer *et al.*, 2001; Mishmar *et al.*, 2003). Traces from this early migration event was partially erased by a vast later expansion wave of L2 and L3 clades dated to 60 000 – 80 000 years BP (Watson *et al.*, 1997; Forster, 2004). Some traces of this early structure, however, still remains among certain hunter–gatherer groups such as the localization of L1c1a to the Pygmy groups of central Africa (Quintana-Murci *et al.*, 2008) and L0d and L0k to the Khoe-San people.

Previous studies reported high frequencies of haplogroups L0d and L0k among Khoe-San groups (Vigilant *et al.*, 1991; Chen *et al.*, 2000; Tishkoff *et al.*, 2007). Haplogroup L0d was found in the !Xun and Khwe at frequencies of 51% and 16%, respectively, while L0k was found at frequencies of 26% in the !Xun and 23% in the Khwe (Chen *et al.*, 2000) (Table 1.2). The same groups were examined by Tishkoff *et al.*, and they reported frequencies of 61% and 22% of L0d and L0k, respectively, in the combined group (Table 1.2) (Tishkoff *et al.*, 2007). In addition, they found L0d at a frequency of 5% in the click speaking Sandawe but not in the Hadzabe population from Tanzania (Tishkoff *et al.*, 2007).

In the Jul'hoansi from Botswana, L0d was found to be the most prevalent haplogroup (96%), while the remaining mtDNA lineages (4%) were resolved into haplogroup L0k (Vigilant *et al.*, 1991) (Table 1.2). L0d and L0k are absent or found at low frequencies in other sub-Saharan African populations. Salas *et al.*, and Perreira *et al.*, respectively,

reported thirteen (out of 307) and eight (out of 109) L0d individuals in Bantu-speaking individuals from Southeastern Africa (Pereira *et al.*, 2001; Salas *et al.*, 2002). There was also a report of an L0d individual in Lake Turkana, Kenya (Watson *et al.*, 1997), one in an African American individual (Allard *et al.*, 2005) and one from Kuwait (Behar *et al.*, 2008).

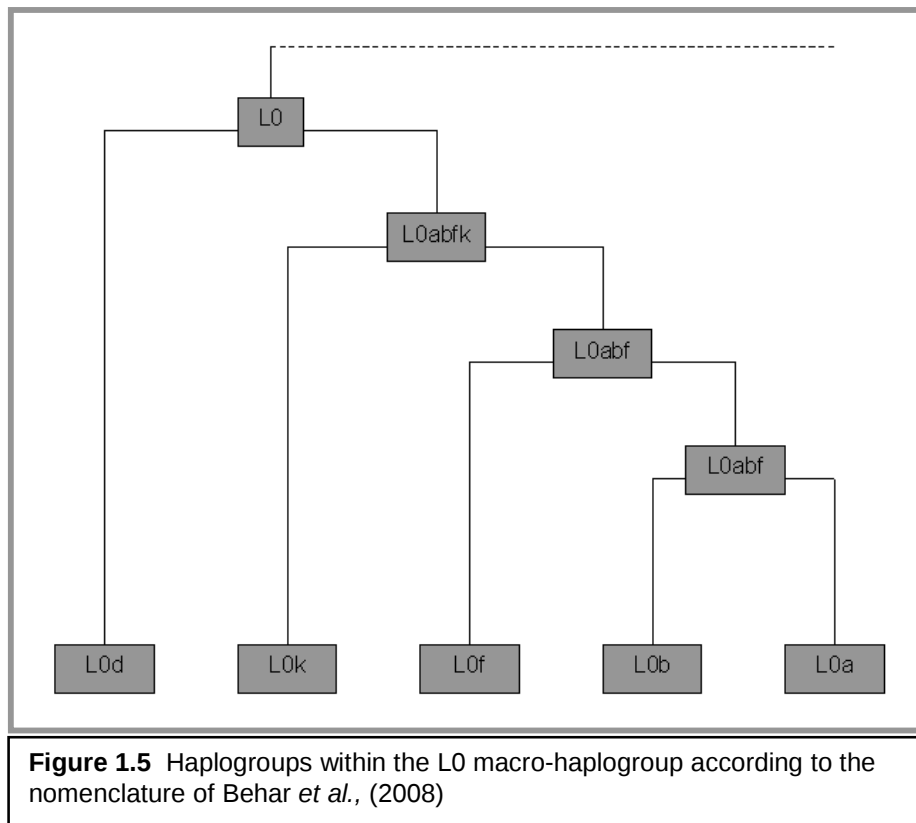


Table 1.2 MtDNA haplogroup frequencies in San populations studied to date

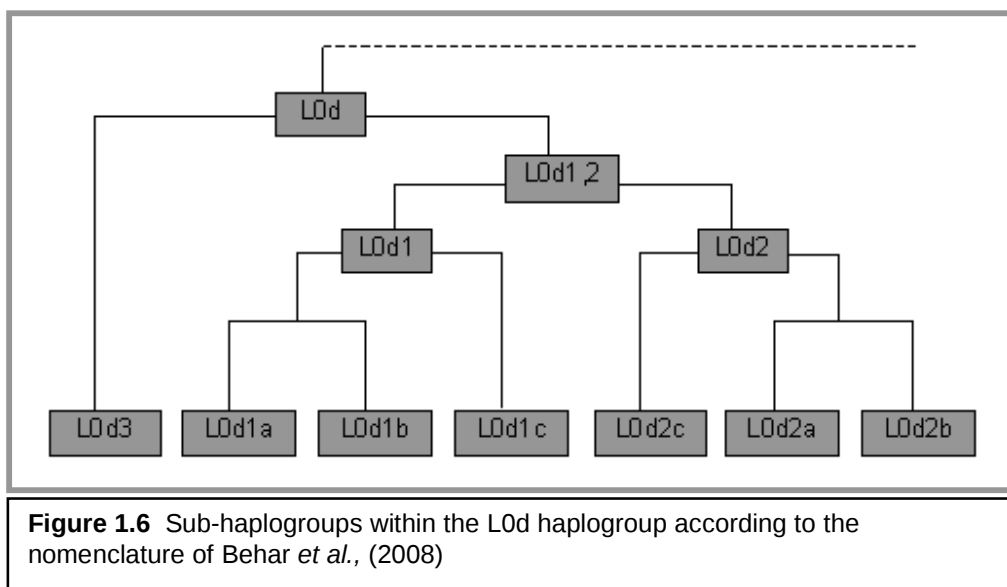
Haplogroup	Haplogroup frequencies			
	Jul'hoansi ¹ (n=24)	!Xun ² (n=43)	Khwe ² (n=31)	!Xun+Khwe ³ (n=18)
L0d	0.958	0.512	0.161	0.611
L0k	0.042	0.256	0.226	0.222
L0a	-	0.023	0.097	-
L1b	-	-	0.032	-
L2*	-	-	-	0.167
L2a	-	-	0.097	-
L2b	-	0.047	0.065	-
L3b	-	0.116	0.032	-
L3e	-	0.047	0.290	-

¹ Vigilant *et al.*, 1991

² Chen *et al.*, 2000

³ Tishkoff *et al.*, 2007

Gonder *et al.*, did whole genome mitochondrial sequencing on selected individuals in the group reported on in Tishkoff *et al.*, (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). They proposed that the L0d clade be split into a San L0d clade and another L0d clade present in Tanzania. Behar *et al.*, have presented an updated phylogeny and nomenclature of mtDNA haplogroups in Africa and have subsequently shown that the L0d haplogroup consists of L0d1, L0d2 and L0d3 and their associated sub-haplogroups (Figure 1.6) (Behar *et al.*, 2008). The Tanzanian L0d sequences fall into a subgroup of L0d3, while San individuals are represented in all seven of the L0d sub-haplogroups (Figure 1.6).



When published data are classified according to the new nomenclature (Table 1.3) it becomes clear that the L0d sub-haplogroups do not have a homogenous distribution across the different San groups and need to be investigated independently.

Table 1.3 Published mtDNA sub-haplogroup frequencies in San populations as fractions of the total number of L0d/k haplotypes in the sample group

Sub-Haplogroup	Haplogroup frequencies				
	Jul'hoansi ¹ (n=24)	!Xun ² (n=33)	Khwe ² (n=12)	!Xun+Khwe ³ (n=15)	Bantu-speakers ^{4,5} (n=21)
L0d1a	0.083	0.091	-	0.067	-
L0d1b	0.708	0.030	-	-	0.190
L0d1c	-	0.485	-	0.467	0.095
L0d2a	0.042	-	-	0.067	0.143
L0d2b	-	-	-	-	0.095
L0d2c	0.125	-	-	0.067	0.238
L0d2d	-	-	-	-	0.095
L0d3	-	-	-	-	0.048
L0dx	-	0.061	0.417	0.067	0.048
Unclassified	-	-	-	-	0.048
L0k1	0.042	0.333	0.583	0.267	-

¹ Vigilant *et al.*, 1991

² Chen *et al.*, 2000

³ Tishkoff *et al.*, 2007

⁴ Salas *et al.*, 2002

⁵ Perreira *et al.*, 2001

Behar *et al.*, proposed two hypotheses to explain how the two ancient lineages L0d and L0k became largely localized to the Khoe-San groups of southern Africa where they remained isolated from other haplogroups, for an extremely long period of between 50 000 and 100 000 years until the development of LSA technologies (Behar *et al.*, 2008).

In the first hypothesis an initial prolonged diffusion of anatomically modern humans from east Africa (200 000 - 100 000 years BP) is followed by a dispersal wave (~100 000 years BP) of a part of the population and the localization of L0d and L0k to southern Africa.

In the second hypothesis an early division in the human population (~200 000 years BP) resulted in the localization of L0 in southern Africa and L1-6 in eastern Africa. The eastern and southern populations continued to evolve separately until a dispersal event (~150 000 - 100 000 years BP) of the L0abf group from the southern population and its merger with the eastern population. This resulted in the southern population composed only of L0d and L0k and the eastern composed of L1-6 and L0abf.

Information about ancestral population sizes and population growth parameters can be used to infer past population demographics and expansion events. Various methods exist

to investigate such events. Star like patterns in median networks is an indication of recent expansions, Watson *et al.*, used median networks to identify star-like patterns in L2 and L3 RFLP and D-loop sequence diversity, which suggested an expansion 60 000 – 80 000 years BP (Watson *et al.*, 1997).

Coalescent theory is another method that can be used to gain information about ancestral population size and growth parameters using extant genetic variation (Kingman, 1982; Hudson, 1990; Griffiths and Tavaré, 1994). Coalescent theory was implemented to predict population expansions in the form of mismatch distributions using sequence data (Rogers and Harpending, 1992) and by testing the validity of summary statistics that predict expansions (Schneider and Excoffier, 1999; Rozas *et al.*, 2003). Coalescent analysis of past African population demographics using RFLP and mtDNA sequence data (Harpending *et al.*, 1993; Sherry *et al.*, 1994; Excoffier and Schneider, 1999; Harpending and Rogers, 2000; Pilkington *et al.*, 2008) revealed that most human populations show significant signs of expansion around 70 000 years BP. An exception, however, were that some hunter-gatherer populations from different continents (including San and Pygmy populations from Africa) did not show these expansion signals. The lack of expansion signal around this time in hunter-gatherers was proposed to be due to post-Neolithic population bottlenecks that led to the loss of previous expansion signals (Excoffier and Schneider, 1999). In this theory, populations that did not go through the Neolithic transition, experienced reduction of effective population sizes because of competing Neolithic farmers that caused fragmentation of the hunter-gatherer habitat (Excoffier and Schneider, 1999).

The coalescence analysis employed in mismatch distributions assumes a single exponentially growing population and contained large degrees of statistical uncertainty. Also, by applying these methods earlier population expansions can be obscured by recent population bottlenecks (Excoffier and Schneider, 1999). Recent improvements in coalescence inference methods led to increased accuracy, without the need to assume a single exponential growth curve (Shapiro *et al.*, 2004; Atkinson *et al.*, 2008). The Bayesian skyline plot (BSP) is a useful coalescence procedure that uses Bayesian inference to infer effective population sizes through time (Drummond *et al.*, 2005; Drummond and Rambaut, 2007).

Atkinson *et al.*, constructed BSP for the four most common African mtDNA macro-haplogroups, L0, L1, L2 and L3 using whole mitochondrial genomes (Atkinson *et al.*, 2009). The four haplogroups revealed very different patterns of growth. The patterns of haplogroups L2 and L3 were significantly different from each other and from haplogroups L0 and L1. Both haplogroups showed signals of rapid expansions albeit at different times. While L2 occurred at relatively low frequencies until a sudden period of fast growth beginning 12 000 – 20 000 years BP, L3 had a rapid expansion phase from its time to most recent common ancestor (TMRCA), 61 000 – 86 000 years BP onwards (Atkinson *et al.*, 2009).

Haplogroups L0 and L1 on the contrary showed slow constant growth over the last 100 000 – 200 000 years BP. Growth patterns throughout the history of these two lineages are not significantly different from each other. This would be expected if both lineages formed part of an early panmictic African population that contributed equally to the current African mtDNA diversity. Various studies, however, support population structure deep in the mtDNA tree based on the localization of L0d and L0k to the Khoe-San speakers of southern Africa (Knight *et al.*, 2003; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). The BSP analysis of Atkinson *et al.*, showed that, if such deep population structure exists, it did not generate considerably different population growth profiles between these early lineages (Atkinson *et al.*, 2009). A BSP for L0d and L0k lineages only compared to the L0 and L1 growth profile also was not significantly different. This indicates that these suggested deep divergence events were probably not connected to substantial changes in available territory or mode of living (Atkinson *et al.*, 2009).

This thesis will use mitochondrial DNA variation from the different included population groups to infer mitochondrial DNA phylogenies and networks. These will be used to group the mtDNA genomes into haplogroups and to compare the haplogroup profiles of the different populations in the study group. Furthermore, different coalescence methods will be employed to look for population expansions and contractions in the different haplogroups as well as in the different population groups.

1.2.2.3 Y-chromosome studies

Similar to the mitochondrial genome, large parts of the Y-chromosome is haplotypic and does not undergo recombination. The Y-chromosome contains the largest non-recombining block in the human genome and is therefore extremely important for evolutionary genetic studies. While studies on mtDNA describe the maternal history of a population, the paternal history can be described through using the Y-chromosome. The first Y-chromosome polymorphism was reported in 1985 (Casanova *et al.*, 1985) but more than a decade elapsed before a well-resolved Y-chromosome tree was available (Underhill *et al.*, 2000; Hammer *et al.*, 2001a; YCC, 2002).

Y-chromosome tree structure

The Y-chromosome tree structure is primarily based on binary polymorphisms, and specific branches are assigned to haplogroups following a hierarchical pattern. Currently the Y-Chromosome tree consists of 20 major clades (Figure 1.7) containing 311 distinct haplogroups defined by 599 mutational events (Karafet *et al.*, 2008). Furthermore, by typing Y-chromosome short tandem repeats (Y-STRs) haplotypes are generated, which are then used for finer resolution within the haplogroups (Underhill and Kivisild, 2007).

The two primary splits in the Y-chromosome tree leads to two branches, Haplogroup A and B, which have a distribution restricted to Africa (Figure 1.7). These two clades are genetically diverse and its haplogroups have different geographical distribution patterns. This suggests population fragmentation, isolation and re-expansions in pre-historic Africa. Haplogroups A and B are associated with the distribution of ancient hunter-gatherer tribes before the expansions of pastoralists (Underhill *et al.*, 2001; Underhill and Kivisild, 2007). The rest of the Y-chromosome tree is defined by the M168 mutation, which represents the most common African lineages (Haplogroup E) as well as all the non-African clades (Figure 1.7).

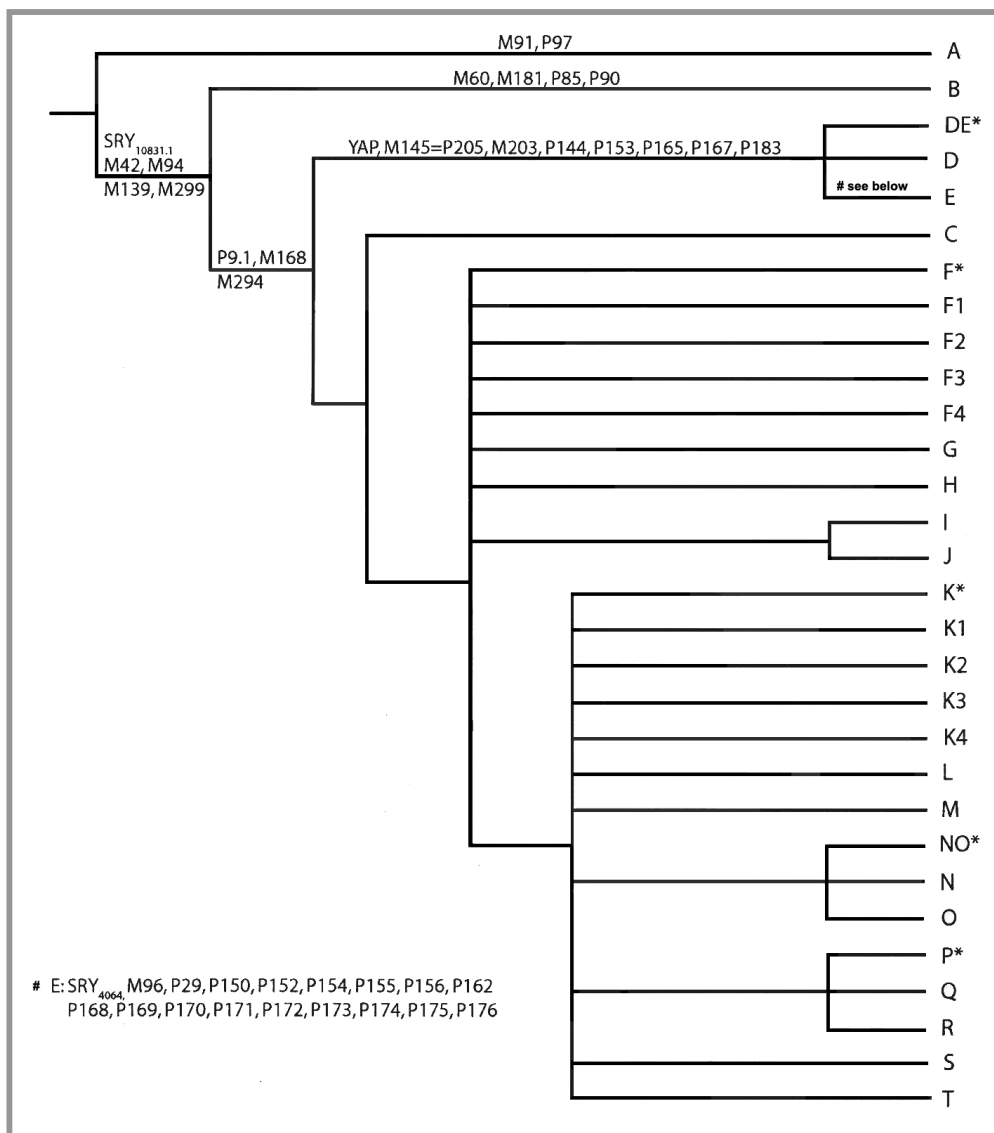
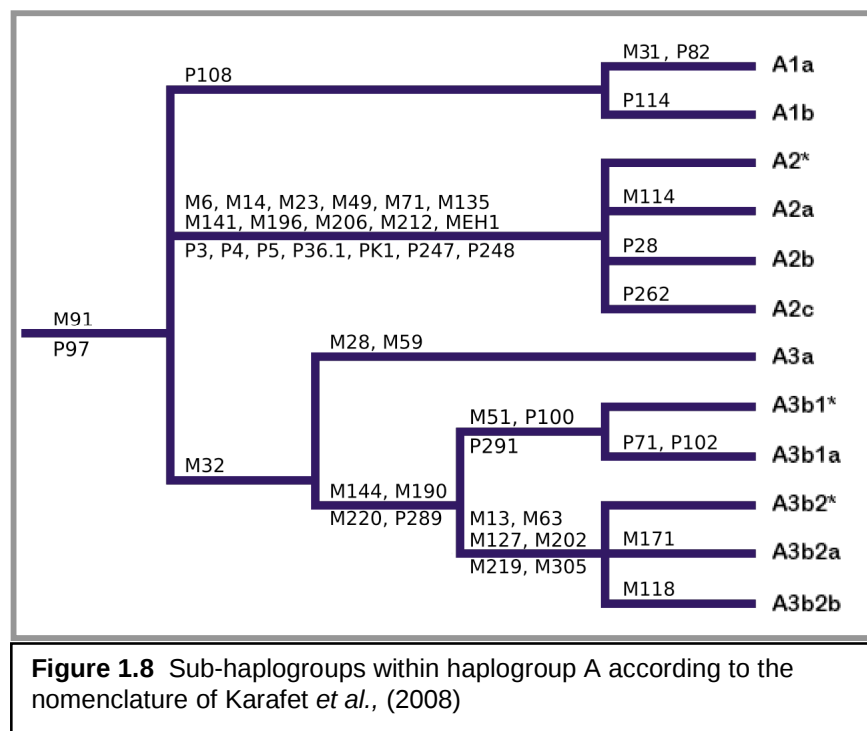


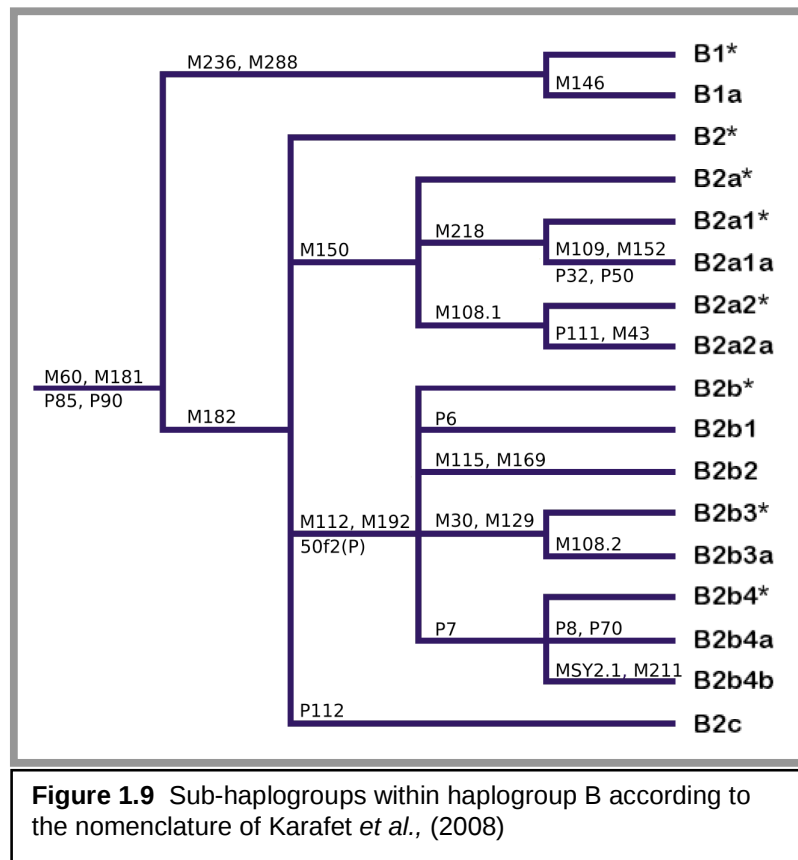
Figure 1.7 Tree showing global Y-chromosome macro-haplogroups according to the nomenclature of Karafet *et al.*, (2008)

Haplogroup A is defined by the M91 and P97 mutations and contains 12 branches determined by 45 (internal) mutations (Figure 1.8). A strict regional distribution is particularly pronounced for haplogroup A. Within Haplogroup A; A1 is found in Mali and Morocco (Underhill *et al.*, 2000; Scozzari *et al.*, 2001), A3b2 is found in east Africa (Sudan, Ethiopia, Tanzania, Kenya) and in lower frequencies in north Cameroon (Scozzari *et al.*, 1999; Underhill *et al.*, 2000; Cruciani *et al.*, 2002; Semino *et al.*, 2002; Knight *et al.*, 2003), while A3b1 and A2 is found exclusively among the Khoe-San (Scozzari *et al.*, 1999; Underhill *et al.*, 2000).



Haplogroup B is defined by four mutations (M60, M181, P85, and P90) and contains 17 branches with 28 internal markers (Figure 1.9) (Karafet *et al.*, 2008). Haplogroup B occur throughout Africa but have high frequencies among Pygmies, Khoe-San and Hadza, with some lineages being restricted to them (Underhill *et al.*, 2000; Cruciani *et al.*, 2002; Semino *et al.*, 2002; YCC, 2002; Knight *et al.*, 2003). There is a clear-cut difference between the B haplogroups associated with the Pygmies, Khoe-San and Hadza vs. all the other African populations. Pygmies, Khoe-San and Hadza populations have mainly Haplogroup B

haplotypes defined by the M112 mutation, while other populations have the M150 mutation. Within haplogroup B-M112, haplogroups B2b2, B2b3 and B2b4b are restricted to the Pygmy populations while B2b1 (P6) and B2b4a (P8) are restricted to Khoe-San groups (Hadza groups were not typed for B-M112 sub-groups). The B2b* ancestral haplotype occurs in both Pygmy and Khoe-San groups (Underhill *et al.*, 2000; Cruciani *et al.*, 2002; Semino *et al.*, 2002; YCC, 2002; Knight *et al.*, 2003).



Eighteen mutations currently define haplogroup E. Haplogroup E is the most mutationally diverse of all the major Y-chromosome clades and contains 83 polymorphisms that define 56 distinct haplogroups (Figure 1.10) (Karafet *et al.*, 2008). The E haplogroups are found at high frequencies in Africa, at moderate frequencies in the Middle East and southern Europe, and has sporadic occurrences in Central and South Asia. Although Haplogroup E groups are widespread all over Africa, the distributions of the numerous distinctive haplogroups are not homogeneous across the continent (Hammer and Horai, 1995;

Hammer *et al.*, 1997; Qamar *et al.*, 1999; Bosch *et al.*, 2001; Hammer *et al.*, 2001a; Underhill *et al.*, 2001; Cruciani *et al.*, 2002; Cruciani *et al.*, 2004).

Haplotypes carrying the mutations M75 (E2) and M33 (E1a) are present at low frequencies across Africa but with different individual distributions. Haplogroups E1b1a and E1b1b is the most frequent and widespread of the E haplogroups (Hammer *et al.*, 2001a; Underhill *et al.*, 2001; Cruciani *et al.*, 2002; Cruciani *et al.*, 2004; Semino *et al.*, 2004).

E1b1a, defined by M2 and seven other mutations is mainly limited to sub-Saharan populations and is associated with the expansion of Bantu-speaking populations (Hammer *et al.*, 1998; Passarino *et al.*, 1998; Scozzari *et al.*, 1999). The E1b1a subgroups have differential distributions and frequencies. The M191 mutation defines the most frequent E-M2 subgroup and is evident of a founder effect that resulted from the Bantu-expansions (Hammer *et al.*, 2001a; Underhill *et al.*, 2001; Cruciani *et al.*, 2002; Cruciani *et al.*, 2004; Semino *et al.*, 2004).

The non-African distribution of haplogroup E is associated with haplogroup E1b1b characterized by the M35 and M215 mutations (Hammer *et al.*, 1998; Semino *et al.*, 2000; Underhill *et al.*, 2001; Semino *et al.*, 2004). This haplogroup, however, also have a widespread African representation (Hammer *et al.*, 2001a; Underhill *et al.*, 2001; Cruciani *et al.*, 2002; Cruciani *et al.*, 2004; Semino *et al.*, 2004). Compared to other E haplogroups, M35 occur at very low frequencies within Bantu speakers but is widely though not uniformly dispersed throughout Africa. Among the different lineages carrying the M35 mutation, haplotypes defined by M78 occurs in east Africa, north Africa, the Middle East and Europe. It is the E-M35 subgroup with the highest frequency and the widest distribution outside Africa. This marker has a northeastern African origin and multiple exodus routes out of Africa have been demonstrated (Cruciani *et al.*, 2007). M123 haplotypes are present in eastern Africa, northeast Africa, the Middle East and southeast Europe but does not reach western Europe. M81 is found at high frequencies only in northern Africa and is almost absent in Europe (with the exception of Sicily and Iberia) (Bosch *et al.*, 2001; Cruciani *et al.*, 2004; Semino *et al.*, 2004). In addition to these differentiated E1b1b lineages there were many haplotypes that were classified as E-M35*, which occurred in high frequencies

particularly in Ethiopian, Kenyan, Tanzanian and Khoe-San groups (Cruciani *et al.*, 2004; Semino *et al.*, 2004).

Recently a new Y-chromosome polymorphism (M293) was discovered, which grouped these previously paraphyletic E-M35 groups into a monophyletic group (Henn *et al.*, 2008). This E-M293 haplogroup has a spread concentrated in eastern and southern Africa with maximum frequencies in Tanzania and southern Africa. In eastern Africa high frequencies of M293 is observed in the Datog (43%), Burunge (28%), Sandawe (24%) and Hadza (11%). The Datog are pastoralists who speak a Southern Nilotic language and the Burunge are Afro-Asiatic agropastoralists. In southern Africa it was observed in the Khwe (31%) and !Xun (11%). The Khwe and the !Xun were the only Khoe-San groups included into the study. Network analysis revealed haplotype sharing and close similarities between Khwe/!Xun haplotypes and Hadza/Sandawe haplotypes. M293 also occurs at low to moderate frequencies in Bantu-speaking populations of eastern and southern Africa, which likely reflects recent admixture with local populations after the Bantu-expansions (Henn *et al.*, 2008).

E-M293 data from the study supported a demic diffusion model correlated with the spread of sheep, cows and pottery along a Tsetse fly free corridor between eastern and southern Africa, 2 000 years BP (Sadr, 1998; Gifford-Gonzalez, 2000; Smith, 2005; Henn *et al.*, 2008). A previous model where pastoralism was transmitted from eastern Africa to southern-central Africa with little to no population movement was thus rejected (Sadr, 1998; Smith, 2005). The new model suggested that a small pastoralist population carrying M293 migrated from east Africa into southern-central Africa with their livestock (Henn *et al.*, 2008). After arriving in southern Africa, these pastoralists could have mixed with local populations, or expanded without substantial genetic exchange with local groups. Without representation of more Khoe-San groups the study, however, could not address the question of how pastoralism spread after it reached south-central Africa. The scale of the migration from east Africa may have been small, minimally four E-M293 male individuals. It is possible that other male individuals who did not carry M293 were also involved. For instance E-M2 individuals could have been involved but it would not be possible to distinguish these from the E-M2 introduced later during the Bantu-expansions. The Henn *et*

al., study thus supports a migration, independent of the Bantu expansion, of east Africans harbouring the E-M293 marker, which initially brought pastoralism to southern Africa (Henn *et al.*, 2008).

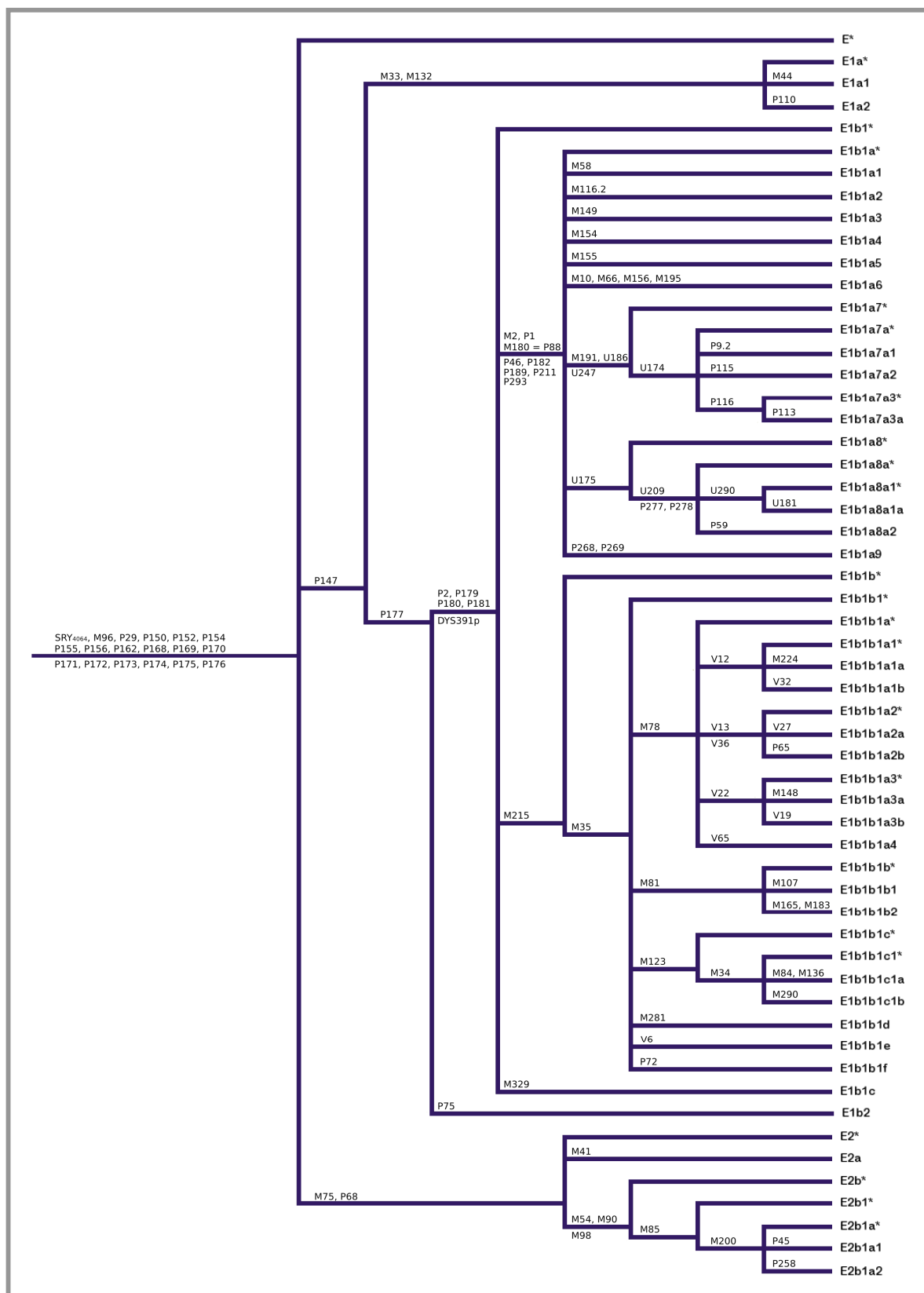


Figure 1.10 Sub-haplogroups within haplogroup E according to the nomenclature of Karafet *et al.*, (2008)

The age of the Y-chromosome tree

The TMRCA of the human Y-chromosome tree has been determined by using microsatellites (Wilson and Balding, 1998; Pritchard *et al.*, 1999) and by sequencing parts of the Y-chromosome (Thomson *et al.*, 2000). Both the sequence based and microsatellite based studies supported a model of exponential growth for the Y-chromosome and also found substantial continental structure in the data. The microsatellite studies estimated a TMRCA between 46 000 and 91 000 years, depending on the different mutation models used. This is a very young date compared to what is predicted by the TMRCA of mitochondria, X-chromosome and autosomes (Horai, 1995; Harding *et al.*, 1997; Harris and Hey, 1999; Kaessmann *et al.*, 1999). The mutation mechanisms of microsatellites are not understood very well and this might lead to dating errors. This young date was, however, confirmed by sequencing three genes in the Y-chromosome. The sequence data was analyzed using a coalescent approach and yielded a TMRCA of 59 000 years (Thomson *et al.*, 2000).

Reasons for this young date might be that the ancestral population was very small or that the Y-chromosome is subject to strong selection. The strong selection might be in the form of positive selection for advantageous mutations (hitchhiking) or negative selection against deleterious mutations (background selection) (Thomson *et al.*, 2000). A low effective population (N_e), which will lead to a younger date, could also have been caused by higher variance in male reproductive success (Wilder *et al.*, 2004).

Y-chromosome studies in the Khoe-San

Published studies on the Khoe-San people covered three sample collections, the Platfontein Khwe and !Xun (Scozzari *et al.*, 1997), a mixed group of Ju!'hoansi and !Xun (Underhill *et al.*, 2000) and a mixed group of Ju!'hoansi, !Xun, Khwe, Nama and Dama (Hammer *et al.*, 1997; Wood *et al.*, 2005). The group of Khwe and !Xun was originally reported in (Scozzari *et al.*, 1997) and reanalyzed or included in various other studies (Scozzari *et al.*, 1999; Cruciani *et al.*, 2002; Jobling and Tyler-Smith, 2003; Knight *et al.*, 2003; Cruciani *et al.*, 2004; Henn *et al.*, 2008). The mixed group of Ju!'hoansi and !Xun was originally reported in (Underhill *et al.*, 2000) and were also subsequently included in various comparative studies (Underhill *et al.*, 2001; Semino *et al.*, 2002; Knight *et al.*,

2003). The mixed Khoe-San group reported in Wood *et al.*, included Khoe-San groups reported on in Hammer *et al.*, (Hammer *et al.*, 1997; Hammer *et al.*, 2001a; Wood *et al.*, 2005). Mainly the African haplogroups (A, B and E) have been found in varying frequencies in the above-mentioned published Khoe-San groups (Table 1.4). These studies revealed that:

- the Khoe-San people carry high frequencies of the most ancient lineages on the Y-chromosome tree (Haplogroup A and B),
- some of these lineages are exclusive to the Khoe-San (A-M51 and A-M23 derived lineages),
- other lineages in these ancient clades (A and B) was also identified in high frequencies in other populations with recent hunter-gatherer ancestry (such as the Pygmy populations and the Hadza and Sandawe of east Africa)
- Khoe-San populations have varying frequencies of Bantu-speaking associated haplogroups and the more isolated populations (such as the Ju'hoansi) have lower frequencies of these haplogroups.
- The Khwe population is different from the other San populations in that they have lower frequencies of haplogroups A and B, higher frequencies of the Bantu-associated E haplogroups and higher frequencies of the E-M35 haplogroup.

Both Y-chromosome binary polymorphisms and Y-chromosome microsatellites will be used in this thesis to assign the Y-chromosomes from the various population groups to haplogroups. Y-chromosome haplogroup profiles will thereafter be compared between the different Khoe-San groups to assess their relatedness. Furthermore haplogroup profiles will be compared to neighboring groups thereby investigating amount of admixture. Additionally, the published high frequencies of specific sub-groups of haplogroup A and B will be evaluated against frequencies seen in Khoe-San groups from this study. The prevalence and spread of the E-M35* haplogroup will also be examined to try to infer the spread of pastoralism.

Table 1.4 Y-chromosome haplogroup frequencies (%) of Khoe-San populations studied to date

Haplogroup	Haplogroup frequencies (%)			
	!Xun ¹ (n=64)	Khwe ¹ (n=26)	Jul'hoansi and !Xun ² (n=39)	Mixed Khoe-San ³ (n=90)
A-M51	28	12	28	22
A-M14	5	-	13	14
A-M114	3	-	3	-
A-P28	-	-	-	11
B-M182	-	-	-	1
B-M112	8	-	28	13
				(P6 = 9 and P7 = 4)
E*-SRY ₄₀₆₄	-	-	-	1
E-M75	-	-	-	1
E-M54	-	-	-	1
E-M85	6	4	-	-
E-M2	23	50	18	14
E-M191	16	-	-	10
E-M154	-	4	-	-
E-M35	11	31	10	7
J-12f2	-	-	-	1
R-M343	-	-	-	2

¹ Scozzari *et al.*, (1997)² Underhill *et al.*, (2000)³ Wood *et al.*, (2005)

Y-chromosome and mtDNA comparative studies

The haploid nature of the mtDNA and Y-chromosome allow us to study the history maternal and paternal lineages separately because of their unilateral transmission. It further allows us to compare their dynamics and deduce female vs. male migration rates and effective population sizes.

Wood *et al.*, investigated the effects of male vs. female gene flow in various African populations (Wood *et al.*, 2005). Mantel tests and AMOVA analysis found strong correlations between Y-chromosome genetic distance and linguistic distance, but no correlation between Y-chromosome genetic distance and geographic distance. Conversely the mitochondrial genetic distances between populations showed weak correlations with both geographic distance as well as linguistic distance. When Bantu speakers were removed, however, the correlation with linguistic variation disappears for the Y-chromosome and strengthens for mtDNA (Wood *et al.*, 2005).

From this it is clear that patterns between different populations vary. Differences in mtDNA and Y-chromosome gene-flow can be extrapolated to sociocultural practices in the populations involved. Seielstad *et al.*, inferred patrilocality in African populations based on the fact that inter-population variation was much higher based on Y-chromosome variation than mtDNA variation (Seielstad *et al.*, 1998). From this higher female than male migration rates were calculated. A study by Hammer *et al.*, however, found contrasting results (Hammer *et al.*, 2001a). The gene flow in this study was male biased and supported a greater mobility of male individuals that led to lower inter-population Y-chromosome distances than mtDNA distances. The discrepancy between the two studies was explained by the fact that the Seielstad *et al.*, study only considered food-producing populations while the study of Hammer *et al.*, included hunter-gatherer populations (Destro-Bisol *et al.*, 2004). In a comparative study between food producers and hunter-gatherers a marked heterogeneity in terms of distribution of the unilaterally transmitted markers was found (Destro-Bisol *et al.*, 2004). While in food producers the gene flow was female biased because of patrilocality, hunter-gatherer populations had a male biased gene-flow. The male biased gene-flow in hunter-gatherers was explained as a combined effect of asymmetric gene flow between the food producers and hunter-gatherers as well as different levels of polygyny and patrilocality between the two groups (Destro-Bisol *et al.*, 2004).

Wood and colleagues (2005) also investigated the paternal and maternal signatures in food-producers and hunter-gatherers by sequencing parts of the mitochondrial genome and Y-chromosome. The resultant data also supported dissimilar male and female histories and differences in hunter-gatherers and food-producers. For mitochondrial data the food producers fit a model of population expansion and the hunter-gatherers a model of population stationarity, while for the Y-chromosome both populations best fit a model of constant population size. The reasons proposed for the dissimilar Y-chromosomal and mtDNA results, were that food-producers in the past had a smaller effective population sizes (N_e) and lower migration rates (m) than hunter-gatherers. Cultural practices that lead to a lower $N_e m$ are polygyny and patrilocality (Wood *et al.*, 2005).

Polygyny leads to variance in reproductive success between males, which lower their N_e relative to females (Low, 1988; Wilder *et al.*, 2004). Generally food-producers are described as more polygynous than hunter-gatherers (Cavalli-Sforza, 1986; Bieseke and Royal, 1999). Additionally, patrilocality can result in lower rates of male migration (Murdock, 1981). Most agricultural societies are patrilocal (Murdock, 1967), but hunter-gatherer groups are bilocal, (spending time living with both the male's and the female's families (Marlowe, 2004)). These processes would have changed and shifted as populations converted from foraging to food-producing lifestyles. This may have played an important role in the distinctive patterns observed for mtDNA and the Y-chromosome.

In the present thesis correlations between physical geographic distances and genetic distances will be done for lineages representing male lines (Y-chromosomes) as well as lineages representing the female lines (mtDNA). Positive correlations between Y-chromosome genetic distances and physical distances are expected if the gene-flow is male biased, as was seen previously in food producers (Seielstad *et al.*, 1998; Destro-Bisol *et al.*, 2004). On the contrary, if gene-flow is female biased, we expect to see correlations between physical geographic distance and mtDNA genetic distance, as was previously seen in hunter-gatherer societies (Hammer *et al.*, 2001a; Destro-Bisol *et al.*, 2004).

1.2.2.4 Autosomal DNA studies

Compared to Y-chromosome and mtDNA phylogenetic studies, studies on the autosomes are complicated because of recombination. This problem can be partly overcome by studying short stretches of linked polymorphisms and inferring haplotypes. The inference of haplotypes has been made easy by the development of various algorithms that use homozygous group frequencies to infer the phase of heterozygous loci (Excoffier and Slatkin, 1995; Stephens *et al.*, 2001; Niu *et al.*, 2002; Scheet and Stephens, 2006). Consequently these short stretches of inferred haplotypes can be treated as lineages in the same way that the non-recombining mtDNA and Y-chromosome DNA are treated. An early example of a autosomal haplotype study is the 2.7 kb region on chromosome 11 that encompass the β -globin gene (Harding *et al.*, 1997). The phylogeny obtained from the 326 haplotypes reflected results from Y-chromosome and mitochondrial studies. The root of the tree was in Africa with many lineages that were exclusive to Africa. Since then several

other loci have been studied, all supporting an African root (Clark *et al.*, 1998; Harris and Hey, 1999; Harding *et al.*, 2000). Similar to mtDNA and Y-chromosome studies, however, the history and dynamics of the lineage under investigation is the history coupled to a certain locus and is reflective of only a small part of the genome. Some of these loci might be heavily influenced by selection, which would violate the assumptions of population genetic models and in the end would not give a true picture of the population history. Ultimately to get the true history of a population or the human species, one should take into account all of the separate loci.

Another way to utilize information contained in the autosomes is to use genotypes of unlinked markers spread over the whole genome, instead of inferred haplotypes. Through using AMOVA analysis on such multilocus genotypes (microsatellites, single nucleotide polymorphisms (SNPs) and insertion/deletions) it was found that 79-94% (depending on the marker type) of variation represents variation between individuals within the same population (Barbujani *et al.*, 1997; Jorde *et al.*, 2000; Romualdi *et al.*, 2002; Rosenberg *et al.*, 2002). This thus means that genotypic variation is not homogenous across the human species but 21-6% of the variation is due to differences between populations and continental groups. This led to the question of whether a genotype from an individual can be correctly assigned to the correct population or continent of origin.

The earliest method that explored this question was by calculating pairwise individual distances based on allele sharing (Bowcock *et al.*, 1994). These distances were then used to construct a tree of genotypes from individuals without taking into account any prior of population origin. The aim was to see if the tree shows clusters according to populations or continents. The tree that resulted correctly assigned 88% of genotypes to continent specific clusters. The population specificity was less precise but 64% of populations formed clusters that included more than half of their individuals (Bowcock *et al.*, 1994). Since then more powerful genotype assigning methods have been developed (Pritchard *et al.*, 2000; Corander *et al.*, 2003; Falush *et al.*, 2003; Francois *et al.*, 2006; Falush *et al.*, 2007).

A widely used technique, implemented in the program STRUCTURE, is based on the Bayesian clustering of individuals into K number of clusters (Pritchard *et al.*, 2000; Falush

et al., 2003; Falush *et al.*, 2007). The user specifies the K number of clusters and the program assign a genotype or a proportion of a genotype to a certain cluster with a certain probability. A signature of population structure will then emerge (if there is structure) through the unequal assignment of individuals or partial genotypes to certain clusters. For instance: If K=2, the program will divide the total variation of the whole study group optimally into two clusters and then assign each individual with a certain probability to each of the two clusters. When there is structure in the sample group, individuals from population x will be preferentially assigned to a certain cluster, for instance cluster 1, while individuals of population y, will be preferentially assigned to cluster 2. If, for instance individuals from population z resulted from a admixture event between population x and y, these individuals will be assigned with certain probabilities to both clusters 1 and 2 depending on the marker contribution from each population into the individual. When an admixture model is assumed, individuals are not assigned to a cluster with a certain probability; rather a part of their genome (made up by the markers included in the study) is assigned to a certain cluster. The procedure usually followed when running STRUCTURE is to assign K clusters from K=2 to K=10 and then test which K number of clusters has the highest likelihood by looking at the posterior likelihood scores or by using the deltaK method that takes into account the rate of change between successive K clusters (Evanno *et al.*, 2005).

The first genotypic studies were based on limited number of markers and individuals (Bowcock *et al.*, 1987; Nei and Livshits, 1989; Bowcock *et al.*, 1991a; Bowcock *et al.*, 1991b). RFLPs were individually typed from isolated DNA, which were cloned or transformed to increase the quantity. These laborious processes limited the experimental size. During the past 20 years, however, techniques rapidly developed that enabled high throughput marker typing. The newest techniques are able to type thousands of markers (Jakobsson *et al.*, 2008; Li *et al.*, 2008; Tishkoff *et al.*, 2009).

Despite the small size of the first studies it was immediately apparent that African and non-African genetic variation represent the earliest diversion in human history (Bowcock *et al.*, 1987; Nei and Livshits, 1989; Bowcock *et al.*, 1991a; Bowcock *et al.*, 1991b; Bowcock *et al.*, 1994). Africans had higher levels of nucleotide diversity compared to non-Africans.

Furthermore, the genetic diversity in non-African populations represents a subset of the genetic diversity in sub-Saharan Africa. Also, more private alleles and haplotypes are observed in Africa than in other regions. All of this strongly supported the Out of Africa model that was suggested by mitochondrial studies. Additionally, these low-resolution studies were already able to distinguish individuals on a continental basis.

Increasing the number of loci, increased the accuracy of the continental assignment of genotypes and facilitated the emergence of sub-clusters which correspond to populations within continents (Rosenberg *et al.*, 2002; Rosenberg *et al.*, 2005; Jakobsson *et al.*, 2008; Li *et al.*, 2008; Tishkoff *et al.*, 2009). Most of these studies utilized the HGDP-CEPH panel (Rosenberg *et al.*, 2002; Rosenberg *et al.*, 2005; Jakobsson *et al.*, 2008; Li *et al.*, 2008). The panel consists of cell lines of 1064 individuals from 51 populations from sub-Saharan Africa, North Africa, Europe, the Middle East, South/Central Asia, East Asia, Oceania, and the Americas (Cann *et al.*, 2002). This data set is freely available and allows a detailed characterization of worldwide genetic variation. The Khoe-San representation in this panel is, however, limited. Only seven individuals from a location south of Tsumkwe in Namibia are included in the panel. These individuals are indicated as “San relatives” and based on the geographic location probably belong to the Ju’hoansi or the ǀX’ao//’ãesi groups.

The study by Rosenberg *et al.*, used 377 autosomal microsatellite loci on the HGDP-CEPH panel (Rosenberg *et al.*, 2002). They found that worldwide variation could be clustered into six clusters of which five correspond to major geographic locations. Furthermore they could infer sub-clusters within these major regions. The sub-Saharan African cluster is optimally divided into four sub-clusters, which represent Bantu-speaking + pre Bantu-speakers, San, Mbuti Pygmy and Biaka Pygmy clusters (Rosenberg *et al.*, 2002; Rosenberg *et al.*, 2005).

Li *et al.*, used 650 000 SNP markers on the HGDP-CEPH panel and also found clustering into the five continental groups at K=5 (Li *et al.*, 2008). At K=6 south/central Asia separates from Europe and the Middle East; and at K=7 the Middle East separates from Europe. Many populations, however, have representation from more than one cluster. This can be an indication of recent admixture or shared ancestry before divergence. Additionally, PCA showed that the largest part of variation (56%) can be summarised as variation between

African and non-African populations. In the population distance tree, African populations lay closest to the root. The San group forms the earliest branch, followed by the Mbuti Pygmy group, the Biaka Pygmy group and thereafter the Bantu-speakers + pre-Bantu-speakers (Li *et al.*, 2008).

Jakobsson *et al.*, typed both SNPs (525 910) and STRs (396) on the HGDP-CEPH panel (Jakobsson *et al.*, 2008). This study found that on the global scale, similar to Rosenberg *et al.*, (2002), populations optimally grouped into six clusters of which five correspond to major geographic regions. The clustering within Africa, however, yielded interesting results. While Rosenberg *et al.*, optimally identified four clusters corresponding to the two Pygmy groups, the San and the Bantu- and pre-Bantu speakers, Jakobsson *et al.*, only identified three clusters. One of the three clusters represented Bantu- and pre-Bantu-speakers grouped. The Bantu-speakers from South Africa showed the largest contributions from the San/Pygmy clusters additional to the Bantu-Speaking cluster, followed by the Kenyans, the Yoruba and the Mandenka. The remaining two clusters were present at highest frequency in the Pygmy and San populations. Aside from small amounts of admixture from the Bantu/pre-Bantu speaking cluster, the Mbuti belonged almost exclusively to one of these clusters. The Biaka predominantly belonged to a third cluster but also had large contributions from the Mbuti cluster. The San contained both the Mbuti and Biaka cluster but with a larger contribution from the Mbuti cluster. It thus appears that the San and Mbuti group are closer related (Jakobsson *et al.*, 2008).

A recent study by Tishkoff *et al.*, included 2 432 African individuals from 113 geographically diverse populations (Tishkoff *et al.*, 2009). For evaluation against non-African groups the HGDP-CEPH panel was also included. The San group representation was better in this study, with a group of !Xun/Khwe samples included in addition to the HGDP-CEPH San samples. Additionally, a group of mixed Cape Coloured individuals were also typed. In these samples 1 327 polymorphic markers (microsatellites and insertion/deletions) were typed. Similar to previous studies, African populations contained the highest levels of genetic diversity. Globally, diversity declines with distance from Africa. Within Africa, the Pygmy and San populations had the highest genetic diversities, while the San groups had the most private alleles. In the tree analysis, the two Khoe-San populations cluster together

and are most distant from the other populations. The Cape Coloured population shows high levels of non-African admixture and are located between African and non-African groups. Using PCA, 72 significant global Principal Components (PCs) were identified. The first PC (19.5%) separates African from non-African populations. The Hadza is separated from other populations at PC3 (3.5%).

Using STRUCTURE analysis, the populations showed clustering according to major geographic region, both on a global scale and within Africa (Tishkoff *et al.*, 2009). Globally 14 ancestral population clusters were identified, while nine of these were found in Africa (Tishkoff *et al.*, 2009). A cluster emerged (at K=5) that is present in the Hadza, and to a lesser extent the Pygmy, San and Sandawe hunter-gatherers. Subsequently (at K=6) the cluster split into a Hadza/Sandawe and Pygmy/Khoe-San cluster. The Mbuti Pygmy and San groups split from the other Pygmy groups at K=11, indicating common ancestry between these groups. Results from this study showed that the San, Hadza, Sandawe and Pygmy populations contain shared genetic variation that distinguishes them from other African populations (Tishkoff *et al.*, 2009). This led to the suggestion that these groups are the remnants of a proto-Khoe-San/Pygmy/Hadza/Sandawe population of hunter-gatherers. MtDNA and Y-chromosome analysis suggest a divergence of >35 000 years BP (Semino *et al.*, 2002; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008; Tishkoff *et al.*, 2009).

The Hadza are genetically the most distinct from the other African groups (Tishkoff *et al.*, 2009), which is consistent with linguistic evidence that the Hadza language is unrelated to other Khoisan languages (Sands, 1998; Güldemann and Elderkin, Forthcoming; Güldemann, In Press). The Hadza is an isolated population that had little interaction with surrounding groups and has maintained their hunter-gatherer lifestyle up to recent times. They show only very low levels of asymmetric gene flow from surrounding groups. The Sandawe on the other hand adopted mixed farming practices and show evidence of bi-directional gene flow with neighboring groups (Newman, 1995). Populations from northern Tanzania, Southern Ethiopia and northern Kenya show evidence of the Sandawe associated genetic cluster (Tishkoff *et al.*, 2009). Aside from the association proven by autosomal DNA results, other commonalities between these two east African groups and

the Khoe-San groups are: the language connection between Sandawe and Khoisan, similarities between Tanzanian and San rock art, the Sandawe formerly performed a trance dance similar to San trance dances and there is evidence of pan San belief system across all of southern Africa to as far north as Zimbabwe (Huffman, 1983; Lewis-Williams, 1986; Güldemann and Elderkin, Forthcoming).

The clustering of the Khoe-San groups with the Pygmies (Tishkoff *et al.*, 2009) suggests that they may have a common genetic history. Pygmy populations might have had a Khoisan related language before it was replaced by Bantu-speaking language. Anthropological support for this theory comes from the shared music styles between the Khoe-San and Pygmy groups (Lomax, 1968; Tishkoff *et al.*, 2009). The San populations show a closer shared genetic ancestry to the Mbuti Pygmy than the Biaka Pygmy groups (Jakobsson *et al.*, 2008; Tishkoff *et al.*, 2009). The Mbuti lives in the Ituri rainforest of the eastern DRC while the Biaka (also called Baka, part of the Mbenga group) live to the west of the Mbuti in Cameroon, Gabon and the Republic of Congo (Figure 1.11). Another main group of Pygmies, the Twa or Ba-Twa and Cwa, live in dispersed groups south-central to the Mbuti and Mbenga (Cavalli-Sforza, 1986). These groups live in swamps and deserts far from the forest, there are no genetic data available for them, and it is not known if they are indigenous to the area or more recent migrants from the forest. It may be that before the Bantu-expansions, these Pygmy groups formed a continuous network of related groups that also had contact and gene-flow with their Khoe-San neighbours to the south and Hadza and Sandawe neighbours to the east.

To summarise; the genetic evidence emerging from the cluster analysis regarding the hunter gatherer populations, support linguistic data that suggest that Khoe-San ancestors may once have extended from Somalia through eastern Africa and into southern Africa and possibly also into western Africa (Ambrose, 1982; Tishkoff *et al.*, 2009; Güldemann and Elderkin, Forthcoming).

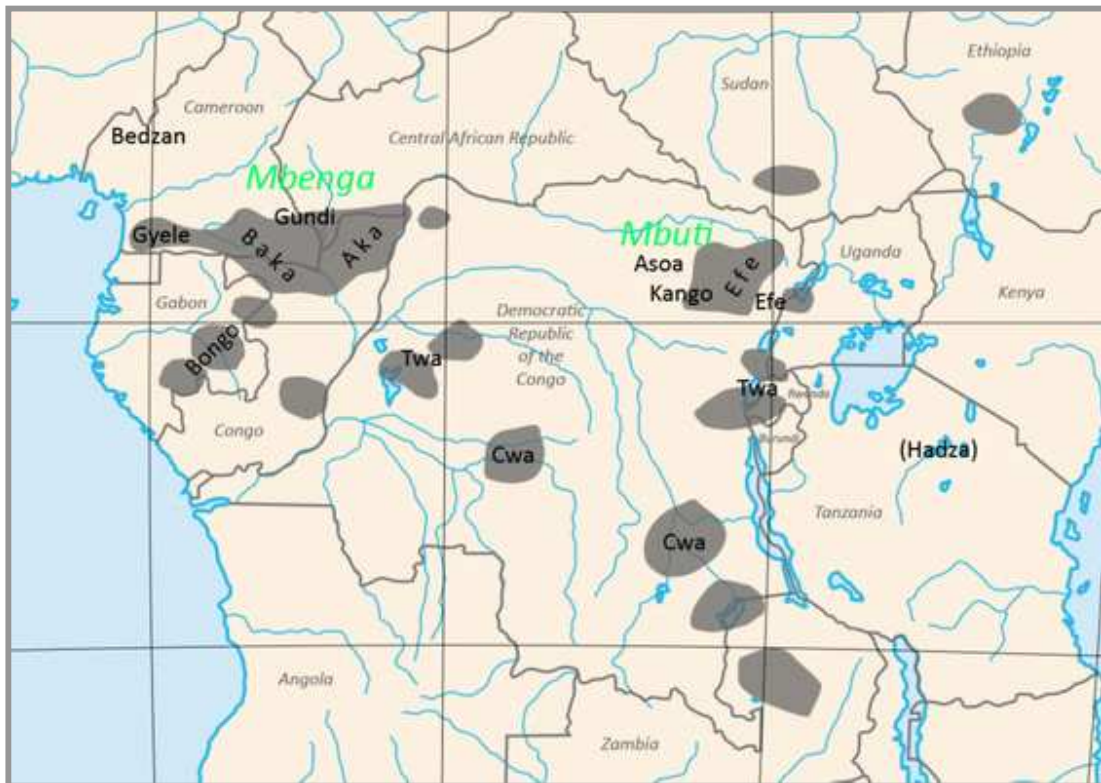


Figure 1.11 Distribution of Pygmies according to Cavalli-Sforza (1986). The Hadza do not form part of the Pygmy groups but are included to indicate proximity.
Map obtained from Wikipedia (<http://en.wikipedia.org/wiki/Pygmy>)

To better understand the evolutionary history of the Khoe-San, this study has made use of a number of autosomal SNPs that were typed in the various representative Khoe-San groups and their neighbours. These SNPs are spread over all of the 22 autosomes and are thus representative of the whole autosomal genome. By typing and analysing these SNPs we expect to find an intermediary picture of population structures and affinities compared to what we will find for the Y-chromosome and mtDNA, which represents the male and female lineage histories.

1.3 Aims

In this thesis the genetic structure of some living Khoe and San populations will be examined making use of different genetic markers (mtDNA, Y-chromosome and autosomal DNA). The study critically examines how females (mtDNA) and males (Y-chromosome) have contributed in shaping the gene pool of Khoe and San populations. The additional investigation of autosomal DNA markers will give an all-inclusive view of the population structures within the Khoe and San. The three genetic systems will also give insight to the amount and mode of admixture from various neighbouring population groups into the Khoe-San groups. An assessment of the ancestral association of San and Khoe populations will be implemented using various analytical methods. The resultant information from the genetic data will then be discussed in conjunction with linguistic, archaeological, historical and anthropological data to contribute to the writing of the history of the Khoe and the San.

In previous sections certain aspects about the presently known history of the Khoe-San were highlighted and elaborated upon. Other disciplines have contributed most of these historical perspectives regarding the Khoe-San and the aim of this thesis is to address these aspects from a genetic point of view. In particular the following fields will be concentrated on:

- Evidence of genetic distinction between the groups that represent the linguistic Ju, Tuu and Khoe divisions

The grouping of the Khoe-San into separate populations is largely based on a linguistic classification system. In sections 1.1.1 and 1.2.1.1 the linguistic classification system is reviewed in detail in conjunction with the demography and geographic localization of the groups involved. The history of the Khoe and San populations based on inference from the linguistic classification is discussed in sections 1.2.1.1 and 1.2.1.2. Linguistics supports a hierarchical relatedness of Khoe-San groups within the three main branches of the Khoisan linguistic family (Ju, Tuu and Khoe). It further supports the possibility that the Ju and Tuu branches may share a very deep common ancestor and were associated with the original

San hunter-gathers, while the Khoe branch was introduced to the area later in conjunction with pastoralism.

This study aims to investigate if the genetic relatedness between the groups correlate with the classification based on linguistics. The genetic relatedness of representatives from the three main Khoisan linguistic branches will be evaluated to see if they are closer related to each other than to representatives from other linguistic branches.

- Evidence of a relationship between physical geographic distance and genetic distance between groups regarding males and females in hunter-gatherer communities

Serological studies (discussed in section 1.2.2.1) suggested relatedness between different Khoe-San groups based on geographical distance rather than linguistics. In this thesis the relationship between genetic distance and physical geographic distance for all three of the genetic systems (mtDNA, Y-chromosome, autosomal) will be investigated. Section 1.2.2.3 discussed that results from previous studies suggested that either Y-chromosomal genetic distance (male line) or mtDNA genetic distance (female line) shows a correlation with geographic distance depending on if the population involved are food-producers or hunter-gatherers. Food-producers practice patrilocality, which limit male migration and cause strong correlations between the Y-chromosome genetic distance and physical distance. The reverse case applies to hunter-gatherers where the mtDNA genetic distance correlates with geographic distance but not the Y-chromosome genetic distance. In this study correlations between genetic and physical geographic distances between the different genetic systems will be considered to identify dissimilarity between the female and male migration histories.

- The genetic affinities of the Khwe population

The Khwe group is discussed in sections 1.1.1.2.2 and 1.1.1.5.6. Although the Khwe speak a Khoe language their classification as a Khoe-San group has been questioned. They phenotypically resemble Bantu-speakers and it is not clear if they are Khoe-San groups with extensive Bantu-speaking admixture, Bantu-speakers that lost their cattle, another pastoralist population closely related to Bantu-speakers who occupied the region before

the Bantu expansions or a mixture of various refugee groups driven from the grazing grounds into the Okavango swamps. Serological studies (section 1.2.2.1.2) found them to be closely related to Bantu-speakers. Y-chromosome studies (discussed in section 1.2.2.3) suggested that the Khwe might be related to east African groups who introduced pastoralism into southern Africa (possibly together with the Khoe languages).

By typing the three genetic systems in the Khwe and comparing their genetic profile with other Khoe-San groups as well as Bantu-speakers this study will aim to establish the genetic identity of the Khwe. Furthermore, Y-chromosome evidence will be assessed to evaluate the claim that the Khwe are descendent from east African populations who introduced pastoralism into southern Africa.

- Whether genetic evidence supports a cultural or demic diffusion of pastoralism. The possibility of a combination of a cultural/linguistic and demic diffusion and the likelihood of gender biased demic diffusion will also be looked at

Both archaeology and linguistics contributed to the theory of how pastoralism spread from the area of northern Botswana towards the south (see sections 1.2.1.3 and 1.2.1.2 for discussion). Y-chromosome genetic studies suggested how pastoralism was introduced into northern Botswana from east Africa (discussed in section 1.2.2.3). Without representation of more Khoe-San groups the study, however, could not address the question of how pastoralism spread after it reached the area around northern Botswana.

Linguistics couples the large amount of variation and dialects in the Khoe language branch to a rapid expansion related to the spread of pastoralism. According to the linguistic theory (see section 1.2.1.2) pastoralism was introduced to northern Botswana by a group from east Africa (link that exist between Khoisan and Sandawe). Thereafter there was a rapid diversification of the language that formed the Kalahari Khoe branches. It is not known if the language expansion and diversification that formed the Kalahari Khoe branches are correlated with the diversification and expansion of the east African immigrant groups. Thus, it is not sure if all the groups that speak the Kalahari Khoe branches are descendant form the east African immigrants or descendant from hunter-gatherers that adopted pastoralism and language from the east African immigrants with limited admixture.

Thereafter pastoralism, the Khoe language and possibly the pastoralist groups themselves spread south into the present day Cape Province of South Africa. Here the KhoeKhoe branch diverged from the Kalahari Khoe branches by incorporating elements from the !Ui language group from the Tuu linguistic division spoken by resident San hunter-gatherers.

The archaeological explanation for the spread of pastoralism (see section 1.2.1.3) is based on the introduction of pottery and sheep remains in the archeological record. Two alternative routes were suggested for the southern spread of the pastoralists based on a demic diffusion model. Certain aspects in the archaeological record, however, suggest that a clear-cut demic diffusion model might not be the best explanation (see section 1.2.1.3).

Neither archaeology nor linguistics can conclusively prove whether the spread of pastoralism is associated with a demic diffusion of populations together with the pastoralist culture or a diffusion of the culture on its own. An intermediate model where only few individuals, perhaps only males, spread and transferred the pastoralist tradition and their language to resident hunter-gatherer groups further south is also possible. A genetic approach using male specific (Y-chromosome) and female specific (mtDNA) markers would be employed in this thesis. A specific Y-chromosome marker was coupled to the introduction of pastoralism, and this marker was strongly associated with the Khwe population (see section 1.2.2.3). Therefore, this Y-chromosome profile as well as the male and female profile of the Khwe will be examined in this study to see if Khwe associated Y-chromosome and/or mtDNA markers are prevalent in the Khoe groups and other southern Khoe-San groups.

- Investigation if population growth signals in genetic data reflects population expansions in the archaeological record

- Following from above... if these signals give an indication of a recent population contraction due to a post-Neolithic population bottleneck induced by pastoralist groups

The archaeological record contributed extensively to the inference of Khoe-San history. In section 1.2.1.3 a broad overview of the Stone Age history from southern Africa is presented. Temporally associated population expansions are discussed in conjunction with

the factors that possibly caused these expansions. In the present theses various methods will be used to infer expansion signals in the mtDNA sequence data. These signals will be dated to specific times in the past and correlated with the archaeological record to identify possible temporal overlaps.

Section 1.2.1.3 discussed how published data explained results from mismatch distributions by inferring post-Neolithic bottlenecks in hunter-gatherer societies induced by pastoralists. In the archaeological community there is also disparity about how in-moving pastoralists affected hunter-gatherer communities (see section 1.2.1.3). This study will investigate mtDNA sequence data for evidence of recent bottlenecks associated with in-moving pastoralists.

2. SUBJECTS AND METHODS

2.1 Subjects

A total of 551 individuals were included in this study for Y-chromosome, mitochondrial and autosomal SNP screening. An additional 161 individuals were included for the validation of a mitochondrial minisequencing panel. All samples were collected with the subjects' informed consent. This study was approved by the Human Research Ethics Committee (Medical) at the University of Witwatersrand, Johannesburg, South Africa (Himla Soodyall, Protocol Number M980553 and Carina Schlebusch, Protocol Number M050902; Appendix A). The study and the participation of San individuals were furthermore approved by the South African San Council and the Working Group of Indigenous Minorities in Southern Africa (WIMSA). The additional 161 DNA samples used for the validation of the minisequencing panel were contributed by Prof. S.W. van der Merwe from the Department of Immunology, University of Pretoria as part of a collaborative project with Prof. Soodyall.

Due to gender, family relations or missing data particular individuals were excluded from certain parts of the analyses. Table 2.1 and Figure 2.1 summarise the numbers of individuals included in the different parts of the project, their place of sampling, their population group and the population group code used throughout this manuscript.

At the time of sampling 10 ml of blood were collected into EDTA tubes or buccal swabs were taken from volunteers. Information from the subjects on their place of birth, the language spoken by them and by their parents and their self-classified ethnicity were collected. This information was used to group individuals.

Table 2.1 Number of individuals in which mtDNA, Y-Chromosome and autosomal variation were examined, their group and group-code, and place of sampling and origin

Group name	Group code	Place of sampling (Country)	Place of origin If different from place of sampling	N (mtDNA)	N (Y-Chr)	N (Autosomal SNPs)
* Karretjie people	KAR	Colesberg (SA)		30	19	25
* Karoo Coloured	COL	Colesberg (SA)		77	35	22
# Cape Coloured	CAC	Wellington (SA)		20	3	20
* ‡Khomani	KHO	Askham (SA)		57	37	-
* Northern Cape Coloured	CNC	Askham (SA)		40	23	-
* //Xegwi	XEG	Chrissiesmeer (SA)		3	3	-
* Duma San	DUM			1		-
# Nama	NAM	Windhoek (NM)		28	14	28
# /Gui, //Gana and Kgalagari	GUG	Kutse Game reserve (BT)		22	19	21
* Naro	NAR	Johannesburg (SA)	Ghanzi (BT)	2	2	-
# Jul'hoansi	JOH	Tsumkwe (NM)		42	28	41
# !Xun	XUN	Omega camp (NM) and Schmidtsdrift (SA)		49	48	45
# Khwe	KWE	Omega camp (NM) and Schmidtsdrift (SA)		18	13	19
# Manyanga	DRC	Luozi (DRC)		14	14	14
# Herero	HER	(NM)		15	15	14
## Sotho, Tswana	SOT	Various (SA)		22	21	See SEB
* Swazi	SWZ	Chrissiesmeer (SA)		5	2	See SEB
## Zulu, Xhosa	ZUX	Various (SA)		36	30	See SEB
## Afrikaner	AFR	Various (SA)		21	13	15
## European	EUR	Various (SA)	Europe and Canada	11	3	15
## Indian	IND	Various (SA)		25	11	25
## South-eastern Bantu- speakers	SEB	Various (SA)				48
Total				538	353	352

AN – Angola
 BT – Botswana
 DRC – Democratic Republic of Congo
 NM – Namibia
 SA – South Africa

* Collected during field trip conducted by author
 # Other samples collected by Prof's H. Soodyall and T. Jenkins



Figure 2.1 Map indicating the place of origin for the Coloured and Khoe-San individuals who participated in the study

The Coloured, Khoe and San groups were collected on specific sampling trips at specific locations. The South-eastern Bantu-speaking individuals, Afrikaner and European individuals were assembled from various sampling groups and originate from various locations. They, together with the Herero and Manyanga, were used as comparative data to test admixture proportions into the San, Khoe and Coloured groups.

During the collection in Colesberg, the KAR samples were collected at the “outspans” of the Karretjie people, while the COL samples were from the Coloured school in the Lowryville township adjacent to Colesberg. The samples collected in and around Askham were also divided into two groups. Individuals who indicated their ethnicity as Coloured, Griqua or Nama were assigned to the CNC group while the individuals who identified themselves and/or their parents as ‡Khomani or “Bushmen” were assigned to the KHO group. The KHO and CNC samples were not collected at the time when autosomal SNP work were being conducted and are therefore absent from the autosomal SNP analyses.

The //Xegwi, Duma and Naro groups had very few representative individuals and were only used in individual based analysis, and not in group based analyses. The CAC group was also excluded from group-based analysis for the Y-chromosome since the group contained only three males. The EUR and AFR groups were combined into the AFE group for Y-chromosome analysis due to low number of males.

The /Gui, //Gana and Kgalagari (GUG) were a mixed group of San and Bantu-speaking individuals who had ancestries from both /Gui and //Gana San groups as well as the Kgalagari Bantu-speaking group.

The 161 additional samples screened for the validation of the minisequencing panel included 156 individuals contributed by Prof. S. van der Merwe and 5 additional individuals from the HGDDR laboratory. The samples contributed by Prof. van der Merwe included 29 Khoe-San individuals (9 ‡Khomani, 11 !Xun, 7 Khwe and 2 unspecified San) and 127 south-eastern Bantu-speakers (SEB) from various ethnic groups. The extra 5 individuals

from the HGDDRU laboratory comprised 3 individuals from Zanzibar and 2 additional SEB individuals.

2.2 Methods

Details of reagents used in molecular methods are available in Appendix B.

2.2.1 DNA extraction

DNA was extracted from either EDTA-blood or buccal swabs.

DNA extraction from EDTA-blood was done using the salting-out method as described in Miller 1988 (Miller *et al.*, 1988) with some modifications. The modified procedure is described as follows. After thawing EDTA-blood tubes, the blood was decanted into centrifugation tubes and filled to the 30 ml mark with chilled Sucrose-Triton-X Lysing buffer. The tubes were inverted several times to mix and centrifuged for 15 min @ 1000 g (4°C). The supernatant was discarded and 20 ml Sucrose-Triton-X Lysing buffer was added to the pellet. Tubes were then vortexed to break up the pellet and put on ice for 5 min. After centrifugation for 10 min @ 1000 g (4°C) the supernatant was again discarded. The pellet was resuspended and digested overnight at 42°C with 1.5 ml T20E5, 0.1 ml 10% SDS and 0.25 ml freshly made Proteinase-K mix. After digestion, 0.5 ml of saturated NaCl was added to each tube, shaken vigorously for 15 s and put on ice for 10 min, followed by centrifugation for 30 min @ 1000 g (4°C). The DNA-containing supernatant was poured into a clean tube and the protein pellet was discarded. Two volumes of room temperature 100% ethanol were added and the tube was gently agitated. The visible DNA was spooled, washed in 70% ice-cold ethanol and transferred to an empty Eppendorf tube. After air-drying for 30 min the DNA was resuspended in 500-1000 µl TE buffer. DNA was allowed to dissolve overnight before quantification.

If no DNA was visible after precipitation with 100% ethanol the following procedure was followed. The tube was left at -20°C overnight. The next day the tube was centrifuged for

20 min @ 1000 g (4°C), the supernatant discarded and 10 ml of 70% ethanol was added to the pellet. The tube was centrifuged for 20 min @ 1000 g (4°C), the supernatant discarded and the pellet allowed to air-dry. After air-drying the DNA was resuspended in 100-200 µl TE buffer. DNA was allowed to dissolve overnight before quantification.

Extraction from buccal swabs was done using the PureGene® Genomic DNA Purification Kit (Gentra Systems) according to the manufacturer's instructions.

DNA was quantified using the NanoDrop ND-1000 Spectrophotometer (Coleman Technologies Inc., LabVIEW®) and diluted to the required concentration using double distilled water (ddH₂O).

2.2.2 MtDNA methods

To assign individuals to mitochondrial haplogroups, two approaches were followed. A mitochondrial minisequencing panel was designed to target specific polymorphisms in the mtDNA coding region. The minisequencing panel allocates the mtDNA to one of the 10 major macro-haplogroups found in mitochondrial variation worldwide. The design and implementation of the minisequencing panel have recently been published (Schlebusch *et al.*, 2009). Secondly HVS-I and HVS-II were sequenced to further classify the haplotype into sub-haplogroups and to be used in phylogenetic and population genetic analyses.

2.2.2.1 MtDNA minisequencing method

The minisequencing procedure is based on a single base extension of an unlabelled primer. The reaction mix contains ddNTPs labeled with four different colours and a fifth colour is used for the internal lane standard (LIZ 120). The primers are designed to bind directly adjacent to the 5' side of the mutation of interest. During the extension cycles the primer is extended by only one basepair, which is the colour labeled ddNTP. Primers are designed to differ in size by attaching poly(GATC) tails to the hybridization part. When the products are separated on the Genetic Analyser an electropherogram of different sized peaks result. The colour of the peak indicates the allele present at the site of interest.

For the design of the minisequencing protocol the ABI PRISM® SNaPshot™ Multiplex Kit was used and the general guidelines of the Protocol were followed with some minor modifications. Whereas the supplier's protocol was originally optimized using POP-4 polymer, our method was optimized using the POP-7 polymer using suggestions proposed by Applied Biosystems in a subsequent bulletin (Applied Biosystems Manual P/N: 4367258).

The minisequencing protocol was designed to distinguish between the seven African L mitochondrial macro-haplogroups as well as the three non-African macro-haplogroups M, N and R. The panel tests for 14 SNP variations that define these 10 macro-haplogroups (Figure 2.2). It was designed in such a way that for every split in the tree there is a SNP that defines both branches in the split. For instance where L1 splits from the rest of the tree there is a SNP defining the L1 branch (Figure 2.2) and a SNP defining the L2-6 clade (Figure 2.2).

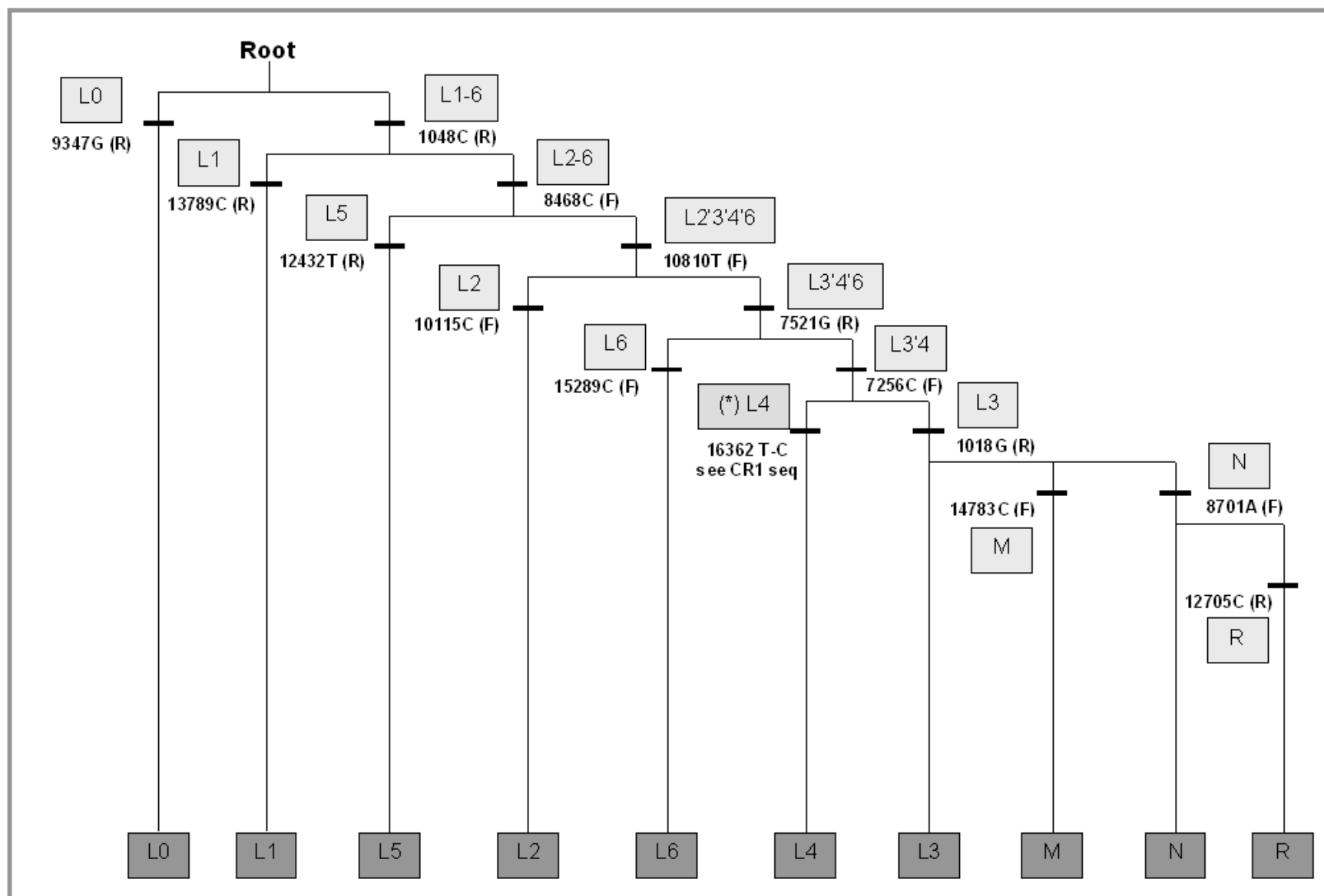


Figure 2.2 Tree showing the 10 mtDNA macro-haplogroups (dark-grey) that are distinguished by typing 14 SNPs (light-grey). L4 (*) is identified by a HVS polymorphism that is not included in the panel. R and F indicate whether the reverse or forward primer orientation was used. Branch nomenclature on the tree is according to Behar *et al.*, (2008)

2.2.2.1.1 PCR-multiplex amplification

The PCR-multiplex preceding the minisequencing reaction consisted of the simultaneous amplification of 6 PCR fragments of various lengths (Table 2.2). The binding sites for the 14 minisequencing primers are contained within these 6 fragments (SNPs that are closely positioned are co-amplified in the same amplicon). Multiplex primers were designed by selecting specific regions and adjusting amplicon lengths in order to correlate annealing temperatures to allow for multiplexing. Only regions without sequence polymorphism or low amounts of sequence polymorphism were considered as possible primer binding sites. Primers for the multiplex were designed using Primer 3 software (Rozen and Skaletsky, 2000) and checked with the Autodimer program (Vallone and Butler, 2004) (Reverse and forward primer sequences in Table 2.2). PCR Multiplex Primers were manufactured and HPLC purified by Metabion. The primers were diluted to 100µM, and stored at -20°C.

The concentration of primers, MgCl₂ and DNA template were optimised. The reaction was not sensitive to variation in DNA concentration as long as amounts above 5 ng were used. In the final optimised PCR procedure, the reaction volume was 25 µl, including 10 ng DNA template, 1 µl of premixed 25x primer mix (see Table 2.2 for final reaction concentrations), 2 U FastStart Taq (Roche Applied Science), 1x FastStart Taq buffer (containing no added MgCl₂), 3,5 mM MgCl₂, 0.3 mM dNTPs and ddH₂O to make up the reaction volume to 25 µl.

Thermal cycling conditions were as follows: Initial step at 95°C for 6 min followed by 35 cycles of denaturation at 95°C for 1 min 30 s, annealing at 60°C for 1 min 30 s and amplification at 72°C for 2 min; final extension for 10 min at 72°C. All the PCR reactions were performed on a 9700 GeneAmp® PCR System (Perkin-Elmer, Applied Biosystems). During optimisation PCR product sizes were checked on a 2% agarose gel with ethidium bromide staining (1 x TBE running buffer, Bromophenol blue Ficoll dye loading buffer, 1Kb DNA ladder size standard (Gibco BRL).

Post PCR purification was done by adding 1 U of Shrimp Alkaline Phosphatase (USB Corporation) and 2 U of Exonuclease I (New England Biolabs) to 5 µl PCR product in a

total reaction volume of 7 µl. The reaction was incubated at 37°C for 1 h followed by 15 min at 75°C for enzyme inactivation.

Table 2.2 Primer sequences, binding sites, amplicon sizes and concentrations for multiplex PCR amplification of 6 fragments

Primer name	Amplicon size (bp)	PCR primer sequences (5' - 3')	Mitochondrial region of primer binding site *	Final Concentration (µM) **
MTSS_1f	210	CCGGCGTAAAGAGTGTTTATAGAT	931-953	0.04
MTSS_1r		TTCTGGCGAGCAGTTTGT	1121-1140	0.04
MTSS_2f	502	CCCTATTCTCAGGCTACACCCTA	7096-7118	0.03
MTSS_2r		TGCATGTGCCATTAAGATATATAGGA	7572-7597	0.03
MTSS_3f	1051	CAGTGAAATGCCCCAACTAAATAC	8359-8382	0.05
MTSS_3r		TGGTATGTGCTTCTCGTGTTAC	9387-9409	0.05
MTSS_4f	868	CTCTTTTAGTATAAATAGTACCGTTAACTCC	9992-10023	0.20
MTSS_4r		TAATTAGGCTGTGGGTGGTTGT	10838-10859	0.20
MTSS_5f	1577	CAGCTATCCATTGGTCTTAGGC	12281-12302	0.20
MTSS_5r		TAGGTAGTTGAGGTCTAGGGCTGTTA	13832-13857	0.20
MTSS_6f	672	CCACGACCAATGATATGAAAAAC	14694-14716	0.03
MTSS_6r		TGTTTGATCCCGTTTCGTG	15347-15365	0.03

* Numbering according to the revised Cambridge reference sequence.

** The final concentration of the primers in the reaction mix

2.2.2.1.2 Minisequencing reaction

Minisequencing primers were designed using Primer 3 software (Rozen and Skaletsky, 2000) and checked with the Autodimer program (Vallone and Butler, 2004) (minisequencing primer sequences in Table 2.3). In minisequencing, primer sizes and different fluorochrome colours are important in the separation and detection of the extension products. Therefore, the primers were designed to be of varying lengths (at least 5 bp) through the addition of poly (dGACT) tails at the 5' end to ensure good separation in the electropherogram (Table 2.3). Minisequencing primers were manufactured and HPLC purified by Metabion.

The minisequencing reaction had a total volume of 5 µl containing 1.5 µl of purified PCR product, 1 µl of ABI PRISM® SNaPshot™ Multiplex Ready Reaction Mix, 1 µl of premixed

5x minisequencing primer mix (see Table 2.3 for final concentrations) and 1.5 µl ddH₂O. Thermal cycling was performed for 35 cycles with denaturation at 96°C for 10 s, annealing at 50°C for 5 s and extension at 60°C for 30 s.

Post extension treatment was done in a total volume of 7 µl containing 5 µl minisequencing reaction product, 0.5 U Shrimp Alkaline Phosphatase (USB Corporation), 1x Shrimp Alkaline Phosphatase buffer and ddH₂O to make up the reaction volume. The reaction was incubated at 37°C for 1 h followed by 15 min at 75°C for enzyme inactivation.

Two µl of cleaned minisequencing reaction product was then mixed with 7.5 µl Hi-Di formamide (Applied Biosystems) and 0.5 µl of GeneScan-LIZ 120 internal size standard (Applied Biosystems). After a denaturing step for 2 min at 95°C followed by cooling to 4°C the fragments were separated on an ABI PRISM® 3130xl Genetic Analyzer (Applied Biosystems) according to ABI PRISM® SNaPshot™ Multiplex Kit instructions and analysed using GeneMapperID v3.2 software (Applied Biosystems). The resultant electropherogram displayed the different sized products (Table 2.4 gives the expected band sizes and peak colours).

Table 2.3 Minisequencing primers used to distinguish haplogroups L0-L6, M, N and R

PCR amplicon Primers (see Table 2.2)	SNP sequence variation	Haplogroup resolved (see tree in Figure 2.2)	Minisequencing primer sequences (5' - 3') *	Minisequencing primer orientation	Mitochondrial region of primer binding site **	Final concentration (μ M) ***
MTSS1 F+R	1018G	L3	<u>(GATC)</u> CAGATATGTTAAAGCCACTTTCGTAGT	R	1019-1045	0.20
MTSS1 F+R	1048C	L1-6	<u>CCC(GATC)</u> ₂ CCAGTTTGGGTCTTAGCTATTGTGT	R	1049-1073	0.10
MTSS2 F+R	7256C	L3'4	<u>(GATC)</u> ₅ CGATGCATACACCACATGAAA	F	7235-7255	0.20
MTSS2 F+R	7521G	L3'4'6	<u>(GATC)</u> ₄ TGACAAAAGTTATGAAATGGTTTTCTAATA	R	7522-7551	0.20
MTSS3 F+R	8468C	L2-6	<u>(GATC)</u> ₆ CCAACTAAAAATATTAAACACAACTACCAC	F	8473-8467	0.20
MTSS3 F+R	8701A	N	<u>(GATC)</u> ₁₁ CTAATCAAACCTAACCTCAAACAAATGATA	F	8671-8700	0.40
MTSS3 F+R	9347G	L0	<u>(GATC)</u> ₉ ATTGGTATATGGTTAGTGTGTTGGTTAG	R	9348-9375	0.20
MTSS4 F+R	10115C	L2	<u>(GATC)</u> ₁₀ AACACCCTCCTAGCCTTACTACTAATAAT	F	10086-10114	0.20
MTSS4 F+R	10810T	L2'3'4'6	<u>(GATC)</u> ₁₂ CAACAATTATATTACTACCACTGACATGACT	F	10779-10809	0.14
MTSS5 F+R	12432T	L5	<u>CC(GATC)</u> ₁₅ CAATGGATTTTACATAATGGGG	R	12433-12454	0.50
MTSS5 F+R	12705C	R	<u>CC(GATC)</u> ₁₄ CGGTAACCTAAGATTAGTATGGTAATTAGGAA	R	12706-12736	0.50
MTSS5 F+R	13789C	L1	<u>C(GATC)</u> ₁₈ CGAGGGCTGTGAGTTTtaggt	R	13790-13810	0.50
MTSS6 F+R	14783C	M	<u>CCC(GATC)</u> ₁₈ CGCAAAATTAACCCCTAATAAAA	F	14759-14782	0.50
MTSS6 F+R	15289C	L6	<u>C(GATC)</u> ₂₀ ACCCTCACACGATTCTTTACCTT	F	15266-15288	0.30

* The non-specific primer tail is underlined and in italic

** Numbering according to the revised Cambridge reference sequence.

*** The final concentration of the primers in the reaction mix

Table 2.4 Chromatogram band profile for identifying haplogroups L0-L6, M, N and R

Mutation	Electropherogram Band size	Haplogroup resolved (See tree in Figure 2.2)	Primer Orientation	Peak color Negative	Peak color Positive
1018G	31	L3	R	t-red	c-black
1048C	36	L1-6	R	a-green	g-blue
7256C	41	L3'4	F	t-red	c-black
7521G	46	L3'4'6	R	t-red	c-black
8468C	55	L2-6	F	t-red	c-black
9347G	64	L0	R	t-red	c-black
10115C	69	L2	F	t-red	c-black
8701A	74	N	F	g-blue	a-green
10810T	79	L2'3'4'6	F	c-black	t-red
12432T	84	L5	R	g-blue	a-green
12705C	89	R	R	a-green	g-blue
13789C	94	L1	R	a-green	g-blue
14783C	99	M	F	t-red	c-black
15289C	104	L6	F	t-red	c-black

2.2.2.2 HVS amplification and sequencing

Mitochondrial sequencing of HVS-I and II were done to cover regions 16024-16569 for HVS-I and 57-630 for HVS-II. The amplification and sequencing were done according to two previously published methods (Vigilant *et al.*, 1989; Behar *et al.*, 2007). Initially the protocol of (Vigilant *et al.*, 1989) was followed and later replaced by the protocol of (Behar *et al.*, 2007). Amplification and sequencing primers are shown in Table 2.5 and procedures in Table 2.6.

Table 2.5 Sequences of primers used to amplify and sequence HVS-I and II

Primer	Description	Primer sequence 5'-3'	Reference
PCR primers			
L15996	PCR forward	CTCCACCATTAGCACCCAAGC	(Vigilant <i>et al.</i> , 1989)
H408	PCR reverse	CTGTTAAAAGTGCATACCGCCA	(Vigilant <i>et al.</i> , 1989)
15876F	PCR forward	TCAAATGGGCCTGTCCTTGTAG	(Behar <i>et al.</i> , 2007)
639R	PCR reverse	GGGTGATGTGAGCCCGTCTA	(Behar <i>et al.</i> , 2007)
Cycle sequencing primers			
L15996	HVS-I forward	CTCCACCATTAGCACCCAAGC	(Vigilant <i>et al.</i> , 1989)
H16401	HVS-I reverse	TGATTTCACGGAGGATGGTG	(Vigilant <i>et al.</i> , 1989)
15946F	HVS-I forward	CAAGGACAAATCAGAGAAAA	(Behar <i>et al.</i> , 2007)
132R	HVS-I reverse	GACAGATACTGCGACATAGG	(Behar <i>et al.</i> , 2007)
L29	HVS-II forward	GGTCTATCACCTCTTAACCAC	(Vigilant <i>et al.</i> , 1989)
H408	HVS-II reverse	CTGTTAAAAGTGCATACCGCCA	(Vigilant <i>et al.</i> , 1989)
639R	HVS-II reverse	GGGTGATGTGAGCCCGTCTA	(Behar <i>et al.</i> , 2007)

Table 2.6 PCR ingredients and cycling conditions for amplification and sequencing of HVS-I and II. Final concentrations of ingredients are shown

Description	Concentrations / Conditions according to:					
PCR						
Ingredients	(Vigilant <i>et al.</i> , 1989)			(Behar <i>et al.</i> , 2007)		
DNA	~50 ng			~50 ng		
FastStart 10x Buffer (with added MgCl ₂)	1 x			1 x		
Primer 1	0.4 µM			0.4 µM		
Primer 2	0.4 µM			0.4 µM		
dNTP's	0.1 mM			0.1 mM		
BSA	1 mg/ml			-		
FastStart Taq (Roche Applied Science)	1 U			1 U		
Total volume	50 µl			50 µl		
Cycling conditions	Temperature (°C)	Time (min:sec)	Cycles	Temperature (°C)	Time (min:sec)	Cycles
Initiation	95	5:00	30	95	5:00	35
Denaturation	94	1:00		95	0:30	
Annealing	56	1:00		55	0:30	
Extension	74	1:00		72	2:00	
Final extension	74	10:00		72	10:00	
Hold	4	Hold		4	Hold	
Cycle sequencing						
Ingredients	(Vigilant <i>et al.</i> , 1989)			(Behar <i>et al.</i> , 2007)		
PCR product	4-8 µl			2 µl		
Big Dye	4 µl			1 µl		
Primer	0.165 µM			0.33 µM		
Total volume	20 µl			10 µl		
Cycling conditions	Temperature (°C)	Time (min:sec)	Cycles	Temperature (°C)	Time (min:sec)	Cycles
Initiation	96	1:00	25	96	1:00	25
Denaturation		0:30		96	0:10	
Annealing		0:15		50	0:05	
Extension		4:00		60	4:00	
Hold		Hold		4	Hold	

PCR cleanup was performed using MultiScreen® PCR_μ96 Plates (Millipore) according to kit instructions. Product sizes of the PCR were checked on a 2% agarose gel with ethidium bromide staining (1 x TBE running buffer, Bromophenol blue Ficoll dye loading buffer, 1Kb DNA ladder (Gibco BRL) size standard). Sequencing reaction cleanup was done using Montage SEQ₉₆ Sequencing Reaction Cleanup Plates (Millipore). All thermal cycling were performed on a 9700 GeneAmp® PCR System (Perkin-Elmer, Applied Biosystems). Sequencing products were separated on an ABI PRISM® 3130xl Genetic Analyzer (Applied Biosystems) and analysed using Sequencing Analysis Software v5.2 (Applied Biosystems)

2.2.2.3 MtDNA data analysis

The designed minisequencing method was used to group samples in their major haplogroups. Further classification was achieved by analysing HVS-I and II.

HVS-I and II sequences were aligned to the control region reference sequence (Andrews *et al.*, 1999) using the Clustal W algorithm (Thompson *et al.*, 1994) implemented in BioEdit v.7.0.5.3 (Hall, 1999). HVS-I and II sequences (15997-16569 and 57-607) were then combined into one sequence of 1124 bp for further analysis. Unique haplotypes were identified using DnaSP v4.10 (Rozas *et al.*, 2003) and variant sites were recorded electronically using S-compare (Nelson, 2006). Using the variant positions together with a phylogenetic approach, haplogrouping was done according to the nomenclature of Behar 2008 (Behar *et al.*, 2008).

Variation in the HVS-II region 303-315 were not considered or reported in any of the analyses. Insertions in the poly C repeat track at position 568-573 were taken as a 1 bp C insertion. All other regions were considered albeit some regions were differentially weighted as outlined in the analysis description.

Phylogenetic tree analyses of sequences were done through Maximum likelihood analysis using PHYML (Guindon *et al.*, 2005). The HKY substitution model with Gamma distributed rates and Invariable sites, received the best likelihood prediction through likelihood ratio

tests using Modeltest 3.7 (Posada and Crandall, 1998) in conjunction with PAUP v4.0b10 (Swofford, 1998) and were implemented in the Maximum likelihood analysis. The tree topology search employed was nearest neighbour interchange (NNI). An approximate likelihood ratio test (aLRT) was computed to determine branch support (Anisimova and Gascuel, 2006). Trees were visualized in MEGA4 (Tamura *et al.*, 2007).

Networks of the sequences were constructed using the Median Joining algorithm (Bandelt *et al.*, 1999) of Network v4.5.0.0 (Fluxus-engineering, 2008). Networks were subjected to maximum parsimony post-analysis using the Steiner maximum parsimony algorithm (Polzin and Daneschmand, 2003) within Network 4.5.0.0. For network analysis the epsilon parameter (Network program parameter for quick calculation of sparse networks), was set to 2 and transversions were weighted 3x the weight of transitions. Furthermore the weight of the 16189 position was reduced 10x and the weight each of the CA repeats at position 523 was reduced 5x per nucleotide in the repeat.

Sequences from other sources included in phylogenetic and network analyses were Neanderthal (Genbank accession number: NC_011137) (Green *et al.*, 2008) and the control region reference sequence (Andrews *et al.*, 1999). Additional L0d sequences published in the literature (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008) were included in the L0d network to compare our results with. Sequences from Gonder *et al.*, and Tishkoff *et al.*, had overlap in some of the subjects and only one of the two in each case were selected (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007).

Time estimates of L0d subgroups were calculated using the Rho statistic (Forster *et al.*, 1996) with the associated standard deviation, sigma (Saillard *et al.*, 2000), using a mutation rate of 2.5×10^{-6} per nucleotide per generation (Ward *et al.*, 1991) (25 yrs per generation; 1124 nucleotides). Time estimates were also calculated using other published mutation rates (i.e. 1.75×10^{-6} per nucleotide per generation (Horai *et al.*, 1995); 4.5×10^{-6} per nucleotide per generation (Forster *et al.*, 1996); 2.1×10^{-6} per nucleotide per generation (Soodyall *et al.*, 1996) but because of its intermediate value the mutation rate of Ward *et al.*, was used in subsequent discussions and analyses (Ward *et al.*, 1991). A generation time of 25 years was used throughout.

Haplogroup isofrequency maps were generated applying the Kriging method (Oliver and Webster, 1990; Xue *et al.*, 2005) incorporated in the Surfer v.8.06.39 program (Golden-Software, 2006). Mitochondrial contour plots were based on frequencies of the L0d/k subgroups on the background of the L0d/k group as a whole. This was done to eliminate the effects that admixture from Bantu-speakers and non-Africans would have on the distribution of the L0d/k subgroups. When frequencies were calculated, sample size effects were corrected by adjusting the total sample sizes in all groups to the same value.

Mismatch distributions of populations and haplogroups were calculated in Arlequin v.3.11 (Excoffier *et al.*, 2005). From these the validity of demographic expansions and the date of expansions were inferred. The demographic expansion scenario is tested through simulating a population going through an expansion and testing whether the actual data is significantly different from the simulated expansion scenario. A non-significant Sum of Squared deviation (SSD) p-value will therefore indicate a population/group of sequences that went through an expansion. Parameters calculated are θ_1 , θ_0 , and τ . Dividing θ_1 by θ_0 give an indication of the magnitude of the expansion while τ gives an indication of the time of the expansion. The mutation rate of 2.5×10^{-6} per nucleotide per generation (Ward *et al.*, 1991) and a generation time of 25 years were used to convert τ (Tau) to T (Time BP when expansion took place) by using the equation $T = (\tau/2\mu) \times \text{generation time}$. In the equation μ is the mutation rate per gene per generation i.e 2.5×10^{-6} per nucleotide per generation (Ward *et al.*, 1991) $\times 1124$ sites results in $\mu = 2.81 \times 10^{-3}$.

The summary statistics; number of sequences, haplotype number, gene diversity (Nei, 1987) and nucleotide diversity (Nei, 1987), for each group were calculated in DnaSP v4.10 (Rozas *et al.*, 2003). Using DnaSP v4.10, the population mutation parameter (θ) was estimated from using segregating sites (θ_s per nucleotide site) as well as the Waterson estimator (W- θ_s per sequence) (Tajima, 1996). From W- θ_s the effective population size (N_e) was estimated by dividing W- θ_s with 2μ where μ is the mutation rate per gene per generation of 2.81×10^{-3} (Ward *et al.*, 1991) as explained in the previous paragraph. The

selective neutrality tests of Tajima's D (Tajima, 1989), Fu's F_s statistic (Fu, 1997) and the R_2 statistic (Ramos-Onsins and Rozas, 2002) were also calculated using DnaSP v4.10.

To visually represent the effective population size changes through time, Bayesian Skyline Plots (BSP) (Drummond *et al.*, 2005) were constructed. For each of the haplogroups, BSPs of effective population size through time were constructed using a Markov Chain Monte Carlo (MCMC) sampling algorithm, as implemented in BEAST v. 1.4.8 (Drummond and Rambaut, 2007). The population size function of the BSP can be implemented using either a piecewise constant or a piecewise linear function of population size change. In the present study, a piecewise linear model made up of 10 control points was used. The general time-reversible (GTR) substitution model with estimated base frequencies and a Gamma + Invariant Sites heterogeneity model was used to infer the ancestral gene trees for each haplogroup. The mean substitution rate was fixed to the rate of Ward *et al.*, (Ward *et al.*, 1991) and a relaxed molecular clock (Uncorrelated Lognormal) was employed. Each MCMC sampling was repeated for 40 000 000 generations, sampled every 4 000, with the first 4 000 000 generations discarded as burn-in. All runs had an effective sample size of at least 1 000 for the parameters of interest. Each independent run was repeated at least twice and results were combined using the LogCombiner v1.4.8 tool included in the BEAST package. BSPs were visualized in TRACER v. 1.4 (Rambaut and Drummond, 2007).

Population pairwise differences were calculated with Arlequin v3.11 (Excoffier *et al.*, 2005) by using F_{st} distances (Reynolds *et al.*, 1983) incorporating the nucleotide correction model of Tamura and Nei (Tamura and Nei, 1993) and a gamma correction of 0.532. An exact test of population differentiation (Raymond and Rousset, 1995) was also calculated using Arlequin v3.11 (Excoffier *et al.*, 2005). The distance matrix was visualized through PCA and cluster analysis in PAST v.1.54 (Hammer *et al.*, 2001b).

The relationship between physical and genetic distances were investigated in the Khoesan and Coloured groups by doing a linear regression using R v.2.5.0 (R-Project, 2006). The regression was applied on a scatter plot resulting from pairwise comparisons of distance matrices based on physical and genetic distances. The linear regression model tests a curved line, a straight line with a gradient and a straight line with a gradient of zero

against one another and assign significance values to each model. Additionally, a Mantel test implemented in Arlequin v3.11 (Excoffier *et al.*, 2005) was also done to test the correlation between the two distance matrices.

The physical distance matrix was constructed by obtaining latitude and longitude information of the different sampling locations from the website “Google Maps Latitude, Longitude Popup” (Gorissen, 2008) and calculating the great circle distance (in km) between the points using the “Latitude/Longitude Distance Calculation” website (Michels, 1997). The physical distance matrix is included in Appendix C.

Inter-population genetic distances were used in Analyses of Molecular Variance (AMOVA), implemented in Arlequin v.3.11 (Excoffier *et al.*, 2005). The distribution of variance among three hierarchical levels was tested in order to assess relationships among groups of populations. The lowest level is the variation contained between individuals within the same population. The next level contains the variation that exists between populations (populations in this case was the groups defined in Table 2.1). The third level contains the variation between groupings of these populations. Different groupings of populations were attempted, which were based on geographic distribution, language and self-identification of populations.

2.2.3 Y-chromosome methods

A total of 353 male samples were typed for Y-chromosome variation. Analyses of the Y-chromosomes were performed at two levels: firstly, haplogroup-defining bi-allelic markers were typed using restriction fragment length polymorphism (RFLP) assays or by using several SNaPshot minisequencing systems designed by the HGDDRU laboratory (Naidoo *et al.*, Unpublished). Secondly, microsatellite repeat-length analysis of short tandem repeat loci (Y-STRs) was done to determine intra haplogroup variation.

The RFLP assays were used initially and were gradually replaced as new minisequencing panels were developed and became available in the HGDDRU laboratory. The two

techniques combined use 83 polymorphisms in the non-recombining part Y-chromosome to assign individuals to 71 haplogroups. Figure 2.3 illustrate the Y-chromosome tree with nomenclature according to Karafet *et al.*, (Karafet *et al.*, 2008) and highlight the mutations used through different methods applied in the HGDDRU laboratory.

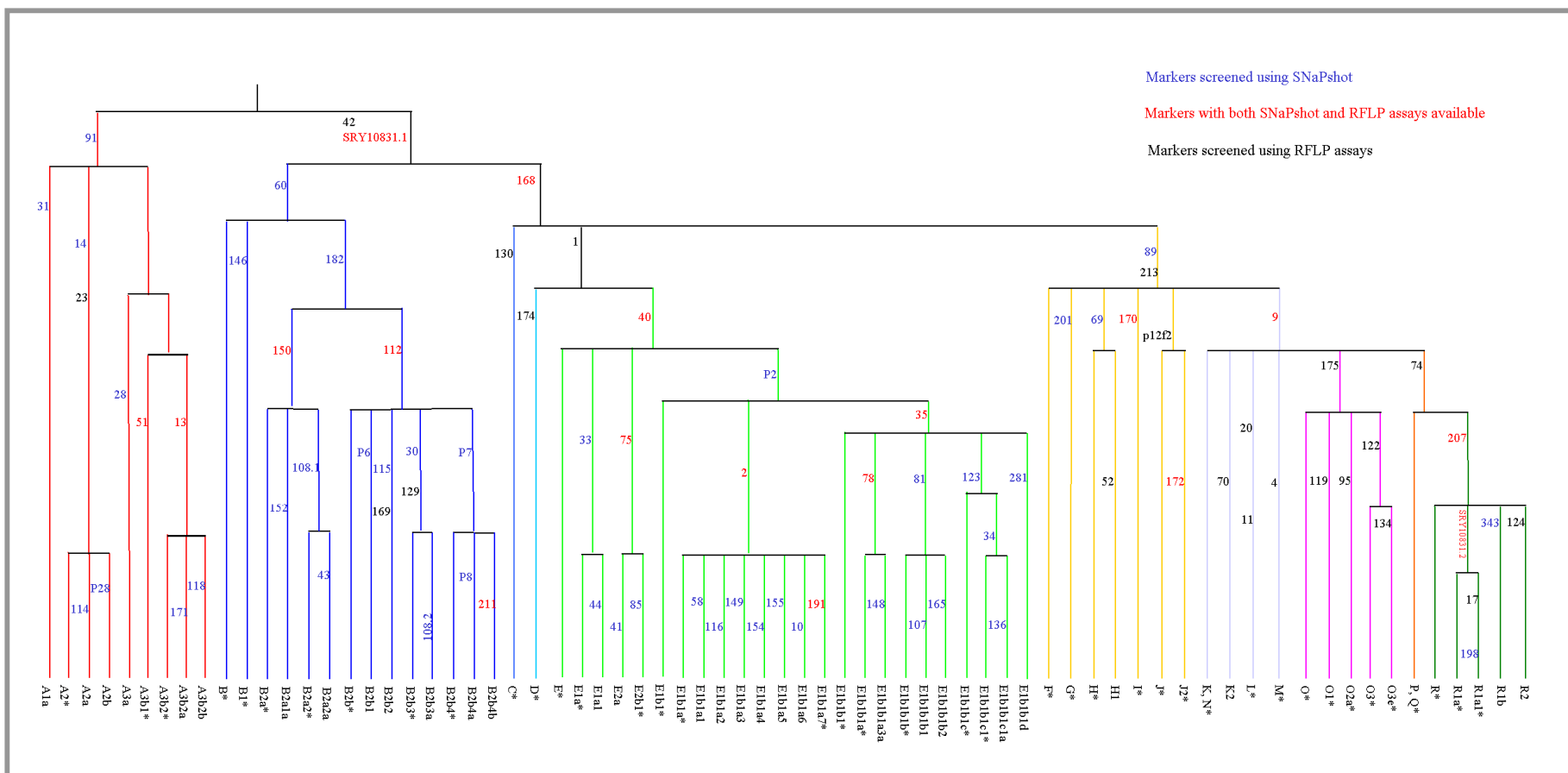


Figure 2.3 The Y-chromosome haplogroup tree with nomenclature according to Karafet *et al.*, (2008) indicating the branch-defining mutations screened for by using SNaPshot minisequencing panels and RFLP assays in the HGDDR laboratory

2.2.3.1 Y-chromosome RFLP

For the assignment of the 353 Y-chromosomes in the sample group a total of 24 of the Y-chromosome RFLPs were typed in a hierarchical manner (listed in Table 2.7) using the tree provided in Figure 2.3. The assays consisted of PCR amplification followed by restriction digests and separation on agarose gels.

The PCR amplification reaction had a total volume of 25 μ l, which contained ~50 ng DNA, 0.1 mM dNTPs, 1 U FastStart Taq polymerase and optimised amounts of $MgCl_2$, BSA, primer and spermidine (Table 2.8). All PCR's were performed on a 9700 GeneAmp® PCR System (Perkin-Elmer, PE Applied Biosystems), with the following thermal cycling conditions: Initial step at 95°C for 6 min followed by 35 cycles of denaturation at 95°C for 30 s, annealing at the appropriate temperature (Table 2.8) for 30 s and amplification at 72°C for 30 s; finishing off with a 7 min final extension at 72°C.

Restriction digests were done in a 30 μ l volume containing 25 μ l PCR product, 1x restriction enzyme buffer, 0.1 U restriction enzyme and in some cases (Table 2.8) added BSA (final concentration = 0.3 mg/ml). Digestion temperatures and reaction specific conditions are listed in Table 2.8. After digestion the fragments were separated on agarose gels of appropriate concentrations (Table 2.8) (1 x TBE running buffer, Bromophenol blue Ficolll dye loading buffer, 1Kb DNA ladder size standard (Gibco BRL). During every RFLP assay, control samples known to be ancestral and derived for the polymorphism were included, as well as a PCR blank control containing no DNA. Separated fragments were visualized under UV light and gel photographs were taken using the G:Box gel documentation system (Vacutec, SynGene, Cambridge, England) and GeneSnap v6.08 software (Synoptics Ltd., SynGene, Cambridge, England).

Table 2.7 SNPs typed in RFLP assays to determine Y-chromosome haplogroup

Initial typing	
SRY10831-1 (G-A)	
YAP (a/p)	
M213 (T-C)	
M168 (C-T)	
African groups	Eurasian groups
Haplogroup A	Haplogroup C
M51 (G-A)	M130 (C-T)
M23 (A-G)	Haplogroup R
Haplogroup B	M9 (C-G)
M112 (G-A)	M207 (A-G)
M150	SRY10831-2 (A-G)
M129 (G-A)	M17 (a/p - G)
M169 (T-C)	Other groups (J,C,I,H,L)
M211 (C-T)	p12f2 (a/p – 88 bp)
Haplogroup E	M172 (T-G)
M2 (A-G)	M52 (A-C)
M191 (T-G)	M170 (A-C)
M75 (G-A)	L-M11 (A-G)
M35 (G-C)	

a/p – absent or present

Table 2.8 Conditions and concentrations used during Y-chromosome RFLP typing

Marker	SRY10831	M51	M23	M168	M150	M112
Mutation	A-G; reversion G-A	G-A	A-G	C-T	C-T	G-A
Haplogroup(s) defined by derived state	SRY10831.1: B - R SRY10831.2: R	A - M51	A - M23	E - R	B - M150	B - M112
PCR stock solutions						
MgCl ₂ (25 mM)	1.5 mM	2 mM	2.5 mM	2.5 mM	1.5 mM	2 mM
primer F (10 uM)	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.4 uM
primer R (10 uM)	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.4 uM
BSA (10 mg/ml)	-	-	-	-	-	-
spermidine (2.5 mM)	-	-	-	-	-	-
PCR conditions	PCR in 50-ul vol.					
annealing temperature (°C)	60	58	58	56	60	61
Digestion	in 50-ul vol.; overlay with oil					
PCR product size (bp)	167	339	327	473	167	227
restriction enzyme	Dra III + BSA	Hind III	Xba I + BSA	Hinf I	Aat II	TspR I + BSA
digestion conditions (°C)	37	37	37	37	37	65
gel detection	3% agarose	2% agarose	2% agarose	3% agarose	3% agarose	2% agarose
ancestral allele - product sizes (bp)	167 (A)	339 (G)	173 + 154 (A)	234 + 105 + 81 + 52 (C)	146 + 21 (C)	155 + 72 (G)
derived allele - product sizes (bp)	112 + 55 (G)	307 + 32 (A)	327 (G)	234 + 186 + 52 (T)	167 (T)	227 (A)
Comments	reverse mut. G-A					
	in HG R					
References						
Reference: polymorphism	(Whitfield <i>et al.</i> , 1995)	(Underhill <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 2000)	(Shen <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 2000)
Reference: primers	(Santos <i>et al.</i> , 1999)	(Underhill <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 2000)	unpublished	unpublished
Reference: PCR-RFLP assay	(Santos <i>et al.</i> , 1999)	unpublished	unpublished	unpublished	unpublished mismatch primer	unpublished

Table 2.8 - continue Conditions and concentrations used during Y-chromosome RFLP typing

Marker	M129	M169	M211	M130 (RPS4Y)	YAP	M2 (sY81)
Mutation	G-A	T-C	C-T	C-T	absence - presence of YAP	A-G
Haplogroup(s) defined by derived state	B - M129	B - M169	B - M211	C - M130	D and E	E - M2
PCR stock solutions						
MgCl ₂ (25 mM)	1.5 mM	1.5 mM	2 mM	1.5 mM	1.5 mM	1.5 mM
primer F (10 uM)	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.2 uM
primer R (10 uM)	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.4 uM	0.2 uM
BSA (10 mg/ml)	-	-	-	-	1 ug/ul	-
spermidine (2.5 mM)	-	-	-	-	-	-
PCR conditions						
annealing temperature (°C)	62	58	58	50	51	58
Digestion						
PCR product size (bp)	255	200	208	91	150 (YAP-) / 450 (YAP+)	148
restriction enzyme	Msp I	Dra I (Roche)	Rsa I	Bsl I	-	Nla III + BSA
digestion conditions (°C)	37	37	37	55	-	37
gel detection	2% agarose	2% agarose	2% agarose	3% agarose	2% agarose	3% agarose
ancestral allele - product sizes (bp)	219 + 36 (G)	106 + 94 (T)	208 (C)	57 + 34 (C)	150 (YAP-)	105 + 43 (A)
derived allele - product sizes (bp)	255 (A)	200 (C)	137 + 71 (T)	91 (T)	450 (YAP+)	148 (G)
Comments						
References						
Reference: polymorphism	(Underhill <i>et al.</i> , 2000)	(Shen <i>et al.</i> , 2000)	(Shen <i>et al.</i> , 2000)	(Bergen <i>et al.</i> , 1999)	(Hammer, 1994)	(Seielstad <i>et al.</i> , 1994)
Reference: primers	(Underhill <i>et al.</i> , 2000)	unpublished	unpublished	(Kayser <i>et al.</i> , 2000)	(Hammer and Horai, 1995)	(Thomas <i>et al.</i> , 1999)
Reference: PCR-RFLP assay	unpublished	unpublished	unpublished	(Kayser <i>et al.</i> , 2000)	-	(Thomas <i>et al.</i> , 1999)

Table 2.8 - continue Conditions and concentrations used during Y-chromosome RFLP typing

Marker	M191b	M35	M75	M213	M170	M52
Mutation	T-G	G-C	G-A	T-C	A-C	A-C
Haplogroup(s) defined by derived state	E - M191	E - M35	E - M75	F - R	I - M170	H - M52
PCR stock solutions						
MgCl ₂ (25 mM)	2.5 mM	2 mM	2 mM	2.5 mM	4 mM	2.5 mM
primer F (10 uM)	0.3 uM	0.4 uM	0.3 uM	0.4 uM	0.4 uM	0.4 uM
primer R (10 uM)	0.3 uM	0.4 uM	0.3 uM	0.4 uM	0.4 uM	0.4 uM
BSA (10 mg/ml)	-	-	-	-	-	-
spermidine (2.5 mM)	-	-	-	-	-	-
PCR conditions						
annealing temperature (°C)	60	58	55	56	59	60
Digestion						
PCR product size (bp)	156	186	189	409	129	164
restriction enzyme	Mbo I	Bsr I	Nla III + BSA	Nla III + BSA	Nla III + BSA	HpyCH4 IV
digestion conditions (°C)	37	65	37	37	37	37
gel detection	3% agarose	2% agarose	3% agarose	2% agarose	3% agarose	3% agarose
ancestral allele - product sizes (bp)	156 (T)	122 + 64 (G)	189 (G)	290 + 119 (T)	109 + 20 (A)	164 (A)
derived allele - product sizes (bp)	129 + 27 (G)	186 (C)	165 + 24 (A)	409 (C)	129 (C)	138 + 26 (C)
Comments						
incomplete digestion						
References						
Reference: polymorphism	(Shen <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 2000)	(Shen <i>et al.</i> , 2000)	(Shen <i>et al.</i> , 2000)	(Shen <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 2000)
Reference: primers	unpublished	unpublished	unpublished	(Underhill <i>et al.</i> , 2001)	unpublished	unpublished
Reference: PCR-RFLP assay	unpublished	unpublished	unpublished	unpublished	unpublished	unpublished
	mismatch primer		mismatch primer		mismatch primer	mismatch primer

Table 2.8 - continue Conditions and concentrations used during Y-chromosome RFLP typing

Marker	p12f2	M172	M11	M9	M207	M17
Mutation	no del - del	T-G	A-G	C-G	A-G	WT-del G
Haplogroup(s) defined by derived state	J - p12f2 (Eu 10)	J - M172 (Eu 9)	L - M11	O - R	R - M207	R-M17
PCR stock solutions						
MgCl ₂ (25 mM)	1.5 mM	2 mM	3.5 mM	1.5 mM	1.5 mM	1.5 mM
primer F (10 uM)	0.3 uM	0.4 uM	0.4 uM	0.2 uM	0.4 uM	0.3 uM
primer R (10 uM)	0.3 uM	0.4 uM	0.4 uM	0.2 uM	0.4 uM	0.3 uM
BSA (10 mg/ml)	-	-	-	-	-	-
spermidine (2.5 mM)	-	-	-	-	-	-
	M2-F and M2-R (0.3 uM each)					
PCR conditions						
annealing temperature (°C)	58	58	58	54	56	56
Digestion						
PCR product size (bp)	p12f2 = 88; M2 = 148	148	215	340	423	124
restriction enzyme	-	Nla III + BSA	Msp I	Hinf I	Dra I (Roche)	Afl III
digestion conditions (°C)	-	37	37	37	37	37
gel detection	2% agarose	2% agarose	3% agarose	3% agarose	2% agarose	3% agarose
ancestral allele - product sizes (bp)	148 + 88 (no del)	148 (T)	215 (A)	181 + 95 + 64 (C)	356 + 77 (A)	124 (+G)
derived allele - product sizes (bp)	148 (del)	122 + 26 (G)	193 + 22 (G)	245 + 95 (G)	423 (G)	101 (-G)
	(co-amplification with M2)					
References						
Reference: polymorphism	(Casanova <i>et al.</i> , 1985)	(Shen <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 1997)	(Underhill <i>et al.</i> , 1997)	(Shen <i>et al.</i> , 2000)	(Underhill <i>et al.</i> , 1997)
Reference: primers	(Rosser <i>et al.</i> , 2000)	(Nebel <i>et al.</i> , 2001)	(Qamar <i>et al.</i> , 2002)	(Underhill <i>et al.</i> , 1997)	(Underhill <i>et al.</i> , 2001)	(Thomas <i>et al.</i> , 1999)
Reference: PCR-RFLP assay	-	(Nebel <i>et al.</i> , 2001)	(Qamar <i>et al.</i> , 2002)	unpublished	unpublished	(Thomas <i>et al.</i> , 1999)
		mismatch primer	mismatch primer			mismatch primer

2.2.3.2 Y-chromosome minisequencing

Seven Y-chromosome minisequencing panels were developed in the HGDDRU laboratory by T. Naidoo (Naidoo *et al.*, Unpublished). The “Y-SNP1” panel resolve some of the basal branches in the Y-chromosome tree (SRY10831.1, M168, M89) and thereafter targets Eurasian haplogroups (Figure 2.3 and Table 2.9). The other 6 panels concentrate on resolving African haplogroups. The “haplogroup E” panel resolve the main haplogroup E branches while the “E1b1a” and “E1b1b” panels focuses on these two common subgroups. The main branches of haplogroup B are resolved by the “Haplogroup B” panel and the “B2b” panel focus on the branches of the B2b subgroup. The subgroups of haplogroup A are fully resolved by one “Haplogroup A” panel (Figure 2.3 and Table 2.9).

The various SNPs in the minisequencing panels, their ancestral and derived states and their electropherogram profiles are listed in Table 2.9.

The methods for implementing these panels are according to T. Naidoo (Naidoo *et al.*, Unpublished). Each panel involves one multiplex PCR amplification followed by a PCR cleanup, minisequencing reactions with labelled ddNTPs, minisequencing reaction cleanup and analysis on a sequencer, similar to the mitochondrial minisequencing methodology described earlier.

Table 2.9 Information on the seven Y-chromosome minisequencing panels used to resolve haplogroups according to Figure 2.3

Marker name	Electropherogram peak number	Ancestral allele	Electropherogram label color	Derived allele	Electropherogram label color
Haplogroup A					
M91	1	T	RED	A	GREEN
M31	2	C	BLACK	G	BLUE
M14	3	A	GREEN	G	BLUE
M114	4	A	GREEN	G	BLUE
P28	5	C	BLACK	T	RED
M28	6	A	GREEN	C	BLACK
M51	7	C	BLACK	T	RED
M13	8	G	BLUE	C	BLACK
M171	9	C	BLACK	G	BLUE
M118	10	A	GREEN	T	RED
Haplogroup B					
M60	1	A	GREEN	T	RED
M146	2	T	RED	G	BLUE
M182	3	C	BLACK	T	RED
M150	4	C	BLACK	T	RED
M152	5	C	BLACK	T	RED
M108	6	A	GREEN	G	BLUE
M43	7	A	GREEN	G	BLUE
M112	8	G	BLUE	A	GREEN
Haplogroup B2b					
P6	1	G	BLUE	C	BLACK
M115	2	C	BLACK	T	RED
M30	3	G	BLUE	A	GREEN
P7	4	T	RED	C	BLACK
P8	5	G	BLUE	A	GREEN
M211	6	C	BLACK	T	RED
Haplogroup E					
M40	1	C	BLACK	T	RED
M33	2	A	GREEN	C	BLACK
M44	3	C	BLACK	G	BLUE
M75	4	G	BLUE	A	GREEN
M41	5	G	BLUE	T	RED
M85	6	G	BLUE	T	RED
P2	7	G	BLUE	A	GREEN
M2	8	T	RED	C	BLACK
M35	9	C	BLACK	G	BLUE
Haplogroup E1b1a					
M58	1	G	BLUE	A	GREEN
M116.1	2	A	GREEN	C	BLACK
M149	3	G	BLUE	A	GREEN
M154	4	A	GREEN	G	BLUE
M155	5	C	BLACK	T	RED
M10	6	T	RED	C	BLACK
M191	7	T	RED	G	BLUE
Haplogroup E1b1b					
M78	1	C	BLACK	T	RED
M148	2	T	RED	C	BLACK
M81	3	C	BLACK	T	RED
M107	4	A	GREEN	G	BLUE
M165	5	T	RED	C	BLACK
M123	6	G	BLUE	A	GREEN
M34	7	G	BLUE	T	RED
M136	8	G	BLUE	A	GREEN
M281	9	C	BLACK	T	RED
Y-SNP1					
SRY10831	2	A	GREEN	G	BLUE
M168	3	C	BLACK	T	RED
M89	4	G	BLUE	A	GREEN
M201	5	G	BLUE	T	RED
M69	6	T	RED	C	BLACK
M170	7	A	GREEN	C	BLACK
M172	8	T	RED	G	BLUE
M9	9	C	BLACK	G	BLUE
M207	10	A	GREEN	G	BLUE
M198	11	C	BLACK	T	RED
M343	12	C	BLACK	A	GREEN

2.2.3.3 Y-chromosome STR

Twelve Y-chromosome STRs on the Y-chromosome non-recombining region were typed using the PowerPlex® Y System (Promega) according to kit instructions with some modification. The kit allows for the co-amplification of 12 Y-STR loci (DYS19, DYS385a/b, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439) in a single multiplex reaction. The multiplex reaction was performed in a total volume of 6.25 µl, including 5 ng DNA, 1 x PowerPlex® Y Buffer, 1 x PowerPlex® Y Primer mix, 5 U FastStart Taq polymerase (Roche Applied Science) and ddH₂O to make up the reaction mix. Thermal cycling conditions are shown in Table 2.10

1 µl of the multiplex PCR reaction product was mixed with 8.5 µl Hi-Di formamide (Applied Biosystems) and 0.5 µl of ILS600 internal size standard (Promega). After a denaturing step for 2 min at 95°C followed by cooling to 4°C the fragments were separated on an ABI PRISM® 3130xl Genetic Analyzer (Applied Biosystems) according to PowerPlex® Y System Kit instructions and analysed using GeneMapperID v3.2 software (Applied Biosystems).

Table 2.10 Y-STR PCR Thermal Cycler Conditions

Temperature (and Ramp Speed)	Time (min:sec)	
95°C	11:00	
96°C	01:00	
94°C (ramp 100%)	00:30	} 10 cycles
60°C (ramp 29%)	00:30	
70°C (ramp 23%)	00:45	
90°C (ramp 100%)	00:30	} 20 cycles
58°C (ramp 29%)	00:30	
70°C (ramp 23%)	00:45	
60°C	30:00	
4°C	∞	

2.2.3.4 Y-chromosome data analysis

Using the bi-allelic polymorphisms the Y-chromosomes in the study group were allocated to haplogroups according to the nomenclature of Karafet *et al.*, (Karafet *et al.*, 2008). Intra-population Y-chromosome variation was calculated using STR-haplotypes to infer Gene Diversities in Arlequin v3.11 (Excoffier *et al.*, 2005).

Networks of STR haplotypes were constructed using the Median Joining algorithm (Bandelt *et al.*, 1999) of Network v4.5.0.0 (Fluxus-engineering, 2008). Networks were subjected to maximum parsimony post-analysis using the Steiner maximum parsimony (MP) algorithm (Polzin and Daneschmand, 2003) within Network 4.5.0.0. For network analysis the epsilon parameter was set to 0 and the median vector criterion was set to “Connection Cost”. Loci were not weighted differently but repeats at the DYS389II locus were modified. DYS389 is a composite locus that contains regions that are phylogenetically informative as well as fast evolving regions that obscure phylogenetic structure. To alleviate this problem DYS389I was subtracted from 389II to give DYS389c, this excludes some of the uninformative data. In further analyses 389I and 389c are then used.

TMRCAs for the haplogroups were estimated from the median joining networks using a mutation rate of $\mu = 6.9 \times 10^{-4}$ per locus per generation with a generation time of 25 years (Zhivotovsky *et al.*, 2004).

Individual STR variation was also subjected to distance analysis using the $\delta\mu^2$ distance measure (Goldstein *et al.*, 1995), as employed in Populations v.1.2.30 (Langella, 2002). The $\delta\mu^2$ statistic is a genetic distance specifically developed for microsatellite loci, incorporating features of the stepwise mutation model. Distance matrices for all haplogroups were calculated and used to construct Neighbour Joining (NJ) trees in MEGA4 (Tamura *et al.*, 2007) and Multidimensional Scaling (MDS) plots in PAST v.1.54 (Hammer *et al.*, 2001b).

To visualize haplogroup frequency distributions, haplogroup isofrequency maps were generated applying the Kriging method (Oliver and Webster, 1990; Xue *et al.*, 2005) incorporated in the Surfer v.8.06.39 program (Golden-Software, 2006).

The relationships among the population groups were analysed using haplogroup frequency data as well as STR haplotype data. Inter population distances were calculated by generating F_{st} distance matrices from haplotype frequency data and R_{st} (Slatkin, 1995) distance matrices for STR haplotypes, both calculated in Arlequin v3.11 (Excoffier *et al.*, 2005). For both data types, population differentiation were examined using an exact test (Raymond and Rousset, 1995) implemented in the Arlequin v3.11. The matrices were visualized through PCA plots and cluster analysis in PAST v.1.54 (Hammer *et al.*, 2001b). The correlation between the two matrices was tested through a Mantel test applied in Arlequin v3.11.

The two genetic distance matrices were also compared to a physical distance matrix (Appendix C). This was accomplished by doing a linear regression using R v.2.5.0 (R-Project, 2006) on the scatter plot resulting from pairwise comparisons of distance matrices based on physical and genetic distances. A Mantel test implemented in Arlequin v3.11 (Excoffier *et al.*, 2005) was also performed to test the correlation between the two genetic distance matrices and the physical distance matrix.

The F_{st} and R_{st} distance matrices were both used in AMOVA analysis, performed in Arlequin v3.11 (Excoffier *et al.*, 2005). The distribution of variance among three hierarchical levels was tested in order to assess relationships among groups of populations. The lowest level is the variation contained between individuals within the same population. The next level contains the variation that exists between populations (populations in this case were the groups defined in Table 2.1). The third level contains the variation between groupings of these populations. Different groupings of populations were attempted, which were based on geographic distribution, language and self-identification of populations.

2.2.4 Autosomal SNP methods

To analyse information contained in the autosomes, 220 autosomal SNPs were specifically selected in the following way:

- 10 SNPs per chromosome (chromosome 1 to 22) were selected.
- The 10 SNPs were selected in two groups of 5 linked SNPs.
- The two groups were completely unlinked from one another (Figure 2.4).

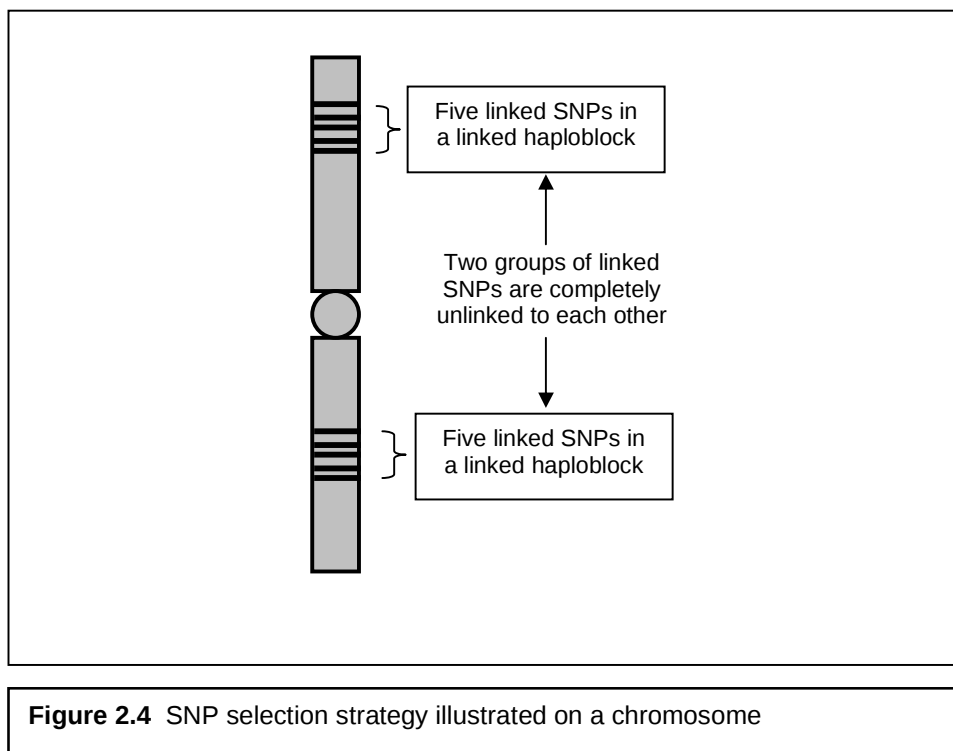


Figure 2.4 SNP selection strategy illustrated on a chromosome

The five SNPs in the five SNP group were selected to be on the same haploblock. To select the SNPs the software SNPbrowserTM v3.1 (Applied Biosystems) were used and both Hapmap and Applied Biosystems (ABI) SNP databases were considered. In the ABI database, haplotype blocks from the African American study group were considered and in the Hapmap database, haplotype blocks from the Yoruba study group were considered. None of these two study groups are Khoe-San but these were the closest related population groups from which sufficient SNP data was available at the time. SNPs were selected to be on the same haploblock in the Yoruba group and preferentially also on the same haploblock in the African American group. The average distance between

consecutive selected SNPs in the same haploblock was 4347 bp (SD =3730.8 ; MIN = 192 bp ; MAX = 22332 bp). The haploblocks that contained the SNPs were not associated with any known coding part of the genome. Therefore neutral genetic variation was targeted and influence of selection minimized. Furthermore SNPs were selected to have a minor allele frequency above 10% in the African population groups in order to try and select SNPs that contain polymorphisms in African populations. The full details for selected SNPs (their chromosome number, the group they sorted into, their alternate name for the analyses, position on chromosome, distances from other SNPs and minimum allele frequencies in the Yoruba and African American groups), are listed in Appendix C.

SNPs were selected in this fashion to allow for multiple types of analyses using the same dataset. Firstly the selection allows for the compilation of multiple different genotype sets with 44 unlinked polymorphisms in each, by selecting one SNP per SNP-group. Analyses of these different SNP sets of 44 SNPs can then be compared with one another to see if similar results are obtained. If 100 such sets are selected, analysed and compared it will be the same as 100 separate studies with 44 unlinked SNPs in each. Furthermore if haplotypes are inferred (using haplotype inferring software) for SNPs in the same haploblock, these haplotypes can be used in different analyses than that utilized for unlinked genotypic SNPs.

All autosomal SNPs were typed by a commercial company (Harvard-Partners Center for Genetics and Genomics, Genotyping Facility, Cambridge, Massachusetts, United States). The company used Sequenom iPLEX SNP genotyping, which allows interrogation of up to 40 assays in one well of a 384 well plate and therefore reduces the per genotype cost. The technique is based on a multiplexed PCR followed by a minisequencing reaction in a single well. iPLEX chemistry involves the extension of minisequencing probes by a single mass-modified dideoxynucleotide using a proprietary enzyme from Sequenom. The size of reaction products is determined directly by MALDI-TOF mass spectrometry, yielding genotype information. Specialized equipment for this work includes a Pre- and Post-PCR Biomek FX liquid handling system, a Multimek liquid handler, a nanoliter plotting robot for spotting the extension products onto chips, and a Bruker Compact mass spectrometer.

Multiplex PCR assays were designed by the company using Sequenom SpectroDESIGNER software (version 3.0.0.3) by inputting sequence containing the SNP site and 100 bp of flanking sequence on either side of the SNP. The SNPs are grouped into multiplexes so that the extended product does not overlap in mass with any other oligonucleotide present in the reaction mix, and where no primer-primer, primer-product non-specific interactions will occur.

Resultant SNP data were downloaded from the company web-based database and edited into formats suitable for computational analyses. Seven of the 220 loci were discarded because of poor assay quality (indicated in Appendix C).

2.2.4.1 Autosomal SNP data analysis (Genotypic)

The panel of 220 SNPs (consisting of two unlinked groups of five linked SNPs per chromosome) was used to generate 100 different random combinations of 44 unlinked SNPs.

The proportion of polymorphic loci, heterozygosity (Weir, 1996a) and gene diversity (Weir, 1996b) of each of the 100 different SNP datasets were calculated for each of the 14 populations analysed as well as for the total dataset using GDA v 1.0 (Lewis and Zaykin, 2001). The averages as well as the standard deviation of these three summary statistics were then calculated across the 100 datasets. The heterozygosity estimate is the proportion of heterozygous individuals in the population and Gene diversity (often referred to as expected heterozygosity), is defined as the probability that two randomly chosen alleles from the population are different.

To test if there was a correlation between the variation between the different runs (standard deviation) and the average heterozygosity of each population, a scatter plot was generated with the average heterozygosity of each population on the Y-axis and the standard deviation (SD) between the heterozygosities in the different datasets on the X-axis. Using R v.2.5.0 (R-Project, 2006), a linear regression was done to find the function that best described the relationship between the points.

Analysis of population structure on the 100 different SNP sets was done using a K-means clustering approach implemented in STRUCTURE v2.2 (Pritchard *et al.*, 2000; Falush *et al.*, 2003; Falush *et al.*, 2007). The STRUCTURE analysis of the 100 sets of 44 SNP were conducted as follows: STRUCTURE runs with 10 iterations at K=1 to K=10 were conducted with a burn-in of 50K and repeats of a 100K for each set. Allele frequencies were correlated and a model with admixture was assumed for all runs. The 10 iterations at each K for each of the 100 SNP sets were then collapsed into 1 consensus run using CLUMPP (Jakobsson and Rosenberg, 2007). Thereafter the 100 sets of random SNPs were collapsed into a consensus run at each K using CLUMPP v1.1.1. Results were visualized using DISTRUCT (Rosenberg, 2002). Figure 2.5 illustrates this process.

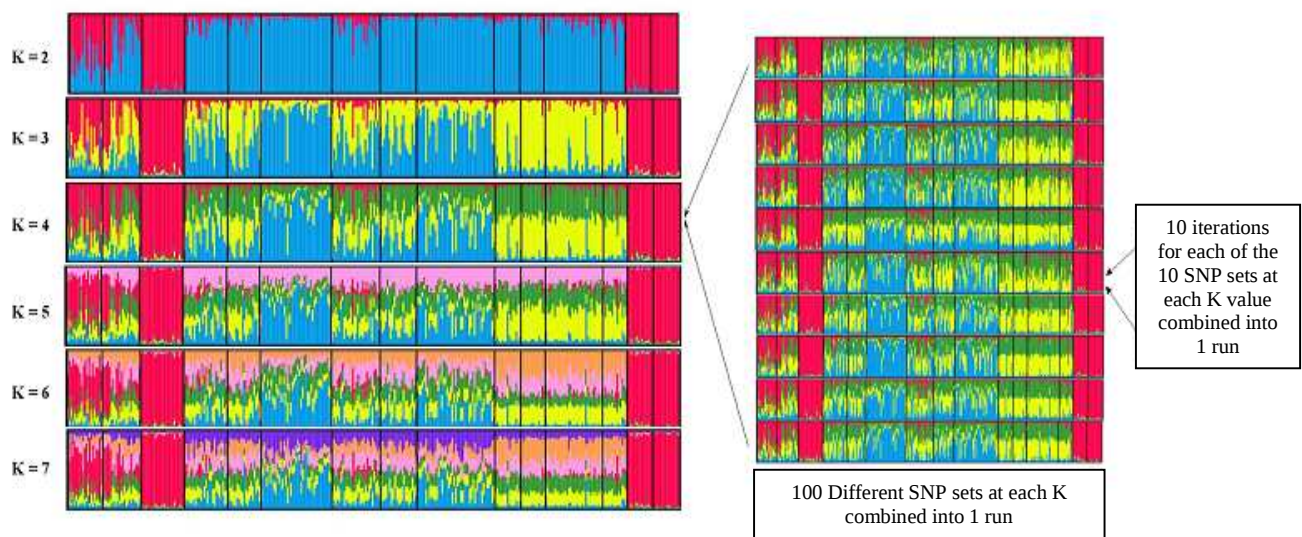


Figure 2.5 Diagram illustrating how STRUCTURE results for 100 SNP sets were condensed into one consensus run. Starting at the right with 10 x 100 sample sets for each K value. The 10 iterations for each of the 100 runs were condensed into one run leaving 100 different SNP sets at each K. These 100 different SNP sets at each K were then condensed into one result at each K value.

The K value with the highest average likelihood and the highest delta K value (Evanno *et al.*, 2005) were calculated and compared to one another to identify the best cluster assignment.

Results of assignments of K=3 were plotted on a triangle plot using R v.2.5.0 (R-Project, 2006), while incorporating the ADE4 R-library.

The variation of assignment to K clusters of the 100 different runs was compared to one another. The mean population assignments of each run at K=2 to K=5 were plotted on a graph using R v.2.5.0 to graphically illustrate the variations between runs. Furthermore the differences between the runs were tested by doing pairwise correlations between runs at K=3 using Pearson's correlation coefficient (r) implemented in PAST v.1.54 (Hammer *et al.*, 2001b).

The same 100 SNP datasets used in the STRUCTURE analysis was also used in distance based analysis. To construct inter-population distance matrices of each of the 100 datasets, Reynolds distance (Reynolds *et al.*, 1983) was used as implemented in Powermarker v3.25 (Liu and Muse, 2005). To condense the 100 different population distance matrices into one output, two alternate approaches were followed. In the first case 100 different NJ trees were constructed in Powermarker v3.25 and were then condensed into one Majority Rule consensus tree using CONSENSE implemented in PHYLIP v.3.65 (Felsenstein, 2004). The tree was then visualized in Dendroscope (Huson *et al.*, 2007). In the second approach an average distance matrix was calculated by taking the average of each pairwise comparison in the 100 distance matrices. This average distance matrix was then used to construct a NJ tree using NEIGHBOUR, implemented in PHYLIP v.3.65 (Felsenstein, 2004). The average population distance matrix was also further used to do PCA in PAST v.1.54 (Hammer *et al.*, 2001b).

The NJ tree consensus tree is useful in illustrating the number of times that a particular branch is supported by the 100 separate trees but the branch length of the tree is not an indication of distance between populations. The average distance matrix consensus tree on the other hand does not tell us the number of times a particular branch is supported by the 100 distance matrices but gives us a good indication of the distances between populations through mean branch lengths.

Inter-individual pairwise distance matrices for the 352 individuals in the 100 different SNP datasets was also constructed using Reynolds distance (Reynolds *et al.*, 1983) in Powermarker v3.25 (Liu and Muse, 2005). The average of the 100 individual distance matrices was calculated by taking the average of each pairwise comparison. The average individual distance matrix was then used for PCA in PAST v.1.54 (Hammer *et al.*, 2001b).

To investigate the relationship of physical distance and genetic distance using autosomal SNPs in the Khoe-San and Coloured populations the composite distance matrix of the 100 datasets (Reynolds distance) (Reynolds *et al.*, 1983) was compared to a physical distance matrix (Appendix C). Pairwise comparisons between physical distance (X-axis) and genetic distance Y-axis was plotted on graphs and a linear regression was done using R v.2.5.0 (R-Project, 2006) to determine the line with the best fit through the points. A Mantel test implemented in Arlequin v.3.11 (Excoffier *et al.*, 2005) was also done to test the correlation between the two distance matrices.

Five random sets from the 100 datasets were chosen to do AMOVA, implemented in Arlequin v3.11 (Excoffier *et al.*, 2005). The average values of the five sets were reported. The distribution of variance among three hierarchical levels was tested in order to assess relationships among groups of populations. The lowest level is the variation contained between individuals within the same population. The next level contains the variation that exists between populations (populations in this case was the groups defined in Table 2.1). The third level contains the variation between groupings of these populations. Different groupings of populations were attempted, which were based on geographic distribution, language and self-identification of populations.

2.2.4.2 Autosomal SNP data analysis (Haplotypic)

For haplotype analysis the five linked SNPs on the same haploblock was used to infer 44 haplotypes consisting of 5 bp each. The haplotypes were inferred separately for each population and each SNP set of 5 using Powermarker v3.25 (Liu and Muse, 2005). The frequencies of the different types of haplotypes in the 44 haplotype loci were calculated in Powermarker v3.25 (Liu and Muse, 2005) and represented in the form of bar charts using Microsoft Excel.

To try and condense the information from the 44 separate haplotype loci two approaches were followed. In the first approach the 88 haplotypes (2 at each loci) of each individual was concatenated into two haplotypes for each individual. The order in which two alleles of the same locus is combined with the two alleles of any other loci will not be important since the alleles at the 44 loci segregate independently in the population. Individuals with >50% missing data at any of the 5-SNP loci were excluded from the analysis. Of the 352 individuals, 298 remained and therefore 596 haplotypes. Since some of the loci were very polymorphic and contained many different haplotypes, the combination of several such loci will lead to high haplotype diversities. Concatenating haplotypes in individuals led to 594 unique haplotypes in the total of 596 haplotypes. The individual haplotypes were then used to construct distance matrices. Both population and individual distance matrices were constructed using the Maximum composite likelihood algorithm in MEGA4 (Tamura *et al.*, 2007). These distance matrices were then used for PCA and cluster analysis in PAST v.1.54 (Hammer *et al.*, 2001b).

In the second approach only the 44 small haplotypes with the highest frequency in each specific population at each of the 44 loci were selected. This was then taken as the 44 representing small haplotypes of each population. The 44 small haplotypes were then concatenated into one haplotype sequence for each population. These 14 population representative sequences were then used to construct a distance matrix using the Maximum composite likelihood method in MEGA4 (Tamura *et al.*, 2007). The distance matrix was then used for PCA and cluster analysis in PAST v.1.54 (Hammer *et al.*, 2001b).

The 44 separate haplotypes that are concatenated into one haplotype will have different evolutionary histories and a single unique tree will not best characterize the phylogenetic representation of the haplotype. An approach should rather be followed where data are not forced into a single tree. An approach that employs this strategy is the Neighbour-Net method (Bryant and Moulton, 2002). Data are decomposed into several splits and represented in the form of a splits graph. Ideal data will yield a tree but data that do not support a single unique tree will yield a tree-like network representing different incompatible phylogenies. Although this method does not force data into a single tree it

gives a good indication how tree-like a dataset is. The dataset with the single representative haplotypes of each population was used to generate a Neighbour-Net network using SplitsTree4 (Huson and Bryant, 2006).

The population matrices of the two different approaches were also used to test the relationship of physical distance and genetic distance using autosomal SNP haplotypes in the Khoe-San and Coloured populations. The two genetic distance matrices were compared to a physical distance matrix (Appendix C). Pairwise comparisons between physical distance (X-axis) and genetic distance (Y-axis) were plotted on graphs and a linear regression was done using R v.2.5.0 (R-Project, 2006) to determine the line with the best fit through the points. A Mantel test implemented in Arlequin v.3.11 (Excoffier *et al.*, 2005) was also applied to test the correlation between the distance matrices.

Linkage Disequilibrium (LD) analyses would have been interesting but was not done for the present study since the marker coverage was very low. High resolution SNP typing of Khoe and San individuals is in process and these studies would give a much better picture of the LD patterns in the Khoe-San.

3. MITOCHONDRIAL-DNA STUDIES

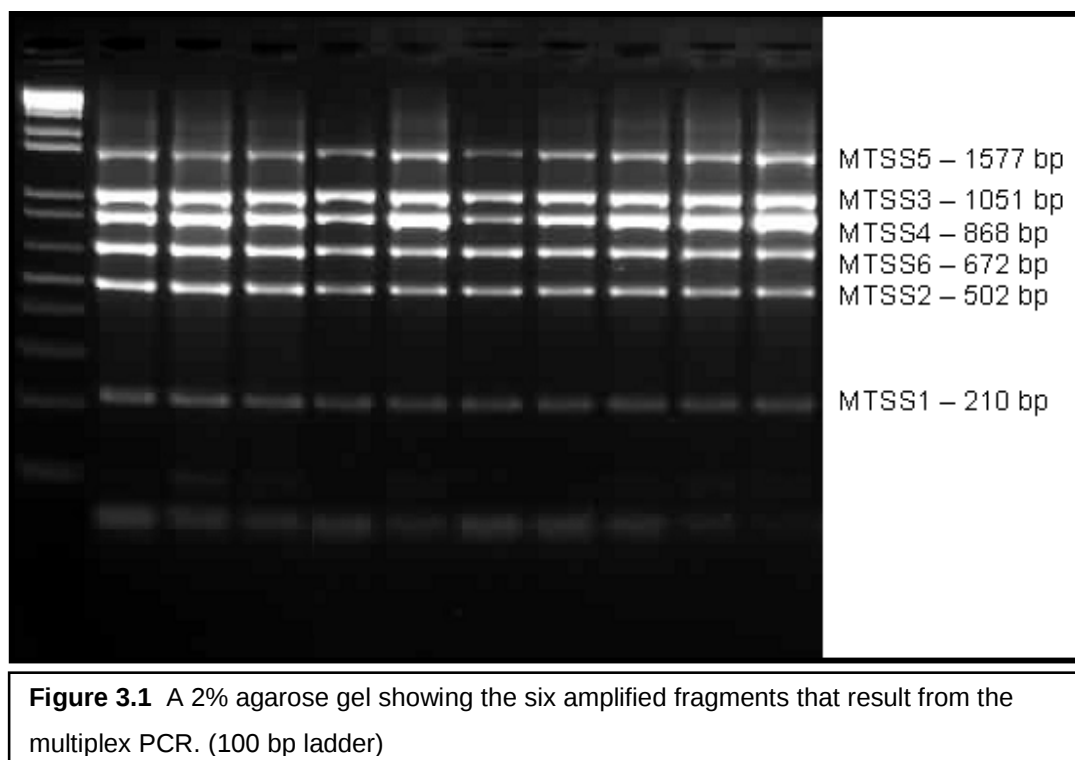
Only few studies thus far concentrated on studying the maternal genetic history of Khoe-San people (Vigilant *et al.*, 1991; Chen *et al.*, 2000; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). These studies covered only three groups of San people, including, the two Ju speaking groups: the !Xun that were originally from Angola (now located in Platfontein, SA) and the linguistically closely related Ju!’hoansi (from northern Botswana and Namibia) and the Khoe-speaking San group, the Khwe (also originally from Angola but now located in Platfontein, SA). All three of the San groups were originally from either Angola or northern Namibia, positioning them in the northern parts of the original distribution of Khoe-San people. This leaves a gap with no studies being done on groups’ representative of the southern San people and the Khoe people. Furthermore, studies published thus far concentrated on studying the L0d and L0k lineages in the San groups as a whole, without looking into the unique histories and distributions of the L0d sub-haplogroups. Data collected in this study have facilitated an understanding of the sub-structure of the L0d and L0k haplogroups and their distribution among various additional groups with Khoe and San ancestry. The following sections will present the analyses of this dataset, compare results to published data and discuss the relevancy to Khoe-San history.

In the first part of this chapter the results from the minisequencing protocol, which have now been published (Schlebusch *et al.*, 2009), will be provided. Thereafter results from the analysis of the HVS will be presented. Firstly the haplogroups assignments, phylogenetic trees and networks assembled from the sequence data will be shown. Subsequently the further analyses of the distribution of L0d/k subgroups (in the form of isofrequency maps) and the analyses of haplogroup expansion and contraction signals (in the form of summary statistics mismatch distributions and Bayesian Skyline Plots (BSPs)) will be provided. All results regarding specific L0d/k subgroups will thereafter be discussed in detail. Next, the results regarding the genetic relationships between the different Khoe-San groups included in the study will be presented and discussed.

3.1 Minisequencing

A minisequencing protocol was designed to distinguish between the seven African mtDNA macro-haplogroups (L0-L6) as well as the three non-African macro-haplogroups M, N and R (Figure 2.2). The panel types 14 SNPs that define these 10 macro-haplogroups. The panel was validated by successfully screening 699 individuals and assigning them into their correct macro-haplogroups. These comprised 538 individuals included for mitochondrial HVS-I and II analysis plus 161 additional individuals as outlined in Chapter 2. Results were compared to HVS based classification using a phylogenetic approach and no inconsistencies were found (Table 3.1).

The PCR amplification of the regions that encompass the 14 SNPs were optimized into one multiplex reaction that amplify these regions in six amplicons of various sizes (Figure 3.1). After the minisequencing reaction the products are separated on the genetic analyzer and an electropherogram displays the different sized products (Figure 3.2 give example electropherograms for haplogroups L0, L1, L3 and M). Due to differences in the electrophoretic mobility the detected band size on the electropherogram differs from the real size (Table 2.4) for certain products. These differences in electrophoretic mobility are influenced by the length of the sequence, the nucleotide composition and the dye that labels the extended primer. The effect of nucleotide composition generally has a higher influence on the shorter fragments. The difference in our actual sizes was at least five nucleotides between successive bands. Even with this difference, however, certain bands still migrated on top of one another. This, however, did not affect our classification of the sequences.



As can be seen in Figure 3.2, bands 2 and 3 (resolving branches L1-6 and L3'4 according to tree in Figure 2.2) migrated on top of one another but because of the difference in the color of the dye for the two polymorphisms (green/blue for L1-6 and red/black for L3'4) they are easily distinguished. Furthermore bands 11 and 12 (resolving haplogroups R and L1 according to tree in Figure 2.2) occasionally migrated on top of one another. When both R and L1 are ancestral there are two green peaks, which in some cases cannot be distinguished (see L3 in Figure 3.2). When either is derived, however, a clear blue peak becomes visible, as can be seen in the L1 picture in Figure 3.2. When the two green peaks appear on top of one another there might be the problem of a null allele in one of the bands, this problem is, however, overcome by the fact that the other bands in the hierarchical typing confirms the position of the two polymorphisms. The final two peaks, 13 and 14 (resolving haplogroups M and L6 according to tree in Figure 2.2), also occasionally overlapped (shown in Figure 3.2; L0). This results in the presence of a single red peak (instead of two). This is further exacerbated by the low peak amplitude of the M peak. When M is derived, however, the black peak can clearly be seen (Figure 3.2; M). As with the case of L1 and R, the presence of other peaks will hierarchically confirm the presence

and state of the M and L6 SNPs. In order to resolve the separation issues at peaks 11 and 12, and 13 and 14, it would be practical to add one to three bases to the tails of the L1 and L6 primers, thereby changing their mobility.

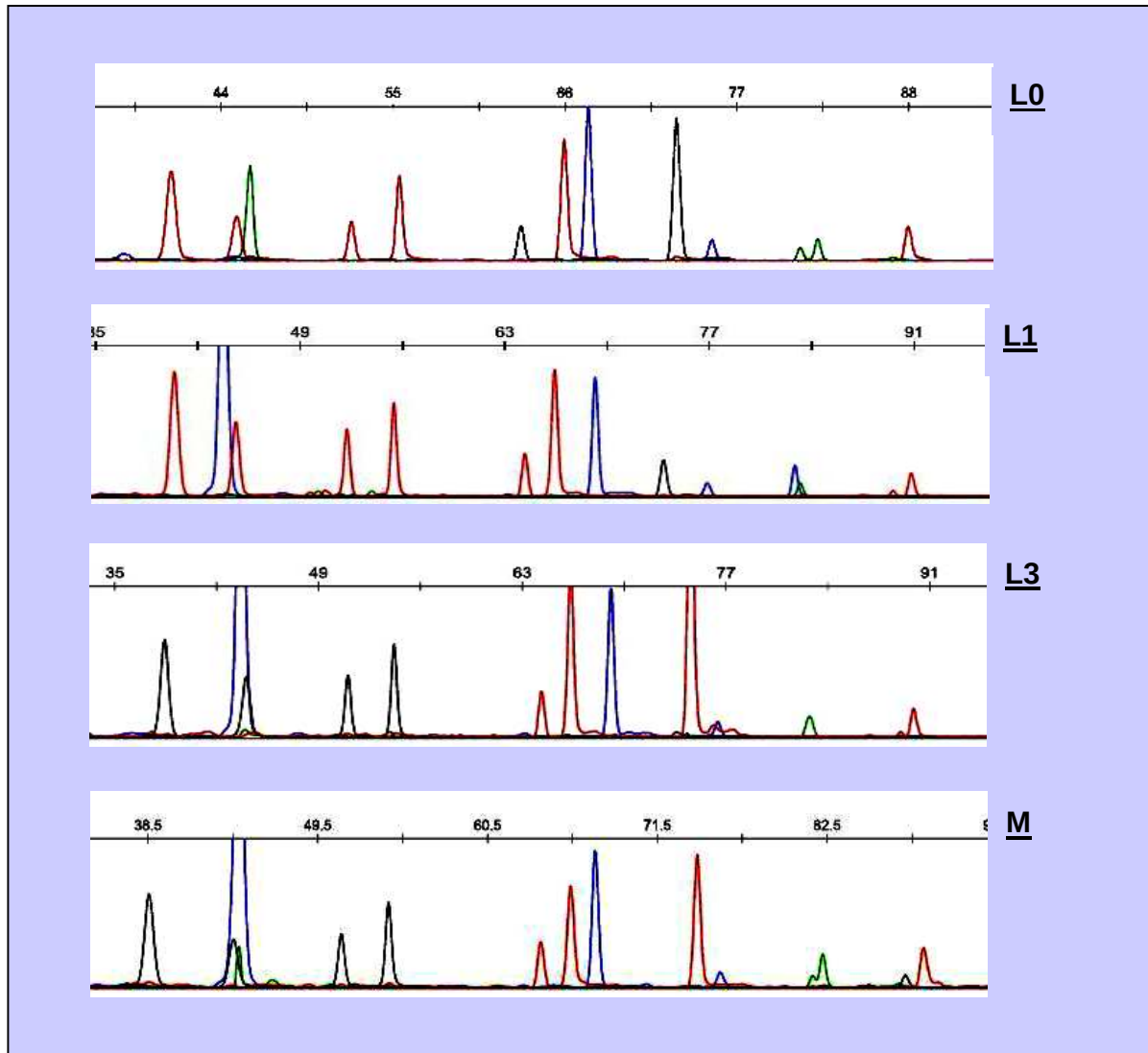


Figure 3.2 Electropherogram examples showing peak profiles of haplogroups L0, L1, L3 and M.

Peaks from left to right, in format:

“polymorphism position (defined haplogroup according to Fig. 2.2 and Table 2.4)” are:

1018G-A (L3), 1048C-T (L1-6), 7256C-T (L3'4'), 7521G-A (L3'4'6'), 8468C-T (L2-6), 9347A-G (L0), 10115T-C (L2), 10398A-G (N), 10810T-C (L2'3'4'6'), 12432C-T (L5), 12705C-T (R), 13789T-C (L1), 14783T-C (M), 15289T-C (L6)

Notwithstanding these minor problems in the panel, 699 individuals were successfully classified to their correct macro-haplogroups. Haplogrouping, based on HVS variation using a phylogenetic approach, were compared to the minisequencing coding region classification and no inconsistencies were found. Table 3.1 summarises the haplogroup classification based on HVS sequences and the macro-haplogroup classification based on the minisequencing coding region classification. There were a few instances where one of the bands in the profile failed or reverted. The band failure was most probably due to polymorphisms in the primer binding sites and the bands that failed were phylogenetically very specific (Table 3.1). None of this, however, had any affect on the classification of the implicated sequences.

Table 3.1 Results of the minisequencing screening and classification of 699 sequences compared to classification based on HVS sequences

MtDNA haplogroup based on HVS sequence	Number of sequences	Macro-Haplogroup identified using minisequencing	Number of sequences	Problems observed during screening
L0a	40	L0	447	► L5 peak fail in haplogroup L0d1b (40% of L0d1b sequences) ► L2 peak fail in L0d3, due to 10114C mutation in L0d3 (92% L0d3 sequences); also occurs in singletons of L0d1c1 and L0d2a
L0d	372			
L0k	35			
L1b	3	L1	18	► L1-6 mutation is positive (incorrectly) in L1c2 (3 sequences)
L1c	15			
L2*	1	L2	71	► L0 peak fail in L2b3 (1 sequence)
L2a	63			
L2b	5			
L2c	2			
L3b	1	L3	92	
L3c	1			
L3d	44			
L3e	38			
L3f	8			
L4	7	L4	7	
L5	3	L5	3	
M	18	M	18	
N	6	N	6	
R	37	R	37	► L2 is positive (incorrectly) in two of the sequences of haplogroup H with the same HVS profile (2 sequences)
Total	699	Total	699	

3.2 HVS-I and II variation

The 538 samples used in mitochondrial analysis were first classified into macro-haplogroups using the minisequencing method. Further, finer scale classification was achieved by analysing HVS-I and II sequences.

A total of 1124 bp in a combined HVS-I and II were analysed (HVS-I: positions 15997-16569 and HVS-II: positions 57-607). There were 205 (18.2%) variable positions in the combined sequence; HVS-I had 122 (21.3%) variable sites while HVS-II had 83 (15.1%). Fourteen sites had three different alleles and seven sites had four different alleles (16093, 16188, 16265, 16266, 16286, 16291, 16293). The transversion : transition ratio was 1 : 5.6. Insertions were observed at four sites (291, 455, 523, 573) while deletions occurred at seven sites (16183, 16179, 16325, 247, 249, 498, 523). All deletions involved 1 bp except the 523 region of HVS-II that contain an “AC” repeat motif that were inserted or deleted in several sequences. Insertions involved 1 bp insertions at 291 and 455, one sequence had a 2 bp insertion at 455. All insertions in the poly C repeat track at position 568-573 were taken as a 1 bp C insertion.

Using the coding polymorphisms implemented in the minisequencing procedure as well as the 205 variable sites from HVS-I and II, the 538 sequences were classified into 18 haplogroups encompassing 245 haplotypes (Figure 3.3 and 3.4). A full haplotype list with HVS-I and II variant sites and their population assignment is included in Appendix E.

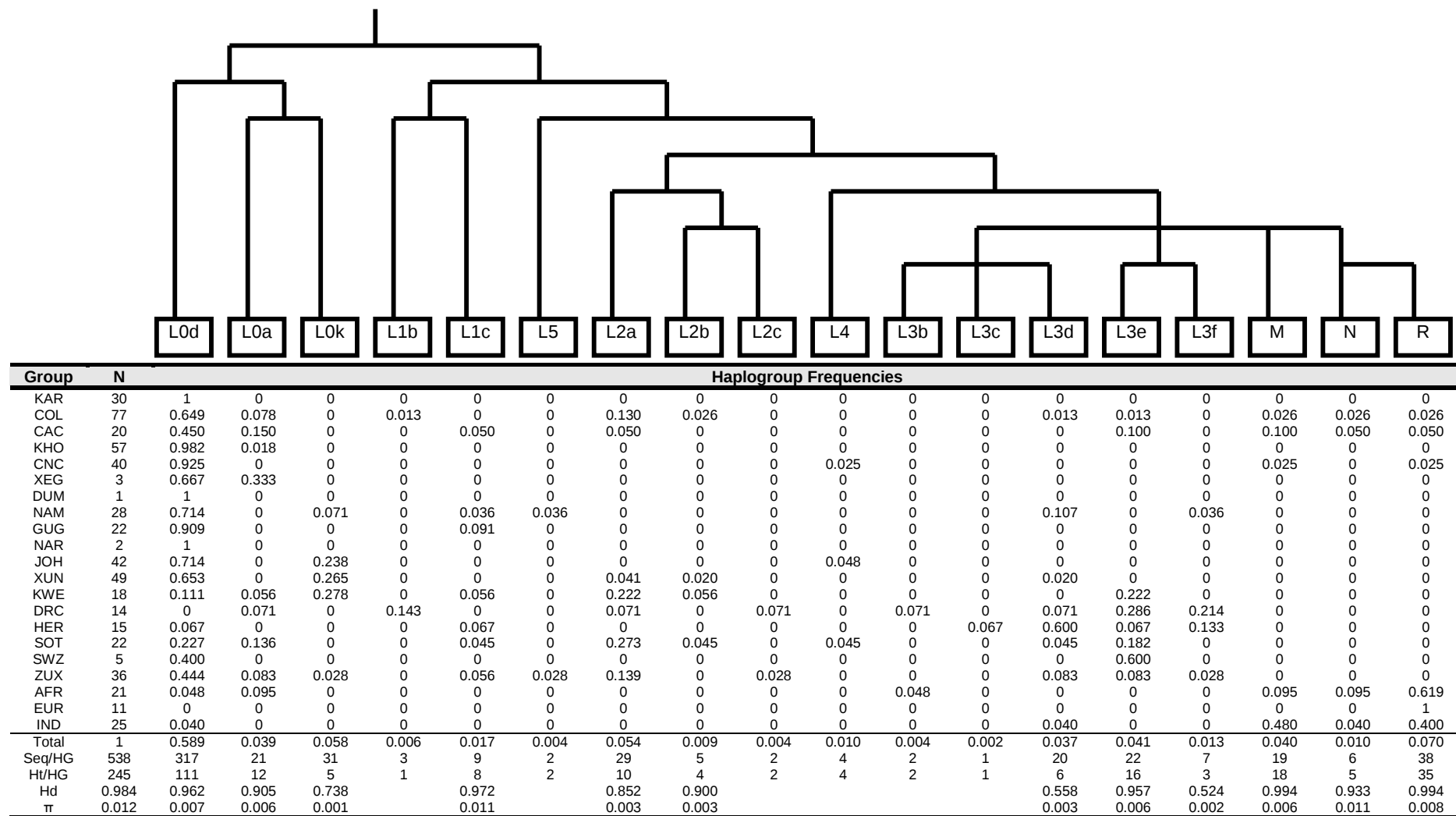


Figure 3.3 Mitochondrial haplogroup tree with nomenclature according to Behar *et al.*, (2008), listing haplogroup frequencies in the different populations in the study group. The number of sequences per haplogroup (Seq/HG), number of haplotypes per haplogroup (Ht/HG), Haplotype Diversities (Hd) and Nucleotide Diversities (π) in the different haplogroups are also indicated.

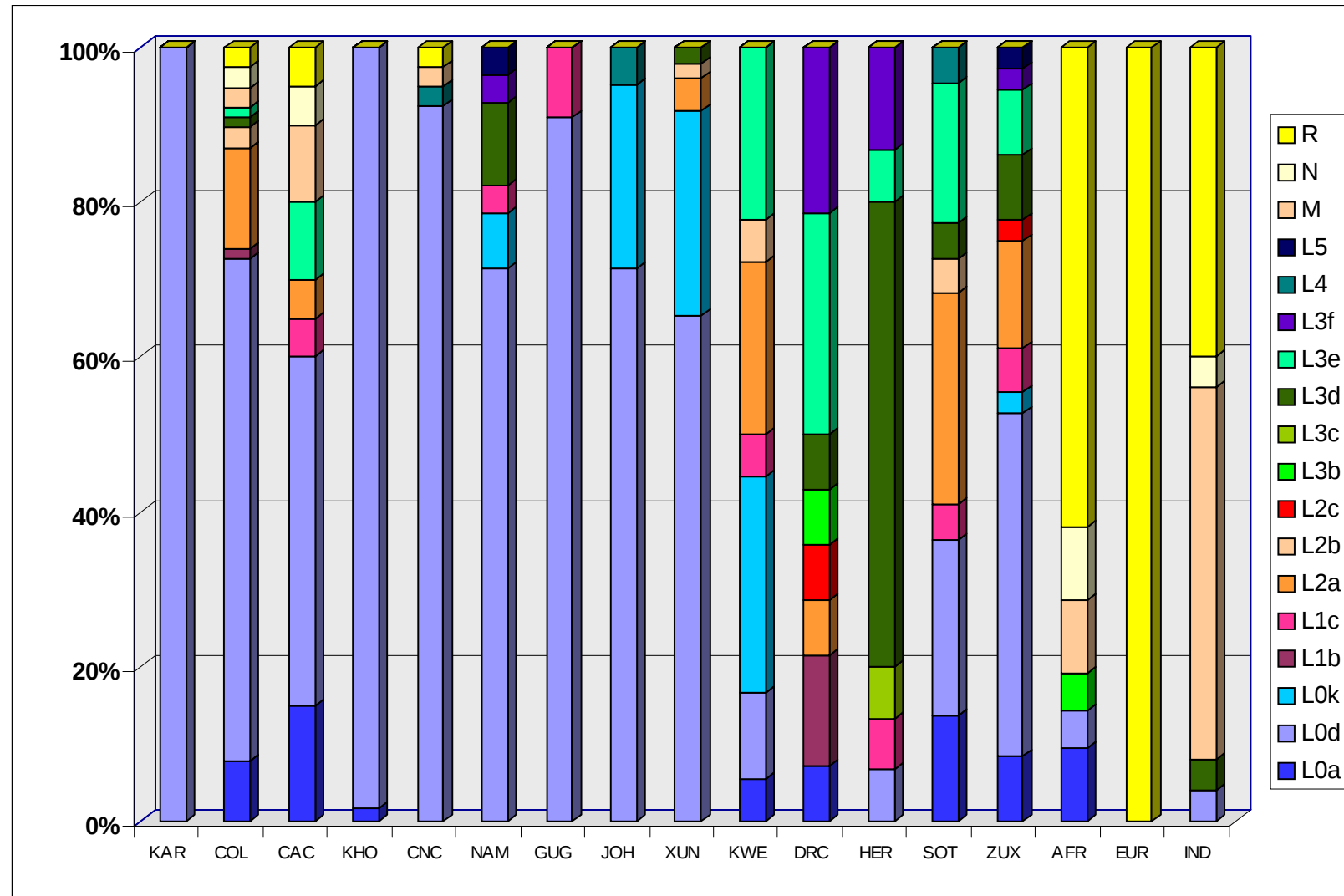


Figure 3.4 Graphical illustration of percentage mitochondrial haplogroup assignment in the populations used in comparative population analysis

3.3 Haplogroup assignment and structure

Haplogroups other than L0d were found at very low frequencies in the total sample group (no other haplogroup >7% total frequency). High frequencies of these non-L0d haplogroups were mostly seen in the comparative groups and not in the Khoe-San or Coloured groups. L0d was the most frequent haplogroup in the total sample comprising 59% of all sequences. L0d had high frequencies in all of the Khoe-San and Coloured groups ranging from 45% in the Cape Coloured to 100% in the Karretjie group. L0d frequencies in the Coloured groups of South Africa (CAC - 45%, COL - 65%, CNC - 93%, KAR - 100%) compared well with frequencies in San (KWE - 11%, XUN - 65%, JOH - 71%, GUG - 91%, KHO - 98%) and Khoe (NAM - 71%) groups.

Relationships of haplotypes within the main haplogroups were assessed using maximum likelihood trees (Figure 3.5 a and b) and parsimony based network analysis (Figure 3.6). Overall, identified sub-haplogroups did group together on the tree (Figure 3.5 a and b). In some instances, however, especially within the L3, L4, M, N and R branches the tree lacked structure and identified sub-groups did not group together. This again illustrate that the high rate of back mutation and lack of variation in the HVS-I and II necessitate the use of coding region variation to indicate and direct the overall classification and structure of haplogroups. Control region variation should then be used for finer within sub-haplogroup structuring.

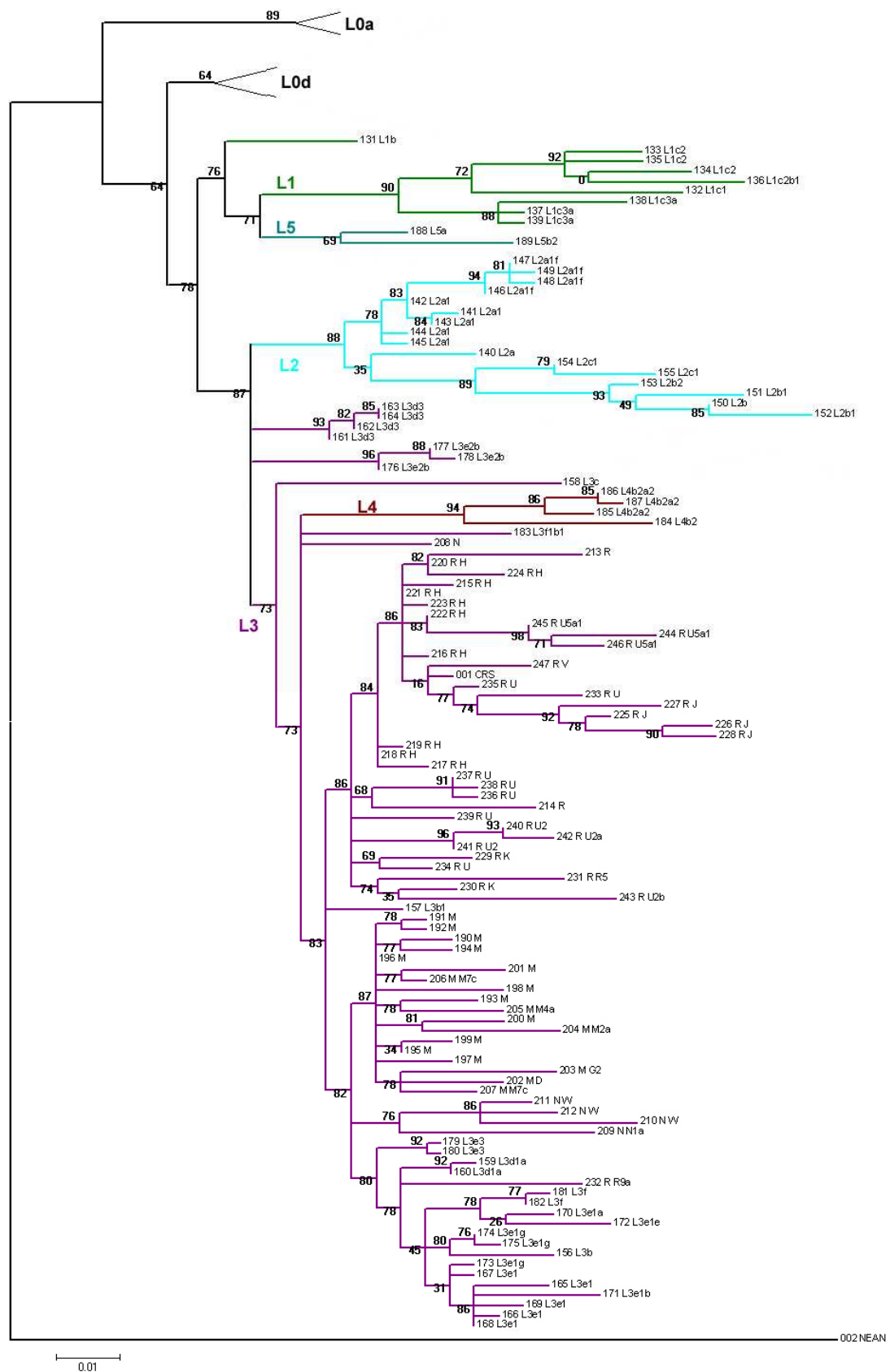


Figure 3.5a Maximum likelihood tree representing the substructure of L1 to L5. Individuals are labeled with numbers corresponding to the haplotype list in Appendix E and their classified haplogroup. A Neanderthal sequence form the outgroup. Branch support (%) was calculated through aLRT.

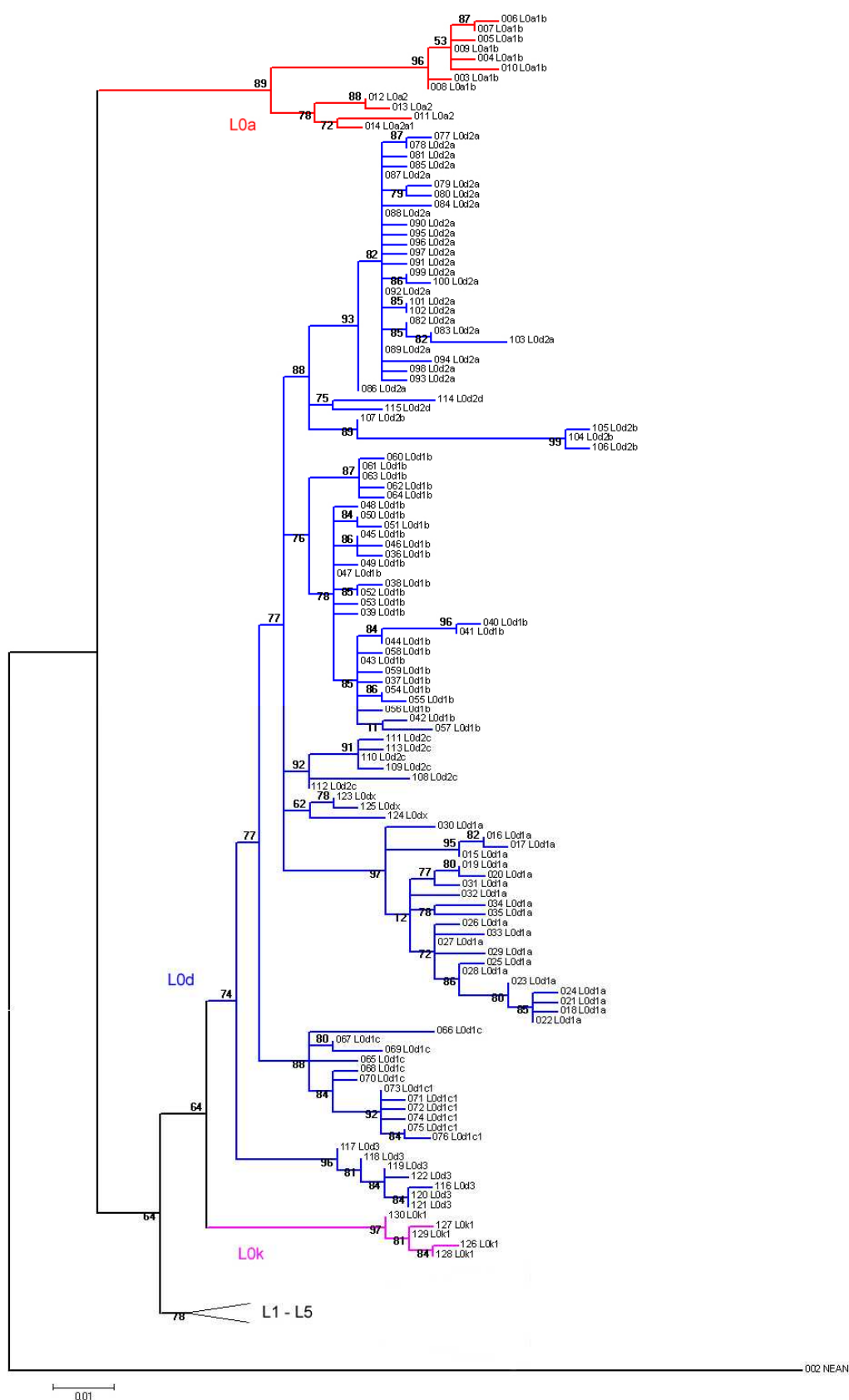


Figure 3.5b Maximum likelihood tree showing the relationships of the different mtDNA haplotypes within haplogroup L0. Individuals are labeled with numbers corresponding to the haplotype list in Appendix E and their classified haplogroup. A Neanderthal sequence form the outgroup. Branch support (%) was calculated through aLRT.

3.3.1 Haplogroup L0d/k

The relationships of haplotypes in the L0 branch are shown in the phylogenetic tree in Figure 3.5b and the in network presented in Figure 3.6. A schematic tree showing the substructure and population frequencies of haplogroups L0d and L0k are shown in Figure 3.7. In addition Figure 3.8 represents the population frequencies of L0d and L0k sub-haplogroups in the form of bar charts.

Coalescent times (Time to Most Recent Common Ancestors - TMRCA) for all the L0d/k subgroups and times at which their lineages diverged from the other lineages were calculated from the network. The ρ and σ values as well as the values in years according to various mutation rates are represented in Table 3.2. Although years according to all of the most widely used rates are represented in Table 3.2, rates according to Ward *et al.*, will be used in the description henceforth (Ward *et al.*, 1991). Figure 3.9 represents coalescent and divergence times of Table 3.2 in a graphic format.

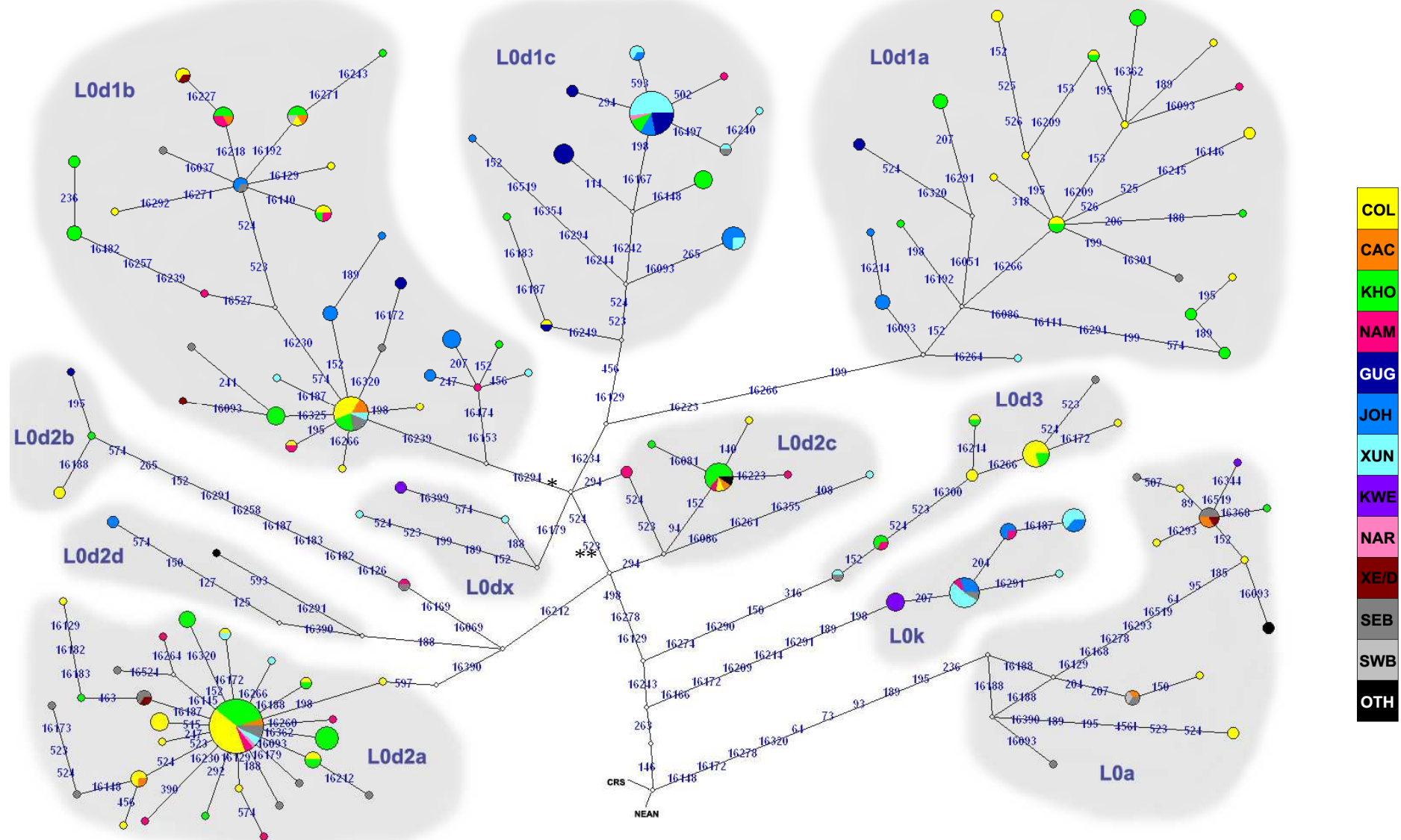


Figure 3.6 Median joining network representing L0 substructure in the different populations of the study group. Stars indicate median vectors that are discussed in the text. CRS – Control Region Sequence, NEAN – Neanderthal – Root. Numbers indicate mutations according to HVS base pair number. Circles represent haplotypes and are proportional to the number of sequences represented. The colour key indicates from which populations different haplotypes originated.

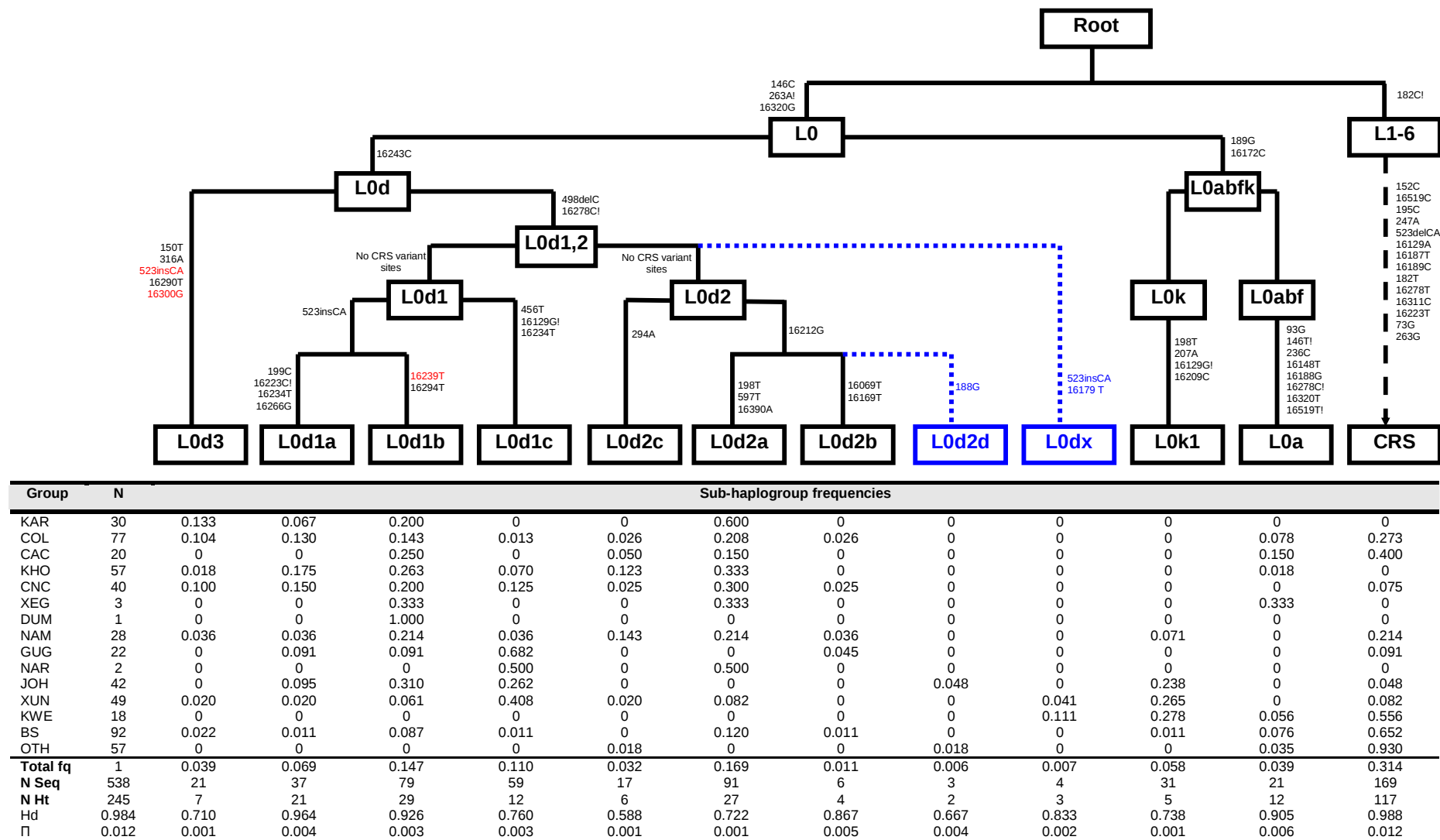


Figure 3.7 L0d structure as published in Behar *et al.*, 2008 (black). Suggested changes according to this thesis are highlighted: In blue – Two new, previously unidentified clades. In red – Mutations suggested to be removed as clade defining mutations. The table summarise the frequencies, Haplotype Diversities (Hd) and Nucleotide Diversities (π) of the various L0d subgroups as well as L0k1 and L0a. BS – Bantu-speaking, OTH - Other

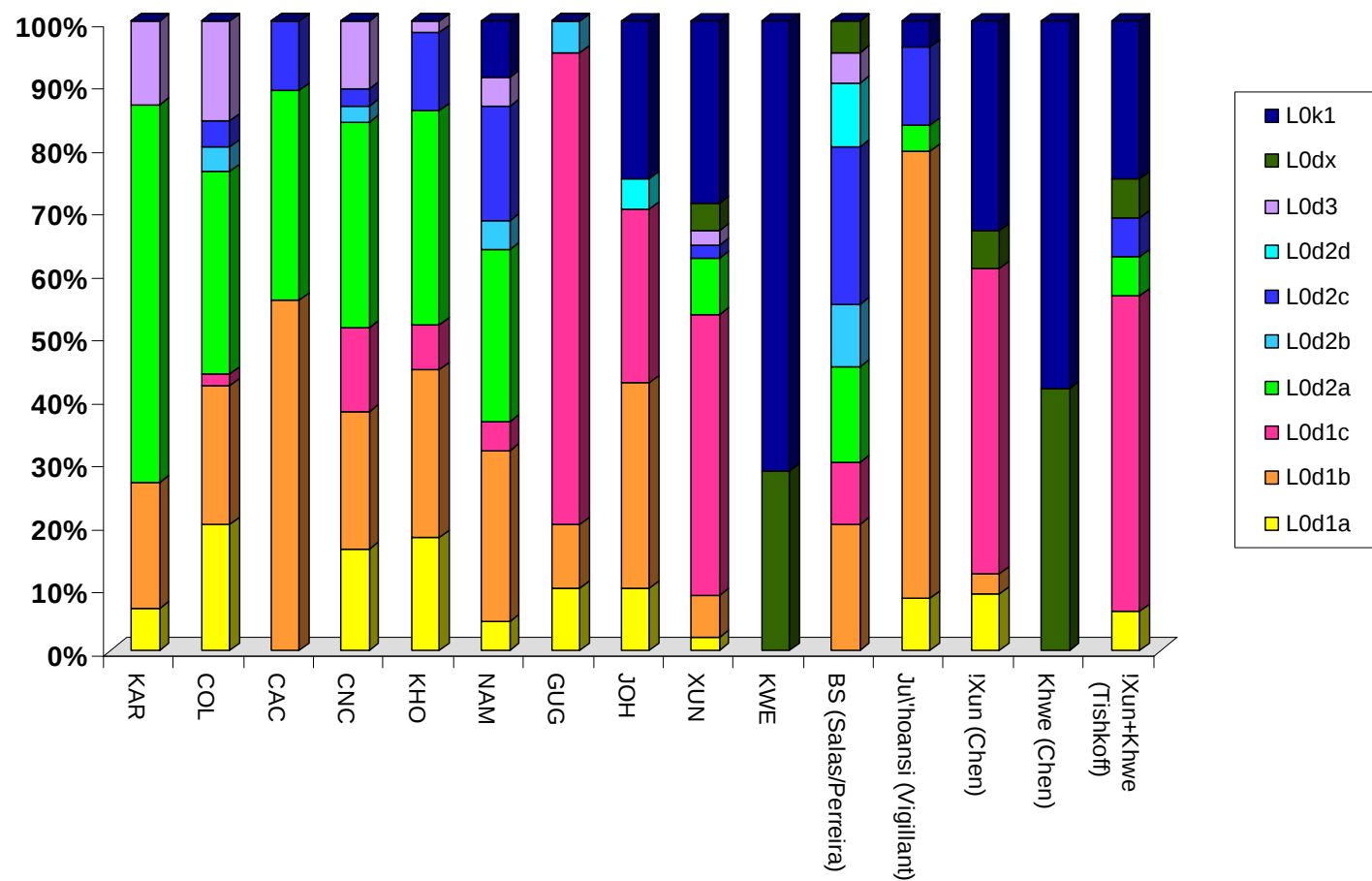


Figure 3.8 Graphical illustration of percentage L0d/k sub-haplogroup assignment in the populations used in comparative population analysis. Published comparative data is according to Table 1.3

Table 3.2 TMRCA calculated for the L0d/k subgroups. Four different mutation rates are applied

Split from other										
Ref			Horai ¹		Soodyall ²		Ward ³		Foster ⁴	
Haplogroup	ρ	σ	Years	SD	Years	SD	Years	SD	Years	SD
L0d	10.8580	2.2635	138002	28768	116247	24233	96601	20138	53668	11188
L0d1a	7.1351	2.0865	90685	26519	76389	22338	63480	18563	35267	10313
L0d1b	4.6709	1.5502	59366	19703	50007	16597	41556	13792	23087	7662
L0d1c	6.7119	2.2906	85306	29113	71858	24523	59714	20379	33175	11322
L0d2a	3.8242	1.7380	48604	22089	40942	18607	34023	15463	18902	8590
L0d2b	8.5000	2.4777	108032	31491	91002	26527	75623	22044	42013	12247
L0d2c	3.2941	1.5519	41867	19724	35267	16615	29307	13807	16282	7671
L0d2d	5.0000	1.8559	63549	23588	53531	19869	44484	16512	24714	9173
L0d3	8.0000	2.6273	101678	33392	85649	28128	71174	23375	39542	12986
L0dx	4.0000	1.5411	50839	19587	42824	16499	35587	13711	19771	7617
L0k1	8.5161	2.8144	108237	35770	91174	30131	75766	25039	42093	13911

TMRCA										
	ρ	σ	Years	SD	Years	SD	Years	SD	Years	SD
L0d	9.8580	2.0307	125292	25810	105541	21741	87705	18067	48725	10037
L0d1	6.4286	1.2576	81706	15984	68825	13464	57194	11189	31775	6216
L0d2	4.8718	1.6199	61919	20588	52158	17343	43343	14412	24080	8007
L0d1a	4.1351	1.1634	52556	14786	44271	12455	36789	10351	20439	5750
L0d1b	3.6709	1.1845	46656	15055	39301	12681	32659	10538	18144	5855
L0d1c	4.7119	1.8019	59887	22902	50446	19291	41921	16031	23290	8906
L0d2a	1.8242	1.0143	23185	12891	19530	10859	16230	9024	9016	5013
L0d2b	6.5000	2.0344	82613	25857	69590	21780	57829	18100	32128	10055
L0d2c	2.2941	1.1867	29157	15083	24561	12705	20410	10558	11339	5866
L0d2d	4.0000	1.5635	50839	19872	42824	16739	35587	13910	19771	7728
L0d3	4.0000	1.7037	50839	21654	42824	18240	35587	15157	19771	8421
L0dx	3.0000	1.1726	38129	14903	32118	12554	26690	10432	14828	5796
L0k1	1.5161	0.9596	19269	12197	16232	10274	13488	8538	7494	4743

Years are calculated from ρ by multiplying with the specific mutation rate

Standard deviation (SD) are calculated from σ by multiplying with the specific mutation rate

¹ Horai *et al.*, (1995)

² Soodyall *et al.*, (1996)

³ Ward *et al.*, (1991)

⁴ Foster *et al.*, (1996)

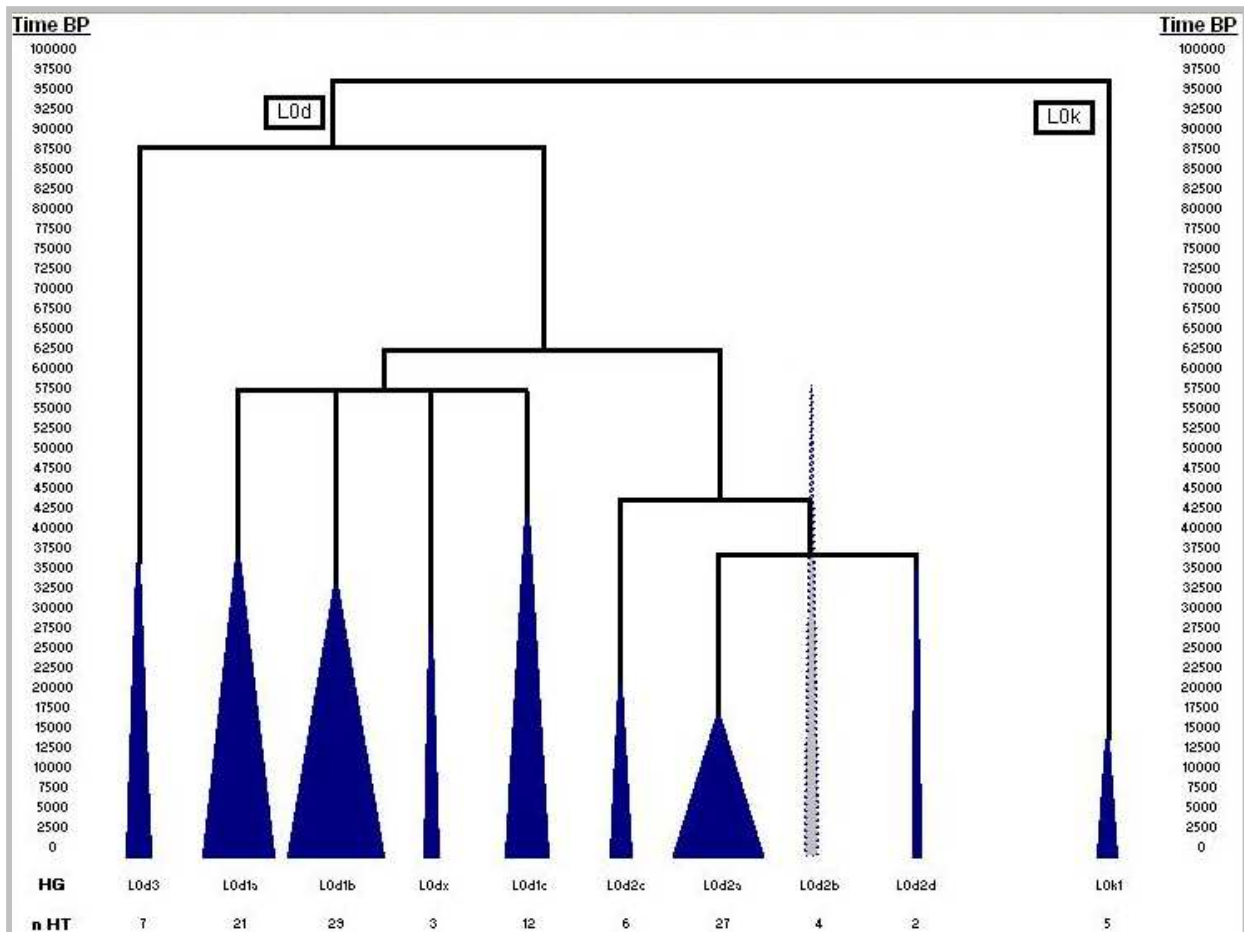


Figure 3.9 Graphic representation of coalescent times and times of divergence of the mtDNA sub-haplogroups of L0d and L0k. The mutation rate estimated by Ward *et al.*, (1991) was used in these estimates

3.3.2 Khoe-San associated haplogroups L0d and L0k – Further analysis

The sub-haplogroups of the Khoe-San associated haplogroups L0d/k were differentially distributed in the different sample groups included in this study (Figure 3.7 and 3.8)

Just by observing the distribution over the different sampling groups in the form of bar-charts (Figure 3.8) one could immediately see the differences. It was especially clear between the southern-San/Coloured/Khoe groups (KAR, COL, CAC, KHO, CNC, NAM) and the San groups located north of them (GUG, JOH, XUN). To further investigate these differential distributions, analysis of sub-haplogroup distribution was done.

Sample groups were arranged in a southeast to northwest direction and coloured with increasing shade from the southeast to the northwest; the resultant distribution of L0d/k subgroups is represented in Figure 3.10. A clear clinal pattern for all of the haplogroups was observed. L0d2a and L0d3 seemed to have a more southeastern distribution (lighter shades), while L0d2b, L0d2c, L0d1a has and intermediate central pattern. L0d1c, L0k1 as well as the few sequences belonging to L0dx and L0d2d, however, was much darker and seem to predominate in the northern groups.

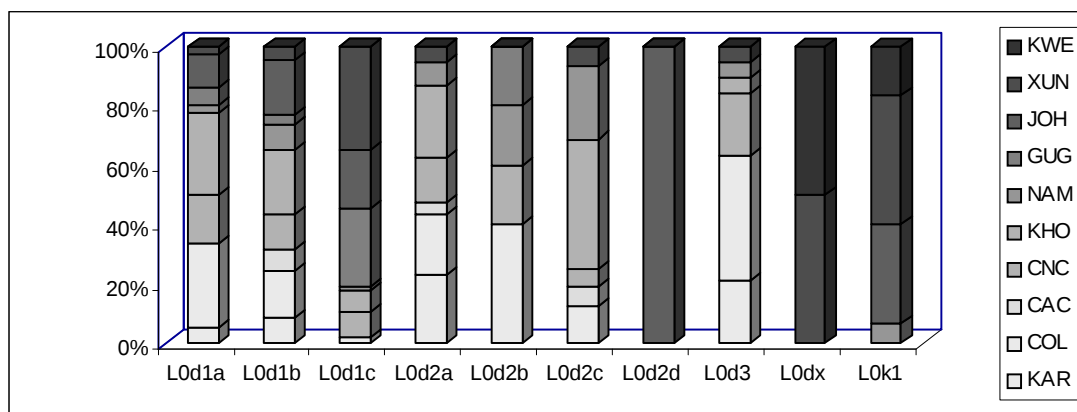


Figure 3.10 Bar-graph indicating the clinal distribution of the L0d/k subgroups. Darker shades are north-western groups and lighter shades are southern groups

To further investigate this apparent clinal distributions contour plots of the haplogroups were constructed with the Surfer v.8.06.39 program and is shown in Figure 3.11.

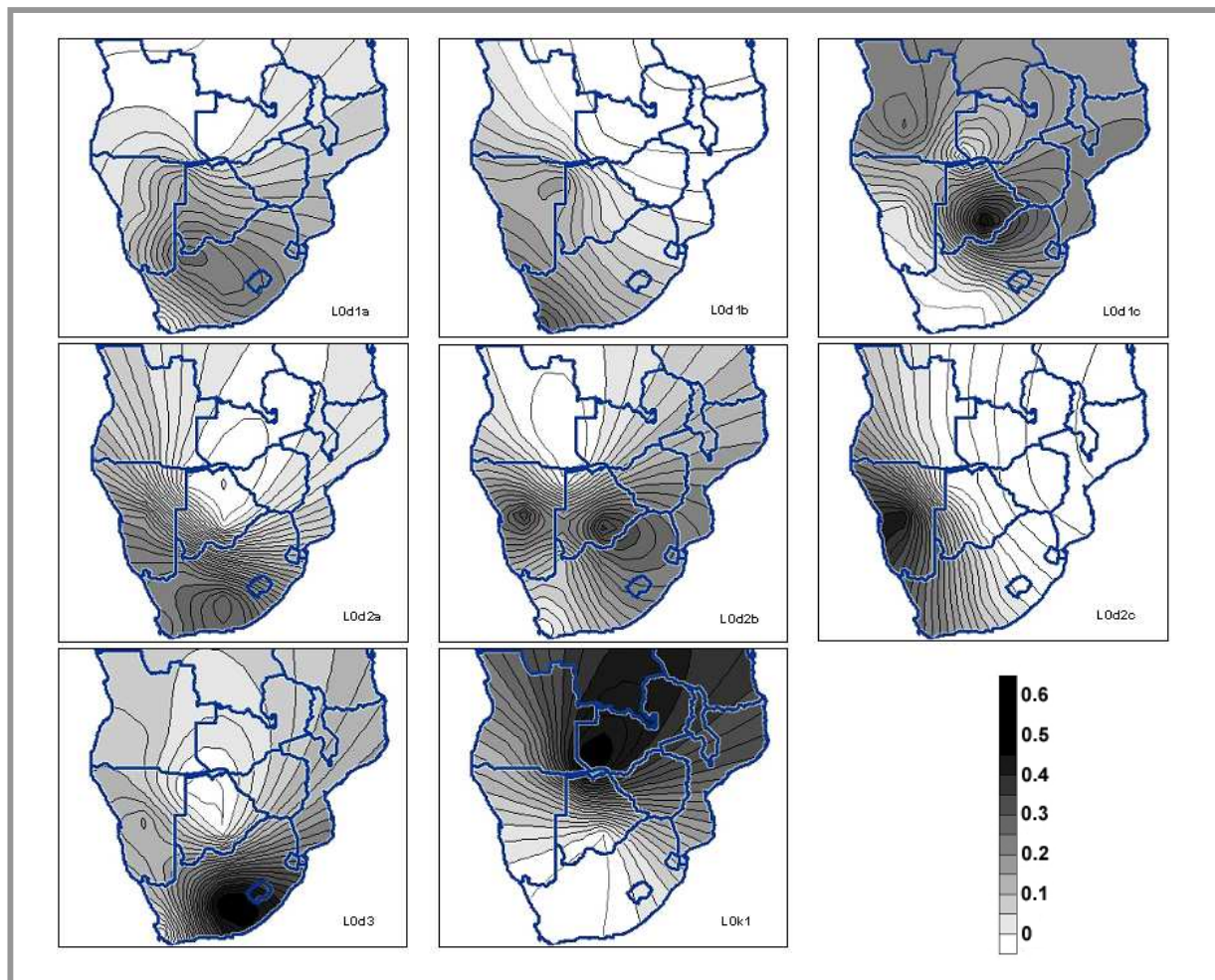
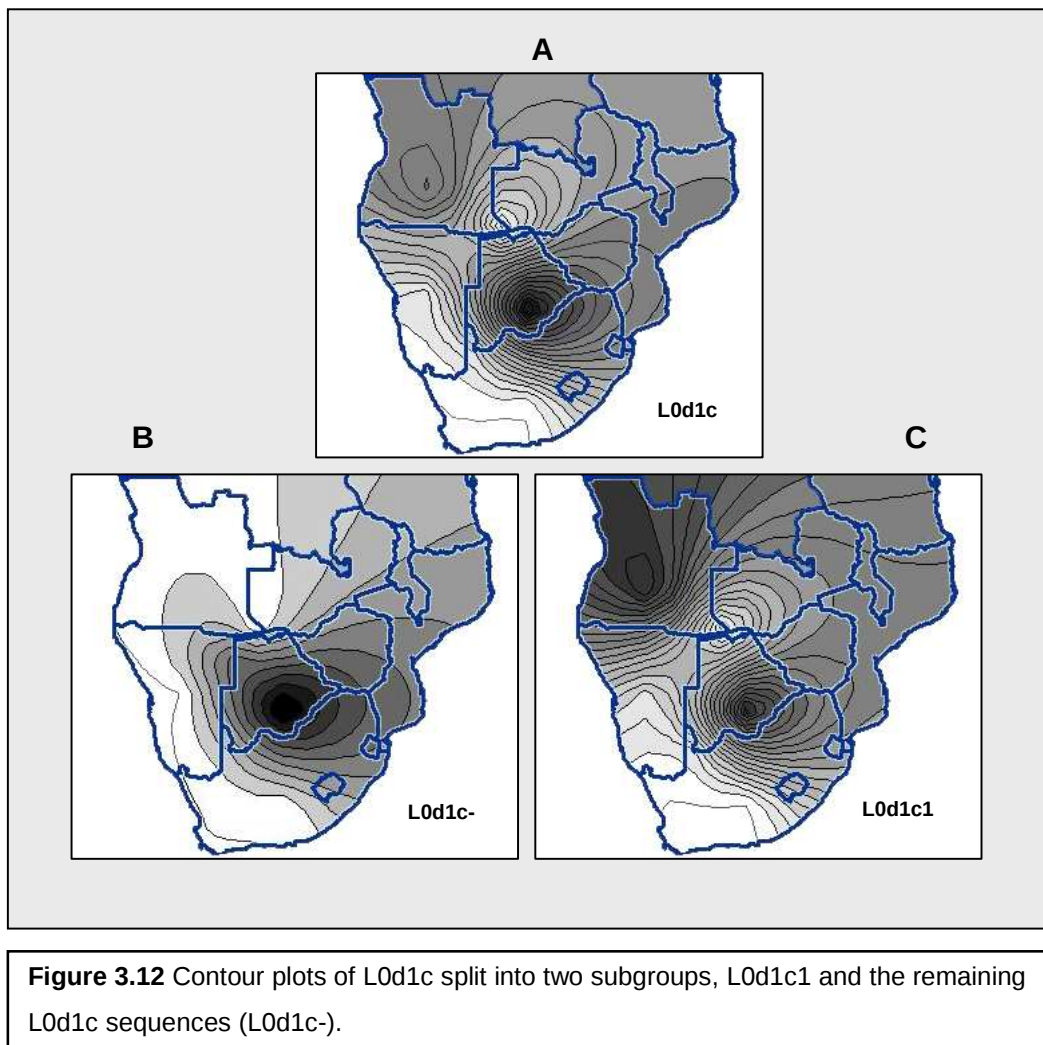


Figure 3.11 Contour plots indicating the frequency distributions of L0d/k subgroups

The contour plots reflected the distributions of the L0d/k subgroups as discussed in section 3.1.3.2. Certain haplogroups (L0d3 and L0d2a) had higher frequencies in the southeast than the northwest; others had a more gradual and central distribution (L0d1a, L0d1b, L0d2c and L0d2b), while some had higher frequencies in the north (L0k1 and to a certain extent L0d1c).

In the contour plots of Figure 3.11, all the haplogroups except L0d2b and L0d1c seemed to have a unimodal distribution with a single point of highest frequency and then decreasing frequencies from there in a clinal fashion. L0d2b showed two peaks represented by the NAM and GUG, however, this sub-haplogroup was observed at too low frequencies to

have any significance. L0d1c also showed a bimodal distribution pattern. To analyse this further the L0d1c group was split up into two groups. The first group was the L0d1c1 sequences as defined by Behar *et al.*, (Behar *et al.*, 2008), and are represented by the star-like expansion pattern within L0d1c in the network (Figure 3.6). The second group L0d1c- was the remaining L0d1c sequences after the L0d1c1 sequences were removed. Contour plots of these groups are presented in Figure 3.12.



From Figure 3.12 it can be seen that L0d1c originally had a unimodal clinal distribution but a subsequent expansion in the L0d1c1 subgroup occurred that caused elevated

frequencies in the XUN. L0d1c1 did not occur in the KWE and were at low frequencies in the JOH. This led to the overall bimodal distribution of L0d1c.

To further analyse the individual haplogroup histories, to test if they had notable expansions and to date these expansions, mismatch distributions of the sub-haplogroup sequences were constructed (Figure 3.13 and Table 3.3)

Table 3.3 Mismatch distribution statistics (haplogroups)

HG	Raggedness index	τ	τ^*	Theta0	Theta0 qt 5%-95%	Theta1	Theta1 qt 5% - 95%	Model (SSD) p-value
L0d1a	0.013	6.285	27 958	0.017	0.000 - 1.325	31.250	14.805 - 99999.000	0.700
L0d1b	0.033	6.805	30 271	0.002	0.000 - 1.130	11.646	6.190 - 84.458	0.230
L0d1c	0.080	5.971	26 561	0.000	0.000 - 0.751	5.188	3.883 - 99999.000	0.160
L0d2a	0.042	1.545	6 873	0.000	0.000 - 0.366	579.075	3.714 - 99999.000	0.680
L0d2b	0.276	17.441		0.000	0.000 - 57.600	55.181	39.615 - 99999.000	0.000#
L0d2c	0.135	0.000		0.000	0.000 - 0.000	428.125	0.000 - 0.000	0.000#
L0d2d	1.000	10.088		0.002	0.000 - 6.161	99999.000	99999.000 - 99999.000	0.000+
L0d3	0.053	4.500	20 018	0.000	0.000 - 0.508	3.483	1.619 - 99999.000	0.600
L0dx	0.528	8.813		0.002	0.000 - 14.400	9.384	5.060 - 99999.000	0.180#
L0k1	0.055	1.393	6 197	0.000	0.000 - 0.028	99999.000	4.918 - 99999.000	0.600
L0a	0.049	15.965	71 019	0.002	0.000 - 4.502	11.197	6.043 - 166.353	0.360
M	0.019	9.559	42 522	0.290	0.000 - 2.406	86.104	47.432 - 99999.000	0.450
R	0.014	10.424	46 370	1.752	0.000 - 2.387	34.883	23.632 - 183.008	0.490

* τ – Time before present that expansion took place (calculation explained in section 2.2.2.3)

expansion hypothesis rejected - 95% CI overlap

+ excluded – too few sequences

HG – Haplogroup, SSD - Sum of Squared deviation

Mismatch distribution statistics and the results for a spatial expansion test are shown in Table 3.3 (the p-value indicate the probability that the simulated SSD (simulated under an expansion scenario) is not significantly different from the observed SSD). All L0d subgroups except L0d2b, L0d2c and L0dx tested positive for expansions. The haplogroups that indicated expansions with the highest significance was L0d1a and L0d2a. Their mismatch distributions showed smooth unimodal distributions with low raggedness values. Their τ (Tau) values, however, differed with the τ value of L0d2a indicating a much more recent expansion. τ values of the L0d1 haplogroups were similar (indicating expansions of around 27 000 years BP) while L0d3 had a smaller τ value and L0d2a and L0k1 the smallest (indicating expansions around 6 000 years BP). Both M and R haplogroups experienced expansions ~ 40 000 to 50 000 years BP.

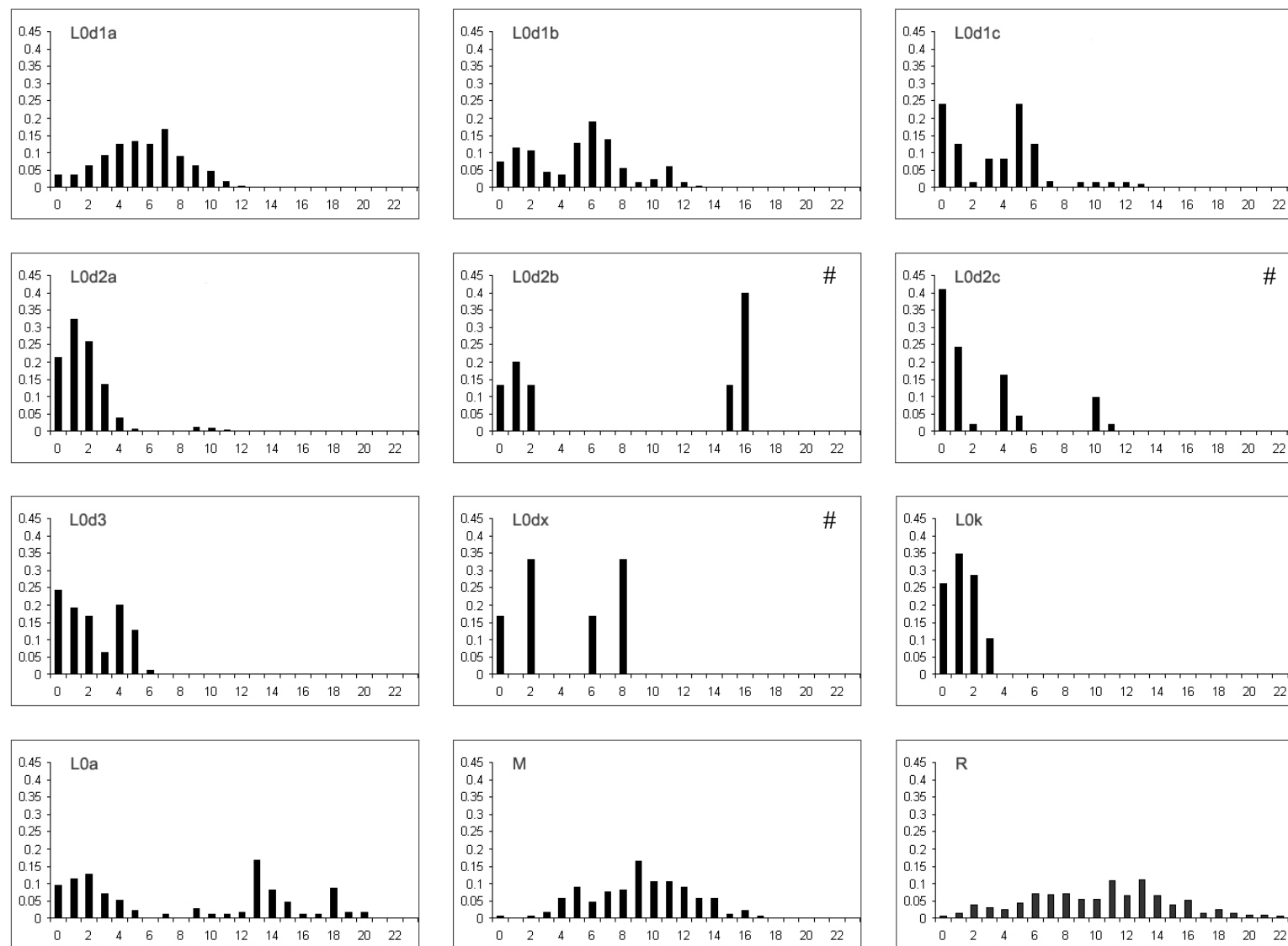


Figure 3.13 Mismatch distributions of L0d/k sub-haplogroups and comparative groups. # expansion hypothesis rejected - 95% CI overlap

Caveats associated with the coalescence analysis employed in mismatch distributions are the assumption of a single exponentially growing population and the large degrees of statistical uncertainty. Also, by applying these methods earlier population expansions can be obscured by recent population bottlenecks (Excoffier and Schneider, 1999). Mismatch distributions have been reported previously to have less ability to predict population expansions than neutrality test summary statistics such as Tajima's D (Tajima, 1989), Fu's Fs (Fu, 1997) and the R2 statistic (Ramos-Onsins and Rozas, 2002). Diversity estimates together with the neutrality tests for the L0d sub-haplogroups are shown in Table 3.4. Also included as comparative samples are other sub-haplogroups in the study group that had more than 10 representative sequences.

Table 3.4 Diversity statistics and neutrality tests of L0d/k subgroups and comparative haplogroups

Group	N seq	N Ht	Hd	π	θ_s	W- θ_s	Ne	Tajima's D	Tajima's D p-value	Fs	Fs p-value	R2	R2 p-value
L0d	317	111	0.962	0.00732	0.01419	15.342	2730	-1.45547	0.028*	-33.984	<0.001***	0.0421	0.069
L0d1a	37	21	0.964	0.00436	0.00584	6.468	1151	-0.84168	0.210	-8.871	0.001**	0.0849	0.120
L0d1b	79	29	0.926	0.00279	0.00551	6.072	1081	-1.51856	0.040*	-17.250	<0.001***	0.0488	0.040*
L0d1c	59	12	0.760	0.00248	0.00388	4.305	766	-1.09831	0.136	-1.628	0.273	0.0714	0.186
L0d2a	91	27	0.722	0.00110	0.00463	5.116	910	-2.29659	<0.001***	-23.082	<0.001***	0.0240	0.005**
L0d2b	6	4	0.867	0.00458	0.00393	4.380	779	1.03370	0.864	1.229	0.701	0.2429	0.738
L0d2c	17	6	0.588	0.00129	0.00239	2.662	474	-1.65319	0.032*	-1.475	0.134	0.1135	0.128
L0d2d	3	2	0.667	0.00357	0.00359	4.000	712	na	Na	na	Na	na	na
L0d3	21	7	0.710	0.00121	0.00124	1.390	247	-0.08107	0.511	-1.287	0.170	0.1327	0.432
L0dx	4	3	0.833	0.00238	0.00244	2.727	485	na	na	na	na	na	na
L0k1	31	5	0.738	0.00110	0.00089	1.001	178	0.58176	0.742	-0.044	0.518	0.1522	0.721
L0a	21	12	0.905	0.00590	0.00526	5.837	1039	0.30935	0.676	-1.317	0.290	0.1475	0.725
L2a	29	10	0.852	0.00332	0.00367	4.074	725	-0.29380	0.430	-0.739	0.382	0.1153	0.472
L3d	20	6	0.558	0.00281	0.00253	2.819	502	0.41429	0.700	2.092	0.849	0.1506	0.662
L3e	22	16	0.957	0.00575	0.00545	6.035	1074	0.08399	0.601	-5.350	0.056	0.1345	0.611
M	19	18	0.994	0.00620	0.01340	14.592	2596	-2.17227	0.004**	-11.280	<0.001***	0.0462	<0.001***
R	38	35	0.994	0.00793	0.01702	19.326	3439	-1.93058	0.010**	-27.889	<0.001***	0.0466	<0.001***

* p-value < 0.05

** p-value < 0.005

*** p-value < 0.001

The effective population size of females (N_e) was estimated from W- θ_s as explained in section 2.2.3. The two non-African macro-haplogroups had very big effective population sizes while the African haplogroup N_e was smaller. In the L0d subgroups the largest N_e was detected in L0d1a, L0d1b and L0d2a while L0d3 and L0k1 had the smallest N_e .

Under neutral expectations with random mating, constant population sizes and no selection π and θ should be equal (Jobling *et al.*, 2004c). Neutrality tests were done to detect deviations from the assumptions of neutrality and constant population size. Significantly negative Tajima D and Fs values and significantly positive R2 values indicate population

growth and/or positive selection. The F_s and R_2 statistic have been reported to detect population expansions very successfully (Ramos-Onsins and Rozas, 2002; Pilkington *et al.*, 2008). F_s is based on the probability of drawing a number of haplotypes that is greater or equal to the observed number of samples drawn from a population of constant size. R_2 is based on the difference between the average number of nucleotide differences and the number of singleton mutations. The R_2 statistic is especially powerful when sample sizes are small (~10) and F_s have a greater ability to detect population expansions when sample sizes are large (~50) (Ramos-Onsins and Rozas, 2002; Pilkington *et al.*, 2008).

In the comparative groups the non-African haplogroups M and R tested positive for population expansion in all three neutrality tests with highly significant P-values. The L0d group as a whole had significant D and F_s values but not for R_2 . R_2 , however, does not perform reliably at large sample sizes (Ramos-Onsins and Rozas, 2002). Of the L0d subgroups L0d2a had the highest significance in all three neutrality tests. L0d1b also attained significance in all three tests while L0d1a had a very significant F_s value but did not reach significance in the Tajima's D and R_2 tests.

While neutrality tests are widely employed to test hypotheses of population expansion events, recent improvements in coalescence inference methods led to increased accuracy, without the need to assume a single exponential growth curve (Shapiro *et al.*, 2004; Atkinson *et al.*, 2008). One of these methods, Bayesian Skyline Plots (BSPs) (Drummond *et al.*, 2005), were employed to further visually represent the changes in N_e through time, were constructed for each haplogroup (Figure 3.14).

The BSPs of all the L0d sub-haplogroups, except L0d1a, indicated a recent increase in N_e (Figure 3.14). L0d1a had an increase that started around 25 000 – 30 000 years BP and a recent decrease that started around 5 000 years BP. L0d1c had a constant population size over a extended period and then similar to L0d1a, started to decrease around 5 000 years BP. Around a 1 000 years BP, however, it increased rapidly. L0d1b had an increase that started around 14 000 years BP and a further increase recently. Despite a shallow coalescence time, L0d2a showed a dramatic increase from 8 000 years BP onwards and a further recent increase. The L0d3 BSP profile that included east African and the Kuwait

haplotypes (L0d3+) showed a slow decline over an extended period followed by a recent increase in N_e . The L0d3 profile that only included the southern African L0d3 haplotypes (L0d3-) showed a more intense decline and an increase that started later than in the L0d3+ profile.

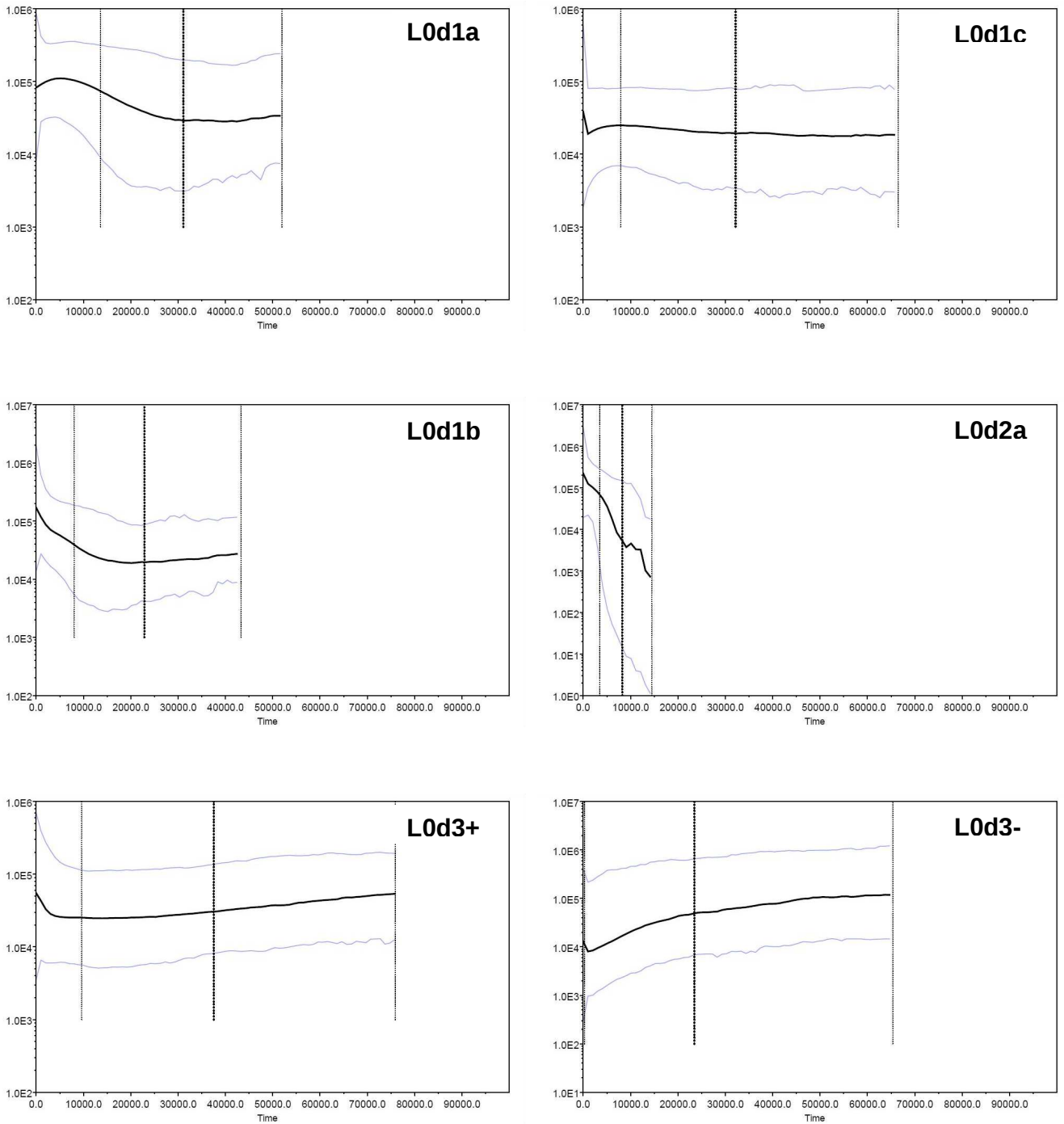


Figure 3.14 Bayesian Skyline plots of haplogroups showing changes in N_e through time. A log scale of N_e is represented on the Y-axis, while years before present is represented on the X-axis, with the present indicated by 0. L0d3+ is L0d3 including the east African and Kuwait sequences. L0d3- includes only L0d3 sequences from the present study. The black bold vertical lines indicate the coalescence date and the lighter vertical lines the 95% confidence intervals for the coalescence. The blue lines indicate the 95% confidence intervals for the plot-lines

3.3.3 Discussion of analyses of Khoe-San associated haplogroups L0d and L0k

In the following section each of the L0d/k sub-haplogroups will be discussed with regard to the different analyses presented in previous sections. The placing of the haplogroup on the network and tree, the TMRCA dating and the frequencies in the different populations will be discussed and compared to published findings. Phylogenetic results and dating might differ from published studies, especially the whole genome sequencing studies, due to different lengths of sequence investigated and different methods employed. These differences are also highlighted in following sections. Furthermore patterns in the network, the geographical spread of the haplogroup together with the evidence of population growth signals are interpreted and linked with evidence from other disciplines such as archaeology, linguistics and ethnography to infer possible histories for the lineages involved.

L0k

While many sequences (31) belonged to haplogroup L0k, they were represented by only five different haplotypes, all belonging to L0k1. The L0k branch on the tree in Figure 3.5b grouped with the L0d branch and not with the L0a branch as was established previously doing whole genome sequencing (Behar *et al.*, 2008). Both branches that separated L0k from L0a, however, had aLRT branch support of just above 60%. Comparing the network (Figure 3.6) with Behar *et al.*, (represented in Figure 3.7) classification (Behar *et al.*, 2008) one can see that the 263 and the 146 mutations separating L0k and L0a on the network should move to be L0 defining mutations. Instead L0a and L0k should group on the same branch with 189G and 16172C defining the common branch (in the network they were on separate L0a and L0k branches). Furthermore the L0k clade defining mutations in the network (16166, 16209, 16214, 16291, 198) compared well with Behar *et al.*, (Behar *et al.*, 2008). The only exception being 207A which also should be a clade defining mutation and then revert in a subgroup of the sequences according to Behar *et al.* In the present network it should thus move to precede the group of KWE haplotypes and then subsequently revert to the ancestral type in this group only.

L0k has two sub-haplogroups L0k1 and L0k2, that separated ~40 000 years BP (Behar *et al.*, 2008). As yet there have been only one report of L0k2 in an individual from Yemen (Behar *et al.*, 2008), while L0k1 was found exclusively in the San groups (Vigilant *et al.*, 1991; Chen *et al.*, 2000; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). All sequences in the current sample group were L0k1 and coalesced 13 488 years BP (+/- 8 538) and diverged from other sequences in L0 75 766 years BP (+/-25 039) (Table 3.2). This is shallow times compared to what was found previously for L0k (39 683+/-8 730 for the coalescence and 142 860+/-11 905 for the split) (Behar *et al.*, 2008).

In the present study, L0k was limited to the northern Khoe-San groups whilst the southern groups and the central Kalahari groups from Botswana showed no presence of the L0k haplogroup (Figure 3.3, 3.4). The northern San groups contained the highest percentages while the Nama contained lower levels (Figure 3.3, 3.4). This can be explained by the fact that the Nama originally came from an area currently known as the northern parts of the Cape Province (SA) and recently moved into the Namibia area (Barnard, 1992). They therefore could be regarded as a southern Khoe-San group rather than a northern group. The L0k1 found in the Nama is thus most likely because of recent gene flow from the San people of northern Namibia (such as the Ju\`hoansi and !Xun).

The frequencies of L0k1 in the !Xun and Khwe from this study (27%, 28%) (Figure 3.3, 3.4) agreed with previous studies that reported frequencies of 26% in the !Xun and 23% in the Khwe (Chen *et al.*, 2000) (Table1.2). Tishkoff *et al.* reported frequencies of 22% in a combined !Xun and Khwe group (Table1.2) (Tishkoff *et al.*, 2007). Our results for the Ju\`hoansi (JOH in Figure 3.3, 3.4), however, differed somewhat from what was found previously where only 4% of the Ju\`hoansi lineages were resolved into haplogroup L0k1 while L0d was previously found to be the most prevalent haplogroup (96%) (Vigilant *et al.*, 1991) (Table1.2). In our study L0k1 represented 24% Ju\`hoansi group and L0d 71%. The two Ju\`hoansi groups were not from the same locations. While the group of the present study was sampled in Tsumkwe, the published group was sampled in Botswana (Dobe) as well as in Namibia.

Since previous studies only reported on the mtDNA haplogroup frequencies of the three northern San groups (!Xun, Khwe and Ju!'hoansi) the low frequency of L0k1 in the Khoe and the absence in the southern San and Coloured groups have never been noted before. Salas *et al.*, however, noted its complete lack in southeastern-Bantu-speakers contrasting with L0d (Salas *et al.*, 2002). Previously, it was thought that the history of L0d and L0k is closely intertwined and synonymous with Khoe-San history (Salas *et al.*, 2002; Behar *et al.*, 2008; Atkinson *et al.*, 2009). From the present study it was clear that, although all groups in this study with Khoe-San ancestry had L0d in common, L0k was only associated with the northern Khoe-San groups (Figure 3.3, 3.4, 3.11).

The history of the L0k1 haplogroup might be closely tied up with the Khwe rather than the rest of the San groups. It was the haplogroup with the highest frequency in the Khwe while in the other northern San groups (!Xun and Ju!'hoansi) it was secondary to L0d groups (Figure 3.3, 3.4) and might have been introduced to these groups through gene flow with the Khwe and other Khoe-speaking San groups. The low L0k1 haplogroup diversities suggest only few founders (Figure 3.3). In the network and tree (Figure 3.6 and 3.5b) it could be seen that all the Khwe sequences belonged to one haplotype and that the Khwe haplotype was ancestral to the haplotypes observed in the !Xun, Ju!'hoansi and Nama. This then suggested that L0k1 was originally a Khwe haplogroup and spread to the other northern San groups, where it diverged further. In the study by Chen *et al.*, L0k1 also was the predominant haplogroup in the Khwe (Table 1.2) (Chen *et al.*, 2000). Furthermore all seven L0k1 sequences identified in the Khwe by Chen *et al.*, was identical to the L0k1 Khwe haplotype of the present study while ten of the eleven L0k1 sequences in the !Xun was derived from the ancestral Khwe haplotype (one !Xun sequence had the Khwe haplotype) (Chen *et al.*, 2000).

It is unclear where the Khwe originally came from. Theories are that they are Khoe-San groups with extensive Bantu-speaking admixture, Bantu-speakers that lost their cattle, another pastoralist population closely related to Bantu-speakers who occupied the region before the Bantu expansions or maybe a mixture of various refugee groups driven from the grazing grounds into the Okavango swamps (Cashdan, 1986). Genetic results from the present study indicated that the maternal lines of the Khwe showed contributions from

southeastern Bantu-speakers and Khoe-San (Figure 3.8). In addition they might have had a unique contribution from an unknown pastoralist or hunter-gatherer population that carried the L0k1 maternal lineage, whose identity has since been lost. The discovery of the L0k2 haplogroup in an individual from Yemen (Behar *et al.*, 2008) suggests that the L0k haplogroups might have had an extensive spread in prehistoric Africa but remnants of the haplogroup in other populations have been lost due to drift or has not been detected due to insufficient sampling.

It would be interesting to know the L0k1 frequency in the other Khoe-speaking San groups. Linguistically, the Khwe belong to one of the three main groups in the western branch of the Khoe-speaking San groups, the other two groups being the Naro and the /Gui and //Gana (Güldemann, 2006b). No L0k1 haplogroups were found in the group of /Gui + //Gana + Kgalagari individuals (GUG), and the two Naro individuals had only L0d haplogroups. Furthermore, serogenetic studies showed that the Naro was genetically more similar to the Ju\’hoansi and ǀX’ao//’ãesi rather than to the Khwe (Jenkins, 1982). To date, no genetic studies have been done on the eastern Khoe-speaking San groups including the Tshua and Shua of eastern Botswana. They have more in common phenotypically to the Khwe, than the western Khoe-speaking San in that they resemble Bantu-speakers (Dornan, 1975; Barnard, 1992). The Tshua and Shua may be genetically closer related to the Khwe even though the Naro and the /Gui and //Gana are linguistically more related. A very interesting linguistic connection is that one of the eastern Khoe-speaking San languages, Hietshware, is the Khoisan language that is closest related to the extinct Kwadi language of western Angola, which in turn is connected to the click language of the Sandawe of eastern Africa (Güldemann, Forthcoming-b; Güldemann and Elderkin, Forthcoming). No traces of L0k have been found in Tanzania and Kenya so far but the presence of L0k2 in Yemen suggests a trans African spread of this haplogroup. If they contain high frequencies of L0k1, such as found in the Khwe, it might be indicative of another ancient hunter-gatherer population that lived northeast of the Khoe-San groups, prior to the spread of the Bantu-speaking-groups. This group might have had linguistic and genetic connections with both the Khoe-San and Sandawe. As discussed in section 1.2.2.4, an ideal candidate for such a group might be the Pygmy groups that lived north of the Khoe-San before the Bantu-expansions. Autosomal studies show genetic similarities

between Mbuti Pygmy groups of east Africa and the Khoe-San. MtDNA studies, however, have found no L0d or L0k in the Pygmy groups studied so far (Quintana-Murci *et al.*, 2008). Mostly Pygmy groups are assigned to a specific L1c haplogroup. The southern Ba-Twa Pygmies have, however, not been studied genetically and it is possible that they might contain maternal genetic connections to the Khwe.

All Pygmy groups lost their original language, which make linguistic connections impossible. The linguistic connection between Hietshware, Kwadi and Sandawe is extremely ancient and barely distinguishable (Güldemann, Forthcoming-b; Güldemann and Elderkin, Forthcoming). The limit of tracing relationships between languages is ~10 000 years. If there is a genetic counterpart to the linguistic connection between the Sandawe and Khoe-San, the age of convergence of genetic lineages cannot much be older than this limit. The TMRCA of all L0k1 lineages from this study was between 7 000 and 19 000 years BP depending on the mutation rate employed (Table 3.2). For L0k1 to be the maternal genetic counterpart to the linguistic connection between the Sandawe and Khoe-San, more sampling of African sequences is necessary to establish if and where other L0k sequences are found in Africa.

L0d

Altogether, 317 sequences were resolved into haplogroup L0d and its sub-haplogroups and 111 unique haplotypes were identified. Haplotypes were grouped into the seven sub-haplogroups according to Behar *et al.*, (Behar *et al.*, 2008) and two extra previously unidentified haplogroups (Figure 3.7). Overall, there was good agreement in the resolution of haplotypes in the present study (Figure 3.6) with the study based on whole genome sequences (Figure 3.7) published by Behar *et al.*, (Behar *et al.*, 2008). The L0d clade-defining mutation, 16243 T-C, is a very stable mutation and did not reoccur or revert in our sample set of 538 sequences.

The L0d haplogroup was estimated to have a coalescence time of 87 705 years BP (+/- 18 067) and diverged from the other L0 groups 96 601 years BP (+/- 20 138) (Table 3.2). This compared well to whole genome studies (Behar *et al.*, 2008), which calculated the coalescence at 100 795 (+/-10 317) and the divergence at 152 384 +/-12,698.

All of the Khoe-San and Coloured groups, with the exception of the Khwe, had L0d as their most frequent haplogroup. The Khwe L0d frequencies were lower than that found in the southeastern Bantu-speaking groups (Figure 3.3 and 3.4). In the remaining Khoe-San and Coloured groups the frequencies of L0d ranged from exclusive (Karretjie people -100%), to very high (‡Khomani - 98%, Coloured-Northern Cape - 93%, /Gui + //Gana + Kgalagari - 90%), moderate (Nama - 71%, Ju\`hoansi - 71%, !Xun - 65%, Coloured-Coleberg - 65%) and lower (Coloured-Wellington - 45%) (Figure 3.3 and 3.4).

Other studies found similar L0d frequencies in the !Xun and the Khwe. Haplogroup L0d was found in the !Xun and Khwe at frequencies of 51% and 16%, respectively (Chen *et al.*, 2000) (Table 1.2) while in the present study the frequencies were 65% and 11%. Tishkoff *et al.* reported frequencies of 61% in a combined !Xun and Khwe group (Table1.2) (Tishkoff *et al.*, 2007).

Again as was noted for L0k, the L0d frequencies in the Ju\`hoansi from our study (71%) (Figure 3.3, 3.4) did not compare to what was published for a different Ju\`hoansi group (96%) (Vigilant *et al.*, 1991) (Table1.2). Although both Ju\`hoansi groups showed little admixture from Bantu-speaking groups (Figure 3.3, 3.4, Table1.2) the Ju\`hoansi from the present study had proportionally more L0k1 and less L0d contribution than the Ju\`hoansi from the published study (Vigilant *et al.*, 1991) (Table1.2). As the published group was sampled in Botswana (Dobe) as well as in Namibia while the group of the present study was sampled in Tsumkwe, it is possible that the Tsumkwe group had more admixture with the neighboring !Xun (Figure 1.2).

Interesting patterns are overlooked by only considering the distribution of the L0d group as a whole among the different Khoe-San and Coloured groups. The distribution of L0d sub-haplogroups in published studies was by no means homogenous for the different groups (Table 1.3). In the present study differential distribution of the L0d sub-haplogroups were also observed and their distributions are visually represented by different contour plots (Figure 3.11). The most striking feature was the absence or low frequencies of L0d1c in the southern groups as well as L0d3 and L0d2a in the northern groups. Furthermore the L0d/k

sub-haplogroups showed that they had different associated histories when comparing their expansion dynamics through looking at mismatch distributions, neutrality tests and Bayesian Skyline Plots (BSPs) (Table 3.3, 3.4 and Figure 3.13). Their varied distribution coupled to dissimilar expansion patterns clearly indicated that the history of L0d as a whole is not homogenous over sub-groups and may not correctly represent individual dynamics. Rather, each sub-haplogroup must be studied separately in accordance with the histories of its carrier population group and region of occurrence.

L0d3

L0d3 have been identified as the oldest L0d clade previously (Behar *et al.*, 2008) and also occurred as the earliest L0d branch on our tree and network (Figure 3.5 and 3.6). The divergence from the other L0d sequences dated to 71 174 years BP (+/- 23 375) (Table 3.2). This was the earliest split in the L0d branch (excluding the problematic L0d2b branch which will be discussed later). L0d3 sequences coalesced 35 587 years BP (+/- 15 157). According to the whole genome study of Behar *et al.*, it separated from the other L0d3 groups ~100 000 years BP (Behar *et al.*, 2008). Only two L0d3 sequences formed part of the whole genome study, one from a San individual and one from an individual from Kuwait, these two sequences coalesce ~31 000 years BP (Behar *et al.*, 2008).

The seven clade-defining mutations (150T, 316A, 523insCA, 16290T, 16300G) that were previously proposed were based on 2 sequences (Behar *et al.*, 2008). The L0d3 clade in the present study is represented by seven haplotypes and based on data displayed in the network (Figure 3.6), it is suggested that 16300G and 523delCA should be removed as clade defining mutations for L0d3 since earlier sequences in our network did not contain these mutations.

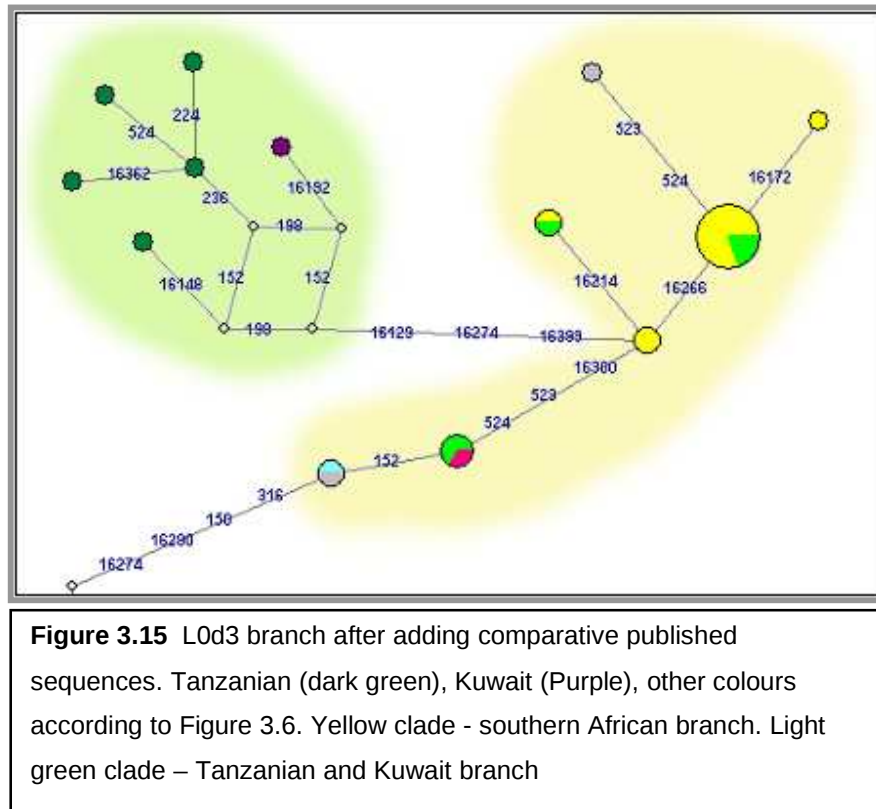
Prior to the Behar *et al.*, (Behar *et al.*, 2008) study, Gonder *et al.*, did whole genome studies on a wide range of African sequences including the Platfontein !Xun and Khwe and the Sandawe from Tanzania (Gonder *et al.*, 2007). The !Xun/Khwe and Sandawe were the only groups that contained L0d sequences (Table 1.2). When the sequences from Gonder *et al.*, were classified according to the classification introduced by Behar *et al.*, all the !Xun/Khwe L0d sequences belonged to L0d1 or 2 (Table 1.3), while all the Sandawe L0d

sequences belonged to L0d3. Since L0d sub-haplogroup classification was not formalized at that time, Gonder *et al.*, coined the L0d1+2 and L0d3 groups they observed in their phylogeny, L0d-South Africa and L0d-Tanzania. According to the Gonder *et al.*, whole genome study the two L0d branches separated ~58 000 years BP, the L0d-South Africa branch coalesced 90 000 years BP and the L0d-Tanzania branch 31 000 years BP (Gonder *et al.*, 2007).

L0d3 sequences (L0d-Tanzania from Gonder *et al.*) were identified in various Khoe-San groups in the present study. When analyzing these L0d3 sequences with the L0d3 sequences from Tanzania and Kuwait (Behar *et al.*, 2008), two clear separate groups were formed. The Tanzania/Kuwait group formed a subgroup of the southern Africa L0d3 group (Figure 3.15). It is suggested that the Tanzanian/Kuwait sequences is coined as L0d3 subgroup, L0d3a, that is defined by the 16129, 16274 reversion and 16399 mutations. The closest related haplotype in the southern Africa branch of L0d3 to the Tanzania/Kuwait branch (L0d3a) occurred in the Karretjie and Coloured groups from Colesberg. A haplotype found in the Karretjie and Coloured groups was directly ancestral to the L0d3a branch. When the Tanzanian, Kuwait and present study L0d3 sequences were put together, the whole L0d3 clade diverged from other L0d sequences ~83 000 years BP while all the L0d3 sequences coalesced 47 000 years BP (Table 3.2). The divergence of the L0d3a sub-haplogroup from the southern Africa haplotypes was dated at ~41 000 years BP. L0d3a sequences converged ~28 000 years BP.

The present study confirmed that the L0d3 branch was not limited to Tanzania (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). L0d3 was present in the southern Khoe-San and Coloured groups but almost absent in the northern groups (Figure 3.7 and 3.8). Although L0d3 had low frequencies compared to the other L0d subgroups, in all the southern groups, its distribution clearly showed a south-north cline. L0d3 had the highest frequency in the southern groups, it declined northwards in the central groups and was absent in the northern groups (only one !Xun individual was assigned to L0d3) (Figure 3.7 and 3.8). The earliest haplotypes in the L0d3 clade were the !Xun individual and a BS individual. The BS individual, however, was from the south, a Zulu from the Drakensberg area where the

Duma San individuals were collected (in close proximity to the Karretjie and Coloured groups).



The low frequency in the northern groups are confirmed by the Gonder *et al.*, and Tishkoff *et al.*, studies that found no L0d3 sequences in their group of !Xun and Khwe (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). Also when the haplogroups of the !Xun and Khwe from the Chen *et al.*, study (Chen *et al.*, 2000) and the Ju!hoansi group from Vigilant *et al.*, (Vigilant *et al.*, 1991) were classified according to the Behar *et al.*, nomenclature (Behar *et al.*, 2008), no L0d3 sequences were observed. When the results from the three studies mentioned above were taken together with results from the present study, a group of 225 !Xun, Khwe and Ju!hoansi were screened and only one !Xun individual (from the present study) contained an L0d3 sequence. This indicates an extremely low incidence of L0d3 in the northern San groups.

Tishkoff *et al.*, discussed the possibility that the linguistic connection between the Sandawe and northern Khoe languages was associated with the L0d genetic connection (Tishkoff *et al.*, 2007). They concluded that the maternal genetic connection between the two groups was very deep (>15 000 years) and it was unlikely that linguistic trace can be detected that far back.

From the present study it was clear that it is unlikely that the linguistic connection of the Sandawe to the northern Khoisan-speaking groups was associated with the L0d3 lineage. Although L0d3 was the exclusive L0d lineage in the Sandawe, it was almost completely absent in the northern Khoisan-speaking groups. In contrast to this absence in the northern groups, the southern groups contained higher L0d3 frequencies and the frequencies were the highest is in the Karretjie group from Colesberg. Furthermore L0d3 sequences were detected in a Bantu-speaking individual from Mozambique (Salas *et al.*, 2002) as well as an individual from northern Kenya (Watson *et al.*, 1997) and an individual from Kuwait (Behar *et al.*, 2008). This suggests an L0d3 spread along the eastern part of Africa forming a connection between the southeastern Khoe-San groups and the Tanzanian Sandawe rather than between the northwestern Khoe-San groups and the Sandawe as the linguistic connection suggests.

To investigate the expansion history associated with the L0d3 haplogroup that led to its geographic spread, the expansion dynamics were investigated. While the expansion hypothesis was not rejected using mismatch distributions, all three neutrality tests used, rejected an expansion for L0d3 (Table 3.4, Figure 3.13 and Table 3.3). BSP analysis (Figure 3.14) indicated that the southern African L0d3 show a steady decline from the coalescence point onwards with a sharp increase starting ~2 000 years BP. When the east Africa sequences were included the decrease was not as severe and the recent expansion started earlier (~4 000 years BP). The recent expansion phases of the haplogroup correlated with the introduction of pastoralism in east Africa and southern Africa, respectively. It is therefore likely that the populations that carried L0d3 either adopted the herding economy or benefited from it.

L0d1 and L0d2

The L0d1,2 branch was separated from their ancestral node by two mutations, 489delC and 16278C! (! Indicates back mutation compared to the Cambridge Reference Sequence) (Figure 3.6 and 3.7). No HVS mutations differentiate between L0d1 and L0d2 (Figure 3.7), they are, however, defined by 5 and 6 coding region mutations, respectively (Behar *et al.*, 2008). When analyzing only the HVS, the whole level separating L0d1 from L0d2, collapses. This can be seen in the network where all subgroups met at two central nodes (marked with * and ** in Figure 3.6). The absence of the 523insCA mutation (seen as an L0d1a-L0d1b defining mutation in Figure 3.7) separated the L0d2 sequences from the L0d1 sequences in the network. In Figure 3.7 L0d1a/b had the 523insCA mutation but L0d1c did not have the 523insCA mutation. In the network all three L0d1 sub-clades as well as the newly identified L0dx had the 523insCA mutation but in L0d1c it was seen as a back-mutation early in the clade. All of this caused the L0d2 groups to converge at node ** in Figure 3.6 and L0d1 groups at node *.

L0d1

In the network (Figure 3.6) L0d1c and L0d1a grouped on one branch because of the common 16234T mutation. According to whole genome sequencing, however, this was not a common ancestral event in these two branches and should rather be separate events on each branch (Figure 3.7). The 523insCA should group L0d1a and L0d1b together and not occur in all three L0d1 clades and then be lost due to back-mutation in L0d1c as seen on the network and described earlier.

Results from the present study indicated that the three L0d1 sub-haplogroups (L0d1a, L0d1b and L0d1c) showed differential spread among the different Khoe-San and Coloured groups and had different associated histories. According to whole genome studies, L0d1 diverged from L0d2 ~90 000 years BP and all L0d1 sequences coalesce 53 000 years BP (Behar *et al.*, 2008). The present study dated the coalescence at 57 000 years BP (Table 3.2).

L0d1a

L0d1a was further defined by 16223C!, 199C and 16266G. The 199C mutation occurred in one other African sequence (one L0dx sequence), four haplogroup M sequences and one haplogroup N sequence. The 16223 and 16266 mutations are both highly reoccurring mutations. The 16266G position mutated to a 16266A further on in the L0d1a clade in a subset of sequences.

The haplotype diversity for L0d1a (0.96) was the highest of all the L0d sub-haplogroups in the study (Figure 3.7). The present study contained 37 HVS sequences that included 21 haplotypes. The 21 haplotypes converged 37 000 years BP (Table 3.2 and Figure 3.9). This is a much later date than the whole genome study indicated (Behar *et al.*, 2008). The whole genome study included three L0d1a sequences (one Khoe-San and two Bantu-speaking), which converged ~18 000 years BP. This can be explained by the fact that the whole genome study did not include any of the haplotypes from the early branches of L0d1a identified in the network compiled from the present study (Figure 3.6).

L0d1a had a central distribution with the highest frequencies in the regions occupied by the #Khomani (Figure 3.7, 3.8 and 3.11). This haplogroup had low frequencies in most of the populations (<20%) but was geographically widespread and present in most groups (Figure 3.7 and 3.8). Even though the L0d1a frequencies were much lower in the northern groups than in the central and southern groups, the northern groups contained the oldest L0d1a haplotypes when looking at the network and trees (Figure 3.5b and Figure 3.6). The two earliest branches of L0d1a contained only individuals from northern groups while the later branches contained mostly central and southern groups with no northern group haplotypes. The northern group haplotypes were not directly ancestral to the central and southern haplotypes but were closer related to the common ancestor (Figure 3.6).

The BSP of L0d1a (Figure 3.14), showed a clear indication of an expansion that started between 25 000 and 30 000 years BP as well as a recent decline in population size from 3 000 – 4 000 years BP to present. The L0d1a network (Figure 3.6) showed a star-like expansion pattern associated with the southern groups. The central haplotype of the pattern was small and derivative sequences accumulated several mutations. This pattern is

indicative of an older expansion in which the central haplotype declined over time and derivative haplotypes accumulated mutations. A mismatch distribution of L0d1a showed a smooth unimodal distribution with a low raggedness index that indicated a single expansion of the haplogroup some time in the past indicated by the τ value (Figure 3.13 and Table 3.3). When τ was converted to years the expansion was dated to ~28 000 years BP. Of the three neutrality tests employed, only the F_s test showed a significant indication of an expansion (Table 3.4). The F_s statistic, however, is used widely and several studies showed it to be an accurate indicator of expansions (Ramos-Onsins and Rozas, 2002; Ramirez-Soriano *et al.*, 2008).

The high genetic diversity of L0d1a, widespread geographic distribution patterns and location specific expansion patterns, suggested population fragmentation, isolation and re-expansion. The expansion of L0d1a is chronologically associated with the start of the LSA (20 000 to 30 000 years BP) in the archaeological record. The archaeological record indicates technological innovation and the emergence of belief systems during this time. Certain sites including Lesotho, southern Cape, Caledon valley, southern Namibia and the southern Kalahari have indications of 'higher energy' human settlement (Deacon and Deacon, 1999; Mitchell, 2002). These sites more or less overlap with the distribution of L0d1a (Figure 3.11). A period of population growth was clearly indicated in the BSP of L0d1a and the carriers of this haplogroup contributed to population expansions during this period. In a period spanning ~25 000 years the L0d1a N_e increased from 20 000 to more than 100 000. The BSP of L0d1a furthermore showed a recent decline that started ~4 000 years BP. The archaeological record, however, indicates a drastic further increase in population size in the last 4 000 years BP. This increase was more prominent from 2 000 years BP onwards when herding was introduced in southern Africa (Deacon and Deacon, 1999; Mitchell, 2002). Reasons for the decline might be that groups carrying the L0d1a haplogroup in high frequencies were out-competed and displaced by other groups that expanded during this stage. These might be population groups moving in from other areas, or drift effects of other haplogroups increasing within the same population.

The L0d1a BSP thus demonstrated the complete opposite picture from L0d3. While L0d3 showed no evidence of a population expansion during the technological innovation period

of the early LSA a clear expansion pattern was observed for L0d1a. The population dynamics of the last 4 000 years BP was also reversed. While the N_e of L0d3 showed a sharp increase during the last 4000 years and at time of introduction of pastoralism, the L0d1a N_e showed a decline. Thus while the populations that carried L0d3 benefited from the introduction of pastoralism (either by directly adopting the lifestyle or benefiting from it through trade relations) the groups that carried L0d1a were negatively affected by pastoralism. It could be that L0d1a was the predominant group in the hunter-gatherer people that were displaced by the pastoralist groups or lifestyle.

L0d1b

The L0d1b clade is further defined by 16239T and 16294T (Figure 3.7). A large subset of sequences that were analysed (5 haplotypes containing 10 sequences), however, did not contain the 16239 mutation. They grouped on the network (Figure 3.6) as a sister group to an early branch in the L0d1b group that had not yet acquired the 16239 mutation. Again it might be that this group, rather than being an early branch without the mutation, lost the 16239 mutation. Again, this hypothesis will have to be ascertained with whole genome sequencing, but for now it is suggested that 16239 should not be used as a clade defining mutation for L0d1b. Only four haplotypes were used in the whole genome classification of L0d1b (Behar *et al.*, 2008) and such a back-mutation was not present in these haplotypes. The 16239 mutation occurred furthermore in four L3e1g sequences and one sequence in haplogroup U and H each. The 16294T mutation is a highly reoccurring mutation.

The coalescence times of the 29 HVS haplotypes from the present study (~33 000 years BP) and the four haplotypes employed in the whole genome study (~35 000 years BP) (Behar *et al.*, 2008) were similar (Table 3.2 and Figure 3.9).

L0d1b had a distribution that is concentrated in the south and declined towards the north (Figure 3.11). The Cape Coloured group had L0d1b as their most prevalent L0d haplogroup, while in the other southern groups it was the second most prevalent (Figure 3.7 and 3.8). Interestingly it was also the most prevalent group in the Ju'hoansi while frequencies in the other northern groups were lower (Figure 3.7 and 3.8). Published studies

also found L0d1b to be the predominant haplogroup in the Ju\’hoansi (Vigilant *et al.*, 1991) while occurring at low frequencies in the !Xun (Chen *et al.*, 2000) (Table 1.3).

The expansion dates for L0d1b indicated by the BSP and mismatch distributions matched the LSA archaeological record of the southern parts of Africa perfectly. According to archaeological sites the population density increased markedly from 13 500 years ago and particularly in the last 4 000 years (Deacon and Deacon, 1999; Mitchell, 2002). This is almost exactly the pattern observed for the L0d1b BSP (Figure 3.14). The first expansion began ~14 000 years BP and the second expansion ~3 000 years BP. In the first expansion the female N_e increased from ~20 000 to ~70 000 in a period of ~12 000 years. The second expansion was more rapid and the female N_e increased from ~70 000 to ~110 000 in a period of ~3 000 years (Figure 3.14). The network also showed several star-like expansion patterns indicating that the haplogroup went through more than one phase of population growth (Figure 3.6). Furthermore, the mismatch distribution indicated more than one expansion (a recent and an older expansion) through a multimodal distribution (Figure 3.13). Additionally, all three neutrality tests significantly supported statistics that indicated expansion (Table 3.4).

From the L0d1b network, the groups involved in the expansions could be identified (Figure 3.6). The older expansion (~14 000 years BP) consisted of two star-like expansion patterns. A smaller expansion that involved mainly the northern groups and a larger expansion involving mainly the southern groups. Derivative haplotypes of the larger expansion, however, included individuals from northern groups indicating a possible migration of individuals from the southern groups to the northern groups. The later expansion pattern (~3 000 years BP) had Ju\’hoansi haplotypes in the center haplotype and southern group haplotypes as derivative haplotypes. The central haplotype, however, was not as big as one would expect from a recent expansion. It might be that a population group that contains high frequencies of the central haplotype was not sampled in this study.

Thus L0d1b was associated with the southern groups as well as the Ju\’hoansi but occurred at low frequencies in the other northern groups. The network suggested several

instances of migration between the southern groups and the Ju\’hoansi. Furthermore the expansion times of the L0d1b haplogroup reflected expansions noted in the archaeological history of southern Africa.

L0d1c

In addition to 16234, L0d1c was further defined by 456T and 16129G!. While 16129 is a highly reoccurring mutation, 456T only appeared in two other sequences (one L0d1b sequence and one L0d2a sequence).

The L0d1c haplogroup contained 59 L0d1c HVS sequences described by 12 unique haplotypes. The 12 haplotypes coalesced ~42 000 years BP and separated from L0d1a/b ~60 000 years BP (Table 3.2 and Figure 3.9). According to the whole genome study L0d1c separates from L0d1a/b ~53 000 years BP and the six genomes studied coalesced 24 000 years BP (Behar *et al.*, 2008).

L0d1c was completely absent or at very low frequencies in the southern groups (Figure 3.11). It increased northwards in the central groups but the highest frequencies were in the northern groups, where it is the predominant L0d group (except in the Ju\’hoansi) (Figure 3.7 and 3.8). Interestingly the L0d1c frequency was lower in the Ju\’hoansi and L0d1b rather was the most prevalent group (Figure 3.7 and 3.8). Results from the present study were also supported by published results, where L0d1c was the predominant group in the !Xun (Chen *et al.*, 2000) and were undetected in the Ju\’hoansi (Vigilant *et al.*, 1991) (Table 1.3).

Behar *et al.*, furthermore classified a sub-clade of L0d1c, namely, L0d1c1 defined by the 16242T, 16167T and 198T mutations (Behar *et al.*, 2008). Four of the six L0d1c haplotypes in the whole genome study belonged to the sub-haplogroup L0d1c1 (Behar *et al.*, 2008). In our study 6 of the 12 haplotypes belonged to L0d1c1, however, most L0d1c sequences fell into L0d1c1 (Figure 3.7). L0d1c1 could be seen on the network as a large star-like pattern at the tip of the L0d1c network, which indicated a recent expansion (Figure 3.6). The L0d1c1 sub-group contained more !Xun haplotypes than the earlier L0d1c haplotypes. When the contour plot of L0d1c was split between the early L0d1c haplotypes and the

L0d1c1 haplotypes it was apparent that the early L0d1c haplotypes had its highest frequencies in the central /Gui + //Gana + Kgalagari group while almost absent in the !Xun (Figure 3.12). L0d1c1 haplotypes, however, had its highest frequency in the !Xun. It therefore seems that L0d1c was originally present in the /Gui and //Gana and then spread to the !Xun before L0d1c1 expanded. L0d1c1 was also present in the ±Khomani and Ju\’hoansi, although at lower frequency.

The low frequency in the Ju\’hoansi was surprising given that the Ju\’hoansi is geographically located between the !Xun and the /Gui + //Gana. The Ju\’hoansi and !Xun lifestyles, however, are vastly different. While the Ju\’hoansi continued to live a foraging lifestyle the !Xun adopted crop cultivation and herding from the local Ovambo population with whom they have lived in close association for centuries (De Almeida, 1965; Barnard, 1992). This can be very clearly observed in the BSP for L0d1c (Figure 3.14). The BSP indicated that the N_e started to decline at the time of the introduction of pastoralism to the area but then turned around dramatically about 1 000 years BP and increased rapidly. It might be that the groups carrying L0d1c did not initially adopt the pastoralist lifestyle and were outcompeted by pastoralists. The situation, however, dramatically switched when the groups (such as the !Xun) adopted pastoralism and this led to a fast increasing N_e . In this short period the N_e doubled from ~20 000 to ~40 000 (Figure 3.14).

In the mismatch distribution the expansion hypothesis was not rejected, but it was the lowest value to be accepted of all the L0d haplogroups (Table 3.3). The τ value, however, indicated that the expansion evaluated was around the start of the LSA and not the recent expansion (Table 3.3). The recent expansion was, however, noticeable in the mismatch graph (Figure 3.13). The recent expansions of L0d3 and L0d1b could also be seen in the mismatch graph but the LSA expansions were evaluated rather (indicated by the τ value) (Figure 3.13 and Table 3.3). Thus the mismatch distributions showed the recent expansions in the mismatch graphs but did not test their significance or note their τ value. Therefore, when the expansion hypothesis for L0d1c was not rejected it was based on an expansion during the LSA transition. The BSP plot (Figure 3.14) showed a slight increase in population size from around 25 000 years BP until 5 000 years BP. This increase,

however, was not comparable to the dramatic increase seen in L0d1a and L0d1b. This could also be seen in the Model (SSD) p-value of the mismatch distribution (Table 3.3), where the value of L0d1c was much lower than the values of L0d1a and L0d1b, although the time frames were more or less the same.

The expansion hypothesis was rejected in all three neutrality tests (Table 3.4). The recent expansion in L0d3 was, however, also not detected by the neutrality tests and it might be that neutrality tests, similar to mismatch distributions, are not sensitive to recent expansions.

To summarise, L0d1c showed slight evidence of a LSA transition population growth. This stage was, however, not as prominent as observed for L0d1a. The reaction of the L0d1c N_e upon the introduction of pastoralism in the area was more complex than seen for the other L0d haplogroups. Initially the groups that carried L0d1c were negatively affected by this stage but the situation turned around resulting in a steep increase in N_e . This turnaround is likely to be due to the adoption of pastoralism and cultivation practices as seen in the !Xun, in whom the L0d1c1 haplogroup was the predominant haplogroup.

L0d2

Three subgroups within L0d2 were previously identified; L0d2a, L0d2b and L0d2c (Behar *et al.*, 2008). L0d2c split first from L0d2a/b (Figure 3.7). The present study had representation across these three L0d2 haplogroups and also identified a fourth group; henceforth called L0d2d (Figure 3.7). L0d2d grouped with L0d2a/b and all three these groups were defined by the 16212G mutation. The 16212 mutation is relatively stable and occurred in only one other sequence in the sample group (haplogroup M) and was seen to revert to the ancestral state in one of the L0d2a sequences. In the present study all L0d2 haplotypes coalesced 43 000 years BP (Table 3.2). The whole genome study calculated coalescence to ~64 000 years BP (Behar *et al.*, 2008).

L0d2a

L0d2a was further defined by 597T and 16390A (Figure 3.6). Behar *et al.*, (Behar *et al.*, 2008) suggested 198T as an L0d2a defining mutation as well but one of the L0d2a

sequences did not contain the mutation. It might, however, be that this sequence contained a back mutation rather than being an ancestral sequence to the other L0d2a sequences as seen in the network (Figure 3.6). The 16390 mutation is a reoccurring mutation that occurred in several other non-L0d2a sequences. The 597 mutation occurred only one other place in the total sample, in one L2a1f sequence.

The 27 L0d2a HVS haplotypes of the present study had a TMRCA of 16 000 years BP (Table 3.2 and Figure 3.9). Coalescence analysis (applied in the BSP - Figure 3.14), however, dated the coalescence of L0d2a at ~8 000 years BP. Based on the eleven haplotypes in the whole genome study, L0d2a coalesced 9 000 years BP (Behar *et al.*, 2008).

The L0d2a haplogroup had a distribution concentrated in the south (Figure 3.11). Its highest frequency was in the Karretjie where it was the most prevalent L0d group (Figure 3.7 and 3.8). It was also the most prevalent L0d group in all the southern groups except the Cape Coloured where it was the second most prevalent subsequent to L0d1b. All the southern groups had either L0d2a or L0d1b as their most prevalent and second most prevalent groups (Figure 3.7 and 3.8). L0d2a was absent in most northern groups and at low frequencies in the !Xun. Interestingly L0d2a was the L0d group that had the highest incorporation into the Bantu-speaking groups (Figure 3.7 and 3.8).

L0d2a formed a large star-like expansion pattern that is indicative of a recent expansion in the population groups represented in the haplogroup. In the network L0d2a had the most pronounced star-like expansion pattern of all the L0d/k haplogroups (Figure 3.6). This indicated a massive expansion associated with the southern groups. All three neutrality tests detected an expansion in L0d2a with the highest associated significance of all the L0d/k subgroups (Table 3.4). The mismatch distribution did not reject an expansion hypothesis and the mismatch graph shows a smooth unimodal curve that indicated a recent expansion (Figure 3.13 and Table 3.3). The τ value dated the expansion to around 7 000 years BP (Table 3.3). Looking at the BSP (Figure 3.14) one can see an immediate dramatic increase in L0d2a N_e from the coalescence date (~8 000 years BP) onwards until present. A further, recent expansion (~1 000 years BP) was also evident. In the span of 8

000 years the N_e increased from between ~5 000 to ~110 000 (Figure 3.14). This remarkable increase necessitates an explanation.

In most haplogroups the introduction of pastoralism apparently led to abrupt increases in population sizes (Figure 3.14). The expansion of L0d2a, however, predated these expansions. L0d2a did have a more recent expansion phase that correlated with the introduction of pastoralism, however, the major part of the L0d2a expansion phase predated the introduction of sheep into the southern regions. The steep expansion in N_e indicated that carriers of L0d2a had a distinct advantage over other L0d haplogroup-carriers during this time and L0d2a N_e increased more rapidly. This rapid increase might be part of the increase noted in the archaeological record that occurred from 14 000 years BP onwards. This was, however, before the coalescence time indicated on the BSP. From archaeological and paleoenvironmental studies we do know that the period between 10 000 BP and 5 000 BP is associated with the reach of maximum temperatures after the LGM and the completion of the rise in sea level. It might be that these events concentrated populations and increased social networking, which led to the spread of technologies between groups in the south. Expansions into new habitats and elaboration of material culture and technology, especially in the Cape Fold Belt and Thukela basin are noted in the archaeological record from 4 000 years BP onwards.

It is difficult to judge which populations contained ancestral haplotypes to the L0d2a expansion haplotype. It seemed that such an ancestral haplotype is present in the Karretjie and Coloured groups from Colesberg. This was based on just one mutation (198) and the fact that this haplotype did not contain the mutation could have been due to a reversion. Whole genome sequencing is important to see if this haplotype was indeed ancestral to the haplotype central to the expansion.

The L0d2a haplotypes had a very shallow coalescence time (~ 8000 years BP) and the L0d2a haplogroup should be much older than this date. L0d2 diverged from the other L0d2 groups ~34 000 years BP (Table 3.2). We, however, do not have representative haplotypes from these earlier times. They might be present in other populations that have not been

studied yet. If these earlier haplotypes were incorporated, one would get a clearer picture of where and when the L0d2a expansion started.

L0d2b

The defining mutations of L0d2b were 16069T and 16169T. Both mutations are stable. The 16069 mutation occurred in no other African sequence but did occur in the European haplogroup J. Only one other sequence in the sample group (an L4b2a2 sequence) contained the 16169 mutation.

Four of the six sequences in L0d2b were separated by a very long branch (9 mutations) from the other two sequences, indicative of a very long separation time between the sequences. The L0d2b node as a whole also had a high coalescence and divergence time, higher than the L0d2 branch as a whole. This is indicative of an inconsistency. A possible explanation might be that the mutations 16182, 16183 (and possibly 16187) were one mutational event. Also the 152 mutation is a highly reoccurring mutation. If the weights of these mutations were decreased it would reduce the coalescence times. Another explanation might be that the terminal nodes in this group were grouped incorrectly within L0d1b. They did contain the 16212, 16069 and 16169 mutations but if you look at the coalescence times and how they compared with other haplogroups, these haplotypes might well be representative of another L0d haplogroup (not a subgroup within L0d2b) (Table 3.2 and Figure 3.9). Whole genome sequencing for these samples needs to be done to precisely assess their relationship to other groups. In the whole genome study (Behar *et al.*, 2008), the one L0d1b sample included, only contained the 16212, 16069 and 16169 mutations and not the additional mutations that defined the three terminal groups in the L0d2b clade on the network.

L0d2b was detected at very low frequencies in the present study (only six sequences in total that represents four haplotypes) (Figure 3.7 and 3.8). The L0d2b haplogroup was represented by only one haplotype in the whole genome study (Behar *et al.*, 2008). It was also not detected previously in the !Xun, Khwe or Ju'hoansi (Vigilant *et al.*, 1991; Chen *et al.*, 2000). In the present study L0d2b was detected at levels < 5% in four groups and had its highest prevalence in the /Gui + //Gana + Kgalagari and Nama (Figure 3.7 and 3.8). Due

to the low frequency of this haplotype it was impossible to draw any conclusions about the history of the haplotype, but this group seemed to be associated more with the central groups in our study group.

L0d2d

L0d2d is a new group suggested in this thesis. It grouped in the same clade as L0d2a and L0d2b (defined by 16212G) and was further defined by 188A-G. It did not contain the 16390, 597 and 16069, 16169 clade-defining mutations of L0d2a and L0d2b. The 188G mutation, however, is a reoccurring mutation and whole genome sequencing of the representative sequences of L0d2d would be necessary to affirm its position in the L0d clade.

Although L0d2d was not identified in the study of Behar *et al.*, (Behar *et al.*, 2008) haplotypes that can be classified as L0d2d using the new nomenclature were reported previously in Bantu-speakers (Salas *et al.*, 2002) and in the !Xun/Khwe (Tishkoff *et al.*, 2007). In the present study L0d2d was confined to the Ju!’hoansi, where it represented 5% of L0d/k haplogroups (Figure 3.7 and 3.8). Interestingly, this rare haplogroup was also found in an Indian individual. Frequencies of the haplogroup were too low to extract any information regarding the history of the haplogroups, however, its distribution did seem to be limited to the northern San groups.

L0d2c

The L0d2c sub-haplogroup consisted of 17 sequences grouped into six haplotypes. The coalescence of the L0d2c haplotypes was dated to ~20 000 years BP. The divergence from L0d2abd was dated to ~29 000 years (Table 3.2 and Figure 3.9). This date was much more recent than the date from the whole genome study. In the whole genome study, four L0d2c haplotypes were included that coalesced ~21 000 years BP and split from L0d2a/b ~64 000 years BP (Behar *et al.*, 2008). The more recent date of the present study can be explained by the relative little HVS variation that defines the L0d2c haplogroup compared to other haplogroups. Only the HVS-II mutation, 294a distinguishes the L0d2c from the

L0d1'2 core haplotype (Figure 3.7). Several coding region mutations, however, separate L0d2c from L0d2abd and also L0d2 from L0d1'2 (Behar *et al.*, 2008).

L0d2c were found at lower frequencies in the sample group (Figure 3.7 and 3.8). L0d2c had its highest frequencies in the !Khomani and Nama, in other groups it was < 5% of L0d/k haplogroups (Figure 3.7 and 3.8). A star-like expansion pattern in the network could be seen and seemed to be associated with the !Khomani group (Figure 3.6). Due to the low frequency of the haplotype the expansion hypothesis was rejected in the mismatch distribution and only one neutrality test detected evidence for an expansion (Table 3.3 and 3.4). A signature of a recent expansion could, however, be observed in the mismatch graph (Figure 3.13). This recent expansion in L0d2c correlated temporally with the recent expansions in L0d1c and L0d3 and is likely to be associated with the introduction of pastoralism.

L0dx

L0dx is the second new haplogroup suggested in this thesis. Its position on the tree was, however, unresolved and can only be affirmed through whole genome sequencing. It did appear that it might group with L0d1a and L0d1b due to the presence of the 523insCA mutation. This cannot, however, be said with certainty due to the instability of this length repeat mutation, therefore the preliminary designation L0dx. This group was further defined by the 16179T mutation, which only reoccurred once in the total sample set, in one L0d2a sequence.

L0dx, was found only in the two northern-most groups, Khwe (11%) and !Xun (4%) (Figure 3.7 and 3.8). L0dx was the only L0d haplogroup found in the Khwe. In the study of Chen *et al.*, L0dx was found at similar frequencies in the !Xun (6%) but at much higher frequency in the Khwe (42%), where it also was the only L0d haplogroup (Chen *et al.*, 2000).

Both the present study and the study of Chen *et al.*, (Chen *et al.*, 2000) thus found L0k and L0dx to be the only non-Bantu-speaking haplogroups in the Khwe. From the network (Figure 3.6) it could be seen that the Khwe all belong to one haplotype and a !Xun haplotype was ancestral to the Khwe haplotype. It therefore seemed that L0dx was an

original !Xun haplotype and through geneflow moved to the Khwe. This was the reverse situation as was seen for L0k. The representative haplotypes of L0dx was, however, very low and more L0dx haplotypes need to be sampled before any deductions can be made with certainty (the !Xun and Khwe L0dx haplotypes reported in Chen *et al.*, (Chen *et al.*, 2000) did not include the 16399 and 574 regions (see network in Figure 3.3 and 3.4) and therefore could not be resolved further).

3.3.4 Summary of haplogroup histories

All the L0d1 haplogroups (L0d1a, L0d1b, and L0d1c) showed signs of expansion during the LSA period that coincides with the development of advanced technologies and belief systems. The two haplogroups with a current southern distribution, L0d1a and L0d1b, had stronger expansion signals than the L0d1c haplogroup, which are currently associated with northern San groups. L0d1a had a growth signal that precedes the L0d1b growth phase by at least 10 000 years.

It was difficult to judge the start of expansion in L0d2a because of the shallow coalescence time of haplotypes. It might be that the L0d2a growth started in the same timeframe as L0d1b. L0d2a and L0d1b was the main groups in the southern populations and might share similar histories. The growth curve in L0d2a was, however, much steeper than in L0d1b. Overall it seemed that the population growth signals of the early and middle LSA had a stronger association with the haplogroups presently found in the southern Khoe-San groups.

In contrast to the above-mentioned haplogroups, L0d3, showed no evidence of the LSA associated expansions. Yet, it also had southern distribution. Drift effects could cause this haplogroup to decrease while other haplogroups in the same populations increased. Another explanation could be that this haplotype was not subjected to similar conditions as the other L0d haplotypes during the early and middle LSA and thus might have only been introduced to these territories after this stage.

Most haplogroups showed expansions during the start of the Iron Age accompanied by the introduction of pastoralism to the southern parts of Africa. An exception was haplogroup L0d1a that showed a decrease during this time. This decrease again could be due to drift or could indicate that the groups carrying L0d1a in high frequencies were negatively affected by this stage. It is historically known that when pastoralists enter a territory they displace the hunter-gatherers to fringe areas, which is unsuitable for their animals. This would then impact on the success of the hunter-gatherer population measured through population growth. From this it is deduced that carriers of L0d1a could possibly have been populations that continued their hunter-gatherer lifestyles and did not adopt pastoralism or enter in to favorable relationships with pastoralists. Initially L0d1c also started to decline, similar to L0d1a, but then turned around. This turnaround might be associated with the recent adoption of pastoralism practices in the !Xun (See discussion above).

3.3.5 Haplogroup contributions from neighboring population groups

In addition to the L0d/k groups in the Khoe-San and Coloured groups there were also a contribution of haplogroups resulting from admixture from Bantu-speakers and Eurasian groups (Figure 3.3 and 3.4). From the groups that represent the people with southern Khoe-San ancestry, the Karretjie and #Khomani groups had almost exclusive L0d maternal lines, while the Coloured people from the Northern Cape also had very high percentages of L0d. The Coloured group with the largest proportion of admixture was the sample group from Wellington, with 20% Eurasian admixture and 35% Bantu-speaking admixture. The Colesberg Coloured group also had large proportions of Bantu-speaking (27%) and Eurasian (8%) admixture. The Coloured group from the Northern Cape had 5% Eurasian admixture and 2.5 % Bantu-speaking admixture. The three Coloured groups were the only groups with Eurasian admixture, the admixture in the remaining Khoe-San groups were due to gene-flow with the Bantu-speaking groups (Figure 3.3 and 3.4).

The Khwe group had the largest input from Bantu-speaking-groups (61%) with the two of the most common southeastern Bantu-speaking associated haplogroups, L2a and L3e making up the largest part (22% each) (Figure 3.3 and 3.4). The Nama had 21.5 % Bantu-

speaker admixture and in this case the Bantu-speaker-admixture was indicative of admixture with southwestern Bantu-speakers with L1c, L3d and L3f haplogroups contributing. The remaining San groups had < 10% Bantu-speaker admixture (Figure 3.3 and 3.4).

3.4 Mitochondrial genetic relationships between different Khoe, San, Coloured and neighboring groups

The previous section investigated the properties and differential distribution of the haplogroups in the various sample groups. To further investigate the population group diversities and their relationship to each other and their neighbours, genetic distances between the groups were considered.

To investigate the genetic differentiation and gene flow between groups Fst values between the different groups were calculated. Table 3.5 give the Fst values and Figure 3.16 and 3.17 give graphical representations of the Fst distance matrix in the form of PCA plots with minimum spanning trees and a cluster analysis tree.

Table 3.5 Mitochondrial population pairwise Fst values

	AFR	CAC	COL	DRC	EUR	CNC	GUG	HER	IND	JOH	KAR	KHO	KWE	NAM	SOT	XUN	ZUX
AFR	0.000																
CAC	0.155*	0.000															
COL	0.240***	0.010	0.000														
DRC	0.078*	0.110*	0.193**	0.000													
EUR	0.031	0.273	0.326**	0.206	0.000												
CNC	0.398***	0.094*	0.031	0.365***	0.495***	0.000											
GUG	0.457***	0.204***	0.149***	0.418***	0.563***	0.128***	0.000										
HER	0.217***	0.216***	0.260***	0.082**	0.363***	0.432***	0.489***	0.000									
IND	0.034	0.213	0.278***	0.110	0.082	0.433***	0.489***	0.245**	0.000								
JOH	0.393***	0.129***	0.089***	0.346***	0.495***	0.078***	0.108***	0.410***	0.435***	0.000							
KAR	0.518***	0.187**	0.081	0.487***	0.633***	0.036**	0.263***	0.557***	0.547***	0.179***	0.000						
KHO	0.468***	0.152**	0.073***	0.450***	0.567***	0.016***	0.182***	0.511***	0.505***	0.122***	0.057***	0.000					
KWE	0.211***	0.055***	0.106***	0.092***	0.320**	0.227***	0.272***	0.193***	0.258***	0.165***	0.319***	0.308***	0.000				
NAM	0.322**	0.033	0.012**	0.257	0.436	0.023**	0.147***	0.307	0.366*	0.057***	0.084***	0.057***	0.128***	0.000			
SOT	0.127**	0.006	0.063*	0.044*	0.229	0.193***	0.268***	0.129***	0.177*	0.208***	0.277***	0.274***	0.024***	0.113***	0.000		
XUN	0.428***	0.178***	0.125***	0.369***	0.523***	0.125***	0.079***	0.428***	0.461***	0.038***	0.204***	0.180***	0.172***	0.102***	0.235***	0.000	
ZUX	0.181**	0.000	0.013*	0.100	0.280	0.095***	0.186***	0.163**	0.230	0.124***	0.152***	0.158***	0.037**	0.029	0.000	0.146***	0.000

Abbreviations:

- * significant difference, $P < 0.05$
- ** significant difference, $P < 0.01$
- *** significant difference, $P < 0.00$

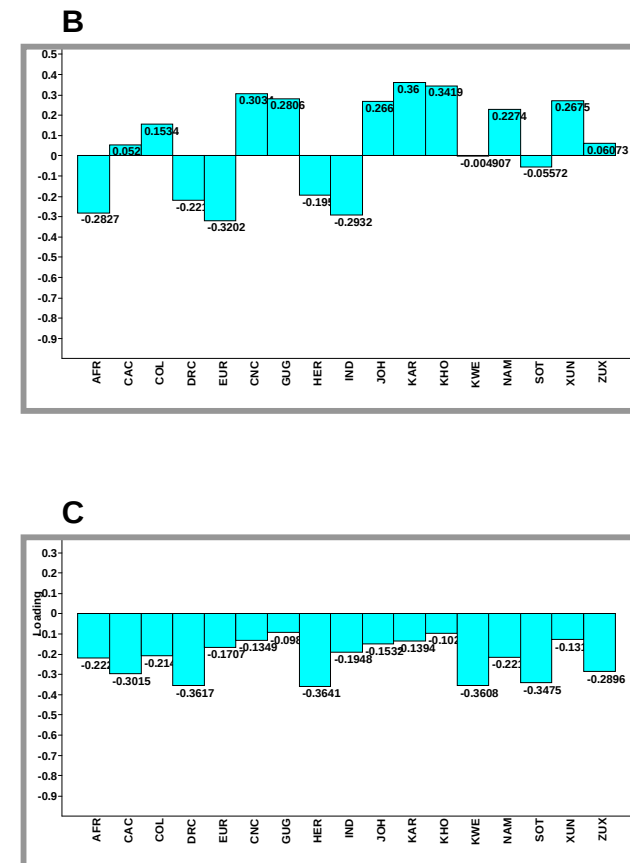
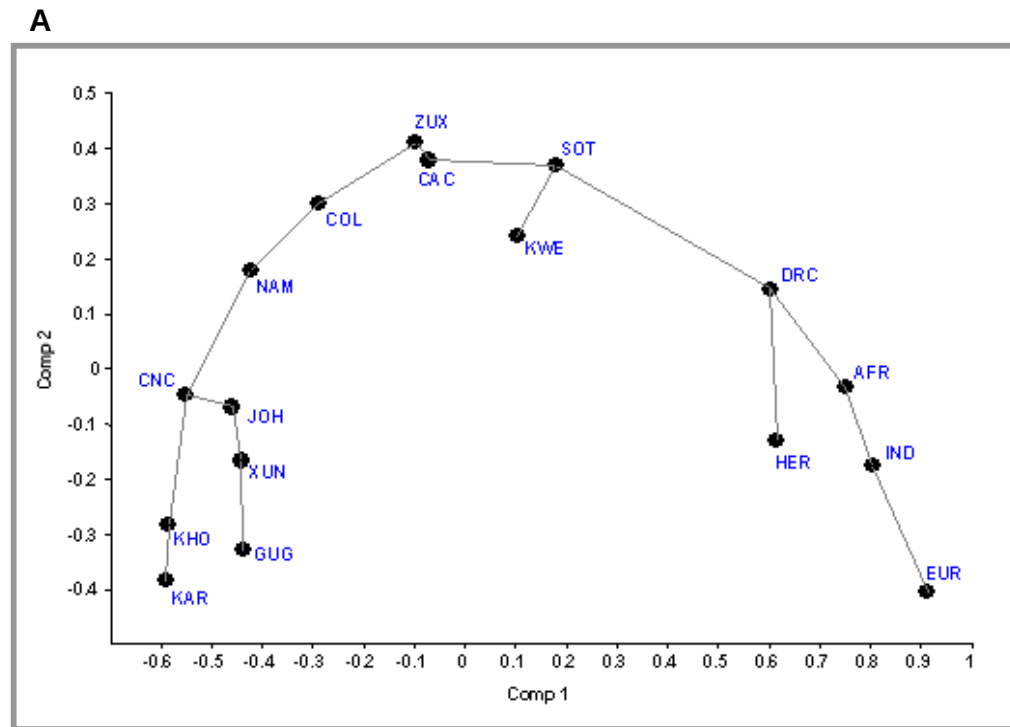
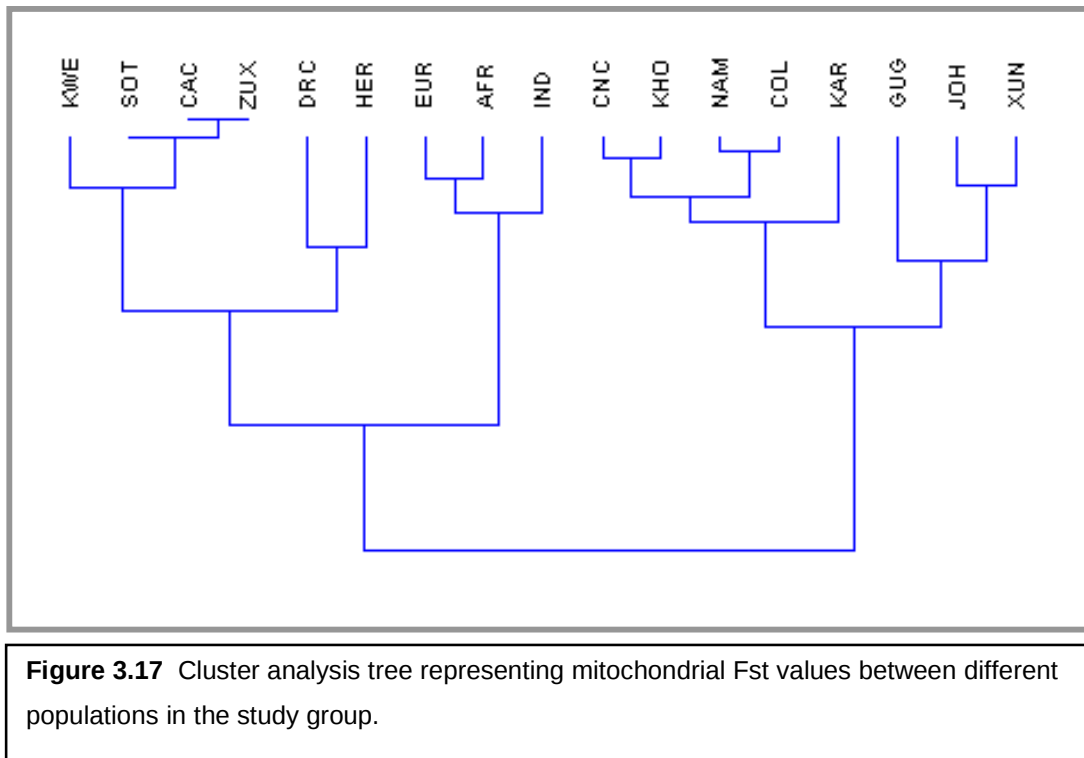


Figure 3.16 **A** – Principal component analysis of F_{st} values between different populations in the study group. A minimum spanning tree connects populations. Component 1 = 73.6% of the variation, Component 2 = 19.2% of the variation. **B** – Loadings for Component 1, **C** – loadings for Component 2.



As can be seen from Figure 3.17, the first split was between populations with considerable amount of Khoe-San ancestry (CNC, KHO, NAM, COL, KAR, GUG, JOH and XUN) and other populations (SOT, ZUX, DRC, HER, EUR, AFR, IND and CAC and KWE). Additional in this last cluster was the KWE and CAC. Evident from their haplogroup frequencies (Figure 3.3 and 3.4) both these groups had high amounts of admixture from Bantu-speakers (KWE) or Europeans and Bantu-speakers (CAC) causing them to group with these groups rather than Khoe-San groups. The PCA plot (Figure 3.16) also summarised this variation in the first component that contained 74% of the total variation. Reflected in the loadings of the first component (Figure 3.16 - B), the Khoe-San and Coloured populations were separated from the non-African and the HER and DRC. The KWE, CAC and also the ZUX and SOT was intermediate because of their various amounts of input/admixture from the Khoe-San populations.

The second component on the PCA plot contained 19 % of the variation (Figure 3.17). This component summarised the variation between Bantu-speakers and the rest of the groups (Figure 3.17 - C).

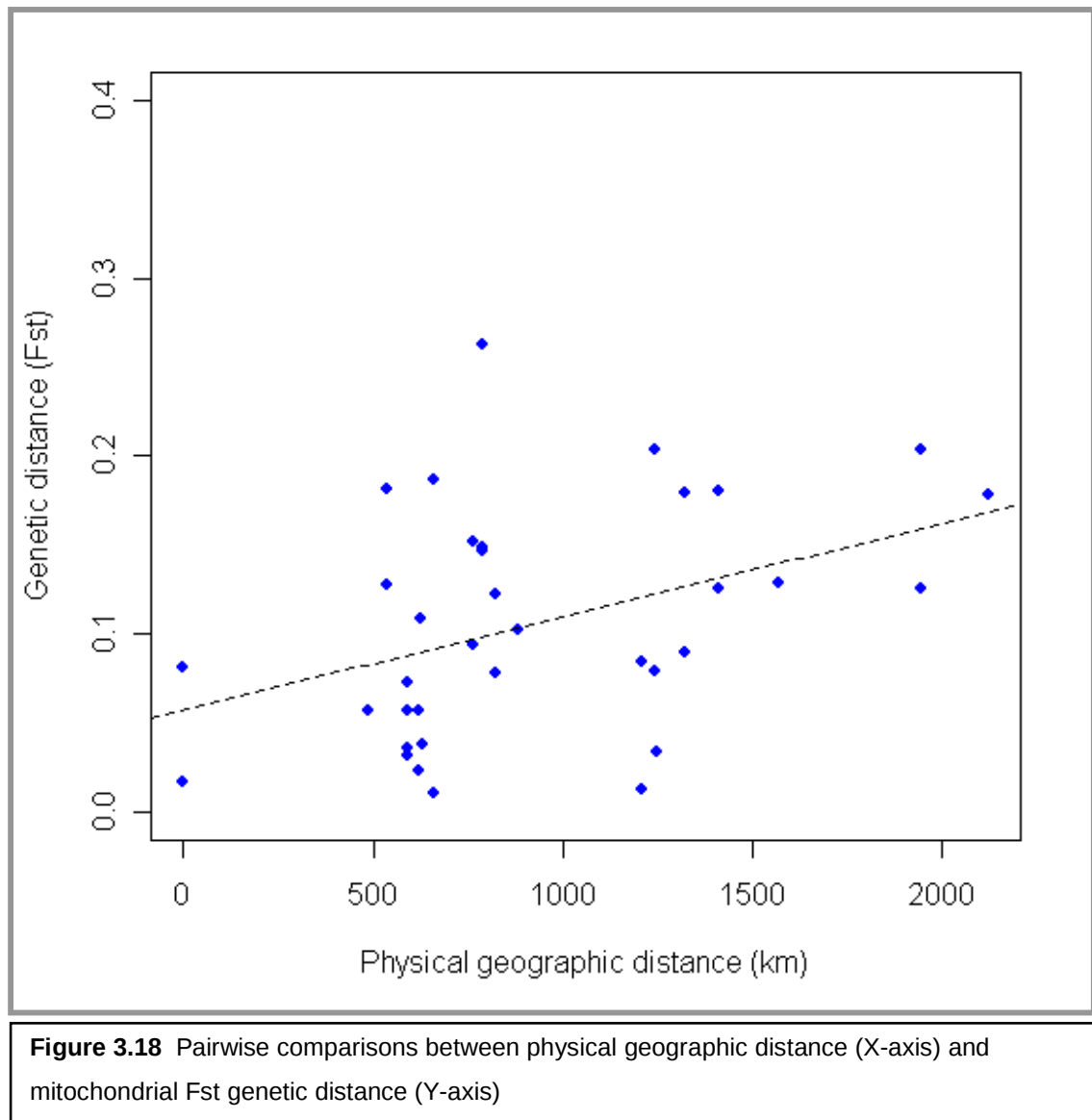
Similarly in the tree (Figure 3.17), the subsequent split was between non-African groups (AFR, EUR, IND) and Bantu-speakers (SOT, ZUX, DRC, HER and CAC, KWE). Thereafter the northern San groups (GUG, JOH, XUN) split from the southern Khoe-San-Coloured groups (CNC, KHO, NAM, COL, KAR).

The northern San groups, JOH and XUN grouped together with GUG on its own. On the other branch, containing the southern Khoe-San-Coloured groups, CNC and KHO grouped together, while COL and NAM grouped together, with KAR forming its own group.

Furthermore the DRC and HER grouped together while the southeastern BS (ZUX and SOT) grouped with the KWE and CAC. This showed the considerable admixture from southeastern Bantu-speakers in CAC and interestingly showed that the KWE group rather grouped with southeastern BS than with the southwestern BS or a central African BS group.

From the graphical visualizations of F_{st} values it appeared that there might be an association between genetic distance and physical distance within the Khoe-San-Coloured groups. Groups that were genetically closely related was also not far removed from one another when looking at physical distance. To see if the genetic distances and physical geographic distances (km) show a correlation within the Khoe-San-Coloured groups, the genetic distance matrix (Table 3.5) was correlated with a physical distance matrix (Appendix C) to see how the two relate to one another (For this part of the analysis the KWE group was not included as part of the Khoe-San-Coloured groups because previous analysis showed their maternal lineages to be genetically more similar to BS than Khoe-San-Coloured). In Figure 3.18 pairwise comparisons between physical distance (X-axis) and genetic distance (Y-axis) was plotted on a graph. A linear regression was done to determine the line with the best fit through the points. The best fit to the graph was a straight line with a slope of 0.00005263 ($p = 0.0149$). Furthermore a Mantel test was

conducted to further affirm the relationship between the two distance matrices. It confirmed the relationship between genetic and physical geographic distance ($p = 0.027$) with 16% of the genetic distance being explained by physical geographic distance and a correlation coefficient (r) of 0.402750 between the two matrices.



Some of the Coloured and Khoe groups (especially CAC, COL, and NAM) had a considerable amount of admixture from Bantu-speakers and or non-African groups (see Figure 3.4). This would be reflected in their Fst values, which would be reproduced in

graphical representations of F_{st} values such as the PCA plot (Figure 3.16) where CAC, COL and NAM grouped closer to these groups. This recent admixture of BS-groups into San groups might obscure historical relationships between Khoe-San groups that existed before the BS expansions and non-African influx. To investigate this historical relationship between putative Khoe-San groups before BS and European admixture all non-L0d/k groups were removed from the sample. The different Khoe-San and Coloured groups were again compared with one another to see if the relationship between them changes (Figure 3.19 and 3.20). It is acknowledged that this might not be a true representation of what the haplogroup structure might have looked like before the influx of BS. It is, however, generally seen that groups with less BS admixture have higher L0d percentages and that in BS-populations the amount of L0d admixture increases to the southern parts of Africa where there was contact with Khoe-San people. For this part of the analysis, it was therefore assumed that L0d and L0k might have been predominant mitochondrial haplogroups of the Khoe-San before the Bantu-expansions and it was investigated how these L0d/k carriers might have been related to one another.

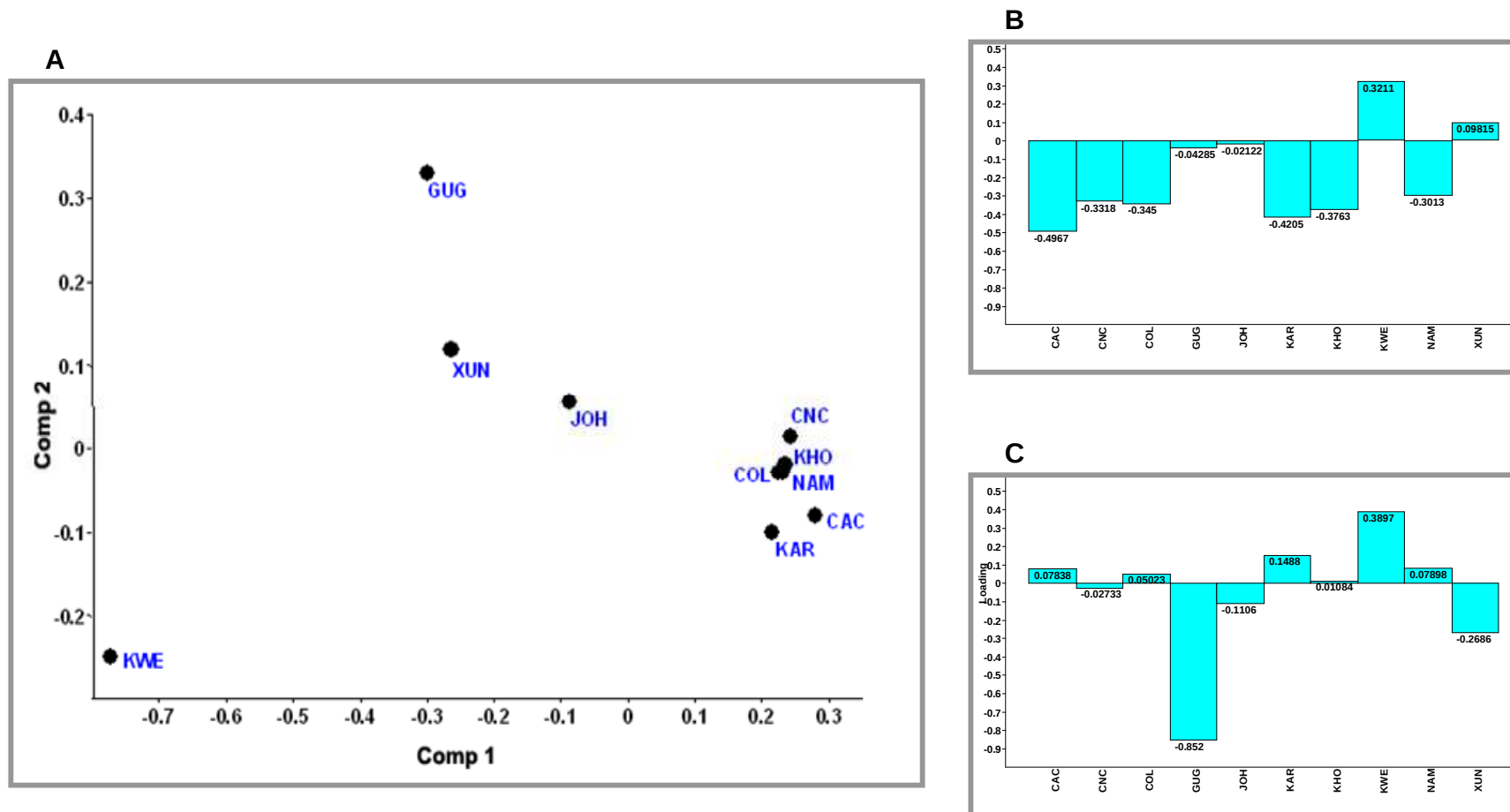
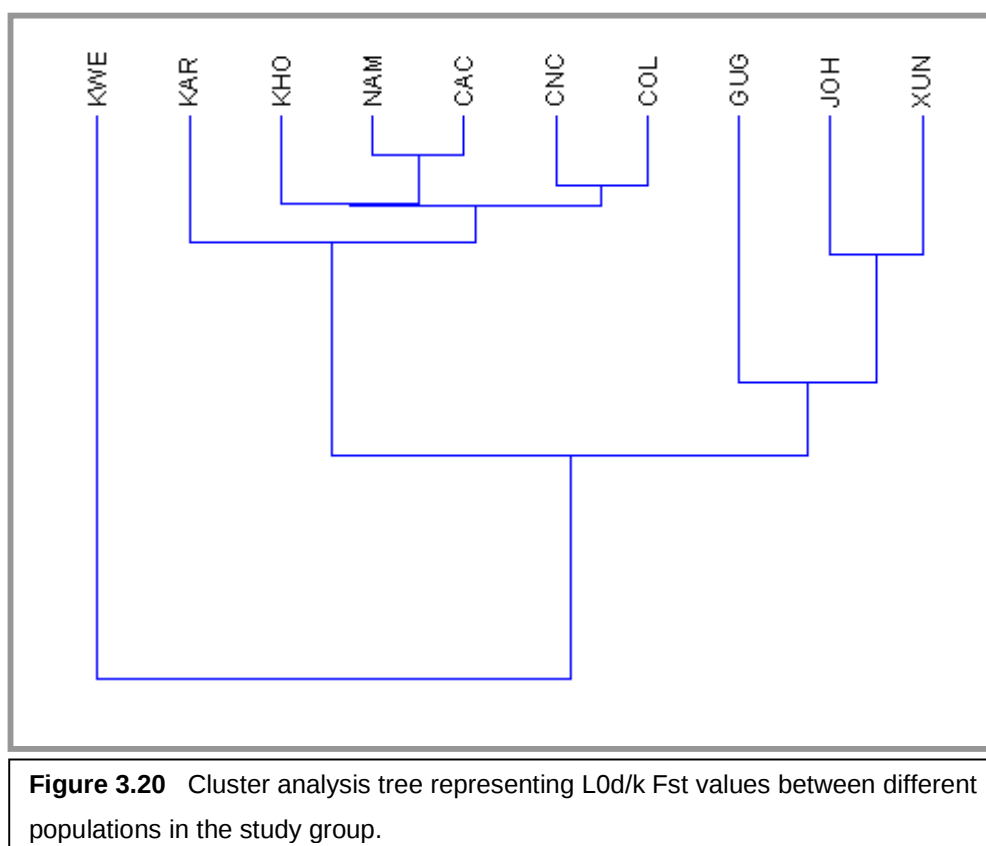


Figure 3.19 **A** – Principal component analysis of L0d/k Fst values between different populations in the study group. Component 1 = 74.5% of the variation, Component 2 = 13.9% of the variation. **B** – Loadings for Component 1, **C** – loadings for Component 2.



In Figure 3.19 and 3.20 the divide between the northern San groups (GUG, JOH, XUN) and southern Coloured and Khoe-San groups (KAR, KHO, NAM, CAC, CNC, COL) can be seen clearly. With the removal of the BS and non-African admixture, CAC moved to the southern San-Khoe-Coloured cluster. The KWE group, however, was still a clear outlier. Groups within the southern San-Khoe-Coloured group clustered very closely together while the northern San (GUG, JOH, XUN) was more distant from one another.

The best fit to the graph of genetic vs. physical distance in this case was still a straight line with a slope ($p = 0.00587$) (Graph included in Appendix F). The slope of the line when non-L0dk sequences were removed was 0.00009086 ($p = 0.00587$) and was steeper than the gradient when non-L0dk sequences were included (0.00005263). The Mantel test also showed a significant relationship between the two distance matrices ($p = 0.034$). In this case 20% of the genetic distance was explained by physical distance and $r = 0.449854$. This correlation was stronger than in the case with non-L0dk sequences included (determination of genetic distance by physical distance = 16% and $r = 0.402750$).

To see what the influence of the presence of L0k is on the separation of northern groups from the southern groups (L0k were found in the northern groups but not in the southern groups) the analysis was repeated with only L0d sequences and is shown in Figure 3.21 and 3.22).

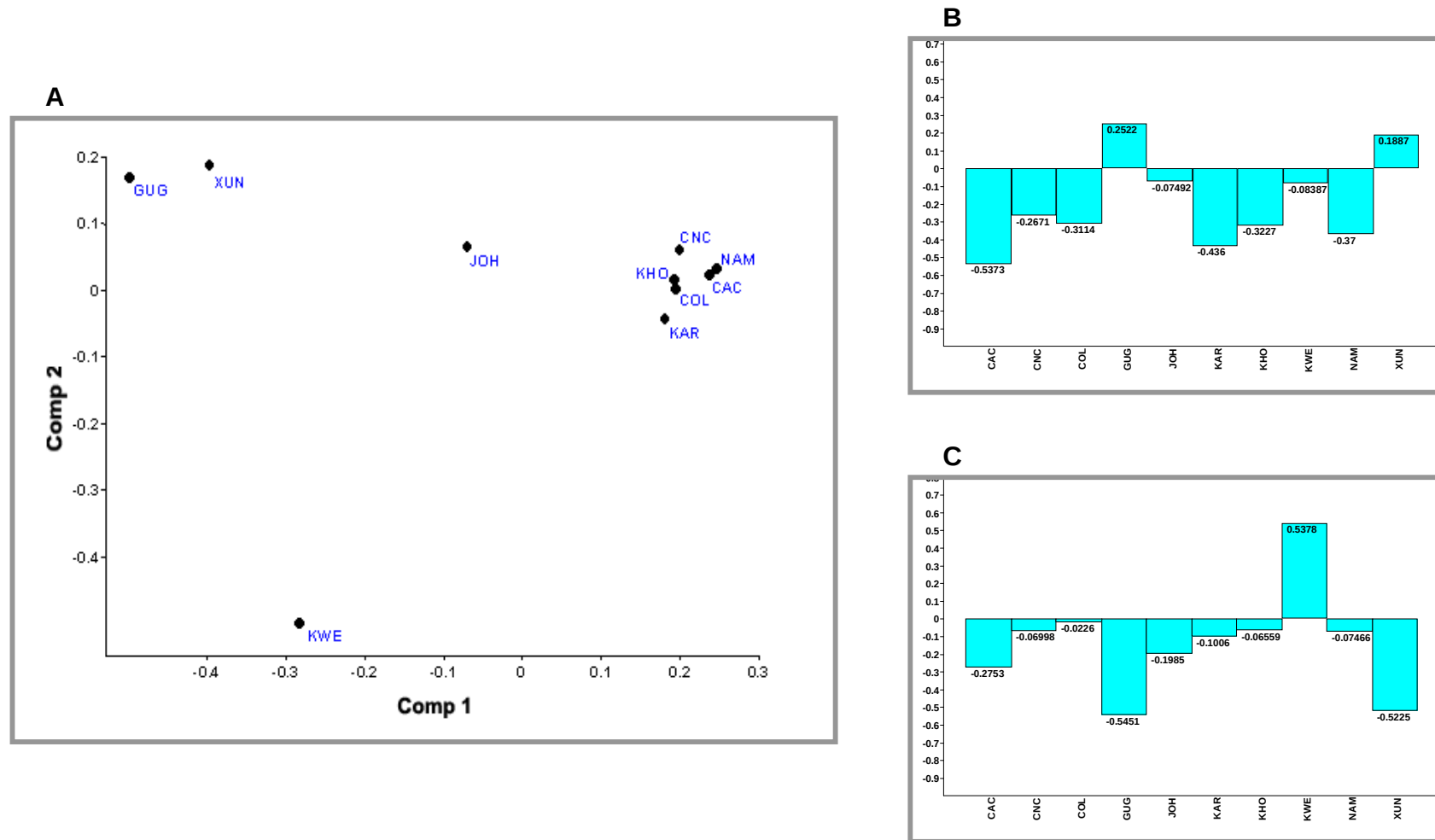
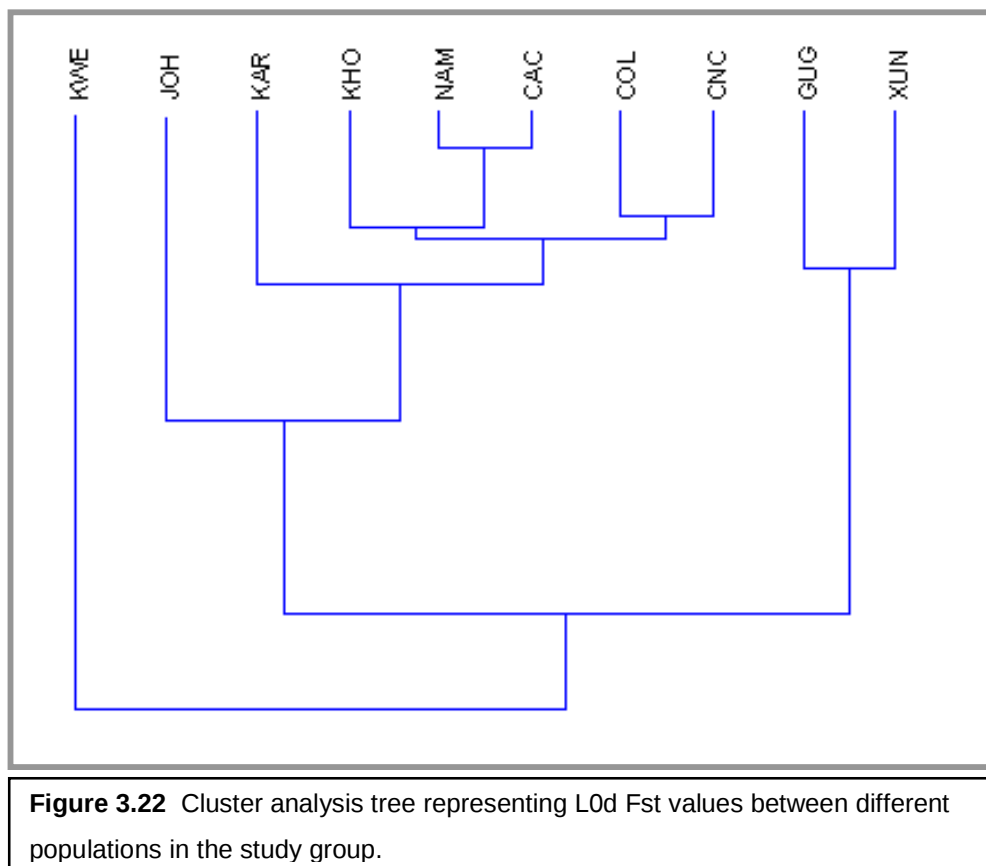


Figure 3.21 **A** – Principal component analysis of L0d Fst values between different populations in the study group. Component 1 = 63.2% of the variation, Component 2 = 27.1% of the variation. **B** – Loadings for Component 1, **C** – loadings for Component 2.



With the removal of L0k, the KWE group was still an outlier but has to be viewed with caution because it is represented by only two L0d sequences. The JOH group now rather grouped with the southern Coloured and Khoe-San groups than with the GUG and XUN but was still removed from them. This could also be seen by observing the L0d haplogroup frequencies (Figure 3.7 and 3.8). JOH had far lower L0d1c frequencies than GUG and XUN but higher L0d1b frequencies. The other southern Coloured and Khoe-San groups still grouped tightly together in a cluster.

In the graph of genetic vs. physical distance for L0d sequences only, the best-fitted line was still a straight line with a slope ($p = 0.00742$) (Appendix F). The slope of the line when only L0d sequences were included is 0.00009472 ($p = 0.00742$) and was steeper when non-L0dk sequences were included (0.00005263) and marginally steeper than the slope for the L0dk sequences (0.00009086). The Mantel test again showed a significant relationship between the two distance matrices ($p = 0.033$). When only L0d sequences were used 19%

of the genetic distance was explained by physical distance and $r = 0.439081$. This correlation was higher than in the case with non-L0dk sequences included and slightly lower than the case where non-L0dk sequences were excluded (For non-L0dk included: “determination of genetic distance by physical distance” = 16% and $r=0.402750$. For non-L0dk excluded “determination of genetic distance by physical distance” = 20% and $r=0.449854$).

To test the apportionment of variation in the different groups, AMOVA analysis was done to see how much variation is contained between defined groups, between the different populations in the study and also within the populations. Table 3.6 give the results of the AMOVA analysis with various different groupings of the highest-level group.

Table 3.6 Results from mitochondrial AMOVA analysis using different groupings on the first level

Grouping	Grouping of first level [Groups]	Between groups	Between populations within groups	Between individuals within populations
A	[afr, eur, ind]			
	[col, cnc, kar, kho, nam, joh, xun, gug, kwe, cac]			
	[drc, her, sot, zux]	21.95	7.7	70.35
B	[afr, eur, ind]			
	[col, cnc, kar, kho, nam, joh, xun, gug]			
	[drc, her, sot, zux]	24.46	6.69	68.85
C	[afr, eur, ind]			
	[col, cnc, kar, kho, nam, joh, xun, gug, drc, her, sot, zux]	29.28	10.03	60.69
D	[col, cnc, kar, kho, nam, joh, xun, gug]			
	[drc, her, sot, zux, afr, eur, ind]	21.82	9.09	69.09
E	[afr, eur, ind]			
	[col, cnc, kar, kho, nam]			
	[drc, her, sot, zux]			
	[gug, joh, xun]	21.16	4.74	74.1
F	[col, cnc, kar, kho, nam, joh, xun, gug]			
	[drc, her, sot, zux]	13.25	5.35	81.39
G	[col, cnc, kar, kho, nam]			
	[joh, xun, gug]	8.3	4.94	86.76
H	[col, cnc, kar]			
	[nam, joh, xun, gug, kho]	1.93	8.42	89.65
I	[col, cnc, kar]			
	[nam]			
	[joh, xun, gug, kho]	1.15	8.8	90.05

For the first analysis, three groups were assigned, namely, BS, non-African and Khoe-San-Coloured (grouping A in Table 3.6). For this analysis 22% of variation was between these three groups, 8% between populations within the groups and 70% between individuals within the populations. CAC and KWE were assigned to the Khoe-San-Coloured (KSC) group but as was seen previously their placement in the Khoe-San-Coloured group is ambiguous because the maternal line in KWE was closer related to BS-groups and CAC had a very admixed origin. When they were left out more of the variation could be assigned to variation between the three main groups (24.5%) (grouping B in Table 3.6). These two groups were left out the subsequent AMOVA analyses.

When two groups were assigned, African and non-African it resulted in an among group variation of 29% the highest of all the AMOVA analyses (grouping C in Table 3.6). Next, the KSC groups were split into northern San and southern KSC (grouping E in Table 3.6). The variation between the resultant four groups (non-African, BS, northern San and southern KSC) was 21% and not much different from BS, KSC and non-African. When non-African groups were left out of the analysis the variation between BS and KSC was 13% between groups, 5 % between populations and 81% within populations (grouping F in Table 3.6). The BS was then left out and the KSC group was split into the northern San group and southern KSC groups (as previous analyses suggested such a split) (grouping G in Table 3.6). In this scenario more of the variation was still contained on group level (8%) vs. population level (5%).

When populations were split into the groups that self identify as Coloured and those that self identify as Khoe-San only 2% variation was explained on the group level (grouping H in Table 3.6). In this case, inter population variance explained more of the variance (8%). Furthermore, if the Khoe was split from the San group to give the populations that self identify as San, Khoe and Coloured even less variation (1%) was explained on group level (grouping I in Table 3.6).

To investigate the expansion dynamics of the different groups involved in the study mismatch distributions of the nucleotide variation in the sequences involved were

constructed. Figure 3.23 show the mismatch distributions of the 10 Khoe-San / Coloured populations and two comparative groups ZUX (Bantu-speaking) and EUR (European).

Other Bantu-speaking and non-African groups (not shown) showed similar unimodal distributions to the two groups used as comparative data. The AFR, however, was slightly multimodal due to admixture from African groups. In the CAC, COL, NAM, JOH, ZUX and EUR the model of demographic expansion was not rejected. The model was rejected in KHO, KAR and KWE ($p=0.01$) as well as CNC ($p=0.05$) and GUG (95% CI of θ_0 and θ_1 overlaps). The model was not rejected in XUN but will be rejected in at a 10% level ($p=0.1$). Table 3.7 show the statistics of the mismatch distributions and the tests for demographic expansion for all of the groups.

The raggedness index can also be used as an indicator of rapid demographic growth (Table 3.7) but does not correlate in all cases with the confidence test for demographic expansions (see the low raggedness values for KAR and CNC – yet expansion model was rejected).

θ_1 divided by θ_0 give an indication of the magnitude of the expansion while τ gives an indication of the time of the expansion. Of the seven groups that does not reject the expansion hypothesis (NAM, JOH, SOT, CAC COL, DRC, ZUX, EUR), the BS groups appeared to have had the greatest expansions (θ_0 is 16 to 19 times smaller than θ_1) while the magnitude in the Khoe-San / Coloured groups was lower (θ_0 is 7 to 10 times smaller than θ_1). τ was converted to time BP (T) when the expansion happened as outlined in section 2.2.2.3 using the mutation rate of 2.5×10^{-6} per nucleotide per generation (Ward *et al.*, 1991) (Table 3.7). The expansion times of the CAC and NAM populations seemed to have happened ~40 000 years BP. The BS populations had earlier expansion times of around 90 000 years BP for ZUX and SOT and 60 000 years BP for DRC. The representative haplogroups of the JOH and COL also showed expansion times of 75 000 – 90 000 years BP. The European expansion time dated to around 40 000 years BP.

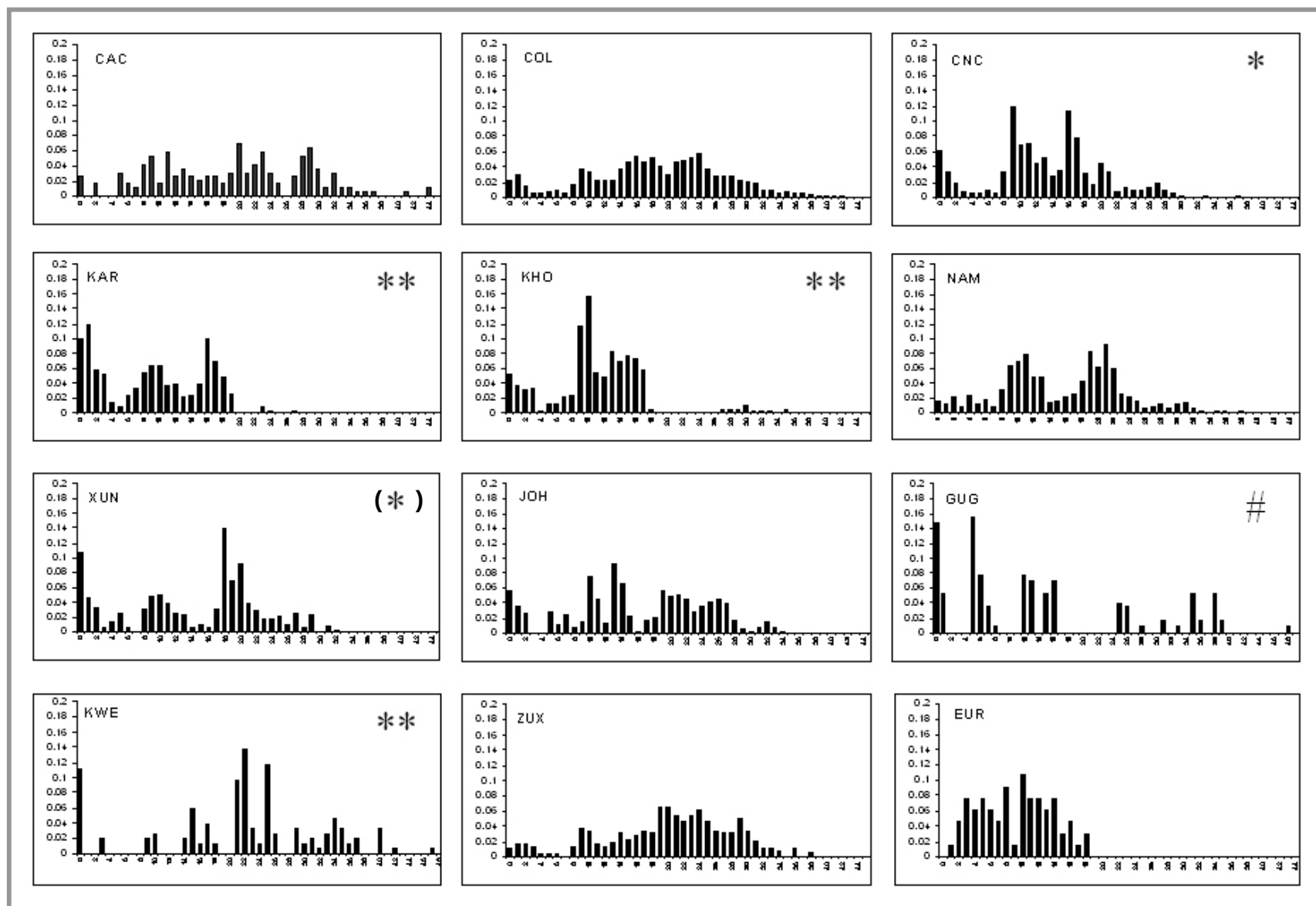


Figure 3.23 Mismatch distributions of populations in the study group. # expansion hypothesis rejected - 95% CI overlap. * Expansion hypothesis rejected on 99% confidence level. ** Expansion hypothesis rejected on 95% confidence level. (*) Expansion hypothesis will be rejected on a 90% confidence level

Table 3.7 Mismatch distribution statistics (Groups)

Group	Raggedness index	Tau	T \$	Theta0	Theta0 qt 5%-95%	Theta1	Theta1 qt 5% - 95%	Model (SSD) p-value
KWE	0.068	20.330		5.462	0.000 - 10.642	66.709	35.889 - 1839.209	0.010**
HER	0.106	19.438		0.004	0.000 - 4.676	20.012	14.233 - 179.074	0.020*
IND	0.006	10.488		0.000	0.000 - 2.496	59.219	32.827 - 99999.000	0.040*
KHO	0.027	12.559		0.007	0.000 - 2.960	37.056	23.384 - 267.368	0.040*
CNC	0.025	10.656		5.054	0.000 - 13.212	49.175	28.076 - 2319.175	0.050*
XUN	0.030	20.305		0.004	0.000 - 4.470	26.216	19.036 - 198.403	0.060(*)
NAM	0.011	9.168	40 783	10.158	0.000 - 27.229	99.863	40.376 - 99999.000	0.180
JOH	0.021	20.098	89 404	2.902	0.000 - 7.597	29.502	17.932 - 109.268	0.230
SOT	0.020	20.805	92 549	5.161	0.000 - 16.367	83.696	50.707 - 99999.000	0.270
GUG	0.076	37.102		0.000	0.000 - 12.057	13.971	7.276 - 99999.000	0.290 [#]
CAC	0.017	9.070	40 347	16.432	0.000 - 37.615	116.484	46.565 - 99999.000	0.440
AFR	0.028	6.938		8.903	0.000 - 24.725	51.567	24.391 - 99999.000	0.490 [#]
COL	0.003	16.949	75 396	5.908	0.000 - 5.950	58.853	36.606 - 162.915	0.630
DRC	0.027	14.445	64 257	4.243	0.000 - 13.044	66.885	41.692 - 99999.000	0.640
ZUX	0.005	20.234	90 009	3.614	0.000 - 11.990	70.107	45.382 - 436.357	0.730
EUR	0.025	9.262	41 201	2.949	0.000 - 4.967	25.376	18.518 - 99999.000	0.810
Mean	0.030	15.167		4.165	0.000 - 11.765	5933.746	5903.052 - 53264.869	0.290
SD	0.028	8.264		4.491	0.000 - 10.013	24240.077	24214.474 - 51097.852	0.280

T – Time before present that expansion took place (calculation explained in section 2.2.2.3)

SSD - Sum of Squared deviation

Expansion hypothesis rejected - 95% CI overlap.

* Expansion hypothesis rejected on 99% confidence level,

** Expansion hypothesis rejected on 95% confidence level.

(*) Expansion hypothesis will be rejected on a 90% confidence level

Summary statistics such as Tajima's D (Tajima, 1989), Fu's Fs (Fu, 1997) and the R2 statistic (Ramos-Onsins and Rozas, 2002), have been reported to have greater sensitivity in detecting population expansions than mismatch distributions (Pilkington *et al.*, 2008). Table 3.8 shows summary statistics in the form of diversity estimates and neutrality tests for the different groups.

Table 3.8 Diversity statistics and neutrality tests for populations in the study group

Group	N seq	N Ht	Hd	π	θ_s	W - θ_s	N_e ($\theta_s/2\mu$)	Tajima's D	Tajima's D p-value	Fs	Fs p-value	R2	R2 p-value
KAR	30	13	0.864	0.00611	0.00756	8.330	1482	-0.82038	0.230	-0.392	0.462	0.0961	0.237
COL	77	49	0.971	0.01245	0.01855	19.941	3548	-1.13707	0.120	-18.138	<0.001***	0.0683	0.152
CAC	20	15	0.963	0.01349	0.01667	18.040	3210	-0.70963	0.248	-1.186	0.310	0.1038	0.199
KHO	57	23	0.947	0.00750	0.01176	12.794	2277	-1.25000	0.089	-2.663	0.209	0.0714	0.137
CNC	40	23	0.927	0.00914	0.01435	15.516	2761	-1.26573	0.087	-3.828	0.116	0.0709	0.078 (*)
XEG	3	3	-	-	-	-	-	-	-	-	-	-	-
DUM	1	1	-	-	-	-	-	-	-	-	-	-	-
NAM	28	23	0.984	0.01035	0.01543	16.703	2972	-1.24601	0.093	-7.691	0.007**	0.0753	0.048*
GUG	22	8	0.853	0.00843	0.01154	13.027	2318	-0.98584	0.174	3.736	0.930	0.1025	0.224
NAR	2	2	-	-	-	-	-	-	-	-	-	-	-
JOH	42	17	0.943	0.00980	0.01087	12.087	2151	-0.32461	0.432	0.638	0.635	0.1056	0.499
XUN	49	23	0.894	0.00934	0.01345	14.578	2594	-1.06493	0.138	-2.155	0.262	0.0750	0.149
KWE	18	9	0.889	0.01468	0.01692	18.316	3259	-0.53662	0.317	3.946	0.940	0.1199	0.355
DRC	14	12	0.978	0.01158	0.01382	15.094	2686	-0.61070	0.290	-1.812	0.171	0.1134	0.169
HER	15	6	0.648	0.00809	0.01094	11.994	2134	-1.12248	0.128	4.011	0.948	0.1031	0.094 (*)
SOT	22	18	0.970	0.01540	0.01887	20.582	3662	-0.77810	0.230	-2.593	0.134	0.1020	0.232
SWZ	5	5	-	-	-	-	-	-	-	-	-	-	-
ZUX	36	31	0.989	0.01377	0.01918	20.983	3734	-1.07663	0.133	-11.823	0.002**	0.0791	0.115
AFR	21	18	0.981	0.01050	0.01486	16.121	2869	-1.08335	0.143	-5.193	0.023*	0.0872	0.066 (*)
EUR	11	11	1.000	0.00757	0.00990	10.925	1944	-1.02790	0.155	-4.759	0.010*	0.0918	0.011*
IND	25	25	1.000	0.00965	0.01924	20.657	3676	-1.96876	0.01*	-17.291	<0.001***	0.0492	<0.001***
All	538	236	0.984	0.01239	0.02967	31.052	5525						

* $p < 0.05$ ** $p < 0.01$ *** $p < 0.001$ (*) $p < 0.01$

Haplotype diversities were high in most of the groups. Groups with lower diversities in ascending order were HER, GUG, KAR, KWE and XUN. Nucleotide diversities were generally higher in African populations than in non-African populations. With the exception of HER, BS populations had higher nucleotide diversities than Coloured-Khoe-San populations. Populations with lower nucleotide diversities included KAR, KHO, HER, GUG, CNC, XUN and JOH. Theta was estimated using segregating sites (θ_s per site) and Watsons- θ_s (W- θ_s per sequence). The female effective population size (N_e) was estimated from W- θ_s as explained in section 2.2.3. Smaller effective population sizes were present in KAR, JOH, HER, GUG, KHO. Under neutral expectations with random mating, constant population sizes and no selection π and θ should be equal (Jobling *et al.*, 2004c). A neutrality test was done to detect deviations from the assumptions of neutrality and constant population size. Significantly positive Tajima's D values indicate balancing selection and / or population subdivision while significantly negative values indicates population growth and /or positive selection (Jobling *et al.*, 2004b). All of the Tajima's D values were negative, only the IND value, however, reached significance indicating population growth. KHO, CNC and NAM would have been significant on a 10% level.

The F_s and the R_2 statistic have been reported to detect population expansions very successfully (Ramos-Onsins and Rozas, 2002; Pilkington *et al.*, 2008). F_s is based on the probability of drawing a number of haplotypes that is greater or equal to the observed number of samples drawn from a population of constant size. R_2 is based on the difference between the average number of nucleotide differences and the number of singleton mutations. The R_2 statistic is especially powerful when sample sizes are small (~10) and F_s have a greater ability to detect population expansions when sample sizes are large (~50) (Ramos-Onsins and Rozas, 2002).

F_s was negative for all samples except HER, GUG and KWE. All three non-African populations (AFR, EUR, IND) had significantly negative values. The only other populations that had significantly negative values were ZUX, NAM, COL. In addition to their significantly negative F_s values IND, EUR and NAM had significantly positive R_2 values as well. AFR almost reached significance and would have been significant on a 10% level. HER would also have been significant on a 10% level. The HER had a very insignificant F_s value, however, F_s does not perform good when sample sizes are small (Ramos-Onsins and Rozas, 2002). CNC is another group that would have reached significance for R_2 on a 10% level; also the CNC F_s P-value was the lowest P-value that did not reach significance. Although ZUX and COL had very significant F_s values it did not reach significant R_2 values. This might be because R_2 performs better at smaller sample sizes (~10) and both ZUX (36) and COL (77) had better sample sizes.

3.4.1 Summary: Genetic Affinities between the Khoe-San and Coloured groups as inferred from mtDNA analysis

Using the maternally transmitted mtDNA marker the affinities between the different Khoe-San and Coloured groups were examined using haplogroup frequencies and the phylogenetic relationships of the haplogroups to one another. Since the associated haplogroups of the Bantu-speakers and non-African groups are very distinct from those commonly found in the Khoe-San, admixture from these populations will have a great influence in the resultant tree that represents relationships between the different population groups. This can be seen in Figure 3.17 where the Bantu-speaking admixture in the Cape

Coloured and Khwe groups cause them to group with the Bantu-speaking-group. The effect of the admixture can also be seen in the PCA plot (Figure 3.16) where the first two components represents Africa vs. non-African variation and Bantu-speaking vs. non-Bantu-speaking variation. While an inclusive comparison is representative of the current genetic composition of the groups studied, it should not be used to make inferences about Khoe-San history and Khoe-San group relations before the Bantu-expansions and the influx of non-African colonists.

In attempt to infer group relations between the Khoe-San and Coloured groups that existed before the pastoralist influx, haplogroups previously associated with Bantu-speakers (and the Bantu-expansions) as well as non-African haplogroups were removed from the Khoe-San and Coloured groups. Remaining Khoe-San associated haplogroups (L0d and L0k) were then used to infer relationships between the Khoe-San groups that might have existed in the past. The resultant PCA plot (Figure 3.19) showed that the southern groups were closely associated with each other while the northern groups are separate from the southern groups. The Khwe was different from all the groups. The association between physical and genetic distance remained and was even stronger than the situation where Bantu-speaking and non-African haplogroups were included.

Due to the possibility that L0k was not part of the original Khoe-San haplogroup pool but rather introduced by other hunter-gatherer groups that were displaced because of the Bantu-expansions (such as previously discussed for the Khwe), L0k was also removed and only L0d based group relations was tested (Figure 3.21 and 3.22). The Khwe group was still the most different from the other groups because it only contained the L0dx haplogroup, which is absent in all other groups except the !Xun where it occurs at low frequency. Interestingly, the Ju\’hoansi group moved closer to the southern groups due to the higher frequency of L0d1b and the lower frequency of L0d1c. The correlation between physical and genetic distance remained for the L0d based group comparison.

Cluster and PCA observations were reaffirmed from AMOVA (Table 3.6). The largest part of the variation in the groups was explained between non-African vs. African variation and Bantu-speaking vs. Khoe-San variation. AMOVA, however, also supported that a

considerable amount of variation could be summarised as variation between northern San groups and southern Khoe-San groups. The current group classifications of Khoe-San vs. Coloured and Khoe vs. San vs. Coloured explained very little of the variation. This classification is thus not supported by a maternal-line genetic analogue.

The following deductions can therefore be made about the maternal line genetic composition of the groups included in the study. Firstly, various levels of admixture from both Bantu-speakers and non-African groups are present in the different groups. Secondly, the Khwe group is different from the other Khoe-San and Coloured groups and might represent remnants of another extinct hunter-gatherer group that were displaced by the Bantu-expansions and became associated with the San. This may include the introduction of the Khoe linguistic group by them, gene-flow of the L0k haplogroup from them to the !Xun and the L0dx haplogroup from the !Xun to the Khwe. Thirdly, there is a distance based genetic relationship between the remaining groups. Fourth, the haplogroup distribution between the southern and northern groups is different and the Nama cluster with and are similar to the southern groups. Fifth, other factors such as the adoption of pastoralism also might have had an important role and the rapid spread of haplogroups associated with populations that accepted pastoralism would have influenced original haplogroup distributions.

4. Y-CHROMOSOME STUDIES

Forty-six bi-allelic polymorphisms and 12 Y-STR markers on the Y-chromosome were examined in 353 unrelated males to investigate the paternal affinities between three Coloured (KAR, COL, CNC), one Khoe (NAM) and five San (KHO, GUG, JOH, XUN, KWE) groups. Their affinities to neighbouring Bantu-speaking (DRC, HER, SOT, ZUX) and non-African populations (IND, AFR and EUR) were examined.

4.1 Haplogroup allocation and geographic distribution

The Y-chromosomes in the sample examined were assigned to 29 haplogroups using the bi-allelic polymorphisms according to the nomenclature of Karafet *et al.*, (Karafet *et al.*, 2008) (Figure 4.1). Eleven major haplogroups were represented at differing frequencies in the groups studied (Figure 4.2). The haplogroup with the highest frequency in the total study group was haplogroup E-M2 (E1b1a*) (20%) followed by haplogroup A-M51 (A3b1) at 15% (Figure 4.1).

By observing the haplogroup distributions in the different population groups in the form of bar-charts (Figure 4.2) a differential distribution in the different Khoe-San and Coloured population groups was noted. To further investigate the geographic distributions of the haplogroups and their sub-haplogroups, contour plots were constructed and are shown in Figure 4.3.

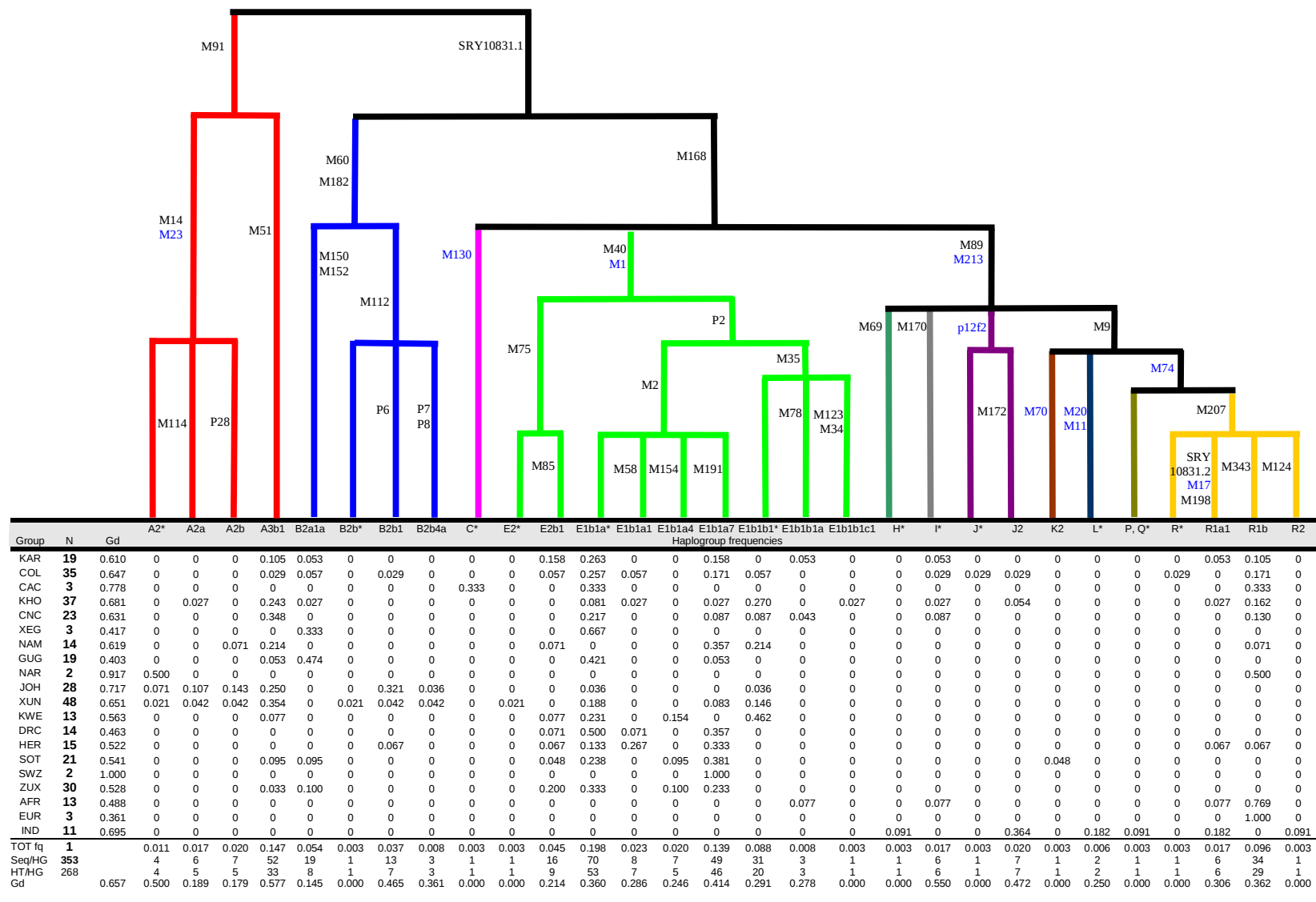


Figure 4.1 Y-chromosome haplogroup tree with nomenclature according to Karafet *et al.*, (2008), listing haplogroup frequencies in the different populations in the study group. The Gene diversities (Gd) in the different haplogroups are also indicated. (Tot fq – Total frequency, Seq/HG – Sequences per haplogroup, HT/HG – Haplotypes per haplogroup, Gd – Gene diversity)

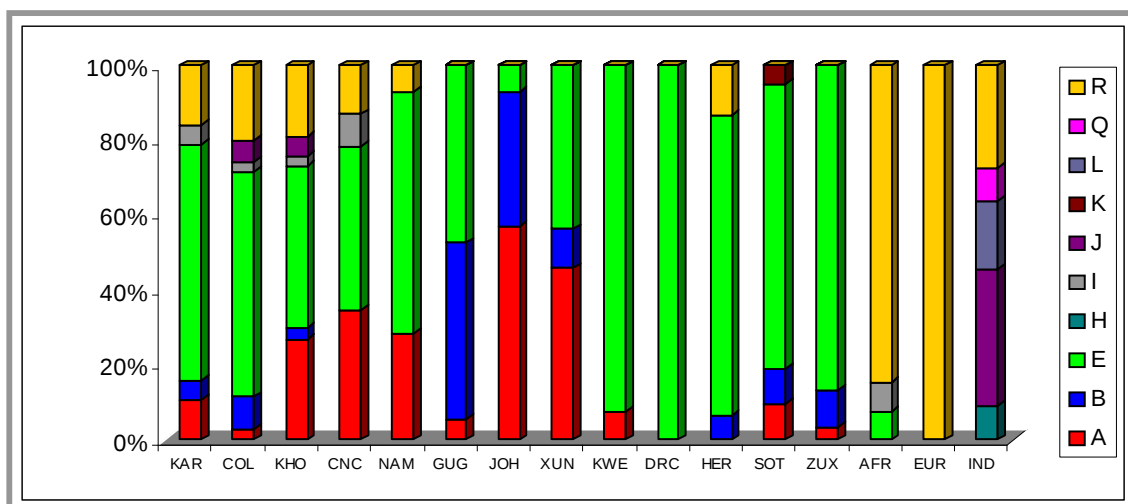


Figure 4.2 Graphical illustration of percentage Y-chromosome haplogroup assignment in the populations used in comparative population analysis

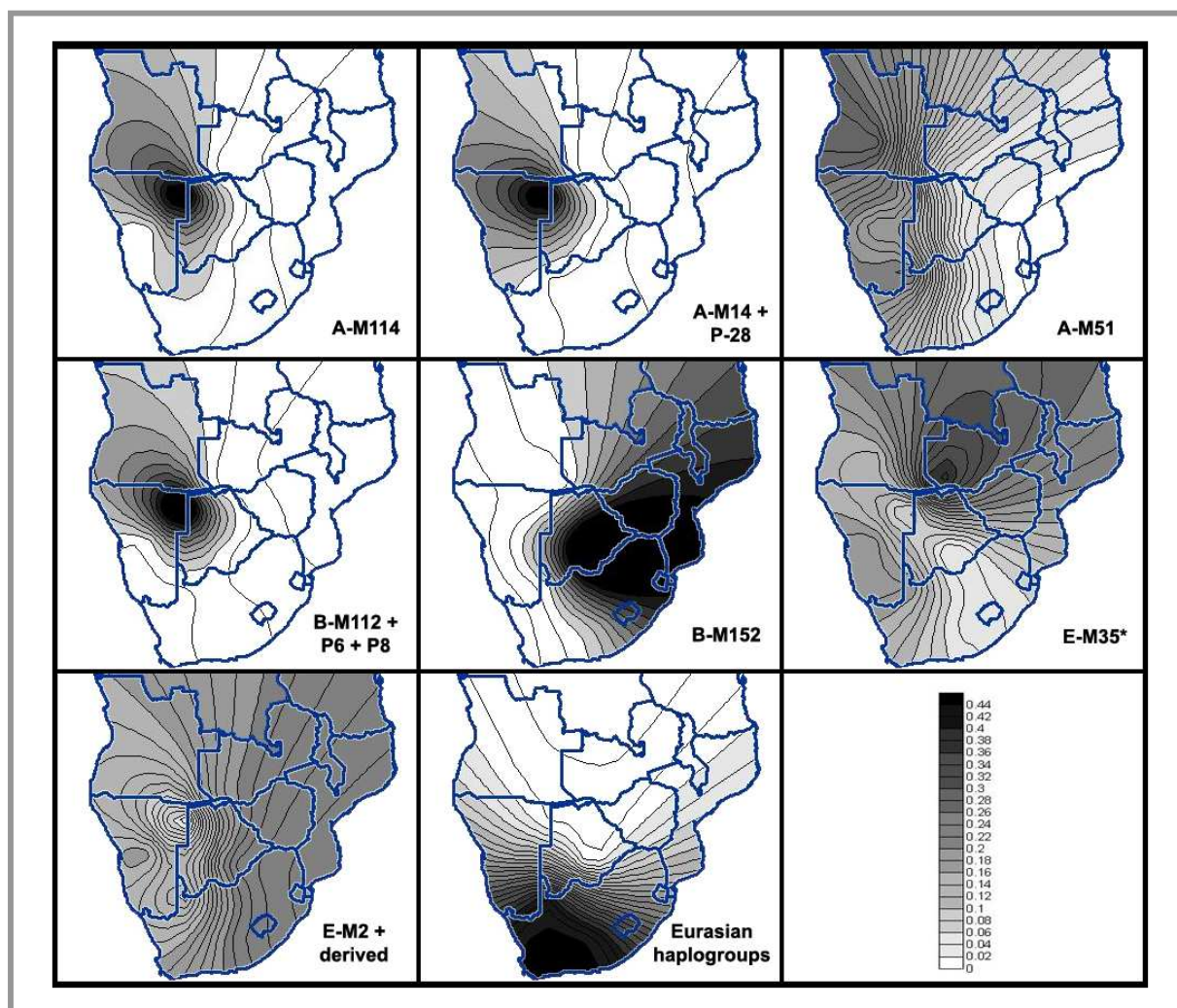


Figure 4.3 Contour plots indicating the frequency distributions of Y-chromosome haplogroups in the Khoe-San and Coloured populations

The contour plots clearly showed that haplogroups had specific geographic patterns (Figure 4.3). A-M114, A-M14-P28 and B-M112-P6-P8 were limited to the northern groups with its highest frequency in the Tsumkwe area (JOH group). A-M51 seemed to have a wide geographic distribution; though its absence in the east of southern Africa should be confirmed by more extensive sampling, as no groups were sampled in this area. The pattern formed by B-M152 was caused by its high frequency in the GUG group and its absence in the XUN, JOH, KWE and NAM. E-M35 also displayed a northern distribution with its highest frequency in the KWE followed by the KHO, but was completely absent in the JOH. It had lower frequencies in the southeast. E-M2 and its derived groups were widely distributed but displayed higher frequencies in the east than in the west. Eurasian haplogroups were distributed in the southern parts of the region indicating the direction from which Eurasian settlers came.

4.2 Haplogroup diversity

The 29 haplogroups were further resolved into 268 Y-STR haplotypes. The full list of haplotypes is included in Appendix G. The genetic diversity of the whole study group was 0.657 (Figure 4.1). The Khoe-San and Coloured groups (with the exception of GUG) generally had higher genetic diversities than BS groups and non-African groups (Figure 4.1). AFR and EUR had low diversities compared to African groups. The GUG group had a lower genetic diversity than all of the groups except EUR. No instance of full haplotype homoplasy between haplogroups was observed in the study samples. The haplotype with the highest frequency is Ht051 (B-M152), which occurred a total of 12 times, of which eight instances were in the GUG group and one each in the XEG, ZUX and KAR. The high occurrence in the GUG might indicate a founder effect / bottleneck in this particular group sampled. This haplotype, however, was found in a KAR as well as a XEG individual indicating a wide geographic spread. It was also found in a BS individual indicating gene flow. Subsequently, eight different haplotypes occurred four times each, representing the second highest haplotype frequency.

4.3 African haplogroup analyses and discussion

To further investigate the structure within the haplogroups, phylogenetic networks, $\delta\mu^2$ distance based NJ trees and MDS plots were constructed from STR profiles (see section 2.3.4 for full description of methodologies employed to construct networks and trees). The ages of different haplogroups were also determined from the networks. Dating of Y-chromosome haplogroups was done with $\mu = 6.9 \times 10^{-4}$ per locus per generation with a generation time of 25 years (Zhivotovsky *et al.*, 2004). The following sections will present these results and results will be discussed in conjunction with results regarding haplogroup geographical distribution and haplotype diversities presented in the previous sections (section 4.1 and 4.2). Furthermore results will be related to the published literature for each haplogroup.

Haplogroup A – Internal structure

The network and NJ tree for haplogroup A clearly separated A-M51 from the other three haplogroups (Figure 4.4 and 4.5). The other three haplogroups did not show the expected topology (see Figure 4.1) but individuals belonging to the same sub-haplogroup did cluster together. An exception is a Naro A-M14 individual that clustered within the A-M114 group.

In the tree A-M51 seemed to have KAR and COL individuals at its root that connected with the other A subgroups (Figure 4.4). The remaining A-M51 samples were split into two branches. The one containing mostly KHO, XUN and JOH individuals and the other one containing mostly KHO and NAM individuals with one GUG and one COL individual.

In the network A-M51 also connected with the other subclades through COL and KAR individuals, however, the two branches within A-M51 seen in the NJ tree were not as apparent (Figure 4.5). Instead there was a reticulation of KHO individuals that formed the base of three branches. The clustering of population groups seen in the NJ tree (KHO + NAM; JOH + XUN) was still present.

The position of the KWE individual was ambiguous, presenting as an early branch at the base of A-M51 in the tree, while being located at the tip of a deeper branch in the network.

To further investigate these relationships the distance matrix used for the NJ tree was also visualised by doing a MDS plot (Figure 4.6). The MDS plot clearly separated the three sub-haplogroups (A-M51, A-M114 and A-M14/A-P28). Furthermore it showed that the COL/KAR A-M51 haplotypes grouped closer to the other A-subgroups than the rest of the A-M51 haplotypes. The COL/KAR, KHO/CNC and NAM haplotypes were located centrally in the A-M51 cluster closest to the other subgroups while the XUN, JOH and KWE A-M51 haplotypes were located on the periphery of the A-M51 cluster, further away from the other subgroups. The uniqueness of the KWE haplotype was also evident since it was removed by a large distance on both the X and Y-axis from the other A-M51 sequences.

The TMRCA of haplogroup A was 65 857 (+/-10 007) years BP. The oldest sub-haplogroup was A-M51, TMRCA = 54 000 (+/- 9 593) years BP, which was five times older than A-M14+P28 (TMRCA = 11 473 +/- 3 960 years BP) and A-M114 (TMRCA = 8 052 +/- 3 102 years BP).

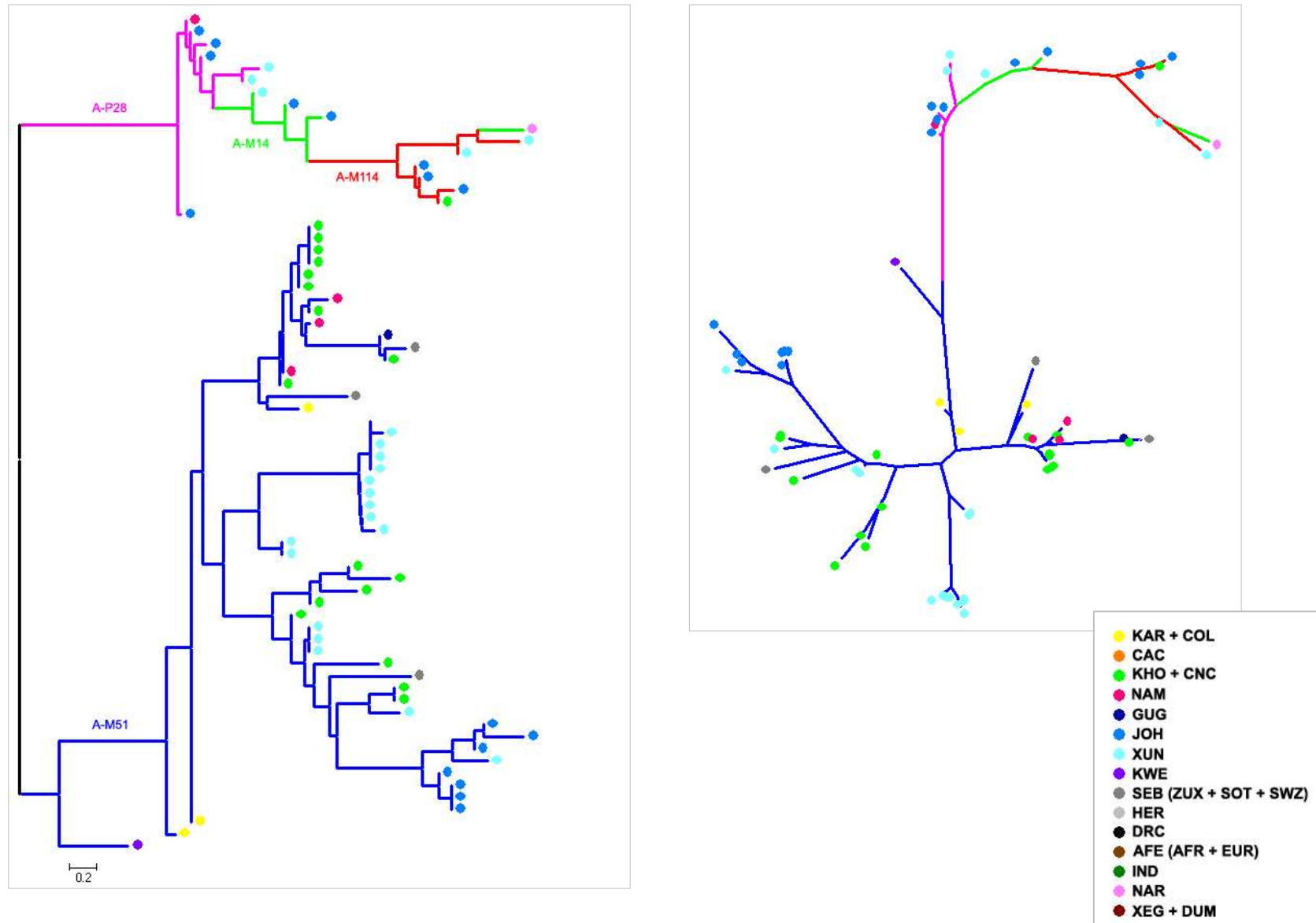


Figure 4.4 Neighbour Joining tree representing the substructure of Haplogroup A. Individuals are colour coded according to the key. The rectangular phylogram and unrooted radial phylogram have sub-haplogroup branches of corresponding colours.

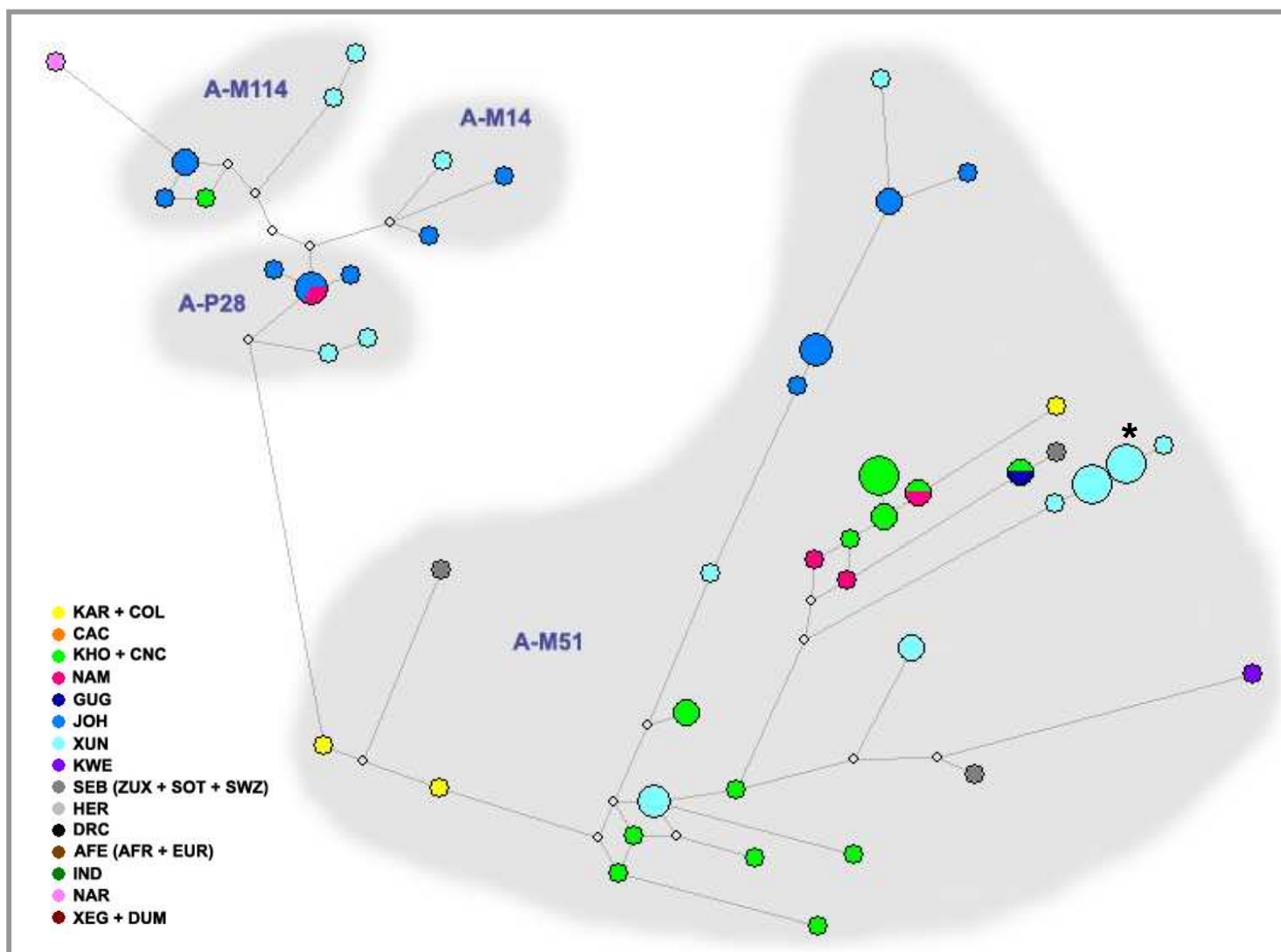


Figure 4.5 Median joining network representing Haplogroup A substructure in the different populations of the study group.

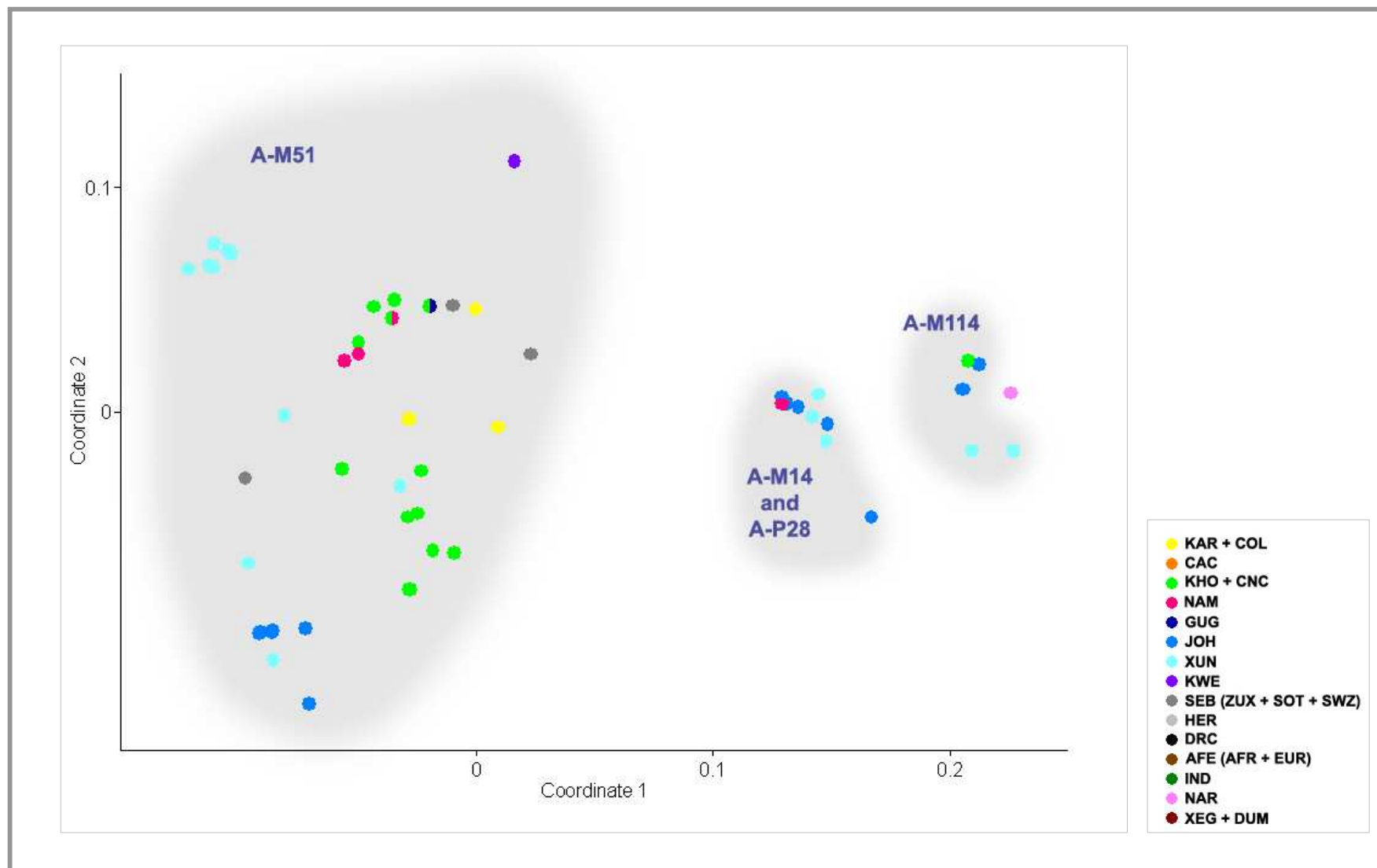


Figure 4.6 MDS plot visualizing the $\delta\mu^2$ distance matrix for haplogroup A (also used for the Neighbour Joining tree). Individuals are colour coded according to the key.

Haplogroup A - Discussion

The oldest haplogroup, haplogroup A, was found at its highest frequencies in the northern San groups (Figure 4.1 and 4.2). Although Coloured and Khoe groups in general had lower frequencies of this haplogroup than San groups, frequencies were mostly higher than found in the Bantu-speakers. Also, groups with higher frequencies of haplogroup A were groups that are known to have had lower admixture with Bantu-speakers and non-African populations. This could be seen in the lower Haplogroup A frequencies in the Karretjie and Colesberg-Coloured groups, who have had substantial paternal line admixture from Bantu-speakers and non-African populations. As discussed in the introduction it is known from historical records that male San individuals, living in the Karoo area, were severely persecuted in the 1700-1800s, while females were relocated to farms. Subsequently it was mostly the local Xhosa males and white farm owners that contributed to the male line genetic variation of the resultant Coloured population. In fact their haplogroup frequencies for haplogroup A were less than that found in some of the Bantu-speakers. Higher frequencies were seen in the ‡Khomani and Coloured-Northern Cape groups. The area that these groups occupied were not as severely targeted by colonists as was the case for the area that the ancestors of the Karretjie and Coloured groups occupied.

Another group with a low frequency of haplogroup A was the /Gui + //Gana + Kgalagari group. While the maternal lines of this group was mostly the Khoe-San associated haplogroup L0d, most of the paternal lines seemed to be from Bantu-speaking individuals. This group was a mixed group with /Gui and //Gana (San) and Kgalagari (Bantu-speakers) ancestry. From results it seemed that the Kgalagari contributed mostly to the male line while the female lines came from the San groups.

The sub-haplogroups within haplogroup A also had different representation patterns among the groups (Figure 4.1 and 4.3). A-M51 was wide-spread with representation in northern as well as southern Khoe-San and Coloured groups. A-M14 and its derived haplogroups, however, seemed to be concentrated in the northern San groups. Except for single males in the ‡Khomani and Nama group, the A-M14 derived haplogroups was completely absent in population groups representative of southern Khoe-San (Figure 4.1 and 4.3). The two haplotypes found in the ‡Khomani and Nama group had type-sharing or were close

neighbours to the Ju\’hoansi haplotypes, indicating that the gene flow to the southern groups came from the Ju\’hoansi group (Figure 4.5).

Published studies have only concentrated on the northern groups, !Xun, Khwe and Ju\’hoansi (Table 1.4). The one study that included Nama did not differentiate haplogroup frequencies from the !Xun, Khwe and Ju\’hoansi. Similar to results from the present study, the published studies found high frequencies of haplogroup A in the !Xun, and Ju\’hoansi but not in the Khwe. A-M51 was the most common A-haplogroup in both the !Xun and mixed Ju\’hoansi and !Xun group (Table 1.4).

From the network and MDS analyses it could be seen that A-M14 and its derived groups cluster closely together and have a lower diversity when compared to A-M51 (Figure 4.5 and 4.6). The A-M51 group had high haplotype diversities and internal structure could also be observed within this haplogroup. Haplotypes from the southern groups had a central position and were closer related to the A-M14 derived haplotypes. Surprisingly the northern !Xun and Ju\’hoansi groups occupied the peripheral regions of the A-M51 cluster and were more distantly related to the A-M14 derived haplotypes. This is unexpected since A-M14 was largely restricted to northern groups and one would think the A-M51 lineages in the northern groups would be more related to the A-M14 lineages. This thus indicates that the A-M14 lineage did not split from the A-M51 lineage while present in the northern San populations. A better explanation would be that there was a very ancient split in haplogroup A before populations had their current distribution and designations. Haplogroup A-M51 then developed its north-south distribution and A-M14 was subsequently incorporated into the northern groups from somewhere else. Another possibility can be that the A-M51 – A-M14 split represent an ancient north-south split in haplogroup A in which A-M51 was a southern haplogroup and A-M14 a northern haplogroup. A subsequent cline then developed within haplogroup A-M51 when some of the A-M51 haplotypes migrated northwards to join A-M14 and differentiated from the southern A-M51 haplotypes.

Haplogroup B – Internal structure

Within haplogroup B, both the network and tree clearly separated B-M152 and B-M112 (and derived subgroups) into two separate groups (Figure 4.7 and 4.8). While B-M152 was predominantly made up of GUG individuals, they all had the same haplotype. Most unique B-M152 haplotypes occurred in BS.

Further, the network and trees showed P8 and P6 as subgroups of B-M112. B-P8 formed a monophyletic clade with three haplotypes (two XUN and one JOH). Interestingly on both the network and tree, two B-P6 haplotypes (one COL and one HER), did not form a monophyletic clade with the rest of the B-P6 haplotypes (all XUN and JOH). On the network these two haplotypes possibly grouped closer to the B-P8 haplotypes than the other B-P6 haplotypes, while on the trees the COL haplotype grouped as a early branch of B-P6 while the HER haplotype grouped at the base of B-P8. To further investigate the relationships between the haplogroup B subgroups a MDS plot of the distance matrix used to construct the NJ tree was created (Figure 4.9).

The MDS plot clearly separated B-M112, B-M152, B-P8 and two clusters in B-P6 from one another. B-M152 and B-P8 each form tight haplotype clusters. B-P6 is, however, divided into two clusters, the one containing the XUN and JOH haplotypes and the other the HER and COL haplotypes.

Haplogroup B had a TMRCA of 54 432 (+10 005) years BP, therefore ~10 000 years younger than Haplogroup A. B-M112 is the oldest haplogroup B subgroup (TMRCA = 36 763 +/- 7 325 years BP), while B-M152 converged at 12 236 +/- 5 512 years BP. Of the two B-M112 subgroups B-P6 (TMRCA = 29 961 +/- 7 436 years BP) was older than B-P8 (TMRCA = 9 058 +/- 3 629 years BP).

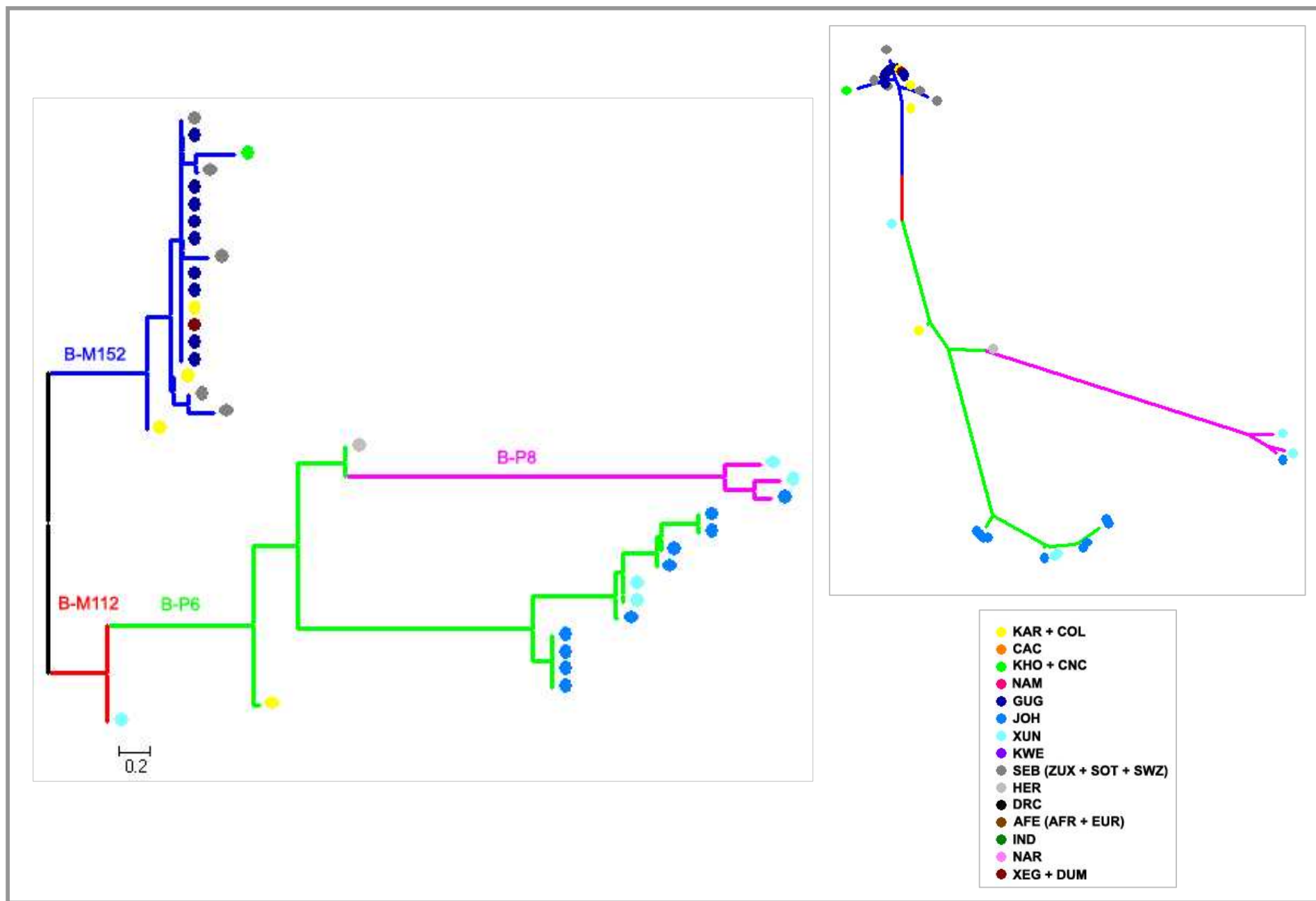


Figure 4.7 Neighbour Joining tree representing the substructure of Haplogroup B. Individuals are colour coded according to the key. The rectangular phylogram and unrooted radial phylogram have sub-haplogroup branches of corresponding colours.

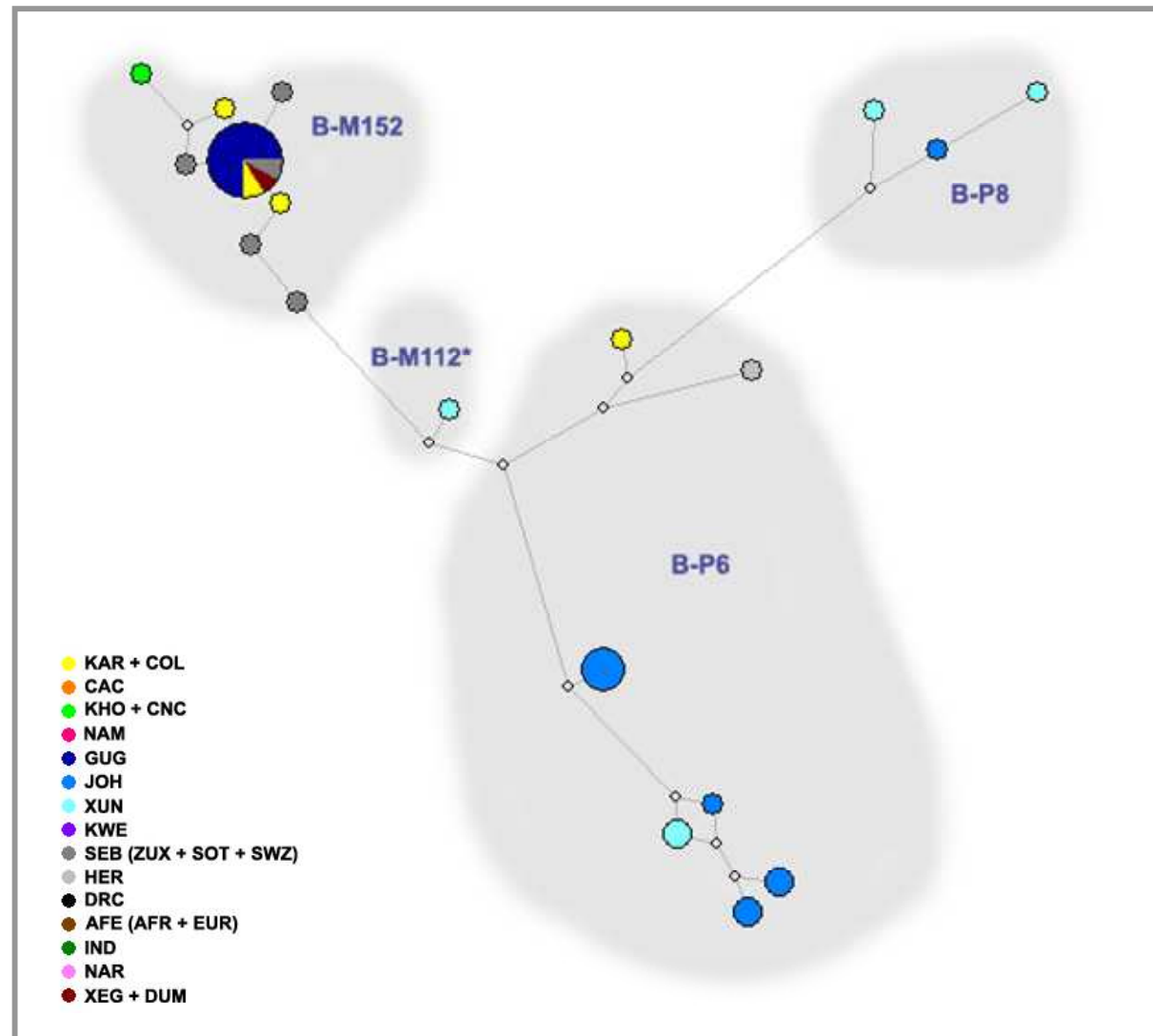


Figure 4.8 Median joining network representing Haplogroup B substructure in the different populations of the study group.

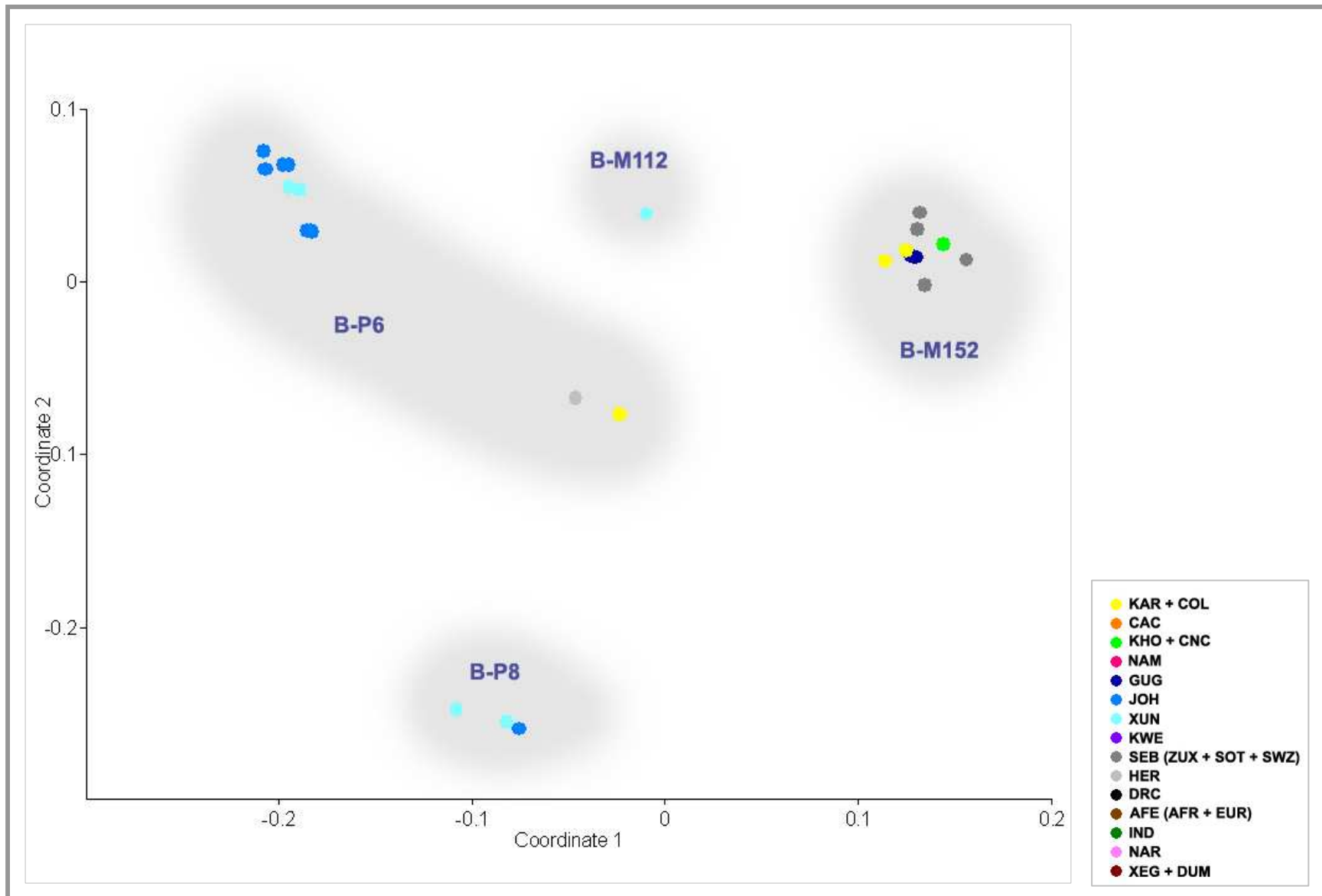


Figure 4.9 MDS plot visualizing the $\delta\mu^2$ distance matrix for haplogroup B (also used for the Neighbour Joining tree). Individuals are colour coded according to the key.

Haplogroup B - Discussion

The Khoe-San associated B-haplogroup, B-M112 and its derived groups B-P6 and B-P7, was found at its highest frequencies in the northern San groups (Figure 4.1 and 4.2). Except for one individual in the Colesberg-Coloured group, this haplogroup was absent from southern Khoe-San and Coloured groups. It was one of the highest frequency haplogroups in the Ju'hoansi with similar frequencies to the A-M14 and A-M51 groups. In the !Xun it had lower frequencies than the A haplogroups. B-M112 haplotypes have also previously been identified in the Pygmy populations of central Africa and the Hadza from east Africa (Underhill *et al.*, 2000; Cruciani *et al.*, 2002; Semino *et al.*, 2002; YCC, 2002).

The fact that B-M112 is present in the Pygmy and Hadza populations and northern San groups but not in the southern San groups was indicative that B-M112 is not a general Khoe-San haplogroup. Rather it could have been a haplogroup present in another ancient hunter-gatherer population north of the northern San groups. This group would have lived south of the Pygmy and Hadza groups before the Bantu-expansion. An ideal candidate would be the Ba-Twa Pygmy group. Further support for this theory comes from the study of rock art. The San rock art tradition ends at the Angola-Namibian border and another rock-art zone begins. This zone is termed the Schematic Art Zone and is significantly different from San rock art (Smith, 2006). The zone stretches into the DRC and Tanzania and is bordered in the north by the Saharan art zone. The Schematic Art Zone has been linked to the Ba-Twa Pygmy group (Smith, 2006). The Ba-Twa groups could have had connections and gene-flow from the San, Hadza and other Pygmy groups. B-M112 is common in the Pygmies and Hadza, therefore B-M112 might also have been a Ba-Twa associated haplogroup in the past that got incorporated into the northern San groups because of gene-flow between the two groups that did not reach the southern San groups. Physical anthropological studies on fossils from this region found no significant overlap with one specific modern group (Morris and Ribot, 2006). Instead, the morphological features for fossils from this region were unique to LSA people from the region itself. The sample size of Pymy representatives were however very small (Morris and Ribot, 2006).

Most of the B-M112 haplotypes in the study group belonged to either B-P6 or B-P8, only one haplotype belong to the ancestral B-M112*. B-P6 and B-P8 have been reported

previously in the Ju\’hoansi (YCC, 2002) and were restricted to Khoe-San populations. In the MDS plot (Figure 4.9) all the B sub-haplogroup haplotypes clustered closely together except B-P6 which formed two clusters, a !Xun / Ju\’hoansi cluster and a cluster of two haplotypes one belonging to a Herero individual and one belonging to a Colesberg-Coloured individual (Figure 4.9). From network and tree analysis it appeared that these two haplotypes were located at the base of the B-P6 and B-P8 split (Figure 4.8). This situation was similar to the Haplogroup A situation where southern Colesberg-Coloured/Karretjie haplotypes were at the base of a haplogroup split in the northern !Xun and Ju\’hoansi.

The B-M152 haplogroup is a Bantu-speaking associated haplogroup and was found at frequencies around 10% in the Bantu-speakers of the present study group (Figure 4.1). It was, however, also found at very high frequencies in the /Gui + //Gana + Kgalagari. The B-M152 representation in the /Gui + //Gana + Kgalagari was, however, from one haplotype and probably indicates a strong recent founder effect in this group (Figure 4.8).

While haplogroup A was found to be the oldest haplogroup with a TMRCA of 65 857 years, haplogroup B dated to 54 432 years. The TMRCA of these two oldest haplogroups falls within the range of the TMRCA for the Y-chromosome (46 000–91 000 years) as determined by microsatellites (Wilson and Balding, 1998; Pritchard *et al.*, 1999).

Haplogroup E – Internal structure

To investigate finer structure within the haplogroup E subclades three different networks and NJ trees were constructed for the following haplogroups and their derivative subclades: Haplogroup E-M75, Haplogroup E-M2 and Haplogroup E-M35.

Haplogroup E-M75

E-M75* was only represented by one XUN sequence, the rest of this haplogroup was contained in the sub-group E-M85. E-M85 was mostly represented by BS and Coloured groups (Figure 4.1). The E-M75 tree and network are represented by Figure 4.10 and 4.11.

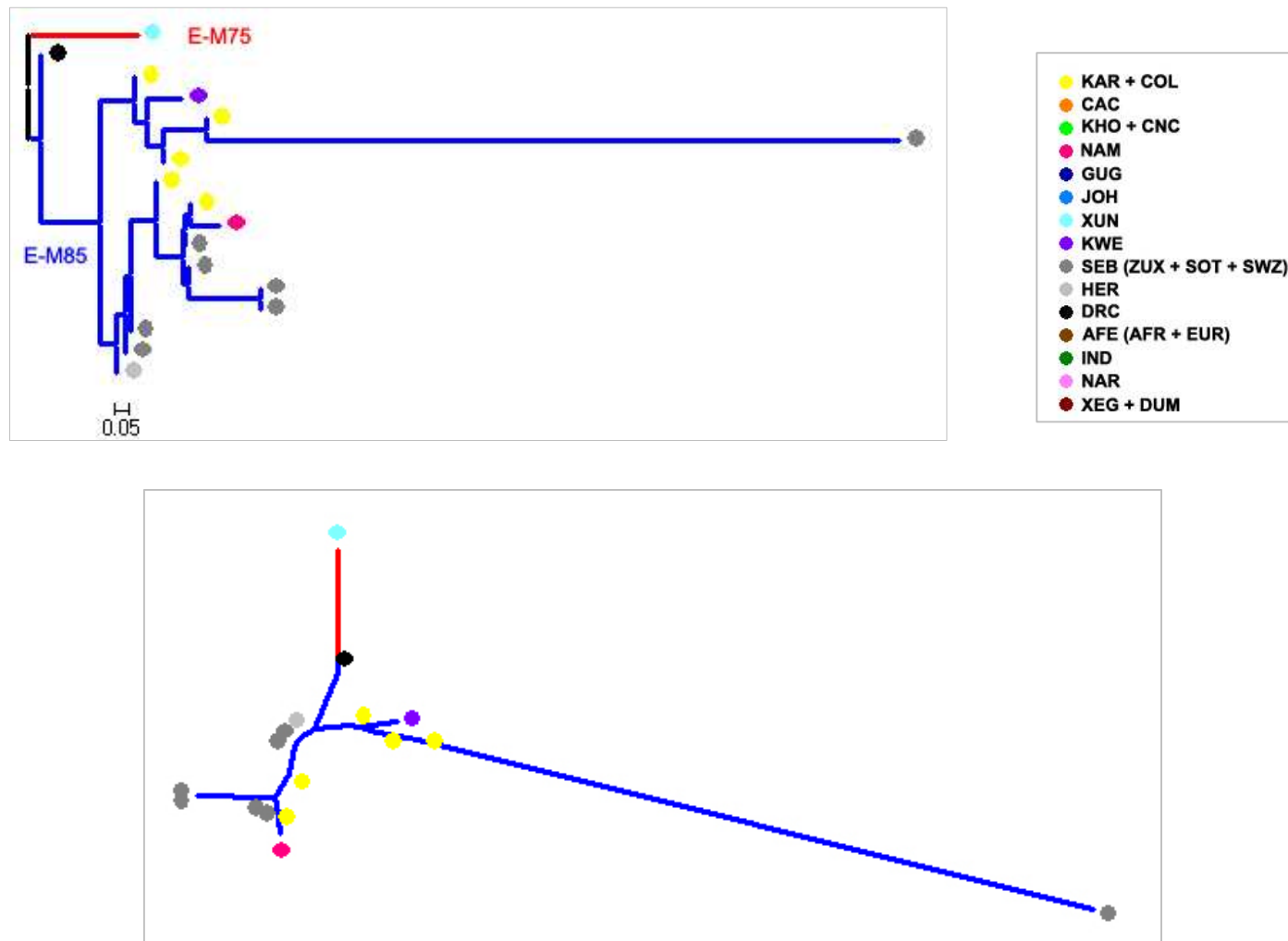


Figure 4.10 Neighbour Joining tree representing the substructure of Haplogroup E-M75. Individuals are colour coded according to the key. The rectangular phylogram and unrooted radial phylogram have sub-haplogroup branches of corresponding colours.

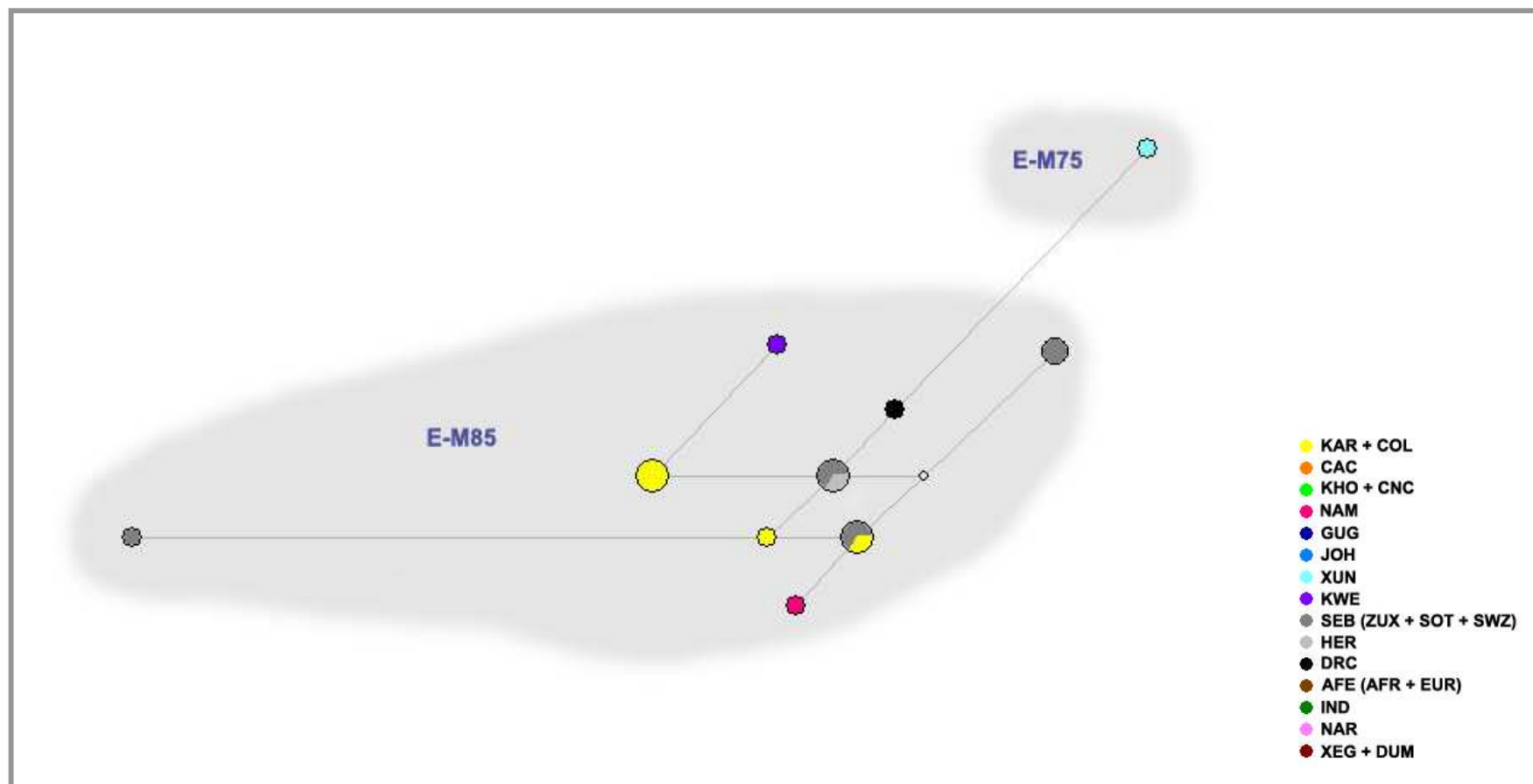


Figure 4.11 Median joining network representing Haplogroup E-M75 substructure in the different populations of the study group.

Haplogroup E-M2

An NJ tree and network shows the internal structure within haplogroup E-M2 (Figure 4.12 and Figure 4.13). The represented sub-haplogroups of E-M2 (E-M58, E-M154 and E-M191) did not form monophyletic clades in the trees or the network. The network also exhibited a high degree of reticulation. Overall there was a high level of partial homoplasy between the haplotypes of these sub-haplogroups.

In the tree E-M191 formed one large monophyletic clade but also had two smaller clades within E-M2, while in the network there was one large clade and one smaller clade within E-M2. E-M58 grouped together in the tree and network with the exception of one sample. E-M154 formed one clade and two separate samples on the tree and two clades in the network.

Members of the different population groups were spread throughout the network, however, members of the same population groups were often neighbours in the network.

Haplogroup E-M2 had a TMRCA of 39 701 +/- 8 263 years BP.

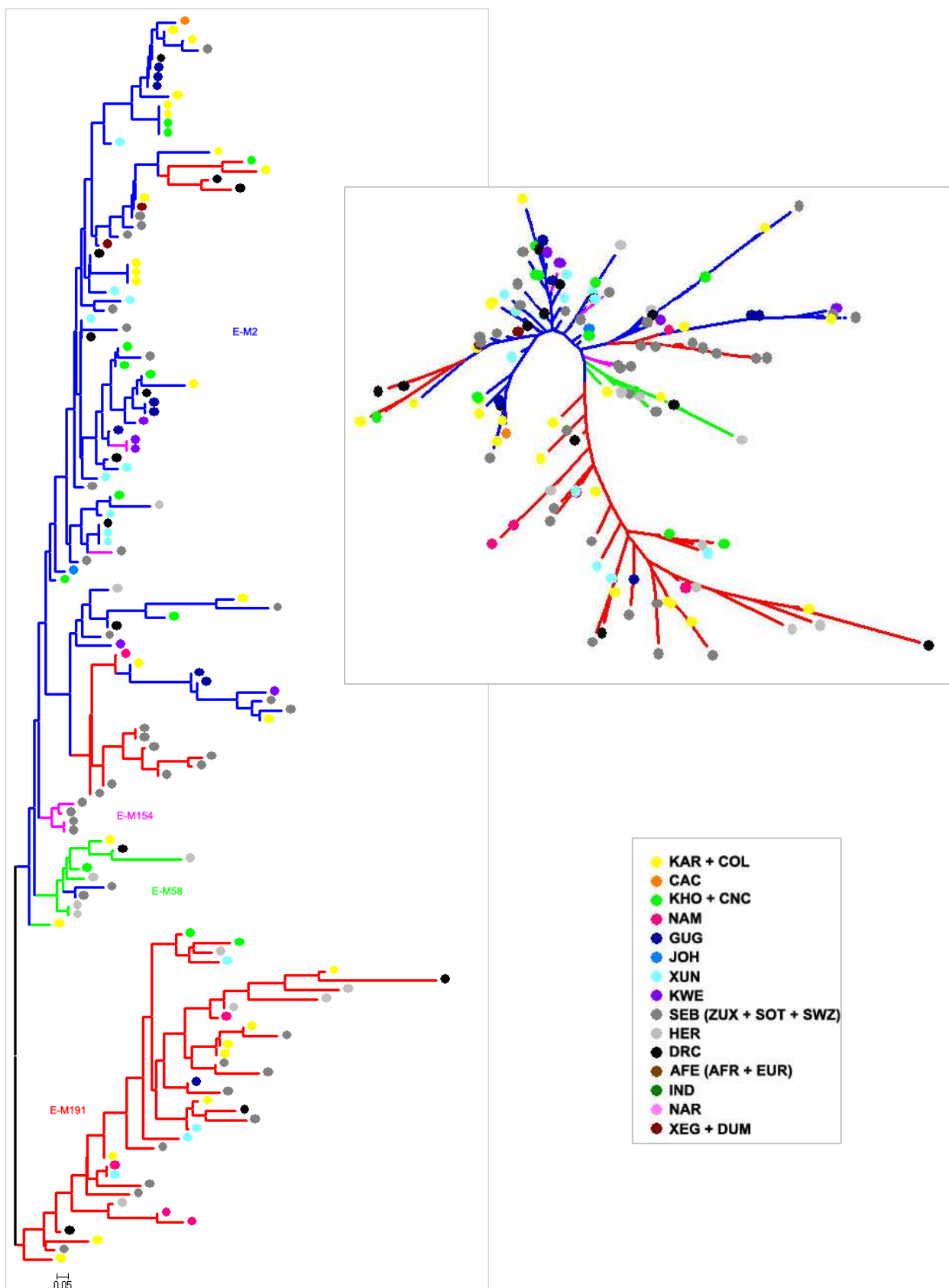


Figure 4.12 Neighbour Joining tree representing the substructure of Haplogroup E-M2. Individuals are colour coded according to the key. The rectangular phylogram and unrooted radial phylogram have sub-haplogroup branches of corresponding colours.

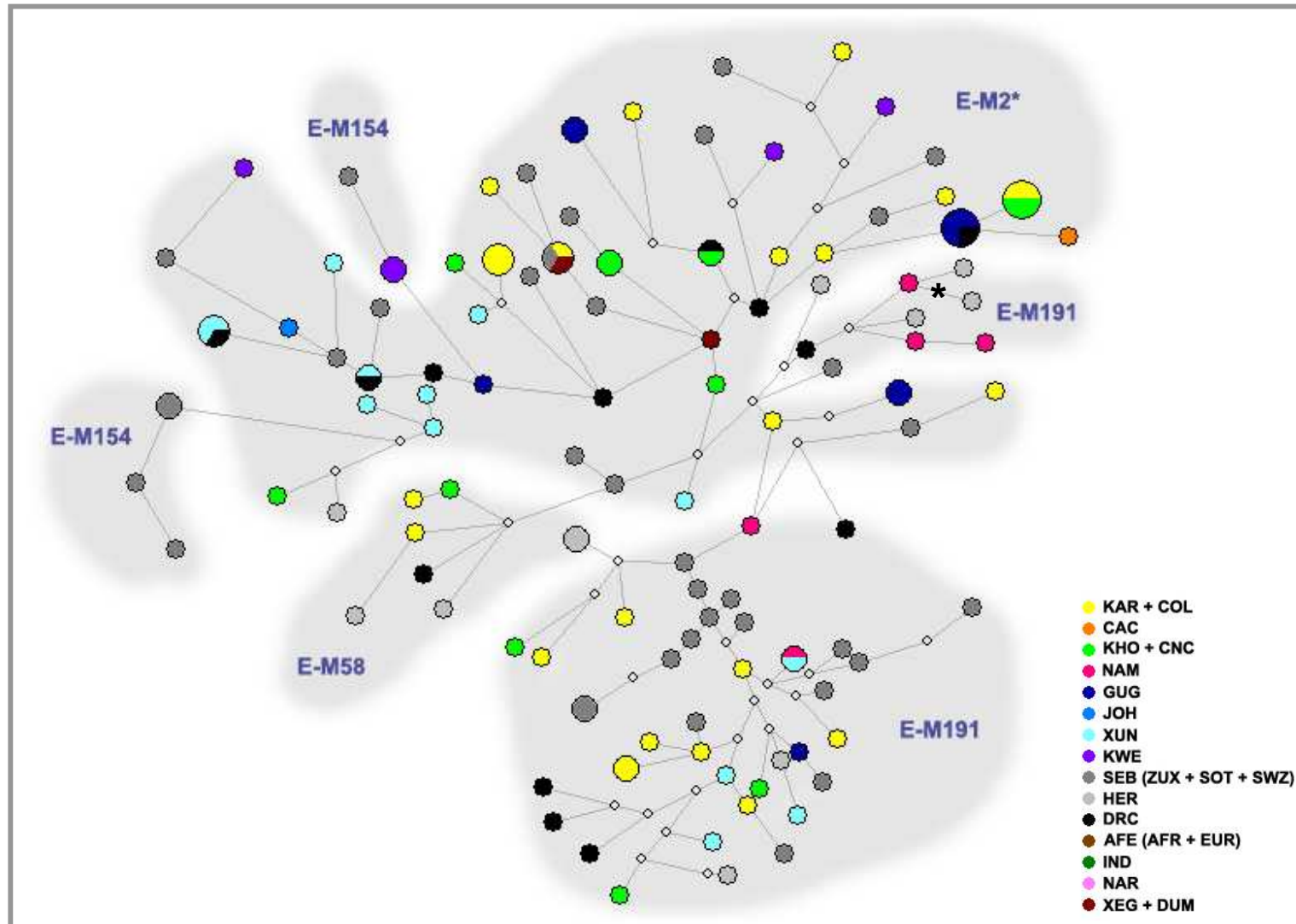


Figure 4.13 Median joining network representing Haplogroup E-M2 substructure in the different populations of the study group.

Haplogroup E-M35

The tree and network (Figure 4.14 and 4.15) separated the two subgroups of E-M35 (E-M78 and E-M34) from the rest of E-M35. There was, however, a clade of E-M35* containing three XUN and one KWE individual that clustered closer to the E-M78 and E-M34 samples. To visualise this relationship better, a MDS plot was constructed using the distance matrix used for the NJ tree (Figure 4.16). The MDS plot also clearly showed that this E-M35 cluster was separate from the rest of E-M35, and grouped closer to E-M34 and E-M78. There were also COL/KAR and a KWE haplotypes that were separated by large distances from the core E-M35* haplotypes.

E-M35* was associated only with Coloured, Khoe and San groups with no representation of BS individuals. The highest representation by far of E-M35 occurred in the KWE group (46%). KHO and NAM also had high frequencies (21-27%) and the COL, CNC and XUN had lower frequencies (5-15%). The GUG group, however, contained no individuals who were E-M35 and the JOH contained only one individual.

The TMRCA for haplogroup E-M35 was 23 205 (+/- 6 303) years BP. The E-M35* haplotypes converged at 17 921 (+/- 5148) years BP while the E-M78 subgroup was younger 8 052 (+/- 3 183) years BP.

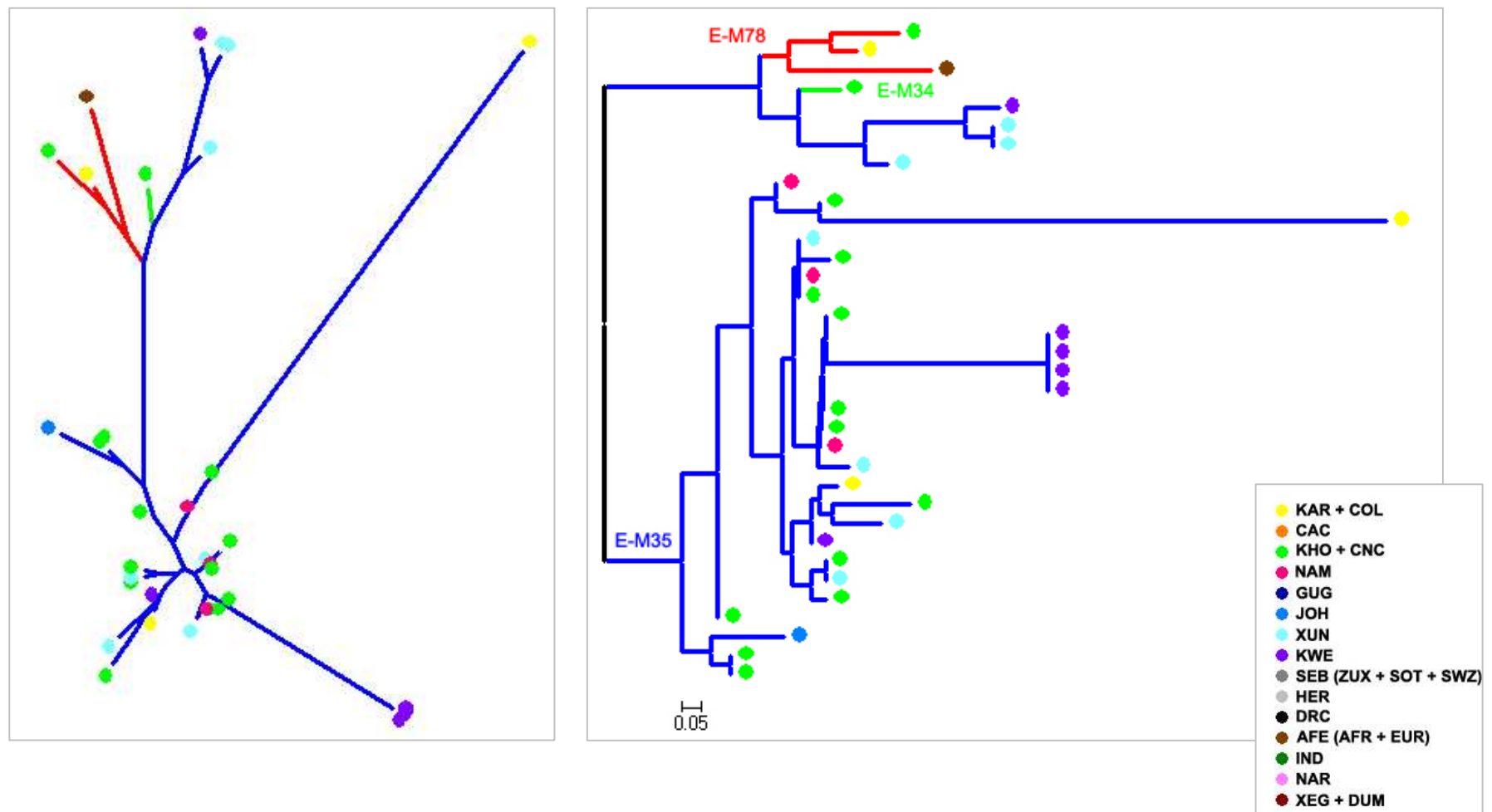


Figure 4.14 Neighbour Joining tree representing the substructure of Haplogroup E-M35. Individuals are colour coded according to the key. The rectangular phylogram and unrooted radial phylogram have sub-haplogroup branches of corresponding colours.

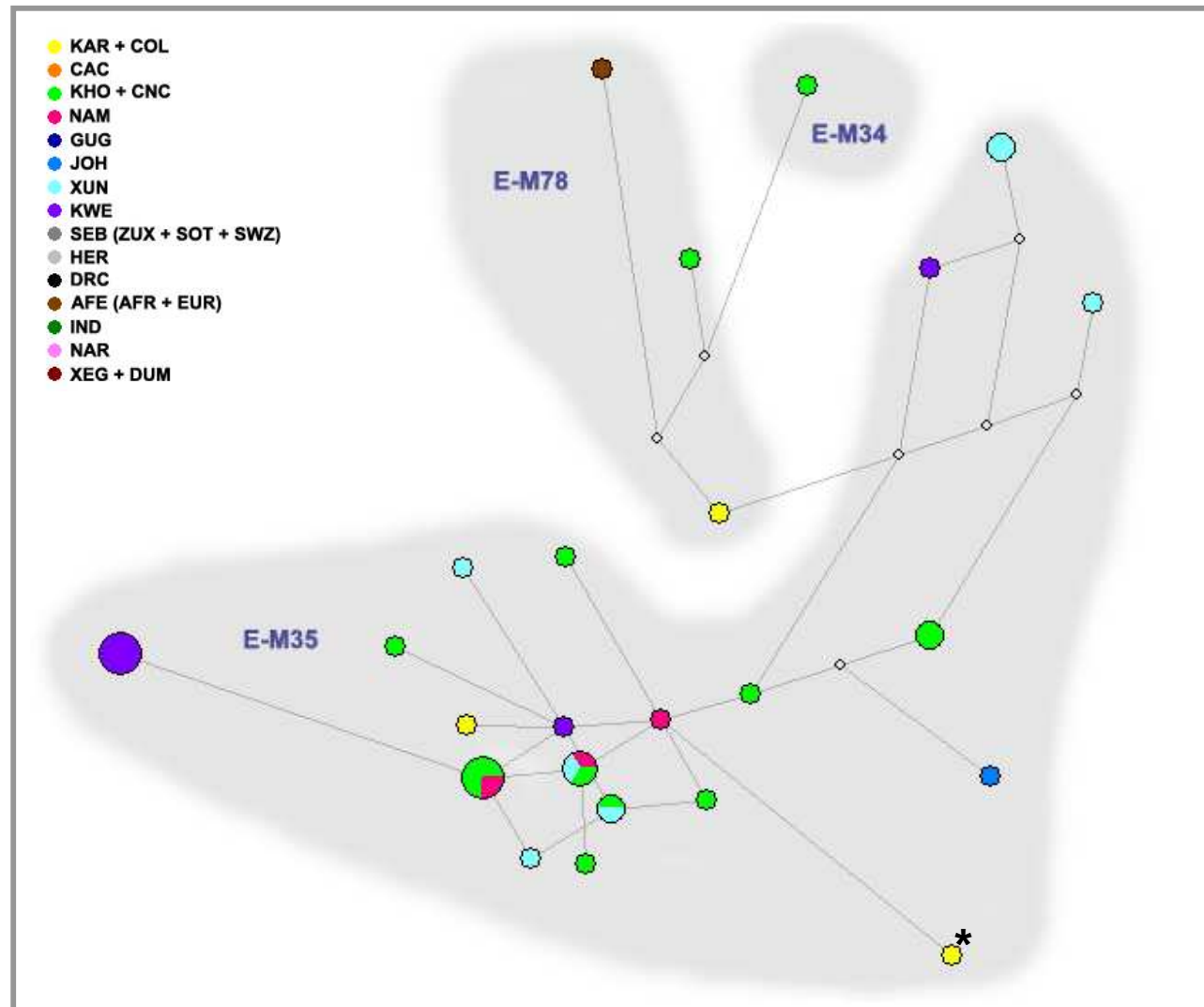


Figure 4.15 Median joining network representing Haplogroup E-M35 substructure in the different populations of the study group.

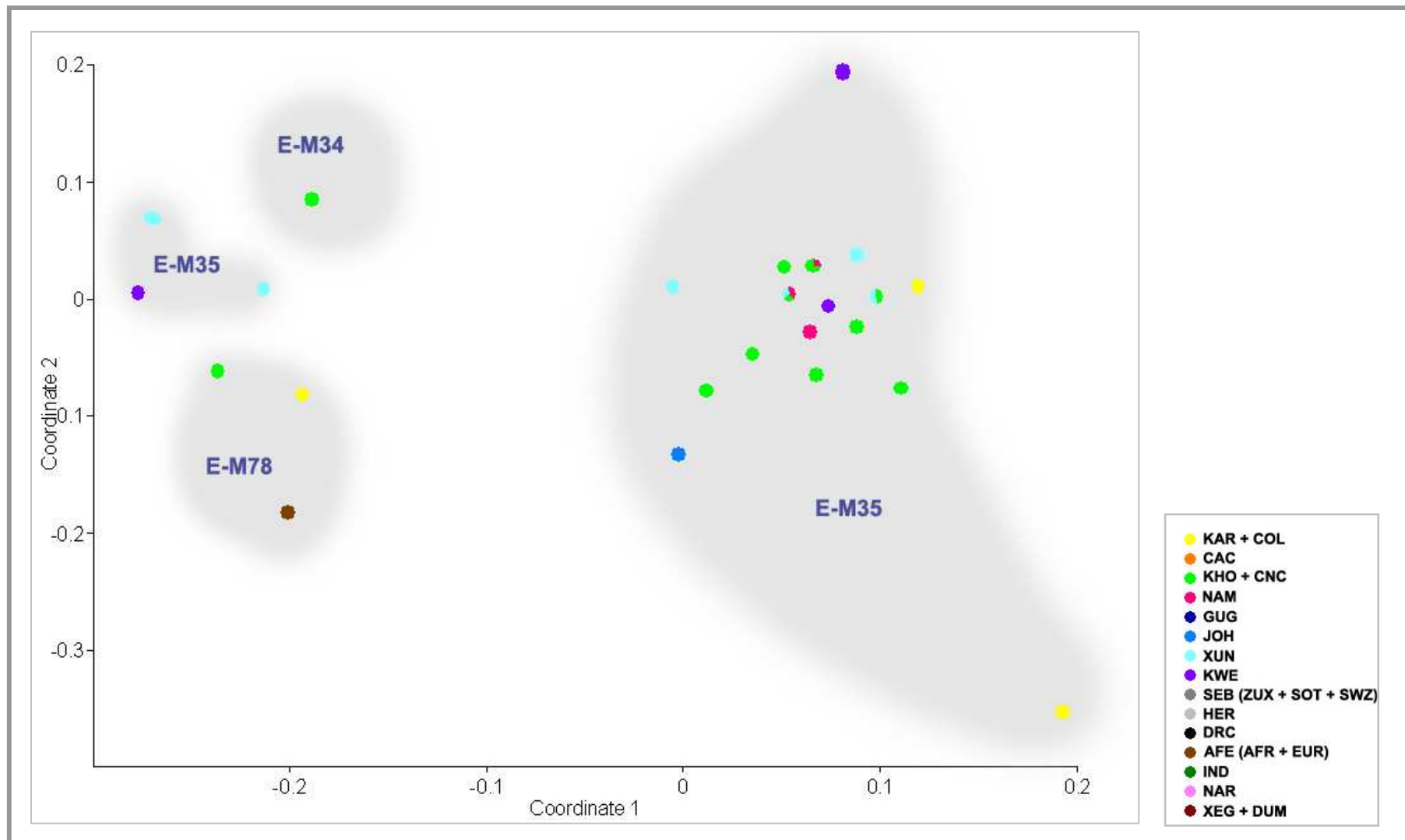


Figure 4.16 MDS plot visualizing the $\delta\mu^2$ distance matrix for haplogroup E-M35 (also used for the Neighbour Joining tree). Individuals are colour coded according to the key.

Haplogroup E - Discussion

The E haplogroups, except E-M35, is associated with Bantu-speakers and presence in Khoe-San and Coloured groups is indicative of Bantu-speaking admixture (see section 1.2.2.3). I will refer to these E haplogroups as BS-haplogroup E. The Bantu-speaking group from the DRC exclusively belonged to BS-haplogroup E (Figure 4.1). The Bantu-speakers from southern Africa had lower frequencies of BS-haplogroup E (76-87%), most likely because of incorporation of hunter-gatherer haplogroups during their migration from east Africa to southern Africa. The non-Bantu-speaking groups with the highest frequencies of BS-haplogroup E was the Colesberg-Coloured and Karretjie groups followed by the /Gui + //Gana + Kgalagari, Khwe, Nama, Coloured-Northern Cape, !Xun, ‡Khomani and the Ju\’hoansi. The high male line Bantu-speaking admixture in the Karretjie, Colesberg-Coloured and /Gui + //Gana + Kgalagari groups was discussed previously. The Ju\’hoansi had only a single individual belonging to BS-haplogroup E. Since it is known that the Ju\’hoansi group was very isolated from outside influence until recently, this furthermore support the notion that BS-haplogroup E was introduced into the Khoe-San groups through recent admixture.

None of the Bantu-speaking groups from the present study had E-M35 representation (Figure 4.1). Most E-M35 haplotypes were not classified into the E-M35 subclades and were E-M35*. The few haplotypes that did belong to the E-M35 subclades (E-M78 and E-M34) were from the Coloured groups and were most likely introduced through admixture from Europeans.

The remaining E-M35* haplotypes were divided into two groups, one group are closely associated with the E-M78 and E-M34 subgroups and the other far larger group was separate from these (Figure 4.15 and 4.16). At the time that the experimental work, which form part of this thesis, was done the E-M293 marker (Henn *et al.*, 2008) was not identified yet. As discussed in the Introduction (see section 1.2.2.3) the E-M293 was found to encompass all the !Xun and Khwe E-M35* haplotypes from that study (Henn *et al.*, 2008). Furthermore, closely related E-M293 haplotypes was identified in the Hadza and Sandawe at high frequencies. The study linked the E-M293 marker to the introduction of pastoralism to the southern parts of Africa (Henn *et al.*, 2008). Without representation of more Khoe-

San groups the study, however, could not address the question of how pastoralism spread after it reached south-central Africa (Henn *et al.*, 2008). The present study offers haplotype frequencies for additional population groups, however, they were not typed with E-M293 and are classified as E-M35*. It is uncertain if all of the E-M35* haplotypes in the present study belong to E-M293. In the MDS plot and network analysis E-M35* clustered into two separate groups (Figure 4.15 and 4.16). Either one or both of these clusters may be E-M293. In the smaller cluster, only Khwe and !Xun were represented, while the larger cluster included haplotypes from all of the Khoe-San and Coloured groups. It may be that both these E-M35* clusters were introduced by a pastoralist group migrating from east Africa.

Henn *et al.*, also typed STRs for their E-M293 samples (Henn *et al.*, 2008). They noted that most of the !Xun and Khwe had DYS389I -10 while the east African populations had a range of repeat amounts at this locus (only one Khwe individual did not have DYS389I-10, it was not stated which repeat number this individual had at DYS389I). In the large haplotype cluster (Figure 4.15 and 4.16), all haplotypes contained DYS389I-10 (except one !Xun individual who had DYS389I -11, which could have been a recent increase). In the smaller cluster all the individuals had 14 or 13 repeats. This smaller cluster could have been an accompanying haplotype of the M293(DYS389I-10) haplotype in its journey from east Africa. It is unlikely that only one haplotype would have migrated south and Henn *et al.*, admitted that it is possible that other male individuals who did not carry M293 were also involved (Henn *et al.*, 2008). Furthermore the Sandawe, Hadza and Datog individuals who had E-M35* had a range of haplotypes both with and without M293(DYS389I-10). It is thus definitely possible that this smaller E-M35* cluster also migrated from east Africa whether it contains M293 or not. Furthermore, the small cluster was located between E-M34 and E-M78 and it is known that these two haplogroups originated in east Africa before spreading to the Middle East and Europe (Semino *et al.*, 2004).

The smaller cluster haplotypes, did not spread to the south, while the larger cluster, DYS389I-10 haplotypes, did (Figure 4.15). It is, however, not likely that the spread of pastoralism was a clear-cut demic or cultural diffusion towards the south. Rather some E-M35* male individuals probably integrated in the southern tribes and took with them the pastoralist practice and possibly also their language. This could be deduced from the

distribution of the E-M35* haplogroup (Figure 4.1 and 4.3). The highest percentage was in the Khwe (46%). The group that introduced pastoralism to the southern parts might well have been the ancestral group to the Khwe population. Aside from their Bantu-speaking admixture the Khwe have a very different Y-chromosome as well as mtDNA profile compared to the other Khoe-San groups. Furthermore the Khwe speak a language from the western Khoe division. It is very important to establish their genetic relationships with the eastern Khoe-speaking San groups (Shua and Tshua), which they phenotypically resemble. As discussed in the mtDNA result section it is one of these eastern Khoe-speaking San groups that harbor the linguistic link to Sandawe through the extinct Kwadi language. The Khwe groups of today are not pastoralists, however, they live in a Tsetse fly invested area. The Shua and Tshua, however, are pastoralists and cultivators or live in close trade relations with Bantu-speakers (Mafisa contracts) in which they tend to their cattle.

Following the Khwe, the groups with the highest E-M35* frequencies were the ‡Khomani (27%) and Nama (21%) (Figure 4.1 and 4.3). Their frequencies were not as high as in the Khwe and unlike the Khwe they contained high frequencies of haplogroups A and B-M112. This suggested not a full population diffusion of the pastoralists but rather incorporation into other resident hunter-gatherer populations. The Nama group adopted the pastoral practice and also speak a Khoe language, however, they still retained a large proportion of original Khoe-San haplogroup A (29%) and had a mtDNA profile similar to the other southern Khoe-San groups. It is difficult to know if the ancestors of the ‡Khomani adopted the pastoral tradition since this grouping of people today have resulted from various disrupted groups. Reports of older individuals, however, indicated that they were hunters and it is known that these groups have spoken the southern San, Tuu division of languages. It might have been that there was a movement of individuals between the southern San hunter-gatherers and the Khoe pastoralists who occupied the same area.

The Colesberg-Coloured (6%) and Coloured-Northern Cape (9%) had lower percentages of E-M35* and the haplogroup was absent from the Karretjie group. The Colesberg-Coloured and to a certain extent the Karretjie and Coloured-Northern Cape would have a large input from the southern Khoe groups such as the Griqua and Cape KhoeKhoe.

Historically, it is known that these groups were pastoralists and spoke Khoe languages. In the Colesberg-Coloured, Karretjie and Coloured-Northern Cape groups, E-M35 frequencies, however, were much lower than in the ‡Khomani, Nama and Khwe groups. This could have been due to the purging of Khoe-San male lineages that happened in the 1700-1800s as discussed previously. When this was taken into account and only Khoe-San associated lineages (A, B-M112, E-M35) were considered (see Figure 4.24) lower frequencies were still seen in these groups. It thus seem that, even though language and pastoralism did transfer from the incoming pastoralists to the southern groups, the male line genetic input in the form of E-M35* declines from north to south and a female line genetic input is almost absent (only two L0k mtDNA haplogroups in Nama, which likely came from recent admixture with the Ju\’hoansi or !Xun)

Incorporation of E-M35* into the !Xun (15%) was lower than into the ‡Khomani and Nama (Figure 4.1). The !Xun also did not adopt the Khoe language nor the pastoralism tradition. Even though the !Xun today have a pastoralist tradition, historical records indicates that they adopted the tradition from neighbouring Bantu-speakers. The relationship between the !Xun and Khwe (here considered to be the remnant of the east African pastoralists) genetic profile was also different from the profile between the Khwe and southern Khoe-San groups. The southern groups adopted language and pastoralism together with an small male-only genetic contribution. The !Xun, on the other hand, did not adopt either the language or pastoralism culture but genetically had more female line (~30% L0k and L0x) than male line (15% E-M35*) contributions from the Khwe.

Similarly the Ju\’hoansi, a group that never adopted pastoralism, had only one E-M35* individual. This Ju\’hoansi E-M35* haplotype had ‡Khomani individuals as its closest neighbours on the network and thus probably did not come from gene-flow with the Khwe directly. The Ju\’hoansi did, however, have substantial (24%) contribution in the female line in the form of L0k and, as discussed in the mtDNA result section, the Khwe L0k haplotypes were ancestral to the !Xun and Ju\’hoansi L0k haplotypes.

Thus to summarise the hypothesis deduced from Y-chromosome and mtDNA results discussed above: The east African pastoralists settled in present-day northern and perhaps

eastern Botswana. Today their remnant genetic variation can be seen in the Khwe and possibly also in the eastern Khoe-speaking San groups. Males from the pastoralist group were incorporated into the southern San groups and transferred their language and pastoralist culture to some of the groups, these groups became known as the Khoe. Some of the other Khoe-speaking San groups might have also been involved and genetic studies on more Botswana San groups is necessary to confirm how exactly the pastoralist culture and language transfer between the Khwe and the Khoe groups took place. The /Gui + //Gana + Kgalagari group from this study did not have E-M35* even though they speak a Kalahari-Khoe language. The E-M35* status of the other central and southern Botswana San groups such as the Naro, !Xõó, ‡Hõã, Tshua, Shua and more /Gui and //Gana groups will be able to help resolve this question. With neighbouring northern San groups, the pastoralists exchanged mostly female and very little genetic male genetic variation. Traditionally the female adopts the culture of the group she relocates to, therefore neither the language nor the pastoralist culture was transferred to the Ju\’hoansi and !Xun (Barnard, 1992).

4.4 Eurasian haplogroups

Haplogroup R – Internal structure

Despite the non-African nature of haplogroup R, this haplogroup exhibited ample representation in the groups studied, especially the Coloured group due to recent admixture. To illustrate haplotype sharing and close neighbours, networks and NJ trees were constructed (Figure 4.17 and 4.18)

The network and tree separated the various subgroups of haplogroup R. Haplogroup R-M343 was most common, with a high degree of reticulation in the network.

Most of the non-Eurasian representation in the R network was from the KHO/CNC and KAR/COL Coloured groups. Furthermore there was one CAC, one NAR, one NAM and two HER individuals. Within R-M343 there was a high degree of type sharing and close neighbours between the KAR/COL and KHO/CNC individuals and the AFR individuals (only three individuals in the network were EUR the rest of the AFE representation were AFR).

Eurasian haplogroups - Discussion

The Eurasian haplogroups found in the study (H, J, I, K, L, P/Q, R) were mostly incorporated into the southern Khoe-San groups by in-moving colonists and slaves from Asia (Figure 4.1). The groups that are presently known as the Coloured population was a result from these unions. The network of the most common Eurasian haplogroup, Haplogroup R, showed many instances of haplotype sharing between Afrikaner and Coloured groups (Figure 4.18). The male and female line contributions in the Coloured populations was very asymmetrical with most of the female lines in the Coloured groups coming from Khoe-San and most of the male lines from the Eurasian and Bantu-speaking input.

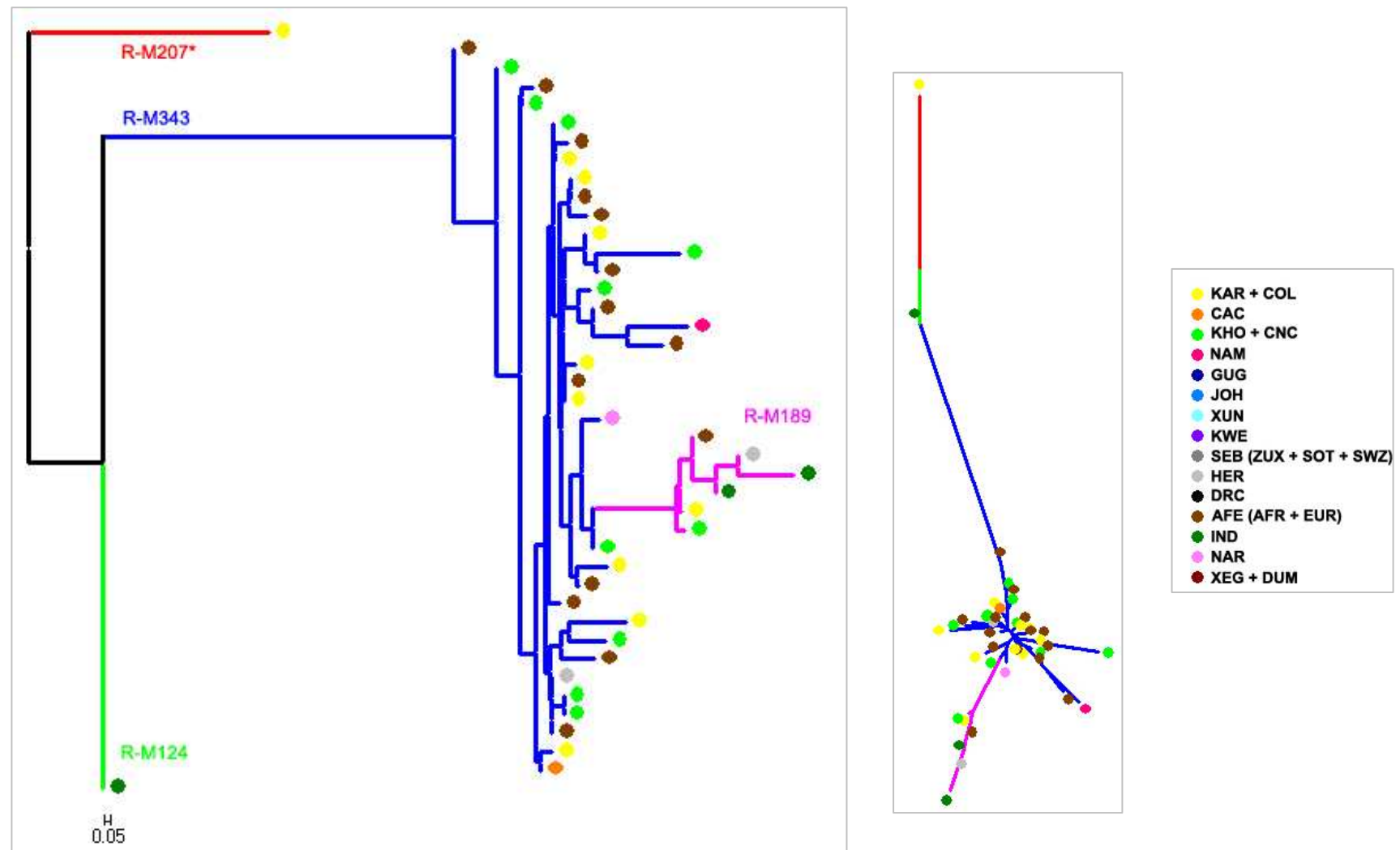


Figure 4.17 Neighbour Joining tree representing the substructure of Haplogroup R. Individuals are colour coded according to the key. The rectangular phylogram and unrooted radial phylogram have sub-haplogroup branches of corresponding colours.

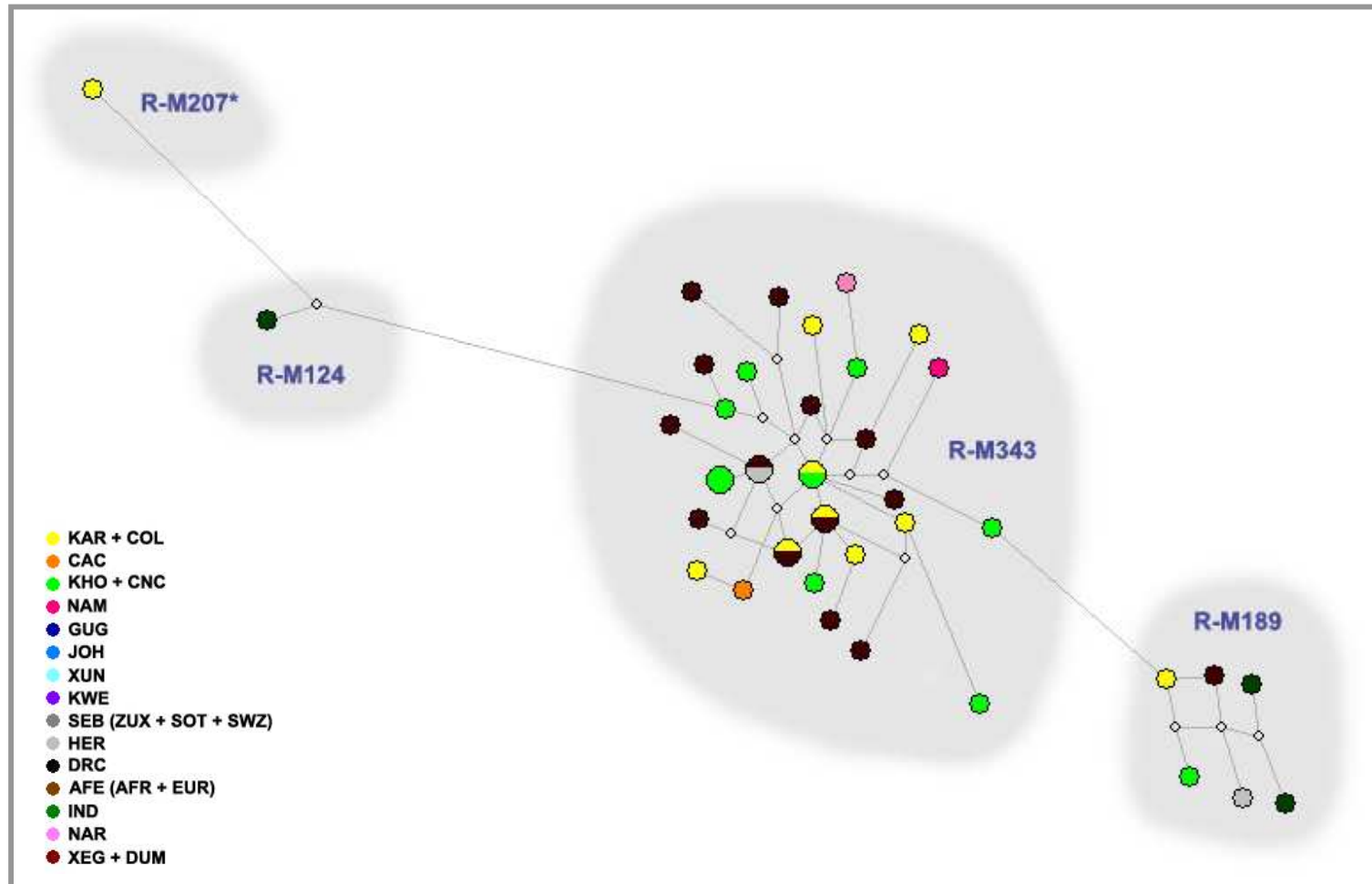


Figure 4.18 Median joining network representing Haplogroup R substructure in the different populations of the study group.

4.5 Analyses of Y-chromosome genetic relationships between different Khoe, San, Coloured and neighbouring groups

The genetic relationships between the 15 populations in the Y-chromosome study group were assessed using exact tests of population differentiation in combination with F_{st} genetic distances on haplogroup frequency data and R_{st} genetic distances on STR haplotype data (Table 4.1). Data from the two types of datasets correlated well when tested using the Mantel test. It confirmed the relationship between haplogroup frequency distances and STR haplotype distances ($p > 0.00001$) with 88% of the haplogroup frequency distance being explained by STR haplotype distance and a correlation coefficient (r) of 0.937 between the two matrices.

The two distance matrices were used to construct PCA plots (Figure 4.19 and 4.20) and trees based on cluster analysis (Figure 4.21 and Figure 4.22).

Table 4.1 Pairwise genetic distances between the 15 study groups calculated from Y-chromosome data

a) Matrix of Fst genetic distances calculated from HG frequency data

	AFE	COL	DRC	CNC	GUG	HER	IND	JOH	KAR	KHO	KWE	NAM	SOT	XUN	ZUX
AFE	0.000														
COL	0.251**	0.000													
DRC	0.640***	0.032	0.000												
CNC	0.341***	0.040	0.140**	0.000											
GUG	0.653***	0.115**	0.183**	0.195***	0.000										
HER	0.419***	0.017	0.053	0.115***	0.241***	0.000									
IND	0.480***	0.129***	0.285***	0.176***	0.318***	0.155***	0.000								
JOH	0.445***	0.131***	0.270***	0.090***	0.286***	0.160***	0.178***	0.000							
KAR	0.317***	0.000	0.027	0.014	0.112**	0.028	0.123***	0.117***	0.000						
KHO	0.280***	0.059**	0.212***	0.012	0.228***	0.129***	0.135***	0.099***	0.062**	0.000					
KWE	0.558***	0.096*	0.218***	0.107*	0.269***	0.189***	0.218***	0.187***	0.104*	0.038	0.000				
NAM	0.437***	0.056	0.156**	0.049	0.299***	0.047*	0.176***	0.120***	0.048	0.036	0.106*	0.000			
SOT	0.478***	0.021	0.010	0.078**	0.147**	0.021	0.190***	0.173***	0.006	0.131***	0.162**	0.031*	0.000		
XUN	0.421***	0.076***	0.156**	0.000	0.203***	0.131***	0.181***	0.061**	0.051***	0.025**	0.086	0.045	0.091**	0.000	
ZUX	0.460***	0.017*	0.004	0.097***	0.108**	0.055**	0.194***	0.189***	0.000	0.147***	0.131***	0.099***	0.000	0.112***	0.000

b) Matrix of Rst genetic distances calculated from STR haplotype data

	AFE	COL	DRC	CNC	GUG	HER	IND	JOH	KAR	KHO	KWE	NAM	SOT	XUN	ZUX
AFE	0.000														
COL	0.236	0.000													
DRC	0.593	0.047	0.000												
CNC	0.250	0.022	0.065	0.000											
GUG	0.632***	0.091***	0.233***	0.150***	0.000										
HER	0.462	0.051	0.033	0.067	0.307***	0.000									
IND	0.220	0.092	0.265	0.104	0.353***	0.206	0.000								
JOH	0.264***	0.067***	0.181**	0.088***	0.226***	0.152**	0.071**	0.000							
KAR	0.219*	0.008	0.065	0.013**	0.139***	0.086*	0.113	0.091***	0.000						
KHO	0.180	0.044	0.155*	0.020	0.199***	0.129*	0.081*	0.086***	0.053***	0.000					
KWE	0.414	0.064	0.146	0.080*	0.233***	0.164*	0.182	0.139***	0.059**	0.049**	0.000				
NAM	0.323	0.034	0.091	0.009	0.236***	0.059	0.110	0.093**	0.059	0.009	0.056*	0.000			
SOT	0.431	0.022	0.013	0.026	0.138***	0.020	0.183	0.132***	0.017	0.113**	0.098	0.062	0.000		
XUN	0.296**	0.032***	0.061*	0.037***	0.210***	0.034***	0.090**	0.075***	0.037***	0.040***	0.056**	0.020*	0.049**	0.000	
ZUX	0.447	0.016	0.036	0.047*	0.129***	0.051	0.209	0.144***	0.026	0.120***	0.108*	0.087	0.000	0.070***	0.000

Abbreviations:

- * significant difference, $P < 0.05$
- ** significant difference, $P < 0.01$
- *** significant difference, $P < 0.001$

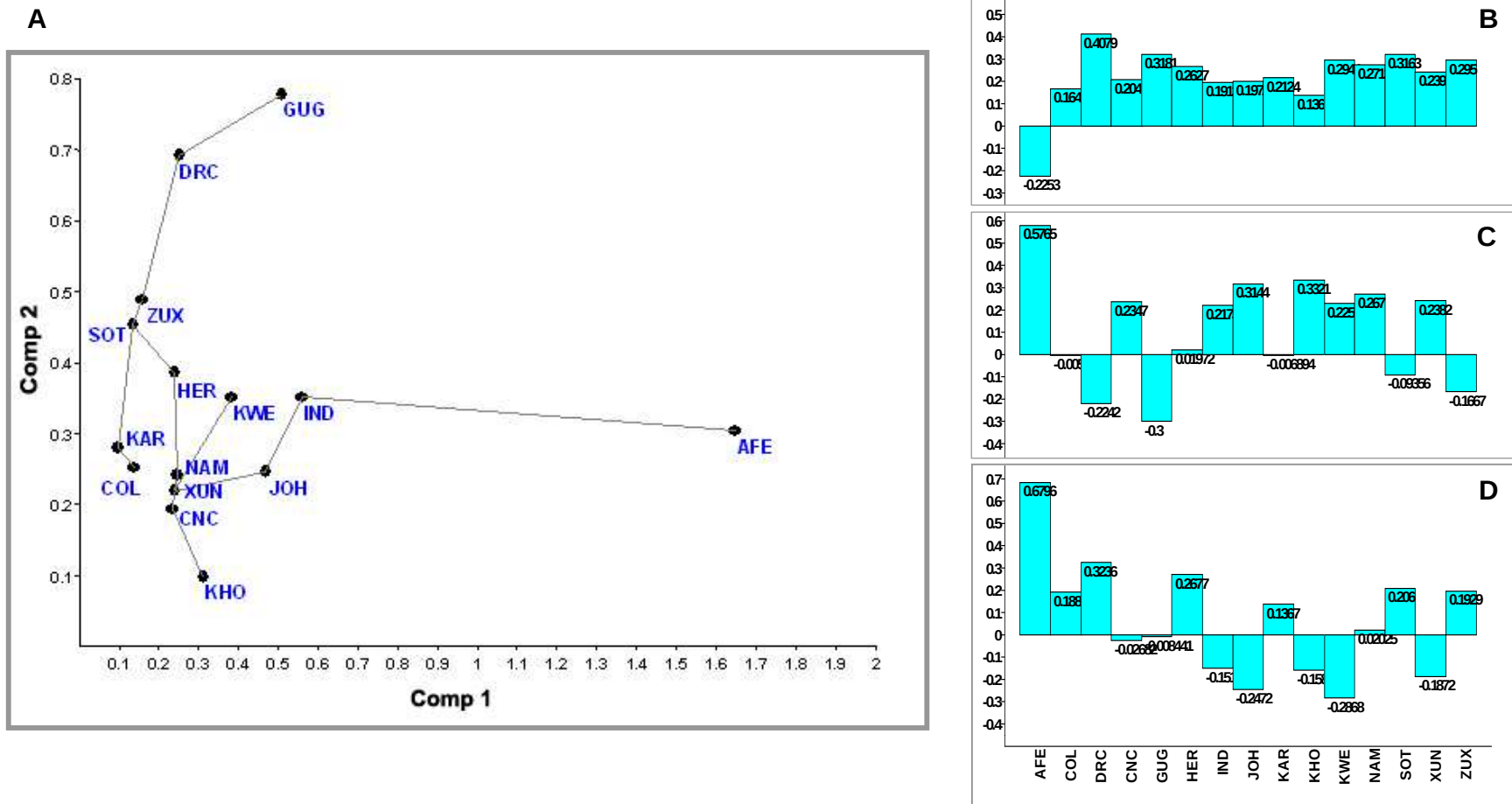


Figure 4.19 **A** – Principal Component Analysis of Y-chromosome Fst values between different populations in the study group. A minimum spanning tree connects populations. Component 1 = 68.7% of the variation, Component 2 = 16.2% of the variation, Component 3 = 7.6% of the variation. **B** – Loadings for Component 1, **C** – loadings for Component 2, **D** – loadings for Component 3

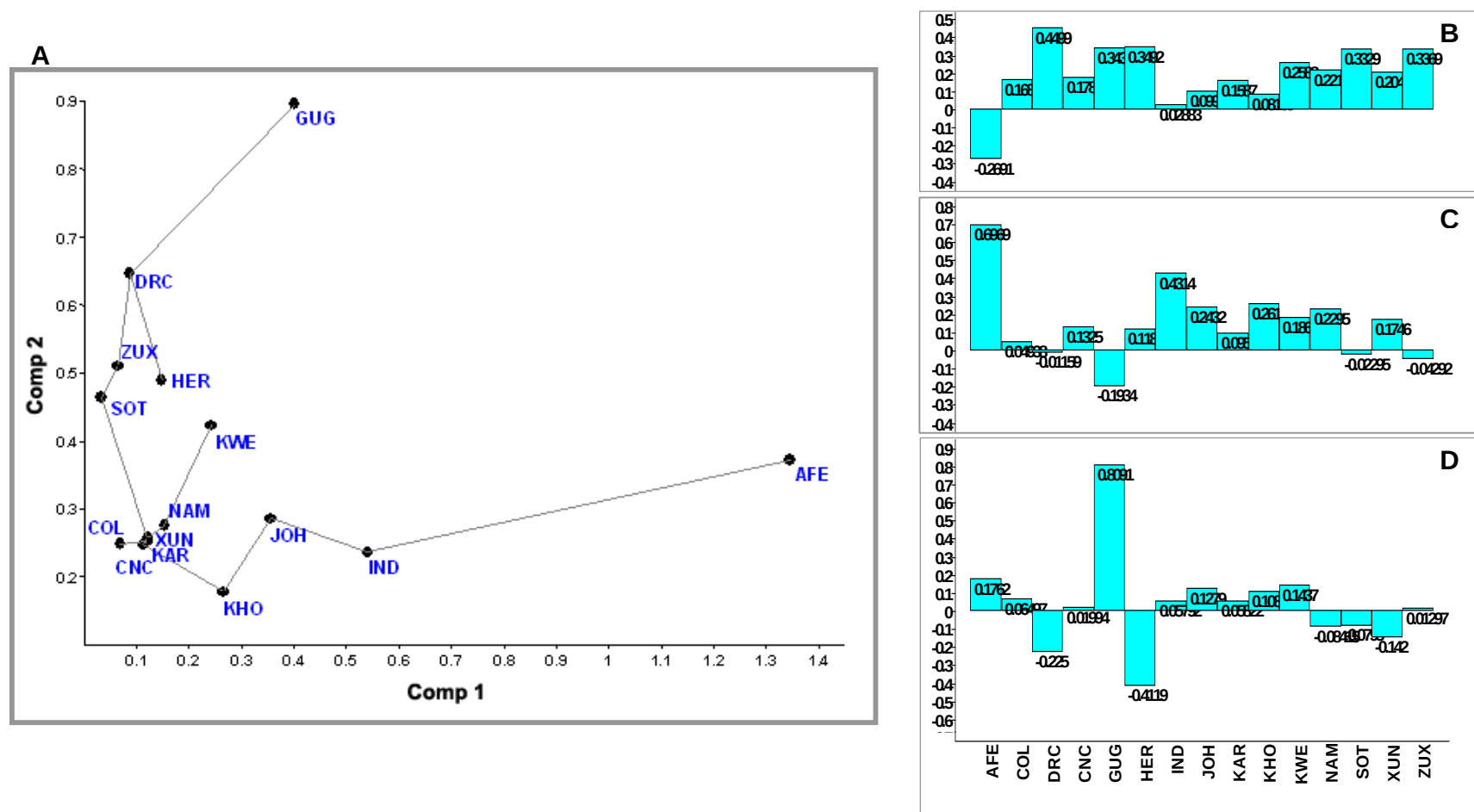


Figure 4.20 **A** – Principal Component Analysis of Y-chromosome Rst values between different populations in the study group. A minimum spanning tree connects populations. Component 1 = 66.9% of the variation, Component 2 = 22.9% of the variation, Component 3 = 5.7% of the variation. **B** – Loadings for Component 1, **C** – loadings for Component 2, **D** – loadings for Component 3

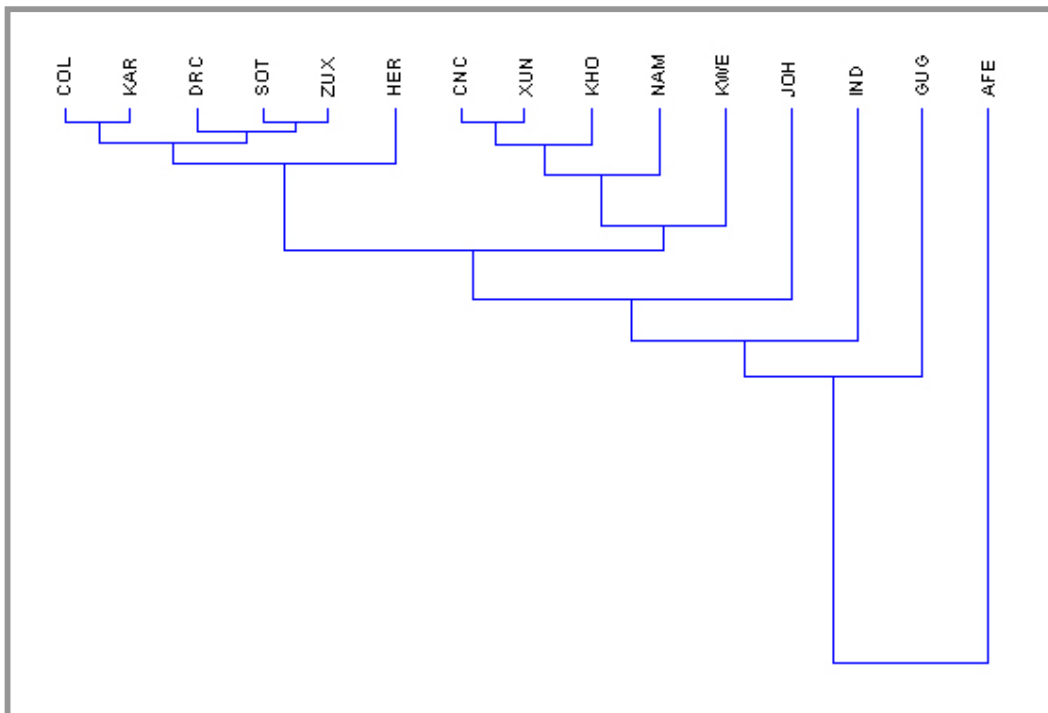


Figure 4.21 Cluster analysis tree representing Y-chromosome F_{st} values between different populations in the study group.

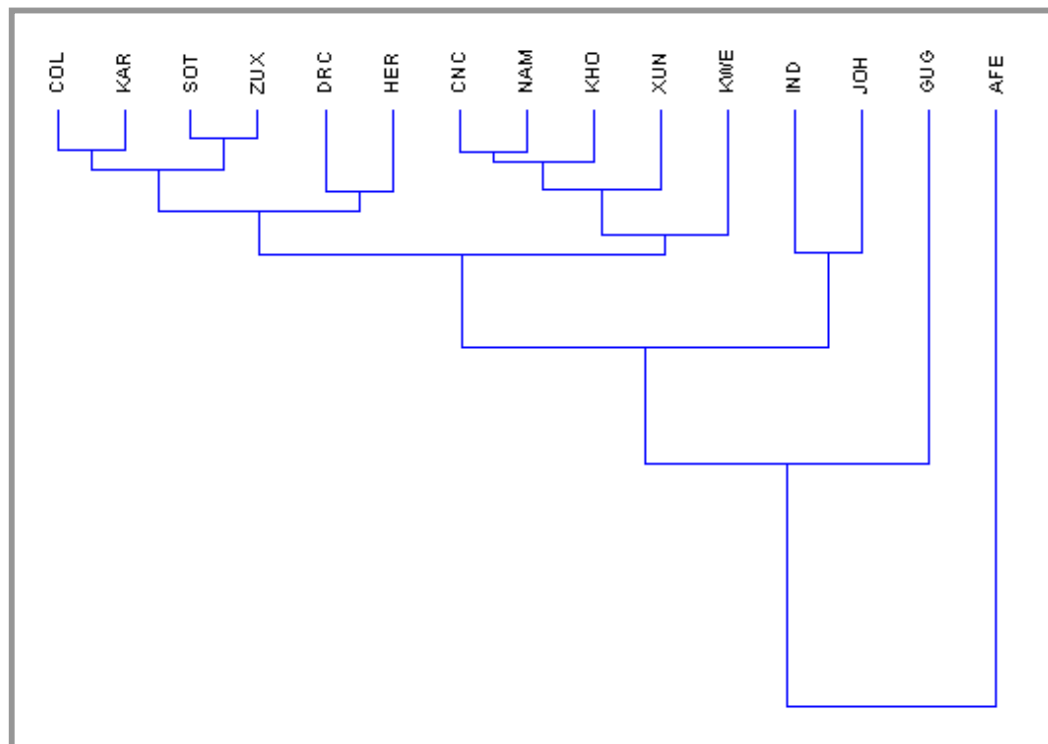


Figure 4.22 Cluster analysis tree representing Y-chromosome R_{st} values between different populations in the study group.

The first component on both PCA plots (Figures 4.19 and 4.20) separated the non-African groups from the African groups and represented 67-68% of the variation. The second component in both cases separated the BS groups from the non-African and Khoe-San-Coloured (KSC) groups. In the Fst PCA plot (Figure 4.19) this component represented 16% of the total variation and in the Rst PCA plot (Figure 4.20) it represented 23%. Component 3 for the Fst PCA plot contained 8% variation and seemed to be a component that separates BS and non-African groups from the KSC group. In the Rst PCA plot the third component (6% variation) separated the GUG group from other groups.

For the cluster analysis, with both haplogroup and haplotype data the AFE, IND, GUG and JOH populations were separated from the other populations. These populations were also significantly different from most other groups in the study group. The separation in the AFE and IND groups was expected. In the GUG group the low level of haplogroup and haplotype diversity together with the relative high frequency of haplogroup B-M152 resulted in isolation from other groups. The JOH group consisted mostly of Haplogroup A and B with a very small contribution of haplogroup E. This separated them from the other groups who all have substantial contributions from haplogroup E.

The rest of the groups formed a monophyletic group with similar internal structures in both datasets. The groups were divided into two branches, the KAR and COL were grouped with the BS groups, while the CNC, NAM, KHO, XUN and KWE grouped on the other branch. KAR and COL had higher contributions from haplogroup E (excluding E-M35), which grouped them closer to the BS groups. In addition, their contributions from haplogroup A and B was similar to the proportions seen in the BS groups. Although the CNC, NAM, KHO, XUN and KWE groups also had high haplogroup E frequencies, a large proportion of their haplogroup E types were E-M35, which was absent from the BS individuals. Furthermore all of these groups except KWE had higher frequencies of haplogroup A than KAR, COL and BS groups. The next level of grouping in this branch indeed excluded KWE in both datasets.

The way in which CNC, NAM, KHO, XUN was grouped further differ between the two sets of data. While the haplogroup data first grouped CNC and XUN together and then

subsequently joined them with KHO and NAM, the haplotype data grouped CNC and NAM and subsequently joined them with KHO and XUN. In the other branch for the haplotype data, KAR and COL grouped together and then joined the southeastern BS group, SOT and ZUL. After that DRC and HER (that also grouped together) joined. In the haplogroup data, HER grouped separately and DRC grouped with SOT and ZUL.

To test the resemblance of the STR-haplotype and haplogroup-frequency based distance to physical distances between Coloured and Khoe-San groups, the genetic distance matrices (Table 4.1) were correlated with a physical distance matrix (Appendix C). In Figure 4.23 pairwise comparisons between physical distance (X-axis) and genetic distance Y-axis was plotted on graphs. A linear regression was done to determine the line with the best fit through the points. Both the graphs of F_{st} and R_{st} vs. physical distance (Figure 4.23) had slightly negative gradients ($-1.451e^{-05}$ and $-1.572e^{-05}$, respectively); these gradients were, however, found to be non-significant ($p = 0.625367$ and $p = 0.533578$). The Mantel test also indicated no associations between the two genetic distances and physical distance that were significantly different from correlation between random datasets generated through permutation tests (F_{st} : $p = 0.625400$ and R_{st} : $p = 0.66770$).

Thus both linear regressions and Mantel tests found no correlation between physical and genetic distance for Y-chromosome haplogroup frequencies and Y-chromosome STRs.

To investigate the possibility that admixture with Eurasian and Bantu-speaking groups erased a gradient of variation in Khoe-San people that existed in the past, genetic distance matrices of groups were again compiled, excluding individuals with Eurasian and Bantu-speaking associated haplogroups. In these matrices only the following haplogroups were included: Haplogroup A and subgroups, haplogroup B-M112 and subgroups, and haplogroup E-M35* (excluding E-M34 and E-M78). The percentages of each of these haplogroups in each of the population groups are represented in Figure 4.24. The PCA and cluster analysis of the resultant R_{st} distance matrix is presented in Figure 4.25 and 4.26.

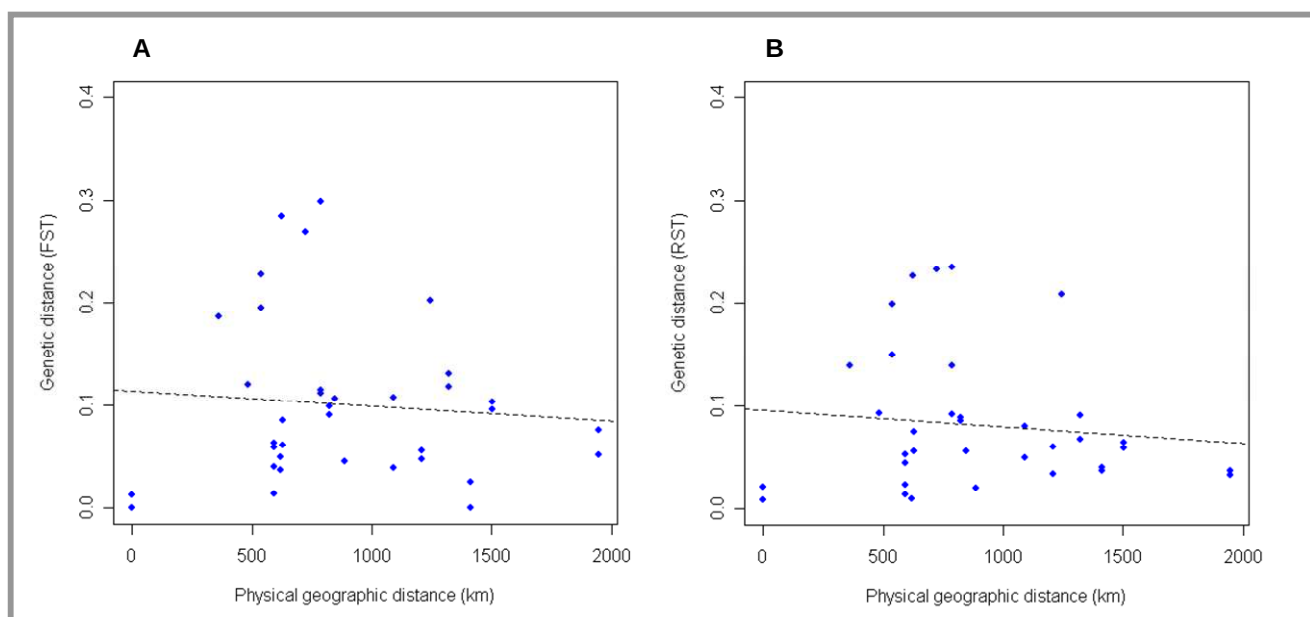


Figure 4.23 Pairwise comparisons between physical geographic distance (X-axis) and Y-chromosome Fst and Rst genetic distance (Y-axis).

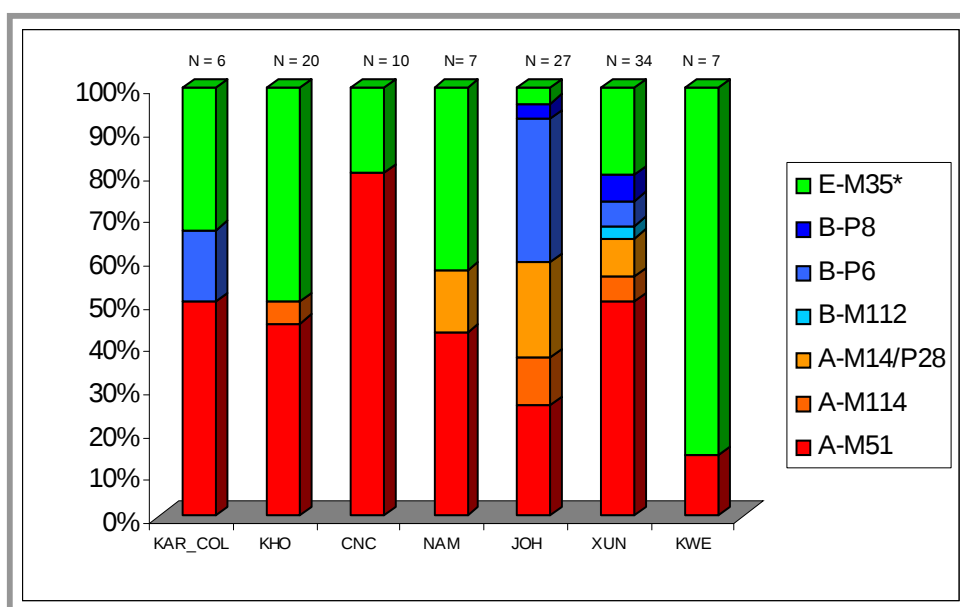


Figure 4.24 Graphical illustration of percentage Y-chromosome haplotype for Khoen-San associated haplogroups in the Khoen-San and Coloured groups.

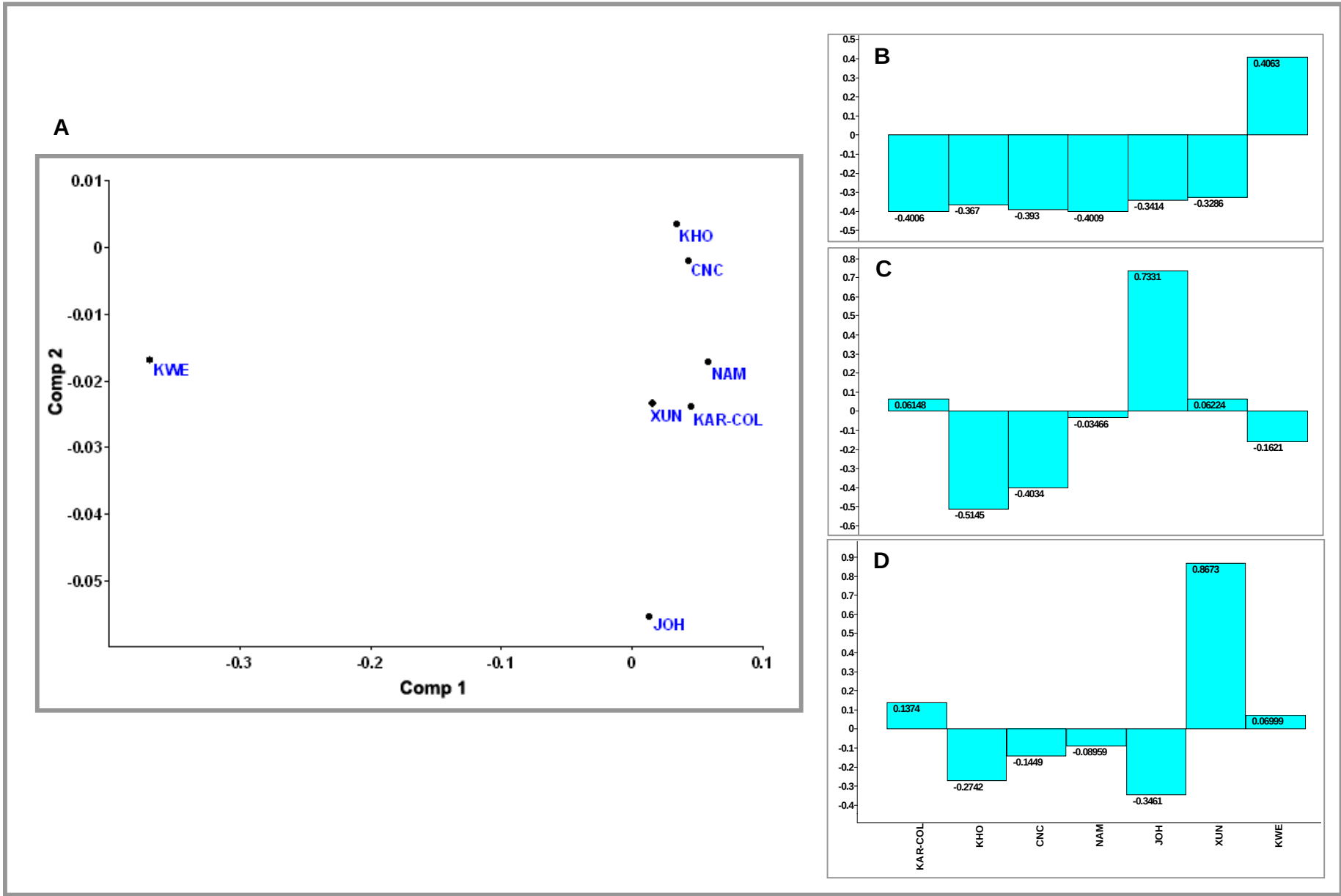
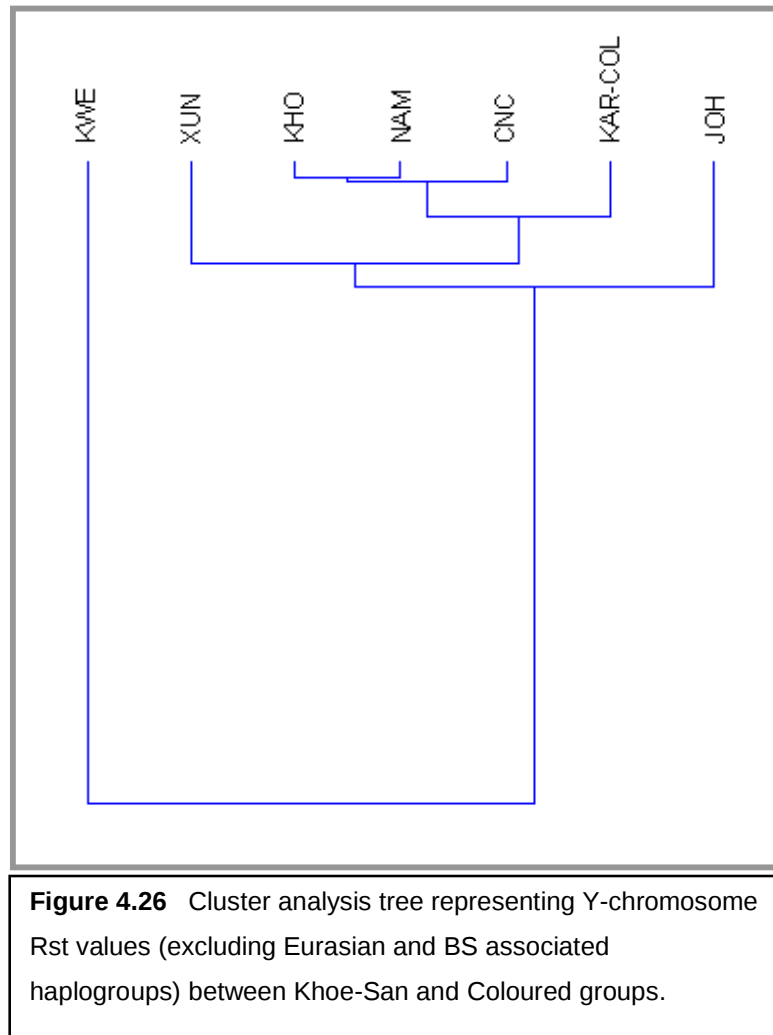


Figure 4.25 Principal component analysis of Y-chromosome Rst values (excluding Eurasian and BS associated haplogroups) between Khoe-San and Coloured groups. A minimum spanning tree connects populations. Component 1 = 97.4% of the variation, Component 2 = 1.5% of the variation, Component 3 = 0.9% of the variation. **B** – Loadings for Component 1, **C** – loadings for Component 2, **D** – loadings for Component 3



For this distance matrix KHO, NAM and CNC clustered closely together (they contained only E-M35 and haplogroup A haplotypes). These three groups were then joined by KAR-COL to form a branch containing all the southern Kho-San and Coloured groups. The XUN subsequently joined this branch and thereafter the JOH group whom had smaller E-M35 and larger Haplogroup B contributions. The KWE group was separate from the other groups and consisted mostly of E-M35 haplotypes with only one haplogroup A haplotype. This is reflected in the PCA plot (Figure 4.25) in which the first component (97% of all variation) separated KWE from the other groups. The subsequent components, 2 and 3, contained only ~1% variation each and separated XUN and JOH from the southern Kho-San-Coloured groups.

In this case the Mantel test still did not find a significant similarity between this Rst distance matrix and the physical distance matrix. The p-value ($p = 0.296$), however, came down and the correlation coefficient had a slightly positive value ($r = 0.008$).

To test the apportionment of variation in the different population groups, AMOVA analysis was done to see how much variation is contained firstly between defined groupings of populations (see Table 4.2), secondly between the different populations in the study and thirdly within the populations. Table 4.2 represent the results of the AMOVA analysis with various different groupings of the first level group.

Overall the AMOVA based on STR data had very high intra-population variances. Very little variation was ascribed to variation among individual populations (generally 2% - 4%) and none to inter-group variation. This illustrated the high variation in STR data and stressed the point that it should only be used for finer mapping within haplogroups.

With the haplogroup frequency data much more variance was ascribed to inter-population and inter-group differences. The inter-populations variances varied between 6-10%. The inter-group variances were the highest (11%) when the non-African populations were split from the African populations (Grouping A – Table 4.2). When the Eurasian, BS and Khoe-San-Coloured (KSC) were grouped into three separate groups, inter-group variance explained 7% of the variation (Grouping B – Table 4.2). When only African groups were considered, there is 5% variance between the KSC groups and the BS groups (Grouping E – Table 4.2). If only the KSC groups are considered and split into northern and southern groups only 0.03 of the variance is explained by the groupings (Grouping F – Table 4.2). The variance explained was less than 1% when the KSC groups were split first into a Khoe-San and Coloured (2 groups) grouping and thereafter into a Khoe, San and Coloured (3 groups) grouping (Grouping G and H – Table 4.2). Of all the different groupings, the only case where inter-group variation was higher than inter-population variation was when the non-African populations were split from the African populations (Grouping A – Table 4.2). In the Eurasian, BS and KSC split the values, however, came close to one another (Grouping B – Table 4.2).

Table 4.2 Results from Y-chromosome AMOVA analysis using different groupings on the first level

Grouping	Grouping of first level [Groups]	RST			Fst		
		Between groups	Between populations within groups	Between individuals within populations	Between groups	Between populations within groups	Between individuals within populations
A	[afe, ind] [col, cnc, kar, kho, nam, joh, xun, gug, kwe, drc, her, sot, zux]	-	2.73	97.92	10.87	9.14	79.99
B	[afr, eur, ind] [col, cnc, kar, kho, nam, joh, xun, gug, kwe, cac] [drc, her, sot, zux]	-	2.85	97.59	7.03	7.90	85.06
C	[afr, eur, ind] [col, cnc, kar, kho, nam] [drc, her, sot, zux] [gug, joh, xun, kwe]	-	2.82	97.44	4.76	8.09	87.15
D	[col, cnc, kar, kho, nam, joh, xun, gug, kwe] [drc, her, sot, zux, afr, eur, ind]	-	2.71	97.48	2.40	10.58	87.02
E	[col, cnc, kar, kho, nam, joh, xun, gug, kwe] [drc, her, sot, zux]	-	2.98	97.34	4.93	6.59	88.48
F	[col, cnc, kar, kho, nam] [joh, xun, gug, kwe]	-	3.87	96.35	0.03	8.15	91.82
G	[col, cnc, kar] [nam, joh, xun, gug, kho]	-	3.81	96.90	0.23	7.70	92.07
H	[col, cnc, kar] [nam] [joh, xun, gug, kho]	-	4.20	97.03	0.19	7.70	92.10

4.5.1 Discussion on the genetic affinities between Khoe-San and Coloured populations from southern Africa

When genetic distances between groups as a whole are compared they depend on the composite haplogroup profile of the group, and the relationships of these haplogroups to one another. Distance trees between groups containing all haplogroups do represent the present day genetic profile of the group. However, as explained in section 3.4, when inferences of the past relationship between Khoe-San groups are attempted, haplogroups that resulted from recent admixture into the groups needs to be removed. For instance

when all haplogroups are included, the Karretjie and Colesberg-Coloured groups cluster with Bantu-Speakers (Figure 4.22), however, when only Khoe-San associated haplogroups (A, B-M112 and E-M35*) are considered, they group with the other southern Khoe-San groups (Figure 4.26).

The cluster analyses of Khoe-San associated haplogroups, group all the southern Khoe-San groups with the !Xun and Ju\’hoansi as separate outsiders. As was seen in the mtDNA group distance analysis the Khwe group is very different from the other groups, even when all Bantu-speaking admixture is removed. The component of the PCA plot that separates the Khwe from the other groups contained 97% of the variation while, the component that separates the !Xun and Ju\’hoansi from the southern groups is very small (1%).

This small amount of variation between southern and northern groups apparent through AMOVA analyses (Table 4.2) was also reflected in the genetic vs. physical distance correlation (Figure 4.23). In neither the case where all haplogroups were included nor where admixed haplogroups were removed were there any indication that genetic and physical distances are related. This is in contrast to the mtDNA results, where there was a correlation between physical and genetic distances. In section 1.2.2.3 the contrasting results from several studies regarding female vs. male gene-flow were discussed. Briefly, what was found is that while in pastoralist groups Y-chromosome based genetic distance is strongly correlated with physical distance, in hunter-gatherer societies Y-chromosome genetic distances are not correlated with physical distance. The lack of correlation between physical and genetic distance of the present study is thus not surprising.

AMOVA analysis also confirmed that, unlike for the maternal lineages, the northern and southern Khoe-San groups are genetically more homogeneous. The distinction between northern and southern Khoe-San contained less than 1% of the variation. Furthermore similar to what was found in the maternal lineages the widely used groupings “San and Khoe”, and “Coloured, Khoe and San” contained <1% variation as well. Significant variation between the male lineages of African and non-African groups was, however, observed in AMOVA analysis (Table 4.2) where 11% of the variation can be ascribed to variation between the continental groups (the same grouping for the female lineages explained 29%

variation). Also, when all the haplogroups and population groups are included for PCA the largest component by far is between non-African and African groups (Figure 4.19 and 4.20). The grouping of Khoe-San+Coloured groups vs. Bantu-speakers contained 5% variation in AMOVA analysis while in the female lineages this grouping explained 22% of the variation (Table 4.2).

5. AUTOSOMAL DNA STUDIES

The results of the typing of the 220 SNP loci are available in the Supplementary electronic data - File A.

5.1 Results and discussion (Genotypes)

For the genotypic part of the autosomal DNA study 100 datasets of 44 unlinked polymorphisms in each were compiled as outlined in Section 2.4. To summarise: The whole dataset contained 220 autosomal SNPs. There were 10 SNPs per chromosome (chromosome 1 to 22), contained in two groups of 5 linked SNPs. The two SNP-groups were completely unlinked from one another (Figure 2.4). To compile datasets for genotypic analysis one SNP per SNP-group (5-linked SNPs) were randomly selected. By selecting one SNP per group for each of the two SNP-groups per chromosome, a set of 44 SNPs are generated. This process was repeated a 100 times to generate 100 different SNP sets, each with 44 unlinked SNPs. These 100 SNP sets with 44 SNPs each formed the dataset for the autosomal genotypic analyses.

5.1.1 Heterozygosity

The proportion of polymorphic loci, heterozygosity and gene diversity for each of the 100 different SNP datasets were calculated for the 14 populations analysed as well as for the total dataset. The averages for these three summary statistics were calculated across the 100 datasets and are shown together with the standard deviation between the 100 datasets in Table 5.1. The heterozygosities in the 14 populations and the total sample set for each of the 100 sample sets are shown as a scatter plot in Figure 5.1.

Higher gene diversities and heterozygosities have been demonstrated for African populations compared to non-African populations (Tishkoff *et al.*, 2009). In the present study the non-African gene diversities and heterozygosities were also low compared to those observed for the African groups (Table 5.1). Unlike previous findings (Tishkoff *et al.*,

2009), however, the Khoe-San populations generally had lower gene diversities than the Bantu-speakers.

To evaluate how heterozygosity correlated with the variation observed between the datasets a scatter plot was generated with the average heterozygosity for each population on the Y-axis and the standard deviation (SD) between the heterozygosities in the different datasets on the X-axis. A linear regression was used to find the function that best described the relationship between the points (Figure 5.2). The linear regression showed that a straight line with a slope of -18.4 best explained the scatter ($p= 0.015$) (Figure 5.2). This suggested that there is a negative relationship between the average heterozygosity and heterozygosity differences between the datasets, i.e., the lower the heterozygosity in a population the higher the differences in heterozygosities among the different sample sets. This indicated that populations with lower heterozygosities (such as the non-African populations) might require more loci to accurately determine correct gene diversity and heterozygosity estimates. When the AFR, EUR and IND data (which are outliers) are omitted, the slope of the regression line was not significant, which might indicate that the differences in the standard variation of heterozygosities between different datasets might also be explained by the differences in non-African vs. African populations.

Table 5.1 Average proportion of polymorphic loci, heterozygosities and gene diversities in each population over the 100 different SNP datasets

Group	N	Ave P	SD P	Ave Gd	SD Gd	Ave Het	SD Het
XUN	45	0.998	0.007	0.384	0.015	0.383	0.017
JOH	41	0.968	0.024	0.361	0.019	0.355	0.020
KWE	19	0.990	0.014	0.407	0.014	0.383	0.018
GUG	21	0.987	0.015	0.383	0.017	0.397	0.020
NAM	28	1.000	0.000	0.413	0.013	0.417	0.016
KAR	25	0.998	0.007	0.397	0.018	0.370	0.018
COL	22	0.991	0.011	0.417	0.014	0.419	0.014
CAC	20	1.000	0.000	0.409	0.014	0.391	0.017
SEB	48	0.996	0.009	0.416	0.014	0.423	0.019
HER	14	0.998	0.007	0.424	0.012	0.406	0.016
DRC	14	0.992	0.014	0.416	0.013	0.407	0.019
AFR	15	0.840	0.046	0.271	0.019	0.256	0.019
EUR	15	0.795	0.042	0.253	0.020	0.262	0.021
IND	25	0.873	0.030	0.276	0.019	0.274	0.021
MEAN		0.961	0.007	0.376	0.010	0.371	0.010

Ave - Average

SD – Standard deviation

P - Proportion of polymorphic loci

Gd - Gene diversity

Het - Heterozygosity

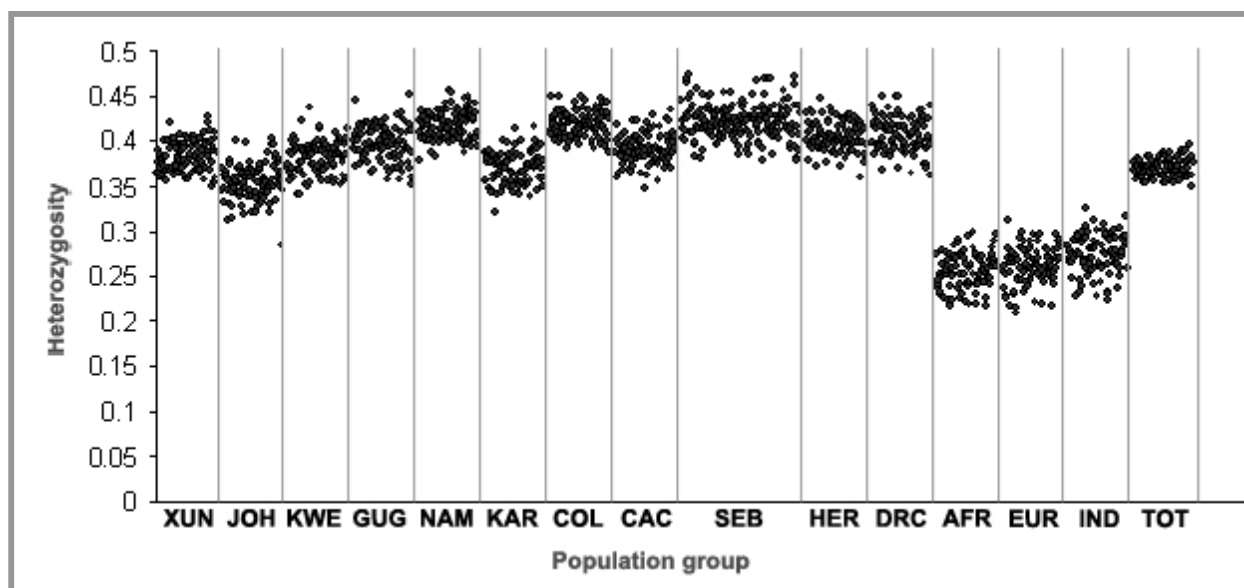


Figure 5.1 Scatter plot of heterozygosities in the 14 populations and the total sample set for each of the 100 sample sets

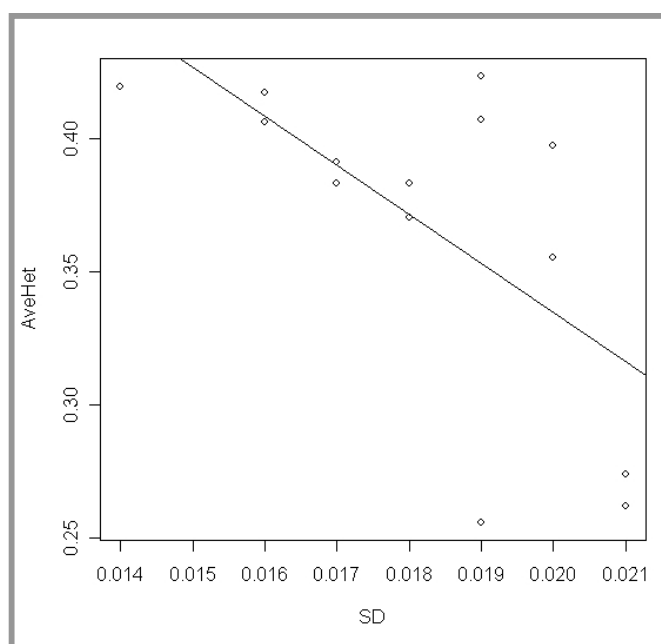


Figure 5.2 Correlation between heterozygosity and the variation observed between the 100 datasets

5.1.2 STRUCTURE analyses

The averaged results of the STRUCTURE runs for the 100 different SNP sets are shown in Figure 5.3 and Table 5.2. The iterations were done for K=1 to K=10. Iterations for K=2 to K=5 are shown.

Similar to previous studies (Jakobsson *et al.*, 2008; Li *et al.*, 2008; Tishkoff *et al.*, 2009), the clustering at K=2 separated African from non-African populations (Figure 5.3 and Table 5.2). The first cluster (blue) predominated in the three non-African populations (AFR, EUR, IND) while the second cluster (yellow) occurred at highest frequencies in African populations. The mixed Coloured populations (CAC, COL) showed a combination of African (yellow) and non-African (blue) contribution. Different amounts of non-African admixture into the Khoe-San and Coloured populations could be observed at K=2 (Figure 5.3 and Table 5.2). Representation from more than one cluster can be an indication of recent admixture or shared ancestry before divergence. The northern San populations (JOH, XUN, GUG, KWE) and Bantu-speakers had very low levels (<10%) from the non-African cluster at K=2 (Table 5.2). This non-African cluster contribution in these groups was most likely because of shared ancestry rather than admixture. This was also apparent if the non-African allocation in the Bantu-speakers was compared to the non-African allocation in the northern San groups. Due to a more recent shared ancestry with non-African groups, the two Bantu-speaking groups (DRC and SEB) had a contribution of around 6% from the non-African cluster while the San groups had a non-African cluster contribution around 3% (Table 5.2). The KWE had similar frequencies to the Bantu-speakers rather than the northern San groups. Similarly the European group (EUR) also had 2% contribution from the African cluster due to shared ancestry while the increased African cluster allocation in the Afrikaner group (AFR) was probably due to recent admixture with African groups (Table 5.2).

The southern Khoe-San and Coloured groups all had more input from the non-African cluster compared to the northern San and Bantu-speakers (Figure 5.3 and Table 5.2). This was consistent with history and with mtDNA and Y-chromosome results. The group with the least amount of admixture (11%) from non-African groups was the Karretjie group

(KAR) (Table 5.2). The Colesberg-Coloured group (COL) that resides next to the Karretjie people had much higher contributions from the non-African cluster (36%). The Cape Coloured group (CAC) had the highest input from the non-African cluster (57%) (Table 5.2). This was also consistent with history, since the CAC group was sampled at Wellington, which is within the region where the original Cape Colony started. It is well known that during the starting years of the colony very high incidences of mixed unions between colonists and local Khoe-San women occurred due to the shortages of female partners.

As K increased ($K > 2$) additional clusters were resolved in the African populations (cluster 2) while the non-African cluster (cluster 1) remained. At $K = 3$, cluster 2 (yellow) predominated in the Khoe-San populations while a third cluster (red) predominated in the BS populations. $K = 3$ thus illustrated the amount of gene-flow between Bantu-speakers and Khoe-San (Figure 5.3 and Table 5.2). Except for the JOH group, the autosomal results supported asymmetric geneflow between the Bantu-speakers and Khoe-San groups with more gene-flow from the Bantu-speakers into the Khoe-San than vice-versa.

The isolated status of the Ju'hoansi group (JOH) was confirmed by autosomal results with a far lower contribution from the Bantu-speakers cluster (13%) than any of the other San groups (Table 5.2). This finding supported the Y-chromosome and mtDNA results (Figure 3.3 and 4.1). Following the JOH, the !Xun group (XUN) had the highest contribution from the Khoe-San cluster (Table 5.2). The contribution from the Bantu-speaking cluster into the XUN was more than double that of the JOH group. As mentioned previously, the !Xun adopted pastoralist practices from surrounding Bantu-speaking groups while the Ju'hoansi maintained their hunter-gatherer lifestyle, isolating them from pastoralists groups.

The Karretjie group (KAR) had the third highest contribution from the Khoe-San cluster. Only the JOH (85%) and XUN (67%) had larger inputs from the Khoe-San cluster than the KAR (55%) (Table 5.2). This finding supported historical records and local opinion that the Karretjie people are descendant from the San groups that once lived in the Karoo (See section 1.1.1.5.3). Their Coloured neighbours (COL) had a much lower input from the Khoe-San cluster (27%) (Table 5.2). Their allocation to the BS-cluster was similar but the

non-African contribution represented by the blue cluster was much higher in the COL. The remaining Coloured group (CAC) had the largest input from the non-African cluster at K=3, while the inputs from the Khoe-San and Bantu-speaking clusters were similar.

In addition to the Ju\’hoansi, !Xun and Karretjie, the Nama (NAM) was the only other group where the Khoe-San cluster (49%) had a greater contribution than the other two clusters (Table 5.2). The Bantu-speaking component in the Nama was larger than the non-African component. This was expected because of the pastoralist culture of the Nama, interaction with the pastoralist Bantu-speakers would not have been uncommon.

The /Gui + //Gana + Kgalagari (GUG) had substantial inputs from the Bantu-speaking cluster (Figure 5.3 and Table 5.2). The Bantu-speaking cluster contributed marginally more than the Khoe-San cluster (Table 5.2). The autosomal results together with the mtDNA (Figure 3.3) and Y-chromosome results (Figure 4.1) therefore illustrated extreme gender biased gene-flow into this mixed group. Autosomal results indicated approximately equal contributions from Khoe-San and Bantu-speakers, while Y-chromosome and mtDNA results illustrated that the male lineages was almost exclusively contributed by Bantu-speakers and the female lineages exclusively by Khoe-San women.

The Khwe (KWE) had the largest input from the Bantu-speaking cluster of all the Khoe-San groups (Figure 5.3 and Table 5.2). This supported previous findings based on the classical blood group markers (See section 1.2.2.1). The KWE did, however, have a larger contribution (35%) from the Khoe-San cluster compared with the Khoe-San contribution into the southern Bantu-speakers (19%). This indicated that the Khwe is not merely a Bantu-speaking group that adopted the hunter-gatherer lifestyle and a Khoisan language.

Higher input from the Khoe-San cluster (~18.6) was seen in the southern Bantu-speakers (SEB and HER) compared to the central African Bantu-speakers (DRC – 11.7%) (Table 5.2). This illustrated the geneflow from resident San groups into the Bantu-speakers when they moved into southern Africa. While the mitochondria indicated much higher gene-flow from the Khoe-San into the SEB than into the HER (Figure 3.3), autosomal results indicated similar frequencies (Table 5.2). This might be an indication that the gene-flow into

the Herero (HER) from the Khoe-San was less female biased. The HER, however, also had less Khoe-San specific Y-chromosome haplogroups (Figure 4.1). It might be that the HER sample size was too small. If not, another cause such as a population bottleneck in the Herero, might explain the pattern. There is evidence that the Herero went through a recent population bottleneck (Excoffier and Schneider, 1999). The low haplotype diversity estimates for both the mtDNA and Y-chromosome results also indicated a possible bottleneck. The original Bantu-speakers that moved to the south might have initially intermixed with the Khoe-San groups. Thereafter, the Herero went through a reduction in population size, which would have caused the loss of many mtDNA and Y-chromosome haplotypes. Subsequently, when the population expanded, the Herero did not intermix with the Khoe-San again. Thus, many of the male and female lineages were lost but the autosomal contribution is still evident.

At $K=4$, the BS cluster was subdivided into two clusters (3-red and 4-green). The red cluster seemed to have lower frequencies than the green cluster in all the Khoe-San and Coloured groups (except the KWE). On the contrary in the BS-groups the green cluster has higher frequencies than the red cluster and this difference was the largest in the DRC group. Higher order clustering ($K=5$ to $K=10$) continued to resolve the BS cluster internally with no apparent substructure between different populations.

The amount of clusters that received the highest average posterior likelihood score across the 100 different SNP sets was $K=3$. The number of clusters, however, with the best delta K score across the runs was $K=2$ (Table 5.3). Although the SD of the likelihoods of $K=2$, $K=3$ and $K=4$ over the 100 different runs did overlap, $K=3$ received the highest likelihood in every single run.

The individual cluster assignments at $K=3$ were also represented in triangle plot (Figure 5.4) with the Khoe-San, Non-African and BS associated clusters on the three different corners of the triangle. From this plot AFR, EUR and IND clearly clustered at the $K=1$ corner while HER, DRC and SEB clustered at the $K=3$ corner and JOH and XUN at the $K=2$ corner. GUG and KWE were positioned on the side of the triangle that separates $K=2$ from $K=3$ while COL and CAC was in the middle of the triangle between the three different

corners. NAM also occurred in the middle of the triangle but was more clustered towards the K2 side. SEB points were more drawn out to the K2 corner than DRC points.

When looking at individual assignments rather than average population assignments (Figure 5.3 and 5.4), it became clear that while certain individuals from admixed groups clearly resulted from admixture between different populations, other individuals had more exclusive cluster assignments. This could especially be seen for certain individuals from the KAR, GUG, NAM and to a lesser extent the KWE group, where some of these individuals clustered amidst the XUN and JOH individuals in the Khoe-San corner of the triangle representation (Figure 5.4). Very few Bantu-speakers clustered towards the Khoe-San corner, and the ones that did were southeastern Bantu-speaking individuals. Certain of the KWE, GUG and COL and to a lesser extent XUN individuals clustered in the Bantu-speaker-corner but only one JOH individual was seen halfway towards the Bantu-speaker corner, the other JOH were in the Khoe-San corner (Figure 5.4). None of the CAC individuals clustered exclusively in the Bantu-speakers or Khoe-San corner and only a few clustered in the non-African corner. Mostly, CAC individuals occurred in the middle of the triangle together with some of the COL individuals, illustrating their individual admixed status (Figure 5.4).

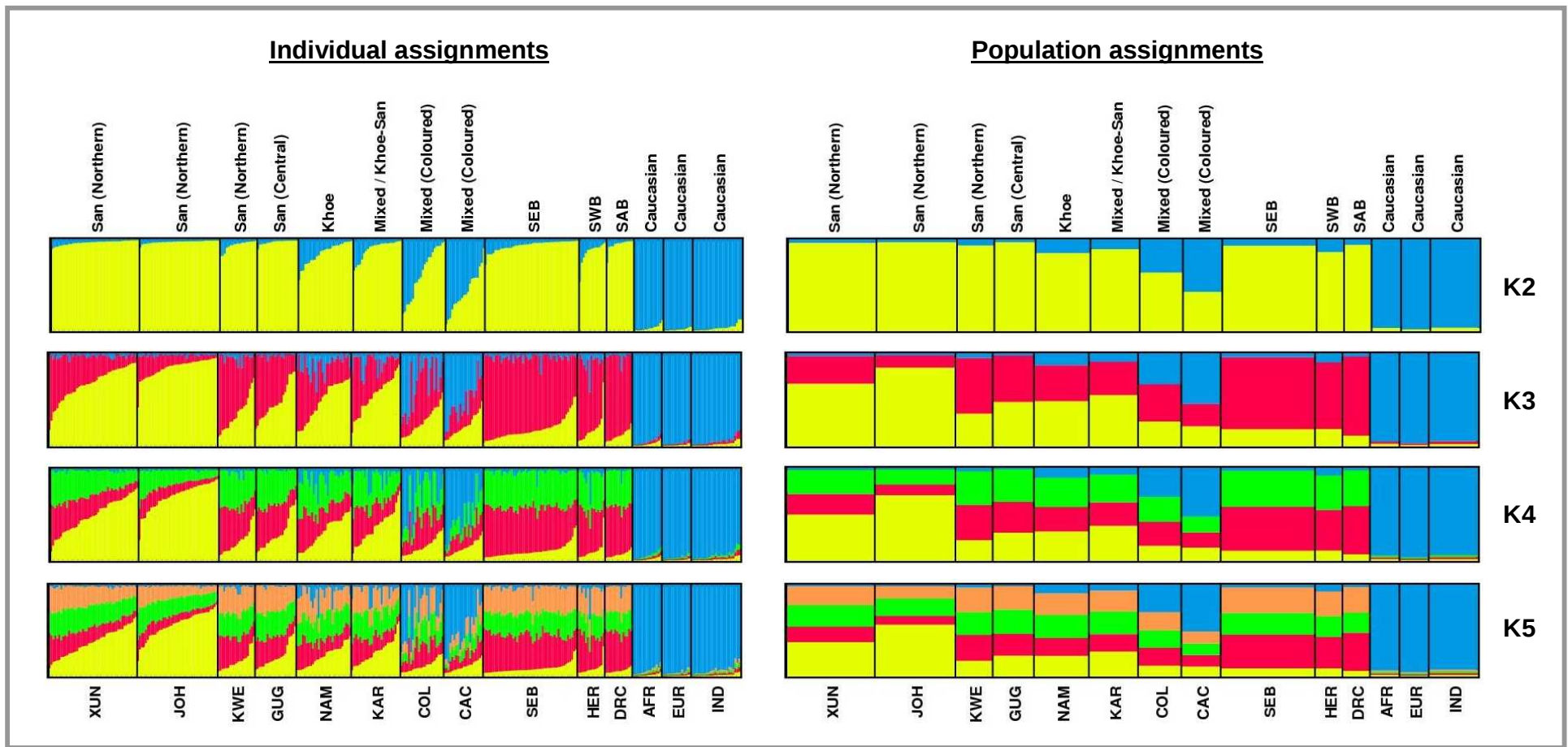


Figure 5.3 Averaged results of the Structure runs of the 100 different SNP sets. K2 to K5 is shown. Individual assignments on the left and population assignments on the right.

Table 5.2 Averaged population cluster assignments of the STRUCTURE runs from the 100 different SNP sets

K	Pop	K1	K2	K3	K4	K5
2	XUN	0.038	0.962			
2	JOH	0.030	0.970			
2	KWE	0.067	0.933			
2	GUG	0.032	0.969			
2	NAM	0.149	0.851			
2	KAR	0.106	0.894			
2	COL	0.361	0.639			
2	CAC	0.571	0.429			
2	SEB	0.068	0.932			
2	HER	0.138	0.862			
2	DRC	0.060	0.940			
2	AFR	0.964	0.036			
2	EUR	0.979	0.021			
2	IND	0.963	0.037			
3	XUN	0.029	0.674	0.296		
3	JOH	0.022	0.846	0.132		
3	KWE	0.048	0.353	0.598		
3	GUG	0.023	0.478	0.498		
3	NAM	0.127	0.487	0.386		
3	KAR	0.091	0.551	0.358		
3	COL	0.332	0.267	0.402		
3	CAC	0.543	0.216	0.242		
3	SEB	0.042	0.185	0.773		
3	HER	0.097	0.187	0.716		
3	DRC	0.035	0.117	0.848		
3	AFR	0.949	0.027	0.024		
3	EUR	0.968	0.016	0.017		
3	IND	0.946	0.026	0.028		
4	XUN	0.022	0.497	0.222	0.259	
4	JOH	0.016	0.705	0.115	0.164	
4	KWE	0.037	0.222	0.380	0.362	
4	GUG	0.017	0.304	0.330	0.349	
4	NAM	0.105	0.317	0.263	0.315	
4	KAR	0.075	0.377	0.252	0.296	
4	COL	0.311	0.162	0.258	0.269	
4	CAC	0.524	0.142	0.163	0.172	
4	SEB	0.033	0.109	0.472	0.387	
4	HER	0.081	0.112	0.436	0.370	
4	DRC	0.028	0.070	0.517	0.385	
4	AFR	0.942	0.021	0.018	0.019	
4	EUR	0.962	0.012	0.013	0.013	
4	IND	0.938	0.020	0.021	0.021	
5	XUN	0.017	0.375	0.168	0.205	0.168
5	JOH	0.013	0.565	0.095	0.138	0.095
5	KWE	0.031	0.171	0.283	0.268	0.283
5	GUG	0.014	0.227	0.236	0.259	0.236
5	NAM	0.093	0.224	0.196	0.238	0.196
5	KAR	0.065	0.273	0.188	0.226	0.188
5	COL	0.298	0.119	0.193	0.200	0.193
5	CAC	0.509	0.108	0.126	0.132	0.126
5	SEB	0.027	0.089	0.367	0.235	0.282
5	SEB	0.030	0.086	0.369	0.283	0.369
5	HER	0.072	0.089	0.345	0.271	0.345
5	DRC	0.024	0.062	0.415	0.281	0.415
5	AFR	0.934	0.017	0.015	0.016	0.015
5	EUR	0.955	0.010	0.011	0.012	0.011
5	IND	0.928	0.016	0.018	0.018	0.018

Table 5.3 Average likelihood and delta-K scores across the 100 runs

K	Ln Likelihood	SD Ln Likelihood	Delta-K	SD Delta-K
2	-16629.9	365.044	980.1158	392.2799
3	-16520.4	362.0634	93.40483	42.02413
4	-16760	372.3144	2.558441	2.593335
5	-16940.9	426.0762	2.205267	2.334452
6	-17078.1	399.9683	1.684401	1.700221
7	-17185.3	417.852	0.958339	0.81949
8	-17302.8	431.8609	0.797023	0.721795
9	-17412.4	428.0192	0.739302	0.599983
10	-17496.5	440.4873	0.573535	0.490531

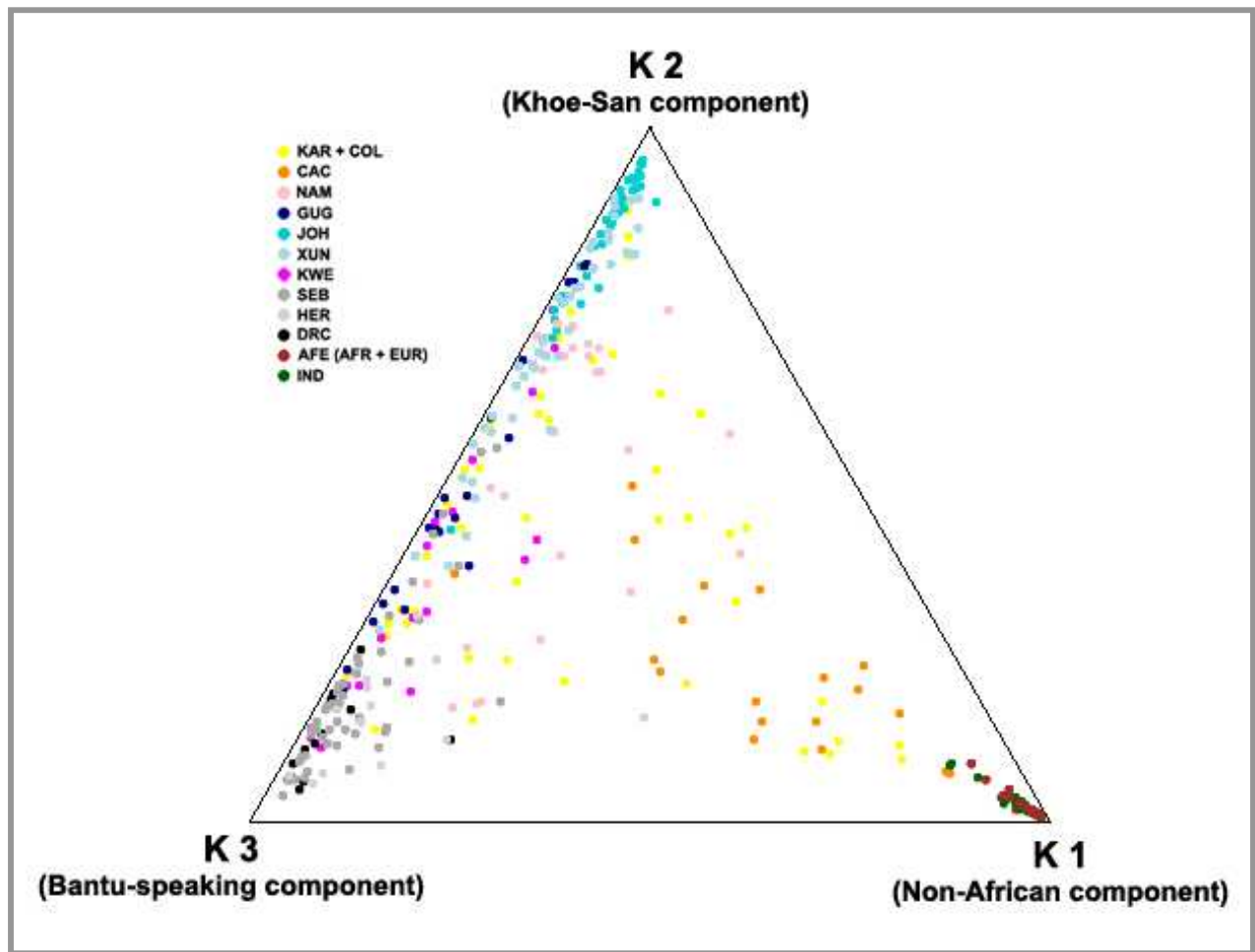


Figure 5.4 Triangle plot of individual cluster assignment at K=3 with the Khoe-San, non-African and BS associated clusters on the three different corners of the triangle

5.1.3 Variation across STRUCTURE datasets

Cluster assignment of populations and individuals over the 100 datasets differed from each other. Figure 5.5 shows a graphical representation of the variation between the population cluster assignments of the 100 datasets. Each dot represents a population K-cluster assignment from one of the 100 datasets.

At $K=2$ the datasets correlated relatively well and the dots form tighter clusters compared to the higher order cluster assignments. Cluster assignments in the different datasets of the non-African and Khoe-San populations were more homogeneous than in the BS and Coloured populations. In most populations clusters were well separated but were more closely associated in the COL and CAC.

At $K=3$ the cluster assignments over different datasets was more heterogeneous than for $K=2$. The non-African populations (blue dots) still had more homogenous cluster assignments over the different datasets than the African populations (red and yellow). The XUN and the JOH had more homogeneous results compared to the other African populations. The red (BS-associated cluster) and yellow (Khoe-San associated cluster) overlapped in many cases. The yellow cluster was well separated from the lower red cluster in the JOH and XUN. In the BS groups the yellow and red clusters were also well separated with the red cluster in this case being the highest assignment. In all the other populations the red and yellow cluster assignments between the different runs overlapped in different extents. This indicated that one might get a wrong picture of a population by just looking at one SNP set. However, by averaging across several SNP sets, as was done here, (Figure 5.1 and Table 5.1) a much more confident deduction could be made for the population cluster assignments.

The higher order clusters $K=4$ and $K=5$ continued to show the higher heterogeneity across datasets. African clusters as well as the African populations compared to the non-African cluster and populations were still more heterogeneous across runs for different datasets.

Pearson's correlation coefficient (r) was calculated for each pair of 100 datasets for each K of $K=3$ and all correlations were significant (Supplementary electronic data - File B). Pairwise correlations (of individual cluster assignments for each K) between the 100 different datasets varied between $r = 0.60$ and $r = 0.91$ with an average of $r = 0.78$.

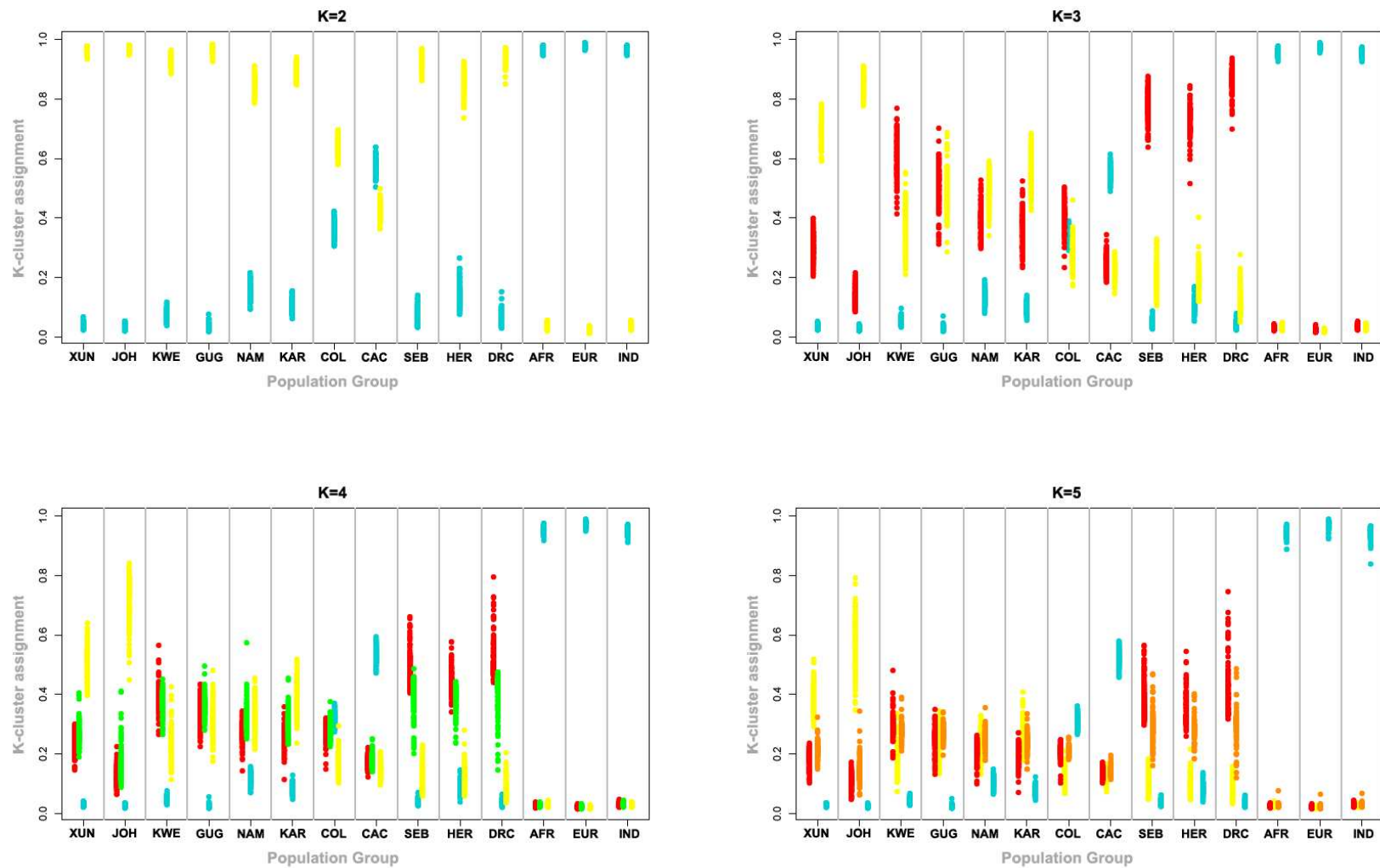


Figure 5.5 Graphical representation of the variation between the population cluster assignments across the 100 runs. Each dot represents a population K-cluster assignment of one of the 100 runs.

5.1.4 Distance based analysis of unlinked SNP sets

The same 100 SNP datasets used in the STRUCTURE analysis was also used in distance based analysis. The 100 population distance matrices were used to construct 100 NJ trees, which were then condensed into a consensus NJ tree (Figure 5.6a). The numbers on the branches indicate the number of times the particular node was supported after computing 100 trees. This approach does not take into account the distances between groups in the form of differences in branch lengths.

A second approach used was to first condense the distance information from the 100 SNP sets into one average population distance matrix (Table 5.4) from which a NJ tree was then constructed (Figure 5.6b). This approach does not indicate the number of times branches were supported by the 100 different datasets. However, the calculated means of the 100 distance matrices is represented on the tree by variable branch lengths.

The overall affinities of the populations using both approaches corresponded reasonably well. Both types of trees that summarised the distance data from the 100 different sample sets clearly divided non-African and African variation with the CAC and COL groups being placed at intermediate positions (Figure 5.6a and 5.6b). Furthermore, Bantu-speaking groups (DRC, SEB, HER) and also northern Khoe-San groups (JOH, XUN, GUG, KWE) formed monophyletic clades on both trees. The KWE assignment to this cluster, however, was weakly supported by the NJ consensus tree (Figure 5.6a). XUN and JOH grouped together, with GUG and KWE being placed closer to the BS populations, most likely due to admixture from these groups (see Figure 5.3). The influence of the non-African component in the KAR and NAM positioned these populations more towards the non-African branch. The high Khoe-San input into these two populations apparent from the STRUCTURE results could not be deduced from the distance trees.

Trees as representations of distance analysis are very sensitive to the influence of admixture in the groups. When data is visually represented in a tree, distances are only optimized in one dimension. This was especially apparent in the KAR and NAM group who

was heavily influenced by the variation contributed by the non-African group. This caused them to group between African and non-African groups and not with the other Khoe-San groups (Figure 5.6a and 5.6b). When STRUCTURE results were considered, however, one could see that these two groups actually had a larger Khoe-San cluster representation than the GUG and KWE (Figure 5.3 and Table 5.2). Yet, the GUG and KWE grouped in a Khoe-San cluster with the JOH and XUN on the tree because of their small contribution of non-African variation (Figure 5.6a and 5.6b). It is thus useful to utilize techniques, such as PCA, that are able to optimize the distance matrix in more than one dimension (Figure 5.7 and 5.8). To extract additional information from the distance matrices, PCA was performed (Figure 5.7)

Table 5.4 Average population distance matrix of autosomal genotypic data

	CAC	COL	IND	KAR	KSB	KSJR	KSNA	KSSK	KSV	SAB	SEB	SWB	WA	WE
CAC	0.000													
COL	0.042	0.000												
IND	0.092	0.157	0.000											
KAR	0.104	0.049	0.273	0.000										
KSB	0.122	0.069	0.293	0.053	0.000									
KSJR	0.158	0.105	0.350	0.046	0.066	0.000								
KSNA	0.085	0.040	0.246	0.021	0.053	0.054	0.000							
KSSK	0.157	0.087	0.345	0.054	0.040	0.057	0.054	0.000						
KSV	0.144	0.083	0.331	0.034	0.047	0.028	0.041	0.037	0.000					
SAB	0.141	0.089	0.314	0.074	0.056	0.108	0.069	0.066	0.078	0.000				
SEB	0.116	0.059	0.290	0.046	0.036	0.074	0.047	0.039	0.049	0.027	0.000			
SWB	0.100	0.050	0.244	0.061	0.051	0.096	0.054	0.067	0.066	0.051	0.035	0.000		
WA	0.107	0.182	0.054	0.289	0.314	0.364	0.259	0.361	0.347	0.334	0.310	0.270	0.000	
WE	0.125	0.195	0.046	0.307	0.336	0.388	0.277	0.382	0.367	0.354	0.329	0.290	0.031	0.000

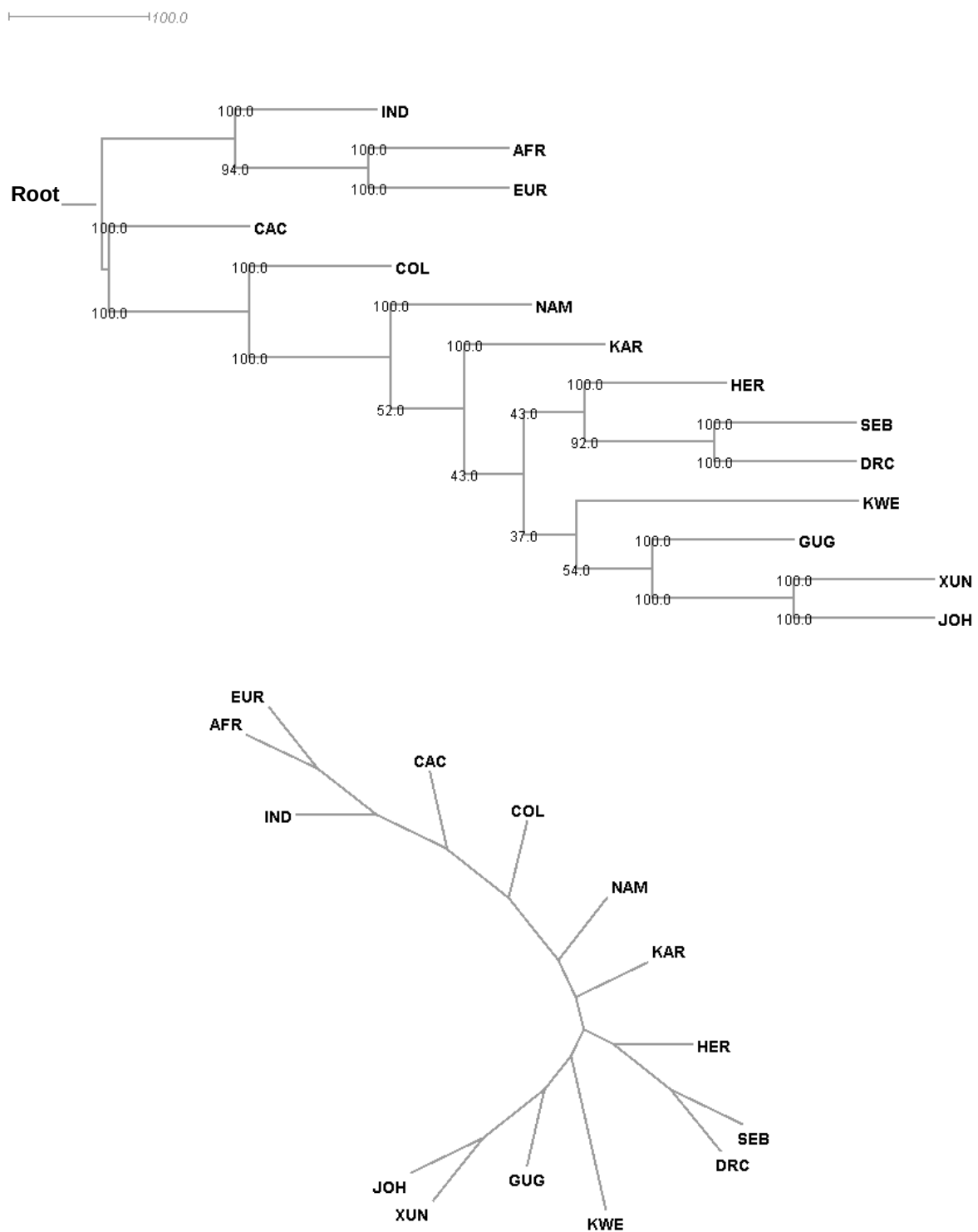


Figure 5.6a The Majority Rule consensus tree constructed from a 100 NJ trees. A rectangular phylogram shows the branch support (indicating the number of times the particular node is supported after computing 100 trees). The root was placed between CAC and the non-African population. The radial phylogram shows the unrooted version of the tree.

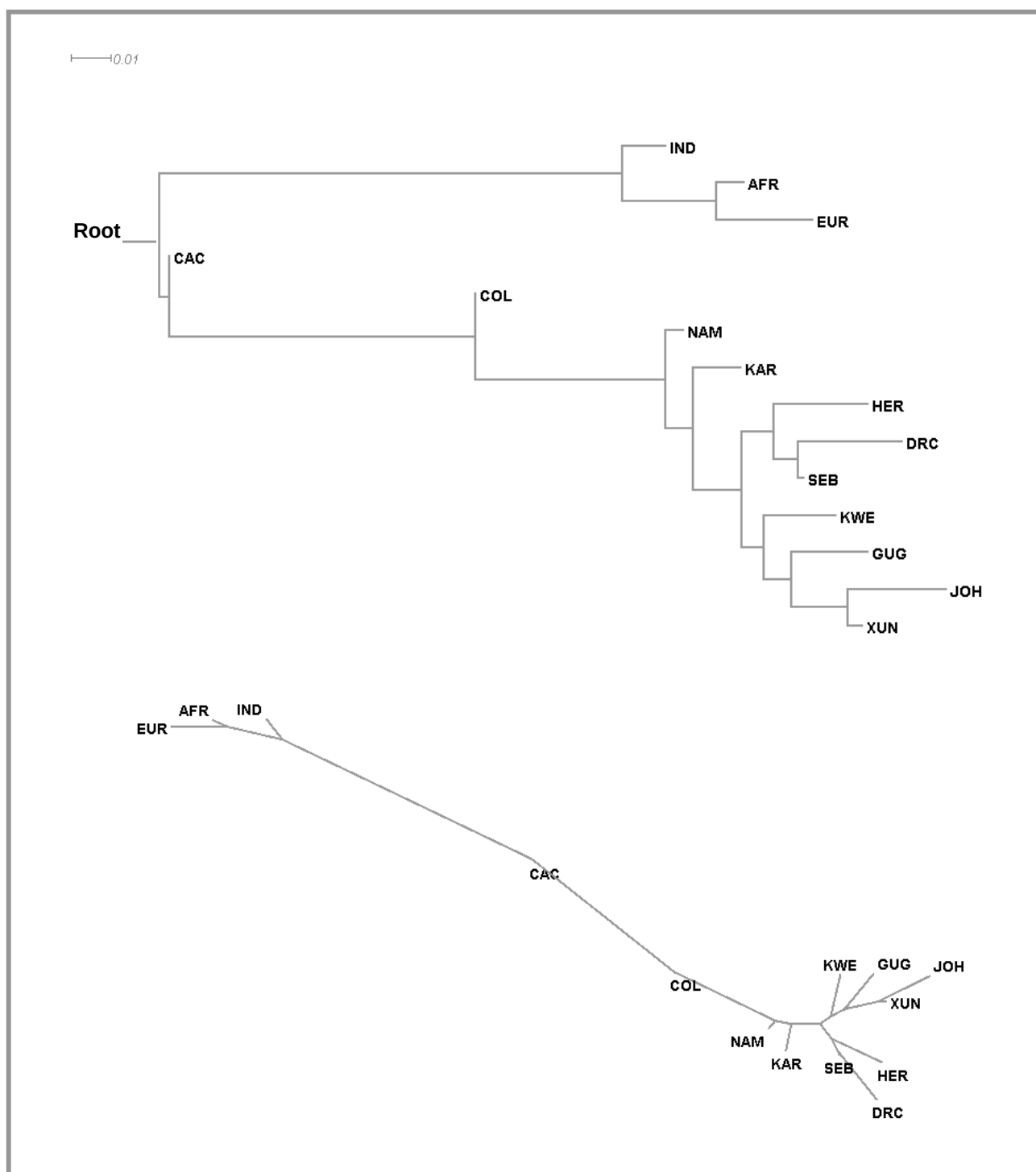


Figure 5.6b The consensus tree constructed from the average of 100 distance matrices. For the rectangular phylogram the root was placed between CAC and the non-African populations. The radial phylogram shows the unrooted version of the tree.

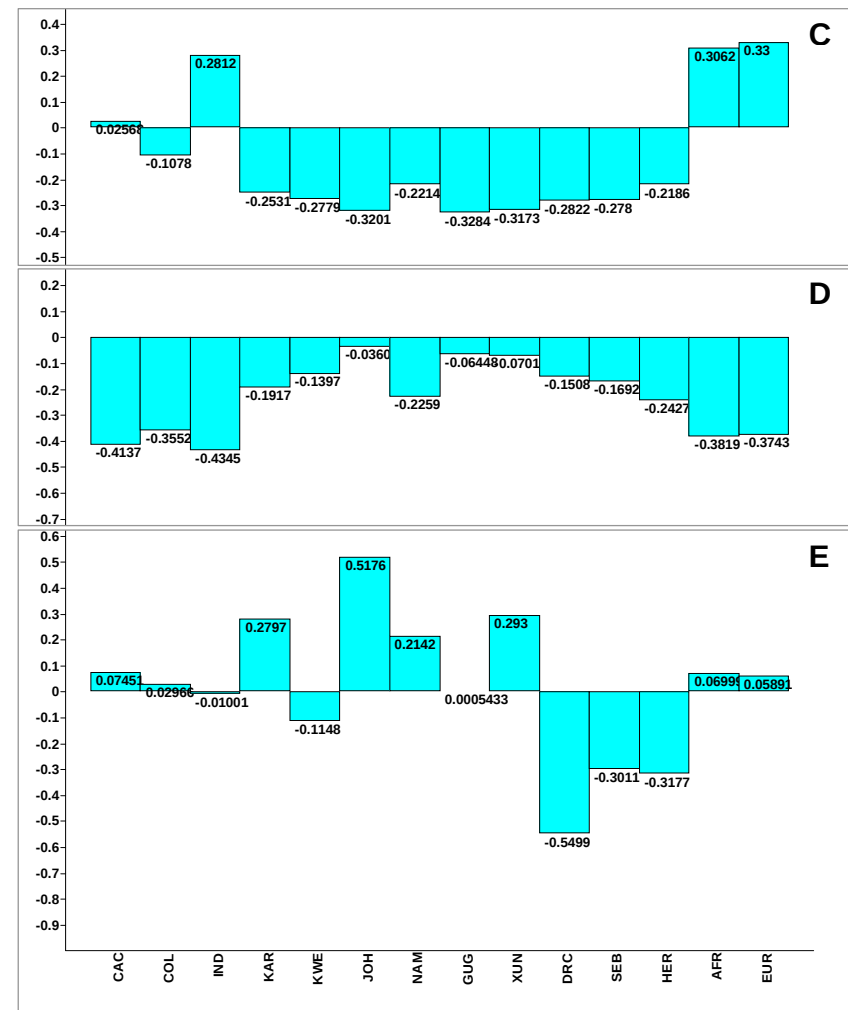
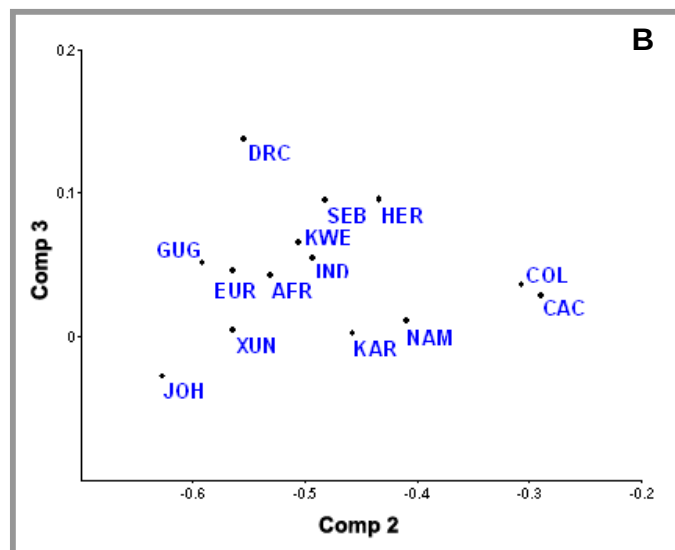
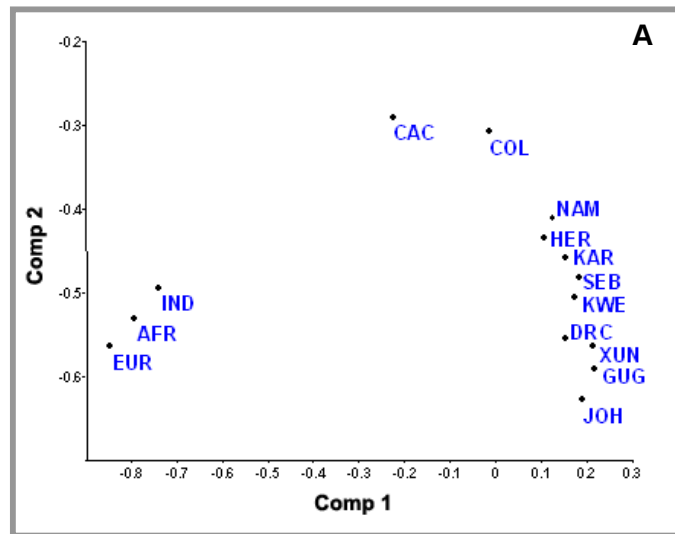


Figure 5.7 A and B – Principal component analysis of autosomal genotypic distances between different populations in the study group. Component 1 = 92.6% of the variation, Component 2 = 5.7% of the variation, Component 3 = 1.1% of the variation (Rest of the components < 0.16 each). **C** – Loadings for Component 1, **D** – loadings for Component 2, **E** – loadings for Component 3

As was observed in other studies (Li *et al.*, 2008; Tishkoff *et al.*, 2009) the first PCA component (93%) summarised the variation present between African and non-African populations (Figure 5.7). This axis separated the African from the non-African populations with varying degrees of admixture in the CAC, COL, SWB and NAM. The very high level of non-African admixture into the CAC group was seen in the first component (Figure 5.7c) and compared well to what was observed in the STRUCTURE result (Figure 5.3 and Table 5.2). Accordingly the other southern groups (COL, NAM, KAR) and the HER also showed high non-African contributions while the JOH, XUN and GUG showed the lowest levels (Figure 5.7c).

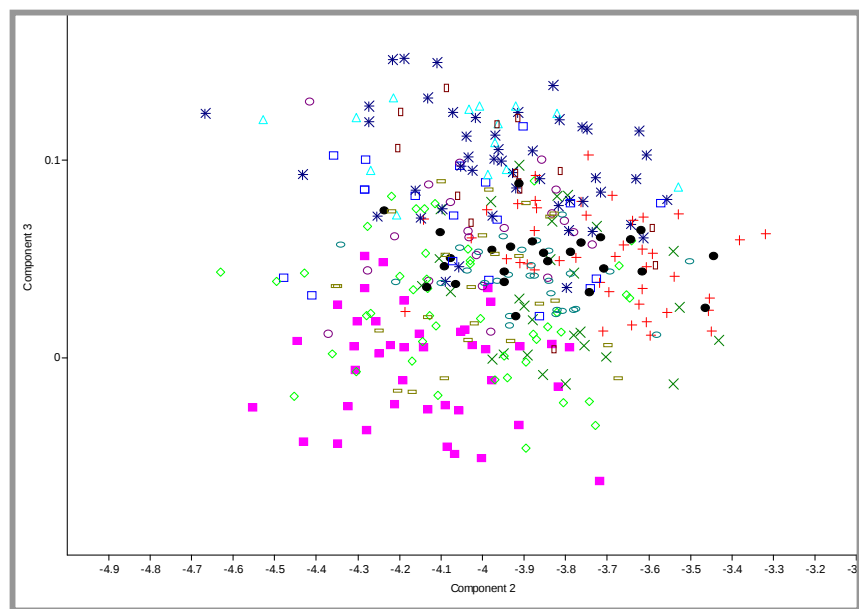
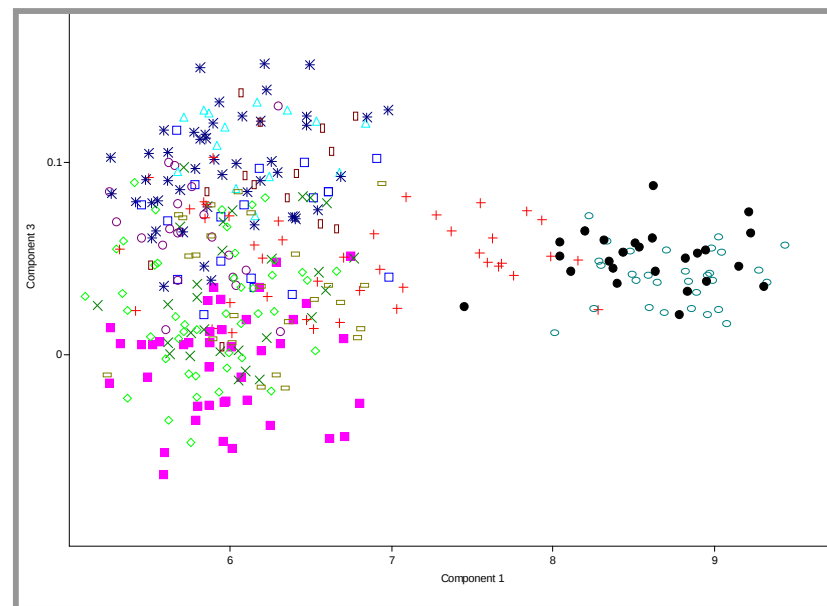
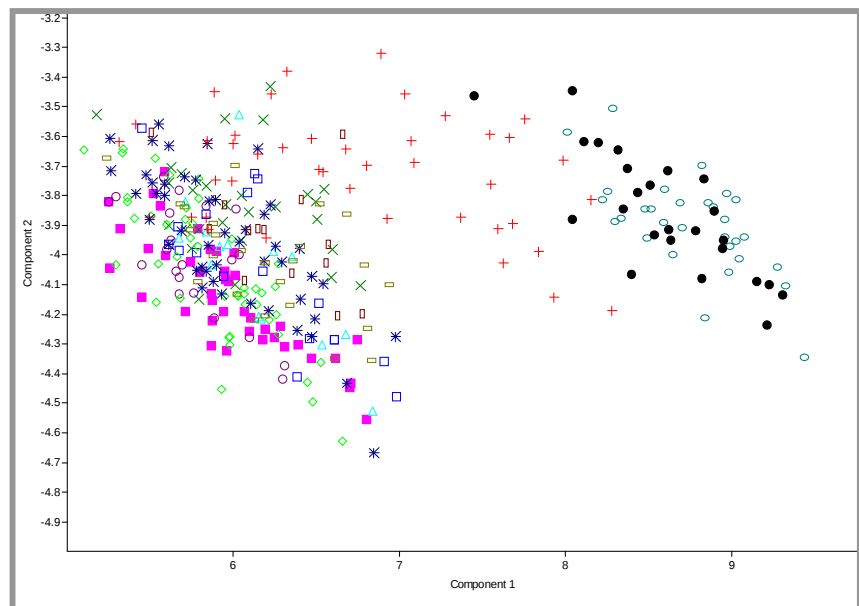
Interestingly, the second component (Figure 5.7) did not separate the Khoe-San groups from the Bantu-speaking groups as was expected based on the STRUCTURE results and seen in the mtDNA and Y-chromosome studies (Figure 3.16, 4.19 and 4.20). Rather the second component (5.7%) (Figure 5.7d) separated the northern San groups (JOH, GUG, XUN) from the southern Khoe-San and Coloured populations (COL, CAC, NAM).

It is only in the third component (1.1%) (Figure 5.7e) that the Bantu-speakers were separated from the San. This might indicate a very ancient split between the northern and the southern Khoe-San groups. On the extremities of the second component (Figure 5.7d) the northern groups (XUN, GUG, JOH) were at the one end and the southern groups (CAC, COL) at the other. NAM was also located with the southern groups but more towards the northern groups. KAR were placed intermediate between the groups. This is interesting since the historically the theory exists that the Karretjie (KAR) are descendant from the /Xam San group while the CAC, COL and NAM are expected to have more Khoe input (see section 1.1.1.5.3). The following hypothesis was formulated from this: The San groups formed an earlier continuum from the northern San groups in the north to the /Xam group in the south. The Khoe ancestral groups that contributed to the CAC, COL and NAM were originally occupying the southern parts of South Africa in the coastal regions where the Cape KhoeKhoe have lived. Upon acquiring the cultural practice of pastoralism from central groups such as the ancestors to the ‡Khomani, these southern groups expanded and moved northwards into the regions occupied by the other hunter-gatherers. To a degree they settled and intermixed with the local hunter-gatherer groups. Later, the Nama

group moved further northwards and had more recent gene-flow with the northern groups. This theory would, however, be very difficult to prove with the disappearance of the cultural identities and languages of the southern Khoe and San groups.

The third component (Figure 5.7e) represented the variation between Bantu-speakers and Khoe-San groups. At extreme ends of the continuum were the Bantu-speakers from the DRC with the lowest amount of Khoe-San input and the Ju\`hoansi (JOH) with the lowest amount of Bantu-speaking input. As was seen in the STRUCTURE results, the Herero (HER) and southeastern Bantu-speakers (SEB) had similar amounts of Khoe-San input (Table 5.2). For the rest of the groups the third component also reflected STRUCTURE results with the KWE showing more Bantu-speaking admixture than the other Khoe-San groups while the XUN and the KAR showed lower amounts than other groups (Figure 5.7e).

PCA plots of the separate individuals rather than the group was also constructed to see if the individual apportionment of variation corresponds to the composite apportionment (Figure 5.8).



◇	XUN
■	JOH
□	KWE
○	GUG
×	NAM
— □	KAR
+	COL
+	CAC
*	SEB
■	HER
△	DRC
○	AFR
○	EUR
●	IND

Figure 5.8 Principal component analysis of the average individual distance matrix. Component 1 = 59.1% of the variation, Component 2 = 13.1% of the variation, Component 3 = 2.3% of the variation. Individuals are colour coded according to the key.

For the individual PCA the apportionment of variation to three axes was not as good as in the group-wise comparison. This was expected as in the group-wise comparison there will be only 14 pairwise comparisons while in the individual comparisons it involves 352 pairwise comparisons. This increased the multi-dimensional space substantially making the reduction into three dimensions more difficult. Overall, the apportionment of variation compared well to the group based analysis. The axis that includes most of the variation (59%) was again the axis that separates African from non-African populations. Similar results were seen for the 650K SNP based study of worldwide variation (Li *et al.*, 2008), where the first component (separating African and non-African variation) comprised 56% of variation. In the microsatellite and insertion/deletion based study of worldwide variation (Tishkoff *et al.*, 2009) the first component only represented 19.5%. The lower representation in the first component of the microsatellite versus SNP studies can be explained by the higher mutation rate of microsatellites, which would lead to convergence of distantly related populations.

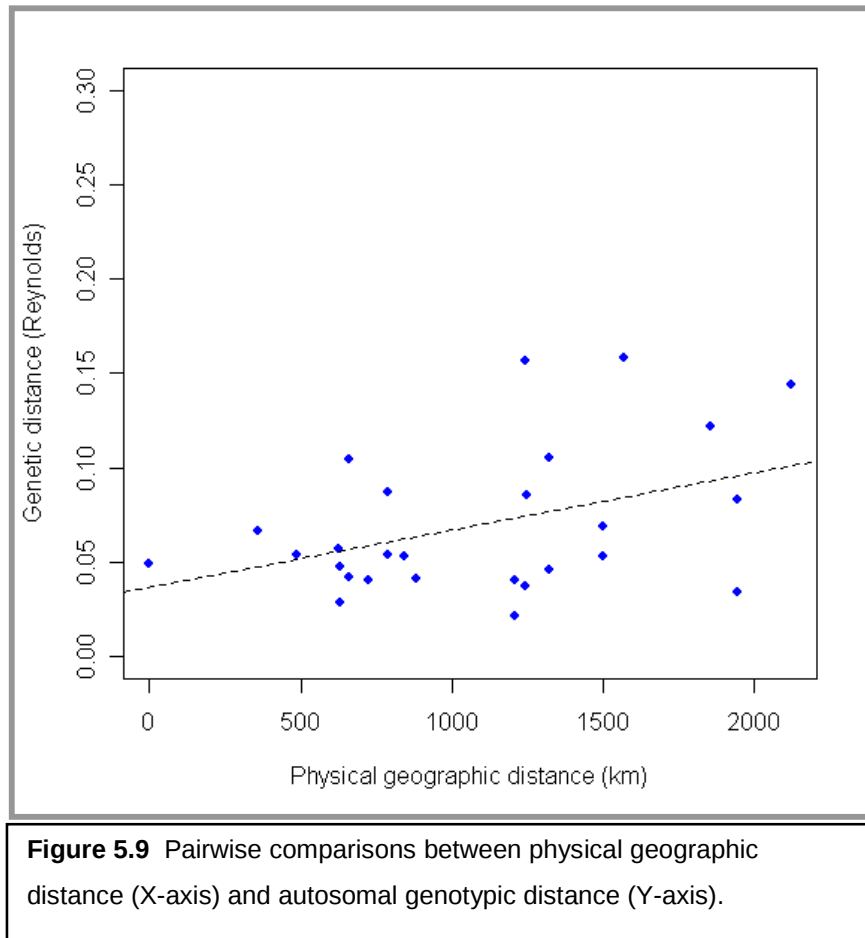
Using the PCA plots of individuals one gets a clearer picture of how individual variation is apportioned and how closely individuals from the same group cluster together. While the first component clearly separated the African from the non-African variation with the Coloured groups in-between, the components representing variation within Africa were more continuous (Figure 5.8). Similar to group based results, the second component for the individual's data (13.2%) separated northern and southern Khoe-San and Coloured groups while the third component (2.3%) separated the Bantu-speakers from the Khoe-San (Figure 5.8). Although one can infer that the second component contained the variation between northern and southern Khoe-San groups, the change was very continuous and more individuals were scattered than observed for the third component. The third component therefore showed a better clustering and separation of individuals from the two different groupings (Khoe-San and Bantu-speakers). Thus even though the second component contained more variation than the third component the second component showed more of a continuum. On this axis there was a gradual decrease in northern San individuals aligned with a gradual increase of southern Coloured and Khoe individuals. This indicated more of a clinal difference between northern and southern Khoe-San groups while the difference between Khoe-San and Bantu-speakers was more abrupt. The above

explanation might be a part of the reason why STRUCTURE did not assign this second variation component as a separate cluster but assigned a cluster for the third component.

To investigate the relationship between the physical geographic distance (km) and genetic distance using autosomal SNPs in the Khoe-San and Coloured populations, the composite distance matrix of the 100 datasets (Reynolds distance) was compared to a physical distance matrix (Appendix C). In Figure 5.9 pairwise comparisons between physical geographic distance (X-axis) and genetic distance Y-axis is plotted on graphs. A linear regression was done to determine the line with the best fit through the points.

The best fit to the points on the graph was a straight line with a slope of 0.00003057 ($p = 0.0258$) (Figure 5.9). A Mantel test also found a correlation between the two distances ($r = 0.421$) that were significantly different from correlation between random datasets generated through permutation tests ($p = 0.0248$). The physical distance explained 17.7% of the genetic distance.

The clinal distribution of genetic variation of northern versus southern Khoe-San groups was thus also illustrated by the correlation of autosomal genetic and physical distance (Figure 5.9). The correlation coefficient between physical distance and genetic distance for unlinked autosomal SNPs was slightly higher than the correlation found for the mtDNA variation ($r = 0.402750$, $p = 0.027$, see section 3.4), while the Y-chromosome studies indicated no significant correlation between physical and genetic distance for either Rst or Fst datasets. For the mtDNA analysis, however, the non-African and Bantu-speaking haplotypes were removed from groups, while this was not possible for the genotypic data. It is therefore expected that the genotypic correlation of physical versus genetic distance would be influenced by the larger input from non-African groups into the southern Khoe-San and Coloured groups.



5.1.5 AMOVA analysis

To test the apportionment of variation at different levels of grouping, AMOVA analysis was used. The degree of variation was tested firstly between defined groups, secondly between the different populations in the study and thirdly within the populations. For the 100 random unlinked autosomal datasets generated, 10 datasets were randomly picked and AMOVA analysis performed on them. Table 5.5 gives the average results of the 10 sets of AMOVA analysis with various different groupings of the first level group.

Table 5.5 Results from autosomal genotypic AMOVA analysis using different groupings on the first level

Grouping	Grouping of first level [Groups]	Between groups	Between populations within groups	Between individuals within populations
A	[afe, ind] [col, kar, cac, nam, joh, xun, gug, kwe, drc, her, seb]	21.00	3.41	75.59
B	[afr, eur, ind] [col, kar, cac, nam, joh, xun, gug, kwe] [drc, her, seb]	12.86	3.01	84.13
C	[afr, eur, ind] [col, kar, cac, nam] [drc, her, seb] [gug, joh, xun, kwe]	11.05	2.07	86.88
D	[col, kar, cac, nam, joh, xun, gug, kwe] [drc, her, seb, afr, eur, ind]	3.87	8.88	87.25
E	[col, kar, cac, nam, joh, xun, gug, kwe] [drc, her, seb]	2.22	3.53	94.25
F	[col, kar, cac, nam] [joh, xun, gug, kwe]	2.73	2.85	94.42
G	[col, kar, cac] [nam, joh, xun, gug, kho]	3.10	3.03	93.87
H	[col, kar, cac] [nam] [joh, xun, gug, kho]	2.32	3.08	94.60

Most of the variation between groups (21%) were explained between African and non-African groups (Grouping A - Table 5.5). When the African groups were split into Bantu-speakers and Khoe-San+Coloured the variation contained by the first level grouping falls to 13% (Grouping B - Table 5.5). The variation on the first level grouping only decreased slightly when the northern and southern Khoe-San+Coloured groups were separated (Grouping C – Table 5.5). Variation between Bantu-speakers and Khoe-San+Coloured groups was only 2.2% (Grouping E – Table 5.5) (when the KWE group was omitted it increased to 2.4% - data not shown). The variation between the southern Khoe-San+Coloured groups and northern Khoe-San groups (2.7%) (Grouping F – Table 5.5) was higher than the variation between Khoe-San+Coloured and Bantu-speaking (Grouping E – Table 5.5) (when the KWE group was omitted it increased to 3%-data not shown). This was in support to findings from the PCA. It was only in the cases when non-African groups were

included, however, that the group-based variation was more than the variation between individual populations (Grouping A, B, C – Table 5.5).

The last two rows in the table shows the classification used today, namely, the division between Coloured and Khoe-San (Grouping G – Table 5.5) and the division between Khoe, San and Coloured (Grouping H – Table 5.5). In both cases the variation between individual populations was almost equal or greater than the variation between groups.

5.2 Results and discussion (Haplotypes)

For haplotype analysis of autosomal data, five linked SNPs on the same haploblock was used to infer 44 short haplotypes consisting of 5 bp each as described in section 2.4. The haplotypes were inferred separately for each population and each SNP set of 5. The full list of inferred haplotypes and their frequencies in the different populations is available in Supplementary Electronic Data – File C.

5.2.1 Inferred haplotypes

The inferred haplotypes for the 44 different loci yielded different results. There were differences in the number of haplotypes per locus, population frequencies and structuring between different populations. A selection of eight haplotype loci with their inferred haplotypes and their frequencies in each of the 14 populations is shown in Figure 5.10. The full set of bar charts of all 44 loci is included in Appendix H.

The number of haplotypes per loci varied from five haplotypes (04-01 in Figure 5.10) to 29 (14-01 in Figure 5.10). In most of the loci, a clear difference in population frequencies could be seen while only few loci failed to show structuring (19-02 in Figure 5.10). The frequencies of representing haplotypes between the African and non-African populations differed in most loci. The non-African populations tended to have smaller subsets of the African haplotypes but one or two haplotypes were predominant in frequency. Some haplotypes showed clear differences between BS and Khoe-San populations (e.g. purple in 13-02, pink in 01-02, yellow in 05-01 in Figure 5.10). These differences, however, were not as pronounced as the differences between African and non-African populations.

The frequency distributions of inferred haplotypes thus clearly illustrated higher African haplotype diversities and that non-African variation represents a subset of African variation (Figure 5.10). This finding corroborate other studies confirming the out of Africa hypothesis (Bowcock *et al.*, 1987; Nei and Livshits, 1989; Bowcock *et al.*, 1991a; Bowcock *et al.*, 1991b; Bowcock *et al.*, 1994; Tishkoff *et al.*, 2009).

5.2.2 Distance analysis

To consolidate the information across the 44 separate haplotype loci, the 88 haplotypes generated for each individual were concatenated into two haplotypes per individual. Individuals with >50% missing data at any locus were excluded from further analysis. Following removal of missing data, 298 individuals were retained in the data yielding 596 haplotypes. The individual haplotypes were then used to construct distance matrices. Both population and individual distance matrices were constructed (the population distance matrix are shown in Table 5.6 and the individual distance matrix is included in Supplementary Electronic Data File D). These distance matrices were then used for PCA for the population matrix (Figure 5.11) and the individual matrix (Figure 5.12).

The concatenation of the different short haplotypes into one long haplotype resulted in high diversities between the individual haplotypes. Since some of the loci were very polymorphic and contained many different haplotypes, the combination of several such loci led to high haplotype diversities. By concatenating haplotypes in individuals led to 594 unique haplotypes in the total of 596 haplotypes.

Table 5.6 Maximum composite likelihood population distances of individual haplotypes

	AFR	CAC	COL	DRC	EUR	GUG	HER	IND	JOH	KAR	KWE	NAM	SEB	XUN
AFR	0.000													
CAC	0.599	0.000												
COL	0.737	0.764	0.000											
DRC	1.095	0.927	0.845	0.000										
EUR	0.377	0.612	0.719	1.044	0.000									
GUG	1.094	0.893	0.786	0.745	1.084	0.000								
HER	0.932	0.851	0.781	0.765	0.912	0.726	0.000							
IND	0.386	0.597	0.709	1.031	0.405	1.048	0.880	0.000						
JOH	1.014	0.828	0.762	0.768	0.973	0.634	0.730	0.969	0.000					
KAR	0.904	0.821	0.748	0.779	0.870	0.685	0.746	0.886	0.631	0.000				
KWE	1.005	0.867	0.792	0.753	0.987	0.676	0.735	0.952	0.678	0.715	0.000			
NAM	0.861	0.818	0.772	0.819	0.832	0.729	0.781	0.845	0.688	0.704	0.759	0.000		
SEB	1.032	0.890	0.817	0.759	0.991	0.716	0.756	0.986	0.726	0.744	0.749	0.786	0.000	
XUN	1.035	0.868	0.780	0.761	1.000	0.648	0.731	1.011	0.600	0.653	0.700	0.705	0.733	0.000

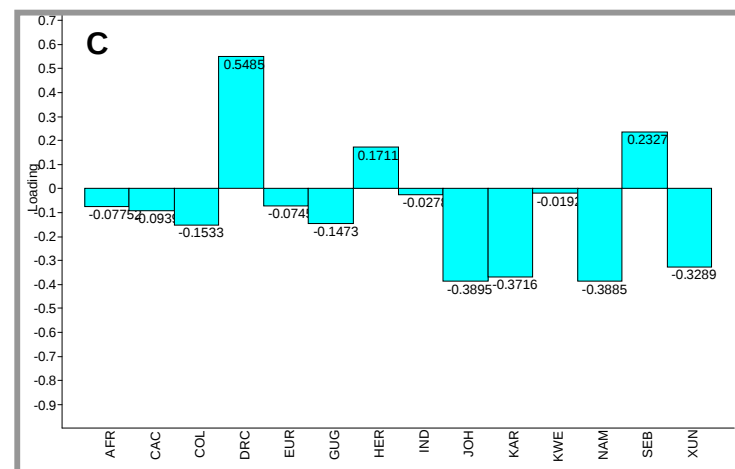
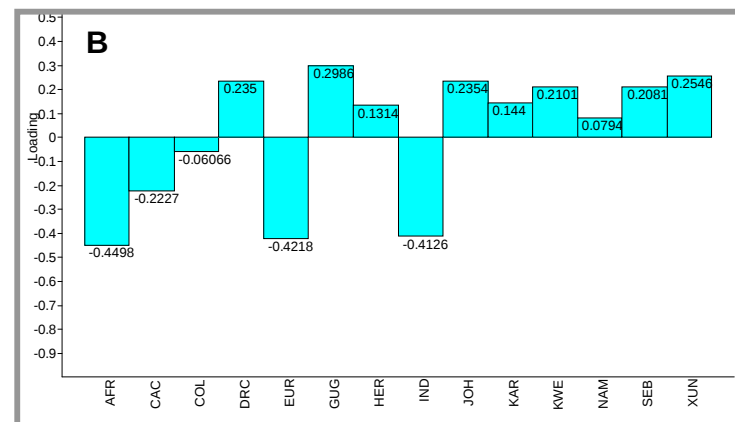
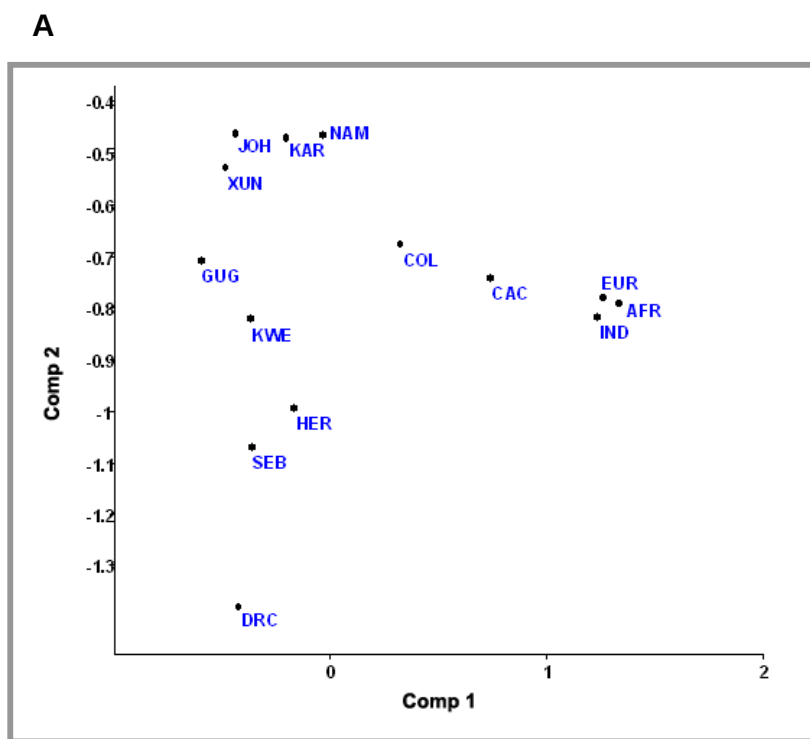


Figure 5.11 A – Principal Component Analysis of autosomal haplotype distance values between different populations in the study group. Component 1 = 54.6% of the variation, Component 2 = 7.1% of the variation. **B** – Loadings for Component 1, **C** – loadings for Component 2.

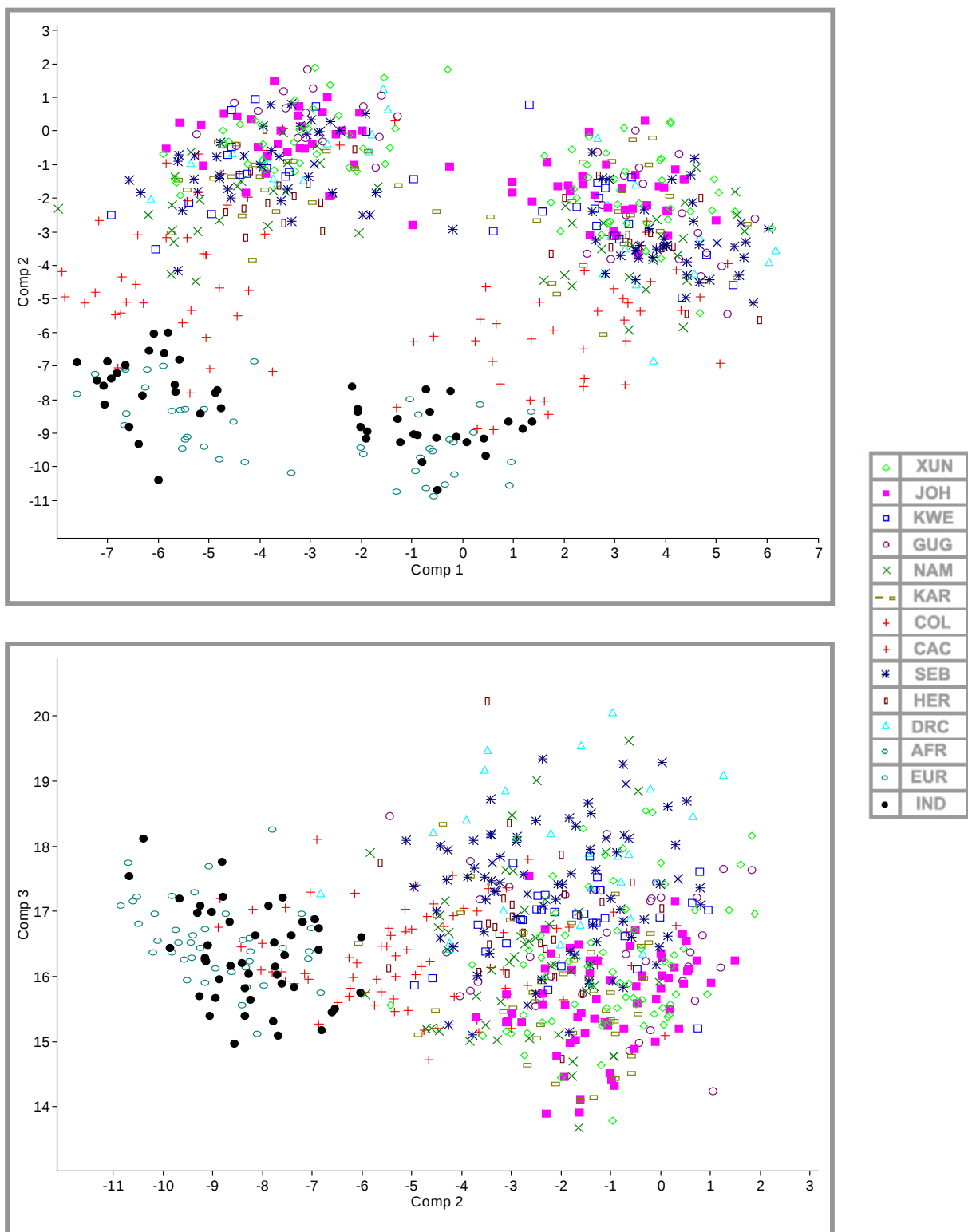


Figure 5.12 Principal Component Analysis of autosomal haplotype distance values between different individuals in the study group. Component 1 = 47.85% of the variation, Component 2 = 27.21% of the variation, Component 3 = 3.5% of the variation (remaining components contains < 1.4% of the variation each)

Group based PCA based on haplotypes again assigned the largest part of variation (PC1=55%) to the African – non-African division (Figure 5.11). Similar to genotypic results the first component illustrated more non-African admixture into the Coloured groups and a relative small non-African component in the northern San groups (Figure 5.11b). Contrary to the genotypic PCA, the PCA based on haplotypes assigned the second component (7%) to the division between Khoe-San groups and Bantu-speakers and not between northern and southern Khoe-San groups. The JOH, NAM, KAR and XUN were separated from the SEB, DRC and HER with the other groups located in between. The remaining components were not informative and also did not differentiate between northern and southern Khoe-San groups (component 3 to 6 contained between 3% and 4% variation, component 7 to 9 contained 2-3% variation and component 11-13, 1 to 2% variation).

Similar results were obtained through the individual based PCA (Figure 5.12). The first axis of the PCA plot, however, separated the different inferred haplotypes in each individual from one another. This was however an artefact resulting from the methodology employed when haplotypes were inferred and thereafter concatenated into one haplotype. When haplotypes were inferred the short haplotypes of 5 bp were sorted alphabetically in each individual by the software program used. For instance; individual 1 would have two haplotypes at a certain locus that would be sorted first AACCC and then AAGCC, for individual 2 it, the haplotypes would be sorted AACCT and then AAGCC. Thus haplotype 1 of individual 1 and 2 and haplotype 2 of individual 1 and 2 would group together. This bias was then reflected when haplotypes are concatenated. In a population comparison this effect would be neutralized.

Thus in the PCA plot in Figure 5.12 only the second and third axis contained useful information. On the second axis non-African individuals were separated from African individuals with CAC and COL in-between. The third axis separated the BS and the Khoe-San individuals. The rest of the axes contained little variation with each representing <1.4% of the variation. The individual PCA plot was useful to observe the variation in each individual. For populations such as GUG, the population as a whole did not associate that strongly with the other Khoe-San groups, however, there were specific GUG individuals that did group with the Khoe-San individuals in the individual PCA plot.

To explain the difference between the genotype and haplotype based PCA the following hypothesis is proposed. It might be that there was more continuous gene-flow over thousands of years between the different Khoe-San groups leading to a clinal distribution of genetic variation with a distance based trend. On the contrary, the Bantu-speaking and Khoe-San gene-pools were isolated for many years before recent admixture. The older continuous gene flow between Khoe-San groups may have broken up many more haplotypes than the recent admixture by between Bantu-speakers and Khoe-San. Thus by inferring and concatenating haplotypes the genotypic signature of the distance based cline between Khoe-San groups were erased. Conversely, many of the haplotypes remained intact when comparing haplotypic variation between Bantu-speakers and Khoe-San.

To alleviate the problem the most common haplotype for each population were selected and used as a population representative haplotype. In this approach only the haplotypes with the highest frequency in each specific population at each of the 44 loci were selected. This was then taken as the 44 representing short haplotypes from each population. The 44 representative short haplotypes for each population were then concatenated into one sequence (long haplotype) for each population. These 14 population representative sequences were then used to construct a distance matrix (Table 5.7). The distance matrix was used to do PCA (Figure 5.13) and cluster analysis (Figure 5.14). This will partially overcome the effect of recent admixture between the groups and level out the difference between the recent and ancient admixture. When this was done, a signature of divergence between the northern and southern Khoe-San groups again emerged (Figure 5.13).

Table 5.7 Maximum composite likelihood population distances of population representative haplotypes

	AFR	CAC	COL	DRC	EUR	GUG	HER	IND	JOH	KAR	KWE	NAM	SEB	XUN
AFR	0.000													
CAC	0.132	0.000												
COL	0.454	0.317	0.000											
DRC	1.268	1.152	0.750	0.000										
EUR	0.093	0.184	0.419	1.626	0.000									
GUG	1.323	1.230	0.790	0.402	1.706	0.000								
HER	1.067	0.940	0.606	0.209	1.185	0.437	0.000							
IND	0.070	0.108	0.454	1.455	0.071	1.421	0.986	0.000						
JOH	1.269	0.963	0.700	0.538	1.239	0.318	0.517	1.202	0.000					
KAR	0.845	0.770	0.427	0.560	1.004	0.277	0.535	0.949	0.323	0.000				
KWE	1.384	1.091	0.814	0.392	1.574	0.262	0.393	1.293	0.430	0.517	0.000			
NAM	0.741	0.712	0.427	0.542	0.781	0.361	0.517	0.858	0.366	0.195	0.440	0.000		
SEB	1.177	1.093	0.721	0.183	1.490	0.235	0.258	1.264	0.397	0.488	0.247	0.463	0.000	
XUN	1.322	1.108	0.756	0.442	1.322	0.247	0.375	1.257	0.205	0.232	0.329	0.320	0.290	0.000

PCA again dedicated the first component to non-African versus African variation (91%) and although both the second and third components were dedicated to Khoe-San versus Bantu-speaking variation, a differentiation between northern and southern Khoe-San groups could be made (Figure 5.13). While the second component (5%) maximized the variation component between Bantu-speakers and southern Khoe-San groups (KAR, NAM, COL) the third component (2%) maximized the variation component between Bantu-speakers and northern Khoe-San groups (XUN, JOH) (Figure 5.13). This indicated that the southern and northern Khoe-San groups' genetic variation were different and had to be optimized against Bantu-speaking variation in different dimensions. Other components contained less than 0.7% of the variation each.

The cluster analysis (Figure 5.14) reflected results from PCA plots. Non-African populations were separated from African populations and the non-African admixture in the CAC and COL caused them to group with the non-African group. The three BS groups grouped together and the KWE and GUG grouped together on an adjacent branch. The KAR grouped with the Khoe-San groups (NAM, JOH, XUN) on one branch. Furthermore the two northern San groups JOH and XUN grouped together while the two southern groups (KAR and NAM) grouped together.

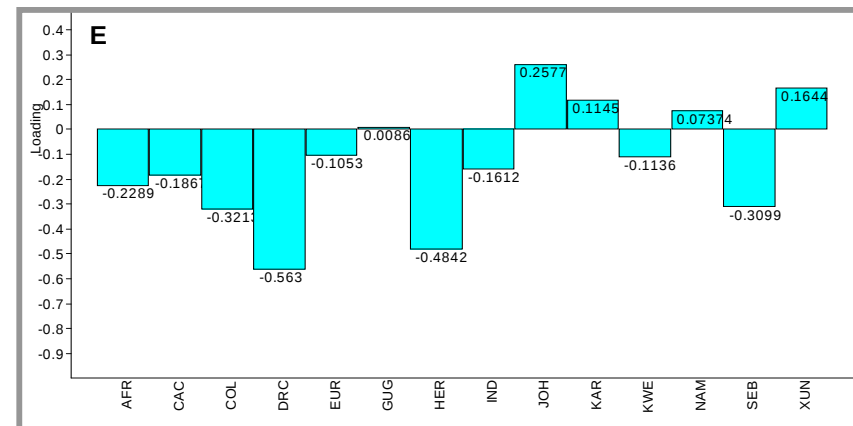
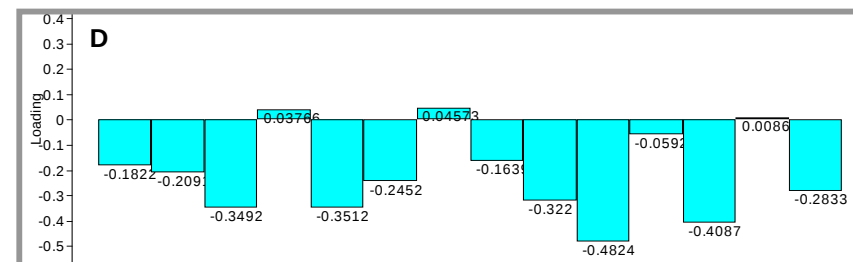
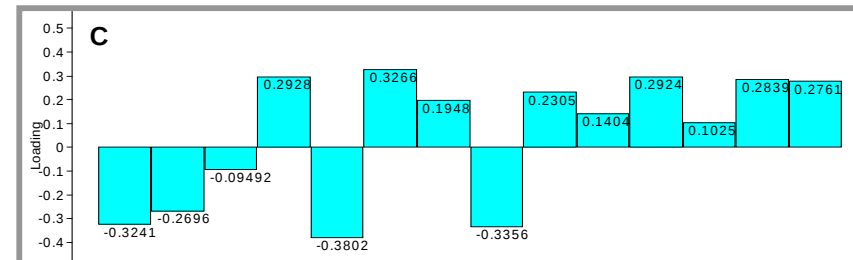
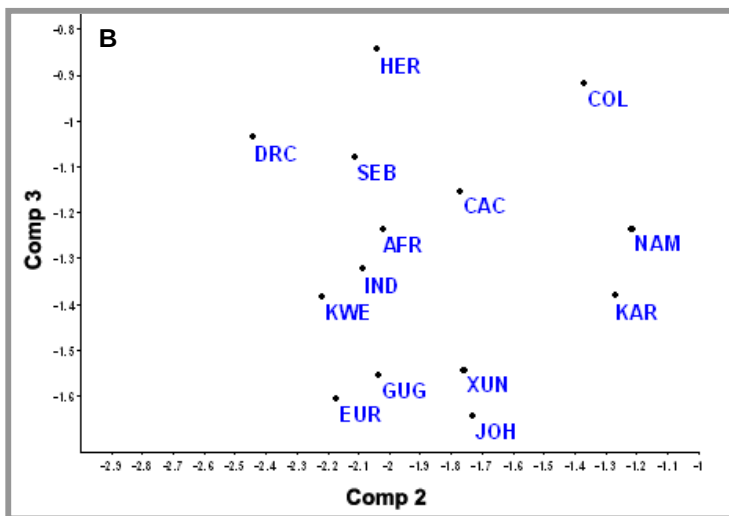
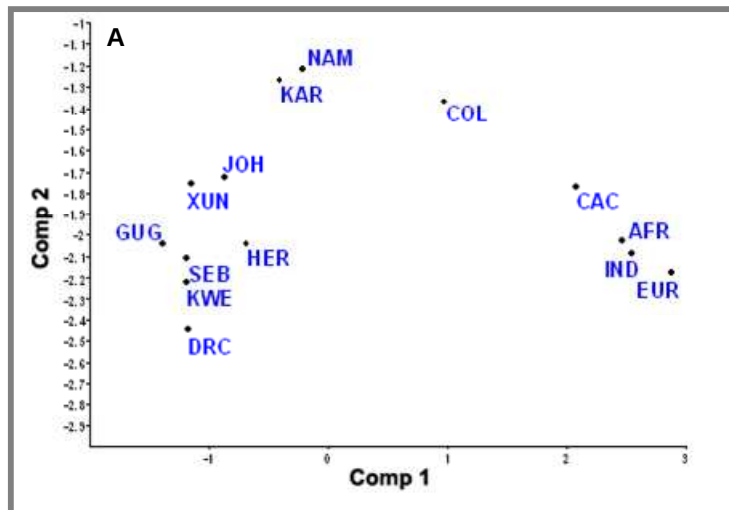


Figure 5.13 A and B – Principal Component Analysis of autosomal representative haplotype distance values between different populations in the study group. Component 1 = 90.9% of the variation, Component 2 = 4.8% of the variation, Component 3 = 2.2% of the variation (remaining components contains < 1.4% of the variation each). **C** – Loadings for Component 1, **D** – loadings for Component 2, **E** – loadings for Component 3

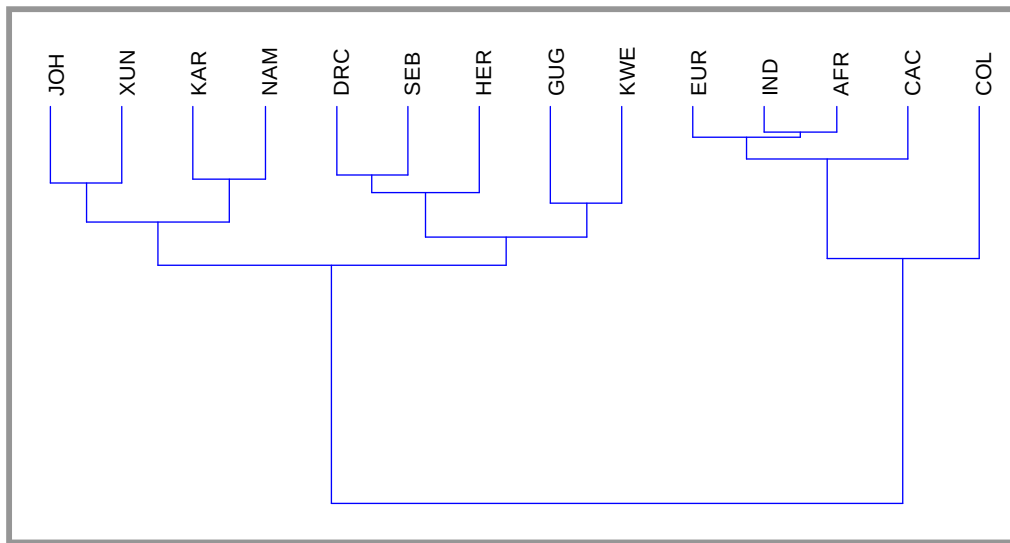


Figure 5.14 Cluster analysis tree illustrating autosomal representative haplotype distance values between different populations in the study group.

The 44 separate haplotypes that were concatenated into one haplotype would have different evolutionary histories, and a single unique tree would not best characterize the phylogenetic representation of the haplotype. To overcome this problem an approach was followed where the data was not forced into a single tree, rather a Neighbour-Net splits decomposition tree was compiled (Figure 5.15). This method gave a good indication of how tree-like the dataset was.

The splits decomposition network clearly showed that there were several trees that explained the relationships between the representative composite haplotypes of the different populations. Although, if only trees were used, which have 95% confidence, the network was reduced to only few reticulations, mainly at the base of the branches supporting BS groups. The African and non-African variation was the most pronounced with the admixed Coloured groups in-between. Furthermore the Bantu-speakers grouped together and the KWE and GUG grouped with them because of the high amounts of admixture. For the remaining Khoe-San groups, the XUN and JOH grouped together, while the NAM and KAR were located more towards the non-African side of the network due to the higher amount of admixture. There were, however, evidence in the reticulations that there were trees that group the NAM and KAR together and also the GUG and KWE with the JOH and XUN.

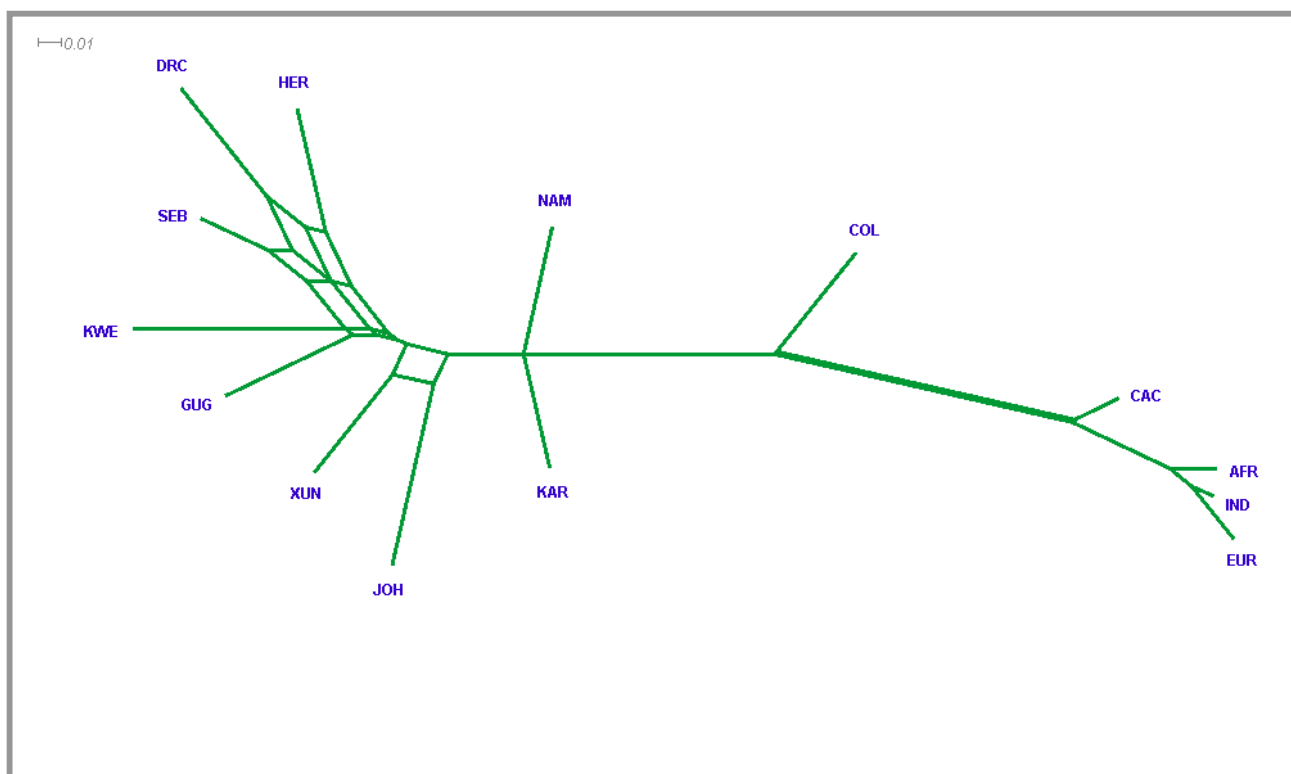
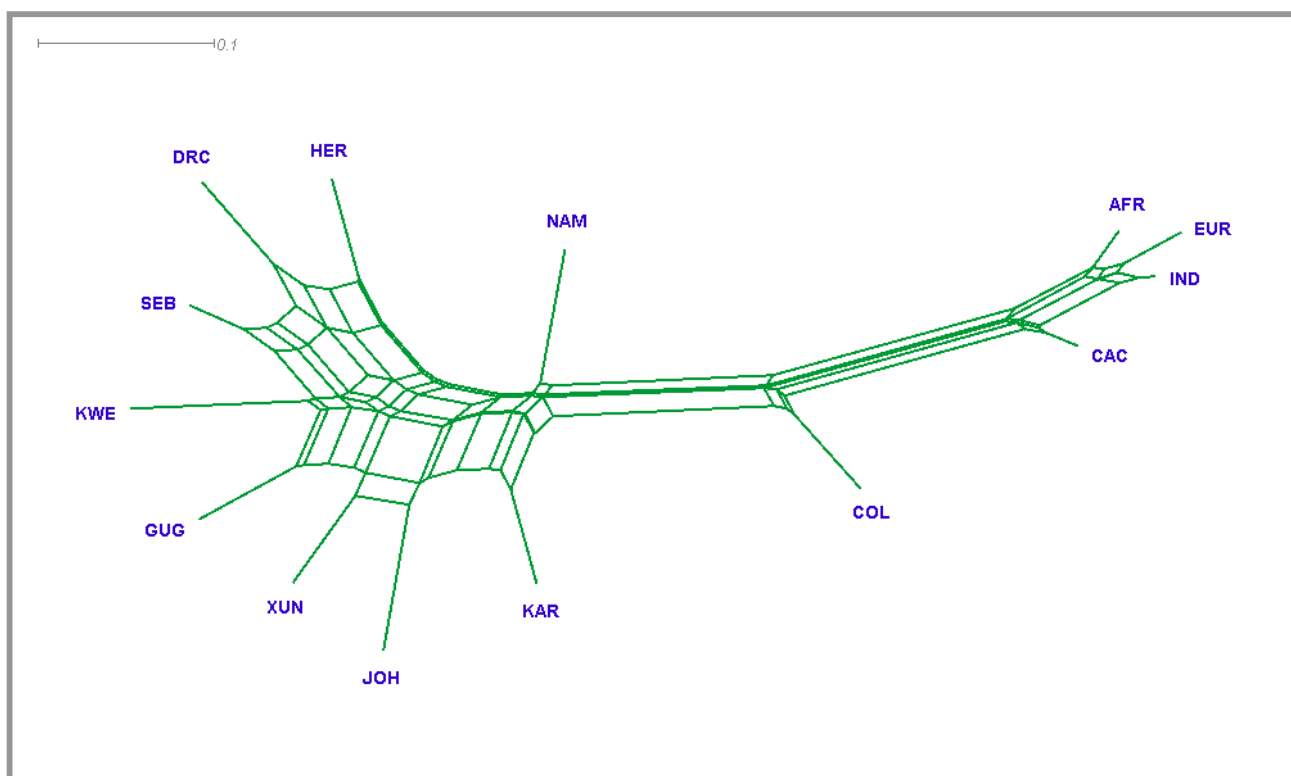


Figure 5.15 A - Splits decomposition network showing the different trees that explain the relationships between the representative composite haplotypes of the different populations.

B - Network resulting when using only trees with 95% confidence.

The relationship of physical distance and genetic distance using the autosomal inferred haplotypes in the Khoe-San and Coloured populations was tested by comparing the genetic distance matrices of both approaches described above to a physical geographic distance matrix. Pairwise comparisons between physical geographic distance (X-axis) and genetic distance based on the individual inferred haplotypes (Figure 5.16 - A) and genetic distance based on the top frequency population representative haplotypes (Figure 5.16 - B) on the Y-axis were plotted on graphs. A linear regression was performed to determine the line with the best fit through the points on both plots.

The best fit to the points on the individual haplotypes graph was a straight line with a slope of 0.000057 ($p = 0.0404$) (Figure 6.8 - A). A Mantel test also found a correlation between the physical and individual based genetic distance ($r = 0.390$) that were significantly different from correlation between random datasets generated through permutation tests ($p = 0.0215$). The physical distance explained 15.2% of the genetic distance.

For the top frequency haplotypes the line that fits the points best was a straight line with a slope of 0.00029 ($p = 0.0072$). The Mantel test also found a significant correlation ($p = 0.0124$) between the two distance matrices ($r = 0.497$). In this case the physical distance explained 24.7 % of the genetic distance

A reduction of the distance based cline seen in the individual based haplotypes versus the population representative haplotypes could also be seen in the comparison of physical versus genetic distance (Figure 5.16) and Mantel test results. There was a stronger correlation between genetic and physical distance (indicating a distance-based cline) in the population representative haplotypes than the individual haplotypes. This is, as explained previously, because the most common haplotypes in a population will not be affected as much by recent admixture.

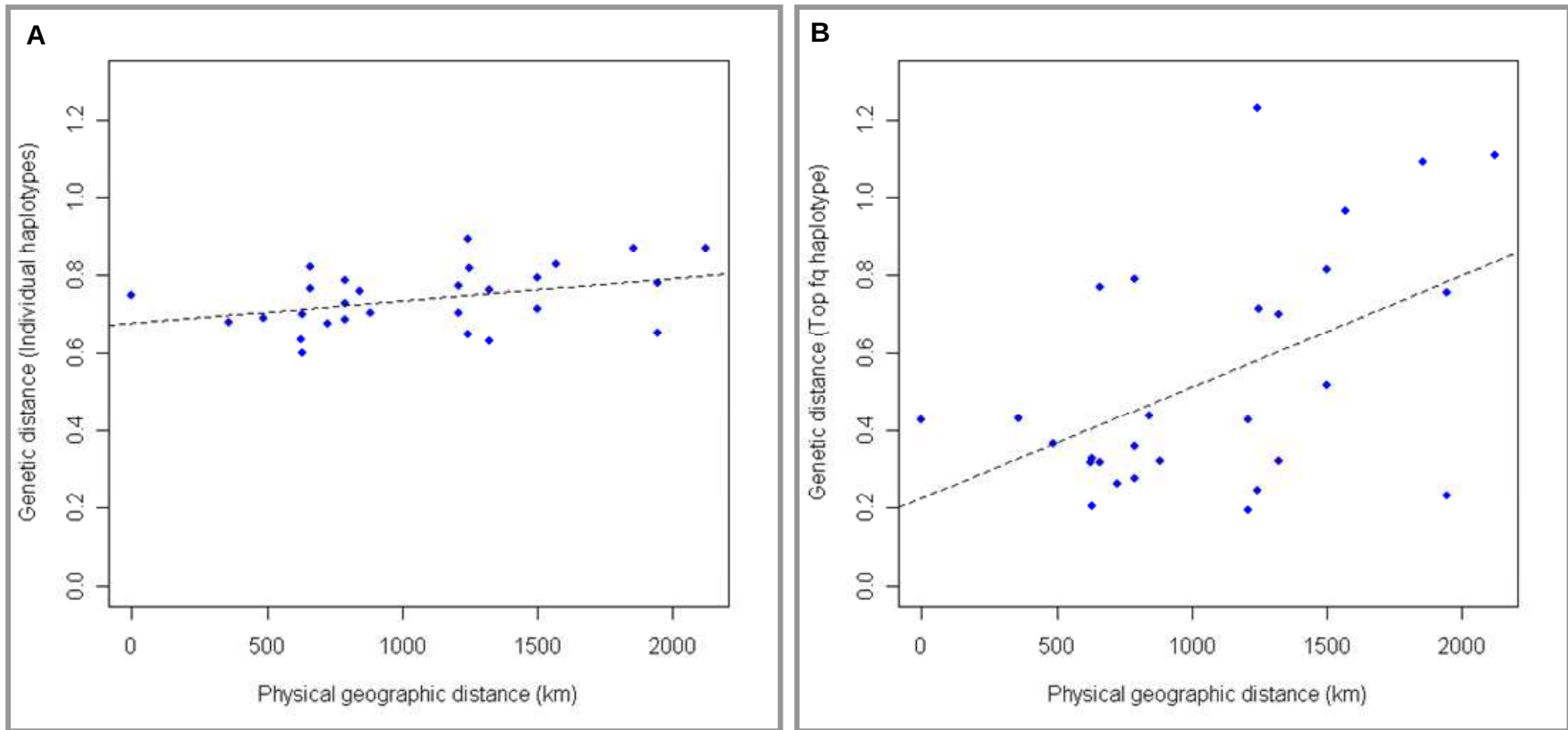


Figure 5.16 Pairwise comparisons between physical geographic distance (X-axis) and autosomal haplotype genetic distance (Y-axis).

A – Using individual haplotypes in genetic distance.

B – Using top frequencies representative haplotypes.

5.3 Summary of autosomal results

STRUCTURE results illustrated different amounts of non-African and Bantu-speaking admixture into the various Khoe-San and Coloured populations. Results supported low levels of contribution from non-Africans to the northern San populations (Ju\’hoansi, !Xun, |Gui + ||Gana, Khwe) and Bantu-speakers. Conversely, the southern Khoe-San and Coloured groups showed evidence of higher non-African admixture. This is consistent with history and with mitochondrial and Y-chromosome results. Furthermore, the southern Bantu-speakers had higher Khoe-San admixture compared to the central African Bantu-speakers, which illustrates the geneflow from resident San groups into the Bantu-speakers when they moved into southern Africa. Excluding the Ju\’hoansi, asymmetric geneflow between the Bantu-speakers and Khoe-San groups were observed with more gene-flow from the Bantu-speakers into the Khoe-San than vice-versa. In support of previous findings based on the classical blood group markers, the Khwe had the highest Bantu-speaker admixture of all the Khoe-San groups. Yet, the Khwe showed a much larger contribution from the Khoe-San compared to the Khoe-San component seen in Bantu-speakers, indicating again that the Khwe is not merely a Bantu-speaking group that adopted the hunter-gatherer lifestyle and a Khoisan language.

While STRUCTURE results could not illustrate a divide between northern and southern Khoe-San groups, the ability of PCA to optimize and reduce different distance components to minimal dimensions illustrated a north-south divide in the Khoe-San. For the genotypic analysis the component summarising the distance between northern and southern groups was in fact larger than the distance component summarising the variation between the Khoe-San and Bantu-speakers. When using population representing haplotypes in the haplotype analyses, the variation between BS vs. northern Khoe-San and BS vs. southern Khoe-San were optimised on different components; confirming the north-south differentiation within the Khoe-San. This north-south divide was also illustrated by the strong association that exist between geographic distance and genetic distance in both the genotypic and haplotypic analyses.

6. GENERAL DISCUSSION

Having had the opportunity to examine three different types of data (mtDNA, Y-chromosome DNA and autosomal DNA) in some Khoe-San and Coloured populations from southern Africa in conjunction with other sub-Saharan African populations using various analytical methods, it is now possible to address some of the specific objectives raised in section 1.3.

The genetic affinities within and between Khoe-San

Linguistic groupings have been used widely to classify the different Khoe-San groups. These studies have suggested that the the Ju, Tuu and Khoe speakers ought to be assigned to three different linguistic families (Table 1.1). These linguistic families are either unrelated or have genealogical relationships that can be traced back in excess of 10 000 years (Güldemann, In Press). The question arising from these linguistic assignments is whether these observations could be corroborated from genetic data. Serogenetic studies conducted by Jenkins and colleagues (Jenkins *et al.*, 1971; Jenkins, 1986) did not find unambiguous correlations between linguistic groupings and genetic clusters (Figure 1.3).

Genetic studies to date that have included “Khoisan” groups have been based on a few groups, notably, the two Ju-speaking groups, the Ju'hoansi and the !Xun, and one Kalahari Khoe group, namely, the Khwe. By including more groups, even though the sample sizes have not been the best in some populations, we found a clinal difference between northern and southern Khoe-San groups in the present study (Figures 3.18, 5.9 and 5.16). The Nama Khoe group has a similar genetic signature to the southern Khoe-San and Coloured groups (Figures 3.16, 3.17, 3.21, 3.22, 4.21, 4.22, 4.26, 5.7, 5.13, 5.14). Haplogroup frequencies differ between northern and southern groups. In both the mtDNA and Y-chromosome studies the northern groups contain haplogroups that are exclusive to them (Figures 3.7, 3.8, 3.11, 4.1, 4.2, 4.3). It is probable that the northern groups had gene-flow with other ancient hunter-gatherer groups north of them that introduced genetic material to them that are not found in the southern groups. Thus the L0k mtDNA haplogroups that was previously defined as a Khoe-San haplogroup are not present in the southern groups. Similarly previously Khoe-San associated Y-chromosome haplogroups A-

M14 and B-M112 mostly occur in northern groups. The pan-Khoe-San associated haplogroups L0d for the mitochondria and A-M51 for the Y-chromosome, have a larger diversity in the southern groups.

Genetic studies based on autosomal and mitochondrial DNA thus did find a difference between Ju speakers and descendants of Tuu speakers. Also, a greater genetic diversity was seen in the pan-Khoe-San associated haplogroups of the Tuu speaker descendants. This mirrors the linguistic profile where the Tuu languages were historically more diverse than the Ju languages. The KhoeKhoe speakers (Nama) clustered with southern Tuu groups, the Kalahari Khoe group (/Gui + //Gana + Kgalagari) clustered with northern groups, while the Khwe (Kalahari Khoe) have some similarity to northern groups but seem to have a unique genetic profile aside from its Bantu-speaking admixture. It therefore seems that the emerging genetic profile reflects the deep division between the Ju and Tuu speakers but that the Khoe language group was introduced later on to some of the Ju and Tuu speakers with some gene flow.

However, to more conclusively establish the genetic relationships between the different linguistically classified Khoe-San groups, bigger sample sizes and the inclusion of additional groups such as the !Xǃó (a Tuu speaking group), and more representation from the Kalahari Khoe group, such as the Naro, the Shua and Tshua and a less admixed group of /Gui and //Gana are needed.

The relationship between geographic and genetic distance in Khoe-San groups

Previous studies found that while in food producers the gene flow between groups was female biased because of patrilocality, hunter-gatherer populations had a male biased gene-flow (Seielstad *et al.*, 1998; Hammer *et al.*, 2001a; Destro-Bisol *et al.*, 2004; Wood *et al.*, 2005). This was observed through the stronger association of geographic distance with mtDNA genetic distance than with Y-chromosome genetic distance.

The present study found similar results to previous studies (Figures 3.18 and 4.23). Khoe-San hunter-gatherer populations had a significant correlation between mtDNA genetic distance and geographic distance, while there was no correlation between Y-chromosome

genetic distance and geographic distance. These results indicates that male movement between groups in the Khoe-San is more prominent than female movement.

The genetic affinities of the Khwe population

Because the Khwe phenotypically resemble Bantu-speakers but speak a Khoisan language it was not certain whether this group genetically resemble Khoe-San groups. Theories put forward was that the Khwe are Khoe-San groups with extensive Bantu-speaking admixture, Bantu-speakers that lost their cattle and language, another pastoralist population closely related to Bantu-speakers who occupied the region before the Bantu expansions or a mixture of various refugee groups driven from the grazing grounds into the Okovango swamps (Cashdan, 1986). Serogenetic studies supported the theory that the Khwe are Bantu-speakers that lost their cattle. Published mtDNA studies showed high amounts of Bantu-speaking admixture (Chen *et al.*, 2000; Tishkoff *et al.*, 2007). It, however, also showed appreciable frequencies of northern San associated haplogroups, L0d and L0k. Henn *et al.*, theorized that the Khwe is a descendant group of the east African pastoralists that introduced sheep into southern Africa (Henn *et al.*, 2008).

Autosomal results from the present study also support high amounts of Bantu-speaking admixture into the Khwe (Figure 5.3 and Table 5.2). The Khwe, however, also contain a large proportion of Khoe-San genetic variation. This Khoe-San genetic component is much larger than the Khoe-San genetic component introduced into other southern Bantu-speaking groups. Their Y-chromosome genetic profile contains high amounts of the east African pastoralist associated marker, supporting the study of Henn *et al.*, (Henn *et al.*, 2008). Besides the Bantu-speaking associated haplogroups, their mtDNA profile contains primarily haplogroup L0k1 and also a newly identified haplogroup L0dx. From network analysis it was deduced that the L0k1 haplogroup was introduced to the northern San groups by the Khwe, while the L0dx haplogroup more likely were transferred from the !Xun to the Khwe. The L0k1 haplogroup is exclusive to the northern San groups. If the Khwe introduced the L0k1 haplogroup into the northern groups it will be interesting to see if any other African hunter-gatherer group contain the L0k1 group. Thus far L0k1 was not found in the Pygmy, Hadza and Sandawe groups. A related haplogroup L0k2 was, however, identified in an individual from Yemen (Behar *et al.*, 2008). This suggests that the L0k

haplogroups might have had an extensive spread in prehistoric Africa but remnants of the haplogroup in other populations have been lost due to drift or has not been detected due to insufficient sampling.

It is therefore likely that the Khwe came from a location north from the traditional San territory and introduced new mtDNA and Y-chromosome haplogroups into the San groups.

The eastern Khoe-speaking San groups, Tshua and Shua, phenotypically resemble the Khwe and it would be interesting to include them in future genetic studies. Groups that occupied the region between east and southern Africa before the Bantu-expansions might be related to the Khwe group. It will therefore also be interesting to include groups such as the Ba-Twa Pygmy group in future genetic studies. Furthermore, comparing east African pastoralist groups containing high frequencies of Y-chromosome haplogroup E-M35 to the Khwe will also be crucial towards pin-pointing their origin.

The spread of pastoralism in southern Africa

Henn *et al.*, suggested that pastoralism was introduced ~2 000 years BP by a group from east Africa to the northern Botswana area (Henn *et al.*, 2008). This group was possibly ancestral to the present day Khwe group since the E-M293 marker associated with the introduction of pastoralism occurs in high frequencies in the Khwe. The Hadza and Sandawe group of east Africa also carry this E-M293 marker. Without representation of more Khoe-San groups in their study, Henn *et al.*, could not address the question of how pastoralism spread after it reached the northern Botswana area (Henn *et al.*, 2008).

The Henn *et al.*, (Henn *et al.*, 2008) study was published after the laboratory work for this thesis was completed and therefore the E-M293 marker was not typed. However, analysis of the E-M35* (DYS389I-10) that most likely is the equivalent of E-M293 were performed. The results showed that it is not likely that the spread of pastoralism was a clear-cut demic or cultural diffusion towards the south. Rather some E-M35* (DYS389I-10) male individuals integrated in the southern tribes and took with them the pastoralist practice and likely also their Khoe-language. The southern San groups that adopted the pastoralist culture and Khoe language had population expansions and became the Khoe (KhoeKhoe speakers).

This theory is supported by the Y-chromosome and mtDNA profile of the representative Khoe group, the Nama. Although the Nama do contain high proportions of E-M35* (DYS389I-10) they still retained a larger proportion of original Khoe-San haplogroup A. Furthermore, their mtDNA and remaining Y-chromosome haplogroup profile is similar to the other southern Khoe-San and Coloured groups.

The present study also identified another E-M35* profile that most likely does not contain the E-M293 marker but possibly also arrived with the group from east Africa. It is unlikely that only one haplotype would have migrated south and Henn *et al.*, admits that it is possible that other male individuals who did not carry M293 were also involved (Henn *et al.*, 2008). Fewer !Xun and Khwe individuals carry this E-M35* profile and this profile did not spread to the southern groups.

A demic diffusion of a few male individuals coupled to cultural diffusion would also explain why there is no ceramic stylistic chain in the archaeological record, which reflects the spread of pastoralism by a Khoe group (Sadr, 1998). Since only male individuals dispersed the ceramic styles would not accompany the pastoralist tradition. Furthermore, a previous study on rock paintings suggested a similar hypothesis (Kinahan, 1995). This theory is based on paintings that indicated male figures that are distinct from the traditional San monochrome trance scenes. They were identified as specialist shamans with higher status. The hypothesis put forward was that these individuals acquired higher status through the acquisition of sheep (Kinahan, 1995). It could also have been that these figures were the immigrant males from east Africa or descendants from them. Due to their high status in the communities they could have transferred their language to the resident San groups as well.

Future research

The typing of the E-M293 in the present study group is crucial. Furthermore genetic characterization of the eastern Khoe-speaking San groups, the Tshua and Shua, of eastern Botswana is important since they phenotypically resembles the Khwe and also might be descendants of the east African pastoralists. Moreover, one of the eastern Khoe-speaking San groups, the Hietshware, is the linguistic link to the extinct language, Kwadi, which in turn links to the east African Sandawe language. An E-M293 characterization in the

remnant hunter-gatherer groups intermediate to the Khwe and the east African groups will also be interesting. Especially the Ba-Twa Pygmy group might harbor interesting genetic commonalities to both east African and the Khoe-San groups.

Do genetic data support population expansions as suggested in the archaeological record?

Archaeological records indicate that certain sites showed increases in population sizes but only for truncated periods during the MSA to LSA transition (30 000 – 20 000 years BP). Around the LGM (~18 000 years BP) population contractions and localized extinctions are recorded. Population densities only increased noticeably from 13 500 years BP after the LGM. Population increases are recorded especially in the last 4 000 years with various technological innovations. Pastoralism was introduced 2 000 years BP and gave rise to further population increases (Deacon and Deacon, 1999; Mitchell, 2002).

MtDNA genetic data is important to see if local population history corresponds to paleoenvironmental history since female associated markers are more sedentary than male associated markers for hunter-gatherer communities. By looking at expansion signals in the mtDNA genetic data through various methods, evidence of the expansions recorded in the archaeological record was found. Different mtDNA haplogroups have different associated expansion signals coupled to different geographic distributions. The localized population increases of the MSA to LSA transition was seen for one haplogroup. Another haplogroup show strong signals of the post LGM population increase. None of the haplogroups showed signals of the population contractions associated with the LGM. Several haplogroups showed further increases in the last 4 000 years. All haplogroups, except one, reacted with a steep population increase upon the introduction of pastoralism.

Many theories in the anthropological and archaeological field hypothesizes that in-moving pastoralists adversely affected hunter-gatherers. The pastoralists occupy resources and marginalize hunter-gatherers. This hypothesis was also used previously to explain mismatch distributions in hunter-gatherers. It was theorized that populations that did not go through the Neolithic transition, experienced reduction of effective population sizes because of competing Neolithic farmers that caused fragmentation of the hunter-gatherer

habitat, These reductions in population size obscured previous population expansion signals (Excoffier and Schneider, 1999).

Most L0d sub-haplogroups showed recent expansion signals associated with the introduction of pastoralism. Only one haplogroup present in low frequency (in both the Khoe and San groups) experienced a population contraction. This contraction is most likely due to drift effects coupled to the steep N_e increase of the other haplogroups. It therefore thus seems that most extant Khoe-San associated haplogroups benefited from the introduction of pastoralism into southern Africa. This might not always have been a direct benefit through the adoption of pastoralist practices, but could be indirect benefits through trade relations with pastoralists.

7. CONCLUSION

The inclusion of three genetic markers with different modes of inheritance (mtDNA maternal inheritance; Y-chromosome DNA paternal inheritance; autosomal DNA bi-parentally inherited), properties (no recombination in mtDNA and Y-chromosome DNA; recombination in autosomal DNA; single locus history for mtDNA and Y-chromosome DNA; multiple unlinked markers for the autosomal DNA) and differences in mutation rate (fast rate in the Y-chromosome; slower rates in the mtDNA and autosomal SNPs) afforded this study the unique opportunity of robustly examining patterns of genetic variation in Khoe-San populations from southern Africa. These data were used to assess the evolutionary history of the Khoe-San in Africa.

Both mtDNA and Y-chromosome studies revealed that the mtDNA lineages (L0d and L0k, found at frequencies of 74% and 14%, respectively) and Y-chromosome haplogroups (haplogroup A found at frequencies of 34%) in the Khoe-San are among the oldest lineages that have survived in the human population and retained in this group at appreciable frequencies. However, differences in frequencies and distribution of sub-haplogroups of the major mtDNA and Y-chromosome haplogroups suggest that the different Khoe-San groups have over the years diverged from an ancestral parental group and acquired their own unique history. Consequently, these findings caution against a haphazard grouping of populations or a pooling of groups into a single group.

Although language as a tool for historical reconstruction has a shallow depth of resolution (~10 000) relative to genetic data (~60 000 – 200 000 years using Y-chromosome DNA and mtDNA), the results from this study were concordant with linguistic data that suggested a deep and ancient divide between northern and southern Khoe-San groups (Güldemann, Forthcoming-a; Güldemann, In Press). This divide was more pronounced in the maternal gene-pool (mtDNA data), where genetic distances between groups strongly correlate with geographic distances. Conversely, no significant correlation was seen between Y-chromosome genetic distances and geographic distances. This pattern could be attributed to female stationarity and male migration between groups.

Y-chromosome data, more specifically, the distribution and frequency of the E-M35 haplogroup, seems to parallel archaeological data with respect to the spread of pastoralism in sub-Saharan Africa (Elphick, 1977; Smith, 1983; Smith, 1992; Sadr, 1998; Mitchell, 2002). Y-chromosome data obtained in the present study and that of Henn *et al.*, (2008) suggests that the present-day group who self identify as Khwe were responsible for the introduction of pastoralism from east Africa into the region of northern Botswana. Also, these data were used to address how pastoralism was introduced to the south. The data tend to favor a coupled cultural-demic model with the movement of a few male individuals that integrated with the existing San tribes south of them and took with them the pastoralist practice and likely also their Khoe-language. This pattern is reflected in the frequency and distribution of E-M35, with highest frequency (46%) in the Khwe and a decrease in frequency towards the south presenting with low frequencies (<10%) in the Karoo Coloured groups. Conversely, none of the mtDNA (female) L0k and L0d lineages observed in the Khwe group was observed in the southern Khoe-San and Coloured groups, suggesting limited or no female movement.

Many of the hypotheses discussed in this thesis were based on the interpretation of results from genetic data examined in the present study. Several of these would be refined, modified or even disproved as more data become available in the future. While the focus of this study was to evaluate the use of various types of genetic markers in reconstructing the history of Khoisan-speaking populations, a more comprehensive comparative analysis of these data with archaeological data were outside the scope of this study, and could be the focus of future studies.

This study has highlighted the place of the Khoe-San in the evolutionary history of African populations. Presently, many Khoe-San groups are still not being respected as individuals with a democratic right to speak for themselves and decide their own destiny. In the span of a few hundred years this group of people has lost so much; they have been massacred, victimized, discriminated against and marginalized by other migrant groups to their homeland region in southern Africa. They are increasingly being affected by social ills such as economic dependency, alcoholism, malnutrition, and societal breakdown. The constant discrimination and humiliation (especially among the younger generations) has had a

profound effect on the way individuals prefer to identify themselves, with a stronger affinity to self identify as Bantu-speakers or Coloured rather than Khoe or San. However, some groups are re-discovering their identity and take pride in the uniqueness of their ancestry. We had the opportunity in this study to take back genetic ancestry test results to many individuals and to share with them the genetic findings from this study. This dialogue with individuals and community, it is hoped, would contribute in spreading the “word” about their unique place in the history of the world and to document their own history.

8. REFERENCES

- Allard MW, Polanskey D, Miller K, Wilson MR, Monson KL and Budowle B (2005). Characterization of human control region sequences of the African American SWGDAM forensic mtDNA data set. *Forensic Sci Int* **148**: 169-79
- Ambrose SH (1982). Berkeley, University of California Press: 104–157.
- Amo T and Brand MD (2007). Were inefficient mitochondrial haplogroups selected during migrations of modern humans? A test using modular kinetic analysis of coupling in mitochondria from cybrid cell lines. *Biochem J* **404**: 345-51
- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, *et al.*, (1981). Sequence and organization of the human mitochondrial genome. *Nature* **290**: 457-65
- Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM and Howell N (1999). Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* **23**: 147
- Anisimova M and Gascuel O (2006). Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* **55**: 539-52
- Atkinson QD, Gray RD and Drummond AJ (2008). mtDNA variation predicts population size in humans and reveals a major Southern Asian chapter in human prehistory. *Mol Biol Evol* **25**: 468-74
- Atkinson QD, Gray RD and Drummond AJ (2009). Bayesian coalescent inference of major human mitochondrial DNA haplogroup expansions in Africa. *Proc Biol Sci* **276**: 367-73
- Balloux F, Handley LJ, Jombart T, Liu H and Manica A (2009). Climate shaped the worldwide distribution of human mitochondrial DNA sequence variation. *Proc Biol Sci* **276**: 3447-55
- Bamshad M, Wooding S, Salisbury BA and Stephens JC (2004). Deconstructing the relationship between genetics and race. *Nat Rev Genet* **5**: 598-609
- Bandelt HJ, Forster P and Rohl A (1999). Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* **16**: 37-48
- Barbujani G, Magagni A, Minch E and Cavalli-Sforza LL (1997). An apportionment of human DNA diversity. *Proc Natl Acad Sci U S A* **94**: 4516-9
- Barnard A (1988). Kinship, language and production: a conjectural history of Khoisan social structure. *Africa* **58**: 29-50
- Barnard A (1992). Hunters and herders of southern Africa - A comparative ethnography of the Khoisan peoples. Cambridge, Cambridge University Press
- Beaumont PB (1980). On the age of Border Cave hominids 1-5. *Palaeontologia Afr* **23**: 131-143
- Behar DM, Rosset S, Blue-Smith J, Balanovsky O, Tzur S, Comas D, Mitchell RJ, *et al.*, (2007). The Genographic Project public participation mitochondrial DNA database. *PLoS Genet* **3**: e104
- Behar DM, Vilems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, *et al.*, (2008). The dawn of human matrilineal diversity. *Am J Hum Genet* **82**: 1130-40
- Bennun N (2004). The Broken String - The last words of an extinct people. London, Penguin Books
- Bergen AW, Wang CY, Tsai J, Jefferson K, Dey C, Smith KD, Park SC, *et al.*, (1999). An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. *Ann Hum Genet* **63**: 63-80
- Bieseke M and Royal K (1999). Africa; Mbuti. The Cambridge encyclopedia of hunters and gatherers. Richard B and Daly R. Cambridge, Cambridge University Press: 210–214.
- Bleek DF (1928). Bushmen of central Angola. *Bantu Studies* **3**: 105-125
- Bleek WHI (1862). A comparative grammar of South African languages. Part I. Phonology. London
- Boonzaier E, Malherbe C, Smith A and Berens P (1996). The Cape Herders: A History of the Khoikhoi of Southern Africa. Cape Town and Johannesburg, David Philip
- Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA and Bertranpetit J (2001). High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet* **68**: 1019-29
- Bouzouggar A, Barton N, Vanhaeren M, d'Errico F, Collcutt S, Higham T, Hodge E, *et al.*, (2007). 82,000-year-old shell beads from North Africa and implications for the origins of modern human behavior. *Proc Natl Acad Sci U S A* **104**: 9964-9
- Bowcock AM, Bucci C, Hebert JM, Kidd JR, Kidd KK, Friedlaender JS and Cavalli-Sforza LL (1987). Study of 47 DNA markers in five populations from four continents. *Gene Geogr* **1**: 47-64
- Bowcock AM, Hebert JM, Mountain JL, Kidd JR, Rogers J, Kidd KK and Cavalli-Sforza LL (1991a). Study of an additional 58 DNA markers in five human populations from four continents. *Gene Geogr* **5**: 151-73

- Bowcock AM, Kidd JR, Mountain JL, Hebert JM, Carotenuto L, Kidd KK and Cavalli-Sforza LL (1991b). Drift, admixture, and selection in human evolution: a study with DNA polymorphisms. *Proc Natl Acad Sci U S A* **88**: 839-43
- Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR and Cavalli-Sforza LL (1994). High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* **368**: 455-7
- Broyhill K, Hitchcock R and Bieseke M (Current). Current situations facing the san peoples of southern africa, Review on Current San Economic and Social Situations for the University of Free State. <http://www.kalaharipeoples.org/downloads/Current%20Situations%20of%20the%20San.pdf>.
- Bryant D and Moulton V (2002). NeighborNet: An agglomerative method for the construction of planar phylogenetic networks. *Algorithms in Bioinformatics*. Guigó R and Guseld D, WABI 2002. **LNCS 2452**: 375-391.
- Campbell AC (1990). Comment on: Foragers, genuine or spurious? by J.S. Solway and R.B. Lee. *Curr Anthropol* **31**: 123-124
- Campbell MC and Tishkoff SA (2008). African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* **9**: 403-33
- Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, et al., (2002). A human genome diversity cell line panel. *Science* **296**: 261-2
- Cann RL, Stoneking M and Wilson AC (1987). Mitochondrial DNA and human evolution. *Nature* **325**: 31-6
- Casanova M, Leroy P, Boucekine C, Weissenbach J, Bishop C, Fellous M, Purrello M, et al., (1985). A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* **230**: 1403-6
- Cashdan E (1986). Hunter-gatherers of the northern Kalahari. *Contemporary Studies on Khoisan*. Vossen R and Keuthmann K. Hamburg, Helmut Buske Verlag. **1**: 145-180.
- Cavalli-Sforza LL (1986). African Pygmies. Orlando (FL), Academic Press
- Cavalli-Sforza LL (1998). The DNA revolution in population genetics. *Trends Genet* **14**: 60-5
- Cavalli-Sforza LL, Menozzi P and Piazza A (1994). The History and Geography of Human Genes. Princeton, Princeton University Press
- Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K and Wallace DC (2000). mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am J Hum Genet* **66**: 1362-83
- Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengard J, Salomaa V, et al., (1998). Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* **63**: 595-612
- Cooke CK (1965). Evidence of human migrations from the rock art of Southern Rhodesia. *Africa* **5**: 263-285
- Corander J, Waldmann P and Sillanpää MJ (2003). Bayesian analysis of genetic differentiation between populations. *Genetics* **163**: 367-74
- Crawhall N (2003). *The rediscovery of N!u and the ǀKhomani Land Claim Process, South Africa. Maintaining the Links: Language Identity and the Land: Proceedings of the Seventh Foundation for Endangered Languages Conference, Broome, Western Australia, Bristol: Foundation for Endangered Languages.*
- Crawhall N (2006). Languages, genetics and archaeology: problems and the possibilities in Africa. *The prehistory of Africa*. Soodyall H. Johannesburg & Cape Town, Jonathan Ball Publishers: 109-124.
- Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, Moral P, Watson E, et al., (2004). Phylogeographic analysis of haplogroup E3b (E-M215). Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet* **74**: 1014-22
- Cruciani F, La Fratta R, Trombetta B, Santolamazza P, Sellitto D, Colomb EB, Dugoujon JM, et al., (2007). Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol Biol Evol* **24**: 1300-11
- Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, et al., (2002). A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* **70**: 1197-214
- De Almeida A (1965). Bushmen and other non-Bantu peoples of Angola. Johannesburg, Witwatersrand University Press for the Institute for the Study of Man in Africa
- De Jongh M (2002). No fixed abode: the poorest of the poor and elusive identities in rural South Africa. *Journal of Southern African Studies* **28**: 441-460
- Deacon HJ and Deacon J (1999). Human Beginnings in South Africa. Uncovering the Secrets of the Stone Age. Cape Town and Johannesburg, David Philip Publishers

- Deacon HJ, Deacon J, Brooker M and Wilson ML (1978). The evidence for herding at Boomplaas Cave in the southern Cape, South Africa. *South African Archaeological Bulletin* **33**: 39-65
- Deacon J (1984). Later Stone Age people and their descendants in southern Africa. *Southern African Prehistory and Paleoenvironments*. Klein R G. Rotterdam, A. A. Balkema: 221-328.
- Deacon J (1996). A Tale of Two Families: Wilhelm Bleek, Lucy Lloyd and the /Xam San of the Northern Cape. *Miscast. Negotiating the Presence of the Bushmen*. Skotnes P. Cape Town, UCT Press: 93-113.
- Denbow JR and Wilmsen EN (1986). Advent and course of pastoralism in the kalahari. *Science* **234**: 1509-15
- Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglia A, Tofanelli S, *et al.*, (2004). Variation of female and male lineages in sub-saharan populations: the importance of sociocultural factors. *Mol Biol Evol* **21**: 1673-82
- Dornan SS (1975). Pygmies and Bushmen of the Kalahari. Cape Town, C. Struik (PTY) LTD.
- Drummond AJ and Rambaut A (2007). BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**: 214
- Drummond AJ, Rambaut A, Shapiro B and Pybus OG (2005). Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol Biol Evol* **22**: 1185-92
- Ehret C (1982). The first spread of food production in southern Africa. *The archaeological and linguistic reconstruction of African history*. Ehret C and Posnansky M. Berkeley, University of California Press: 158-181.
- Ehret C and Posnansky M (1982). The archaeological and linguistic reconstruction of African history. California, University of California Press
- Elphick R (1977). Kraal and castle. New Haven, Yale University Press
- Elphick R (1985). Khoikhoi and the founding of White South Africa. Johannesburg, Raven Press
- Elson JL, Turnbull DM and Howell N (2004). Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. *Am J Hum Genet* **74**: 229-38
- Engelbrecht JA (1936). The Korana: an account of their customs and their history. Cape Town, Miller
- Estermann C, Ed. (1976). *The ethnography of southwestern Angola, Volume 1: The non-Bantu peoples: the Ambo ethnic group*. New York, Africana Publishing Company.
- Evanno G, Regnaut S and Goudet J (2005). Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. *Mol Ecol* **14**: 2611-20
- Excoffier L, Laval G and Schneider S (2005). Arlequin ver. 3.0: An integrated software package for population genetics data analysis. *Evol Bioinform Online* **1**: 47-50
- Excoffier L and Schneider S (1999). Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc Natl Acad Sci U S A* **96**: 10597-602
- Excoffier L and Slatkin M (1995). Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**: 921-7
- Excoffier L and Yang Z (1999). Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol Biol Evol* **16**: 1357-68
- Falush D, Stephens M and Pritchard JK (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**: 1567-87
- Falush D, Stephens M and Pritchard JK (2007). Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol Ecol Notes* **7**: 574-578
- Felsenstein J (2004). PHYLIP (Phylogeny Inference Package) version 3.6. *Distributed by the author. Department of Genome Sciences, University of Washington, Seattle*
- Fluxus-engineering (2008). Fluxus Technology Ltd. **2008-2009**.
- Forster P (2004). Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philos Trans R Soc Lond B Biol Sci* **359**: 255-64; discussion 264
- Forster P, Harding R, Torroni A and Bandelt HJ (1996). Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet* **59**: 935-45
- Francois O, Ancelet S and Guillot G (2006). Bayesian clustering using hidden Markov random fields in spatial population genetics. *Genetics* **174**: 805-16
- Fu YX (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* **147**: 915-25
- Garrigan D and Hammer MF (2006). Reconstructing human origins in the genomic era. *Nat Rev Genet* **7**: 669-80
- Gifford-Gonzalez D (2000). Animal disease challenges to the emergence of pastoralism in sub-Saharan Africa. *Afr Archaeol Rev* **17**: 95-139

- Golden-Software I (2006). Surfer Demo, Golden Software Inc. **2007-2009**.
- Goldstein DB, Ruiz Linares A, Cavalli-Sforza LL and Feldman MW (1995). Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci U S A* **92**: 6723-7
- Gonder MK, Mortensen HM, Reed FA, de Sousa A and Tishkoff SA (2007). Whole-mtDNA genome sequence analysis of ancient African lineages. *Mol Biol Evol* **24**: 757-68
- Gordon R (1984). The !Kung in the Kalahari exchange: an ethnohistorical perspective. Past and present in hunter-gatherer studies. Schrire C. Orlando, FL, Academic Press: 195-224.
- Gordon R (1986). Once again: How many Bushmen are there? The past and future of !Kung ethnography: critical reflections and symbolic perspectives, essays in honour of Lorna Marshall. Biesele M, Gordon R and Lee R. Hamburg, Helmut Buske Verlag: 53-68.
- Gordon RG, Ed. (2005). Ethnologue: Languages of the World. Online version: <http://www.ethnologue.com/>. Dallas, Texas, SIL International.
- Gorissen P (2008). Google Maps Latitude, Longitude Popup. **2008**.
- Green RE, Malaspinas AS, Krause J, Briggs AW, Johnson PL, Uhler C, Meyer M, *et al.*, (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell* **134**: 416-26
- Greenberg JH (1963). The languages of Africa. Bloomington, Indiana, Indiana University Press
- Greenberg JH (1972). Linguistic evidence concerning Bantu origins. *J Afr Hist* **13**: 189-216
- Griffiths RC and Tavaré S (1994). Simulating probability distributions in the coalescent. *Theor Popul Biol* **46**: 131-159
- Grine FE, Bailey RM, Harvati K, Nathan RP, Morris AG, Henderson GM, Ribot I, *et al.*, (2007). Late Pleistocene human skull from Hofmeyr, South Africa, and modern human origins. *Science* **315**: 226-9
- Grun R, Shackleton NJ and Deacon HJ (1990). Electron-Spin-Resonance Dating of Tooth Enamel From Klasies River Mouth Cave. *Curr Anthropol* **31**: 427-432
- Guenther M (1996). From 'Lords of the Desert' to 'Rubbish People': The Colonial and Contemporary State of the Nharo of Botswana. Miscast. Negotiating the Presence of the Bushmen. Skotnes P. Cape Town, UCT Press.
- Guenther MG (1986). Acculturation and assimilation of the Bushmen of Botswana and Namibia. Contemporary Studies on Khoisan. Vossen R and Keuthmann K. Hamburg, Helmut Buske Verlag. **1**: 347-373.
- Guindon S, Lethiec F, Duroux P and Gascuel O (2005). PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Res* **33**: W557-9
- Güldemann T (2006a). The San languages of southern Namibia: linguistic appraisal with special reference to J. G. Krönlén's Njuusaa data. *Anthropological Linguistics* **48**: 369-395
- Güldemann T (2006b). Structural isoglosses between Khoekhoe and Tuu: the Cape as a linguistic area. Linguistic areas: convergence in historical and typological perspective. Matras Y, McMahon A and Vincent N. Hampshire, Palgrave Macmillan: 99-134.
- Güldemann T (2007). Clicks, genetics, and "proto-world" from a linguistic perspective. University of Leipzig Papers on Africa. Leipzig, Institut für Afrikanistik, Universität Leipzig.
- Güldemann T (Forthcoming-a). Greenberg's "case" for Khoisan: the morphological evidence. Problems of linguistic-historical reconstruction in Africa. Vossen R and Ibriszimow D. Köln, Rüdiger Köppe.
- Güldemann T (Forthcoming-b). Person-gender-number marking from Proto-Khoe-Kwadi to its descendents: a rejoinder with particular reference to language contact. Festschrift for Bernd Heine. König C and Vossen R. London, Routledge.
- Güldemann T (In Press). Changing profile when encroaching on hunter-gatherer territory: towards a history of the Khoe-Kwadi family in southern Africa. Hunter-gatherers and linguistic history: a global perspective. Güldemann T, McConwell P and Rhodes R. Cambridge, Cambridge University Press.
- Güldemann T and Elderkin ED (Forthcoming). On external genealogical relationships of the Khoe family. Khoisan Languages and Linguistics: the Riezler Symposium 2003. Brenzinger M and König C. Köln, Rüdiger Köppe.
- Hall TA (1999). BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids. Symp. Ser.* **41**: 95-98
- Hammer MF (1994). A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. *Mol Biol Evol* **11**: 749-61
- Hammer MF and Horai S (1995). Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet* **56**: 951-62

- Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, Griffiths RC, *et al.*, (1998). Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol* **15**: 427-41
- Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H and Zegura SL (2001a). Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* **18**: 1189-203
- Hammer MF, Spurdle AB, Karafet T, Bonner MR, Wood ET, Novelletto A, Malaspina P, *et al.*, (1997). The geographic distribution of human Y chromosome variation. *Genetics* **145**: 787-805
- Hammer O, Harper DAT and Ryan PD (2001b). PAST: Palaeontological Statistics software package for education and data analysis. *Palaeontologia Electronica* **4**: 9
- Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, *et al.*, (1997). Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am J Hum Genet* **60**: 772-89
- Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, *et al.*, (2000). Evidence for variable selective pressures at MC1R. *Am J Hum Genet* **66**: 1351-61
- Harpending H and Rogers A (2000). Genetic perspectives on human origins and differentiation. *Annu Rev Genomics Hum Genet* **1**: 361-85
- Harpending HC, Sherry ST, Rogers AR and Stoneking M (1993). The genetic structure of ancient human populations. *Curr Anthropol* **34**: 483-496
- Harris EE and Hey J (1999). X chromosome evidence for ancient human histories. *Proc Natl Acad Sci U S A* **96**: 3320-4
- Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, *et al.*, (2008). Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A* **105**: 10693-8
- Henshilwood CS (1996). A revised chronology for pastoralism in southernmost Africa: New evidence of sheep at ca. 2000 B.P. from Blombos Cave, South Africa. *Antiquity* **70**: 945-949
- Henshilwood CS, d'Errico F, Yates R, Jacobs Z, Tribolo C, Duller GA, Mercier N, *et al.*, (2002). Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science* **295**: 1278-80
- Hoernle AW, Ed. (1985). The social organization of the Nama and other essays. Johannesburg, Witwatersrand University Press.
- Horai S (1995). Evolution and the origins of man: clues from complete sequences of hominoid mitochondrial DNA. *Southeast Asian J Trop Med Public Health* **26 Suppl 1**: 146-54
- Horai S, Hayasaka K, Kondo R, Tsugane K and Takahata N (1995). Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A* **92**: 532-6
- Hudson RR (1990). Gene genealogies and the coalescent process. *Oxf Surv Evol Biol* **7**: 1-14
- Huffman TN (1982). Archaeology and the ethnohistory of the African Iron Age. *Ann Rev Anthropol* **11**: 133-150
- Huffman TN (1983). The trance hypothesis and the rock art of Zimbabwe. New approaches to southern African rock art. Lewis-Williams J D, South African Archaeological Society: 49-53.
- Huson DH and Bryant D (2006). Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol* **23**: 254-67
- Huson DH, Richter DC, Rausch C, Dezulian T, Franz M and Rupp R (2007). Dendroscope: An interactive viewer for large phylogenetic trees. *BMC Bioinformatics* **8**: 460
- Ingman M and Gyllenstein U (2007). Rate variation between mitochondrial domains and adaptive evolution in humans. *Hum Mol Genet* **16**: 2281-7
- Ingman M, Kaessmann H, Paabo S and Gyllenstein U (2000). Mitochondrial genome variation and the origin of modern humans. *Nature* **408**: 708-13
- Jakobsson M and Rosenberg NA (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics* **23**: 1801-6
- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, *et al.*, (2008). Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**: 998-1003
- Jenkins T (1974). Blood group Abantu population and family studies. *Vox Sang* **26**: 537-50
- Jenkins T (1982). Human evolution in southern Africa. The Unfolding Genome. Bonne-Tamir B. New York, Alan R. Liss Inc.: 227-253.
- Jenkins T (1986). The prehistory of the San and Khoikhoi as recorded in their blood. Contemporary Studies on Khoisan. Vossen R and Keuthmann K. Hamburg, Helmut Buske Verlag. **2**: 51-77.

- Jenkins T (1988). The peoples of southern Africa. Studies in diversity and disease. Raymond Dart Lectures. Lecture 24. Pines N J. Johannesburg, Institute for the Study of Man in Africa, Witwatersrand University Press.
- Jenkins T and Corfield V (1972). The red cell acid phosphatase polymorphism in Southern Africa: population data and studies on the R, RA and RB phenotypes. *Ann Hum Genet* **35**: 379-91
- Jenkins T and Dunn DS (1981). Haematological genetics in the tropics. Part 1: Tropical Africa. *Clin Haematol* **10**: 1029-50
- Jenkins T, Harpending HC, Gordon H, Keraan MM and Johnston S (1971). Red-cell-enzyme polymorphisms in the Khoisan peoples of Southern Africa. *Am J Hum Genet* **23**: 513-32
- Jenkins T and Nurse GT (1972). Blood group gene frequencies. *S Afr Med J* **46**: 560
- Jenkins T, Zoutendyk A and Steinberg AG (1970). Gammaglobulin groups (Gm and Inv) of various Southern African populations. *Am J Phys Anthropol* **32**: 197-218
- Jobling MA, Hurles ME and Tyler-Smith C (2004a). Human Evolutionary Genetics. Origins, Peoples & Disease. New York, Garland Publishing
- Jobling MA, Hurles ME and Tyler-Smith C (2004b). Making inferences from diversity. Human Evolutionary Genetics. Origins, Peoples & Disease. New York, Garland Publishing: 164.
- Jobling MA, Hurles ME and Tyler-Smith C (2004c). Measuring and summarizing genetic variation. Human Evolutionary Genetics. Origins, Peoples & Disease. New York, Garland Publishing: 155.
- Jobling MA and Tyler-Smith C (2000). New uses for new haplotypes the human Y chromosome, disease and selection. *Trends Genet* **16**: 356-62
- Jobling MA and Tyler-Smith C (2003). The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* **4**: 598-612
- Johnston HH (1913). A survey of the ethnography of Africa: and the former racial and tribal migrations of that continent. *Journal of the Royal Anthropological Institute* **XLIII**: 391-392
- Jorde LB, Watkins WS and Bamshad MJ (2001). Population genomics: a bridge from evolutionary history to genetic medicine. *Hum Mol Genet* **10**: 2199-207
- Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT and Batzer MA (2000). The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. *Am J Hum Genet* **66**: 979-88
- Kaessmann H, Heissig F, von Haeseler A and Paabo S (1999). DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat Genet* **22**: 78-81
- Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL and Hammer MF (2008). New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res* **18**: 830-8
- Kayser M, Brauer S, Weiss G, Underhill PA, Roewer L, Schiefenhover W and Stoneking M (2000). Melanesian origin of Polynesian Y chromosomes. *Curr Biol* **10**: 1237-46
- Kinahan J (1995). A new archaeological perspective on nomadic pastoralist expansion in south-western Africa. *Azania* **29/30**: 211-226
- Kingman JFC (1982). On the genealogy of large populations. *J Appl Probab* **19A**: 27-43
- Kivisild T, Shen P, Wall DP, Do B, Sung R, Davis K, Passarino G, *et al.*, (2006). The role of selection in the evolution of human mitochondrial genomes. *Genetics* **172**: 373-87
- Klein RG (1986). The prehistory of stone age herders in the Cape Province of South Africa. *South African Archaeological Society, Goodwin Series* **5**: 5-12
- Klein RG (2000). The human career: Human biological and cultural origins. Chicago, University of Chicago Press
- Klein RG, Avery G, Cruz-Urbe K, Halkett D, Parkington JE, Steele T, Volman TP, *et al.*, (2004). The Ysterfontein 1 Middle Stone Age site, South Africa, and early human exploitation of coastal resources. *Proc Natl Acad Sci U S A* **101**: 5708-15
- Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, *et al.*, (2003). African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr Biol* **13**: 464-73
- Korsman SA and Plug I (1992). Archeological evidence and ethnographic analogy - interpreting prehistoric social behaviour at Honingklip in the eastern Transvaal. *S Afr J Ethnol* **15**: 120-126
- Lahr MM and Foley RA (1998). Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *Am J Phys Anthropol* **Suppl 27**: 137-76
- Landsteiner K (1901). Über Agglutinationserscheinungen normalen menschlichen. *Wiener Klin. Wochenschr.* **14**: 1132-1134
- Langella O (2002). Populations v.1.2.30. **2008**.

- le Roux W and White A, Eds. (2004). Voices of the San. Cape Town, Kwela Books.
- Lee RB (1979). The !Kung San: men, women, and work in a foraging society. Cambridge, Cambridge University Press.
- Lewis-Williams JD (1986). Beyond style and portrait: A comparison of Tanzanian and southern African rock art. Contemporary Studies on Khoisan. Vossen R and Keuthmann K. Hamburg, Helmut Buske Verlag. **2**: 95-122.
- Lewis PO and Zaykin DV (2001). Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d16c). Free program distributed by the authors over the internet.
- Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, *et al.*, (2008). Worldwide human relationships inferred from genome-wide patterns of variation. *Science* **319**: 1100-4
- Liu K and Muse SV (2005). PowerMarker: an integrated analysis environment for genetic marker analysis. *Bioinformatics* **21**: 2128-9
- Lomax A (1968). Folk Song Style and Culture. Washington, DC, National Association for the Advancement of Science: 16–18, 26, 91–92.
- Lombard M (2008). From testing times to high resolution: The Late Pleistocene Middle Stone Age of South Africa and beyond. *Goodwin Series* **10**: 180-188
- Low BS (1988). Measures of polygyny in humans. *Curr Anthropol* **29**: 189–194
- Maca-Meyer N, Gonzalez AM, Larruga JM, Flores C and Cabrera VM (2001). Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* **2**: 13
- Macaulay V, Richards M, Hickey E, Vega E, Cruciani F, Guida V, Scozzari R, *et al.*, (1999). The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* **64**: 232-49
- Marlowe FW (2004). Is human ovulation concealed? Evidence from conception beliefs in a hunter-gatherer society. *Arch Sex Behav* **33**: 427–432
- Marshall J and Ritchie C (1984). Where are the Ju/wasi of Nyae Nyae? Changes in a Bushman society: 1958-1981. Cape Town, Centre for African Studies, University of Cape Town (Communications No.9)
- Marshall L (1960). !Kung Bushmen Bands. *Africa* **30**: 325-355
- Marshall L (1976). The !Kung of Nyae Nyae. Cambridge, Harvard University Press
- Mazel A (1996). In pursuit of San Pre-colonial History in the Natal Drakensberg: A Historical Overview. Miscast. Negotiating the Presence of the Bushmen. Skotnes P. Cape Town, UCT Press: 191-195.
- McDougall I, Brown FH and Fleagle JG (2005). Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* **433**: 733-6
- Merriwether DA, Clark AG, Ballinger SW, Schurr TG, Soodyall H, Jenkins T, Sherry ST, *et al.*, (1991). The structure of human mitochondrial DNA variation. *J Mol Evol* **33**: 543-55
- Meyer S, Weiss G and von Haeseler A (1999). Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* **152**: 1103-10
- Michels C (1997). Latitude/Longitude Distance Calculation. **2008**.
- Miller-Ockhuizen A and Sands BE (1999). !Kung as a linguistic construct. *Language & Communication* **19**: 401-413
- Miller SA, Dykes DD and Polesky HF (1988). A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res* **16**: 1215
- Mishmar D, Ruiz-Pesini E, Golik P, Macaulay V, Clark AG, Hosseini S, Brandon M, *et al.*, (2003). Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci U S A* **100**: 171-6
- Mitchell PJ (2002). The Archaeology of Southern Africa. Cambridge, Cambridge University Press
- Mitchell PJ (2008). Developing the archaeology of Marine Isotope Stage 3. *Goodwin Series* **10**: 52-65
- Morris AG (1992). Biological relationships between Upper Pleistocene and Holocene populations in southern Africa. Continuity or Replacement: Controversies in Homo sapiens evolution. Brauer G and Smith F H. Rotterdam, Balkema: 131-143.
- Morris AG (2002). Isolation and the Origin of the Khoisan: Late Pleistocene and Early Holocene Human Evolution at the Southern End of Africa. *Hum Evol* **17**: 231-240
- Morris AG (2003). The Myth of the East African 'Bushmen'. *S Afr Arch Bull* **58**: 85-90
- Morris AG (2005). Prehistory in blood and bone: An essay on the reconstruction of the past from genetics and morphology. *Transactions of the Royal Society of South Africa* **60**: 111-114
- Morris AG (2008). Searching for 'real' Hottentots: the Khoekhoe in the history of South African physical anthropology *Southern African Humanities* **20**: 221-233
- Morris AG and Ribot I (2006). Morphometric cranial identity of prehistoric Malawians in the light of sub-Saharan African diversity. *Am J Phys Anthropol* **130**: 10-25

- Murdock GP (1967). *Ethnographic atlas*. Pittsburgh (PA), University of Pittsburgh Press
- Murdock GP (1981). *Atlas of World Cultures*. Pittsburgh (PA), University of Pittsburgh Press
- Naidoo T, Schlebusch CM, Makkan H, Patel P, Mahabeer R, Erasmus JC and Soodyall H (Unpublished). Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations.
- Nebel A, Filon D, Brinkmann B, Majumder PP, Faerman M and Oppenheim A (2001). The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. *Am J Hum Genet* **69**: 1095-112
- Nei M (1987). *Molecular Evolutionary Genetics*. New York, USA, Columbia University Press
- Nei M and Livshits G (1989). Genetic relationships of Europeans, Asians and Africans and the origin of modern Homo sapiens. *Hum Hered* **39**: 276-81
- Nelson RM (2006). S-Compare. **2006-2008**.
- Newman JL (1995). *The peopling of Africa*. New Haven, CT, Yale University Press
- Niu T, Qin ZS, Xu X and Liu JS (2002). Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* **70**: 157-69
- Nurse GT (1983). Population movement around the northern Kalahari. *African Studies* **42**: 153-63
- Nurse GT and Jenkins T (1977). Serogenetic studies on the Kavango peoples of South West Africa. *Ann Hum Biol* **4**: 465-78
- Nurse GT, Lane AB and Jenkins T (1976). Sero-genetic studies on the Dama of South West Africa. *Ann Hum Biol* **3**: 33-50
- Nurse GT, Weiner JS and Jenkins T (1985). *The Peoples of Southern Africa and their Affinities*. New York, Oxford University Press
- Oliver MA and Webster R (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information Systems* **4**: 313
- Parkington JE (1984). Soaqua and Bushmen: hunters and robbers. Past and present in hunter-gatherer studies. Schrire C. New York, Academic Press: 151-174.
- Parkington JE, Yates R, Manhire A and Halkett D (1986). The social impact of pastoralism in the southwestern Cape. *Journal of Anthropological Archaeology* **5**: 313-329
- Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer M and Santachiara-Benerecetti AS (1998). Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am J Hum Genet* **62**: 420-34
- Penn N (1996). "Fated to Perish": The Destruction of the Cape San. Miscast. Negotiating the Presence of the Bushmen. Skotnes P. Cape Town, UCT Press: 81-91.
- Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ and Amorim A (2001). Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* **65**: 439-58
- Phillipson D (1993). *African Archaeology*. Cambridge, UK, Cambridge Univ Press
- Pijper A (1932). Blood-groups of Bushmen. *S Afr Med J* **6**: 35-37
- Pijper A (1935). Blood groups in the Hottentots. *S Afr Med J* **9**: 192-195
- Pilkington MM, Wilder JA, Mendez FL, Cox MP, Woerner A, Angui T, Kingan S, *et al.*, (2008). Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa. *Mol Biol Evol* **25**: 517-25
- Polzin T and Daneschmand SV (2003). On Steiner trees and minimum spanning trees in hypergraphs. *Operations Res Lett* **31**: 12-20
- Posada D and Crandall KA (1998). MODELTEST: testing the model of DNA substitution. *Bioinformatics* **14**: 817-8
- Potgieter EF (1955). The disappearing Bushmen of Lake Chrissie: a preliminary survey. Pretoria, J.L. van Schaick
- Prins F (Unknown). A glimpse into Bushman presence in the Anglo Boer War. http://www.chrissiesmeer.co.za/the_sun.html
- Pritchard JK, Seielstad MT, Perez-Lezaun A and Feldman MW (1999). Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* **16**: 1791-8
- Pritchard JK, Stephens M and Donnelly P (2000). Inference of population structure using multilocus genotype data. *Genetics* **155**: 945-59
- Przeworski M, Hudson RR and Di Rienzo A (2000). Adjusting the focus on human variation. *Trends Genet* **16**: 296-302
- Qamar R, Ayub Q, Khaliq S, Mansoor A, Karafet T, Mehdi SQ and Hammer MF (1999). African and Levantine origins of Pakistani YAP+ Y chromosomes. *Hum Biol* **71**: 745-55

- Qamar R, Ayub Q, Mohyuddin A, Helgason A, Mazhar K, Mansoor A, Zerjal T, *et al.*, (2002). Y-chromosomal DNA variation in Pakistan. *Am J Hum Genet* **70**: 1107-24
- Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, *et al.*, (2008). Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci U S A* **105**: 1596-601
- R-Project (2006). The R-Project for statistical computing, CRAN project. **2006-2009**.
- Rambaut A and Drummond AJ (2007). Tracer v1.4.
- Ramirez-Soriano A, Ramos-Onsins SE, Rozas J, Calafell F and Navarro A (2008). Statistical power analysis of neutrality tests under demographic expansions, contractions and bottlenecks with recombination. *Genetics* **179**: 555-67
- Ramos-Onsins SE and Rozas J (2002). Statistical properties of new neutrality tests against population growth. *Mol Biol Evol* **19**: 2092-100
- Raymond M and Rousset F (1995). An exact test for population differentiation. *Evolution Int J Org Evolution* **49**: 1280-1283
- Reed FA and Tishkoff SA (2006). African human diversity, origins and migrations. *Curr Opin Genet Dev* **16**: 597-605
- Reynolds J, Weir BS and Cockerham CC (1983). Estimation of the Coancestry Coefficient: Basis for a Short-Term Genetic Distance. *Genetics* **105**: 767-779
- Richards M, Corte-Real H, Forster P, Macaulay V, Wilkinson-Herbots H, Demaine A, Papiha S, *et al.*, (1996). Paleolithic and neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* **59**: 185-203
- Risch N, Burchard E, Ziv E and Tang H (2002). Categorization of humans in biomedical research: genes, race and disease. *Genome Biol* **3**: comment2007
- Rogers AR and Harpending H (1992). Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* **9**: 552-69
- Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry ST, Stoneking M, *et al.*, (2002). Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res* **12**: 602-12
- Rosenberg NA (2002). Distruct: a program for the graphical display of structure results.
- Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK and Feldman MW (2005). Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLoS Genet* **1**: e70
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA and Feldman MW (2002). Genetic structure of human populations. *Science* **298**: 2381-5
- Rosser ZH, Zerjal T, Hurler ME, Adojaan M, Alavantic D, Amorim A, Amos W, *et al.*, (2000). Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* **67**: 1526-43
- Rozas J, Sanchez-DelBarrio JC, Messeguer X and Rozas R (2003). DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496-7
- Rozen S and Skaletsky H (2000). Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* **132**: 365-86
- Ruiz-Pesini E, Mishmar D, Brandon M, Procaccio V and Wallace DC (2004). Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**: 223-6
- Sadr K (1997). Archaeology and the Bushman Debate. *Curr Anthropol* **38**: 104-112
- Sadr K (1998). The First Herders at the Cape of Good Hope. *Afr Archaeol Rev* **15**: 101-132
- Saillard J, Forster P, Lynnerup N, Bandelt HJ and Norby S (2000). mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* **67**: 718-26
- Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, *et al.*, (2002). The making of the African mtDNA landscape. *Am J Hum Genet* **71**: 1082-111
- Sands B (1998). Language, Identity and Conceptualization Among the Khoisan. Köln, Rudiger Kupper. **Bd 15**: 266–283.
- Sands BE, Miller AL and Brugman J (2007). The Lexicon in Language Attrition: The Case of N!uu. Selected Proceedings of the 37th Annual Conference on African Linguistics. Payne D L and Peña J. Somerville, MA, Cascadilla Proceedings Project: 55-65.
- Santos FR, Pandya A, Tyler-Smith C, Pena SD, Schanfield M, Leonard WR, Osipova L, *et al.*, (1999). The central Siberian origin for native American Y chromosomes. *Am J Hum Genet* **64**: 619-28
- Schapera I (1930). *The Khoisan Peoples of South Africa: Bushmen and Hottentots*. London, George Routledge and Sons

- Scheet P and Stephens M (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* **78**: 629-44
- Schlebusch CM, Naidoo T and Soodyall H (2009). SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis* **30**: 3657-64
- Schneider S and Excoffier L (1999). Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* **152**: 1079-89
- Schultze L, Ed. (1928). *Zur Kenntnis des Körpers der Hottentotten und Buschmänner*. Zoologische und Anthropologische Ergebnisse einer Forschungsreise im westlichen und zentralen Südafrika.
- Scozzari R, Cruciani F, Malaspina P, Santolamazza P, Ciminelli BM, Torroni A, Modiano D, *et al.*, (1997). Differential structuring of human populations for homologous X and Y microsatellite loci. *Am J Hum Genet* **61**: 719-33
- Scozzari R, Cruciani F, Pangrazio A, Santolamazza P, Vona G, Moral P, Latini V, *et al.*, (2001). Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. *Hum Immunol* **62**: 871-84
- Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D, Arredi B, *et al.*, (1999). Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet* **65**: 829-46
- Sealy J and Yates R (1994). The chronology of the introduction of pastoralism to the Cape, South Africa. *Antiquity* **68** 58-67
- Seielstad MT, Hebert JM, Lin AA, Underhill PA, Ibrahim M, Vollrath D and Cavalli-Sforza LL (1994). Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Hum Mol Genet* **3**: 2159-61
- Seielstad MT, Minch E and Cavalli-Sforza LL (1998). Genetic evidence for a higher female migration rate in humans. *Nat Genet* **20**: 278-80
- Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, *et al.*, (2004). Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* **74**: 1023-34
- Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, *et al.*, (2000). The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. *Science* **290**: 1155-9
- Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL and Underhill PA (2002). Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet* **70**: 265-8
- Semino O, Torroni A, Scozzari R, Brega A and Santachiara Benerecetti AS (1991). Mitochondrial DNA polymorphisms among Hindus: a comparison with the Tharus of Nepal. *Ann Hum Genet* **55 (Pt 2)**: 123-36
- Shapiro B, Drummond AJ, Rambaut A, Wilson MC, Matheus PE, Sher AV, Pybus OG, *et al.*, (2004). Rise and fall of the Beringian steppe bison. *Science* **306**: 1561-5
- Sharp J and Douglas S (1996). Prisoners of their Reputation? The Veterans of the 'Bushman' Battalions in South Africa. *Miscast. Negotiating the Presence of the Bushmen*. Skotnes P. Cape Town, UCT Press: 323-329.
- Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, *et al.*, (2000). Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci U S A* **97**: 7354-9
- Sherry ST, Rogers AR, Harpending H, Soodyall H, Jenkins T and Stoneking M (1994). Mismatch distributions of mtDNA reveal recent human population expansions. *Hum Biol* **66**: 761-75
- Silberbauer GB (1965). Report to the Government of Bechuanaland on the Bushman Survey. Gabarone, Bechuanaland Government
- Slatkin M (1995). Hitchhiking and associative overdominance at a microsatellite locus. *Mol Biol Evol* **12**: 473-80
- Smith A (2005). The concepts of 'Neolithic' and 'Neolithisation' for Africa? *Before Farming* **1**: 1- 6
- Smith A, Malherbe C, Guenther M and Berens P (2000). The Bushmen of Southern Africa. Cape Town, David Phillips Publishers
- Smith AB (1983). Prehistoric Pastoralism in the Southwestern Cape, South Africa. *World Archaeology* **15**: 79-89
- Smith AB (1986). Competition, Conflict and Clientship: Khoi and San Relationships in the Western Cape. *Goodwin Series* **5**: 36-41
- Smith AB (1992). Origins and Spread of Pastoralism in Africa. *Annual Review of Anthropology* **21**: 125-141

- Smith AB (1995). Einiqualand: Studies of the Orange River Frontier. Cape Town, University of Cape Town Press
- Smith AB, Sadr K, Gribble J and Yates R (1991). Excavations in the South-Western Cape, South Africa, and the Archaeological Identity of Prehistoric Hunter-Gatherers within the Last 2000 Years. *The South African Archaeological Bulletin* **46**: 71-91
- Smith BW (2006). Reading rock art and writing genetic history. *The Prehistory of Africa - Tracing the lineage of modern man*. Soodyall H. Johannesburg & Cape Town, Jonathan Ball Publishers: 76-96.
- Soodyall H and Jenkins T (1992). Mitochondrial DNA polymorphisms in Khoisan populations from southern Africa. *Ann Hum Genet* **56 (Pt 4)**: 315-24
- Soodyall H, Vigilant L, Hill AV, Stoneking M and Jenkins T (1996). mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. *Am J Hum Genet* **58**: 595-608
- Stephens M, Smith NJ and Donnelly P (2001). A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**: 978-89
- Steyn HP (1984). Southern Kalahari San Subsistence Ecology: A Reconstruction. *The South African Archaeological Bulletin* **39**: 117-124
- Stoneking M (2000). Hypervariable sites in the mtDNA control region are mutational hotspots. *Am J Hum Genet* **67**: 1029-32
- Stoneking M and Soodyall H (1996). Human evolution and the mitochondrial genome. *Curr Opin Genet Dev* **6**: 731-6
- Stow GW (1905). The native races of South Africa. London, Swan Sonnen-schein
- Stynder DD (2009). Craniometric evidence for South African Later Stone Age herders and hunter-gatherers being a single biological population. *Journal of Archaeological Science* **36**: 798-806
- Stynder DD, Ackermann RR and Sealy JC (2007a). Craniofacial variation and population continuity during the South African Holocene. *Am J Phys Anthropol* **134**: 489-500
- Stynder DD, Ackermann RR and Sealy JC (2007b). Early to mid-Holocene South African Later Stone Age human crania exhibit a distinctly Khoesan morphological pattern. *S Afr J Sci* **103**: 349-352
- Swofford DL (1998). PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sunderland, Massachusetts, Sinauer Associates
- Tajima F (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585-95
- Tajima F (1996). The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* **143**: 1457-65
- Tamura K, Dudley J, Nei M and Kumar S (2007). MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol Biol Evol* **24**: 1596-9
- Tamura K and Nei M (1993). Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* **10**: 512-26
- Ten Raa R (1970). The couth and the uncouth: ethnic, social and linguistic division among the Sandawe of central Tanzania. *Anthropos* **65**: 127-153
- Thomas MG, Bradman N and Flinn HM (1999). High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. *Hum Genet* **105**: 577-81
- Thompson JD, Higgins DG and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-80
- Thomson R, Pritchard JK, Shen P, Oefner PJ and Feldman MW (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci U S A* **97**: 7360-5
- Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, *et al.*, (2007). History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol* **24**: 2180-95
- Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, *et al.*, (2009). The genetic structure and history of Africans and African Americans. *Science* **324**: 1035-44
- Tobias PV (1985). History of physical anthropology in Southern Africa. *Am J Phys Anthropol* **28**: 1-52
- Torroni A, Achilli A, Macaulay V, Richards M and Bandelt HJ (2006). Harvesting the fruit of the human mtDNA tree. *Trends Genet* **22**: 339-45
- Torroni A, Bandelt HJ, D'Urbano L, Lahermo P, Moral P, Sellitto D, Rengo C, *et al.*, (1998). mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* **62**: 1137-52

- Torroni A, Huoponen K, Francalacci P, Petrozzi M, Morelli L, Scozzari R, Obinu D, *et al.*, (1996). Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**: 1835-50
- Torroni A, Rengo C, Guida V, Cruciani F, Sellitto D, Coppa A, Calderon FL, *et al.*, (2001). Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* **69**: 1348-56
- Traill A (1973). 'N4 or S7': another Bushman language. *African Studies* **32**: 25-32
- Traill A (1996). !Khwaka Hhouiten Hhouiten - "The Rush of the Storm" : The linguistic death of /Xam. *Miscast. Negotiating the Presence of the Bushmen*. Skotnes P. Cape Town, UCT Press: 171-183.
- Traunmüller H (2003). Clicks and the idea of a human protolanguage. *Phonum* **9**: 1-4
- Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, Davis RW, *et al.*, (1997). Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* **7**: 996-1005
- Underhill PA and Kivisild T (2007). Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annu Rev Genet* **41**: 539-64
- Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley RA, Oefner PJ, *et al.*, (2001). The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* **65**: 43-62
- Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, *et al.*, (2000). Y chromosome sequence variation and the history of human populations. *Nat Genet* **26**: 358-61
- Wadley L (2007). The Middle Stone Age and Later Stone Age. A Search for Origins: Science, History and South Africa's 'Cradle of Humankind'. Bonner P, Esterhuysen A and Jenkins T. Johannesburg, Wits University Press: 122-135.
- Walker NJ (1995). The archaeology of the San: the Late stone age of Botswana. Speaking for the Bushmen. Sanders A J G M. Gabarone, The Botswana Society: 54-87.
- Wallace DC (1995). 1994 William Allan Award Address. Mitochondrial DNA variation in human evolution, degenerative disease, and aging. *Am J Hum Genet* **57**: 201-23
- Vallone PM and Butler JM (2004). AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques* **37**: 226-31
- Walter RC, Buffler RT, Bruggemann JH, Guillaume MM, Berhe SM, Negassi B, Libsekal Y, *et al.*, (2000). Early human occupation of the Red Sea coast of Eritrea during the last interglacial. *Nature* **405**: 65-9
- Vansina JC (1990). Paths in the rainforest. Towards a history of political tradition in equatorial Africa. London, Currey
- Ward RH, Frazier BL, Dew-Jager K and Paabo S (1991). Extensive mitochondrial diversity within a single Amerindian tribe. *Proc Natl Acad Sci U S A* **88**: 8720-4
- Watson E, Forster P, Richards M and Bandelt HJ (1997). Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* **61**: 691-704
- Weir BS (1996a). Genetic data analysis II. Sunderland, MA, Sinauer Associates, Inc: 141-150.
- Weir BS (1996b). Genetic data analysis II. Sunderland, MA, Sinauer Associates, Inc
- Westphal EOJ (1963). The Linguistic Prehistory of Southern Africa: Bush, Kwadi, Hottentot, and Bantu Linguistic Relationships. *Africa: Journal of the International African Institute* **33**: 237-265
- Westphal EOJ (1971). The click languages of southern and eastern Africa. Current trends in linguistics, 7: Linguistics in Sub-Saharan Africa. Sebeok T A. The Hague, Mouton: 367-420.
- Westphal EOJ (1974). Notes on A. 'Traill: N4 or S7?' (with a reply by A. Traill). *African Studies* **33**: 243-255
- White TD, Asfaw B, DeGusta D, Gilbert H, Richards GD, Suwa G and Howell FC (2003). Pleistocene Homo sapiens from Middle Awash, Ethiopia. *Nature* **423**: 742-7
- Whitfield LS, Sulston JE and Goodfellow PN (1995). Sequence variation of the human Y chromosome. *Nature* **378**: 379-80
- Vigilant L, Pennington R, Harpending H, Kocher TD and Wilson AC (1989). Mitochondrial DNA sequences in single hairs from a southern African population. *Proc Natl Acad Sci U S A* **86**: 9350-4
- Vigilant L, Stoneking M, Harpending H, Hawkes K and Wilson AC (1991). African populations and the evolution of human mitochondrial DNA. *Science* **253**: 1503-7
- Wilder JA, Mobasher Z and Hammer MF (2004). Genetic evidence for unequal effective population sizes of human females and males. *Mol Biol Evol* **21**: 2047-57
- Wilmsen EN (1989). Land filled with flies: A political economy of the Kalahari. Chicago, Chicago University Press
- Wilmsen EN, Denbow JR, Bicchieri MG, Binford LR, Gordon R, Guenther M, Lee RB, *et al.*, (1990). Paradigmatic History of San-Speaking Peoples and Current Attempts at Revision [and Comments and Replies]. *Curr Anthropol* **31**: 489-524

- Wilson IJ and Balding DJ (1998). Genealogical inference from microsatellite data. *Genetics* **150**: 499-510
- Vinnicombe P (1976). People of the eland: rockpaintings of the Drakensberg Bushmen as a reflection of their life and thought. Pietermaritzburg, University of Natal Press
- Vogel JO (1994). Eastern and south-central African Iron Age. Encyclopedia of precolonial Africa. Vogel J O. Walnut Creek, Alta-Mira Press: 439–444.
- Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, *et al.*, (2005). Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* **13**: 867-76
- Vossen R (1998). Historical Classification of Khoe (Central Khoisan) Languages of Southern Africa. *African Studies* **57**: 93-106
- Wright JB (1971). Bushman raiders of the Drakensberg, 1840-1870. Pietermaritzburg, University of Natal Press
- Xue FZ, Wang JZ, Hu P and Li GR (2005). The "Kriging" model of spatial genetic structure in human population genetics. *Yi Chuan Xue Bao* **32**: 219-33
- Yao YG, Kong QP, Man XY, Bandelt HJ and Zhang YP (2003). Reconstructing the evolutionary history of China: a caveat about inferences drawn from ancient DNA. *Mol Biol Evol* **20**: 214-9
- YCC (2002). A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* **12**: 339-48
- Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, *et al.*, (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet* **74**: 50-61
- Zivovogel D (1955). Notes on the language of the Eastern Transvaal Bushmen. The disappearing Bushmen of Lake Chrissie: a preliminary survey. Potgieter E F. Pretoria, J.L. van Schaick.
- Zoutendyk A, Kopec AC and Mourant AE (1955). The blood groups of the Hottentot. *Am J Phys Anthropol* **13**: 691-698

9. APPENDICES

Appendix A: Ethics approval

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

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Schlebusch

CLEARANCE CERTIFICATE **PROTOCOL NUMBER M050902**

PROJECT Genetic Variation in Khoisan-Speaking
Populations from Southern Africa

INVESTIGATORS Miss CM Schlebusch

DEPARTMENT Human Genetics

DATE CONSIDERED 05.09.30

DECISION OF THE COMMITTEE* Approved unconditionally

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

DATE 05.10.06 **CHAIRPERSON** 
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable

cc: Supervisor : Prof H Soodyall

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10005, 10th Floor, Senate House, University.

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

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

Division of the Deputy Registrar (Research)

COMMITTEE FOR RESEARCH ON HUMAN SUBJECTS (MEDICAL)

Ref: R14/49 Soodyall

CLEARANCE CERTIFICATE

PROTOCOL NUMBER M980553

PROJECT

Human Genetics Diversity and Disease

INVESTIGATORS

Dr H Soodyall

DEPARTMENT

Specialist Scientist, SAIMR

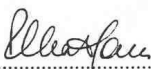
DATE CONSIDERED

980529

DECISION OF THE COMMITTEE *

Approved unconditionally

DATE 980615

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

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

c c Supervisor: Prof T Jenkins

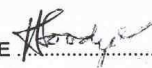
Dept of Human Genetics, SAIMR

Works2\lain0015\HumEth97.wdb\W 980553

DECLARATION OF INVESTIGATOR(S)

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

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

DATE 11/3/98.....**SIGNATURE** .....

PROTOCOL NO.: M 980553

PLEASE QUOTE THE PROTOCOL NUMBER IN ALL ENQUIRIES

Human Research Ethics Committee (Medical)
(formerly Committee for Research on Human Subjects (Medical))

Secretariat: Research Office, Room SH10005, 10th floor, Senate House • Telephone: +27 11 717-1234 • Fax: +27 11 339-5708
Private Bag 3, Wits 2050, South Africa

University
of the Witwatersrand,
Johannesburg



9 September 2004

Dr H Soodyall
Director: MRC/NHLS/Wits Human Genomic Diversity Disease
Research Unit
University
By fax 489-9226

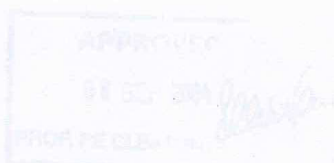
Dear Dr Soodyall

RE: RENEWAL OF PROTOCOL M980553

This letter serves to confirm that the Chairman of the Human Research Ethics Medical has approved the renewal of the abovementioned study for 5 years.

Yours sincerely,

Anisa Keshav
Secretary
Human Research Ethics Committee (Medical)



Appendix B: Recipes for reagents and solutions used

Sucrose-Triton X Lysing buffer

10 ml 1 M Tris-HCl pH8
5 ml 1 M MgCl₂
10 ml Triton-X 100
Make up to 1 L with dH₂O and autoclave
Add 109.5 g sucrose just before use
Keep chilled at 4°C

1 M Tris-HCl

121.1 g Tris
1 L dH₂O
Autoclave

1 M MgCl₂

101.66 g MgCl₂
500 ml dH₂O
Autoclave

T20E5

20 ml 1M Tris-HCl
10 ml 0.5M EDTA pH8
Make up to 1 L with dH₂O and autoclave

0.5 M EDTA

93.06 g EDTA
500 ml dH₂O
pH to 8.0 with NaOH and autoclave

10% SDS

10 g SDS
100 ml dH₂O
Autoclave

Proteinase K (10 mg/ml)

100 mg Proteinase K stock (100 mg/ml)*
10 ml ddH₂O

*Available from Roche Diagnostics

Proteinase-K mix

For 16 extractions:
400 µl 10% SDS
16 µl 0.5 M EDTA
2.8 ml autoclaved dH₂O
Add 800 µl Proteinase K (10 mg/ml stock) just before use

Saturated NaCl

100 ml autoclaved dH₂O

Slowly add 40 g NaCl until absolutely saturated (some NaCl will precipitate out)

Before use, agitate and let NaCl precipitate out

1 X TE buffer

10 ml 1 M Tris-HCl pH8

2 ml 0.5 M EDTA

Make up to 1 L with dH₂O and autoclave

10 X TBE buffer

108 g Tris

55 g Boric acid

7.44 g EDTA

Make up to 1 L with dH₂O and autoclave

1 X TBE (1:10 dilution)

40 ml 10 X TBE

Make up to 200ml with ddH₂O

Bromophenol blue Ficoll dye

50 ml dH₂O

50 g sucrose

1.86 g EDTA

0.1 g bromophenol blue

10 g Ficoll

Dissolve

Adjust volume to 100 ml with dH₂O, stir overnight

pH to 8.0

Filter through Whatmann filter paper

Store at room temperature

10 mg/ml Ethidium bromide (EtBr)

Add 1 g of ethidium bromide to

100 ml of ddH₂O

Stir for several hours until completely dissolved

Store wrapped in aluminum foil at 4°C

1kb size standard

285 µl 1kb ladder (GibcoBRL)

143 µl Ficoll dye

2 400 µl 1 X TE

10 mg/ml BSA

1 g BSA

10 ml ddH₂O

Aliquot into 1 ml amounts and store at -20°C

2.5mM dNTPs

Use 100 mM premade stocks of dATP, dGTP, dCTP and dTTP (GibcoBRL)

10 µl of each stock dNTP + 360 µl sterile ddH₂O = 400 µl of 2.5 mM dNTPs

2.5 mM Spermidine (Sigma)

Add 6.887 ml ddH₂O to 1 g of Spermidine to make a 1 M stock

Dilute 1 in 400 to 2.5 mM for use

Appendix C: Physical distance matrix (in km) between Khoe-San and Coloured groups

	CAC	COL	CNC	GUG	JOH	KAR	KHO	KWE	NAM	XUN
CAC	0.000									
COL	659.7942	0.000								
CNC	761.8532	591.0919	0.000							
GUG	1241.469	787.2487	537.4522	0.000						
JOH	1570.437	1321.059	821.7263	624.5947	0.000					
KAR	659.7942	0.000	591.0919	787.2487	1321.059	0.000				
KHO	761.8532	591.0919	0.000	537.4522	821.7263	591.0919	0.000			
KWE	1854.39	1501.823	1092.547	722.4106	359.1735	1501.823	1092.547	0.000		
NAM	1248.44	1208.281	618.8586	787.1903	484.7312	1208.281	618.8586	843.9047	0.000	
XUN	2120.967	1944.929	1412.054	1244.793	629.6031	1944.929	1412.054	629.6031	884.9136	0.000

Appendix D: Details of SNP used in autosomal analyses

Name in thesis	Chromosome	Group on chromosome	SNP ID (hCV)	SNP ID (rs)	Base Position	Distance from previous marker	Yoruba MAF	Afr American MAF
chr01-1-1	1	1	hCV29985869	rs7523071	185581438		32	N/A
* chr01-1-2	1	1	hCV8349921	rs1445667	185582609	1171	32	N/A
chr01-1-3	1	1	hCV8352908	rs1445670	185587536	4927	42	N/A
chr01-1-4	1	1	hCV30688593	rs6660605	185594475	6939	32	N/A
chr01-1-5	1	1	hCV30688596	rs6666285	185594691	216	44	N/A
chr01-2-1	1	2	hCV26908697	rs6702432	243839090		26	N/A
chr01-2-2	1	2	hCV28021091	rs7366424	243845655	6565	27	N/A
chr01-2-3	1	2	hCV30447094	rs7555211	243848391	2736	36	N/A
chr01-2-4	1	2	hCV12075636	rs1954187	243851004	2613	9	N/A
* chr01-2-5	1	2	hCV30382617	rs10399826	243861576	10572	38	N/A
chr02-1-1	2	1	hCV15781272	rs2373901	40769550		50	N/A
chr02-1-2	2	1	hCV8809743	rs882007	40781807	12257	20	26
chr02-1-3	2	1	hCV29048038	rs6755751	40782253	446	47	N/A
chr02-1-4	2	1	hCV1296264	rs3851315	40785189	2936	37	N/A
chr02-1-5	2	1	hCV26252361	rs11124754	40787764	2575	32	N/A
chr02-2-1	2	2	hCV29410368	rs6743609	78370721		27	N/A
chr02-2-2	2	2	hCV29410366	rs6715934	78379067	8346	23	N/A
chr02-2-3	2	2	hCV16127289	rs2839828	78382969	3902	39	N/A
chr02-2-4	2	2	hCV11464467	rs1837144	78383601	632	50	N/A
chr02-2-5	2	2	hCV11464466	rs1816652	78388857	5256	17	N/A
chr03-1-1	3	1	hCV11749718	rs1987888	4053654		24	N/A
chr03-1-2	3	1	hCV8827003	rs1087817	4063576	9922	33	N/A
chr03-1-3	3	1	hCV626367	rs317575	4063809	233	N/A	N/A
chr03-1-4	3	1	hCV626362	rs317530	4069293	5484	34	40
chr03-1-5	3	1	hCV626353	rs317534	4074043	4750	49	N/A
chr03-2-1	3	2	hCV27956340	rs4624549	189144204		48	N/A
chr03-2-2	3	2	hCV3244174	rs2590451	189147479	3275	42	N/A
chr03-2-3	3	2	hCV1058808	rs567713	189151423	3944	47	N/A
chr03-2-4	3	2	hCV15917716	rs2679506	189154725	3302	28	N/A
chr03-2-5	3	2	hCV3244161	rs522833	189160082	5357	27	N/A
chr04-1-1	4	1	hCV2967242	rs9998475	13325188		26	N/A
chr04-1-2	4	1	hCV2967234	rs1352786	13326354	1166	26	42
chr04-1-3	4	1	hCV1192506	rs1948354	13334081	7727	26	N/A
chr04-1-4	4	1	hCV1192503	rs6837122	13335534	1453	24	41
chr04-1-5	4	1	hCV7562999	rs1032358	13338502	2968	19	N/A
chr04-2-1	4	2	hCV29608728	rs10084822	172054953		29	N/A
chr04-2-2	4	2	hCV30204213	rs9312493	172061519	6566	31	N/A
chr04-2-3	4	2	hCV30114165	rs10004230	172066255	4736	17	N/A
chr04-2-4	4	2	hCV8242322	rs1403213	172075840	9585	43	N/A
* chr04-2-5	4	2	hCV30600558	rs10002204	172096780	20940	21	N/A
chr05-1-1	5	1	hCV7447360	rs1366370	66593667		39	N/A
chr05-1-2	5	1	hCV27915872	rs755877	66593979	312	45	N/A
chr05-1-3	5	1	hCV7447351	rs1593948	66594316	337	47	N/A
chr05-1-4	5	1	hCV11824955	rs7715561	66598715	4399	37	N/A
chr05-1-5	5	1	hCV2937282	rs919308	66604140	5425	17	N/A
* chr05-2-1	5	2	hCV26117944	rs165073	163963822		31	N/A
chr05-2-2	5	2	hCV7522487	rs1363174	163978188	14366	N/A	N/A
chr05-2-3	5	2	hCV1393057	rs250597	163980289	2101	30	N/A

chr05-2-4	5	2	hCV30220715	rs10515884	163985604	5315	41	N/A
chr05-2-5	5	2	hCV7522494	rs1421905	163990354	4750	38	N/A
chr06-1-1	6	1	hCV30355724	rs9505359	809219		22	N/A
chr06-1-2	6	1	hCV1819928	rs884126	815244	6025	27	N/A
chr06-1-3	6	1	hCV8773399	rs885450	815563	319	N/A	N/A
chr06-1-4	6	1	hCV1819934	rs873560	820559	4996	24	N/A
chr06-1-5	6	1	hCV1819941	rs6916756	825467	4908	23	N/A
chr06-2-1	6	2	hCV30164637	rs6912046	79193277		45	N/A
chr06-2-2	6	2	hCV15868784	rs2223722	79197714	4437	46	N/A
chr06-2-3	6	2	hCV7546896	rs926654	79202638	4924	36	43
chr06-2-4	6	2	hCV30416745	rs9361404	79205477	2839	21	N/A
chr06-2-5	6	2	hCV29496547	rs9448411	79208314	2837	32	N/A
chr07-1-1	7	1	hCV3253650	rs2592859	35206935		31	N/A
chr07-1-2	7	1	hCV1071172	rs731015	35212110	5175	25	N/A
chr07-1-3	7	1	hCV16249550	rs2541911	35216715	4605	37	N/A
chr07-1-4	7	1	hCV16249554	rs2250212	35221258	4543	7	N/A
chr07-1-5	7	1	hCV3253622	rs2592848	35230892	9634	22	N/A
chr07-2-1	7	2	hCV30792597	rs7806350	144859843		49	N/A
chr07-2-2	7	2	hCV7434566	rs1523729	144867554	7711	27	37
chr07-2-3	7	2	hCV15843844	rs2888245	144871885	4331	24	N/A
chr07-2-4	7	2	hCV7435229	rs1523723	144877013	5128	20	N/A
* chr07-2-5	7	2	hCV30792607	rs6954212	144880096	3083	28	N/A
chr08-1-1	8	1	hCV8947909	rs871565	18152103		39	N/A
chr08-1-2	8	1	hCV8947923	rs1493029	18165651	13548	29	N/A
chr08-1-3	8	1	hCV8947937	rs902960	18168085	2434	38	N/A
* chr08-1-4	8	1	hCV29066331	rs7846103	18170309	2224	22	N/A
chr08-1-5	8	1	hCV16075982	rs2131422	18178912	8603	23	N/A
chr08-2-1	8	2	hCV11456221	rs2385226	126751178		17	N/A
chr08-2-2	8	2	hCV2761265	rs4871628	126752121	943	27	N/A
chr08-2-3	8	2	hCV8449160	rs7838054	126753324	1203	27	N/A
chr08-2-4	8	2	hCV2761254	rs1159478	126757397	4073	N/A	N/A
chr08-2-5	8	2	hCV2761245	rs7460157	126761038	3641	22	N/A
chr09-1-1	9	1	hCV1617703	rs10966574	24919668		42	N/A
chr09-1-2	9	1	hCV3157880	rs7025715	24924491	4823	37	N/A
chr09-1-3	9	1	hCV1617701	rs7871011	24925087	596	47	N/A
chr09-1-4	9	1	hCV26305217	rs4085752	24931125	6038	14	N/A
chr09-1-5	9	1	hCV8767627	rs1461333	24936349	5224	42	N/A
chr09-2-1	9	2	hCV11489339	rs1927239	123675437		21	N/A
chr09-2-2	9	2	hCV16242136	rs2489161	123678034	2597	28	N/A
chr09-2-3	9	2	hCV995477	rs562239	123679804	1770	21	N/A
chr09-2-4	9	2	hCV29392986	rs4836945	123689332	9528	21	N/A
chr09-2-5	9	2	hCV16069779	rs2768818	123690135	803	28	N/A
chr10-1-1	10	1	hCV29522539	rs9663972	60527538		20	N/A
chr10-1-2	10	1	hCV31345052	rs6481457	60531364	3826	42	N/A
chr10-1-3	10	1	hCV908092	rs733341	60533393	2029	46	N/A
chr10-1-4	10	1	hCV31345143	rs11006373	60539023	5630	45	N/A
chr10-1-5	10	1	hCV31345171	rs7921026	60541895	2872	27	N/A
chr10-2-1	10	2	hCV11207816	rs7094944	109799612		37	N/A
chr10-2-2	10	2	hCV1798848	rs10509859	109803462	3850	23	30
chr10-2-3	10	2	hCV1798849	rs1125798	109808286	4824	25	31
chr10-2-4	10	2	hCV1798851	rs7073564	109813235	4949	23	N/A
chr10-2-5	10	2	hCV1798854	rs1556592	109819760	6525	35	42
chr11-1-1	11	1	hCV29137013	rs7124156	13198502		42	N/A
chr11-1-2	11	1	hCV9600088	rs900141	13204100	5598	20	N/A

chr11-1-3	11	1	hCV1870543	rs900142	13204831	731	22	N/A
chr11-1-4	11	1	hCV30567849	rs7117211	13205223	392	32	N/A
chr11-1-5	11	1	hCV7667097	rs7107711	13212114	6891	43	N/A
chr11-2-1	11	2	hCV11481013	rs2042599	127235817		34	N/A
chr11-2-2	11	2	hCV11481007	rs1812931	127240375	4558	30	N/A
chr11-2-3	11	2	hCV7504970	rs1364777	127242208	1833	27	N/A
chr11-2-4	11	2	hCV2890056	rs1107869	127249002	6794	27	N/A
chr11-2-5	11	2	hCV31697360	rs10893778	127253038	4036	27	N/A
chr12-1-1	12	1	hCV7562390	rs917589	3412660		17	N/A
chr12-1-2	12	1	hCV7562396	rs917587	3412936	276	16	N/A
chr12-1-3	12	1	hCV2649193	rs2878578	3413587	651	47	N/A
chr12-1-4	12	1	hCV29394818	rs6489468	3421275	7688	34	N/A
chr12-1-5	12	1	hCV2649182	rs7961141	3424976	3701	45	N/A
chr12-2-1	12	2	hCV2801082	rs855228	101400231		29	N/A
chr12-2-2	12	2	hCV7570428	rs855224	101405390	5159	35	N/A
chr12-2-3	12	2	hCV7570434	rs855218	101409109	3719	35	N/A
chr12-2-4	12	2	hCV7570449	rs855211	101413277	4168	32	N/A
chr12-2-5	12	2	hCV3061163	rs35746	101417107	3830	47	N/A
chr13-1-1	13	1	hCV1620102	rs4769191	21547069		35	N/A
chr13-1-2	13	1	hCV7556053	rs1323170	21547219	150	19	N/A
chr13-1-3	13	1	hCV30332355	rs4770238	21548179	960	37	N/A
chr13-1-4	13	1	hCV1620098	rs9316743	21548512	333	45	N/A
chr13-1-5	13	1	hCV7556051	rs1323172	21550247	1735	44	N/A
chr13-2-1	13	2	hCV509921	rs978089	85554112		21	N/A
* chr13-2-2	13	2	hCV509923	rs4910994	85559270	5158	41	N/A
chr13-2-3	13	2	hCV7508241	rs1029143	85563006	3736	41	36
chr13-2-4	13	2	hCV30569079	rs9594117	85578891	15885	20	N/A
chr13-2-5	13	2	hCV9462203	rs1413441	85580898	2007	19	N/A
chr14-1-1	14	1	hCV15790014	rs2383584	33849679		21	N/A
chr14-1-2	14	1	hCV29357552	rs7143582	33852799	3120	33	N/A
chr14-1-3	14	1	hCV1453684	rs1958572	33858595	5796	47	N/A
chr14-1-4	14	1	hCV1453694	rs1958574	33867066	8471	15	N/A
chr14-1-5	14	1	hCV1453706	rs1958579	33870654	3588	13	26
chr14-2-1	14	2	hCV3244666	rs1241743	91751928		40	N/A
chr14-2-2	14	2	hCV3244664	rs1241745	91752315	387	36	N/A
chr14-2-3	14	2	hCV3244656	rs1956413	91753943	1628	44	N/A
chr14-2-4	14	2	hCV11666013	rs1956414	91758924	4981	40	N/A
chr14-2-5	14	2	hCV7585435	rs1741443	91774327	15403	47	N/A
chr15-1-1	15	1	hCV8926261	rs722150	31201795		N/A	42
chr15-1-2	15	1	hCV9960323	rs4780082	31202774	979	23	N/A
chr15-1-3	15	1	hCV11671510	rs1988447	31204618	1844	14	N/A
chr15-1-4	15	1	hCV29223603	rs7181962	31204650	32	46	N/A
chr15-1-5	15	1	hCV9960256	rs8023846	31211066	6416	35	N/A
chr15-2-1	15	2	hCV9708740	rs920921	66573339		41	N/A
chr15-2-2	15	2	hCV9708750	rs1373697	66577067	3728	35	N/A
chr15-2-3	15	2	hCV9708758	rs895133	66580703	3636	34	N/A
chr15-2-4	15	2	hCV15809641	rs2084032	66582870	2167	37	N/A
chr15-2-5	15	2	hCV9708767	rs895131	66583554	684	27	N/A
chr16-1-1	16	1	hCV11624551	rs1848824	61630443		44	N/A
chr16-1-2	16	1	hCV2281952	rs153322	61631942	1499	50	N/A
chr16-1-3	16	1	hCV2281956	rs153341	61644707	12765	23	N/A
chr16-1-4	16	1	hCV29048177	rs1605960	61655814	11107	27	N/A
chr16-1-5	16	1	hCV2281807	rs198007	61678146	22332	28	N/A
chr16-2-1	16	2	hCV1446720	rs1510205	84851316		17	N/A

chr16-2-2	16	2	hCV26612563	rs2883250	84859632	8316	22	N/A
chr16-2-3	16	2	hCV1521231	rs2696815	84859844	212	40	N/A
chr16-2-4	16	2	hCV31422718	rs717482	84862498	2654	16	N/A
chr16-2-5	16	2	hCV8898710	rs1027910	84866445	3947	17	N/A
chr17-1-1	17	1	hCV7596153	rs2007643	52084308		38	N/A
chr17-1-2	17	1	hCV2641904	rs6503752	52088714	4406	38	N/A
chr17-1-3	17	1	hCV2297056	rs714832	52093256	4542	29	N/A
chr17-1-4	17	1	hCV29726775	rs10491158	52099116	5860	17	N/A
chr17-1-5	17	1	hCV7596175	rs1019117	52103128	4012	27	N/A
chr17-2-1	17	2	hCV29084850	rs7222022	66763060		38	N/A
chr17-2-2	17	2	hCV2574303	rs2158906	66769428	6368	21	N/A
chr17-2-3	17	2	hCV6785	rs724856	66776439	7011	N/A	N/A
chr17-2-4	17	2	hCV16151352	rs2190461	66787482	11043	48	N/A
chr17-2-5	17	2	hCV26366590	rs6501466	66789877	2395	36	N/A
chr18-1-1	18	1	hCV15866228	rs2940757	34847593		32	N/A
chr18-1-2	18	1	hCV15873100	rs2958610	34848055	462	45	N/A
chr18-1-3	18	1	hCV7458925	rs1509219	34852830	4775	32	N/A
chr18-1-4	18	1	hCV30437001	rs9304198	34854813	1983	17	N/A
chr18-1-5	18	1	hCV28986615	rs8083419	34856469	1656	47	N/A
chr18-2-1	18	2	hCV703794	rs165130	73464384		37	35
chr18-2-2	18	2	hCV3033039	rs905443	73464575	191	40	N/A
chr18-2-3	18	2	hCV738436	rs165128	73464782	207	18	N/A
chr18-2-4	18	2	hCV3033031	rs9952646	73470415	5633	37	N/A
chr18-2-5	18	2	hCV11740294	rs2407139	73472582	2167	21	N/A
chr19-1-1	19	1	hCV29353745	rs7256520	36812013		42	N/A
chr19-1-2	19	1	hCV29353740	rs8100570	36814702	2689	42	N/A
chr19-1-3	19	1	hCV9608785	rs892210	36817849	3147	N/A	N/A
chr19-1-4	19	1	hCV31999332	rs8112540	36818052	203	31	N/A
chr19-1-5	19	1	hCV29353735	rs8101359	36822617	4565	33	N/A
chr19-2-1	19	2	hCV8710254	rs1654338	43228193		24	N/A
chr19-2-2	19	2	hCV2380087	rs734204	43231828	3635	25	30
chr19-2-3	19	2	hCV8710218	rs941037	43235466	3638	25	N/A
chr19-2-4	19	2	hCV2825747	rs1725467	43235743	277	25	N/A
chr19-2-5	19	2	hCV8710210	rs1725504	43238719	2976	35	N/A
chr20-1-1	20	1	hCV29643199	rs6085916	7112725		21	N/A
chr20-1-2	20	1	hCV8954459	rs1033604	7126839	14114	22	N/A
chr20-1-3	20	1	hCV8954461	rs1016264	7128675	1836	N/A	N/A
chr20-1-4	20	1	hCV30166365	rs6133401	7129330	655	22	N/A
chr20-1-5	20	1	hCV29751510	rs6117693	7135874	6544	22	N/A
chr20-2-1	20	2	hCV3249308	rs2424383	21514717		12	N/A
chr20-2-2	20	2	hCV2808411	rs1014889	21518837	4120	31	45
chr20-2-3	20	2	hCV8890830	rs1014890	21519183	346	N/A	N/A
chr20-2-4	20	2	hCV2808414	rs1074606	21521722	2539	30	N/A
chr20-2-5	20	2	hCV30615207	rs6035902	21530338	8616	36	N/A
chr21-1-1	21	1	hCV534386	rs150210	18284183		37	N/A
chr21-1-2	21	1	hCV569469	rs197562	18285788	1605	26	N/A
chr21-1-3	21	1	hCV2959026	rs2824593	18290483	4695	22	N/A
chr21-1-4	21	1	hCV534395	rs158077	18294083	3600	49	N/A
chr21-1-5	21	1	hCV9488991	rs1505265	18296572	2489	40	N/A
chr21-2-1	21	2	hCV31154097	rs8131079	24467571		37	N/A
chr21-2-2	21	2	hCV29107061	rs7280999	24469539	1968	37	N/A
chr21-2-3	21	2	hCV3236789	rs1024318	24475083	5544	22	29
chr21-2-4	21	2	hCV3236790	rs1910605	24480779	5696	49	40
chr21-2-5	21	2	hCV2787192	rs1910622	24483502	2723	49	N/A

chr22-1-1	22	1	hCV2463578	rs137462	31940397		N/A	23
chr22-1-2	22	1	hCV29637994	rs9306274	31940642	245	32	N/A
chr22-1-3	22	1	hCV2221204	rs137472	31945136	4494	32	28
chr22-1-4	22	1	hCV2221205	rs118033	31945610	474	18	N/A
chr22-1-5	22	1	hCV2221209	rs137475	31951775	6165	27	41
chr22-2-1	22	2	hCV15796647	rs2413378	34796843		17	N/A
chr22-2-2	22	2	hCV1088397	rs715550	34797853	1010	25	35
chr22-2-3	22	2	hCV1088400	rs715546	34798045	192	24	N/A
chr22-2-4	22	2	hCV30226396	rs7286844	34804171	6126	20	N/A
chr22-2-5	22	2	hCV1088401	rs739203	34805381	1210	42	N/A
					AVE	4347.0		
					STD	3730.8		
					Min	192		
					Max	22332		

* Excluded - Poor Quality

Appendix E: Haplotype list of HVR I and HVR II variation

HT	N	HG	HVR 1 Variant sites	HVR 2 Variant sites	KAR	COL	CAC	KHO	CNC	XEG	DUM	NAM	GUG	NAR	JOH	XUN	KWE	DRC	HER	SOT	SWZ	ZUX	AFR	EUR	IND
Ht_001	1	CRS																							
Ht_002	1	NEAN	16037 A-G; 16078 A-G; 16129 G-A; 16139 A-T; 16148 C-T; 16154 T-C; 16169 C-T; 16182 A-C; 16183 A-C; 16189 T-C; 16209 T-C; 16223 C-T; 16230 A-G; 16234 C-T; 16244 G-A; 16256 C-A; 16258 A-G; 16262 C-T; 16263 insA; 16278 C-T; 16299 A-G; 16311 T-C; 16320 C-T; 16362 T-C; 16400 C-T; 16519 T-C	73 A-G; 146 T-C; 150 C-T; 152 T-C; 189 A-G; 200 A-G; 243 A-G; 245 T-C; 247 G-A; 262 C-T; 263 A-G; 417 G-A; 438 C-T; 520 delCACAC; 547 A-G																					
Ht_003	2	L0a1b	16093 T-C; 16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-G; 16311 T-C; 16320 C-T	93 A-G; 95 A-C; 152 T-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 523 delAC																		2			
Ht_004	1	L0a1b	16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-C; 16311 T-C; 16320 C-T	93 A-G; 95 A-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 523 delAC		1																			
Ht_005	1	L0a1b	16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-G; 16311 T-C; 16320 C-T; 16368 T-C	93 A-G; 95 A-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 523 delAC			1																		
Ht_006	1	L0a1b	16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-G; 16311 T-C; 16320 C-T	89 T-C; 93 A-G; 95 A-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 507 T-C; 523 delAC																	1				
Ht_007	1	L0a1b	16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-G; 16311 T-C; 16320 C-T	89 T-C; 93 A-G; 95 A-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 523 delAC		1																			
Ht_008	1	L0a1b	16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-G; 16311 T-C; 16320 C-T	93 A-G; 95 A-C; 152 T-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 523 delAC		1																			
Ht_009	6	L0a1b	16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-G; 16311 T-C; 16320 C-T	93 A-G; 95 A-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 523 delAC			2			1									2		1				
Ht_010	1	L0a1b	16129 G-A; 16148 C-T; 16168 C-T; 16172 T-C; 16187 C-T; 16188 C-G; 16189 T-C; 16223 C-T; 16230 A-G; 16278 C-T; 16293 A-G; 16311 T-C; 16320 C-T; 16344 C-T; 16519 T-C	93 A-G; 95 A-C; 185 G-A; 189 A-G; 236 T-C; 247 G-A; 263 A-G; 523 delAC												1									

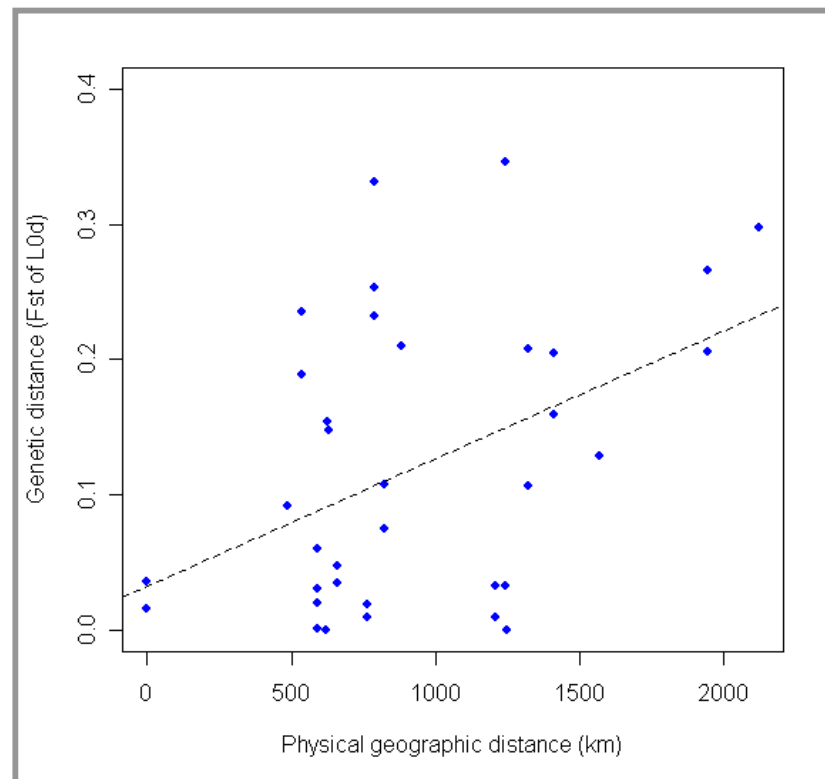
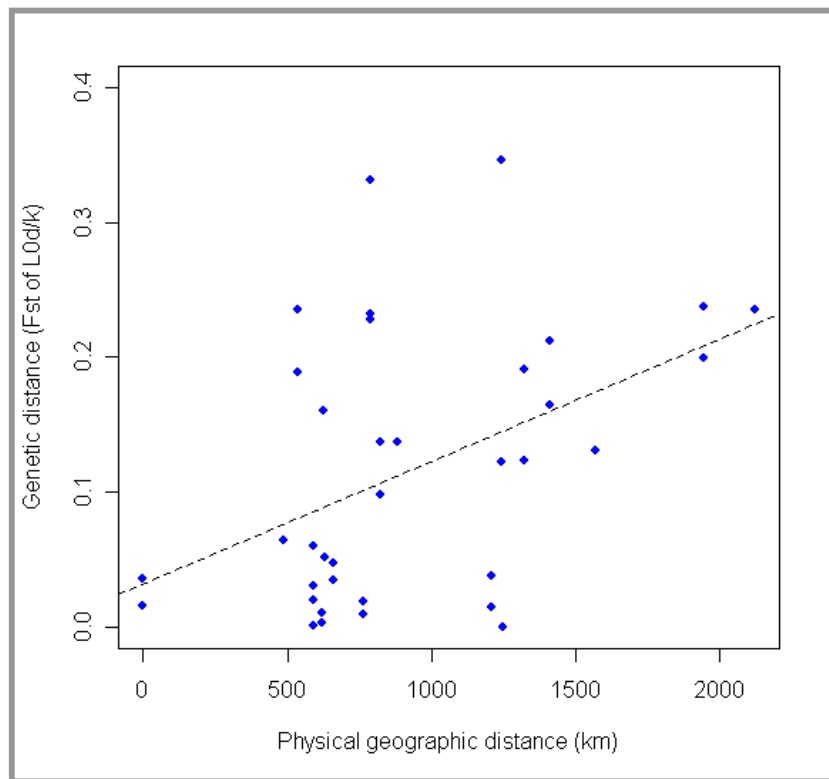
[illegible]

Ht_126	8	L0k1	16166 A-C; 16172 T-C; 16189 T-C; 16209 T-C; 16214 C-T; 16223 C-T; 16230 A-G; 16278 C-T; 16291 C-G; 16311 T-C; 16519 T-C	73 A-G; 146 T-C; 152 T-C; 189 A-G; 195 T-C; 198 C-T; 204 T-C; 207 G-A; 247 G-A; 523 delAC												3	5									
Ht_127	1	L0k1	16166 A-C; 16172 T-C; 16187 C-T; 16189 T-C; 16209 T-C; 16214 C-T; 16223 C-T; 16230 A-G; 16278 C-T; 16291 C-A; 16311 T-C; 16519 T-C	73 A-G; 146 T-C; 152 T-C; 189 A-G; 195 T-C; 198 C-T; 207 G-A; 247 G-A; 523 delAC													1									
Ht_128	4	L0k1	16166 A-C; 16172 T-C; 16187 C-T; 16189 T-C; 16209 T-C; 16214 C-T; 16223 C-T; 16230 A-G; 16278 C-T; 16291 C-G; 16311 T-C; 16519 T-C	73 A-G; 146 T-C; 152 T-C; 189 A-G; 195 T-C; 198 C-T; 204 T-C; 207 G-A; 247 G-A; 523 delAC								1				3										
Ht_129	13	L0k1	16166 A-C; 16172 T-C; 16187 C-T; 16189 T-C; 16209 T-C; 16214 C-T; 16223 C-T; 16230 A-G; 16278 C-T; 16291 C-G; 16311 T-C; 16519 T-C	73 A-G; 146 T-C; 152 T-C; 189 A-G; 195 T-C; 198 C-T; 207 G-A; 247 G-A; 523 delAC								1				4	7						1			
Ht_130	5	L0k1	16166 A-C; 16172 T-C; 16187 C-T; 16189 T-C; 16209 T-C; 16214 C-T; 16223 C-T; 16230 A-G; 16278 C-T; 16291 C-G; 16311 T-C; 16519 T-C	73 A-G; 146 T-C; 152 T-C; 189 A-G; 195 T-C; 198 C-T; 247 G-A; 523 delAC													5									
Ht_131	3	L1b	16126 T-C; 16187 C-T; 16189 T-C; 16223 C-T; 16264 C-T; 16270 C-T; 16278 C-T; 16311 T-C; 16519 T-C	73 A-G; 152 T-C; 182 C-T; 185 G-T; 195 T-C; 247 G-A; 263 A-G; 357 A-G; 523 delAC		1													2							
Ht_132	1	L1c1	16129 G-A; 16172 T-C; 16173 C-T; 16188 C-A; 16189 T-C; 16223 C-T; 16256 C-T; 16278 C-T; 16293 A-G; 16294 C-T; 16311 T-C; 16360 C-T; 16368 T-C; 16519 T-C	73 A-G; 151 C-T; 152 T-C; 182 C-T; 186 C-A; 189 A-G; 195 T-C; 198 C-T; 247 delG; 263 A-G; 297 A-G; 316 G-A; 523 delAC																		1				
Ht_133	2	L1c2	16129 G-A; 16187 C-T; 16189 T-C; 16214 C-T; 16223 C-T; 16265 A-C; 16278 C-T; 16286 C-A; 16291 C-T; 16294 C-T; 16311 T-C; 16360 C-T; 16519 T-C; 16527 C-T	73 A-G; 151 C-T; 152 T-C; 182 C-T; 186 C-A; 189 A-C; 195 T-C; 198 C-T; 247 G-A; 263 A-G; 297 A-G; 316 G-A; 513 G-A								2														
Ht_134	1	L1c2	16172 T-C; 16187 C-T; 16189 T-C; 16223 C-T; 16265 A-C; 16278 C-T; 16286 C-G; 16294 C-T; 16311 T-C; 16360 C-T; 16519 T-C; 16527 C-T	73 A-G; 151 C-T; 152 T-C; 182 C-T; 186 C-A; 189 A-C; 195 T-C; 198 C-T; 247 G-A; 263 A-G; 297 A-G; 316 G-A; 385 A-G; 471 T-C; 523 delAC																1						
Ht_135	1	L1c2	16108 C-T; 16129 G-A; 16187 C-T; 16189 T-C; 16260 C-T; 16265 A-C; 16278 C-T; 16286 C-A; 16294 C-T; 16311 T-C; 16360 C-T; 16519 T-C; 16527 C-T	73 A-G; 151 C-T; 152 T-C; 182 C-T; 186 C-A; 189 A-C; 195 T-C; 198 C-T; 247 G-A; 263 A-G; 297 A-G; 316 G-A			1																			
Ht_136	1	L1c2b1a	16071 C-T; 16129 G-A; 16145 G-A; 16187 C-T; 16189 T-C; 16213 G-A; 16223 C-T; 16234 C-T; 16265 A-C; 16278 C-T; 16286 C-G; 16294 C-T; 16311 T-C; 16360 C-T; 16365 C-T; 16527 C-T	73 A-G; 151 C-T; 152 T-C; 182 C-T; 186 C-A; 189 A-C; 195 T-C; 198 C-T; 247 G-A; 263 A-G; 297 A-G; 316 G-A													1									

[illegible]

[illegible]

Appendix F: Graphs – Physical vs. Genetic distance (L0d/k sequences and L0d sequences)



Appendix G: Haplotype list of 12 marker Y-STR panel

HT	HG	DYS19	DYS390	DYS391	DYS392	DYS393	DYS385a	DYS385b	DYS389I	DYS389II	DYS437	DYS438	DYS439	KAR	COL	CAC	KHO	CNC	XEG	NAM	GUG	NAR	JOH	XUN	KWE	DRC	HER	SOT	SWZ	ZUX	AFR	EUR	IND	TOT
Ht001	A-M114	12	21	10	14	11	15	17	13	27	16	10	14				1																	1
Ht002	A-M114	12	21	10	14	11	15	18	13	27	16	10	13										2											2
Ht003	A-M114	12	21	10	14	11	15	18	13	27	16	10	14										1											1
Ht004	A-M114	12	21	10	14	11	17	17	13	27	16	10	11											1										1
Ht005	A-M114	12	21	10	15	11	17	17	13	27	16	10	11											1										1
Ht006	A-M14	12	21	10	13	11	14	17	14	29	15	10	11											1										1
Ht007	A-M14	12	21	10	13	11	15	17	14	29	16	11	11										1											1
Ht008	A-M14	12	21	10	13	11	16	16	14	29	16	10	10										1											1
Ht009	A-M14	12	22	10	14	11	17	18	13	27	16	9	13									1												1
Ht010	A-M51	14	21	10	10	13	14	18	13	30	15	11	12	1																				1
Ht011	A-M51	14	23	10	11	13	14.2	15	14	30	14	9	11												1									1
Ht012	A-M51	15	18	10	10	13	14	16	13	29	15	11	13				1																	1
Ht013	A-M51	15	18	10	10	13	15	16	12	30	15	10	11				2																	2
Ht014	A-M51	15	19	10	10	13	15	15	13	30	17	11	11					1																1
Ht015	A-M51	15	19	10	10	13	15	16	12	29	16	11	11					1																1
Ht016	A-M51	15	19	10	10	13	15	16	13	30	15	11	11											3										3
Ht017	A-M51	15	20	10	10	13	15	16	12	29	16	11	11					1																1
Ht018	A-M51	15	21	10	10	13	14	17	12	29	15	11	12		1																			1
Ht019	A-M51	15	22	10	10	13	14	19	14	30	15	12	12																1					1
Ht020	A-M51	15	22	10	10	13	16	17	14	31	15	12	13	1																				1
Ht021	A-M51	16	18	10	10	13	14	16	12	30	14	11	10											1										1
Ht022	A-M51	16	18	11	10	13	14	18	12	28	14	11	10										1											1
Ht023	A-M51	16	19	10	10	13	15	16	13	30	14	11	11					1																1
Ht024	A-M51	16	19	10	9	13	14	20	12	28	14	11	10										1											1
Ht025	A-M51	16	19	11	10	13	14	15	13	29	14	9	11															1						1
Ht026	A-M51	16	20	11	10	13	15	17	12	30	17	11	11				1																	1
Ht027	A-M51	16	22	10	10	13	15	17	13	30	14	12	13				1			1														2
Ht028	A-M51	16	22	11	10	13	17	17	12	29	14	11	12					1			1													2
Ht029	A-M51	16	22	11	10	13	17	18	12	29	14	11	12															1						1
Ht030	A-M51	16	22	11	11	14	13.2	14.2	13	30	14	11	11											4										4
Ht031	A-M51	16	22	11	11	14	13.2	14.2	13	30	14	11	12											4										4

[illegible]

[illegible]

[illegible]

343

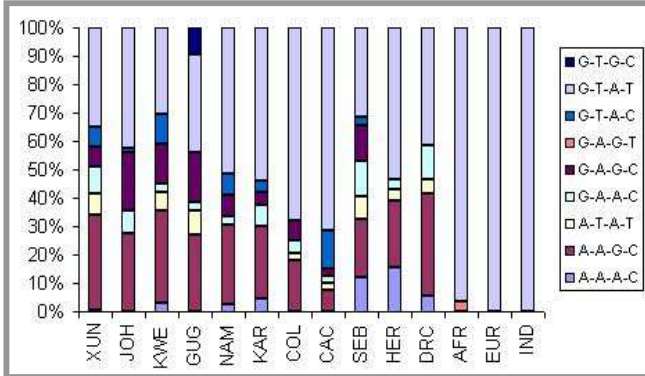
[illegible]

345

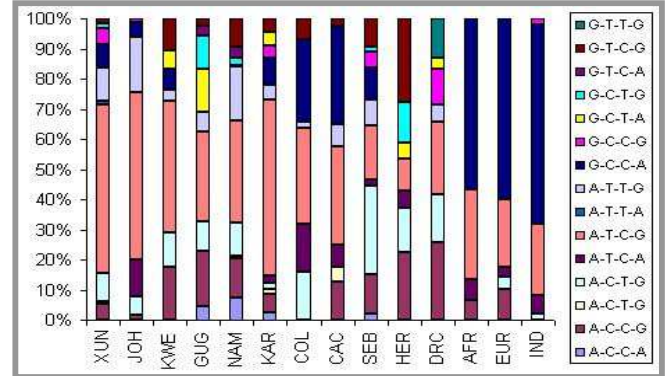
346

Appendix H: Bar charts showing haplotype frequencies for 44 inferred short haplotypes

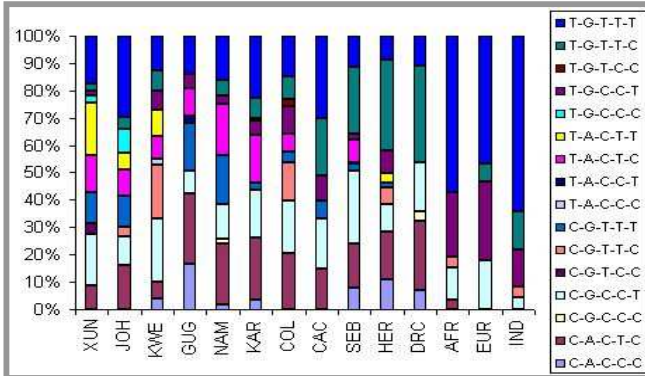
01-01



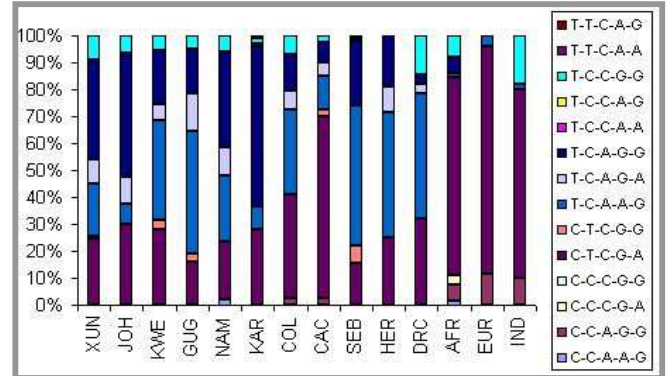
01-02



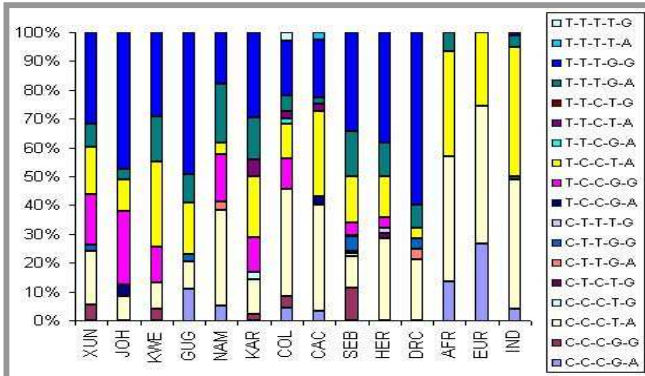
02-01



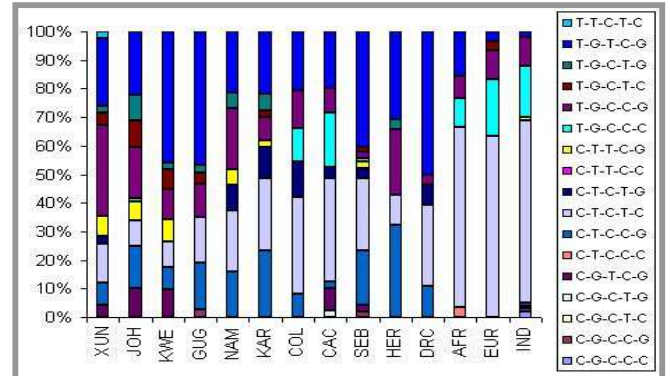
02-02



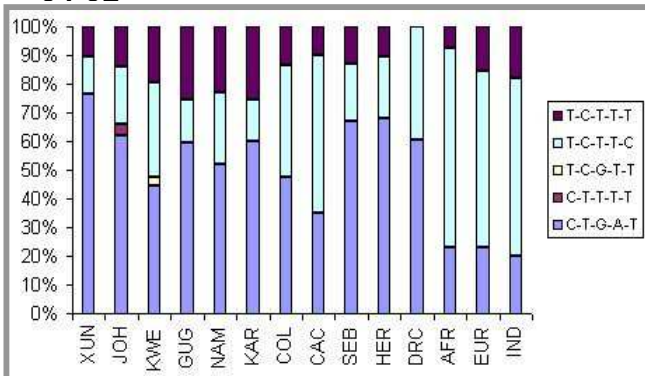
03-01



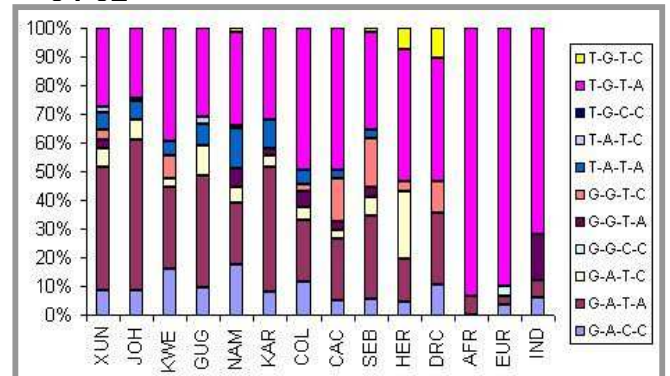
03-02



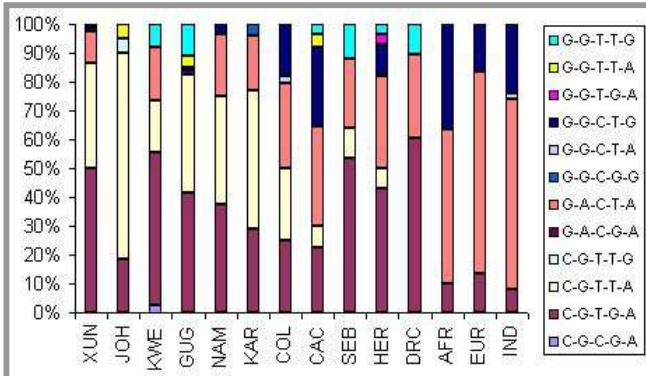
04-01



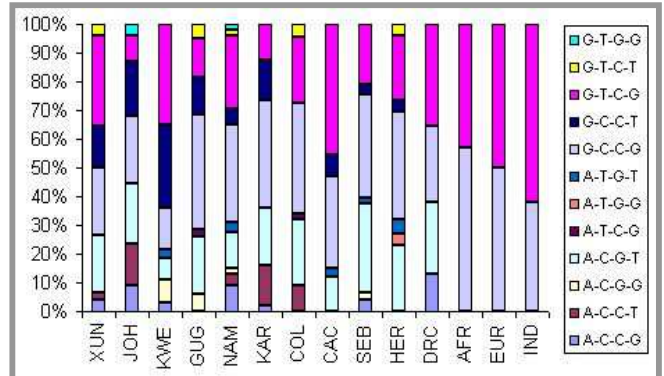
04-02



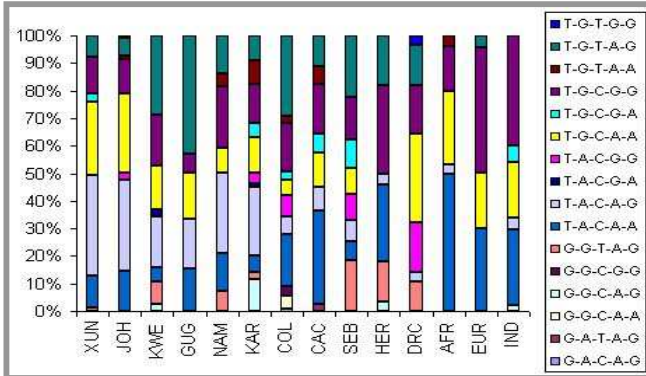
05-01



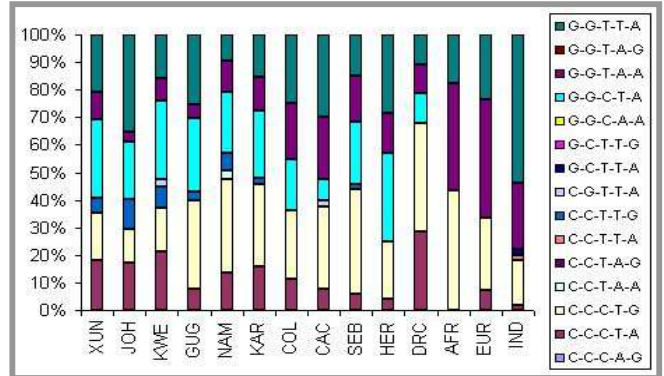
05-02



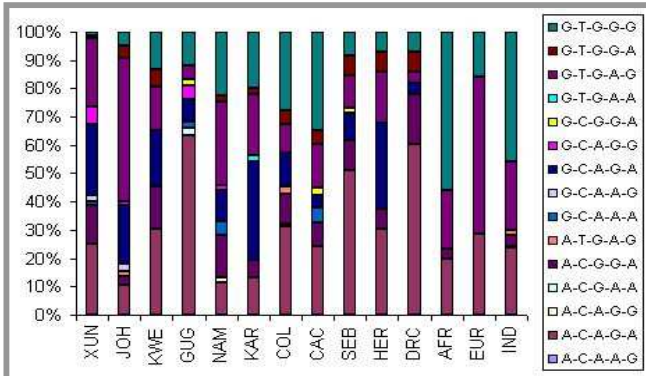
06-01



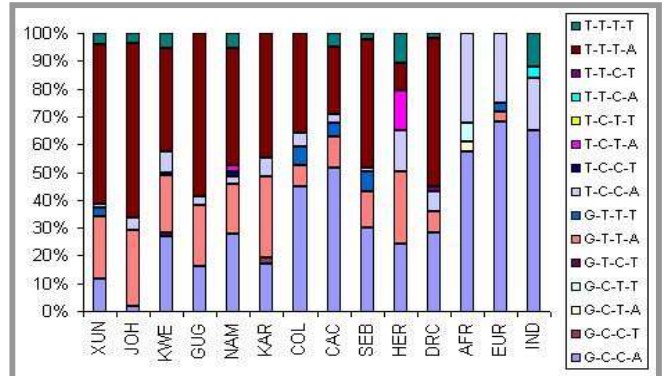
06-02



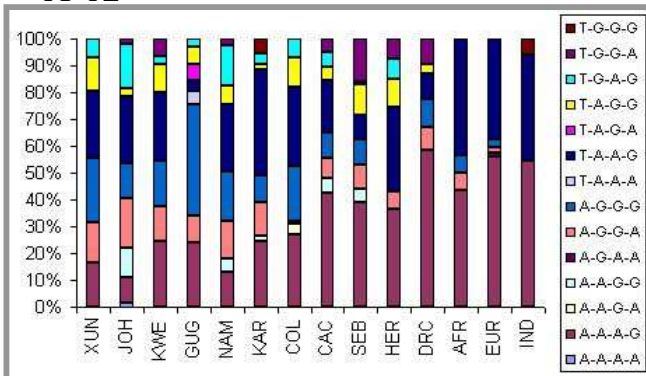
07-01



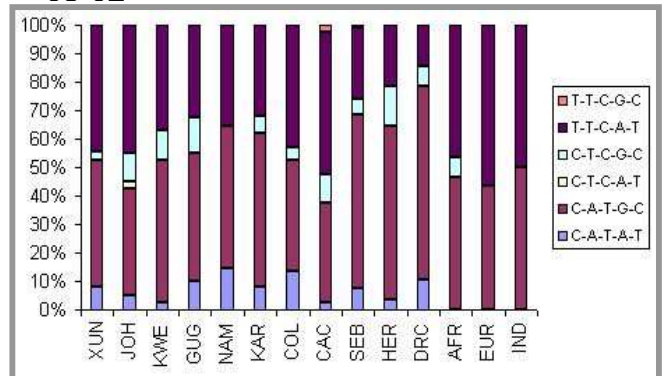
07-02



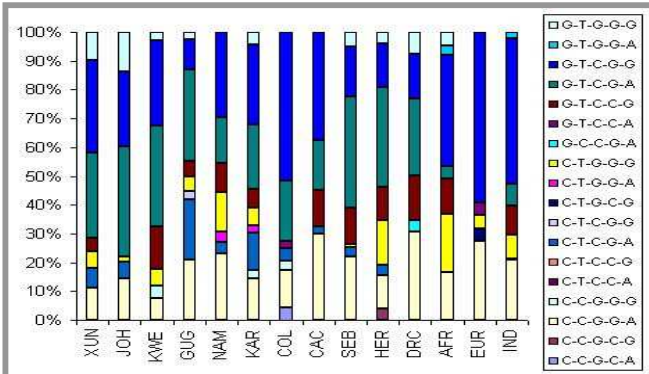
08-01



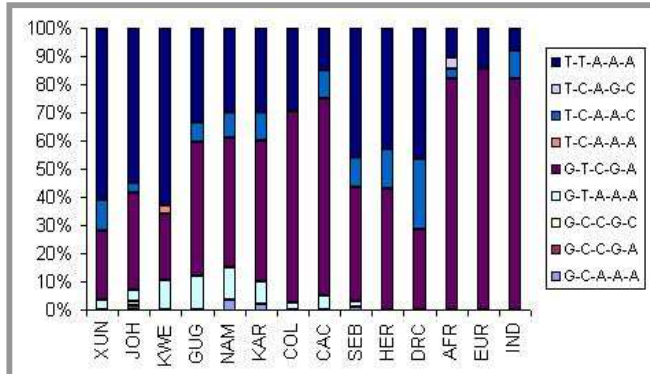
08-02



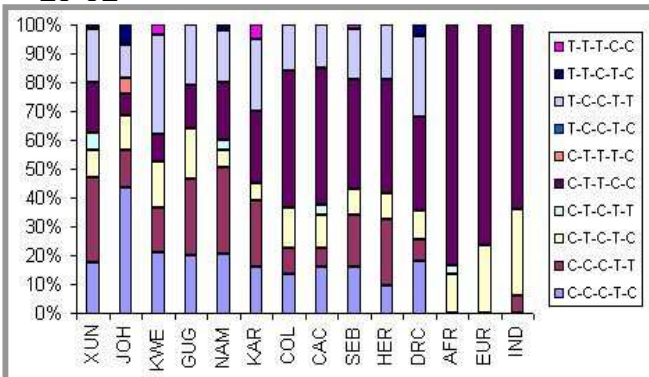
09-01



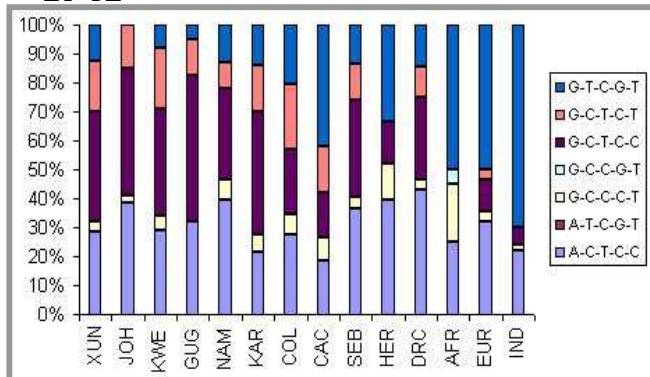
09-02



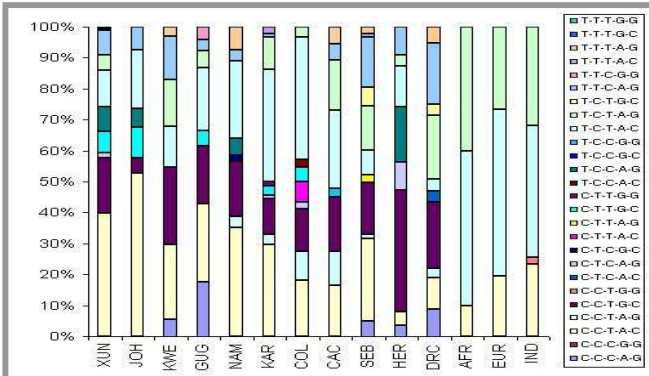
10-01



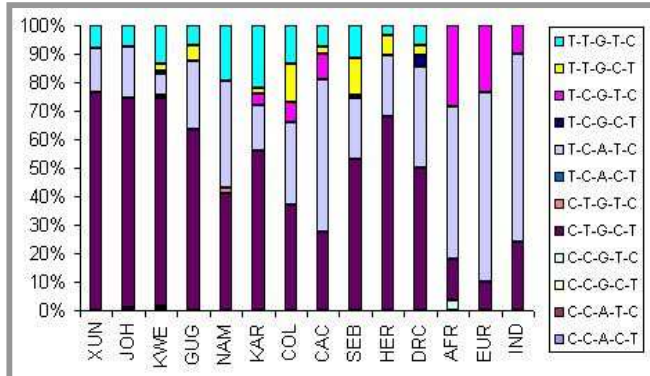
10-02



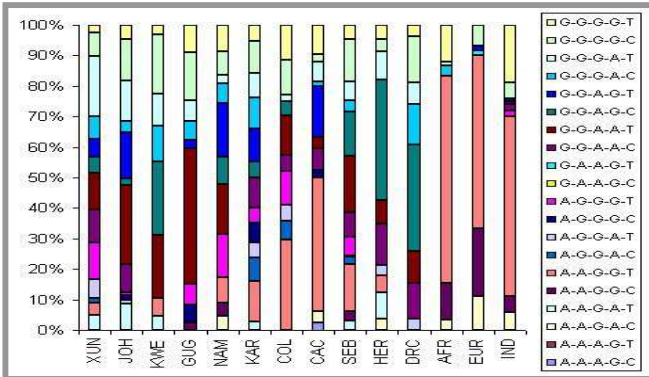
11-01



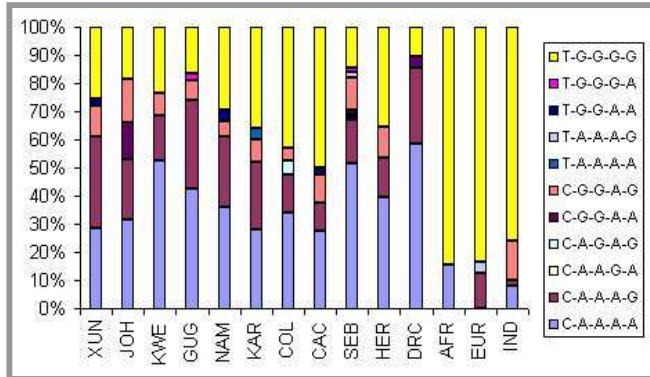
11-02



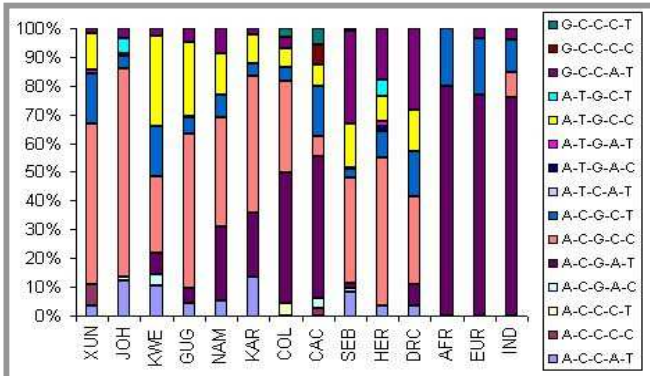
12-01



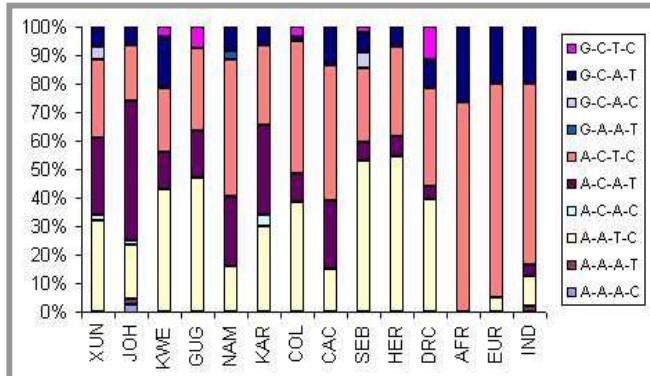
12-02



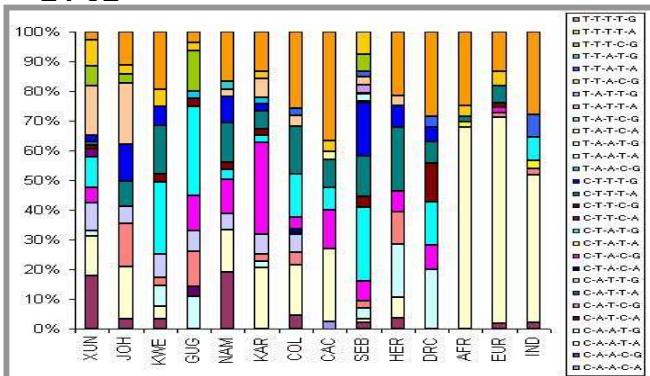
13-01



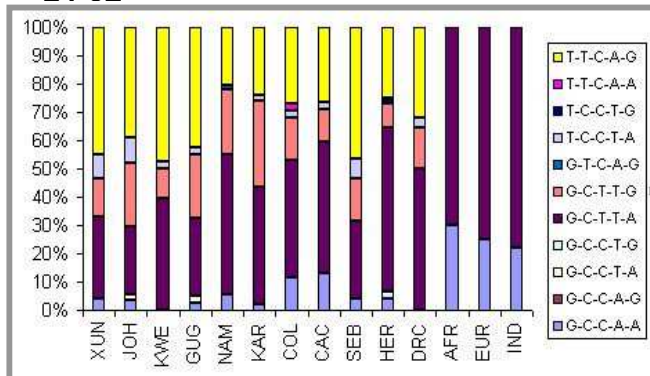
13-02



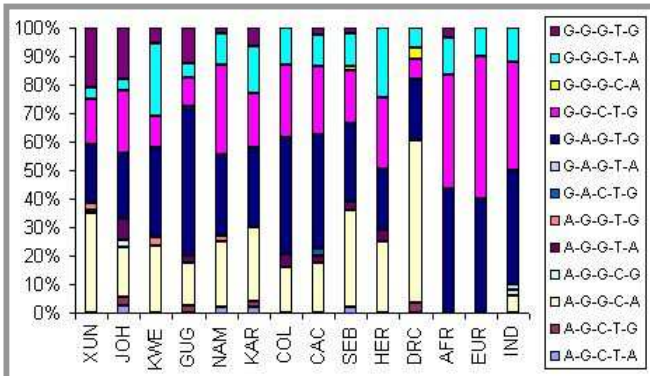
14-01



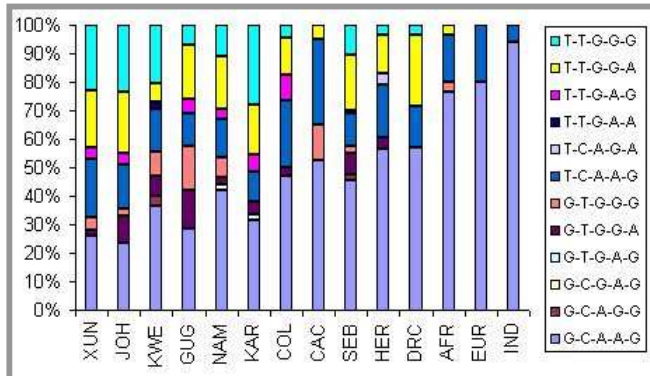
14-02



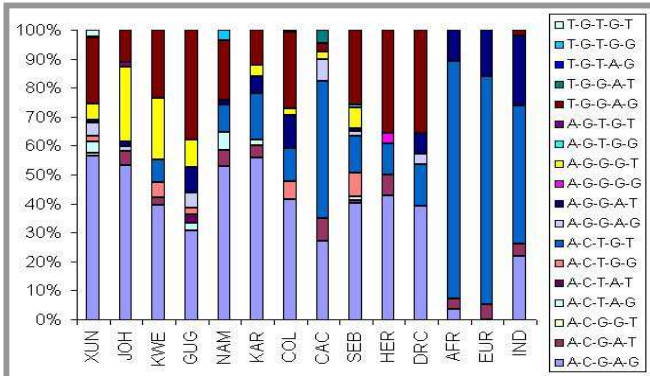
15-01



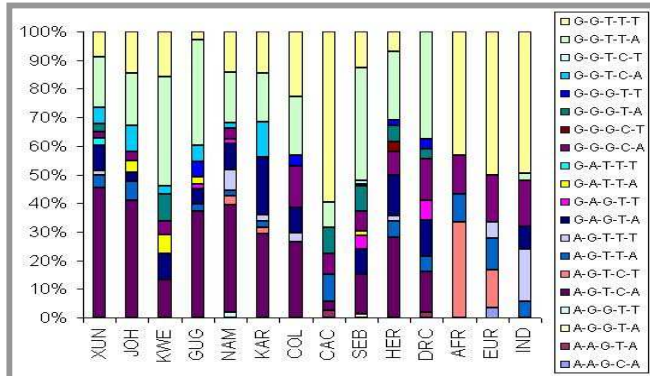
15-02



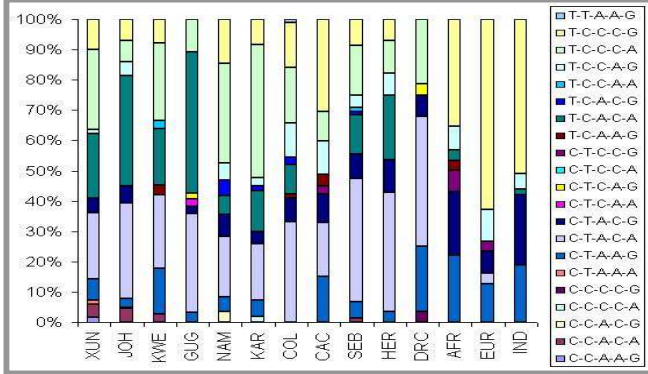
16-01



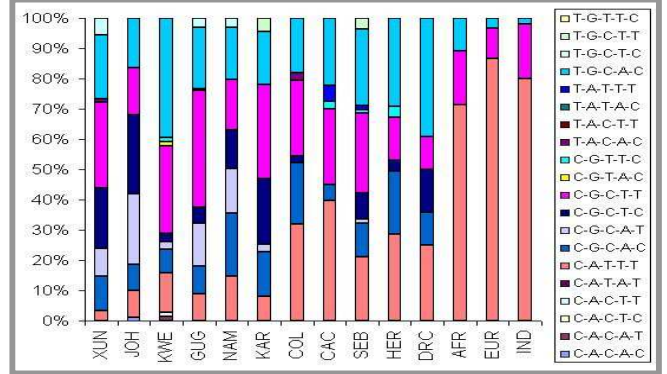
16-02



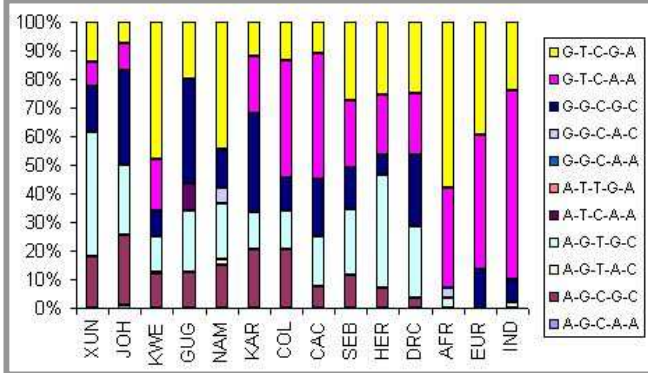
17-01



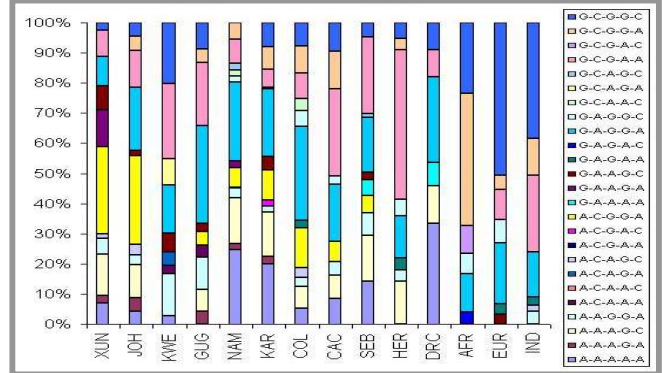
17-02



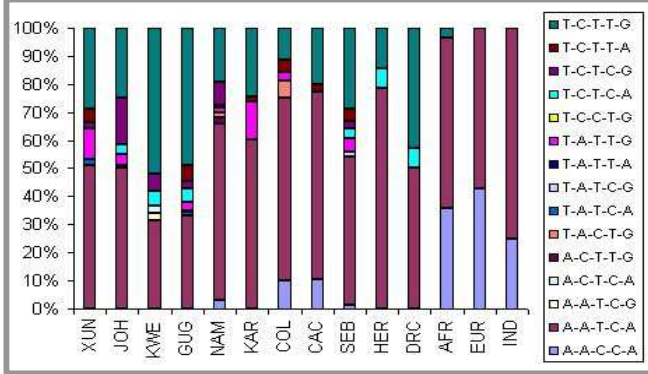
18-01



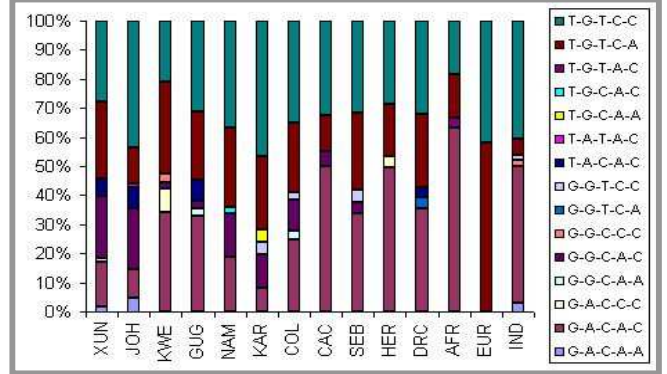
18-02



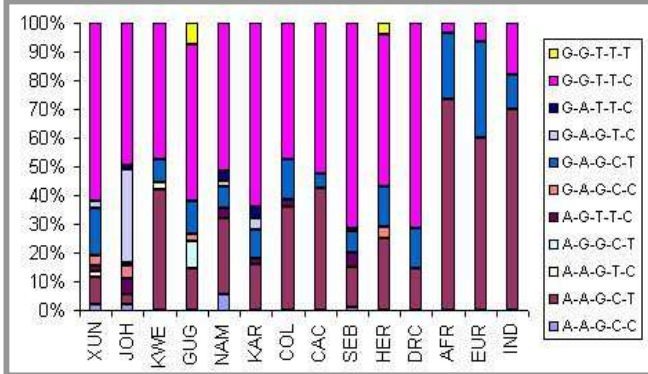
19-01



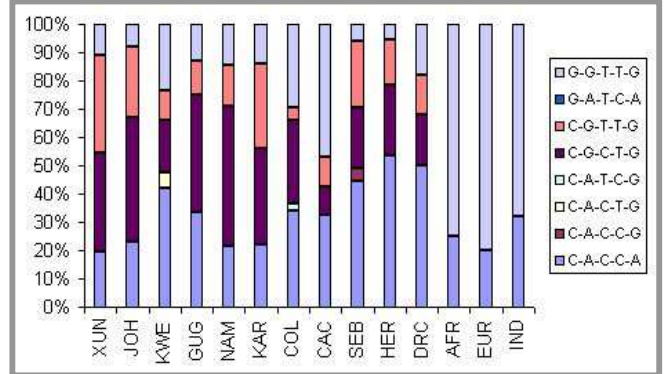
19-02



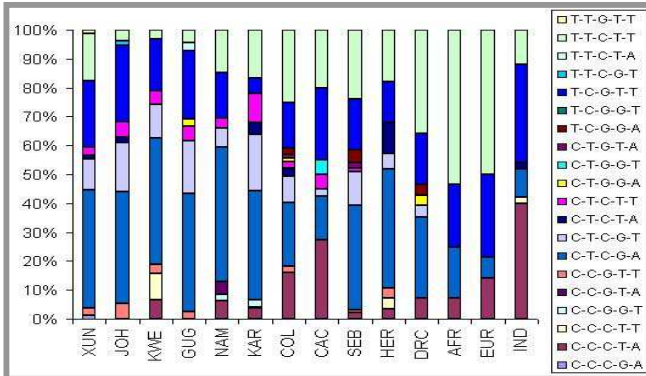
20-01



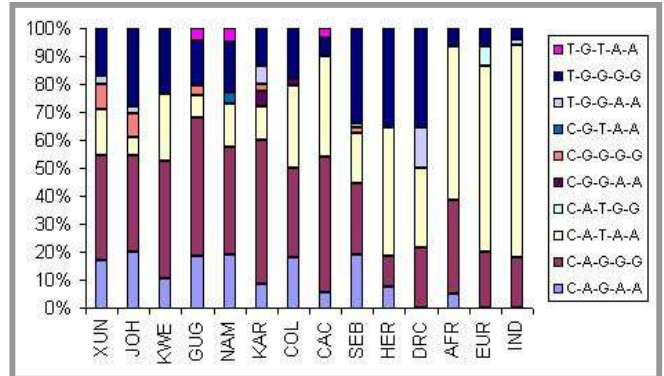
20-02



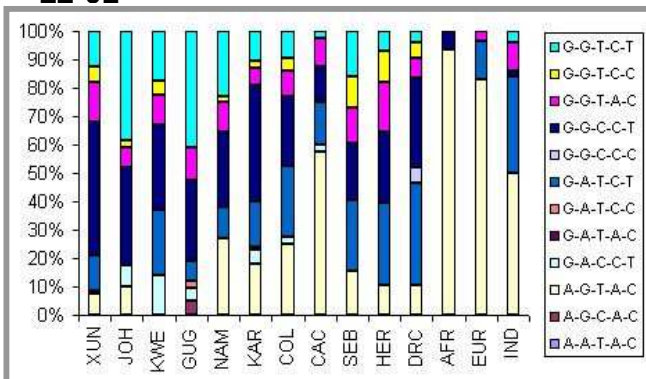
21-01



21-02



22-01



22-02

