

https://dspace.org/

Phylogeography of Y chromosome haplogroups A & B in Africa

Naidoo, Thijessen

2015-04-22
http://hdl.handle.net/10539/17506

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

PHYLOGEOGRAPHY OF Y CHROMOSOME HAPLOGROUPS A & B IN AFRICA

Thijessen Naidoo

A dissertation submitted to the Faculty of Health Sciences, University of the Witwatersrand, Johannesburg, in fulfilment of the requirements for the degree of Master of Science in Medicine

Johannesburg, 2014

DECLARATION

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

Thijessen Naidoo



14th day of October, 2014

ABSTRACT

Evolution and historical events over the past 300 000 years have contributed in shaping the gene pool of sub-Saharan African populations. By examining patterns of Y chromosome variation, through the screening of single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs), the present study aimed to characterise the phylogeography of ancient African Y chromosome haplogroups found in populations across sub-Saharan Africa, as well as understand the genetic affinities of these populations.

In order to screen the large number of the markers required, seven multiplex single base extension assays were developed. These were used to refine the resolution of Y chromosomes commonly found in Africa, but also included a few markers to delineate the common non-African Y chromosome haplogroups, following a hierarchical screening process. In total, 1667 males were screened, and these data were compiled together with comparative published data.

The resultant SNP and STR dataset was used in illustrating, more specifically, the phylogeographies of haplogroups A and B. The wide geographic distribution of haplogroup A, together with its position at the root of the phylogeny and high diversity, support an early diversification of the haplogroup into its subclades, which subsequently spread across Africa. The distribution of major haplogroup B subclades, however, are possibly due to post-glacial migrations in the case of haplogroup B-M112, and recent population expansions, leading to the common presence of haplogroup B-M152 across sub-Saharan Africa. The spread of haplogroup E, however, created the biggest impact on

African populations; with its expansion likely resulting in the diminished presence of many of the subclades of haplogroups A and B.

The Y chromosome compositions of present sub-Saharan African populations are, thus, the result of several diversification events, followed by migration, and mixing of population groups, over the course of modern human existence.

ACKNOWLEDGEMENTS

I would like to thank the sample donors who participated in this research. I am also grateful to Professor Trefor Jenkins and colleagues in the Division of Human Genetics for assistance with fieldwork and processing of samples.

I would like to acknowledge the following people and groups for their advice and aid: Cristian Capelli and George Busby for providing the R script for allelic variance; David Soria, Mattias Jakobsson, and especially Daniel Platt, for their assistance with BATWING analysis; and the Wits Core Cluster for the use of their computing facilities.

During my studies I was supported by the Medical Research Council, the German Academic Exchange Service and the University of the Witwatersrand. This research was supported by grants awarded to Professor Himla Soodyall from the South African Medical Research Council, the Palaeontological Scientific Trust, the National Research Foundation, the University of the Witwatersrand, the National Geographic Society and the National Health Laboratory Service Research Trust.

My deepest gratitude goes to Professor Himla Soodyall for her guidance and support all through these years, and for her assistance in compiling this dissertation. A sincere thank you also goes out to former and current members of the HGDDRL, for their friendship and help.

Finally, I would like to express my love and appreciation for my family, who have continually pushed and encouraged me to complete this endeavor.

TABLE OF CONTENTS

DECLARATION II
ABSTRACT III
ACKNOWLEDGEMENTS V
TABLE OF CONTENTS VI
LIST OF FIGURES VIII
LIST OF TABLES X
LIST OF ABBREVIATIONS XI
1 INTRODUCTION 1
1.1 Molecular anthropology and the haploid marker 1
1.2 Y chromosome nomenclature and the Y phylogeny 2
1.3 Y chromosome haplogroups of Africa 3
1.3.1 Haplogroup A 5
1.3.2 Haplogroup B 8
1.3.3 Haplogroup E 9
1.3.4 Non-African haplogroups in Africa 12
1.4 Screening of markers 13
1.5 Human Phylogeography 15
1.6 Study Objectives 16
2 SUBJECTS AND METHODS 17
2.1 Subjects 17
2.2 Methods 24
2.2.1 DNA extraction 24
2.2.2 Y chromosome molecular methods 24
2.3 Data Analyses 36
2.3.1 Population affinities 36
2.3.2 Phylogeographic analyses 37
3 RESULTS 40
3.1 SBE assay development 40
3.1.1 SNP Selection and Screening Strategy: 40
3.1.2 PCR Optimization 43
3.1.3 SBE Optimization 43
3.1.4 Validation of SBE assays 44
3.2 Population Affinity Analysis 44
3.3 Phylogeography 57
3.3.1 Haplogroup A 57
3.3.2 Haplogroup B 75
4 DISCUSSION 98
4.1 Y chromosome SBE assay development and optimization 98
4.1.1 Marker-specific issues 99
4.2 Population affinities in sub-Saharan Africa 100
4.3 The phylogeography of haplogroup A in sub-Saharan Africa 101
4.3.1 Haplogroup A-M14 103
4.3.2 Haplogroup A-M51 104
4.3.3 Haplogroup A-M13 105
4.4 The phylogeography of haplogroup B in sub-Saharan Africa 107
4.4.1 Haplogroup B-M150 108
4.4.2 Haplogroup B-M112 109
4.5 Potential impact of the Last Glacial Maximum 113

4.6 Expansion of haplogroup E and its effect on sub-Saharan African diversity 114
4.6.1 Haplogroup E-M2 115
4.6.2 Haplogroup E-M35 116
4.7 Admixed populations in sub-Saharan Africa 118
4.8 Study limitations and potential future directions 120
5 CONCLUSION 122
6 REFERENCES 123
7 APPENDICES 134
7.1 Appendix A 134
7.2 Appendix B 140
7.3 Appendix C 144
7.4 Appendix D 147
7.5 Appendix E 153
7.6 Appendix F 157
7.7 Appendix G 168
7.8 Appendix H 174

LIST OF FIGURES

Figure 1.1: Schematic of the human Y chromosome (Hurles and Jobling, 2001) 2
Figure 1.2: An abbreviated Y chromosome phylogeny showing the major Y chromosome haplogroups, from Karafet, et al. (2008). 4
Figure 1.3: (A) Revised backbone of the human Y chromosome phylogeny, based on 146 newly discovered bi-allelic variants (Cruciani, et al., 2011). (B) Comparison of the Y chromosome phylogeny backbones as reported in Cruciani, et al. (2011) (left) and Karafet, et al. (2008). 6
Figure 1.4: An abbreviated haplogroup A phylogeny (for the full ISOGG 2013 haplogroup A phylogeny, see appendix A). 7
Figure 1.5: An abbreviated haplogroup B phylogeny (for the full ISOGG 2013 haplogroup B phylogeny, see appendix B). 9
Figure 1.6: Y chromosome haplogroup E1b1 phylogeny as proposed in Trombetta, et al. (2011). 11
Figure 1.7: Diagram illustrating the principle and steps of the SBE technique. The detection primer anneals to the PCR product immediately upstream of the SNP. A fluorescently labelled ddNTP is then attached to the primer during extension. 14
Figure 3.1: Electropherogram and phylogeny of (A) YSNP1, (B) Hg-A, (C) Hg-B, (D) Hg-B2b, (E) Hg-E, (F) Hg-E1b1a, and (G) Hg-E1b1b1 42
Figure 3.2: MDS plot of Fst distances between populations. See List of Abbreviations for a description of population codes. 50
Figure 3.3: MDS plot representing Rst distances between populations. See List of Abbreviations for a description of population codes. 51
Figure 3.4: Cluster analysis tree representing Fst distances between populations. See List of Abbreviations for a description of population codes. 52
Figure 3.5: Cluster analysis tree representing Rst distances between populations. See List of Abbreviations for a description of population codes. 53
Figure 3.6: (A) Phylogeny of haplogroup A-L419 (A1b1) and its subclades, with TMRCA estimates indicated by the boxes surrounding the markers used in the BATWING analysis. Frequency distributions of (B) haplogroup A-M14*, (C) A-M6, (D) A-M51, and (E) A-M13. 60
Figure 3.7: RM-MJ network of A-M14 based on a 3 SNP-15 STR haplotype (M6-M114-P28-DYS19-DYS389I-DYS389c-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439-DYS448-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 2 and Epsilon = 0. 62
Figure 3.8: RM-MJ network of A-M51 based on a 14 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 2 and Epsilon = 0. 66
Figure 3.9: RM-MJ network of A-M13 based on a 10 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439) with reduction threshold = 2 and Epsilon = 0. 73
Figure 3.10: (A) Phylogeny of haplogroup B-M182 (B2) and its subclades, with TMRCA estimates indicated by the boxes surrounding the markers used in the BATWING analysis. Frequency distributions of (B) B-M150 (C) B-M112, (D) B-P6, and (E) B-P7*. 82
Figure 3.11: RM-MJ network of B-M150 based on a 15 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-

DYS439-DYS448-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 1 and Epsilon = 0. 85
Figure 3.12: RM-MJ network of B-M150 based on a 10 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439) with reduction threshold = 1 and Epsilon = 0. 86
Figure 3.13: RM-MJ network of B-M112 based on a 6 SNP – 14 STR haplotype (P6-P7-M115-M30-P8-M211-DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS448-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 2 and Epsilon = 0.....94
Figure 3.14: RM-MJ network of B-M112 based on a 9 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438) with reduction threshold = 2 and Epsilon = 0. 95

LIST OF TABLES

Table 2.1: Details of populations examined in the present study 19
Table 2.2: PCR primer sequences, amplicon lengths and final concentrations 26
Table 2.3: SBE primer sequences and final concentrations 30
Table 2.4: Y chromosome SNPs genotyped using RFLP assays in the present study 32
Table 2.5: Y chromosome SNPs and respective haplogroups included in the SBE assays .33
Table 3.1: Y chromosome haplogroup frequencies for sub-Saharan African populations ..46
Table 3.2: Fst and Rst AMOVA analysis based on various groupings 56
Table 3.3: Haplogroup A-M14 frequencies in sub-Saharan African populations 59
Table 3.4: Allelic variance and haplotype gene diversity estimates of haplogroup A-M14, based on population groups, regions and subclades. 63
Table 3.5: Haplogroup A-M51 frequencies in sub-Saharan African populations 65
Table 3.6: Allelic variance and haplotype gene diversity estimates of haplogroup A-M51 based on population groups and regions 67
Table 3.7: Haplogroup A-M13 frequencies in sub-Saharan African populations 70
Table 3.8: Allelic variance and haplotype gene diversity estimates of haplogroup A-M13 based on population groups and regions 74
Table 3.9: Haplogroup B-M150 frequencies in sub-Saharan African populations 77
Table 3.10: Allelic variance and haplotype gene diversity estimates of haplogroup B-M150 based on population groups and regions 88
Table 3.11: Haplogroup frequencies of B-M112 and its subclades in the populations studied 90
Table 3.12: Allelic variance and haplotype gene diversity estimates of haplogroup B-M112 based on population groups, regions and subclades 97

LIST OF ABBREVIATIONS

BATWING:	Bayesian Analysis of Trees With Internal Node Generation
CACB:	Central African Central Bantu speakers
CAR:	Central African Republic
CAU:	Central African Ubangian speakers
CI:	confidence interval
CKS:	Central Khoe-San
CMAL:	Cape Malay
DAM:	Dama
ddH_2O:	double distilled water
ddNTP:	dideoxynucleoside-triphosphate
DHPLC:	denaturing High Performance Liquid Chromatography
EACB:	East African Central Bantu speakers
EACU:	East African Cushitic speakers
EAN:	East African Nilotic speakers
EUR:	European Descent
FPP:	food-producing population
HAD:	Hadza
HGDDRL:	Human Genomic Diversity and Disease Research Laboratory
HGP:	hunter-gatherer population
IND:	Indian Descent
ISOGG:	International Society of Genetic Genealogy
KBAD:	Admixed Khoe-San/Bantu speakers
kya:	thousand years ago
LGM:	Last Glacial Maximum
MCMC:	Markov chain Monte Carlo
MDS:	multidimensional scaling
ML:	maximum likelihood
mtDNA:	mitochondrial DNA
NKS:	Northern Khoe-San
NMBC:	Nigerian mixed Benue-Congo speakers
NRY:	non-recombining region of the Y chromosome

NWB:	Northwest Bantu speakers
PCR:	polymerase chain reaction
RFLP:	restriction fragment length polymorphism
RM-MJ:	reduced-median and median-joining
SAC:	South African Coloured
SACB:	Southern African Central Bantu speakers
SAND:	Sandawe
SBE:	single base extension
SEB:	Southeastern Bantu speakers
SKS:	Southern Khoe-San
SMM:	stepwise mutation model
SNP:	single nucleotide polymorphism
STR:	short tandem repeat
SWB:	Southwestern Bantu speakers
SWCB:	Southwestern African Central Bantu speakers
TMRCA:	time to most recent common ancestor
UEP:	unique event polymorphism
WAM:	West African Mandinka
WAMA:	West African mixed Atlantic speakers
WPYG:	Western Pygmy
YAP:	Y-Alu Polymorphism
YCC:	Y Chromosome Consortium
YHRD:	Y chromosome Haplotype Reference Database

1 INTRODUCTION

1.1 Molecular anthropology and the haploid marker

Anthropology is the study of humanity, our physical and cultural similarities and differences over time. To this end scholars have utilized many tools within the disciplines of human anatomy, archaeology, linguistics and history to reconstruct the human historical record. During the past century, however, the field of genetics has witnessed an upsurge in its application to anthropological questions, documenting the similarities and differences between people, and offering a clearer image of humanity's past. Given that human genome variation evolves over time due to several factors - among them mutation, genetic drift, migration and selection - the genome has retained some of the record of these historical and evolutionary events. Recently, whole genome approaches have become useful in answering questions related to the origin and diversification of modern humans (Lawson, et al., 2012). This is due to higher resolution screening and better computational analyses. For over two decades, however, the uniparentally inherited marker systems – mitochondrial DNA (mtDNA) and the non-recombining region of the Y chromosome (NRY) (Fig. 1.1) – have been used to answer questions regarding human origins and the subsequent demographic events leading to current peoples and populations (Underhill and Kivisild, 2007). The utility of haploid systems such as mtDNA and the NRY when examining origins and affinities is due to a lack of meiotic recombination, unlike the rest of the genome. This results in the inheritance of intact haplotypes through generations, which change only by mutation, thus preserving a simpler record of their history. Additionally, their uniparental modes of inheritance – maternal in the case of mtDNA; paternal in the

case of the NRY – allows for the elucidation of specifically female or male contributions to the shaping of the human gene pool. The NRY and mtDNA continue to be important sources of information; especially in light of their detailed and well-defined human phylogenies, which clearly represent both the clinality and discreteness that characterises human diversity.

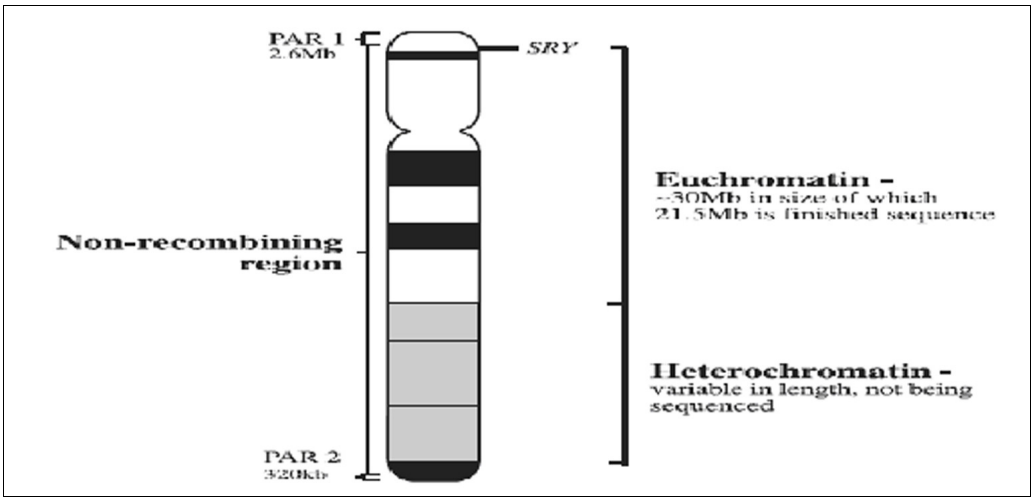

Figure 1.1: Schematic of the human Y chromosome (Hurles and Jobling, 2001)

1.2 Y chromosome nomenclature and the Y phylogeny

While the first polymorphism on the Y chromosome was recorded in 1985 (Casanova, et al., 1985), it was only toward the end of the twentieth century that enough markers were compiled to generate well-resolved Y chromosome phylogenies (Underhill, et al., 2000; Hammer, et al., 2001; Underhill, et al., 2001). The usage of multiple independent naming systems, however, made it difficult to compare results between them. A standard nomenclature was established by the Y Chromosome Consortium (YCC) (Y Chromosome Consortium [YCC], 2002), which resolved the global pattern of Y chromosome variation

into 18 major haplogroups that were classified using capital letters A through to R. This was revised by Karafet, et al. (2008) to a Y chromosome haplogroup phylogeny that contained 311 branches delineated by approximately 600 markers (primarily bi-allelic) and included an additional two haplogroups (S and T), increasing the major haplogroup number to 20 (Fig. 1.2). Currently, the Y chromosome phylogeny is kept up to date by the International Society of Genetic Genealogy (ISOGG), as new publications and the discovery of more variation improve its resolution.

Due to the lower effective population size of the Y chromosome – 25% of any autosome and 33% of the X chromosome – it is known to be more susceptible to genetic drift (Jobling and Tyler-Smith, 2003). The increased effect of drift results in accelerated differentiation of Y chromosomes from different populations (Seielstad, Minch and Cavalli-Sforza, 1998). This has resulted in strong concordance between Y chromosome haplogroups and the geographic distributions of populations and adds to the Y chromosome's utility in reconstructing the history and migrations of humans over time.

1.3 Y chromosome haplogroups of Africa

Of the 20 major Y chromosome haplogroups, only three exhibit substantial presence throughout Africa, *viz.* haplogroups A, B and E. While haplogroups A and B represent the oldest splits in the Y chromosome phylogeny, it is haplogroup E that has attained highest frequencies, with some subclades spreading out of Africa into the Middle East and Europe.

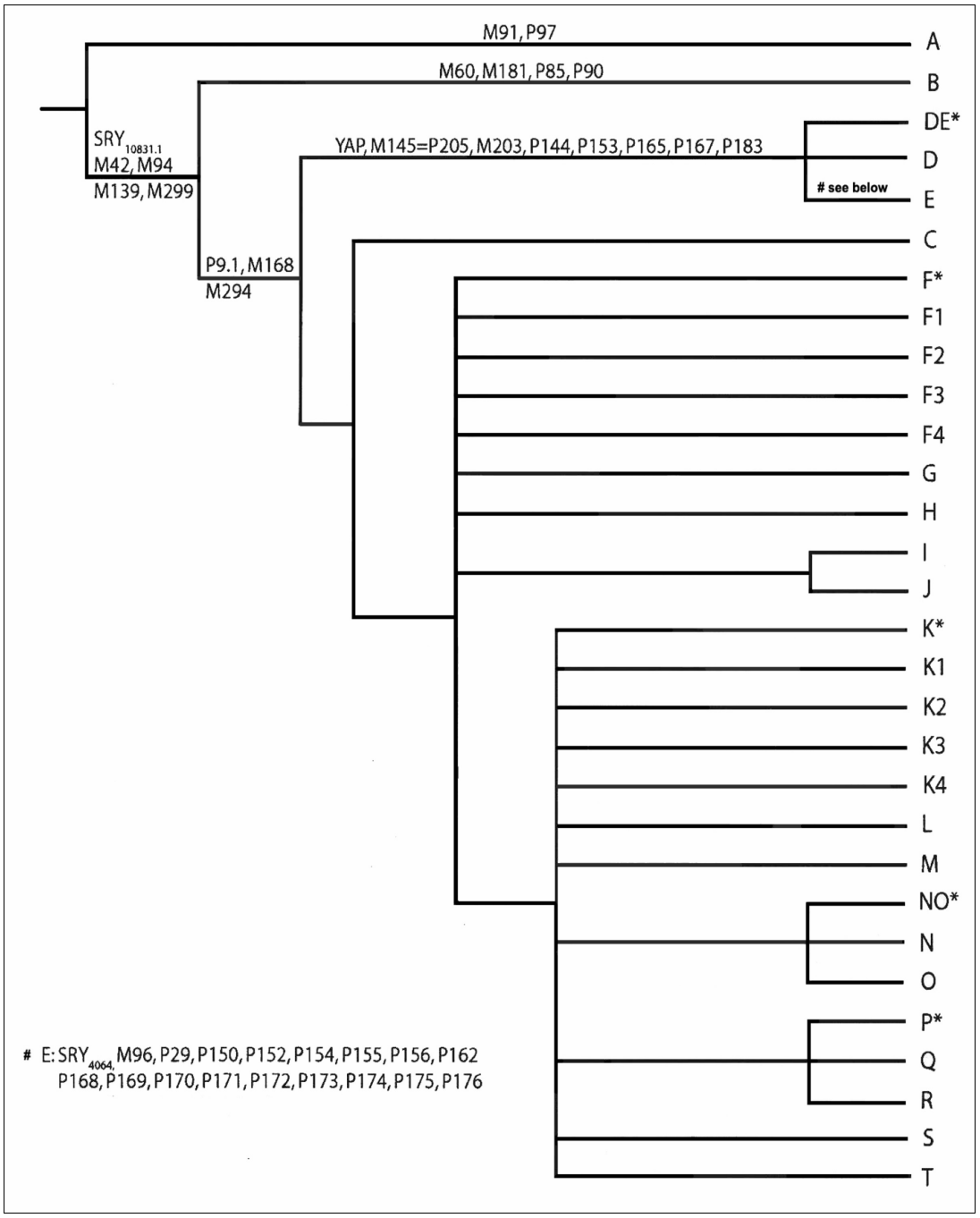

Figure 1.2: An abbreviated Y chromosome phylogeny showing the major Y chromosome haplogroups, from Karafet, et al. (2008).

1.3.1 Haplogroup A

Haplogroup A was the first to branch off the Y chromosome phylogeny, and is regarded as the oldest haplogroup. Initially it was defined by the presence of the M91 mutation (Underhill, et al., 2000). Karafet, et al. (2008) then reinforced the branch by introducing the P97 mutation alongside M91. All lineages within haplogroup A were considered monophyletic (defined by M91 and P97) until 2011, when Cruciani, et al. (2011) resequenced sections of the Y chromosome in a subset of individuals, which resulted in the detection of 146 bi-allelic variants (Fig. 1.3A). These variants, when incorporated into the existing Y chromosome phylogeny, drastically modified its basal backbone (Fig. 1.3B). These modifications showed that haplogroup A was, instead, composed of a number of polyphyletic clades, with M91 and P97 phylogenetically equivalent to one of these clades. The deepest split separated haplogroup A1b from a monophyletic clade comprised of haplogroups A1a, A2, A3 and BT.

Haplogroup A1a then separated out, leaving a monophyletic clade containing A2, A3 and BT. Haplogroup A1b has since been renamed to A0 (ISOGG 2012) in order to preserve as much of the current Y chromosome haplogroup nomenclature as possible. Furthermore Mendez, et al. (2013) discovered a new haplogroup, named A00, which was even older than haplogroup A0. One of the consequences of the modified phylogeny was the deeper time to most recent common ancestor (TMRCA) at 338 thousand years ago (kya).

The subclades of haplogroup A (Fig. 1.4) exhibit strong geographic structuring. Haplogroup A00 was initially found in an African American individual; however, it is also found at low frequencies in Cameroon (Mendez, et al., 2013).

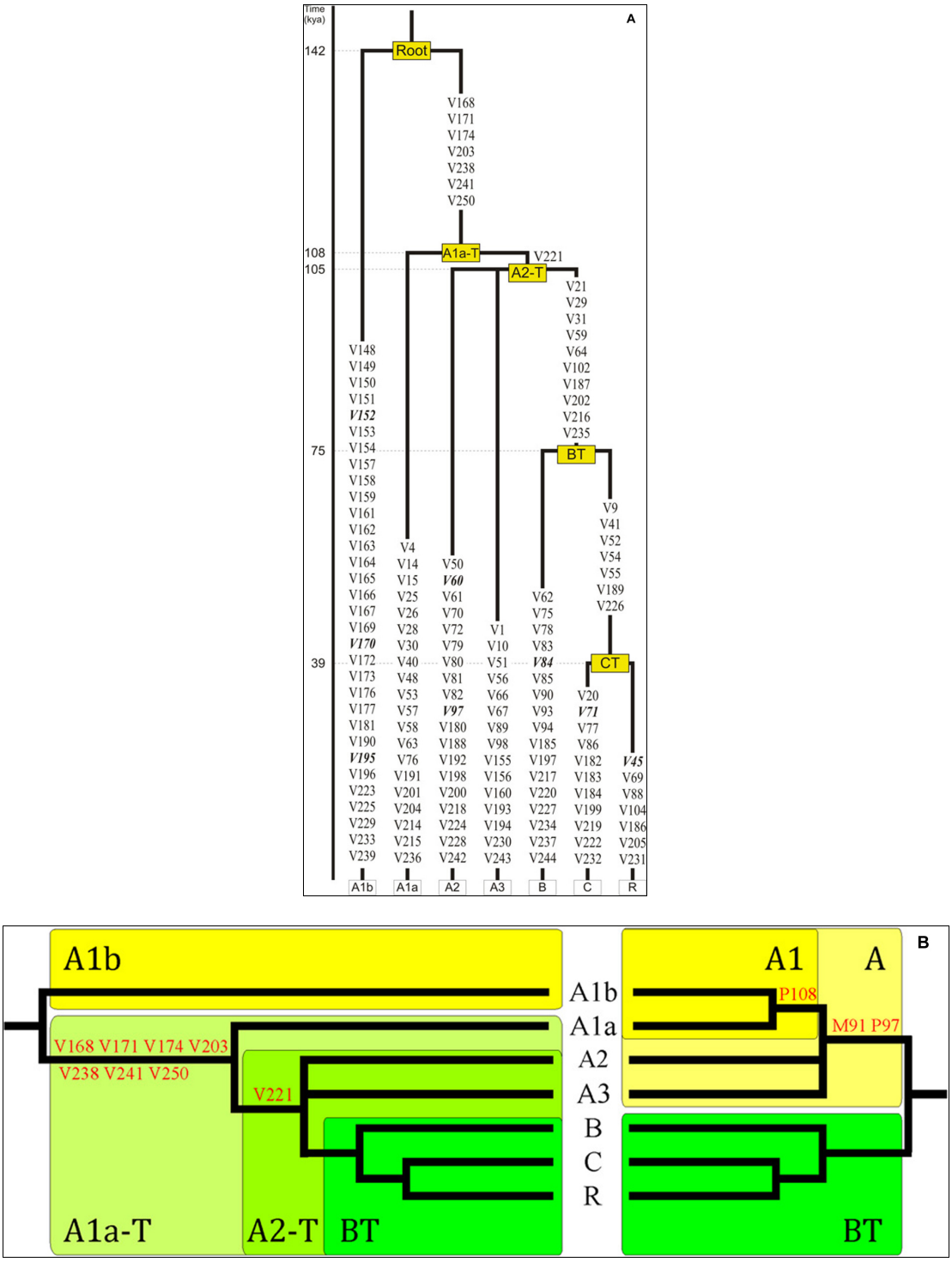

Figure 1.3: (A) Revised backbone of the human Y chromosome phylogeny, based on 146 newly discovered bi-allelic variants (Cruciani, et al., 2011). (B) Comparison of the Y chromosome phylogeny backbones as reported in Cruciani, et al. (2011) (left) and Karafet, et al. (2008).

While haplogroup A0 (A0a1a, based on the P114 mutation) has been found in Cameroon and at low frequency in northern Algeria (Cruciani, et al., 2011), haplogroup A1a (A-M31) has been found across west Africa (Underhill, et al., 2000; Cruciani, et al., 2002; Gonçalves, et al., 2003; Wood, et al., 2005; Rosa, et al., 2007; Cruciani, et al., 2011). Haplogroup A1b1b2b (A-M13) has its strongest presence in east Africa - Sudan, Ethiopia, Tanzania, Kenya - (Underhill, et al., 2000; Cruciani, et al., 2002; Semino, et al., 2002; Luis, et al., 2004; Wood, et al., 2005; Tishkoff, et al., 2007; Hassan, et al., 2008) and at lower frequencies in Cameroon (Cruciani, et al., 2002; Wood, et al., 2005). In contrast, haplogroup A1b1b2a (A-M51) occurs at its highest frequencies among the Khoe-San populations of southern Africa (Underhill, et al., 2000; Cruciani, et al., 2002; Wood, et al., 2005; Naidoo, et al., 2010) as well at lower frequencies in neighbouring Southeastern Bantu speakers (Naidoo, et al., 2010) and the South African Coloured population (Quintana-Murci, et al., 2010). Haplogroup A1b1a1 (A-M14), thought to be found exclusively among the Khoe-San (Underhill, et al., 2000; Cruciani, et al., 2002; Wood, et al., 2005), has been found in an ancestral form among central African Pygmy populations (Batini, et al., 2011a).

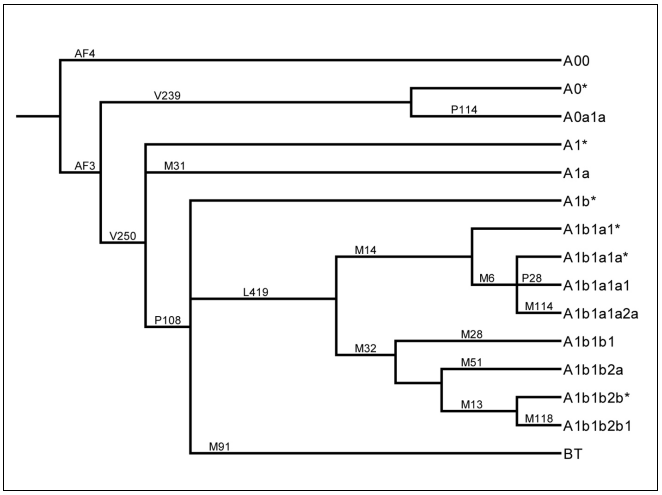

Figure 1.4: An abbreviated haplogroup A phylogeny (for the full ISOGG 2013 haplogroup A phylogeny, see appendix A).

1.3.2 Haplogroup B

After haplogroup A, haplogroup B is regarded as one of the oldest haplogroups. While Karafet, et al. (2008) defined the haplogroup using M60 (Underhill, et al., 2000), M181 (Underhill, et al., 2001), P85 and P90 (Karafet, et al., 2008), the resequencing done by Cruciani, et al. (2011), led to the unearthing of 17 more bi-allelic markers, strengthening support for the branch. The subclades of haplogroup B (Fig. 1.5) occur throughout sub-Saharan Africa but also display a high level of geographic structure. Haplogroup B1 (B-M236) has been found at low frequencies in Cameroon, Burkina Faso (Cruciani, et al., 2002) and Mali (Underhill, et al., 2000). Most of the diversity present in haplogroup B, however, is contained within the haplogroups B2a (B-M150) and B2b (B-M112). Haplogroup B2a1a (B-M152) is commonly observed in Africa, enjoying widespread distribution at moderate to high frequencies in central Africa, east Africa and southern Africa (Underhill, et al., 2000; Cruciani, et al., 2002; Beleza, et al., 2005; Wood, et al., 2005; Naidoo, et al., 2010; Batini, et al., 2011a), possibly as a result of the Bantu Expansion (Diamond, 1997). It is believed that around 5000 years ago, Bantu speakers from a region that is now northwest Cameroon / southern Nigeria, began a migration that resulted in the spread of Bantu languages throughout sub-Saharan Africa. Haplogroup B-M112, while found at lower frequencies overall, compared to B-M150, occurs at high frequencies among hunter-gatherer populations such as Pygmies (Eastern and Western), Khoe-San and Hadza, with some lineages segregating specifically to each of these groups (Underhill, et al., 2000; Cruciani, et al., 2002; Knight, et al., 2003; Tishkoff, et al., 2007; Batini, et al., 2011a).

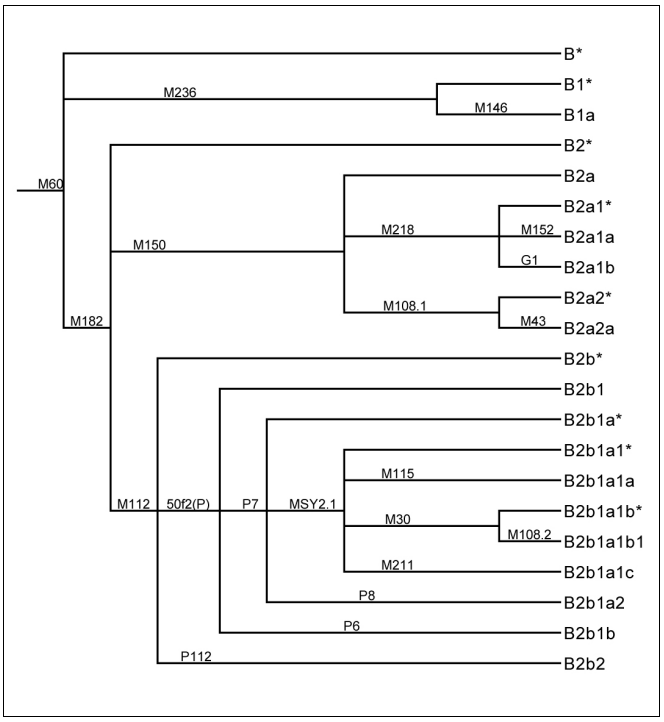

Figure 1.5: An abbreviated haplogroup B phylogeny (for the full ISOGG 2013 haplogroup B phylogeny, see appendix B).

1.3.3 Haplogroup E

While haplogroup E is the most common African haplogroup, it is the sister clade to the Asian haplogroup D, as both have the Y Alu Polymorphism (YAP) in common. While individuals bearing the ancestral paragroup, DE*, have been found in Africa (Weale, et al., 2003; Rosa, et al., 2007) and Asia (Shi, et al., 2008) it is commonly accepted that haplogroup E arose in Africa; possibly in east Africa. The branch is defined, currently, by as many as 25 mutations (ISOGG 2013), and while 56 distinct sub-haplogroups were recorded in 2008 (Karafet, et al., 2008), the past few years has seen substantial restructuring and additions within haplogroup E. It has obtained the highest frequencies of any major haplogroup in Africa; however, most of these numbers are accounted for by members of the E1b1 (E-P2) subclade (Fig. 1.6). Haplogroup E1a (E-M33) has been

observed primarily in north and west Africa (Underhill, et al., 2000; Bosch, et al., 2001; Cruciani, et al., 2002; Semino, et al., 2002; Gonçalves, et al., 2003; Arredi, et al., 2004; Luis, et al., 2004; Wood, et al., 2005; Rosa, et al., 2007), with minor occurrences in Portugal (Gonçalves, et al., 2005) and Italy (Battaglia, et al., 2009) as well. Its highest frequency was found in Mali (Underhill, et al., 2000; Wood, et al., 2005). Haplogroup E2 (E-M75) is present throughout sub-Saharan Africa at low to moderate frequencies (Underhill, et al., 2000; Cruciani, et al., 2002; Gonçalves, et al., 2003; Luis, et al., 2004; Wood, et al., 2005; Rosa, et al., 2007). While no basal E1b1* (E-P2*) examples have been found, surviving branches of this haplogroup, E1b1a (E-V38) and E1b1b (E-M215), have been found throughout Africa. Haplogroup E1b1a, previously defined by M2 (which now defines E1b1a1), occurs at its highest frequencies and diversity in west Africa (Cruciani, et al., 2002; Semino, et al., 2002; Gonçalves, et al., 2003; Rosa, et al., 2007).

While its prevalence decreases clinally from west to east across Africa, E-V38 and its subclades are still found at frequencies above 50% in many parts of central, east and southern Africa (Underhill, et al., 2000; Cruciani, et al., 2002; Wood, et al., 2005). The mutations V38 and V100 currently define haplogroup E1b1a (Trombetta, et al., 2011), but the lineages within the E-M2 subclade are still the most commonly found. The most significant of these lineages, whose high frequencies characterise the Bantu Expansion, are haplogroup E1b1a1a1f1a1 (E-U174) - on an E-M191 background - and haplogroup E1b1a1a1g (E-U175) (de Filippo, et al., 2011).

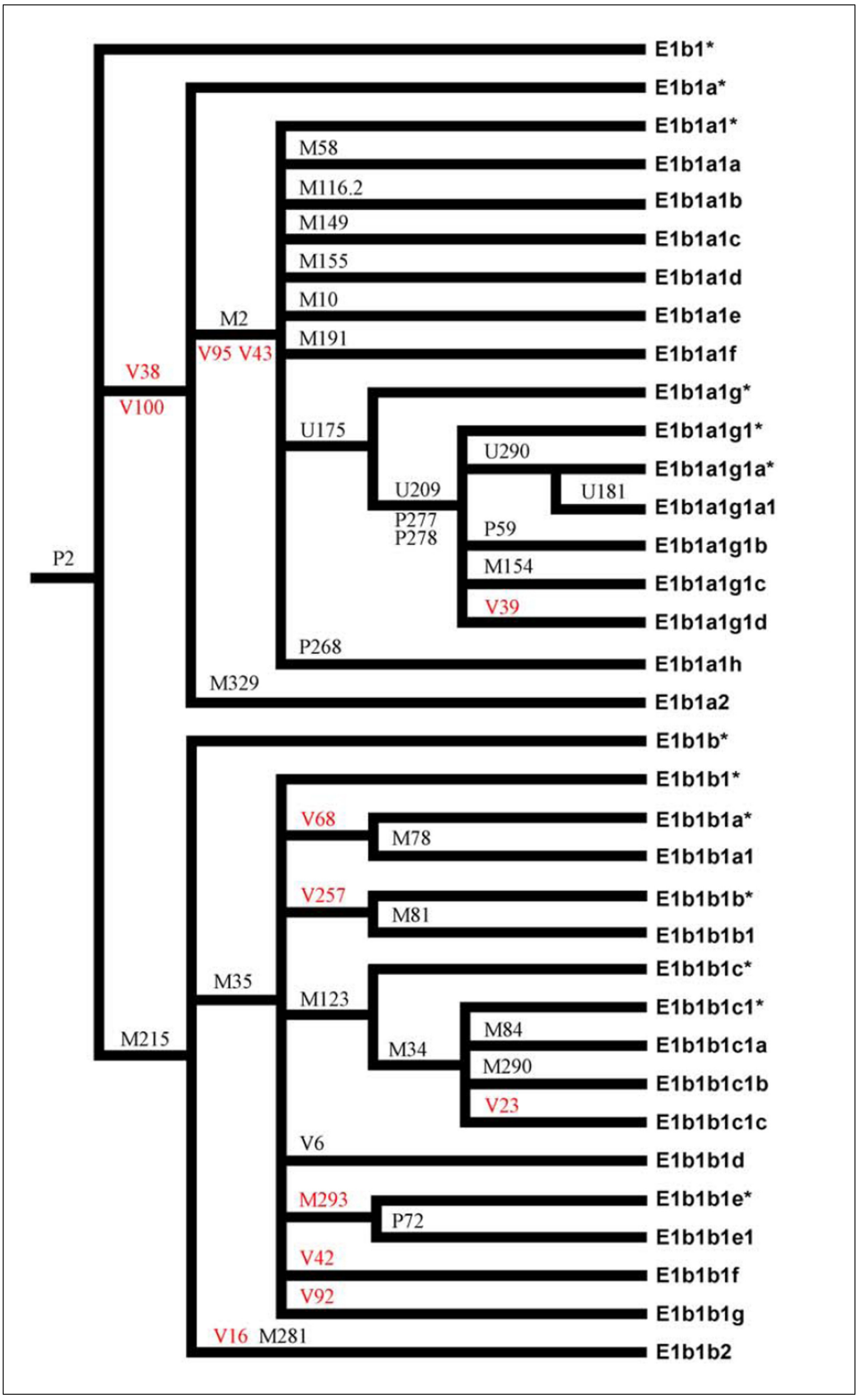

Figure 1.6: Y chromosome haplogroup E1b1 phylogeny as proposed in Trombetta, et al. (2011).

Haplogroup E1b1b, more particularly its major subclade E1b1b1 (E-M35.1), is the only African haplogroup to have attained appreciable frequencies outside of Africa. Much of its European and Middle Eastern distribution is accounted for by haplogroup E1b1b1a1 (E-M78) (Cruciani, et al., 2007), though this haplogroup is thought to have arisen in northeast Africa. The northwest African presence of E1b1b1 is represented by E1b1b1b1a (E-M81), which is associated mainly with Berber populations (Bosch, et al., 2001). It has found its way into Europe as well, with a minor presence in the Iberian Peninsula. In 2008, the M293 marker was discovered (Henn, et al., 2008), which was found to characterise a major proportion of haplogroup E1b1b1 variation from east Africa into southern Africa. Haplogroup E1b1b1b2b (E-M293) was found to occur at high frequencies among certain Tanzanian populations and Southern African Khoe-San populations, and is possibly associated with the spread of pastoralism from east Africa into southern Africa (Henn, et al., 2008).

1.3.4 Non-African haplogroups in Africa

While haplogroups A, B and E comprise the majority of Y chromosome variation in Africa, some haplogroups which diverged outside of Africa, and found more commonly among Eurasians, have been found at varying frequencies among African populations. This could be due to relatively recent admixture, as is the case with South African Coloured populations (Quintana-Murci, et al., 2010) and many north African populations (Arredi, et al., 2004), or due to an ancient back-migration. It has been suggested that an ancient back-migration could account for the presence of a subclade of haplogroup R1b, R-V88, in north Africa (Algeria, Egypt and Morrocco) and central Africa (Chad, Cameroon, Niger and Nigeria) at appreciable frequencies, especially in northern Cameroon where it occurs at

frequencies of 9-95% among certain populations (Cruciani, et al., 2002; Cruciani, et al., 2010).

Henceforth any reference to a haplogroup will be based on marker nomenclature (e.g. A-M51) to avoid confusion and complicated names (e.g. E1b1b1b2a1a1).

1.4 Screening of markers

Over the years a number of different techniques such as denaturing High Performance Liquid Chromatography (dHPLC) and resequencing were used to discover Y chromosome polymorphisms (Underhill, et al., 2000; Hammer, et al., 2003; Karafet, et al., 2008). These methods, together with other polymerase chain reaction (PCR) and restriction fragment length polymorphism (RFLP) assays, have been utilised in the genotyping of Y chromosomes in several populations. In most labs, however, these techniques were not adopted due to equipment limitations and cost. This prompted the development of other genotyping techniques (Sobrino, Brión and Carracedo, 2005); which have increased the resolution at which Y chromosomes are examined. Of these methods, single base extension (SBE) – otherwise known as minisequencing – was chosen for use in this study. The principle of the method lies in the extension of a "detection" primer that has annealed immediately 3' of the single nucleotide polymorphism (SNP) of interest using a fluorescently labelled dideoxynucleoside-triphosphate (ddNTP) that is complementary to the SNP of interest (Syvänen, 1999) (Fig. 1.7). Due to its convenience and relative affordability, SBE is now used in many genetic and forensic applications. In this study SBE assays were used to refine the resolution of Y chromosome haplogroups commonly found in Africa, having also incorporated a few bi-allelic markers to delineate the common

non-African Y chromosomes following a hierarchical screening process. While many Y chromosome bi-allelic markers have not attained the frequencies needed to be classified as SNPs, for convenience the abbreviation, SNP, will be used from hereon when referring to bi-allelic markers.

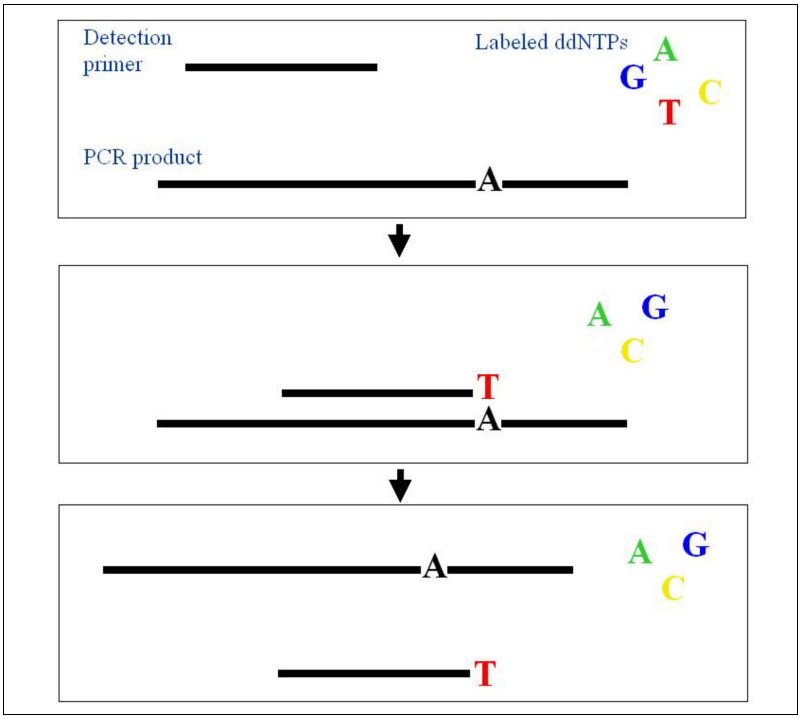

Figure 1.7: Diagram illustrating the principle and steps of the SBE technique. The detection primer anneals to the PCR product immediately upstream of the SNP. A fluorescently labelled ddNTP is then attached to the primer during extension.

1.5 Human Phylogeography

In 1987, Avise and colleagues attempted to unite the fields of phylogenetics and population genetics and in the process the concept of phylogeography resulted. Phylogeography is the study of the principles and processes governing the geographical distribution of genealogical lineages (Avise, et al., 1987), and was conceived to be "the phylogenetic analysis of geographically contextualised genetic data for testing hypotheses regarding the causal relationship among geographic phenomena, species distributions, and the mechanisms driving speciation" (Avise, et al., 1987). While this was an attempt to bridge the gap between micro-evolutionary processes and macro-evolutionary differences among species, phylogeography was found to be extremely useful in the examination of intra-species diversity and distribution, including that of modern humans.

Initially, mtDNA was the molecule of choice for phylogeographic application; however the Y chromosome has also been aptly utilized. The rapid discovery of markers on the Y chromosome within the last decade has allowed for the application of phylogeographic methods to human Y chromosome diversity over most of the globe; especially Europe (Di Giacomo, et al., 2004; Underhill, et al., 2010; Myres, et al., 2011), Asia (Sahoo, et al., 2006; Shi, et al., 2008), and Africa (Cruciani, et al., 2002; Cruciani, et al., 2004; Semino, et al., 2004; Batini, et al., 2011a).

1.6 Study Objectives

Evolution and recent historical events over the past 200 000 to 300 000 years have contributed in shaping the gene pool of sub-Saharan African populations. Using patterns of Y chromosome variation, it is possible to examine how males have contributed in shaping the gene pool among sub-Saharan African populations. The present study aimed to characterise the phylogeography of ancient African Y chromosome haplogroups, especially those within haplogroups A and B, through the examination and analysis of high resolution genetic data; generated from the screening of SNP and short tandem repeat (STR) markers in sub-Saharan African males.

The project entailed:

- Optimizing multiplex PCR and SBE assays for the high-throughput genotyping of Y chromosome markers.
- Screening for SNP and STR markers in 1667 male individuals to resolve Y chromosome haplogroups and haplotypes in a number of sub-Saharan African populations.
- Examining the range distribution of Y chromosome haplogroups, and inferring their evolutionary histories and the genetic affinities of sub-Saharan African populations.

2 SUBJECTS AND METHODS

2.1 Subjects

DNA samples from 1667 males from sub-Saharan Africa were analysed in the present study (Table 2.1). All DNA samples were collected with the subjects' informed consent, and this research was approved by the Human Research Ethics Committee (Medical) at the University of the Witwatersrand, Johannesburg (Protocol Numbers. M090576 / M050906; appendix C). Sampling was done by collecting peripheral blood in EDTA-tubes or through buccal swabs, from volunteers.

The sample included at least 105 ethnicities from 16 countries. However, due to the low sample numbers of some ethnicities, individuals were assigned a group identity (Table 2.1) for use in some of the analyses. The group identity was based on geographic, ethnic and/or linguistic self-affiliation. To illustrate, the !Xun and Ju|'hoansi, both Northern Khoisan (Ju) speakers, were assigned to the Northern Khoe-San group; while the Biaka and Mbenzele, who are Pygmy populations in the Central African Republic (CAR), were assigned to the Western Pygmy group.

Additional data from ongoing projects in the Human Genomic Diversity and Disease Research Laboratory (HGDDRL), as well as published Y chromosome data on other sub-Saharan African populations (Cruciani, et al., 2002; Semino, et al., 2002; YCC, 2002; Knight, et al., 2003; Luis, et al., 2004; Moran, et al., 2004; Wood, et al., 2005; Gonçalves, Spínola and Brehm, 2007; Rosa, et al., 2007; Tishkoff, et al., 2007; Hassan, et al., 2008;

Berniell-Lee, et al., 2009; Coelho, et al., 2009; de Filippo, et al., 2010; Gomes, et al., 2010; Batini, et al., 2011a) were included to increase sample numbers or as comparative data in the analyses (see Appendix D).

Table 2.1: Details of populations examined in the present study

Assigned Group ID	N	Country	Ethnicity	Language Group	n
Central African Central Bantu	164	Congo	Bembe	Central Bantu	1
			Mixed Central Bantu speakers	Central Bantu	2
		Democratic Republic of Congo	Bamboma	Central Bantu	2
			Fulero	Central Bantu	1
			Havu	Central Bantu	1
			Kango	Central Bantu	1
			Kanyok	Central Bantu	1
			Lemfu	Central Bantu	1
			Luba	Central Bantu	22
			Lunda	Central Bantu	3
			Manyanga	Central Bantu	93
			Mbala	Central Bantu	2
			Nande	Central Bantu	1
			Ndibu	Central Bantu	6
			Ngongo	Central Bantu	1
			Nombe	Central Bantu	1
			Ntandu	Central Bantu	4
			Pelende	Central Bantu	1
			Pende	Central Bantu	3
			Shi	Central Bantu	1
			Songe	Central Bantu	2
			Songo	Central Bantu	2
			Suku	Central Bantu	1
			Swahili	Central Bantu	3
			Yombe	Central Bantu	3
		Rwanda	Banyarwanda	Central Bantu	5

Table 2.1 cont.: Details of populations examined in the present study

Assigned Group ID	N	Country	Ethnicity	Language Group	n
Central African Ubangian	87	Central African Republic	Banda	Ubangian	1
			Gbaya	Ubangian	32
			Kpatili	Ubangian	3
			Lagba	Ubangian	3
			Nzakara	Ubangian	30
			Sangha-Sangha	Ubangian	11
			Yakoma	Ubangian	1
			Yakpa	Ubangian	1
			Zande	Ubangian	4
		Democratic Republic of Congo	Ngbandi	Ubangian	1
Northwest Bantu	32	Central African Republic	Mpiemo	Northwest Bantu	3
		Democratic Republic of Congo	Kusu	Northwest Bantu	1
			Libinza	Northwest Bantu	1
			Lokele	Northwest Bantu	1
			Mongo	Northwest Bantu	5
			Ngombe	Northwest Bantu	3
			Ntomba	Northwest Bantu	1
			Ohendo	Northwest Bantu	1
			Sakata	Northwest Bantu	2
			Sengele	Northwest Bantu	1
			Teke	Northwest Bantu	1
			Tetela	Northwest Bantu	4
			Yaka	Northwest Bantu	2
			Yansi	Northwest Bantu	6
Western Pygmy	41	Central African Republic	Biaka	Northwest Bantu	23
			Mbenzele	Northwest Bantu	18

Table 2.1 cont.: Details of populations examined in the present study

Assigned Group ID	N	Country	Ethnicity	Language Group	n
Nigerian mixed Benue-Congo	7	Nigeria	Igbo	Benue-Congo	4
			Urhobo	Benue-Congo	1
			Yoruba	Benue-Congo	2
East African Central Bantu	177	Tanzania Uganda	Swahili	Central Bantu	17
			Banyarwanda	Central Bantu	1
			Fumbira	Central Bantu	11
			Ganda	Central Bantu	2
			Hororo	Central Bantu	21
			Kiga	Central Bantu	77
			Konjo	Central Bantu	6
			Nyankole	Central Bantu	40
			Soga	Central Bantu	1
			Tooro	Central Bantu	1
East African Nilotic	1	Uganda	Acholi	Nilotic	1
South East African Central Bantu	4	Malawi	Ngoni	Central Bantu	1
			Nyanja	Central Bantu	2
			Tonga	Central Bantu	1
Southern African Central Bantu	71	Zambia	Bemba	Central Bantu	17
			Kaonde	Central Bantu	3
			Kunda	Central Bantu	1
			Mambwe	Central Bantu	1
			Mwanga	Central Bantu	3
			Ngoni	Central Bantu	3
			Nkoya	Central Bantu	1
			Nsenga	Central Bantu	2
			Nyanja	Central Bantu	25
			Tonga	Central Bantu	13
			Tumbuka	Central Bantu	2

Table 2.1 cont.: Details of populations examined in the present study

Assigned Group ID	N	Country	Ethnicity	Language Group	n
Southeastern Bantu	343	Tanzania	Tsonga	Southeastern Bantu	1
		Lesotho	Sotho	Southeastern Bantu	7
			Southern Ndebele	Southeastern Bantu	1
		Mozambique	Tsonga	Southeastern Bantu	1
		South Africa	Hlubi	Southeastern Bantu	1
			Pedi	Southeastern Bantu	13
			Sotho	Southeastern Bantu	81
			Mixed Southeastern Bantu speakers	Southeastern Bantu	57
			Southern Ndebele	Southeastern Bantu	6
			Swazi	Southeastern Bantu	4
			Tsonga	Southeastern Bantu	9
			Tswana	Southeastern Bantu	24
			Venda	Southeastern Bantu	4
			Xhosa	Southeastern Bantu	36
			Zulu	Southeastern Bantu	63
		Zambia	Lozi	Southeastern Bantu	27
			Shona	Southeastern Bantu	2
		Zimbabwe	Northern Ndebele	Southeastern Bantu	3
			Shona	Southeastern Bantu	3
Southwestern Bantu	112	Namibia	Herero	Southwestern Bantu	45
			Himba	Southwestern Bantu	25
			Ovambo	Southwestern Bantu	42
Dama	30	Namibia	Dama	Central Khoisan (Khoi)	29
		South Africa	Dama	Central Khoisan (Khoi)	1
Central Khoe-San	74	Angola	Khwe	Central Khoisan (Khoi)	51
		Botswana	Naro	Central Khoisan (Khoi)	2
		Namibia	Nama	Central Khoisan (Khoi)	17
		South Africa	Nama	Central Khoisan (Khoi)	4

Table 2.1 cont.: Details of populations examined in the present study

Assigned Group ID	N	Country	Ethnicity	Language Group	n
Khoe-San/Bantu admixed	21	Botswana	ǀGui-ǁGhana-Kgalagari	Central Khoisan / Southeastern Bantu	21
Northern Khoe-San	121	Angola	!Xun	Northern Khoisan (Ju)	80
		Namibia	Jul'hoansi	Northern Khoisan (Ju)	41
Southern Khoe-San	41	South Africa	/Xam	Southern Khoisan (Tuu)	1
			Khomani	Southern Khoisan (Tuu)	40
South African Coloured	148	South Africa	South African Coloured	Indo-European	148
South African Cape Malay	17	South Africa	Cape Malay	Indo-European	17
European Descent	156	South Africa	South African White	Indo-European	156
Indian Descent	20	South Africa	South African Indian	Indo-European + Dravidian	20

2.2 Methods

Details of reagents and recipes used are listed in Appendix E and G.

2.2.1 DNA extraction

DNA from EDTA-blood was extracted using the salting-out method (Miller, Dykes and Polesky, 1988) and the Gentra Puregene Buccal Cell Kit (Qiagen, Germany) was used to extract DNA from buccal swabs according to the manufacturer's instructions.

DNA was quantified using the NanoDrop ND-1000 Spectrophotometer (Lab-VIEW®, Coleman Technologies Inc, FL, USA) and diluted to the required concentrations using double distilled water (ddH_2O).

2.2.2 Y chromosome molecular methods

Individuals in the sample were screened for Y chromosome variation by genotyping both haplogroup-defining SNPs and STR markers.

i. SNP genotyping

Seven multiplex SBE assays, which incorporated 60 Y chromosome SNPs described in the YCC phylogeny 2003 (Jobling and Tyler-Smith, 2003), were developed. These resolved 61 Y chromosome haplogroups (Table 2.5).

Primer design

The sequences of the regions encompassing the polymorphisms were taken from GenBank. The PCR and SBE primers were designed using Primer3 software (Rozen and Skaletsky, 2000), before aligning them to human genomic sequences using the NCBI BLAST alignment tool to confirm template specificity. The screening software, AutoDimer (Vallone and Butler, 2004) was used to check for primer-dimer and hairpin loop formation. HPLC-purified primers were purchased via Roche from Metabion, diluted to 100μM, and frozen.

PCR primer lengths ranged from 20 to 27 mers; and GC percentage varied between 30% and 60%. Amplicons were designed to differ slightly in size to distinguish them following agarose gel electrophoresis to check the success of the PCR. In total, 53 pairs of PCR primers were designed encompassing all 60 SNPs (Table 2.2). Fewer pairs of primers were needed, as some SNPs were co-amplified on the same amplicons (M13 and M14; M40 and M41; M58 and M155; M123 and M281; M81 and M154; M85, M148 and M149).

Poly-C or Poly-GACT tails of differing lengths were added to the 5' end of most SBE primers (Table 2.3), so as to differentiate between them during capillary electrophoresis. SBE primer lengths ranged from 25 to 80 mers, and differed in size from each other by four to five mers.

Table 2.2: PCR primer sequences, amplicon lengths and final concentrations

SNP	Primer (5' – 3')		Fragment size (mers)	Concentration (µM)	Assay
	Forward	Reverse			
M170	CTAGTATGCTTCACACAAATGCG	GACCACACAAAAACAGGTCCTC	390	0.08	YSNP1
M207	AAGGGCAAGCAAAATAGCAATAC	TGTTCGCTGCTACGAATCTTT	363	0.08	
M201	CATGGGTAATTCGGTTGTTACC	CTAAACATCATGGTGTGACGAAC	331	0.08	
M168	GGTTTGAATGAGACTGGGTCA	TGGTAATCTCATAGGTCTCTGACTG	295	0.08	
M343	AGGTAGGAGGATCCAAAGCTGA	CACCTTTGTCCTCTTGCTCTTT	276	0.08	
M9	TGCAGCATATAAAACTTTCAGGAC	TTCCTCATTTTTGAAGCTCGTG	241	0.08	
M69	TCAGCCATTTCACCAAAACTCT	CTGAAGAAACAAACCTACCTGGAA	233	0.08	
M89	CCAAGCTGGTGAGTCTTATCCT	GCAGAATAGCTGCTCAGGTACA	215	0.08	
M172	CAGAAGATGCCCCATTATATCCT	ACTCCATGTTGGTTTGGAACAG	208	0.08	
M198	TAGGCACTTGGGAACTTACACTC	TTCTTGTGATAGCATGCCGTTT	178	0.08	
SRY10831	CATCCAGTCCTTAGCAACCATT	AATGACACAAGGCACCACATAG	163	0.08	
P28	TTTTGAGAGAAGACAAGGGGGATA	TTGGAGGGACATTATTCTCCTGA	559	0.20	Hg-A
M13-M14[a]	ATCACGCCCCTCTCATTGTC	AGCTCTAGATAAAAGCACATTGACAC	457	0.08	
M91	GATCACAAAGACCTGGACAGATTACA	AACGGAAATGCCAAGAATCGTA	429	0.05	
M31	GCTGAAACAATAGTTCTTCACAATGG	CAGTCCTATGCATAATGCCGTGT	400	0.03	
M114	GCCTTGATTTTCTTCGTACTTCATAAG	CCAGTTTCTCACTGAGTTCATTCCTT	370	0.06	
M28	GGGCTTCAGTTCTTGACGCTAC	CCGTCTTAATTTTGCGGTATTCAA	329	0.04	
M51	AAACCACACCTGTCTTACCAGAGC	CTGTTCCCCAGTTTTCAATCTCC	293	0.04	
M171	GGCTGTGTGGAGTATGTGTTGG	CAAATATCCTGCCCCAGCTTAGT	217	0.04	
M118	TCCCTTGAAATTAAGGACAACAACT	CATTCTTCTCAACCAGCTGACACT	167	0.06	
M43	GACTCCATAAGCAAAAGGTCATCAA	AAAAGAAGTTGAGGACTGGAGCA	518	0.12	Hg-B
M112	GGCCATGCTAACAGAGATCTGAC	CACAGTTCAATTCTTGTCTGTTGC	493	0.08	
M152	AAGCAAAAAGCTCCTTCTGAGGT	CAGAAGGTGGATCAGGGTAGAAA	381	0.08	
M182	CATTTTTGTGTCAGGTATCCTTTGT	CAAGACGGCGTATCAACTCAAG	368	0.12	
M108	GCTTTTCTAACACCACCCATGAC	TATGTGATAGAGGTGGCTTTAAGTGG	342	0.08	
M150	CCAGGCTAGCAGTGGAGATGAA	AGGGTGGACTGCTGACCTACTTT	312	0.08	
M146	TTACAGGTGGAATGGGGTGTTAC	GAGAAGAACTGCCTTCCATGACATA	279	0.08	
M60	CCTGATGTGGACTCAACCTTGTA	TGTTCATTATGGTTCAGGAGGAG	250	0.08	

Table 2.2 cont. : PCR primer sequences, amplicon lengths and final concentrations

SNP	Primer (5' – 3')		Fragment size (mers)	Concentration (µM)	Assay
	Forward	Reverse			
M211	CACTGCACACACTACACTGACCAC	ATGTTGATTGGGTAGAGCCCTTT	386	0.06	Hg-B2b
P6	TATTAGGGAAATCACTCAGGATGGT	TCTACGAATGTTTAACCTCAGATACC	343	0.12	
P8	AGTTGTTGGAAAGCCTCTGTTCA	TGATACTAGACGTGGCATCTTGTCA	313	0.06	
M115	TGCCATGCTTGTTTCTTAATCCA	AACTATGTTGCACATCAGCCTCA	270	0.12	
P7	GGCCAAAGCCTAGAATGAAATTG	AAGTGCTTGTCCAAGGCAGTATAA	228	0.16	
M30	ACAAATCATGAGCTTACAGAACCTG	GGCACAGCCAGATAACCCTACA	200	0.12	
M40-M41[a]	TAGCTGGTATGACAGGGGATGAT	GGGTAGGATAGGCTAGCTATTACGC	435	0.08	Hg-E
M2	GGAGAAGAACGGAAGGAGTTCTAA	ACTTGCCAGAGACTTCCAGTTTG	372	0.12	
M85[b]	GAACTGGCATCCAATACTAGCTGA	TCACCTCTTTTGTATTGGCTTCTTC	350	0.08	
P2	TGGTCTGGTAACACCCATAAAGGT	GCAGTTTTCTCAGATGCTTCTCCTA	335	0.08	
M35	GCCTAAAGAGCAGTCAGAGTAGAATG	GAGAATGAATAGGCATGGGTTCA	303	0.08	
M75	GTCACATTCCACACATCAAGAAAACT	GTGAATCTCTGCCCAGAAAAGAAAA	274	0.12	
M44	ATTGGATATGGAAGCCAGTCTCA	ATGTGTTTGAGGACCACCCTAGA	250	0.12	
M33	GGCTTCTGTTCAATTTTCCTTTGAT	TTATTTGTTGAAGCCCCCAAGAG	223	0.08	
M10	GTTCAAGACAATGAAGGGAGAGACT	TGACATTGACCTGCAGCATAGG	520	0.08	Hg-E1b1ba
M191	GAGCAAGTACAGCGAGCAGTAAG	GGTTTAACACAATGCAGGTCAATTC	480	0.08	
M154[b]	CAATGGAGGCTATAGGTGATTGC	CTGTTTGTTCATGGAGATGTCTGTA	461	0.08	
M149[b]	GAACTGGCATCCAATACTAGCTGA	TCACCTCTTTTGTATTGGCTTCTTC	350	0.06	
M116	TATGAAGTACGAAGAAAATCAAGGCTA	TGGGTAGAAAAACTGCAAGTAGATGA	328	0.12	
M58-M155[a]	TGGGCCTGACCTCTTAACTTGTA	CATAATAAGCTAAGAAACATCCAGCC	293	0.06	
M81[b]	CAATGGAGGCTATAGGTGATTGC	CTGTTTGTTCATGGAGATGTCTGTA	461	0.10	Hg-E1b1b1
M123-M281[a]	CTAATTCATGCTCTCAGGGGAAA	ATAACCTCTGGAAGTGTGTCTTTACCT	404	0.10	
M107	AATCCCACCTCACATACACATAAGC	AGGGGTTGACAAGAAAAAGGAATA	386	0.06	
M148[b]	GAACTGGCATCCAATACTAGCTGA	TCACCTCTTTTGTATTGGCTTCTTC	350	0.08	
M78	ATGGCTGTATGGGTTTCTTTGACT	CGGAATATGGACAGTCATCGTATT	330	0.08	
M165	CAAGTCAGCAAGGAGTAGGTGGA	TTGCACTGACACAAGTTATCTCCCT	293	0.08	
M34	GATAACCTCATTGTGGAGAGCACTT	ATGCTAAAGCAAGTAACCCTGTGG	254	0.10	
M136	ACCAACCGTATTACCTTCTCCTCA	CATGAGTCCAAAGTATAGTGGGCTA	226	0.10	

[a] Due to the close proximity of the SNPs, a single amplicon was used.

[b] Again, a single amplicon was used for these SNPs, however, in different assays.

Multiplex PCRs

Primer design was verified by performing simplex PCR, using a GeneAmp PCR system 9700 (Life Technologies, CA, USA), for each of the primer pairs. Thereafter, the multiplex PCRs were optimized to work with DNA at a concentration of 10ng/µl (see Table A2), and were catalysed using *FastStart Taq* DNA Polymerase (Roche, Basel, Switzerland). Relative primer concentrations were adjusted in order to obtain balanced amplification of amplicons within each multiplex. The thermal cycler programs were as follows: one cycle at 95ºC for six minutes, 35 cycles at 95ºC for 30 seconds, 54 ºC (for YSNP1), 55ºC (for Hg-A, Hg-B, Hg-B2b, Hg-E, and Hg-E1b1a) or 61ºC (for Hg-E1b1b1) for 30 seconds, extending at 72 ºC for 30 seconds, and a final extension of 72ºC for ten minutes. In order to confirm that the multiplex PCRs produced the required amplification products, 5µl of each multiplex PCR product was electrophoresed on a 2% Metaphore® agarose gel (Cambrex, NJ, USA).

Multiplex SBE

Excess PCR primers and dNTPs were eliminated from the PCR product mixture, following amplification, using an enzymatic purification method. One unit of Exonuclease I (*Exo I*) (New England Biolabs, MA, USA) and 0.5 units of Shrimp Alkaline Phosphatase (SAP) (USB, OH, USA) were added to five microlitres of amplification product, and the resultant mixture incubated for one hour at 37ºC, followed by 15 minutes at 75ºC.

The multiplex SBE reactions were performed in a final volume of 5µl, comprised of 1.5µl purified amplification product, 1.5µl of double distilled water, 1µl of SNaPshot Multiplex Ready Reaction Mix (Life Technologies, CA, USA), and 1µl of SBE primer mix, specific

to the assay being conducted (see table 2.3 for final primer concentrations). The thermal cycler program was as follows: 96ºC for 10 seconds, 50ºC for 5 seconds, and 60ºC for 30 seconds, for 35 cycles.

Following the SBE reaction, excess ddNTPs were removed through the addition of 0.5U of SAP to the 5µl SBE product. The mixture was incubated for one hour at 37ºC, followed by 15 minutes at 75 ºC.

Capillary electrophoresis

Following post-extension treatment, 2µl of SBE product was mixed with 0.5µl of the internal size standard, GS120LIZ (Life Technologies, CA, USA), and 7.5µl Hi-Di formamide (Life Technologies, CA, USA). This was then run on a 3130*xl* Genetic Analyzer (Life Technologies, CA, USA). The SNaPshot protocol was originally optimized for use with POP-4 polymer; modifications recommended by Applied Biosystems were incorporated for use with POP-7 polymer (Applied Biosystems Manual P/N: 4367258). The resultant electropherograms (Fig. 3.1) were analyzed using GeneMapperID v3.2 software (Life Technologies, CA, USA).

Assay validations

Some of the markers used in the SBE assays were validated using a set of control samples, previously screened using RFLP assays. Those markers for which samples of known haplogroup were unavailable were sequenced in order to confirm the presence of the polymorphism. The above procedures were discussed in Naidoo, et al. (2010) (see Appendix F).

Table 2.3: SBE primer sequences and final concentrations

SNP		SBE Primer (5' – 3')	Size (mers)	Concentration (µM)	Assay
SRY10831	FW	$(C)_3$CTCTTGTATCTGACTTTTTCACACAGT	30	0.10	YSNP1
M168	FW	$(C)_{12}$TGGAGTATGTGTTGGAGGTGAGT	35	0.40	
M89	RV	$(GACT)_2(C)_{10}$CAACTCAGGCAAAGTGAGAGAT	40	0.40	
M201	FW	$(GACT)_2(C)_9$AGATCTAATAATCCAGTATCAACTGAGG	45	0.40	
M69	FW	$(GACT)_4(C)_{11}$GGAGGCTGTTTACACTCCTGAAA	50	0.40	
M170	FW	$(GACT)_4(C)_9$ACTATTTTATTTACTTAAAAATCATTGTTC	55	0.80	
M172	FW	$(GACT)_7(C)_{12}$CCAAACCCATTTTGATGCTT	60	0.40	
M9	FW	$(GACT)_8(C)_{11}$AAACGGCCTAAGATGGTTGAAT	65	0.40	
M207	FW	$(GACT)_8(C)_{11}$GCAAATGTAAGTCAAGCAAGAAATTTA	70	0.80	
M198	FW	$(GACT)_9(C)_9$TCAGTATACCAATTAATATTTTTGAAAGAG	75	0.80	
M343	FW	$(GACT)_{13}(C)_9$AGAGTGCCCTCGTGTTCCA	80	0.40	
M91	FW	CCTACATTGCTATTCTGTTTTTTTT	25	0.60	Hg-A
M31	RV	$(C)_8$CCACTGCTGTTCTGTCTACCA	29	0.60	
M14	RV	$(C)_5$CTTCATTAACTTTTTTAAACTGCTTATA	33	0.60	
M114	RV	$(C)_{15}$AGCTGTACAAGGCTCTTCAAAT	37	0.60	
P28	FW	$(C)_{14}$GGTTAAAAGAAAAAAGCTCTCAGATAG	41	0.40	
M28	RV	$(C)_{27}$TCGAGGTCCTCTGGCATC	45	0.50	
M51	RV	$(C)_{29}$CTCTGATCCCTGTTGGAAGC	49	0.50	
M13	FW	$(C)_{31}$GTAGGTTAAGGGCAAGACGGTTA	54	0.60	
M171	RV	$(C)_{32}$AGGTCTCTGACTGTTCAGTTTTATT	57	0.50	
M118	RV	$(C)_{35}$CAGCTGACACTTGTGTTTTCTTTATA	61	0.20	
M60	FW	$(C)_3$TTACATTTCAAAATGCATGACTTAAAG	30	0.40	Hg-B
M146	RV	$(C)_{11}$CTAAAACCCAGTGTTAATTACCCG	35	0.80	
M182	FW	$(C)_{13}$CTTAAAGCAGTGGTTAATGTAAACAAA	40	0.80	
M150	FW	$(C)_{22}$TGCCCACACACACAGATAGAAGT	45	0.80	
M152	FW	$(C)_{23}$GCTTTCTCCTGATAATGTTCTTCTTCT	50	0.80	
M108	RV	$(C)_{27}$CTTTTCTCTGACATTCAGGTATAGTTTC	55	0.30	
M43	FW	$(C)_{39}$CTCTTTCCATGGCCAACAAAC	60	0.40	
M112	FW	$(C)_{38}$AAAGAGGTGAGATAAAAACAAAGCAGT	65	0.40	

Table 2.3 cont.: SBE primer sequences and final concentrations

SNP		SBE Primer (5' – 3')	Size (mers)	Concentration (µM)	Assay
P6	FW	$(C)_3$TCAATAGAGGTTTCCACAGTTAAGTCT	30	0.10	Hg-B2b
M115	FW	$(C)_3$CAGAGTTTAAATTAGTATTTGATTTCACATTA	35	0.80	
M30	FW	$(C)_{14}$ATCATGTTTTAAGTCCTGACATCTGT	40	0.10	
P7	FW	$(C)_{21}$CCATCACCTGGGTAAAGTGAATTA	45	0.40	
P8	FW	$(C)_{27}$GCAGCTCACCTTTCATTTAGGTC	50	0.20	
M211	FW	$(C)_{30}$TAGGCAAAAGGATGTTAACAACAAG	55	0.80	
M40	RV	$(C)_{10}$TCTTCACCCTGTGATCCGCT	30	0.60	Hg-E
M33	FW	CGATCTGTTCAGTTTATCTCATAAGTTACTAGTTA	35	0.30	
M44	RV	$(C)_{11}$AGGAAATCTCCTAACCTTCTAGTACACTG	40	0.40	
M75	FW	$(C)_{20}$AAAAGACAATTATCAAACCACATCC	45	0.10	
M41	FW	$(C)_{30}$TGGCCAACATGGTGAAACTG	50	0.50	
M85	RV	$(C)_{24}$GCTTGTGTTCTATTAAGTGTAGTTTTGTTAG	55	0.20	
P2	RV	$(GACT)_8(C)_8$AGGTGCCCCTAGGAGGAGAA	60	0.40	
M2	RV	$(GACT)_9(C)_6$CCCTTTATCCTCCACAGATCTCA	65	0.60	
M35	RV	$(GACT)_{10}(C)_9$TTCGGAGTCTCTGCCTGTGTC	70	0.80	
M58	FW	$(C)_5$ATTTATTGTCTTCTGCAGAATTGGC	30	0.10	Hg-E1b1a
M116	FW	$(C)_5$GCTTTCTGAAAAAATAATTTCAAACTGATA	35	0.40	
M149	FW	$(C)_9$CTAACAAAACTACACTTAATAGAACACAAGC	40	0.10	
M154	RV	$(C)_{16}$GTGTTACATGGCCTATAATATTCAGTACA	45	0.40	
M155	RV	$(C)_{23}$AATTCAGAATATTTCATCTCTGGTCAC	50	0.40	
M10	FW	$(C)_{26}$AATTTTTTTGTTTATTCCCAATGATCTTA	55	0.50	
M191	FW	$(GACT)_5(C)_{10}$ATTTACATTTTTTTCTTTACAACTTGACTA	60	0.40	
M78	FW	AATTGATACACTTAACAAAGATACTTCTTTC	31	0.80	Hg-E1b1b1
M148	RV	$(C)_7$TTTCTAGGTAACGTATGTAGACATTTCTG	36	0.80	
M81	FW	$(C)_{15}$AGAGGTAAATTTTGTCCTTTTTTGAA	41	0.80	
M107	FW	$(C)_{18}$TAAGCCAACGTATTAACCTTCTAATTTC	46	0.20	
M165	RV	$(C)_{20}$AAATATTTCAGGTAAAACCACTCTATTAGTA	51	0.40	
M123	FW	$(C)_{27}$AAAGTCACAGTATCTGAACTAGCATATCA	56	0.80	
M34	FW	$(GACT)_7(C)_{13}$GCCTGGCTTCCACCCAGGAG	61	0.20	
M136	RV	$(GACT)_8(C)_{12}$GGTGAGCAGCATTGAGGAAGAC	66	0.10	
M281	RV	$(GACT)_8(C)_{11}$AGGTTGCACAAACTCAGTATTATTAAAC	71	0.80	

FW = forward orientation RV = reverse orientation

ii. Other SNP genotyping assays

While most of the individuals were resolved to their terminal haplogroup branches using the abovementioned SBE assays, 50 were resolved using RFLP assays where those SNPs were not included in the SBE assays (Table 2.4). Also, 181 individuals from !Xun, Dama, Herero, Himba, Khwe and Ovambo populations which were genotyped previously using only RFLP assays were included in the study, in order to increase sample sizes and improve representation of some of these populations. As such, the level of haplogroup resolution for these individuals may, potentially, be lower than that of the other samples in the dataset. Details of reagents and conditions for the RFLP assays are listed in Appendix G. Additionally, the SNPs M6 and M49 were genotyped using ABI Taqman® assays in a 7900HT Fast Real-Time PCR System (Life Technologies, CA, USA). Details of reagents and conditions for the Taqman® assays (Life Technologies, CA, USA) are listed in Appendix G. In total, 72 Y chromosome haplogroups were screened for.

Table 2.4: Y chromosome SNPs genotyped using RFLP assays in the present study

Marker	Haplogroup	Aternate Name
M11	L	L-M11
M130	C	C-M130
M175	O	O-M175
M74	PQ	PQ-M74
P12F2	J	J-P12F2

iii. STR screening

STR markers were genotyped in order to generate haplotypes for use in examining intra-haplogroup variation.

Table 2.5: Y chromosome SNPs and respective haplogroups included in the SBE assays

Assay	SNP	Haplogroup	Alternate Name	Assay	SNP	Haplogroup	Alternate Name
YSNP1	SRY10831	BT (or R1a1)		**Hg-E**	M40	E	E-M40
	M168	CT			M33	E1a	E-M33
	M89	F	F-M89		M44	E1a1	E-M44
	M201	G	G-M201		M75	E2	E-M75
	M69	H	H-M69		M41	E2a	E-M41
	M170	I	I-M170		M85	E2b1	E-M85
	M172	J2	J-M172		P2	E1b1	E-P2
	M9	K	K-M9		M2	E1b1a1	E-M2
	M207	R	R-M207		M35.1	E1b1b1	E-M35.1
	M198	R1a1a	R-M198				
	M343	R1b	R-M343	**Hg-E1b1a**	M58	E1b1a1a1a	E-M58
					M116.2	E1b1a1a1b	E-M116.2
Hg-A	M91	BT			M149	E1b1a1a1c	E-M149
	M31	A1a	A-M31		M154	E1b1a1a1g1c	E-M154
	M14	A1b1a1	A-M14		M155	E1b1a1a1d	E-M155
	P28	A1b1a1a1	A-P28		M10	E1b1a1a1e	E-M10
	M114	A1b1a1a2a	A-M114		M191	E1b1a1a1f1a	E-M191
	M28	A1b1b1	A-M28				
	M51	A1b1b2a	A-M51	**Hg-E1b1b1**	M78	E1b1b1a1	E-M78
	M13	A1b1b2b	A-M13		M148	E1b1b1a1c1	E-M148
	M171	A3b2a*	A-M171		M81	E1b1b1b1a	E-M81
	M118	A1b1b2b1	A-M118		M107	E1b1b1b1a1	E-M107
					M165	E1b1b1b1a2a	E-M165
Hg-B	M60	B	B-M60		M123	E1b1b1b2a	E-M123
	M146	B1a	B-M146		M34	E1b1b1b2a1	E-M34
	M182	B2	B-M182		M136	E1b1b1b2a1a1	E-M136
	M150	B2a	B-M150		M281	E1b1b2	E-M281
	M152	B2a1a	B-M152				

Table 2.5 cont.: Y chromosome SNPs and respective haplogroups included in the SBE assays

Assay	SNP	Haplogroup	Alternate Name	Assay	SNP	Haplogroup	Alternate Name
	M108	B2a2 (or B2b1a1b1)	B-M108.1 (or B-M108.2)				
	M43	B2a2a	B-M43				
	M112	B2b	B-M112				
Hg-B2b	P7	B2b1a	B-P7				
	M115	B2b1a1a	B-M115				
	M30	B2b1a1b	B-M30				
	M211	B2b1a1c	B-M211				
	P8	B2b1a2	B-P8				
	P6	B2b1b	B-P6				

STR genotyping was performed in 1236 individuals, using the AmpFlSTR® YFiler™ PCR Amplification Kit (Life Technologies, CA, USA) which included 17 STR markers – DYS19, DYS385a, DYS385b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and GATA H4. In addition, 250 individuals were genotyped using the PowerPlex® Y System (Promega, WI, USA) which included 12 STR markers - DYS19, DYS385a, DYS385b, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439. These were done prior to the adoption of AmpFlSTR® YFiler™ protocol in the lab. Details of reagents and conditions for the AmpFlSTR® YFiler™ and PowerPlex® Y multiplex systems are listed in Appendix G.

The 181 low resolution individuals mentioned in section 2.2.2.ii were also not screened using either of the above two kits. They were genotyped for 7 STR markers previously - DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 – using an in-house protocol. Due to their low resolution, these samples were excluded from some analyses.

The DYS389 STR locus is composite in nature; DYS389I is contained within DYS389II (Kayser, et al., 1997). The locus contains phylogenetically informative regions as well as fast evolving regions that obscure structure. To account for this DYS389I was subtracted from DYS389II to give DYS389c, thus excluding some of the obscuring information. DYS389I and DYS389c were then used in the analyses. The DYS385a/b duplicated loci were excluded from the analyses due to difficulties in distinguishing between the two loci, as they are co-amplified. A result of this co-amplification is the inability to individually assign allele sizes to the each of the loci (alleles are called as an ordered pair). Thus, their use could potentially affect the results of the analyses (Balaresque, et al., 2006).

2.3 Data Analyses

The SNPs genotyped in the abovementioned assays were used to allocate the Y chromosomes in the dataset to haplogroups according to the nomenclature of ISOGG 2013, and haplogroup frequencies were calculated.

2.3.1 Population affinities

Relationships among the populations examined in the present study were analysed based on haplogroup frequency data and STR haplotype data. In order to include the maximal number of samples and populations in the analysis, a seven-STR haplotype (DYS19, DYS389I, DYS389c, DYS390, DYS391, DYS392, and DYS393) was used. Inter-population distances were calculated by generating a pairwise Fst distance matrix from haplogroup frequency data and a pairwise Rst (Slatkin, 1995) distance matrix from STR haplotype data. Both matrices were calculated, and checked for correlation using a Mantel test in Arlequin v3.5.1.2 (Excoffier and Lischer, 2010). The matrices were then visualized through multidimensional scaling (MDS) plots and cluster analysis in PAST v.2.10 (Hammer, Harper and Ryan, 2001). Exact tests of population differentiation (Raymond and Rousset, 1995) were performed using Arlequin v3.5.1.2, based on both haplogroup frequency data and STR data. Apportionment of variation in the different population groups was tested by conducting AMOVA analysis. This was also implemented in Arlequin v3.5.1.2.

2.3.2 Phylogeographic analyses

Haplogroup frequency spatial surfaces were generated using the Kriging procedure (Delfiner, 1976) in SAGA-GIS v2.0.6 (www.saga-gis.org). Haplotype diversity was estimated using Arlequin v3.5.1.2 while mean STR allelic variance was estimated using R v2.1.3.1 (Busby, G. and Capelli, C., personal communication, 2011). The haplotype diversity and variance analyses were conducted using a ten-STR haplotype (DYS19, DYS389I, DYS389c, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439).

Phylogenetic relationships within haplogroups were investigated by sequentially performing reduced-median and median-joining (RM-MJ) procedures (Bandelt, et al., 1995; Bandelt, Forster and Röhl, 1999; Polzin and Daneshmand, 2003) using Network 4.6.0.0 (Fluxus-engineering.com, http://www.fluxus-engineering.com/sharenet.htm). STR markers were weighted proportionally to the inverse of STR allelic variance (Cruciani, et al., 2004), while SNPs (when used) were given maximum weighting to compensate for the orders of magnitude differences in mutation rate between STRs and SNPs. The haplotypes were coloured according to the legends in the figures, and circle sizes were proportional to absolute frequency, with the smallest representing n = 1. The networks were constructed based on a 15 STR haplotype (DYS19-DYS389I-DYS389c-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439-DYS448-DYS456-DYS458-DYS635-GATA H4) – plus SNPs if present within the haplogroup – or a reduced 10 STR haplotype (DYS19-DYS389I-DYS389c-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439) which allowed for the incorporation of lower resolution comparative data into the analyses. The exceptions to this included the A-M51 network and the B-M112 networks, which

omitted DYS448 (A-M51) and DYS439 (B-M112). This was due to the presence of null alleles at these loci in the populations used for the networks.

TMRCA for haplogroups were estimated using the Bayesian Analysis of Trees With Internal Node Generation program (BATWING) (Wilson, Weale and Balding, 2003). This estimate was done assuming the population growth model with exponential growth from an initially constant-size population without population structure. Weakly informative priors were used for N, the effective population size before expansion [gamma(1,0.0001): mean = 10000, SD = 10000]; α, the rate of growth per generation [gamma(2,400): mean = 0.005, SD = 0.0035]; and β, the time in coalescent units when exponential growth began [gamma(2,1): mean = 2, SD = 1.41] (Wilson, Weale and Balding, 2003; Xue, et al., 2006). Forty unique event polymorphisms (UEPs) were included in the analysis. The same mutation rate was assumed for all UEP loci. A stepwise mutation model (SMM) was assumed for the STR loci (DYS19-DYS389I-DYS389c-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439), with locus-specific mutation rates used [gamma(35,15225), gamma(34,13474), gamma(49,13445), gamma(31,14747), gamma(38,14621), gamma(6,14553), gamma(14,13399), gamma(3,9808), gamma(51,9782), gamma(12,9787)], which were taken from the Y chromosome Haplotype Reference Database (YHRD), release 37 (Willuweit, Roewer, International Forensic Y Chromosome User Group 2007, www.yhrd.org/). Generation time was assumed to be 25 years, and median values of posterior data were used in the estimations. In total, 2.4 million Markov chain Monte Carlo (MCMC) samples were collected. A 300k burn-in was used. Equilibration was determined by examining '*lltimes'* and effective ancestral population, '*N*' (Haber, et al., 2011), and by examining median and 95% confidence intervals (CI) of *N* and *lltimes* for various subsets of sequential sample segments.

Additional unpublished population data from the HGDDRL, listed in appendix D, was included in the TMRCA analysis.

3 RESULTS

3.1 SBE assay development

3.1.1 SNP Selection and Screening Strategy:

Seven multiplex SBE assays, which incorporated 60 Y chromosome markers described in the YCC Phylogeny 2003 (Jobling and Tyler-Smith, 2003) were developed, which resolved 61 Y chromosome haplogroups. The first multiplex, YSNP1, consisted of the markers SRY10831, M168, M89, M201, M69, M170, M172, M9, M207, M198, and M343 (Fig. 3.1A). YSNP1 resolved Y chromosomes into either the African haplogroups (A, B, or E) or Eurasian haplogroups found occasionally in African males. NB: The marker, SRY10831, initially resolves haplogroup BR; while its reversion is used to define haplogroup R1a.

Any sample found to harbour the ancestral state at all markers within YSNP1 was screened using the multiplex assay, Hg-A. This multiplex consisted of the markers, M91, M31, M14, M114, P28, M28, M51, M13, M171, and M118 (Fig. 3.1B); and was used to resolve the sub-clades of haplogroup A. Samples found to be derived at SRY10831, but ancestral at all other markers within YSNP1 were screened using the multiplex assay, Hg-B. This multiplex consisted of M60, M146, M182, M150, M152, M108, M43, and M112 (Fig. 3.1C), and resolved the sub-clades of haplogroup B. Those samples with the derived allele at M112 were screened further using the multiplex assay, Hg-B2b, which contained the markers P6, M115, M30, P7, P8, and M211 (Fig. 3.1D), providing resolution of

haplogroup B2b samples to the terminal branches of the phylogeny. While M108 recurs in haplogroup B2b, resolving haplogroup B2b1a1b1 its presence in the Hg-B multiplex assay would be sufficient to resolve both its occurrences in haplogroup B; negating the need to include it in the Hg-B2b assay.

Those samples found to be derived at SRY10831 and M168, while remaining ancestral at all other markers within YSNP1, could be assigned to haplogroups C, D, or E. These samples were then screened using the Hg-E multiplex assay, which consisted of M40, M33, M44, M75, M41, M85, P2, M2, and M35 (Fig. 3.1E). Samples found to be derived for M2 or M35 would fall into haplogroups E1b1a or E1b1b1, respectively. E1b1a Y chromosomes were further resolved using the assay, Hg-E1b1a; a multiplex comprised of the markers M58, M116, M149, M154, M155, M10, and M191 (Fig. 3.1F). Those Y chromosomes assigned to haplogroup E1b1b1 were screened further using the multiplex assay, Hg-E1b1b1, which consisted of the markers M78, M148, M81, M107, M165, M123, M34, M136, and M281 (Fig. 3.1G). When a terminal haplogroup could not be assigned to a sample after screening with the SBE assays, further screening was done using RFLP and Taqman® assays. This hierarchical screening approach facilitated the resolution of the relevant haplogroup in an individual after one, two, or at most, three reactions, depending on the haplogroup present.

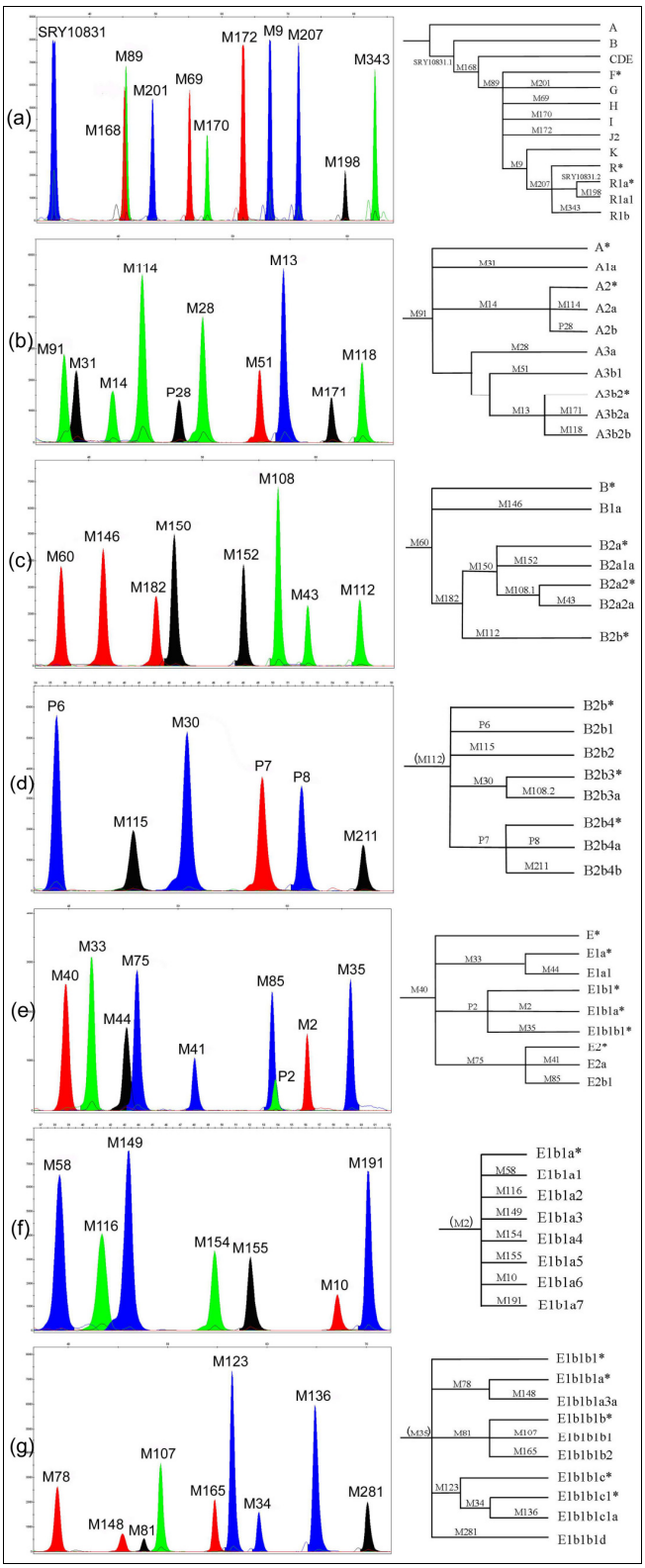

Figure 3.1: Electropherogram and phylogeny of (A) YSNP1, (B) Hg-A, (C) Hg-B, (D) Hg-B2b, (E) Hg-E, (F) Hg-E1b1a, and (G) Hg-E1b1b1

3.1.2 PCR Optimization

While PCR primer concentrations were initially 0.02μM – 0.04μM, these were increased or decreased incrementally, in order to obtain relatively equal amplification of amplicons in the multiplex PCR (table 2.2). The marker P28, in the Hg-A assay, initially experienced low amplification after multiplexing. This was rectified by increasing the final concentration of the P28 PCR primers to 0.2μM, and decreasing the buffer concentration to 0.8X. The annealing temperature was also optimized to ensure maximum product yield and to minimize formation of spurious products. A spurious amplification product was found to occur in the Hg-E1b1b1 assay; which was eliminated by increasing the annealing temperature to 61°C.

3.1.3 SBE Optimization

The SBE primers designed for the seven SBE assays, ranging from 25 to 80 bases in length, were designed to differ by four to five bases within each assay (Table 2.3). This did not always reflect in the electropherogram, with a lack of uniform separation in most of the assays. This resulted in a few extension products (e.g. M85 and P2 in Hg-E, M168 and M89 in Hg-YSNP1) co-migrating (Fig. 3.1). Fortunately this did not interfere with interpretation of results. The estimated lengths of extension products in the electropherogram (based on mobility) differed from the designed lengths, on average by four bases.

While the generation of aspecific peaks did occur occasionally, this was usually due to insufficient purification of the PCR products resulting in the incorporation of the PCR primers or dNTPs into the SBE reaction. The presence of one permanent aspecific peak did

occur, however, in the Hg-B2b assay (a red peak between the P8 and M211 peaks). This peak seemed to be linked to the P7 primer.

To intensify peak heights, the number of cycles in the SBE reaction program was increased from 25 to 35.

3.1.4 Validation of SBE assays

The seven SBE assays were validated using samples whose haplogroup status was previously determined. Additionally, sequencing was performed to confirm the presence of alleles for 15 mutations not screened for before the use of these SBE assays. These included M14, M114, M152, P6, P7, P8, M33, M44, M85, M58, M154, M34, M201, M198, and M343.

3.2 Population Affinity Analysis

The genetic relationships among the populations were assessed in the present study. Populations with less than 15 individuals were excluded from the analysis, which resulted in 25 population groups, including those from published comparative data. Haplogroup frequencies are listed in Table 3.1. Haplogroups that were not present in the sample were not included in the table, while some haplogroups were condensed into their ancestral branches for the population affinity analysis. Some of these haplogroups will be examined in more detail in later sections. The African-specific haplogroups A and B accounted for 5.7% and 10% of the total sample, respectively. While they were scattered across sub-Saharan Africa, both showed only a minor presence in the west African sample.

Haplogroup E comprised the majority of Y chromosomes in the total sample (~74%) and was widespread across sub-Saharan Africa. The Asian haplogroup C, notably, was found at low frequencies in East African Central Bantu speakers (0.4%) and East African Nilotic speakers (2.0%), aside from its presence in South African Coloured and South African Indian populations. With regard to the other non-African haplogroups, members of the super-haplogroup F, haplogroups F*, G, H, I and J (excluding super-haplogroup K), occurred at low frequencies in sub-Saharan Africa (0.1% - 1.4%), with most of these Y chromosomes found in the non-African populations groups (Cape Malay, European descent and Indian descent). Still, haplogroup I found its way into numerous southern African populations.

Similar to haplogroup F and its subclades, most Y chromosomes belonging to haplogroup K and its subclades, K*, L, O, P, R and T were found in the non-African population groups. The presence of haplogroup R-M343 at low to moderate frequencies in central African, west African (1.2% - 6.9%) and southern African (0.3% - 16.2%) populations was a notable exception.

The subclades of haplogroup A often displayed differing regional distributions. Haplogroup A-M13 was found mainly in east African populations at low frequencies (0.4% - 3%), though rising to 26% in East African Nilotic speakers. It was also found in Central African Ubangian speakers at 3.5%. Its sister-clade, haplogroup A-M51, however, was found only in southern African populations at frequencies ranging from 4.7% to 36.6%. The highest frequencies occurred in the Khoe-San populations (Central – 18.9%; Northern – 31.4%, Southern – 36.6%).

Table 3.1: Y chromosome haplogroup frequencies for sub-Saharan African populations

Population	Code	N	A M31	A P108	A M14	A M51	A M13	B M60	B M150	B M112	DE YAP	E M40	E M33	E P2	E M2	E M35	E M75	C M130	F M89	G M201	H M69	I M170	J P12F2	J M172	K M9	L M11	T M70	O M175	P M74	R M207	R M198	R M343
West African																																
Nigerian Mixed Benue-Congo speakers	NMBC	19													100.0																	
West African Mandinka	WAM	28											3.6		89.3	7.1																
West African Mixed Atlantic speakers	WAMA	143	4.9					0.7			0.7		7.7	2.1	71.3	10.5	0.7															1.4
Central African																																
Central African Central Bantu speakers	CACB	164						0.6	1.8	0.6					91.5	0.6	3.7															1.2
Northwest Bantu speakers	NWB	853		0.6					6.9	0.2		1.2	0.2		79.4	0.1	6.1			0.1												5.2
Central African Ubangian speakers	CAU	87		1.2	3.5		3.5	2.3	1.2	3.5					74.7	2.3	1.2															6.9
Western Pygmy	WPYG	102			2.9				4.9	49.0		5.9			35.3																	2.0
East African																																
East African Central Bantu speakers	EACB	241					0.4		2.1	2.1	5.8				80.5	6.6	1.2	0.4					0.4							0.4		
East African Cushitic speakers	EACU	24								25.0	29.2				4.2	41.7																
East African Nilotic speakers	EAN	150					26.0		16.7	9.3	4.0				12.7	18.0	10.7	2.0									0.7					
Hadzabe	HAD	54								53.7					27.8	18.5																
Sandawe	SAND	67					3.0		1.5	13.4	4.5				44.8	32.8																
Southern African																																
Southeastern Bantu speakers	SEB	343				4.7			12.0	0.3		0.3			69.4	1.2	10.8					0.3			0.3		0.3				0.3	0.3
Southern African Central Bantu speakers	SACB	71							4.2						87.3		8.5															
Southwestern African Central Bantu speakers	SWCB	39							2.6						84.6		10.3															2.6
Southwestern Bantu speakers	SWB	309							4.5	2.6		1.0			84.1	0.3	3.2					0.7									0.3	3.2
Dama	DAM	30			3.3	6.7				3.3					66.7		10.0					3.3										6.7
Admixed Khoe-San/Bantu-speakers	KBAD	21				4.8			52.4						42.9																	
Central Khoe-San	CKS	74			2.7	18.9				1.4					41.9	28.4	4.1															2.7
Northern Khoe-San	NKS	121			21.5	31.4				17.4					21.5	7.4	0.8															
Southern Khoe-San	SKS	41			2.4	36.6			2.4						22.0	26.8						2.4									4.9	2.4
South African Coloured	SAC	148				7.4			4.1	2.0					35.8	6.8	3.4	2.0		2.0	0.7	8.1		0.7		0.7		2.0	0.7	0.7	6.8	16.2
Cape Malay	CMAL	17																				11.8			11.8			35.3			29.4	11.8
European Descent	EUR	156											0.6		0.6	9.0			0.6	5.8		18.0	1.9	3.9			1.3				6.4	51.9
Indian Descent	IND	20																5.0	5.0		10.0			15.0	5.0	10.0				10.0	35.0	5.0
Total		3322	0.2	0.2	1.1	2.9	1.4	0.1	5.3	4.6	0.9	0.6	0.5	0.1	62.5	5.3	4.5	0.2	0.2	0.4	0.1	1.4	0.1	0.3	0.1	0.1	0.1	0.3	0.0	0.1	1.1	5.4

Haplogroup A-M14 was, again, found among the southern African Khoe-San speakers (upto to 21.5% in the Northern Khoe-San) and the Dama at 3.3%, while a low level presence in central Africa (2.9% - 3.5%) was also noted, though these were in different subclades. Haplogroups A-M31 and A-P108 appeared to be the rarest members of haplogroup A, of those subclades found in the sample. Haplogroup A-M31 was found only in the West African mixed Atlantic speakers at moderate frequency (4.9%), while A-P108 was found at low frequencies in central African Northwest Bantu speakers and Central African Ubangian speakers (0.6% and 1.2% respectively).

Haplogroup B, while twice as common as haplogroup A, was represented in sub-Saharan Africa by its two main subclades, B-M150, and B-M112. While neither of these showed much of a presence in west Africa, both were found throughout central, east and southern Africa. Despite both their substantial distributions, a clear difference was observed in the distributions of haplogroups B-M150 and B-M112 among population groups. While haplogroup B-M112 was found at its highest frequencies (Sandawe – 13.4%; Northern Khoe-San – 17.4%; Hadza – 53.7%) in Khoe-San speakers (east African and southern African), East African Cushitic speakers (25%) and Western Pygmy populations (49%), haplogroup B-M150 was found more commonly among Bantu speakers and other agriculturalist populations. An exception to this was its very high frequency (52.4%) in the admixed Khoe-San/Bantu speakers from Botswana.

As mentioned previously, haplogroup E was the most commonly occurring haplogroup in sub-Saharan Africa. Most of its distribution, however, was due to the near ubiquitous presence of its subclade, haplogroup E-M2, which reached a frequency of 62.5% in the total sample. Its highest frequencies were found in Bantu speaking populations and other

members of the Niger-Congo linguistic group, though appreciable levels were found in most other sub-Saharan African populations. The other haplogroup E subclades were found at much lower frequencies overall, with only haplogroups E-M35 and E-M75 reaching 5.3% and 4.5% respectively. Most E-M35 Y chromosomes were found in east African populations, and also reached appreciable levels in southern African Khoe-San populations. Haplogroup E-M75 reached moderate frequencies in central, east and southern Africa, with only a minor presence in west Africa. The highest frequencies were seen in Bantu speakers and Nilotic speakers.

Exact tests of population differentiation were used in combination with Fst genetic distances on haplogroup frequency data and Rst genetic distances on STR haplotype data (See appendix H for the matrices). In order to include the maximal number of samples and populations in the analysis, a seven-STR haplotype (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393) was used to generate the Rst distance matrix. The data from the two types of datasets (Fst and Rst) correlated well when compared using a Mantel test. The relationship between the Fst distances and Rst distances were confirmed ($p < 0.00001$) with a correlation coefficient (r) of 0.755 between the two matrices, while 57% of the Fst distance was explained by the Rst distance.

The two distance matrices were used to construct MDS plots (Figs. 3.2 and 3.3) and trees based on cluster analysis (Figs. 3.4 and 3.5)

In both the Fst and Rst MDS plots, the non-African populations (Cape Malay, European descent, Indian descent), separated from the African populations, though with greater separation in the Rst MDS plot. The Bantu speakers and their linguistically related

populations - Ubangian speakers, Nigerian mixed Benue-Congo speakers, Mandinka, and mixed Atlantic speakers - all clustered closely together, with the exception of the the Western Pygmy who were either Bantu or Ubangian speakers. Finally the Khoisan speakers and east African populations (non-Bantu speakers) clustered loosely in the centre of the plots, together with the South African Coloured population (as well as the aforementioned Western Pygmy). The clustering appeared to be stronger in the Rst MDS plot.

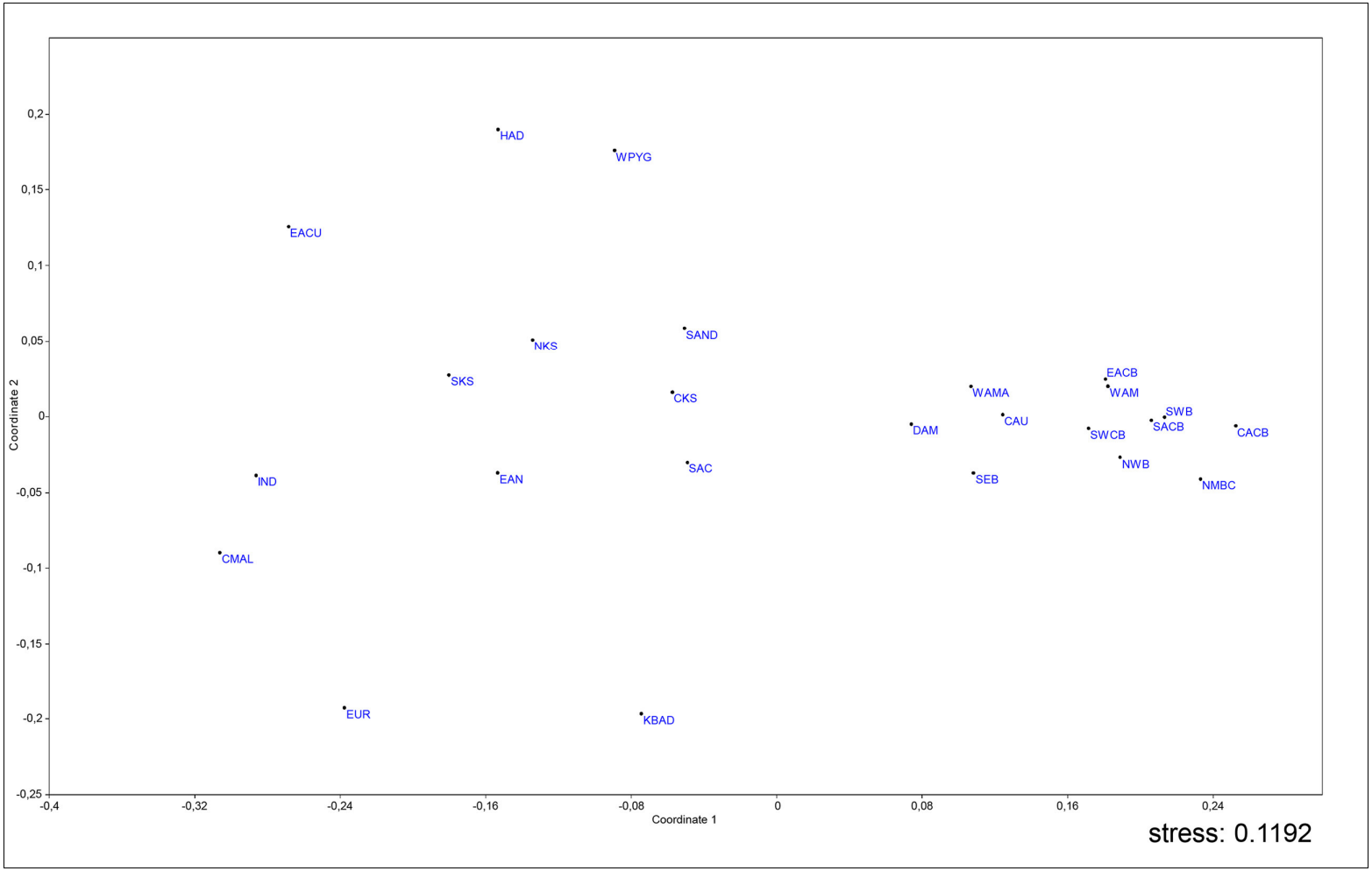

Figure 3.2: MDS plot of Fst distances between populations. See List of Abbreviations for a description of population codes.

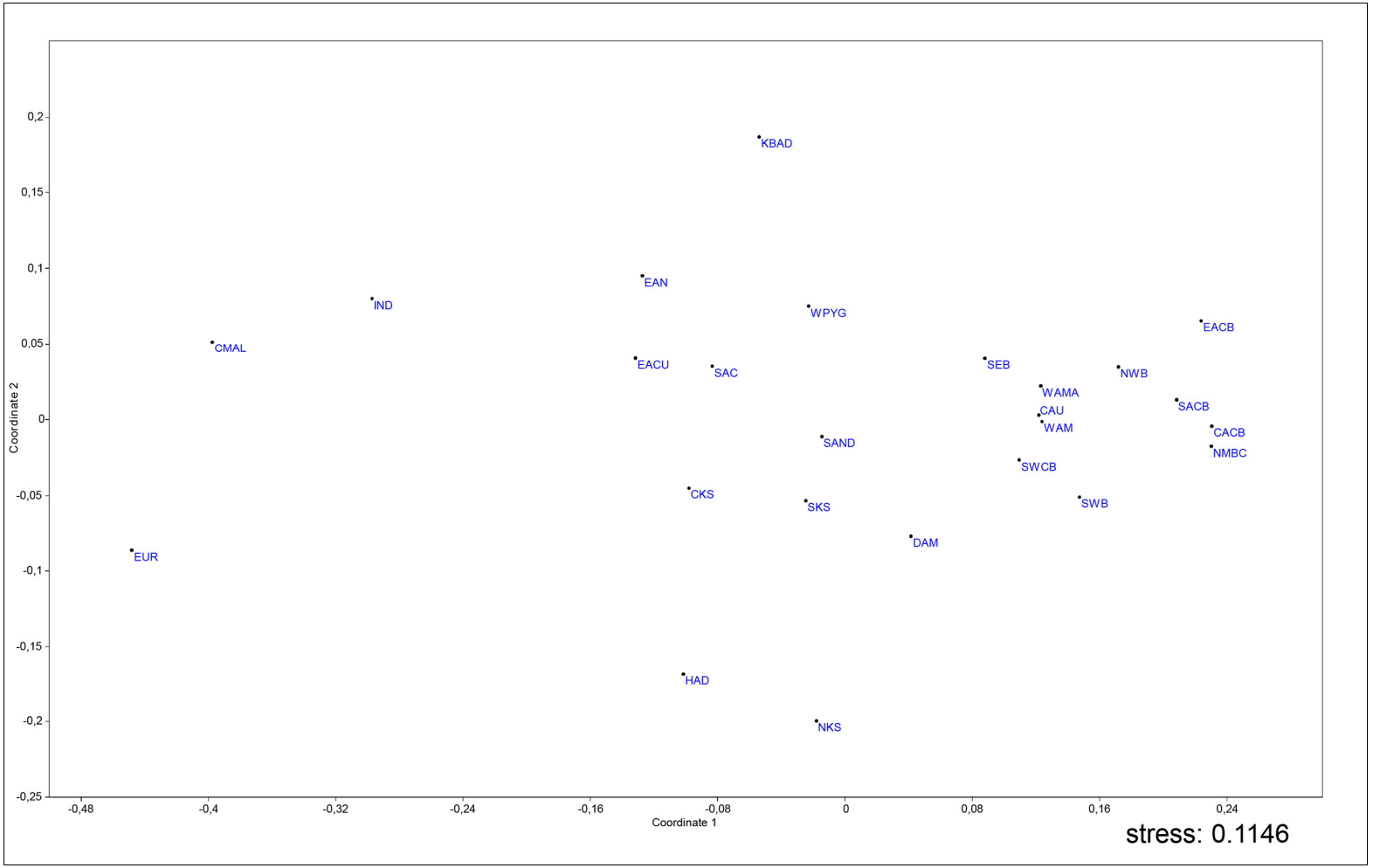

Figure 3.3: MDS plot representing Rst distances between populations. See List of Abbreviations for a description of population codes.

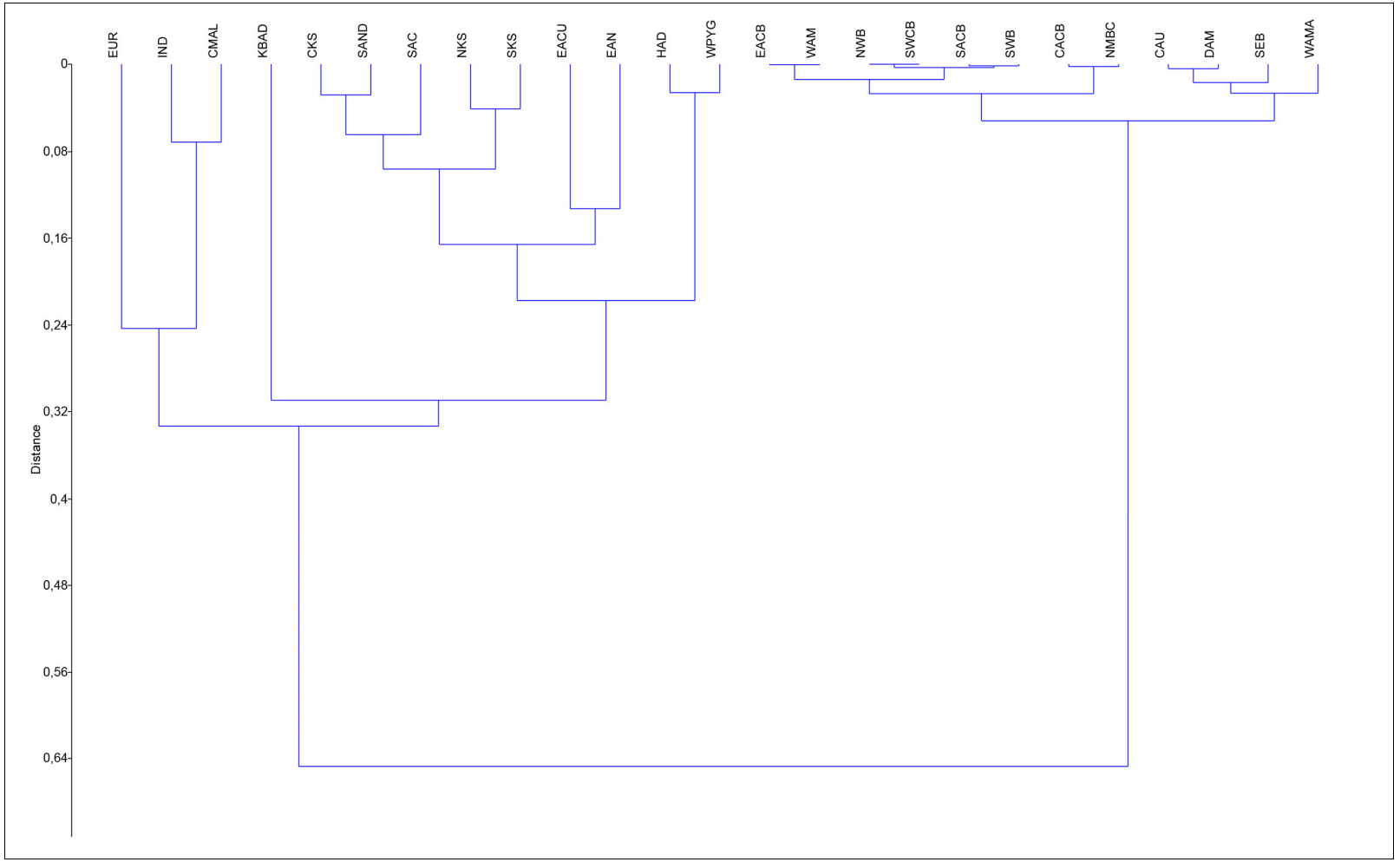

Figure 3.4: Cluster analysis tree representing Fst distances between populations. See List of Abbreviations for a description of population codes.

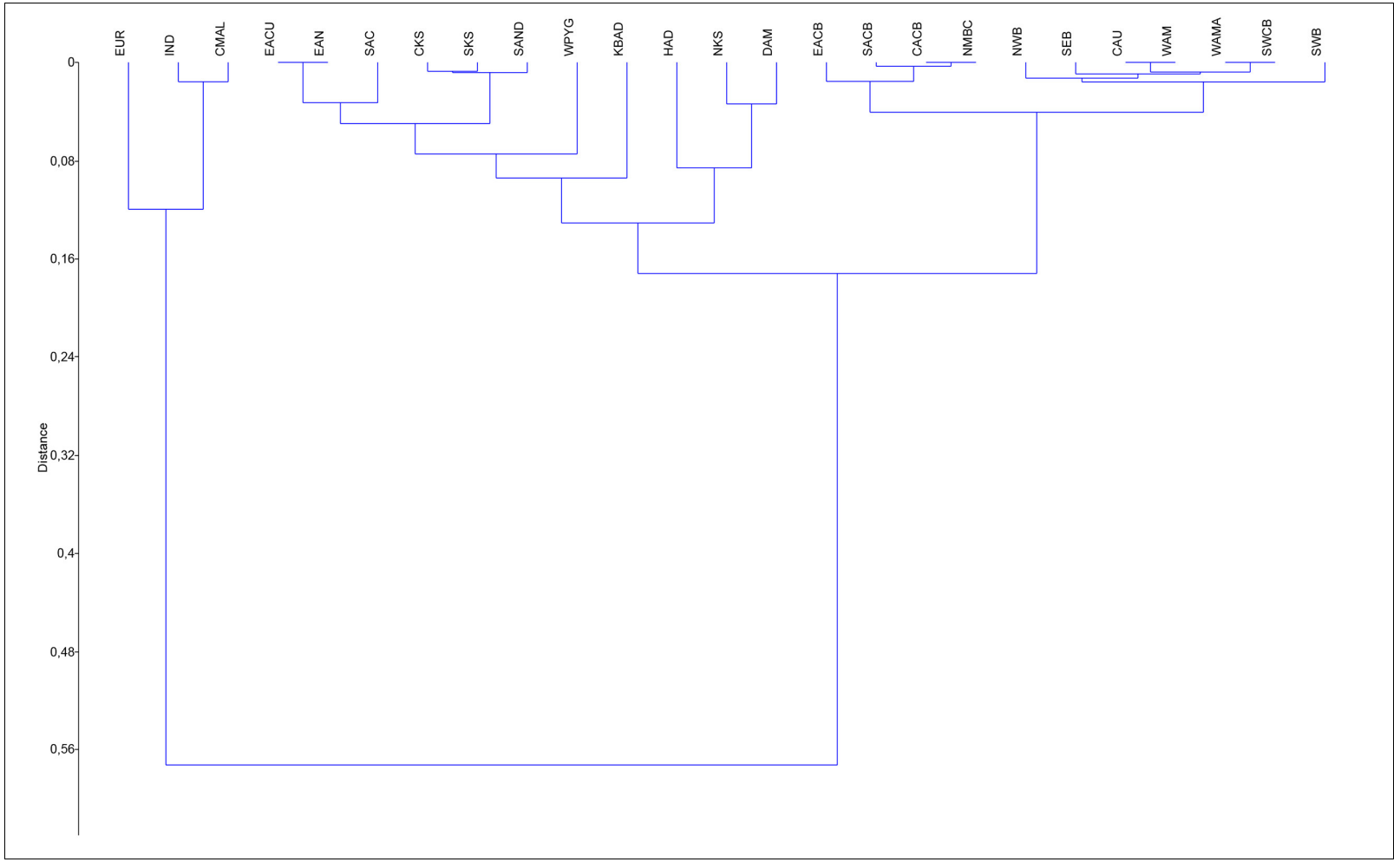

Figure 3.5: Cluster analysis tree representing Rst distances between populations. See List of Abbreviations for a description of population codes.

For the cluster analysis, while the Fst tree displayed two major clades, with the European, Indian and Cape Malay populations grouping with the non-Bantu speaking populations, the Rst tree separated these non-African populations out as a third major clade. As with the MDS plots, both Fst and Rst distances grouped the Bantu speakers and related populations into one clade. One of the key differences between the Fst and Rst trees was the placing of the Dama population. In the Fst tree, they were placed among the Bantu speaker clade. In the Rst tree, however, the Dama were placed in the non-Bantu speaker clade. The trees also displayed differences in structure within the non-Bantu-speaker clade. While the Nilotic and Cushitic populations grouped together in both the trees, the positions of the other populations within the clade were not consistent between the trees.

The exact tests of population differentiation based on haplogroup frequencies, indicated that most populations were significantly different from each other, with the exception of the Bantu speakers and their linguistically related populations (Ubangian speakers, Nigerian mixed Benue-Congo speakers, Mandinka, and mixed Atlantic speakers); and the Dama and the South African Coloured population ($p = 0.17141 \pm 0.0131$). Exact tests of population differentiation based on STR data, however, appeared to be more sensitive. Thus, even the Bantu speakers appeared to be mostly significantly different from each other (see Appendix H).

To test the apportionment of variation in the different population groups, AMOVA analysis was conducted to see how much variation was contained among defined groupings of populations (inter-group), then within these defined groupings (inter-population), and finally within the populations themselves (intra-population) (Table 3.2). The inter-group variation was highest when the groups reflected the major clades of the Rst clustering tree

(grouping A) in both the AMOVA based on haplogroup frequency data (Fst) (26.59%) and the AMOVA based on STR haplotype data (Rst) (15.80%). When groupings were based on the Fst clustering tree (Grouping B), the inter-group variation from the Fst AMOVA was lower (20.47%), and decreased to 11.46% in the Rst AMOVA. When grouping based on linguistic classification, values for inter-group variation were slightly higher than those based on the Fst clusering tree (Fst AMOVA – 22.74%; Rst AMOVA – 12.38%). In order to ascertain the effect of Bantu speakers on apportionment of variation, AMOVA was performed, excluding the Bantu-speaker group. This resulted in a drastic decrease in inter-group variation (Fst AMOVA – 7.99%; Rst AMOVA – 4.92%). Only grouping based on geographic region resulted in a similar drastic lowering of inter-group variation (Fst AMOVA – 4.95%; Rst AMOVA – 5.51%).

Table 3.2: Fst and Rst AMOVA analysis based on various groupings

#	Grouping	Based on	Fst (% of variation)			Rst (% of variation)		
			Among groups	Among populations within groups	Among individuals within populations	Among groups	Among populations within groups	Among individuals within populations
A	[CMAL, EUR, IND] [CKS, DAM, EACU, EAN, HAD, KBAD, NKS, SAC, SAND, SKS, WPYG] [CACB, CAU, EACB, NMBC, NWB, SACB, SEB, SWB, SWCB, WAM, WAMA]	Rst Clustering (Fig. 3.4)	26.59	6.17	67.24	15.80	4.63	79.58
B	[CKS, EACU, EAN, HAD, KBAD, NKS, SAC, SAND, SKS, WPYG] [CMAL, EUR, IND] [CACB, EACB, NMBC, NWB, SACB, SWB, SWCB, WAM] [CAU, DAM, SEB, WAMA]	Fst Clustering (Fig. 3.3)	20.47	6.78	72.75	11.46	5.13	83.40
C	[CMAL, EUR, IND] [CACB, CAU, NWB, WPYG] [NMBC, WAM, WAMA] [EACB, EACU, EAN, HAD, SAND] [CKS, DAM, KBAD, NKS, SAC, SACB, SEB, SKS, SWB, SWCB]	Geographic Region	4.95	18.54	76.51	5.51	9.24	85.25
D	[CMAL, EUR, IND, SAC] [CACB,EACB, NWB, SACB, SEB, SWB, SWCB] [CKS, DAM, HAD, KBAD, NKS, SAND, SKS] [CAU] [EACU] [EAN] [NMBC] [WAM] [WAMA] [WPYG]	Linguistic Classification	22.74	5.75	71.51	12.38	4.89	82.73
E	[CMAL, EUR, IND, SAC] [CKS, DAM, HAD, KBAD, NKS, SAND, SKS] [CAU] [EACU] [EAN] [NMBC] [WAM] [WAMA] [WPYG]	Linguistic Classification (Excl. Bantu-speakers)	7.99	14.01	78.00	4.92	11.00	84.08

3.3 Phylogeography

The previous section examined the relationships among some sub-Saharan African populations, on the basis of shared Y chromosome heritage. This section will focus in detail on the evolution and distribution of ancient Y chromosome haplogroups, more specifically, haplogroups A and B and the subclades they are comprised of; in order to gain insight into their origins and subsequent spread across Africa.

3.3.1 Haplogroup A

Following publication of Cruciani, et al. (2011), large-scale changes have been made to the Y chromosome phylogeny, and in particular, haplogroup A in 2012. The discovery of those mutations, and subsequent others (Scozzari, et al., 2012; Mendez, et al., 2013) have resulted in a refining of the basal branching structure of the phylogeny along with an updating of all haplogroup A names.

Haplogroup A is now a polyphyletic group, whose branches are strongly supported by the discovery of numerous SNPs (Cruciani, et al., 2011; Scozzari, et al., 2012; Mendez, et al., 2013). Its most basal member is haplogroup A00 (Mendez, et al., 2013); which split from the monophyletic clade, A0-T. Within A0-T, haplogroup A0 then separated out leaving A1-T; which is comprised of the clades, A1 and BT. Haplogroup A1 has been resolved further into haplogroups A1a (A-M31) and A1b (A-P108).

i. Haplogroup A-M14

Haplogroup A-M14 is a subclade of haplogroup A-P108. While not as rare as haplogroups A-P114 and A-M31, haplogroup A-M14 was found in only 60 individuals (Table 3.3). The majority of these (50) were found in southern Africa within Khoe-San or Khoisan speaking populations. Appreciable numbers were found in the Jul'hoansi from Namibia, the !Xun from Angola (and South Africa) and in an admixed Khoe-San/Bantu population, the |Gui+||Ghana+Kgalagari, from Botswana. Notably, it was also found to be present in small numbers in central Africa (10 individuals).

This presence of haplogroup A-M14 outside of southern Africa was made more surprising by the fact that these A-M14 chromosomes exhibited ancestral states at markers (M6, and M49) found to be derived in all southern African A-M14 chromosomes. This confirmed the existence of at least one independent subclade of haplogroup A-M14 outside of southern Africa. In central Africa, these A-M14* chromosomes (Fig. 3.6B) were found among the Baka of Cameroon and Gabon, and the Gbaya of CAR. The two Forro of São Tomé & Príncipe were possibly also members of A-M14*, though they were not screened for markers beyond M14. The southern African cohort of A-M14 consisted solely of M6-derived chromosomes. While these were also derived for M49, for convenience, M6 will be used to designate the haplogroup. Haplogroup A-M6 (Fig. 3.6C), however, was resolved into 3 subclades, A-M114, A-P28 and A-M6*. These subclades exhibited little population specificity within Khoe-San populations, however, they occurred most frequently in the Northern Khoe-San (the !Xun and Jul'hoansi). Haplogroup A-M114 was also found in a single Khomani individual, while haplogroup A-P28 singletons occurred in the Dama and the Nama. The 10 |Gui+||Ghana+Kgalagari individuals were not screened for M114 and P28, and so were not placed into subclades.

Table 3.3: Haplogroup A-M14 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg A2		Subclades of A2 (%)			
				n	%	A2-M14*	A2-M6*	A2-M114	A2-P28
B[2]	Angola	!Xun	64	5	7.8			3.1	
A	Angola	!Xun	80	12	15.0		1.3	5.0	8.8
E[3]	Botswana	ǀGui-ǁGhana-Kgalagari	65	10	15.4		15.4		
A	Botswana	Naro	2	1	50.0		50.0		
E	Cameroon	Baka	63	2	3.2	3.2			
A	Central African Republic	Gbaya	32	3	9.4	9.4			
E[4]	Gabon	Baka	33	3	9.1	9.1			
A	Namibia	Dama	29	1	3.4				3.4
A	Namibia	Jul'hoansi	41	14	34.1		7.3	9.8	17.1
A	Namibia	Nama	19	1	5.3				5.3
E	Namibia	San	5	1	20.0		20.0		
C	Namibia	Tsumkwe San	11	4	36.4		27.3		9.1
D[5]	São Tomé and Príncipe	Forro	68	2	2.9				
A	South Africa	Khomani	46	1	2.2			2.2	

[1] A = Present Study; B = Cruciani, et al. (2002); C = YCC (2002); D = Gonçalves, Spínola and Brehm (2007); E = Batini, et al. (2011a).

[2] Cruciani, et al. (2002) did not screen markers M6 and P28.

[3] Batini, et al. (2011a) did not screen for markers M114 and P28.

[4] Haplogroup A samples from Berniel-Lee, et al. (2009) were screened further in Batini, et al. (2011a).

[5] Chromosomes from Gonçalves, Spínola and Brehm (2007) were not sub-classified, due to the low number of markers screened.

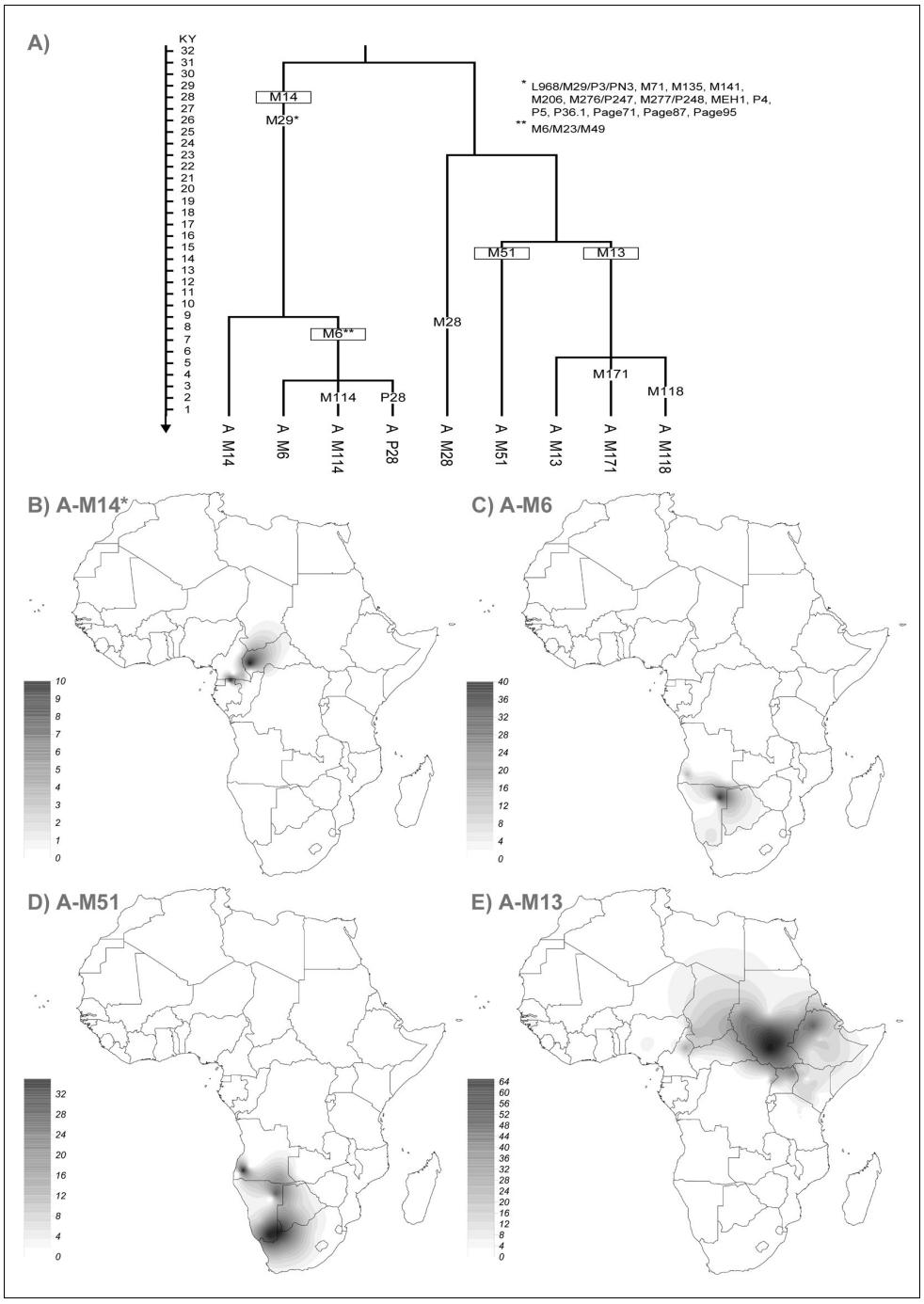

Figure 3.6: (A) Phylogeny of haplogroup A-L419 (A1b1) and its subclades, with TMRCA estimates indicated by the boxes surrounding the markers used in the BATWING analysis. Frequency distributions of (B) haplogroup A-M14*, (C) A-M6, (D) A-M51, and (E) A-M13.

To better examine the phylogenetic relationships among the A-M14 chromosomes in African populations, a RM-MJ network was constructed based on three SNP-15 STR haplotypes (Fig. 3.7). The network displayed two distinct branches of A-M14*, each containing one of two populations, Baka Pygmies, and Gbaya, respectively. The A-M6 branches showed that most A-M6* haplotypes were closely related; apart from one, which appeared to be related to the root of the A-M114 cluster. The most common subclade, haplogroup A-P28, appears to have undergone a demographic expansion, characterised by the star-like cluster of haplotypes. Upon incorporating |Gui+||Ghana+Kgalagari individuals into a low-resolution A-M14 network (data not shown); they clustered closely with the other Khoe-San groups.

Mean allelic variance and haplotype diversity indices were calculated for the major population groups, geographic regions, and for haplogroups A-M14, A-M14* and A-M6 (Table 3.4). Regionally, central Africa showed the highest levels of variance (0.455) yet the lowest haplotype diversity (0.607), yet the variance estimates within Central African Ubangian speakers and Western Pygmy groups, which comprised the central Africa sample, were extremely low (0.033 and 0.000 respectively). Southern Africa (allelic variance: 0.311; haplotype diversity: 0.879) and southwest Africa (allelic variance: 0.225; haplotype diversity: 0.886) both showed comparable levels of variance. The allelic variance of the paragroup A-M14* (0.455) was also higher than haplogroup A-M6 (0.286).

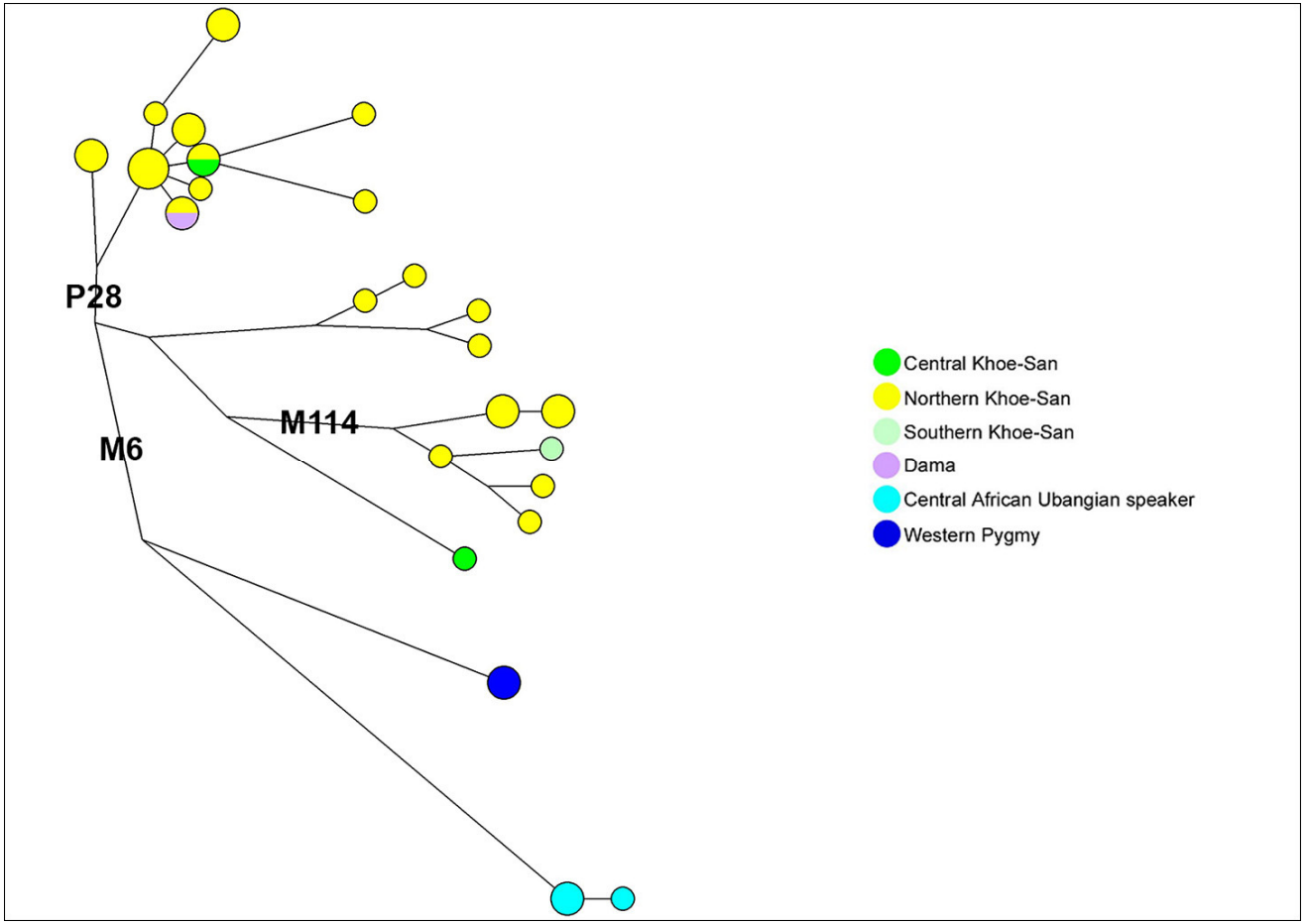

Figure 3.7: RM-MJ network of A-M14 based on a 3 SNP-15 STR haplotype (M6-M114-P28-DYS19-DYS389I-DYS389c-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439-DYS448-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 2 and Epsilon = 0.

With a TMRCA of 28.2 kya (CI 95%: 18.3 – 44.4 kya), the mutations defining haplogroup A-M14 was substantially older than those defining its subclade A-M6, which had a TMRCA of 7.5 kya (CI 95%: 4.7 – 12.4 kya) (Fig. 3.6A).

Table 3.4: Allelic variance and haplotype gene diversity estimates of haplogroup A-M14, based on population groups, regions and subclades.

Group	n	Allelic Variance (CI 95%)		Haplotype Gene Diversity (sd)	
Central African Ubangian	3	0.033	0.000 - 0.033	0.667	0.314
Central Khoe-San	2	0.3	0.000 - 0.300	1.000	0.500
Dama	1	-	-	-	-
Khoe-San/Bantu admixed	10	0.233	0.149 - 0.300	0.822	0.097
Northern Khoe-San	26	0.239	0.172 - 0.302	0.914	0.039
Southern Khoe-San	1	-	-	-	-
Western Pygmy	5	0.000	0.000 - 0.000	0.000	0.000
central Africa	8	0.455	0.213 - 0.536	0.607	0.164
southwest Africa	28	0.225	0.158 - 0.287	0.886	0.049
southern Africa	12	0.311	0.205 - 0.425	0.879	0.075
A-M14	48	0.369	0.291 - 0.453	0.944	0.019
A-M6	40	0.286	0.235 - 0.339	0.933	0.026
A-M14*	8	0.455	0.213 - 0.536	0.607	0.164

ii. Haplogroup A-M51

While haplogroup A-M51 was observed to be the dominant haplogroup A subclade in southern Africa, it was found in only 159 individuals (Table 3.5). Its distribution was also restricted, solely, to southern Africa (Fig. 3.6D), from the west in Angola to the east in Mozambique, and south into South Africa. Haplogroup A-M51 reached its highest frequencies in Khoe-San populations such as the !Xun, the Jul'hoansi, the Khomani of South Africa, and the Nama of Namibia (and South Africa); thus exhibiting a relatively even distribution across Northern, Central and Southern Khoe-San groups. It was also observed at low frequencies in the |Gui+||Ghana+Kgalagari, the South African Coloured population, the Dama from Namibia, and in a number of Southeastern Bantu speakers. Within these Bantu speaking groups, haplogroup A-M51 was found at varying frequencies, with the highest found in the Sotho population at 7.4%, and the lowest in Mozambique at 0.3%.

The phylogenetic relationships among the A-M51 chromosomes in southern African populations were examined using a RM-MJ network, which was constructed based on a 14 STR haplotype (Fig. 3.8). At this level of resolution, the high diversity within haplogroup A-M51 became apparent, not only overall but also within all of the populations represented. Only a few shared haplotypes were present, with the South African Coloureds displaying haplotypes in common with Southern Khoe-San, Northern Khoe-San and Southeastern Bantu speakers. Also, the major |Gui+||Ghana+Kgalagari haplotype was shared equally with Southeastern Bantu speakers.

Table 3.5: Haplogroup A-M51 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg A-M51	
				n	%
B	Angola	!Xun	64	18	28.1
A	Angola	!Xun	80	28	35.0
B	Angola	Khwe	26	3	11.5
A	Angola	Khwe	51	7	13.7
D	Botswana	|Gui-||Ghana-Kgalagari	65	4	6.2
A	Botswana	|Gui-||Ghana-Kgalagari	21	1	4.8
D	Botswana	mixed Bantu-speakers	15	1	6.7
D	Mozambique	mixed Central Bantu speakers	303	1	0.3
A	Namibia	Dama	29	2	6.9
A	Namibia	Ju\'hoansi	41	10	24.4
A	Namibia	Nama	19	6	31.6
D	Namibia	San	5	1	20.0
C	Namibia	Tsumkwe San	11	2	18.2
A	South Africa	Khomani	46	15	32.6
A	South Africa	mixed Southeastern Bantu speakers	63	2	3.2
A	South Africa	Nama	11	2	18.2
A	South Africa	Pedi	127	2	1.6
A	South Africa	Sotho	189	14	7.4
A	South Africa	South African Coloured	313	21	6.7
A	South Africa	Swazi	56	2	3.6
A	South Africa	Tsonga	136	1	0.7
A	South Africa	Tswana	187	7	3.7
D	South Africa	Xhosa	65	1	1.5
A	South Africa	Xhosa	175	3	1.7
A	South Africa	Zulu	402	5	1.2

[1] A = Present Study; B = Cruciani, et al. (2002); C = YCC (2002); D = Batini, et al. (2011a)

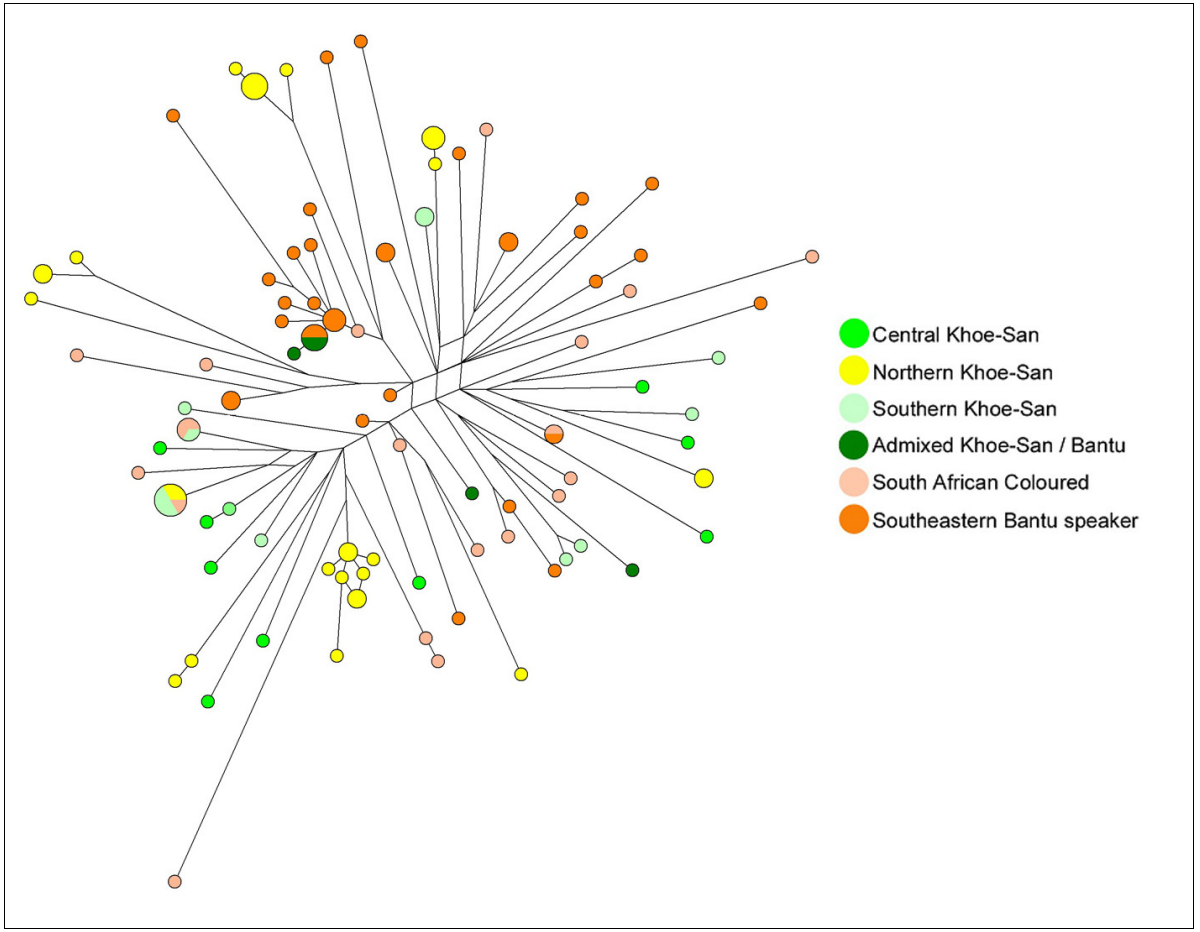

Figure 3.8: RM-MJ network of A-M51 based on a 14 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 2 and Epsilon = 0.

Few major clusters appeared, apart from ones specific to either Northern Khoe-San or Southeastern Bantu speakers (this cluster also contained the haplotype shared with the |Gui+||Ghana+Kgalagari).

Mean allelic variance and haplotype diversity indices were calculated for the major population groups, geographic regions, and for haplogroup A-M51 (Table 3.6). Most population groups exhibited similar variance and diversity estimates, apart from the |Gui+||Ghana+Kgalagari, which were lowest (allelic variance: 0.420; haplotype diversity: 0.700). Regionally, southern Africa (allelic variance: 0.636; haplotype diversity: 0.977) was similar to southwest Africa (allelic variance: 0.671; haplotype diversity: 0.953), with both only slightly lower than the overall estimate (allelic variance: 0.678; haplotype diversity: 0.983).

The TMRCA of haplogroup A-M51 was estimated at 14.5 kya (CI 95%: 10.4 – 21.4 kya) (Fig. 3.6A).

Table 3.6: Allelic variance and haplotype gene diversity estimates of haplogroup A-M51 based on population groups and regions

Group	n	Allelic Variance (CI 95%)		Haplotype Gene Diversity (sd)	
Central Khoe-San	9	0.636	0.305 - 0.903	1.000	0.052
Khoe-San/Bantu admixed	5	0.420	0.000 - 0.630	0.700	0.218
Northern Khoe-San	28	0.613	0.522 - 0.685	0.926	0.025
South African Coloured	21	0.610	0.393 - 0.922	0.967	0.026
South East African Central Bantu	1	-	-	-	-
Southeastern Bantu	38	0.581	0.409 - 0.777	0.963	0.021
Southern African Central Bantu	1	-	-	-	-
Southern Khoe-San	15	0.689	0.530 - 0.782	0.943	0.040
southeast Africa	1	-	-	-	-
southwest Africa	35	0.671	0.573 - 0.786	0.953	0.018
southern Africa	82	0.636	0.531 - 0.759	0.977	0.009
A-M51	118	0.678	0.599 - 0.776	0.983	0.005

iii. Haplogroup A-M13

Haplogroup A-M13 was found to be the most common subclade of haplogroup A, and occurred in 287 individuals (Table 3.7). While most common in east Africa (Fig. 3.6E), especially Ethiopia and Sudan, its distribution reached as far as Guinea-Bissau on the west coast, albeit in only one individual. In Sudan, haplogroup A-M13 was found at its highest frequencies in East African Nilotic speakers such as the Dinka, Shilluk, Borgu and Nuer, as well the Nuba, who speak both Eastern Sudanic and Kordofanian languages. In Ethiopia, the Ethiopian Jews were found to harbour the highest frequencies, with low to moderate frequencies in various other Afro-Asiatic speakers. In Uganda and Kenya, it was also found most often in Nilotic speakers such as the Karamajong and the Maasai. Outside of east Africa, haplogroup A-M13 numbers fell, with only 24 individuals found in central Africa, and 11 individuals found in west Africa. Notably, the Nilotic thread continued into central Africa, where A-M13 was found in the Alur of the DRC, and in linguistically related Nilo-Saharan populations such as the Kanuri from Cameroon. In west Africa, 10 individuals were found in Nigeria, while one was present in Guinea-Bissau. Haplogroup A-M13 has two known subclades, A-M171 and A-M118, which were screened for in the present study. Haplogroup A-M171 was found in two Ubangian speakers from CAR, while haplogroup A-M118 was found in a Ugandan Central Bantu speaker.

The phylogenetic relationships among the A-M13 chromosomes in African populations were examined using a RM-MJ network, which was constructed based on a 10 STR haplotype (Fig. 3.9). It appeared that while most East African Nilotic speakers were found to cluster closely with each other, the group still harboured a diverse array of A-M13 haplotypes. All Nigerian haplotypes, while distinct from each other, clustered closely, together with primarily central African groups. Afro-Asiatic groups i.e. the Chadic,

Cushitic, Semitic and Omotic groups were present throughout the network, with little clustering within each of these populations. A 15 STR network (data not shown) revealed a breaking up of the Nilotic cluster.

Mean allelic variance and haplotype diversity indices were calculated for the major population groups, geographic regions, and for haplogroup A-M13 (Table 3.8). Regionally, east Africa (allelic variance: 0.381; haplotype diversity: 0.973) and west Africa (allelic variance: 0.380; haplotype diversity: 0.978) showed similar levels of variance with central Africa (allelic variance: 0.283; haplotype diversity: 0.978) being slightly lower. Within east Africa the highest levels of variation was found in East African Cushitic speakers (allelic variance: 0.393; haplotype diversity: 0.933). The East African Nilotic speakers, while containing the most numbers of A-M13 individuals displayed an estimated variance of only 0.298.

Table 3.7: Haplogroup A-M13 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg A-M13 n	Hg A-M13 %
B	Cameroon	Fulbe	17	2	11.8
K	Cameroon	Kanuri	12	3	25.0
K	Cameroon	Mandara	30	3	10.0
F	Cameroon	Mandara	28	4	14.3
B	Cameroon	mixed Adamawa-speakers	18	1	5.6
B	Cameroon	mixed Chadic-speakers	15	1	6.7
B	Cameroon	mixed Nilo-Saharan-speakers	9	2	22.2
K	Cameroon	Tupuri	9	1	11.1
F	Cameroon	Tupuri	9	2	22.2
A	Central African Republic	Banda	1	1	100.0
A	Central African Republic	Lagba	3	1	33.3
A	Central African Republic	Nzakara	30	1	3.3
F	Democratic Republic of Congo	Alur	9	2	22.2
F	Democratic Republic of Congo	Hema	18	1	5.6
F	Democratic Republic of Congo	Mbuti	47	1	2.1
D	Egypt	Arabs	147	4	2.7
F	Egypt	Egyptian	92	3	3.3
K	Ethiopia	Amhara	49	5	10.2
C	Ethiopia	Amhara	48	7	14.6
F	Ethiopia	Amhara	18	3	16.7
K	Ethiopia	Dawro	78	6	7.7
B	Ethiopia	Ethiopian Jews	22	9	40.9
E	Ethiopia	mixed Ethiopians	242	41	16.9
F	Ethiopia	mixed Semitic-speakers	20	1	5.0
K	Ethiopia	Oromo	37	7	18.9
C	Ethiopia	Oromo	78	8	10.3
F	Ethiopia	Oromo	9	1	11.1

Table 3.7 cont.: Haplogroup A-M13 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg A-M13 n	Hg A-M13 %
G	Guinea-Bissau	Balanta	26	1	3.8
K	Kenya	Elmolo	23	3	13.0
F	Kenya	Kikuyu & Kamba	42	1	2.4
F	Kenya	Luo	9	1	11.1
K	Kenya	Maasai	81	5	6.2
F	Kenya	Maasai	26	7	26.9
D	Kenya	mixed Bantu-speakers	29	4	13.8
K	Kenya	Samburu	34	1	2.9
K	Kenya	Turkana	53	7	13.2
K	Nigeria	Idoma	47	1	2.1
K	Nigeria	Igala	40	2	5.0
K	Nigeria	Tiv	54	1	1.9
K	Nigeria	mixed Nigerians	-	6	-
I	Sudan	Beja	42	2	4.8
I	Sudan	Borgu	26	9	34.6
I	Sudan	Dinka	26	16	61.5
I	Sudan	Fur	32	10	31.3
I	Sudan	Gaalien	50	3	6.0
I	Sudan	Hausa	32	4	12.5
I	Sudan	Masalit	32	6	18.8
K	Sudan	mixed Sudanese	35	2	5.7
I	Sudan	Nuba	28	13	46.4
I	Sudan	Nuer	12	4	33.3
I	Sudan	Shilluk	15	8	53.3
H	Tanzania	Datog	35	1	2.9
H	Tanzania	Sandawe	68	3	4.4
D	Tanzania	Wairak (Iraqw)	43	3	7.0
F	Uganda	Ganda	26	2	7.7

Table 3.7 cont.: Haplogroup A-M13 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg A-M13 n	Hg A-M13 %
J	Uganda	Karamojong	118	39	33.1
A	Uganda	Kiga	77	1	1.3

[1] A = Present Study; B = Cruciani, et al. (2002); C = Semino, et al. (2002); D = Luis, et al. (2004); E = Moran, et al. (2004); F = Wood, et al. (2005); G = Rosa, et al. (2007); H = Tishkoff, et al. (2007); I = Hassan, et al. (2008); J = Gomes, et al. (2010); K = Batini, et al. (2011a)

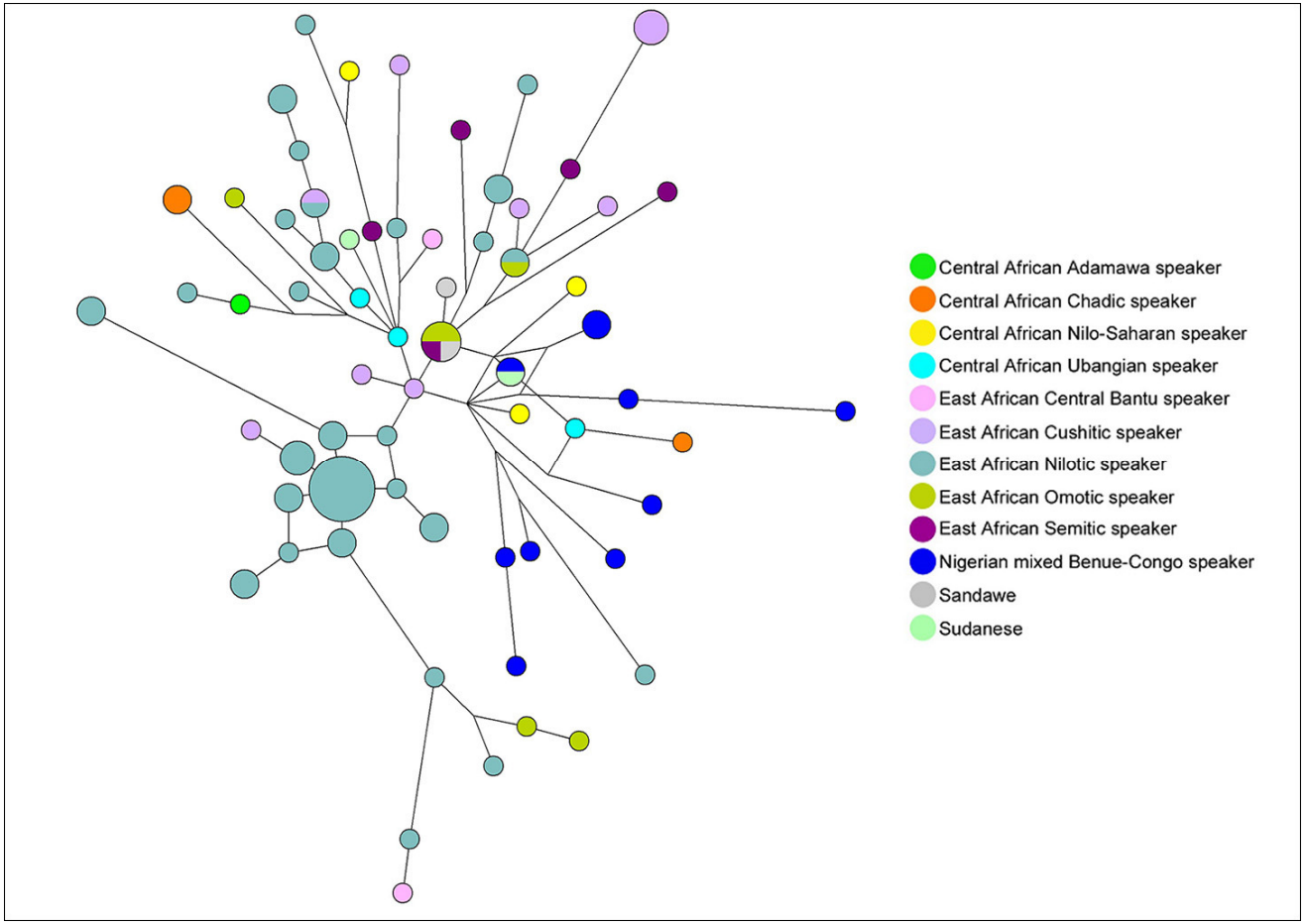

Figure 3.9: RM-MJ network of A-M13 based on a 10 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439) with reduction threshold = 2 and Epsilon = 0.

The TMRCA of haplogroup A-M13 was estimated at 14.5 kya (CI 95%: 10.4 – 21.4 kya) (Fig. 3.6A).

Table 3.8: Allelic variance and haplotype gene diversity estimates of haplogroup A-M13 based on population groups and regions

Group	n	Allelic Variance (CI 95%)		Haplotype Gene Diversity (sd)	
Central African Chadic	3	0.267	0.000 - 0.267	0.667	0.314
Central African Gur	1	-	-	-	-
Central African Nilo-Saharan	3	0.400	0.000 - 0.467	1.000	0.272
Central African Ubangian	3	0.167	0.000 - 0.267	1.000	0.272
East African Central Bantu	2	0.000	0.000 - 1.100	1.000	0.500
East African Cushitic	10	0.393	0.231 - 0.537	0.933	0.077
East African Nilotic	49	0.298	0.225 - 0.366	0.943	0.024
East African Omotic	6	0.147	0.067 - 0.197	0.933	0.122
East African Semitic	5	0.320	0.120 - 0.550	1.000	0.127
Nigerian Mixed Benue-Congo	10	0.382	0.231 - 0.566	0.978	0.054
Sandawe	2	0.050	0.000 - 0.050	1.000	0.500
Sudanese	2	0.250	0.000 - 0.250	1.000	0.500
central Africa	10	0.283	0.200 - 0.349	0.978	0.054
east Africa	76	0.381	0.329 - 0.434	0.973	0.010
west Africa	10	0.380	0.232 - 0.563	0.978	0.054
A-M13	96	0.409	0.365 - 0.455	0.982	0.007

3.3.2 Haplogroup B

Unlike haplogroup A, haplogroup B is monophyletic, with support for the branch strengthened by the number of mutations defining it (Underhill, et al., 2001; Karafet, et al., 2008; Cruciani, et al., 2011). That the majority of individuals within haplogroup B are contained in the subclades B-M150 and B-M112, was reflected strongly in the present study, with none of the other haplogroups, B-M236*, B-M146 and B-M182*, found. We, however, did find individuals in haplogroup B (derived for M60) that did not belong in any of its known subclades. Two were found among Ubangian speakers from CAR, and one in a Manyanga individual from DRC.

i. Haplogroup B-M150

Haplogroup B-M150 was found to be relatively common, and occurred in 528 individuals across sub-Saharan Africa in over 100 different populations (Table 3.9). Most haplogroup B-M150 individuals were found in southern Africa (Fig. 3.10B), with very high frequency among the ǀGui-ǁGhana-Kgalagari of Botswana. It was also quite common among Southeastern Bantu speakers across southern Africa. B-M150 was also found quite often in central Africa, with low to moderate frequencies reached in many populations across Cameroon and DRC, including the Pygmy populations (Biaka from CAR, Baka and Bakola from Cameroon, and Mbuti from DRC). In east Africa most B-M150 chromosomes were found in Uganda among the Karamajong. Only a few B-M150 chromosomes were found in west Africa, in Mali, Nigeria and Benin.

Haplogroup B-M152 was the most commonly observed subclade of B-M150. In fact, of the B-M150 chromosomes found in the present study, all were observed to be haplogroup B-M152, throughout sub-Saharan Africa. In the published literature, while only a few attempted to discriminate lineages within B-M150, it is possible that most of those found were indications of a presence of B-M152. Of those samples, found to be ancestral for the M152 mutation (or M109 in some cases), most belonged to B-M218* in the Karamajong of Uganda, together with a surprising absence of B-M152/M109. This population was also found to harbour B-G1, a newly discovered sister clade to B-M152.

Individuals bearing B-M150* chromosomes were found within the Mbuti in eastern DRC, the Tupuri of Cameroon, the Baka of CAR, the Luo and the Kikuyu and Kamba in Kenya, and the Dogon of Mali. Of these, however, only the Mbuti (Cruciani, et al., 2002) were tested for M218. Due to the strong presence of B-M218* in east Africa, among the Karamajong, it is possible that some or all east African M150* Y chromosomes may be derived for M218. The non-B-M218 subclades of B-M150 *viz.* B-M108.1 and its subclade B-M43 were only found in a few individuals in Ethiopia and Mali, respectively, as reported in Underhill, et al. (2000).

To examine the phylogenetic relationships among the B-M150 chromosomes in African populations, RM-MJ networks were constructed based on 15 STR haplotypes (Fig. 3.11) and 10 STR haplotypes (Fig. 3.12). The reduced 10 STR haplotype incorporated the lower resolution comparative data into the analyses. Due to the large number of populations containing B-M150 chromosomes, samples were split based on either country (15 STR) or broad geographic region (10 STR).

Table 3.9: Haplogroup B-M150 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg B-M150		Subclades of B-M150 (%)		
				n	%	B-M150*	B-M218*	B-M152
J[2]	Angola	Ngangela	11	1	9.1	9.1		
J	Angola	Nyaneka-Nkhumbi	75	3	4.0	4.0		
J	Angola	Ovimbundu	96	7	7.3	7.3		
L[2]	Benin	mixed Beninese	125	1	0.8	0.8		
A	Botswana	ǀGui-ǁGhana-Kgalagari	21	11	52.4			52.4
L	Botswana	ǀGui-ǁGhana-Kgalagari	65	35	53.8	53.8		
L	Botswana	mixed Southeastern Bantu speakers	15	4	26.7	26.7		
L	Cameroon	Baka	63	3	4.8	4.8		
G	Cameroon	Bakola	33	2	6.1			6.1
L	Cameroon	Bakola	30	2	6.7	6.7		
I[2]	Cameroon	Bakola	22	4	18.2	18.2		
L	Cameroon	Bassa	42	1	2.4	2.4		
B	Cameroon	Ewondo	29	3	10.3			10.3
L	Cameroon	Ewondo	26	3	11.5	11.5		
B	Cameroon	Fali	39	7	17.9			17.9
L	Cameroon	Fali	35	8	22.9	22.9		
L	Cameroon	Kanuri	12	1	8.3	8.3		
L	Cameroon	Mandara	30	1	3.3	3.3		
G	Cameroon	Mandara	28	1	3.6			3.6
B	Cameroon	mixed Adamawa-speakers	18	2	11.1		5.6	5.6
E[2]	Cameroon	mixed Bantu-speakers	14	1	7.1	7.1		
L	Cameroon	mixed Cameroonian	290	4	1.4	1.4		
B	Cameroon	mixed Chadic-speakers	15	1	6.7			6.7
L	Cameroon	Ngumba	31	6	19.4	19.4		
G	Cameroon	Ngumba	31	7	22.6			22.6
I	Cameroon	Ngumba	24	8	33.3	33.3		
B	Cameroon	Ouldeme	21	1	4.8			4.8

Table 3.9 cont.: Haplogroup B-M150 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg B-M150		Subclades of B-M150 (%)		
				n	%	B-M150*	B-M218*	B-M152
L	Cameroon	Ouldeme	10	2	20.0	20.0		
G	Cameroon	Ouldeme	13	4	30.8			30.8
L	Cameroon	Tupuri	9	1	11.1	11.1		
G	Cameroon	Tupuri	9	1	11.1	11.1		
G	Central African Republic	Baka	18	1	5.6	5.6		
G	Central African Republic	Biaka	31	1	3.2			3.2
A	Central African Republic	Biaka	24	1	4.2			4.2
L	Central African Republic	Biaka	21	1	4.8	4.8		
B	Central African Republic	Biaka	20	1	5.0			5.0
A	Central African Republic	Nzakara	30	1	3.3			3.3
L	Congo	Babinga	20	1	5.0	5.0		
L	Congo	Beti	36	1	2.8	2.8		
L	Congo	Teke	38	1	2.6	2.6		
A	Democratic Republic of Congo	Bamboma	2	1	50.0			50.0
A	Democratic Republic of Congo	Luba	22	1	4.5			4.5
B	Democratic Republic of Congo	Mbuti	12	1	8.3	8.3		
L	Democratic Republic of Congo	Mbuti	33	3	9.1	9.1		
G	Democratic Republic of Congo	Mbuti	47	5	10.6	10.6		
A	Democratic Republic of Congo	Teke	1	1	100.0			100.0
A	Democratic Republic of Congo	Yansi	6	1	16.7			16.7
A	Democratic Republic of Congo	Yombe	3	1	33.3			33.3
G	Egypt	Egyptian	92	2	2.2			2.2
C[2]	Ethiopia	Amhara	48	1	2.1	2.1		
L	Ethiopia	Dawro	78	1	1.3	1.3		
F[2]	Ethiopia	mixed Ethiopians	242	2	0.8	0.8		
I	Gabon	Akele	50	1	2.0	2.0		
I	Gabon	Benga	48	2	4.2	4.2		

Table 3.9 cont.: Haplogroup B-M150 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg B-M150		Subclades of B-M150 (%)		
				n	%	B-M150*	B-M218*	B-M152
I	Gabon	Duma	46	2	4.3	4.3		
I	Gabon	Eshira	42	6	14.3	14.3		
I	Gabon	Eviya	24	5	20.8	20.8		
I	Gabon	Fang	60	1	1.7	1.7		
I	Gabon	Galoa	47	1	2.1	2.1		
I	Gabon	Kota	53	2	3.8	3.8		
I	Gabon	Makina	43	6	14.0	14.0		
I	Gabon	Ndumu	36	1	2.8	2.8		
I	Gabon	Nzebi	57	4	7.0	7.0		
I	Gabon	Obamba	47	2	4.3	4.3		
I	Gabon	Punu	58	5	8.6	8.6		
I	Gabon	Shake	43	6	14.0	14.0		
I	Gabon	Tsogo	60	5	8.3	8.3		
G	Kenya	Kikuyu & Kamba	42	1	2.4	2.4		
G	Kenya	Luo	9	2	22.2	11.1		11.1
G	Kenya	Maasai	26	2	7.7			7.7
E	Kenya	mixed Bantu speakers	29	1	3.4	3.4		
L	Kenya	Samburu	34	1	2.9	2.9		
G	Mali	Dogon	55	4	7.3	7.3		
L	Mali	mixed Malian	54	1	1.9	1.9		
L	Mozambique	mixed Central Bantu speakers	303	29	9.6	9.6		
A	Namibia	Herero	45	1	2.2			2.2
A	Namibia	Himba	25	3	12.0			12.0
L	Nigeria	Igala	40	2	5.0	5.0		
L	Nigeria	Tiv	54	1	1.9	1.9		
E	Rwanda	Hutu	69	2	2.9	2.9		

Table 3.9 cont.: Haplogroup B-M150 frequencies in sub-Saharan African populations

Reference[1]	Country	Population	N	Hg B-M150		Subclades of B-M150 (%)		
				n	%	B-M150*	B-M218*	B-M152
E	Rwanda	Tutsi	94	1	1.1	1.1		
A	South Africa	Khomani	46	1	2.2			2.2
A	South Africa	mixed Southeastern Bantu speakers	63	5	7.9			7.9
A	South Africa	Pedi	127	29	22.8			22.8
A	South Africa	Sotho	189	20	10.6			10.6
A	South Africa	South African Coloured	313	16	5.1			5.1
A	South Africa	Southern Ndebele	36	2	5.6			5.6
A	South Africa	Swazi	56	5	8.9			8.9
A	South Africa	Tsonga	136	13	9.6			9.6
A	South Africa	Tswana	187	53	28.3			28.3
A	South Africa	Venda	110	23	20.9			20.9
L	South Africa	Xhosa	65	3	4.6	4.6		
A	South Africa	Xhosa	175	11	6.3			6.3
A	South Africa	Zulu	402	32	8.0			8.0
H[2]	Tanzania	Mbugwe	15	1	6.7	6.7		
H	Tanzania	Sandawe	68	1	1.5	1.5		
D[2]	Tanzania	Sukuma	32	3	9.4	9.4		
E	Tanzania	Wairak (Iraqw)	43	1	2.3	2.3		
G	Tanzania	Wairak (Iraqw)	9	1	11.1			11.1
K	Uganda	Karamojong	118	25	21.2		16.9	
A	Uganda	Nyankole	40	1	2.5			2.5
A	Zambia	Bemba	17	1	5.9			5.9
A	Zambia	Kaonde	3	1	33.3			33.3
A	Zambia	Lozi	27	3	11.1			11.1
A	Zambia	Nyanja	25	1	4.0			4.0
A	Zimbabwe	mixed Southeastern Bantu speakers	75	6	8.0			8.0
G	Zimbabwe	Shona	49	5	10.2			10.2

[1] A = Present Study ; B = Cruciani, et al. (2002) ; C = Semino, et al. (2002); D = Knight, et al. (2003); E = Luis, et al. (2004); F = Moran, et al. (2004); G = Wood, et al. (2005); H = Tishkoff, et al. (2007); I = Berniell-Lee, et al. (2009); J = Coelho, et al. (2009); K = Gomes, et al. (2010); L = Batini, et al. (2011a)

[2] Chromosomes from Batini, et al. (2011a), Berniell-Lee, et al. (2009), Coehlo, et al. (2009), Knight, et al. (2003), Luis, et al. (2004) Moran, et al. (2004), Semino, et al. (2002) and Tishkoff, et al. (2007) were only screened for M150.

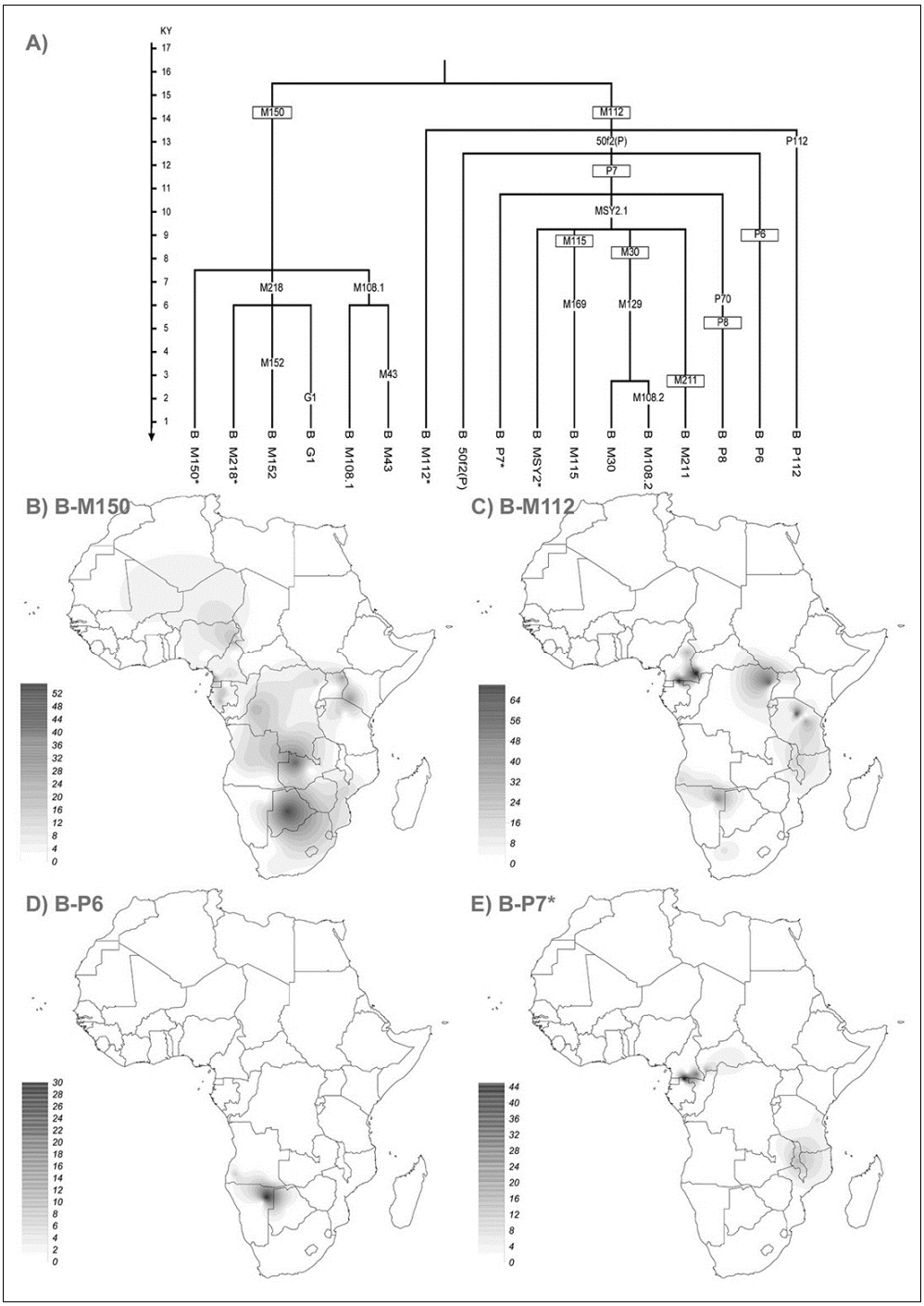

Figure 3.10: (A) Phylogeny of haplogroup B-M182 (B2) and its subclades, with TMRCA estimates indicated by the boxes surrounding the markers used in the BATWING analysis. Frequency distributions of (B) B-M150 (C) B-M112, (D) B-P6, and (E) B-P7*.

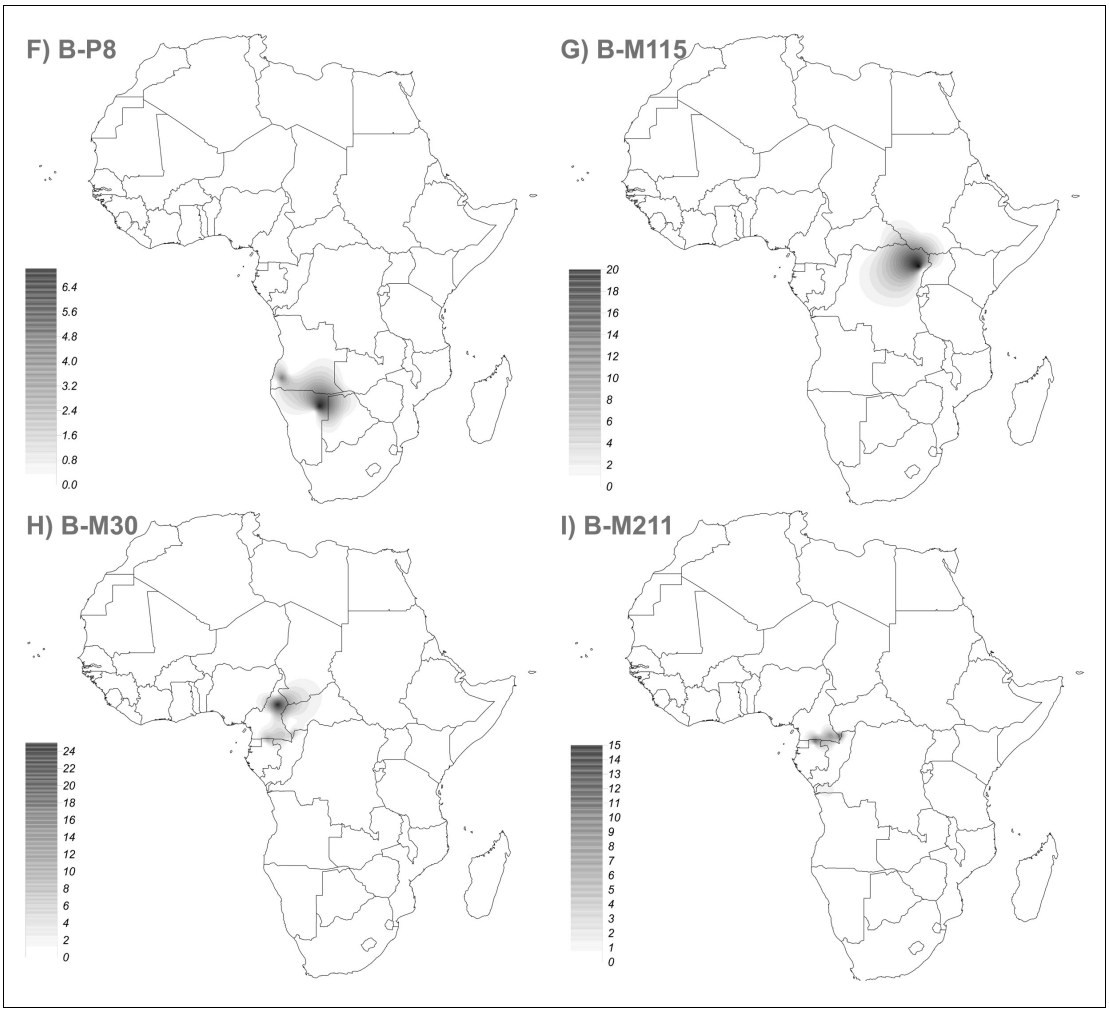

Figure 3.10 cont.: Frequency distributions of (F) B-P8 (G) B-M115, (H) B-M30, and (I) B-M211.

It was clear that even at the higher level of resolution most South African haplotypes were closely related to each other, with a number of these haplotypes occurring at high frequency. Botswanan B-M150 haplotypes, while closely related to South African haplotypes, were unique to Botswana, with most individuals found within a few haplotypes. Shared haplotypes also occurred between South African individuals and those in Zambia and Zimbabwe. Countries such as Uganda and Cameroon contained a number of haplotypes displaying high diversity. When the level of resolution was lowered, which allowed an increase of samples from central Africa; the levels of expansion within southern Africa were reinforced. Notably, southwest African samples were found primarily in one cluster, which also displayed shared haplotypes with southern, central and east Africans. Even at the lower resolution, the high diversities of haplotypes in central and east Africa were apparent. While haplotype sharing did occur between southern, central and east Africa, it was quite minimal.

Mean allelic variance and haplotype diversity indices were calculated for the major population groups, geographic regions for haplogroup B-M150 (Table 3.10).
While west Africa appeared to exhibit the highest levels of variance (allelic variance: 0.433; haplotype diversity: 1.000), this was based on only 3 individuals, and so is unlikely to be significant. Central Africa (allelic variance: 0.348; haplotype diversity: 0.980) and east Africa (allelic variance: 0.316; haplotype diversity: 0.949) followed, while southern Africa (allelic variance: 0.171; haplotype diversity: 0.863), and southwest Africa (allelic variance: 0.162; haplotype diversity: 0.936) showed relatively low levels of variance within their samples.

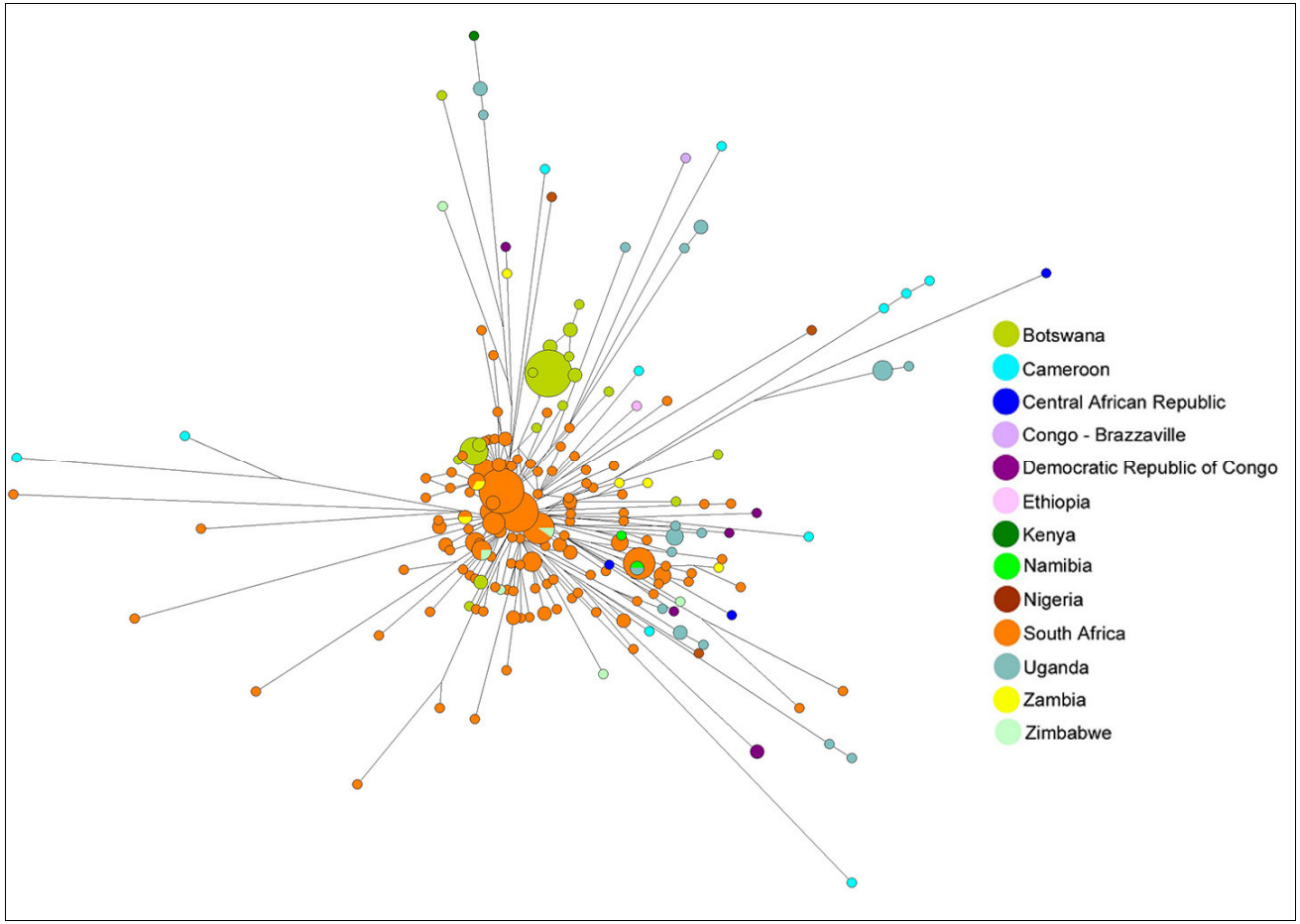

Figure 3.11: RM-MJ network of B-M150 based on a 15 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439-DYS448-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 1 and Epsilon = 0.

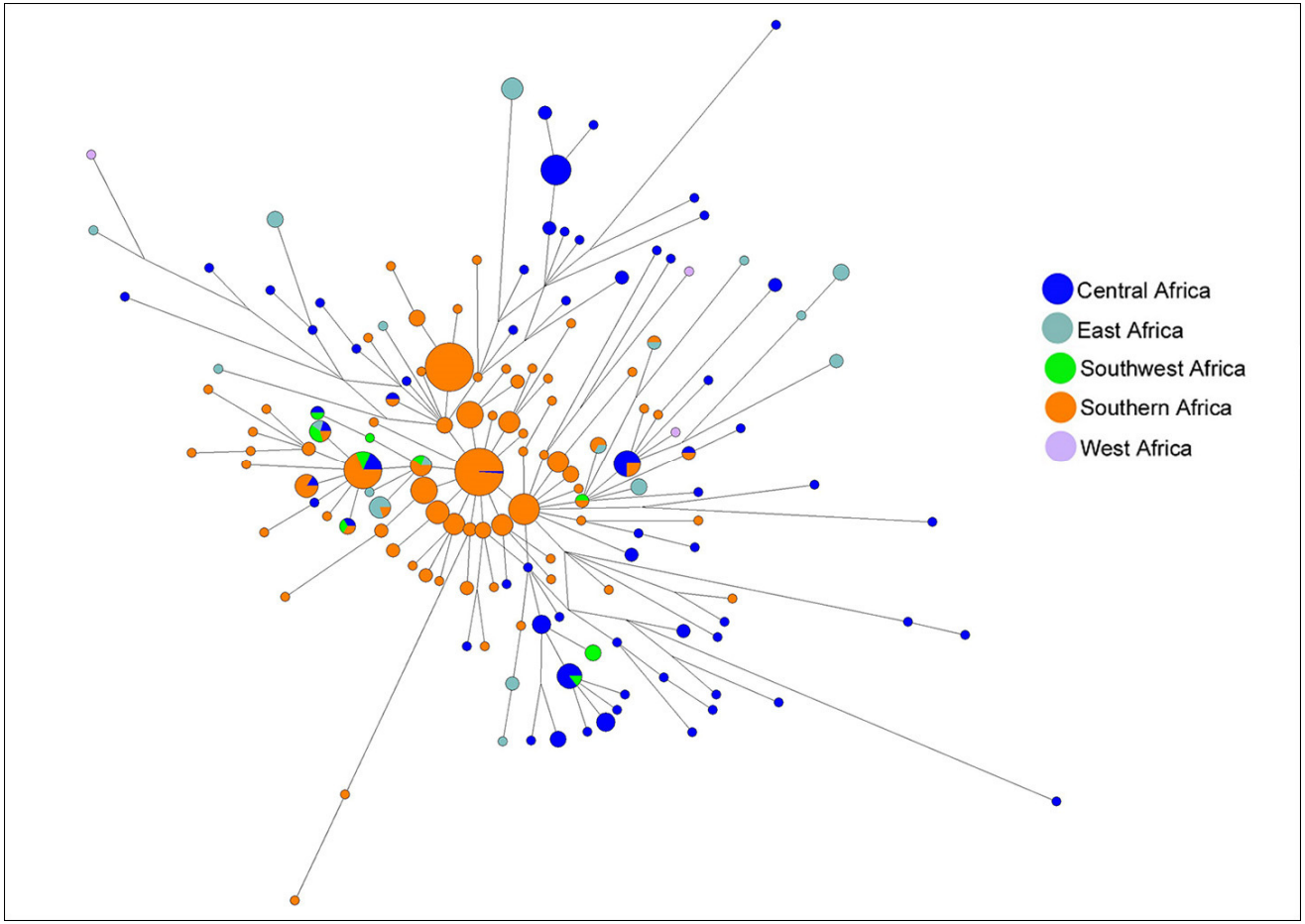

Figure 3.12: RM-MJ network of B-M150 based on a 10 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS439) with reduction threshold = 1 and Epsilon = 0.

Population groups within the regions usually appeared to have lower levels of variance than the regions they were found in. Exceptions to this included the Western Pygmy group who had the highest variance overall (allelic variance: 0.505; haplotype diversity: 0.985) and the South African Coloured populations (allelic variance: 0.417; haplotype diversity: 0.942). While the |Gui-||Ghana-Kgalagari of Botswana showed highest frequencies of B-M150, they also exhibited one of the lowest levels of diversity (allelic variance: 0.061; haplotype diversity: 0.588).

The TMRCA of haplogroup B-M150 was estimated at 14.2 kya (CI 95%: 11.3 – 17.9 kya) (Fig. 3.10A).

Table 3.10: Allelic variance and haplotype gene diversity estimates of haplogroup B-M150 based on population groups and regions

Group	n	Allelic Variance (CI 95%)		Haplotype Gene Diversity (sd)	
Central African Adamawa	8	0.064	0.000 - 0.121	0.464	0.200
Central African Central Bantu	3	0.133	0.000 - 0.267	1.000	0.272
Central African Chadic	3	0.200	0.000 - 0.200	0.667	0.314
Central African Gur	1	-	-	-	-
Central African Nilo-Saharan	1	-	-	-	-
Central African Ubangian	1	-	-	-	-
East African Central Bantu	5	0.270	0.020 - 0.430	0.900	0.161
East African Nilotic	26	0.274	0.224 - 0.319	0.920	0.027
East African Omotic	1	-	-	-	-
Eastern Pygmy	2	0.000	0.000 - 0.000	0.000	0.000
Khoe-San/Bantu admixed	46	0.061	0.041 - 0.084	0.588	0.071
Nigerian mixed Benue-Congo	3	0.433	0.000 - 0.700	1.000	0.272
Northwest Bantu	69	0.263	0.213 - 0.325	0.974	0.009
Sandawe	1	-	-	-	-
South African Coloured	16	0.417	0.137 - 0.821	0.942	0.041
South West African Central Bantu	1	-	-	-	-
Southeastern Bantu	206	0.151	0.110 - 0.210	0.832	0.026
Southern African Central Bantu	7	0.214	0.086 - 0.329	1.000	0.076
Southern Khoe-San	1	-	-	-	-
Southwestern Bantu	12	0.161	0.107 - 0.181	0.924	0.058
Western Pygmy	12	0.505	0.250 - 0.689	0.985	0.040
central Africa	100	0.348	0.282 - 0.434	0.980	0.006
east Africa	33	0.316	0.265 - 0.368	0.949	0.019
southwest Africa	13	0.162	0.106 - 0.182	0.936	0.051
southern Africa	276	0.171	0.133 - 0.224	0.863	0.019
west Africa	3	0.433	0.000 - 0.700	1.000	0.272
B-M150	425	0.268	0.235 - 0.307	0.939	0.009

ii. Haplogroup B-M112

While haplogroup B-M112 is quite widespread and has been observed in central, east and southern Africa (Fig. 3.10C, Table 3.11), the majority, around two thirds of all B-M112 chromosomes, have been found in groups that recently were, or still are, hunter-gatherer populations (HGPs). These include the Baka and Bakola from Cameroon and Gabon, the Biaka and Mbenzele from CAR, the Mbuti from DRC, the !Xun from Angola (and South Africa), the Ju|'hoansi from Namibia, as well as the Hadza and Sandawe from Tanzania. Many of these populations have high frequencies of B-M112 and its subclades (Table 3.11). Many of the pastoral or agricultural groups (food-producing populations = FPPs) in whom B-M112 occurs are often from the same countries as their hunter-gatherer neighbours.

The subclades of haplogroup B-M112 exhibited substantial geographic structure and population specificity (Table 3.11 and Fig. 3.10). Haplogroups B-P6 and B-P8 (Fig. 3.10D and F) occurred mostly in the !Xun and Ju|'hoansi of Angola and Namibia, respectively. Haplogroup B-P6 chromosomes were also found among the South African Coloured population; as well the Namibian Herero and the |Gui-||Ghana-Kgalagari from Botswana; though these were rare occurrences. A solitary B-P8 chromosome was found in the Angolan Khwe. Haplogroup B-M115 (Fig. 3.10G) was extremely specific, having been found only in the Mbuti of northeastern DRC.

Table 3.11: Haplogroup frequencies of B-M112 and its subclades in the populations studied

Reference[1]	Country	Population	N	Hg B-M112		Subclades of B-M112 (%)							
				n	%	M112*	P6	M115	M30*	M108.2	P7*	P8	M211
A	Angola	!Xun	80	9	11.3	1.3	6.3					3.8	
B	Angola	!Xun	64	5	7.8	7.8							
A	Angola	Khwe	51	1	2.0							2.0	
H[4]	Angola	Kuvale	26	3	11.5	11.5							
H	Angola	Nyaneka-Nkhumbi	75	3	4.0	4.0							
K	Botswana	Khoe-San/Bantu admixed	65	1	1.5		1.5						
K	Cameroon	Baka	63	23	36.5				6.3		23.8 [2]		6.3
K	Cameroon	Bakola	30	3	10.0	6.7			3.3				
E[5]	Cameroon	Bakola	33	1	3.0	3.0							
B[6]	Cameroon	mixed Nilo-Saharan	9	2	22.2				22.2				
E	Central African Republic	Baka	18	12	66.7						66.7		
A	Central African Republic	Biaka	24	16	66.7	4.2			29.2		12.5		20.8
B	Central African Republic	Biaka	20	6	30.0	5.0			5.0				20.0
K	Central African Republic	Biaka	21	4	19.0	4.8			4.8				9.5
E	Central African Republic	Biaka	31	15	48.4	3.2					45.2		
B	Central African Republic	Lissongo	4	1	25.0					25.0			
A	Central African Republic	Mbenzele	18	6	33.3				5.6		27.8		
K	Central African Republic	Mbenzele	42	5	11.9				2.4		9.5 [2]		
A	Central African Republic	Nzakara	30	1	3.3						3.3		
A	Central African Republic	Sangha-Sangha	11	2	18.2	18.2							
A	Democratic Republic of Congo	Manyanga	93	1	1.1								1.1
B	Democratic Republic of Congo	Mbuti	12	3	25.0	16.7		8.3					
E	Democratic Republic of Congo	Mbuti	47	20	42.6	21.3					21.3		
K	Democratic Republic of Congo	Mbuti	33	17	51.5	27.3		24.2					
G[3]	Gabon	Baka	33	21	63.6				9.1		45.5 [2]		9.1
G	Gabon	Duma	46	1	2.2						2.2		
G	Gabon	Shake	43	1	2.3						2.3		
K	Kenya	Maasai	81	5	6.2	4.9					1.2		
A	Malawi	mixed Central Bantu speakers	8	1	12.5						12.5		
K	Mozambique	mixed Central Bantu speakers	303	14	4.6	2.3					2.3		
A	Namibia	Jul'hoansi	41	12	29.3		24.4					4.9	
A	Namibia	Dama	29	1	3.4	3.4							
A	Namibia	Herero	45	2	4.4	2.2	2.2						

Table 3.11 cont.: Haplogroup frequencies of B-M112 and its subclades in the populations studied.

Reference[1]	Country	Population	N	Hg B-M112		Subclades of B-M112 (%)							
				n	%	M112*	P6	M115	M30*	M108.2	P7*	P8	M211
K	Namibia	San	5	3	60.0		40.0					20.0	
D[5]	Rwanda	Hutu	69	1	1.4	1.4							
D	Rwanda	Tutsi	94	13	13.8	13.8							
A	South Africa	mixed Southeastern Bantu speakers	63	2	3.2	3.2							
A	South Africa	Tswana	187	1	0.5	0.5							
A	South Africa	Tsonga	136	2	1.5	1.5							
A	South Africa	Pedi	127	1	0.8	0.8							
A	South Africa	Nama	11	1	9.1	9.1							
A	South Africa	South African Coloured	313	6	1.9		1.3				0.6		
F	Tanzania	Burunge	24	6	25.0	25.0							
F	Tanzania	Datog	35	1	2.9	2.9							
F	Tanzania	Hadza	57	29	50.9	50.9							
C[5]	Tanzania	Hadza	23	12	52.2	52.2							
A	Tanzania	mixed Tanzanians	34	4	11.8	5.9					5.9		
F	Tanzania	Sandawe	68	9	13.2	13.2							
F	Tanzania	Sukuma	30	2	6.7	6.7							
C	Tanzania	Sukuma	32	2	6.3	6.3							
F	Tanzania	Turu	20	1	5.0	5.0							
E	Tanzania	Wairak (Iraqw)	9	1	11.1						11.1		
D	Tanzania	Wairak (Iraqw)	43	2	4.7	4.7							
J[5]	Uganda	Karamojong	118	13	11.0	11.0							
I[5]	Zambia	Bisa	33	1	3.0	3.0							
A	Zimbabwe	mixed Southeastern Bantu speakers	75	1	1.3	1.3							

[1] A = Present Study; B = Cruciani, et al. (2002); C = Knight, et al. (2003); D = Luis, et al. (2004); E = Wood, et al. (2005); F = Tishkoff, et al. (2007); G = Berniel-Lee, et al. (2009); H = Coelho, et al. (2009); I = de Filippo, et al. (2010); J = Gomes, et al. (2010); K = Batini, et al. (2011a)

[2] Those chromosomes found by Batini, et al. (2011a) to be in MSY* have been included under B-P7*.

[3] While Berniell-Lee, et al. (2009) only screened for 50f2, most chromosomes were resolved further by Batini, et al. (2011a).

[4] While Coehlo, et al. (2009) only screened for M112, all chromosomes were resolved further by Batini, et al. (2011a).

[5] Chromosomes from De Filippo, et al. (2010), Gomes, et al. (2010), Knight, et al. (2003), Luis, et al. (2004), and Wood, et al. (2005) were not fully sub-classified due to the low number of markers screened, and thus were placed mainly in B-M112* - and B-P7* in the case of Wood, et al. (2005).

[6] Cruciani, et al. (2002) did not screen markers P6 and P7, so its B-M112* values are possibly inflated.

Haplogroups B-M30 and B-M211 (Fig. 3.10H and I) were commonly found among Western Pygmy populations from Cameroon, CAR and Gabon. Additionally, haplogroup B-M30 was found in two Nilo-Saharan speakers in Cameroon, and in a single Lissongo individual from CAR. Haplogroup B-M211 was also present in a Manyanga individual from southwestern DRC. The paragroups of B-M112* and B-P7* showed wider geographic distribution, and were found in populations across central, east and southern Africa.

To better examine the phylogenetic relationships among the B-M112 chromosomes in African populations, RM-MJ networks were constructed based on six SNP – 14 STR haplotypes (Fig. 3.13) and nine STR haplotypes (Fig. 3.14) with the former containing only the highest resolved chromosomes (the reduced nine STR haplotype incorporated the lower resolution comparative data into the analyses). At the higher resolution, the high diversity and population structure of B-M112 was apparent, with low levels of haplotype sharing.

A lowering of resolution to the nine-STR level resulted in slightly increased haplotype sharing. The Hadza shared a few haplotypes with a Tanzanian Datog individual and a Tanzanian Sukuma individual observed to be in haplogroup B-P7*. In addition, the Western Pygmy (from Cameroon and Gabon) shared a haplotype with a Manyanga individual from DRC, while two Hadza shared a haplotype with a !Xun individual. The final shared haplotype was between an Mbuti individual and a Southeastern Bantu speaker.

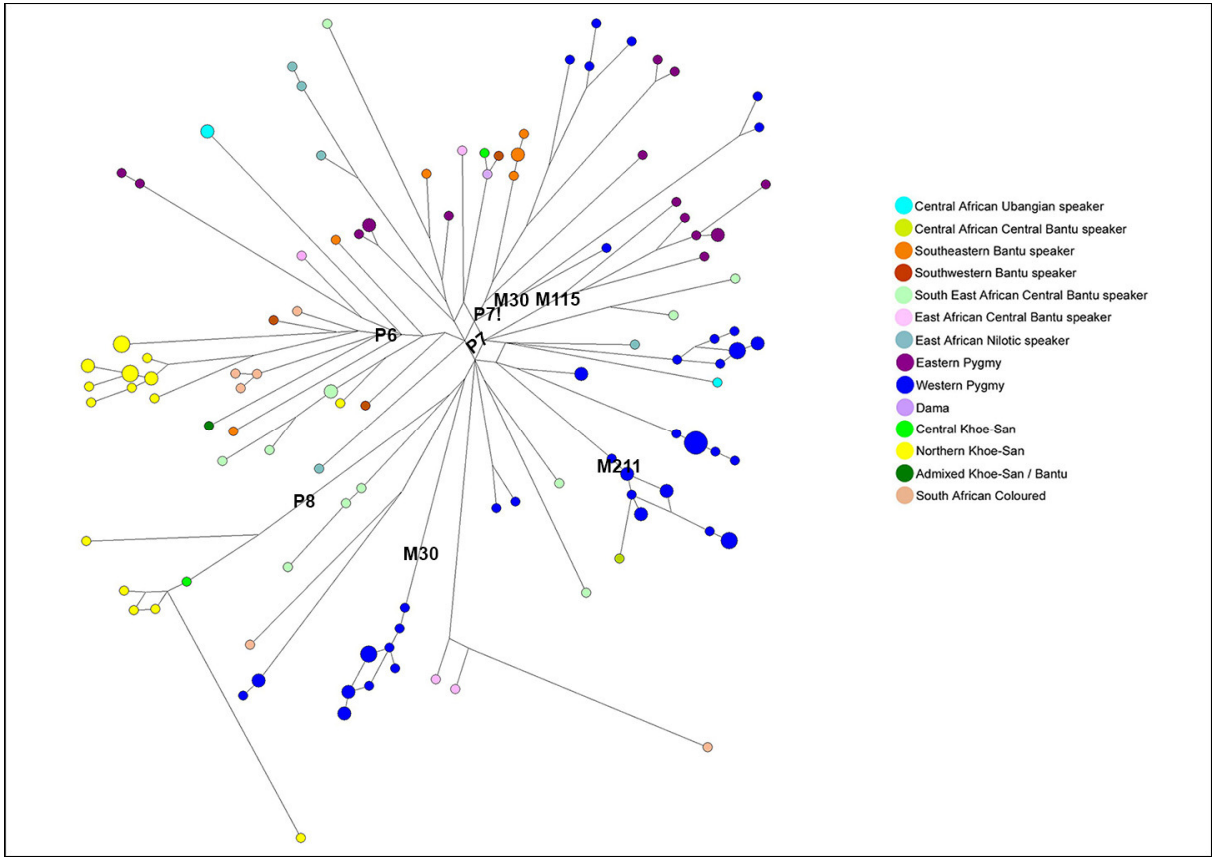

Figure 3.13: RM-MJ network of B-M112 based on a 6 SNP – 14 STR haplotype (P6-P7-M115-M30-P8-M211-DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438-DYS448-DYS456-DYS458-DYS635-GATA H4) with reduction threshold = 2 and Epsilon = 0.

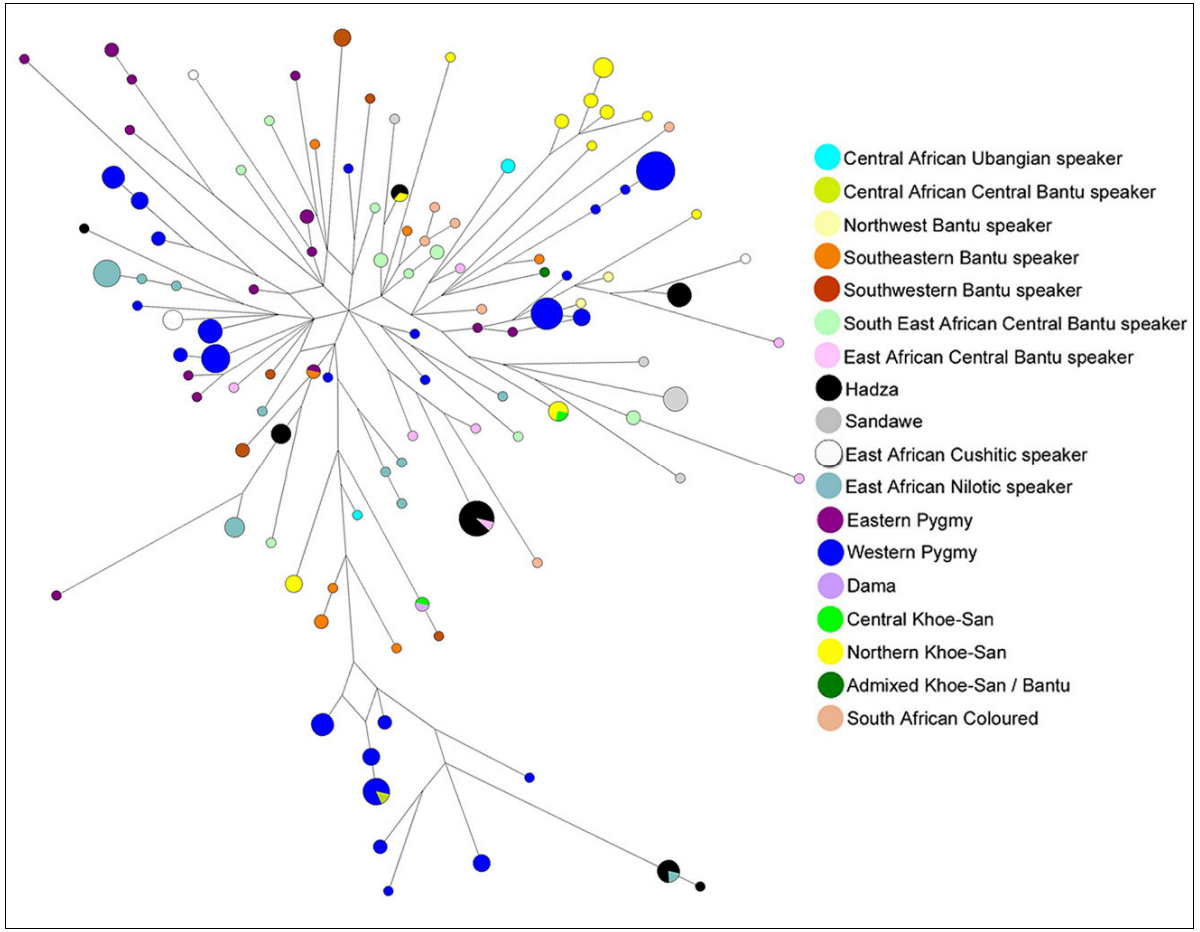

Figure 3.14: RM-MJ network of B-M112 based on a 9 STR haplotype (DYS19-DYS389I-DYS389II-DYS390-DYS391-DYS392-DYS393-DYS437-DYS438) with reduction threshold = 2 and Epsilon = 0.

Mean allelic variance and haplotype diversity indices were calculated for the major population groups, geographic regions, and for haplogroup B-M112 and its subclades (Table 3.12). Regionally, east Africa and southwestern Africa showed the highest levels of variance. Notably, the mean STR allelic variance exhibited in the Hadza and Sandawe (0.495 and 0.349 respectively) were relatively low in comparison to the east African regional estimate (1.011). The levels of variance within the B-M112 subclades appeared to correlate with the estimated TMRCAs of each subclade (Fig. 3.10A), with the exceptions of haplogroups B-M30 (allelic variance: 0.073; haplotype diversity: 0.314) and B-M211 (allelic variance: 0.061; haplotype diversity: 0.742), which showed lower variance than expected. The paragroup, B-M112*, contained the highest level of variation (allelic variance: 0.998; haplotype diversity: 0.974) in comparison to the defined subclades. East African B-M112* chromosomes showed highest variance (1.032) when examined regionally, while central African B-P7* chromosomes (allelic variance: 0.762; haplotype diversity: 0.906) showed highest variance for the paragroup, regionally.

The TMRCAs of haplogroup B-M112 and its subclades were estimated using BATWING (Fig 3.10A). With a TMRCA of 14.2 kya (CI 95%: 11.3 – 17.9 kya), the mutations defining haplogroup B-M112 were estimated to be around 2400 years older than those defining its subclade, haplogroup B-P7, which had a TMRCA of 11.8 kya (CI 95%: 9.4 – 14.9 kya). Haplogroup B-P6 arose independent of haplogroup B-P7, with a TMRCA of 8.9 kya (CI 95%: 5.9 – 12.9 kya), with similar times for both haplogroup B-M115 at 8.7 kya (CI 95%: 5.2 – 12.1 kya) and haplogroup B-M30 at 8.2 kya (CI 95%: 4.9 – 11.8 kya). Haplogroup B-P8 appeared to be a little younger with a TMRCA of 5.2 kya (CI 95%: 2.8 – 9.3 kya), though haplogroup B-M211 was the youngest at 2.7 kya (CI 95%: 1.4 – 5.7 kya).

Table 3.12: Allelic variance and haplotype gene diversity estimates of haplogroup B-M112 based on population groups, regions and subclades

Group	n	Allelic Variance (CI 95%)		Haplotype Gene Diversity (sd)	
Central African Central Bantu	1	-	-	-	-
Central African Ubangian	3	0.259	0.000 - 0.259	0.667	0.314
Central Khoe-San	2	0.000	0.000 - 1.778	1.000	0.500
Dama	1	-	-	-	-
East African Central Bantu	7	0.794	0.460 - 1.048	1.000	0.076
East African Cushitic	6	1.052	0.000 - 1.433	0.600	0.215
East African Nilotic	19	0.810	0.591 - 1.012	0.842	0.069
Eastern Pygmy	16	0.838	0.573 - 1.070	0.983	0.028
Hadza	29	0.495	0.327 - 0.644	0.796	0.049
Khoe-San/Bantu admixed	1	-	-	-	-
Northern Khoe-San	21	0.881	0.471 - 1.095	0.929	0.031
Northwest Bantu	2	0.167	0.000 - 0.167	1.000	0.500
Sandawe	9	0.349	0.000 - 0.639	0.583	0.183
South African Coloured	6	1.067	0.100 - 1.348	1.000	0.096
South East African Central Bantu	13	0.422	0.224 - 0.614	0.962	0.041
Southeastern Bantu	7	0.434	0.217 - 0.497	0.952	0.096
Southwestern Bantu	8	0.800	0.440 - 0.980	0.857	0.108
Western Pygmy	85	0.715	0.640 - 0.781	0.932	0.013
central Africa	107	0.845	0.749 - 0.943	0.955	0.009
east Africa	70	1.011	0.889 - 1.123	0.939	0.013
southeast Africa	13	0.420	0.225 - 0.609	0.962	0.041
southwest Africa	31	1.024	0.758 - 1.193	0.953	0.018
southern Africa	15	0.817	0.390 - 1.188	0.991	0.028
B-M112	236	1.073	0.999 - 1.149	0.985	0.002
B-M112*	109	0.998	0.909 - 1.089	0.974	0.006
B-P6	21	0.571	0.338 - 0.697	0.943	0.031
B-P7*	59	0.811	0.706 - 0.933	0.938	0.016
B-M115	7	0.291	0.106 - 0.481	0.952	0.096
B-M30	18	0.073	0.000 - 0.183	0.314	0.138
B-P8	6	0.119	0.000 - 0.237	0.600	0.215
B-M211	16	0.061	0.042 - 0.074	0.742	0.073
central African B-M112*	18	0.505	0.423 - 0.563	0.961	0.030
central African B-P7*	48	0.762	0.668 - 0.842	0.906	0.022
east African B-M112*	67	1.032	0.911 - 1.147	0.939	0.013
east African B-P7*	3	0.407	0.000 - 0.704	1.000	0.272
southern African B-M112*	8	0.448	0.212 - 0.522	0.964	0.077
southern African B-P7*	2	0.000	0.000 - 0.222	1.000	0.500
southeast African B-M112*	7	0.513	0.063 - 0.788	0.905	0.103
southeast African B-P7*	6	0.241	0.085 - 0.296	0.933	0.122
southwest African B-M112*	9	0.852	0.559 - 0.978	0.889	0.091

4 DISCUSSION

4.1 Y chromosome SBE assay development and optimization

While a number of genotyping techniques were used in the production of Y chromosome haplogroup data for the study, SBE (Syvänen, 1999) was used to generate most of the data. This method was chosen due to its convenience and relative affordability; and allowed us to examine Y chromosomes at a high resolution. The seven SBE assays that were developed were used to resolve Y chromosomes into haplogroups indigenous to Africa, or into a few common Eurasian lineages.

During development and optimization of the assays, it became apparent that estimated lengths of extension products differed from the designed lengths. This difference was ascribed to the migration rate of primers (which was influenced by their actual length), possible secondary structure (Konrad and Pentoney, 1993), mobility of the dye attached (Tu, et al., 1998), and the use of POP-7® polymer. The difference was especially stark for the M91 primer, a 25 base primer which appeared 11 bases longer. Despite these observations, profiles generated by all the assays were usually easily interpreted.

Since most aspecific peaks were due to insufficient purification, they usually did not permanently disrupt the interpretation of results. The one permanent aspecific peak in the Hg-B2b assay also caused no disruptions. Its link to the P7 primer may have been due to a problem that occurred during the synthesis of that primer. It was also more visible when overall peak height was decreased.

The increase in cycle number of the SBE reaction program resulted in an improvement in overall peak height. Variability of peak heights within some assays, however, was unavoidable, despite the adjustment of relative SBE primer concentrations. This was possibly influenced by the efficiency of interaction between SBE primers and template sequences.

4.1.1 Marker-specific issues

The marker M91, in the Hg-A assay, is a homopolymer length variant associated with a single base deletion in a poly-T tract (Underhill, et al., 2001). While the use of SBE in the screening of homopolymer variants is not common, the detection of the M91 mutation using the SBE method was successful. This was reaffirmed phylogenetically (Capelli, et al., 2003; Salas, et al., 2007) by the presence of this mutation exclusively in samples belonging to subclades of haplogroup A.

The validation process resulted in the redesign of just two SBE primers, P28 and M35. The initial P28 SBE primer did not pick up the mutation, likely due to non-specific primer binding; while the initial M35 primer resulted in an extremely low peak height when the mutation was present. This was possibly due to the preferential amplification of the ancestral allele, or a lower efficiency of binding by the original SBE primer.

Finally, in haplogroups B-P7 and B-M211, P7 showed the presence of two different extension products; displaying both the ancestral and derived states, simultaneously. This also occurred in haplogroup B-P8, with P8, additionally, exhibiting the same property. The presence of both states was confirmed when sequencing was performed. It is thus likely

that all samples in haplogroups B-P7*, B-P8 and B-M211 will display two peaks at the relevant markers. This was, likely, a consequence of these markers being located within paralogous sequence variants (Hammer, et al., 2003). It should be noted that such mutations are more susceptible to back-mutation through gene conversion, as it was with P25 (Adams, et al., 2006). For this reason more stable markers that resolve these subclades of B-P7 would be preferable.

4.2 Population affinities in sub-Saharan Africa

When examining the population affinities of sub-Saharan African populations using two measures of genetic distance, Fst and Rst, it was found that while both measures displayed strong similarities, a few differences were observed. The grouping of the non-African populations (in the MDS and cluster analysis) with the non-Bantu speakers, based on Fst distance, contrasted with a strong separation of African and non-African populations when using Rst distance. In both cases, however, the strongest cluster appeared to be the Bantu speakers and related populations. This is, likely, primarily due to the high frequencies of E-M2 in these populations.

Another notable difference between the Fst and Rst trees was the placing of the Dama population. In the Fst tree, they were placed among the Bantu speaker clade. This may have been due to these populations also exhibiting high frequencies of E-M2. In the Rst tree, however, the Dama were placed in the non-Bantu speaker clade, possibly indicating that while this population has high frequencies of the haplogroups characterising Bantu speaking populations, they did not necessarily derive them, at least solely, from Bantu speaking populations.

While east and southern African non-bantu speakers, and the Western Pygmy population, grouped together based on both distances, this may have been caused by the strong association observed in the Bantu speakers. This was further corroborated by the AMOVA analysis. The removal of the Bantu speakers group, when using linguistic classification, resulted in a large decrease (both Fst and Rst) in inter-group variation.

4.3 The phylogeography of haplogroup A in sub-Saharan Africa

While the resolution of the Y chromosome phylogeny has steadily increased over the past decade (YCC, 2002; Karafet, et al., 2008), its basal backbone had remained the same. The branching pattern and the distribution of haplogroups were used to support an African origin of anatomically modern humans (Underhill, et al., 2000), while populations from southern and east Africa were thought to have contained the oldest lineages (Hammer, et al., 2001; Semino, et al., 2002). It was only following the work of Cruciani, et al. (2011), who sought more SNPs to test the robustness of the basal phylogenetic structure, that the backbone was called into question. Their findings not only unearthed the polyphyletic nature of haplogroup A, but also pushed the estimated age of the oldest human Y chromosomes back to 142 kya. This newly estimated age was easier to reconcile with the coalescent age for mtDNA (Soares, et al., 2009) and with plausible scenarios of modern human origins. Following that, the discovery of haplogroup A00 by Mendez, et al. (2013), pushed the oldest human Y chromsosome further back to 338 kya. Haplogroups A00, A-P114, A-M31 and A-P108*, the deepest clades of the Y chromosome phylogeny, were found in central and northwest Africa. However, their frequencies were much lower than most other major haplogroups.

Haplogroup A00 was found, initially in an African American individual (Mendez, et al., 2013). Following its discovery, a more comprehensive search placed it among a few members of the Mbo of Cameroon (Mendez, et al., 2013).

According to the 2013 ISOGG Y chromosome phylogeny, A-P114 is a subclade of A0, haplogroup A0a1a. Haplogroup A0 is now defined by a number of newly discovered mutations, the first of which was P305. While much of this data has not yet been published, haplogroup A-P114 has been found in the Berbers of Algeria (Cruciani, et al., 2011), the Bakola of Cameroon (Batini, et al., 2011a; Cruciani, et al., 2011) and in Ghana (Scozzari, et al., 2012). No A-P114 chromosomes were found in the present study.

Haplogroup A-M31, while also not found in the present study, appeared more frequently than A-P114 in the published literature. With a northwest African distribution, it has been found in Mali (Underhill, et al., 2000; Wood, et al., 2005), in a number of populations in Guinea-Bissau (Gonçalves, et al., 2003; Rosa, et al., 2007), in Gambia among the Mandinka (Wood, et al., 2005), as well as in Cabo Verde (Gonçalves, et al., 2003) and in the Berbers of Morocco (Cruciani, et al., 2002; Cruciani, et al., 2011).

While P108 is now regarded as a basal mutation for most of the Y chromosome phylogeny, thus defining super-haplogroup A1b (or A1b-T), members of the paragroup, A-P108* (A1b*) have been found among central African populations in the Congo, Cameroon and Gabon (Batini, et al., 2011a). While we did not test for P108, the present study discovered a Y chromosome that was ancestral for all tested haplogroup A markers (apart from M91), whose STR haplotype matched A-P108* haplotypes. This were found in a Sangha-Sangha individual from CAR.

According to Cruciani, et al. (2011), while the discovery and correct placement of these lineages on the Y chromosome phylogeny would open up new perspectives on modern human evolution, more data would be needed to make stronger inferences.

4.3.1 Haplogroup A-M14

Haplogroup A-M14 was long thought to be exclusive to southern African Khoe-San populations (Underhill, et al., 2001; Wood, et al., 2005). It was in 2007, however, that the first indication of haplogroup A-M14 in central Africa occurred (Gonçalves, Spínola and Brehm, 2007); though, this discovery appeared to pass unnoticed (Batini, et al., 2011a). Batini, et al. (2011a) subsequently reported the finding of A-M14 among Baka Pygmies in Cameroon and Gabon. The present study reinforced the presence of A-M14 in central Africa (in the Gbaya). The fact that these A-M14 Y chromosomes were ancestral for markers M6 and M49, confirmed the existence of at least one independent A-M14 subclade outside of Khoe-San populations. The long branch lengths between the Baka, Gbaya populations which harboured A-M14* (Fig. 3.7), however, hint at the possibility that there may be more than one subclade within A-M14*.

Based on the absence of derived states for markers, M6 and M49, Batini, et al. (2011a) regarded the A-M14* Y chromosomes found in the Baka as an older lineage than those found in southern African Khoe-San populations. This was used as evidence of an evolutionary link or interaction between Western Pygmies and Khoe-San, whereby the presence of A-M14 in the Khoe-San was due to their interactions with a Pygmy group. This remains a distinct and plausible possibility, and the discovery of more central African A-M14 Y chromosomes strengthens its case, since the presence of different A-M14*

lineages in the Gbaya supports a central African origin for Khoe-San A-M14. The high allelic variance overall for A-M14*, with very low levels found in Baka and Gbaya individually (Table 3.4), indicate that, historically, haplogroup A-M14 may have been more frequent and more diverse.

4.3.2 Haplogroup A-M51

Haplogroup A-M51, together with its sister-clade A-M13, has evolved on the backbone of P108 (A1b), M32 (A1b1b) and M144 (A1b1b2; defined by seven other markers). It is also resolved into 2 further subclades: A-P291 and A-P102 (Karafet, et al., 2008). While haplogroup A-M51 was the most common haplogroup A subclade in southern Africa, it's numbers were still quite low in comparison to other haplogroups. The majority (~74%) of A-M51 Y chromosomes found were within or descended from Khoe-San populations. With high frequencies within the Khoe-San and low frequencies in other surrounding populations, it is very likely that haplogroup A-M51 originated within the Khoe-San, as supported by previous findings (Underhill, et al., 2000; Cruciani, et al., 2002; Wood, et al., 2005). The uniform distribution of A-M51 among the Northern, Central and Southern Khoe-San suggests that it arose prior to the splitting of these groups, or that gene flow between the groups was ongoing for a long period of time. While the Southern Khoe-San exhibited the highest allelic variance (Table 3.6), the Central Khoe-San and Northern Khoe-San values were not much different, making it difficult to pinpoint a point of origin within the Khoe-San.

Frequencies for haplogroup A-M51 among Southeastern Bantu speakers remained well below 5%, with the exception of Sotho speakers, in which it reached 7.4%. While at a low

frequency in Southeastern Bantu speaking populations, A-M51 occurred more frequently than another likely Khoe-San derived haplogroup, B-M112. This is an indication that A-M51 Y chromosomes in Southeastern Bantu speakers found their way into these populations through their interactions with Central and Southern Khoe-San populations, who have little to no presence of haplogroup B-M112. The numerous unique A-M51 haplotypes (Fig. 3.8) within Southeastern Bantu speakers, however, point to them having derived these Y chromosomes from populations not found in our dataset. Notably, the Southeastern Bantu speakers did share a haplotype with the ǀGui-ǁGhana-Kgalagari of Botswana. It is, thus, possible that this admixed Khoe-San/Bantu group may have contributed to A-M51 in Southeastern Bantu speakers. This, however, is not definitive. While the ǀGui-ǁGhana-Kgalagari appear to have a clear maternal link to Khoe-San ancestry, with ~90% frequency of L0d (Schlebusch, Lombard and Soodyall, 2013), their paternal ancestries appear to be derived primarily from Bantu speakers, with high frequencies of haplogroups B-M150 and E-M2 (Table 3.1). It is also quite possible that they derived their A-M51 Y chromosomes from Bantu speakers as well; especially when the shared haplotype clusters closely with numerous unique Southeastern Bantu speaker haplotypes.

4.3.3 Haplogroup A-M13

Haplogroup A-M13 was the dominant haplogroup A subclade in east Africa; with a low level presence in central and west Africa. Overall, the largest proportion of A-M13 Y chromosomes was found among Nilo-Saharan speaking populations (43%), with most of them being Nilotic speakers (37%). Afro-Asiatic speaking populations then comprised 27.1% of A-M13 Y chromosomes, with most major linguistic phyla represented, *viz.*

Chadic (4.2%), Cushitic (8.5%), Egyptian (1.1%), Omotic (2.1%), and Semitic (11.3%). Based on the high numbers of both Nilo-Saharan and Afro-Asiatic A-M13 Y chromosomes, as well as comparative levels of diversity, it is difficult to determine the source of haplogroup A-M13, out of these two likely groups. The moderate to high frequencies of A-M13 among both Afro-Asiatic and Nilo-Saharan populations in east Africa, as well as the presence of closely related and shared haplotypes (Fig. 3.9) are suggestive of extensive contact and gene flow over a long period of time. This is corroborated through autosomal data (Tishkoff, et al., 2009).

While, Nilo-Saharans, and in particular Nilotic speakers, hold the highest frequency of haplogroup A-M13 by far, this may be due to recent events. The biggest cluster on the RM-MJ network (Fig. 3.9) is composed of very closely related high-frequency East African Nilotic haplotypes that show a distinctive expansion pattern. This is strongly indicative that a recent rapid expansion occurred which resulted in the generation of a large proportion of the Nilotic A-M13; which would imply that while Nilotic populations may contain ancient forms of A-M13, they occur at a lower frequency than the more recent forms; an observation corroborated by Gomes, et al. (2010).

While east Africa contains the most A-M13 Y chromosomes, the high diversity of central and west African forms, together with the potentially bi-directional nature of migration along the Sahel (Tishkoff, et al., 2009) increases the potential for the haplogroup to have arose anywhere along the belt. The presence of related haplogroup A-M28 (A3a) in east Africa (Underhill, et al., 2000; Wood, et al., 2005), however, further supports its claim to A-M13. The lack of comparative data for the subclades of haplogroup A-M13, haplogroups A-M171 and A-M118, make the analysis of these lineages difficult. While A-

M171 was found in two central African Ubangian speakers, and A-M118 was found in a Ugandan Central Bantu speaker in the present study, little other data exist on these subclades apart from the findings of Underhill, et al. (2000) who found A-M118 in six Ethiopians and A-M171 in a Sudanese individual. Haplogroup A-M171 was recently removed from the ISOGG Y chromosome phylogeny, as until then it was only found in a single individual and so was regarded as a private mutation. Its presence in Central African Ubangian speakers now makes the case for it to be re-added to the phylogeny.

4.4 The phylogeography of haplogroup B in sub-Saharan Africa

While much younger than the oldest members of haplogroup A (Cruciani, et al., 2011), the monophyletic haplogroup B is still regarded as one of the oldest and most diverse of haplogroups. While most of this variation is encompassed within its two most common subclades, B-M150 and B-M112, individuals in other subclades have been found. Haplogroup B-M236* (B1*) was found in two Bamileke individuals from Cameroon (Cruciani, et al., 2002), while haplogroup B-M146 (B1a) was found in a Malian individual (Underhill, et al., 2000) and in a Mossi individual from Burkina Faso (Cruciani, et al., 2002). Haplogroup B-M182, which gives rise to both B-M150 and B-M112, also has a paragroup, B-M182*. This has been found at low frequencies among the Pygmy populations of Cameroon, CAR, and DRC (Wood, et al., 2005). The existence of individuals that do not belong to any of its known subclades affirms the existence of more diversity to uncover within the haplogroup. While the major subclades of haplogroup B are common throughout sub-Saharan Africa; the presence of a number of basal groups such as B-M236*, B-M182* and B-M60* in central Africa point to this region as the original source of haplogroup B.

4.4.1 Haplogroup B-M150

Haplogroup B-M150, or more specifically its subclade B-M152, appeared to be the most common branch of haplogroup B in sub-Saharan Africa. While the highest frequencies were found among the |Gui-||Ghana-Kgalagari of Botswana and Southeastern Bantu speakers, the low levels of variance (Table 3.10) in southern Africa and in the |Gui-||Ghana-Kgalagari indicate that these high frequencies may be due to the rapid expansion of related haplotypes. Consequently, haplogroup B-M152 is regarded as a signature of the Bantu Expansion (Gomes, et al., 2010); though it should be noted that Gomes, et al. (2010) made use of a different marker (M109) which delineates the same haplogroup. While B-M152 is found in central Africa at low to moderate frequencies, its variance in the region is substantially higher than in southern Africa. Due to its rarity in west and east Africa, a case could, thus, be made for the origin of B-M152 in central Africa.

Since only a few studies have discriminated lineages within haplogroup B-M150 (apart from B-M152), it remains difficult to infer the origins of these other subclades; including the origin of the ancestral B-M150 Y chromosome. The presence of B-M150* (x M218) within the Mbuti, however, raises the question of whether the original B-M150 Y chromosomes were introduced from Pygmy populations into neighbouring pastoral / agricultural groups, before undergoing rapid expansion.

While the TMRCA for B-M150 was estimated at 14.2 kya, this was based primarily on B-M152 chromosomes. As a result, this is unlikely to be a clear reflection of the age of haplogroup B-M150. A more accurate estimate would require the discovery of more individuals belonging to other B-M150 subclades.

4.4.2 Haplogroup B-M112

The second most common haplogroup B subclade was that of B-M112, a very diverse collection of haplogroups, more so than the previously discussed subclades of haplogroups A and B. Defined by the SNP M112 (Underhill, et al., 2000), the haplogroup has been resolved further into eleven subclades (Batini, et al., 2011a; Scozzari, et al., 2012) through the discovery of the defining markers: M30, M108.2, M115, M129, M169 (Underhill, et al., 2000), MSY2.1, P6, P7, P8 (Hammer, et al., 2001), M211 (Underhill, et al., 2001), P70, 50f2(P), M192, P112 (Karafet, et al., 2008) and V341 (Scozzari, et al., 2012) (Fig. 3.10A). The initial splitting of B-M112 resulted in the formation of haplogroups B-P112 and B-50f2(P) (Scozzari, et al., 2012). Due to the rarity of B-P112, most of the known diversity in B-M112 is found in B-50f2(P). Within B-50f2(P), haplogroups B-P7 and B-P6 diverged from the ancestral B-50f2(P) independently. B-P7 then appears to be ancestral to the other subclades, *viz.,* B-MSY2.1 and B-P8. Haplogroups B-M115, B-M30, and B-M211 cluster within B-MSY2.1 (Batini, et al., 2011a). Finally B-M108.2 is derived from B-M30, though this haplogroup has only been observed in a single individual (Underhill, et al., 2000; Cruciani, et al., 2002).

Many of the subclades of B-M112 displayed strong geographic structure. While the B-M112 paragroup, B-M112*, occurs throughout sub-Saharan Africa, it is found at its highest frequencies in the Hadza from east Africa. The demonstrated haplotype sharing between one of the major B-M112* Hadza haplotypes and a B-P7* Central Bantu speaker (with the next closest haplotype also belonging to B-P7*), however, calls into question the high frequency of B-M112* in east Africa and in turn the low frequencies of B-P7 in the region. The presence of B-M112* in the Southeastern Bantu speakers (0.5 - 3.2%) of

southern Africa and the southwestern Bantu speakers (2.2% - 11.5%) of Angola and Namibia could have resulted from recent male gene flow from Khoe-San populations into these Bantu speakers as they migrated into southern Africa following recent the Bantu expansion. This is supported in Coelho, et al. (2009) who found significant levels of admixture in the Kuvale and Nyaneka-Nkhumbi from Khoe-San sources, based on both Y chromosome and mtDNA. The Southeastern Bantu speakers also show admixture with Khoe-San populations, based on the moderate frequencies of the mtDNA haplogroup L0d (Schlebusch, Naidoo and Soodyall, 2009), low frequencies of the Y chromosome haplogroup A3b1 in these populations (Naidoo, et al., 2010), and a recent genome-wide autosomal study that supported Khoe-San admixture into Southeastern Bantu speakers (Schlebusch, et al., 2012). While B-M112* exhibited high levels of diversity, the networks (Fig.3.13) showed some association of Y chromosome haplotypes among east Africans, southeastern Africans, southwestern Africans and Mbuti Pygmies. This suggests a range distribution of B-M112* spanning a region between east Africa and southern Africa. The Western Pygmy haplotypes appear to be isolated; however, their strong divergence from the Ubangian haplotypes indicate historically higher diversity than the present day in central African B-M112*. Due to the high diversity and wide distribution of B-M112* across sub-Saharan Africa, it was not possible to locate where B-M112 arose.

Haplogroup B-P7 has maintained its greatest presence, along with most of its subclades in central Africa, with a scattering of ancestral forms in east and southeastern Africa. Its presence in southwestern Africa is constrained to one of its subclades, haplogroup B-P8, and only two chromosomes found in southern Africa. It displays highest diversity within central Africa, having evolved a number of subclades within the region. Haplogroup B-MSY2.1 (Batini, et al., 2011a) is the major subclade of B-P7 and is the foundation upon

which haplogroups B-M115, B-M30, and B-M211 were derived. Its high frequency in central Africa and distinctive absence elsewhere make the case for the evolution of B-MSY2.1 in central Africa, possibly in the Baka who maintain high frequencies of both B-MSY2.1 and the ancestral B-P7* (Berniell-Lee, et al., 2009; Batini, et al., 2011a). The place of origin for the ancestral B-P7, however, is more difficult to discern. While central Africa shows higher frequencies and high diversity of B-P7*, the eastern and southeastern African forms show comparable diversity. The stronger presence of B-P7 subclades in central Africa provides support for it as the place of origin; however, the levels of diversity in southeastern Africa imply that we cannot rule out the presence of as yet undiscovered subclades within the paragroup.

The localised distributions of haplogroups B-M115, B-M30, B-M211 and B-P6 point to them having arisen and evolved in or close to their current locale. Found only in the Mbuti of DRC, haplogroup B-M115 is characterised by the M115 and M169 mutations (Underhill, et al., 2000) on a background of P7 and MSY2.1 (Batini, et al., 2011a). A separation time of 10 to 15 kya between Western and Eastern Pygmy populations has been suggested (Batini, et al., 2011a) based on divergence between B-M115 and B-M30/B-M211. If no male-mediated gene flow occurred between the groups after 10 000 years, this would imply that the B-MSY2.1 chromosomes from which haplogroup B-M115 evolved disappeared from the Mbuti population through genetic drift. Both haplogroups B-M30 and B-M211 were found primarily among the Western Pygmy populations in central Africa. The TMRCA for haplogroup B-M30 also appeared to be older than expected, in comparison to its allelic variance (0.073). The older estimate for B-M30, however, seems more congruent, allowing time for at least two observed mutational reversions to have occurred, i.e. to P7 and M129 (Batini, et al., 2011a). The low allelic variances of B-M30

and B-M211 could then be due to a loss of variation as a result of a bottleneck, which has been observed at other loci (Patin, et al., 2009; Batini, et al., 2011b). While found primarily in the Northern Khoe-San groups of Angola and Namibia, the presence of haplogroup B-P6 in other populations is unlikely to be due to interactions with the !Xun and Ju|'hoansi, as indicated by the distances of the |Gui-||Ghana-Kgalagari and South African Coloured haplotypes to the main Northern Khoe-San clusters (Fig. 3.13). The South African Coloured individuals likely trace a portion of their ancestry to the Khoe-San populations who had previously inhabited the Cape of South Africa (Nurse, Weiner and Jenkins, 1985; Schlebusch, et al., 2012).

While the absence of haplogroup B-P8 elsewhere would point to it being autochthonous to the Khoe-San of southwestern Africa, we examined several scenarios to evaluate its ancestry. The first possibility is that haplogroup B-P7 existed previously among the Khoe-San forebears at higher frequency and diversity, with this haplogroup drifting to extinction, leaving only its subclade, haplogroup B-P8 behind. Another possibility is that B-P8 evolved from chromosomes derived from interactions between the Khoe-San and central African populations, and lastly, haplogroup B-P8 could have evolved from chromosomes derived from interactions between the Khoe-San and east African populations; these last two options being possible due to the higher frequency and diversity of haplogroup B-P7 in central and east Africa. The presence of B-P7* in southern Africa was accounted for by only two chromosomes. One of these B-P7* chromosomes appeared to be related to east and southeastern African B-P7* chromosomes. The most common B-P8 haplotype clustered close to the Sandawe B-M112* chromosomes of Tanzania (Fig. 3.14). While this could be regarded as homoplasy, the hidden presence of B-P7 chromosomes among Hadza populations brings into question the homogenous nature of haplogroup B-M112 among

Tanzanian Khoisan speakers. Another potential source would be the Western Pygmy populations of Gabon and Cameroon, which contained B-P7* chromosomes that were relatively close haplotypically to the most common B-P8 haplotype (Fig. 3.14), with one Northwest Bantu speaker from Gabon even harbouring the distinctive DYS439 NULL allele associated mainly with B-P8 chromosomes. As with haplogroup A-M14, Batini, et al. (2011a) used the common presence of B-P8 to show evidence of gene flow between Khoe-San and Pygmy populations. They estimated the split between the Khoe-San- and Pygmy-specific B-P7 lineages at around three to four kya; which falls within the range of our estimate for the TMRCA of haplogroup B-P8.

4.5 Potential impact of the Last Glacial Maximum

An examination of TMRCA estimates for the major A and B subclades found in the study resulted in an interesting observation. With the exception of haplogroup A-M14, haplogroups A-M13, A-M51, B-M150 and B-M112 appeared to have very similar TMRCAs (14.2 – 14.5 kya). Seeing that A-M13 and A-M51 are sister clades (with the same being true for B-M150 and B-M112), common TMRCAs between them may reflect the point at which they split from each other (Platt, D. E. – personal communication, 2012). However, the same reasoning cannot be used to explain similar TMRCAs across haplogroups (e.g. A-M51 versus B-M112). These TMRCAs appear to coincide with the end of the Last Glacial Maximum (LGM). During the LGM (19 – 26 kya) the climate in Africa was generally colder and more arid, which caused desert and semi-desert areas to expand (Clark, et al., 2009). This led to the abandonment of certain areas and population contractions into still habitable refugia. The end of the LGM brought about increased precipitation and re-colonisation of previously uninhabitable zones (Barham and Mitchell,

2008). It might be that these haplogroups emerged following the subsiding of the LGM. The possibility also exists that the population contractions and bottlenecks caused by the LGM are responsible for the similar TMRCAs, which estimate the point of population expansion, and not the ages of the haplogroups.

4.6 Expansion of haplogroup E and its effect on sub-Saharan African diversity

While haplogroup E may be the youngest of the major African haplogroups, it has clearly contributed substantially to the shaping of the current demography. Haplogroup E-M2 is its most frequent subclade at ~63% overall (Table 3.1), and accounts for most of its distribution in sub-Saharan Africa. Haplogroups E-M35 (5.3%) and E-M75 (4.5%) then account for most of the rest of its distribution, with other forms of haplogroup E occurring at negligible frequencies. Even those of haplogroups DE-YAP* and E-M40* in Table 3.1 are likely to be slightly inflated due to the markers screened in some of the published data that these frequencies were obtained from (Tishkoff, et al., 2007; Berniell-Lee, et al., 2009; Coelho, et al., 2009).

When considering the origins of haplogroup E, the frequency of its ancestral paragroup (E-M40/M96*) provides little assistance due to its rarity. It is also possible that any presence of the paragroup is representative of, as yet, unknown subclades (de Filippo, et al., 2011). While there is still some debate over whether haplogroup E arose in Asia or Africa; the discovery of ancestral DE* in Africa (Weale, et al., 2003; Rosa, et al., 2007), and the

pushing back of the age estimate of haplogroup E (Karafet, et al., 2008) lend support to an African origin.

4.6.1 Haplogroup E-M2

While the mutations V38 and V100 currently define haplogroup E1b1a (Trombetta, et al., 2011), the lineages within the E-M2 subclade are still the most frequent. The highest frequencies of the haplogroup were found in Bantu speakers in central, east and southern Africa, as well as in populations speaking related Niger-Congo languages in west and central Africa. The high frequency and diversity of E-M2 in west African populations was used as evidence for an origin and early expansion of the haplogroup in the region (Rosa, et al., 2007). This was further reinforced by de Filippo, et al. (2011), who showed that west African populations displayed high frequencies of E-M2*, which exhibited a clinal reduction toward east and southern Africa. Thus, the high frequencies of E-M2 in Bantu speaking populations across sub-Saharan Africa were representative of identified subclades of E-M2, viz. E-U174 and E-U175 (de Filippo, et al., 2011). While haplogroup E-M2 may have been responsible for the strong clustering of Bantu speakers and linguistically related populations (Figs. 3.2 – 3.5), at both haplogroup and haplotype level, genetic structure was missed by not screening for markers U174 and U175. Based on the above, the early expansion of haplogroup E-M2 in west Africa may be responsible for the drastic reduction in frequency of ancient Y chromosome haplogroups A00, A-P114, A-M31 and A-P108*. This may have been compounded by the more recent Bantu Expansion, which resulted in the spread of haplogroups E-U174 and E-U175 across sub-Saharan Africa. Their spread may have likely resulted in restrictions to the movement and demic growth of haplogroup A and B subclades, including those now lost to us through extinction.

4.6.2 Haplogroup E-M35

Haplogroup E-M35 is the only African haplogroup to attain appreciable frequencies outside of Africa, and while found throughout sub-Saharan Africa, it was nowhere near as commonly found as the related haplogroup E-M2. While a few individuals were found in central Africa, haplogroup E-M35 has usually been found in northwest Africa, east Africa and southern Africa. Although the northwest African presence was represented primarily by its subclade, haplogroup E-M81 (Bosch, et al., 2001) in Berber populations, Rosa, et al. (2007) found several E-M78 individuals from Guinea-Bisseau in west Africa among Atlantic speakers. While a few E-M35* individuals were also found, M81 was not screened for, and so may have been present. As E-M78 was also found at low frequency in northwest Africa (Bosch, et al., 2001), this may be a likely source for the presence of haplogroup E-M35 in the region. In east Africa, haplogroup E-M35 was found in a number of populations, with the highest frequencies among Cushitic speakers (41.7%) and the Sandawe (32.8%). Relatively low frequencies, however, were found among east African Bantu speakers (6.6%). This, together with the low frequencies found among other Bantu speakers from central and southern Africa, may indicate that E-M35 found its way into Bantu speakers through admixture. It is in east Africa, however, that E-M35 is believed to have originated, due to a number of factors, including the presence of its direct ancestral clade (E-M215), its high frequency and its diversity in the region (Cruciani, et al., 2004; Semino, et al., 2004). In southern Africa, haplogroup E-M35 was found mainly in Khoe-San populations, with the highest frequencies among Central Khoe-San and Southern Khoe-San groups (28.4% and 26.8%, respectively). Haplogroup E-M35 was also found among South African Whites, though this presence is mostly derived from their European origins. While found at low frequencies in the South African Coloured population (6.8%),

this group may derive E-M35 chromosomes from both their Khoe-San and European ancestors (Nurse, Weiner and Jenkins, 1985; Schlebusch, 2010).

A major proportion of haplogroup E-M35 variation in east and southern Africa, however, has been delineated by the M293 marker (Henn, et al., 2008). Henn, et al. (2008) found high frequencies of haplogroup E-M293 among the Datog, Burunge, Sandawe and Hadza in Tanzania, while also observing it in the Khwe and !Xun of their southern African sample. It also occurred at low frequencies in the Bantu speakers of east and southern Africa (Henn, et al., 2008). While our study did not screen for M293, it may be that a major proportion of the southern African cohort of E-M35 is comprised of E-M293, seeing as Henn, et al. (2008) found all their Khwe and !Xun E-M35* chromosomes to be derived for M293. The presence of E-M293 in east and southern Africa was used to associate the spread of pastoralism from east Africa to southern Africa, based on a demic diffusion model (Henn, et al., 2008).

Based on an examination of haplogroup E-M35* in various southern African Khoe-San populations, Schlebusch (2010) deduced that the spread of pastoralism toward the south was, however, unlikely to have been a clear-cut demic or cultural diffusion. Haplogroup E-M35* was found at its highest frequency in the Khwe (Central Khoe-San) at 46%, and so the group that introduced pastoralism to southern Africa may have been the ancestors of the Khwe. Today, however, the Khwe do not practice a pastoral lifestyle, due to them living in a tsetse fly infested area (Schlebusch, 2010).

Following the Khwe, the next highest frequencies of E-M35* were found in the Khomani (Southern Khoe-San) at 27%, and the Nama (Central Khoe-San) at 21%. The presence of high frequencies of haplogroup A-M51 in the Nama and Khomani, and haplogroup B-

M112 in the Nama, however, suggested that there was not a full demic diffusion of pastoralists into southern Africa, but rather incorporation into the indigenous hunter-gatherer populations (Schlebusch, 2010). Nonetheless, it is clear that at least the Nama adopted pastoral practices and, like the Khwe, speak a Khoe language. Further south, Coloured populations, whose Khoe-San ancestors may have been the southern Khoe groups (e.g. Griqua and Cape KhoeKhoe) exhibited much lower frequencies of E-M35* (6-9%). This suggests that while the Khoe language and pastoralism reached South Africa, the male genetic input was lower than further north.

The presence of haplogroup E-M35* in the Northern Khoe-San was represented almost solely by the !Xun in Angola (Schlebusch, 2010). With a frequency of 15%, the !Xun showed lower levels of E-M35* than the Khomani or Nama; though higher than the South African Coloured populations. Notably, the !Xun did not adopt the Khoe language or pastoralism from the interactions with the Khwe; though pastoralism in the !Xun eventually arrived through contact with neighbouring Bantu speakers (Schlebusch, 2010). Their interactions with the Khwe, however, were evidenced not only through the E-M35* lineages, but also through high levels (~30%) of mtDNA maternal lineages derived from the Khwe (Schlebusch, 2010). Thus, the spread of haplogroup E-M35* (or likely E-M293) from east Africa through southern Africa did not directly parallel the spread of pastoralism.

4.7 Admixed populations in sub-Saharan Africa

As a result of human movements and migrations over the last 400 to 600 years, there have been increased levels of gene flow between previously separated populations (Winkler, Nelson and Smith, 2010). This gene flow between populations, known as admixture, has

produced numerous unique admixed populations around the world, with some residing in Africa.

The South African Coloured population is one of these unique admixed populations, displaying some of the highest levels of recent admixture (Quintana-Murci, et al., 2010). From the paternal perspective, the South African Coloured population displayed ancestral contributions from African, European and Asian sources. The presence of haplogroups A-M51 (7.4%) and B-M112 (2.0%, primarily its subclade B-P6) indicated some level of Khoe-San input to the South African Coloured paternal ancestry. Haplogroups E-M2 (35.8%), B-M150 (4.1%) and E-M75 (3.4%) were likely introduced from South African Southeastern Bantu speakers. It should be noted however, that the strong presence of E-M2 in southern African Khoe-San populations, introduces the possibility that, at least some of the South African Coloured E-M2 was derived from Khoe-San sources as well. A clearer delineation of the subclades within the South African Coloured E-M35 cohort would potentially find that those chromosomes were derived from both Khoe-San and European forebears. The European input was represented mainly by relatively high frequencies of haplogroups R-M343 (16.2%) and I-M170 (8.1%); while Asian haplogroups, such as C-M130 (2.0%), H-M69 (0.7%), J-M172 (0.7%), L-M11 (0.7%) and O-M175 (2.0%), occurred at low levels. These results appeared to correspond to those of Quintana-Murci, et al. (2010), who came to similar conclusions regarding the paternal ancestry of the South African Coloured population.

As shown previously by the presence of haplogroups A-M51 and B-M112 in southern African Bantu speakers, admixture also occurred between them and Khoe-San populations. The frequencies of these lineages, however, were relatively low; especially compared to

the levels of haplogroup E-M2 found in Khoe-San populations. This was a strong indication that male-mediated gene flow between these groups was heavily biased in a Bantu speaker to Khoe-San direction, and reinforces previous findings (Hammer, et al., 2001; Destro-Bisol, et al., 2004; Wood, et al., 2005). The levels of haplogroup E-M2 also differed between Khoe-San populations, with the Central Khoe-San (41.9%) exhibiting almost double the frequencies found in Southern Khoe-San (22.0%) and Northern Khoe-San (21.5%) populations. A prior study (Scozzari, et al., 1997) had shown that the Khwe, exhibited E-M2 frequencies of 50%. While the Khwe were, linguistically, closer to Khoe speaking San, phenotypically, they were similar to neighbouring Bantu-speakers (Nurse and Jenkins, 1977; Schlebusch, 2010). This, together with high levels of haplogroup E-M35 marks the Khwe as an enigmatic group with a complex history, and potentially another unique admixed population. It should also be noted that the Khoe-San represent the deepest population divergence of humankind (Schlebusch, et al., 2012). Their ancient roots allow for the possibility that they derived their E-M2 chromosomes prior to the Bantu Expansion.

4.8 Study limitations and potential future directions

As with any study, there are limitations to the inferences that can be made from the results obtained. While the non-recombining nature of the NRY allows it to preserve a simpler record of its history; this also means that the NRY is regarded as a single locus, and so can only provide a limited amout of information pertaining to population processes (Jobling and Tyler-Smith, 2003). It should also be noted that the SBE assays used in the study were designed, based on an older version of the Y chromosome phylogeny. The phylogeny has since been updated numerous times and currently contains many more markers and

lineages. Future studies of African Y chromosome haplogroup phylogeography will, likely, take this into account and examine the distributions of these additional lineages. An example of where this would be useful would be the screening of E-M2 subclades, such as E-U174 and E-U175, in southern African Khoe-San populations, in order to investigate the origins of their haplogroup E-M2 chromosomes. While de Filippo et al. (2011) studied the distribution of these lineages in African populations, their Khoe-San sample was limited and did not contain haplogroup E-M2. The present study has shown that E-M2 is found at moderate frequencies in Khoe-San populations, which raises the questions of where and when they were introduced into these groups.

5 CONCLUSION

The examination of the phylogeography of African Y chromosome haplogroups A and B, as well as an assessment of the genetic affinities of the major population groups included in the study have allowed us the opportunity to elucidate how modern human evolution and population movements have contributed in shaping the gene pool among sub-Saharan African populations; especially from the male perspective.

The wide geographic distribution of haplogroup A, together with its position at the root of the Y chromosome phylogeny, and its extremely high levels of diversity support both an early diversification of the haplogroup and early dispersal of its members throughout Africa. The spread of the major members of haplogroup B, however, are possibly due to post-glacial population movements, in the case of haplogroup B-M112, and more recent population expansions, which have led to the common presence of haplogroup B-M152 across sub-Saharan Africa. It is the spread of haplogroup E, however, which has created the biggest impact on African populations. The expansion of its subclade, E-M2, has likely resulted in the diminished presence of many of the subclades of haplogroups A and B.

The Y chromosome compositions of present sub-Saharan African populations are, thus, the result of several diversification events, followed by migration, and mingling of population groups, over the last 300 000 years. It could be said, that most of sub-Saharan Africa belongs to a unique admixed population.

6 REFERENCES

Adams, S.M., King, T.E., Bosch, E. and Jobling, M.A., 2006. The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal marker P25 through gene conversion. *Forensic science international*, 159(1), 14-20.

Arredi, B., Poloni, E.S., Paracchini, S., Zerjal, T., Fathallah, D.M., Makrelouf, M., Pascali, V.L., Novelletto, A. and Tyler-Smith, C., 2004. A predominantly neolithic origin for Y-chromosomal DNA variation in North Africa. *American journal of human genetics,* 75(2)**,** 338-345.

Avise, J.C., Arnold, J., Ball, R.M., Bermingham, E., Lamb, T., Neigel, J.E., Reeb, C.A. and Saunders, N.C., 1987. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. *Annual review of ecology and systematics,* 18(1)**,** 489-522.

Bandelt, H.J., Forster, P., Sykes, B.C. and Richards, M.B., 1995. Mitochondrial portraits of human populations using median networks. *Genetics,* 141(2)**,** 743-753.

Bandelt, H.J., Forster, P. and Röhl, A., 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular biology and evolution,* 16(1)**,** 37-48.

Barham, L. and Mitchell, P., 2008. *The first Africans: African archaeology from the earliest toolmakers to most recent foragers.* Cambridge: Cambridge University Press.

Balaresque, P., Sibert, A., Heyer, E. and Crouau-Roy, B., 2006. Unbiased interpretation of haplotypes at duplicated microsatellites. *Annals of human genetics*, 71(2):209-219.

Batini, C., Ferri, G., Destro-Bisol, G., Brisighelli, F., Luiselli, D., Sánchez-Diz, P., Rocha, J., Simonson, T., Brehm, A., Montano, V., Elwali, N.E., Spedini, G., D'amato, M.E., Myres, N., Ebbesen, P., Comas, D. and Capelli, C., 2011a. Signatures of the preagricultural peopling processes in sub-Saharan Africa as revealed by the phylogeography of early Y chromosome lineages. *Molecular biology and evolution,* 28(9)**,** 2603-2613.

Batini, C., Lopes, J., Behar, D.M., Calafell, F., Jorde, L.B., Van Der Veen, L., Quintana-Murci, L., Spedini, G., Destro-Bisol, G. and Comas, D., 2011b. Insights into the demographic history of African Pygmies from complete mitochondrial genomes. *Molecular biology and evolution,* 28(2)**,** 1099-1110.

Battaglia, V., Fornarino, S., Al-Zahery, N., Olivieri, A., Pala, M., Myres, N.M., King, R.J., Rootsi, S., Marjanovic, D., Primorac, D., Hadziselimovic, R., Vidovic, S., Drobnic, K., Durmishi, N., Torroni, A., Santachiara-Benerecetti, A.S., Underhill, P.A. and Semino, O., 2009. Y-chromosomal evidence of the cultural diffusion of agriculture in southeast Europe. *European journal of human genetics : EJHG,* 17(6)**,** 820-830.

Beleza, S., Gusmão, L., Amorim, A., Carracedo, A. and Salas, A., 2005. The genetic legacy of western Bantu migrations. *Human genetics,* 117(4)**,** 366-375.

Bergen, A.W., Wang, C.Y., Tsai, J., Jefferson, K., Dey, C., Smith, K.D., Park, S.C., Tsai, S.J. and Goldman, D., 1999. An Asian-Native American paternal lineage identified by RPS4Y resequencing and by microsatellite haplotyping. *Annals of human genetics,* 63(1)**,** 63-80.

Berniell-Lee, G., Calafell, F., Bosch, E., Heyer, E., Sica, L., Mouguiama-Daouda, P., Van der Veen, L., Hombert, J.-M., Quintana-Murci, L. and Comas, D., 2009. Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. *Molecular biology and evolution*, 26(7), 1581-1589.

Bosch, E., Calafell, F., Comas, D., Oefner, P.J., Underhill, P.A. 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. *American journal of human genetics,* 68(4)**,** 1019-1029.

Capelli, C., Tschentscher, F. and Pascali, V.L., 2003. "Ancient" protocols for the crime scene? Similarities and differences between forensic genetics and ancient DNA analysis. *Forensic science International*, 131(1), 59-64.

Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., Bishop, C., Fellous, M., Purrello, M., Fiori, G. and Siniscalco, M., 1985. A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science (New York, N.Y.),* 230(4732)**,** 1403-1406.

Clark, P.U., Dyke, A.S., Shakun, J.D., Carlson, A.E., Clark, J., Wohlfarth, B., Mitrovica, J.X., Hostetler, S.W. and Mccabe, A.M., 2009. The Last Glacial Maximum. *Science (New York, N.Y.),* 325(5941)**,** 710-714.

Coelho, M., Sequeira, F., Luiselli, D., Beleza, S. and Rocha, J., 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC evolutionary biology*, 9:80.

Cruciani, F., Santolamazza, P., Shen, P., Macaulay, V., Moral, P., Olckers, A., Modiano, D., Holmes, S., Destro-Bisol, G., Coia, V., Wallace, D.C., Oefner, P.J., Torroni, A.,

Cavalli-Sforza, L.L., Scozzari, R. and Underhill, P.A., 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *American journal of human genetics,* 70(5)**,** 1197-1214.

Cruciani, F., La Fratta, R., Santolamazza, P., Sellitto, D., Pascone, R., Moral, P., Watson, E., Guida, V., Colomb, E.B., Zaharova, B., Lavinha, J., Vona, G., Aman, R., Cali, F., Akar, N., Richards, M., Torroni, A., Novelletto, A. and Scozzari, R., 2004. Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *American journal of human genetics*, 74(5): 1014-1022.

Cruciani, F., La Fratta, R., Trombetta, B., Santolamazza, P., Sellitto, D., Colomb, E.B., Dugoujon, J.-M., Crivellaro, F., Benincasa, T., Pascone, R., Moral, P., Watson, E., Melegh, B., Barbujani, G., Fuselli, S., Vona, G., Zagradisnik, B., Assum, G., Brdicka, R., Kozlov, A.I., Efremov, G.D., Coppa, A., Novelletto, A. and Scozzari, R., 2007. Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Molecular biology and evolution,* 24(6)**,** 1300-1311.

Cruciani, F., Trombetta, B., Massaia, A., Destro-Bisol, G., Sellitto, D. and Scozzari, R., 2011. A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *American journal of human genetics,* 88(6)**,** 814-818.

De Filippo, C., Heyn, P., Barham, L., Stoneking, M. and Pakendorf, B., 2010. Genetic perspectives on forager-farmer interaction in the Luangwa valley of Zambia. *American journal of physical anthropology*, 141(3):382-394.

De Filippo, C., Barbieri, C., Whitten, M., Mpoloka, S.W., Gunnarsdóttir, E.D., Bostoen, K., Nyambe, T., Beyer, K., Schreiber, H., De Knijff, P., Luiselli, D., Stoneking, M. and Pakendorf, B., 2011. Y-chromosomal variation in sub-Saharan Africa: insights into the history of Niger-Congo groups. *Molecular biology and evolution,* 28(3)**,** 1255-1269.

Delfiner, P. 1976. Linear estimation of non-stationary spatial phenomena. In: M. Guarascio, M. David and C. Huijbregts, eds. 1976. *Advanced geostatistics in the mining industry: proceedings of the NATO Advanced Study Institute held at the Istituto di Geologia Applicata of the University of Rome, Italy, 13-25 October 1975.* Dordrecht: D. Riedel Publishing Company.

Destro-Bisol, G., Donati, F., Coia, V., Boschi, I., Verginelli, F., Caglià, A., Tofanelli, S., Spedini, G. and Capelli, C., 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Molecular biology and evolution,* 21(9)**,** 1673-1682.

Di Giacomo, F., Luca, F., Popa, L.O., Akar, N., Anagnou, N., Banyko, J., Brdicka, R., Barbujani, G., Papola, F., Ciavarella, G., Cucci, F., Di Stasi, L., Gavrila, L., Kerimova, M.G., Kovatchev, D., Kozlov, A.I., Loutradis, A., Mandarino, V., Mammi', C., Michalodimitrakis, E.N., Paoli, G., Pappa, K.I., Pedicini, G., Terrenato, L., Tofanelli, S., Malaspina, P. and Novelletto, A., 2004. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. *Human genetics,* 115(5)**,** 357-371.

Diamond, J., 1997. *Guns, Germs, and Steel.* New York: W.W.Norton & Company.

Excoffier, L. and Lischer, H.E.L., 2010. Arlequin suite ver 3.5: a new series of programs to perform population genetics analyses under Linux and Windows. *Molecular ecology resources,* 10(3)**,** 564-567.

Gomes, V., Sánchez-Diz, P., Amorim, A., Carracedo, A. and Gusmão, L., 2010. Digging deeper into East African human Y chromosome lineages. *Human genetics,* 127(5)**,** 603-613.

Gonçalves, R., Rosa, A., Freitas, A., Fernandes, A., Kivisild, T., Villems, R. and Brehm, A., 2003. Y-chromosome lineages in Cabo Verde Islands witness the diverse geographic origin of its first male settlers. *Human genetics,* 113(6)**,** 467-472.

Gonçalves, R., Freitas, A., Branco, M., Rosa, A., Fernandes, A.T., Zhivotovsky, L.A., Underhill, P.A., Kivisild, T. and Brehm, A., 2005. Y-chromosome lineages from Portugal, Madeira and Açores record elements of Sephardim and Berber ancestry. *Annals of human genetics,* 69(4)**,** 443-454.

Gonçalves, R., Spínola, H. and Brehm, A., 2007. Y-chromosome lineages in São Tomé e Príncipe islands: evidence of European influence. *American journal of human biology : the official journal of the Human Biology Council,* 19(3)**,** 422-428.

Haber, M., Platt, D.E., Badro, D.A., Xue, Y., El-Sibai, M., Bonab, M.A., Youhanna, S.C., Saade, S., Soria-Hernanz, D.F., Royyuru, A., Wells, R.S., Tyler-Smith, C. and Zalloua, P.A., 2011. Influences of history, geography, and religion on genetic structure: the Maronites in Lebanon. *European journal of human genetics : EJHG,* 19(3)**,** 334-340.

Hammer, M.F., Karafet, T.M., Redd, A.J., Jarjanazi, H., Santachiara-Benerecetti, S., Soodyall, H. and Zegura, S.L., 2001. Hierarchical patterns of global human Y-chromosome diversity. *Molecular biology and evolution,* 18(7)**,** 1189-1203.

Hammer, M.F., Blackmer, F., Garrigan, D., Nachman, M.W. and Wilder, J.A., 2003. Human population structure and its effects on sampling Y chromosome sequence variation. *Genetics*, 164(4), 1495-1509.

Hammer, Ø., Harper, D.A.T. and Ryan, P.D., 2001. PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia electronica,* 4(1)**,** 9 pp.

Hassan, H.Y., Underhill, P.A., Cavalli-Sforza, L.L. and Ibrahim, M.E., 2008. Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history. *American journal of physical anthropology*, 137(3): 316-323.

Henn, B.M., Gignoux, C., Lin, A.A., Oefner, P.J., Shen, P., Scozzari, R., Cruciani, F., Tishkoff, S.A., Mountain, J.L. and Underhill, P.A., 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proceedings of the National Academy of Sciences of the United States of America,* 105(31)**,** 10693-10698.

Hurles, M.E. and Jobling, M.A., 2001. Haploid chromosomes in molecular ecology: lessons from the human Y. *Molecular ecology,* 10(7)**,** 1599-1613.

Jobling, M.A. and Tyler-Smith, C., 2003. The human Y chromosome: an evolutionary marker comes of age. *Nature reviews genetics,* 4(8)**,** 598-612.

Karafet, T.M., Mendez, F.L., Meilerman, M.B., Underhill, P.A., Zegura, S.L. and Hammer, M.F., 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome research,* 18(5)**,** 830-838.

Kayser, M., Caglià, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C. and Roewer, L., 1997. Evaluation of Y-chromosomal STRs: a multicenter study. *International journal of legal medicine,* 110(3)**,** 125-133, 141-149.

Kayser, M., Brauer, S., Weiss, G., Underhill, P.A., Roewer, L., Schiefenhövel, W. and Stoneking, M., 2000. Melanesian origin of Polynesian Y chromosomes. *Current biology : CB,* 10(20)**,** 1237-1246.

Knight, A., Underhill, P.A., Mortensen, H.M., Zhivotovsky, L.A., Lin. A.A., Henn, B.M., Louis, D., Ruhlen, M. and Mountain, J.L., 2003. African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Current biology*, 13(6):464-473.

Konrad, K.D. and Pentoney, S.L. Jr., 1993. Contribution of secondary structure to DNA mobility in capillary gels. *Electrophoresis*, 14(5-6), 502-508.

Lawson, D.J., Hellenthal, G., Myers, S. and Falush, D., 2012. Inference of population structure using dense haplotype data. *PLoS genetics,* 8(1)**,** e1002453.

Luis, J.R., Rowold, D.J., Regueiro, M., Caeiro, B., Cinnioğlu, C., Roseman, C., Underhill, P.A., Cavalli-Sforza, L.L. and Herrera, R.J., 2004. The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *American journal of human genetics*, 74(3):532-544.

Mendez, F.L., Krahn, T., Schrack, B., Krahn, A.-M., Veeramah, K.R., Woerner, A.E., Fomine, F.L.M., Bradman, N., Thomas, M.G., Karafet, T.M. and Hammer, M.F., 2013. An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *American journal of human genetics,* 92(3)**,** 454-459.

Miller, S.A., Dykes, D.D. and Polesky, H.F., 1988. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids research,* 16(3)**,** 1215.

Moran, C.N., Scott, R.A., Adams, S.M., Warrington, S.J., Jobling, M.A., Wilson, R.H., Goodwin, W.H., Georgiades, E., Wolde, B. and Pitsiladis, Y.P., 2004. Y chromosome haplogroups of elite Ethiopian endurance runners. *Human Genetics*, 115(6): 492-497.

Myres, N.M., Rootsi, S., Lin, A.A., Järve, M., King, R.J., Kutuev, I., Cabrera, V.M., Khusnutdinova, E.K., Pshenichnov, A., Yunusbayev, B., Balanovsky, O., Balanovska, E., Rudan, P., Baldovic, M., Herrera, R.J., Chiaroni, J., Di Cristofaro, J., Villems, R., Kivisild, T. and Underhill, P.A., 2011. A major Y-chromosome haplogroup R1b Holocene era founder effect in Central and Western Europe. *European journal of human genetics : EJHG,* 19(1)**,** 95-101.

Naidoo, T., Schlebusch, C.M., Makkan, H., Patel, P., Mahabeer, R., Erasmus, J.C. and Soodyall, H., 2010. Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. *Investigative genetics,* 1(1)**,** 6.

Nurse, G., Weiner, J. and Jenkins, T., 1985. *The Peoples of Southern Africa and their Affinities.* New York: Oxford University Press.

Nurse, G.T. and Jenkins, T., 1977. Serogenetic studies on the Kavango peoples of South West Africa. *Annals of human biology,* 4**,** 465-478.

Patin, E., Laval, G., Barreiro, L.B., Salas, A., Semino, O., Santachiara-Benerecetti, S., Kidd, K.K., Kidd, J.R., Van Der Veen, L., Hombert, J.-M., Gessain, A., Froment, A., Bahuchet, S., Heyer, E. and Quintana-Murci, L., 2009. Inferring the demographic history of African farmers and pygmy hunter-gatherers using a multilocus resequencing data set. *PLoS genetics,* 5(4)**,** e1000448.

Polzin, T. and Daneshmand, S.V., 2003. On Steiner trees and minimum spanning trees in hypergraphs. *Operations research letters,* 31(1)**,** 12-20.

Qamar, R., Ayub, Q., Mohyuddin, A., Helgason, A., Mazhar, K., Mansoor, A., Zerjal, T., Tyler-Smith, C. and Mehdi, S.Q., 2002. Y-chromosomal DNA variation in Pakistan. *American journal of human genetics,* 70(5)**,** 1107-1124.

Quintana-Murci, L., Harmant, C., Quach, H., Balanovsky, O., Zaporozhchenko, V., Bormans, C., Van Helden, P.D., Hoal, E.G. and Behar, D.M., 2010. Strong maternal Khoisan contribution to the South African Coloured population: a case of gender-biased admixture. *American journal of human genetics,* 86(4)**,** 611-620.

Raymond, M. and Rousset, F., 1995. An exact test for population differentiation. *Evolution,* 49(6)**,** 1280-1283.

Rosa, A., Ornelas, C., Jobling, M.A., Brehm, A. and Villems, R., 2007. Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC evolutionary biology,* 7**,** 124.

Rosser, Z.H., Zerjal, T., Hurles, M.E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., Barbujani, G., Beckman, G., Beckman, L., Bertranpetit, J., Bosch, E., Bradley, D.G., Brede, G., Cooper, G., Côrte-Real, H.B., De Knijff, P., Decorte, R., Dubrova, Y.E., Evgrafov, O., Gilissen, A., Glisic, S., Gölge, M., Hill, E.W., Jeziorowska, A., Kalaydjieva, L., Kayser, M., Kivisild, T., Kravchenko, S.A., Krumina, A., Kucinskas, V., Lavinha, J., Livshits, L.A., Malaspina, P., Maria, S., Mcelreavey, K., Meitinger, T.A., Mikelsaar, A.V., Mitchell, R.J., Nafa, K., Nicholson, J., Nørby, S., Pandya, A., Parik, J., Patsalis, P.C., Pereira, L., Peterlin, B., Pielberg, G., Prata, M.J., Previderé, C., Roewer, L., Rootsi, S., Rubinsztein, D.C., Saillard, J., Santos, F.R., Stefanescu, G., Sykes, B.C., Tolun, A., Villems, R., Tyler-Smith, C. and Jobling, M.A., 2000. Y-chromosomal

diversity in Europe is clinal and influenced primarily by geography, rather than by language. *American journal of human genetics,* 67(6)**,** 1526-1543.

Rozen, S. and Skaletsky, H., 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods in molecular biology*, 132, 365-386.

Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S., Trivedi, R., Endicott, P., Kivisild, T., Metspalu, M., Villems, R. and Kashyap, V.K., 2006. A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proceedings of the National Academy of Sciences of the United States of America,* 103(4)**,** 843-848.

Salas, A., Bandelt, H.-J., Macaulay, V. and Richards, M.B., 2007. Phylogeographic investigations: the role of trees in forensic genetics. *Forensic science international,* 168(1), 1-13.

Schlebusch, C.M., Naidoo, T. and Soodyall, H., 2009. SNaPshot minisequencing to resolve mitochondrial macro-haplogroups found in Africa. *Electrophoresis,* 30(21)**,** 3657-3664.

Schlebusch, C.M., 2010. *Genetic variation in Khoisan-speaking populations from southern Africa.* Ph.D. University of the Witwatersrand.

Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G.B., Soodyall, H. and Jakobsson, M., 2012. Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. *Science (New York, N.Y.),* 338(6105)**,** 374-379.

Schlebusch, C.M., Lombard, M. and Soodyall, H., 2013. MtDNA control region variation affirms diversity and deep sub-structure in populations from southern Africa. *BMC evolutionary biology,* 13**,** 56.

Scozzari, R., Cruciani, F., Malaspina, P., Santolamazza, P., Ciminelli, B.M., Torroni, A., Modiano, D., Wallace, D.C., Kidd, K.K., Olckers, A., Moral, P., Terrenato, L., Akar, N., Qamar, R., Mansoor, A., Mehdi, S.Q., Meloni, G., Vona, G., Cole, D.E., Cai, W. and Novelletto, A., 1997. Differential structuring of human populations for homologous X and Y microsatellite loci. *American journal of human genetics,* 61(3)**,** 719-733.

Scozzari, R., Massaia, A., D'atanasio, E., Myres, N.M., Perego, U.A., Trombetta, B. and Cruciani, F., 2012. Molecular dissection of the basal clades in the human Y chromosome phylogenetic tree. *PloS one,* 7(11)**,** e49170.

Seielstad, M.T., Minch, E. and Cavalli-Sforza, L.L., 1998. Genetic evidence for a higher female migration rate in humans. *Nature genetics,* 20(3)**,** 278-280.

Semino, O., Santachiara-Benerecetti, A.S., Falaschi, F., Cavalli-Sforza, L.L. and Underhill, P.A., 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *American journal of human genetics,* 70(1)**,** 265-268.

Semino, O., Magri, C., Benuzzi, G., Lin, A.A., Al-Zahery, N., Battaglia, V., Maccioni, L., Triantaphyllidis, C., Shen, P., Oefner, P.J., Zhivotovsky, L.A., King, R., Torroni, A., Cavalli-Sforza, L.L., Underhill, P.A. and Santachiara-Benerecetti, A.S., 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. *American journal of human genetics,* 74(5)**,** 1023-1034.

Shen, P., Wang, F., Underhill, P.A., Franco, C., Yang, W.H., Roxas, A., Sung, R., Lin, A.A., Hyman, R.W., Vollrath, D., Davis, R.W., Cavalli-Sforza, L.L. and Oefner, P.J., 2000. Population genetic implications from sequence variation in four Y chromosome genes. *Proceedings of the National Academy of Sciences of the United States of America,* 97(13)**,** 7354-7359.

Shi, H., Zhong, H., Peng, Y., Dong, Y.-L., Qi, X.-B., Zhang, F., Liu, L.-F., Tan, S.-J., Ma, R.Z., Xiao, C.-J., Wells, R.S., Jin, L. and Su, B., 2008. Y chromosome evidence of earliest modern human settlement in East Asia and multiple origins of Tibetan and Japanese populations. *BMC Biology,* 6**,** 45.

Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics,* 139(1)**,** 457-462.

Soares, P., Ermini, L., Thomson, N., Mormina, M., Rito, T., Rohl, A., Salas, A., Oppenheimer, S., Macaulay, V. and Richards, M.B., 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. *American journal of human genetics,* 84(6)**,** 740-759.

Sobrino, B., Brión, M. and Carracedo, A., 2005. SNPs in forensic genetics: a review on SNP typing methodologies. *Forensic science international,* 154(2-3)**,** 181-194.

Syvänen, A.C., 1999. From gels to chips: "minisequencing" primer extension for analysis of point mutations and single nucleotide polymorphisms. *Human mutation,* 13(1)**,** 1-10.

Tishkoff, S.A., Gonder, M.K., Henn, B.M., Mortensen, H., Knight, A., Gignoux, C., Fernandopulle, N., Lema, G., Nyambo, T.B., Ramakrishnan, U., Reed, F.A. and Mountain, J.L., 2007. History of click-speaking populations of Africa inferred from

mtDNA and Y chromosome genetic variation. *Molecular biology and evolution*, 24(10):2180-2195.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.-M., Doumbo, O., Ibrahim, M., Juma, A.T., Kotze, M.J., Lema, G., Moore, J.H., Mortensen, H., Nyambo, T.B., Omar, S.A., Powell, K., Pretorius, G.S., Smith, M.W., Thera, M.A., Wambebe, C., Weber, J.L. and Williams, S.M., 2009. The genetic structure and history of Africans and African Americans. *Science (New York, N.Y.),* 324(5930)**,** 1035-1044.

Trombetta, B., Cruciani, F., Sellitto, D. and Scozzari, R., 2011. A new topology of the human Y chromosome haplogroup E1b1 (E-P2) revealed through the use of newly characterized binary polymorphisms. *PLoS one,* 6(1)**,** e16073.

Tu, O., Knott, T., Marsh, M., Bechtol, K., Harris, D., Barker, D. and Bashkin, J., 1998. The influence of fluorescent dye structure on the electrophoretic mobility of end-labeled DNA. *Nucleic Acids Research*, 26(11), 2797-2802.

Underhill, P.A., Jin, L., Lin, A.A., Mehdi, S.Q., Jenkins, T., Vollrath, D., Davis, R.W., Cavalli-Sforza, L.L. and Oefner, P.J., 1997. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome research,* 7(10)**,** 996-1005.

Underhill, P.A., Shen, P., Lin, A.A., Jin, L., Passarino, G., Yang, W.H., Kauffman, E., Bonne-Tamir, B., Bertranpetit, J., Francalacci, P., Ibrahim, M., Jenkins, T., Kidd, J.R., Mehdi, S.Q., Seielstad, M.T., Wells, R.S., Piazza, A., Davis, R.W., Feldman, M.W., Cavalli-Sforza, L.L. and Oefner, P.J., 2000. Y chromosome sequence variation and the history of human populations. *Nature genetics,* 26(3)**,** 358-361.

Underhill, P.A., Passarino, G., Lin, A.A., Shen, P., Mirazón Lahr, M., Foley, R.A., Oefner, P.J. and Cavalli-Sforza, L.L., 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Annals of human genetics,* 65(1)**,** 43-62.

Underhill, P.A. and Kivisild, T., 2007. Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Annual review of genetics,* 41**,** 539-564.

Underhill, P.A., Myres, N.M., Rootsi, S., Metspalu, M., Zhivotovsky, L.A., King, R.J., Lin, A.A., Chow, C.-E.T., Semino, O., Battaglia, V., Kutuev, I., Järve, M., Chaubey, G., Ayub, Q., Mohyuddin, A., Mehdi, S.Q., Sengupta, S., Rogaev, E.I., Khusnutdinova, E.K., Pshenichnov, A., Balanovsky, O., Balanovska, E., Jeran, N.,

Augustin, D.H., Baldovic, M., Herrera, R.J., Thangaraj, K., Singh, V., Singh, L., Majumder, P., Rudan, P., Primorac, D., Villems, R. and Kivisild, T., 2010. Separating the post-Glacial coancestry of European and Asian Y chromosomes within haplogroup R1a. *European journal of human genetics : EJHG,* 18(4)**,** 479-484.

Vallone, P.M. and Butler, J.M., 2004. AutoDimer: a screening tool for primer-dimer and hairpin structures. *Biotechniques*, 37, 226-231.

Weale, M.E., Shah, T., Jones, A.L., Greenhalgh, J., Wilson, J.F., Nymadawa, P., Zeitlin, D., Connell, B.A., Bradman, N. and Thomas, M.G., 2003. Rare deep-rooting Y chromosome lineages in humans: lessons for phylogeography. *Genetics,* 165(1)**,** 229-234.

Wilson, I.J., Weale, M.E. and Balding, D.J., 2003. Inferences from DNA data: population histories, evolutionary processes and forensic match probabilities. *Journal of the Royal Statistical Society: series A (statistics in society),* 166(2)**,** 155-188.

Winkler, C.A., Nelson, G.W. and Smith, M.W., 2010. Admixture mapping comes of age. *Annual review of genomics and human genetics,* 11**,** 65-89.

Wood, E.T., Stover, D.A., Ehret, C., Destro-Bisol, G., Spedini, G., Mcleod, H., Louie, L., Bamshad, M., Strassmann, B.I., Soodyall, H. and Hammer, M.F., 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *European journal of human genetics : EJHG,* 13(7)**,** 867-876.

Xue, Y., Zerjal, T., Bao, W., Zhu, S., Shu, Q., Xu, J., Du, R., Fu, S., Li, P., Hurles, M.E.,

Yang, H. and Tyler-Smith, C., 2006. Male demography in East Asia: a north-south contrast in human population expansion times. *Genetics*, 172(4):2431-2439.

Y Chromosome Consortium, 2002. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome research,* 12(2)**,** 339-348.

7 APPENDICES

7.1 Appendix A

ISOGG 2013 Y-DNA Haplogroup A and its subclades
http://www.isogg.org/tree/ISOGG_HapgrpA.html

Y-DNA Haplogroup A and its Subclades - 2013

The entire work is identified by the Version Number and date given on the Main Page. Directions for citing the document are given at the bottom of the Main Page.
Version History Last revision date for this specific page: 19 August 2013

Because of continuing research, the structure of the Y-DNA Haplogroup Tree changes and ISOGG does its best to keep the tree updated with the latest developments in the field. The viewer may observe other versions of the tree on the Web. Email Alice Fairhurst if the differences need clarification or if you find any broken links on this page.

LINKS: Main Page Y-DNA Tree Trunk SNP Index Papers/Presentations Cited Glossary Listing Criteria
CLADE/SUBCLADE SYMBOLS: Added Redefined
SNP SYMBOLS: Not on 2012 tree Confirmed within subclade *Provisional* *Private* *Investigation*

Y Root (Y-Adam)

- A00 AF4, AF5, AF6, AF7, AF8, AF9, AF10, AF13, L1086, L1087, L1088, L1091, L1092, L1094, L1096, L1097, L1102, L1103, L1104, L1106, L1107, L1108, L1109, L1110, L1111, L1113, L1114, L1115, L1117, L1119, L1122, L1126, L1131, L1133, L1134, L1138, L1139, L1140, L1141, L1144, L1146, L1147, L1148, L1151, L1152, L1154, L1156, L1157, L1158, L1159, L1160, L1161, L1163, L1233, L1234, L1236
- A0-T AF3, L1085, L1089, L1090, L1093, L1095, L1098, L1099, L1101, L1105, L1116, L1118, L1120, L1121, L1123, L1124, L1125, L1127, L1128, L1129, L1130, L1132, L1135, L1136, L1137, L1142, L1143, L1145, L1150, L1155, L1235
- • A0 L529.2, L896, L982, L984, L990, L991, L993, L995, L997, L998, L999, L1000, L1001, L1006, L1008, L1010, L1012, L1016, L1018, L1055, V148, V149, V154, V165, V166, V167, V172, V173, V176, V177, V190, V196, V223, V225, V229, V233, V239
- • • A0* -
- • • A0a L979, L980, L987, L996, L1011, L1015, L1017
- • • • A0a* -
- • • • A0a1 L1070, L1072, L1073, L1075, L1076, L1078, L1079, L1080, L1081, L1082, V150, V153, V157, V158, V159, V162, V164, V170
- • • • • A0a1* -
- • • • • A0a1a P114, V151, V152, V161.1, V169, V181, V195
- • • • • A0a1b L1289
- • • • A0a2 L981, L983, L988, L994, L1007, L1014, V203
- • • A0b L92.2, L1035, L1036, L1037, L1038, L1039, L1040, L1041, L1042, L1043, L1044, L1045, L1046, L1047, L1048, L1049, L1050, L1051, L1052, L1054, L1056, L1057, L1058
- • A1 L985, L986, L989, L1002, L1003, L1004, L1005, L1009, L1013, L1053, L1084, L1112, L1153, P305, V161.2, V168, V171, V174, V238, V241, V250
- • • A1* -
- • • A1a M31, P82, V4, V14, V15, V25, V26, V28, V30, V40, V48, V57, V58, V63, V191, V201, V204, V215
- • • A1b P108, V221

• • • • A1b* -
• • • • A1b1 L419/PF712
• • • • • A1b1* -
• • • • • A1b1a L602, V50, V82, V198, V224
• • • • • • A1b1a* -
• • • • • • A1b1a1 M14, M23, L968/M29/P3/PN3, M71, M135, M141, M206, M276/P247, M277/P248, MEH1, P4, P5, P36.1, Page71, Page87, Page95
• • • • • • • A1b1a1* -
• • • • • • • A1b1a1a M6, M49/Page41, M196
• • • • • • • • A1b1a1a* -
• • • • • • • • A1b1a1a1 P28
• • • • • • • • A1b1a1a2 L963
• • • • • • • • • A1b1a1a2* -
• • • • • • • • • A1b1a1a2a M114, M212
• • • • • • • • • A1b1a1a2b P262
• • • • • A1b1b M32
• • • • • • A1b1b* -
• • • • • • A1b1b1 M28
• • • • • • A1b1b2 L427, L430, M144, M190/Page35/PF1373, M220, M305/Page17, P289/PF1372, Page50, PF1365/V160
• • • • • • • A1b1b2* -
• • • • • • • A1b1b2a M51/Page42, M229, M239/Page89, P71, P100
• • • • • • • • A1b1b2a* -
• • • • • • • • A1b1b2a1 P291
• • • • • • • • • A1b1b2a1 -
• • • • • • • • • A1b1b2a1a P102
• • • • • • • A1b1b2b M13, M63, M127, M202, M219, *Page53, Page77/PF1364/V10*
• • • • • • • • A1b1b2b* -
• • • • • • • • A1b1b2b1 M118
• • • • BT L413/PF1409/V31, L418, L438, L440, L604/PF1243, L957, L962, L969, L970/PF1065, L971, L977, M42, M91, L1060/PF1021, L1061/PF1101, L1062/PF302, M94/PF1081, M139, M299, P97, Page65.1/SRY1532.1/SRY10831.1, V29/PF1408, V59/PF1411, V64/PF1412, V102/PF1406, V187/PF1403, V202, V235/PF1410

Private SNPs are being removed from the tree and placed in the following category:
Private SNPs - After having been investigated, these SNPs have not met the population distribution criteria for placement on the tree. Either too few confirmed positive testers have been found OR multiple confirmed testers were confined to either a single surname or to a small group of related males.

- *M59* found in the same sample from Ethiopia. Listed 15 February 2012.
- *M171* found in a single A3a2 (the former A3b2, positive for M13 or its equivalents) sample from

Sudan. Listed 15 February 2012.

- *L1100* and *L1149* are located downstream of L1086. Reported in Perry surname. Listed 5 March 2013.

SNPs under Investigation - Additional testing is needed to confirm adequate positive samples and/or correct placement on the tree.

- *L600, L601, L603* are downstream from V50. These SNPs were found in an A-V50 individual by Thomas Krahn. Listed 15 February 2012
- *V53, V214, V236*. These SNPs were found in an A-M31+ sample by Cruciani (2011). Listed 15 February 2012
- *Page110* is downstream of Page41, Page71, Page87, and Page95, but it needs to be tested to find out where it belongs in relation to other SNPs. It may be private. Listed 15 February 2012
- *V60, V61, V70, V72, V79, V80, V81, V180, V188, V192, V200, V218.1, V228, V242*. These SNPs were found in an A-P262 sample by Cruciani (2011). Listed 15 February 2012
- *V97* is found dowstream of M212. This SNP was found in an A-M114 sample by Trombetta (2010) and has been found derived in A-P262+ sample. Listed 15 February 2012
- *V1, V51/PF1369, V56/PF1371, V66/PF1370, V155, V156/PF1362, V193/PF1358, V194/PF1368, V230, V243/PF1367*. These SNPs were found in an A-M13* sample by Cruciani (2011). Listed 15 February 2012
- *V89/PF1359, V98*. These SNPs were found in an A-M13* by Trombetta (2010). Listed 15 February 2012
- *V67* is found downstream of V161,2. Listed 15 February 2012.
- *L411/PF11, L412, L414, L420, L421, L422, L423, L424, L425, L426, L428, L429, L436/PF962, L437/PF965, L439, L441, L442*. These SNPs were found in an A-M13 WTY participant. They are being tested in more samples to determine where they belong in the tree. Listed 15 February 2012
- *Page52.3* has been found derived in the current hg A-P262 WTY sample from the U of A. It is ancestral in A-L896 and in A-M13. Listed 15 February 2012.
- *PK1* may be philogenically equivalent to BT. Listed 10 July 2012.
- *Page10* may be philogenically equivalent to M144. Listed 10 July 2012.
- *V139.1* and *V141* are located at approximately M14. Listed 26 February 2013.
- *L949, L950, L951, L952, L954, L955, L956, L958, L959, L961, L964, L965, L966, L973, L974, L975L976* are located at approximately P262. Listed 26 February 2013.
- *PF825.1* is located at approximately V148. Listed 26 June 2013.

NOTES:

- Identical SNPs that were discovered separately are listed in alphabetical order, not necessarily in the order of discovery, and separated by "/". Examples: M29/P3/PN3, M276/P247 and M277/P248.
- PK1 (formerly A2) and M63 (formerly A3b2) were removed from the tree due to the uncertainty of

their placement.

- A0 is being used to emphasize that this constitutes a haplogroup distinct from A1 and downstream clades, although for practical reasons, the letter A will still be used in its name, following the precedent of the addition of L0 and its subclades to the mtDNA phylogenetic tree. Listed 15 February 2012.
- There is some uncertainty about the position of M14, M29/P3/PN3, M71, M135, M141, M206, M276/P247, M277/P248, MEH1, P4, P5, P36.1, Page71, Page87, and Page95, since although they have been considered equivalent to M23 in the past, they have not yet been tested in the A2* sample in which V50, V82, etc, are derived and M23 is ancestral. Listed 15 February 2012.
- BT is shown on this tree, though it is not considered to be a part of Haplogroup A, in order to make it clear that, as a sibling clade with A1b1 and A1b2, BT and all other haplogroups are downstream of A1b. Listed 15 February 2012.
- The naming conventions of the subclades at the beginning of the tree come from the supplement to Mendez et al (2013). Listed 5 March 2013.

Y-DNA haplogroup A contains lineages deriving from the earliest branching in the human Y chromosome tree. The oldest branching event, separating A0-P305 and A1-V161, is thought to have occurred about 140,000 years ago. Haplogroups A0-P305, A1a-M31 and A1b1a-M14 are restricted to Africa and A1b1b-M32 is nearly restricted to Africa. The haplogroup that would be named A1b2 is composed of haplogroups B through T. The internal branching of haplogroup A1-V161 into A1a-M31, A1b1, and BT (A1b2) may have occurred about 110,000 years ago. A0-P305 is found at low frequency in Central and West Africa. A1a-M31 is observed in northwestern Africans; A1b1a-M14 is seen among click language-speaking Khoisan populations. A1b1b-M32 has a wide distribution including Khoisan speaking and East African populations, and scattered members on the Arabian Peninsula.

References:

Batini et al, Signatures of the Preagricultural Peopling Processes in Sub-Saharan Africa as Revealed by the Phylogeography of Early Y Chromosome Lineages. (abstract) Molecular Biology and Evolution. 28 (9): 2603-2613, 2011.

Behar et al, Genome-Wide Structure of the Jewish People. Nature, 446:238-42, 2010.

Cinnioglu et al, Excavating Y-chromosome Haplotype Strata in Anatolia. (pdf) Human Genetics. 114:127-148, 2004.

Cruciani et al, A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes. American Journal of Human Genetics, 70:1197-1214, 2002.

Cruciani et al, A Revised Root for the Human Y Chromosomal Phylogenetic Tree--The Origin of Patrilineal Diversity in Africa. The American Journal of Human Genetics, doi:10.1016/j.ajhg.2011.05.002, 2011.

Deng et al, Evolution and Migration History of the Chinese Population Inferred from the Chinese Y-chromosome Evidence. (pdf) Journal of Human Genetics, 49:339-348, 2004.

Francalacci et al, Low-Pass DNA Sequencing of 1200 Sardinians Reconstructs European Y-Chromosome

Phylogeny. Science: Vol. 341 no. 6145, pp. 565-569, DOI: 10.1126/science.1237947, 2 August 2013.
Karafet et al, New Binary Polymorphisms Reshape and Increase Resolution of the Human Y-Chromosomal Haplogroup Tree. Abstract. Genome Research, published online April 2, 2008. Supplementary Material.
Mendez et al, An African American Paternal Lineage Adds an Extremely Ancient Root to the Human Y Chromosome Phylogenetic Tree. (abstract) American Journal of Human Genetics 28 February 2013.
Mohyuddin et al, Detection of Novel Y SNPs Provides Further Insights into Y Chromosomal Variation in Pakistan. Journal of Human Genetics, 2006.
Rozen et al, Remarkably Little Variation in Proteins Encoded by the Y Chromosome's Single-Copy Genes, Implying Effective Purifying Selection. American Journal of Human Genetics. 2009 December 11; 85(6): 923-928.
Schuster et al, Complete Khoisan and Bantu Genomes from Southern Africa. (abstract) Nature 463, 943-947, 18 February 2010.
Semino et al, Ethiopians and Khoisan Share the Deepest Clades of the Human Y-Chromosome Phylogeny. (pdf) American Journal of Human Genetics, 70:265-268, 2002.
Shen et al, Reconstruction of Patrilineages and Matrilineages of Samaritans and other Israeli Populations from Y-Chromosome and Mitochondrial DNA Sequence Variation. (pdf) Human Mutation, 24:248-260, 2004.
Valone et al, Y SNP Typing of African-American and Caucasian Samples Using Allele-Specific Hybridization and Primer Extension. (pdf) Journal of Forensic Science, 49:4, July 2004.

Additional Resources:
ISOGG Wiki - What you need to know about Genetic Genealogy.
Y Haplogroup A, Bonnie Schrack.
The African DNA Project (A), Dr. Ana Oquendo Pabon.

Corrections/Additions made since 1 January 2013:

- Added PF712, PF1081, PF1364, PF1365, PF1372, PF1373, PF1403, PF1408, PF1411, PF1412 to tree; added PF11, PF962, PF965, PF1358, PF1359, PF1362, PF1367, PF1368, PF1369, PF1370, PF1371 to Investigation on 4 January 2013.
- Added L963, L1060/PF1021, L1061/PF1101, L1062/PF302, M229, M239/Page89, V102/PF1406 to tree; moved L427, L430, L602, PF1364/V10 from Investigation to tree on 25 February 2013.
- Added L949, L950, L951, L952, L954, L955, L956, L958, L959, L961, L964, L965, L966, L973, L974, L975, L976, V139.1, V141 to Investigation on 26 February 2013.
- Added PF1409; added Mendez et al (2013) on 1 March 2013.
- Removed Ana Oquendo Pabon as a contact for Haplogroup A on 3 March 2013.
- Added .2 to V161; added note on naming conventions on 5 March 2013.
- Added AF3, AF4, AF5, AF6, AF7, AF8, AF9, AF10, AF13 on 5 March 2013.
- Moved L979, L980, L981, L983, L987, L988, L994, L1007, L1011, L1014, L1015, L1017 from Investigation to tree on 5 March 2013.
- Moved V15, V26, V28, V30, V40, V48, V57, V58, V63, V150, V151, V153, V157, V158, V159,

V162, V169, V170, V181, V191, V195, V201, V204, V215 from Investigation to tree on 5 March 2013.

- Added L529.2, L896, L982, L984, L986, L990, L991, L993, L995, L996, L997, L998, L999, L1000, L1001, L1005, L1006, L1008, L1010, L1012, L1016, L1018, L1055, L1084, L1085, L1086, L1087, L1088, L1089, L1090, L1091, L1092, L1093, L1094, L1095, L1096, L1097, L1098, L1099, L1101, L1102, L1103, L1104, L1105, L1106, L1107, L1108, L1109, L1110, L1111, L1112, L1113, L1114, L1115, L1116, L1117, L1118, L1119, L1120, L1121, L1122, L1123, L1124, L1125, L1126, L1127, L1128, L1129, L1130, L1131, L1132, L1133, L1134, L1135, L1136, L1137, L1138, L1139, L1140, L1141, L1142, L1143, L1144, L1145, L1146, L1147, L1148, L1150, L1151, L1152, L1153, L1154, L1155, L1156, L1157, L1158, L1159, L1160, L1161, L1163, L1233, L1234, L1235, L1236 on 5 March 2013.
- Added L92.2, L996, L1035, L1036, L1037, L1038, L1039, L1040, L1041, L1042, L1043, L1044, L1045, L1046, L1047, L1048, L1049, L1050, L1051, L1052, L1054, L1056, L1057, L1058 L1070, L1072, L1073, L1075, L1076, L1078, L1079, L1080, L1081, L1082, L1289, V161.1, V164 on 5 March 2013.
- Added L1100, Ll149 to Private on 5 March 2013.
- Added PF1065, PF1243, PF1410 on 8 March 2013.
- Added Batini et al (2011) on 22 March 2013.
- Added Schuster et al on 27 March 2013.
- Added PF825.1 to Investigation on 26 June 2013.
- Added Francalacci et al (2013) on 19 August 2013.

Contact Person for Haplogroup A: Bonnie Schrack.

Back to **Main Page**
Back to **Y-DNA Tree Trunk**
Back to **SNP Index**
Back to **Papers/Presentations Cited**
Back to **Glossary**
Back to **Listing Criteria**

Copyright 2013, International Society of Genetic Genealogy. All Rights Reserved.

7.2 Appendix B

ISOGG 2013 Y-DNA Haplogroup B and its subclades
http://www.isogg.org/tree/ISOGG_HapgrpB.html

Y-DNA Haplogroup B and its Subclades - 2013

The entire work is identified by the Version Number and date given on the Main Page. Directions for citing the document are given at the bottom of the Main Page.

Version History Last revision date for this specific page: 3 April 2013

Because of continuing research, the structure of the Y-DNA Haplogroup Tree changes and ISOGG does its best to keep the tree updated with the latest developments in the field. The viewer may observe other versions of the tree on the Web. Email Alice Fairhurst if the differences need clarification or if you find any broken links on this page.

LINKS: Main Page Y-DNA Tree Trunk SNP Index Papers/Presentations Cited Glossary Listing Criteria

CLADE/SUBCLADE SYMBOLS: Added Redefined

SNP SYMBOLS: Not on 2012 tree Confirmed within subclade *Provisional* *Private* *Investigation*

- **B** M60, M181/Page32, M247/P85, P90
  - **B*** -
  - **B1** M236, M288
    - **B1*** -
    - **B1a** M146
  - **B2** M182
    - **B2*** -
    - **B2a** M150, *Page18*
      - **B2a*** -
      - **B2a1** M218
        - **B2a1*** -
        - **B2a1a** M109, M152/Page60, P32, P50
        - **B2a1b** G1
      - **B2a2** M108.1
        - **B2a2*** -
        - **B2a2a** M43, P111
    - **B2b** M112
      - **B2b*** -
      - **B2b1** M192, 50f2(P)
        - **B2b1*** -
        - **B2b1a** P7_1, P7_2, P7_3
          - **B2b1a*** -
          - **B2b1a1** MSY2.1
            - **B2b1a1*** -
            - **B2b1a1a** M115, M169
            - **B2b1a1b** M30, M129
              - **B2b1a1b*** -
              - **B2b1a1b1** M108.2
            - **B2b1a1c** M211
          - **B2b1a2** P8_1, P8_2, P8_3, P70

• • • • **B2b1b** P6
• • • **B2b2** P112, V341

SNPs under Investigation - Additional testing is needed to confirm adequate positive samples and/or correct placement on the tree.

- *V244* is located at approximately B-M60. Listed 8 March 2013.
- *V85, V90, V94, V220, V234* and *V237* are located at approximately B-M182. Listed 8 March 2013.
- *V93* is located at approximately B-M150 and upstream from V75 and V197. Listed 8 March 2013.
- *V78* is located at approximately downstream from B-M150 and upstream from B-V75 and B-V197. Listed 8 March 2013.
- *V75* and *V197* are located at approximately downstream from B-V78 but upstream from B-M218 and B-M108.1. Listed 8 March 2013.
- *V62* and *V227* are located at approximately downstream from B-M218 but upstream to B-V254 and B-M109. Listed 8 March 2013.
- *V254* is located at approximately downstream from B-V62 and B-V227. Listed 8 March 2013.
- *V84* and *V217* are located at approximately at B-M109. Listed 8 March 2013.
- *Page86* is located between M112 and M211. Listed 22 March 2013.
- *Page72* is located at approximately B-M112. Listed 22 March 2013.

NOTES:

- Identical SNPs that were discovered separately are listed in alphabetical order, not necessarily in the order of discovery, and separated by "/". Examples: P257/U6, L31/S149.
- B-P50 is possibly a downstream subclade of B-M109 with equivalent subclade SNPs V83 and V185. Listed 8 March 2013.
- The position of G1 relative to V62 and V227 is uncertain. Listed 8 March 2013.

Y-DNA haplogroup B, like Y-DNA haplogroup A, is seen only in Africa and is scattered widely, but thinly across the continent. B is thought to have arisen approximately 50,000 years ago. These haplogroups have higher frequencies among hunter-gather groups in Ethiopia and Sudan, and are also seen among click language-speaking populations. The patchy, widespread distribution of these haplogroups may mean that they are remnants of ancient lineages that once had a much wider range but have been largely displaced by more recent population events.

Some geographic structuring is seen between the sub-groups B2a (B-M150) and B2b (B-M112). Sub-group B2b is seen among Central African Pygmies and South African Khoisan. Sub-group B2a is seen among Cameroonians, East Africans, and among South African Bantu speakers. B2a1a (B-M109) is the most commonly seen sub-group of B2a. About 2.3% of African-Americans belong to haplogroup B - with 1.5% of them belonging to the sub-group B2a1a.

References:

Batini et al, Signatures of the Preagricultural Peopling Processes in Sub-Saharan Africa as Revealed by the Phylogeography of Early Y Chromosome Lineages. (abstract) Molecular Biology and Evolution. 28 (9): 2603-2613, 2011.
Cruciani et al, A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes. American Journal of Human Genetics, 70:1197-1214, 2002.
Gomes et al, Digging Deeper into East African Human Y Chromosone Lineages, Human Genetics, 127(5):603-13, 2010.
Karafet et al, New Binary Polymorphisms Reshape and Increase Resolution of the Human Y-Chromosomal Haplogroup Tree. Abstract. Genome Research, published online April 2, 2008. Supplementary Material.
Regueiro et al, Iran: Tricontinental Nexus for Y-Chromosome Driven Migration. (abstract) Human Heredity, Vol. 61, No 3, 132-143, 2006.
Rozen et al, Remarkably Little Variation in Proteins Encoded by the Y Chromosome's Single-Copy Genes, Implying Effective Purifying Selection. American Journal of Human Genetics. 2009 December 11; 85(6): 923-928.
Schuster et al, Complete Khoisan and Bantu Genomes from Southern Africa. (abstract) Nature 463, 943-947, 18 February 2010.
Scozzari et al, Molecular Dissection of the Basal Clades in the Human Y Chromosome Phylogenetic Tree, PLoS ONE, Vol. 7, Issue 11, e4917, 2012.
Semino et al, Ethiopians and Khoisan Share the Deepest Clades of the Human Y-Chromosome Phylogeny. (pdf) American Journal of Human Genetics, 70:265-268, 2002.
Valone et al, Y SNP Typing of African-American and Caucasian Samples Using Allele-Specific Hybridization and Primer Extension. (pdf) Journal of Forensic Science, 49:4, July 2004.

Additional Resources:
ISOGG Wiki - What you need to know about Genetic Genealogy.
B Haplogroup Project .
The African DNA Project (B), Dr. Ana Oquendo Pabon.

Corrections/Additions made since 1 January 2013:

- Added Aaron R. Brown as contact person and Scozzari et al (2012) on 3 March 2013.
- Added V62, V75, V78, V84, V85, V90, V93, V94, V197, V217, V220, V227, V234, V237, V244, V254, V341 to Investigation; added three (3) Notes on 8 March 2013.
- Added Batini et al (2011) on 22 March 2013.
- Moved V341 from Investigation to tree; removed Page72, Page86 from tree and added to Notes on 22 March 2013.
- Moved Page72, Page86 from Notes to Investigation; removed obsolete note regarding 50f2(P) and M192 on 24 March 2013.
- Added Schuster et al on 27 March 2013.
- Corrected subclade indents on 3 April 2013.

Contact Peoplefor Haplogroup B: Aaron R. Brown and Ana Oquendo Pabon

Back to **Main Page**
Back to **Y-DNA Tree Trunk**
Back to **SNP Index**
Back to **Papers/Presentations Cited**
Back to **Glossary**
Back to **Listing Criteria**

Copyright 2013 International Society of Genetic Genealogy. All Rights Reserved.

7.3 Appendix C

Ethics Approval

UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG
Division of the Deputy Registrar (Research)

HUMAN RESEARCH ETHICS COMMITTEE (MEDICAL)
R14/49 Professor Himla Soodyall

CLEARANCE CERTIFICATE	**M090576**
PROJECT	Human Genetic Diversity and Disease (Previously M980553)
INVESTIGATORS	Professor Himla Soodyall.
DEPARTMENT	Diversity and Disease Research Unit
DATE CONSIDERED	09.05.29
DECISION OF THE COMMITTEE*	Annual Renewal Approved

Unless otherwise specified this ethical clearance is valid for 5 years and may be renewed upon application.

DATE 27/09/2009 **CHAIRPERSON** .....
(Professor PE Cleaton-Jones)

*Guidelines for written 'informed consent' attached where applicable
cc: Supervisor : Professor T Jenkins

DECLARATION OF INVESTIGATOR(S)

To be completed in duplicate and **ONE COPY** returned to the Secretary at Room 10004, 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

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

1 March 2007

Mr T Naidoo
Human Genetics
NHLS

Dear Mr Naidoo:

Re: Protocol M050906: Phylogeography of Y-ChromosomeHaplogroups A, B and E in Africa
Approval date: 05.10.23

According to our records, your ethics application **Protocol M050906** was approved. In order to comply with international requirements that Ethics Committees should monitor research, an annual statement from investigators is needed. For this we are using a simple report form based on that of the Medical Research Council Ethics Committee.

The Human Research Ethics Committee (Medical) therefore requests that you complete the attached report form on the abovementioned protocol and any others of yours approved during 2005, by no later than 28 March 2007. More than one protocol may be listed on the same form.

Should you have any problems regarding the above request, please inform Ms A Keshav (Committee Secretariat) in writing

Yours sincerely

Professor PE Cleaton-Jones
Chairman
Human Research Ethics Committee (Medical)

MRC/NHLS/WITS HUMAN GENOMIC
DIVERSITY AND DISEASE RESEARCH UNIT
(HGDDRU)

Division of Human Genetics, National Health Laboratory Service, P O Box 1038, Johannesburg, 2000
Room 149 Watkins Pitchford Building, Corner of Hospital and De Korte Streets, Braamfontein
Tel: (011) 489-9237 (Laboratory) Prof Himla Soodyall: (011) 489-9208 FAX: (011) 489-9226

22 February 2012

Professor P. Cleaton-Jones
Chairman
Committee for Research on Human Subjects

RE: RENEWAL OF ETHICS PROTOCOL M050906

Dear Professor Cleaton-Jones

I am writing to request your approval in renewing my ethics clearance certificate related to my previous application entitled "Phylogeography of Y-chromosome haplogroups A, B and E in Africa", Protocol Number M050906.

While the lab-work for the project has been completed, the dissertation is currently being finalized. It is required that a valid ethics clearance certificate be in place upon submission of the dissertation and any published work related to it.

Thanking you in advance for considering this request for renewal of my ethics clearance certificate.

Yours sincerely

Mr. Thijessen Naidoo
MRC/NHLS/Wits Human Genomic Diversity and Disease
Research Unit
Tel: 011 489 9237

and

Prof. Himla Soodyall
Director: MRC/NHLS/Wits Human Genomic Diversity and Disease
Research Unit
Tel: 011 489 9208

7.4 Appendix D

Table A1: Details of additional populations and comparative data examined in the present study

Country	Ethnicity	Language Group	n	Source
Benin	mixed Beninese population	-	125	Batini, et al. (2011a)
Cameroon	mixed Cameroonian	-	290	Batini, et al. (2011a)
Cameroon	Fali	Adamawa	35	Batini, et al. (2011a)
Cameroon	Fali	Adamawa	39	Cruciani, et al. (2002)
Cameroon	mixed Adamawa	Adamawa	18	Cruciani, et al. (2002)
Cameroon	Fulani (Fulbe)	Atlantic	17	Cruciani, et al. (2002)
Cameroon	mixed Bantu speakers	Bantu	14	Luis, et al. (2004)
Democratic Republic of Congo	Hema	Central Bantu	18	Wood, et al. (2005)
Rwanda	Hutu	Central Bantu	69	Luis, et al. (2004)
Rwanda	Tutsi	Central Bantu	94	Luis, et al. (2004)
Cameroon	Mandara	Chadic	30	Batini, et al. (2011a)
Cameroon	Mandara	Chadic	28	Wood, et al. (2005)
Cameroon	mixed Chadic	Chadic	15	Cruciani, et al. (2002)
Cameroon	Ouldeme	Chadic	21	Cruciani, et al. (2002)
Cameroon	Uldeme	Chadic	10	Batini, et al. (2011a)
Cameroon	Uldeme	Chadic	13	Wood, et al. (2005)
Cameroon	Tupuri	Gur	9	Batini, et al. (2011a)
Cameroon	Tupuri	Gur	9	Wood, et al. (2005)
Cameroon	Kanuri	Nilo-Saharan	12	Batini, et al. (2011a)
Cameroon	mixed Nilo-Saharan	Nilo-Saharan	9	Cruciani, et al. (2002)
Democratic Republic of Congo	Alur	Nilotic	9	Wood, et al. (2005)
Angola/South Africa	Khwe	Central Khoisan (Khoi)	26	Cruciani, et al. (2002)
Namibia	Nama	Central Khoisan (Khoi)	2	HGDDRL
South Africa	Nama	Central Khoisan (Khoi)	7	HGDDRL
Kenya	Kikuyu & Kamba	Central Bantu	42	Wood, et al. (2005)

Table A1 cont.: Details of additional populations and comparative data examined in the present study

Country	Ethnicity	Language Group	n	Source
Kenya	mixed Bantu speakers	Bantu	29	Luis, et al. (2004)
Tanzania	Mbugwe	Central Bantu	15	Tishkoff, et al. (2007)
Tanzania	Sukuma	Central Bantu	32	Knight, et al. (2003)
Tanzania	Sukuma	Central Bantu	30	Tishkoff, et al. (2007)
Tanzania	Swahili	Central Bantu	17	HGDDRL
Tanzania	Turu	Central Bantu	20	Tishkoff, et al. (2007)
Uganda	Ganda	Central Bantu	26	Wood, et al. (2005)
Sudan	Hausa	Chadic	32	Hassan, et al. (2008)
Ethiopia	Oromo	Cushitic	37	Batini, et al. (2011a)
Ethiopia	Oromo	Cushitic	78	Semino, et al. (2002)
Ethiopia	Oromo	Cushitic	9	Wood, et al. (2005)
Sudan	Beja	Cushitic	42	Hassan, et al. (2008)
Tanzania	Burunge	Cushitic	24	Tishkoff, et al. (2007)
Tanzania	WaFiome	Cushitic	2	Tishkoff, et al. (2007)
Tanzania	Wairak (Iraqw)	Cushitic	43	Luis, et al. (2004)
Tanzania	Wairak (Iraqw)	Cushitic	9	Wood, et al. (2005)
Kenya	Elmolo	Cushitic	23	Batini, et al. (2011a)
Kenya	Luo	Nilotic	9	Wood, et al. (2005)
Kenya	Maasai	Nilotic	81	Batini, et al. (2011a)
Kenya	Maasai	Nilotic	26	Wood, et al. (2005)
Kenya	Samburu	Nilotic	34	Batini, et al. (2011a)
Kenya	Turkana	Nilotic	53	Batini, et al. (2011a)
Sudan	Borgu	Nilotic	26	Hassan, et al. (2008)
Sudan	Dinka	Nilotic	26	Hassan, et al. (2008)
Sudan	Masalit	Nilotic	32	Hassan, et al. (2008)
Sudan	Nuer	Nilotic	12	Hassan, et al. (2008)
Sudan	Shilluk	Nilotic	15	Hassan, et al. (2008)
Tanzania	Datog	Nilotic	35	Tishkoff, et al. (2007)
Uganda	Karamojong	Nilotic	118	Gomes, et al. (2010)

Table A1 cont.: Details of additional populations and comparative data examined in the present study

Country	Ethnicity	Language Group	n	Source
Ethiopia	Dawro	Omotic	78	Batini, et al. (2011a)
Ethiopia	Amhara	Semitic	49	Batini, et al. (2011a)
Ethiopia	Amhara	Semitic	48	Semino, et al. (2002)
Ethiopia	Amhara	Semitic	18	Wood, et al. (2005)
Ethiopia	Ethiopian Jews	Semitic	22	Cruciani, et al. (2002)
Ethiopia	mixed Semitic	Semitic	20	Wood, et al. (2005)
Sudan	Gaalien	Semitic	50	Hassan, et al. (2008)
Democratic Republic of Congo	Mbuti	Central Sudanic	33	Batini, et al. (2011a)
Democratic Republic of Congo	Mbuti	Central Sudanic	12	Cruciani, et al. (2002)
Democratic Republic of Congo	Mbuti	Central Sudanic	47	Wood, et al. (2005)
Egypt	Arabs	Semitic	147	Luis, et al. (2004)
Egypt	Egyptian	Erythraic	92	Wood, et al. (2005)
Ethiopia	mixed Ethiopians	-	242	Moran, et al. (2004)
Tanzania	Hadza	Hadza	23	Knight, et al. (2003)
Tanzania	Hadza	Hadza	57	Tishkoff, et al. (2007)
Botswana	ǀGui-ǁGhana-Kgalagari	Central Khoisan / Southeastern Bantu	65	Batini, et al. (2011a)
Mali	mixed Malian population	-	54	Batini, et al. (2011a)
Nigeria	Idoma	Benue-Congo	47	Batini, et al. (2011a)
Nigeria	Igala	Benue-Congo	40	Batini, et al. (2011a)
Nigeria	Tiv	Benue-Congo	54	Batini, et al. (2011a)
Nigeria	Yoruba	Benue-Congo	13	Tishkoff, et al. (2007)
Angola/South Africa	!Xun	Northern Khoisan (Ju)	64	Cruciani, et al. (2002)
Namibia	San	Northern Khoisan (Ju)	5	Batini, et al. (2011a)
Namibia	Tsumkwe San	Northern Khoisan (Ju)	11	YCC (2002)
Cameroon	Bassa	Northwest Bantu	42	Batini, et al. (2011a)
Cameroon	Ewondo	Northwest Bantu	26	Batini, et al. (2011a)
Cameroon	Ewondo	Northwest Bantu	29	Cruciani, et al. (2002)
Cameroon	Fang	Northwest Bantu	4	Berniell-Lee, et al. (2009)

Table A1 cont.: Details of additional populations and comparative data examined in the present study

Country	Ethnicity	Language Group	n	Source
Cameroon	Ngumba	Northwest Bantu	31	Batini, et al. (2011a)
Cameroon	Ngumba	Northwest Bantu	24	Berniell-Lee, et al. (2009)
Cameroon	Ngumba	Northwest Bantu	31	Wood, et al. (2005)
Central African Republic	Lissongo	Northwest Bantu	4	Cruciani, et al. (2002)
Congo	Teke	Northwest Bantu	38	Batini, et al. (2011a)
Congo	Beti	Northwest Bantu	36	Batini, et al. (2011a)
Gabon	Akele	Northwest Bantu	50	Berniell-Lee, et al. (2009)
Gabon	Bekwil	Northwest Bantu	5	Berniell-Lee, et al. (2009)
Gabon	Benga	Northwest Bantu	48	Berniell-Lee, et al. (2009)
Gabon	Duma	Northwest Bantu	46	Berniell-Lee, et al. (2009)
Gabon	Eshira	Northwest Bantu	42	Berniell-Lee, et al. (2009)
Gabon	Eviya	Northwest Bantu	24	Berniell-Lee, et al. (2009)
Gabon	Fang	Northwest Bantu	60	Berniell-Lee, et al. (2009)
Gabon	Galoa	Northwest Bantu	47	Berniell-Lee, et al. (2009)
Gabon	Kota	Northwest Bantu	53	Berniell-Lee, et al. (2009)
Gabon	Makina	Northwest Bantu	43	Berniell-Lee, et al. (2009)
Gabon	Mbaouin	Northwest Bantu	1	Berniell-Lee, et al. (2009)
Gabon	Ndumu	Northwest Bantu	36	Berniell-Lee, et al. (2009)
Gabon	Nzebi	Northwest Bantu	57	Berniell-Lee, et al. (2009)
Gabon	Obamba	Northwest Bantu	47	Berniell-Lee, et al. (2009)
Gabon	Okande	Northwest Bantu	6	Berniell-Lee, et al. (2009)
Gabon	Orungu	Northwest Bantu	21	Berniell-Lee, et al. (2009)
Gabon	Punu	Northwest Bantu	58	Berniell-Lee, et al. (2009)
Gabon	Shake	Northwest Bantu	43	Berniell-Lee, et al. (2009)
Gabon	Teke	Northwest Bantu	48	Berniell-Lee, et al. (2009)
Gabon	Tsogo	Northwest Bantu	60	Berniell-Lee, et al. (2009)
Tanzania	Sandawe	Sandawe	68	Tishkoff, et al. (2007)
Sao Tome & Principe	Forro	-	68	Goncalves, et al. (2007)
South Africa	South African Coloured	Indo-European	164	HGDDRL

Table A1 cont.: Details of additional populations and comparative data examined in the present study

Country	Ethnicity	Language Group	n	Source
Malawi	mixed Central Bantu speakers	Central Bantu	4	HGDDRL
Mozambique	mixed Central Bantu speakers	Central Bantu	303	Batini, et al. (2011a)
Angola	mixed Central Bantu speakers	Central Bantu	28	Coelho, et al. (2009)
Angola	Ngangela	Central Bantu	11	Coelho, et al. (2009)
South Africa	Mixed Southeastern Bantu speakers	Southeastern Bantu	6	HGDDRL
South Africa	Northern Ndebele	Southeastern Bantu	1	HGDDRL
South Africa	Pedi	Southeastern Bantu	114	HGDDRL
South Africa	Shona	Southeastern Bantu	1	HGDDRL
South Africa	Sotho	Southeastern Bantu	108	HGDDRL
South Africa	Southern Ndebele	Southeastern Bantu	30	HGDDRL
South Africa	Swazi	Southeastern Bantu	52	HGDDRL
South Africa	Tsonga	Southeastern Bantu	127	HGDDRL
South Africa	Tswana	Southeastern Bantu	163	HGDDRL
South Africa	Venda	Southeastern Bantu	105	HGDDRL
South Africa	Xhosa	Southeastern Bantu	65	Batini, et al. (2011a)
South Africa	Xhosa	Southeastern Bantu	139	HGDDRL
South Africa	Zulu	Southeastern Bantu	339	HGDDRL
Zimbabwe	mixed Southeastern Bantu speakers	Southeastern Bantu	69	HGDDRL
Zimbabwe	Shona	Southeastern Bantu	49	Wood, et al. (2005)
Botswana	mixed Bantu speakers	Bantu	15	Batini, et al. (2011a)
Zambia	Bisa	Central Bantu	33	de Filippo, et al. (2010)
South Africa	Khomani	Southern Khoisan (Tuu)	6	HGDDRL
Angola	Kuvale	Southwestern Bantu	26	Coelho, et al. (2009)
Angola	Nyaneka-Nkhumbi	Southwestern Bantu	75	Coelho, et al. (2009)
Angola	Ovimbundu	Southwestern Bantu	96	Coelho, et al. (2009)
Sudan	Fur	Fur	32	Hassan, et al. (2008)
Sudan	mixed Sudanese	-	35	Batini, et al. (2011a)
Sudan	Nuba	Eastern Sudanic / Kordofanian	28	Hassan, et al. (2008)
Mali	Dogon	Dogon	55	Wood, et al. (2005)

Table A1 cont.: Details of additional populations and comparative data examined in the present study

Country	Ethnicity	Language Group	n	Source
Guinea-Bissau	Mandinka	Mande	45	Rosa, et al. (2007)
Guinea-Bissau	Balanta	Atlantic	26	Rosa, et al. (2007)
Guinea-Bissau	Bijagos	Atlantic	21	Rosa, et al. (2007)
Guinea-Bissau	Felupe Djola	Atlantic	50	Rosa, et al. (2007)
Guinea-Bissau	Fulbe	Atlantic	59	Rosa, et al. (2007)
Guinea-Bissau	Nalu	Atlantic	17	Rosa, et al. (2007)
Guinea-Bissau	Papel	Atlantic	64	Rosa, et al. (2007)
Cameroon	Baka	Ubangian	63	Batini, et al. (2011a)
Cameroon	Baka	Ubangian	5	Berniell-Lee, et al. (2009)
Cameroon	Bakola	Northwest Bantu	30	Batini, et al. (2011a)
Cameroon	Bakola	Northwest Bantu	22	Berniell-Lee, et al. (2009)
Cameroon	Bakola	Northwest Bantu	33	Wood, et al. (2005)
Central African Republic	Baka	Ubangian	18	Wood, et al. (2005)
Central African Republic	Biaka	Northwest Bantu	21	Batini, et al. (2011a)
Central African Republic	Biaka	Northwest Bantu	20	Cruciani, et al. (2002)
Central African Republic	Biaka	Northwest Bantu	31	Wood, et al. (2005)
Central African Republic	Mbenzele	Northwest Bantu	42	Batini, et al. (2011a)
Congo	Babinga	Ubangian	20	Batini, et al. (2011a)
Gabon	Baka	Ubangian	33	Berniell-Lee, et al. (2009)
Tanzania	Zanzibari	Central Bantu	100	HGDDRL
Madagascar	Malagasy	Austronesian	100	HGDDRL

7.5 Appendix E

Recipes for reagents and solutions used

Sucrose-Triton X Lysing buffer

10 ml 1 M Tris-HCl pH8

5 ml 1 M $MgCl_2$

10 ml Triton-X 100

Make up to 1 L with dH_2O 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 dH_2O

Autoclave

1 M $MgCl_2$

101.66 g $MgCl_2$

500 ml dH_2O

Autoclave

T20E5

20 ml 1M Tris-HCl

10 ml 0.5M EDTA pH8

Make up to 1 L with dH_2O and autoclave

0.5 M EDTA

93.06 g EDTA

500 ml dH_2O

pH to 8.0 with NaOH and autoclave

10% SDS

10 g SDS

100 ml dH_2O

Autoclave

Proteinase K (10 mg/ml)

100 mg Proteinase K stock (100 mg/ml)*

10 ml ddH_20

*Available from Roche Diagnostics

Proteinase-K mix

For 16 extractions:

400 µl 10% SDS

16 µl 0.5 M EDTA

2.8 ml autoclaved dH_2O

Add 800 µl Proteinase K (10 mg/ml stock) just before use

Saturated NaCl

100 ml autoclaved dH_2O

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 dH_2O and autoclave

10 X TBE buffer

108 g Tris

55 g Boric acid

7.44 g EDTA

Make up to 1 L with dH_2O and autoclave

1 X TBE (1:10 dilution)

40 ml 10 X TBE

Make up to 200ml with ddH_20

Bromophenol blue Ficoll dye

50 ml dH_2O

50 g sucrose

1.86 g EDTA

0.1 g bromophenol blue

10 g Ficoll

Dissolve

Adjust volume to 100 ml with dH_2O, 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 ddH_2O

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

2.5mM dNTPs

Use 100 mM premade stocks of dATP, dGTP, dCTP and dTTP (GibcoBRL)

10 µl of each stock dNTP + 360 µl sterile ddH_2O = 400 µl of 2.5 mM dNTPs

7.6 Appendix F

Naidoo, et al. (2010)

Naidoo *et al. Investigative Genetics* 2010, **1**:6
http://www.investigativegenetics.com/content/1/1/6



METHODOLOGY Open Access

Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations

Thijessen Naidoo*, Carina M Schlebusch, Heeran Makkan, Pareen Patel, Rajeshree Mahabeer, Johannes C Erasmus, Himla Soodyall

Abstract

Background: The ability of the Y chromosome to retain a record of its evolution has seen it become an essential tool of molecular anthropology. In the last few years, however, it has also found use in forensic genetics, providing information on the geographic origin of individuals. This has been aided by the development of efficient screening methods and an increased knowledge of geographic distribution. In this study, we describe the development of single base extension assays used to resolve 61 Y chromosome haplogroups, mainly within haplogroups A, B and E, found in Africa.

Results: Seven multiplex assays, which incorporated 60 Y chromosome markers, were developed. These resolved Y chromosomes to 61 terminal branches of the major African haplogroups A, B and E, while also including a few Eurasian haplogroups found occasionally in African males. Following its validation, the assays were used to screen 683 individuals from Southern Africa, including south eastern Bantu speakers (BAN), Khoe-San (KS) and South African Whites (SAW). Of the 61 haplogroups that the assays collectively resolved, 26 were found in the 683 samples. While haplogroup sharing was common between the BAN and KS, the frequencies of these haplogroups varied appreciably. Both groups showed low levels of assimilation of Eurasian haplogroups and only two individuals in the SAW clearly had Y chromosomes of African ancestry.

Conclusions: The use of these single base extension assays in screening increased haplogroup resolution and sampling throughput, while saving time and DNA. Their use, together with the screening of short tandem repeat markers would considerably improve resolution, thus refining the geographic ancestry of individuals.

Background

The Y chromosome has demonstrated its utility, for a number of years, in shedding light on human history and identifying population affinities. Given that human genome variation evolves over time due to several factors - among them mutation, genetic drift, migration and selection - the genome has retained some of the record of these historical and evolutionary events. This record is more easily read from the Y chromosome due to the lack of recombination along most of its length and a strict paternal mode of transmission. Consequently, the Y chromosome has become a marker of the male contribution to the shaping of human populations and their histories.

A standard nomenclature established by the Y Chromosome Consortium [1] resolved the global pattern of Y chromosome variation into 18 major haplogroups that were classified using capital letters A through to R. This has recently been revised by Karafet *et al.* [2] to a Y chromosome haplogroup phylogeny that contains 311 branches delineated by approximately 600 markers (primarily bi-allelic) and includes an additional two haplogroups (S and T), increasing the major haplogroup number to 20. The frequency and distribution of these haplogroups shows good concordance with the geographic distribution of populations. This, together with high levels of population differentiation, [3] have added

* Correspondence: thijessen.naidoo@nhls.ac.za
Human Genomic Diversity and Disease Research Unit, Division of Human Genetics, School of Pathology, University of the Witwatersrand and the National Health Laboratory Services, Johannesburg, South Africa



© 2010 Naidoo et al; licensee BioMed Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

value to the Y chromosome as a tool for reconstructing the history and migrations of humans over time.

While Y chromosome short tandem repeats (STRs) are now used routinely in forensic analysis [4], the use of bi-allelic markers - mainly single nucleotide polymorphisms (SNPs) - which designate Y chromosome haplogroups, is advancing steadily due to their ability to provide information on the geographic origin of individuals. Their use, however, is hindered by the paucity of simple screening methods and insufficient knowledge of their global distribution. However, this has improved in recent years [5]

A number of assays for the rapid screening of Y chromosome haplogroups have been developed [6-9]. These were targeted primarily at resolving the major haplogroups found in European populations. While these studies have included in their assays a few SNPs to resolve the major Y chromosome haplogroups commonly found in sub-Saharan Africa, they do not contain the markers needed to resolve the majority of terminal branches of the Y chromosome phylogeny that exist among African populations.

In the present study, we report on the development of single base extension (SBE) assays used to refine the resolution of Y chromosome haplogroups commonly found in Africa, having also incorporated a few SNPs to delineate the common non-African Y chromosomes following a hierarchical screening process. SBE, due to its convenience and relative affordability, is now used in many genetic and forensic applications. Following the validation of the assays, we applied these methods in order to resolve the Y chromosome haplogroups in 683 male subjects, primarily from southern Africa. Haplogroup frequencies for the populations analysed were then calculated.

Results and discussion

SNP selection and screening strategy

Seven multiplex SBE assays, which incorporated 60 Y chromosome markers described in the YCC Phylogeny 2003 [10], were developed which resolved 61 Y chromosome haplogroups. The first multiplex, YSNP1, consisted of the markers SRY10831, M168, M89, M201, M69, M170, M172, M9, M207, M198 and M343 (Figure 1a). YSNP1 resolved Y chromosomes into either the African haplogroups (A, B or E) or Eurasian haplogroups found occasionally in African males [unpublished restriction fragment length polymorphism (RFLP) data]. Note: the marker, SRY10831, initially resolves haplogroup BR, while its reversion is used to define haplogroup R1a.

Any sample found to harbour the ancestral state at all markers within YSNP1 was screened using the multiplex assay, Hg-A. This multiplex consisted of the markers, M91, M31, M14, M114, P28, M28, M51, M13, M171 and M118 (Figure 1b) and was used to resolve the sub-clades of haplogroup A. Samples found to be derived at SRY10831, but ancestral at all other markers within YSNP1, were screened using the multiplex assay, Hg-B. This multiplex consisted of M60, M146, M182, M150, M152, M108, M43 and M112 (Figure 1c) and resolved the sub-clades of haplogroup B. Those samples with the derived allele at M112 were screened further, using the multiplex assay, Hg-B2b, which contained the markers P6, M115, M30, P7, P8 and M211 (Figure 1d), providing resolution of haplogroup B2b samples to the terminal branches of the phylogeny. While M108 recurs in haplogroup B2b resolving haplogroup B2b3a, its presence in the Hg-B multiplex assay should be sufficient to resolve both its occurrences in haplogroup B, thus negating the need to include it in the Hg-B2b assay.

Those samples found to be derived at SRY10831 and M168, while remaining ancestral at all other markers within YSNP1, could be assigned to haplogroups C, D or E. These samples were then screened using the Hg-E multiplex assay, which consisted of M40, M33, M44, M75, M41, M85, P2, M2 and M35 (Figure 1e). Samples found to be derived for M2 or M35 would fall into haplogroups E1b1a or E1b1b1, respectively. E1b1a Y chromosomes were further resolved using the assay, Hg-E1b1a; a multiplex comprised of the markers M58, M116, M149, M154, M155, M10 and M191 (Figure 1f). Those Y chromosomes assigned to haplogroup E1b1b1 were screened further using the multiplex assay, Hg-E1b1b1, which consisted of the markers M78, M148, M81, M107, M165, M123, M34, M136 and M281 (Figure 1g). When a sample was found to be ancestral for the M40 polymorphism, it was screened for the mutations that defined haplogroups C (M130) and D (M174) separately. This hierarchical screening approach facilitated the resolution of the relevant haplogroup in an individual after one, two, or at most, three reactions, depending on the haplogroup present.

Polymerase chain reaction (PCR) optimization

While PCR primer concentrations were initially 0.02 μM - 0.04 μM, these were increased or decreased incrementally in order to obtain a relatively equal amplification of amplicons in the multiplex PCR (see Table 1). The marker P28, in the Hg-A assay, initially experienced low amplification after multiplexing. This was rectified by increasing the final concentration of the P28 PCR primers to 0.2 μM, and decreasing the buffer concentration to 0.8×. The annealing temperature was also optimized in order to ensure maximum product yield and to minimize the formation of spurious products. A spurious amplification product was found to occur in the Hg-E1b1b1 assay which was eliminated by increasing the annealing temperature to 61°C.

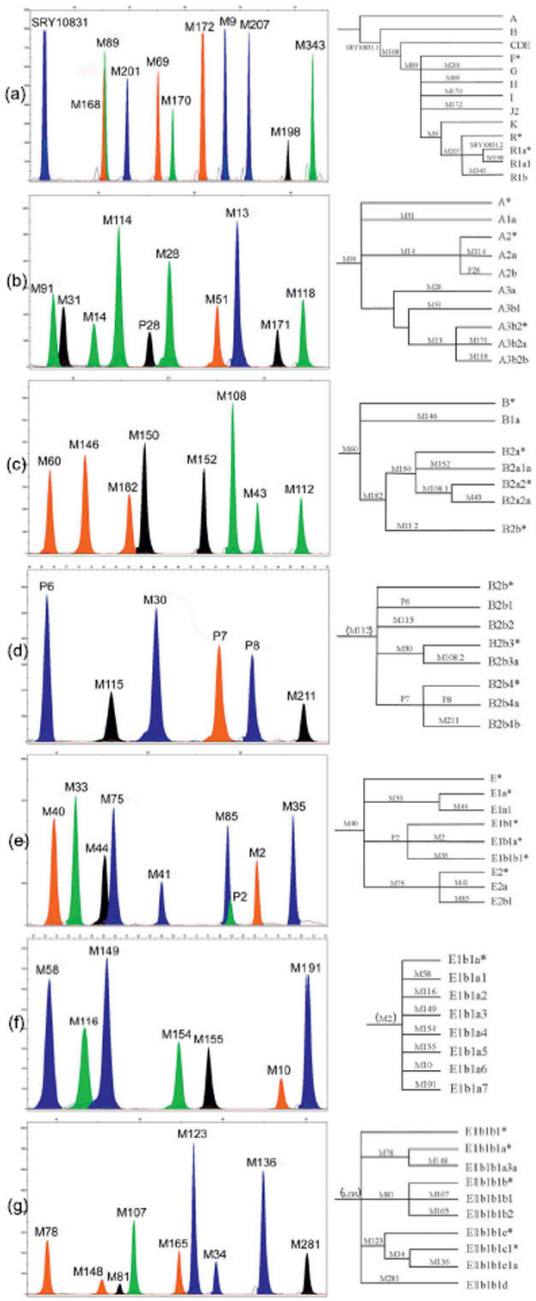

Figure 1 Electropherogram and phylogeny of (a) YSNP1, (b) Hg-A, (c) Hg-B, (d) Hg-B2b, (e) Hg-E, (f) Hg-E1b1a, and (g) Hg-E1b1b1. The markers shown in parentheses are there to indicate the hierarchical nature of the phylogenies, and are not a part of the panels they are shown with.

Table 1 Polymerase chain reaction (PCR) primer sequences, amplicon lengths and final PCR primer concentrations used in the study

SNP	Primer (5' - 3')		Fragment size (mers)	Concentration (μM)	Assay
	Forward	Reverse			
M170	CTAGTATGCTTCACACAAATGCG	GACCACACAAAAACAGGTCCTC	390	0.08	YSNP1
M207	AAGGGCAAGCAAAATAGCAATAC	TGTTCGCTGCTACGAATCTTT	363	0.08	
M201	CATGGGTAATTCGGTTGTTACC	CTAAACATCATGGTGTGACGAAC	331	0.08	
M168	GGTTTGAATGAGACTGGGTCA	TGGTAATCTCATAGGTCTCTGACTG	295	0.08	
M343	AGGTAGGAGGATCCAAAGCTGA	CACCTTTGTCCTCTTGCTCTTT	276	0.08	
M9	TGCAGCATATAAAACTTTCAGGAC	TTCCTCATTTTTGAAGCTCGTG	241	0.08	
M69	TCAGCCATTTCACCAAAACTCT	CTGAAGAAACAAACCTACCTGGAA	233	0.08	
M89	CCAAGCTGGTGAGTCTTATCCT	GCAGAATAGCTGCTCAGGTACA	215	0.08	
M172	CAGAAGATGCCCCATTATATCCT	ACTCCATGTTGGTTTGGAACAG	208	0.08	
M198	TAGGCACTTGGGAACTTACACTC	TTCTTGTGATAGCATGCCGTTT	178	0.08	
SRY10831	CATCCAGTCCTTAGCAACCATT	AATGACACAAGGCACCACATAG	163	0.08	
P28	TTTTGAGAGAAGACAAGGGGGATA	TTGGAGGGACATTATTCTCCTGA	559	0.20	Hg-A
M13-M14*	ATCACGCCCCTCTCATTGTC	AGCTCTAGATAAAAGCACATTGACACC	457	0.08	
M91	GATCACAAAGACCTGGACAGATTACA	AACGGAAATGCCAAGAATCGTA	429	0.05	
M31	GCTGAAACAATAGTTCTTCACAATGG	CAGTCCTATGCATAATGCCGTGT	400	0.03	
M114	GCCTTGATTTTCTTCGTACTTCATAAG	CCAGTTTCTCACTGAGTTCATTCCTT	370	0.06	
M28	GGGCTTCAGTTCTTGACGCTAC	CCGTCTTAATTTTGCGGTATTCAA	329	0.04	
M51	AAACCACACCTGTCTTACCAGAGC	CTGTTCCCCAGTTTTCAATCTCC	293	0.04	
M171	GGCTGTGTGGAGTATGTGTTGG	CAAATATCCTGCCCCAGCTTAGT	217	0.04	
M118	TCCCTTGAAATTAAGGACAACAACT	CATTCTTCTCAACCAGCTGACACT	167	0.06	
M43	GACTCCATAAGCAAAAGGTCATCAA	AAAAGAAGTTGAGGACTGGAGCA	518	0.12	Hg-B
M112	GGCCATGCTAACAGAGATCTGAC	CACAGTTCAATTCTTGTCTGTTGC	493	0.08	
M152	AAGCAAAAAGCTCCTTCTGAGGT	CAGAAGGTGGATCAGGGTAGAAA	381	0.08	
M182	CATTTTTGTGTCAGGTATCCTTTGT	CAAGACGGCGTATCAACTCAAG	368	0.12	
M108	GCTTTTCTAACACCACCCATGAC	TATGTGATAGAGGTGGCTTTAAGTGG	342	0.08	
M150	CCAGGCTAGCAGTGGAGATGAA	AGGGTGGACTGCTGACCTACTTT	312	0.08	
M146	TTACAGGTGGAATGGGGTGTTAC	GAGAAGAACTGCCTTCCATGACATA	279	0.08	
M60	CCTGATGTGGACTCAACCTTGTA	TGTTCATTATGGTTCAGGAGGAG	250	0.08	
M211	CACTGCACACACTACACTGACCAC	ATGTTGATTGGGTAGAGCCCTTT	386	0.06	Hg-B2b
P6	TATTAGGGAAATCACTCAGGATGGT	TCTACGAATGTTTAACCTCAGATACCG	343	0.12	
P8	AGTTGTTGGAAAGCCTCTGTTCA	TGATACTAGACGTGGCATCTTGTCA	313	0.06	
M115	TGCCATGCTTGTTTCTTAATCCA	AACTATGTTGCACATCAGCCTCA	270	0.12	
P7	GGCCAAAGCCTAGAATGAAATTG	AAGTGCTTGTCCAAGGCAGTATAA	228	0.16	
M30	ACAAATCATGAGCTTACAGAACCTG	GGCACAGCCAGATAACCCTACA	200	0.12	
M40-M41*	TAGCTGGTATGACAGGGGATGAT	GGGTAGGATAGGCTAGCTATTACGC	435	0.08	Hg-E
M2	GGAGAAGAACGGAAGGAGTTCTAA	ACTTGCCAGAGACTTCCAGTTTG	372	0.12	
M85†	GAACTGGCATCCAATACTAGCTGA	TCACCTCTTTTGTATTGGCTTCTTC	350	0.08	
P2	TGGTCTGGTAACACCCATAAAGGT	GCAGTTTTCTCAGATGCTTCTCCTA	335	0.08	
M35	GCCTAAAGAGCAGTCAGAGTAGAATG	GAGAATGAATAGGCATGGGTTCA	303	0.08	
M75	GTCACATTCCACACATCAAGAAAACT	GTGAATCTCTGCCCAGAAAAGAAAA	274	0.12	
M44	ATTGGATATGGAAGCCAGTCTCA	ATGTGTTTGAGGACCACCCTAGA	250	0.12	
M33	GGCTTCTGTTCAATTTTCCTTTGAT	TTATTTGTTGAAGCCCCCAAGAG	223	0.08	
M10	GTTCAAGACAATGAAGGGAGAGACT	TGACATTGACCTGCAGCATAGG	520	0.08	Hg-E1b1ba

Table 1 Polymerase chain reaction (PCR) primer sequences, amplicon lengths and final PCR primer concentrations used in the study *(Continued)*

M191	GAGCAAGTACAGCGAGCAGTAAG	GGTTTAACACAATGCAGGTCAATTC	480	0.08	
M154†	CAATGGAGGCTATAGGTGATTGC	CTGTTTGTTCATGGAGATGTCTGTA	461	0.08	
M149†	GAACTGGCATCCAATACTAGCTGA	TCACCTCTTTTGTATTGGCTTCTTC	350	0.06	
M116	TATGAAGTACGAAGAAAATCAAGGCTA	TGGGTAGAAAAACTGCAAGTAGATGA	328	0.12	
M58-M155*	TGGGCCTGACCTCTTAACTTGTA	CATAATAAGCTAAGAAACATCCAGCCT	293	0.06	
M81†	CAATGGAGGCTATAGGTGATTGC	CTGTTTGTTCATGGAGATGTCTGTA	461	0.10	Hg-E1b1b1
M123-M281*	CTAATTCATGCTCTCAGGGGAAA	ATAACCTCTGGAAGTGTGTCTTTACCT	404	0.10	
M107	AATCCCACCTCACATACACATAAGC	AGGGGTTGACAAGAAAAAGGAATA	386	0.06	
M148†	GAACTGGCATCCAATACTAGCTGA	TCACCTCTTTTGTATTGGCTTCTTC	350	0.08	
M78	ATGGCTGTATGGGTTTCTTTGACT	CGGAATATGGACAGTCATCGTATT	330	0.08	
M165	CAAGTCAGCAAGGAGTAGGTGGA	TTGCACTGACACAAGTTATCTCCCT	293	0.08	
M34	GATAACCTCATTGTGGAGAGCACTT	ATGCTAAAGCAAGTAACCCTGTGG	254	0.10	
M136	ACCAACCGTATTACCTTCTCCTCA	CATGAGTCCAAAGTATAGTGGGCTA	226	0.10	

* Due to the close proximity of the single nucleotide polymorphisms (SNPs), a single amplicon was used.
† Again, a single amplicon was used for these SNPs, but in different assays.

SBE optimization

The SBE primers designed for the seven SBE assays, ranging from 25 to 80 bases in length, were designed to differ by four to five bases within each assay (see Table 2). This was not always reflected in the electropherogram, with a lack of uniform separation in most of the assays. This resulted in a few extension products (for example, M85 and P2 in Hg-E, M168 and M89 in Hg-YSNP1) co-migrating (see Figure 1). Fortunately, this did not interfere with the interpretation of results. The estimated lengths of extension products in the electropherogram (based on mobility) differed from the designed lengths, on average by four bases. This difference was ascribed to the migration rate of the primer (which was influenced by its actual length), possible secondary structure [11], mobility of the dye attached [12] and the use of POP-7® polymer. This was especially apparent for the M91 primer, a 25-base primer which sized, on average, 11 bases larger. Despite these observations, profiles generated by all the assays were usually easily interpreted.

While the generation of aspecific peaks did occur occasionally, this was usually due to insufficient purification of the PCR products resulting in the incorporation of the PCR primers or deoxyribonucleotide triphosphates (dNTPs) into the SBE reaction. The presence of one permanent aspecific peak did occur, however, in the Hg-B2b assay (a red peak between the P8 and M211 peaks). This peak seemed to be linked to the P7 primer, perhaps due to a problem during its synthesis, and was usually more visible when overall peak height was decreased.

In order to intensify peak heights, the number of cycles in the SBE reaction program was increased from 25 to 35. While overall peak height improved, variability of peak heights within some assays was unavoidable, despite the adjustment of relative SBE primer concentrations. This was possibly influenced by the efficiency of interaction between SBE primers and template sequences.

Validation of SBE assays

The seven SBE assays were validated using samples whose haplogroup status was previously determined. A total of 683 samples were then screened. Additionally, sequencing was performed to confirm the presence of alleles for 15 mutations not screened for before the use of these SBE assays. These included M14, M114, M152, P6, P7, P8, M33, M44, M85, M58, M154, M34, M201, M198 and M343.

The marker M91, in the Hg-A assay, is a homopolymer length variant associated with a single base deletion in a poly-T tract [13]. While the use of SBE in the screening of homopolymer variants is not common, the detection of the M91 mutation using the SBE method was successful. This was reaffirmed phylogenetically [14,15] by the presence of this mutation exclusively in samples belonging to subclades of haplogroup A.

The validation process resulted in the redesign of just two SBE primers, P28 and M35. The initial P28 SBE primer did not pick up the mutation, probably due to non-specific primer binding, while the initial M35 primer resulted in an extremely low peak height when the mutation was present. This was possibly due to the

Table 2 Single base extension (SBE) primer sequences and final concentrations used in the study, grouped by assay

SNP		SBE primer (5' - 3')	Size (mers)	Concentration (μM)	Assay
SRY10831	FW	$(C)_3$CTCTTGTATCTGACTTTTTCACACAGT	30	0.10	YSNP1
M168	FW	$(C)_{12}$TGGAGTATGTGTTGGAGGTGAGT	35	0.40	
M89	RV	$(GACT)_2(C)_{10}$CAACTCAGGCAAAGTGAGAGAT	40	0.40	
M201	FW	$(GACT)_2(C)_9$AGATCTAATAATCCAGTATCAACTGAGG	45	0.40	
M69	FW	$(GACT)_4(C)_{11}$GGAGGCTGTTTACACTCCTGAAA	50	0.40	
M170	FW	$(GACT)_4(C)_9$ACTATTTTATTTACTTAAAAATCATTGTTC	55	0.80	
M172	FW	$(GACT)_7(C)_{12}$CCAAACCCATTTTGATGCTT	60	0.40	
M9	FW	$(GACT)_8(C)_{11}$AAACGGCCTAAGATGGTTGAAT	65	0.40	
M207	FW	$(GACT)_8(C)_{11}$GCAAATGTAAGTCAAGCAAGAAATTTA	70	0.80	
M198	FW	$(GACT)_9(C)_9$TCAGTATACCAATTAATATTTTTGAAAGAG	75	0.80	
M343	FW	$(GACT)_{13}(C)_9$AGAGTGCCCTCGTGTTCCA	80	0.40	
M91	FW	CCTACATTGCTATTCTGTTTTTTTT	25	0.60	Hg-A
M31	RV	$(C)_8$CCACTGCTGTTCTGTCTACCA	29	0.60	
M14	RV	$(C)_5$CTTCATTAACTTTTTTAAACTGCTTATA	33	0.60	
M114	RV	$(C)_{15}$AGCTGTACAAGGCTCTTCAAAT	37	0.60	
P28	FW	$(C)_{14}$GGTTAAAAGAAAAAAGCTCTCAGATAG	41	0.40	
M28	RV	$(C)_{27}$TCGAGGTCCTCTGGCATC	45	0.50	
M51	RV	$(C)_{29}$CTCTGATCCCTGTTGGAAGC	49	0.50	
M13	FW	$(C)_{31}$GTAGGTTAAGGGCAAGACGGTTA	54	0.60	
M171	RV	$(C)_{32}$AGGTCTCTGACTGTTCAGTTTTATT	57	0.50	
M118	RV	$(C)_{35}$CAGCTGACACTTGTGTTTTCTTTATA	61	0.20	
M60	FW	$(C)_3$TTACATTTCAAAATGCATGACTTAAAG	30	0.40	Hg-B
M146	RV	$(C)_{11}$CTAAAACCCAGTGTTAATTACCCG	35	0.80	
M182	FW	$(C)_{13}$CTTAAAGCAGTGGTTAATGTAAACAAA	40	0.80	
M150	FW	$(C)_{22}$TGCCCACACACACAGATAGAAGT	45	0.80	
M152	FW	$(C)_{23}$GCTTTCTCCTGATAATGTTCTTCTTCT	50	0.80	
M108	RV	$(C)_{27}$CTTTTCTCTGACATTCAGGTATAGTTTC	55	0.30	
M43	FW	$(C)_{39}$CTCTTTCCATGGCCAACAAAC	60	0.40	
M112	FW	$(C)_{38}$AAAGAGGTGAGATAAAAACAAAGCAGT	65	0.40	
P6	FW	$(C)_3$TCAATAGAGGTTTCCACAGTTAAGTCT	30	0.10	Hg-B2b
M115	FW	$(C)_3$CAGAGTTTAAATTAGTATTTGATTTCACATTA	35	0.80	
M30	FW	$(C)_{14}$ATCATGTTTTAAGTCCTGACATCTGT	40	0.10	
P7	FW	$(C)_{21}$CCATCACCTGGGTAAAGTGAATTA	45	0.40	
P8	FW	$(C)_{27}$GCAGCTCACCTTTCATTTAGGTC	50	0.20	
M211	FW	$(C)_{30}$TAGGCAAAAGGATGTTAACAACAAG	55	0.80	
M40	RV	$(C)_{10}$TCTTCACCCTGTGATCCGCT	30	0.60	Hg-E
M33	FW	CGATCTGTTCAGTTTATCTCATAAGTTACTAGTTA	35	0.30	
M44	RV	$(C)_{11}$AGGAAATCTCCTAACCTTCTAGTACACTG	40	0.40	
M75	FW	$(C)_{20}$AAAAGACAATTATCAAACCACATCC	45	0.10	
M41	FW	$(C)_{30}$TGGCCAACATGGTGAAACTG	50	0.50	
M85	RV	$(C)_{24}$GCTTGTGTTCTATTAAGTGTAGTTTTGTTAG	55	0.20	
P2	RV	$(GACT)_8(C)_8$AGGTGCCCCTAGGAGGAGAA	60	0.40	
M2	RV	$(GACT)_9(C)_6$CCCTTTATCCTCCACAGATCTCA	65	0.60	
M35	RV	$(GACT)_{10}(C)_9$TTCGGAGTCTCTGCCTGTGTC	70	0.80	
M58	FW	$(C)_5$ATTTATTGTCTTCTGCAGAATTGGC	30	0.10	Hg-E1b1a
M116	FW	$(C)_5$GCTTTCTGAAAAAATAATTTCAAACTGATA	35	0.40	

Table 2 Single base extension (SBE) primer sequences and final concentrations used in the study, grouped by assay *(Continued)*

M149	FW	$(C)_9$CTAACAAAACTACACTTAATAGAACACAAGC	40	0.10	
M154	RV	$(C)_{16}$GTGTTACATGGCCTATAATATTCAGTACA	45	0.40	
M155	RV	$(C)_{23}$AATTCAGAATATTTCATCTCTGGTCAC	50	0.40	
M10	FW	$(C)_{26}$AATTTTTTTGTTTATTCCCAATGATCTTA	55	0.50	
M191	FW	$(GACT)_5(C)_{10}$ATTTACATTTTTTTCTTTACAACTTGACTA	60	0.40	
M78	FW	AATTGATACACTTAACAAAGATACTTCTTTC	31	0.80	Hg-E1b1b1
M148	RV	$(C)_7$TTTCTAGGTAACGTATGTAGACATTTCTG	36	0.80	
M81	FW	$(C)_{15}$AGAGGTAAATTTTGTCCTTTTTTGAA	41	0.80	
M107	FW	$(C)_{18}$TAAGCCAACGTATTAACCTTCTAATTTC	46	0.20	
M165	RV	$(C)_{20}$AAATATTTCAGGTAAAACCACTCTATTAGTA	51	0.40	
M123	FW	$(C)_{27}$AAAGTCACAGTATCTGAACTAGCATATCA	56	0.80	
M34	FW	$(GACT)_7(C)_{13}$GCCTGGCTTCCACCCAGGAG	61	0.20	
M136	RV	$(GACT)_8(C)_{12}$GGTGAGCAGCATTGAGGAAGAC	66	0.10	
M281	RV	$(GACT)_8(C)_{11}$AGGTTGCACAAACTCAGTATTATTAAAC	71	0.80	

SNP, single nucleotide polymorphism; FW, forward orientation; RV, reverse orientation

preferential amplification of the ancestral allele, or a lower efficiency of binding by the original SBE primer.

Finally, in haplogroups B2b4* and B2b4b, P7 showed the presence of two different extension products, displaying both the ancestral and derived states, simultaneously. This also occurred in haplogroup B2b4a, with P8, additionally, exhibiting the same property. The presence of both states was confirmed when sequencing was performed. Thus, it is likely that all samples in haplogroups B2b4*, B2b4a and B2b4b will display two peaks at the relevant markers. This was, probably, a consequence of these markers being located within paralogous sequence variants [16]. It should be noted that such mutations are more susceptible to back-mutation through gene conversion, as it was with P25 [17]. For this reason, more stable markers that resolve these subclades of B2b4 would be preferable.

Haplogroup assignment using SBE assays

The sample of 683 males screened using the seven SBE assays was assigned to 26 of the 61 haplogroups that the assays collectively resolved (see Table 3 and Figure 2). The subclades of haplogroup A were found most commonly in the KS, at a frequency of 44.3%. Haplogroup A3b1 was the commonest (28.4%) and was also found to be present in the BAN at low levels (5.0%). The haplogroups A2*, A2a and A2b were found to be unique to the KS at frequencies of 2.7%, 4.4% and 8.7%, respectively. Haplogroup B was present at moderate frequencies in both the KS and the BAN. Its subclades, however, displayed differing distributions, with haplogroup B2a1a occurring at a substantially higher frequency in the BAN (16.0%) than in the KS (0.5%). The situation was reversed with regard to haplogroup B2b, with its subclades together constituting 10.9% of KS individuals, as compared to 0.3% in the BAN. Haplogroup E was the most common haplogroup in the BAN group (78.1%), with its subclades E1b1a* and E1b1a7 occurring at frequencies of 34.1% and 25.9%, respectively. While both these haplogroups occurred in the KS at 13.1% and 7.7%, respectively, the most frequent E subclade amongst the KS was E1b1b1* (15.8%). Haplogroups E2* and E2b1 were found at much higher frequencies (10.2% versus 1.6%) in the BAN compared with the KS. The haplogroups shared between the BAN and KS, described above, showed extremely significant differences in frequency between the two groups (Fisher's exact test: $P < 0.0001$; for haplogroup E2: $P = 0.0001$). Both the KS and the BAN showed low levels (3.3% and 0.6%, respectively) of assimilation of the Eurasian Y chromosome haplogroups I, K* (x R), R1a1, and R1b.

Y chromosomes in the SAW sample were resolved into macro-haplogroup F (89.2%) of which haplogroup R (58.0%) and haplogroup I (17.8%) together accounted for 75.8%. Haplogroup E comprised the rest of the SAW at 10.8%, with its subclade E1b1b1a found at a frequency of 7.6%. These low to moderate levels of E1b1b1 illustrate the spread of the haplogroup and its subclades into southern Europe and the Middle East, where they are often found. Only two SAW samples, however, clearly belonged to African haplogroups (E1a1 and E1b1a*).

The haplogroup distributions and their relative frequencies in the BAN and the KS were consistent with previous studies which included these populations [18-20], while those of the SAW were found to correlate strongly with the Western European populations from which the majority are derived [21].

Table 3 Y chromosome haplogroup frequencies in south eastern Bantu speakers, Khoe-San, and South African Whites

Marker	Haplogroup	South eastern Bantu-speakers	(%)	Khoe-San	(%)	South African Whites	(%)
M14	A2* (x A2a, A2b)			5	2.7		
M114	A2a			8	4.4		
P28	A2b			16	8.7		
M51	A3b1	17	5.0	52	28.4		
M152	B2a1a	55	16.0	1	0.5		
M112	B2b* (x B2b1, B2b2, B2b3, B2b4)	1	0.3	2	1.1		
P6	B2b1			13	7.1		
P8	B2b4a			5	2.7		
M40	E* (x E1a, E1b1, E2)	1	0.3				
M44	E1a1					1	0.6
M2	E1b1a* (x E1b1a1, E1b1a2, E1b1a3, E1b1a4, E1b1a5, E1b1a6, E1b1a7)	117	34.1	24	13.1	1	0.6
M58	E1b1a1	11	3.2	2	1.1		
M154	E1b1a4	10	2.9	2	1.1		
M191	E1b1a7	89	25.9	14	7.7		
M35	E1b1b1* (x E1b1b1a, E1b1b1b, E1b1b1c, E1b1b1d)	5	1.5	29	15.8	2	1.3
M78	E1b1b1a					12	7.6
M34	E1b1b1c1* (x E1b1b1c1a)			1	0.5	1	0.6
M75	E2* (x E2a, E2b1)	4	1.2	1	0.5		
M85	E2b1	31	9.0	2	1.1		
M89	F* (x G, H, I, J2, K)†					4	2.5
M201	G					9	5.7
M170	I			1	0.5	28	17.8
M172	J2					6	3.8
M9	K* (x R)†	2	0.6			2	1.3
M198	R1a1			2	1.1	10	6.4
M343	R1b			3	1.6	81	51.6
	TOTAL	343		183		157	

† Included in the frequencies of haplogroups F* and K* are those lineages not included in the SBE assays, but typed using PCR or RFLP assays, namely P12F2 and M70.

Conclusions

Seven SBE assays containing 60 SNP markers were designed, which allowed for the rapid assignment of samples to Y chromosome haplogroups, more especially those belonging to the major African haplogroups A, B and E. The assays were designed based on markers found in the YCC phylogeny 2003 [10]. Since then many more markers have been discovered that have further resolved the phylogeny [2]. If needed, these new markers could be incorporated into the current SBE assays, thereby increasing resolution. The use of the current SBE assays in screening Y chromosomes, however, has still resulted in increased haplogroup resolution and sample throughput, and at the same time was quicker and made use of less DNA.

Based on the abovementioned haplogroup frequencies, the KS and the BAN populations are discernable from each other with the KS exhibiting significantly higher frequencies of haplogroups A2, A3b1, B2b and E1b1b1. However, the BAN are identifiable by the strong presence of haplogroups B1a1a, E1b1a and E2. The SAW were appreciably different from both the KS and BAN. There were, however, considerable levels of admixture between the populations (especially the KS and BAN) due to their history of interaction. Consequently, while the elucidation of Y chromosome haplogroups is useful in African populations for both anthropological and forensic purposes, their use together with Y chromosome STR screening would considerably improve resolution and thus refine an individual's geographic ancestry.

Methods

DNA samples

DNA samples from 683 individuals with diverse ethnic backgrounds were analysed in the present study. All DNA samples were collected with the subjects' informed

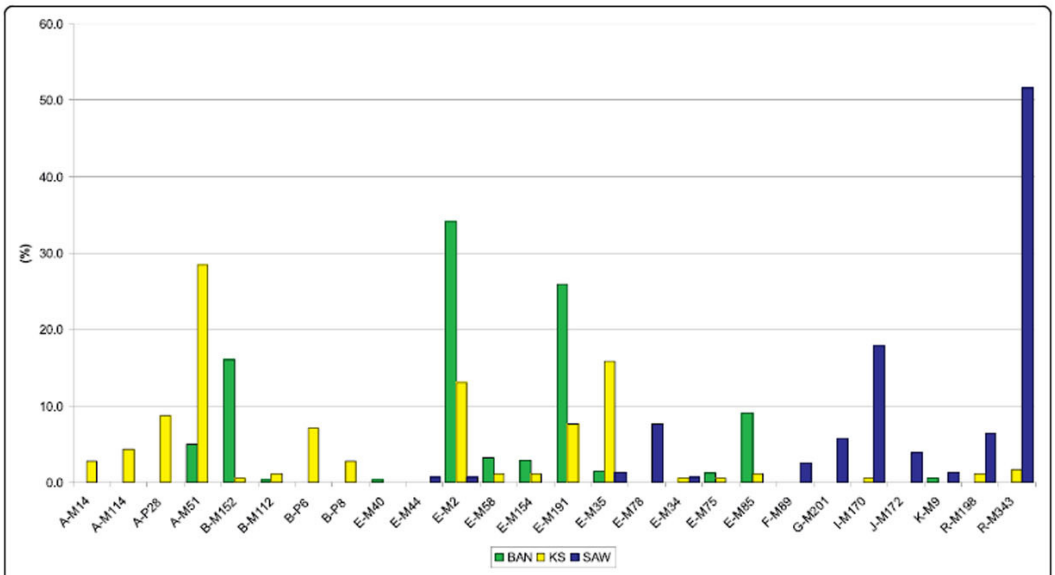

Figure 2 Clustered column graphs of haplogroup frequencies in south eastern Bantu-speakers (BAN), Khoe-San (KS) and South African Whites (SAW).

consent, and this research was approved by the Human Research Ethics Committee (Medical) at the University of the Witwatersrand, Johannesburg (Protocol No. M050906). The sample included 383 south eastern Bantu speakers (BAN), 183 Khoe-San (KS) and 157 South African Whites (SAW).

DNA extraction

DNA from EDTA-blood was extracted using the salting-out method described by Miller *et al.* [22] and the Gentra Puregene Buccal Cell Kit (Qiagen, Germany) was used to extract DNA from buccal swabs according to the manufacturer's instructions. DNA was quantified using the NanoDrop ND-1000 Spectrophotometer (LabVIEW®, Coleman Technologies Inc, FL, USA) and diluted to 10 ng/μL using double distilled water.

Primer design

The sequences of the regions encompassing the polymorphisms were taken from GenBank. The PCR and SBE primers were designed using Primer3 software [23], before aligning them to human genomic sequences using the National Center for Biotechnology Information basic local alignment search tool (BLAST) in order to confirm template specificity. The screening software, AutoDimer [24] was used to check for primer-dimer and hairpin loop formation. High-performance liquid chromatography-purified primers were purchased via Roche from Metabion (Martinsried, Germany), diluted to 100 μM and frozen.

PCR primer lengths ranged from 20 to 27 mers and GC percentage varied between 30% and 60%. Amplicons were designed to differ slightly in size in order to distinguish them following agarose gel electrophoresis to check the success of the PCR. In total, 53 pairs of PCR primers were designed encompassing all 60 SNPs (see Table 1). Fewer pairs of primers were needed, as some SNPs were co-amplified on the same amplicons (M13 and M14; M40 and M41; M58 and M155; M123 and M281; M81 and M154; M85, M148 and M149).

Poly-C or Poly-GACT tails of differing lengths were added to the 5' end of most SBE primers (Table 2), so as to differentiate between them during capillary electrophoresis. SBE primer lengths ranged from 25 to 80 mers, and differed in size from each other by 4-5 mers.

Multiplex PCRs

Primer design was verified by performing simplex PCR, using a GeneAmp PCR system 9700 (Applied Biosystems, CA, USA), for each of the primer pairs. Thereafter, the multiplex PCRs were optimized to work with DNA at a concentration of 10 ng/μl (see Table 4), and were catalysed using FastStart Taq DNA Polymerase (Roche, Basel, Switzerland). Relative primer concentrations were adjusted in order to obtain balanced amplification of amplicons within each multiplex. The thermal cycler programmes were as

Table 4 Polymerase chain reaction (PCR) reagent mixtures for multiplex single base extension assays

PCR reagent	YSNP1	Hg-A	Hg-B	Hg-B2b
	Concentration	Concentration	Concentration	Concentration
DNA	10 ng/μL	10 ng/μL	10 ng/μL	10 ng/μL
Buffer (10×)	1×	0.8×	1×	1×
$MgCl_2$ (25 mM)	2 mM	4 mM	3 mM	3 mM
dNTPs 2.5 mM)	300 μM	200 μM	200 μM	200 μM
Forward primer (10 μM)	See Table 1	See Table 1	See Table 1	See Table 1
Reverse primer (10 μM)	See Table 1	See Table 1	See Table 1	See Table 1
FastStart Taq (5 U/μL)	1 U	1 U	1 U	1 U
ddH_2O	Made up to 25 μL	Made up to 25 μL	Made up to 25 μL	Made up to 25 μL

PCR reagent	Hg-E	Hg-E1b1a	Hg-E1b1b1
	Concentration	Concentration	Concentration
DNA	10 ng/μL	10 ng/μL	10 ng/μL
Buffer (10×)	1.25×	1.25×	1.25×
$MgCl_2$ (25 mM	3 mM	3 mM	3 mM
dNTPs (2.5 mM)	200 μM	200 μM	200 μM
Forward primer (10 μM)	See Table 1	See Table 1	See Table 1
Reverse primer (10 μM)	See Table 1	See Table 1	See Table 1
FastStart Taq (5 U/μL)	1 U	1 U	1 U
ddH_2O	Made up to 25 μL	Made up to 25 μL	Made up to 25 μL

dNTP, deoxyribonucleotide triphosphate; ddH_20, double distilled water

follows: one cycle at 95°C for 6 min, 35 cycles at 95°C for 30 s, 54°C (for YSNP1), 55°C (for Hg-A, Hg-B, Hg-B2b, Hg-E and Hg-E1b1a) or 61°C (for Hg-E1b1b1) for 30 s, extending at 72°C for 30 s and a final extension of 72°C for 10 min. Following the optimization procedures, all multiplex PCRs produced the required amplification products at adequate yields. This was confirmed by running 5 μL of multiplex PCR product on a 2% Metaphore® agarose gel (Cambrex, NJ, USA).

Multiplex SBE

Excess PCR primers and dNTPs were eliminated from the PCR product mixture, following amplification, using an enzymatic purification method. One unit of Exonuclease I (*Exo I*) and 0.5 units of Shrimp Alkaline Phosphatase (SAP) were added to 5 μL of amplification product and the resultant mixture incubated for 1 h at 37°C, followed by 15 min at 75°C.

The multiplex SBE reactions were performed in a final volume of 5 μL, comprised of 1.5 μL purified amplification product, 1.5 μL of double distilled water, 1 μL of SNaPshot Multiplex Ready Reaction Mix (Applied Biosystems) and 1 μL of SBE primer mix, specific to the assay being conducted (see Table 2 for final primer concentrations). The thermal cycler programme was as follows: 96°C for 10 s, 50°C for 5 s and 60°C for 30 s for 35 cycles.

Following the SBE reaction, excess dideoxyribonucleotide triphosphates (ddNTPs) were removed through the addition of 0.5 U of SAP to the 5 μL SBE product. The mixture was incubated for 1 h at 37°C, followed by 15 min at 75°C.

Capillary electrophoresis

Following post-extension treatment, 2 μL of SBE product was mixed with 0.5 μL of the internal size standard, GS120LIZ (Applied Biosystems) and 7.5 μL Hi-Di formamide (Applied Biosystems). This was then run on a 3130*xl* Genetic Analyzer (Applied Biosystems). The SNaPshot protocol was originally optimized for use with POP-4 polymer; modifications recommended by Applied Biosystems were incorporated for use of the POP-7 polymer (Applied Biosystems Manual P/N: 4367258). The resultant electropherograms (Figure 1) were analysed using GeneMapperID v3.2 software (Applied Biosystems).

Assay validations

Some of the markers used in the SBE assays were validated using a set of control samples, previously screened using RFLP assays. Those markers for which samples of known haplogroup were unavailable were sequenced in order to confirm the presence of the polymorphism. After the screening of the 683 samples, Fisher's exact tests were performed using GraphPad InStat version 3.10 32 bit for Windows (GraphPad Software, CA, USA, http://www.graphpad.com), in order to test significance of differences in haplogroup frequency between the BAN and KS.

Abbreviations
BAN: Bantu speaker; dNTP: deoxyribonucleotide triphosphate; ddNTP: dideoxyribonucleotide triphosphate; ddH20: double distilled water; KS: Khoe-San; PCR: polymerase chain reaction; RFLP: restriction fragment length polymorphism; SAP: shrimp alkaline phosphatase; SAW: South African White; SBE: single base extension; SNP: single nucleotide polymorphism; STR: short tandem repeat.

Acknowledgements
We are grateful to all subjects who participated in our research. This study was supported by grants awarded to: HS from the South African Medical Research Council, the Palaeontological Scientific Trust, the National Research Foundation and the University of the Witwatersrand; HS and TN from the National Health Laboratory Service Research Trust; and TN was supported by the National Research Foundation, the German Academic Exchange Service and the University of the Witwatersrand. We are also grateful to Professor Trefor Jenkins and colleagues in the Division of Human Genetics for assistance with fieldwork and processing of samples.

Authors' contributions
TN conceived the study, participated in its design, carried out screening of samples and drafted the manuscript. CMS and HM were involved in sample collection, screening, and assisted with the manuscript. PP, RM and JCE were involved in sample collection and screening. HS participated in its design and contributed to the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 4 February 2010 Accepted: 1 September 2010
Published: 1 September 2010

References

1. Y Chromosome Consortium: **A nomenclature system for the tree of human Y-chromosomal binary haplogroups.** *Genome Res* 2002, **12**:339-348.
2. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF: **New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree.** *Genome Res* 2008, **18**:830-838.
3. Seielstad MT, Minch E, Cavalli-Sforza LL: **Genetic evidence for a higher female migration rate in humans.** *Nat Genet* 1998, **20**:278-280.
4. Daniels DL, Hall AM, Ballantyne J: **SWGDAM developmental validation of a 19-locus Y-STR system for forensic casework.** *J Forensic Sci* 2004, **49**:668-683.
5. Hammer MF, Behar DM, Karafet TM, Mendez FL, Hallmark B, Erez T, Zhivotovsky LA, Rosset S, Skorecki K: **Extended Y chromosome haplotypes resolve multiple and unique lineages of the Jewish priesthood.** *Hum Genet* 2009, **126(5)**:707-717.
6. Brion M, Sobrino B, Blanco-Verea A, Lareu MV, Carracedo A: **Hierarchical analysis of 30 Y-chromosome SNPs in European populations.** *Int J Legal Med* 2005, **119**:10-15.
7. Lessig R, Zoledziewska M, Fahr K, Edelmann J, Kostrzewa M, Dobosz T, Kleemann WJ: **Y-SNP-genotyping - a new approach in forensic analysis.** *Forensic Sci Int* 2005, **154**:128-136.
8. Onofri V, Alessandrini F, Turchi C, Pesaresi M, Buscemi L, Tagliabracci A: **Development of multiplex PCRs for evolutionary and forensic applications of 37 human Y chromosome SNPs.** *Forensic Sci Int* 2006, **157**:23-35.
9. Sanchez JJ, Børsting C, Hallenberg C, Buchard A, Hernandez A, Morling N: **Multiplex PCR and minisequencing of SNPs--a model with 35 Y chromosome SNPs.** *Forensic Sci Int* 2003, **137**:74-84.
10. Jobling MA, Tyler-Smith C: **The human Y chromosome: an evolutionary marker comes of age.** *Nat Rev Genet* 2003, **4**:598-612.
11. Konrad KD, Pentoney SL Jr: **Contribution of secondary structure to DNA mobility in capillary gels.** *Electrophoresis* 1993, **14**:502-508.
12. Tu O, Knott T, Marsh M, Bechtol K, Harris D, Barker D, Bashkin J: **The influence of fluorescent dye structure on the electrophoretic mobility of end-labeled DNA.** *Nucleic Acids Res* 1998, **26**:2797-2802.
13. Underhill PA, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL: **The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations.** *Ann Hum Genet* 2001, **65**:43-62.
14. Capelli C, Tschentscher F, Pascali VL: **'Ancient' protocols for the crime scene? Similarities and differences between forensic genetics and ancient DNA analysis.** *Forensic Sci Int* 2003, **131**:59-64.
15. Salas A, Bandelt H-J, Macaulay V, Richards MB: **Phylogeographic investigations: the role of trees in forensic genetics.** *Forensic Sci Int* 2007, **168**:1-13.
16. Hammer MF, Blackmer F, Garrigan D, Nachman MW, Wilder JA: **Human population structure and its effects on sampling Y chromosome sequence variation.** *Genetics* 2003, **164**:1495-1509.
17. Adams SM, King TE, Bosch E, Jobling MA: **The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal marker P25 through gene conversion.** *Forensic Sci Int* 2006, **159**:14-20.
18. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, *et al*: **A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes.** *Am J Hum Genet* 2002, **70**:1197-1214.
19. Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonné-Tamir B, Bertranpetit J, Francalacci P, *et al*: **Y chromosome sequence variation and the history of human populations.** *Nat Genet* 2000, **26**:358-361.
20. Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF: **Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes.** *Eur J Hum Genet* 2005, **13**:867-876.
21. Rosser ZH, Zerjal T, Hurles ME, Adojaan M, Alavantic D, Amorim A, Amos W, Armenteros M, Arroyo E, Barbujani G, *et al*: **Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language.** *Am J Hum Genet* 2000, **67**:1526-1543.
22. Miller SA, Dykes DD, Polesky HF: **A simple salting out procedure for extracting DNA from human nucleated cells.** *Nucleic Acids Res* 1998, **16**:1215.
23. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365-386.
24. Vallone PM, Butler JM: **AutoDimer: a screening tool for primer-dimer and hairpin structures.** *Biotechniques* 2004, **37**:226-231.

doi:10.1186/2041-2223-1-6
Cite this article as: Naidoo *et al.*: Development of a single base extension method to resolve Y chromosome haplogroups in sub-Saharan African populations. *Investigative Genetics* 2010 **1**:6.

Submit your next manuscript to BioMed Central and take full advantage of:

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central

7.7 Appendix G

Table A2: PCR reagent mixtures for Multiplex SBE Assays.

PCR reagent	YSNP1	Hg-A	Hg-B	Hg-B2b
	Concentration	Concentration	Concentration	Concentration
DNA	10ng/µl	10ng/µl	10ng/µl	10ng/µl
Buffer (10X)	1X	0.8X	1X	1X
$MgCl_2$ (25mM)	2mM	4mM	3mM	3mM
dNTPs 2.5mM)	300µM	200µM	200µM	200µM
Forward primer (10µM)	See Table 2.2	See Table 2.2	See Table 2.2	See Table 2.2
Reverse primer (10µM)	See Table 2.2	See Table 2.2	See Table 2.2	See Table 2.2
FastStart Taq (5U/µl)	1U	1U	1U	1U
ddH_2O	Made up to 25µl	Made up to 25µl	Made up to 25µl	Made up to 25µl

PCR reagent	Hg-E	Hg-E1b1a	Hg-E1b1b1
	Concentration	Concentration	Concentration
DNA	10ng/µl	10ng/µl	10ng/µl
Buffer (10X)	1.25X	1.25X	1.25X
$MgCl_2$ (25mM	3mM	3mM	3mM
dNTPs (2.5mM)	200µM	200µM	200µM
Forward primer (10µM)	See Table 2.2	See Table 2.2	See Table 2.2
Reverse primer (10µM)	See Table 2.2	See Table 2.2	See Table 2.2
FastStart Taq (5U/µl)	1U	1U	1U
ddH_2O	Made up to 25µl	Made up to 25µl	Made up to 25µl

Table A3: Conditions and concentrations used during Y chromosome RFLP typing

Marker	M11	M130	M175	M74	P12f2
Mutation	A - G	C - T	no del - del	G - A	no del - del
Haplogroup	L-M11	C-M130	O-M175	PQ-M74	J-P12f2
PCR stock solutions					
PCR Buffer (10X)	1X	1X	1X	1X	1X
MgCl2 (25 mM)	3.5 mM	1.5 mM	1.5 mM	1.5 mM	1.5 mM
dNTPs (25 mM)	100 µM	100 µM	100 µM	100 µM	100 µM
primer Forward (10 mM)	0.4 µM	0.4 µM	0.2 µM	0.3 µM	0.3 µM
primer Reverse (10 mM)	0.4 µM	0.4 µM	0.2 µM	0.3 µM	0.3 µM
FastStart Taq (5 U/µl)	1 U	1 U	1 U	1 U	1 U M2-F and M2-R (0.3 uM each)
PCR conditions					
annealing temperature (ºC)	58	50	60	60	58
Digestion					
PCR product size (bp)	215	91	444	385	p12f2 = 88; M2 = 148
restriction enzyme	Msp I	Bsl I	Mbo II	Rsa I	-
digestion conditions (ºC)	37	55	37	37	-
gel detection	3% agarose	3% agarose	2% agarose	3% agarose	2% agarose
ancestral allele - product sizes (bp)	215 (A)	57 + 34 (C)	370 + 74 (no del)	385 (G)	148 + 88 (no del)
derived allele - product sizes (bp)	193 + 22 (G)	91 (T)	444 (del)	195 + 190 (A)	148 (del) (co-amplification with M2)

Table A3 cont.: Conditions and concentrations used during Y chromosome RFLP typing

Marker	M11	M130	M175	M74	P12f2
References					
Reference: polymorphism	Underhill, et al. (1997)	Bergen, et al. (1999)	Shen, et al. (2000)	Shen, et al. (2000)	Casanova, et al. (1985)
Reference: primers	Qamar, et al. (2002)	Kayser, et al. (2000)	Underhill, et al. (2000)	Underhill, et al. (2000)	Rosser, et al. (2000)
Reference: PCR-RFLP assay	Qamar, et al. (2002)	Kayser, et al. (2000)	unpublished	unpublished	-

Table A4: Reagents and conditions used during Y chromosome Taqman® assays M6 and M49

Real-Time		**Volume (µl)**
PCR reagents	TaqMan® SNP Genotyping Mastermix	2.50
	TaqMan® Primer mix	0.25
	DNA (5ng)	1.00
	Water	1.25
		5.00

PCR conditions	**Enzyme Activation**	**40 cycles**	
time	10 min (hold)	15 sec	1 min
temperature	95 °C	95 °C	60 °C

Table A5: Reagents and conditions used during Y-STR screening using AmpFlSTR® YFiler™ PCR Amplification kit

PCR reagents	**Volume (µl)**
DNA sample (1ng/µl)	1.00
AmpflSTR Yfiler PCR reaction mix	2.30
AmpflSTR Yfiler Primer set	1.25
AmpliTaq Gold® DNA polymerase	0.20
ddH_2O	1.50
	6.25

Temperature (and Ramp Speed)	**Time (minutes)**	
95°C	11:00	
94°C (ramp 100%)	01:00	30 cycles
61°C (ramp 100%)	01:00	
72°C (ramp 100%)	01:00	
60°C	80:00	
4°C		∞

Detection is performed on a 3130xl Genetic Analyzer (Life Technologies, CA, USA)

Visualisation Reagents	**Volume (µl)**
Hi-Di® Formamide	8.7
GS500 LIZ	0.3
PCR product OR Ladder	1.0
	10.0

Table A6: Reagents and conditions used during Y-STR screening using PowerPlex® Y System kit

PCR reagents	**Volume (µl)**
DNA sample (1-2 ng/µl)	1.000
PowerPlex® Y Buffer (10X)	0.625
PowerPlex® Y Primer mix	0.625
AmpliTaq Gold® DNA polymerase	0.125
ddH_2O	3.875
	6.250

Temperature (and Ramp Speed)	**Time (minutes)**	
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		∞

Detection is performed on a 3130xl Genetic Analyzer (Life Technologies, CA, USA)

Visualistion Reagents	**Volume (µl)**
Hi-Di® Formamide	8.5
ILS 600	0.5
PCR product	1.0
	10.0

Table A7: Reagent suppliers

Reagent	Supplier
Metaphore® agarose gel	(Cambrex, NJ, USA)
SNaPshot Multiplex Ready Reaction Mix	(Life Technologies, CA, USA)
GS120LIZ	(Life Technologies, CA, USA)
Taqman® Primer mix	(Life Technologies, CA, USA)
TaqMan® SNP Genotyping Mastermix	(Life Technologies, CA, USA)
AmpFlSTR® YFiler™ PCR Amplification Kit	(Life Technologies, CA, USA)
AmpliTaq Gold® DNA polymerase	(Life Technologies, CA, USA)
Hi-Di® Formamide	(Life Technologies, CA, USA)
Exonuclease I (*Exo I*)	(New England Biolabs, MA, USA)
FastStart Taq DNA Polymerase	(Roche, Basel, Switzerland)
PowerPlex® Y System	(Promega, WI, USA)
ILS 600	(Promega, WI, USA)
Shrimp Alkaline Phosphatase (SAP)	(USB, OH, USA)

7.8 Appendix H

Table A8: Fst matrix with p-values (haplogroup frequency data)

	CMAL	CACB	CAU	CKS	DAM	EACB	EACU	EAN	EUR
CMAL	0.000								
CACB	2.453***	0.000							
CAU	0.781***	0.061***	0.000						
CKS	0.335***	0.530***	0.181***	0.000					
DAM	0.499***	0.153***	0.004	0.099**	0.000				
EACB	1.189***	0.032***	0.014*	0.267***	0.044*	0.000			
EACU	0.337***	2.320***	0.775***	0.211***	0.541***	1.058***	0.000		
EAN	0.218***	0.771***	0.356***	0.141***	0.255***	0.532***	0.133***	0.000	
EUR	0.233***	1.351***	0.663***	0.371***	0.505***	0.992***	0.388***	0.290***	0.000
HAD	0.460***	1.311***	0.489***	0.239***	0.364***	0.705***	0.186***	0.198***	0.500***
IND	0.071*	2.146***	0.696***	0.289***	0.438***	1.073***	0.282***	0.179***	0.253***
KBAD	0.481***	1.172***	0.378***	0.246***	0.273***	0.535***	0.538***	0.185***	0.550***
NMBC	1.620***	0.002	0.070*	0.365***	0.153**	0.045	1.517***	0.579***	1.035***
NKS	0.247***	0.723***	0.320***	0.082***	0.188***	0.488***	0.200***	0.144***	0.333***
NWB	1.171***	0.026***	0.011*	0.332***	0.024	0.019***	1.195***	0.657***	1.091***
SAND	0.377***	0.539***	0.172***	0.028*	0.124***	0.242***	0.152***	0.125***	0.417***
SEB	0.720***	0.082***	0.029**	0.171***	0.005	0.043***	0.734***	0.357***	0.732***
SAC	0.163***	0.366***	0.126***	0.051***	0.060**	0.230***	0.227***	0.117***	0.175***
SACB	1.486***	0.003	0.033*	0.329***	0.064*	0.020*	1.435***	0.520***	0.982***
SKS	0.265***	1.218***	0.429***	0.034*	0.253***	0.636***	0.196***	0.145***	0.332***
SWCB	1.086***	0.010	0.014	0.249***	0.024	0.013	1.079***	0.431***	0.818***
SWB	1.494***	0.010*	0.017*	0.392***	0.054*	0.014**	1.476***	0.677***	1.165***
WAM	1.200***	0.003	0.021	0.244***	0.069*	0.001	1.102***	0.462***	0.874***
WAMA	0.735***	0.092***	0.017*	0.138***	0.022	0.023**	0.667***	0.352***	0.682***
WPYG	0.435***	0.791***	0.325***	0.251***	0.242***	0.493***	0.298***	0.215***	0.490***

Table A8 cont.: Fst matrix with p-values (haplogroup frequency data)

	HAD	IND	KBAD	NMBC	NKS	NWB	SAND	SEB
CMAL								
CACB								
CAU								
CKS								
DAM								
EACB								
EACU								
EAN								
EUR								
HAD	0.000							
IND	0.398***	0.000						
KBAD	0.486***	0.407***	0.000					
NMBC	0.889***	1.294***	0.961***	0.000				
NKS	0.165***	0.207***	0.260***	0.518***	0.000			
NWB	0.789***	1.084***	0.455***	0.050*	0.595***	0.000		
SAND	0.149***	0.325***	0.267***	0.372***	0.146***	0.327***	0.000	
SEB	0.498***	0.651***	0.222***	0.103**	0.323***	0.023***	0.188***	0.000
SAC	0.205***	0.149***	0.152***	0.286***	0.091***	0.261***	0.078***	0.135***
SACB	0.851***	1.277***	0.686***	0.034	0.488**	0.007	0.339***	0.037**
SKS	0.302***	0.220***	0.302***	0.744***	0.041**	0.710***	0.131***	0.388***
SWCB	0.673***	0.926***	0.539***	0.054	0.396***	0.000	0.260***	0.023
SWB	0.908***	1.363***	0.638***	0.027	0.618***	0.005*	0.381***	0.043***
WAM	0.691***	1.000***	0.654***	0.029	0.415***	0.020	0.239***	0.058*
WAMA	0.470***	0.657***	0.362***	0.092*	0.316***	0.043	0.128***	0.041***
WPYG	0.026*	0.381***	0.349***	0.578***	0.160***	0.551***	0.177***	0.346***

Table A8 cont.: Fst matrix with p-values (haplogroup frequency data)

	SAC	SACB	SKS	SWCB	SWB	WAM	WAMA	WPYG
CMAL								
CACB								
CAU								
CKS								
DAM								
EACB								
EACU								
EAN								
EUR								
HAD								
IND								
KBAD								
NMBC								
NKS								
NWB								
SAND								
SEB								
SAC	0.000							
SACB	0.240***	0.000						
SKS	0.090***	0.742***	0.000					
SWCB	0.184***	0.000	0.558***	0.000				
SWB	0.289***	0.001	0.857***	0.000	0.000			
WAM	0.209***	0.010	0.560***	0.010	0.005	0.000		
WAMA	0.130***	0.057**	0.373***	0.038*	0.052***	0.023	0.000	
WPYG	0.168***	0.535***	0.332***	0.438***	0.597***	0.470***	0.351***	0.000

Abbreviations:

*significant difference P<0.05

**significant difference P<0.01

***significant difference P<0.001

Table A9: Exact test of population differentiation (haplogroup frequency data)

	CMAL	CACB	CAU	CKS	DAM	EACB	EACU
CACB	0.00000+-0.0000						
CAU	0.00000+-0.0000	0.00017+-0.0002					
CKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000				
DAM	0.00000+-0.0000	0.00051+-0.0004	0.07173+-0.0116	0.00034+-0.0003			
EACB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000		
EACU	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	
EAN	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
EUR	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
HAD	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
IND	0.00494+-0.0013	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
KBAD	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
NMBC	0.00000+-0.0000	1.00000+-0.0000	0.92937+-0.0035	0.00025+-0.0002	0.24842+-0.0061	0.85531+-0.0098	0.00000+-0.0000
NKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
NWB	0.00000+-0.0000	0.00142+-0.0003	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SAND	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SEB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00145+-0.0008	0.00000+-0.0000	0.00000+-0.0000
SAC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.17141+-0.0131	0.00000+-0.0000	0.00000+-0.0000
SACB	0.00000+-0.0000	0.50709+-0.0071	0.00041+-0.0001	0.00000+-0.0000	0.00269+-0.0008	0.00085+-0.0004	0.00000+-0.0000
SKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.02807+-0.0033	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SWCB	0.00000+-0.0000	0.44917+-0.0074	0.18435+-0.0106	0.00000+-0.0000	0.22367+-0.0081	0.02680+-0.0050	0.00000+-0.0000
SWB	0.00000+-0.0000	0.20119+-0.0099	0.00008+-0.0001	0.00000+-0.0000	0.00120+-0.0007	0.00000+-0.0000	0.00000+-0.0000
WAM	0.00000+-0.0000	0.12619+-0.0070	0.45179+-0.0112	0.00013+-0.0001	0.00952+-0.0013	0.48000+-0.0138	0.00000+-0.0000
WAMA	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000

Table A9 cont.: Exact test of population differentiation (haplogroup frequency data)

	EAN	EUR	HAD	IND	KBAD	NMBC
CACB						
CAU						
CKS						
DAM						
EACB						
EACU						
EAN						
EUR	0.00000+-0.0000					
HAD	0.00000+-0.0000	0.00000+-0.0000				
IND	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000			
KBAD	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000		
NMBC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00005+-0.0000	
NKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
NWB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.68117+-0.0203
SAND	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00089+-0.0002
SEB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00418+-0.0010	0.39165+-0.0178
SAC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00116+-0.0007	0.05665+-0.0097
SACB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.51601+-0.0029
SKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SWCB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.48744+-0.0025
SWB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.96127+-0.0040
WAM	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.70627+-0.0028
WAMA	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.55969+-0.0107
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00040+-0.0003	0.00000+-0.0000	0.00000+-0.0000	0.00002+-0.0000

Table A9 cont.: Exact test of population differentiation (haplogroup frequency data)

	NKS	NWB	SAND	SEB	SAC	SACB	SKS
CACB							
CAU							
CKS							
DAM							
EACB							
EACU							
EAN							
EUR							
HAD							
IND							
KBAD							
NMBC							
NKS							
NWB	0.00000+-0.0000						
SAND	0.00000+-0.0000	0.00000+-0.0000					
SEB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000				
SAC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000			
SACB	0.00000+-0.0000	0.48060+-0.0178	0.00000+-0.0000	0.27232+-0.0118	0.00000+-0.0000		
SKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00058+-0.0004	0.00000+-0.0000	
SWCB	0.00000+-0.0000	0.80420+-0.0083	0.00000+-0.0000	0.33088+-0.0164	0.00034+-0.0003	0.62513+-0.0076	0.00000+-0.0000
SWB	0.00000+-0.0000	0.00011+-0.0001	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.37234+-0.0128	0.00000+-0.0000
WAM	0.00000+-0.0000	0.00602+-0.0020	0.00079+-0.0005	0.01399+-0.0024	0.00491+-0.0011	0.02470+-0.0020	0.00000+-0.0000
WAMA	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000

Table A9 cont.: Exact test of population differentiation (haplogroup frequency data)

	SWCB	SWB	WAM	WAMA
CACB				
CAU				
CKS				
DAM				
EACB				
EACU				
EAN				
EUR				
HAD				
IND				
KBAD				
NMBC				
NKS				
NWB				
SAND				
SEB				
SAC				
SACB				
SKS				
SWCB				
SWB	0.62088+-0.0144			
WAM	0.05521+-0.0032	0.05385+-0.0073		
WAMA	0.00099+-0.0006	0.00000+-0.0000	0.87150+-0.0037	
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000

Table A10: Rst matrix with p-values (STR data)

	CMAL	CACB	CAU	CKS	DAM	EACB	EACU	EAN	EUR
CMAL	0.000								
CACB	1.352***	0.000							
CAU	0.655***	0.025**	0.000						
CKS	0.298***	0.311***	0.144***	0.000					
DAM	0.478***	0.113***	0.023	0.087**	0.000				
EACB	0.962***	0.022**	0.032**	0.327***	0.100**	0.000			
EACU	0.215***	0.405***	0.152***	0.037	0.113**	0.291***	0.000		
EAN	0.269***	0.339***	0.163***	0.061***	0.179***	0.301***	0.000	0.000	
EUR	0.068*	1.268***	0.754***	0.316***	0.622***	1.045***	0.318***	0.306***	0.000
HAD	0.488***	0.373***	0.166***	0.095***	0.082**	0.334***	0.100**	0.148***	0.567***
IND	0.016	0.922***	0.448***	0.173***	0.296***	0.679***	0.117**	0.163***	0.171***
KBAD	0.499***	0.343***	0.165***	0.126**	0.226***	0.246***	0.095*	0.068*	0.606***
NMBC	1.363***	0.000	0.031	0.266***	0.122**	0.015	0.311***	0.320***	1.409***
NKS	0.335***	0.249***	0.152***	0.118***	0.034*	0.295***	0.188***	0.295***	0.436***
NWB	0.792***	0.010**	0.008	0.230***	0.062**	0.018***	0.237***	0.243***	0.808***
SAND	0.244***	0.142***	0.038**	0.008	0.011	0.155***	0.024	0.052***	0.345***
SEB	0.534***	0.047***	0.015*	0.106***	0.048*	0.053***	0.102**	0.113***	0.571***
SAC	0.176***	0.239***	0.101***	0.025*	0.068**	0.230***	0.024	0.041***	0.206***
SACB	1.118***	0.001	0.020***	0.253***	0.087**	0.009	0.305***	0.278***	1.165***
SKS	0.385***	0.192***	0.094***	0.007	0.052*	0.230***	0.061*	0.108***	0.560***
SWCB	0.804***	0.045**	0.023	0.100**	0.065*	0.092***	0.180***	0.174***	0.851***
SWB	0.942***	0.034***	0.019**	0.162***	0.033*	0.072***	0.256***	0.250***	0.867***
WAM	0.907***	0.014	0.000	0.121***	0.064*	0.048*	0.161***	0.156***	0.960***
WAMA	0.792***	0.029**	0.009	0.137***	0.063**	0.054***	0.178***	0.163***	0.808***
WPYG	0.250***	0.182***	0.079***	0.103***	0.050*	0.140***	0.063*	0.102***	0.432***

Table A10 cont.: Rst matrix with p-values (STR data)

	HAD	IND	KBAD	NMBC	NKS	NWB	SAND	SEB
CMAL								
CACB								
CAU								
CKS								
DAM								
EACB								
EACU								
EAN								
EUR								
HAD	0.000							
IND	0.325***	0.000						
KBAD	0.260***	0.201**	0.000					
NMBC	0.307***	0.808***	0.516	0.000				
NKS	0.089***	0.242***	0.235	0.181***	0.000			
NWB	0.293***	0.548***	0.152	0.022	0.300***	0.000		
SAND	0.055**	0.145**	0.073*	0.112**	0.086***	0.098***	0.000	
SEB	0.187***	0.337***	0.067*	0.066*	0.221***	0.016***	0.034**	0.000
SAC	0.128***	0.088**	0.059*	0.239***	0.154***	0.160***	0.011	0.068***
SACB	0.314***	0.705***	0.317	0.005	0.194***	0.006	0.102***	0.037**
SKS	0.100***	0.230***	0.134**	0.128**	0.096***	0.156***	0.008	0.072***
SWCB	0.214***	0.486***	0.167**	0.072*	0.134***	0.022*	0.035*	0.013
SWB	0.240***	0.636***	0.227	0.062*	0.175***	0.022***	0.062***	0.024***
WAM	0.185***	0.538***	0.223	0.027	0.149***	0.008	0.031	0.007
WAMA	0.218***	0.510***	0.123**	0.051*	0.197***	0.009*	0.048**	0.003
WPYG	0.141***	0.120**	0.104**	0.163***	0.151***	0.121***	0.035**	0.068***

Table A10 cont.: Rst matrix with p-values (STR data)

	SAC	SACB	SKS	SWCB	SWB	WAM	WAMA	WPYG
CMAL								
CACB								
CAU								
CKS								
DAM								
EACB								
EACU								
EAN								
EUR								
HAD								
IND								
KBAD								
NMBC								
NKS								
NWB								
SAND								
SEB								
SAC	0.000							
SACB	0.189***	0.000						
SKS	0.067**	0.145***	0.000					
SWCB	0.096***	0.061**	0.042*	0.000				
SWB	0.139***	0.043***	0.089***	0.005	0.000			
WAM	0.110***	0.024	0.049*	0.000	0.013	0.000		
WAMA	0.104***	0.034**	0.083***	0.000	0.012**	0.000	0.000	
WPYG	0.051***	0.105***	0.089***	0.121***	0.145***	0.097***	0.108***	0.000

Abbreviations:

*significant difference	P<0.05
**significant difference	P<0.01
***significant difference	P<0.001

Table A11: Exact test of population differentiation (STR data)

	CMAL	CACB	CAU	CKS	DAM	EACB	EACU
CACB	0.00000+-0.0000						
CAU	0.00000+-0.0000	0.00000+-0.0000					
CKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000				
DAM	0.01829+-0.0042	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000			
EACB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000		
EACU	0.00012+-0.0001	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	
EAN	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
EUR	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
HAD	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
IND	0.05979+-0.0050	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.01929+-0.0027	0.00000+-0.0000	0.00010+-0.0001
KBAD	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
NMBC	0.05612+-0.0044	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.01992+-0.0039	0.00000+-0.0000	0.00001+-0.0000
NKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
NWB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SAND	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SEB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SAC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SACB	0.00006+-0.0001	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SWCB	0.00056+-0.0002	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00005+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SWB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WAM	0.01724+-0.0024	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00196+-0.0006	0.00000+-0.0000	0.00000+-0.0000
WAMA	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000

Table A11 cont.: Exact test of population differentiation (STR data)

	EAN	EUR	HAD	IND	KBAD	NMBC
CACB						
CAU						
CKS						
DAM						
EACB						
EACU						
EAN						
EUR	0.00000+-0.0000					
HAD	0.00000+-0.0000	0.00000+-0.0000				
IND	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000			
KBAD	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000		
NMBC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.05419+-0.0032	0.00000+-0.0000	
NKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
NWB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SAND	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SEB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SAC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00016+-0.0002	0.00000+-0.0000	0.00000+-0.0000
SACB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SWCB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00195+-0.0009	0.00000+-0.0000	0.00081+-0.0005
SWB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WAM	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.01867+-0.0024	0.00000+-0.0000	0.01348+-0.0028
WAMA	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000

Table A11 cont.: Exact test of population differentiation (STR data)

	NKS	NWB	SAND	SEB	SAC	SACB	SKS
CACB							
CAU							
CKS							
DAM							
EACB							
EACU							
EAN							
EUR							
HAD							
IND							
KBAD							
NMBC							
NKS							
NWB	0.00000+-0.0000						
SAND	0.00000+-0.0000	0.00000+-0.0000					
SEB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000				
SAC	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000			
SACB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000		
SKS	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	
SWCB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
SWB	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WAM	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WAMA	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000

Table A11 cont.: Exact test of population differentiation (STR data)

	SWCB	SWB	WAM	WAMA
CACB				
CAU				
CKS				
DAM				
EACB				
EACU				
EAN				
EUR				
HAD				
IND				
KBAD				
NMBC				
NKS				
NWB				
SAND				
SEB				
SAC				
SACB				
SKS				
SWCB				
SWB	0.00000+-0.0000			
WAM	0.00000+-0.0000	0.00000+-0.0000		
WAMA	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	
WPYG	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000	0.00000+-0.0000