



Nº ordem 17/D/2007

## TESE DE DOUTORAMENTO

apresentada na

UNIVERSIDADE DA MADEIRA

Para obtenção do grau de Doutor

Alexandra Rosa

Phylogenetic structure of Guinea-Bissau ethnic groups for mitochondrial DNA and Y  
chromosome genetic systems

Júri:

Reitor da Universidade da Madeira

Doutor António Salas Ellacuriaga, Universidade de Santiago de Compostela

Doutor Richard Villems, Universidade de Tartu

Doutora Ornella Semino, Universidade de Pavia

Doutor Duarte Nuno Pessoa Vieira, Universidade de Coimbra

Doutor António Manuel Dias Brehm, Universidade da Madeira

Doutora Maria Manuela de Medeiros Lima, Universidade dos Açores





Nºordem 17/D/2007

## TESE DE DOUTORAMENTO

apresentada na  
UNIVERSIDADE DA MADEIRA

Para obtenção do grau de Doutor

Alexandra Rosa

Phylogenetic structure of Guinea-Bissau ethnic groups for mitochondrial DNA and Y  
chromosome genetic systems

Júri:

Reitor da Universidade da Madeira

Doutor António Salas Ellacuriaga, Universidade de Santiago de Compostela

Doutor Richard Villems, Universidade de Tartu

Doutora Ornella Semino, Universidade de Pavia

Doutor Duarte Nuno Pessoa Vieira, Universidade de Coimbra

Doutor António Manuel Dias Brehm, Universidade da Madeira

Doutora Maria Manuela de Medeiros Lima, Universidade dos Açores



Laboratório de Genética Humana



UNIVERSITY OF TARTU  
INSTITUTE OF MOLECULAR AND CELL BIOLOGY

**FCT** Fundação para a Ciência e a Tecnologia

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E ENSINO SUPERIOR Portugal





Scientific supervisors:

Professor António Brehm – PhD

Professor of Genetics at the University of Madeira

Director of the Human Genetics Laboratory - University of Madeira

Vice-Rector of the University of Madeira

Professor Richard Villems – PhD, Doctor of Sciences

Professor of Archaeogenetics, Department of Evolutionary Biology at  
the Institute of Molecular and Cell Biology - University of Tartu

President, Estonian Academy of Sciences

The current dissertation is based on the following publications referred to in the text:

**Rosa A**, Brehm A, Kivisild T, Metspalu E, Villems R(2004). MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region. *Ann Hum Gen* 68: 340-352.

**Rosa A**, Ornelas C, Brehm A, Villems R (2006) Population data on 11 STRs from Guiné-Bissau. *Forensic Sci Int* 157: 210-217.

**Rosa A**, Ornelas C, Jobling MA, Brehm A, Villems R (2007) Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol Biol* 7: 124.

(appended in Supplementary material)



## **Index**

<i>Abbreviations</i>	7
<i>Definition of basic terms and concepts</i>	8
<i>Abstract/Resumo</i>	11
<b>Chapter One – Introduction</b>	13
<b>Chapter Two - Literature overview</b>	15
1 – Basic concepts of the phylogenetic and phylogeographic approaches	15
1.1 - Phylogenetic trees and networks	15
1.1.1 - Determination of a phylogenetic root	17
1.2 - Calibration of the molecular clock and coalescence estimates	18
1.3 - DNA polymorphisms as molecular markers for phylogenetic studies	19
1.4 - Phylogeographic perspective of human uniparental genetic systems	20
2 – Mitochondrial DNA phylogenetic analysis	22
2.1 - Structure and organization of mtDNA	22
2.2 - On the origin of mitochondria and mtDNA	25
2.3 – Distinctive features of mtDNA	26
2.3.1 – Maternal inheritance	26
2.3.2 - Non-recombining transmission and homoplasmy	26
2.3.3 – Mutation rate, homoplasmy and multiple hits	28
2.4 – The role of selection in the mtDNA genome	29
2.5 – Calibration of the mtDNA molecular clock	30
2.6 – Phylogenetic classification and nomenclature of mtDNA haplogroups	32
2.7 - Worldwide variation of mtDNA haplogroups uncovers origins and settling processes	35
2.7.1 – Phylogeography of the African mtDNA variation	41
3 – Phylogenetic analysis of the Y chromosome	50
3.1 - Structure and organization of the Y chromosome	50
3.2 - Evolution of the Y chromosome	53
3.3 – Distinctive features of the Y chromosome	55
3.3.1 - Haploidy and paternal inheritance	55
3.3.2 - Absence of recombination on the NRY	56
3.3.3 – The role of selection in the Y chromosome	57
3.4 - Calibration of the Y chromosome molecular clock	57
3.5 - Phylogeny and nomenclature of Y chromosome haplogroups	60
3.6 – The origin and worldwide dispersal of Y chromosomes	63

3.6.1 – Phylogeography of the African paternal variation	68
4 – Complementary sources of evidence	76
4.1 - Environmental records	76
4.2 - Archaeology and anthropology in the pre-history	78
4.3 - Historical and ethnical background	82
4.4 - Linguistic affiliation	85
4.5 - Records of autosomal genetic systems	87
<b>Chapter Three - Aims of the study</b>	<b>89</b>
<b>Chapter Four - Material and Methods</b>	<b>91</b>
5 – Sampling procedure	91
6 – DNA typing	91
6.1 - DNA extraction	91
6.2 - PCR amplification	92
6.3 – Electrophoresis on agarose and polyacrylamide gels	93
6.4 – Purification of PCR products	93
6.5 – Automatic sequencing	94
6.6 – Typing of Y chromosome microsatellites	94
6.7 - Typing of Restriction Fragment Length Polymorphisms – RFLPs	95
7 - Data analysis	95
7.1 - Phylogenetic assignment	95
7.2 - Definition of populational units	96
7.3 - Phylogenetic networks	96
7.4 - Coalescence time estimates	97
7.5 – Haplotype exact matches	98
7.6 - Statistical parameters	98
7.6.1 - Frequency calculation	98
7.6.2 - Genetic diversity	99
7.6.3 – $F_{ST}$ statistics	99
7.6.4 - Exact test of population differentiation	100
7.6.5 – Graphical display of results – PCA	100
7.6.6 - Analysis of MOlecular VAriance (AMOVA)	101
7.6.7 – Mismatch distribution	101
7.6.8 - Neutrality tests	101

<b>Chapter Five - Results and Discussion</b>	103
8 - MtDNA analysis in the population of Guinea-Bissau – a phylogenetic approach	103
8.1 - Principal Component Analysis	116
8.2 - Analysis of Molecular Variance	119
9 - Statistical parameters from mtDNA nucleotidic sequences	119
9.1 - Mismatch distribution	121
10 - A phylogenetic perspective of Y chromosome pool in Guinea-Bissau population	123
10.1 - Principal Component Analysis	131
10.2 - Analysis of Molecular Variance	133
10.3 – Statistical parameters for Y chromosome microsatellite variation	134
11 - Combined analysis of Y chromosome and mtDNA haplogroups	136
12 - Analysis of autosomal genetic markers in Guinea-Bissau ethnic groups	138
<b>Chapter Six – Conclusions</b>	141
<i>References</i>	149
<i>Acknowledgements</i>	171
<i>Supplementary material</i>	173



## Abbreviations

AMH	Anatomically Modern Human
bp/kb(p)/Mb(p)	base pair / thousand (kilo) base pairs / mega (million) base pairs
<i>ca.</i>	<i>circa</i> , about
(r)CRS	(revised) Cambridge Reference Sequence
D-loop	displacement loop/control region of mtDNA
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
HVS-I/HVS-II	first/second hypervariable segment of mtDNA
Indel(s)	polymorphism of insertion-deletion
ky(a)	thousand/kilo years (ago)
LGM	Last Glacial Maximum
(T)MRCA	(time to the) Most Recent Common Ancestor
mtDNA	mitochondrial DNA
my(a)	million years (ago)
np(s)	nucleotide position(s)
PCR	Polymerase Chain Reaction
RFLP	Restriction Fragment Length Polymorphism
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
rRNA	ribosomal ribonucleic acid
tRNA	transfer ribonucleic acid

Definition of basic terms and concepts:

Allele(s) – Variant recognizable form(s) of a gene or DNA sequence at a specific chromosomal location.

Bottleneck – The reduction of genetic diversity that results from a drastic reduction in population size.

Clade – An evolutionary branch.

Cladistics – The science of reconstructing evolutionary relationships by identifying common ancestors through the sharing among taxa of “derived” characteristics, rather than the sharing of “ancestral” characteristics.

Coalescence time (age) - Time, sufficient to generate observed genetic variation of a phylogenetic tree, as a rule somewhat shorter than time when particular clade arose; coalescence age is usually considered as a time since the beginning of expansion of a monophyletic clade.

Effective population size ( $N_e$ ) - The number of adults contributing gametes to the next generation, on average a third of the actual size (“census” size) of a human population in present, according to the empirical principle (see, e.g. Cavalli-Sforza *et al.* 1994).

Fixation - The process by which one allele increases in a population until all other alleles go extinct and the locus becomes monomorphic.

Founder effect – Reduced genetic diversity in a population founded by a small number of individuals.

Founder haplotype - Common ancestral haplotype to which all haplotypes under concern coalesce to.

Genetic drift – Random fluctuations in the frequency of haplotypes in a finite population owed to stochastic sampling from one generation to the next, which may accelerate differentiation of groups, in particular in small populations.

Genetic Marker - Random mutations in the DNA sequence which act as genetic milestones.

Haplotype – The combination of allelic states of a set of polymorphic markers lying on the same DNA molecule, e.g. a chromosome or region of a chromosome. Of mtDNA (= lineage), sequence footprint for the characterized polymorphisms encompassing all identical sequences. Of Y chromosome,



defined by the pattern of length variation of STRs of a particular chromosome. The difference of a single genetic marker delineates a distinct haplotype.

Haplogroup - Monophyletic clade of haplotypes that share characteristic sequence polymorphisms (genetic mutations or “markers”), and derive from a single ancestral founder haplotype. It is usually defined by relatively slowly mutating markers and thus has more phylogenetic stability than haplotypes.

Homoplasy – Sharing of identical character states that cannot be explained by inheritance from the common ancestor of a group of taxa.

Lineage – A group of taxa sharing a common ancestor to the exclusion of other taxa.

Monophyletic – Relating to a clade, consisting of an ancestor and all of its descendants.

Mutation - Transmission error in DNA, fixed after the replication of DNA.

Natural selection – evolutionary process of differential contribution of individuals to the following generations, favouring the transmission of beneficial mutations and limiting the transmission of deleterious ones.

Paraphyletic – A grouping that shares a common ancestor to the exclusion of many other lineages but does not include all descendants of that common ancestor.

Parsimony – The principle that the best explanation is that which requires the least number of causal factors.

Phylogeny – Representation of the origin and evolution of a set of organisms or lineages, where the ancestral relationships and pathways of transmission from parents to offsprings are depicted.

Phylogeography - Analysis of the geographical distribution of the different clades of a phylogeny.

Recombination – The emergence of new combinations of alleles due to meiotic crossing-over.

Star-like phylogeny - Phylogeny of a set of sequences that follows, in their length distribution, a Poisson mode of distribution and coalesces into an ancestral haplotype, i.e. each extant taxon is derived independently from the common ancestor of all taxa.



## Abstract

The maternal and paternal genetic profile of Guineans is markedly sub-Saharan West African, with the majority of lineages belonging to L0-L3 mtDNA sub-clusters and E3a-M2 and E1-M33 Y chromosome haplogroups. Despite the sociocultural differences among Guinea-Bissau ethnic groups, marked by the supposedly strict admixture barriers, their genetic pool remains largely common. Their extant variation coalesces at distinct timeframes, from the initial occupation of the area to later inputs of people. Signs of recent expansion in mtDNA haplogroups L2a-L2c and NRY E3a-M2 suggest population growth in the equatorial western fringe, possibly supported by an early local agricultural centre, and to which the Mandenka and the Balanta people may relate. Non-West African signatures are traceable in less frequent extant haplogroups, fitting well with the linguistic and historical evidence regarding particular ethnic groups: the Papel and Felupe-Djola people retain traces of their putative East African relatives; U6 and M1b among Guinea-Bissau Bak-speakers indicate partial diffusion to Sahel of North African lineages; U5b1b lineages in Fulbe and Papel represent a link to North African Berbers, emphasizing the great importance of post-glacial expansions; exact matches of R1b-P25 and E3b1-M78 with Europeans likely trace back to the times of the slave trade.

## Resumo

O perfil genético materno e paterno dos Guineenses é característico das populações do Oeste sub-Sahariano, observando-se que a maioria das linhagens pertence aos haplogrupos mitocondriais L0-L3 e do cromossoma Y E3a-M2 e E1-M33. Apesar das diferenças sócio-culturais entre os grupos étnicos da Guiné-Bissau, marcadas por supostas barreiras de miscigenização, a sua estrutura genética é semelhante. A diversidade actual coalesce em distintas escalas temporais, desde a colonização inicial da área aos subseqüentes episódios imigratórios. Índícios de expansão recente nos haplogrupos mitocondriais L2a-L2c e NRY E3a-M2 sugerem um crescimento populacional no Oeste Africano, possivelmente suportado por um centro local de agricultura, no qual os Mandenka e Balanta podem ter desempenhado um papel preponderante. Haplogrupos menos frequentes representam influências não Oeste africanas, corroborando as evidências linguísticas e históricas de determinados grupos: os Papel e Felupe-Djola retêm traços genéticos de possíveis origens Este Africanas; as linhagens U6 e M1b, presentes em membros da família linguística Bak, indicam difusão parcial de linhagens Norte Africanas; mtDNAs U5b1b encontrado em Fulbes e Papel sugerem uma ligação aos Berberes, e enfatizam a importância das expansões pós-glaciais; haplotipos R1b-P25 e E3b1-M78, observados em Europeus e idênticos aos encontrados em Guineenses, possivelmente remontam ao tráfico de escravos.



## Chapter One

### Introduction

The mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome constitute haploid and non-recombining genetic systems of uniparental inheritance, whose particular features makes them valuable molecular records. In the last decades, both systems have proved to be powerful tools for investigating the demographic history of humankind, assuming that the present variability can unveil episodes in which our maternal and paternal ancestors were involved.

Throughout time, the main molecular differentiation of anatomically modern humans occurred during their processes of dispersal into different continents and regions, and therefore the subsets of variation tend to be associated to particular geographic areas and populations. The phylogeographic approach is then applied in order to better understand past demographic phenomena such as range expansion, genetic drift (founder-effects and bottlenecks) and population subdivision. The lineage-based approach attempts to unravel the history of genetic lineages of shared ancestry (known as haplogroups), while the population-based approach focuses on the prehistory of individual populations, geographical regions or on population migrations, by using human population groups as the unit of study.

The history of uniparental systems sheds light over a novel dimension but may not accurately reflect the history of a population because of drift effects in a single locus. Indeed, sex-specific scenarios may rather diverge than converge because of differential demographic histories throughout time. It has therefore become clear that studies of mtDNA variation need to be complemented with data on the male-specific Y chromosome, and ideally with autosomal data. In that sense, researchers are able to unravel socio-cultural effects that might have influenced the extant pool such as polygyny, the effects of matrilocality versus patrilocality or the social stratification dictated by ethnolinguistic affiliation. This is particularly important when dealing with African populations deeply-structured on ethnic social constraints.

The particular interest of our work is to understand the origin and rise of the genetic diversity of the main ethnic groups in Guinea-Bissau, together with the processes that might have shaped it. The first section gives us an overview of the basic principles of the phylogenetic and phylogeographic approaches. A description of particular features of both mtDNA and Y chromosome genetic systems follows, driving the reader to the central focus of the literature overview - the general phylogenetic topology and worldwide variation of mtDNA and Y chromosome haplogroups. The analysis is nevertheless multidisciplinary and is performed in the context of genetic, linguistic, historical and demographic evidences, described in more detail in the last part of the overview. The synthesis will serve as background for to interpret the variation in the studied groups.



## Chapter Two

### Literature overview

#### 1 – Basic concepts of the phylogenetic and phylogeographic approaches

##### 1.1 - Phylogenetic trees and networks

A phylogenetic approach (from the Greek: *phylon* = race and *genetic* = birth) is the classification of taxa based on how closely they are related in evolutionary terms. The existing variation and pattern of relationship of lineages are expressed by the construction of phylogenetic trees as attempts to arrange and order the evolutionary relationships between different variants in a relevant and meaningful way (for in-depth descriptions see <sup>Li 1997, Page and Holmes 1998, Graur and Li 2000, Salemi and Vandamme 2003</sup>). Individual differences at the molecular level of DNA sequences can constitute the raw information to relate the entities. Each coalescent node with descendants represents the hypothetical or “real” Most Recent Common Ancestor (MRCA) of the divergent lines (the latter when characters evolve sufficiently fast to be tracked in extant populations), where edge lengths are proportional to divergence time. A character state is assigned to each node, and each split partitions a set of sequences into two mutually exclusive sets. The tree building requires determining the tree topology (branching order and, if of interest, determination of a root), the evolutionary time (branch lengths) and the ancestral types, as the overall likelihood and reliability. There are three main methods of constructing phylogenetic trees from molecular data: distance-based such as neighbour-joining (NJ; Saitou and Nei 1987, Studier and Keppler 1988), parsimony-based such as maximum parsimony (MP; Fitch 1971, 1977; Swofford 1993), and character-state-based such as maximum likelihood (ML; <sup>Felsenstein 1988</sup>). Traditional tree building methods are however of limited success in human DNA analysis due to intraspecific short distance between individuals. The evolution of DNA characters may well be polytomous at many branching points, what in tree building would be reduced to dichotomous earlier branching or artificially resolved polytomies (as in NJ and MP, respectively; <sup>Bandelt *et al.* 2000, Graur and Li 2000</sup>). Homoplasies (similar characters produced by convergent evolution) and variable substitution rates among sites also create incompatibilities in the classical trees: a high number of equally likely topologies can be drawn from the same dataset and conflicts are mostly solved by chance, so that one can select an incorrect topology.

All the plausible trees are better summarized in networks where links connect the nodes under the assumption of maximum parsimony <sup>(Page and Holmes 1998, Graur and Li 2000, Salemi and Vandamme 2003)</sup>. These are expected to better model reality since the actual evolutionary history may not be particularly tree-like. The homoplastic parallel mutations and the state reversal of characters are represented by reticulations

- equally likely pathways of evolution that unite divergent haplotypes. Fast algorithms of network construction have been set to remove the network's least likely links from all the generated possible trees (reduced-median, RM; applied to studies of mtDNA by <sup>Bandelt et al. 1995</sup>). The algorithm is applied in a hypothetical sequential decomposition of informative characters, e.g. partitioning the groups of haplotypes character-by-character, where correlated sites across haplotypes are combined in one character. A binary matrix of presence or absence of the mutation compared to a reference sequence is built from which a network with 0-1 vectors will be constructed (Figure 1). Each reduction step employs parsimony and frequency criteria, allied to the knowledge of site mutation rate <sup>(Bandelt et al. 2000)</sup>. This approach, usually applied for small samples sizes ( $n < 50$ ) contains all the equally likely trees and can assist in identifying sequencing errors, which manifest themselves in implausible network substructures <sup>(Bandelt et al. 2002)</sup>.



Figure 1 – One-step subnetworks of the vectors 000, 011, 101: a) six-linked network with three unobserved intermediate nodes (in black); b) generation of a median vector (001) by parsimony criteria. *In* Bandelt et al. <sup>(1995)</sup>.

An alternative to limit the levels of reticulation in multi-state markers and large datasets (several hundreds) is the median-joining method (Figure 2; <sup>Bandelt et al. 1999</sup>). The algorithm is based on selective inclusion of the most likely trees in a minimum spanning network. Unobserved hypothetical nodes can be added to shorten the overall length and make it closer to the most parsimonious. The difficulties in determining the most parsimonious phylogeny might be also overwhelmed by assigning different relative weights to the mutations <sup>(Richards et al. 1998b, Helgason et al. 2000)</sup>, associated to their occurrence rates, and essentially by a good level of resolution in the basal topology <sup>(Torrioni et al. 1996, Macaulay et al. 1999b, Bandelt et al. 2000, Chen et al. 2000, Kivisild et al. 2002)</sup>.

A step-by-step guide to hand construction of median networks is given by Bandelt et al. <sup>(2000)</sup>, in a pre- and post-processing parameterized strategy entitled “speedy construction and greedy reduction”. At the same time, the use of MJ after RM algorithm is advised, where RM reduces the homoplasy in the matrix and the reticulation in the network. The tree is built from “cliques”, sets that include all the pairwise compatible characters. The ancestral-derived states are taken from the median majority consensus, with priority defined by relative mutation rate of the sites. Its speedy construction



involves selection of extreme characters, as the terminal links and the ones that determine cubes. The initial central node contains all the sampled types, but with the analysis of a character,  $n$  sequences are popped out. The characters that could fit without additional recurrent evolutionary paths are placed at first. The shelling procedure amounts for collapsing peripheral types into the central node, by

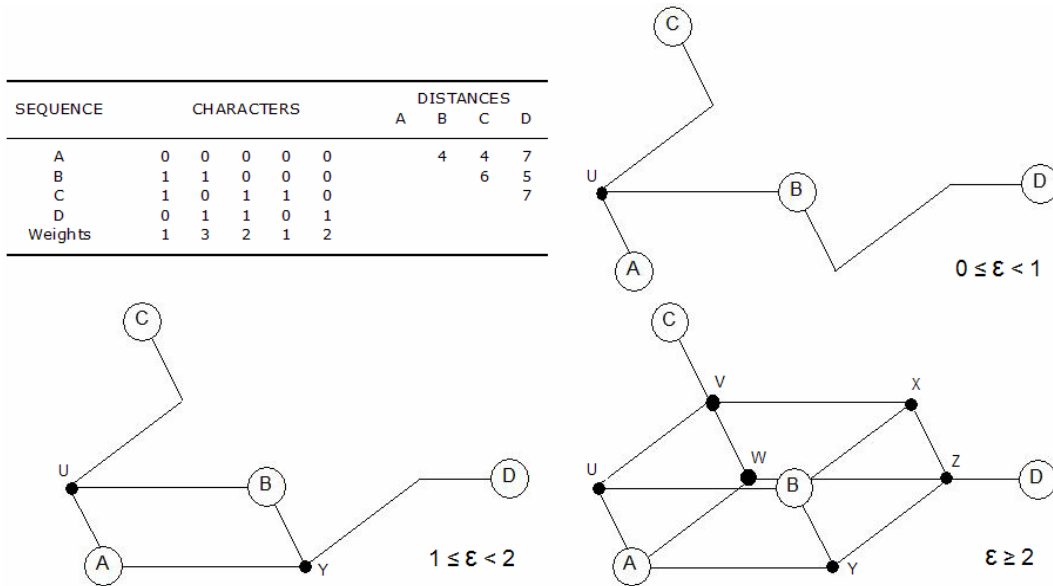


Figure 2 – Median-joining networks constructed from the data of the table with three different settings of the parameter  $\epsilon$ ; inferred sequence types U, V, W, X, Y, and Z are added to the growing network as median vectors. Adapted from Bandelt *et al.* <sup>(1999)</sup>.

partitioning the characters, and ends when one it is not possible to discern any of the two states and a sequence remains as the median. “Greedy reduction” involves a final post-processing that undoes excessively recurrent cases by operating in non-peripheral clades, but not in cliques.

### 1.1.1 - Determination of a phylogenetic root

In the unrooted trees given by NJ and MP tree-building methods, the ancestor is taken as unknown and no evolutionary relationships are assumed between members, thus having no relation to a timeline. On the other hand, a rooted phylogenetic tree is an evolutionary directed tree, with a unique taxa defined as the MRCA (Figure 3). The temporal stratification of branching events is possible since taxa are oriented with respect to evolutionary time, defining relationships from ancestral to descendant divergent nodes and leaves of the tree. The root is specified by means of an outgroup, known to have separated from the common phylogenetic lineage before the existence of the MRCA of the group

under study. In human DNA studies it is common to use chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla*) or *Neanderthal* sequences, or alternatively a subset of the intraspecific variation. If external evidence is not available, a node in the tree from which the distance to all terminal nodes is minimal, is specified as the midpoint (mid-point rooting approach), having as principal a constant evolutionary rate.

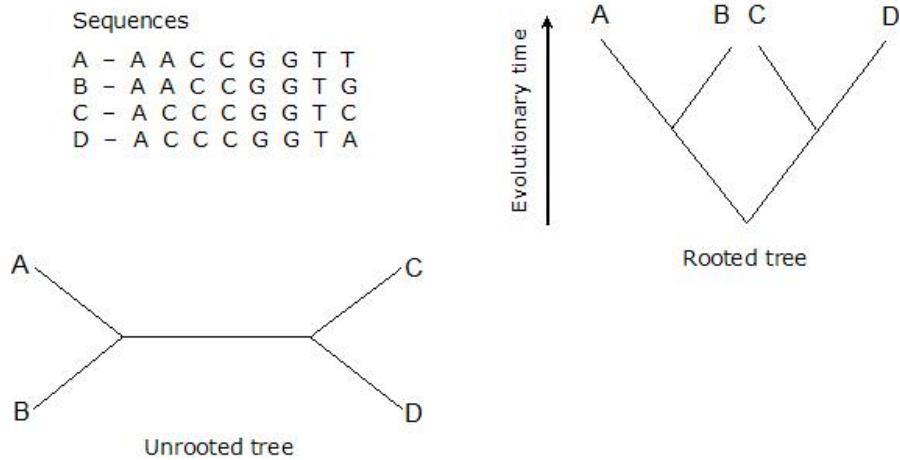


Figure 3 - Unrooted and rooted trees generated from sequence A-D dataset. Adapted from Page and Holmes (1998).

## 1.2 - Calibration of the molecular clock and coalescence estimates

The introduction of molecular data with the advent of genetic markers made it possible to calculate divergence times, vital for reconstructing and interpreting the origin and demographic history of species. DNA studies reveal then an independent chronology for combining with non-genetic sources of evidence (as in Forster 2004). The molecular clocks, an abstraction of early observations on protein evolution (e.g. Gillespie 1991), are used with the purpose of contextualizing the molecular variation generated in time, by associating an absolute timescale to the sequence diversity. The mutation evolution is governed by a time-continuous Markov-chain which implies that new mutations are completely independent from those already existing in the system (see details in Bandelt et al. 2006). One of the crucial points is the calibration of the molecular clock for a particular genetic system, dependent on the mutation rate of the underlying polymorphisms. Several approaches take into account quantitatively checkable parameters, such as: 1) constant mutation rate of different lineages and; 2) neutrality of the phylogenetic markers. The clock calibration can be performed directly from familial and pedigree analysis or indirectly by determining the accumulation of mutations in a timescale, using historical records, palaeontological or archaeological evidences about the split between different population groups/species as a basis, although none is without controversy.

Discordance of results given by familial/pedigree and evolutionary studies is well documented (Heyer *et al.* 1997, 2001; Macaulay *et al.* 1997; Pritchard *et al.* 1999; Forster *et al.* 2000; Kayser *et al.* 2000a; Holtkemper *et al.* 2001; Zhivotovsky *et al.* 2004). These differences were nevertheless not unexpected since the approaches look at opposite ends of the evolutionary process. As a consequence, we can see that 1) the “slower” mutations are only detected in the evolutionary studies while the “fast” are relatively overexpressed in the others, 2) state reversions are identified at the population level but are treated as new mutations in pedigrees and 3) the familial/pedigree analysis does not consider a non-uniform generation time through thousands of years of evolution, while in the latter variation is filtered by long-lasting natural selection (Howell *et al.* 2003a, Zhivotovsky *et al.* 2004). Under such circumstances, the mutation rates obtained by evolutionary studies on the basis of dated outgroups have been proved to be more suitable for evolutionary purposes, while the family/pedigree estimates are recommended for analysis of recent history, as it may possibly manifest itself at the tips of the phylogenetic trees (Pääbo 1996, Macaulay *et al.* 1997, Zhivotovsky *et al.* 2006).

Once agreed upon a mutation rate, the coalescence calculations estimate the time of divergence of a closely-related set of lineages that share mutational motifs. The intra-clade frequency and diversity of the sampled variation is evaluated in a backwards sense, by accessing the sequence dissimilarities since they last shared a common ancestral haplotype. The correct specification of the phylogenetic topology is then essential, since branching lengths are proportional to the average number of mutational changes between the root/nodal haplotype and every individual haplotype in the network. A  $\rho$  statistic is applied for obtaining a mutational time ( $\rho = \mu t$ , where  $\mu$  refers to mutation rate and  $t$  to time; Forster *et al.* 1996) and is then converted to absolute time by the use of the calibrated clock. As a “model-free” approach,  $\rho$  disregards the prehistoric demography and population structure affecting the molecular evolution. Nevertheless, until effective models have evolved, its use is more robust than models inaccurately characterizing the demography in the distant past (Bandelt *et al.* 2006). The  $\rho$  standard error is function of the structure and number of branches available for estimation (Saillard *et al.* 2000), in theory only correctly estimated from the real genealogy (Bandelt *et al.* 2006).

### 1.3 - DNA polymorphisms as molecular markers for phylogenetic studies

The occurrence of mutations, mere changes in the nucleotidic sequence of the DNA molecule, is the basis of individual variation. Over evolutionary time, the accumulation of different allelic variants results in the genetic diversity characterizing extant human populations. The sources of mutation are multiple, both spontaneous when errors in DNA replication happen, and induced, by physical or chemical agents. For the mutations to be transmitted to the progeny they have to occur in the germline and should not be lethal.

For the use of molecular markers with phylogenetic purposes an extended knowledge on their mutational differences is crucial, namely in terms of rates and processes (see sections 2.3.3 and 3.4). In that sense, two classes of markers can be distinguished: 1) “slow-evolving” markers are essential for resolving basal branches of the phylogenetic tree, usually defining haplogroups; 2) the “fast-evolving” markers are more useful for determining inner variation within a haplogroup, measured among the most recent branches of the phylogeny.

Most of the evolutionary studies make use of *Single Nucleotide Polymorphisms* (SNPs) and *Short Tandem Repeats* (STRs or microsatellites) polymorphisms. The SNPs refer to the simplest difference between two homologous DNA sequences – a base substitution (transition/transversion) or insertion/deletion (indel) of one base pair. In a general sense, SNPs are responsible for more than 90% of the genomic variants (Collins *et al.* 1997; ~7 millions of SNPs described in HapMap database). Due to pragmatic reasons their study has been standardized for Restriction Fragments Length Polymorphism (RFLPs) research, since many of those mutations are able to either prevent enzymatic hydrolysis of DNA by a (battery of) restriction enzyme at a known position of genome, or to generate new restriction sites.

The STR loci are repetitive elements of motifs of 3–7 bp in length (Fregeau and Fourney 1993, Smith 1995). These repeats are distributed throughout the human genome (except for mtDNA that does not contain and constitute a rich source of highly polymorphic markers, distinguished by the number of copies of the repeat unit (from 3 to 49 repeats/locus, de Knijff *et al.* 1997). The ancestral and derived states are then ascertained by comparison with a reference (provided and unambiguously identified). When combined with the binary SNP markers the “fast-evolving” STRs (or microsatellites) are able to discriminate the inner variety of basal branches. Notice however that the extra-nuclear and autonomous genome of mitochondria does not contain STRs.

#### 1.4 - Phylogeographic perspective of human uniparental genetic systems

The phylogeographic approach for human genetic variation analyses the spatial distribution of clades within a phylogeny, in parallel to their accumulated frequency, diversity and age estimates (Avisé *et al.* 1987, Avisé 2000). By performing a systematic comparison of those variables, alternative interpretations are left open to integrate patterns of modern diversity with the probable regional sources of variation and migration routes, as well as cultural and climatic events that inevitably have strongly contributed to the shaping of the present-day gene pool.

Due to their specific features, the human mtDNA and Y chromosome are the two genetic systems reflecting, respectively, a maternal and paternal perspective of the modern *Homo sapiens* origin and demographic processes throughout the world. Both systems manifestly tend to show

geographic and, often, also ethno-linguistic clustering patterns (e.g. <sup>Rosser *et al.* 2000; Underhill *et al.* 2000, 2001a; Pereira *et al.* 2001b; Richards *et al.* 2002; Salas *et al.* 2002; Semino *et al.* 2002, 2004; Destro-Bisol *et al.* 2004; Wood *et al.* 2005</sup>), presumably because mutations have accumulated along closely related radiating lineages. In general, the clines of paternal variation are much more evident than that in their maternal counterpart, suggesting that females have experienced higher rates of migration and gene flow compared to males and/or lower rates of genetic drift due to sex differences in effective population sizes (<sup>Perez-Lezaun *et al.* 1999, Oota *et al.* 2001, Fagundes *et al.* 2002, Dupanloup *et al.* 2003, Malyarchuk *et al.* 2004</sup>). Social habits can underline these differences, under sex-biased admixture (e.g. polygyny-monogamy transition in <sup>Dupanloup *et al.* 2003</sup>) and/or the phenomena of patrilocality, under which men are expected to live closer to their birthplaces while women move to their husband's natal domicile (e.g. <sup>Kayser *et al.* 2003, Wen *et al.* 2004, Hamilton *et al.* 2005, Wood *et al.* 2005</sup>). Both phenomena are not mutually exclusive, either one might be true for different areas and periods of time. Surprisingly, the patrilocal pattern still represents ~70% of the cases of modern societies (<sup>Murdock 1967, Burton *et al.* 1996, Seielstad *et al.* 1998</sup>), with larger differences accumulating in between more distant parts of the globe. In that sense, sex-specific scenarios may rather diverge than converge because of differential demographic histories, with mtDNA expected to have lower differentiation with geographical distances (<sup>Seielstad *et al.* 1998, Jorde *et al.* 2000, Oota *et al.* 2001</sup>). Indeed geography rather than language seems to be a better predictor of Y chromosomal affinities in Europeans, Americans and Austronesians (<sup>Rosser *et al.* 2000, Hurles *et al.* 2002, Zegura *et al.* 2004</sup>). However these may not be straightforward, since there are evidences of spread of Y chromosomal lineages without evident mtDNA counterpart (e.g. haplogroup NO counter-clock route from inner Asia/southern Siberia to east Europe, <sup>Rootsi *et al.* 2007</sup>; YAP+ chromosomes in Asia, Tibet and Andaman islands from an African source, <sup>Underhill *et al.* 2000, Tajima *et al.* 2004, Wen *et al.* 2004, Hammer *et al.* 2006</sup>). In Africa, where the miscegenation system is mostly dictated by social constraints, the paternal variation is apportioned among both geographic and ethnolinguistic units (<sup>Destro-Bisol *et al.* 2004, Wood *et al.* 2005</sup>).

Under episodes of demographic expansion, some of the genetic types tend to become more frequent. If the conditions that allowed expansion persist (e.g. technological improvement or climatic stabilization), the next generation(s) will further increment their frequency and diversity until new mutations arise and the original variants start to decay. Starlike phylogenies are then formed and testify for the extent of the expansion and the timescale of its evolution (<sup>Forster 2004</sup>). On the other hand, for variants that have supposedly arrived in a new territory, a founder analysis (e.g. <sup>Forster *et al.* 1996, Richards *et al.* 2000</sup>) allows to identify a founder haplotype – an ancestral node which is present or phylogenetically reconstructed both in the source and in the destination area – and from there evaluate the accumulated variance. Ideally, the coalescence age to the founder(s) type(s) would reflect the arrival of the migrating group. However, if the population is small and does not disperse upon arrival, the coalescence times of the founder types may underestimate their entrance time. On the opposite, more massive migrations carry a considerable amount of variation, so that the extant variation is a sum of

different periods in their demographic history, and will therefore overestimate the dates. The strategy to be outlined requires then a deep phylogenetic analysis in a width of a relevant geographic scope.

## 2 - Mitochondrial DNA phylogenetic analysis

### 2.1 - Structure and organization of mtDNA

Mitochondria are cytoplasmatic organelles responsible for the energy production of the cells. The energy-generating oxidative phosphorylation (OXPHOS) pathway physically takes place in the mitochondria, including fatty acid  $\beta$ -oxidation, the urea cycle and the common pathway for ATP production – the respiratory chain. These organelles also play a part in intracellular signalling and apoptosis, in intermediate metabolisms such as the Krebs or tricarboxylic acid cycles, and in the metabolism of amino acids, lipids, cholesterol, steroids and nucleotides <sup>(Chinnery 2006)</sup>. Mitochondrial DNA – mtDNA - represents an extranuclear genome whose content and size varies in different living organisms. When compared to the conserved organization in metazoan organisms <sup>(Saccone *et al.* 1999)</sup> most of mtDNA genes have been lost in mammals. The crucial set of genes for OXPHOS pathway, transcription and replication processes, have nevertheless been kept, with most of the molecules containing 12 to 20 protein-coding genes. The economically built human mtDNA genome is reflected in the almost lack of non-coding regions <sup>(Anderson *et al.* 1981)</sup>. Nevertheless, nuclear genes also code for a much larger variety of mitochondrial peptides <sup>(Shoubridge 2001)</sup>, synthesized in the cytoplasm with a mitochondrial targeting sequence.

The number of mitochondria in human cells may vary largely according to the cell type and size. In energy-dependending tissues, thousands of mitochondria can be found, each with two to ten mtDNA copies in its matrix. Somatic cells are estimated to host 1000 to 10000 mtDNA molecules <sup>(Lightowlers *et al.* 1997)</sup>. In the case of germline cells, the mtDNA content of the mature oocytes averages the 200,000 molecules <sup>(Chen *et al.* 1995a, Steuerwald *et al.* 2000, Reynier *et al.* 2001, Santos *et al.* 2006)</sup>, with each mitochondrion containing a single DNA molecule <sup>(Piko and Matsumoto 1976, Piko and Taylor 1987)</sup>, while sperm cells mid-piece counts with 50 to 75 mitochondria (50 to 1200 mtDNA molecules, <sup>Diez-Sanchez *et al.* 2003</sup>) in charge of their mobility. In Figure 4 the cellular content of mitochondria and the internal structure of these organelles are evidenced by fluorescence methods.

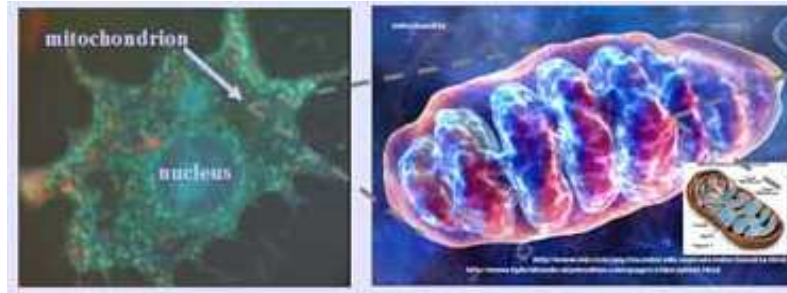
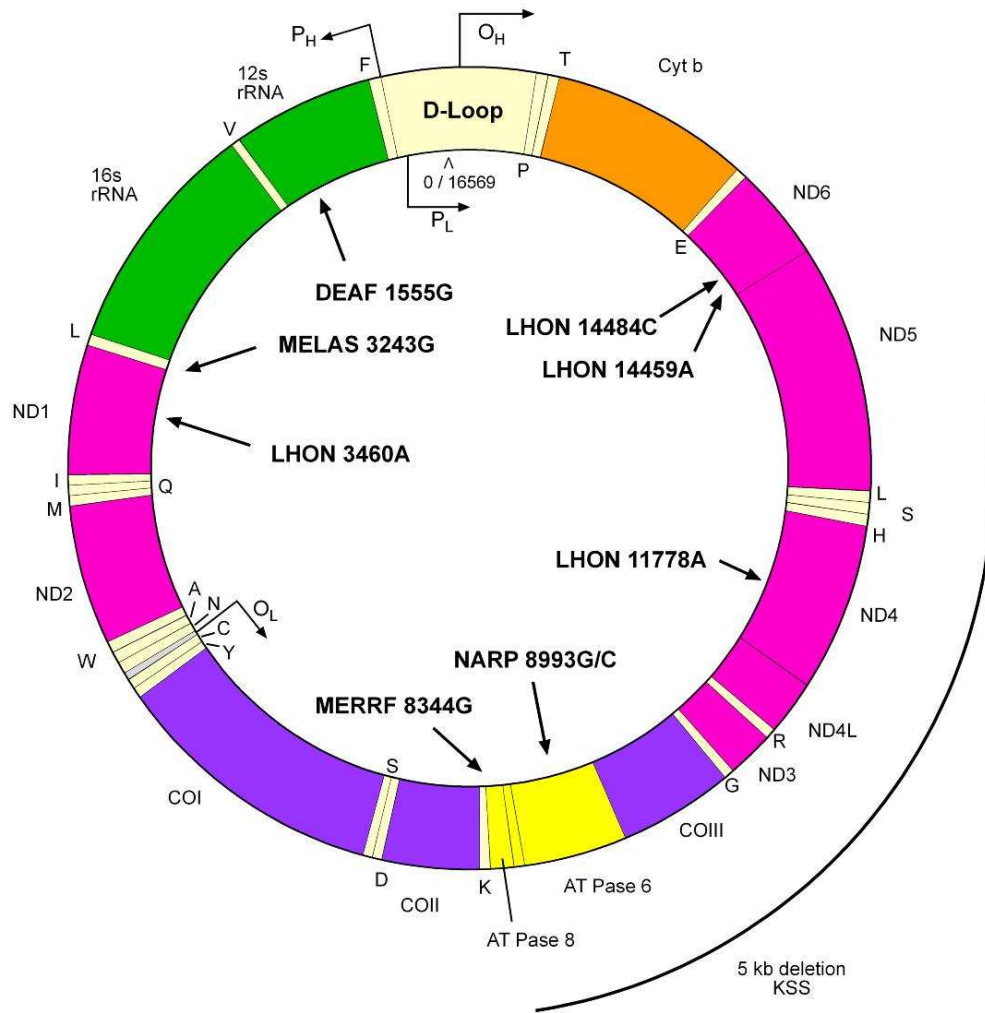








Figure 4 – Cellular location and internal structure of the mitochondria, revealed by fluorescent methods. *In* [www.microscopy.fsu.edu/cells/animals/mitochondria.html](http://www.microscopy.fsu.edu/cells/animals/mitochondria.html).

The human mitochondrial genome is a circular double-stranded DNA molecule about 16.6 kb long, whose sequence was first fully described 26 years ago <sup>(Anderson *et al.* 1981)</sup>. This particular sequence from a human tumour line is known as “Cambridge reference sequence”, in order to indicate that the sequence was resolved in MRC Laboratory for Molecular Biology, Cambridge, UK. For the revised version of human mtDNA “reference” sequence see Andrews *et al.* <sup>(1999)</sup>. Most of its length comprises contiguous coding regions, encoding for 13 polypeptides involved in OXPHOS electron transport system (ETS), 22 tRNAs and 2 rRNAs, essential to protein synthesis (Figure 5a). The largest non-coding region of human is the control region or D-loop segment (D standing for displacement), an extension of 1.1Kb comprised in-between np 16024 and np 576 with regulatory elements for the replication and transcription processes <sup>(Lightowlers *et al.* 1997)</sup>. Three short segments here comprised and generally named hypervariable sequences – HVS-I, HVS-II AND HVS-III – have a highly variable sequence in comparison to the rest of the genome <sup>(Brandstatter *et al.* 2004a)</sup>. The control region contains heavy- (H) and light- (L) strand promoters, the multiple origins of the H-strand replication (OH<sub>n</sub>), three conserved sequence blocks (CSBI, II and III) and the termination-associated sequences (TAS, see Figures 5a and 5b). The multiple origins of replication are known to relate with different modes of replication synthesis, namely mtDNA maintenance under steady-state conditions or as a response to physiological demands (e.g. <sup>Coskun *et al.* 2003</sup>). The mtDNA is intertwined, in punctuated structures of the matrix of the mitochondrial inner membrane called nucleoids, in close proximity to the ETS <sup>(Clayton 1992)</sup>.



	Complex I genes (NADH dehydrogenase)		Complex III genes (ubiquinol: cytochrome c oxidoreductase)		Transfer RNA genes
	Complex IV genes (cytochrome c oxidase)		Complex V genes (ATP synthase)		Ribosomal RNA genes

a)



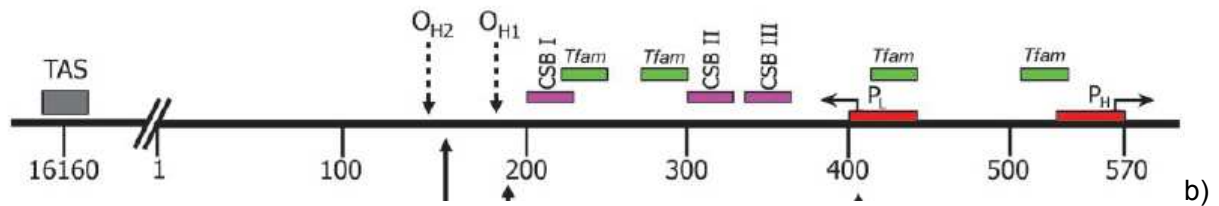


Figure 5 – Genomic map of human mtDNA. Schematic diagrams of a) the full molecule and b) the mtDNA control region. H and L stand for heavy and light strands respectively, given the asymmetric distribution of G and C nucleotides, with H being the G-rich one. The genes of the seven subunits of OXPHOS (ND1, 2, 3, 4L, 4, 5 and 6), one subunit of complex III (Cyt b), three subunits of complex IV (COI, COII AND COIII), two ribosomal RNAs (12SrRNA and 16SrRNA), 22 tRNAs and D-loop regions are evidenced. Gene products encoded by the L-strand are shown in the inner complete circle while the products of the H-strand, are shown in the outer circle. The location of promoters (P<sub>L</sub> and P<sub>H</sub>) for transcription and replication origins (O<sub>H1</sub>, O<sub>H2</sub>, as in <sup>Coskun *et al.* 2003</sup> are shown by arrows. In (b) the functional sequences are indicated with boxes. TAS, termination association sequence; CSB I, II, III, conserved sequence block I, II, and III; Tfam, Tfam binding sites. Adapted from MITOMAP and Coskun *et al.* <sup>(2003)</sup>.

## 2.2 - On the origin of mitochondria and mtDNA

It is generally accepted that mitochondrion have originated about  $1.5 \times 10^9$  years ago, as an endosymbiotic prokaryotic organism who chose an proto-eukaryotic cell as a host, providing extra energy in return for a safer environment <sup>(Margulis 1970, 1975)</sup>. The establishment of this relationship is probably associated with a global disaster at two billion years ago, when atmospheric oxygen levels started to rise due to the cyanobacteria activity <sup>(Holland 1994)</sup>. Under these conditions, anaerobic life forms that were unable to protect themselves from the toxicity of oxygen, or unable to find a suitable anaerobic microenvironment, probably became extinct. The mitochondrial traces resembling modern bacteria, such as the circular shape of the genome and the discrete origin of replication support the hypothesis of an endosymbiosis <sup>(Vellai *et al.* 1998)</sup>. Molecular phylogenies have also provided evidence of a single and monophyletic bacterial origin for these modern cellular organelles <sup>(Gray *et al.* 1999)</sup>, with mtDNA genes having the closest link to those of the present-day alpha-proteobacteria <sup>(Gray *et al.* 1999, Karlberg *et al.* 2000, Andersson *et al.* 2003)</sup>.

In this process of co-evolution, some of the mitochondrial genes have been transferred to the nucleus as orthologous though non-functional genes <sup>(Lopez *et al.* 1994)</sup>. The analysis of the human nuclear genome has revealed more than 600 nuclear inserts of mtDNA (*numts*; <sup>Olson and Yoder 2002, Tourmen *et al.* 2002, Bensasson *et al.* 2003, Mishmar *et al.* 2004, Ricchetti *et al.* 2004, Hazkani-Covo and Graur 2007</sup>) of variable length from ~40 bp nearly up to the whole mitochondrial genome. Their sequence homology to the mtDNA ranges from 78 to nearly 100%, thus they are supposed to result from temporarily different insertional events, where the lower level of identity means they have resided in the nucleus for a longer time. The integration of

*numts* seems then to be an ongoing process that shapes nuclear genomes (Bensasson *et al.* 2001, Ricchetti *et al.* 2004). Furthermore, human-specific *numts* preferentially target coding or regulatory sequences and can therefore generate mutagenic alterations found to be associated with diseases in humans (e.g. thrombospondin homologous gene, an angiogenesis inhibitor that retards tumour growth; Bogdanov *et al.* 1999, Borensztajn *et al.* 2002, Turner *et al.* 2003).

On the other hand, a large number of mitochondrial proteins are imports coded by nuclear sequences. The modern mitochondrial proteome is then a result of both reductive and expansive processes, where many ancestral mitochondrial genes have been simply purged and lost when neutralized by the organism's environment (Karlberg *et al.* 2000, Andersson *et al.* 2003), but the loss is supposedly tolerated by a compound uptake from the cytosol (Andersson *et al.* 1998, Berg and Kurland 2000).

## 2.3 – Distinctive features of mtDNA

### 2.3.1 – Maternal inheritance

The maternal transmission constitutes the mtDNA hallmark, since it is inherited from the cytoplasm of the egg (Giles *et al.* 1980, Stoneking 1994, Stoneking and Soodyall 1996, Wallace *et al.* 1999). A mutation altering the mtDNA of a woman's oocyte will henceforth characterize her descendants, what enables researchers to trace related lineages back through time without the confounding effects of biparental inheritance. At the time of fertilization the paternal mtDNA is comparatively low numbered (Michaels *et al.* 1982, Diez-Sanchez *et al.* 2003), in a average proportion of 100 carried in the sperm's tail to 10<sup>4</sup> in the oocytes, making paternal contribution highly unlikely. Moreover there is an active system that eliminates the spermal mtDNAs. It is believed to happen at early stages of the embryogenesis, by ubiquination of the mid-piece of the sperm cell (Hopkin 1999, Sutovsky *et al.* 2004). Although there can be leakage of paternal mtDNA in the case of poor quality oocytes (St John *et al.* 2004) or dysfunctions of the system of paternal mtDNA destruction (Schwartz and Vissing 2002), the populational genetics (Macaulay *et al.* 1999a) and pedigree studies (e.g. Bendall *et al.* 1996, Parsons *et al.* 1997, Soodyall *et al.* 1997, Jazin *et al.* 1998, Howell *et al.* 2003a) and further analysis of patients with mitochondrial myopathies (Davis *et al.* 1997, Taylor *et al.* 2003, Schwartz and Vissing 2004) showed no evidence this would happen in normal conditions, leaving the exclusive inheritance to the maternal component.

### 2.3.2 - Non-recombining transmission and homoplasmy

The mammalian mtDNA does not recombine as nuclear chromosomes do (Olivo *et al.* 1983, Merriwether *et al.* 1991, Stoneking 1994, Stoneking and Soodyall 1996, Wallace *et al.* 1999), meaning allelic association for mitochondrial markers therein. Therefore, the genetic material acts as a single locus and allows the drawing of a

unique matrilinear genealogy, where all copies trace back to a common ancestor. As it represents one quarter of the effective population size of autosomal loci (three-fold less than X-chromosome), this genome is more prone to random fluctuations of allele frequencies under genetic drift, increasing its sensitivity to detect such phenomena.

The lack of recombination in human mtDNA has been questioned for several times, endangering the interpretation of the human mtDNA variation. If to occur, recombination could be possible as the molecular machinery is present (Thyagarajan *et al.* 1996) although it is still unclear to what extent mitochondria within a cell are able to fuse and exchange contents (Ono *et al.* 2001, Legros *et al.* 2004). A statistical analysis of pairwise linkage disequilibrium, as function of distance between sites, suggested recombination since it declined with increasing markers distance (Awadalla *et al.* 1999). Similar hypothesis was proposed by a phylogenetic study of coding region homoplasies (Eyre-Walker *et al.* 1999) since their frequency at the same np was much higher than expected under a single rate of synonymous mutations. The quality of both data has however been questioned (Kivisild and Villems 2000, Kumar *et al.* 2000) and their reanalysis gave no significant results (Arctander 1999, Jorde and Bamshad 2000, Parsons and Irwin 2000). Other later studies found no evidence of recombination of the molecule (Ingman *et al.* 2000, Elson *et al.* 2001, Piganeau and Eyre-Walker 2004). Recently, a case of observed recombination was reported in the only known human with maternal and paternal DNA (Schwartz and Vissing 2002, Kraysberg *et al.* 2004). The patient's muscle tissue contained approximately 0.7% of recombined types. Recombination is however a very rare phenomena, and if it would occur in homoplasmic cells mtDNAs will not differ from the original (Pakendorf and Stoneking 2005).

The presence of only one type of mtDNA throughout the tissues of an organism— homoplasmy – is thought to be essential to the normal functioning of mitochondria (Hirata *et al.* 2002), as it allows a coordinate gene expression of mitochondrial and nuclear genes. When a new mutation arises and there is a complete replacement of the existing mtDNA variant, the new homoplasmic state guarantees the normal functioning of cellular respiration, unless the variant turns out to be deleterious. The state of heteroplasmy can nevertheless happen in cases of incomplete switch, so that in a generation time two or more variants become co-existent (Poulton *et al.* 1998 and references therein). But again, the mechanism of genetic bottleneck in the oogenesis, at first reducing the number of molecules (Hopkin 1999) and then increasing it in mature oocytes (Thorburn and Dahl 2001) seems to preserve homoplasmy. A more recent study proposes that the new type can clonally expand in the cell (Coskun *et al.* 2003), over what one can preclude that the pathway of energy production will be significantly altered. However, a mechanism of overproduction of mtDNA is initiated to deal with the chronic deficit, in cells where the mutant induced the deficiency. A selective amplification can occur and mutants may become predominant if they do not generate defective gene products.

### 2.3.3 – Mutations rate, homoplasy and multiple hits

Human mtDNA mutation rate is about ten to hundred-fold faster than that of nuclear DNA <sup>(Brown *et al.* 1979, Cann *et al.* 1987, Pesole *et al.* 1999, Ingman and Gyllensten 2001)</sup> therefore providing more information on the phylogeny within the species than equivalent DNA segments in the nucleus. The main reasons for such rate are the lack of the sophisticated DNA repair mechanisms and perhaps the absence of protective histones <sup>(Bogenhagen 1999)</sup> like those of nuclear genome, plus the high exposure to free radicals as a result of the OXPHOS pathway that increases oxidative damage of DNA.

Within the mtDNA molecule the mutation rates are long shown to vary widely between the regions and at nps within a region <sup>(Hasegawa *et al.* 1993, Wakeley 1994, Macaulay *et al.* 1997, Torroni *et al.* 1998, Macaulay *et al.* 1999b, Finnilä *et al.* 2001, Mishmar *et al.* 2003, Kivisild *et al.* 2006b)</sup>. The rate of substitution is higher for the control region, where transitions are generally much more frequent than transversions <sup>(Meyer *et al.* 1999</sup> and references therein). In parallel, 1) synonymous nps in protein-coding genes and peripheral domains of the D-loop evolve five to ten times faster than the remaining control region domains; 2) the rate of synonymous sites is quite uniform over the mitochondrial genome, whereas the rate of nonsynonymous sites differs considerably between genes <sup>(Pesole *et al.* 1999, Kivisild *et al.* 2006b)</sup>, 3) synonymous sites and rRNA evolve *ca.* 20 times and tRNAs *ca.* 100 times more rapidly than the equivalents in nuclear genome <sup>(Pesole *et al.* 1999)</sup>. A calculation based on the comparison of a migrant group and its source population estimated a rate of  $1.8 \times 10^{-7}$  for the control region segment nps 16090-16365 <sup>(Forster *et al.* 1996, 2001)</sup>. In the coding region the rate is quite uniform for transitions ( $3.5 \times 10^{-8}$ /site/year) but not for transversions ( $2.1 \times 10^{-9}$  and  $4.1 \times 10^{-10}$  substitutions/site/year for synonymous and tRNA mutations, respectively; <sup>Kivisild *et al.* 2006b)</sup>). An excess of rRNA and nonsynonymous base substitutions among “hotspots” of recurrent mutations was observed, mostly involving guanine to adenine transitions. A distinct mutational pattern among and within the control region and protein-coding region might have functional and structural underlying reasons <sup>(Tamura 2000, Kivisild *et al.* 2006b)</sup>. For example, the physical structure of the D-loop formation, in which the H-strand is displaced by the nascent L-strand and made to be in a single-stranded state <sup>(Reyes *et al.* 1998)</sup>, can be suggested as a causal factor; depurination, the most frequent spontaneous alteration that occurs in DNA under physiological conditions <sup>(Loeb and Preston 1986)</sup>, might explain the higher mutability of purines in the single-stranded H, in addition to that a repairing mechanism works only for double-stranded DNA as it requires a complementary template.

The inferences from control region can become problematic due to its high rate of mutation and because it is subject to saturation due to excess homoplasy <sup>(Tamura and Nei 1993, Bandelt *et al.* 2006)</sup>. Mutations on the same sites can arise due to distinct and independent events, showing up as parallel mutations in different lineages, more frequently in the so-called “hotspots”. Multiple hits of recurrent

mutations can then create ambiguous interpretations of the phylogenetic analysis, and it is essential that these are correctly identified (Bandelt *et al.* 2006). However, the basal structure of the phylogeny (its “skeleton”) is not much affected by recurrent mutations, because the level of resolution between the branches is sufficient, usually on the account of slow-evolving diagnostic coding region sites (see section 2.6). In fact, phylogenetically more stable coding region mutations offer supplementary power to distinguish recurrent mutations, for example within HVS-I (Bandelt *et al.* 2006).

## 2.4 – The role of selection in mtDNA genome

According to the neutral theory of molecular evolution (Kimura 1968, 1983), the fixation of stochastically rising mutations depends primarily on random genetic drift. Thus, demographic history supposedly plays a determinant role in the accumulation of mutations along radiating female lineages and the role of positive selection is negligible. As the theory assumes that the rate of evolution depends solely on mutation rate, neutrality can be tested by estimating the fixation difference between synonymous (neutral) and non-synonymous mutations at intra and inter-species level (Graven *et al.* 1995, Nachman 1998, Nielsen and Weinreich 1999). Several studies have observed an excess of non-synonymous mutations in the more recent branches of the phylogeny (Excoffier and Yang 1999, Elson *et al.* 2004, Ruiz-Pesini *et al.* 2004, Kivisild *et al.* 2006b). The most straightforward interpretation of that phenomenon is that negative (purifying) selection has acted on the human mtDNA-encoded proteins during evolution (reviewed by Gerber *et al.* 2001, Elson *et al.* 2004, Kivisild *et al.* 2006b), but has not yet purified the slightly deleterious mutations from the youngest offshoots of the phylogenetic tree that might have introduced into the mtDNA pool, in particular during the phases of fast expansion of populations (Excoffier 1990, Merriwether *et al.* 1991). It might be the case in European mtDNA haplogroup J, characterized by mutations associated with LHON disease (Torroni *et al.* 1997; Carelli *et al.* 2002, 2006; Howell *et al.* 2003b). Non-recombining mtDNA acts as a single locus where neutral or slightly deleterious substitutions can “hitchhike” their frequencies up or down as they are in linkage with sites under strong selectional pressure. Therefore, these mutations contribute to the so-called “Muller’s ratchet” (Muller 1964, Lynch 1996), a slow but inexorable accumulation of slightly deleterious mutations, made possible by the absence of recombination of this particular genome.

Continental differences of mtDNA variation has been sometimes interpreted as shaped by climatic adaptations so that geographical distribution of particular mitochondrial haplogroups has been influenced by positive selection (Torroni *et al.* 1994a, Mishmar *et al.* 2003, Ruiz-Pesini *et al.* 2004). Many of the changes associated with the adaptations were found in the coding region like in ATP6 gene, even though it is believed to be one of the most conserved mtDNA proteins (Mishmar *et al.* 2003). However, Elson *et al.* (2004) and Kivisild *et al.* (2006b) found no significant differences in relating climate and the rate of non-synonymous changes for mtDNA haplogroups. The former studies seem to have erroneously

compared region-specific haplogroups of different diversity levels e.g. older lineages of Africans and younger of the Arctic populations, so that an excess of nonsynonymous mutations was predictable for the more recent variants (Elson *et al.* 2004, Kivisild *et al.* 2006b). To consider the neutrality of mtDNA markers is a crucial subject in order to use mtDNA phylogenetic system in interpreting human dispersals. As no significant differences were found for different mtDNA lineages, positive selection should be assumed not to play a significant role in shaping the present variation of mtDNA and mtDNA markers can be taken, at least overwhelmingly, as neutral.

## 2.5 – Calibration of the mtDNA molecular clock

Genetic dating, especially using uniparental markers such as mtDNA and the non-recombining portion of the Y chromosome, plays a crucial role for reconstructing the evolution and spread of modern humans. In evolutionary studies, the calibration of the molecular clock makes use of an outgroup, analysing the mean accumulated differences between equivalent sequences in the light of their distance to the MRCA (established independently from a different source, as a rule from fossil evidence or from known from archaeology events). Regions of known colonization time allow to calculate an average rate of human mtDNA divergence, in a founder analysis perspective in which the archaeological records provide the evidence for the initial settlement. The molecular diversity of the population “derived-by-migration” is compared to that of the source population and founder types are identified by their frequency and accumulated variance. This model assumes the present-day variability in the source population to be similar to the original one, and considers low level of back migrants and parallel mutations. The populations of New Guinea, Australia and America have been used for such calibrations: their specific mtDNA clusters estimate a 2-4% per my divergence rate for the complete mtDNA molecule (Wilson *et al.* 1985, Torroni *et al.* 1994b). The control region calibration based on the data of Beringian expansion (Forster *et al.* 1996, Saillard *et al.* 2000) obtained a corresponding value of 36% per my, later averaged and converted into one mutation every 20180 years (Forster *et al.* 2001) for the HVS-I segment nps 16090-16365.

Analogous mtDNA sequences of chimpanzee (Vigilant *et al.* 1991, Ingman *et al.* 2000, Maca-Meyer *et al.* 2001, Mishmar *et al.* 2003, Gonder *et al.* 2006, Kivisild *et al.* 2006b) and *Neanderthal* (Krings *et al.* 1997, Ovchinnikov *et al.* 2000), as well as mtDNA segments inserted in the human nuclear genome (*numts*; e.g. Kivisild *et al.* 2006b), have been also used as outgroups serving the purposes of mtDNA phylogenetics. In brief, because of their location, *numts* evolve at a nuclear lower rate and appear to be “frozen” in comparison to their mitochondrial counterparts (Fukuda *et al.* 1985). Therefore, their application is wide: phylogenetic markers, if there is enough sequence divergence at the aimed inter- or intraspecific level; infer ancestral states or root mitochondrial phylogenies if paralogous and their derivatives are known (Zischler *et al.* 1995); set the baseline

for the study of nuclear mutation on which other evolutionary factors operate; study mutagenic phenomenon responsible for a variety of genetic diseases. Returning to the calibration of the mtDNA molecular clock, different attempts have obtained slightly different estimates. For example, Ingman *et al.* (2000) and Mishmar *et al.* (2003) obtained respectively  $1.7 \times 10^{-8}$  and  $1.26 \times 10^{-8}$  substitutions/site/year for the average rate of sequence evolution of mtDNA coding region in Europeans (control region not considered due to higher probability of reverse mutations). The main difference between the studies resides in the age considered for the human-chimpanzee split (5 mya in Ingman *et al.* 2000, according to Andrews 1992, Kumar and Hedges 1998, 6.5 mya in Mishmar *et al.* 2003, with 500 ky of lineage sorting, over 6 my for the split following Goodman *et al.* 1998). Under the last estimate, which corresponds to 5140 years per substitution in the whole coding region, the age of the MRCA of human mtDNAs was inferred to be of  $198 \pm 19$  ky (Mishmar *et al.* 2003).

It has been more recently noticed that the purifying selection has left its mark in the mtDNA phylogeny (Elson *et al.* 2004, Kivisild *et al.* 2006b). The deeper branches are relatively impoverished in non-synonymous substitutions compared with synonymous ones, hinting for that the deleterious mutations can survive in the short term but are eventually weeded out in the long run. The kind and quantity of the decay of non-synonymous mutations is not yet clear but is rather represented by a sigmoid-like curve (Bandelt *et al.* 2006) than exponential (Ho *et al.* 2005). At first, the appropriateness of using the average molecular clock over all mtDNA sites in dating events in human population history seems to be undermined. Kivisild *et al.* (2006) pioneered in using solely the mutation rate of synonymous transitions for the calibration of the molecular clock, in which one mutation occurs within every  $6764 \pm 140$  years. Not surprisingly, the ultimate coalescence age of  $\sim 160 \pm 22$  ky and the TMRCA for the many nodes in the tree are generally younger than when calculated based in both non-synonymous and synonymous transitions. Nevertheless, the phylogenetic approach for analyzing mtDNA sequence data at intraspecies level remains viable because reconstruction of the basic branches is robust and the excess of non-synonymous substitutions affects mainly the terminal branches of the tree (Kivisild *et al.* 2006b).

More direct approaches that do not require historical or outgroup data consider the familial/pedigree data. In those studies, the mutation rate is estimated from a defined genealogy, screening for new mutations in a scale of few generations. The given estimates of  $\sim 0.47 \times 10^{-6}$  substitutions/site/year are 10-fold higher than the calculated from the phylogenies (e.g. Howell *et al.* 1996, Parsons *et al.* 1997, Howell *et al.* 2003a) and at first, may question the accuracy of dating past divergences on the basis of phylogenetic rates (Pääbo 1996). Although other analysis have obtained comparable rates to those given by evolutionary studies (Macaulay *et al.* 1997, Soodyall *et al.* 1997, Jazin *et al.* 1998), pedigree rates should be of careful use because these include mildly deleterious mutations that will not be fixed by selection (Forster *et al.* 2002). The clue for understanding the discrepancy between phylogenetic and pedigree-based rates comes from the highly heterogeneous mutation rates, with the existence of mutational hotspots

(Richards *et al.* 1998b, Excoffier and Yang 1999, Meyer *et al.* 1999, Heyer *et al.* 2001). Furthermore, using pedigree rates for dating known historical demographic events such as peopling of different continents offers predictions clearly incompatible with reality (see Sigurgardottir *et al.* 2000 and Bandelt *et al.* 2006 for critical assessments).

The clock-wise evolution of mtDNA has been questioned for several times, concerning either the control-region alone (Ingman *et al.* 2000) or particular mtDNA clades (Torroni *et al.* 2001a). While the former claim could be dismissed because of the reliability of the test and tools employed, the latter may constitute a one-off deviation, rather pointing to unsatisfactory models for mtDNA sequence evolution (Howell *et al.* 2004, Bandelt *et al.* 2006). The clock-like behaviour of the basal Eurasian mtDNA variation, evaluated in Macaulay *et al.* (2005), was found to adequately describe the data. Conversely, in the particular case of the African L2a clades, a complex demographic history with population subdivision might have produced the differences, since mutation and fixation rates are function of the effective population size (Salas *et al.* 2002, Howell *et al.* 2004). In the well-argued opinion of Bandelt *et al.* (2006), occasional concerns that the molecular clock might be elusive and not tick regularly for human mtDNA should not hinder us from attempting on calibration. These authors envision that future recalibrations of the mutation rate might discard non-synonymous substitutions altogether but will embrace mutations at slowly evolving sites of the control region, in order that the two spectra of positional rates are somewhat comparable.

Besides the interpretive gap between molecular evolution and prehistory of their carriers, pitfalls do exist in the confidence of the underlying topology and stipulation and calibration of the mtDNA molecular clock (Bandelt *et al.* 2003, reviewed in Bandelt *et al.* 2006). The phylogeny may suffer of low phylogenetic resolution, either on the account of small fragments of mtDNA, poor or misapplied phylogenetic method or even sequencing errors, compromising any estimation. The partial saturation at highly recurrent characters is also a main issue because it will swamp phylogenetic signals and eventually lead to a near total loss of meaningful inferences. For instance, hotspot sites in the control region account for partial saturation roughly after 60 kya (Bandelt *et al.* 2006). For the same reason, the calibration of the molecular clock using distant outgroup information is not fully satisfactory. Finally, the decay of non-synonymous mutations, occurring at a still unknown fashion, can also be overestimated, again because of partial saturation (Bandelt *et al.* 2006).

## 2.6 – Phylogenetic classification and nomenclature of mtDNA haplogroups

The research of mtDNA as a molecular marker was pioneered by Wesley Brown and Douglas Wallace in the late 1970s, with the intention of describing the origin of AMH (Brown 1980). Early studies of coding region polymorphisms, carried out in “low”- and “high”-resolution by analysing the “cutting” patterns of sets of 5-6 or 12-14 restriction enzymes (RFLPs) respectively, were found to be useful for the purposes of human population genetics by establishing the torso of the mtDNA tree (e.g. Denaro *et al.*



1981; Johnson *et al.* 1983; Cann *et al.* 1987; Scozzari *et al.* 1988; Soodyall and Jenkins 1992, 1993; Torroni *et al.* 1992; Chen *et al.* 1995b). The results suggested a radiating phylogeny of global variation, with a single central haplotype as a putative indication that modern human populations might have shared a common evolutionary history for a very long time. Soon after the analysis of mtDNA HVS-I proved to define a similar classification and to be more informative (Richards *et al.* 1996), and became of routine use in phylogenetic studies. Subsequent works adopted a synthesis of the two patterns of variation – that for HVS-I and of the coding region RFLPs (e.g. Macaulay *et al.* 1999b). The basal topology of an mtDNA network is overwhelmingly based on coding region SNPs, where homoplasy is rarer (lower mutation rate), having therefore minimal effects on the construction. The topology becomes more complex with the introduction of control region substitutions that define the internal variability of individual haplotypes. A most parsimonious phylogeny might be overwhelmed by assigning different relative weights to the mutations, associated to their occurrence rates, and essentially by a good level of resolution in the basal topology, ideally a combination of control and coding region data (Torroni *et al.* 1996, Macaulay *et al.* 1999b, Bandelt *et al.* 2000, Chen *et al.* 2000, Kivisild *et al.* 2002).

Independent lineages in the phylogeny – haplotypes - are defined by the accumulation pattern of mutations, where the polymorphic sites are identified relative to a consensus sequence (Cambridge Reference Sequence CRS, Anderson *et al.* 1981, revised by Andrews *et al.* 1999). These haplotypic variants cluster in clades with common mutations – haplogroups - simplifying a hierarchical classification (see phylogeny scheme in Figure 6). Furthermore, the comparison of trees presented by different authors demanded a common (universal) nomenclature system to label the branches. The study of Native Americans by Torroni *et al.* (1993) initiated the currently accepted nomenclature, by describing four basal clusters in alphabetical order – haplogroups A, B, C and D. The classification has nevertheless become a continuous process were new data emerge within short periods of time, permitting frequent adjustments and a better resolution. To the capital letters representing the haplogroups, subclusters are attributed additional symbols (alternating letters and numbers, e.g. L0a1). The \* symbol is used to refer to paragroups, different yet unidentified clades, that in principal can even be MRCAs.

Despite the good correspondence of data on both HVS-I region and coding region RFLPs, the estimation of coalescence ages and network construction were made problematic because of inconsistencies of phylogenetic significance, namely the high and very variable substitution rate between sites (Howell *et al.* 1996, Excoffier and Yang 1999, Kivisild *et al.* 2006b) and the saturation due to excess homoplasy (Tamura and Nei 1993, Bandelt *et al.* 2006). From today's perspective, one can conclude that the combined HVS – (limited) coding region information allowed to establish a robust general topology of the mtDNA tree, as far as its main branches and sub-branches were concerned. But it did not allow to go further. Most of the present work concerning the improvements of phylogeny and estimates of temporal layers make use of complete sequencing of mtDNA genomes, with more than 2000 molecules fully described to date (Ingman *et al.* 2000, Finnilä *et al.* 2001, Maca-Meyer *et al.* 2001, Richards and Macaulay 2001, Torroni

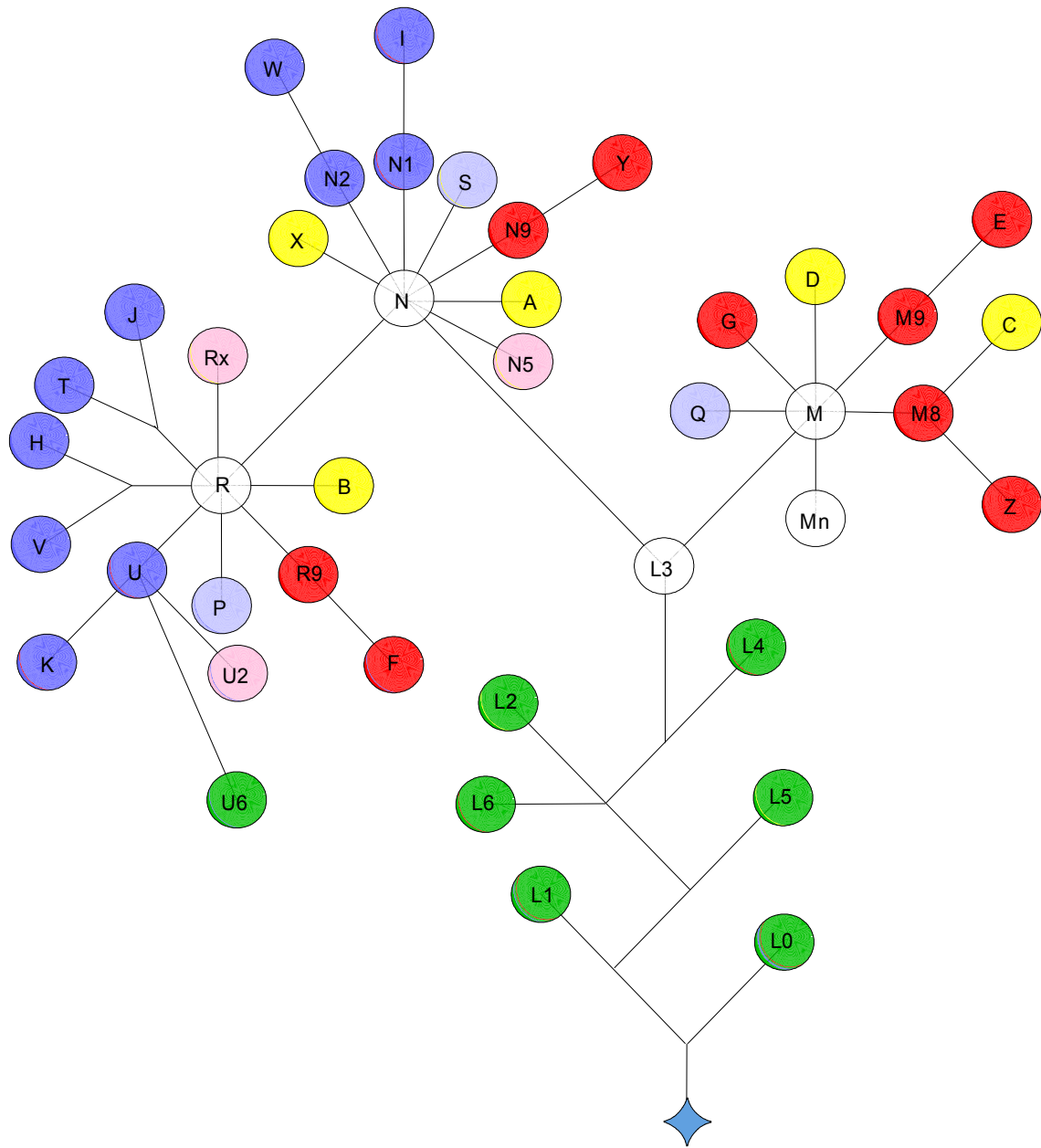


Figure 6 – Skeleton of the global phylogeny of mtDNA. Colors denote haplogroups of specific distribution in: dark blue - West Eurasia; light blue – Near Oceania; green – Africa; pink – Indian sub-continent; red – East Eurasia. The clades in yellow refer to haplogroups found in East Eurasians and Americans. The Mn cluster comprises sub-branches found in Indians, East Eurasians and Near Oceanians. The branching pattern near the root is as indicated in Mishmar *et al.* (2003), Macaulay *et al.* (2005), Torroni *et al.* (2006), using the two available complete mtDNA sequences from chimpanzees (Horai 1995). Based on data from Finnilä *et al.* (2001), Maca-Meyer *et al.* (2001), Herrnstadt *et al.* (2002), Kivisild *et al.* (2002), Kong *et al.* (2003), Palanichamy *et al.* (2004), Friedlaender *et al.* (2005), Macaulay *et al.* (2005), and Sun *et al.* (2006).

et al. 2001a, Herrnstadt *et al.* 2002, Kivisild *et al.* 2002, Ingman and Gyllensten 2003, Kong *et al.* 2003, Mishmar *et al.* 2003, Achilli *et al.* 2004, Palanichamy *et al.* 2004, Tanaka *et al.* 2004, Achilli *et al.* 2005, Friedlaender *et al.* 2005, Macaulay *et al.* 2005, Merriwether *et al.* 2005, Thangaraj *et al.* 2005, Accetturo *et al.* 2006, Gonder *et al.* 2006, Kivisild *et al.* 2006b, Olivieri *et al.* 2006, Underhill and Kivisild 2007).

Trees of complete mtDNA genomes allow defining more clearly the topology within clades, in particular nested sub-clades (e.g. U5 motif in Saami and North Africans, Achilli *et al.* 2005), and high-resolution dissection of earlier unresolvable haplogroups (e.g. the typically European haplogroup H; Achilli *et al.* 2004, Loogväli *et al.* 2004, Pereira *et al.* 2006, Roostalu *et al.* 2007). Furthermore, these allow unambiguous identification and phylogeographic cataloguing of true basal lineages, allowing inter alia, to trace the settlement process in Eurasia alongside the southern coast of the supercontinent to Melanesia and Australia (described in detail in section 2.7). Highly diverse African maternal lineages, and of high regional specificity have been disclosed in recent years by the use of high resolution studies and complete sequences. Re-rooting of the tree exposed L1 in its original definition as a paraphyletic clade. Haplogroup L0 turned out to be one of the earliest offshoots of the mtDNA variation, a sister clade of the branch that holds all other mtDNA haplogroups of extant AMHs (Mishmar *et al.* 2003, Kivisild *et al.* 2004, Gonder *et al.* 2006, Torroni *et al.* 2006). The real complexity of the mutational pattern near the root only began to emerge in recent years (Kivisild *et al.* 2004, Bandelt *et al.* 2006, Kivisild *et al.* 2006b) but more elucidative studies are forthcoming (Behar MD, Villems R, *et al.* ms submitted). The previously named L1e in Salas *et al.* (2004), Stevanovitch *et al.* (2004) and Knight *et al.* (2003) was redefined as haplogroup L5 with a position intermediate relative to L1 and L2'L3 (Shen *et al.* 2004, Gonder *et al.* 2006). The derived allele at 3594 was before taken as the defining marker of L3 variation (Watson *et al.* 1997) but nowadays the coding region information distinguishes the L4 earliest branch which also harbours the derived state (Kivisild *et al.* 2004). The same authors found as well that paragroup L3 includes sub-haplogroups L3h, L3i, L3x and L3w and that the previous L3g is indeed a sister-clade of L4a, then reclassified as L4g. Many recent studies focus on specific branches of the phylogeny (e.g. Achilli *et al.* 2004, 2005; Friedlaender *et al.* 2005; Macaulay *et al.* 2005; Kivisild *et al.* 2006a; Olivieri *et al.* 2006), therefore contributing partially to a broader understanding. Six coding region and a control region mutation further define L6, a sister-clade of L2 and L3'L4 variation (Kivisild *et al.* 2004).

## 2.7 - Worldwide variation of modern human mtDNA uncovers origins and settling processes

The analysis of mtDNA variation hit the popular consciousness with the publicity given to the debate on modern human origins. In parallel to the technological improvements, the scientific community assisted to an exponential progress in the field since multiple hypotheses were to be tested. Two antagonistic models were initially in the highlights to explain the non-African variation: the “multiregionalism” versus the “recent Out-of-Africa”. Multiregionalists defended an independent and parallel evolution of AMH from archaic ancestors in the different parts of the world, largely based on

archaeological findings and anatomical features supposedly shared by ancient (here – Middle Palaeolithic) fossil crania and modern humans from the same geographical location<sup>(Weidenreich 1943, Wolpoff 1989, Wolpoff and Caspari 1997)</sup>. The “Out-of-Africa” hypothesis assumed AMH to have arisen in Africa, where evolution proceeded for 50-100 kya, and from where it spread over other regions completely replacing regionally different earlier humans<sup>(Lewin 1987; Stringer and Andrews 1988; Foley 1998; Stringer 2000, 2003)</sup>.

The root of the tree obtained by early mtDNA studies was found to be shared by a high number of individuals of worldwide distribution<sup>(Excoffier and Langaney 1989)</sup> and therefore sometimes disputed as a support for the multiregional model<sup>(Templeton 1992)</sup>. Although initially targeted with strong criticism, Allan Wilson’s group was among the first to consistently suggest an African origin for all humankind maternal lineages, in studies of worldwide samples with the highest resolution to their date<sup>(Cann *et al.* 1987, Vigilant *et al.* 1991)</sup>. As it has been shown in these two papers, there is a basal deep split between a clearly exclusive African branch and the other one, encompassing the remaining African and non-African types. Furthermore, it has been observed that mtDNA lineages had the highest diversity in Africa, implicating that the particular ancestral variant (“mitochondrial Eve”) has been present in Africa much earlier than elsewhere. The following mtDNA evidences indeed supported the “Out-of-Africa” scenario (e.g.<sup>Chen *et al.* 1995b, Horai *et al.* 1995, Jorde *et al.* 1995, Watson *et al.* 1997, Ingman *et al.* 2000</sup>). In sum, the molecular identity of the mtDNA that is the MRCA of the present-day global pool of all mtDNAs, *i.e.* of the “African Eve”, is explained by coalescence theory, while its extant diversity is the result of molecular evolution, shaped further by demographic history of AMHs, including multiple worldwide dispersals, regional expansions and contractions and other events<sup>(Torrioni *et al.* 1994c; Macaulay *et al.* 1999b; Forster *et al.* 2001, 2002; Maca-Meyer *et al.* 2001; Salas *et al.* 2002; Kivisild *et al.* 2002)</sup>.

The maternal lineages of all living humans coalesce in a Southeast or East Africa cradle at about 160-200 kya (Figure 7a; RFLP and/or HVS-I analysis,<sup>Stoneking 1994, Horai *et al.* 1995, Watson *et al.* 1997, Kivisild *et al.* 1999a, Quintana-Murci *et al.* 1999, Stoneking 2000</sup> and complete sequences,<sup>Ingman *et al.* 2000, Maca-Meyer *et al.* 2001, Gonder *et al.* 2006, Torrioni *et al.* 2006</sup>), at a time frame coinciding with the palaeontological data, currently found for the emergence of early AMH<sup>(Day and Stringer 1982; Rightmire 1989, 2006; Grun *et al.* 1990; Rightmire and Deacon 2001; White *et al.* 2003; McDougall *et al.* 2005)</sup>. As indicated by comparisons of complete human mtDNA sequences with the chimpanzee outgroup<sup>(Mishmar *et al.* 2003, Kivisild *et al.* 2006b, Torrioni *et al.* 2006)</sup> the first emerging subsets of variation are haplogroup L0 and the branch common to all of the remaining variation. The next main diversification refers to the split between monophyletic L1 (includes haplogroups L1b and L1c) and its sister branch encompassing the African haplogroups L2’6 and all the non-African variation (see simplified schematic topology in section 2.7.1, Figure 8). It is widely accepted that the African environment was fragmented *ca.* 70 kya<sup>(Lahr and Foley 1994, 1998)</sup>, at around the same time when material culture and personal decorations testify for a rapid development of modern human behaviour in different parts of Africa<sup>(Henshilwood *et al.* 2002, Mellars 2002)</sup>. The maternal component is believed to have acquired new variation during the periods of isolation, so that the diversity outside of Africa can be

considered as a result of diversification within Africa 80–60 kya (Figure 7b; Forster 2004).

Haplogroup L3 arose, likely within the eastern African mtDNA pool, at about 65–75 kya and diverged into a multitude of subclades *in situ* (Figure 7b; Salas *et al.* 2002, Kivisild *et al.* 2004, Macaulay *et al.* 2005, Torroni *et al.* 2006, Kivisild *et al.* 2006b, Behar MD, Villems, R *et al.* ms submitted). Out of this plethora of African-specific L3 lineages just two – M and N (~60–65 kya; Forster *et al.* 2001, Kong *et al.* 2003, Mishmar *et al.* 2003, Macaulay *et al.* 2005) – have made their way out of Africa, giving rise to the mtDNA pool of all non-Africans. The basal N clade seems to have rapidly evolved into several branches, including haplogroup R (~60 kya, Figure 7c; Kivisild *et al.* 2003, Macaulay *et al.* 2005). As a consequence, all over Eurasia, America, Australia and Oceania autochthonous descendants of these three macrohaplogroups can be found (Richards *et al.* 1998a, 2000, 2003; Kivisild *et al.* 1999b; Macaulay *et al.* 1999b, 2005; Ingman and Gyllensten 2003; Kivisild *et al.* 2003; Metspalu *et al.* 2004, 2006; Friedlaender *et al.* 2005, 2007; Torroni *et al.* 2006; Sun *et al.* 2006; Hudjashov *et al.* 2007). However, there is no clear hint where exactly the two macrohaplogroups M and N *de facto* arose: so far not a single mtDNA lineage intermediate between basal L3, on one hand, and either basal M or basal N, on the other, has been sampled neither in Africa nor anywhere else, with both of the derived M and N haplogroups being quite distant from L3 node (four nucleotidic substitutions for M and five for N). An East African source might be considered, since a large and diverse population seems to have persisted in this area (Gonder *et al.* 2006), and, for instance, modern Ethiopians exhibit a highly variable pool of lineages around L3 node, when compared to other sub-Saharanans (Kivisild *et al.* 2004). Alternatively, because there are no “pre-M” or “pre-N” lineages in Ethiopia, whereas South Asia is very rich in autochthonous basal sub-clades of both M and N (including R), one can argue that these lineages arose outside Africa, on the way to, or already within, South Asia and elsewhere.

In the nowadays limelight of the debate are the questions about how many “Out-of-Africa” migrations have happened and which routes have been used. Two hypothetical independent routes are considered for the “Out-of-Africa” spread: i) over the Sinai Peninsula, through the Levant, with a further spread in the direction of Central Asia; ii) via Ethiopia and the Horn of Africa, by crossing the southern part of Red Sea around Bab-el-Mandeb, and further towards South Asia. In brief, the support for the southern route relies on archaeological artefacts of Middle Palaeolithic along a southern route to Australia (Lahr and Foley 1994; Stringer 2000, 2003; Bowler *et al.* 2003; Leavesley and Chappell 2004). The northern passage was favored by a theoretical historical link between the first Upper Palaeolithic stone blade technologies in the Levant (the “Aurignacian” period) and similar blade technologies in northern Africa (the “Dabban” sites, Bar-Yosef 2002), suggesting a movement at about 40 kya (Mellars 2004). Recent archaeological findings in the Kostenki sites, Russia, were found to be a of an early Upper Palaeolithic technocomplex, with no European analog and no obvious root in the local Russian Middle Palaeolithic and thus not “Aurignacian” nor transitional. Anikovich *et al.* (2007) argue that they represent a pioneering group of modern humans, implying that the first ones to colonize European Russia may not have spread from Levant via central Europe but instead came from interior western Asia via the Caucasus Mountains or

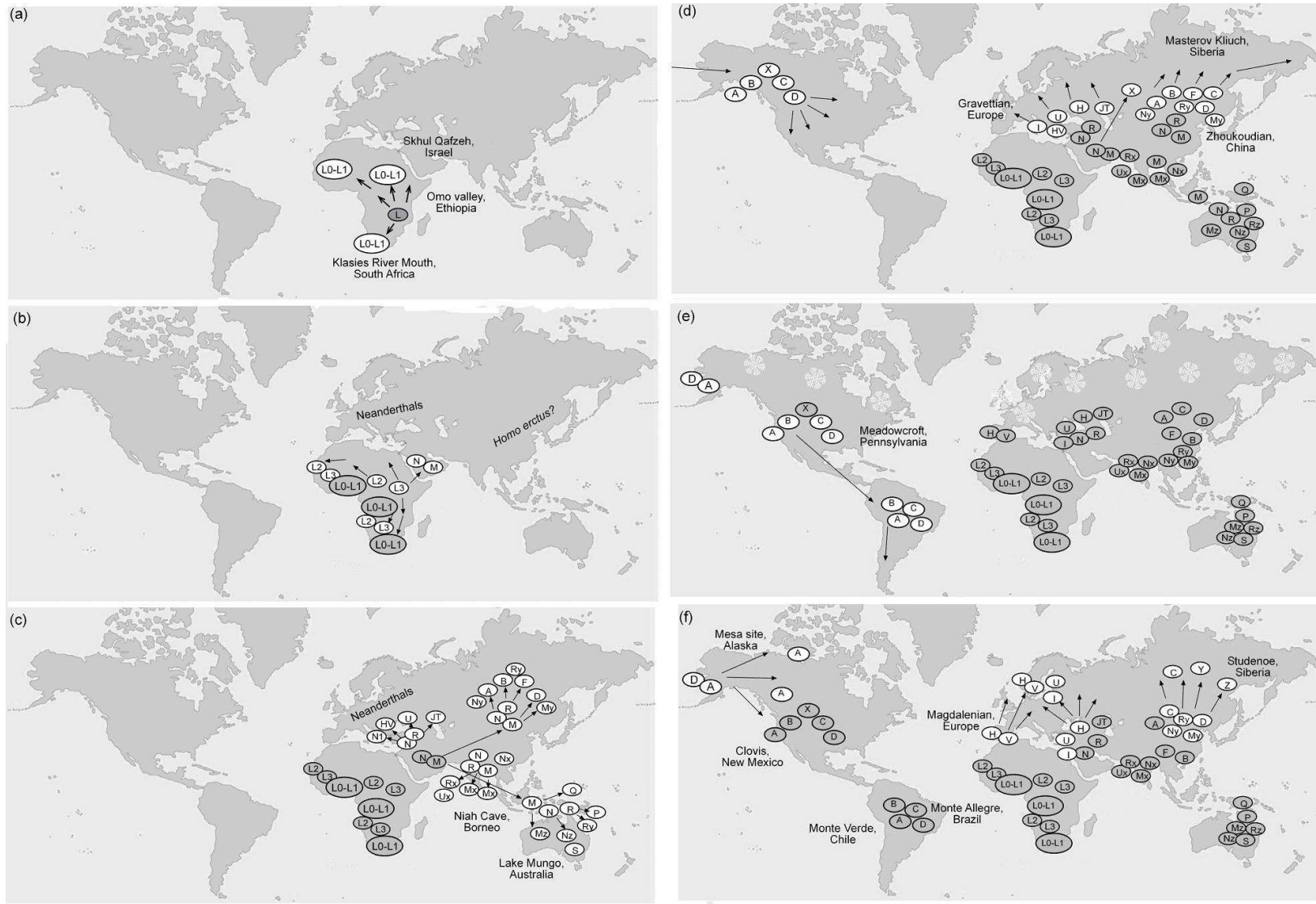


Figure 7 – Evolution, expansion and migration followed by mtDNA haplogroups across the world: a) 200-100 kya; b) 80-60 kya; c) 60-30 kya; d) 30-20 kya; e) 20-15 kya; f) 15-2 kya. The hypothetic scheme results from the analysis of the present-day mtDNA pool of local populations (redrawn from Forster 2004 with additional information from Kong *et al.* (2003), Metspalu *et al.* (2004), Palanichamy *et al.* (2004), Friedlaender *et al.* (2005), Macaulay *et al.* (2005), Merriwether *et al.* (2005), Kivisild *et al.* (2006b), Sun *et al.* (2006), and Hudjashov *et al.* (2007). In the considered geographic regions the extant pool is represented by haplogroups: Mx – M2-M6, M18, M25, M30-M40; Nx – N5; Rx – R5-R8, R30 and R31; Ux – U2a-U2c; My – M7-M12G and M13; Ny – N9; Ry – R9 and R11; Mz – M27-29 in Melanesia (M21 and M22 in New Guinea), and M42 in Australia; Nz – N12 ; Rz – R21 in New Guinea and P and R12 in Australia. The arrows indicate the direction of the migration, not the precise routes followed by the modern humans in the processes of “Out-of-Africa” and colonization of the several continents.

from further eastern or central Asia.

There is increasing genetic evidence corroborating for a single “Out-of-Africa” dispersal that has followed a coastal route towards southeast Asia, and has reached Australia 45-50 kya (Figure 7c). The fact that the basal variation of the L3-derived haplogroups M, N (and R) is still present among the ancestors of indigenous populations along the southern coast of Asia as well as those of Melanesia and Australia, became apparent from complete genome studies, revealing a plethora of independent basal lineages all over the area: among Aboriginal Australians (van Holst Pellekaan *et al.* 2006, Hudjashov *et al.* 2007), Melanesians (Merriwether *et al.* 2005, Friedlaender *et al.* 2007, Hudjashov *et al.* 2007 and references therein), island southeast Asia (Pierson *et al.* 2006, Hill *et al.* 2007), Papuans (Forster *et al.* 2001), Andaman islanders (Endicott *et al.* 2003, Thangaraj *et al.* 2005), the Orang Asli Malaysians (Macaulay *et al.* 2005, Hill *et al.* 2006), and in the Indian continent in particular (Kivisild *et al.* 1999b, 2003, 2006b; Metspalu *et al.* 2004, 2006; Palanichamy *et al.* 2004; Sun *et al.* 2006). Furthermore, since the three root types are distributed all over the southern route (Figure 7c-f) one may conclude that this pioneer migration, as well as the settlement of areas alongside the southern flanks of Eurasia and beyond, happened relatively rapidly (less than 5.2 ky according to Hudjashov *et al.* 2007).

The mtDNA lineages in southwestern Eurasia gained footholds and assisted to a range expansion further north into the inner parts of the continent, probably at about 45 kya, radiating into numerous region specific lineages (Figure 7c-d): South Asians harbour nowadays a panoply of M basal lineages (Metspalu *et al.* 2004, Sun *et al.* 2006, Chaubey *et al.* 2007), N5 (within N; Palanichamy *et al.* 2004), R5-R8, R30 and R31 (within R, Kivisild *et al.* 1999a, 2003; Palanichamy *et al.* 2004; Quintana-Murci *et al.* 2004); haplogroups A and N9 (including Y; within N), B’R11 and R9 (including F; within R) and several lineages assigned to macrohaplogroup M broadly characterize the northeastern Asians (Kong *et al.* 2003); haplogroups E and M7 (within M) and B4a and R9 (within R) are of typical Southeast Asian distribution (Forster *et al.* 2001, Kivisild *et al.* 2002, Merriwether *et al.* 2005) although substantial overlap exists. Other mtDNA types followed a fast coastal route and evolved by mutations to autochthonous haplogroups Q, S and P in Papua New Guinea (Forster *et al.* 2001) and into further subdivisions in Australia (Ingman and Gyllensten 2003, Friedlaender *et al.* 2005, Friedlaender *et al.* 2007, Hudjashov *et al.* 2007). New Guineans and Australians seem to share the same M and N founders, dating from the African exodus, and a furthermore characteristic variant not found elsewhere (Hudjashov *et al.* 2007). Taken together, the fact that the ancestral node is shared by Australians and Melanesians and the existence of specific branches, these authors argue for a single founder group settling the whole region, with posterior substantial isolation. The Asian A, B, C and D holders, actually the first ones to be described (Torroni *et al.* 1993, 1994c) have later crossed the Beringia Strait to populate the Americas (Forster *et al.* 1996), supposedly when the sea-level was substantially lower at about 14-37 kya allowing a wide land bridge (Figure 7d; Forster *et al.* 1996, Smith *et al.* 1999, Silva *et al.* 2002, Rubicz *et al.* 2003).

An early offshoot of the pool carried “out-of-Africa” eventually led to the peopling of West Eurasia (Figure 7c). It likely experienced assisted to a lengthy pause until the climate improved and their ancestors were able to enter the Levant, Anatolia and Europe. Haplogroup R then radiated in the

Near and Middle East and West Eurasia, to give a plethora of sub-branches, starting perhaps as early as at the beginning of Upper Palaeolithic (Figure 7c). At present, the family of macro-haplogroup R in West Eurasia is represented by haplogroups R0 (previous pre-HV, <sup>Torrioni *et al.* 2006</sup>), JT and U. In addition, three minor N1, N2 and X branches derive directly from the basal node of haplogroup N. Taken together, these variants make up 98% of the mtDNA pool characteristic of the extant Europeans (<sup>Torrioni *et al.* 1996; Richards *et al.* 1998b, 2002; Macaulay *et al.* 1999b</sup>) and which is shared with Near Easterns. Here, it is important to notice that the mtDNA pool of western Eurasians, as far as the basal M, N variation is concerned, is quite narrow compared with that one observes further eastwards, in fact starting from the Indus Basin (<sup>Chaubey *et al.* 2007</sup>). It is best explained by a side role that the Near East (and subsequently Europe) have played during the pioneer phase of the “Out-of-Africa” settlement of Eurasia (<sup>Torrioni *et al.* 2006</sup>).

Ice Ages, in particular the last one, have played a major role in shaping human diversity (<sup>Forster 2004</sup>). About 20 kya the Last Glacial Maximum (LGM) was reached, forcing humans to retreat southwards into refugia. As a result, genetic diversity was likely significantly reduced in the northern regions of Eurasia (Figure 7e). Distribution and age estimates of H1, H3, V and U5 support a repopulation of North Europe from the Franco-Cantabrian refuge after the LGM (Figure 7e; <sup>Torrioni *et al.* 1998, 2001b; Achilli *et al.* 2004, 2005; Loogväli *et al.* 2004; Pereira *et al.* 2005</sup>). The massive and rapid expansion is testified by the radiating phylogeny of H1 and H3 (<sup>Achilli *et al.* 2004</sup>) which dates are overlapping with the radiocarbon dates of the re peopling of north-western Europe ~16 ky (<sup>Gamble *et al.* 2004</sup>).

It is not clear yet to what extent the genetic legacy of the Palaeolithic population is reflected in the present European mtDNA pool because later immigrations from the Near East could have replaced the descendants of the first settlers (<sup>Ammerman and Cavalli-Sforza 1984</sup>), whereas back migrations, from Europe to Anatolia and to the inner Near and Middle East, may have carried European acquired mtDNA variation to the Near East (<sup>Richards *et al.* 2000</sup>). Hot debate between geneticists led to a proposal that about three-quarters of the European maternal inheritance originate from the indigenous Mesolithic or Palaeolithic contributors, as opposed to the Neolithic newcomers (<sup>Richards *et al.* 1996, 1997, 2000; Cavalli-Sforza and Minch 1997; Richards and Sykes 1998</sup>). It is important to stress that this rough estimate is in fact close to the conclusions drawn from classical markers, using principal component analysis (<sup>Cavalli-Sforza *et al.* 1994</sup>), regardless of the fact that the latter does not offer time dimension. Therefore, most of researchers agree today that the genetic legacy of the Neolithic farmers that spread in Europe ~5-10 kya represents a minor share in the maternal heritage of the present-day Europeans. This conclusion is also supported by the Y chromosome system (<sup>Semino *et al.* 2000, Rootsi *et al.* 2004</sup>). Returning to mtDNA, it has been suggested that younger sub-haplogroups seem to have been carried by farmer migrants, in particular J1b1, J2a and T1a and perhaps the rarer R1, R2 and N1a (Figure 7f; <sup>Richards *et al.* 2000</sup>). The work of Haak *et al.* (<sup>2005</sup>) based on Neolithic remains proved that N1a was present in Europe at about 7 kya, among the carriers of the typically Neolithic Central European Linear Pottery culture, at more than



a hundred times higher frequencies than observed today. This supports the idea that only a minor fraction of the Neolithic migrant genetic pool has reached the present-day, later excessively diluted by the genetic legacy of Mesolithic Europeans. However, the process of Neolithization in Europe may have used several routes and spread scenarios – including one alongside the Danube Basin, the other alongside the Mediterranean coast.

### 2.7.1 – Phylogeography of the African mtDNA variation

It has been shown that the genetic diversity of mtDNA in Africa is considerably greater than elsewhere, with the most prevalent haplogroups having variable distribution and sub-structuring when geography and ethnolinguistic affiliations are considered (Watson *et al.* 1997, Chen *et al.* 2000, Pereira *et al.* 2001b, Salas *et al.* 2002, Destro-Bisol *et al.* 2004). Nevertheless mtDNA variation seems to be more structured by geography, as it is the case of West Africa (Gonzalez *et al.* 2006). The latter authors even suggest that languages have spread within Africa mainly as a cultural imposition over an already genetically diverse landscape. The analysis of spatial distribution of lineages allowed to reveal and to better understand demographic changes within a Middle and Late Stone Age chronological frame up to the present day.

The following intends to summarize the present-day knowledge about African mtDNA variation, with proposed origins and coalescence ages of the haplogroups and its subclusters. We describe as well the region/population specific clades and the more relevant migrational events that can be traced in a genetic basis. For convenience, the coding region and HVS-I positions defining the main clades of African variation are summarized in Figure 8 and Table 1, respectively. Note, however, that time estimates do not exactly parallel, especially if based on HVS-I against coding region, or even against the recently proposed calculation based solely on coding region synonymous mutations (Kivisild *et al.* 2006b).

The human mtDNA phylogeny coalesces in a time depth of about 150-200 kya (Horai *et al.* 1995, Ingman *et al.* 2000, Maca-Meyer *et al.* 2001, Gonder *et al.* 2006, Torroni *et al.* 2006). One of the earliest offshoot of the phylogenetic tree, haplogroup L0 further includes sub-haplogroups L0a, L0d, L0f and L0k (see Figure 8; Mishmar *et al.* 2003, Salas *et al.* 2004, Gonder *et al.* 2006, Kivisild *et al.* 2006b, Torroni *et al.* 2006). As already indicated above, in order to avoid mis-interpretation of the earlier literature, it is important to note that these clades were previously reported as branches of the “original” L1 (Watson *et al.* 1997, Salas *et al.* 2002).

L0d is the first individual sub-clade to derive from the L0 node. Its distribution appears to be restricted to Khoisan people in South Africa and to Tanzanian populations (Vigilant *et al.* 1991, Soodyall and Jenkins 1992, Bandelt and Forster 1997, Wallace *et al.* 1999, Pereira *et al.* 2001b, Salas *et al.* 2002, Kivisild *et al.* 2004, Gonder *et al.* 2006). Sub-haplogroup L0k, a branch of L0afk in Figure 8, is found exclusively among South African Khoisan (Chen

*et al.* 2000, Salas *et al.* 2002). More than a half of the extant mtDNA pool of Khoisan speakers is constituted by indigenous L0d and L0k lineages, possible relics of a widespread and ancient proto-Khoisan population (Bandelt and Forster 1997, Chen *et al.* 2000, Pereira *et al.* 2001b, Salas *et al.* 2002). Conversely, L0d presence in southeast Bantu may be due to assimilation or recurrent gene flow from Khoisan, though L0k is not so far sampled in Bantu-speaking populations. The present topology of haplogroup L0 allows to identify L0k as a sister clade to the one that included both L0f and L0a (Kivisild *et al.* 2006b).

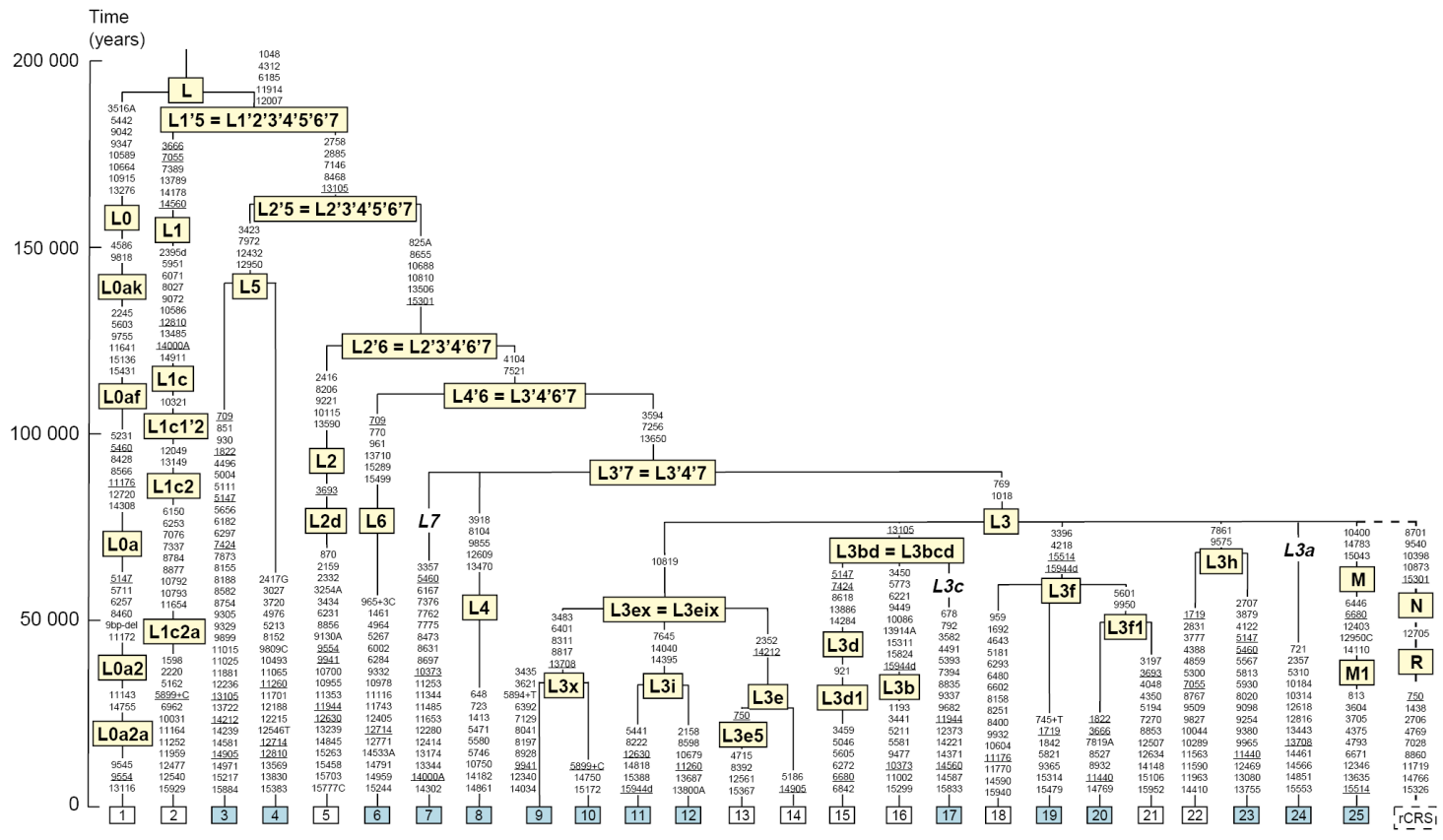
The marginally frequent lineages classified as L0f are more divergent and frequent only in East Africa, with their highest incidence in Tanzanians (Watson *et al.* 1997, Pereira *et al.* 2001b, Salas *et al.* 2002, Knight *et al.* 2003). L0a lineages are widely spread through eastern, central and southern Africa, encompassing almost a quarter of maternal lineages there (Soodyall and Jenkins 1992, 1993; Chen *et al.* 1995b; Watson *et al.* 1997; Macaulay *et al.* 1999b; Salas *et al.* 2002, 2004; Kivisild *et al.* 2004), while only single examples of L0a are found in the West Africans (Graven *et al.* 1995, Rosa *et al.* 2004). The L0a mtDNA variants coalesce back to a common node at Palaeolithic times in East Africa (Watson *et al.* 1997, Salas *et al.* 2002). The L0a1 subclade has an eastern and southeastern African distribution, with the root type coalescing at ~33 kya (HVS-I estimate, Salas *et al.* 2002). A 9-bp deletion in the COII-tRNA<sup>Lys</sup> intergenic region characterizes the L0a2 lineages, that are believed to represent the dispersal of the Bantu-speakers (Soodyall *et al.* 1996) from its source region nearby the Congo Basin (Soodyall *et al.* 1996, Chen *et al.* 2000).

L1 variation was found to coalesce at about 140-150 kya (see Figure 8; Torroni *et al.* 2006, Behar MD, Villems R, *et al.* ms submitted) or at about 112 ky if calculations are according to Kivisild's molecular clock (Kivisild *et al.* 2006b, Underhill and Kivisild 2007). One of its daughter clades, haplogroup L1b is concentrated in western Africa, particularly along the coastal areas (Graven *et al.* 1995, Mateu *et al.* 1997, Watson *et al.* 1997, Rando *et al.* 1998, Rosa *et al.* 2004) and peaks also in Mauritania (~19%, Gonzalez *et al.* 2006) and Senegal (Graven *et al.* 1995, Rando *et al.* 1998). Its sister clade L1c is mainly found in Central and West Africa (Watson *et al.* 1997, Rando *et al.* 1998, Destro-Bisol *et al.* 2004, Rosa *et al.* 2004), with a subgroup common in Biaka Pygmies (Wallace *et al.* 1999). Both L1b and L1c have a proposed origin in Central Africa, therefore their presence in West Africa suggests a westwards expansion. The extant variation of L1b suggests much later expansion than that for its sister clade L1c (Table 1; Salas *et al.* 2002). The reason for that can be that earlier branches of L1b have not survived or have not been sampled yet, because it is very unlikely that the nodal L1 survived at a time depth of 30 kya to give rise to an expansion. This is likely a clear-cut example where a large pattern of variation become extinct, leaving no other progeny except a clade that began expanding at about 30 kya (Behar MD, Villems R, *et al.* ms submitted). This bottleneck and re-expansion in West Africa seem to have shaped the evolution of L1b. Their spread to Northwest Africa was probably more recent, in the Neolithic or during the times of the slave trade (Rando *et al.* 1998).

Haplogroup L5 (Kivisild *et al.* 2004; previously referred to as L1e in Pereira *et al.* 2001b, Salas *et al.* 2002, 2004) has been observed at low frequency only in eastern Africa (Salas *et al.* 2002) and Egypt (Stevanovitch *et al.* 2004), with minor gene flow into the Mbuti Pygmies (Salas *et al.* 2002). The central African Pygmies-specific intra-population divergence in both L1c and L5 may signal a “relict” status, similar to that proposed for the Khoisan (Chen *et al.* 2000).

Haplogroup L2 has a pan-African distribution and together with L3 comprises ~70% of the summary sub-Saharan variation of mtDNA (Chen *et al.* 1995b, 2000; Graven *et al.* 1995; Watson *et al.* 1997; Salas *et al.* 2002). Chen *et al.* (2000) and Torroni *et al.* (2001a) dissected haplogroup L2 into sub-clades L2a, L2b, L2c and L2d. Haplogroup L2a is the most frequent and widespread mtDNA cluster in Africa, making its geographic origin very difficult to identify. Furthermore, its structure represents a phylogenetic challenge, if to rely on HVS I variation only because of reticulation of mutations. Salas *et al.* (2002) identified sub-clusters of variation assuming the main reticulations on HVS-I nps 16189 and 16192, positions of rapid transition known to undergo forward and reverse mutations (Howell and Smejkal 2000). The clade appears to be further subdivided by a more stable HVS-I 16309 transition. Howell *et al.* (2004) considers the ancestral L2a already with 16309 A:G transition on the contrary of Salas *et al.* (2002), but precise comparisons are not possible as coding region information is lacking in the earlier work. The deepest types of this clade (L2a- $\alpha$  in Salas *et al.* 2002) are most common in East Africa and of putative origin at about 55 kya. However, the coalescence age and diversity of the L2a sub-clades in West Africa are similar to the ones in the East. A possible origin is then placed between East and West, followed by separate dispersals along the Sahel corridor after the LGM (main shared founder types ~14 kya, Salas *et al.* 2002). Recent star-like demographic bursts in L2a1a and L2a2 and their expansion to southeast Africans is most likely associated with the expansion(s) of the Bantu-speaking populations during the sub-Saharan agricultural spread and later (Pereira *et al.* 2001b, Salas *et al.* 2002).

L2b-L2d haplotypes are largely confined to West and Central West Africa. L2b mtDNAs are absent in East Africa (Watson *et al.* 1997, Krings *et al.* 1999), rare in the southern populations of the continent (Vigilant *et al.* 1991, Chen *et al.* 2000, Pereira *et al.* 2001b) but common in Senegal (Chen *et al.* 1995b, Rando *et al.* 1998). L2c mtDNAs have a similar distribution and coalescence time as their sister clade L2b, being also dominant in West Africa. An expansion time in the scale of 18 ky (Chen *et al.* 2000), indicates possible expansion in West Africa together with L1b (Chen *et al.* 1995b, Rando *et al.* 1998). However, their predecessors might have been present in the area long before the expansion that gave rise to the present variation. At an approximate age of 120 kya (based on HVS-I calculation, see Table 1; Salas *et al.* 2002, Rosa *et al.* 2004) it seems unlikely that L2d have diverged in West Africa. Given the period of potential drift and extinction, the data are more consistent with its Central African origin. A single type of subclade L2d1 in the Bubi of Bioko and its absence in southeastern Africans may link to its origin (Mateu *et al.* 1997, Salas *et al.* 2002).



TRENDS in Genetics

Figure 8 – A maximum-likelihood phylogeny of African mtDNA haplogroups, with a calibrated time-scale show on left (in years). The phylogeny shows the most parsimonious reconstruction of coding region mutation in the nps 577-16023 range from 25 entire mtDNAs (references in Torroni *et al.* <sup>2006</sup>, using two chimpanzee sequences to root the whole tree (as described in Macaulay *et al.* <sup>2005</sup>). Mutations are scored relative to the revised Cambridge Reference Sequence (rCRS; Andrews *et al.* <sup>1999</sup>) and therefore variant nucleotides do not necessarily constitute derived states; the rCRS is shown by the dashed lines. Suffixes indicate transversions (to A, G, C or T) or indels (C, d); recurrent mutations are underlined. The naming of haplogroups follows the scheme of Richards *et al.* <sup>(1998b)</sup>, opting for the most compact notation. L3a and L3c denoting novel L3 branches from this study should not be confused with the obsolete definitions of ‘L3a’ and ‘L3c’ (=U6a) introduced by Watson *et al.* <sup>(1997)</sup>. The ancestral motifs for the novel haplogroups L7 (corresponding to L4g in Kivisild *et al.* <sup>2004</sup>), L3a, and L3c (in italics) have not been determined. However, note that L4 and L7 (L4g) were clustered together in Kivisild *et al.* <sup>(2004)</sup> because of their sharing of the (unstable) control-region mutation 16362. In addition, observe that there are minor differences with Kivisild *et al.* <sup>(2006b)</sup> in characterizing the mutations for some haplogroups. MtDNA sequences in sky-blue colour boxes are from Ethiopian subjects, whereas those in white boxes are from Nigerian (sequences 13, 18 and 21) and Dominican (sequences 1–2, 5, 14–16 and 22) subjects. Complete mtDNA sequences, including their control regions, are available in GenBank, accession numbers DQ341058–DQ341082. *In* Torroni *et al.* <sup>(2006)</sup>.

Table 1 - HVS-I sequence motifs used for haplogroup classification and correspondent coalescence ages (as in Salas *et al.* <sup>2002, 2004</sup>)

Haplogroup	HVS-I sequence <sup>a</sup>	TMRCA (ky)	SE (ky)
L0a	129-148-172-187-188G-189-223-230-311-320	40,4	16,3
L0a1	129-148-168-172-187-188G-189-223-230-311-320	33,4	16,6
L0a2	148-172-187-188G-189-223-230-311-320	8,3	3,7
L0d	129-187-189-223-230-243-311	49,6	13,5
L1b	126-187-189-223-264-270-278-311	30,6	16,3
L1c	129-187-189-223-278-294-311-360	59,7	11,8
L2	223-278-390	70,1	15,3
L2a	223-278-294-390	55,2	19,4
L2b	114A-129-213-223-278-390	31,6	11,2
L2c	nr	27,5	7,3
L2d	223-278-390-399	121,9	34,2
L3	223	61,3	11,7
L3b	124-223-278-362	21,6	6,9
L3d	124-223	30,3	8,5
L3e1	223-327	32,2	11,5
L3e2	223-320	37,4	18,4
L3e3	223-265T	14,2	4,5
L3e4	223-264	24,2	10,4
L3f	209-223-311	36,3	12,8
L3i <sup>b</sup>	153-223	nd	nd
L3x <sup>b</sup>	169-223-311	68,6	21,5
L3w <sup>b</sup>	223-260-311	5,8	4,1
L3h <sup>c</sup>	nr	nd	nd
L4a <sup>b</sup>	223-260	51,2	17,2
L4g <sup>d</sup>	223-293T-311-355-362	45,1	12,5
L5 <sup>e</sup>	129-148-166-187-189-223-311	83,0	24,9
L6 <sup>b</sup>	223-224-278-311	38,9	25,6
M1 <sup>f</sup>	129-189-249-311	36,8	7,1
U5b1b <sup>g</sup>	189-270	8,6	2,4
U6 <sup>f</sup>	172-219	37,5	4,3

<sup>a</sup> motif relative to rCRS minus 16000 bp; <sup>b</sup> newly described in Kivisild *et al.* <sup>(2004)</sup>; <sup>c</sup> newly described in Rosa *et al.* <sup>(2004)</sup>; <sup>d</sup> renamed in Kivisild *et al.* <sup>(2004)</sup>, L3g in Salas *et al.* <sup>(2002)</sup>; <sup>e</sup> misassigned as L1e in Salas *et al.* <sup>(2002, 2004)</sup>; <sup>f</sup> Olivieri *et al.* <sup>(2006)</sup>; <sup>g</sup> Achilli *et al.* <sup>(2005)</sup>; nr - not recognizable from the closest paraphyletic cluster in the basis of HVS-I sequence; nd - not determined.

The variation classified as haplogroup L6 is largely confined to and frequent in Yemenis (~12%). Its East African origin is likely, given its presence in Ethiopians and the fact that its sister clades are all diverse and frequent there (Kivisild *et al.* 2004). It is noteworthy that L6, has a very narrow phylogeography, although the *ca.* 110 ky coalescence with its L3'4 sister clades (Torrioni *et al.* 2006). However, its own coalescence is only around 23 ky (Behar MD, Villems R, *et al.* ms submitted) presumably because the past variation was wiped out or actually never expanded thanks to drift in very small and isolated communities. In any case we may still be missing the homeland of L6, given large areas of missing sampling in East Africa, e.g. Somalia.

Haplogroup L4 is a sister clade of L3, typical for East and Northeast Africa although present at low frequencies (Watson *et al.* 1997, Krings *et al.* 1999, Kivisild *et al.* 2004). The L4a motif has been found in Sudan and Ethiopia though misclassified earlier as L3e4 in Salas *et al.* (2002). Following the review by Torrioni *et al.* (2006), haplogroup L7 (corresponding to L4g in Kivisild *et al.* 2004) is also considered as a "sister clade" of L3 and L4 (Figure 8), if to ignore shared by L4 and L7 transition at np 16362, taken as diagnostic by Kivisild *et al.* (2004). The sister clade L4g/L7 displays one particular motif quite frequent in Tanzania (Salas *et al.* 2002, Kivisild *et al.* 2004, Gonder *et al.* 2006).

An East African origin at about 65-75 kya years ago is pointed out for superhaplogroup L3 (Salas *et al.* 2002, based on HVS-I information; Kivisild *et al.* 2006b, coding region using "synonymous clock"; Macaulay *et al.* 2005, Torrioni *et al.* 2006 and Behar MD, Villems R, *et al.* ms submitted, using Mishmar's clock), the cluster that further harbors all the non-Africans mtDNAs. It is widespread in Africa and provides evidence for a mainly sub-Saharan expansion of its sub-clades (Watson *et al.* 1997, Salas *et al.* 2002) with a gradient of decreasing frequency and diversity from East to West Africa. From 20% of L3\* undefined eastern lineages in Salas *et al.* (2002), three novel L3 subclades (L3i, L3x and L3w, Kivisild *et al.* 2004) were described in Ethiopia and Yemeni samples, with another one - L3i - potentially recognized in the Sudanese of (Krings *et al.* 1999).

Both L3b and L3d are prevalent in the West quadrant of sub-Saharan Africa and their split from a common node occurred at about 20-30 kya (Vigilant *et al.* 1991, Soodyall and Jenkins 1993, Watson *et al.* 1997, Rando *et al.* 1998, Pereira *et al.* 2001b, Salas *et al.* 2002, Rosa *et al.* 2004). A subset of L3b with sequence motif 124-223-278 is common among Bantu speakers of southwestern Africa and thus is another possible marker of the Bantu expansion (Watson *et al.* 1997). The occurrence of a western African-specific subcluster of L3b, coalescing in a lineage with transitions at nps 124-223 (but not 278), may suggest that L3 has reached Western Africa by 30 kya (Watson *et al.* 1997).

The L3e cluster has been subdivided into L3e1, L3e2, L3e3 and L3e4, based of HVS-I information (Bandelt *et al.* 2001). The oldest branches of L3e are thought to have arisen in areas

neighbouring Central Africa/Sudan ca. 45 kya, from where spread throughout sub-Saharan Africa, comprising by now about one third of L3 types of these people <sup>(Bandelt *et al.* 2001, Salas *et al.* 2002)</sup>. Although it probably arose in central areas of the continent, L3e1 became frequent in southeast Bantu-speakers. Again a link to an eastern Bantu route may justify its presence in Kenya and southern regions. Within L3e2, the L3e2b lineages constitute the most frequent and widespread type of L3e, primarily found in West Africa (range expansion ~9 kya, <sup>Salas *et al.* 2002</sup>). Together with L3e2a, these were supposedly successful hitchhikers of the population movement in the Sahara during the Great Wet Phase of the early Holocene and subsequently Wet Phase <sup>(Muzzolini 1993, Bandelt *et al.* 2001)</sup>. Meanwhile, L3e4 is essentially restricted to Atlantic West Africa, signalling much later dispersals and local expansion events with the rise of food production and the iron smelting (see section 4).

As in Salas *et al.* <sup>(2002)</sup>, L3f retains all L3\* lineages with HVS-I 16209 mutation. The spread zone of haplogroup L3f appears to be mostly in East Africa, being the most frequent L3 type in Ethiopia <sup>(Kivisild *et al.* 2004)</sup>. The few matches to L3f1 founder lineages in Central and West Africa <sup>(Salas *et al.* 2002, Rosa *et al.* 2004)</sup> point to an early rather than recent dispersal of the lineages while in East Africa it supposedly started to expand ~10 kya <sup>(Kivisild *et al.* 2004)</sup>.

Although L3h lacks a distinctive HVS-I motif, it can be classified by the 9575 coding region substitution. The subset of variation with motif 16129-16223-16256A-16311-16362 was first reported in the context of the present Guinean survey <sup>(Rosa *et al.* 2004)</sup>. Similar haplotypes are found in Cape Verde and Niger/Nigeria at low frequencies (~1%), but it reaches its highest known frequency in the Ejamat people in Guinea (8%, <sup>Rosa *et al.* 2004</sup>). Other close variants of this putative sub-clade are found in Ethiopian Amharans, though they lack substitutions at nps 16129 and 16362 <sup>(Kivisild *et al.* 2004)</sup>.

The present distribution and coalescence ages for the deepest branches in the mtDNA tree testify for the early modern human presence in East and South Africa, probably the result of a moderate range expansion <sup>(Watson *et al.* 1997, Salas *et al.* 2002)</sup>. The starlike phylogeny and wide distribution of many subclades within L2 and L3 testify for major demographic expansion(s) not earlier than 60 kya. The subsequent population fragmentation and re-expansion at Late Stone Age induced the clades to evolve into regional-specific clusters, and thus had a major impact on the modern sub-Saharan mitochondrial phylogeographic structuring. Haplogroups L1b, L3b and L3d suggest that West Africa has been occupied at least since ~20-30 kya <sup>(Salas *et al.* 2002)</sup>.

One has to consider as well that an earlier pattern of the distribution of mtDNA haplogroups may have been significantly altered by subsequent demographic processes. In this context, Salas and colleagues <sup>(2002)</sup> emphasize the role of the LGM conditions as well as the migrations of Bantu-speaking

people. In sub-Saharan, the “Last Glacial Aridity Maximum” (LGAM) climatic alterations culminated in the reduction of woodlands and savannas to a small fraction of the Congo basin at about 14.5 kya (see section 4.1; Adams and Faure 1997). These may have acted as a refuge area from which modern humans later dispersed: haplogroup L2a was fractioned east and westwards and L1b possibly expanded to the west.

The Bantu migrations are among, if not the most important recent demographic upheavals in African history, supposed to have started at about 3kya or slightly earlier from a central source in the vicinity of Cross River Valley (western Central Africa; Huffman 1982, Phillipson 1993). The movement is associated with the transition of a hunter-gathering to agricultural lifestyle and the advent of iron-smelting, therefore promoting a populational growth. Two main spread routes followed east- and westwards in direction to the south. There are genetic evidences from both mtDNA and Y chromosome systems that testify for the strong impact of the Bantu migrations on the gene pool of sub-Saharan Africa almost to the point of erasing the pre-existent one. It is likely that this expansion was the main mechanism explaining the spread of haplogroups L0a2 (Bandelt *et al.* 1995, Chen *et al.* 1995b) and L3b (Watson *et al.* 1997) and fragments of haplogroups L2, L3e and L5 (Alves-Silva *et al.* 2000, Bandelt *et al.* 2001, Pereira *et al.* 2001b) from West, Central, and East Africa towards the south. Such “Bantu-markers” are therefore helpful in specifying the routes and the pattern of admixture of their carriers with the local populations on their southward migrations. The more ancestral L0-L1 types eventually become a minority, except maybe in the ancestors of the Khoisan-speakers (Bushmen of South Africa) and the Biaka (West Pygmies in Central Africa; Vigilant *et al.* 1991, Watson *et al.* 1997, Chen *et al.* 2000), interpreted as the surviving footprints of the ancient variants. Overall, the African diversity observed nowadays combines levels of ancient population differentiation with that reflecting more recent gene flow episodes (Kivisild *et al.* 2006a).

A few non-L mtDNA haplogroups can be found in the African continent as well. Among them haplogroup M1 is mostly restricted to East Africa (Passarino *et al.* 1998, Quintana-Murci *et al.* 1999, Richards *et al.* 2003, Kivisild *et al.* 2004) despite occasional occurrences in West and Northwest Africa (Torroni *et al.* 1996, Rando *et al.* 1998, Rosa *et al.* 2004) and Nile Valley (Krings *et al.* 1999). The analysis of complete mtDNA sequences identified a basal split in the African M1 phylogeny, giving rise to M1a and M1b sister clades at about 28.8 and 23.4 kya, respectively (Olivieri *et al.* 2006), each encompassing several independent basal branches. While M1a ranges the entire geographical distribution of M1, the sub-haplogroup M1b (defined by a transition at 16185, previously named M1c in Kivisild *et al.* 2004) virtually covers the haplogroup’s variation in Northwest Africa and the Near East. It is therefore difficult to interpret M1b as an East African derivate and it might be better explained as a branch that followed a trajectory in the southern basin of the Mediterranean (Olivieri *et al.* 2006). An ancient arrival of M1 to Africa (or to its vicinity) is supported by the fact that none of the numerous M haplogroups in Asia harbour any of the M1-characteristic mutations (Kong *et al.* 2006, Sun *et al.* 2006, Chaubey *et al.* 2007) and by the lack of other Asian-specific clades within M in Africa,



as would be expected in case of more recent arrival. This ancient back migration from Asia to Africa had been already suggested by analysis of HVS-I lineages (Quintana-Murci *et al.* 1999, Richards *et al.* 2003, Forster 2004, Kivisild *et al.* 2004).

Haplogroup U6 is said to be autochthonous of North Africans (Corte-Real *et al.* 1996, Macaulay *et al.* 1999b) but is also present in eastern Africans (Kivisild *et al.* 2004), a situation that parallels that of haplogroup M1. This haplogroup is seen as the first Palaeolithic return to Africa of ancient Caucasoid lineages (40-50 kya, Rando *et al.* 1998, Olivieri *et al.* 2006). In fact, similar phylogenetic lineages have been sampled in Eurasia and the Near East (Di Rienzo and Wilson 1991) so that it has most likely migrated from the Near East-Mediterranean area and dispersed to Northwest and East Africa. The most representative of its clades, U6a (~38 ky based on mtDNA complete sequences of Olivieri *et al.* (2006), displays an increasing frequency and diversity pattern towards Northwest Africa, supporting the idea of a local drift. The most frequent motif 16172-16189-16219-16278 is believed to have started to expand ~11 kya (Rando *et al.* 1998), with partial diffusion to the Sahel (Rando *et al.* 1998, Rosa *et al.* 2004, Coia *et al.* 2005).

The temporal overlap of haplogroups M1, U6 (Olivieri *et al.* 2006) and U5 (Richards *et al.* 2000, Achilli *et al.* 2005) with the events that led to the peopling of Europe by AMHs, raises the possibility that their molecular ancestors lived in the same broad geographical area of Southwest Asia (possibly in separate regional enclaves) and that they later expanded towards the Near East and through Levant (Olivieri *et al.* 2006). U6 and M1 differentiated into their subclades while in the Mediterranean area, with the carriers of U6 and M1b mtDNA genomes inhabiting broadly the same geographic areas.

The main radiation of U5 took place in Europe, where it reached in early Upper Palaeolithic, most probably from Middle East/Caucasus region (~40-50 kya, Richards *et al.* 2000). An unexpected finding linked the Saami of the Scandinavia to the Berbers of North Africa and the sub-Saharan Fulbe, in a relatively recent branch of U5b1b ~9 kya (Achilli *et al.* 2005). If we parallel the situation to that of H and V sub-haplogroups, which attain their highest diversity in Iberians and Moroccan Berbers as post-glacial signatures (Achilli *et al.* 2004), it might have happened that both haplogroups have crossed the Strait of Gibraltar. The Franco-Cantabrian refuge area is then seen as the source of late-glacial expansions of hunter-gatherers that repopulated northern Europe, which also contributed to the mtDNA pool of North Africans, likely including the ancestors of Berbers (Achilli *et al.* 2005). Specific sub-clades of U5b were found at very low frequencies across sub-Saharan West Africa (Rando *et al.* 1998, 1999; Rosa *et al.* 2004; Coia *et al.* 2005) giving support to the hypothesis of hitchhiking episodes of North Africans that have crossed the Sahara.

### 3 –Phylogenetic analysis of the Y chromosome

#### 3.1 - Structure and organization of the Y chromosome

The Y chromosome represents a nuclear chromosome whose biological importance relies on the sex-determining <sup>(Ford *et al.* 1959, Jacobs and Strong 1959)</sup> and male fertility roles (e.g. <sup>Tiepolo and Zuffardi 1976, Levy and Burgoyne 1986, Ma *et al.* 1993, Vogt 1997, Wyckoff *et al.* 2000</sup>). As it is of haploid nature and has no homologous chromosome to recombine with, it is expected to be transmitted from father to son unchanged, defining paternal lineages. This is the case of 90% of its length – the Non-Recombining region of the Y chromosome (see NRY in Figure 9, also refer to as Male-Specific region - MSY). The remaining portion - the pseudoautosomal region (PAR) – is located in the telomeres and shows partial homology to the X-chromosome, being therefore prone to recombination <sup>(Cooke *et al.* 1985, Simmler *et al.* 1985, Freije *et al.* 1992, Li and Hamer 1995, Lien *et al.* 2000)</sup>. The PAR is divided into two flanking segments of less than 3 Mb of its approximately 67 Mb total length (Figure 9). Throughout this text the term Y chromosome is many times used as synonymous of NRY, because molecular markers to be discussed are chosen from this region of the chromosome.

For a long time thought to contain a large amount of junk-DNA, and thus to be quite non-polymorphic, the structural and functional features of the Y-chromosome started to be better explored in the last decade. Until recent years the nucleotidic composition was only accessed for the AZFa and AZFc portions (Figure 9; <sup>Sun *et al.* 2000, Kuroda-Kawaguchi *et al.* 2001</sup>), because of its medical interest, associated with male infertility (e.g. <sup>Vogt 1998, 2004, 2005; Lin *et al.* 2005</sup>). The first detailed physical map of Y chromosome was published by Tilford and collaborators in 2001. A couple of years later the sequence of 97% of the NRY roughly revealed about 160 transcription units, of which half encode for 27 proteins or protein families (12 expressed ubiquitously and 11 are testis-specific, <sup>Skaletsky *et al.* 2003</sup>). Many genes in NRY are Y-specific while others have known functional homologous in the X-chromosome, and therefore might have functions crucial in both sexes (see housekeeping genes in Figure 9; <sup>Lahn and Page 1997</sup>). The knowledge of its sequence provided a starting point for detailed study of Y chromosome diversity in humans, their implicit mutational processes and the role of such DNA portion in human disease. The

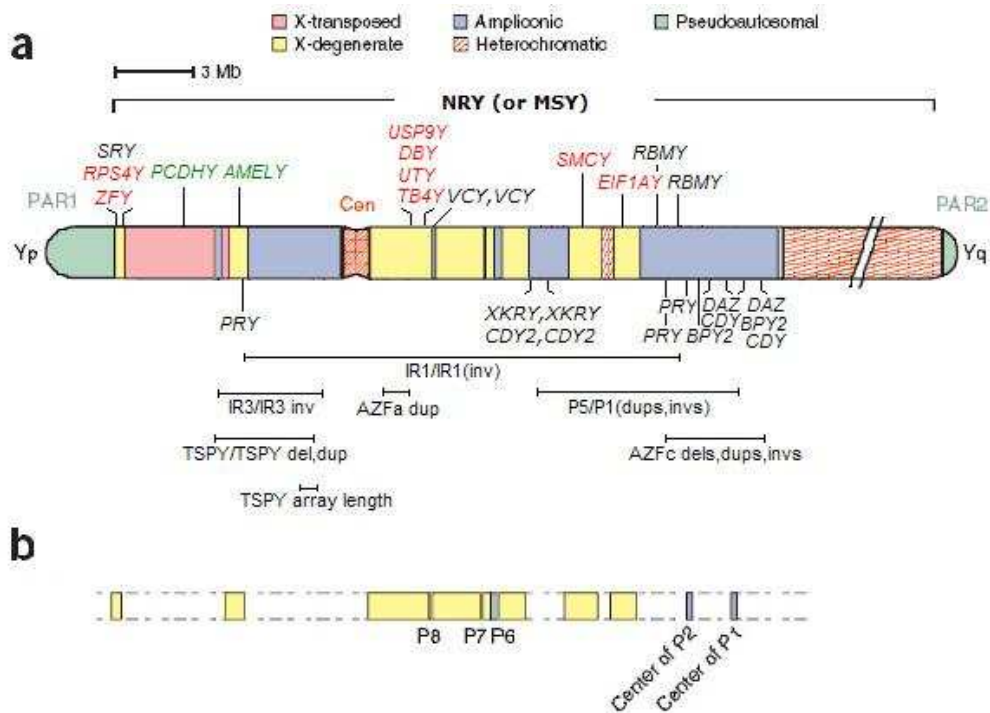


Figure 9 – Schematic representation of the human Y chromosome. a) Structure of the reference Y chromosome, based on sequence information that was largely derived from a haplogroup-R individual in Skaletsky *et al.* <sup>(2003)</sup>. The scheme includes both short and long arms (Yp and Yq), the pseudoautosomal regions 1 and 2 (PAR1 and PAR2), centromere (Cen), heterochromatic and NRY regions. The potential structural variation in the NRY is emphasized, by the location of three classes of euchromatic sequences (X-degenerated, X-transposed and ampliconic), inverted repeats (IR) and palindromes (P). Deletion, duplication and inversion segments are identified (“del”, “dup” and “inv”, respectively). Approximate locations of the some genes are shown (in italic), selected according to the resolution of this map. Genes named at the top of the chromosome have active X-chromosome homologues whereas the ones in the bottom lack known X homologues. The genes in red are widely expressed housekeeping genes; genes in black are expressed in the testis only; and genes in green are expressed neither widely, nor testis specifically. With the exception of the SRY (sex-determining region Y) gene, all the testis-specific Y genes are multicopy. A 3-Mb bar indicates the scale of the diagram. b) Structural elements conserved between human and chimpanzee Y chromosomes are shown according to their position in the reference human Y chromosome. Adapted from schemes in Lahn *et al.* <sup>(2001)</sup>, Jobling and Tyler-Smith <sup>(2003)</sup>, Skaletsky *et al.* <sup>(2003)</sup>, and Repping *et al.* <sup>(2006)</sup>.

more recently sequenced Y chromosome in chimpanzees <sup>(Hughes *et al.* 2005, Kuroki *et al.* 2006)</sup> constitutes the second well characterized mammalian chromosome for comparative analysis.

The Y chromosome includes both heterochromatic and euchromatic portions, with the totality of genes located in the later. Besides the large block of heterochromatic sequences found in the centromeric region of each nuclear chromosome <sup>(Schueler *et al.* 2001;</sup> about 1Mb in the Y chromosome, <sup>Tyler-Smith *et al.* 1993)</sup>, the male-specific chromosome contains a block of approximately 40 Mb that roughly

comprises the bulk of the distal arm (Caspersson *et al.* 1970, Pearson *et al.* 1970). A third heterochromatic block demarks a region of approximately 400kb with 3000 tandem repeats of 125 bp, that interrupts the euchromatin in the proximal Yq (Skaletsky *et al.* 2003). The entire sequence of euchromatin comprises 14.5 Mb in the Yq long arm and 8 Mb in the Yp short arm, plus two minor gaps of ~1.5 Mb. Three classes of Y chromosome elements are described: X-degenerated, X-transposed and ampliconic sequences (Figure 9). The X-degenerated elements, which are scattered with the X-transposed portions, are dotted with single-copy genes or pseudogene homologues of 27 different X-linked genes, registering a 60-96% nucleotidic identity with those. These genes have been interpreted as surviving relics of ancient autosomes from where both sexual chromosomes co-evolved (Lahn and Page 1999). In half of the cases, the pseudogenes display sequence similarity to both exons and introns of the functional X homologue, while the remaining seem to be transcribed functional genes, that encode for very similar though not identical protein isoforms (Skaletsky *et al.* 2003). All the 12 ubiquitously expressed Y chromosome genes are located in the degenerated areas. On the other hand, only one of the genes expressed in the testis is X-degenerated (Skaletsky *et al.* 2003).

The X-Y transposition supposedly happen 3-4 mya, in a large-scale event after the human-chimpanzee divergence (Page *et al.* 1984, Mumm *et al.* 1997, Schwartz *et al.* 1998, Rozen *et al.* 2003). The NRY X-transposed segments have a 99% homology with DNA sequences in the Xq21 long arm. However, the chromosome was likely targeted with a subsequent inversion of the male-specific region in short-arm that cleaved the X-blocks into two non-contiguous segments (Figure 9; Mumm *et al.* 1997, Schwartz *et al.* 1998). The X-transposed sequences do not participate in the X-Y crossing-over during male meiosis, distinguishing them from the PARs at the telomeric regions of the human X and Y chromosomes. In their combined length of 3.4Mb, only two genes were identified and a high prevalence of repeat elements was detected (Skaletsky *et al.* 2003), stating for the low informational density of the transposed regions.

The seven ampliconic segments are dispersed across the Yq arm and the proximal part of Yp, in a combined length of 10.2Mb (Figure 9). The long repeat units, designated as amplicons, display a marked sequence similarity within and between segments in both arms of the chromosome (as much as 99.9%; Rozen *et al.* 2003, Skaletsky *et al.* 2003). This class harbours the highest density of NRY genes, comprising nine distinct coding families of predominant or exclusive expression in the testis. The most pronounced structural features of the ampliconic regions of Yq are the eight palindromes of large extension (cumulatively ~5.7Mb, one quarter of the Y chromosome male-specific region; see Figure 9), where eight of the multi-copy gene families have members and six gene families are exclusively located (Skaletsky *et al.* 2003). In addition, the amplicons include five sets of more widely spaced inverted repeats (referred as IR; Schwartz *et al.* 1998, Tilford *et al.* 2001) and a variety of long tandem arrays, namely the prominent NORF (no long open reading frame) and TSPY clusters. The first arrays owe their name to

a great diversity of spliced but apparently non-coding transcription units, while the latter encode for the TSPY protein (see <sup>Skaletsky *et al.* 2003</sup> for further details).

### 3.2 - Evolution of the Y chromosome

The sexual chromosomes, in mammals represented by the XY system, have as their putative ancestors a pair of autosomes (Figure 10; <sup>Ohno 1967, Bull 1983, Graves and Schmidt 1992</sup>) at an evolutionary timescale of 300 mya (<sup>Lahn and Page 1999, Lahn *et al.* 2001, Skaletsky *et al.* 2003</sup>). The latter studies on modern X-Y gene pairs have suggested those as surviving “fossils” where extensive sequence identity between ancestral X and Y chromosomes once existed. The differentiation of the two sex chromosomes supposedly started only when crossing-over between the chromosomes ceased, since there was a strong correlation between the age of individual X-Y gene pairs and the locations of their X members on the human X chromosome (<sup>Lahn and Page 1999</sup>). The first genes to diverge were probably the SRY (human sex-determining region Y) gene and its SOX3 homologue, which persists on the mammalian X-chromosome (<sup>Stevanovic *et al.* 1993, Foster and Graves 1994, Lahn and Page 1999</sup>).

Four evolutionary events are believed to have contributed to the human sex chromosome evolution, the first at about 300 mya and the last at 30 mya (<sup>Lahn and Page 1999, Lahn *et al.* 2001</sup>). The sequential suppression of recombination and consequent extension of the non-recombining portions are reflected in the mentioned X chromosome stepwise age increase along its length and the existence of four ‘evolutionary strata’ (<sup>Lahn and Page 1999</sup>). In the case of X-degenerated genes and pseudogenes, a single molecular process is rather likely: the region-by-region suppression of crossing-over, probably because of sequence inversions in the Y chromosome (<sup>Graves 1996, Lahn and Page 1999, Stefansson *et al.* 2005</sup>). In the absence of recombination a monotonic functional decline was triggered and therefore the few coding genes in these regions appear as resistant in the absence of sexual recombination (<sup>Lahn *et al.* 2001, Skaletsky *et al.* 2003</sup>).

On the other side, the ampliconic sequences arose from a handful of genomic sources and mechanisms (Figure 10; <sup>Skaletsky *et al.* 2003</sup>), although they represent almost identical copies in palindromes. Under a molecular reasoning, the near identity of the DAZ gene copies, residing exclusively on the arms of palindromes P1 and P2 may at first suggest that the gene amplification has occurred only within the last 200 ky (<sup>Agulnik *et al.* 1998</sup>). However, subsequent comparative sequence analysis of human and chimpanzee ampliconic sequences showed that the Y chromosome palindromes predate the speciation, with a series of autosomal transpositions and subsequent amplification having occurred during primate evolution (e.g. DAZ genes derive from DAZL autosomal transcription unit still present in chromosome 3; <sup>Saxena *et al.* 1996, Skaletsky *et al.* 2003</sup>). While the divergence between orthologous palindromes is owed to the accumulation of neutral mutations in separated

lineages of humans and chimpanzees (about 1.44%, <sup>Rozen et al. 2003</sup>), the little intraspecific arm-to-arm divergence (0.021-0.028%, <sup>Rozen et al. 2003</sup>) suggests that the paired arms of the palindromes evolved in concert. In fact, the nearly identical copies with remarkably uniform patterns of tissue expression might have gene conversion as the underlying mechanism: the Y chromosome itself repairs the mutations by non-reciprocal transfer of information between similar gene pairs in the male-specific portion (<sup>Rozen et al. 2003, Bosch et al. 2004</sup>). The sequences displaying intrachromosomal identities of >99.9% represent a large and distinct subset of NRY euchromatin, that comprises the eight palindromes as well as large portions of the IR2 and IR3 inverted repeats, and where the gene conversion is supposedly engaged

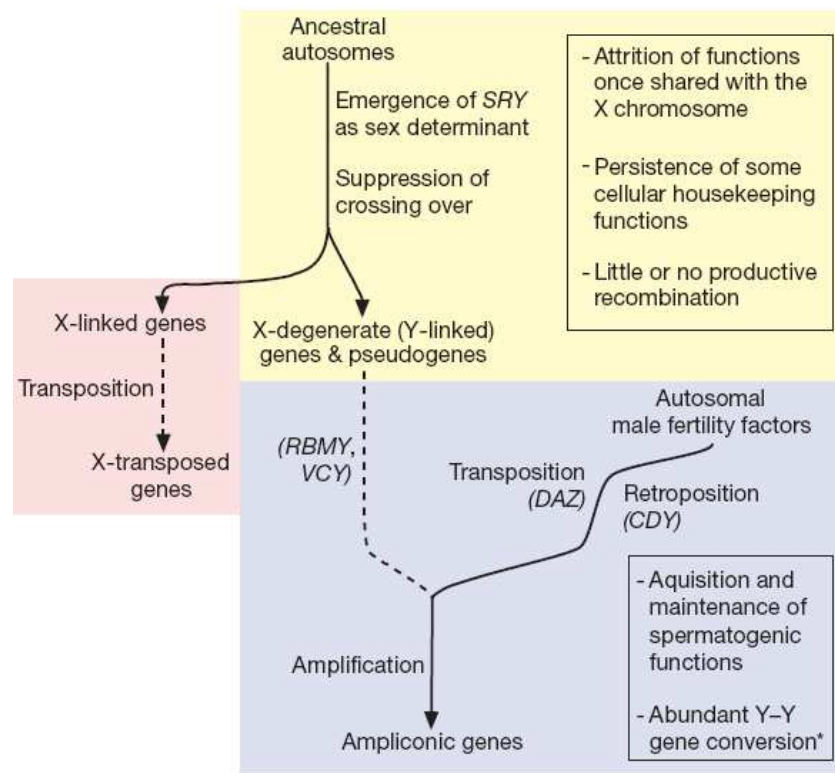


Figure 10 – Molecular evolutionary pathways and processes that gave rise to the genes in the three NRY sequence classes: X-degenerated genes (yellow) derive from an autosomal pair ancestral to both X and Y chromosomes (and enlarged by subsequent fusion with other autosomes or autosomal segments, <sup>Watson et al. 1991</sup>); X-transposed genes (in pink) derive from X-linked genes, by turn derived from the ancestral autosomal pair; Ampliconic genes (blue) originate from three converging processes, namely amplification of X-degenerated genes (e.g. RBMY, VCY), transposition and amplification of autosomal genes (DAZ) and retroposition and amplification of autosomal genes (CDY). The boxes enumerate dominant themes in gene evolution. The asterisk denotes that Y-Y gene conversion is apparently common in ampliconic sequences that exhibit intrachromosomal identities higher than 99.9% (<sup>Skaletsky et al. 2003, Rozen et al. 2003</sup>). In <sup>Skaletsky et al. (2003)</sup>.

in a routine basis (Rozen *et al.* 2003, Skaletsky *et al.* 2003). Additional evidence for the process is detected in particular SNPs of the recent genealogy of human NRY. As for the X-transposed elements these seem to be of more recent acquisition, at approximately 3-4 mya (Rozen *et al.* 2003, Skaletsky *et al.* 2003).

Very many large-scale structural rearrangements in non-ampliconic portions of the human Y chromosome have accumulated since the humans and chimpanzees diverged (Hughes *et al.* 2005). However, none of these differences is polymorphic among extant human Y chromosomes (Repping *et al.* 2006). On the other hand, there is little similarity between the ampliconic structure of the human and chimpanzee Y chromosomes (Repping *et al.* 2006), except for the conserved P6, P7, P8 palindromes and the centres of palindromes P1 and P2 (Rozen *et al.* 2003, Hughes *et al.* 2005). The examination of structural variation across a worldwide genealogical tree (Underhill *et al.* 2000, YCC 2002) better explained the mutational dynamics underlying the structural polymorphisms and showed that the high mutation rate of large-scale mutations in palindromic regions ( $2.3-4.4 \times 10^{-4}$  mutations per father-to-son Y transmission) is the main force driving structural polymorphism among human Y chromosomes. The limited variation in Y-linked genes raises the possibility of selective constraints (Repping *et al.* 2006).

### 3.3 – Distinctive features of Y chromosome

#### 3.3.1 - Haploidy and paternal inheritance

In healthy men, the sex-defining Y chromosome appears in a single copy. The major structural rearrangements on its evolutionary history prevented pairing and recombination with the X chromosome in most of its length, and allowed the male-specific region to extend along the Y chromosome (Lahn and Page 1999, Lahn *et al.* 2001, reviewed in Skaletsky *et al.* 2003). The heritage process is by that restricted to men, characterizing paternal lineages where the descendants inherit the mutations. The single-locus behaviour is only altered by the accumulation of mutations over time and thus keeps a direct historical record, in contrast to autosomes, where biparental inheritance and recombination make it virtually impossible to reconstruct precise genealogical lineages of molecular descent. In a somewhat simplified way, one may assume that with a sex ratio of 1:1, the effective population size of Y chromosome in a population is one-quarter of the autosomes, a third of the X-chromosomes and equivalent to that of mtDNA.

### 3.3.2 - Absence of recombination on the NRY

As mentioned, the NRY is passed intact over to the next male generation since it is assumed to escape close pairing and crossing-over with the X-chromosome in the meiotic process. The genetic markers are thus in perfect linkage originating haplotypes, this is particular combination of allelic states. There are nevertheless situations under which recombination could occur, as episodes of gene conversion between paralogues (Rozen *et al.* 2003). However, these are not considered under the conventional definition of recombination by crossing over. In the pathological situation (47, XYY) that incises 1/1000 men (Walzer and Gerald 1975) recombination is possible but since the copies are identical, there are no major consequences in the molecular sequence. In reality such males are able to eliminate one Y from the germline (Chevret *et al.* 1997). Y-like segments have been detected in autosomes, possible asymptomatic translocations that may recombine (Cooke and Noel 1979, Andersson *et al.* 1988). Fortunately such cases are recognized by a robust phylogeny.

The non-recombining elements tend to decay rapidly because there is no obvious mechanism to regenerate or stop the accumulation of slightly deleterious mutations (caused by selective forces as “Muller’s ratchet”, and “hitchhiking” when linked to selectively favoured alleles; Muller 1964, Rice 1987). Once a region has become isolated by haploidy, it can no longer be repaired by piecing together unchanged parts of the homologous chromosome and is at the mercy of genetic drift. Provided that the X and the Y chromosomes were identical some 300 mya, one may assume that the Y chromosome has lost 1393 of the supposed original 1438 genes, retaining about a half of a gene *per* Mb of its genome, compared to *ca.* 10 genes *per* Mb in the X chromosome (Graves 2004). However, the frequent occurrence of amplification within the chromosome it is likely to compensate for the inevitable decay, built into haploid genetic systems. The arrangement of nearly all genes essential for spermatogenesis (such as DAZ genes) in multiple copies of paralogous repeats has supposedly evolved to protect against harmful mutations, or at least retard the erosion, through beneficial Y-Y gene conversion (Rozen *et al.* 2003, Bosch *et al.* 2004). It might even be that the lack of crossing-over with a homologue makes such intra-chromosomal events more frequent, resulting in a degree of identity among sequence pairs that rivals that of autosomal homologues chosen at random from the human populations (Jobling and Tyler-Smith 2003). Assuming a steady-state balance between new mutations, that create differences between arms, and gene conversion episodes that erase the differences, it is possible to estimate a rate of  $2.2 \times 10^{-4}$  conversions per duplicated locus per 20-year human generation (Rozen *et al.* 2003). Graves (2004) has stated the existence of palindromic structures, able to make internal loops within the structure of a single Y chromosome, where gene conversion compensates the lack of recombination. The process can however resurrect inactive copies.



### 3.3.3 – The role of selection in the Y chromosome

Selection acts as a force shaping Y haplotypes' diversity, in particular if the coding region is affected. The Y chromosome is then subject to purifying selection when, for example, the inactivation or loss of Y genes produces XY females and hermaphrodites or male infertility (e.g. <sup>Sun *et al.* 1999</sup>). When referring to this haploid system makes no sense to consider balancing selection, and frequency-dependent selection has not been testified. The concern centres rather on the potential influence of positive selection, with advantageous changes becoming fixed in the population. Because of the lack of recombination any selection will affect the entire chromosome and produce an increment on frequency of a lineage more rapidly than would be expected by drift. One has to regard however past and present differential selection of Y lineages, with neutral variants becoming advantageous or disadvantageous within the timescale of evolution. The haplotypes used for evolutionary purposes were subject to association studies with those phenotypes suspiciously under selection (<sup>Jobling *et al.* 1998, Paracchini *et al.* 2000, Passarino *et al.* 2001, Quintana-Murci *et al.* 2001, McElreavey and Quintana-Murci 2003</sup>). Many of the studies showed no association or when found, often could not be reproduced (e.g. <sup>Kuroki *et al.* 1999; Carvalho *et al.* 2003, 2004</sup>), or could be explained by population structure (<sup>Previdere *et al.* 1999</sup>), strengthening the idea about the neutrality of the evolutionary markers employed. A few associations seem robust but can be explained by plausible mechanisms, not selective constraints. As an example, an inversion polymorphism in haplogroup P prevents the ectopic recombination between genes that produce many XX males (<sup>Jobling *et al.* 1998</sup>), resulting on a lower frequency of these males phylogenetically assigned to such haplogroup.

### 3.4 - Calibration of the Y chromosome molecular clock

Because of the very slow molecular evolution of the Y chromosome (identical to that of autosomes, <sup>Nachman and Crowell 2000</sup>), the direct use of base substitutional mutations for introducing a temporal scale into the phylogenetic reconstructions, has been and still is largely impossible for a reason that AMH, as a species, is relatively young. Therefore, the methods widely and successfully used in interspecies studies are far beyond reach in calibrating the events within the human Y-chromosomal phylogeny. Furthermore, the only reliable outgroup for such calibration – the chimpanzee – separated from humans about 6 mya (<sup>Goodman *et al.* 1998</sup>), a time interval tens of times larger than that of the existence of AMH, and which makes the calibration prone to large errors. On the other hand, the large size of the human Y chromosome permits, in theory, that enough mutations accumulate to construct a precise phylogenetic tree. Recent and future progress, principally of DNA sequencing techniques, may open new experimental approaches that would allow employing direct calibration of the molecular clock, by counting the accumulated single nucleotide changes in the Y

chromosome since the MRCA, making use of outgroups (great apes) or even demographic events that are known from the archaeological record. However, to the best of my knowledge there is so far no published paper in the Y chromosome literature, where the topology of the phylogenetic tree of it has been supplied with time scale, provided by SNP-based temporal calibration of the tree (see forthcoming review of Underhill and Kivisild <sup>(2007)</sup>).

Instead, temporal estimates are obtained using an entirely different genetic system that of the very rapidly evolving STRs. Unlike SNPs, which can be treated as unique events in the genealogy of such a young species as humans are, the accumulation of variation in STRs length is not unique, often taking place as parallel events in any of the sub-clusters of the Y chromosome phylogeny. The STR-based polymorphisms have been detected already long ago in autosomal chromosomes and much of their basic characteristics, such as molecular mechanisms that create length change (gain or loss of repeat units by replication slippage), have been postulated using autosomal STRs as models (e.g. Kornberg *et al.* 1964, Levinson and Gutman 1987, McMurray 1995, Chakraborty *et al.* 1997, Kruglyak *et al.* 1998, Huang *et al.* 2002, Lai and Sun 2003; and other references in the review of Nikitina and Nazarenko 2004). Such mechanisms, as well as their frequencies (i.e. the rates of length changes) are apparently identical in autosomes and sex chromosomes, including the Y chromosome (e.g. Zhivotovsky *et al.* 1997, 2004; Brinkmann *et al.* 1998; Xu *et al.* 2000; International Human Genome Sequencing Consortium 2001; Subramanian *et al.* 2003).

The calibration of the Y chromosome molecular clock then requires a reliable estimate of the evolutionary mutation rate of Y-STR loci. Different approaches and estimations have however been reported: deep rooting pedigrees ( $2.0 \times 10^{-3}$  mutation/generation, Heyer *et al.* 1997); father/son pair analysis ( $2.6 \times 10^{-4}$  or  $2.8 \times 10^{-3}$  mutation/generation; Forster *et al.* 2000 and Kayser *et al.* 2000b, respectively); sperm analysis ( $2 \times 10^{-3}$  mutation/generation, Holtkemper *et al.* 2001); evolutionary studies ( $0.7 \times 10^{-3}$ , Pritchard *et al.* 1999,  $0.69 \times 10^{-3}$ , Zhivotovsky *et al.* 2004). It soon become obvious that the clock rates offered by familial and pedigree studies differ rather profoundly from those calculated from evolutionary studies. In this respect the picture is analogous to that found for mtDNA, as it has been described before. A general unsatisfaction started to insurge, especially when a panoply of factors other than the scale of evolution could justify the deviation of estimates. The high mutation rates might arise from asymmetrical sister-chromatid exchange, or replication slippage facilitated by the secondary structure of the repeats. Alternatively, the lower estimates (Forster *et al.* 2000) can be achieved if only one-step mutation rate is considered, the "fast" microsatellite markers are neglected or a population with more ancient split is analysed (Zhivotovsky *et al.* 2004). In addition, the rarity of mutations leading to the large standard errors of father-son comparisons and the confusing factor of non-paternity in deep-rooting pedigrees, further contribute to the ambiguities of the estimators.

Zhivotovsky *et al.* <sup>(2004)</sup> proposed a model of microsatellite evolution where multistate STR mutations are weighted in the effective mutation rate (Slatkin 1995, Zhivotovsky and Feldman 1995). The approach analysed the variation of ten tri- and tetra-STRs on Y chromosome and autosomes for the New

Zealand Maori and Bulgarian Gypsies, populations whose known divergence time was established by archaeological and historical data (Marushiakova and Popov 1997, Diamond and Bellwood 2003). Since the underlying mutational mechanism seems to be the same (Mountain *et al.* 2002), the averaging of both Y chromosome and autosomal STRs mutation rates supposedly decreases the random effects of sampling, and therefore an average estimate of  $0.69 \pm 0.13 \times 10^{-3}$  mutations/25 years should represent an appropriate basis for dating populational events.

Further uncertainty is introduced if to consider that the microsatellite diversity and mutation rate tend to be locus-specific (Kayser *et al.* 2000b) and vary also within the haplogroups (Carvalho-Silva *et al.* 1999, Dupuy *et al.* 2004), therefore affecting selectively the coalescence estimates. One can not exclude that mutation rates at a certain locus may vary among haplogroups due to differences in allele repeat scores, or that these are even population-specific (Zhivotovsky *et al.* 2004). However, the data of Dupuy *et al.* (2004) did not provide details on the way which STR-mutation rate increases as function of repeat score. A simulation by removing the loci under the “large allele-size argument” did not influenced significantly the previously estimated  $0.69 \times 10^{-3}$  (Zhivotovsky and Underhill 2005).

A more accurate estimate should nevertheless combine the dynamics of haplogroups and the evolution of their microsatellite variation, such as the rapid extinction of newly arisen microsatellite alleles (Zhivotovsky *et al.* 2006). By definition, when a SNP mutation defines a new haplogroup, there is zero STR variation in it, but over time, different STR haplotypes accumulate and radiate from the central ancestor (Jobling and Tyler-Smith 2003, Figure 11). Meanwhile, the accumulated STR diversity is being continuously removed by genetic drift (mostly bottlenecks) during stochastic fluctuations in haplogroup's frequency over generations. A mathematical modelling under these evolutionary assumptions resulted in that the rate of accumulation of microsatellite variation is about 3.6 times lower than those predicted from the germline mutation rate, and thus reflects the removal process

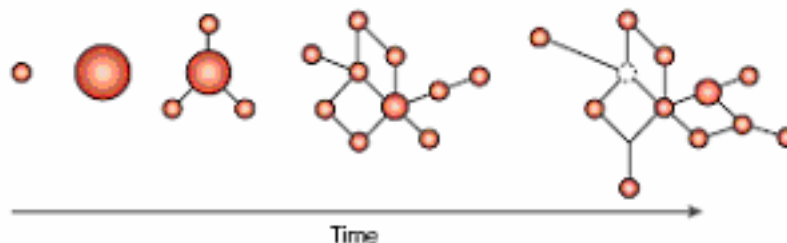


Figure 11 – Network of microsatellite variation accumulation with time from a single ancestor. In Jobling and Tyler-Smith (2003).

(Zhivotovsky *et al.* 2006). In that sense, each haplogroup has its own demographic history but that is not discernible from the current information. Several parameters can act to produce the inter-haplogroup differences, namely fluctuations in the effective population size, where fastly expanding populations tend to show a higher rate of variance increase. The average estimates do not reflect the true “state-of-the-affairs” and are simplistic tools for the studies (Pakendorf and Stoneking 2005). Nevertheless, Zhivotovsky *et al.* (2006) concluded that germline mutation rate can be used for probability calculation in forensic and disease studies, whereas the evolutionary effective mutation rate are still the more appropriate for evolutionary studies.

### 3.5 - Phylogeny and nomenclature of Y chromosome haplogroups

The advances in mutation-detection technology and consequent increment of knowledge ended the established idea of low level of polymorphism in the Y chromosome (review in Jobling and Tyler-Smith 2003). To the dozen of molecular markers initially known (Jobling *et al.* 1997) other tens were added (Underhill *et al.* 1997) by the analysis of populations from different regions on the globe. More than 300 well-characterized SNPs are currently described for their use in Y-phylogenetics (Shen *et al.* 2000, Underhill *et al.* 2000, Hammer *et al.* 2001, YCC 2002, Jobling and Tyler-Smith 2003, Underhill 2003). Y-SNPs are considered to be of single evolutionary occurrence in the human history (Unique Event Polymorphisms, UEPs) and their low mutation rate ( $\sim 2 \times 10^{-8}$  per base per generation, Nachman and Crowell 2000) makes them a preferential choice for constructing unique basal phylogenies. The robust and developing phylogeny of Y chromosome is hierarchical and accepted to be unique, where the cumulative occurrence of SNPs defines the position of the haplogroups in the tree. To the level of resolution achieved in 2003 only five homoplasies were described (YCC 2002, Jobling and Tyler-Smith 2003) but these are easily identified by the disposition of adjacent markers. Indel length variants that do not result in a medical condition (AZF and DAZ3/4 loci; Vogt 1998, 2005; Fernandes *et al.* 2002b, 2004; Repping *et al.* 2003, 2004) can also become frequent enough to be considered polymorphisms, as the 2kb deletion in marker 12f2 characterizing haplogroup J (Casanova *et al.* 1985). Newly found markers are continuously enhancing phylogenetic resolution of the tree.

The mutation rate of Y-STRs is much higher than that of biallelic markers ( $\sim 0.7 \times 10^{-3}$  mutations/generation, see references in section 3.4), making them useful to determine the intrahaplogroup diversity and to attribute molecular ages (de Knijff 2000). The Y chromosome multi-allelic markers, first used by Litt and Luty (1989), produce an haploid profile of male DNA and therefore are also informative for gene mapping, human identification in forensics (Jobling *et al.* 1997, Jones and Ardren 2003) and in paternity and kinship studies (Kimpton *et al.* 1993, Hammond *et al.* 1994, Kayser *et al.* 2004).

A bit less than a decade ago, several laboratories started to study, in parallel, many different populations from diverse regions and ethnic affiliations. As an outcome, a number of different

nomenclature systems were created and published in the literature (e.g. <sup>Su *et al.* 1999, Jobling and Tyler-Smith 2000, Semino *et al.* 2000, Underhill *et al.* 2000, Hammer *et al.* 2001, Karafet *et al.* 2001</sup>). At the same time, ambiguities remained in the phylogenetic position of many markers and the consequent haplogroup assignment. The need for a consensus nomenclature emerged, and after a couple of years of close attention in the subject, such a nomenclature was finally developed by the Y Chromosome Consortium (YCC) in 2002 <sup>(YCC 2002)</sup>. In analogy to what was done for mtDNA trees, homologous NRY sequences of gorillas, chimpanzees and orangutans provided an outgroup for rooting of the tree and the emerging hierarchical tree allowed to determine the likely ancestral state <sup>(Underhill *et al.* 2000, Hammer *et al.* 2001)</sup>. The term “haplogroup”, applied to the Y-chromosomal phylogenetic tree, refers to NRY lineage defined by one or more SNPs.

Presented in Figure 12 is the phylogenetic tree of the human Y chromosome based on the state-of-art knowledge in 2003 (i.e. YCC2003 tree in Jobling and Tyler-Smith 2003), though it does not include all the details known to that moment <sup>(Underhill and Kivisild 2007)</sup>. The phylogeny is split into major haplogroups from A to R with a total of 245 markers distributed among 153 branching subclusters. The refined typing adds numbers and small letters sequentially (e.g. A1b). The potentially paraphyletic lineages, interior nodes on the tree not included in the sub-clades of a clade, were named paragroups and assigned a “\*” symbol by the YCC <sup>(2002)</sup>, R1\* for instance. A designation such as R(xR1a) indicates the partial typing of markers, in this case excludes those belonging to R1a. In this standardized nomenclature system most markers are designated with prefix (identifying laboratory, where the marker was first found), and quite often in the text the lineages are mentioned by haplogroup plus terminal mutation – e.g. R1-M173. The use of a mutation-based nomenclature, referring to the last defining marker is used for the sake of clarity in the comparisons. In terms of nomenclature the YCC proposal introduced many advantages: i) placed the haplogroups in an hierarchical order; ii) standardized haplogroup names and, therefore, allowed to discard of a large number of different, partially incompatible nomenclatures and haplogroup names ; iii) made it easier to include new markers when found. A year later Jobling and Tyler-Smith <sup>(2003)</sup> published corrections and minor changes to the original consensus. Nevertheless, a certain fraction of earlier published low resolution data became obsolete because of an inevitable ambiguity in their assignment into phylogenetically deeper topology, unless further refinement has been carried out. Comparison of YCC 2002 and YCC 2003 topologies reveals a few differences, one of them with specific value to foresee the future changes. Namely, in YCC 2002 haplogroups N-LLY22g and O-M175 are depicted as independent basal clades, deriving from the central node defined by M9, while in YCC 2003 the novel M214 SNP was introduced, revealing that haplogroups O and N are sister clades within their ancestral clade NO. The topology of the Y chromosome tree (Figure 12) allows to see that, for example, the node defined by M89 (at our present knowledge = M213; =P14) gives rise to independent clades G-M201, H-M52(=M69); I-M170 (=P19); J-12f2.1 and to paragroup F\* and the branch defined by M9. Therefore, our current phylogenetic knowledge assumes multifurcation at the internal node defined by M89.

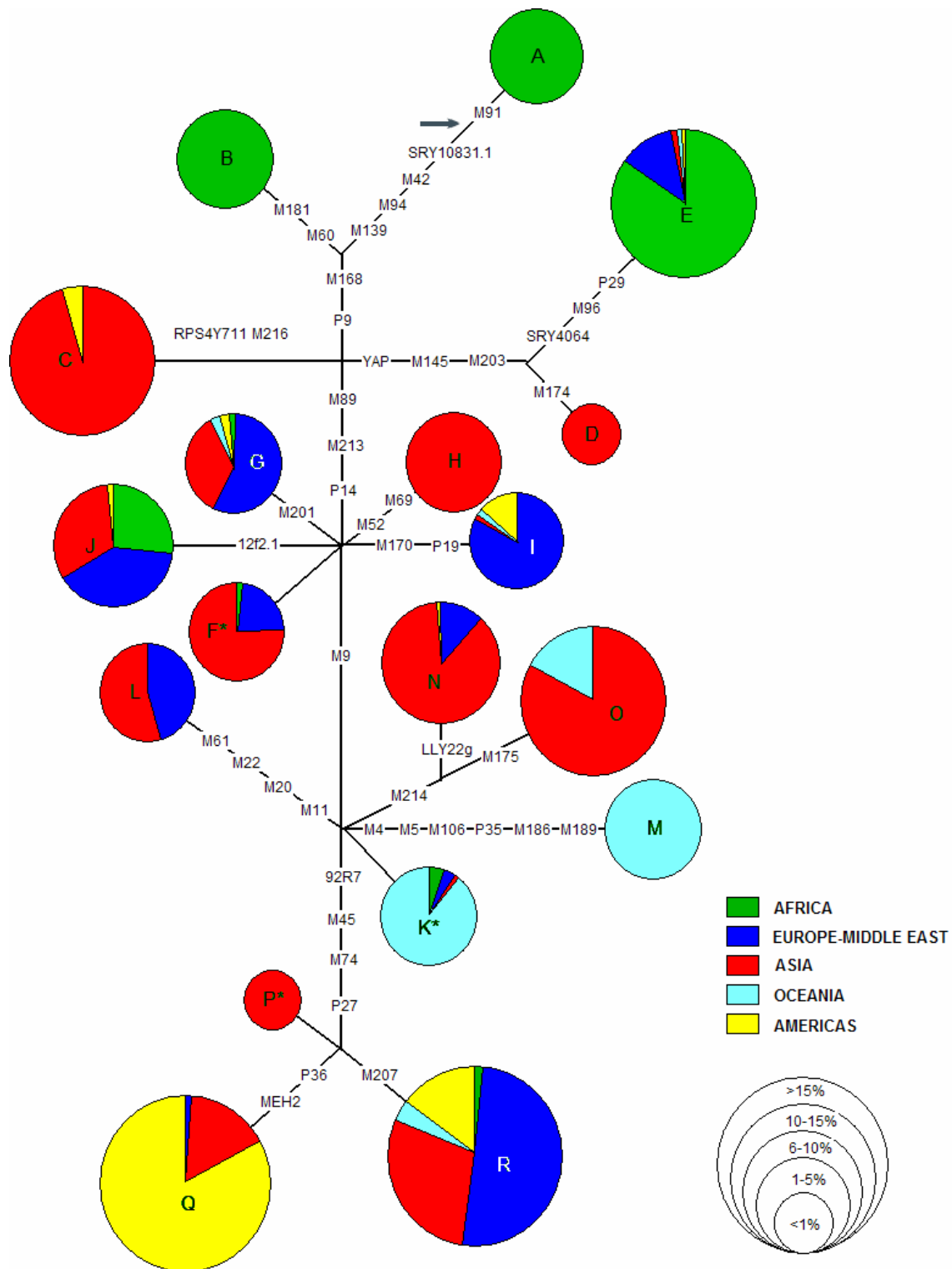


Figure 12 – Evolutionary tree of the major 18 haplogroups/paragroups and their continental distribution. The phylogenetic root is denoted by an arrow. The mutation events are label by the marker name. The size of the pie-charts represents their overall proportion, correspondent to the displayed frequency classes. The colored portions refer to the frequency in the geographic regions. Based on the total frequency of worldwide Y chromosomes in Hammer and Zegura<sup>(2002)</sup>, and adapted with information from the YCC<sup>(2002)</sup> and Jobling & Tyler-Smith<sup>(2003)</sup>.

However, it is quite likely that this multifurcation hides so far unknown internal bifurcations. Indeed, one such has been recently discovered, bringing together haplogroups I and J (Underhill and Kivisild 2007), and one may anticipate that in future, richer internal structuring of the Y chromosome tree will be available. From the phylogeographic point of view, such improvements are valuable in the reconstruction of ancient migration patterns. Indeed, the joint NO clade established by M214 marker allowed to suggest a major counter-clockwise spread of haplogroup N from East Asia to North Europe, based on phylogeography of rare NO\* chromosomes (Rootsi et al. 2007). However, it would be incorrect to insist that the current topology of the Y chromosomal variation is erroneous: it is very likely correct, but it has a considerable potential to be further refined.

### 3.6 - The origin and worldwide dispersal of Y chromosomes

The global phylogeographic analysis of Y chromosomes shows their high geographic specificity and suggests an African, presumably East African, origin for the modern humans (e.g. Underhill et al. 2000, 2001a; Hammer et al. 2001; Ke et al. 2001). Under a model of constant population size, the coalescence time estimates to the oldest root in the phylogeny are considerably older than posterior estimates assuming exponential growth ( $147 \pm 51$  kya Hammer et al. 1998 and  $90.4 \pm 20.1$  ky Hammer and Zegura 2002 against 59 kya, 95% confidence interval 40-140 kya, Thomson et al. 2000 or STR-based 46-91 kya, 95% confidence interval 16-126 kya, Pritchard et al. 1999).

It should be stressed that there is no theoretical need to expect that coalescence ages for maternally inherited mtDNA and paternally inherited Y chromosome should coincide. Nor should one expect that the coalescence ages of these two sex-related genetic systems should be even close to the palaeontologically estimated emergence of AMH, based on fossil evidence. Note that the assumed split between the ancestor of modern humans and *Neanderthals* and the emergence of AMH (e.g. Lahr and Foley 2004) date much later than the coalescence of diversity present in autosomal chromosomes - about a half a million – million years ago (Fullerton et al. 1997).

Haplogroups A-M91 and B-M60 are the two most profound clades of the Y chromosome phylogeny, with the root of the tree placed somewhere between them. Their present distribution is almost exclusive of sub-Saharan Africans, at low or moderate frequencies of about 6-7% (e.g. Underhill et al. 2000, Semino et al. 2002, Wood et al. 2005, Rosa et al. 2007). However, in the “relic” hunter-gatherers !Kung and Khwe (Scozzari et al. 1997, 1999; Knight et al. 2003; Wood et al. 2005) and Central African Pygmies (Underhill et al. 2000, Wood et al. 2005) these lineages comprise more than a half of the paternal genetic pool, what has been interpreted as the survival of the ancestral pool of modern humans (Underhill et al. 2001a). The phylogenetic position and accumulated variation of A-M91 and B-M60 lineages, when combined with the anthropological and archaeological data (Nurse et al. 1985), are suggestive of an early diversification and dispersal of human

populations within Africa, possibly main hitchhikers of the initial pan-continental dispersal(s) (Figure 13). However, one needs to treat with certain caution the 5 years ago suggested coalescence time estimates of roughly 40 ky (Hammer and Zegura 2002), not the least because unpublished data of some laboratories suggest by far deeper split between haplogroups A and B (Underhill and Kivisild 2007).

Since the extant phylogenetic tree of NRY lineages and their phylogeographic distribution are assumed to shed light over the past evolutionary events, both those that have reshaped the initial phylogeographic pattern and those that occurred over a long time span, and are independent from historic and archaeological evidence, the Y chromosome research can assist in interpreting concurrent hypothesis on events affecting human variation (e.g. Quintana-Murci *et al.* 1999; Underhill *et al.* 2000, 2001a, 2003; Hammer *et al.* 2001; Cavalli-Sforza and Feldman 2003). The worldwide dispersal of modern Y chromosomes has been interpreted in the light of the “Out-of-Africa” (Cann *et al.* 1987, Stringer and Andrews 1988, Stringer 2003) and “multiregional” models (Excoffier and Langaney 1989, Wolpoff 1989, Templeton 1997). The father-son inherited portions elected the “Out-of-Africa” hypothesis, where the modern extant Y chromosomes trace their ancestry to a limited number of African forefathers who successfully left Africa relatively recently, and eventually replaced archaic lineages elsewhere in the world. The theory found support in the highest level of mean pairwise differences among haplotypes of African populations, while the remaining variation has arisen from multiple founder effects and subsequent episodes of bottlenecks and expansions (Calafell *et al.* 1998, Hammer *et al.* 2001, Underhill *et al.* 2001a, Yu *et al.* 2002). The central node of the “Out-of-Africa” expansion is the marker M168 (=P9), basal to all clades found outside Africa (Figure 13). Lineages derived from M168 chromosomes are frequent in Africa as well (for further discussion see chapter 3.6.1). Different ages are pointed for the last common ancestor of all non-African Y chromosomes (of about 45 ky, Thomson *et al.* 2000, Underhill *et al.* 2000 or 69 ky Hammer and Zegura 2002), with the most recent estimate suggesting its split from haplogroup B at about 80 kya, with further rapid diversification at around 60 kya (Underhill and Kivisild 2007).

For quite some time the debate has centred over the routes of dispersal of modern humans, in the long-lasting peopling process of Eurasia and other continents, and their corresponding chronology. As for the mtDNA genetic system, the distribution of Y-chromosomal lineages was interpreted considering: i) an early southern migration, perhaps following a coastal route around the northern edge of the Indian Ocean over the Horn of Africa before 50 kya (Lahr and Foley 1994; Stringer 2000, 2003; Bowler *et al.* 2003; Leavesley and Chappell 2004; Mellars 2006), ii) a later northern migration into Eurasia over Sinai via the Levantine corridor (Bar-Yosef *et al.* 1986, 2002). Despite many subsequent demographic changes, distinct remnants of early Y chromosome carriers and their phylogeography seem to favour the southern route, although more in-depth analyses are needed. A diverse set of basal C-RPS4Y, F-M89 and K-M9 founder lineages, that have likely arisen on the way to Asia after an earlier departure from Africa of their ancestral chromosomes, were shown to congregate along the southern Asian corridor of migration and to have further diverge towards Oceania and the Americas (Figure 13; Karafet *et al.* 1999, 2001, 2002; Kayser *et al.* 2000a, 2003; Underhill *et al.* 2000, 2001b; Bamshad *et al.* 2001; Capelli *et al.* 2001; Hammer *et al.* 2001, 2006; Wells *et al.* 2001; Lell *et al.* 2002;



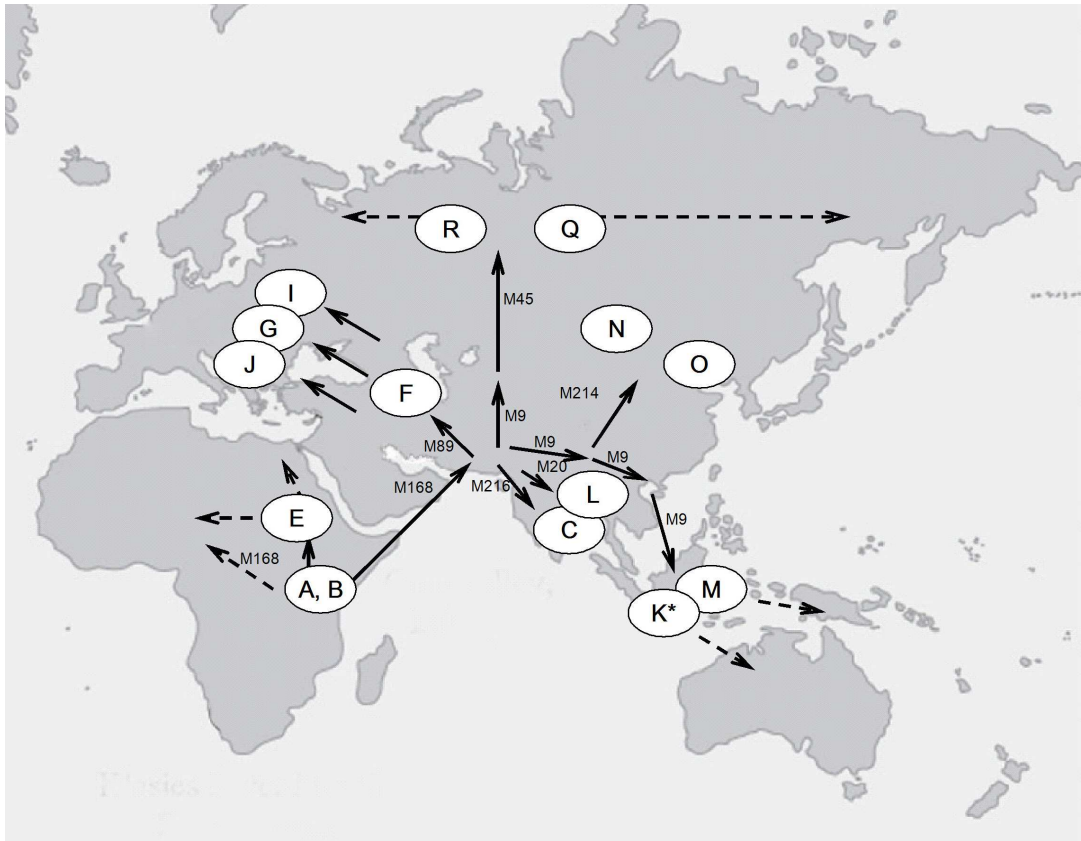


Figure 13 – Schematic reconstruction of the origin and worldwide dispersal of Y chromosome superhaplogroup F-M89, with subsequent diversification of M9 lineages. The full line arrows do not indicate precise migration routes but the direction of the movement. Dashed arrows indicate subsequent movements of the clades. Adapted from Underhill<sup>(2003)</sup> and Rootsi<sup>(2004)</sup>.

Redd *et al.* 2002; Kivisild *et al.* 2003; Zegura *et al.* 2004; Scheinfeldt *et al.* 2006). For instance, the C-RPS4Y clade, considered among the oldest lineage in Asia and the Southwest Pacific was likely introduced with the first settlers perhaps as early as 50 ky<sup>(Underhill 2003)</sup>. On the other hand, the package of Y chromosome founder lineages in West Eurasia is reduced to F-M89 and K-M9<sup>(Rosser *et al.* 2000, Semino *et al.* 2000, Underhill *et al.* 2000, Hammer *et al.* 2001, Wells *et al.* 2001, Hammer and Zegura 2002, Kivisild *et al.* 2003)</sup>, supporting the idea that the “Out-of-Africa” migration first reached Southwest Asia and from there dispersed both east and westwards, consistent with the single coastal route scenario. Further, a number of deep-rooting subclusters like C\*-RPS4Y, F\*-M89, H-M69, (within F), L-M20 (within K) and R2 are frequent and largely restricted to the Indian subcontinent<sup>(Bamshad *et al.* 2001, Kivisild *et al.* 2003)</sup>, a region that seems to have played a pivotal role in late Pleistocene genetic differentiation of the western and eastern Eurasian gene pools.

After the initial settlement of Eurasia, its genetic composition, as far as the Y chromosome is concerned, did not remain constant. Small groups of modern humans, holders of the founders C, F and K clades, started to split into several isolated groups and developed region-specific variants

(reviewed in Underhill 2003). Haplogroup C, with a patchy distribution, is represented in the Indian subcontinent only by the C\* lineages (Bamshad *et al.* 2001, Kivisild *et al.* 2003) while in southeast Asia, C\* chromosomes are in combination with C3-M217 (actually the only C lineage in inland Asia and Native Americans; Karafet *et al.* 1999, 2001, 2002; Underhill *et al.* 2000; Hammer *et al.* 2001, 2006; Wells *et al.* 2001; Hammer and Zegura 2002; Lell *et al.* 2002; Kayser *et al.* 2003; Zegura *et al.* 2004) and in East Indonesia, Melanesia and Polynesia are together with the native C2-M38 (Underhill *et al.* 2000; Kayser *et al.* 2000a, 2003; Hammer *et al.* 2001, 2006; Redd *et al.* 2002; Scheinfeldt *et al.* 2006).

Superhaplogroup F, characterized by M89 derived state, comprises the subsequent diversity of the phylogenetic tree, with many region-specific derivatives (e.g. the Near and Middle Eastern haplogroups J-12f2.1 and G-M201 and the European haplogroup I-M170; Figure 13). The superhaplogroup K-M9 harbors numerous branches that have followed different trajectories and gave rise to more specific subdivisions: haplogroup L-M20 in Southwest Asia and K\*-M9, K5-M230 and M-M4 types in Oceania and New Guinea (Underhill *et al.* 2000; Kayser *et al.* 2000a, 2003; Capelli *et al.* 2001; Hammer and Zegura 2002; Jobling and Tyler-Smith 2003; Hammer *et al.* 2006; Scheinfeldt *et al.* 2006); O-M175 lineages are of ubiquitous distribution from southeastern and eastern Asia up to Siberia (Underhill *et al.* 2000, 2003; Capelli *et al.* 2001; Karafet *et al.* 2001, 2002, 2005; Kayser *et al.* 2003; Hammer *et al.* 2006; Rootsi *et al.* 2007) but rather infrequent in Polynesia and Melanesia; the Central Asian haplogroups P-92R7 and Q-PN36 are typical of Siberians, with Q\* and Q3 lineages particularly found in Amerinds (Karafet *et al.* 1999, 2002; Underhill *et al.* 2000, 2001a; Wells *et al.* 2001; Hammer and Zegura 2002, Lell *et al.* 2002, Zegura *et al.* 2004, Hammer *et al.* 2006); R1a is in general common in Central and South Asia (Karafet *et al.* 1999, 2001, 2002; Underhill *et al.* 2000; Hammer *et al.* 2001, 2006; Wells *et al.* 2001; Kayser *et al.* 2003, Kivisild *et al.* 2003) but can also represent nearly half of the pool of some East European populations (Karafet *et al.* 1999, Rosser *et al.* 2000, Semino *et al.* 2000). The K\*-M9 is among the most common macrohaplogroups in Australians and Melasians (e.g. Kayser *et al.* 2003, Hammer *et al.* 2006), comprising in the later a considerable heterozygosity of local clades, detected only recently when adding new binary markers (K6-P79, K7-P117, M2-P87 and M2a-P22; Scheinfeldt *et al.* 2006). Actually, all the native Near Oceanic haplogroups seem to have developed *in situ* at about 30-45 kya, at a comparable date to the predicted for ancient mtDNA expansions (Friedlaender *et al.* 2005, Merriwether *et al.* 2005) and the earliest settlements in the region at approximately 40 – 50 kya (e.g. Leavesley and Chappell 2004).

The haplogroup D lineages, that have accumulated M174 mutation in a YAP+ background, are nowadays confined predominantly to Japanese and, to lesser extent, Tibetans (Hammer and Horai 1995; Karafet *et al.* 1999, 2001; Underhill *et al.* 2000; Wells *et al.* 2001; Hammer *et al.* 2006). Curiously, their presence in Andaman islanders suggest that their earliest (perhaps Palaeolithic) phylogeography in Asia might have been considerably wider than it is at the present and state for a founder effect that has been lost (Su *et al.* 1999, 2000; Tajima *et al.* 2004; Wen *et al.* 2004). The newly identified P47 mutation (Hammer *et al.* 2006) establishes a fourth Asian D lineage that marks most chromosomes that were previously ancestral D\*-M174 from Central Asia (e.g. Karafet *et al.* 2001).

Together with genetic drift that has operated over the ancestral variation, several later episodes of gene flows further shaped the Y chromosome diversity. For example, haplogroup J-12f2.1

in India (Kivisild *et al.* 2003) represents a more recent arrival from Near/Middle East chromosomes (Nebel *et al.* 2001, Underhill *et al.* 2001a, Semino *et al.* 2004). Separate and distinct genetic contributions to modern Japanese are evident in the coalescent analysis the short tandem repeat accumulated diversity: haplogroups D and C seem to have begun their expansion in Japan at about 20 and 12 kya, respectively, while haplogroup O2b1-M47z began its expansion only after 4 kya (Hammer *et al.* 2006). Another surprising link has been established between the diverse southern India/Sri Lanka C\* lineages and those of the Australian aborigines, where these represent nearly half of the paternal pool (Karafet *et al.* 1999, Kayser *et al.* 2001): the aboriginal microsatellite diversity forms a tight subcluster, possibly affiliated with a subset of the diverse Indian chromosomes (Redd *et al.* 2002). However, the recent work by Hudjashov *et al.* (2007) identified a new Y marker M347, which distinguishes all Australian C types from Indian or other Asian C types. Together with no affinities found for other lineages of the paternal variation, this adds weight to the rejection of the Huxley's hypothesis of Indian-Australian connection (Huxley 1870).

The reduction of the Y chromosome genetic package to the F-M89 and K-M9 founder lineages, most likely occurred during the westward migration to West Eurasia and Europe. The lineages found in the present-day pool although not participants of the initial migrations likely reflect the dispersal of their precursors, particularly those of the most frequent haplogroups I-M170, J-12f2.1 (within F) and R-M207 (within K, Figure 13; Rosser *et al.* 2000; Semino *et al.* 2000, 2004; Underhill *et al.* 2000, 2003; Hammer *et al.* 2001, Wells *et al.* 2001, Bosch *et al.* 2001, Hammer and Zegura 2002). From the Central Asian P node, the Eurasian R-M207 supposedly expanded westwards and further developed its specific branches (Figure 13). Therefore, the M173-bearing chromosomes in Europe are considered to delineate an ancient expansion from Asia during the Upper Palaeolithic ~30kya (Semino *et al.* 2000, Underhill *et al.* 2001a, Wells *et al.* 2001). Both R1b3-M269 and R1a1-M17 became very common in West Eurasia although harboring opposite clines (Scozzari *et al.* 2001). A male-mediated counter-clockwise migratory route from Southeast Asia towards Northwestern Europe in the Late Pleistocene-Holocene, and thus more recent, is testified in haplogroup N-M231 (Rootsi *et al.* 2007).

Though drastically reduced and remaining limited in size throughout the LGM, the populations experienced a subsequent size expansion, as indicated by the starlike genealogy of the surviving paternal lineages (Underhill *et al.* 2000). From the refugia, the contracted groups of AMH started to spread with the warmer and more humid and stable climate of late Pleistocene and Holocene. As shown in Rootsi *et al.* (2004) haplogroup I, a "genuine European" variant of the Y chromosome divides into different subclades that have likely participated in the recolonization of Europe, from refugia in Francocantabria and East Europe and/or Balkans. The advent of Neolithization starting about 10-12 kya in the Near East, was another main impellor of demographic expansion. It has been suggested that farming societies, usually large and settled, exhibited changes in haplogroup frequencies owing to drift. These were supposedly slow processes and led to clinal structures, a possible scenario for China and Eurasia in the Holocene (Rosser *et al.* 2000, Jobling and Tyler-Smith 2003, Quintana-Murci *et al.* 2004). The spread of

agriculture was not likely caused by a complete population replacement, or solely by cultural transmission, meaning that neither clines of particular lineages, nor relatively deep branches in Europe allow to directly estimate its contribution (Barbujani and Goldstein 2004). The extant population is then regarded as a hybrid among past contributions. The estimated Near Eastern, and thus taken as the Neolithic contribution is, according to some authors, large and decreases as one moves from east to west (from nearly 80% in the Balkans lowering down to a minimum of 15-34% in Iberia and other Western Europe regions; Chikhi *et al.* 2002, Dupanloup *et al.* 2004). These values might be somewhat overestimated for different reasons, but as a tendency, they reflect the likely scenario, put forward already some time ago (for a comprehensive overview see Cavalli-Sforza *et al.* (1994).

### 3.6.1 - Phylogeography of the African paternal variation

Since the earlier studies, populations in the African continent have shown to have the deepest clades of Y chromosome phylogeny (Scozzari *et al.* 1997, 1999; Underhill *et al.* 2000; Hammer *et al.* 2001; Cruciani *et al.* 2002; Hammer and Zegura 2002; Semino *et al.* 2002; Knight *et al.* 2003; Weale *et al.* 2003; Luis *et al.* 2004; Wood *et al.* 2005; Rosa *et al.* 2007). Sub-Saharanans are today characterized by the presence of haplogroups A-M91, B-M60 and the predominant haplogroup E-SRY4064, all clusters sharing the ancestral state relative to the M89 molecular marker (phylogeny depicted in Figure 14).

The deep A-M91 branch is frequent among East and South Africa people, with its Khoisan-specific sub-clades, in a way similar to that of the matrilinear ancestry of this hunter-gatherers that includes the haplogroup L0a, a deep branch in mtDNA phylogeny (Chen *et al.* 1995b, 2000; Knight *et al.* 2003; Destro-Bisol *et al.* 2004). Within haplogroup A, the most widespread and common variant is A3-M32, of typical assignment in Sudanese and Ethiopians (Underhill *et al.* 2000, Semino *et al.* 2002). Its sub-haplogroup A3b1-M51 is represented in South Africa Khoisan (Scozzari *et al.* 1997, 1999; Underhill *et al.* 2000; Wood *et al.* 2005) while A3b2-M13 is found at high proportion in East Africa and at low frequency in Cameroonians (Cruciani *et al.* 2002). On the opposite, the A1-M31 lineages have showed a patchy distribution in rather different geographic regions: Mali (Underhill *et al.* 2000), Guinea-Bissau (Rosa *et al.* 2007), Bakola Pygmies (Wood *et al.* 2005) and Moroccan Berbers (Scozzari *et al.* 2001). The frequency of A2-M14 lineages is of approximately 15% in the Kung and mixed Khoisan (Scozzari *et al.* 1997, 1999; Underhill *et al.* 2000; Wood *et al.* 2005).

Haplogroup B-M60, another deep-coalescing branch in the tree, is found throughout sub-Saharan at marginal proportions (Scozzari *et al.* 1997, 1999; Underhill *et al.* 2000, 2001a; Semino *et al.* 2002; Arredi *et al.* 2004; Wood *et al.* 2005). The B2b-M112 cluster, almost Pygmy-specific, defines the clear-cut difference of these people and all the other Africans (Underhill *et al.* 2000, Cruciani *et al.* 2002, Wood *et al.* 2005). The molecular ages for Y

chromosome A and B lineages differentiation are according to some authors of about  $42.8 \pm 23.0$  and  $36.8 \pm 13.6$  ky, respectively <sup>(Hammer and Zegura 2002)</sup> supporting their ancestry back to the first spreading events of modern human in Africa, and having survived throughout subsequent major expansions. However, as already mentioned above, these time estimates should be taken with caution and more recent estimates offer considerably earlier primary bifurcation within human Y chromosomal lineages still in circulation <sup>(Underhill and Kivisild 2007)</sup>.

The lineages that have acquired an Alu element (known as YAP marker) divide into a pair of sister haplogroups: haplogroup E-SRY4064 and haplogroup D-M174 (phylogeny in Figure 14), the first that has remained mainly within Africa and diversified for the last 50 ky <sup>(Bosch *et al.* 2001, Hammer *et al.* 2001, Underhill *et al.* 2001a)</sup> and the latter of an exclusive Asian assignment <sup>(Hammer and Horai 1995, Su *et al.* 2000, Tajima *et al.* 2004, Wen *et al.* 2004)</sup>. The E-SRY4064 cluster has a non-homogeneous distribution in Africa and the Mediterranean area (Figure 14a) and is actually the most frequent clade in sub-Saharan Africa, where it likely spread from East to West at about 50 kya, and posteriorly to the other quadrants <sup>(Scozzari *et al.* 1999, Underhill *et al.* 2000, Bosch *et al.* 2001)</sup>.

The E\*-SRY4064 refers to a paragroup that gathers a variety of “unspecified-by-SNPs” SRY4064 derivatives. This fraction of chromosomes has been considerably diminished by now, thanks to the new phylogenetically informative markers found. The E1-M33 haplotypes are of low and spotty distribution over the continent, with a pattern suggesting its West-Central African origin (Figure 14b) and subsequent introgression in other populational groups: in the Mali people represents 33% of the Y chromosomes <sup>(Underhill *et al.* 2000)</sup> while in the Burkina-Faso region sums 10% of the pool <sup>(Scozzari *et al.* 1997, 1999)</sup>; exceptionally in the Cameroonian Fulbe the E1-M33 chromosomes peak at 53% <sup>(Scozzari *et al.* 1997, 1999)</sup>; a proportion between 0.7 to 3% traduces the reality for Northwest Arabs and Berbers <sup>(Scozzari *et al.* 1997, 1999; Bosch *et al.* 2001; Arredi *et al.* 2004)</sup>; in Sudan the frequency reaches the 3% <sup>(Underhill *et al.* 2000)</sup>. The E2-M75 types are reported in sample sets of Sudanese and Ethiopians <sup>(Underhill *et al.* 2000)</sup> and are again frequent in the Burkina-Faso Rimaibe people <sup>(Scozzari *et al.* 1997, 1999)</sup>, the North Cameroonian <sup>(Cruciani *et al.* 2002)</sup>, the Central African Mbuty <sup>(Underhill *et al.* 2000, Wood *et al.* 2005)</sup>, and the South African Bantu and Khoisan-speakers <sup>(Underhill *et al.* 2000, Cruciani *et al.* 2002, Wood *et al.* 2005)</sup>. Its distribution pattern in North and Equatorial Africa is depicted in Figure 14c.

The P2 transition <sup>(Hammer *et al.* 1997)</sup> further clusters the Y chromosome variation into haplogroup E3a-M2 in sub-Saharan Africans and haplogroup E3b-M35 in North and East Africa, Mediterranean basin and Middle East (Figure 14d-g; <sup>Underhill *et al.* 2000; Semino *et al.* 2002, 2004; Cruciani *et al.* 2004</sup>). The chromosomes with no further defining marker are placed in paragroup E3\*, reported for East and West African people <sup>(Scozzari *et al.* 1997, 1999; Semino *et al.* 2002; Wood *et al.* 2005; Rosa *et al.* 2007)</sup>.

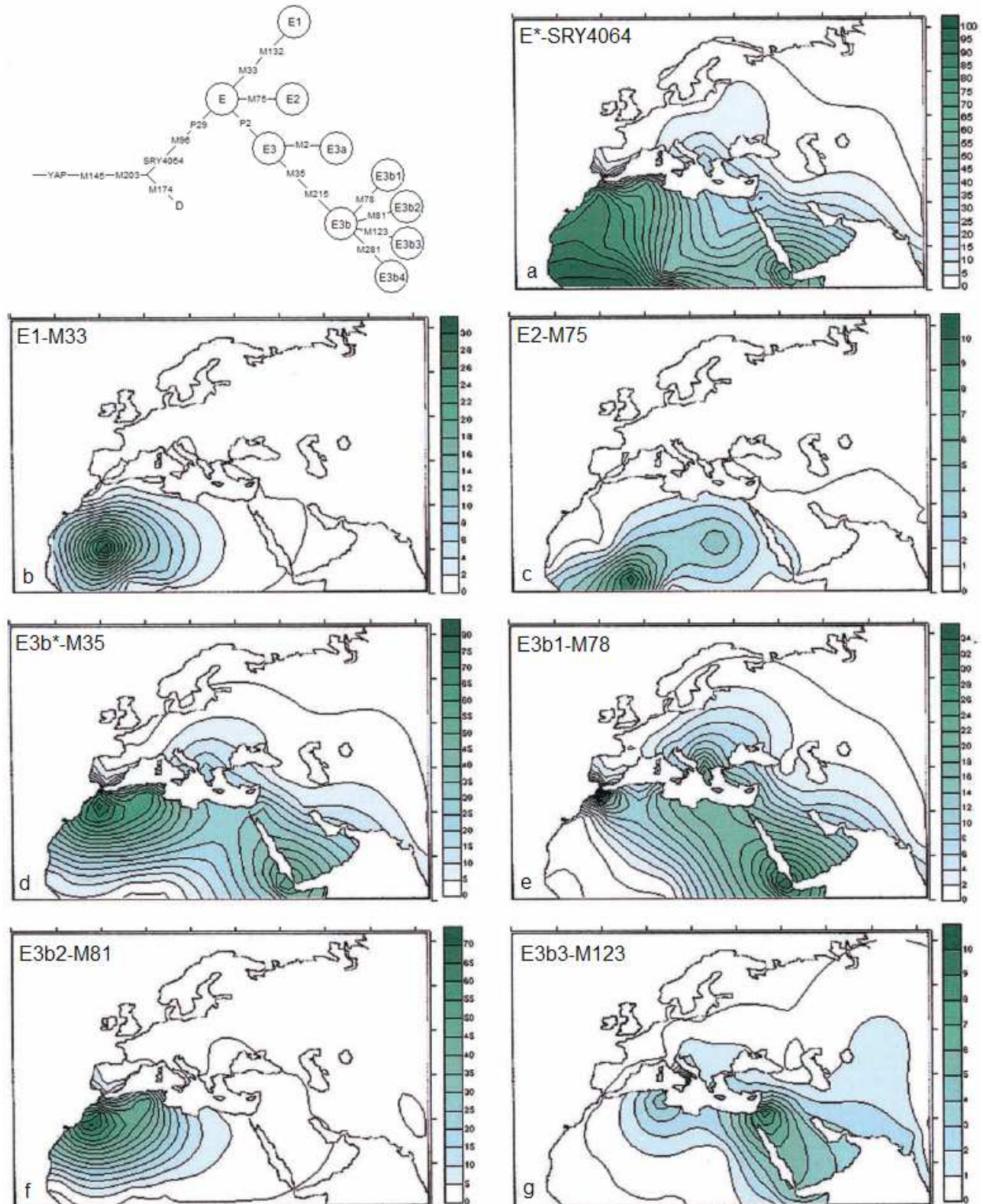


Figure 14 - Phylogeny and frequency distributions of haplogroup E and its main subclades. The numbering of mutations is according to the YCC (YCC 2002; Jobling and Tyler-Smith 2003). Haplogroup-frequency surfaces (individual scales shown in the left of each panel) were graphically reconstructed as described in Semino *et al.* (2004), using population datasets of references therein. In Semino *et al.* (2004).



The haplogroup E3a-M2 has been proposed to have dispersed widely and rather recently through subequatorial Africa, and is likely to signal Bantu dispersals (Passarino *et al.* 1998, Underhill *et al.* 2001a) though its older existence (approximately 19 ky, Semino *et al.* 2004). It comprises more than 65% of the West African paternal pool, peaking at 80-90% in Senegal Mandenka, Burkina-Faso Fulbe (Scozzari *et al.* 1997, 1999; Semino *et al.* 2002; Wood *et al.* 2005) and a few Central African groups (Scozzari *et al.* 1999). The eastwards movement(s) of people in the Sahelian region are reflected in a clinal decrease of E3a-M2 frequency, with the Kenyan Bantu representing a linguistic boundary relative to Ethiopians and Sudanese since the E3a-M2 frequency drops from nearly 50% in the first to zero in the latter (Underhill *et al.* 2000, Semino *et al.* 2002, Wood *et al.* 2005). More than a half of the NRY lineages in South Africa Bantu speakers are classified as E3a-M2, also with a significant introgression into the pool of Khoisan hunter-gatherers (45-58%; Underhill *et al.* 2000, Cruciani *et al.* 2002). Interestingly, a direct link to South Cameroonians is based on exact matches of two particular microsatellite haplotypes (Underhill *et al.* 2001a, Cruciani *et al.* 2002). An expansion time of 3-5 ky has been estimated on the basis of five microsatellites (Thomas *et al.* 2000) for an event that has shaped the genetic landscape below the desert and almost erased the Paleolithic imprints, thus making haplogroups A and B to be rare (Underhill *et al.* 2001a, Semino *et al.* 2002). On the extensively analysed Bantu dispersals there seems to have been a higher male than female line drift, resulting in a reduced incoming paternal variability (Salas *et al.* 2002). The cause may partially be the higher assimilation of females (and thus of mtDNA lineages) in the indigenous populations along the Bantu migration routes, larger effective population size for men in the dispersing groups, or the socio-cultural patterns of admixture (Salas *et al.* 2002). Further refinement of E3a-M2 phylogeny would be certainly helpful. In this respect, one clarifying cluster to unveil population sub-structuring is haplogroup E3a7-M191. Given its clinal distribution, opposite to that of E3a-M2 (from 23% in Cameroon to 1% in Senegal pool; in the case of the Pygmies 40% of the M2 members; Underhill *et al.* 2001a, Cruciani *et al.* 2002, Semino *et al.* 2002), a Central-Western Africa origin and a later demic expansion to West Africa have been hypothesized (Cruciani *et al.* 2002).

Contrary to E3a-M2, the E3b-M35 chromosomes are more common in East Africans, and also present in North Africans and the South African Khoisan (Figure 14d, distribution in South African populations not shown; Underhill *et al.* 2000, Cruciani *et al.* 2002, Wood *et al.* 2005). Haplogroup E3b has probably arisen in eastern sub-Saharan Africa 30 kya as indicated their highest microsatellite diversity and a variety of undifferentiated E3b\* lineages in the East (Bosch *et al.* 2001, Cruciani *et al.* 2004). The later expansion to the Near East and northern Africa happened most likely at the end of the Pleistocene (Underhill *et al.* 2001a). The E3b\*-M35 lineages appear to be confined almost exclusively to the sub-Saharan populations, with their highest levels in the Kenya Massai, ethnic groups in the Democratic Republic of Congo (Wood *et al.* 2005) and the !Kung (Cruciani *et al.* 2002, Luis *et al.* 2004), the people that also harbor the highest STR variability of the paragroup (Cruciani *et al.* 2004). Its several derived states have a patchy pattern of distribution, indicating

independent phenomena of local genetic drift, possibly including founder effects. The following can be mentioned:

- E3b1-M78 occurs commonly in northern and eastern Africans, western Asians and many populations in Europe, mostly Mediterraneans (Figure 14e; <sup>Underhill *et al.* 2000; Bosch *et al.* 2001; Cruciani *et al.* 2002, 2004; Semino *et al.* 2002, 2004; Arredi *et al.* 2004</sup>). Single occurrences are reported in Senegalese <sup>(Semino *et al.* 2002, Wood *et al.* 2005)</sup> and Kenyans but the lineages seem not to have diffused further south <sup>(Cruciani *et al.* 2004, Wood *et al.* 2005)</sup>. The geographically broad distribution suggests that haplogroup E3b1-M78 encompasses a collection of sub-haplogroups with very different evolutionary histories, which nevertheless coalesce back to a putative East African root at about 23 kya <sup>(Cruciani *et al.* 2004)</sup>. On the basis of Y-STR particular alleles, these authors found the internal diversification of E3b1-M78 to diagnose population substructuring: i) cluster E3b1- $\alpha$ , characterized by the rare nine-repeat allele at A7.1, is very common in the Balkans and declines west towards Iberia. It was most likely carried by the Neolithic or the post-Neolithic migrants from the Balkans at about 8 kya; ii) cluster  $\beta$  is characterized by the DYS413\*23/21 form and the rare 10-repeat allele at DYS439, being common and probably autochthonous of Northwest Africans (approximately 14% of their pool, more than 80% of the E-M78 local variation). The estimated TMRCA for E3b1- $\beta$  is of about 5 ky; iii) the cluster E3b1- $\gamma$  is identified by the short DYS19\*11 repeat allele and is of East African prevalence, coalescing at the basal node of E3b1 at about 10 kya. Outside of this area it has been observed only in Egyptians and Moroccan Arabs; iv) the residual haplotypes have been defined as cluster E3b1- $\delta$ , widespread throughout all regions of M35 distribution, albeit at very low frequencies. The lineages here included were supposedly involved in the first dispersals of M78 chromosomes from a putative East African source to North Africa and the Near East regions, at about 15 kya. More recently, the molecular dissection of E3b1-M78 cluster, by analyzing about 60 kb of the NRY portion of chromosomes on each of the four STR-based clades, allowed the identification of six novel SNPs and six new clusters <sup>(Cruciani *et al.* 2006)</sup>. The UEPs showed a striking correspondence with the microsatellite clusters  $\gamma$  and  $\delta$  (markers E-V32 and E-V13, respectively). Conversely, the evidence on E3b1- $\alpha$  and E3b1- $\beta$  confirmed those as monophyletic clusters but their defining binary markers yet to be discovered. The better defined phylogenetic context offers the opportunity to explore the origin and distribution of the chromosomes, in addition to the previously defined but rather uninformative M148 and M224 <sup>(Underhill *et al.* 2000, 2001a; Arredi *et al.* 2004; Shen *et al.* 2004)</sup>. As a consequence of the continuing refinement of the Y chromosome genealogy, most of the terminal haplogroups in the tree are found to be restricted to a specific population and/or geographic region.



- E3b2-M81 is exclusive of and excessive in North Africans, comprising 33 to 76% of their Y-chromosomal pool (Figure 14f; Scozzari *et al.* 1999, 2001; Bosch *et al.* 2001; Cruciani *et al.* 2002, 2004; Semino *et al.* 2004; Wood *et al.* 2005). Interestingly, the area of its distribution matches the present area of the Berber-speaking populations, with representatives in Mali, Senegal and Sudan (Underhill *et al.* 2000, Semino *et al.* 2002) corroborating for a close haplogroup-ethnic group parallelism. The E3b2 origin in North West Africa is estimated at 5.6 kya, with a possibly expansion to East Africa on the last 2 kya (Cruciani *et al.* 2004). A recent gene flow to Iberia has also been testified, a minor contribution of the Arab conquest and long-time presence in the Peninsula (Bosch *et al.* 2001, Cruciani *et al.* 2004).
- E3b3-M123 lineages are found in Ethiopians, North Africans, Near Easterns and some European populations (Figure 14g; Underhill *et al.* 2000; Semino *et al.* 2002, 2004; Cinnioglu *et al.* 2004; Cruciani *et al.* 2004; Wood *et al.* 2005). Although their origin remains unclear, the Near East has been hypothesized as the source of variation, since the East African distribution of E3b3 is basically restricted to Ethiopia, and these lineages have been found in the large majority of the Near Eastern datasets, where they display a higher variance (Underhill *et al.* 2000, Cinnioglu *et al.* 2004, Cruciani *et al.* 2004, Semino *et al.* 2004).
- Solely in Ethiopians, haplogroup E3b4-M281 encompasses 38% of the YAP+ variation of the sample set of Semino *et al.* (2002), whilst M2 lineages are virtually absent.

Cruciani *et al.* (2002) hypothesized on a signature of backflow from Eurasia into North Cameroon, exemplified by a derived form of haplogroup R. These lineages harbour the SNP derived allele at M173, characteristic to and defining haplogroups R1a and R1b, but not the SRY10831.2 or M269 widely present among West Eurasians (Semino *et al.* 2000, Underhill *et al.* 2000, Bosch *et al.* 2001, Wells *et al.* 2001). Representatives of this cluster classified as R1\*, were later identified in the Bantu of southern Cameroon, the Rwanda Hutu and people in Oman and Egypt (Scozzari *et al.* 1999, Luis *et al.* 2004). However, since a set of lineages, labeled as “star”, is not necessarily phylogenetically cladistic (i.e. with their unique joint MRCA), then one should be careful in speculate further on its spotty phylogeography. The Asian origin is anyway quite plausible as the putative source of both European and Cameroonian M173, since most of the M9 variation is Asian (Cruciani *et al.* 2002). Salas *et al.* (2002) have suggested an association of R1\* Y chromosomes with the Eurasian mtDNA haplogroups U6 and H, found in the Fulbe population of Nigeria (Salas *et al.* 2002). A recent North African flow into North Cameroon mediated by Fulbe or other pastoralists could explain the presence of two ancient West Eurasian haplogroups in a restricted region, even prior to 4 kya, On the other hand, such timescale seems unlikely to Cruciani *et al.* (2002) because the nowadays frequent West Eurasian M269 or M173 variants were not found among the North Cameroonians, neither the post-LGM or Neolithic types. The African M173 Y

chromosomes may then be relics of an ancient back migration from Asia to Africa, of some branches of an Upper Paleolithic clade that has emerged ~30 kya <sup>(Underhill *et al.* 2001a)</sup>.

According to Semino *et al.* <sup>(2002)</sup>, other M89 haplogroups within super-haplogroup F that were found in their Ethiopian dataset can signal “back-to-Africa” migrations. The haplogroups G-M201 and J-12f2.1, common in Central Europe and of Middle Eastern distribution have been more specifically associated to the Neolithic expansion <sup>(Semino *et al.* 1996, Rosser *et al.* 2000, Semino *et al.* 2000)</sup>. Their presence in North Africa, especially in Egypt was interpreted as a southern branch of the Neolithic demic diffusion, originating from the Near East <sup>(Luis *et al.* 2004)</sup>. Yet, apart from the Ethiopian exception the two mentioned clusters appear to be absent in sub-Saharan <sup>(Underhill *et al.* 2000, Cruciani *et al.* 2002)</sup>. In the same context, it is worth to indicate that the Ethiopians and Cameroon Fulbe carry also K2-M70 chromosomes in their paternal pool, a rather minor clade but that according to the present understanding of the phylogeny, derives directly from the M9 central node <sup>(Cruciani *et al.* 2002, Luis *et al.* 2004)</sup>.

For widely distributed haplogroups, further phylogenetic relationships can be analyzed by the microsatellite typing, which allows associating particular haplotypes to geographic regions, and defining the coalescence time of the cluster’s variation. For instance in African variants, haplogroup A3-M32 shows no haplotype sharing across Cameroonian and Ethiopian populations, E3b-M35 are distinct in Ethiopia and South Africa and B2-M182 haplotypes found in Pygmies are not common among the Khoisan <sup>(Cruciani *et al.* 2002)</sup>. Therefore, though a shared haplogroup is virtually an unambiguous proof of a common MRCA for two or more populations, enough time has passed to generate the further variation characteristic of the extant populations, by molecular evolution of STRs (repeat gain and loss), probably associated by distinct founder events and random genetic drift in general. In the cases of differences generated by drift, these may arise fast provided that influential demographic factors like bottlenecks/founder effects are involved. The Y-STR analysis has also shown that the differences among non-Africans are mostly intrapopulational, whereas the Africans exhibit the highest interpopulation variability (e.g. <sup>Jorde *et al.* 2000</sup>).

Unfortunately, one must accept that several areas of the continent are not yet covered by sampling and analysis, or are only superficially investigated, including a large portion of the Saharan/Sahelian belt, North Africa from Tunisia to Egypt and South African regions. Clearly, additional studies are necessary for a better understanding of the origin and distribution of many interesting lineages, only partially known thus far. The haplotypes observed at low frequency could represent important signatures of pre-agricultural settlements that have been overwhelmed by the strong demographic impact of the farmers.

Y chromosome studies among European (Semino *et al.* 1996, Rosser *et al.* 2000, Semino *et al.* 2000, Chikhi *et al.* 2002, Quintana-Murci *et al.* 2004), American (Zegura *et al.* 2004) and Austronesian (Hurles *et al.* 2002) people have in general showed that geographic distances correlate better with their genetic component than with their linguistic affiliation. Yet, it may not be a common rule, because for instance in Siberians, the opposite has been observed (Karafet *et al.* 2002). However, Siberia is a rather specific case because of its wide area, contrasting with very low populational density throughout tens of thousands of years. According to some authors, when African populations are tested for associations between genetics, linguistics and geography, the genetic diversity is apportioned among both variables (Poloni *et al.* 1997, Scozzari *et al.* 1999, Salas *et al.* 2002). Others found the Y-chromosome genetic variation to better correlate with linguistics while the mtDNA component is weakly correlated with both linguistics and geography (Wood *et al.* 2005). However, if the Bantu speakers are taken out of the picture, the Y chromosome-linguistics “link” seems to fade, while it strengthens for mtDNA (Wood *et al.* 2005). These data suggest that patterns of differentiation and gene flow in Africa have differed for men and women in the recent evolutionary past. Sex-biased rates of admixture and migration and/or language borrowing between expanding farmers and local hunter-gatherers might have played then a striking role in influencing the patterns of genetic variation (Destro-Bisol *et al.* 2004). In fact, although the Bantu people occupy a wide region, some are genetically closer to each other irrespective of geography, than geographically restricted language groups are. On the other hand, very close genetic similarities can be also found in linguistically different groups that inhabit the same area, suggesting gene exchange without language change (Rando *et al.* 1998, Scozzari *et al.* 1999, Cerny *et al.* 2004). Furthermore, the role of polygyny and patrilocality is sustained by the evidence of a differential pressure of genetic drift and gene flow on maternal and paternal lineages (Destro-Bisol *et al.* 2004, Wood *et al.* 2005). As a consequence, local differentiation of the Y chromosomes is enhanced and by contrast, mtDNA is expected to show lower geographical clustering (Jobling and Tyler-Smith 2003). In that sense, the groups of people targeted in the studies are largely defined by cultural and historical aspects. Since demographic differences in relation to gender may modulate disparities in the uniparental inherited genetic systems, their combined analysis is essential.

## 4 – Complementary sources of evidence

In order to fulfil the natural curiosity about our origins we have to consider many different and independent sources of evidence, from a full range of disciplines. Such diverse approaches and records contribute with complementary and independent interpretations, concordant or not, or even new hypothesis, for strengthening the credibility of the conclusions <sup>(Harpending *et al.* 1998, Underhill *et al.* 2001b, Cavalli-Sforza and Feldman 2003)</sup>. Although the rooted phylogenies from molecular genetics studies provide in theory an absolute chronology for the events, related to the branching pattern of phylogenetic trees, the challenge is in integrating data with the chronology obtained from disciplines such as archaeology and the dynamics of palaeovegetation and palaeoclimatology, where radioisotopes and several other physical methods are accurately quantifiable, though not always universal and applicable <sup>(Beck *et al.* 2001)</sup>. A cross reference between records permits a placing in time so that we can relate, for instance, a population fluctuation with a climatic change and/or a technological innovation. The following topics try to gather the most relevant information for the purpose of Guinea-Bissau population genetics.

To geographically contextualize Guinea-Bissau it is worth to briefly mention some considerations. The Republic is located in the West African coast, in a territory of 36 thousand square kilometres, surrounded by Senegal at its North and Guinea at East and Southeast. The country also integrates more than 40 islands that constitute the archipelago of Bijagós. The territory is in close proximity to the so-called Sudanese belt (an almost uninterrupted strip of land of sparsely forested savannas south of the Sahara, from the Atlantic Ocean to the Red Sea) and the Sahel (a narrow strip of arid land at the most northern region of the belt). The Senegambia region has an historical connotation to geography, to specify the territory extending from Senegal to Gambia.

### 4.1 - Environmental records

Ancient African migrations are complex, owing to historical fluctuations in geology, ecology and climate, that have affected population expansions and contractions <sup>(Lahr and Foley 1998)</sup>. During the life span of *Homo sapiens* as a species, the African climate has been profoundly influenced by the Sahara desert, the largest in the world, which covers one quarter of Africa and counts only with few oasis inhabitants at the present. With the latitudinal increase of humidity, the vegetation at both its north and south boundaries changes from desert and subdesert scrubs to grasslands and xeric trees like acacias and baobabs. Then the wooded savannas, on which Guinea-Bissau is included nowadays, give place to the tropical rain forest, with a usual sharp boundary (Figure 15a, based on reconstruction of <sup>Adams and Faure 1997</sup>).

The Sahelian zone in West Africa has been ever since affected by major climatic alterations and these changes had an important impact on the peopling of the area. Palaeovegetational studies have demonstrated that in-between 20-13 <sup>14</sup>C kya (about 23-14.5 calendar kya) the continent was extremely arid, with the desert at its widest (Adams 1997). At this so-called Last Glacial Aridity Maximum, the Sahara reached hundreds of kilometres further south (Figure 15b), according to ancient sand dune distributions. A belt of semi-desert appeared to the south of the present-day desert margin (Hooghiemstra *et al.* 1992) and there was a rainforest retreat in the entire equatorial zone, replaced by savanna and grasslands (Hamilton 1988, Jahns 1995). The rainforests were reduced in much of the Congo basin, confined to “pockets” in close proximity to rivers, namely in Nigeria and Sierra Leone. These may have acted as refugia, from where people have later dispersed. However, the position and extent of forested refugial areas is controversial due to the sparsity of evidence. Nevertheless, tropical rainforest “pockets” have probably existed quite near to the Guinea-Bissau grassland area.

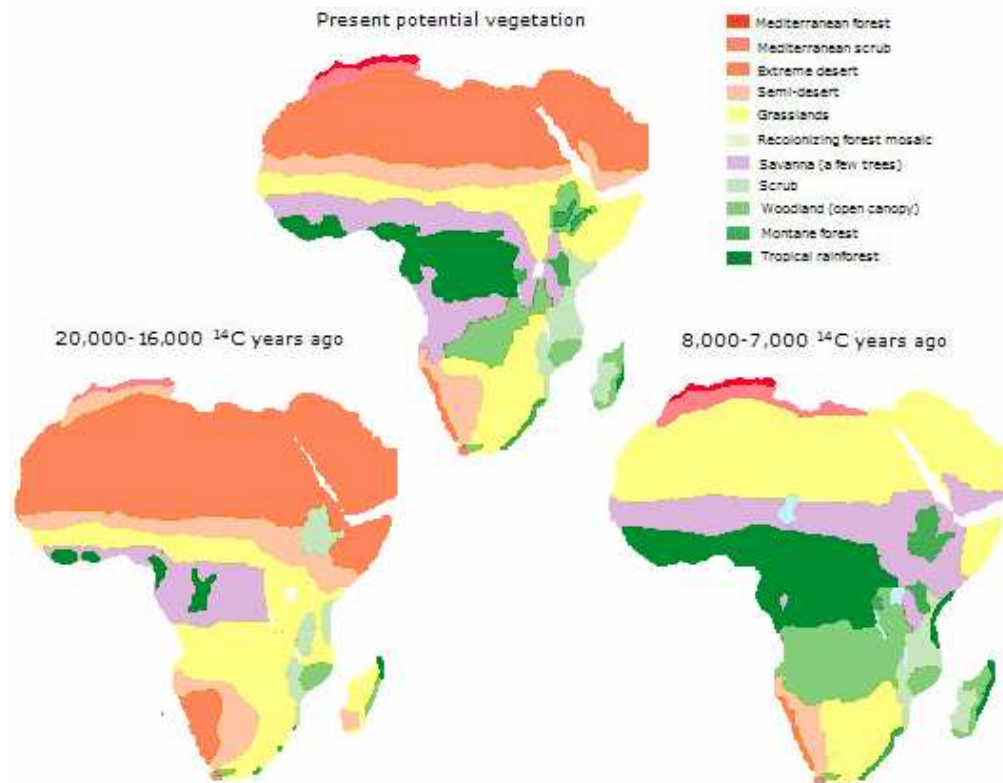


Figure 15 – Distribution of vegetation zones in the African continent at several time intervals, based on reconstruction by Adams (1997). a) present-day potential vegetation; b) putative distribution 20-16 <sup>14</sup>C kya; c) putative distribution 8-7 <sup>14</sup>C kya. Based in reconstruction from Adams and Faure (1997).

A slight moistening of the climate and temperature rise has been noticed in the pollen records of Central and East Africa around 14 <sup>14</sup>C kya (Hamilton 1988). Forest remnants in the northern parts of the Gulf of Guinea and in several mountainous areas indicate that the tropical rainforest today concentrated in the lowlands, has extended more widely. Around 7-8 <sup>14</sup>C (9 kya) the Sahara went through a period of maximum wetness (Aumassip *et al.* 1994; Figure 15c), becoming habitable with savannah, grassland rivers and large lakes. At that time, a border savannah- rainforest is supposed to have existed in the location of Guinea-Bissau. A posterior progressive desertification ~4kya, that drove to the nowadays distribution of climatic and ecological zones, forced the pastoralists in the desert to move northwards, to the Maghreb), and southwards to the Sahel and the savannahs (Clark 1980).

## 4.2 - Archaeology and anthropology of the pre-history

### *The Middle Stone Age*

The fossil and artefact evidence in Africa are unfortunately weak on a scarce material background as the soil chemical and physical characteristics are unfavourable to preserve skeletons (e.g. no remains in Zaire) or conserve human-shaped objects and landscapes. Recent findings of *Homo sapiens sapiens* fossils unearthed near the Ethiopian village of Herto fill a gap in the record of our direct ancestors, by dating between 154-160 kya (White *et al.* 2003). Before, fossils with modern human morphology have been found in Omo-Kibish (Ethiopia), with a suggested age of 130 ky determined by associated shells (Butzer *et al.* 1969), and there was also evidence of a AMH marine diet at ca. 125 ky in East Africa (Stringer 2000). However, in February 2005 a re-dating of 195 ky ( $\pm$  5ky) for Omo I and Omo II skulls was proposed, based on geological testimonies of Kibish deposit rocks (McDougall *et al.* 2005). Thus far, these fossils are considered as being the oldest links of the emergence of AMHs. Other important sites include the Border caves and the Klasies River Mouth (South Africa) dated at the earliest of 100-110 kya (Grun *et al.* 1990, Rightmire and Deacon 2001).

The Middle Stone Age industries, named after tool assemblages, seem to have begun considerable earlier in Africa than elsewhere and to have had here a more consistent technological sophistication (Tattersall *et al.* 2000). Fossil remains from that time indicate three major human groups in Africa: i) the ancestral Khoisanids Bushmen and Hottentots, that once extended to East Africa (Huffman 1982, Newman 1995); ii) the ancestral Negroid in West Africa, who today live in tropical areas and much of East and South Africa (Tattersall *et al.* 2000); iii) the ancestral Caucasoid-related groups in North Africa (Camps 1974). Archaeological discoveries in southern and eastern Africa suggest that, at approximately 80-60 kya, certain African groups assisted to a major increase in the complexity of their technological, economic, social and cognitive behaviour. For instance, signs of modern behaviour like geometric

notches in bone <sup>(Henshilwood *et al.* 2002)</sup> date *ca.* 80 kya in the South of the continent. Major demographic expansions are believed to have happened at the time, probably triggered by the innovative technology and environmental changes and eventually lead to the “Out-of-Africa” expansion <sup>(Mellars 2006)</sup>. The Hofmeyer’s skull that has been deposited in South Africa for 36 ky is morphometrically more similar to modern humans of European Upper Palaeolithic than to recent South Africans or Europeans <sup>Grine *et al.* 2007</sup>. Thus, it seems that modern sub-Saharan and Europeans share a relatively recent common ancestor that has likely expanded out of East Africa 60 kya <sup>(Mellars 2006)</sup>.

### *The Late Stone Age*

Until recently the presence of modern humans in North West Africa was only evident from 40 kya onwards <sup>(Alimen H 1987)</sup>. New evidence of personal decorations suggests that people with “modern behaviour” existed in NW Africa about 100 kya <sup>(Vanhaerens *et al.* 2006, Bouzougar *et al.* 2007)</sup>, though not much more is known. Industrial sites of the Late Stone Age period (*ca.* 25ky) were discovered in Maghreb with important assemblages related to the Iberomarusian culture (22-9 kya, <sup>Camps 1974</sup>). An equally large number of sites were associated to the Epipalaeolithic (microlithic) complex that suppressed the Iberomarusian – the Capsian. This pre-agricultural culture evolved locally or through diffusion from the Near East <sup>(Camps-Fabrer 1989)</sup> and was centred on the eastern parts of the Atlas plateau <sup>(Camps 1974, Lubell 1975)</sup>.

The human activity in the African equatorial belt is manifested in rock shelters of about 35 kya in Cameroon and Equatorial Guinea <sup>(Mercader and Marti 2003)</sup>. Although less well dated, Late Stone Age industries appeared throughout the sub-Sahara, possibly as early as 50 kya <sup>(Foley and Lahr 1997)</sup>. Other few though undated sites in Ghana, Nigeria and Burkina-Faso all produced microlithic industries but little is known about the economies and the Late Pleistocene sequence of events in the West Africa savannahs and rain forests <sup>(Clark 1994)</sup>. The archaeological sequence in Ounjougou (Mali) is quite exceptional for sub-Saharan West Africa. Analysis of lithic elements seems to point for an almost continuous occupation of the area from 70 kya until 20 kya <sup>(Phillipson 1993, Newman 1995, Cornelissen 2002, Rasse *et al.* 2004)</sup>. A cultural flow, from the southeast of sub-Saharan Africa and to the Sahara, could explain the diffusion of the microlithic industries all the way through West Africa. Sites are initially observed in Cameroon at Shum Laka (~31-33 kya), then at the Ivory Coast in Bingerville (~15-16 kya), in Nigeria in Iwo Eleru (~13 kya), and finally in Northern Mali (~12 kya, Ounanian culture; <sup>Clark 1994</sup>). The repopulation of the area seemed to have happened only in Early Holocene ~9kya. It has been considered as an early agricultural occupation, where pottery and seed grinding implements date at least since the eighth millennium BC, and are the oldest artefacts of this type known at present in sub-Saharan Africa <sup>(Clark 1994)</sup>.

## *Introduction of agriculture*

The post-LGAM reoccupation of the Sahara by animals and humans happened about 9 kya. These were however rare and widely dispersed, mainly in close proximity to oases <sup>(Camps 1974)</sup> perhaps indicating sparseness of human presence during the reoccupation phase. The Sahara's "wet phase" at 9-8 kya is also coincident with hunter's rock engravings <sup>(Mori 1974, Hassan 1978)</sup> that state for a non-arid climate, under which several "Neolithic"<sup>1</sup> cultures started to flourish. The Near East was the earliest and better characterized region of the agricultural origin for the spread into Egypt 9.5-7 kya, where Caucasoid and Negroid people seem to have coexisted <sup>(Dutour et al. 1988)</sup>. The already domesticated cereals found a similar climate in the Nile Valley proximity and together with cattle created the foundations for a food-producing economy that supplanted the hunter-gathering lifestyle. The shift to agriculture supported a larger number of people, with a better nutrition, ensuring higher life expectancy and increased fertility <sup>(Diamond 2002)</sup>. The numerical growth drove ultimately to waves of migrations, believed to be the responsible for the spread of the agricultural knowledge, in sub-Saharan Africa in particular.

The Middle Eastern "agricultural package" did not succeed under the sub-Saharan climatic conditions, thus delaying the shift to full-scale food production economy. The Late Stone Age microlithic cultures continued until new cultural elements replaced the hunting and gathering economy <sup>(Clark 1994)</sup>. As mentioned before, this author hypothesized on an early agricultural occupation since 10 kya on the basis of archaeological evidence. Other authors point to more recent time, though not later than 6 kya since pottery and ground stone artefacts started to occur in likely small and isolated communities in Nigeria, Sierra Leone, Burkina-Faso and Ghana <sup>(Atherton 1972)</sup>. Their direct association to agricultural practices remains to be proved <sup>(Calvocoressi and David 1979, Phillipson 1993)</sup>. However, at around the same time, centres at the Sahel zone were cultivating specific packages of crops: besides the predominant sorghum, finger and pearl millets, the savannah complex included cowpea, gourd and baobab (Figure 16). At the West African forest margin African rice, oil palm and tuber yams were the major crops, the latter known from ~3.5 ky (e.g. Kintampo culture in Ghana, <sup>Stahl 1985</sup>).

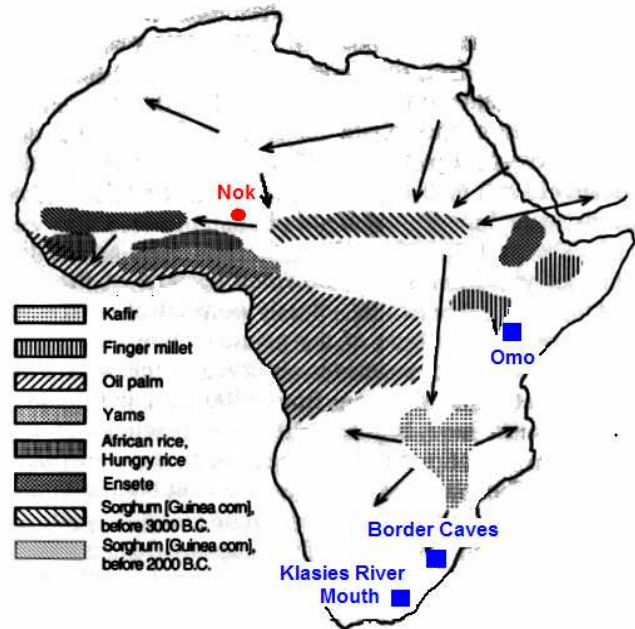
The coastal resources of West and Equatorial Africa, namely in Mauritania and Senegal have been likely exploited already at 4 kya <sup>(Sutton 1982)</sup>. A clear sign of domesticated cattle is evident in figures dating of the Bovidian period ~5.5 kya, where human Negroids of morphological and cultural Fulani resemblance are also represented <sup>(Smith 1982, Dupuy 1999)</sup>. The progressive desertification forced a retreat

<sup>1</sup> The term Neolithic is generally used to refer to the advent of agriculture associated to the change in tool usage, pottery manufacture and increased human sedentism. As it proved difficult to identify the beginning and end of the cultural phases in the African archaeological record, the rejection of the term has been adopted by some archaeologists <sup>(Sinclair et al. 1993)</sup> and geneticists <sup>(Jobling et al. 2004)</sup>.



of the domestic animals to the south, to the savannas of the Sudanese belt, their presence in West Africa and Ethiopia dating at about 3.5 and 3 kya (Shaw 1980).

Figure 16 – Cultivated local plants in African Neolithic (Shaw 1980). Arrows refer to the introductory movement of the domesticated cultivars. The squares indicate the archaeological sites and early AMH remains, referred to in text. Adapted from Cavalli-Sforza *et al.* (1994).



### Iron Age

Elsewhere in Equatorial Africa the hunting and gathering continued up to the coming Iron Age. The iron smelting techniques, which were developed in West Asia ~3.5 kya, reached Egypt and Nubia with the Assyrians ~2.7 kya and quickly supplanted stone, copper and bronze material for the tool-making (Phillipson 1993). In the nowadays territory of Nok (Nigeria) is situated the earliest large iron-smelting centre (~2.5 kya) in sub-Saharan Africa. Soon many others flourished in sub-Saharan East Africa but only a few sites are documented in the western quadrant (e.g. Kissi Burkina-Faso ca. 2.1-2.7 kya and Fiko at Mali's Dogon Plateau ca. 2 kya (Huysecom 2002, Magnavita 2003, Serneels 2005, Robion-Brunner *et al.* 2006). With the advent of both agriculture and iron-smelting, farming economies spread further and centralised political systems emerged ~1.5 kya giving rise to the Late Iron Age expansion. Though biological and archaeological evidences show that there were major expansions towards South and Central Africa before the Iron Age, the main impulse for the Bantu movements was most likely the art of iron production at this time (Cavalli-Sforza *et al.* 1994).

### 4.3 - Historical and ethnical background

The historical records that have reached our days state that in the last two millennia early states began to develop in sub-Saharan Africa. For instance, the Ghana Empire in between Southeast Mauritania and Mali, constituted by Mande farmers (the vast group on which Mandenka are included) and Berber pastoralists, is the oldest known occidental African Kingdom (since the 8<sup>th</sup> century). Soon after, around the 12<sup>th</sup> century, the Ghanian state was destroyed by the Arabs and the region assisted to the urge of the Mali and the Songhai “Black Kingdoms” from the 14<sup>th</sup> to the 16<sup>th</sup> century. The Mali Empire insurged from the revolt of the Mandenka over the Sussu people. The Songhai state was also under rule of a Mande emperor that brought conquests and political reforms. The Sahel was then integrated in a context of commercial reasons, with a route corridor along Mauritania, Mali, Burkina-Faso and Niger, up to Sudan (Almada, 1964). The main objects of the trans-Saharan trade were gold (mined in the Sudan), copper, ivory and salt for North African and Middle Eastern manufactured goods (Cavalli-Sforza *et al.* 1994).

One of the earliest documented inputs of people to Senegambia refers to the massive arrival of **Fulbe** in Futa-Toro on the 8<sup>th</sup> century (Carreira and Meireles 1959) from a Central African epicentre. Other parallel Fulbe migrations occurred at that time, namely to Sudan, Upper Niger and North Nigeria. The Fulani are nowadays mostly nomadic pastoral communities that nevertheless keep somewhat separate of the local agricultural populations. Exceptions were made in order to freely circulate in other’s people territory, where cattle and women were paid as a tribute (Almada 1964). From the 12<sup>th</sup> century on, these people started to expand to Mali and by the 15<sup>th</sup> reached Upper Senegal, Upper Gambia and Guinea-Bissau Beafada territory, up to Sierra Leone. Some of them miscigenized with the Serer from Senegal giving rise to today’s Toucoulers. On the 16<sup>th</sup> century Pastoral Fulbe arrived again slowly but en masse, from Futa Toro and Sahel, dominating the region.

Together with Fulbe, the **Mandenka** arrived to Guinea-Bissau on the 16<sup>th</sup> century. Both with a broad distribution in the Sahel, and presumed to be of non-Bantu origin (Teixeira da Mota 1954) became the most prevalent groups in the territory. The Mandenka are a Mande people physically and culturally descendent from the Mali Empire which controlled the trans-Saharan trade from the Middle East to West Africa. Later in the 19<sup>th</sup> century their occupation on Western Africa continued with the aims of conquering new territories and propagating Maomet religion, after being islamized by the Fulbe. In fact, the Futa-Djalon state located in between Guinea-Bissau and Sierra Leone was born in the previous century when Fulani Muslims rose against the non-Muslims.

The denomination “Brame” is mainly used for the Papel, the Manjaco and the Mancanha. Thought to have been one single people, on the 15th century were know by separate names. Their languages show high affinities and all share religious beliefs and even ceremonies <sup>(Carreira and Quintino 1964)</sup>. In a broader sense the “Bramés” can also include the Balanta, the Djolas and the Beafada.

The **Balanta** are among the today’s most numerous groups and have quickly spread over other ethnic group territories in the first quarter of the 20<sup>th</sup> century, especially in the southern part of the country. Their origin is uncertain, apart from the suggestion of “Sudanese” *sensu latu* and even Bantu affinities. From the cultural and somatic point of view the Balanta are probably closer to Bantu-speakers than to the Sudanese family (see next chapter, <sup>Quintino 1969</sup>). Moreover, there are language affinities between the Balanta and most of the Bantu languages and, as the Bantu, can be the result of Camite invaders from Asia admixture with local North Africans at the end of the Pleistocene.

The **Felupe**, the **Djola** and the Baiotes are in truth all Djolas, with the Europeans being the responsible on giving the names that are today known. According to oral tradition these people came from Sudan in the 15<sup>th</sup> or 16<sup>th</sup> centuries. The Beafada also call themselves Djolas though the rather heterogeneous group has an oral tradition of Mali origins <sup>(Lopes 1999)</sup>.

Teixeira da Mota <sup>(1954)</sup> considers **Nalú** as the autochthonous population of the region they occupy today, being there much before the 15<sup>th</sup> century.

The **Bijagós**, inhabitants of the archipelago with the same name, can represent a separate branch of Djolas but can also relate with Papel or Nalú. Quintino <sup>(1964)</sup> proposes strong cultural resemblances with Egyptians but research has so far reached low scientific resolution, their origin being extremely uncertain.

The admixture of Berbers with native populations of West Africa dates back at least to the 8th century A.D. by the times of the Ghanian Empire. In 1086 Omniade hordes conquered North-Western Africa and again pushed the Berber Almoravids from South Morocco and Mauritania to the Senegal region <sup>(Moreira 1964)</sup>. By the 15<sup>th</sup> century the Portuguese arrived to Guinea-Bissau and found the nowadays known ethnic groups already settled. The European occupation brought down many of the ethnic barriers, upcoming an intense cultural contact and higher level of miscegenation. However, the Portuguese constitute a very small portion of the nowadays inhabitants, caused by the exodus of Portuguese settlers that took place after Guinea-Bissau gained independence.

The tribes are often endogamous, and therefore their genetic pool is mostly determined by cultural factors. However, the geographical distances can here play an important part. For example, although belonging to the same great group of Fulbe, the Senegalese Peul and Nigeria Fulani have distinct genetic pools (Cavalli-Sforza *et al.* 1994). There are nevertheless indications of frequent miscegenation of the numerically superior Fula and Mandenka men with Balanta, Manjaco and Papel women, where offsprings are included in the father's lineage (Carreira and Meireles 1959), in a cultural expansion of islamized people called "Sudanization". The process of 'Balantization' of Papel started about 100 years ago (Carreira and Meireles 1959), with the matting of Papel men and Balanta women the most frequent.

Frequent movements, specially pressing the populations to coastal areas have been registered even in a limited territory as Guinea-Bissau. The internal wars between ethnic groups for land conquering, the slave capture purposes and the spread of the religious beliefs were since ever the main responsible. The animists have initiated migrations for demographic expansion, not for cultural diffusion purposes but because of the fall of the socio-political organization (as a consequence of the European administrative system) and the limitations of the natural resources. On the 19th-20th centuries the Balanta and the Fulbe moved from northern parts to Guinea-Bissau coast due to climate changes (Teixeira da Mota 1954). Demographic data from 1950 indicates an increase of Balanta and Brame migration against a decrease of Fula and Mandenka people mobility. Due to their complex history the major ethnic groups in Guinea-Bissau do not follow a clear present-day settlement pattern but maintain a certain degree of identity (Figure 17).

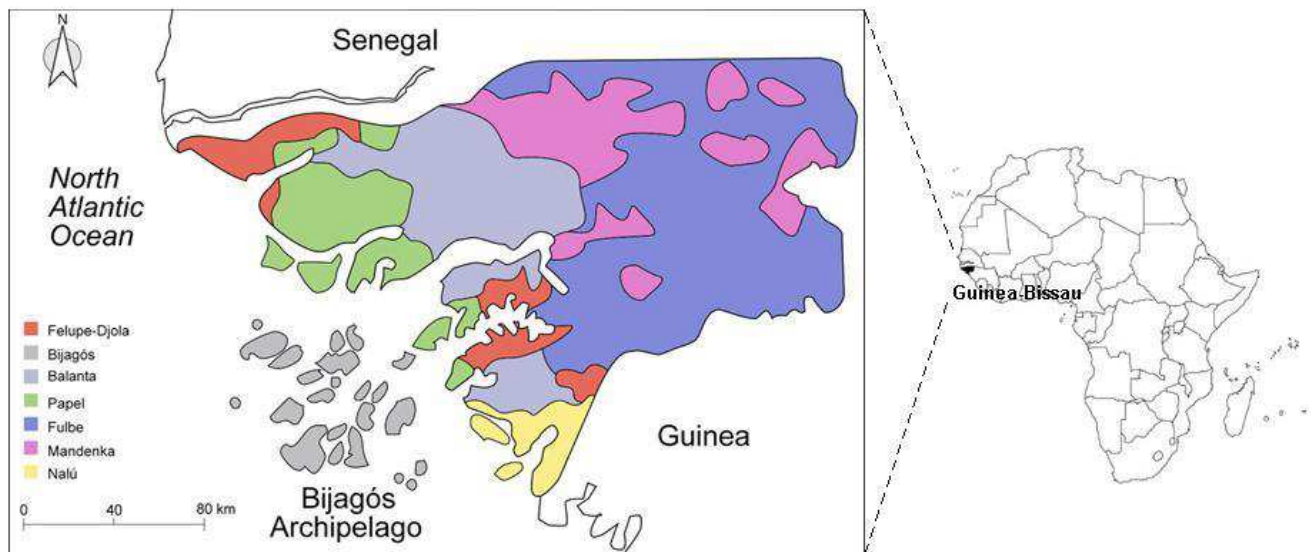


Figure 17 – Present-day settlement pattern of the main Guinea-Bissau ethnic groups considered in this study. To note that the boundaries may not entirely correspond to the precise distribution as overlapping areas do exist. Based on information from Moreira (1964).

Further data on ethnic groups, including their religious affiliations are presented in Table 1 in Rosa *et al.* (2004). Most people are farmers with traditional religious beliefs (animism); 45% are Muslim, principally the Fulbe and Mandenka peoples; and less than 8% are Christian, most of whom are Roman Catholic. Today's perspective shows a clear pattern of increasing number of Fulbe and Mandenka, while Felupe-Djola, Baiote, Cassanga and Beafada tend to decrease (2002 official census; information on Gordon and Raymond 2005).

#### 4.4 - Linguistic affiliation

The deepest ancestral relationships of languages spoken in Africa are probably beyond linguistic reconstruction as languages evolve much faster than genes and can, after all, be replaced by entirely different ones in few generations. Moreover, linguistic barriers may strengthen the genetic isolation between groups speaking different languages. Furthermore, even if the gene flow between populations does exist, it does not have to be accompanied by language replacement,

Greenberg (1963) proposed a classification system for the present-day autochthonous African languages, subdivided in four major phyla: Niger-Kordofanian, Nilo-Saharan, Afro-Asiatic and Khoisan (see distribution pattern in Figure 18). The sub-Saharan linguistic families are supposed to have arisen



Figure 18 – Distribution of major linguistic families in Africa (according to classification of Greenberg 1963). Based on illustration from *African languages: an introduction* (2000), Heine *et al.* (eds).

between the Sahara and the Equatorial forest <sup>(Blench 1993)</sup>, with Niger-Congo and Nilo-Saharan sharing a common ancestor <sup>(Phillipson 1993)</sup>. The branch of Niger-Kordofanians comprises the Kordofanian languages spoken in Sudan and the diverse Niger-Congo phylum, with more than a thousand languages and over 180 million geographically dispersed speakers. The wide distribution can be owed to the expansion of iron-working agriculturalists <sup>(Diamond and Bellwood 2003)</sup> over hunter-gatherers found in their path, what would lead to correlation of linguistics and genetics. The most outstanding diversity is stated for the Niger-Congo region, the putative cradle of Bantoid languages <sup>(Johnston 1919, Greenberg 1974)</sup>. Interestingly, the two main sub-groups of Bantu language correspond to the western and eastern dispersal routes <sup>(Vansina 1994)</sup>, so that the linguistic term covers the biological reality. The likely homelands of the largest language families appear to be near the centers of agricultural innovation, from where they could have moved along with the culture <sup>(Renfrew 1987, Bellwood 2001)</sup>.

In Guinea-Bissau population, the indigenous languages Balanta, Fulbe, Manjaco, Mandenka and Papel count with the higher number of speakers. In contrast, the Portuguese official language, the Kriol (a Portuguese-based creole language) and other European languages are spoken by only 14%. The autochthonous languages are within the Niger-Congo Atlantic (the scheme in Figure 19 elucidates for the inner subdivisions and relative proximity of the ethnicities, based on information from <sup>Gordon and Raymond 2005</sup>). To note however that the subdivisions have no linear levels of classification and hierarchy that may define a consistent cladogram. Tracing back the common nodes of languages, the majority of the tribes in Guinea-Bissau are thought to be related to “Sudanese” or even Ethiopian ancestors.

The present-day distribution of the “Sudanese” family goes from river Senegal and Niger to Upper Nile <sup>(Quintino 1967)</sup>. According to Stuhlmann <sup>(1910)</sup>, this family derives from a Bantu branch, separated in the Pleistocene near the Nile. For the movement contributed the arrival of Camite hordes from Asia in successive waves, expelling the natives. Mandenka, Fulbe, Manjaco, Papel and Beafadas are considered members of the occidental Sudanese family. For Felupe-Djola and Baiotes, oral sources even say they come from Sudan on the 15<sup>th</sup> and 16<sup>th</sup> centuries, showing affinities with Bijagós and Papel. The proposed affinities of Balanta with Sudanese go back 2 ky, separated with the first spread of Camite hordes <sup>(Quintino 1969)</sup>. The Fulani languages, that include the Fulbe speakers, are the tongue of several million people inhabiting an area from Senegal to a region East of Lake Chad. The Fula show the typical “glottal catch” which characterizes the whole group. The Mande group consists of languages prevalent in the Niger valley, Liberia, and Sierra Leone, such as Mende in Liberia and Malinke in Mali. The approaches to group Manding languages summed four to five million speakers in nine West African countries, the Bambara dialect prevalent in the east and the Mandinka in western Senegal, Gambia and Guinea-Bissau <sup>(Sullivan 2004)</sup>. Galtier <sup>(1981)</sup> includes Sussu with Mandenka based on his lexico-statistic calculations.

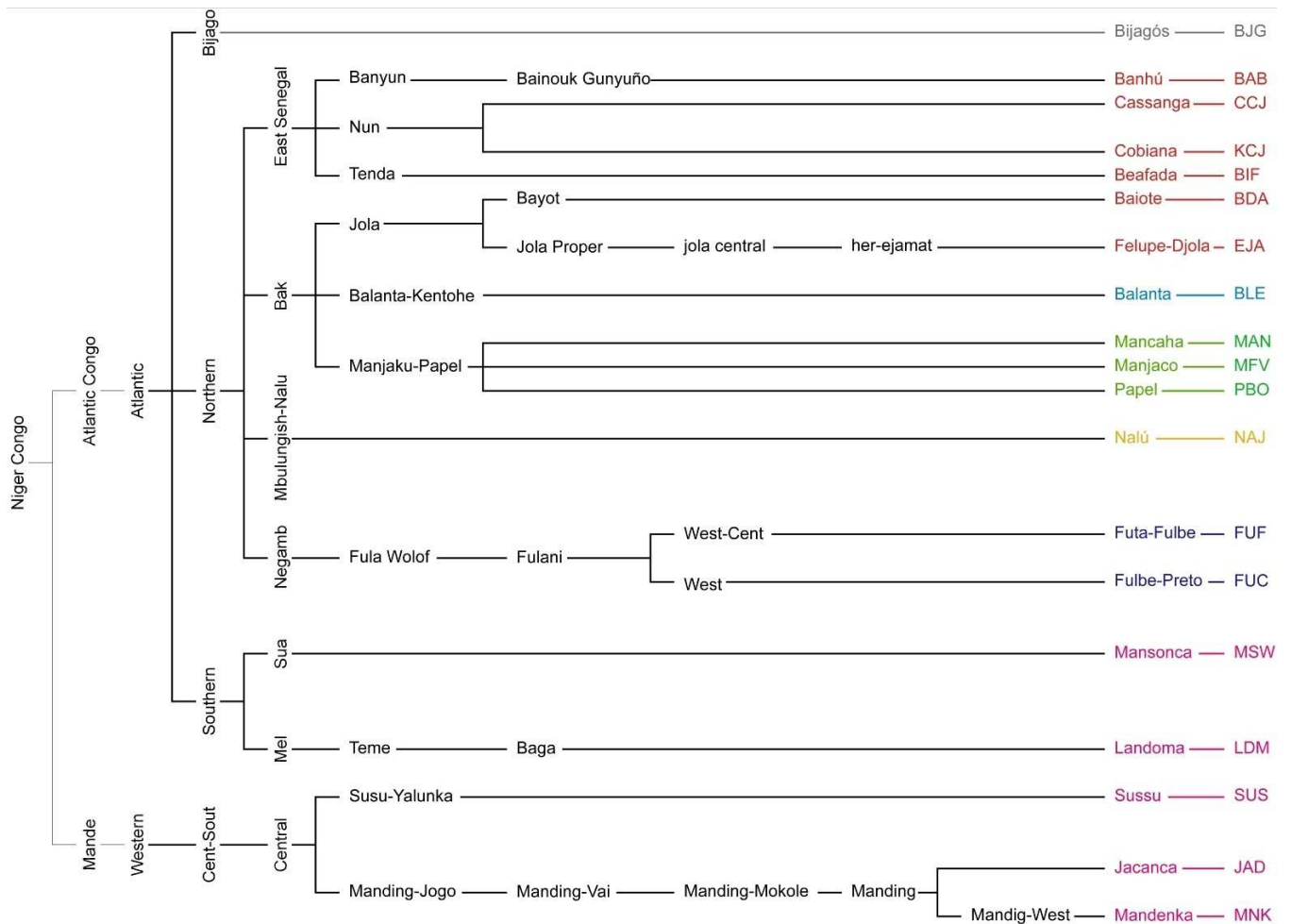


Figure 19 – Linguistic affiliation of Guinea-Bissau ethnic groups. The subdivisions do not correspond to a precise tree of languages; the linguistic families are according to information in Gordon and Raymond <sup>(2005)</sup>.

#### 4.5 - Records of autosomal genetic systems

In a chapter dedicated to Africa, Cavalli-Sforza and colleagues <sup>(1994)</sup> gathered data on classical genetic markers of worldwide populations (see publication for details on populations, genetic loci and references). The phylogenetic NJ tree suggests that the most important genetic gradient in Africa is a North vs. South axis <sup>(Cavalli-Sforza et al. 1994)</sup>. The major cluster of sub-Saharan populations reveals as outliers the Pygmies, the Khoisanids, the Chadic Sara and the Senegalese Serer, Wolof and Peul. In a parallel Principal Component Analysis two major clusters of Central-South Bantu-Nilotic and West African populations are evidenced. The Bantu are clustered together while West Africans appear more dispersed. Surprisingly, the Senegalese Peul are outliers while the Nigeria Fulani, all Fulbe speakers,

join the West African cluster together with Mande. The major West African cluster in the tree may be the outcome of a single agricultural expansion earlier than the Bantu expansion.

Synthetic maps of several genes define a Sahelian strip and a trans-Saharan connection up to Egypt, reinforcing the hypothesis of a West African ancestry to all agricultural expansion in Central-South Africa ~5-4 kya (Cavalli-Sforza *et al.* 1994). To the authors West Africa seemed to be the first part of the continent to experience population increase related to farming. The genetic data support two independent populational expansions: one in West Senegal and one in Nigeria/Mali/Burkina-Faso, the latter giving rise to Mande. The supposedly pre-Bantu agricultural expansion in West Africa is reflected in particular pools of unexpected high or low gene frequencies: HLA-A\*28 and HLA-B\*7 have peak regions in Burkina-Faso, HLA-B\*17 shows high frequency in all sub-Saharanans while HLA-B\*35 highest proportion is found in West Africa; the PhosphoGlucoMutase PGM2\*1 has a suggestive distribution, of a western stream of the Bantu spread; the transferrin TF allele D is frequent among West Africa and Bantu western stream, and is of probable local origin (see references in Cavalli-Sforza *et al.* 1994).

Studies of other single genes reveal particularities by representing signatures of farming-imposed diseases, mostly infectious ones due to sedentism and animal hosts (zoonoses, Cockburn 1971) and transmission by vectors attracted by waste (e.g. malaria and cholera). The emergence of *Plasmodium falciparum* as a major human pathogen is actually coincident with the beginning of agriculture (Coluzzi *et al.* 2002). Malaria is then the strongest known force for evolutionary selection in the recent history of the human genome that induces a remarkable wide range of erythrocyte variants (Kwiatkowski 2005). Favorable forms of genes under natural selection to malaria were shown to be peaking in West and tropical Africa: ACP1-B, hemoglobin C (Agarwal *et al.* 2000), FY\*O allele of the Duffy blood-group system (Miller *et al.* 1976, Hamblin and Di Rienzo 2000, Hamblin *et al.* 2002), G6PD-A (Tishkoff *et al.* 2001, Sabeti *et al.* 2002). Their amplified frequencies in regions where *Plasmodium vivax* and *Plasmodium falciparum* are nowadays absent is indicative of an ancient presence, as the malarial vectors evolve in agricultural centers but humans were selected for not to be intermediate hosts of the sexual cycle of the parasite. Kwiatkowski (2005) and colleagues believe that there are even ethnic differences in susceptibility to the disease, with different people developing different genetic variation for protection. Studies in Fulani in Burkina Faso (Modiano *et al.* 1996) and Mali (Dolo *et al.* 2005) have documented lower prevalence of the parasitemia and fewer clinical attacks in these people than in the neighbor groups. The differences are primarily genetic, with their higher levels of antimalarial antibodies (Modiano *et al.* 1998, 1999) and fewer protective globin variants (Modiano *et al.* 2001).

It is expected that natural selection has had a role in determining the frequencies of “lactase persistence” (the ability to digest milk in the adulthood), specially in the descendants of those populations that have traditionally practiced cattle domestication. Such lactose tolerance has shown to be low in African non-pastoralists (~5-20%) but common in pastoralists as the Tutsi and the Fulani (~90% and ~50%, respectively; Durham 1992, Swallow 2003). A recent study of Tishkoff *et al.* (2007) found a positive phenotypic association to three SNPs in LCT gene of East and South Africans.



## Chapter Three

### Aims of the study

With the present study we intend to:

- 1) Improve the knowledge about mtDNA and Y-chromosome haplogroup variation in the present-day Guinea-Bissau ethnic groups. Many genetic studies have been biased towards economically more advanced African countries that have their own research and medical centers, while populations from politically unstable regions remain under sampled. By filling the gaps in the datasets, we look forward to contribute for deeper and integrative phylogenetic studies;
- 2) Analyse the inter-ethnic genetic variation to learn more about the genetic relation within populations, in the West African landscape. When surveying African populations it makes particular sense to evaluate the genetic pool in ethnic populational units on the light of their social, linguistic and religious constraints;
- 3) Unveil traces of the colonizing genetic variation and the reshaping effects of the “Last Maximum Aridity” climatic period and the advent of agriculture, this last on the putative account of a large West African agricultural centre;
- 4) Apply a phylogeographic approach to reconstruct long distance gene flows in space in time, for to evaluate the culturally proposed origins in East African progenitors and North Africa trans-Saharan migrants;

In addition we aimed to investigate whether we can identify gender differences in the migrational inputs and the social patterns of admixture, with particular attention to the phenomena of “Balantization” and “Sudanization”.



## Chapter Four

### Material and Methods

#### 5 – Sampling procedure

The sampling procedure took place in Guinea-Bissau military camps and villages throughout the country, with the permission of the Chairman of the Joint Chiefs of Staff and the field intervention and support of the Guinea-Bissau Ministry of Health. A detailed explanation of the project, followed by a brief and individual interview, aimed the informed consent of the participants and the ascertainment of their ethnic background. The classification on ethnicities was based on information on both parental sides, whose ancestors were known to belong exclusively to a specific ethnic group for the last three generations.

A total of 372 unrelated male individuals of Guinea-Bissau constitute the sample group for the present study. Their distribution by ethnic group is as follows: 62 Balanta (BLE), 6 Baiote (BDA), 1 Banhú (BAB), 19 Beafada (BIF), 22 Bijagó (BJG), 8 Brame (BRA), 6 Cassanga (CCJ), 18 Djola (EJA), 38 Fula (FUL), 19 Fula-Preto/Forro (FUC), 19 Futa-Fula (FUF), 1 Fula-Toranca (FUT), 1 Jancanca (JAD), 1 Landoma (LAN), 19 Mancanha (MAN), 30 Mandenka (MNK), 27 Manjaco (MFV), 18 Mansonca (MSW), 26 Nalú (NAJ), 23 Papel (PBO) and 8 Sussu (SUD). All samples were typed for the mitochondrial lineages (see Table S1). A subset of 282 individuals, representative of the major groups was used for Y chromosome SNP analysis. Of those, 215 were analyzed for the Y-STRs (see Table S9).

Five ml of blood were collected by venipuncture and conserved in EDTA (K<sub>2</sub>) vacutainer tubes (Venoject II - Terumo®) at 4°C, until shipped to the Human Genetics Laboratory (LGH) - University of Madeira for further analysis. The chemical components of the tubes avoid the coagulation and assured the conditions for no degradation of nucleic acids.

#### 6 – DNA typing

##### 6.1 - DNA extraction

At the facilities of the LGH, blood samples were separated into cellular fractions by low-speed centrifugation (1000 and 3000 rpm) using a Biofuge 13 centrifuge (Heraeus Instruments). DNA extraction was done from the leukocitary fraction of whole blood using Chelex®-100 resin extraction (Lareu *et al.* 1994). The procedure aims to remove non-wanted cellular components, by alternating steps of temperature change and centrifuging. The cellular lysis is promoted by thermal shock, while the resin

binds to DNA-inhibitors (metal ions as haemoglobin iron, catalases, among others; Walsh *et al.* 1991), stabilized by the addition of a buffer. The 200µl aliquots were conserved at -20°C, for later use. By measuring the absorbance at 260nm in GeneQuant II (Pharmacia Biotech) the amount of extracted DNA was estimated to be of 90-100 ng/µl.

## 6.2 - PCR amplification

The Polymerase Chain Reaction (PCR; Saiki *et al.* 1988) is an *in vitro* procedure where the amount of segment of interest of DNA is exponentially increased by the action of a DNA polymerase, given that the *in vivo* conditions of replication are simulated (Mullis *et al.* 1992). The enzyme promotes the synthesis of a DNA strand using the complementary strand as a template.

The mtDNA hypervariable segment I (HVS-I) of the control region was amplified by PCR using primer pair F15907 (light-strand np 15907-15928) and R16547 (heavy-strand np 16525-16547; designed in EBC). The reaction mix consisted of 1X reaction buffer (75 mM Tris-HCl pH 8.8 at 25°C, 20mM (NH)<sub>2</sub>SO<sub>4</sub>, 0.01% Triton X-100, 0.5% Ficoll 400, 1mM Tartrazine), 2.5mM MgCl<sub>2</sub> (Solis Biodyne), 0.2pmol each primer (DNA Technology A/S and EuroGentec), 0.1mM each dNTP (Promega), 0.75U FIREPol Taq DNA Polymerase (Solis Biodyne) and 2 to 3µl (~100ng) of extracted DNA. The PCRs were carried out in a "Biometra UNO II" thermocycler with the following temperature profile: initial denaturation step of 94°C for 3 min, 36 cycles of amplification with denaturation at 94°C for 15 s, annealing at 52°C for 20 s, extension at 72°C for 50 s, and a final extension of 72°C for 5 min.

Primers and PCR conditions for HVS-II and coding region sites (to be determined by RFLPs or sequencing) are also described in Appendix 1, with extension time adjusted to the fragment size. The reagents on the PCR mixes followed the proportions described above. A few samples were selected for full sequencing of the 16.6kb molecule, with the aim of clarifying their phylogeny, but the typing is still ongoing. The primers for those are described in Rieder *et al.* (1998) to which temperature profile had to be optimized.

The typing of Y chromosome made use of NRY SNP-markers, and their hierarchical classification in the YCC phylogeny (YCC 2002). The polymorphisms surveyed were: YAP (Hammer and Horai 1995), 92R7 (Mathias *et al.* 1994), SRY4064, SRY10831 (Whitfield *et al.* 1995), PN2 (Hammer 1995), P25 (Hammer *et al.* 2000), M40 (Hammer *et al.* 1998), M2, M9, M10, M13, M14, M31, M32, M33, M35, M44, M60, M75, M78, M81, M89, M91, M116, M123, M130, M155, M168, M173, M174 and M191 (Underhill *et al.* 2000, 2001a). The reaction mix included the same reagents in the same proportions as for mtDNA PCR reaction. The temperature variation occurred in a "Biometra UNO II" thermocycler, with different profiles according to the Y chromosome SNP markers (see Appendix 2).

As for the Y-STRs, a multiplex reaction for eleven markers (DYS19, DYS389I-II, DYS390, DYS391, DYS392, DYS393, DYS385, DYS437, DYS438 and DYS439; Appendix 3) was carried out

with Powerplex® Y-System (Promega) as described in the manufacturer's instructions. A mix of nuclease-free water, Gold ST★R 1X Buffer (50mM KCl, 10mM Tris-HCl pH 8.3 – 25°C, 1.5mM MgCl<sub>2</sub>, 0.1% Triton® X-100, 0.2mM each dNTP, 0.16mg/ml BSA), PowerPlex® Y 1X Primer Pair Mix and AmpliTaq Gold® DNA polymerase 2.75U was made, and that volume added to 2.5µl (~100ng) of template DNA. The temperature profile was as follows: 95°C 1 min - 96°C 2 min (94°C 1 min - 60°C 1 min - 70°C 1.5 min) for 10 cycles, (90°C 1 min – 58°C 1 min – 70 °C 1.5 min) for 22 cycles, and final extension of 60°C 30 min. For some samples (haplotypes H168 to H220, Table S9), markers were typed individually with published primers and conditions: DYS19, DYS389I/II, DYS390, DYS391, DYS392 and DYS393 (Kayser *et al.* 1997), DYS385 (Schneider *et al.* 1998) and DYS439 (Ayub *et al.* 2000; further details in Appendix 3). Additional GATA STR A7.1 (DYS460, White *et al.* 1999) was surveyed in E3b1 Y chromosomes.

### 6.3 – Electrophoresis on agarose and polyacrylamide gels

The assumption of electrophoretic analysis relies on the migration of a molecule when a voltage gradient is applied, given its electric properties. In the case of DNA the phosphate group is negatively charged and thus makes the molecule to be attracted to the cathode. The progression state is inversely proportional to the fragment size.

The quality of the amplified fragments was visually estimated in 2% agarose gels with ethidium bromide staining (0.5 µg/ml) in a UV-Transilluminator and subsequently photographed (Grab-IT Annotating Grabber 2.53, UVP). Due to the higher level of accuracy in separating DNA fragments by their size, 9% polyacrylamide gels with silver staining (Luis and Caeiro 1995) were used to access the RFLP state of markers indicated in Appendixes 3 and 4. The electrophoretic runs varied between 30 min to 1h 30 min, and 80-100V in agarose and 180-200V for acrylamide gels. For the case of Y chromosome YAP marker, referring to an insertion polymorphism, the molecular allele could be directly asferred from the gel.

### 6.4 – Purification of PCR products

The amplified fragments to be sequenced were submitted to a purifying treatment with 1U Shrimp Alkaline Phosphatase (SAP) and 1U Exonuclease I (ExoI, Werle *et al.* 1994), following the conditions indicated by the manufacturer (Tested User Friendly TM USB; 37°C 20 min, 80°C 15 min). This step intends to eliminate from the mix components that may interfere in the next procedures, namely non-bound dNTPs and oligonucleotidic primers.

## 6.5 – Automatic sequencing

Depending on the quality of the PCR product, a suitable amount varying from 2 to 5µl was used as template for the sequencing reaction, with the DYEnamic ET\* Terminator Cycle Sequencing Kit (Amersham Biosciences). For mtDNA HVS-I, forward and/or reverse primers were used (nps 15975 and 16494 np, respectively, 5pM). The manufacturer conditions were followed: 96°C 4 min – (96°C 10 s – 50°C 5 s – 60°C 4 s) x 30 cycles – 60°C 10 min. For all sequences that presented a homopolymeric cytosine stretch at HVS-I nps 16184-16193, an additional reaction was performed on both strands. The typing of mtDNA coding region sites was done using the primers described in Appendix 4. For the Y-SNPs the strand to be sequenced was selected according to the positionment of the primer relative to the mutation (see Appendix 5).

An ethanol precipitation protocol with sodium acetate/EDTA buffer, dextran and ethanol on various concentrations (90 and 70% v/v; Sambrook *et al.* 1989), aimed a subsequent purification of the product of the sequencing reaction. Alternatively, GFX purification columns (Pharmacia) with Sephadex G-50 (Amersham Biosciences) were used for cleansing. After addition of formamide, the samples were loaded into a polymeric gel and the capillary electrophoresis was held in automatic sequencers ABI PRISM™ 310 Genetic Analyser (Applied Biosystems) and MegaBACE 1000 (Amersham Biosciences).

The sequences on both control and coding regions were aligned with the Cambridge Reference Sequence (CRS; Anderson *et al.* 1981, revised in Andrews *et al.* 1999), using the software Wisconsin Package Version 10.0 (Genetics Computer Group (GCG) 2005). HVS-I nps 16024-16365 was unambiguously accessed for the totality of the samples, with mutation recorded by their positions in the CRS minus 16000 bp. When informative, HVS-II nps 150-330 were sequenced (for haplotypes in Table S1a). Y chromosome biallelic polymorphisms not accessible by means of enzymatic restriction were determined by sequencing (as indicated in Appendix 2) and then compared to the ancestral/derived allelic state (YCC 2002).

## 6.6 – Typing of Y chromosome microsatellites

The genotyping of Y-STRs was carried out in ABI PRISM™ 310 Genetic Analyser along with Genescan 2.1 analysis software (Applied Biosystems). The PCR products were directly loaded into the gel, together with formamide and ILS600 size standard. The fluorescent labels allowed accessing the relative size of the fragment, allowing the assignment of alleles (Appendix 3). The typing followed the International Society of Forensic Genetics - ISFG guidelines for Y-STR analysis (Gill *et al.* 2001).

## 6.7 - Typing of Restriction Fragment Length Polymorphisms - RFLPs

Most of the mtDNA HVS-I sequence data had a motif that allowed the unequivocal assignment of the samples to a specific haplogroup. However, when the data led to inconsistent definition of the haplogroup and/or its subclusters, the putative members were screened for further diagnostic markers in the coding region, mostly by RFLP assays with the appropriate restriction enzymes. In such cases, the nucleotidic substitution determines the loss or gain of a restriction site, by altering the palindromic sequence recognized by the enzyme.

The restriction of amplified products was done according to the manufacturer's instructions (Fermentas and New England BioLabs). The fragments were resolved in agarose and acrilamide gels, with different polymeric percentages depending on their size. The following mtDNA polymorphic restriction sites were screened: 323 *HaeIII*, 1715 *DdeI*, 2348 *MboI*, 2759 *RsaI*, 3594 *HpaI*, 3693 *MboI*, 4158 *AluI*, 4686 *AluI*, 5585 *AluI*, 5656 *NheI*, 7056 *AluI*, 8615 *MboI*, 10084 *TaqI*, 10321 *AluI*, 10394 *DdeI*, 10398 *AluI*, 10806 *Hinfl*, 11438 *MboI*, 11641 *HaeIII*, 12308 *Hinfl*, 13804 *HaeIII*, 13958 *HaeIII*, 14766 *MseI* and 14868 *MboI* (further details on Appendixes 1 and 4). The state of the NRY markers 92R7, M2, M9, M13, M31, M33, M35, M40, M81, M130, M168 and M174 was assigned by RFLP analysis with appropriate restriction endonucleases (Appendixes 2 and 5).

## 7 - Data analysis

### 7.1 - Phylogenetic assignment

All mtDNA sequences were first classified on the basis of HVS-I motifs relative to rCRS. Whenever necessary HVS-II typing or RFLP data were added for clarifying the assignment. The haplogroup classification was based on the phylogenetic analyses and nomenclature of African and European mtDNAs as in Chen *et al.* <sup>(1995b, 2000)</sup>, Torroni *et al.* <sup>(1997, 2001a)</sup>, Watson *et al.* <sup>(1997)</sup>, Rando *et al.* <sup>(1998, 1999)</sup>, Macaulay *et al.* <sup>(1999b)</sup>, Quintana-Murci *et al.* <sup>(1999)</sup>, Alves-Silva *et al.* <sup>(2000)</sup>, Richards *et al.* <sup>(2000)</sup>, Bandelt *et al.* <sup>(2001)</sup>, Pereira *et al.* <sup>(2001b)</sup>, Richards and Macaulay <sup>(2001)</sup>, Salas *et al.* <sup>(2002)</sup>, Mishmar *et al.* <sup>(2003)</sup>, Rosa *et al.* <sup>(2004)</sup>, Bandelt *et al.* <sup>(2006)</sup>, Gonder *et al.* <sup>(2006)</sup> and Kivisild *et al.* <sup>(2006b)</sup>. Regarding the Y chromosome system, the nomenclature and phylogenetic relationships of SNP-lineages followed the recommendations of YCC <sup>(2002)</sup>. The Y-STRs were designated according to the number of repeated units (as proposed by Kayser *et al.* 1997 and de Knijff *et al.* 1997, with the exception of locus DYS389 as in Roewer *et al.* 2000).

## 7.2 - Definition of populational units

The populational units of many genetic studies are usually identified by their country of residence or linguistic group. We believe however that the genetic basis was defined long before the present political boundaries. Therefore, in the present work we chose to define a prime role to ethnicities, which is in much associated to linguistics and dictates the social pattern of miscegenation.

In order to have a dealable number of units with reasonable size, some ethnic groups were clustered when anthropological and linguistics affinities allowed it (following Almeida 1939, 1964; Barros 1947, Carreira 1962, 1983; Carreira and Quintino 1964; Hair 1967; Quintino 1967, 1969; Diallo 1972 and Lopes 1999). Seven units were formed, as shown in the Results and Discussion section. Some groups were left unpooled: the Balanta for whom a Sudanese origin has been suggested and the Bijagós mainly due to their particular geographical location. Data on several other African populations were taken from the literature for comparison with the maternal and paternal Guinean lineages (Tables S2, S3, S10 and S11). This selection was based on the level of resolution, if possible including as many typed markers as in our analysis.

## 7.3 - Phylogenetic networks

Extensively described in section 1.1, networks of mtDNA lineages were built by hand, checked with the software Network 4.1.1.2 (Fluxus Technology Ltd.), with a combined setting of RM and NJ algorithms (Bandelt *et al.* 1995, 2000) and then drawn in NetViz 6.5 (NetViz Corporation). To note that networks were first constructed for each haplogroup separately and then combined to show the overall topology. Phylogenetic relationships departed from the most parsimonious HVS-I nucleotide variation, assigning higher priority to the coding-region information. Reticulations were resolved based on parsimony and frequency criteria, by successive cleansing of uninformative sites, and subsequently collapsing and decomposing steps (as described earlier in Bandelt *et al.* 1995). Further resolution was achieved by assigning information on the mutability of the different sites (Hasegawa *et al.* 1998, Richards *et al.* 1998b), transversions and indels requiring further processing: HVS-I nucleotidic positions 16129, 16189, 16223, 16278, 16294, 16311 and 16362 were down weighted in the general analysis (in particular nps 16093 and 16274 for L1c and nps 16093, 16230, 16274, 16292, 16309, 16319 and 16355 for L2a), in order to solve reticulations. Transversions from A to C at np 16183 were disregarded. Although the built phylogenies had no outgroup for rooting, their construction was based in the structure of published networks in relation to non-human primate mtDNA sequences (e.g. Mishmar *et al.* 2003, Kivisild *et al.* 2006b).

The network of biallelic polymorphisms in NRY followed the hierarchical topology of YCC 2002. For the Y chromosome microsatellites, the intra-haplogroup variation was best represented by



networks of microsatellite haplotypes, built with Network 4.1.1.1 (Fluxus Engineering). Reduce median and median joining algorithms were sequentially applied to 7 loci haplotypes (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392 and DYS393) as described in Bandelt *et al.* <sup>(1995, 1999)</sup>, to achieve the most parsimonious topology. Singletons were excluded from the analysis of E3a\*-M2 and the threshold level of 2 was set. STR weighting was done according to Helgason *et al.* <sup>(2000)</sup>.

#### 7.4 - Coalescence time estimates

The networks in section 8 constitute the raw material for the coalescence estimates of mtDNA haplogroups. Therefore, the accuracy of their topology (adequately reconstructing the founder lineages and pathways of evolution) defines the reliability of the dating. The coalescence ages were calculated within several monophyletic clades as described in Forster *et al.* <sup>(1996)</sup> and Saillard *et al.* <sup>(2000)</sup>. In a genealogy of  $n$  sequences with a specific root and  $k$  links, the  $l_i$  number of observed mutations along the  $i$ th link is taken as an age estimator. The links are scaled to time and each interior node corresponds to a coalescence event. The average distance to the likely founder haplotype is for that represented as  $\rho = (n_1l_1 + n_2l_2 + \dots + n_kl_k)/n$ . The conversion of  $\rho$  into absolute time assumes that a transition within HVS-I 16090-16365 np corresponds to 20180 years, with a generation time of 25 years <sup>(Forster *et al.* 1996, 2001)</sup>. Transitions, transversions or indels outside this frame and length polymorphisms in the C-run (np 16183-16194) were disregarded. Saillard *et al.* <sup>(2000)</sup> denotes  $\sigma^2$  standard deviation index of the phylogeny-based  $\rho$  ( $\sigma^2 = (n_1^2l_1 + n_2^2l_2 + \dots + n_m^2l_m)/n^2$ ;  $n_i$  and  $l_i$  are respectively the number of individuals and the number of observed mutations along the  $i$ th link). The coalescence time is given by  $\rho \pm \sigma$ , being  $\sigma$  of relevant importance in cases where large time intervals do not allow a discrimination of temporal events.

The coalescence age of Y chromosome haplogroups was estimated from the proportion of microsatellite variability within each clade. The TMRCA of haplogroups A1, E1\*, E3a\*, E3b1 and E3b2 was estimated under a stepwise mutation model, by calculating the average squared distance (ASD), a measure linearly related to coalescence time <sup>(Slatkin 1995, Goldstein *et al.* 1995)</sup>. For a set of 10 Y-STRs (the included in the multiplex kit, except DYS385), the squared difference of allele lengths for each microsatellite was determined between each individual haplotype and the one assumed to be the modal haplotype. To note that modal haplotype is built up with the most frequent allele of each microsatellite. For Y chromosomes belonging to the same haplogroup, the mean values were then averaged over the loci and divided by effective mutation rate ( $6.9 \pm 1.3 \times 10^{-4}$  mutations per locus per generation of 25 years <sup>(Zivotovsky *et al.* 2004)</sup>). Confidence intervals were calculated according to the methodology outlined in Thomas *et al.* <sup>(1998)</sup>.

## 7.5 – Haplotype exact matches

A match was determined by the number of times similar haplotypes occurred in datasets of unrelated individuals. A compilation of more than 20,000 HVS-I sequences done by Professor R. Villems's working team at the Estonian Biocentre was used for the search of individuals harbouring the same mtDNA HVS-I profile as the Guinean haplotypes. It contains a worldwide random collection of unrelated individuals on various ethnic backgrounds, taken from the available scientific literature and unpublished works. The populations included in public databases were not of interest for the purpose of our analysis. Similarly, Y-STRs haplotypes were surveyed for exact matches in YHRD ([www.yhrd.org](http://www.yhrd.org), <sup>Willuweit and Roewer 2007</sup>), a curated and frequently updated database containing more than 51,253 haplotypes in 447 worldwide populations (as of July 2007), plus additional searches in published literature. The allelic state of both 8 and 10 STRs (respectively “minimal” and “extended” haplotypes) were surveyed for exact matches.

## 7.6 - Statistical parameters

By the use of statistical tests the data is summarized in a standardized way which facilitates the interpretation. Inevitably some information is lost, but it is among the commonest ways of comparing both populations and loci. Data needs to be handled with care to avoid the drawing of unwarranted conclusions, in what concerns exploratory methods. As observed data have usually some deviations from the expected one, either due to chance fluctuation or wrong hypothesis, a bias ascertainment is most of the times needed. The size of the deviations defines then the point where to start doubting the hypothesis. For testing reliabilities, probabilities  $P$  play a major role by establishing the 0.05 limit for mere chances.

Although not without controversy, the parameters described below better represent the genetic systems and have proved to be more informative. Most of the statistical analysis was performed with the software Arlequin 2.000 <sup>(Schneider *et al.* 2000)</sup>.

### 7.6.1 - Frequency calculation

Frequency is estimated by the  $x_i$  observed times of  $i$  alleles/haplotypes in a sample of  $n$  gene copies. Not normally distributed, the frequency depends on the sample size and is associated to a certain group of people at a defined time. Methods based on frequency comparisons are built up after a matrix of correlation inferred with binomial variances and covariances for alleles of the same gene, and setting to zero all correlations between alleles of different unlinked genes. For the populations to

use in comparisons the average allele/haplotype/haplogroup frequency within the same geographic location was weighted by sample size for a cluster of  $s$  populations.

### 7.6.2 - Genetic diversity

Nei's gene diversity  $H$  (also called as heterozygosity for diploid loci; <sup>Nei 1987</sup>) defines the probability of two sequences, chosen at random from a population, being distinct. The index and its associated variance were similarly calculated for different levels from mitochondrial DNA sequence and Y-STR alleles to haplogroup data on both mtDNA and Y chromosome systems as follows:

$$\hat{H} = \frac{n}{n-1} \left(1 - \sum_{i=1}^k p_i^2\right)$$

$$V(\hat{H}) = \frac{2}{n(n-1)} \left\{ 2(n-2) \left[ \sum_{i=1}^k p_i^3 - \left(\sum_{i=1}^k p_i^2\right)^2 \right] + \sum_{i=1}^k p_i^2 - \left(\sum_{i=1}^k p_i^2\right)^2 \right\}$$

$H$  varies from zero to one, reaching the last when there is a high degree of polymorphism and almost all types are different from each other. Under neutral evolution, the level of diversity can reach mutation-drift equilibrium, where new alleles formed by mutation are balanced by the ones eliminated by drift. If selection is absent the  $H$  parameter becomes equivalent to the mean nucleotide diversity (Tajima 1983, Nei 1987).

### 7.6.3 – $F_{ST}$ statistics

The underlying molecular basis of the polymorphisms makes possible to determine evolutionary distances based on the number of differences. The simplest genetic distance between two populations for one biallelic gene would be the difference of the allelic frequencies  $p_i$  and  $p_j$  (<sup>Nei 1987</sup>). In this study the evolutionary relatedness of population units, in terms of their genetic structure, was measured with  $F_{ST}$  statistic, pairwise comparisons also known as coancestry coefficients. The genetic distance evaluates the average gene frequency and its associated variance and may incorporate the molecular distances among haplogroups (matrix of defining mutations; Slatkin's linearized  $F_{ST}$ ). The  $F_{ST}$  index varies from zero to one, where larger values indicate greater evolutionary distance and lower ones suggest lower distance (possibly with considerable gene flow between populational units). Therefore, an obtained value of 0.3 means 30% of the allelic frequency is attributed to the interpopulational differences, while the remaining variation is found within the population. Again, the null hypothesis is of no difference between the units. The distribution of pairwise

$F_{ST}$  is obtained by permutation tests and  $P$  significance level is the proportion of permutations with a larger or equal value of the observed  $F_{ST}$  ( $P < 0.05$  reject the hypothesis).

For the pairwise comparison, either of populations or allele-by-allele locus analysis, the  $F_{ST}$  index equals distance values, though a linearization to divergence time has to be performed (Slatkin 1995). Reynolds and collaborators concluded that  $F_{ST}$  is more satisfactory than Nei's distance when only a few new mutations emerge in the evolutionary time period examined, as it is the situation for the evolution of modern humans (Reynolds *et al.* 1983).

#### 7.6.4 - Exact test of population differentiation

Population haplogroup frequencies were compared in permutating pairs using the exact tests of population differentiation, to test the null hypothesis of identical haplogroup frequencies in the units under comparison. A random distribution of  $k$  different haplotypes/haplogroups frequencies among  $r$  populations is tested as in Raymond and Rousset (1995), in a Fisher's-type  $r \times k$  contingency table.  $P$ -values based on 10,000 Markov steps are reported. The Markov chain sequential steps explore all potential states of the observed table and estimate the probability of obtaining a less or equally likely table, under the null hypothesis of panmixia.

#### 7.6.5 – Graphical display of results – PCA

The genetic information can be displayed in a more comprehensible manner by graphical means. The Principal Component Analysis is a dimension-reduction method which seeks to explain the variance of multivariate data by a smaller number of variables - the principal components (PCs). Individual axes are sequentially extracted and associated to a percentage of variation, as linear functions of the original measurement data, in this case built from the haplogroup frequencies. After reducing the variables to PCs the coordinates are plotted in a dimensional graphic representing the genetic landscape. For  $n$  populations,  $n-1$  dimensions are required to fully represent their pairwise distances with a reduced loss of information. Usually the two-dimensional Cartesian diagram exhibits a good display of the relative genetic similarities, satisfactory when the percentages of retained variance are around 60-75%. The addition of a third PC is graphically possible but never as clear as the two-dimensional map. The PCAs tend to show the same inferences as the clustering trees though cannot be taken as independent methods of analysis.

In the present work bi-axial PCAs were constructed from the relative frequency vectors of mtDNA and Y chromosome haplogroup composition with the software MVSP Version 3.13m (Kovach

Computing Services), for both Guinea-Bissau ethnic clusters and to compare our Guinean sample with other African data.

#### 7.6.6 - Analysis of Molecular Variance (AMOVA)

In the AMOVA test a genetic structure of groups of populations is tested by an analysis of gene frequency variance (Excoffier *et al.* 1992), that takes into account the molecular relationship of the alleles/haplotypes (Long 1986). The total variance is hierarchically partitioned into  $\sigma$  covariance components due to intergroup  $\sigma^2_a$ , inter-population  $\sigma^2_b$  and intra-population  $\sigma^2_c$  differences (Weir 1996, Excoffier 2000). An algorithm leads to a fixation index  $F_{ST}$ , identical to F-statistics weighted over loci (Weir and Cockerham 1984) and that can be seen as correlation coefficient. The significance of the fixation is tested using a non-parametric permutation approach described in Excoffier *et al.* (1992), consisting in permuting haplotypes, individuals or populations (in this study 1000 permutations were set). The expectation of the estimator is zero but can result, by chance, in slightly negative values due to the absence of genetic structure. A grouping strategy using geographic, linguistic and religious criteria was used for AMOVA tests of our data and other compiled from the literature.

#### 7.6.7 – Mismatch distribution

The mismatch distribution (MD) refers to the nucleotide pairwise differences between individuals (Tajima 1983, 1993), useful when dealing with discrete data as SNPs, RFLPs or STRs. Easily determined in mtDNA sequences, MD reflects the sequence diversity within the sample and its shape of distribution might reveal demographic episodes. It is usually multimodal in populations at demographic equilibrium, reflecting the stochastic shape of the network, but it can assume a unimodal bell-shaped curve in populations that have had recent populational expansion (Slatkin and Hudson 1991, Rogers and Harpending 1992). Demographic inferences from mismatch analysis have to consider time of expansion, size before expansion, gene flow and sub-structuring of populations. Deviations from unimodal shapes may reflect the significant role of one or more of those factors over a population (Marjoram and Donnelly 1994). We have determined the MD for each mitochondrial haplogroup. Irregularities of the mismatch distribution were tested by the significance of raggedness index (Harpending 1994).

#### 7.6.8 - Neutrality tests

The Tajima's test of selective neutrality is based on the infinite-site model of mutational occurrence without recombination, and it is to be applied in short DNA sequences or haplotypes, to

infer about natural selection. Two estimators of the mutation parameter are compared in a  $D$  statistics, relating nucleotide diversity  $\pi$  and the number of segregating sites  $S$  (Tajima 1989). The significance of the test is done by randomizing the samples under the null hypothesis of selective neutrality and population equilibrium. The  $P$  value is ascertained as the portion of random  $F$ -statistics less or equal to the observed. In the case of selectively neutral genetic systems, significant values of  $D$  can be generated by expansion, bottleneck or heterogeneity of mutation rates among DNA sequences (Tajima 1993, Tajima 1996). The Fu's  $F_S$  test captures other facets of the data, analyzing the null hypothesis of constant population size in equilibrium of drift and mutation (Fu 1997).

## Chapter Five

### Results and Discussion

#### 8 - MtDNA analysis in the population of Guinea-Bissau – a phylogenetic approach

The presence of the most prevalent sub-Saharan African mitochondrial haplogroups and sub-haplogroups within L0-L3 was revealed for Guinea-Bissau (93.8%) while the minor variants were distributed among non-L clusters, namely M1b1 (1.3%), U5b1b (2.7%) and U6a (2.2%). The profile of haplotypes for each ethnic group is presented in Supplementary Material – Table S1, while the frequency of the classes is shown in Table S3 and in Rosa *et al.* <sup>(2004)</sup> (Table 2). Due to the large size of the dataset, the referred article included only the skeletons of phylogenetic trees of the haplogroups. Here we show a phylogenetic network of clades defined by HVS-I and partial coding region RFLP data, together with the ethnic affiliation of the 372 individuals (Figure 20). A meta-analysis was performed with the purpose of integrating the findings, by comparing the populations used by Rosa *et al.* <sup>(2004)</sup> with meanwhile published individual studies (see populational units in Supplementary Material – Table S2 and Figure S1 - and respective haplogroup frequencies in Table S3). By choosing to deepen the phylogenetic analysis several points are of interest to outstand:

Haplogroup L0a1 is the only subclade of L0 in Guinea-Bissau, inserted in a context of West Africans L0a1 lineages with rather low frequencies (below 5%; Rando *et al.* 1998, Brehm *et al.* 2002, Rosa *et al.* 2004, Jackson *et al.* 2005, Cerny *et al.* 2006, Ely *et al.* 2006). The geographic range of L0a spans East Africa from Egypt to Kenya (3-19%, Table S3; Watson *et al.* 1997, Brandstatter *et al.* 2004b, Kivisild *et al.* 2004), and has further frequency peaks in the Bantu people of Cameroon <sup>(Destro-Bisol *et al.* 2004)</sup>, Central Africa Chadic-speakers (up to ~22% in Chadic Mafa, <sup>Cerny *et al.* 2004</sup>), and in Mozambicans (15-29%; <sup>Pereira *et al.* 2001b, Salas *et al.* 2002</sup>). It is interesting to note that in our dataset only the Balanta, the group that claims a Sudanese origin <sup>(Quintino 1969)</sup>, shows an increased frequency of this clade (11%). In case of L0a1, we are dealing with a rather ancient daughter group that has derived from L0a in Palaeolithic times ~33kya <sup>(Salas *et al.* 2002)</sup>. However, the relatively young coalescent age of L0a1 in Guineans (6.4±2.6 ky, Table S4) suggests that only a small subset of such variation reached Guinea-Bissau during the Holocene, probably at the post-LGAM with the more favorable conditions for migration. Their lineages are shared with other African regions but surprisingly only at the level of the founder type. Several exact matches of the Guinean GB4 HVS-I motif, detected in eight East African populations, Chad and West African Cape Verde, Senegal Mandenka, Sierra Leone and Mali (see Table S5) may represent a possible route of migration of

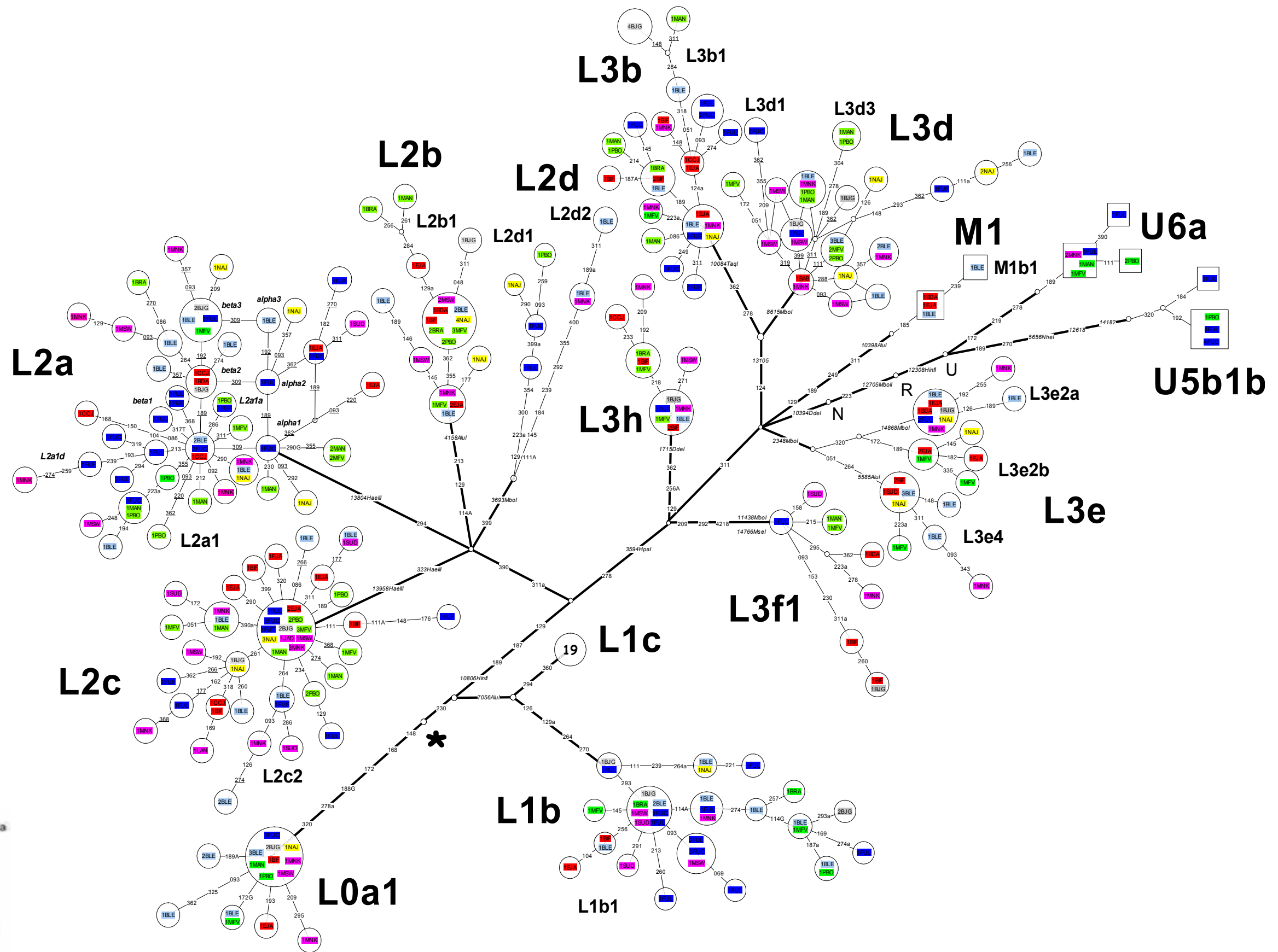


Figure 20 - MtDNA median network of Guinea-Bissau haplogroups based on HVS-I (minus 16000 bp) and partial coding-region RFLP information. Numbers along the links indicate transitions, transversions (with suffixes), state reversions (letter "a") or RFLP site changes (*in italic*) so that the rCRS derived states are indicated. Recurrent mutations relative to each haplogroup are underlined. Mutations are placed on a hierarchical level where the basal define haplogroups (bold links and letters) and the ones at the endings refer to individual haplotypes. Sub-clades are signalled with bold letters with lower size font. The codes and numbers in the circles represent respectively the ethnic group (as in section 5 and Table S2) and the number of individuals harbouring the particular haplotypes, with correspondent colour to the ethnic assignment. Length polymorphisms in the C-run (16184-16193) and 16183 A to C transversion were disregarded for the network construction. In the present network the following typographic errors in Rosa *et al.* (2004) were correct: in L1b 187A should be 187a; the L2a\* lineage 182-189-362 lacks 270; in L2a1, 317 should be 317T and the node 213-264A is indeed 213-294; in L2c the lineages 261-318 and 093-126-274 include two individuals; L2d basal sample is in fact included in L2d. L1c haplotypes are described in detail in Figure 22.



L0a1 from East to West Africa. The spread probably happened in a short timeframe since no further sharing was detected for the lineages radiating in several populations. In the context of mtDNA lineages we will refer to exact matches as lineages harboring the same HVS-I motif. Back to L0a1 variation, the accumulated diversity and associated coalescence time in Guinea-Bissau are most likely a reflection of the arrival of the eastern founders. Alternatively, we can hypothesize on a prior existence of L0a1 in West Africa, with limited spread and a reduction of diversity through a bottleneck, and later expansion in the Holocene. The first hypothesis of a more recent founder effect seems more likely, as otherwise common lineages would have been sampled in a wider geographic range. The doubt remains if the Guinea-Bissau L0a lineages represent a post-LGAM recovery of people or if it otherwise relates to the agriculture-promoted expansions. The lack of the L0a2 clade, associated with the 9bp deletion, characteristic of the widespread Bantu speaking populations <sup>(Soodyall *et al.* 1996)</sup> suggests that L0a has at least two distinct phylogeographic patterns in Central and West Africa.

The West African L1b is represented by its basal founder type, a deriving branch found in Balanta, Nalú and Fulbe, with HVS-I transitions 16111-16239 and state reversion at np 16264, and the L1b1 daughter-clade defined by additional HVS-I mutation in 16293 (Figure 20). The L1b1 sub-branch with mutations at nps 16114A and 16274 is common in Senegalese Mandenka and Wolof <sup>(Graven *et al.* 1995, Rando *et al.* 1998, Salas *et al.* 2002)</sup> and frequent in Balanta and Papel. The proposed origin of expansion of L1b in Central Africa ~30.5 kya <sup>(Salas *et al.* 2002)</sup> implies its westward spread to West Africa. Some scholars hypothesize on a later bottleneck and re-expansion to West Africa ~17 kya that has reshaped the earlier phylogenetic pattern <sup>(Chen *et al.* 1995b, Rando *et al.* 1998)</sup>, at a time coincident with the gradual return to wetter climatic conditions <sup>(Burke *et al.* 1971)</sup>. The coalescence estimate of L1b lineages in Guinea-Bissau (Table S4) corroborates for its early presence in West Africa, even prior to lineages present in Central Africans. On the other hand, if we pay attention to the L1b1 founder node - first central one excluding the lineages with np 16114A-16274 (Table S4, TMRCA~16ky) - this could have been among the movements triggered by the climatic improvement. The suggestion is bolstered by that the few matches of the basal L1b variation are restricted to West Africa (GB11 and GB24, the latter found in Fulbe) and identical haplotypes in cluster 16114A-16274 are found mostly in Senegalese Mandenka, while L1b1 haplotypes lacking motif 16114A-16274 match in both Central and West Africa (Table S5). The proportion of L1b lineages among Fulbe is not surprising given their high prevalence in other Fulani

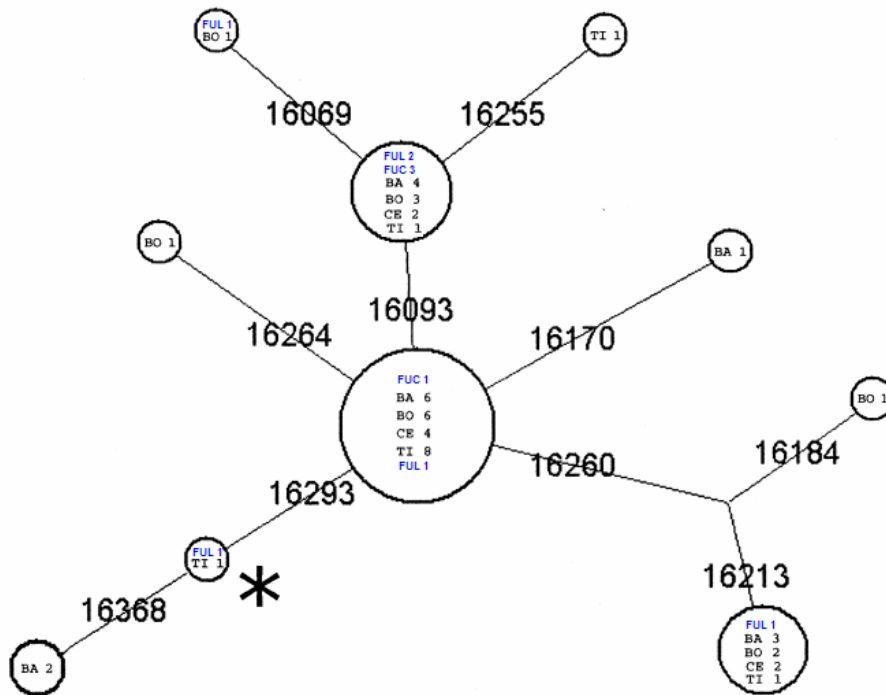


Figure 21 - Network of the L1b haplotypes in Fulani people of West and Central Africa. The \* denotes the MRCA of the variation defined by mutations at nps 16126, 16187, 16189, 16223, 16264, 16270, 16278, and 16311. The abbreviations are as follows: BA - Fulani from Banfora; BO - Fulani from Bongor; CE - Fulani from Tcheboua; TI - Fulani from Tindangou; FUL and FUC – Fulani from Guinea-Bissau. The number following the abbreviations corresponds to the number of cases. Circle sizes are proportional to the haplotype frequency. Adapted from Cerny *et al.* (2006).

populations (Watson *et al.* 1996, Cerny *et al.* 2006). It is here interesting to note the similarities displayed by the Fulani people living in different territories in West and Central Africa (Figure 21). Three of the L1b haplotypes were found to be exclusive of Fulani (GB7, GB8 and GB20; Table S5). Most of the matches have a West African distribution: Senegalese Mandenka registered the highest number of lineages in common with Balanta (GB13, GB14 and GB21); Bijagós match one Wolof type (GB17); Sierra Leone people share one lineage with Felupe-Djola and Papel (GB21). The more basal GB23 and GB24 lineages are pan-African and also of non-African distribution, and represent central nodes of regional diversity.

Haplogroup L1c is among the mitochondrial lineages that can make use of larger and tailor-made datasets and increased level of resolution to shed light on its origin and processes underlying its present distribution. This clade is characterized by a high internal diversity, and since its first description by Rando *et al.* (1998) has been targeted with substantial revision of its phylogeny, based on control-region (Salas *et al.* 2002, Destro-Bisol *et al.* 2004, Batini *et al.* 2007) or coding-region analysis (Gonder *et al.* 2006, Gonzalez *et al.* 2006, Kivisild *et al.* 2006b). The

network in Rosa *et al.* (2004) was further resolved using additional information on sites of phylogenetic importance (see Figure 22). The classification of L1c and its subclades L1c1, L1c1a and L1c3 as in Salas *et al.* (2002), and L1c3a1 and L1c3b1 as in González *et al.* (2006). Our phylogeny is consistent with basal branches determined by full-sequencing studies of the molecule, except for Guinea-Bissau L1c1\* and L1c1a mtDNAs which harbor the substitution at np 14148 but lack the one at np 11899 (Bandelt *et al.* 2006, Gonder *et al.* 2006, Kivisild *et al.* 2006b). The available data does not consistently elucidate the phylogenetics, but the subject surely deserves further attention, for instance in a particular substitution at np 12879, to our knowledge not described to date. For details on complete sequences of 2 L1c Guinean mtDNAs see forthcoming work of Behar MD, Villems R, *et al.* (ms submitted).

The highest frequencies of mtDNA haplogroup L1c are among Cameroon, Central African Republic and Republic of Congo people (up to 96% in Mbenzele Pygmies; see Table S3; Cerny *et al.* 2004, Destro-Bisol *et al.* 2004, Batini *et al.* 2007) whereas is less frequent in other quadrants of the continent (Watson *et al.* 1997, Rando *et al.* 1998, Salas *et al.* 2002, Destro-Bisol *et al.* 2004, Plaza *et al.* 2004, Rosa *et al.* 2004, Coia *et al.* 2005, Jackson *et al.* 2005, Gonzalez *et al.* 2006). The Central African origin has been proposed, possibly retaining indigenous signatures of a phase common to the ancestors of Western Pygmies and Bantu people, while more specific sub-clades mark their divergence (Salas *et al.* 2002, Batini *et al.* 2007). Except for the Loko people, where L1c peaks at about 13% of L1c\* lineages (Jackson *et al.* 2005), haplogroup L1c in West Africa averages the 5-8% (Table S3; e.g. Guinea-Bissau Nalú, Balanta and Fulbe, Rosa *et al.* 2004; Sierra Leone Temne and Limba, Jackson *et al.* 2005; Senegalese Serer, Rando *et al.* 1998; Cape Verdians, Brehm *et al.* 2002; and Mauritians, Gonzalez *et al.* 2006).

Although not without controversy, the distinct L1c lineages seem to tell different evolutionary histories. Considered to be of Western Pygmy origin and posteriorly transmitted to non-Pygmy groups (mostly Volta-Congo speakers; Destro-Bisol *et al.* 2004, Gonzalez *et al.* 2006), L1c1a mtDNAs were found in a Guinean Balanta and a Papel (Table S1, GB37) and are also represented in Mali (Gonzalez *et al.* 2006). Sub-haplogroup L1c3 seems to be ubiquitously distributed from Senegal to Cameroon (Salas *et al.* 2002, references in Gonzalez *et al.* 2006) and is shown by all Guineans, with the exception of Bijagós and Nalú. Other than in Felupe-Djola and Papel, the HVS-I GB30 and GB36 motifs (respectively; see Table S5) have been sampled in Bambara, curiously a Mande group. The L1c3a1 haplotypes found in the Guinean Bak-speakers are also common among Yoruba, Hausa and other Afro-Asiatic speakers (Gonzalez *et al.* 2006). The L1c3b1 lineage in Mandenka probably matches a lineage in Yoruba, African-Brazilian and Dominican, though classified in Salas *et al.* (2002) as L1c1\*. None of the allegedly Bantu L1c1b, L1c1c or L1c2 (Batini *et al.* 2007) were found in our sample set. The variation of our Guinean L1c dataset coalesces at 108.3 kya ( $\pm 21.9$ ) which is consistent

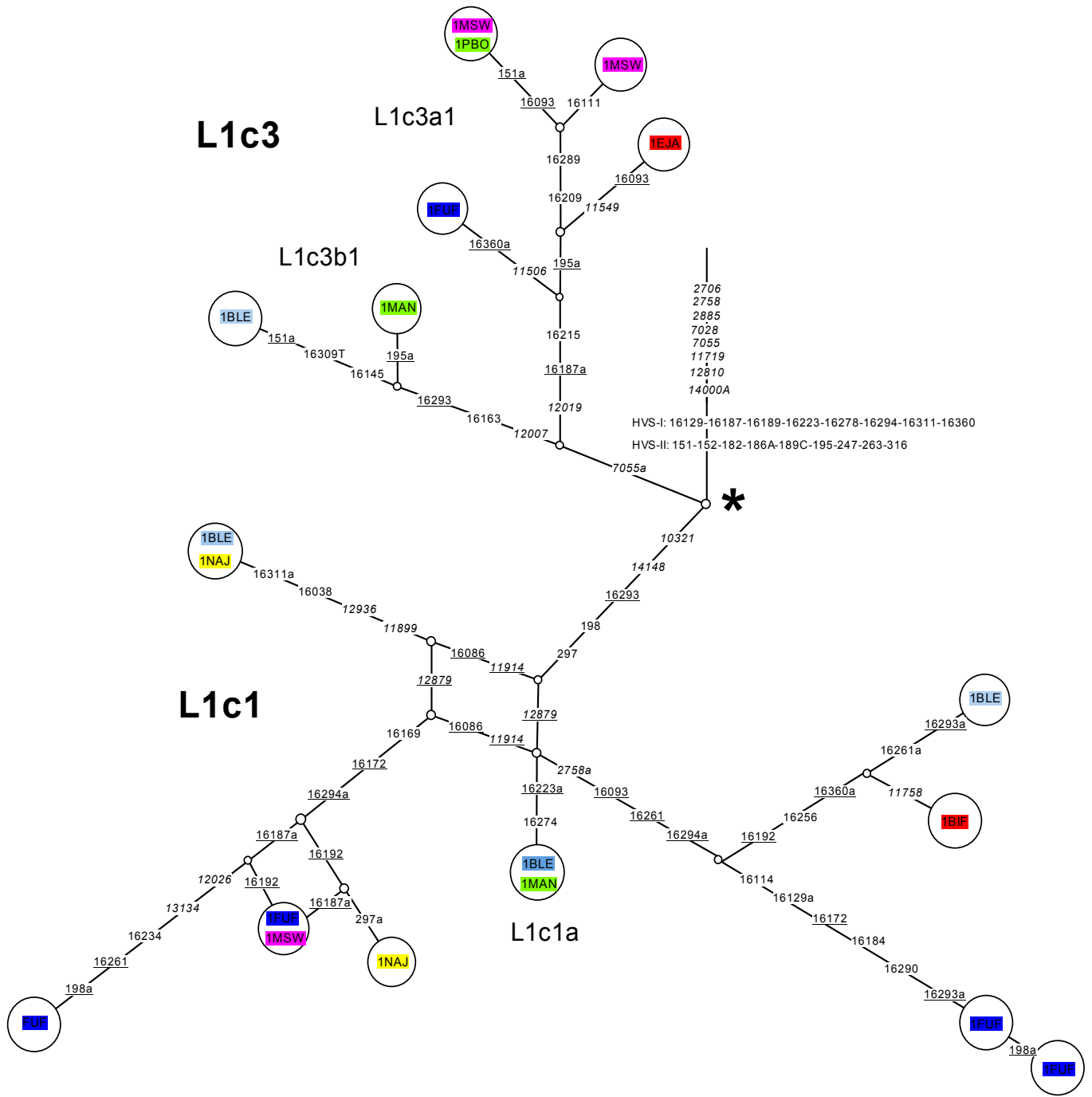


Figure 22 – Median network of the L1c haplotypes in Guinea-Bissau people, based on HVS-I, HVS-II and partial coding region information. Substitutions are shown along the links, with suffixes indicating transversions, “a” state reversions and italicized coding-region changes, relative to rCRS. Recurrent mutations within the L1c clade are underlined. The root is designated by “\*”. The abbreviation and color code of ethnic groups codes are as in Figure 20. Length polymorphisms in the C-run (16184-16193) and 16183 A to C transversion were disregarded for the network construction.

with the timeframe in Gonder *et al.* (2006), Kivisild *et al.* (2006b), Batini *et al.* (2007) but considerably older than previous estimates on smaller datasets (Salas *et al.* 2002). The absence of basal L1c types in our dataset leads us to hypothesize on the westwards expansion of the lineages relatively late in the evolution of the haplogroup. The suggestion is further corroborated by the Central-West distribution and a relatively recent coalescence age of the L1c3 subclade (Salas *et al.* 2002).

The wide continental spread of L2a justifies why the lineages in Fulbe, Felupe-Djola and Bijagós (GB57, GB62 and GB76) have matches in a broad geographical range (Tables S3 and S5). The Mandenka L2a haplotypes are common not only to Senegalese and Senegalese Mandenka (GB71, GB80) but to other geographically close Mande populations (Malinke, GB53) and Mozabites (GB71). Other than in Balanta, the L2a- $\alpha$ 3 GB44 motif traces a corridor of exact matches from Sudan and Somalia through Central areas until West African Malinke and Sierra Leone Limba and Temne. In parallel, the Balanta GB59 is present in a Moroccan Arab, stating for the relationship of North Africans and these sub-Saharan. Here again multiple matches of Fulbe lineages (L2a1- $\beta$ 3 GB57, Table S5) tell of their populational history with contacts with various people and inputs at various timescales. Other L2a1- $\beta$ 3 matches are restricted to West Africa (B56 and GB58). Haplotype GB39 L2a- $\alpha$ 2 appears as unspecific to other West Africans when matching only with eastern people in Lake Turkana, Tuareg, Nubian, and a Saudi Arabian (Table S5). The Fulbe L2a proportion of 22% contrasts with the frequency of other Fulani people in West-Central areas (not sampled to 13%, Cerny *et al.* 2006) but nevertheless three mtDNA lineages are exactly alike haplotypes in Central Africa (Table S5, L2a- $\alpha$ 1 GB50, L2a1d GB68 and GB70 and L2a1-  $\beta$ 1 GB75), some even shared with Fulani people.

The distribution of L2c demonstrates a clear West African cluster with localized expansion, where the highest frequency (39.1% in Senegal Mandenka and 22.4% in Guinean Mandenka, Table S3; Graven *et al.* 1995, Rosa *et al.* 2004) and extent of shared lineages are not unexpected: the largest number of matches is with Cape Verdians; three haplotypes in L2c are common to Felupe-Djola and Sierra Leone Temne (GB94, GB107, GB114; Table S5); Balanta and Fulbe share lineages with Mali and Senegal Mandenka (GB92 and GB110); GB117 is very common and spread over multiple people - all West African residents, Chad's Masa, and Northwest Saharans and Mauritians; Mande in Senegal have matches with Papel, Fulbe and Mandenka of Guinea-Bissau (GB111, GB112). Several one-step lineages from the central type are interestingly shown by Felupe-Djola and Papel (Figure 20), the members of supposedly "late" eastern arrivals that have most likely acquired these lineages in West Africa or on their way westwards, given its absence in the East.

The diverse set of haplogroups L2a and L2c sequences constitute the most frequent types in our sample but with different distribution within ethnolinguistic groups (26% in Bijagós and Balanta to 38% in Fulbe and Mandenka; Table S3). The almost starlike phylogeny of L2c and the many one-step derivatives in L2a1 sub-clades may reveal signatures of expansion, from a limited number of founder haplotypes shared by groups of different linguistic affiliations (Figure 20). These could have been protagonists of expansions coinciding with the gradual climatic amelioration and even participated in the dispersal of L2a clades ~14kya from a source in between East and West Africa (Salas *et al.* 2002), although being priority frequent and having a previously defined diversity (e.g. L2a1-β1 TMRCA ~30 ky, L2c TMRCA ~20 ky; Table S4). In fact, unless a drastic founder effect occurred, the age of a clade predates, sometimes considerably, the age of the ethnically defined population where it is found. Because ethnic definitions almost always include linguistic aspect, such definitions do not go usually deeper than 10-15 ky, due to the limitations in reconstructing deep-rooted language trees.

Guineans retain one of the highest proportions of L2b, a cluster largely restricted to West Africa (Table S3; Chen *et al.* 1995b, Rando *et al.* 1998, Salas *et al.* 2002). Among the best represented are the Nalú, the Felupe-Djola and the Papel, these last curiously harbouring both the basal and the more derived lineages (Figure 20). The Guinean haplotypes coalesce at approximately 39 kya, while the expansion of L2b1 begun more recently at about 9 kya (Table S4). On the light of the theory about an eastern homeland of Papel and Djola, and in parallel to that suggested for L2c lineages, it is more likely that the L2b mtDNAs were acquired on the way to, or in West Africa, as these are absent in eastern people. We note that L2b is absent in Fulbe of Guinea-Bissau though quite frequent in other Fulani and therefore defines a certain degree of “inter-Fulani” distinctiveness (Table S3). Most matches in L2b are with Cape Verde, Senegal, Mali and Wolof not to mention the particular links of Mandenka to Ethiopians (Table S5, GB83).

Haplotypes belonging to haplogroup L2d are represented by single individuals (except GB120) and do not show a common founder sequence (Figure 20). The L2d lineages are more common in Central and West Africa (Table S3) coalescing at a common node ~120 kya, probably in Central Africa (Salas *et al.* 2002). The distantly separated clades in Guinea-Bissau coalesce at about the same time and thus tell of their ancestral survival through episodes of genetic drift (Table S4). Other than Balanta and Mandenka, the L2d2 GB120 haplotype matches in Mende and Temne sample sets (Jackson *et al.* 2005), while L2d1 GB123 motif was found in a Guinean Fulbe and Saharawis (Rando *et al.* 1998, Table S5).

Both L3b and L3d are most frequent in the western-central part of the continent (Table S3). Haplogroup L3b is expressive in Bijagós, Fulbe and Felupe-Djola pools (10-14%) inserted in an average panorama of 10% in West Africans. Their near absence in Felupe-Djola's homeland makes us believe in a local amplification of frequency in a small founder group. GB127 and GB134 are particular links of Guinean Balanta to Northwest African Mozabites and Moroccan (Table S5), suggesting together with exact matches in L2a and L2b for the relationships of these people. Felupe-Djola and Mandenka have an intriguing lineage motif of Fulani prevalence (GB129), which is actually the only L3b1 lineage matching both in West and Central Africa. Other L3b mtDNAs are found only in western people, except for the widespread central haplotype of variation (GB137, Figure 20). L3d lineages are in turn restricted to sub-Saharan, with a range of 5-13% in West-Central Africa and occasional peaks in Central Mafa and Fulani (~20%, Table S3). The estimated coalescence of about 30 kya ( $\pm 8.5$ ; *Watson et al. 1997*, *Rando et al. 1998*) overlaps with our estimate of  $42.7 \pm 10.8$  ky for the L3d diversity in Guinea-Bissau (Table S4). The unsolved reticulations wait however for further phylogenetic resolution. We note that the lineages with longer evolutionary time are present in Balanta and Nalú, the Guineans groups who also have higher frequency of this clade. Geography is the main denominator of matches with few particular lineages found elsewhere (Table S5), e.g. L3d1 GB157 in Mandenka similar to lineages in Central Fulbe and Ethiopians. The vast majority of lineages are matching within West Africa corner, with GB159 being particularly frequent in Sierra Leone (*Jackson et al. 2005*).

L3e4 is observed predominantly in Atlantic West Africa and thought to be a protagonist of local expansion events with the rise of food production and the iron-smelting (*Bandelt et al. 2001*). In Guinea-Bissau moderate frequencies of L3e4 are found in Balanta and Felupe-Djola (8% and 4% respectively, Table S3), with the more derived lines shared by Balanta and Mandenka (Figure 20). As its founder type is shared by 7 individuals and only a few lineages emerge, L3e4 is placed among the haplogroups of more recent presence in Guinea-Bissau (Table S4, TMRCA of  $11.0 \pm 5.2$  ky). Curiously, the GB170 basal cluster is not only shared with West Africans but also with inhabitants of Mozambique and Sudan (Table S5). The L3e2b cluster is the most widespread type of L3e, and together with L3e2a are considered successful hitchhikers of the population movement in the Sahara during the Great Wet Phase (early Holocene) and subsequent Wet Phase (*Muzzolini 1993*, *Bandelt et al. 2001*). The haplogroups L0a1 and L3h could have also participated in such movement. In fact, Guinea-Bissau L3e2a GB162 and L3e3b GB166 motifs match widely in the continent. For L3e2a, the more ancient mtDNA lineages are shown by Mandenka and Balanta while L3e2b is mainly a Felupe-Djola and Papel cluster (Figure 20) with probable links to their homeland mirrored in East African similar lineages (Table S5).

The distribution of the haplogroup L3f comprises all East Africa (1-33%, Table S3; Watson *et al.* 1996, 1997; Krings *et al.* 1999; Brandstatter *et al.* 2004b; Kivisild *et al.* 2004) and again is present in particular groups of people of Chad and Cameroon (up to 39% in Chad Kotoko). The proportions in Guinea-Bissau average the 2% (maximum 6% in Felupe-Djola) while in the neighbouring Senegal, Mali and Sierra Leone ranges 6-11% (Rando *et al.* 1998, Jackson *et al.* 2005, Ely *et al.* 2006). The lineages tell of an ancient migration from East Africa to more central areas of the continent. For instance, one Fulbe lineage (GB178) shows exact matches with sequences from a wide range of East-African populations in Somalia (Table S5, Watson *et al.* 1996, Watson *et al.* 1997), Ethiopia (Kivisild *et al.* 2004), Egypt (unpublished) and a Peul in Mali (Gonzalez *et al.* 2006) that testify for their wide territory in the Sahel, although in the Fulani “world” L3f is just represented in the Tcheboua of South Cameroon (Table S3, Cerny *et al.* 2006). The coalescence  $49.4 \pm 16.2$  ky determined for Guinea-Bissau cluster (Table S4) is within the previously estimated error range of  $39.4 \pm 10.4$  ky (Salas *et al.* 2002).

The proportion of haplogroup L3h in Felupe-Djola is among the highest found (8%, Table S3; Watson *et al.* 1996, 1997; Brehm *et al.* 2002; Rosa *et al.* 2004). This very rare cluster has been shown to be present in East Africa (although lacking nps 16129-16362, Kivisild *et al.* 2004) and Niger/Nigeria (HVS-I motif in Watson *et al.* 1997 allowed to classify it into L3h) but not in West Africans. The L3h lineages in Guinea-Bissau may then represent an input of an eastern subset that has derived in isolation from its original types, since no matches or common founders are traceable. The high percentages among Felupe-Djola and Papel are according to their tradition of East African descendants (though not exactly establishing a direct link with Sudan). However, it is hard to interpret its association to the proposed very recent arrival of the Felupe-Djola in the 15<sup>th</sup> century, since Guinean L3h extant lineages are shared by people of different ethnolinguistic affiliation and show no common types in East Africa. Nevertheless, the coalescence age of  $14.1 \pm 8.4$  ky (Table S4) does not necessarily reflect their arrival, since a certain level of diversity could have been present already among the migrants. No exact matches were found except in Cape Verde (GB184, Table S5) which is most probably of Guinean descendants. As the haplogroup L3h is in general very rare, it could have escaped sampling in West Africa, or it can otherwise represent a direct input from eastern people.

None of the Bantu-associated markers L0a2 9bp-del ColI/tRNA<sup>Lys</sup> (Soodyall *et al.* 1996), the L2a1 HVS-I motif 16192 (Pereira *et al.* 2001b), the haplotype 16124-16223-16278 in L3b (Watson *et al.* 1997), the L3e1a characterized by mutation 16185 (Bandelt *et al.* 2001) and L5 (previously L1e in Pereira *et al.* 2001, renamed in Kivisild *et al.* 2004) were found in the Guinean sample. This suggests that either Bantu migrations contributed very little to the gene pool of Guineans or



that they had a distinct gene pool from that associated with the southwards migrants. The lack of Bantu branches of the Niger-Congo linguistic family, among a plethora of languages spoken in Guinea-Bissau, is more congruent with the first hypothesis.

The M1 lineages relate again only the Balanta and the Felupe-Djola with a predominantly East African clade (Figure 20 and Table S3). The haplotypes match however with those of one Iranian (Metspalu *et al.* 2004), two Saudi Arabians (Metspalu *et al.*, *unp data*), one West Saharan (Rando *et al.* 1998) and two Mozabites (Corte-Real *et al.* 1996), which is not surprising given the occasional occurrences of M1 in West and Northwest Africa (Torroni *et al.* 1996, Rando *et al.* 1998). We have reassigned our samples as M1b defined by Olivieri *et al.* 2006 (previously M1c in Kivisild *et al.* 2004), on the basis of a np16185 mutation present only in Morocco and Northwest Africa (Plaza *et al.* 2003, Olivieri *et al.* 2006) and several coding regions posteriorly typed (GB185 with the mtDNA molecule completely sequence, unpublished data). The M1b is a North African branch that has followed a trajectory on the southern coast of the Mediterranean, from the Near East to North West Africa namely to Morocco. The trans-Saharan spread is not likely a product of recent gene flow since a random assortment of other North West African mtDNAs would have likely been carried by the migrants as well. We have to bear in mind that Mauritania/Senegal and Mali border seems to be an important barrier to southward gene flow of the North African Euroasiatic haplogroups to sub-Sahara (Gonzalez *et al.* 2006). It is nevertheless intriguingly found in Guinean groups who harbour eastern African mtDNA variants.

Haplogroup U6 is seen as the first Palaeolithic return to Africa of ancient Caucasoid lineages from a Near-Eastern/ Mediterranean area (40-50 kya; Rando *et al.* 1998, Olivieri *et al.* 2006). The increasing frequency and diversity of its most representative clade U6a towards Northwest Africa supports the idea of a local expansion, with rather frequent distribution in Algerian Berbers, Moroccans and Mauritians (Table S3; Corte-Real *et al.* 1996, Rando *et al.* 1998, Macaulay *et al.* 1999b, Plaza *et al.* 2003). The most frequent motif 172-189-219-278 is believed to have registered a partial diffusion to Sahel ~11kya (Rando *et al.* 1998, Coia *et al.* 2005) and was observed in three different haplotypes in Fulbe, Mandenka and Manjaco (Papel related group, Figure 20). The basal haplotype GB191 matches widely in Africa suggesting a relation to an ancient Berber expansion (Table S5). The particular contact of North Africans with Guinean neighbours has been historically documented (Moreira 1964) and hypothesized based on exact matches of other haplogroups above mentioned.

Nine Fulbe and a Papel mtDNAs fall into two haplotypes of haplogroup U5b, which otherwise exhibits the main radiation in Europe (Richards *et al.* 2000). The U5b1b lineage with HVS-I-motif 16189-16192-16270-16320 and nps 7385 and 10927 (additionally typed to the markers in Rosa *et al.* 2004) is one or few steps away from a common and widespread type in

Europe, but also among Moroccans, Saharawis and Tunisians (Figure 23; Plaza *et al.* 2003, Tambets *et al.* 2004, Achilli *et al.* 2005) though none exact match was registered with these populations. A recent sizable European admixture by the time of the slavetrade is highly unlikely as other frequent European variants are absent. A recent finding linked the Saami U5 in Scandinavia to that found on North African Berbers and a sub-Saharan Fulbe by the extremely young branch of U5b1b ~9kya (Achilli *et al.* 2005), emphasizing the great importance of post-glacial expansions in shaping the genetic diversity of modern humans, in this case from the Franco-Cantabrian refuge. These post-glacial signatures most likely have crossed the Strait of Gibraltar and further derived into local clusters. The main evidence for that is that the Guinean GB188 haplotype appears in the datasets of Wolof and Serer (Table S5; Rando *et al.* 1998), Malinke (Ely *et al.* 2006), Fulani in Burkina-Faso and Chad (Cerny *et al.* 2006), Moroccans (Pennarun *et al.*, *unp data*) and North Cameroonians (Coia *et al.* 2005) indicative of a founder lineage in West Africa. The one-step derivatives in North Africa make it a more likely source for the episodes crossing the Sahara. Again, as in haplogroup U6, the linguistic correlation suggests that the spread of the haplotype in Senegambia might be related to the movement of Berber or Fulani populations (Achilli *et al.* 2005). A link to Fulani carriers is also likely under the suggestion of their ancient origin in the more northerly mountain massifs of the Central Sahara (Dupuy 1999). This suggestion corroborates to a certain extent with Cruciani *et al.* hypothesis of Eurasian influence to West Sub-Saharan Africa (based on Y chromosome evidence but with no detected analogy in mtDNA, Cruciani *et al.* 2002, Coia *et al.* 2005).

The L1b lineages shared between Papel and Balanta (see Figure 20) may at the first glance call our attention to the 'Balantization' process (Carreira and Meireles 1959), in which Balanta mtDNAs are integrated in the Papel father's ethnic group. In such line of thought, the 'Balantization' could have contributed to mask any existent East African lineages. Other shared lineages in L2a1-β2 (np 16264), L2a1-β1 (np 16355) L2c (np 16234, np 16264 and np 16390) and L3d could testify for recent influence of the Balanta and Papel women over Mandenka and Fulbe ethnic groups, known as "Sudanization" (Carreira and Meireles 1959). However, since these processes occurred in a very recent timescale (the last 4-5 generations), they would have only become evident in the present-day gene pool if there were a sizable influx of very distinct pools of maternal variability. These lineages are of West African prevalence, and therefore may probably be part of a largely common ancestral maternal pool, outlined before the definition of most of the presently known ethnolinguistic groups and certainly much before the beginning of the social processes discussed above. This is also evident in the inter-ethnically shared lineages at the basal level of many haplogroups.

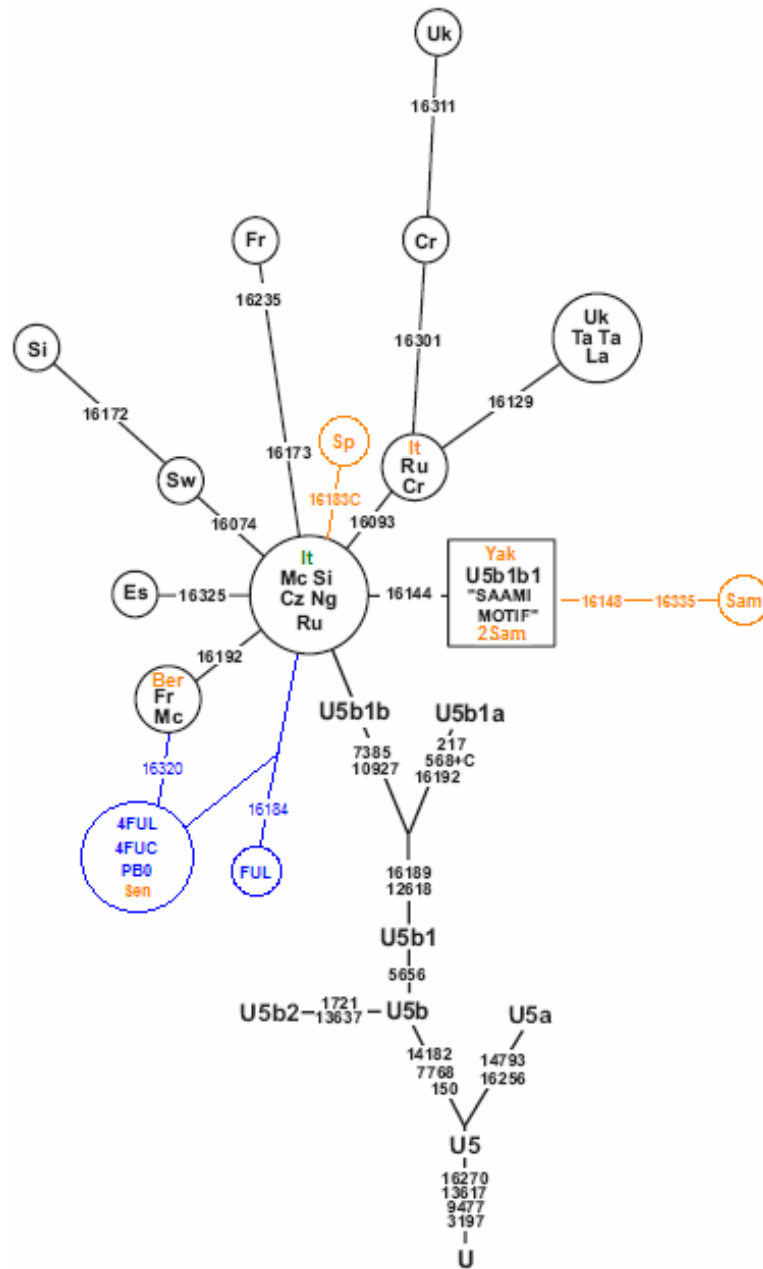


Figure 23 - Phylogenetic network of U5b1b lineages based on HVS-I sequences and their position in the phylogeny of haplogroup U (adapted from figure 3A in Tambets *et al.* <sup>(2004)</sup>). Sequence information from Finnilä *et al.* <sup>(2001)</sup> and Herrnstadt *et al.* <sup>(2002)</sup> has been used for the coding region and HVS-II. The nucleotide positions at which the nodes differ from rCRS <sup>(Anderson *et al.* 1981; Andrews *et al.* 1999)</sup> are listed along links. Suffixes are specific only for transversions and “+” indicates an insertion. U5b1b1 haplotypes include the labeled “Saami motif” whose further refined in Tambets *et al.* <sup>(2004)</sup> (Figure 3, panel B). The population codes are as follows: Cr - Croats, Cz - Czechs, Es - Estonians, Fr - French, La - Latvians, Mc - Moroccans, Ng - Nogays, Ru - Russians, Si - Sicilians, Sw – Swedes, Ta - Tatars, Uk – Ukrainians (denoted in black, Tambets *et al.* 2004); It - Italy, Sp - Spain, Sam - Saami, Yak - Yakut, Sen - Senegal, Ber – Berber <sup>(Achilli *et al.* 2005)</sup>, denoted in orange); FUL, FUC - Guinea-Bissau Fulbe, PBO – Guinea-Bissau Papel <sup>(Rosa *et al.* 2004)</sup>, denoted in blue).

## 8.1 - Principal Component Analysis

In order to place Guinea-Bissau maternal lineages in a continental context of mtDNA variation, the haplogroup frequencies of African populations mentioned in the literature (Table S3) were submitted to a PC analysis. A vast analysis of the main West African ethnic groups has been carried by González *et al.* (2006), where the authors concluded that it is difficult to distinguish between a geographic and a linguistic component underlying the genetic differentiation of groups. In that sense, and as our main focus is towards West Africans, we here chose to consider the genetic differentiation among ethnolinguistic units.

The resolution of the graphic display for the gathered 64 populations is however not the most adequate since very tight clusters are formed by the West African and Central/East African units (Supplementary material - Figure S2). We then selectively excluded populations with particular haplogroup composition, very different from others because of their outlying position compared to the rest of Africans (Khoisan-speakers) or due to the influx of lineages from other continents (North Africans), to better understand Guinea-Bissau affinities. Left with 46 populational units, the 1<sup>st</sup> PC establishes a clear-cut between West Africans and

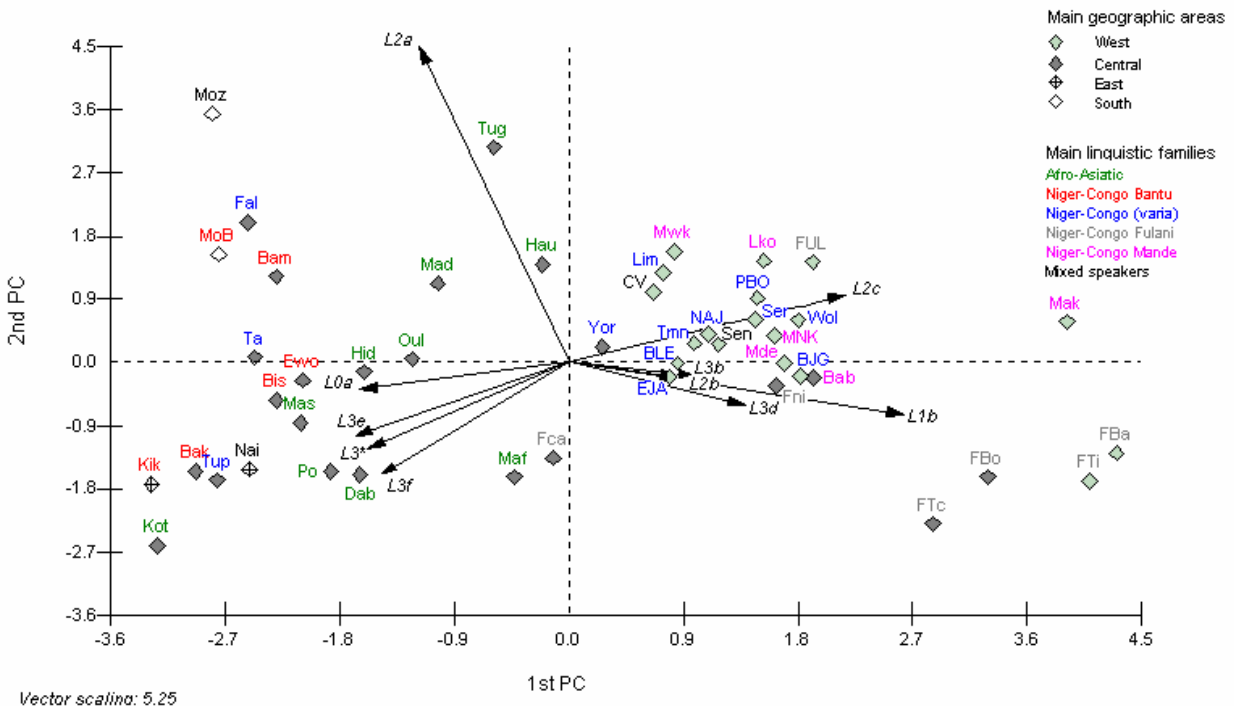


Figure 24 – Principal Component Analysis of sub-Saharan populations based on mtDNA haplogroup frequencies. Calculations based on the frequencies of 17 haplogroups for 46 populational units. The 1<sup>st</sup> PC and 2<sup>nd</sup> PC retain 34.1% and 14.0% of the variance, respectively. Population codes are as in Table S2.

other sub-Saharan, except for the Central African Fulani groups that show more (though vague) similarities to the West on the main account of their high frequency of haplogroups L1b and L3d (Figure 24). On the responsibility of L0a, L3\*, L3f and L3e the East African Nairobi and Kikuyu are placed together with Central Africans. Neither linguistics nor geography are good predictors for the placement of Central and South African populations: Chad's Chadic speakers Hide, Mafa, Kotoko and Masa show a sparse distribution in the plot; North Cameroonians and Chad Chadic speakers seem to be most closely related with Ewondo and Bassa Bantu people. All Bantu and Adamawa speakers (labeled as Volta-Congo speakers in <sup>(Gonzalez et al. 2006)</sup> are placed apart in the PC plot, among various Afro-Asiatic speakers. Although inhabitants of North and South Cameroon, three pairs of ethnic groups display curious affinities: i) Bakaka and Tupuri, ii) Bamileke and Fali (also close to Mozambique Bantu), iii) Ewondo, Tali and Bassa. The South Cameroon Bamileke affinities with North Cameroon Fali and Mozambique Bantu can actually represent Bantu both because of the genetic legacies *in situ* and due to the Bantu migrations. As already stated in <sup>(Gonzalez et al. 2006)</sup>, the Afro-Asiatic Tuareg, Hausa and Yoruba in Niger-Nigeria are genetically closer to the West Atlantic-Mande than to their linguistic counterparts. Samples from Guinea-Bissau seem to be genetically related to their geographic neighbors in Senegal. All the Guinea-Bissau ethnic groups form a cluster linked by short distances: Bijagós are curiously close to the Bambara, a Mande-speaking group, at the first glance excluding their claimed connections to the Djola, Papel or Nalú mothers; both Papel and Mandenka display similarities with the nearby Senegal Serer and Wolof driven by the L2c proportion, with the latter also in close proximity to Mende of Sierra Leone. The Balanta, Felupe-Djola and Nalú are close to Senegalese (mixed) and Temne. In our analysis, the Bambara Mande-speakers cluster together with Sierra Leone Mende and Guinea Bissau Mandenka. The Fulbe of Guinea-Bissau are distinct of both the Cerny et al. <sup>(2006)</sup> nomads and the Cameroonians Fulbe <sup>(Destro-Bisol et al. 2004)</sup>, driven by the proportion of L2c to the nearness of Sierra Leone's Loko, curiously Mande speakers, while other Fulani are shown as "outliers" under the influence of their high L1b proportion. This finding is in agreement with the non-differentiation among the nomad Fulani but from all the other settled Fulani <sup>(Cerny et al. 2006)</sup>. The Niger/Nigeria Fulbe <sup>(Watson et al. 1996)</sup> are nevertheless more related to Guinea-Bissau by their maternal pool. Although the Fulani spring from an originally nomadic population, differences may have accumulated with changes in the lifestyle and mobility of these people, nowadays mostly settled agriculturalists but also some cattle herders' nomads.

By plotting the Guinean ethnic groups alone, their relative positions change considerably as the effect of other populations is annulled (Figure 25). The distinct coordinates of these people are influenced by various mtDNA types: the Nalú reflect relatively high proportion of L2b and L3d; the Fulbe harbor higher frequency of Eurasian

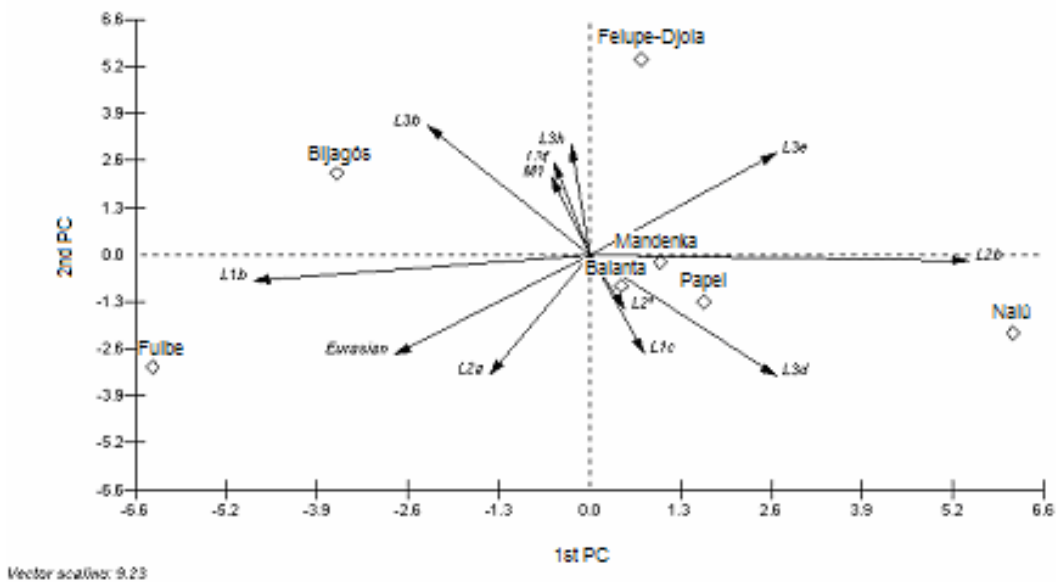


Figure 25 – Principal Component Analysis of Guinea-Bissau ethnic groups based on mtDNA haplogroup frequencies. The 1<sup>st</sup> PC and 2<sup>nd</sup> PC retain 43.82% and 23.91% of the variance, respectively.

haplogroups together with L1b and L2a; Bijagós are mainly driven on L3b, on the 1<sup>st</sup> axis closer to Fulbe; L3e defines Felupe-Djola's position.

When the variability is classified in haplogroups, the resulting diversity in our sample is  $H=0.9019\pm 0.0053$  (Rosa *et al.* 2004). In general, the indexes of molecular diversity tend to decrease following an east-to-west cline, but in Guinea-Bissau it is slightly higher than in other West Africans (Table S3). This is again not surprising given the multiple contacts and inputs that these peoples were subjected, being part of a “melting pot” in a corridor of commercial networks. The  $F_{ST}$  distances are not significant only for Hausa, Serer, Mali, and Sierra Leone Limbe and Temne (Table S6). As in Cerny *et al.* (2006), none of the mtDNA pools of the Fulani nomads was differentiated, but are distinct from that of the settled Fulani. It is interesting to note that among Guineans  $F_{ST}$  pairwise comparisons, Bijagós are the only indistinct from all the considered Fulani groups, independently of their homeland. Within our sample units we have found that the maternal pool of Fulbe is significantly different from Felupe-Djola, Balanta, Papel and Nalú ( $P<0.05$ ). Some discrepancies are found between the P-values assigned to  $F_{ST}$ -base pairwise comparisons and the relative distances plotted in the PC graphics. These differences were already mentioned by González *et al.* (2006) that attributed the heterogeneous results due to small samples sizes, global versus binary relationships and the extent of variance captured by the first two principal components.

## 8.2 - Analysis of Molecular Variance

The AMOVA test based on haplogroup frequencies of several African populations attributed 83.25% of variance to the differences within populations, 6.10% among populations within groups and 10.64% among five main geographic groups (Table S7a,  $P=0.0000$ ). The results continue to be significant and with comparable distribution of indexes in a linguistic grouping of the main families and subfamilies. These results are in agreement with the previously obtained data indicating that African geography plays an important role in defining the differences but linguistic affiliations cannot be disregarded (Salas *et al.* 2002, Wood *et al.* 2005, Gonzalez *et al.* 2006). If we consider only the sub-Saharanans (!Kung and Pygmies excluded from the analysis) the index among groups decreases considerably for both geographic (91.26-4.74-4.01) and linguistic criteria (93.0-2.49-4.51), leaving us in a more homogeneous panorama of sub-Sahara. As shown before by Wood and colleagues, when the Bantu mtDNA pool is taken out of the picture, there is an increase of within group variance (Wood *et al.* 2005). The proportion of variance attributed to the intrapopulation level continues to rise if we narrow the criteria towards West Africa and Niger-Congo speakers (97.89-1.42-0.70 and 97.75-1.57-0.68, respectively), revealing their overall homogeneity and no preferential association to geography or linguistics. The fractioning of variance is in accordance with González *et al.* (2006) clustering on West African linguistics (98.8-1.42-0.50) but not with geographic criteria that attributes 1.9% to the variance among groups  $P<0.001$ ). Out of curiosity, we tested the Fulani on their geographic areas and obtained a significant association of the intergroup and intragroup variation (97.76-0.89-1.35;  $F_{CT}=0.01346$ ,  $P=0.03617\pm 0.00709$ ;  $F_{SC}=0.00903$ ,  $P=0.04399\pm 0.00594$ ) when discriminating Central African, Burkina-Faso and Guinea-Bissau Fulani. The structuring of Guinea-Bissau ethnic groups had no statistical significance under religious or geographic criteria (Table S7b) with the vast majority of variation within the populations (>99%). The linguistic grouping with Bijagós and Mande against all Niger-Congo revealed significant for the variance between populations within groups ( $F_{SC}=0.0093$ ,  $P=0.0137\pm 0.0031$ ), most likely due to the high heterogeneity within the Niger-Congo speakers.

## 9 - Statistical parameters from mtDNA nucleotidic sequences

The demographic history of populations is thought to be reflected in various parameters of intrapopulation variation of their mitochondrial pool, including haplotype diversity, mean number of pairwise differences and the mismatch distribution (Harpending *et al.* 1993). Therefore, we have extended the same line of reasoning to our analysis.

The mtDNA HVS-I sequence comparison for the 372 Guineans led to the identification of 192 haplotypes, defined by the combinations of 93 segregating sites (24.7% of polymorphism in HVS-I between nps 16024-16400), a high ratio found only in eastern countries as Sudan and Ethiopia (Salas *et al.* 2002). The substitutions summed 102, with 92 transitions and 10 transversions, resulting in a nucleotide diversity of  $0.0218 \pm 0.0112$ . Of the distinguished lineages 93 are single occurrences in our sample whereas others occurred more frequently among ethnic groups (see Table S1) or had sequence matches elsewhere (Table S5). The most common haplotype motif is GB117 (haplogroup L2c) with a frequency of 6.7% in the total sample. The random match probability of our HVS-I mtDNA dataset, calculated as the sum of the squares of the haplotype frequencies (Stoneking *et al.* 1991), is of 1:70 (1.4%). Similar calculations have been computed for HVS-I and HVS-II databases from Sierra Leone (1:52, Monson *et al.* 2003), Mozambique (1:28, Pereira *et al.* 2001b) and Nairobi (1:83, Brandstatter *et al.* 2004b) telling of its strong utility for mtDNA testing.

The molecular diversity of mtDNA sequences is outstanding for haplogroups L1c and L2d, with the more elevated mean pairwise number of differences (Table S8a), which is not surprising and is corroborant with the network pattern of distantly separated clades. The same lineages retain the highest nucleotide diversity that roughly suggests their ancestrality. On the contrary, the nucleotide diversity is the lowest in the haplogroups with low average number of pairwise differences, namely L0a and L3h, that seem to have founder types with few (and thus recent) arising types. The Nei's gene diversity among the populations of Guinea-Bissau is generally similar to that of to previously determined values in African populations (Salas *et al.* 2002), except for higher diversity in haplogroup L1b ( $D=0.9260$ , sd  $0.0255$ ) and lower index in haplogroups L2b and L2c ( $D=0.6404$ , sd  $0.0934$ ;  $D=0.8311$ , sd  $0.0490$ ). The diversity of haplogroup L3h, not estimated before, is of  $D=0.6923$  (sd  $0.1187$ ) in the present analysis. The Tajima's D selection test is meant to evaluate deviations from neutrality but in the case of mtDNA, a genetic system assumed to be neutral, it likely reflects fluctuations in population size. Therefore, with the statistically significant negative values for haplogroups L0a, L2a and L2c (Table S8a) these are expected to have experienced expansion, as hypothesized when interpreting the networks. Fu's  $F_S$  test is by turn more sensitive to fluctuations, further unveiling haplogroups L1b, L3b, L3d ( $0.001 < P < 0.002$ ) and L2b ( $P \sim 0.05$ ) as candidates of expansion.

If we calculate the same parameters for the ethnic clusters we obtain the highest mean pairwise difference for Bijagós and Balanta, while the lowest is displayed by the Felupe-Djola (Table S8b). The sequence diversity is the lowest in Bijagós, not surprising given that these people have colonized the archipelago and, to a certain extent may have experienced a founder effect and have remained more isolated by the islander condition. On the opposite, the highest gene diversity is seen in Felupe-Djola, Balanta and Mandenka, for



the first accounting the co-existence of East African and West African types and in both the latter the presence of ancient West African lineages. Tajima's  $D$  is only significant for expansion in Papel ( $P=0.0138$ ). Though not significant for the Bijagós and the Balanta, those present "close-to-zero" values while the Mandenka show the most negative value ( $D= -1.3601$ ). This is suggestive of limited expansion for the first, possibly due to the founder effect and a higher degree of isolation in the archipelago, possibly favoring endogamic practices in the Bijagós, and more pronounced expansion for the Mandenka. On the contrary, the Fu's  $F_s$  indicates expansion for all except Bijagós ( $P=0.1430$ ).

### 9.1 - Mismatch distribution

We have calculated the mismatch distributions of the mtDNA haplogroups as the nucleotide pairwise differences in HVS-I sequences (Figure 22). The interval of mismatch differences is usually between 0-15 nps, depending on geographical region – it is relatively higher among the mtDNA lineages of Africans, which indicates the greater diversity associated with the older age of African gene pool, but low in Europe, where the value of this parameter rarely exceeds 7 (e.g. <sup>Pereira *et al.* 2001a</sup>). The average number of pairwise differences for the West African samples ranges from  $M=5.62$  in Mandenka to  $M=8.49$  in Songhai (<sup>Salas *et al.* 2002</sup>). The  $M=8.2281\pm 3.8233$  for Guinea-Bissau mtDNAs (Kimura-2P, parameter  $\gamma=0.26$ ; <sup>Meyer *et al.* 1999</sup>) is thus among the highest in West Africa.

The tendency of African populations is to display ragged and multimodal distributions, supporting the idea of their being more ancient and stationary or more diversified. When the Guinean mismatch distribution is plotted, the haplogroups with younger coalescence age show unimodal distributions, centered at none or one difference between sequence pairs, exemplified here by clusters L0a1 and L3h (Figure 26). In older clades there is a shift towards larger number of differences between lineages - in haplogroups L1b, L2a and L2c the peak of mismatch moves to 2 or 3 stepwise differences. Their unimodal bell-shape indicates an expansion with step-by-step accumulation of mutations, and their mode depending on the time passed since the expansion. As discussed before, it seems that favorable conditions for population growth have acted over these mtDNA types, with haplogroup L2c being the most starlike of the sampled variants. If the L3b1 lineages and 16148-16293-16362 in L3d are excluded, the slightly bimodal patterns of both become unimodal, thus considered a consequence of haplogroup sub-structuring (data not shown).

The distributions can become multimodal as a result of constant population size for a longer period (under genetic drift lineages present at lower frequencies have higher chance to become extinct) or multiple bottlenecks/narrow founder events and expansions (only

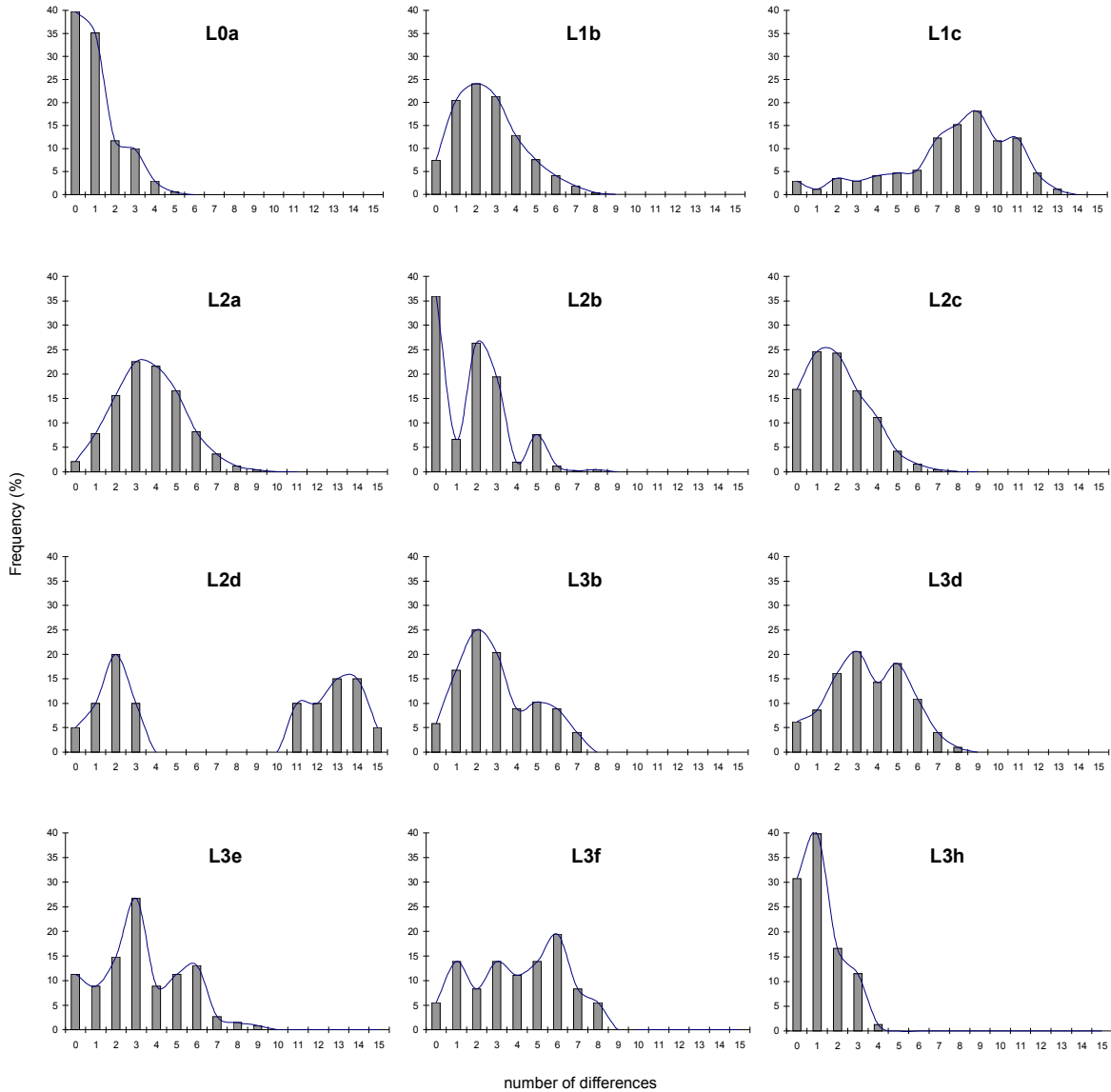


Figure 26 – Mismatch distribution of mtDNA haplogroups found in Guinea-Bissau based on HVS-I sequences.

subsets of variation persist and will define new expansion founders). On the other hand, the variants can escape sampling due to the small sample sizes. Haplogroup L1c has a ragged and multimodal distribution stating for the mutationally distant haplotypes, with the lack of intermediate variants. The capture of well-separated sub-clades of L1c is probably not due to the low sample size but rather due to a rich palette of separate sub-clades, arisen long ago, so that the subset of captured in sampling lineages reflects very different evolutionary trajectories within L1c. The distribution of L3e lineages is slightly multimodal but changes into unimodal by separating the clades L3e2a, L3e2b and L3e4 (data not shown), tellers of

different phylogenetic episodes. Similar mismatch distribution in haplogroup L2b is on the account of the L2b1 and L2b\* distinct lineages, with more than 2 steps difference in between. The curve of haplogroup L3f could be indicative of different ethnic histories, for instance of particular haplotypes in Felupe-Djola and Papel, but we have to assume that many lineages have probably escaped detection (N=9). Finally, haplogroup L2d shows two curves on the responsibility of two independent subdivisions that have accumulated differences over a long timescale – clades L2d1 and L2d2. Departures from normal distribution were tested by the significance of raggedness index <sup>(Harpending 1994)</sup> and did not attain statistical significance for any haplogroup (Table S8a). In sum, the information is redundant with the interpretation of the network topology, useful for corroborating the hypothesized.

## 10 - A phylogenetic perspective of Y chromosome pool in Guinea-Bissau population

The paternal genetic pool of Guinean ethnic groups accessed by the profile of Y chromosome haplogroups is characterized by a high homogeneity ( $D=0.4700$ ,  $sd\ 0.0333$ ) typical of sub-Saharan West Africans (see Table S11). Consistent with previous reports about West African Y chromosomes <sup>(Scozzari *et al.* 1997, 1999; Underhill *et al.* 2000; Semino *et al.* 2002; Wood *et al.* 2005)</sup>, haplogroup E3a\*-M2 is the most frequent clade in every of the considered ethnolinguistic units, ranging from 58.0% in Felupe-Djola to 82.2% in Mandenka (Figure 28; Rosa *et al.* 2007). The M2 genetic marker has been proposed to trace the routes of agriculturalists, especially in the context of the Bantu expansions <sup>(Passarino *et al.* 1998, Underhill *et al.* 2001a)</sup>. Nonetheless, its high proportion and diversity in West Africa (Figure 27) indicates an early local origin and expansion, in the last 19 ky <sup>(Semino *et al.* 2004)</sup>, with a high chance of representing a major populational growth under the cultivation “know-how”. In fact, several authors believe in the early existence of a West African agricultural centre prior to the Bantu-expansions <sup>(Cavalli-Sforza *et al.* 1994, Jobling *et al.* 2004)</sup>, perhaps as early as 9-6 kya <sup>(Atherton 1972, Calvocoressi and David 1979, Clark 1994)</sup>. Husbandry practices allowed to support a higher number of people and have promoted either an expressive expansion that overwhelmed the previous pool, or a more modest growth in a background of reduced diversity (after the LGAM savanna retreat or the malaria epidemics, product of agriculture; <sup>Adams and Faure 1997, Kwiatkowski 2005</sup>).

Mandenka and Balanta show the highest frequency (except Bijagós) and diversity levels of haplogroup E3a\*-M2 in Guinea-Bissau (Figure 28 and Table S13; 82.2%,  $R_{ST}=0.5208$ ,  $sd\ 0.2979$ ; 73.1%,  $R_{ST}=0.5166$ ,  $sd\ 0.2895$ ) attesting for a more divergent founder pool and more marked expansion. As no Bantu-speakers inhabit today the area, and none expressive westward migration of this people is documented, the ancestors of

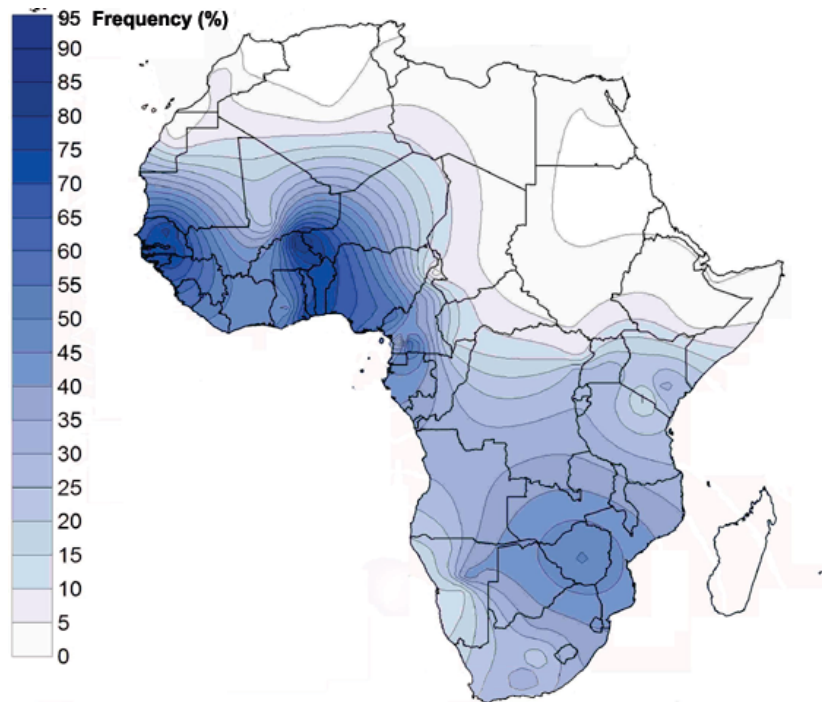


Figure 27 – African spatial distribution of haplogroup E3a-M2. Frequency scale (in percentage) is shown on the left. Data according to population datasets described in Tables S10 and S11.

Mandenka and Balanta are, among Guineans, candidates to have promoted or relate to the people that have promoted the lifestyle transition, or at least to have profited from it (Rosa *et al.* 2007). On the basis of such socioeconomic knowledge the Mande people pioneered with the foundation of economically centralized states, based on trade and agriculture and with the later aid of iron-smelting techniques, more than millennia ago – the historic empires of Ghana, Mali and Songhai (Newman 1995). The Guinean E3a\*-M2 pool, studied in the light of STR variation within the clade, coalesces at a TMRCA of  $20.5 \pm 4.7$  kya (Table S12), which is in accordance with previously calculated ages (Semino *et al.* 2004). If to consider haplogroup E3a\*-M2 coalescence estimates for the different ethnic groups, the Balanta and the Mandenka exhibit the oldest TMRCA in our dataset (Table S13,  $29.0 \pm 6.9$  and  $23.5 \pm 4.4$  kya, respectively; Rosa *et al.* 2007). As mentioned before, the Balanta cultural and physical affinities with Bantu suggest a common origin at the end of the Pleistocene near the Nile (Quintino 1969), where they could have jointly learnt the agricultural techniques. However, one should keep in mind that coalescence ages as those indicated above antedate by far the beginning of agriculture not only in the Nile Valley but anywhere. The E3a\*-M2 pool in Bijagós and Fulbe is less diverse (Table S13) signaling either a genetic bottleneck or a more recent expansion from a less diverse subset of male founders, arriving to western regions. The data are consistent with a less diverse profile of E3a-M2 among Central Africans, and thus the first documented presence of Fulbe in West Africa after the 8<sup>th</sup> century (Carreira and Meireles 1959).

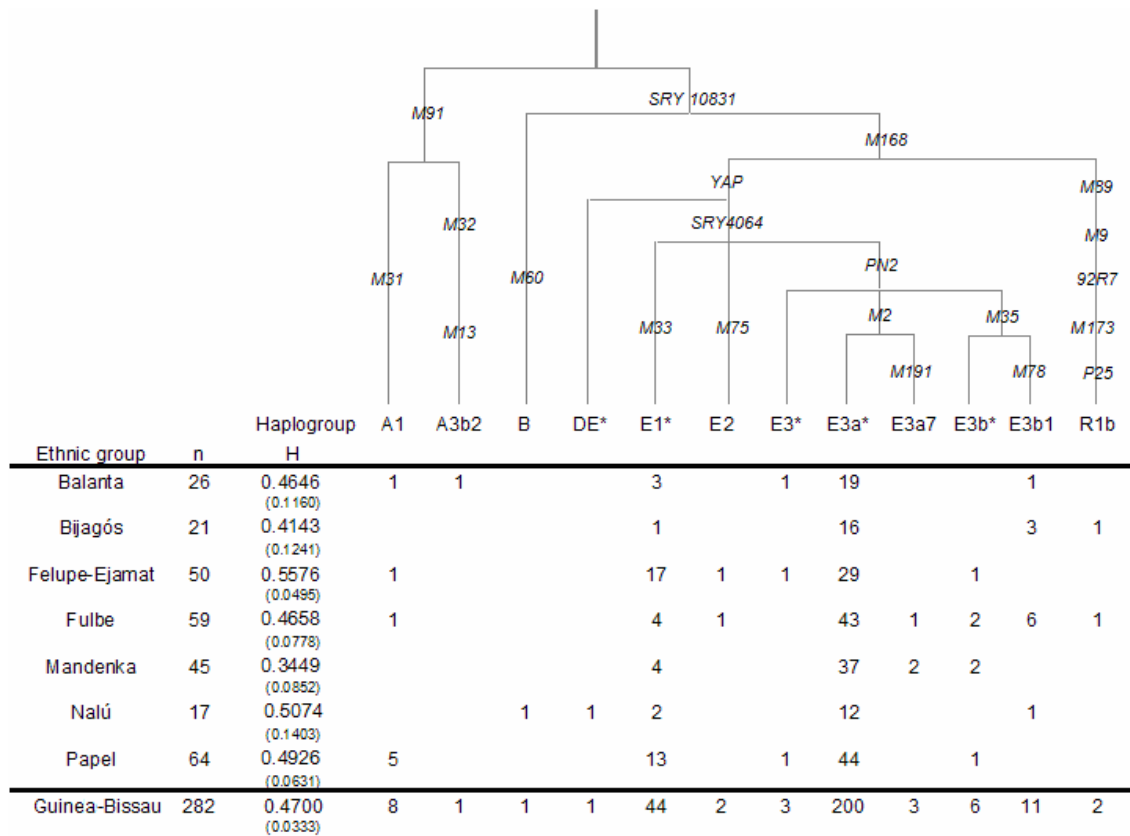


Figure 28 – Parsimonious Y chromosome phylogeny of the twelve haplogroups found in Guinea-Bissau and their distribution among ethnic groups. Absolute numbers are shown for the total sample and ethnical clusters. Haplogroup nomenclature and defining mutations assayed in this study, shown along the branches of the phylogeny, are as proposed by YCC<sup>(2002)</sup> and Jobling and Tyler-Smith<sup>(2003)</sup>. The bold link indicates the root, determined by comparisons with primates (Underhill *et al.* 2000, Hammer *et al.* 2001). In Rosa *et al.* (2007).

The Mandenka people in Guinea-Bissau share E3a\* microsatellite haplotypes with all other groups in Guinea-Bissau and do not match with those out of Central-West Africa (Tables S9 and S14, except H67 in Mozambique), suggesting a localized expansion. Several exact matches of E3a\*-M2 haplotypes were found between Guinean Fulbe and Equatorial Guineans, Angolans, Mozambicans and Xhosa (Alves *et al.* 2003, Arroyo-Pardo *et al.* 2005, Willuweit and Roewer 2007) mirroring the broad distribution of these people. The Felupe-Djola, the Balanta and the Papel share each one particular haplotype with Mozambique and Angola (Table S14a H46, H49 and H127) but this is rather a consequence of a west-to-east flow of Y chromosomes, possibly carried by migrants during the spread of agriculture. Many other matches with Europeans were registered for E3a\*-M2 haplotypes (Tables S14a and S14b), most likely descendants of incoming slaves given the ancestral absence of this haplogroup in Europe. Caution should be however taken when performing the haplotypic search in public

databases: in case of the lack of haplogroup assignment one cannot guarantee that two identical STR-defined haplotypes refer to the same SNP-defined lineage. As recognized before, microsatellite alleles should not be taken as surrogates of UEPs (e.g. <sup>Cruciani *et al.* 2004, 2006; DiGiacomo *et al.* 2004</sup>). The E3a7-M191 lineages present in one Fulbe and two Mandenka are, on the other hand, a testimony of a Central African lineage that followed a trajectory to the west <sup>(Underhill *et al.* 2000, Cruciani *et al.* 2002, Semino *et al.* 2002)</sup>.

The Felupe-Djola and Papel groups exhibit the highest diversity of Y chromosome haplogroups (Figure 28;  $D=0.5576$ ,  $sd\ 0.0495$  and  $D=0.4926$ ,  $sd\ 0.0631$ , respectively) together with traces of the deepest rooting phylogenetic clades in our dataset – haplogroups A-M91, E2-M75 and E3\*-PN2 - also with occasional occurrences in Fulbe and Balanta. These minor imprints may represent genetic flow from Sahel's more central and eastern areas, in particular the E3\*-PN2 which is common in Ethiopia <sup>(Cruciani *et al.* 2002, Semino *et al.* 2002)</sup>. If to interpret their paternal pool in cultural grounds, it is then relevant to mention the arrival of Djola from Sudan in the 15<sup>th</sup>-16<sup>th</sup> century claimed by their oral tradition <sup>(Quintino 1969)</sup>. For the Papel, also curiously affiliated to the Bak-speakers, their Y chromosomes may either represent a late arrival of eastern migrants that have kept a more discrete identity, or survivors of a local ancient pool through bottleneck episodes and expansions <sup>(Rosa *et al.* 2007)</sup>.

Haplogroup E1\*-M33 is of putative West-Central African origin and major distribution (from 10% in Fulbe to 45% in Dogon, with an outstanding frequency of 53% in Cameroonian Fulbe, Table S11; <sup>Scozzari *et al.* 1997, 1999; Wood *et al.* 2005</sup>). These lineages are surprisingly frequent in Guinean Felupe-Djola and Papel (34% and 20%, see Figure 28). Assuming a recent arrival of these people from an eastern source, the high proportion of E1\*-M33 lineages can only be explained by a founder effect – a limited number of migrants have increased the frequency of an West African regional clade when in contact with local populations <sup>(Rosa *et al.* 2007)</sup>. In fact Felupe-Djola and Papel E1 haplotypes form a central cluster of one-step difference from each other, among the Guinean diversity. The E1\*-M33 microsatellite diversity in our dataset coalesces at  $18.7\pm 3.6$  kya (Table S12), slightly older than the obtained for the sample of Semino *et al.* <sup>(2004)</sup> in North Africa and the Mediterranean area ( $14.3\pm 3.7$ ky), not surprising if considering the sub-Saharan origin and expansion of the clade. The relatively lower frequency of haplogroup E3a\*-M2 in Djola (58.8%) and Papel (68.8%) suggests their shallow time of permanence in West Africa and/or that their genetic flow with West Africans has not been sufficient for a greater homogeneity among the peoples.

Haplogroups A-M91 and B-M60 are among the two most basal clades of the Y chromosome phylogenetic tree, associated with the earliest AMH paternal diversification of

lineages and putative hitchhikers of the first pan-African dispersals of hunter-gatherers. The Guinean A-M91 lineages are nevertheless included in the West African A1-M31 subcluster, not in A3-M32 of East and South African distribution (Underhill *et al.* 2000, Semino *et al.* 2002, Rosa *et al.* 2007). The Y chromosome sub-haplogroup A1 is so far reported at low frequencies of about 2-5% in Mali (Table S11; Underhill *et al.* 2000), Gambia/Senegal Mandenka, North African Berbers (Scozzari *et al.* 2001) and Bakola Pygmy (Wood *et al.* 2005). Although A-M91 is among the oldest Y chromosome lineages in modern humans, STRs in its subset sampled in Guinea-Bissau coalesce at about  $9.8 \pm 2.9$  ky (Table S12), stating for a recent gene flow from eastern regions and expansion into West Africa, rather narrow in diversity (i.e. creating sharp founder effect). It is therefore not surprisingly more frequent among the Papel and Balanta people (Figure 28; Rosa *et al.* 2007). The E3\* lineages and the only A3b2-M13 in Guinea-Bissau dataset may trace the Balanta to their Sudanese-speakers relatives (Cruciani *et al.* 2002) since these clades were found to be common among Ethiopians (Semino *et al.* 2002, Shen *et al.* 2004). The clade B-M60 is observed in almost all sub-Saharan collections at marginal proportions (Scozzari *et al.* 1997, Underhill *et al.* 2000, Semino *et al.* 2002) and was present in one Guinea-Bissau Nalú. The low resolution of its typing does not allow further inferences.

E3b\*-M35 lineages are of greater prevalence in the eastern quadrant of Africa (Table S11), also peaking in frequency and diversity in the Democratic Republic of Congo (Wood *et al.* 2005) and the South African !Kung (Scozzari *et al.* 1997, 1999). Its presence at ca. 5% present in Senegal (Semino *et al.* 2002) and the ca. 2% found in Guinea-Bissau may also represent loose relationships with population in North Africa, since it is also widespread at rather low frequencies in North African Arabs and Berbers (<5%; Bosch *et al.* 2001, Scozzari *et al.* 2001, Arredi *et al.* 2004, Semino *et al.* 2004). In our dataset, the paragroup clusters together Felupe-Djola and Papel (about 2%) and is also present among Fulbe and Mandenka people (approximately 4%, Figure 28). Although calculated with a very limiting number of samples, the coalescence age of  $16.9 \pm 5.9$  kya (Table S12) is within the range of estimated for the E3b\* Y-chromosomal lineages in North Africa (9-19 ky in Arredi *et al.* 2004, using the same molecular clock as in Zivotovsky *et al.* 2004).

Guinea-Bissau haplogroup E3b1-M78 attains the highest frequency so far reported for West Africans (about 4%, Table S11). A scenario of eastern prevalence and North and West African spread reflects the African distribution of E3b1 lineages, not to mention the frequency of about 7% in Near Easterns and Mediterranean Europeans (Underhill *et al.* 2000, 2001a; Cruciani *et al.* 2002, 2004; Semino *et al.* 2002, 2004; Wood *et al.* 2005). In Guineans the MRCA time estimate of  $11.5 \pm 3.1$  ky (Table S12) is concordant with the estimate in Semino *et al.* (2004) ( $14.9 \pm 4.1$  ky), but escapes the upper time interval obtained by Cruciani *et al.* (2004) (21.1-25.4 ky) and the

lower one by Arredi *et al.* <sup>(2004)</sup> (5.42-10.71ky). The different estimates have an underlying cause in that the microsatellite haplotypes within E3b1-M78 are considered to represent different clusters with variable frequency in different regions (sub-clusters  $\alpha$ ,  $\beta$ ,  $\gamma$  and  $\delta$ ; Cruciani *et al.* 2004). The haplotype E3b1- $\beta$  DYS349 allele 10, particularly widespread among Moroccan Arabs was found in one Guinean Fulbe and one Bijagó (H162 and H164, Table S9), telling of a contribution of North West Africans that have crossed the Sahara. The hypothesis of more recent European arrival, at the colonizing period post-15<sup>th</sup> century, appears less likely since none of the remaining E3b1 haplotypes harbor the “European” A7.1 allele 9, and all a panoply of rather frequent haplogroups in Europe is absent (except 2 R1b-P25). Nevertheless, three Fulbe E3b1-M78 haplotypes (H155 and H156 for 10 Y-STR loci plus H157 for 8 Y-STR loci, Tables S14a and S14b) match Spanish haplotypes <sup>(Zarrabeitia *et al.* 2003)</sup> and samples in central Portugal, Macedonia, Romania and Poland (YHRD, <sup>Willuweit and Roewer 2007</sup>), hint for possible European paternal ancestry. Intriguingly, both H155 and H156 profiles present the A7.1 allele 12 which is quite frequent in Equatorial Guinea though no exact matches were registered <sup>(Arroyo-Pardo *et al.* 2005)</sup>.

A recent European admixture, at the times of the slavetrade, is a likely explanation for the two R1b-P25 lineages found in Fulbe and Bijagós (Figure 28, Table S9). The European source of R1b chromosomes has been stated as of great expression for the nearby Cape Verdians <sup>(Brehm *et al.* 2002)</sup>. The haplotype H165 has an exact match of 10 Y-STR loci with 71 worldwide populations, of which 57 are Eurasian (nine matches with Portuguese samples, YHRD <sup>Willuweit and Roewer 2007</sup>). Concerning the R1b H166 profile, exact matches are only found when reducing the search criteria to 7 Y-STR loci (Table S14b), matching in three European populations and two out of 300 samples of Reunion islands (known to have a European-permeable society), not to mention Austronesians. Their introduction in Guinea-Bissau territory by North African pastoral immigrants can not be ruled out, though. Here, the R1b lineages (in a proportion of 3-12%) were most likely acquired due to the long-term reported contacts with Europe, mainly Iberia <sup>(Bosch *et al.* 2001)</sup>. However, given that no exact matches with North Africans were established and no other highly frequent North African haplogroups were detected, we consider more likely the European origin for the Guinean R1b chromosomes. The M173 and P25 derived states of our samples rule out the relationship to the R1\*-M173 lineages found in Cameroon, Oman, Egypt and Rwanda, adduced to support the “Back-to-Africa” demographic scenarios <sup>(Cruciani *et al.* 2002, Luis *et al.* 2004)</sup>.

The intriguing profile of genetic markers found in a Nalú individual allowed us to classify it in the rare and deep-rooting paragroup DE\* (Figure 28), so far described only in five Nigerians <sup>(Weale *et al.* 2003)</sup>. The DE\* Y chromosomes represent a coalescent “missing link”



paraphyletic to haplogroup D-E variation or both. However, the Y-STR profile of the Guinean sample is one-step away of the allelic state described for Nigerians (Table S9, DYS390\*21, DYS388 not tested; *Weale et al. 2003*), therefore suggesting a private common ancestor but not elucidating the phylogenetics. The subject deserves further attention, either in checking for state reversion of M174 (no inner markers typed) and searching for new polymorphisms.

The relationships of haplotypes of several Y-chromosomal clusters are depicted by means of networks in Figure 29 (a-e). However, these networks are not highly informative from the phylogenetic point of view, because of multiple reticulations, no clear definition of founder nodes and no apparent ethnic association of subsets of the paternal variation. In the E3a\*-M2 network (Figure 29a), one particular fact called our attention – several haplotypes are shared among Fulbe, Mandenka, Balanta and Papel, precisely the ones involved in the phenomena of “Sudanization” and “Balantization”. Under a strict pattern of patrilocal miscegenation we should expect genetic flow of mtDNAs but not of Y chromosomes. As suggested for the mtDNA counterpart, we are in the context of a largely common paternal variability, defined much earlier than the ethnolinguistic groups. This is therefore a hindrance in tracing the relevance of these socio-cultural processes. Furthermore, we have to consider that the network construction was based only in the information of 7 Y-STR loci, and that the nature of these fast mutating markers makes them prone to reversions. The diversification in E3a\*-M2 background is old enough so that particular lineages can be “equal-by-state” instead of “equal-by-descendant”.

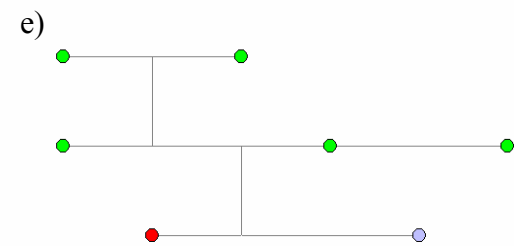
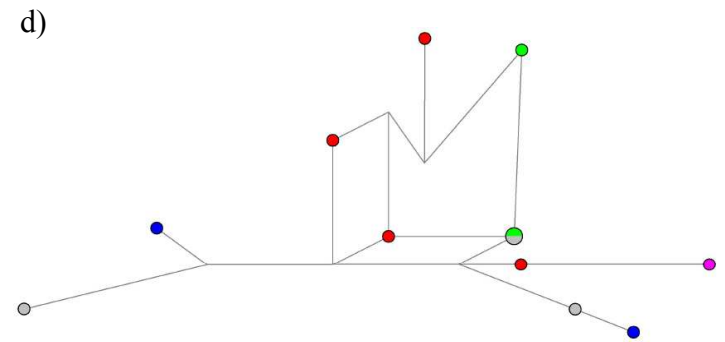
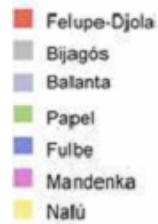
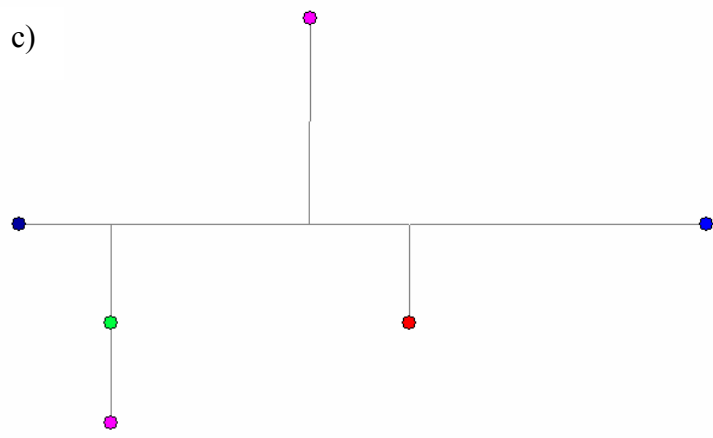
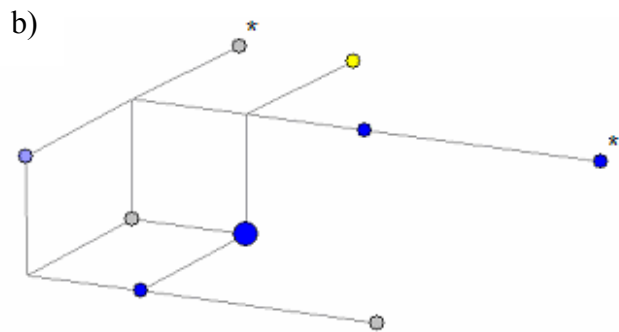
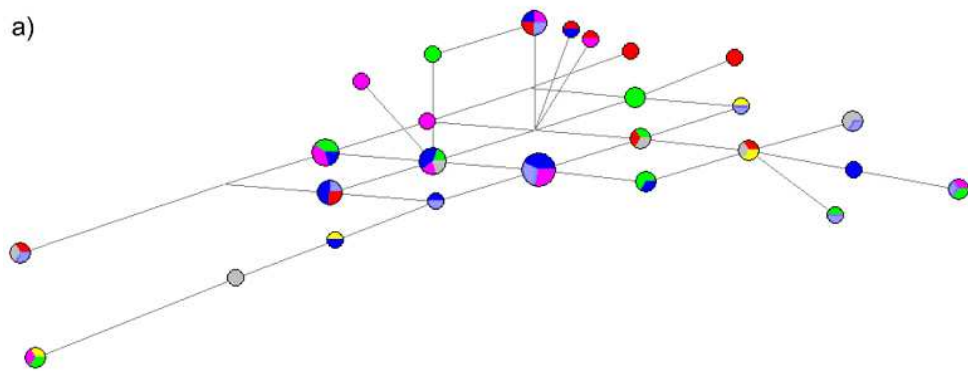


Figure 29 – Networks of Y-STR haplotypes for several Y chromosome haplogroups, based on the information of 7 loci. Loci were weighted according to Helgasson *et al.* <sup>(2000)</sup> and built with combined use of reduced-median and median-joining criteria <sup>(Bandelt *et al.* 1995, 2000)</sup>. a) haplogroup E3a\*-M2, includes 75 individuals distributed among 26 haplotypes (52 singletons excluded); b) the scheme for haplogroup E3b1-M78 represents 11 individuals included in 9 haplotypes, with singletons included; c) haplogroup A1; d) haplogroup E1\* and e) haplogroup E3b\*, all including singletons. Node size is proportional to the number of individuals.

## 10.1 - Principal Component Analysis

The coordinates of the first two Principal Components extracted from the Y chromosome haplogroup frequencies place the North, East and West African populations in a geographic perspective, forming independent and tighter groups (Figure 30; see Table S10 and Figure S3). Haplogroups E3b2-M81, E3b1-M78 and J-12f2 are the responsible for the coordinates of North Africans, positioned almost as outliers in the African paternal landscape. The 1<sup>st</sup> PC clearly separates the Afro-Asiatic speakers from other linguistic families, independently of their geographic location. The West African Y chromosomes are clustering on the major significance of haplogroup E3a\*-M2, with a less significant contribution of E1-M33. Central and South African people are more dispersed in the plot. The Central African Pygmies and South African Khoisan coordinates are on the account of their particular lineages of R-M207, A3-M32 and B2-M182 lineages. A linguistic correlation can be hypothesized for the closeness of Bantu-speakers that nevertheless inhabit distinct quadrants of the continent, driven by the frequency of sub-haplogroup E3a7-M191.

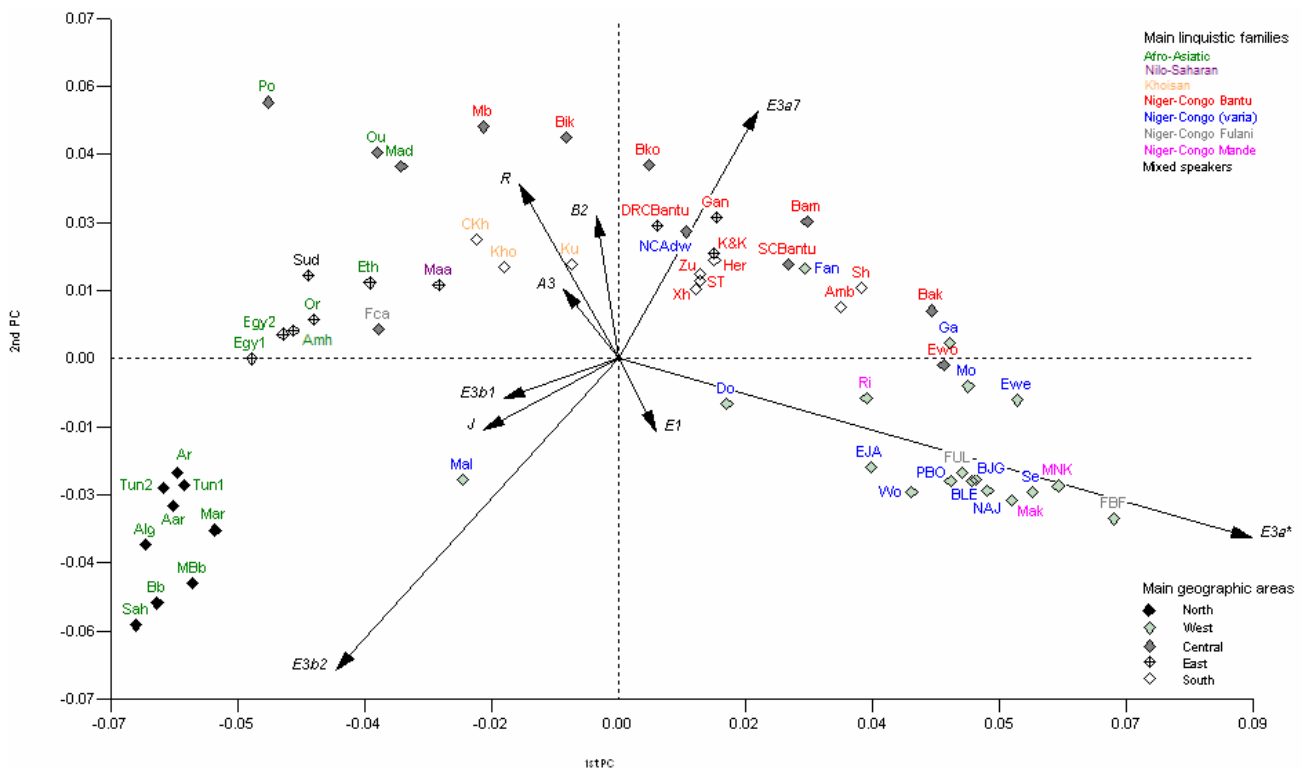


Figure 30 – Principal Component Analysis for several African populations based on Y chromosome haplogroup frequencies. The 1<sup>st</sup> PC captures 42.6% of the variance and 16.9% are under the responsibility of the 2<sup>nd</sup> PC. For codes and further details on population datasets see Supplementary material - Table S10.

(Wood *et al.* 2005). The ethnic groups in Guinea-Bissau are among the West African cluster, together with Gambia and Senegal people, with whom share numerous population groups (e.g. Fulbe and Mandenka). It is worth noting that the Guinean Fulbe are integrated in the paternal variation of other Guinea-Bissau people and thus have a distinct genetic pool of those Fulani in Burkina-Faso and Cameroon. However, we have to be aware that the picture may be refined with the further phylogenetic discernment of haplogroup E3a\*-M2.

A second PCA solely considering the Guinean ethnic units on the present survey aimed to reduce the influence of E3a\*-M2 and at the same time promote that of minor Y chromosome clusters (see Figure 31). The paternal genetic structure of both Felupe-Djola and Papel is displayed as more discrete than that of other people, due to their high proportion of E1-M33. The position of the Mandenka is clearly defined by the E3a\*-M2 chromosomes, not surprising in the light of an assumedly indigenous West African population which harbors the highest diversity of this clade (Table S13). Again, as verified for the maternal genetic pool, the PCA tells of higher similarities of Bijagós and Fulbe people than with the other Guinean ethnic groups, not supporting the idea that Bijagós are relatives of Djola, Papel or even Nalú (Teixeira da Mota 1954).

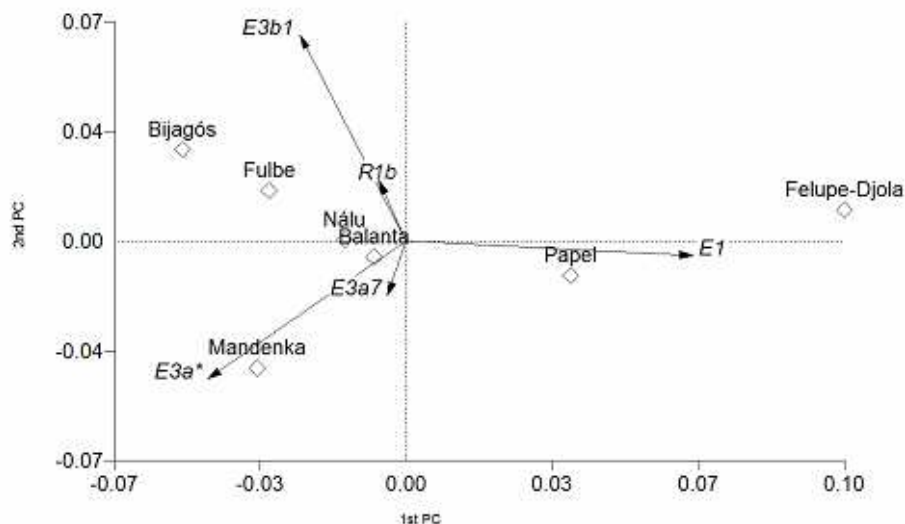


Figure 31 - Principal Component Analysis for Guinea-Bissau ethnic clusters, based on Y chromosome haplogroup frequencies. The PCA captures 87% of the variance with 74% and 13% attributed to the 1<sup>st</sup> and 2<sup>nd</sup> PC, respectively.

A pairwise  $F_{ST}$  analysis based on Y chromosome haplogroup frequencies outstated the non-significant differences among the Bantu-speakers in Central, East and South Africa (Table S15; e.g. Ghanian Ewe, Ga and Fante do not yield statistically significant  $F_{ST}$  indexes

when compared to South Cameroon Bantu and Bamileke or South African Herero, Ambo and Shona). This is likely due to the spread of Bantu people, with replacement of local hunter-gatherers or gene flow and language replacement of the inhabitants *en route*. In Figure 27, the paternal pool of Ewondo Bantu-speakers is shown to have larger affinities to the West African variability. Their proximity is in agreement with non-significant  $F_{ST}$  distances to Bijagós, Balanta, Nalú, Wolof and Mossi (Table S15). The similarity of the paternal pool of Senegalese, Wolof and Mandenka is again confirmed by the  $F_{ST}$  analysis. Also in accordance with the graphic display, and justifying their “misplacement” out of geographic and/or linguistic context, are the  $F_{ST}$  significant differences of Cameroonian Fulbe, Mbuti and Biaka Pygmies, Mali, Dogon, North Cameroon Chadic and Khoisans against all the considered populations. The Felupe-Djola is the only Guinean group with statistically significant differences from Bijagós (Table S15;  $F_{ST}=0.0947$ ,  $P=0.0268$ ), Fulbe ( $F_{ST}=0.0812$ ,  $P=0.0040$ ) and Mandenka ( $F_{ST}=0.1071$ ,  $P=0.0040$ ). The Bijagós and curiously the Burkina-Faso Fulbe do not exhibit significant  $F_{ST}$  from any Guinean group except Felupe-Djola. An exact test of population differentiation further distinguishes the Papel from Bijagós, Fulbe and Mandenka ( $P<0.05$ , data not shown). The results are generally in agreement with the PCAs, with the interpretation of higher dissimilarity of the paternal pool of Felupe-Djola and Papel among other Guineans.

## 10.2 - Analysis of Molecular Variance

When grouping the paternal profiles of the compiled African datasets on Northwest, Northeast, West, Central, East, and South Africa, 15.35% of the variance is apportioned among the geographic groups (Table S16a,  $P=0.0000$ ). The intrapopulation variation retains the largest portion of 72.07% while the remaining is attributed to the differences among populations within groups. The linguistic structuring of the same samples into the Afro-Asiatic, Niger-Congo and Khoisan families is, in terms of inter-group, intra-group and intrapopulation variance, comparable to that of the geographic criteria (67.46-14.89-17.65). However, the percentage of variance among populations within the groups decreases considerably (8.45%) if the languages families are further subdivided (Table S16a), meaning that populations speaking related languages harbor similar genetic pools. The sub-Saharan geographic variance (75.93-14.69-9.11) suggests lower genetic differences among groups, but that is not so when language families are considered (74.69-9.16-16.15). The linguistic criteria seems to play then a higher role in defining the paternal variability, since the proportion retained among groups is higher than that among populations within groups. The sub-Saharan index among groups tells however of a more homogeneous genetic pool

among Niger-Congo speakers (85.66-7.04-7.30). In West Africans the intrapopulation component retains the larger variance (86.31%) if we do not include Cape Verdians, where the non-African diversity is known to be large (Goncalves *et al.* 2003). Significant differences among Niger-Congo West Africans are only found if we consider linguistic subdivisions for the non-Fulani and non-Mande people (89.85-3.10-7.05). The AMOVA test yielded no significant results for the grouping of Guinea-Bissau ethnic units under geographic, linguistic or religious criteria (Table S16b). The intrapopulation variance is statistically significant for all the criteria, retaining more than 97% of the variance. When compared to the AMOVA test for the mtDNA counterpart, we verify that the intrapopulation index is usual lower, while the variance among groups retains a higher proportion. This can testify for a more significant role of the linguistic and geographic criteria in defining the paternal genetic pool, possibly interpreted in terms of patrilocality and sex-biased patterns of admixture. In addition, it inevitably reflects the state-of-the-art of the refinement of the Y chromosome phylogeny, and new biallelic markers may allow achieving higher phylogeographic and intraethnic resolution.

### 10.3 – Statistical parameters for Y chromosome microsatellite variation

The allelic range and allelic frequencies of eleven Y chromosome microsatellite analyzed for the Guinea-Bissau samples (Table 1a and 1b in Rosa *et al.* 2006) are in agreement with the determined for sub-Saharan Africans (Kayser and Sajantila 2001, Gusmao *et al.* 2001, Trovoada *et al.* 2001, Alvarez *et al.* 2002, Pereira *et al.* 2002, Leat *et al.* 2004, Arroyo-Pardo *et al.* 2005). Alleles DYS19\*15, DYS390\*21, DYS392\*11, DYS437\*14 and DYS438\*11 held high frequencies in the Guineans while in the other African populations used for comparison these appear at moderate or low frequencies. In addition, the Guinean proportion of allele DYS393\*14 (about 60%) is to our knowledge the highest reported (Rosa *et al.* 2006). The haplotype distribution of DYS385 ranges from alleles 13 to 21, where the most frequent haplotypes 15/16, 16/16 and 16/17 (~16% for each combination) are either absent or weakly represented outside of Africa (e.g. Bosch *et al.* 2002, Gusmao *et al.* 2002, Zarrabeitia *et al.* 2003, Quintana-Murci *et al.* 2004).

All loci show a unimodal distribution (see frequency profile in Table 1, Rosa *et al.* 2006), including DYS392, which is bimodal in some populations (e.g. Bosch *et al.* 2000, Gonzalez-Neira *et al.* 2000, Kayser and Sajantila 2001, Alves *et al.* 2003, Carvalho *et al.* 2003, Zarrabeitia *et al.* 2003, Arroyo-Pardo *et al.* 2005). Loci DYS19 and DYS389II exhibit the highest allelic diversity in this survey ( $D=0.7182$  and  $0.7239$ ), not to consider the equivalent DYS385 heterozygosity ( $H=0.9031$ ). Together with DYS393, these four loci held higher haplotype diversity than the European populations, and are thus reveal to be more informative for discriminating between African individuals. On the

other hand, DYS391 and DYS392 display the lowest allelic diversities, supporting their limited utility in forensic caseworks involving sub-Saharanans, as previously suggested (Leat *et al.* 2004).

The eleven Y-STR profile of 164 individuals results in 157 distinct haplotypes ( $H=0.9997\pm 0.0011$ ; H165 to HB220 not considered), with the highest frequency of two individuals (Table S9). The Y-STR discriminatory power reached for the Guinean haplotypes is higher than for other populations with similar or higher number of loci analyzed: Europeans (11 loci  $D=0.9983$ , Roewer *et al.* 2001; 14 loci  $D=0.9992$ , Kayser *et al.* 2003; 19 loci  $D=0.9988$ , Bosch *et al.* 2002; Japanese 14 loci  $D=0.9987$  Uchihi *et al.* 2003; North Africans, 12 loci  $D=0.9605-0.9821$  Quintana-Murci *et al.* 2004). For most of the data on African populations included for comparisons, 8 or 9 Y-STRs (the “minimal haplotype”) or even less markers are available, thus limiting comparisons with our data. When the minimal set of 8 loci is considered, the haplotype diversity in Guineans decreases to  $0.9981\pm 0.0010$  (142 haplotypes), comparatively higher than data on Europeans ( $D=0.9972$ , Roewer *et al.* 2001), Afro-Americans ( $D=0.998$ , Kayser *et al.* 2003) and other sub-Saharanans ( $D=0.9900$ ; Trovoada *et al.* 2001, Alvarez *et al.* 2002). Previously published haplotypic data on ten Y-STR loci were used for an analysis of molecular variance (AMOVA) in selected populations (North Africa, Arredi *et al.* 2004; Equatorial Guinea, Arroyo-Pardo *et al.* 2005; Mozambique, Alves *et al.* 2003; North Portugal, Gusmao *et al.* 2002; and Spain, Zarrabeitia *et al.* 2003). Criteria defining three geographic regions attributed the vast majority of variance to the intrapopulation level (99.1%). Although not statistically significant, the among-group variance is of 0.34% ( $P=0.08504\pm 0.00762$ ) while the intragroup component displayed 0.56%. According to an exact test of population differentiation (10000 steps of Markov chain) the six considered populations are distinct.

In order to evaluate the discriminatory power of the extended haplotype, the haplotypic diversity was determined for sets of ten markers (minimal haplotype plus one marker). The additional marker causes a variation in haplotype diversity as follows: DYS437 ( $H=0.9982\pm 0.0010$ , 143 haplotypes), DYS438 ( $H=0.9986\pm 0.0009$ , 146 haplotypes) and DYS439 ( $H=0.9994\pm 0.0008$ , 153 haplotypes). The level of discrimination obtained by additional typing of DYS439 confirms its usefulness for forensic purposes (Gusmao *et al.* 2001, Bosch *et al.* 2002, Beleza *et al.* 2003), by lowering the random match probability. We should however notice that an increase on haplotype diversity when adding new markers is dependent not only on the locus diversity but on the degree of gametic association between markers and the haplotypes previously defined (Beleza *et al.* 2003).

For DYS19, DYS389I, DYS389II, DYS390, DYS391 and DYS392 the haplotype combinations 15-13-30-21-10-11 and 15-13-31-21-10-11 are quite common in our data (Table S9 H46-H55 and H60-H68, respectively). The examples mentioned above are, in fact,

all classified within haplogroup E3a\*-M2. We have verified that depending on the haplogroup assignment, alleles are not randomly associated; therefore particular haplotypic combinations persist while other loci have accumulated differences. In our particular data set, the allelic/haplotype assignment can therefore be a relevant predictor of haplogroup assignment: allele DYS438\*8 is only present in our A1 Y chromosomes; all seven A1 Y chromosomes are DYS392\*11-DYS437\*14-DYS438\*8 and the DYS389I\*13 and DYS391\*11 alleles are frequent (5 out of 7); the combination DYS437\*17, DYS438\*10 and DYS19\*15 is frequent in E1\* lineages (but also present in haplogroup E3a); the six E3b\* Y chromosomes showed no variation at DYS19\*13 and DYS437\*14; all of the E3b1 individuals harbored the profile DYS389I\*13-DYS438\*10. These are most likely close related lineages resulting from founder effects, especially in haplogroup A1 where most of the haplotypes are of Papel individuals. One may also associate these differences with mutation rate differences in single STR loci from SNP-haplogroup to haplogroup as well as inter-loci differences according to the allele repeat score <sup>(Carvalho-Silva *et al.* 1999, Dupuy *et al.* 2004)</sup>. However, in evolutionary studies the genetic sampling over hundreds of generations could lead to differences in repeat variation between loci within a haplogroup and between haplogroups at the same locus, regardless of their effective mutation rates, simply because the evolutionary trajectories at each locus are a random event which can and most likely will differ in different haplogroups <sup>(Zhivotovsky and Underhill 2005)</sup>. The Guinean dataset is far too small for any consistent interpretation in this subject. We would however like to alert for the interest of larger studies and with deeper resolution, in particular for African haplogroups given that there is a clear ascertainment bias towards non-African Y chromosomes. This may elucidate us on each haplogroup's history, contributing for a better resolution of many African lineages, for example within haplogroup E3a-M2.

## 11 - Combined analysis of Y chromosome and mtDNA haplogroups

The mtDNA and Y chromosome genetic analysis of Guineans was performed in the same male samples and thus renders the possibility of analyzing their combined maternal and paternal ancestry (matrix of pie charts depicted in Figure 32). The most outstanding characteristic is that the mtDNA variation is sparsely distributed among several L0-L3 haplogroups and sub-haplogroups, while the Y chromosome lineages belong mostly to E3a\*-M2. Not surprisingly, the largest percentage of the pool is retained by a combination of Y chromosome haplogroups E3a and E1 and mtDNA haplogroups L1b, L2a, L2b, L2c, L3b, L3d, all of major West Africa distribution and common to all Guinean ethnic groups, telling of



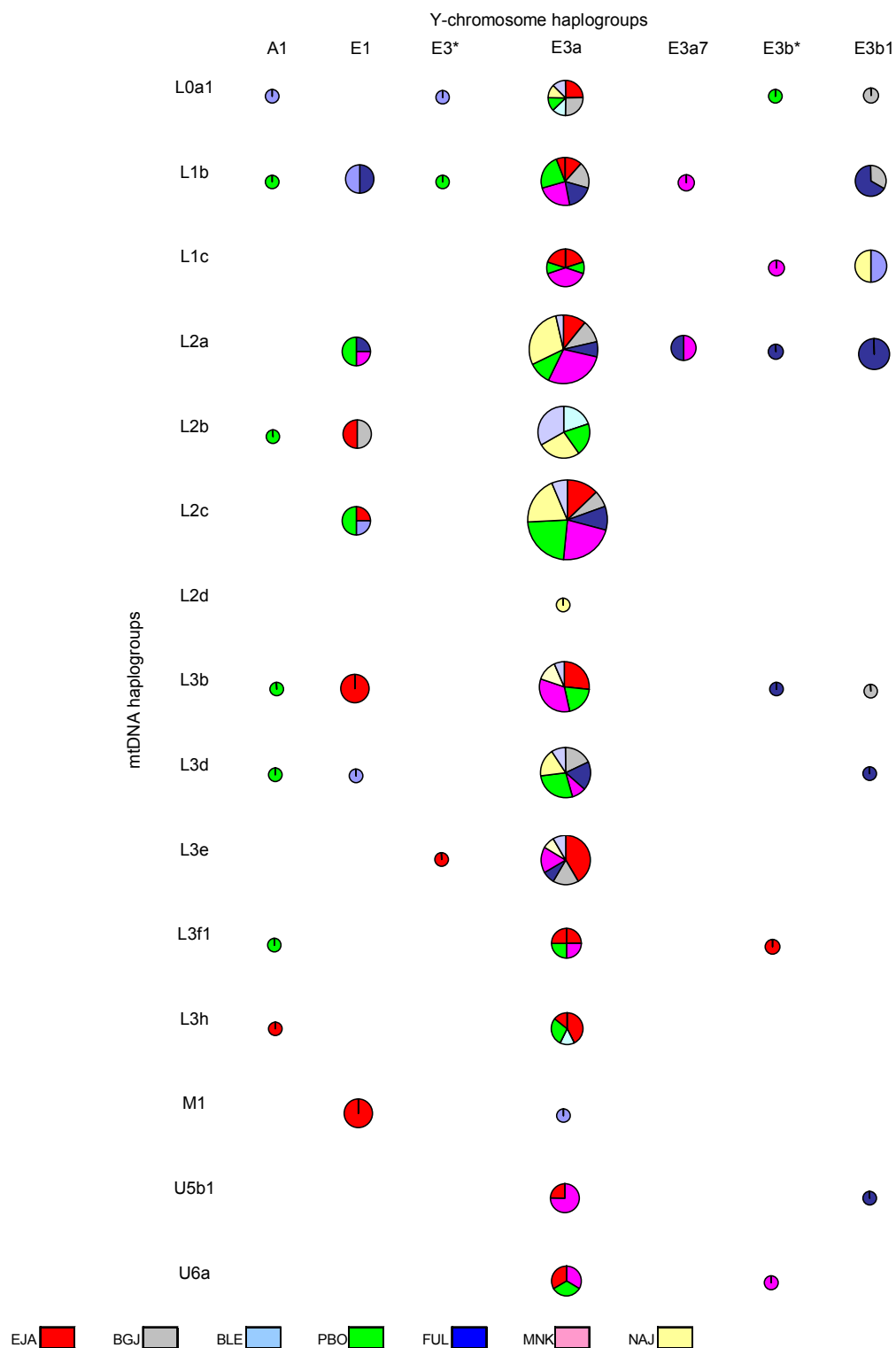


Figure 32 – Distribution of Y chromosome and mtDNA haplogroups of 214 male individuals belonging to the 7 Guinea-Bissau ethnic groups (assigned by colors as indicated, includes only individuals with determined extended microsatellite haplotype). The area of each chart is on linear scale to the number of individuals carrying the particular combination, the smallest representing single individuals.

their common evolutionary history and/or marked genetic flow. The Central-African E3a7 lineages displayed by Nalú, Balanta and Mandenka are curiously in combination with L1b1 and L2a1- $\beta$ 1 which can be Central African lineages in their origin, even though no exact matches are registered. A considerable proportion of lineages tell of non-West African maternal ancestors who have been integrated in the West African mtDNA pool, associated to E3a-M2 Y chromosomes: the East African L0a, L3e, L3f1 and L3h are common among Felupe-Djola while North African M1, U5b1 and U6 lineages are particularly frequent among Balanta and Fulbe. Though less frequent, the opposite is also detected, suggesting admixture of non-West African Y chromosomes with West African mtDNAs: in Papel, haplogroups A1 and E3\* are associated to several West African mtDNAs; one Felupe-Djola has a E2\* Y chromosome and a L2c mtDNA; Fulbe E2\* and E3b\* are shown in a L2c background. Nevertheless the most important finding is that Balanta, Papel and Felupe-Djola are the only people in Guinea-Bissau to show “pure” East African inheritance (L0a, L3e, L3f1 and L3h mtDNAs, combined with A1, A3b2, E3\* and E3b\* Y chromosomes), further supporting their East African origin. The E3b1 and R1b Y chromosomes testify for a most likely North African and/or European genetic imprint. The Fulbe displaying an E3b1 Y chromosome in combination with a U5b1 mtDNA is most likely of North African origin, though no firm interpretations can be made.

## 12 - Analysis of autosomal genetic markers in Guinea-Bissau ethnic groups

Our work team has previously surveyed fifteen autosomal STR (Powerplex 16, Promega) for a subset of 100 Guinea-Bissau individuals (Goncalves *et al.* 2002). All the ten population units (Beafada, Sussu and Mansonca here considered independently) were found to be in Hardy-Weinberg equilibrium and thus to be representative of the population (Fernandes A, *com. pess.*). Although no significant differences were found in overall allelic or haplotypic frequencies, when the ethnic groups were separately analyzed, the following could be distinguished on the basis of one or two loci: Bijagós differ from Papéis, Beafada (included in Felupe-Djola), Fulbe and Mandenka; Balanta are distinct from Felupe-Djola e Papéis; Felupe-Djola are by turn different from Fulbe and Mandenka; Papel differ from Mandenka; interestingly Mansonca and Sussu, for mtDNA and Y chromosome analysis included in the same group, were distinguished from Mandenka on the basis of D13S317 and VWA, respectively. In sum, Mandenka do not differ only from Beafada and Balanta. A NJ tree isolates Bijagós in the first branching (data not shown), indicating the differences in their pool. The next split includes a well-defined cluster with Felupe-Djola, Papel and Beafada,

while Fulbe show in an isolated branch. On the basis of their autosomal STR variation, the more closely related groups are Mandenka and Balanta.

Assuming that the autosomal genetic systems are gender-independent and reflect a combination of both maternal and paternal transmission we chose to contrapose the results with the mtDNA and Y chromosome data. In fact, there seems to be concordance in several aspects:

- Bijagós are distinct of the majority of the groups, perhaps due to insularity factors;
- Relatedness of Mandenka and Balanta, a putative consequence of a long term shared history;
- Papel and Felupe-Djola share large affinities and both differ from Mandenka, possibly on the account of lineages tracing back to East Africa.

Autosomal alleles which are not of frequent African assignment were found, for example the D3S1358\*13 in two Bijagós and two Felupe-Djola. It is interesting to note that one of this Bijagós harbors an R1b Y chromosome (H165) and mitochondrial L2a3 lineage (GB58), most certainly descending from a European father and a sub-Saharan mother. The 10 Y-STR profile further confirms its source as matches numerous European populations in YHRD database. As for the other Bijagó the maternal component is typically sub-Saharan (L0a1, GB4) while the paternal counterpart can either be European or North African (E3b1, H161; DYS439\*12). The Felupe-Djola samples harbor both mtDNA and Y chromosomes very common in the sub-Saharan West corner (GB85–H167 and GB114–H99). Nevertheless, the first has the D3S1358\*12,13 alleles not found among any other Guinean and possibly tracing a different origin. In fact, it is known that when analyzing genetic systems of uniparental inheritance one of the sides of the progenitor's story is lost, but can be conserved in autosomes due to recombination. Therefore, even if the parental types do not show Caucasoid or Berber variants, a link can be established. Alternatively, this can be a case of allelic reversion in a STR "fast-mutating" marker.

The D21S11\*29.2, typically present at low frequency in US Caucasoids and North African Berbers <sup>(Holt *et al.* 2000, Budowle *et al.* 2001)</sup>, was found in Bijagós and Beafada. The latter harbors an A1 Y chromosome (H2 in Table S9) which could have migrated from North Africa, though no exact matches were found. Similar enumerations can continue for VWA\*12 found in Mandenka and Fulbe, two groups historically related with Berber people, and whose autosomal allele is common in Madeira and Azores <sup>(Fernandes *et al.* 2002a, Velosa *et al.* 2002)</sup>. The Fulbe harboring the GB191 mtDNA lineage matches several North Africans and non-Africans (see Table S5), a putative link to the non-West African origin of its VWA\*12 allele.



## Chapter Six

### Final remarks

From the perspectives of both the extant maternal and paternal genetic pools, the Guinea-Bissau people can undoubtedly be classified as West sub-Saharan Africans. The majority of the genetic lineages of every one of the ethnolinguistic groups considered was found to belong to the West African specific and most prevalent sub-clusters of mtDNA L0-L3 and Y chromosome E3a-M2 and E1-M33 haplogroups <sup>(Rosa et al. 2004, 2007)</sup>. Their profile is, in terms of frequency and diversity levels, comparable to those of their neighboring populations, with whom they form clusters encompassing short genetic distances. The genetic background of the ethnic groups in Guinea-Bissau shows no preferential association to geography, linguistics or religion, and though diverse, seems to be uniformly distributed among many of the sub-Saharan populations, particularly those in Central-West Africa <sup>(Rosa et al. 2004, 2007)</sup>.

Comparing the diversity of both uniparental systems is, however, hindered because of the very different mutational properties of their SNPs and Y microsatellites, and because of SNP ascertainment bias on the Y chromosome. Therefore, caution is needed when interpreting the results. When paying attention to the high paternal homogeneity, on the account of the E3a-M2 Y chromosomes, one may speculate on a comparatively lower population size, gender differences in reproductive success or on a major expansion from a pool of limited paternal genetic diversity (in terms of haplogroups), that has overwhelmed less frequent variants. A possible scenario to cause such variation relates to the social practices of male polygamy and patrilocal exogamy, which predominate in most of the ethnic societies in Guinea-Bissau. In that sense the mtDNA variability tends to be maintained and flows among the groups (progeny included in the father's ethnic group) while the paternal variability tends to decrease over time, with a biased transmission of only a limited portion of the pool. However, we believe rather that the overall paternal homogeneity is at least partially illusory due to different rates and modes of evolution of the Y chromosomal polymorphisms, with subclades of phylogenetic importance still to be discriminated, in particular within haplogroup E3a\*-M2.

One should not expect the extant genetic pool, even that of autochthonous people, to directly reflect the demographic events within a geographical region since its initial occupation to recent historical episodes. Also, the time of colonization can not generally be inferred from the coalescence time of the genetic lineages since, unless a drastic founder

effect has occurred the age of the clades often greatly predates the age of the ethnically defined population(s) in which it is found. For instance, haplogroups L1c and L2d are the clades of deepest coalescence in Guinea-Bissau (~110-120 kya; <sup>Rosa *et al.* 2004</sup>). The absence of founder lineages in our dataset plus the proposed Central African origin for such clades (Salas *et al.* 2002, Batini *et al.* 2007) leads us to hypothesize on their westwards expansion relatively late in the evolution of the haplogroups; thus the migrants already carried along substantial molecular diversity, arriving in size(s), capable to maintain it. Their ragged multimodal mismatch distribution, on the account of distantly separated sub-clades, further supports their ancestry.

In Bandelt, Macaulay and Richards' words "only when reconstructed and dated ancestral types appear to have given rise to essentially autochthonous branches of the phylogeny, with approximately equal coalescence time, then one could speak of founder types at the colonizing event" (Bandelt *et al.* 2006). The archaeological findings support a permanent occupation of West Africa by modern humans from 30-40 kya onwards (Mercader and Martí 2003) or possibly even earlier (Phillipson 1993, Newman 1995, Foley and Lahr 1997, Cornelissen 2002). Such archaeological evidence is concordant with the coalescence ages of indigenous mtDNA clades, namely L1b, L2b, L3b and L3d (<sup>Rosa *et al.* 2004</sup>). The Guinean E3a-M2 and E1-M33 NRY variation, coalescing 20-30 kya at the most (<sup>Rosa *et al.* 2007</sup>), is much less consistent within the non-genetic evidence on the initial colonization of West Africa, but again these differences are likely inherent to the systems.

The climatic oscillations between 40-12 kya have caused the expansion and fragmentation of the equatorial forest (<sup>Adams 1997, Lahr and Foley 1998, Cornelissen 2002</sup>), an ecological scenario able to reduce the genetic diversity of AMHs, and generate a fragmented pattern of population distribution, ultimately with genetic sub-structuring, even if to assume earlier more homogeneous spread of the variation. The first inhabitants of West Africa were supposedly dispersed in small and isolated hunter-gatherer groups, their genetic pool unlikely to have been uniform. South of the Sahelian strip, in the vicinity of Guinea-Bissau, a vegetation zone was conserved throughout the climatic oscillations 23-15 kya to 9 kya (<sup>Adams and Faure 1997</sup>), and therefore could have acted as a refugium, conserving genetic diversity. We observe today an overall similarity in the mtDNA and Y chromosome profile of West Africans, which could be due to a common basis and co-evolution of the genetic diversity that has emerged from the refugium. The return to moister and warmer conditions culminated in the Sahara's wet phase ~9kya (<sup>Aumassip *et al.* 1994</sup>), which likely promoted population growth and massive displacement of people, reaching previously uninhabited areas and allowing contact and admixture with isolated before populations (<sup>Camps 1974, Hassan 1978, Dutour *et al.* 1988, Clark 1994</sup>).

Many of the haplogroups and sub-haplogroups in our sample coalesce at about 20 kya and show signs of expansion. From a limited number of founders that are shared between populations of different ethnolinguistic affiliations, haplogroup L2c and subclades of L2a show high haplotype diversity and an almost starlike topology (with many one-mutational step derived haplotypes in L2a1), suggestive of population growth. The Tajima's selection test, here designed to evaluate fluctuations in population size since the hypothesis of selection is neglected, gave significant negative results for L0a, L2a and L2c, whereas the less conservative Fu's  $F_S$  further indicated L1b, L3b1 and L3d as candidates of expansion. The unimodal and bell-shaped pairwise mismatch distribution of these haplogroups is consistent evidence of population growth <sup>(Harpending *et al.* 1993, Schneider and Excoffier 1999)</sup>. However, the size of African human populations should have slowly increased in the Pleistocene, limited by the available fauna. At about 14 kya, when animals were driven near extinction by hunting, the impetus for people to adopt cultivation as a subsistence strategy was triggered and foraging was progressively abandoned <sup>(Cohen 1989)</sup>. By 6 kya centers in the Sahel were cultivating local crops, West Africa at that time being classified as a temperate sub-tropical zone, usually selected as agricultural centres. The widely represented E3a-M2 is a likely genetic marker for the agricultural expansions in sub-Saharan Africa in the last 2-3 ky, especially in the context of the Bantu migrations <sup>(Passarino *et al.* 1998, Underhill *et al.* 2001a)</sup>. Although clear founders are absent, its high frequency and microsatellite diversity in Guineans and nearby populations hint at an early West African origin and expansion (many lineages differ by one mutational step and are shared by the ethnic groups), sometime in the last 20 ky. This fits well with suggestions of a local agricultural centre perhaps as early as 6-9 kya <sup>(Atherton 1972, Calvocoressi and David 1979, Clark 1994)</sup>, with better nutrition supporting a larger number of people. The advent of cultivation obviously drove to numerical growth but one cannot distinguish which subsets of the extant variation have been generated by the post-LGAM return to more beneficial conditions or by the subsequent shift to agriculture. More likely, the population growth happened in a continuous timescale since the climatic return to more stable conditions and became more accentuated with the introduction of agriculture and iron-smelting techniques, with gene flow obscuring earlier diversity and any differences accumulated during the isolation periods.

Less frequent clades summing up to 6% of each uniparental genetic profile may tell about non-West African influences. The mtDNA L3e2a and L3e2b lineages are thought to be successful hitchhikers of population movements in the Sahara in the early Holocene and the Great Wet Phase <sup>(Muzzolini 1993, Bandelt *et al.* 2001)</sup>. The NRY A1-M31 lineages and mtDNA L0a1 and L3h, East African in their origin, could have participated in such movements and have reached West Africa, where they have differentiated in isolation and have given rise to

specific subsets. Additionally, the L3e4 clade sampled among Guinean people and coalescing at about 11 kya, was associated with the West African expansion due to the rise of food-production and iron-smelting (Bandelt *et al.* 2001). Haplogroups U6a and M1b, of North West African origin and prevalent distribution (Corte-Real *et al.* 1996, Macaulay *et al.* 1999b, Plaza *et al.* 2003, Olivieri *et al.* 2006), have likely crossed the Sahara in more ancient times and not at the time of contacts reported in history (Moreira 1964), otherwise a random assortment of Eurasian haplogroups that exist in Northwest Africa at a fairly high frequency would have been carried by the migrants. The particular GB191 U6a lineage is thought to have registered partial diffusion to the Sahel at about 11 kya (Rando *et al.* 1998, Coia *et al.* 2005). We have to bear in mind that the Mauritania/Senegal and Mali border seems to be an important barrier to southward gene flow of the North African Euroasiatic haplogroups to sub-Saharan regions (Gonzalez *et al.* 2006). Although U5b reaches its main radiation in Europe, the U5b1b mtDNAs in Guinea-Bissau are one or few steps away from a widespread type in Europeans, Moroccans, West Saharans and Tunisians (Plaza *et al.* 2003, Rosa *et al.* 2004, Tambets *et al.* 2004, Achilli *et al.* 2005). These are supposedly post-glacial signatures of lineages that have crossed the strait of Gibraltar towards Northwest Africa, and further develop into local clusters, one of which is in West Africa (Rando *et al.* 1998, Rosa *et al.* 2004, Cerny *et al.* 2006, Ely *et al.* 2006). Its sub-Saharan spread could have been mediated by the Berbers or by Berber related people, like the Fulani (Rosa *et al.* 2004, Achilli *et al.* 2005). The E3b1- $\beta$  haplotypes are the only paternal lineages that we can with more confidence identify as North African contributions (Cruciani *et al.* 2004, Rosa *et al.* 2007). Therefore, evidence of gene flow from East and North African populations is found in both maternal and paternal pools of Guinea-Bissau. Even though their precise origin can not be indicated, and a North African source cannot be disregarded, the exact matches of R1b-P25 and a few E3b1-M78 are with Europeans. These may relate to the times of the slave trade and are in agreement with historical records, which describe a predominantly male presence that nevertheless did not leave a strong imprint in Guineans (Teixeira da Mota 1954). Note that no European mtDNA haplotypes, that could in principle have been introduced at the time of the slave trade, were found in our dataset.

The evolutionary relationships based on uniparentally transmitted polymorphisms are primarily concerned with the history of genes and not of populations. Therefore, any suggestions on a population-based phylogeographic approach are mere hypothesis built on the genetic evidence, and corroborated or not by non-genetic data. Inferences are made even more difficult in African populations, where ethnicities are deeply structured by social patterns of admixture (like endogamy or patrilocal exogamy), which blur the mtDNA and NRY inheritance within the ethnic units. The genetic background of the Guinean Mandenka shows a high frequency of clades testifying to expansion, namely the mtDNA L2c which has an almost starlike phylogeny (Rosa *et al.* 2004) and the E3a-M2 Y chromosomes, which harbor in the



Mandenka the highest microsatellite diversity among Guineans <sup>(Rosa *et al.* 2007)</sup>. While the Fu's  $F_s$  index indicates population growth in the Mandenka, the more stringent Tajima's  $D$  is marginally significant. Although the Mandenka are among the last well-defined ethnic groups to have arrived to Guinea-Bissau <sup>(Carreira and Quintino 1964)</sup>, these are West African indigenous people which could have been involved in a marked population increase, possibly due to their food-producing economy. Their ancestors may even relate to the people who instigated farming expertise in West Africa. This is strengthened by the historical records that tell of Guinean Mandenka as physically and culturally descendants of the Mande, the protagonists of agricultural population expansions in the Niger/Mali/Burkina-Faso region <sup>(Cavalli-Sforza *et al.* 1994)</sup> and rulers of the West African Black Empires based on trade and agriculture <sup>(Fage 1995)</sup>.

While some studies suggest linguistic affinities between Balanta and the Sudanese family, their spread related to that of Cushitic migrants <sup>(Quintino 1964)</sup>, others hypothesize on their common origin with Bantu, near the Nile in the Late Pleistocene <sup>(Stuhlmann 1910)</sup>. Minor traces of East African lineages corroborate the non-genetic evidence in claiming their easternmost origin. The mtDNA pool of the Balanta shows an increased frequency of sub-haplogroup L0a1. The subset of L0a1 variation that has reached Guinea-Bissau coalesces relatively recently at 7 kya, and exhibits a corridor of matches in a possible East-to-West route of migration only at the level of the founder haplotype <sup>(Rosa *et al.* 2004)</sup>. Their coalescence time might then reflect their arrival in the Holocene (when post-LGAM conditions were more favorable for migrations in the Sahelian strip). Again, in L2a- $\alpha$ 3 the GB44 motif traces a corridor of matches from East to West Africa. A link of Balanta and Sudanese-speakers is traceable in A3b2-M13 and E3\* Y chromosomes <sup>(Rosa *et al.* 2007)</sup>, found to be frequent among Sudanese and Ethiopians <sup>(Underhill *et al.* 2000, Semino *et al.* 2002)</sup>. The high frequency and microsatellite diversity of E3a\*-M2 in the Balanta attests to a more pronounced expansion of the paternal variation, comparatively to other Guineans (except Mandenka), a population increase which could be related to the farming practices. Even if there are no firm archaeological indications that early Holocene sorghum or millets were being domesticated, the spread of the Sudanic people at that time may be an example of farming/language dispersal <sup>(Ehret 1997, Ehret 2003)</sup>. This dispersal could have extended to all the Sahara, including West Sahara, with later introgressions to the Niger-Congo speakers <sup>(Bellwood 2005)</sup>. Under such model, and together with the genetic evidence, the Balanta's Sudanese origin gains relevance. A common origin with the Bantu, one of the most notable people in the sub-Saharan agricultural context, may suggest that different peoples jointly learnt agricultural techniques, and thus be a support for the expansion observed in the paternal pool of the Balanta. Particular relationships of these Guineans and North Africans are found in exact matches in

haplogroup L2a, L2b and L3b, typical of West Africa and the North African M1b (Plaza *et al.* 2003, Rosa *et al.* 2004, Olivieri *et al.* 2006)

Although the Bantu people are known to have been the main drivers of the sub-Saharan agricultural spread towards South Africa, the absence of mtDNA Bantu-associated markers (Soodyall *et al.* 1996, Watson *et al.* 1997, Bandelt *et al.* 2001, Pereira *et al.* 2001b) in the West African sample sets suggests either that the Bantu contributed very little to the gene pool of Guineans or that they had a distinct gene pool from that associated with the southwards migrations (Rosa *et al.* 2004). Since none or few Bantu-speakers today inhabit West Africa, and the Mandenka and Balanta display evidence of a particularly marked recent population growth, these are the best candidates among Guineans to reflect the demographic effects of the agriculturalist lifestyle, putatively related to the people that introduced early cultivation practices into West Africa or at least those that have experienced a particular benefit from food production. The autosomal STR profiles of Balanta and Mandenka share large affinities (Goncalves *et al.* 2002), possibly as a consequence of a long-term shared history.

There is an intriguing line of evidence in mtDNA haplogroups typically frequent among the Fulani (e.g. L1b), with a few Fulani-exclusive haplotypes. Moreover, exact matches among mtDNA Fulani lineages inhabiting a broad geographic area in West-Central Africa were found in a background of several haplogroups (Watson *et al.* 1996, Destro-Bisol *et al.* 2004, Rosa *et al.* 2004, Cerny *et al.* 2006). This most likely tells of a common ancestry, with lower differentiation of the maternal pool among nomadic Fulani, while the settled communities tend to accumulate differences by gene flow with their geographic neighbors (Cerny *et al.* 2006). The high proportion of haplogroup L2c in Guinean Fulbe contrasts with other Fulani and better relates them with their West African neighbors. This feature is not evident in their Y chromosomal haplogroups, probably on the account of the lower resolution of lineages. Nonetheless, both maternally and paternally inherited pools of Fulbe are significantly different from other Guineans, especially from those showing non-West African traces. Although geography is the main dictator of matches, the Fulbe mtDNAs and Y chromosomes match widely from West-Central to South Africa, thus supporting the broad distribution and multiple origins of their ancestors (Rosa *et al.* 2004, 2007). The U5b1b haplotypes found in our Fulbe sample set are a curious link to the European post-glacial population recovery (Achilli *et al.* 2005). The non-random distribution of haplogroup U5 in the Fulani people suggests a correlation between genetic and linguistic affiliation, and provides evidence of the link between these people and North Africans. Again, E3b1- $\beta$  is a Northwest African contribution (Cruciani *et al.* 2004) to the Guinean Fulbe paternal variation.

Minor imprints of more eastern and central areas of the Sahel are represented in particular mtDNA and NRY haplotypes found in the Papel and Felupe-Djola people (e.g.

L3e2a, L3e2b and L3h, and A1-M31, E2-M75 and E3-PN2, respectively). Curiously these are the ethnic groups in Guinea-Bissau with an oral tradition claiming an origin in Sudan<sup>(Quintino 1969)</sup> and, together with the Balanta, are both affiliated to the Niger-Congo Bak-speakers. Their paternal pool is among the more diverse in our dataset, with the lowest frequency of E3a-M2 among Guineans<sup>(Rosa et al. 2007)</sup> indicating a shallow residence time in the territory and/or that the paternal genetic flow has not allowed greater homogeneity. The increased frequency of NRY haplogroup E1-M33 in Papel and Djola, responsible for a higher dissimilarity of their pool in a PCA of all Guineans, is possibly an amplifying effect of West African lineages on founder groups arriving from an easternmost source<sup>(Rosa et al. 2007)</sup>. Such unconventional frequencies of West African lineages due to genetic drift might have their parallel in mtDNA L3b and M1 in Felupe-Djola, U6 in Papel and L2b in both<sup>(Rosa et al. 2004)</sup>. Furthermore, these groups retain statistically significant differences from other Guineans in maternal, paternal and autosomal pool<sup>(Goncalves et al. 2002, Rosa et al. 2004)</sup>.

In the Guinean genetic context, the mtDNA profile of the Bijagós displays a marked similarity to that of the Fulbe, with many shared lineages. This is also true for the Y chromosomal pool, hindering any claimed similarity to the Djola, Papel or Nalú<sup>(Teixeira da Mota 1954)</sup> or even Egyptians<sup>(Quintino 1964)</sup>. It is interesting to note that these are the only people in Guinea-Bissau whose mitochondrial phylogenetic pool does not exhibit significant differences from those of other Fulani, further suggesting a genetic proximity of both groups. The HVS-I sequence diversity is the lowest in the Bijagós<sup>(Rosa et al. 2004)</sup>, not surprising given that these are the main inhabitants of the archipelago, possibly arriving through one or few founder effects, with a subsequent isolation due to their islander condition.

Besides finding that a high frequency of haplogroup L2b better distinguishes the Nalú maternal pool from other Guineans, no consistent inferences on their genetic pool were achieved by our analysis. They are integrated in the West African genetic diversity, in Guinea-Bissau only retaining significant differences from the Fulbe maternal profile, and do not show any mtDNA or Y chromosomal lineages linked to the East African variation. Although this does not constitute as a firm genetic evidence, it nonetheless does not contradict the idea that they are an Guinea-Bissau autochthonous population<sup>(Teixeira da Mota 1954)</sup>. The DE\* Y chromosome in their pool seems to descend from a private lineage in Nigerians<sup>(Weale et al. 2003)</sup> and may represent a coalescent node of variation paraphyletic to haplogroup D, E or both.

The extant mtDNA and Y chromosome pool is a net outcome of many past events, where episodes in recent history are not readily detectable, and are often not the main focus of attention of population geneticists. Therefore, analyzing the effects of 'Balantization' and 'Sudanization' processes in the extant genetic pool is not straightforward. These processes

are very recent, taking place in the last four or five generations <sup>(Carreira and Meireles 1959)</sup>, and would have only become evident in the modern variability if there was overwhelming intense and directional gene flow between very distinct pools. These phenomena in Guineans involve lineages of a largely common West African pool, established before the definition of the ethnolinguistic units, and certainly long before this process of directed admixture. Despite the obvious sociocultural differences among Guinea-Bissau ethnic groups, marked by the supposedly strict admixture barriers, their genetic pool remains largely shared, because of common ancestry and/or a common history of genetic admixture without language shift.

## References

- Accetturo M *et al* (2006) Human mtDNA site-specific variability values can act as haplogroup markers. *Hum Mutat* 27 (9):965-974.
- Achilli A *et al* (2004) The Molecular Dissection of mtDNA Haplogroup H Confirms That the Franco-Cantabrian Glacial Refuge Was a Major Source for the European Gene Pool. *Am J Hum Genet* 75 (5):910-918.
- Achilli A *et al* (2005) Saami and Berbers - an unexpected mitochondrial DNA link. *Am J Hum Genet* 76 (5):883-886.
- Adams JM (1997) Global land environments during the last interglacial. Oak Ridge National Laboratory: TN, USA.
- Adams JM and Faure H (1997). Review and Atlas of Palaeovegetation: Preliminary land ecosystem maps of the world since the Last Glacial Maximum. February 2007. [<http://www.escl.ornl.gov/ern/qen/adams1.html>].
- Agarwal A *et al* (2000) Hemoglobin C associated with protection from severe malaria in the Dogon of Mali, a West African population with a low prevalence of hemoglobin S. *Blood* 96 (7):2358-2363.
- Agunlik AI *et al* (1998) Evolution of the DAZ gene family suggests that Y-linked DAZ plays little, or a limited, role in spermatogenesis but underlines a recent African origin for human populations. *Hum Mol Genet* 7 (9):1371-1377.
- Alimen H (1987) Evolution du climat et des civilisations depuis 40000 ans du nord au sud du Sahara occidental, Premières conceptions confrontées aux données récentes. *Bull L'Assoc Franç L'Étude Quaternaire* 4:215-227.
- Almada AA (1964) Tratado breve dos rios da Guiné e Cabo Verde. 2nd ed. Editorial LIAM: Lisbon.
- Almeida A (1939) Sobre a etno-economia da Guiné Portuguesa. Vol. 15. *Bol Cult Guiné Port*: Lisbon.
- Alvarez S *et al* (2002) STR data for nine Y-chromosomal loci in Guinea Equatorial (central Africa). *Forensic Sci Int* 127 (1-2):142-144.
- Alves C *et al* (2003) Evaluating the informative power of Y-STRs: a comparative study using European and new African haplotype data. *Forensic Sci Int* 134:126-133.
- Alves-Silva J *et al* (2000) The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* 67 (2):444-461.
- Ammerman AJ, Cavalli-Sforza LL (1984) The Neolithic transition and the genetics of populations in Europe. Princeton University Press: Princeton, New Jersey.
- Anderson S *et al* (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290 (5806):457-465.
- Andersson M *et al* (1988) Y:autosome translocations and mosaicism in the aetiology of 45,X maleness: assignment of fertility factor to distal Yq11. *Hum Genet* 79 (1):2-7.
- Andersson SG *et al* (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396 (6707):133-140.
- Andersson SG *et al* (2003) On the origin of mitochondria: a genomics perspective. *Philos Trans R Soc Lond B Biol Sci* 358 (1429):165-177.
- Andrews P (1992) Evolution and environment in the *Hominoidea*. *Nature* 360 (6405):641-646.
- Andrews RM *et al* (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 23 (2):147.
- Arctander P (1999) Mitochondrial recombination? *Science*. 284 (5423):2090-2091.
- Arredi B *et al* (2004) A predominantly Neolithic origin for Y-chromosomal DNA variation in North Africa. *Am J Hum Genet* 75 (2):338-345.
- Arroyo-Pardo E *et al* (2005) Genetic variability of 16 Y-chromosome STRs in a sample from Equatorial Guinea (Central Africa). *Forensic Sci Int* 20 (149):109-113.
- Atherton JH (1972) Excavations at Kamabai and Yagala Rock Shelters, Sierra Leone. *West Afr J Archaeol* 2:39-74.
- Aumassip G *et al* (1994) Le milieu saharien aux temps préhistoriques. In: Aumassip G (ed) Milieux, hommes et techniques du Sahara préhistorique. Problèmes actuels. L'Harmattan: Paris, pp 9-29.
- Avise J (2000) Phylogeography: The History and Formation of Species. Harvard University Press: Cambridge, Massachusetts.
- Avise J *et al* (1987) Intraspecific phylogeography: the molecular bridge between population genetics and systematics. *Ann Rev Ecol Syst* 18:489-522.
- Awadalla P, Eyre-Walker A, and Smith JM (1999) Linkage disequilibrium and recombination in hominid mitochondrial DNA. *Science* 286 (5449):2524-2525.
- Ayub Q *et al* (2000) Identification and characterization of novel human Y-chromosomal microsatellites from sequence database information. *Nucleic Acids Res* 28 (2):e8.
- Bamshad M *et al* (2001) Genetic evidence on the origins of Indian caste populations. *Genome Res* 11 (6):994-1004.

- Bandelt H-J and Forster P (1997) The myth of bumpy hunter-gatherer mismatch distributions. *Am J Hum Genet* 61 (4):980-983.
- Bandelt H-J *et al* (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. *Mol Biol Evol* 16 (1):37-48.
- Bandelt H-J, Macaulay V, and Richards M (2000) Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol Phylogenet Evol* 16 (1):8-28.
- Bandelt H-J *et al* (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann Hum Genet* 65 (Pt 6):549-563.
- Bandelt H-J *et al* (2002) The fingerprint of phantom mutations in mitochondrial DNA data. *Am J Hum Genet* 71 (5):1150-1160.
- Bandelt H-J, Macaulay V, Richards M (2003) What molecules can't tell us about the spread of languages and the Neolithic. In: Bellwood P, Renfrew C (eds) *Examining the Farming/Language Dispersal Hypothesis*. McDonald Institute for Archaeological Research: Cambridge, pp 99-111.
- Bandelt H-J *et al* (2006) Estimation of mutation rates and coalescence times: some caveats. In: Bandelt H-J, Macaulay V, Richards M (eds) *Human Mitochondrial DNA and the Evolution of Homo sapiens*. Springer-Verlag: Berlin, Heidelberg, pp 149-179.
- Bar-Yosef O (2002) The Upper Paleolithic revolution. *Annu Rev Anthropol* 31:363-393.
- Bar-Yosef O *et al* (1986) New data on the origin of Modern Man in the Levant. *Curr Anthropol* 27:63-64.
- Barbujani G and Goldstein DB (2004) Africans and Asians abroad: genetic diversity in Europe. *Annu Rev Genomics Hum Genet* 5:119-150.
- Barros A (1947) A invasão fula da circunscrição de Bafatá. Queda dos beafadas e Mandingas, tribos "Gabungabé". Vol. 15. 6 ed. Bissau.
- Batini C *et al* (2007) Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. *Mol Phylogenet Evol* 43 (2):635-644.
- Beck JW *et al* (2001) Extremely large variations of atmospheric <sup>14</sup>C concentration during the last glacial period. *Science* 292 (5526):2453-2458.
- Beleza S *et al* (2003) Extending STR markers in Y chromosome haplotypes. *Int J Legal Med.* 117 (1):27-33.
- Bellwood P (2001) Early agriculturalists population diasporas? Farming, languages and genes. *Annual Review in Anthropology* 30:181-207.
- Bellwood P (2005) *First farmers: the origins of agricultural societies*. Blackwell Publishing: Oxford.
- Bendall KE *et al* (1996) Heteroplasmic point mutations in the human mtDNA control region. *Am J Hum Genet* 59 (6):1276-1287.
- Bensasson D *et al* (2001) Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol* 16 (6):314-321.
- Bensasson D, Feldman MW, and Petrov DA (2003) Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *J Mol Evol* 57 (3):343-354.
- Berg OG and Kurland CG (2000) Why mitochondrial genes are most often found in nuclei. *Mol Biol Evol* 17 (6):951-961.
- Blench R (1993) New developments in the classification of Bantu languages and their historical implications. In: Barreteau D and Graffenried CV (eds) *Datation et Chronologie dans le Bassin du Lac Tchad*. ORSTOM: Paris, pp 147-160.
- Bogdanov AJ *et al* (1999) Treatment of experimental brain tumors with trombospondin-1 derived peptides: An in vivo imaging study. *Neoplasia* 1:438-445.
- Bogenhagen DF (1999) Repair of mtDNA in vertebrates. *Am J Hum Genet* 64 (5):1276-1281.
- Borensztajn K *et al* (2002) Characterization of two novel splice site mutations in human factor VII gene causing severe plasma factor VII deficiency and bleeding diathesis. *Br J Haematol* 117:168-171.
- Bosch E *et al* (2000) Y chromosome STR haplotypes in four populations from northwest Africa. *Int J Legal Med* 114 (1-2):36-40.
- Bosch E *et al* (2001) High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula. *Am J Hum Genet* 68 (4):1019-1029.
- Bosch E *et al* (2002) High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions. *Forensic Sci Int.* 125 (1):42-51.
- Bosch E *et al* (2004) Dynamics of a human interparalog gene conversion hotspot. *Genome Res* 14 (5):835-844.

- Bouzouggar A *et al* (2007) 82,000-year-old shell beads from North Africa and implications for the origins of modern human behavior. *Proc Natl Acad Sci U S A* 104 (24):9964-9969.
- Bowler JM *et al* (2003) New ages for human occupation and climatic change at Lake Mungo, Australia. *Nature* 421 (6925):837-840.
- Brandstatter A, Niederstatter H, and Parson W (2004a) Monitoring the inheritance of heteroplasmy by computer-assisted detection of mixed basecalls in the entire human mitochondrial DNA control region. *Int J Legal Med* 118 (1):47-54.
- Brandstatter A *et al* (2004b) Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database. *Int J Legal Med* 118 (5):294-306.
- Brehm A *et al* (2002) Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade. *Ann Hum Genet* 66 (Pt 1):49-60.
- Brinkmann B *et al* (1998) Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am J Hum Genet* 62 (6):1408-1415.
- Brown WM (1980) Polymorphism in mitochondrial DNA of humans as revealed by restriction endonuclease analysis. *Proc Natl Acad Sci U S A* 77 (6):3605-3609.
- Brown WM, George M, Jr., and Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. *Proc Natl Acad Sci U S A* 76 (4):1967-1971.
- Budowle B *et al* (2001) STR primer concordance study. *Forensic Sci Int* 124 (1):47-54.
- Bull JJ (1983) Evolution of Sex Determining Mechanisms. Benjamin Cummings: Menlo Park, California.
- Burke K, Durotoye AB, and Whiteman AJ (1971) A dry phase south of the Sahara 20000 years ago. *West Afr J Archaeol* 1:1-8.
- Burton ML *et al* (1996) Regions based on social structure. *Curr Anthropol* 37:87-123.
- Butzer KW, Brown FH, and Thurber DL (1969) Horizontal sediments of the lower Omo Valley: the Kibish Formation. *Quaternaria* 11:15-29.
- Calafell F *et al* (1998) Short tandem repeat polymorphism evolution in humans. *Eur J Hum Genet* 6 (1):38-49.
- Calvocoressi D and David N (1979) A new survey of radiocarbon and thermoluminescence dates for West Africa. *J Afr Hist* 20:1-29.
- Camps G (1974) Les Civilisations Préhistorique de l'Afrique du Nord et du Sahara. Paris, Doin.
- Camps-Fabrer, H (1989). Capsien et Natoufien au Proche-Orient. In: Colloque L'homme maghrebien et son environnement depuis 100.000 ans. Trav. Du CAPMO: Algeria, pp 71-104.
- Cann RL, Stoneking M, and Wilson AC (1987) Mitochondrial DNA and human evolution. *Nature* 325 (6099):31-36.
- Capelli C *et al* (2001) A predominantly indigenous paternal heritage for the Austronesian-speaking peoples of insular Southeast Asia and Oceania. *Am J Hum Genet* 68 (2):432-443.
- Carelli V *et al* (2002) Respiratory function in cybrid cell lines carrying European mtDNA haplogroups: implications for Leber's hereditary optic neuropathy. *Biochim Biophys Acta* 1588 (1):7-14.
- Carelli V *et al* (2006) Haplogroup effects and recombination of mitochondrial DNA: novel clues from the analysis of Leber hereditary optic neuropathy pedigrees. *Am J Hum Genet* 78 (4):564-574.
- Carreira A (1962) O fundamento dos etnónimos na Guiné Portuguesa. *Revista Garcia da Horta* 10:305-323.
- Carreira A (1983) Migrações nas Ilhas de Cabo Verde. 2nd ed. Instituto Caboverdeano do Livro: Lisbon.
- Carreira A and Meireles M (1959) Notas sobre os movimentos migratórios da população natural da Guiné-Portuguesa. *Bol Cult Guiné Port* XIV (53):7-20.
- Carreira A and Quintino FR (1964) Antroponímia da Guiné Portuguesa. Memórias da Junta de Investigação do Ultramar. Junta de Investigações do Ultramar: Lisbon.
- Carvalho CM *et al* (2003) Lack of association between Y chromosome haplogroups and male infertility in Japanese men. *Am J Med Genet* 116 (2):152-158.
- Carvalho CM *et al* (2004) Y-chromosome haplotypes in azoospermic Israeli men. *Hum Biol* 76 (3):469-478.
- Carvalho-Silva DR *et al* (1999) Divergent human Y-chromosome microsatellite evolution rates. *J Mol Evol* 49 (2):204-214.
- Casanova M *et al* (1985) A human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. *Science* 230 (4732):1403-1406.
- Caspersson T *et al* (1970) Identification of human chromosomes by DNA-binding fluorescent agents. *Chromosoma* 30 (2):215-227.

- Cavalli-Sforza LL and Feldman MW (2003) The application of molecular genetic approaches to the study of human evolution. *Nat Genet* 33 Suppl: 266-275.
- Cavalli-Sforza LL and Minch E (1997) Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 61 (1):247-254.
- Cavalli-Sforza LL, Menozzi P, Piazza A (1994) *The History and Geography of Human Genes*. Princeton University Press: Princeton.
- Cerny V *et al* (2004) mtDNA sequences of Chadic-speaking populations from northern Cameroon suggest their affinities with eastern Africa. *Ann Hum Biol* 31 (5):554-569.
- Cerny V *et al* (2006) MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations. *Hum Biol* 78 (1):9-27.
- Chakraborty R *et al* (1997) Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. *Proc Natl Acad Sci U S A* 94 (3):1041-1046.
- Chaubey G *et al* (2007) Peopling of South Asia: investigating the caste-tribe continuum in India. *Bioessays* 29 (1):91-100.
- Chen X *et al* (1995a) Rearranged mitochondrial genomes are present in human oocytes. *Am J Hum Genet* 57 (2):239-247.
- Chen YS *et al* (1995b) Analysis of mtDNA variation in African populations reveals the most ancient of all human continent-specific haplogroups. *Am J Hum Genet* 57 (1):133-149.
- Chen YS *et al* (2000) mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. *Am J Hum Genet* 66 (4):1362-1383.
- Chevret E *et al* (1997) Meiotic behavior of sex chromosomes investigated by three-colour FISH on 35,142 sperm nuclei from two 47,XYX males. *Hum Genet* 99 (3):407-412.
- Chikhi L *et al* (2002) Y genetic data support the Neolithic demic diffusion model. *Proc Natl Acad Sci U S A* 99 (17):11008-11013.
- Chinnery PF (2006) Mitochondrial DNA in *Homo sapiens*. In: Bandelt H-J, Macaulay V, Richards M (eds) *Human Mitochondrial DNA and the Evolution of Homo sapiens*. Springer: pp 3-15.
- Cinnioglu C *et al* (2004) Excavating Y-chromosome haplotype strata in Anatolia. *Hum Genet* 114 (2):127-148.
- Clark JD (1980) Human Populations and Cultural Adaptations in the Sahara and Nile during Prehistoric Times. In: Williams MA, Faure H (eds) *Quaternary Environments and Prehistoric Occupation in Northern Africa*. Rotterdam, pp 527-582.
- Clark JD (1994) Africa: From the appearance of *Homo sapiens sapiens* to the beginnings of food production. In: De Laet SJ *et al* (eds) *Volume I - Prehistory and the Beginnings of Civilization*. Routledge: New York, pp 191-206.
- Clayton DA (1992) Transcription and replication of animal mitochondrial DNAs. *Int Rev Cytol* 141:217-232.
- Cockburn TA (1971) Infectious diseases in ancient populations. *Curr Anthropol* 12:45-62.
- Cohen MN (1989) *Health and Rise of Civilization*. Yale University Press: New Haven.
- Coia V *et al* (2005) mtDNA variation in North Cameroon: lack of Asian lineages and implications for back migration from Asia to sub-Saharan Africa. *Am J Phys Anthropol* 128 (3):678-681.
- Collins FS, Guyer MS, and Charkravarti A (1997) Variations on a theme: cataloging human DNA sequence variation. *Science* 278 (5343):1580-1581.
- Coluzzi M *et al* (2002) A polytene chromosome analysis of the *Anopheles gambiae* species complex. *Science* 298:1415-1418.
- Cooke HJ and Noel B (1979) Confirmation of Y:autosome translocation using recombinant DNA. *Hum Genet* 50 (1):39-44.
- Cooke HJ, Brown WR, and Rappold GA (1985) Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. *Nature* 317 (6039):687-692.
- Cornelissen E (2002) Human responses to changing environments in Central Africa between 40,000 and 12,000 BP. *World Prehist* 16:197-235.
- Corte-Real HB *et al* (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* 60 (Pt 4):331-50.
- Coskun PE, Ruiz-Pesini E, and Wallace DC (2003) Control region mtDNA variants: longevity, climatic adaptation, and a forensic conundrum. *Proc Natl Acad Sci U S A* 100 (5):2174-2176.
- Cruciani F *et al* (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 70 (5):1197-1214.



- Cruciani F *et al* (2004) Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet* 74:1014-1022.
- Cruciani F *et al* (2006) Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. *Hum Mutat* 27 (8):831-832.
- Davis RE *et al* (1997) Mutations in mitochondrial cytochrome c oxidase genes segregate with late-onset *alzheimer* disease. *Proc Natl Acad Sci U S A* 94 (9):4526-4531.
- Day MH, Stringer CB (1982) A reconsideration of the Omo-Kibish remains and the erectus-sapiens transition. In: De Lumley MA (ed) *L'Homo erectus et la place de l'homme de Tautavel parmi les hominides fossiles*. (Centre National de la Recherche Scientifique: Nice, pp 814-846.
- de Knijff P (2000) Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am J Hum Genet* 67 (5):1055-1061.
- de Knijff P *et al* (1997) Chromosome Y microsatellites: population genetic and evolutionary aspects. *Int J Legal Med* 110 (3):134-149.
- Denaro M *et al* (1981) Ethnic variation in *Hpa* I endonuclease cleavage patterns of human mitochondrial DNA. *Proc Natl Acad Sci U S A* 78 (9):5768-5772.
- Destro-Bisol G *et al* (2004) Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol* 21 (9):1673-1682.
- Di Rienzo A and Wilson AC (1991) Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc Natl Acad Sci U S A* 88 (5):1597-601.
- Diallo T (1972) Les institutions politiques du Fouta Djallon. *Initiat Études Afr* 28.
- Diamond J (2002) Evolution, consequences and future of plant and animal domestication. *Nature* 418 (6898):700-707.
- Diamond J and Bellwood P (2003) Farmers and their languages: the first expansions. *Science* 300:597-602.
- Diez-Sanchez C *et al* (2003) Mitochondrial DNA content of human spermatozoa. *Biol Reprod* 68 (1):180-185.
- DiGiacomo F *et al* (2004) Y chromosomal haplogroup J as a signature of the post-Neolithic colonization of Europe. *Hum Genet* 115 (5):357-371.
- Dolo A *et al* (2005) Difference in susceptibility to malaria between two sympatric ethnic groups in Mali. *Am J Trop Med Hyg* 72 (3):243-248.
- Dupanloup I *et al* (2003) A recent shift from polygyny to monogamy in humans is suggested by the analysis of worldwide Y-chromosome diversity. *J Mol Evol* 57 (1):85-97.
- Dupanloup I *et al* (2004) Estimating the impact of prehistoric admixture on the genome of Europeans. *Mol Biol Evol* 21 (7):1361-1372.
- Dupuy BM *et al* (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum Mutat* 23 (2):117-124.
- Dupuy C (1999) Les apports de l'archéologie et de l'ethnologie a la connaissance de l'histoire ancienne des Peuls. In: Botte R, Boutrais J, Schmitz (eds) *Figure Peules*. Karthala: Paris, pp 53-72.
- Durham WH (1992) *Coevolution: Genes, Culture, and Human Diversity*. Stanford University Press: Stanford, California.
- Dutour O, Vernet R, Aumassip G (1988) Le peuplement préhistorique du Sahara. In: Aumassip G *et al* (eds) *Milieu, hommes et techniques du Sahara préhistorique. Problèmes actuels*. L'Harmattan: Paris, pp 39-52.
- Ehret C (1997) African languages: a historical survey. In: Vogel J (ed) *Encyclopaedia of Precolonial Africa*., pp 159-166.
- Ehret C (2003) Language family expansions: broadening our understandings of cause from an African perspective. In: Renfrew C, McMahon A, Trask L (eds) *Time depth in historical linguistics*. McDonald Institute for Archaeological Research: Cambridge.
- Elson JL *et al* (2001) Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age. *Am J Hum Genet* 68 (3):802-806.
- Elson JL, Turnbull DM, and Howell N (2004) Comparative genomics and the evolution of human mitochondrial DNA: assessing the effects of selection. *Am J Hum Genet* 74 (2):229-238.
- Ely B *et al* (2006) African-American mitochondrial DNAs often match mtDNAs found in multiple African ethnic groups. *BMC Biol* 4:34.
- Endicott P *et al* (2003) The genetic origins of the Andaman Islanders. *Am J Hum Genet* 72 (1):178-184.
- Excoffier L (1990) Evolution of human mitochondrial DNA: evidence for departure from a pure neutral model of populations at equilibrium. *J Mol Evol* 30 (2):125-139.
- Excoffier L (2000) Analysis of Population Subdivision. In: Balding D *et al* (eds) *Handbook of Statistical Genetics*. Wiley & Sons Ltd.

- Excoffier L and Langaney A (1989) Origin and differentiation of human mitochondrial DNA. *Am J Hum Genet* 44 (1):73-85.
- Excoffier L and Yang Z (1999) Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. *Mol Biol Evol* 16 (10):1357-1368.
- Excoffier L, Smouse PE, and Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* 131:479-491.
- Eyre-Walker A, Smith NH, and Smith JM (1999) How clonal are human mitochondria? *Proc R Soc Lond B Biol Sci* 266 (1418):477-483.
- Fage J (1995) A history of Africa. Routledge: London.
- Fagundes NJ *et al* (2002) Genetic, geographic, and linguistic variation among South American Indians: possible sex influence. *Am J Phys Anthropol* 117 (1):68-78.
- Felsenstein J (1988) Phylogenies from molecular sequences: inference and reliability. *Annu Rev Genet* 22:521-565.
- Fernandes AT *et al* (2002a) Genetic profile of the Madeira Archipelago population using the new PowerPlex16 System kit. *Forensic Sci Int* 125 (2-3):281-283.
- Fernandes S *et al* (2002b) High frequency of DAZ1/DAZ2 gene deletions in patients with severe oligozoospermia. *Mol Hum Reprod* 8 (3):286-298.
- Fernandes S *et al* (2004) A large AZFc deletion removes DAZ3/DAZ4 and nearby genes from men in Y haplogroup N. *Am J Hum Genet* 74 (1):180-187.
- Finnilä S, Lehtonen MS, and Majamaa K (2001) Phylogenetic network for European mtDNA. *Am J Hum Genet* 68 (6):1475-1484.
- Fitch W (1971) Toward defining the course of evolution: Minimum change for a specific tree topology. *Syst Zool* 20:406-416.
- Fitch W (1977) On the problem of discovering the most parsimonious tree. *Am Nat* 111:1169-1175.
- Fluxus Technology Ltd. Network 4.1.1.2 (2004). October 2006. [www.fluxus-engineering.com].
- Foley R (1998) The context of human genetic evolution. *Genome Res* 8 (4):339-347.
- Foley R and Lahr MM (1997) Mode 3 technologies and the evolution of modern humans. *Cambridge Archeol J* 7:3-36.
- Ford CE *et al* (1959) A sex-chromosome anomaly in a case of gonadal dysgenesis (Turner's syndrome). *Lancet* 1 (7075):711-713.
- Foster JW and Graves JA (1994) An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene. *Proc Natl Acad Sci U S A* 91 (5):1927-1931.
- Forster P (2004) Ice Ages and the mitochondrial DNA chronology of human dispersals: a review. *Philos Trans R Soc Lond B Biol Sci* 359 (1442):255-264.
- Forster P *et al* (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet* 59 (4):935-945.
- Forster P *et al* (2000) A short tandem repeat-based phylogeny for the human Y chromosome. *Am J Hum Genet* 67 (1):182-196.
- Forster P *et al* (2001) Phylogenetic star contraction applied to Asian and Papuan mtDNA evolution. *Mol Biol Evol* 18 (10):1864-1881.
- Forster P *et al* (2002) Continental and subcontinental distributions of mtDNA control region types. *Int J Legal Med* 116 (2):99-108.
- Fregeau CJ and Fourney RM (1993) DNA typing with fluorescently tagged short tandem repeats: a sensitive and accurate approach to human identification. *Biotechniques* 15 (1):100-119.
- Freije D *et al* (1992) Identification of a second pseudoautosomal region near the Xq and Yq telomeres. *Science* 258 (5089):1784-1787.
- Friedlaender J *et al* (2005) Expanding Southwest Pacific mitochondrial haplogroups P and Q. *Mol Biol Evol* 22 (6):1506-1517.
- Friedlaender J *et al* (2007) Melanesian mtDNA complexity. *PLoS ONE* 2:e248.
- Fu YX (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147:915-925.
- Fukuda M *et al* (1985) Mitochondrial DNA-like sequences in the human nuclear genome. Characterization and implications in the evolution of mitochondrial DNA. *J Mol Biol* 186 (2):257-266.
- Fullerton SM *et al* (1997) The genetic ancestry of modern humans: inferences from the analysis of DNA sequence diversity at the beta-globin locus. *Am J Hum Biol* 9:118.

- Galtier G (1981) Problèmes dialectologiques et phonographématiques des parlers mandingues. *Mandenkan I. Printemps* 1981:39-58.
- Gamble C *et al* (2004) Climate change and evolving human diversity in Europe during the last glacial. *Philos Trans R Soc Lond B Biol Sci* 359 (1442):243-253.
- Genetics Computer Group (GCG) (2005) Wisconsin Package Version 10.0. Madison, Wisconsin.
- Gerber AS *et al* (2001) Does non-neutral evolution shape observed patterns of DNA variation in animal mitochondrial genomes? *Annu Rev Genet* 35:539-566.
- Giles RE *et al* (1980) Maternal inheritance of human mitochondrial DNA. *Proc Natl Acad Sci U S A* 77 (11):6715-6719.
- Gill P *et al* (2001) DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *Forensic Sci Int* 124 (1):5-10.
- Gillespie JH (1991) The causes of molecular evolution. Oxford University Press: New York.
- Goldstein DB *et al* (1995) Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc Natl Acad Sci U S A* 92 (15):6723-6727.
- Goncalves R *et al* (2002) Genetic profile of a multi-ethnic population from Guiné-Bissau (West African coast) using the new PowerPlex 16 System kit. *Forensic Sci Int* 129 (1):78-80.
- Goncalves R *et al* (2003) Y-chromosome lineages in Cabo Verde Islands witness the diverse geographic origin of its first male settlers. *Hum Genet* 113 (6):467-472.
- Gonder MK *et al* (2006) Whole mtDNA Genome Sequence Analysis of Ancient African Lineages. *Mol Biol Evol.* 24(3):757-768.
- Gonzalez AM *et al* (2006) Mitochondrial DNA variation in Mauritania and Mali and their genetic relationship to other Western Africa populations. *Ann Hum Genet* 70 (Pt 5):631-657.
- Gonzalez-Neira A *et al* (2000) Distribution of Y-chromosome STR defined haplotypes in Iberia. *Forensic Sci Int* 110 (2):117-126.
- Goodman M *et al* (1998) Toward a phylogenetic classification of Primates based on DNA evidence complemented by fossil evidence. *Mol Phylogenet Evol* 9 (3):585-598.
- Gordon, Raymond G., Jr. (ed.), 2005. *Ethnologue: Languages of the World*, Fifteenth edition. Dallas, Tex.: SIL International. March 2007. [<http://www.ethnologue.com/>].
- Graur D, Li W-H (2000) *Fundamentals of Molecular Evolution*. 2nd ed. Sinauer Associates, Inc.: Massachusetts.
- Graven L *et al* (1995) Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol Biol Evol* 12 (2):334-345.
- Graves JA (1996) Breaking laws and obeying rules. *Nat Genet* 12 (2):121-122.
- Graves JA (2004) The degenerate Y chromosome - can conversion save it? *Reprod Fertil Dev* 16 (5):527-534.
- Graves JA and Schmidt MM (1992) Mammalian sex chromosomes: design or accident? *Curr Opin Genet Dev* 2 (6):890-901.
- Gray MW, Burger G, and Lang BF (1999) Mitochondrial evolution. *Science* 283 (5407):1476-1481.
- Greenberg JH (1963) Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Language*. MIT Press: Cambridge, pp 73-113.
- Greenberg JH (1974) Bantu and its closest relatives. *Studies in African Linguistics* Suppl. 5:115-124.
- Grine FE *et al* (2007) Late Pleistocene human skull from Hofmeyr, South Africa, and modern human origins. *Science* 315 (5809):226-229.
- Grun R, Beaumont PB, and Stringer CB (1990) ESR dating evidence for early modern humans at Border Cave in South Africa. *Nature* 344 (6266):537-539.
- Gusmao L, Alves C, and Amorim A (2001) Molecular characteristics of four human Y-specific microsatellites (DYS434, DYD437, DYS438, DYS439) for population and forensic studies. *Ann Hum Genet* 65 (Pt 3):285-291.
- Gusmao L *et al* (2002) Forensic evaluation and population data on the new Y-STRs DYS434, DYS437, DYS438, DYS439 and GATA A10. *Int J Legal Med* 116 (3):139-147.
- Haak W *et al* (2005) Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science* 310 (5750):1016-1018.
- Hair PE (1967) Ethnolinguistic continuity on the Guinea Coast. *J Afr Hist* VIII: 247-268.
- Hamblin MT and Di Rienzo A (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66 (5):1669-1679.
- Hamblin MT, Thompson EE, and Di Rienzo A (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* 70 (2):369-383.

- Hamilton AC (1988) Guenon evolution and forest history. A primate radiation: evolutionary biology of the African guenon, Gautier-Hion A. Cambridge University Press: Cambridge, UK, pp 13-34.
- Hamilton G, Stoneking M, and Excoffier L (2005) Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *Proc Natl Acad Sci U S A* 102 (21):7476-7480.
- Hammer MF (1995) A recent common ancestry for human Y chromosomes. *Nature* 378 (6555):376-378.
- Hammer MF and Horai S (1995) Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet* 56 (4):951-962.
- Hammer MF and Zegura SL (2002) The Human Y Chromosome Haplogroup Tree: Nomenclature and Phylogeography of its Major Divisions. *Annu Rev Anthropol* 31:303-321.
- Hammer MF *et al* (1997) The geographic distribution of human Y chromosome variation. *Genetics* 145 (3):787-805.
- Hammer MF *et al* (1998) Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol Biol Evol* 15 (4):427-441.
- Hammer MF *et al* (2000) Jewish and Middle Eastern non-Jewish populations share a common pool of Y-chromosome biallelic haplotypes. *Proc Natl Acad Sci U S A* 97 (12):6769-6774.
- Hammer MF *et al* (2001) Hierarchical patterns of global human Y-chromosome diversity. *Mol Biol Evol* 18 (7):1189-1203.
- Hammer MF *et al* (2006) Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *J Hum Genet* 51 (1):47-58.
- Hammond HA *et al* (1994) Evaluation of 13 short tandem repeat loci for use in personal identification applications. *Am J Hum Genet* 55:175-189.
- Harpending H (1994) Gene frequencies, DNA sequences, and human origins. *Perspect Biol Med* 37 (3):384-394.
- Harpending H *et al* (1993) The genetic structure of ancient human populations. *Current Anthropology* 34 (4):483-496.
- Harpending H *et al* (1998) Genetic traces of ancient demography. *Proc Natl Acad Sci U S A* 95 (4):1961-1967.
- Hasegawa M *et al* (1993) Toward a more accurate time scale for the human mitochondrial DNA tree. *J Mol Evol* 37 (4):347-354.
- Hasegawa M, Cao Y, and Yang Z (1998) Preponderance of slightly deleterious polymorphism in mitochondrial DNA: nonsynonymous/synonymous rate ratio is much higher within species than between species. *Mol Biol Evol* 15 (11):1499-1505.
- Hassan FA (1978) Archaeological Explorations of the Siwa Oasis Region, Egypt. *Curr Anthropol* 19:146-148.
- Hazkani-Covo E and Graur D (2007) A comparative analysis of *numt* evolution in human and chimpanzee. *Mol Biol Evol* 24 (1):13-18.
- Helgason A *et al* (2000) mtDNA and the origins of the Icelanders: deciphering signals of recent population history. *Am J Hum Genet* 66:999-1016.
- Henshilwood CS *et al* (2002) Emergence of modern human behavior: Middle Stone Age engravings from South Africa. *Science* 295 (5558):1278-1280.
- Herrnstadt C *et al* (2002) Reduced-median-network analysis of complete mitochondrial DNA coding-region sequences for the major African, Asian, and European haplogroups. *Am J Hum Genet* 70 (5):1152-1171.
- Heyer E *et al* (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. *Hum Mol Genet* 6 (5):799-803.
- Heyer E *et al* (2001) Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. *Am J Hum Genet* 69 (5):1113-1126.
- Hill C *et al* (2006) Phylogeography and ethnogenesis of aboriginal Southeast Asians. *Mol Biol Evol* 23 (12):2480-2491.
- Hill C *et al* (2007) A mitochondrial stratigraphy for island southeast Asia. *Am J Hum Genet* 80 (1):29-43.
- Hirata S *et al* (2002) Spermatozoon and mitochondrial DNA. *Reprod Med Biol* 1 (1):41-47.
- Ho SY *et al* (2005) Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. *Mol Biol Evol* 22 (7):1561-1568.
- Holland HD (1994) Early Proterozoic atmospheric change. In: Bengtson S (ed) Early life on Earth. Columbia University Press: New York, pp 237-244.
- Holt CL *et al* (2000) Practical applications of genotypic surveys for forensic STR testing. *Forensic Sci Int* 112 (2-3):91-109.
- Holtkemper U *et al* (2001) Mutation rates at two human Y-chromosomal microsatellite loci using small pool PCR techniques. *Hum Mol Genet* 10 (6):629-633.

- Hooghiemstra H *et al* (1992) Vegetational and climatic changes at the northern fringe of the Sahara 250,000-5,000 years BP. *Rev Palaeobot & Palynol* 74:1-53.
- Hopkin K (1999) Death to sperm mitochondria. *Sci Am* 280 (3):21.
- Horai S (1995) Evolution and the origins of man: clues from complete sequences of hominoid mitochondrial DNA. *Southeast Asian J Trop Med Public Health* 26 (Suppl 1):146-154.
- Horai S *et al* (1995) Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. *Proc Natl Acad Sci U S A* 92 (2):532-536.
- Howell N and Smejkal CB (2000) Persistent heteroplasmy of a mutation in the human mtDNA control region: hypermutation as an apparent consequence of simple-repeat expansion/contraction. *Am J Hum Genet* 66 (5):1589-1598.
- Howell N, Kubacka I, and Mackey DA (1996) How rapidly does the human mitochondrial genome evolve? *Am J Hum Genet* 59 (3):501-509.
- Howell N *et al* (2003a) The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates. *Am J Hum Genet* 72 (3):659-670.
- Howell N *et al* (2003b) Low penetrance of the 14484 LHON mutation when it arises in a non-haplogroup J mtDNA background. *Am J Med Genet A* 119 (2):147-151.
- Howell N *et al* (2004) African Haplogroup L mtDNA Sequences Show Violations of Clock-Like Evolution. *Mol Biol Evol* 21(10):1843-1854
- Huang Q, Xu F-H, and Shen H (2002) Mutation pattern at dinucleotide microsatellite loci in humans. *Am J Hum Genet* 70:625-634.
- Hudjashov G *et al* (2007) Revealing the prehistoric settlement of Australia by Y chromosome and mtDNA analysis. *Proc Natl Acad Sci U S A* 104 (21):8726-8730.
- Huffman TN (1982) Archaeology and the ethnohistory of the African Iron Age. *Annu Rev Anthropol* 11:133-150.
- Hughes JF *et al* (2005) Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* 437 (7055):100-103.
- Hurles ME *et al* (2002) Y chromosomal evidence for the origins of oceanic-speaking peoples. *Genetics* 160 (1):289-303.
- Huxley TH (1870) On the geographical distribution of the chief modifications of Mankind. *J Ethnol Soc London* 2:404-412.
- Huysecom E (2002) Palaeoenvironment and human population in West Africa: an international research project in Mali. *Antiquity* 76:335-336.
- Ingman M and Gyllensten U (2001) Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. *J Hered* 92 (6):454-461.
- Ingman M and Gyllensten U (2003) Mitochondrial genome variation and evolutionary history of Australian and New Guinean aborigines. *Genome Research* 13:1600-1606.
- Ingman M *et al* (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408:708-713.
- International Human Genome Sequencing Consortium (2001) A physical map of the human genome. *Nature* 409 (6822):934-941.
- Jackson BA *et al* (2005) Mitochondrial DNA genetic diversity among four ethnic groups in Sierra Leone. *Am J Phys Anthropol* 128 (1):156-163.
- Jacobs PA and Strong JA (1959) A case of human intersexuality having a possible XXY sex-determining mechanism. *Nature* 183 (4657):302-303.
- Jahns, S (1995) Pollen analysis of a deep-sea core from the Congo Basin vegetational and climatic change in West Africa during the Upper Pleistocene. 14th INQUA Congress, Berlin,
- Jazin E *et al* (1998) Mitochondrial mutation rate revisited: hot spots and polymorphism. *Nat Genet* 18 (2):109-110.
- Jobling MA and Tyler-Smith C (2000) New uses for new haplotypes: the human Y chromosome, disease and selection. *Trends Genet* 16 (8):356-362.
- Jobling MA and Tyler-Smith C (2003) The human Y chromosome: an evolutionary marker comes of age. *Nat Rev Genet* 4 (8):598-612.
- Jobling MA, Pandya A, and Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med* 110 (3):118-124.
- Jobling MA *et al* (1998) A selective difference between human Y-chromosomal DNA haplotypes. *Curr Biol* 31 (8):1391-1394.

- Jobling MA, Hurles ME, Tyler-Smith C (2004) *Human Evolutionary Genetics - Origins, Peoples & Disease*. Garland Publishing: New York.
- Johnson MJ *et al* (1983) Radiation of human mitochondria DNA types analyzed by restriction endonuclease cleavage patterns. *J Mol Evol* 19 (3-4):255-271.
- Johnston HH (1919) *A Comparative Study of the Bantu and Semi-Bantu Languages*. Oxford University Press: Oxford.
- Jones AG and Ardren WR (2003) Methods of parentage analysis in natural populations. *Mol Ecol* 12 (10):2511-2523.
- Jorde LB and Bamshad M (2000) Questioning evidence for recombination in human mitochondrial DNA. *Science* 288 (5473):1931.
- Jorde LB *et al* (1995) Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *Am J Hum Genet* 57 (3):523-538.
- Jorde LB *et al* (2000) The distribution of human genetic diversity: A comparison of mitochondrial, autosomal, and Y-chromosomal data. *Am J Hum Genet* 66.
- Karafet TM *et al* (1999) Ancestral Asian source(s) of new world Y-chromosome founder haplotypes. *Am J Hum Genet* 64 (3):817-831.
- Karafet TM *et al* (2001) Paternal population history of East Asia: sources, patterns, and microevolutionary processes. *Am J Hum Genet* 69 (3):615-628.
- Karafet TM *et al* (2002) High levels of Y-chromosome differentiation among native Siberian populations and the genetic signature of a boreal hunter-gatherer way of life. *Hum Biol* 74 (6):761-789.
- Karafet TM *et al* (2005) Balinese Y-chromosome perspective on the peopling of Indonesia: genetic contributions from pre-Neolithic hunter-gatherers, Austronesian farmers, and Indian traders. *Hum Biol* 77 (1):93-114.
- Karlberg O *et al* (2000) The dual origin of the yeast mitochondrial proteome. *Yeast* 17 (3):170-187.
- Kayser M and Sajantila A (2001) Mutations at Y-STR loci: implications for paternity testing and forensic analysis. *Forensic Sci Int* 118 (2-3):116-121.
- Kayser M *et al* (1997) Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med* 110 (3):125-133.
- Kayser M *et al* (2000a) Melanesian origin of Polynesian Y chromosomes. *Curr Biol* 10 (20):1237-1246.
- Kayser M *et al* (2000b) Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. *Am J Hum Genet* 66 (5):1580-1588.
- Kayser M *et al* (2001) Independent histories of human Y chromosomes from Melanesia and Australia. *Am J Hum Genet* 68 (1):173-190.
- Kayser M *et al* (2003) Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. *Am J Hum Genet* 72 (2):281-302.
- Kayser M *et al* (2004) A comprehensive survey of human Y-chromosomal microsatellites. *Am J Hum Genet* 74 (6):1183-1197.
- Ke Y *et al* (2001) African origin of modern humans in East Asia: a tale of 12,000 Y chromosomes. *Science* 292 (5519):1151-1153.
- Kimpton CP *et al* (1993) Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods Appl* 3 (1):13-22.
- Kimura M (1968) Evolutionary rate at the molecular level. *Nature* 217 (129):624-626.
- Kimura M (1983) *The Neutral Theory of Molecular Evolution*. Cambridge University Press: Cambridge.
- Kivisild T and Villems R (2000) Questioning evidence for recombination in human mitochondrial DNA. *Science* 288 (5473):1931.
- Kivisild T *et al* (1999a) Deep common ancestry of Indian and western-Eurasian mitochondrial DNA lineages. *Curr Biol* 9 (22):1331-1334.
- Kivisild T *et al* (1999b) The place of the Indian mitochondrial DNA variants in the global network of maternal lineages and the peopling of the Old World. In: Papiha SS, Deka R, Chakraborty R (eds) *Genomic diversity*. Kluwer Academic/Plenum Publishers: pp 135-152.
- Kivisild T *et al* (2002) The emerging limbs and twigs of the East Asian mtDNA tree. *Mol Biol Evol* 19 (10):1737-1751.
- Kivisild T *et al* (2003) The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *Am J Hum Genet* 72:313-332.
- Kivisild T *et al* (2004) Ethiopian mitochondrial DNA heritage: tracking gene flow across and around the gate of tears. *Am J Hum Genet* 75 (5):752-770.
- Kivisild T *et al* (2006a) The World mtDNA phylogeny. In: Bandelt H-J, Macaulay V, Richards M (eds) *Human mitochondrial DNA and the evolution of Homo sapiens*. Springer-Verlag: Berlin-Heidelberg.

- Kivisild T *et al* (2006b) The role of selection in the evolution of human mitochondrial genomes. *Genetics* 172 (1):373-387.
- Knight A *et al* (2003) African Y chromosome and mtDNA divergence provides insight into the history of click languages. *Curr Biol* 13 (6.):464-473.
- Kong QP *et al* (2003) Phylogeny of East Asian mitochondrial DNA lineages inferred from complete sequences. *Am J Hum Genet* 73 (3):671-676.
- Kong QP *et al* (2006) Updating the East Asian mtDNA phylogeny: a prerequisite for the identification of pathogenic mutations. *Hum Mol Genet* 15 (13):2076-2086.
- Kornberg A *et al* (1964) Enzymatic synthesis of deoxyribonucleic acid, XVI. Oligonucleotides as templates and the mechanism of their replication. *Proc Natl Acad Sci U S A* 51:315-23.
- Kovac WL, Kovach Computing Services. MVSP – A multi-variate statistical package for Windows ver. 3.13m (2004). January 2007. [www.kovcomp.co.uk/mvsp/index.html].
- Kraytsberg Y *et al* (2004) Recombination of human mitochondrial DNA. *Science* 304 (5673):981.
- Krings M *et al* (1997) *Neanderthal* DNA sequences and the origin of modern humans. *Cell* 90 (1):19-30.
- Krings M *et al* (1999) mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration? *Am J Hum Genet* 64 (4):1166-1176.
- Kruglyak S *et al* (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc Natl Acad Sci U S A* 95 (18):10774-10778.
- Kumar S and Hedges SB (1998) A molecular timescale for vertebrate evolution. *Nature* 392:917-920.
- Kumar S *et al* (2000) Questioning evidence for recombination in human mitochondrial DNA. *Science* 288 (5473):1931.
- Kuroda-Kawaguchi T *et al* (2001) The AZFc region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nat Genet* 29 (3):279-286.
- Kuroki Y *et al* (1999) Spermatogenic ability is different among males in different Y chromosome lineage. *J Hum Genet* 44:289-292.
- Kuroki Y *et al* (2006) Comparative analysis of chimpanzee and human Y chromosomes unveils complex evolutionary pathway. *Nat Genet* 38 (2):158-167.
- Kwiatkowski DP (2005) How malaria has affected the human genome and what human genetics can teach us about malaria. *Am J Hum Genet* 77 (2):171-192.
- Lahn BT and Page DC (1997) Functional coherence of the human Y chromosome. *Science* 278 (5338):675-680.
- Lahn BT and Page DC (1999) Four evolutionary strata on the human X chromosome. *Science* 286 (5441):964-967.
- Lahn BT, Pearson NM, and Jegalian K (2001) The human Y chromosome, in the light of evolution. *Nat Rev Genet* 2 (3):207-216.
- Lahr MM and Foley RA (1994) Multiple dispersals and modern human origins. *Evol Anthropol* 3:48-60.
- Lahr MM and Foley RA (1998) Towards a theory of modern human origins: geography, demography, and diversity in recent human evolution. *Am J Phys Anthropol* Suppl (27):137-176.
- Lahr MM and Foley RA (2004) Human evolution writ small. *Nature* 431:1043-1044.
- Lai Y and Sun F (2003) The relationship between microsatellite slippage mutation rate and the number of repeat units. *Mol Biol Evol* 20 (12):2123-2131.
- Lareu MV *et al* (1994) Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats: UK and Galician Caucasian population surveys and usefulness in paternity investigations. *Forensic Sci Int* 66 (1):41-52.
- Leat N, Benjeddou M, and Davison S (2004) Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa. *Forensic Sci Int* 144 (1):73-75.
- Leavesley M and Chappell J (2004) Buang Merabak: additional early radiocarbon evidence of the colonization of the Bismarck Archipelago, Papua New Guinea. *Antiquity* 78:301.
- Legros F *et al* (2004) Organization and dynamics of human mitochondrial DNA. *J Cell Sci* 117 (Pt 13):2653-2662.
- Lell JT *et al* (2002) The dual origin and Siberian affinities of Native American Y chromosomes. *Am J Hum Genet* 70 (1):192-206.
- Levinson G and Gutman GA (1987) Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol* 4 (3):203-221.
- Levy ER and Burgoyne PS (1986) The fate of XO germ cells in the testes of XO/XY and XO/XY/XYY mouse mosaics: evidence for a spermatogenesis gene on the mouse Y chromosome. *Cytogenet Cell Genet* 42 (4):208-213.
- Lewin PK (1987) A unique ancient Egyptian mummified head, demonstrating removal of the brain from the foramen magnum. *Paleopathol News* (57):12-13.

- Li L and Hamer DH (1995) Recombination and allelic association in the Xq/Yq homology region. *Hum Mol Genet* 4 (11):2013-2016.
- Li W (1997) Molecular Evolution. Sinauer Associates, Inc.: Sunderland.
- Lien S *et al* (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am J Hum Genet* 66 (2):557-566.
- Lightowlers RN *et al* (1997) Mammalian mitochondrial genetics: heredity, heteroplasmy and disease. *Trends Genet* 13 (11):450-455.
- Lin YW *et al* (2005) Polymorphisms associated with the DAZ genes on the human Y chromosome. *Genomics* 86 (4):431-438.
- Litt M and Luty JA (1989) A hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. *Am J Hum Genet* 44 (3):397-401.
- Loeb LA and Preston BD (1986) Mutagenesis by apurinic/apyrimidinic sites. *Annu Rev Genet* 20:201-30.
- Long JC (1986) The allelic correlation structure of Gainj- and Kalam-speaking people. I. The estimation and interpretation of Wright's F-statistics. *Genetics* 112 (3):629-647.
- Loogväli EL *et al* (2004) Disuniting Uniformity: A Pied Cladistic Canvas of mtDNA Haplogroup H in Eurasia. *Mol Biol Evol* 21(11):2012-2021.
- Lopes C (1999) Kaabunké, espaço, território e poder na Guiné-Bissau, Gâmbia e Casamance pré-coloniais. Comissão Nacional para as Comemorações dos Descobrimentos Portugueses: Lisboa.
- Lopez JV *et al* (1994) *Numt*, a recent transfer and tandem amplification of mitochondrial DNA to the nuclear genome of the domestic cat. *J Mol Evol* 39:174-190.
- Lubell D (1975) The Prehistoric cultural ecology of Capsian escargotières. *Libya* 23 (43):121.
- Luis JR and Caeiro B (1995) Application of two STRs (VWA and TPO) to human population profiling: survey in Galicia. *Hum Biol* 67 (5):789-795.
- Luis JR *et al* (2004) The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations. *Am J Hum Genet* 74:532-544.
- Lynch M (1996) Mutation accumulation in transfer RNAs: molecular evidence for Muller's ratchet in mitochondrial genomes. *Mol Biol Evol* 13 (1):209-220.
- Ma K *et al* (1993) A Y chromosome gene family with RNA-binding protein homology: candidates for the azoospermia factor AZF controlling human spermatogenesis. *Cell* 75 (7):1287-1295.
- Maca-Meyer N *et al* (2001) Major genomic mitochondrial lineages delineate early human expansions. *BMC Genet* 2 (1):13.
- Macaulay VA *et al* (1997) mtDNA mutation rates--no need to panic. *Am J Hum Genet* 61 (4):983-990.
- Macaulay VA, Richards M, and Sykes B (1999a) Mitochondrial DNA recombination - no need to panic. *Proc R Soc Lond B Biol Sci* 266:2037-2039.
- Macaulay VA *et al* (1999b) The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* 64 (1):232-249.
- Macaulay VA *et al* (2005) Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science* 308 (5724):1034-1036.
- Magnavita S (2003) The beads of Kissi, Burkina Faso. *J Afr Archaeol* 1:127-146.
- Malyarchuk B *et al* (2004) Differentiation of mitochondrial DNA and Y chromosomes in Russian populations. *Hum Biol* 76 (6):877-900.
- Margulis L (1970) Recombination of non-chromosomal genes in *Chlamydomonas*: assortment of mitochondria and chloroplasts? *J Theor Biol* 26 (2):337-342.
- Margulis L (1975) Symbiotic theory of the origin of eukaryotic organelles; criteria for proof. *Symp Soc Exp Biol* 29:21-38.
- Marjoram P and Donnelly P (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. *Genetics* 136 (2):673-683.
- Marushiakova E, Popov V (1997) Gypsies (Roma) in Bulgaria. Studium zur Tsiganologie und Folkloristik (band 18). Peter Lang: Frankfurt.
- Mateu E *et al* (1997) A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and Sao Tome, Gulf of Guinea. *Ann Hum Genet* 61 (Pt 6):507-518.
- Mathias N, Bayes M, and Tyler-Smith C (1994) Highly informative compound haplotypes for the human Y chromosome. *Hum Mol Genet* 3 (1):115-123.
- McDougall I, Brown FH, and Fleagle JG (2005) Stratigraphic placement and age of modern humans from Kibish, Ethiopia. *Nature* 433 (7027):733-736.



- McElreavey K and Quintana-Murci L (2003) Male reproductive function and the human Y chromosome: is selection acting on the Y? *Reprod Biomed Online* 7 (1):17-23.
- McMurray CT (1995) Mechanisms of DNA expansion. *Chromosoma* 104 (1):2-13.
- Mellars P (2002) Archaeology and the Origins of Modern Humans: European and African Perspectives. In: Crow TJ (ed) *The speciation of modern Homo sapiens*. British Academy: London, U.K., pp 31-47.
- Mellars P (2004) Neanderthals and the modern human colonization of Europe. *Nature* 432 (7016):461-465.
- Mellars P (2006) Why did modern human populations disperse from Africa ca. 60,000 years ago? A new model. *Proc Natl Acad Sci U S A* 103 (25):9381-9386.
- Mercader J, Marti R (2003) The Middle Stone Age occupation of Atlantic central Africa: new evidence from Equatorial Guinea and Cameroon. In: Mercader J (ed) *Under the Canopy*. Rutgers University Press: New Brunswick, pp 93-118.
- Merriwether DA *et al* (1991) The structure of human mitochondrial DNA variation. *J Mol Evol* 33 (6):543-555.
- Merriwether DA *et al* (2005) Ancient mitochondrial M haplogroups identified in the Southwest Pacific. *Proc Natl Acad Sci U S A* 102 (37):13034-13039.
- Metspalu M *et al* (2004) Most of the extant mtDNA boundaries in South and Southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans. *BMC Genet* 5 (1):26.
- Metspalu M, Kivisild T, Bandelt H-J, Richards M, Villems R (2006) The pioneer settlement of modern humans in Asia. In: Bandelt H-J, Macaulay V, Richards M (eds) *Human Mitochondrial DNA and the Evolution of Homo sapiens*. Springer-Verlag: Berlin Heidelberg.
- Meyer S, Weiss G, and von Haeseler A (1999) Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. *Genetics* 152 (3):1103-1110.
- Michaels GS, Hauswirth WW, and Laipis PJ (1982) Mitochondrial DNA copy number in bovine oocytes and somatic cells. *Dev Biol* 94 (1):246-251.
- Miller LH *et al* (1976) The resistance factor to *Plasmodium vivax* in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med* 295 (6):302-304.
- Mishmar D *et al* (2003) Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* 100 (1):171-176.
- Mishmar D *et al* (2004) Mitochondrial DNA-like sequences in the nucleus (NUMTs): insights into our African origins and the mechanism of foreign DNA integration. *Hum Mutat* 23 (2):125-133.
- MITOMAP: A Human Mitochondrial Genome Database. February 2007. [<http://www.mitomap.org>].
- Modiano D *et al* (1996) Different response to *Plasmodium falciparum* malaria in west African sympatric ethnic groups. *Proc Natl Acad Sci U S A* 93 (23):13206-13211.
- Modiano D *et al* (1998) Baseline immunity of the population and impact of insecticide-treated curtains on malaria infection. *Am J Trop Med Hyg* 59 (2):336-340.
- Modiano D *et al* (1999) Interethnic differences in the humoral response to non-repetitive regions of the *Plasmodium falciparum* circumsporozoite protein. *Am J Trop Med Hyg* 61 (4):663-667.
- Modiano D *et al* (2001) The lower susceptibility to *Plasmodium falciparum* malaria of Fulani of Burkina Faso (west Africa) is associated with low frequencies of classic malaria-resistance genes. *Trans R Soc Trop Med Hyg* 95 (2):149-152.
- Monson *et al* (2003) The mtDNA population database: an integrated software and database resource for forensic comparison. Forensic Science Communications Online.
- Moreira JM (1964) Os Fulas da Guiné Portuguesa na panorâmica geral do mundo Fula. *Bol Cult Guiné Port* XIX:417-432.
- Mori F (1974) The earliest Saharan rock engravings. *Antiquity* 48:87-92.
- Mountain JL *et al* (2002) SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res* 12 (11):1766-1772.
- Muller HJ (1964) The relation of recombination to mutational advance. *Mutat Res* 1 (1):2-9.
- Mullis K *et al* (1992) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. 1986. *Biotechnology* 24:17-27.
- Mumm S *et al* (1997) Evolutionary features of the 4-Mb Xq21.3 XY homology region revealed by a map at 60-kb resolution. *Genome Res* 7 (4):307-314.
- Murdock GP (1967) *Ethnographic Atlas*. University of Pittsburgh Press: Pittsburgh.
- Muzzolini A (1993) The emergence of a food-producing economy in the Sahara. In: Shaw T *et al.* (ed) *The archaeology of Africa: food, metals and towns*. Routledge: London, pp 227-239.
- Nachman MW (1998) Deleterious mutations in animal mitochondrial DNA. *Genetica* 103 (1-6):61-69.

- Nachman MW and Crowell SL (2000) Estimate of the mutation rate per nucleotide in humans. *Genetics* 156 (1):297-304.
- Nebel A *et al* (2001) The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. *Am J Hum Genet* 69 (5):1095-1112.
- Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press: New York.
- NetViz LCC Corporation. NetViz 6.5 (2002). June 2006 [www.netviz.com].
- Newman JL (1995) *The peopling of Africa*. Yale University Press: New Haven, CT.
- Nielsen R and Weinreich DM (1999) The age of nonsynonymous and synonymous mutations in animal mtDNA and implications for the mildly deleterious theory. *Genetics* 153 (1):497-506.
- Nikitina TV and Nazarenko SA (2004) Human microsatellites: mutation and evolution. *Russian Journal of Genetics* 40:1065-1079.
- Nurse GT, Weiner JS, Jenkins T (1985) The San yesterday and today. In: Harrison GA (ed) *The peoples of southern Africa and their affinities*. Clarendon Press: Oxford.
- Ohno S (1967) *Sex chromosomes and sex-linked genes*. Springer: Berlin.
- Olivieri A *et al* (2006) The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa. *Science* 314 (5806):1767-1770.
- Olivo PD *et al* (1983) Nucleotide sequence evidence for rapid genotypic shifts in the bovine mitochondrial DNA D-loop. *Nature* 306 (5941):400-402.
- Olson LE and Yoder AD (2002) Using secondary structure to identify ribosomal *numts*: cautionary examples from the human genome. *Mol Biol Evol* 19 (1):93-100.
- Ono T *et al* (2001) Human cells are protected from mitochondrial dysfunction by complementation of DNA products in fused mitochondria. *Nat Genet* 28 (3):272-275.
- Oota H *et al* (2001) Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat Genet* 29 (1):20-21.
- Ovchinnikov IV *et al* (2000) Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature* 404 (6777):490-493.
- Pääbo S (1996) Mutational hot spots in the mitochondrial microcosm. *Am J Hum Genet* 59 (3):493-496.
- Page DC *et al* (1984) Occurrence of a transposition from the X-chromosome long arm to the Y-chromosome short arm during human evolution. *Nature* 311 (5982):119-123.
- Page RD, Holmes EC (1998) *Molecular Evolution. A Phylogenetic Approach*. Blackwell Science Ltd: London.
- Pakendorf B and Stoneking M (2005) Mitochondrial DNA and human evolution. *Annu Rev Genomics Hum Genet*. 6:165-83.
- Palanichamy MG *et al* (2004) Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia. *Am J Hum Genet* 75 (6):966-978.
- Paracchini S *et al* (2000) Y-chromosomal DNA haplotypes in infertile European males carrying Y-microdeletions. *J Endocrinol Invest* 23 (10):671-676.
- Parsons TJ and Irwin JA (2000) Questioning evidence for recombination in human mitochondrial DNA. *Science* 288 (5473):1931.
- Parsons TJ *et al* (1997) A high observed substitution rate in the human mitochondrial DNA control region. *Nat Genet* 15 (4):363-368.
- Passarino G *et al* (1998) Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am J Hum Genet* 62 (2):420-434.
- Passarino G *et al* (2001) Y chromosome binary markers to study the high prevalence of males in Sardinian centenarians and the genetic structure of the Sardinian population. *Hum Hered* 52 (3):136-139.
- Pearson PL, Bobrow M, and Vosa CG (1970) Technique for identifying Y chromosomes in human interphase nuclei. *Nature*. 226 (5240):78-80.
- Pereira L *et al* (2001a) Y-chromosome mismatch distributions in Europe. *Mol Biol Evol* 18 (7):1259-1271.
- Pereira L *et al* (2001b) Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* 65 (Pt 5):439-458.
- Pereira L *et al* (2002) Bantu and European Y-lineages in sub-Saharan Africa. *Ann Hum Genet* 66 (Pt 5-6):369-378.
- Pereira L *et al* (2005) High-resolution mtDNA evidence for the late-glacial resettlement of Europe from an Iberian refugium. *Genome Res* 15 (1):19-24.

- Pereira L *et al* (2006) Evaluating the forensic informativeness of mtDNA haplogroup H sub-typing on a Eurasian scale. *Forensic Sci Int* 159 (1):43-50.
- Perez-Lezaun A *et al* (1999) Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y-chromosome short tandem repeats and mtDNA. *Am J Hum Genet* 65 (1):208-219.
- Pesole G *et al* (1999) Nucleotide substitution rate of mammalian mitochondrial genomes. *J Mol Evol* 48 (4):427-434.
- Phillipson DW (1993) African Archaeology. 2nd ed. Cambridge University Press: Cambridge.
- Pierson MJ *et al* (2006) Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. *Mol Biol Evol* 23 (10):1966-1975.
- Piganeau G and Eyre-Walker A (2004) A reanalysis of the indirect evidence for recombination in human mitochondrial DNA. *Heredity* 92 (4):282-288.
- Piko L and Matsumoto L (1976) Number of mitochondria and some properties of mitochondrial DNA in the mouse egg. *Dev Biol* 49 (1):1-10.
- Piko L and Taylor KD (1987) Amounts of mitochondrial DNA and abundance of some mitochondrial gene transcripts in early mouse embryos. *Dev Biol* 123 (2):364-374.
- Plaza S *et al* (2003) Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann Hum Genet* 67 (4):312-328.
- Plaza S *et al* (2004) Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. *Hum Genet* 115 (5):439-447.
- Poloni ES *et al* (1997) Human genetic affinities for Y-chromosome P49a,f/TaqI haplotypes show strong correspondence with linguistics. *Am J Hum Genet* 61 (5):1015-1035.
- Poulton J, Macaulay V, and Marchington DR (1998) Mitochondrial genetics '98 is the bottleneck cracked? *Am J Hum Genet* 62 (4):752-757.
- Previdere C *et al* (1999) Y-chromosomal DNA haplotype differences in control and infertile Italian subpopulations. *Eur J Hum Genet* 7:733-736.
- Pritchard JK *et al* (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol Biol Evol* 16 (12):1791-1798.
- Quintana-Murci L *et al* (1999) Genetic evidence of an early exit of *Homo sapiens sapiens* from Africa through eastern Africa. *Nat Genet* 23 (4):437-441.
- Quintana-Murci L, Krausz C, and McElreavey K (2001) The human Y chromosome: function, evolution and disease. *Forensic Sci Int* 118 (2-3):169-181.
- Quintana-Murci L *et al* (2004) Where West meets East: The complex mtDNA landscape of the Southwest and Central Asian corridor. *Am J Hum Genet* 74:827-845.
- Quintino F (1964) Sobrevivências da Cultura etiópica no ocidente africano. *Bol Cult Guiné Port* XIX:5-35.
- Quintino F (1967) Os povos da Guiné. *Bol Cult Guiné Port* XXII:5-40.
- Quintino F (1969) Os povos da Guiné. *Bol Cult Guiné Port* XXIV:861-915.
- Rando JC *et al* (1998) Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations. *Ann Hum Genet* 62 (Pt 6):531-550.
- Rando JC *et al* (1999) Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann Hum Genet* 63 (Pt 5):413-428.
- Rasse M *et al* (2004) La séquence pléistocène supérieur d'Ounjougou (Pays dogon, Afrique de l'Ouest): évolution géomorphologique, enregistrements sédimentaires et changements culturels. *Quaternaire* 15/4:329-341.
- Raymond C and Rousset F (1995) An exact test for population differentiation. *Evolution* 49:1280-1283.
- Redd AJ *et al* (2002) Gene flow from the Indian subcontinent to Australia. Evidence from the Y chromosome. *Curr Biol* 12 (8):673-7.
- Renfrew C (1987) Archaeology and Language: the Puzzle of Indo-European Origins. Jonathan Cape: London.
- Repping S *et al* (2003) Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nat Genet* 35 (3):247-251.
- Repping S *et al* (2004) A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1.8-Mb deletion in the azoospermia factor c region. *Genomics* 83 (6):1046-1052.
- Repping S *et al* (2006) High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat Genet* 38 (4):463-467.
- Reyes A *et al* (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol Biol Evol* 15 (8):957-966.

- Reynier P *et al* (2001) Mitochondrial DNA content affects the fertilizability of human oocytes. *Mol Hum Reprod* 7 (5):425-429.
- Reynolds J, Weir BS, and Cockerham CC (1983) Estimation of the coancestry coefficient, Basis for a short term genetic distance. *Genetics* 105:767-779.
- Ricchetti M, Tekai F, and Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS Biol* 2 (9):1313-1324.
- Rice WR (1987) Genetic hitch-hiking and the evolution of reduced genetic activity of the Y sex chromosome. *Genetics* 116:161-167.
- Richards M and Macaulay V (2001) The mitochondrial gene tree comes of age. *Am J Hum Genet* 68 (6):1315-1320.
- Richards M and Sykes B (1998) Reply to Barbujani *et al*. *Am J Hum Genet* 62:491.
- Richards M *et al* (1996) Paleolithic and Neolithic lineages in the European mitochondrial gene pool. *Am J Hum Genet* 59 (1):185-203.
- Richards M *et al* (1997) Reply to Cavalli-Sforza and Minch. *Am J Hum Genet* 61:251-254.
- Richards M *et al* (1998a) Phylogeography of mitochondrial DNA in western Europe. *Ann Hum Genet* 62 (Pt 3):241-260.
- Richards M, Oppenheimer S, and Sykes B (1998b) mtDNA suggests Polynesian origins in Eastern Indonesia. *Am J Hum Genet* 63 (4):1234-1236.
- Richards M *et al* (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* 67 (5):1251-1276.
- Richards M *et al* (2002) In search of geographical patterns in European mitochondrial DNA. *Am J Hum Genet* 71 (5):1168-1174.
- Richards M *et al* (2003) Extensive female-mediated gene flow from sub-Saharan Africa into near eastern Arab populations. *Am J Hum Genet* 72 (4):1058-1064.
- Rieder MJ *et al* (1998) Automating the identification of DNA variations using quality-based fluorescence re-sequencing: analysis of the human mitochondrial genome. *Nucleic Acids Res* 26 (4):967-973.
- Rightmire GP (1989) Middle Stone Age humans from eastern and southern Africa. The Human Revolution. Princeton University Press: Princeton.
- Rightmire GP and Deacon HJ (2001) New human teeth from Middle Stone Age deposits at Klasies River, South Africa. *J Hum Evol* 41 (6):535-544.
- Rightmire GP *et al* (2006) Human foot bones from Klasies River main site, South Africa. *J Hum Evol* 50 (1):96-103.
- Robion-Brunner C *et al* (2006) A thousand years of iron metallurgy on the Dogon plateau (Mali). [<http://cohesion.rice.edu/CentersAndInst/SAFA/emplibrary/Robionetal,C.SAfA2006.pdf>].
- Roewer L *et al* (2000) A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males. *Forensic Sci Int* 114 (1):31-43.
- Roewer L *et al* (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Sci Int* 118 (2-3):106-113.
- Rogers AR and Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol Biol Evol* 9 (3):552-569.
- Roostalu U *et al* (2007) Origin and expansion of haplogroup H, the dominant human mitochondrial DNA lineage in West Eurasia: the Near Eastern and Caucasian perspective. *Mol Biol Evol* 24 (2):436-448.
- Rootsi S (2004) Human Y-chromosomal variation in European populations.
- Rootsi S *et al* (2004) Phylogeography of Y-chromosome haplogroup I reveals distinct domains of prehistoric gene flow in Europe. *Am J Hum Genet* 75 (1):128-137.
- Rootsi S *et al* (2007) A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. *Eur J Hum Genet* 15 (2):204-211.
- Rosa A *et al* (2004) MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region. *Ann Hum Genet* 68 (4):340-352.
- Rosa A *et al* (2006) Population data on 11 Y-chromosome STRs from Guiné-Bissau. *Forensic Sci Int* 157 (2-3):210-217.
- Rosa A *et al* (2007) Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective. *BMC Evol Biol* 7:124.
- Rosser ZH *et al* (2000) Y-chromosomal diversity in Europe is clinal and influenced primarily by geography, rather than by language. *Am J Hum Genet* 67 (6):1526-1543.
- Rozen S *et al* (2003) Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423 (6942):873-876.

- Rubicz R *et al* (2003) Mitochondrial DNA variation and the origins of the Aleuts. *Hum Biol* 75 (6):809-835.
- Ruiz-Pesini E *et al* (2004) Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* 303 (5655):223-226.
- Sabeti PC *et al* (2002) Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419 (6909):832-837.
- Saccone C *et al* (1999) Evolutionary genomics in Metazoa: the mitochondrial DNA as a model system. *Gene* 238 (1):195-209.
- Saiki RK *et al* (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science* 239 (4839):487-491.
- Saillard J *et al* (2000) mtDNA variation among Greenland Eskimos: the edge of the Beringian expansion. *Am J Hum Genet* 67 (3):718-726.
- Saitou N and Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* 4 (4):406-425.
- Salas A *et al* (2002) The making of the African mtDNA landscape. *Am J Hum Genet* 71 (5):1082-1111.
- Salas A *et al* (2004) The African diaspora: mitochondrial DNA and the Atlantic slave trade. *Am J Hum Genet* 74 (3):454-465.
- Salemi M, Vandamme A-M (2003) *The Phylogenetic Handbook - A Practical Approach to DNA and Protein Phylogeny*. Cambridge University Press: Cambridge.
- Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular cloning: a laboratory manual*. Second ed. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.
- Santos TA, El SS, and St John JC (2006) Mitochondrial content reflects oocyte variability and fertilization outcome. *Fertil Steril* 85 (3):584-591.
- Saxena R *et al* (1996) The DAZ gene cluster on the human Y chromosome arose from an autosomal gene that was transposed, repeatedly amplified and pruned. *Nat Genet* 14 (3):292-299.
- Scheinfeldt L *et al* (2006) Unexpected NRY chromosome variation in Northern Island Melanesia. *Mol Biol Evol* 23 (8):1628-1641.
- Schneider PM *et al* (1998) Tandem repeat structure of the duplicated Y chromosomal STR locus DYS385 and frequency studies in the German and three Asian populations. *Foren Sci Int* :61-70.
- Schneider S and Excoffier L (1999) Estimation of past demographic parameters from the distribution of pairwise differences when the mutation rates vary among sites: application to human mitochondrial DNA. *Genetics* 152 (3):1079-1089.
- Schneider S, Roessli D, Excoffier L (2000) Arlequin version 2.000: a software for population genetics data analysis.
- Schueler MG *et al* (2001) Genomic and genetic definition of a functional human centromere. *Science* 294 (5540):109-115.
- Schwartz A *et al* (1998) Reconstructing hominid Y evolution: X-homologous block, created by X-Y transposition, was disrupted by Yp inversion through LINE-LINE recombination. *Hum Mol Genet* 7 (1):1-11.
- Schwartz M and Vissing J (2002) Paternal inheritance of mitochondrial DNA. *N Engl J Med* 347 (8):576-580.
- Schwartz M and Vissing J (2004) No evidence for paternal inheritance of mtDNA in patients with sporadic mtDNA mutations. *J Neurol Sci* 218 (1-2):99-101.
- Scozzari R *et al* (1997) mtDNA and Y chromosome-specific polymorphisms in modern Ojibwa: implications about the origin of their gene pool. *Am J Hum Genet* 60 (1):241-244.
- Scozzari R *et al* (1988) Genetic studies on the Senegal population. I. Mitochondrial DNA polymorphisms. *Am J Hum Genet* 43 (4):534-544.
- Scozzari R *et al* (1999) Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. *Am J Hum Genet* 65 (3):829-846.
- Scozzari R *et al* (2001) Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region. *Hum Immunol* 62 (9):871-884.
- Seielstad MT, Minch E, and Cavalli-Sforza LL (1998) Genetic evidence for a higher female migration rate in humans. *Nat Genet* 20 (3):278-280.
- Semino O *et al* (1996) A view of the Neolithic demic diffusion in Europe through two Y chromosome-specific markers. *Am J Hum Genet* 59 (4):964-968.
- Semino O *et al* (2000) The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective. *Science* 290 (5494):1155-1159.

- Semino O *et al* (2002) Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *Am J Hum Genet* 70 (1):265-268.
- Semino O *et al* (2004) Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the Neolithization of Europe and later migratory events in the Mediterranean area. *Am J Hum Genet* 74 (5):1023-1034.
- Serneels V (2005) An ongoing research project on iron production and use in the Dogon Country, Mali. *Historical Metallurgy Soc News* 60:1-3.
- Shaw TC (1980) Agricultural origins in Africa. In: Sherratt A (ed) *The Cambridge Encyclopedia of Archaeology*: Cambridge, pp 179-184.
- Shen P *et al* (2000) Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci U S A* 97 (13):7354-7359.
- Shen P *et al* (2004) Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. *Hum Mut* 24(3):248-260.
- Shoubridge EA (2001) Nuclear genetic defects of oxidative phosphorylation. *Hum Mol Genet* 10:2277-2284.
- Sigurdottir S *et al* (2000) The mutation rate in the human mtDNA control region. *Am J Hum Genet* 66 (5):1599-1609.
- Silva WA, Jr. *et al* (2002) Mitochondrial genome diversity of Native Americans supports a single early entry of founder populations into America. *Am J Hum Genet* 71 (1):187-192.
- Simmler MC *et al* (1985) Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes. *Nature* 317 (6039):692-697.
- Sinclair PJJ, Shaw T, Andah B (1993) Introduction. In: Shaw T *et al* (eds) *The Archaeology of Africa: Food, Metals and Towns*. Routledge: London, pp 1-31.
- Skaletsky H *et al* (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423 (6942):825-837.
- Slatkin M (1995) A measure of population subdivision based on microsatellite allele frequencies. *Genetics* 139:457-462.
- Slatkin M and Hudson RR (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129 (2):555-562.
- Smith DG *et al* (1999) Distribution of mtDNA haplogroup X among Native North Americans. *Am J Phys Anthropol* 110 (3):271-284.
- Smith PE (1982) The Late Palaeolithic and Epi-Palaeolithic of Northern Africa. In: Clark J.D. (ed) Vol 1. *From the Earliest Times to c. 500 B.C.* Cambridge University Press: Cambridge, pp 342-409.
- Smith RN (1995) Accurate size comparison of short tandem repeat alleles amplified by PCR. *Biotechniques* 18 (1):122-128.
- Soodyall H and Jenkins T (1992) Mitochondrial DNA polymorphisms in Khoisan populations from southern Africa. *Ann Hum Genet* 56 (Pt 4):315-324.
- Soodyall H and Jenkins T (1993) Mitochondrial DNA polymorphisms in Negroid populations from Namibia: new light on the origins of the Dama, Herero and Ambo. *Ann Hum Biol* 20 (5):477-485.
- Soodyall H *et al* (1996) mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans. *Am J Hum Genet* 58 (3):595-608.
- Soodyall H *et al* (1997) The founding mitochondrial DNA lineages of Tristan da Cunha Islanders. *Am J Phys Anthropol* 104 (2):157-166.
- St John JC, Lloyd R, and El Shourbagy S (2004) The potential risks of abnormal transmission of mtDNA through assisted reproductive technologies. *Reprod Biomed Online* 8 (1):34-44.
- Stahl AB (1985) Reinvestigation of Kintampo 6 Rock Shelter, Ghana: Implications for the Nature of Culture Change. *Afr Archaeol Rev* 3:117-150.
- Stefansson H *et al* (2005) A common inversion under selection in Europeans. *Nat Genet* 37 (2):129-137.
- Steuerwald N *et al* (2000) Quantification of mtDNA in single oocytes, polar bodies and subcellular components by real-time rapid cycle fluorescence monitored PCR. *Zygote* 8 (3):209-215.
- Stevanovic M *et al* (1993) SOX3 is an X-linked gene related to SRY. *Hum Mol Genet* 2 (12):2013-2018.
- Stevanovitch A *et al* (2004) Mitochondrial DNA Sequence Diversity in a Sedentary Population from Egypt. *Ann Hum Genet* 68 (1):23-39.
- Stoneking M (1994) Mitochondrial DNA and human evolution. *J Bioenerg Biomembr* 26 (3):251-259.
- Stoneking M (2000) Hypervariable sites in the mtDNA control region are mutational hotspots. *Am J Hum Genet* 67 (4):1029-1032.
- Stoneking M and Soodyall H (1996) Human evolution and the mitochondrial genome. *Curr Opin Genet Dev* 6 (6):731-736.

- Stoneking M *et al* (1991) Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am J Hum Genet* 48 (2):370-382.
- Stringer C (2000) Coasting out of Africa. *Nature* 405 (6782):24-25.
- Stringer C (2003) Human evolution: Out of Ethiopia. *Nature* 423 (6941):692-693.
- Stringer C and Andrews P (1988) Genetic and fossil evidence for the origin of modern humans. *Science* 239 (4845):1263-1268.
- Studier JA and Keppler KJ (1988) A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol* 5:729-731.
- Stuhlmann F (1910) Handwerk und Industrie in Ostafrika, kulturgeschichtliche Betrachtungen. Friedrichsen: Hamburg.
- Su B *et al* (1999) Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last ice age. *Am J Hum Genet* 65 (6):1718-1724.
- Su B *et al* (2000) Polynesian origins: insights from the Y chromosome. *Proc Natl Acad Sci U S A* 97 (15):8225-8228.
- Subramanian S, Mishra RK, and Singh L (2003) Genome-wide analysis of microsatellite repeats in humans: their abundance and density in specific genomic regions. *Genome Biol* 4 (2):R13.
- Sullivan TD (2004) A preliminary report of existing information on the Manding languages of West Africa. *SIL Electronic Survey Reports* 2004-2005.
- Sun C *et al* (1999) An azoospermic man with a de novo point mutation in the Y-chromosomal gene USP9Y. *Nat Genet* 23 (4):429-432.
- Sun C *et al* (2000) Deletion of azoospermia factor a (AZFa) region of human Y chromosome caused by recombination between HERV15 proviruses. *Hum Mol Genet* 9 (15):2291-2296.
- Sun C *et al* (2006) The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes. *Mol Biol Evol* 23 (3):683-690.
- Sutovsky P *et al* (2004) Degradation of paternal mitochondria after fertilization: implications for heteroplasmy, assisted reproductive technologies and mtDNA inheritance. *Reprod Biomed Online* 8 (1):24-33.
- Sutton JE (1982) Archaeology in West Africa: A Review of the Recent Work and a Further List of Radiocarbon Dates. *J Afr Hist* 23:291-314.
- Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet* 37:197-219.
- Swofford DL (1993) PAUP: Phylogenetic Analysis Using Parsimony. Illinois Natural History Survey: Champaign.
- Tajima A *et al* (2004) Genetic origins of the Ainu inferred from combined DNA analyses of maternal and paternal lineages. *J Hum Genet* 49 (4):187-193.
- Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105 (2):437-460.
- Tajima F (1989) DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* 123 (1):229-240.
- Tajima F (1993) Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* 135 (2):599-607.
- Tajima F (1996) The amount of DNA polymorphism maintained in a finite population when the neutral mutation rate varies among sites. *Genetics* 143 (3):1457-1465.
- Tambets K *et al* (2004) The Western and Eastern Roots of the Saami—the Story of Genetic "Outliers" Told by Mitochondrial DNA and Y Chromosomes. *Am J Hum Genet* 74 (4):661-682.
- Tamura K (2000) On the estimation of the rate of nucleotide substitution for the control region of human mitochondrial DNA. *Gene* 259 (1-2):189-197.
- Tamura K and Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol* 10:512-526.
- Tanaka M *et al* (2004) Mitochondrial genome variation in eastern Asia and the peopling of Japan. *Genome Res* 14 (10a):1832-1850.
- Tattersall I, Delson E, Van Couvering J, Brooks AS (2000) Encyclopedia of Human Evolution and Prehistory. 2nd ed. Garland Publishing: New York.
- Taylor RW *et al* (2003) Genotypes from patients indicate no paternal mitochondrial DNA contribution. *Ann Neurol* 54 (4):521-524.
- Teixeira da Mota A (1954) Guiné Portuguesa. Agência Geral do Ultramar: Lisbon.
- Templeton AR (1992) Human origins and analysis of mitochondrial DNA sequences. *Science* 255 (5045):737.
- Templeton AR (1997) Out of Africa? What do genes tell us? *Curr Opin Genet Dev* 7 (6):841-847.
- Thangaraj K *et al* (2005) Reconstructing the origin of Andaman Islanders. *Science* 308 (5724):996.

- Thomas MG *et al* (1998) Molecular instability in the COII-tRNA(Lys) intergenic region of the human mitochondrial genome: multiple origins of the 9-bp deletion and heteroplasmy for expanded repeats. *Philos Trans R Soc Lond B Biol Sci* 353 (1371):955-965.
- Thomas MG *et al* (2000) Y chromosomes traveling south: the Cohen modal haplotype and the origins of the Lemba - the "Black Jews of Southern Africa. *Am J Hum Genet* 66 (2):674-686.
- Thomson R *et al* (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc Natl Acad Sci U S A* 97 (13):7360-7365.
- Thorburn DR and Dahl HH (2001) Mitochondrial disorders: genetics, counseling, prenatal diagnosis and reproductive options. *Am J Med Genet* 106 (1):102-114.
- Thyagarajan B, Padua RA, and Campbell C (1996) Mammalian mitochondria possess homologous DNA recombination activity. *J Biol Chem* 271 (44):27536-27543.
- Tiepolo L and Zuffardi O (1976) Localization of factors controlling spermatogenesis in the nonfluorescent portion of the human Y chromosome long arm. *Hum Genet* 34 (2):119-124.
- Tilford CA *et al* (2001) A physical map of the human Y chromosome. *Nature* 409 (6822):943-945.
- Tishkoff SA *et al* (2001) Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. *Science* 293 (5529):455-462.
- Tishkoff SA *et al* (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39 (1):31-40.
- Torroni A *et al* (1992) Native American mitochondrial DNA analysis indicates that the Amerind and the Nadene populations were founded by two independent migrations. *Genetics* 130 (1):153-162.
- Torroni A *et al* (1993) Asian affinities and continental radiation of the four founding Native American mtDNAs. *Am J Hum Genet* 53 (3):563-590.
- Torroni A *et al* (1994a) Mitochondrial DNA analysis in Tibet: implications for the origin of the Tibetan population and its adaptation to high altitude. *Am J Phys Anthropol* 93 (2):189-199.
- Torroni A *et al* (1994b) Mitochondrial DNA "clock" for the Amerinds and its implications for timing their entry into North America. *Proc Natl Acad Sci U S A* 91 (3):1158-1162.
- Torroni A *et al* (1994c) mtDNA and Y-chromosome polymorphisms in four Native American populations from southern Mexico. *Am J Hum Genet* 54 (2):303-318.
- Torroni A *et al* (1996) Classification of European mtDNAs from an analysis of three European populations. *Genetics* 144 (4):1835-1850.
- Torroni A *et al* (1997) Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. *Am J Hum Genet* 60 (5):1107-1121.
- Torroni A *et al* (1998) mtDNA analysis reveals a major late Paleolithic population expansion from southwestern to northeastern Europe. *Am J Hum Genet* 62 (5):1137-1152.
- Torroni A *et al* (2001a) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* 69 (6):1348-1356.
- Torroni A *et al* (2001b) A signal, from human mtDNA, of postglacial recolonization in Europe. *Am J Hum Genet* 69 (4):844-852.
- Torroni A *et al* (2006) Harvesting the fruit of the human mtDNA tree. *Trends Genet* 22 (6):339-345.
- Tourmen Y *et al* (2002) Structure and chromosomal distribution of human mitochondrial pseudogenes. *Genomics* 80:71-77.
- Trovada MJ *et al* (2001) Evidence for population sub-structuring in Sao Tome e Principe as inferred from Y-chromosome STR analysis. *Ann Hum Genet* 65 (Pt 3):271-283.
- Turner C *et al* (2003) Human genetic disease caused by de novo mitochondrial-nuclear DNA transfer. *Hum Genet* 112:303-309.
- Tyler-Smith C *et al* (1993) Localization of DNA sequences required for human centromere function through an analysis of rearranged Y chromosomes. *Nat Genet* 5 (4):368-375.
- Uchihi R *et al* (2003) Haplotype analysis with 14 Y-STR loci using 2 multiplex amplification and typing systems in 2 regional populations in Japan. *Int J Legal Med* 117 (1):34-38.
- Underhill PA (2003) Inferring Human History: Clues from Y-Chromosome Haplotypes. Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory Press: pp 487-493.
- Underhill PA and Kivisild T (2007) Use of Y chromosome and mitochondrial DNA population structure in tracing human migrations. *Ann Rev Genetics* 41: *in press*.



- Underhill PA *et al* (1997) Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. *Genome Res* 7 (10):996-1005.
- Underhill PA *et al* (2000) Y chromosome sequence variation and the history of human populations. *Nat Genet* 26 (3):358-361.
- Underhill PA *et al* (2001a) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65 (1):43-62.
- Underhill PA *et al* (2001b) Maori origins, Y-chromosome haplotypes and implications for human history in the Pacific. *Hum Mutat* 17 (4):271-280.
- Van Holst Pellekaan SM *et al* (2006) Mitochondrial genomics identifies major haplogroups in Aboriginal Australians. *Am J Phys Anthropol* 131 (2):282-294.
- Vanhaereny M *et al* (2006) Middle Paleolithic shell beads in Israel and Algeria. *Science* 312 (5781):1785-1788.
- Vansina J (1994) A Slow Revolution: Farming in Subequatorial Africa. *Azania* 29-30.
- Vellai T, Takacs K, and Vida G (1998) A new aspect to the origin and evolution of eukaryotes. *J Mol Evol* 46 (5):499-507.
- Velosa RG, Fernandes AT, and Brehm A (2002) Genetic profile of the Azores Archipelago population using the new PowerPlex 16 system kit. *Forensic Sci Int* 129 (1):68-71.
- Vigilant L *et al* (1991) African populations and the evolution of human mitochondrial DNA. *Science* 253 (5027):1503-1507.
- Vogt PH (1997) Human Y chromosome deletions in Yq11 and male fertility. *Adv Exp Med Biol* 424:17-30.
- Vogt PH (1998) Human chromosome deletions in Yq11, AZF candidate genes and male infertility: history and update. *Mol Hum Reprod* 4 (8):739-744.
- Vogt PH (2004) Molecular genetics of human male infertility: from genes to new therapeutic perspectives. *Curr Pharm Des* 10 (5):471-500.
- Vogt PH (2005) Azoospermia factor (AZF) in Yq11: towards a molecular understanding of its function for human male fertility and spermatogenesis. *Reprod.Biomed.Online*10 (1):81-93.
- Wakeley J (1994) Substitution-rate variation among sites and the estimation of transition bias. *Mol Biol Evol* 11 (3):436-442.
- Wallace DC, Brown MD, and Lott MT (1999) Mitochondrial DNA variation in human evolution and disease. *Gene* 238 (1):211-230.
- Walsh PS, Metzger DA, and Higuchi R (1991) Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *Biotechniques* 10 (4):506-513.
- Walzer S and Gerald PS (1975) Social class and frequency of XYY and XXY. *Science* 190 (4220):1228-1229.
- Watson E *et al* (1996) mtDNA sequence diversity in Africa. *Am J Hum Genet* 59 (2):437-444.
- Watson E *et al* (1997) Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* 61 (3):691-704.
- Watson JM *et al* (1991) Sex chromosome evolution: platypus gene mapping suggests that part of the human X chromosome was originally autosomal. *Proc Natl Acad Sci U S A* 88 (24):11256-11260.
- Weale ME *et al* (2003) Rare deep-rooting Y chromosome lineages in humans: lessons for phylogeography. *Genetics* 165 (1):229-234.
- Weidenreich F (1943) The "Neanderthal man" and the ancestors of "Homo sapiens". *Am Anthropologist* 45:39-48.
- Weir BS and Cockerham CC (1984) Estimating F-statistics for the analysis of population structure. *Evolution* 38:1358-1370.
- Weir BS (1996) Genetic Data Analysis II: Methods for Discrete Population Genetic Data. Sunderland, MA, USA.
- Wells RS *et al* (2001) The Eurasian heartland: a continental perspective on Y-chromosome diversity. *Proc Natl Acad Sci USA* 98 (18):10244-10249.
- Wen B *et al* (2004) Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. *Am J Hum Genet* 74 (5):856-865.
- Werle E *et al* (1994) Convenient single-step, one tube purification of PCR products for direct sequencing. *Nucleic Acids Res* 22 (20):4354-4355.
- White PS *et al* (1999) New, male-specific microsatellite markers from the human Y chromosome. *Genomics* 57:433-437.
- White TD *et al* (2003) Pleistocene *Homo sapiens* from Middle Awash, Ethiopia. *Nature* 423 (6941):742-747.
- Whitfield LS, Sulston JE, and Goodfellow PN (1995) Sequence variation of the human Y chromosome. *Nature* 378 (6555):379-380.

- Willuweit S and Roewer L, on behalf of the International Forensic Y Chromosome User Group (2007) Y chromosome haplotype reference database (YHRD): Update, Forensic Science International. *Genetics* 1(2):83-87. March 2007 [www.yhrd.org].
- Wilson A *et al* (1985) Mitochondrial DNA and two perspectives on evolutionary genetics. *Biol J Linnean Soc* 26:375-400.
- Wolpoff M (1989) Multiregional evolution: the fossil alternative to Eden. In: Mellars P, Stringer C (eds) *The human revolution: Behavioral and biological perspectives on the origins of modern humans*. Edinburgh University Press: Edinburgh.
- Wolpoff M, Caspari R (1997) *Race and Human Evolution*. Simon & Schuster: New York.
- Wood ET *et al* (2005) Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* 13 (7):867-876.
- Wyckoff GJ, Wang W, and Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. *Nature* 403 (6767):304-309.
- Xu X, Peng M, and Fang Z (2000) The direction of microsatellite mutations is dependent upon allele length. *Nat Genet* 24 (4):396-399.
- YCC – Y Chromosome Consortium (2002) A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res* 12 (2):339-348.
- Yu M *et al* (2002) A new haplogroup pattern displayed in Fujian Han in China. *J Hum Genet* 47:95-98.
- Zarrabeitia MT *et al* (2003) Spanish population data and forensic usefulness of a novel Y-STR set (DYS437, DYS438, DYS439, DYS460, DYS461, GATA A10, GATA C4, GATA H4). *Int J Legal Med* 117 (5):306-311.
- Zegura SL *et al* (2004) High-resolution SNPs and microsatellite haplotypes point to a single, recent entry of Native American Y chromosomes into the Americas. *Mol Biol Evol* 21 (1):164-175.
- Zhivotovsky LA and Feldman MW (1995) Microsatellite variability and genetic distances. *Proc Natl Acad Sci U S A* 92 (25):11549-11552.
- Zhivotovsky LA and Underhill PA (2005) On the evolutionary mutation rate at Y-chromosome STRs: comments on paper by Di Giacomo *et al.* (2004). *Hum Genet* 116 (6):529-532.
- Zhivotovsky LA, Feldman MW, and Grishechkin SA (1997) Biased mutations and microsatellite variation. *Mol Biol Evol* 14 (9):926-933.
- Zhivotovsky LA *et al* (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *Am J Hum Genet* 74 (1):50-61.
- Zhivotovsky LA, Underhill PA, and Feldman MW (2006) Difference between evolutionarily effective and germ line mutation rate due to stochastically varying haplogroup size. *Mol Biol Evol* 23 (12):2268-2270.
- Zischler H *et al* (1995) A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. *Nature* 378 (6556):489-492.

## Acknowledgments

This study was carried out in the Human Genetics Laboratory, University of Madeira and the Department of Evolutionary Biology at the Institute of Molecular and Cell Biology - University of Tartu.

My deepest gratitude goes to my scientific supervisors Professor António Brehm and Professor Richard Villems, and to Professor Toomas Kivisild for their invaluable guidance and critical reviewing, essential for the success of the project and completion of the manuscript.

I am most grateful to all my co-workers for the fascinatingly hard and friendly way I was accepted in the working teams, and for the inspiring atmosphere of discussion and healthy criticism on scientific matters. I feel indebted to Kristiina Tambets, Siiri Rootsi, Jüri Parik and Erwan Pennarun: you all contributed to my scientific and personal maturation and, more importantly, were my family for three years. I would like to express my sincere gratitude to Ana Teresa Fernandes and Rita Gonçalves from the HGL for their kind help and advices. I would like to thank Ille Hilpus and Jaan Lind for the technical assistance.

It is to outstand the unconditional support and endless understanding of my parents and family, throughout the last years' "rollercoaster".

Thank you my long-term friends, for making justice to Dumas' "one for all, and all for one".

To you Nuno, I hope to be able to reward you for your persistence and endless patience, especially during this last stressful stage. It is for me a privilege and definitely my main impellor, that you have chosen me to share the most positive and joyful temperament on Earth!

I am thankful to blood donors in Guinea-Bissau. The collection of samples has been possible thanks to permissions from Chairman of the Joint Chiefs of Staff and the Ministry of Health of the Republic of Guinea-Bissau. Professor António Brehm received a grant from the Regional Government of Madeira (Portugal) and conducted the blood collection in the field, where AMI – Assistência Médica Internacional gave local support. The Fundação para a Ciência e Tecnologia has granted me with the PhD grant SFRH/BD/12173/2003.



Supplementary material















Table S2 – African datasets used for comparison of mtDNA genetic pool

Geographic region/Ethnic group	Abbreviation	N	Linguistic			Reference	
			Family	Sublevel			
<i>Northwest Africa</i>							
Mauritania	Mauritanians	Mau	94	Mixed	Mixed	Rando <i>et al.</i> 1998, González <i>et al.</i> 2006 <sup>a</sup>	
West Sahara	Saharawis	Sah	25	Afro-Asiatic	Semitic	Rando <i>et al.</i> 1998	
Morocco	Arabs	MAr	350	Afro-Asiatic	Semitic	Rando <i>et al.</i> 1998, Pennarun <i>et al.</i> (unpublished)	
	Berbers	MBb	268	Afro-Asiatic	Berber	Rando <i>et al.</i> 1998, Pennarun <i>et al.</i> (unpublished)	
Algeria	Arabs	AAr	55	Afro-Asiatic	Semitic	Pennarun <i>et al.</i> (unpublished)	
	Berbers	ABb	64	Afro-Asiatic	Berber	Pennarun <i>et al.</i> (unpublished)	
	Mozabites	Mzt	85	Afro-Asiatic	Berber	Côrte-Real <i>et al.</i> 1996	
<i>Northeast Africa</i>							
Egypt	Egyptians	Egy1	192	Afro-Asiatic	Mixed	Metspalu <i>et al.</i> (unpublished)	
	Egyptians	Egy2	107	Afro-Asiatic	Mixed	Krings <i>et al.</i> 1999	
Sudan	Nubians	Nub	148	Nilo-Saharan	Eastern	Krings <i>et al.</i> 1999	
<i>West Africa</i>							
Cape Verde	Cape Verdeans	CV	292	Creole	Portuguese-based	Brehm <i>et al.</i> 2002	
Senegal	Mandenka	Mak	110	Niger-Congo	Manding	Graven <i>et al.</i> 1995	
	Senegalese	Sen	50	Mixed	Mixed	Rando <i>et al.</i> 1998	
Mali	Wolof	Wol	48	Niger-Congo	Atlantic-Wolof	Rando <i>et al.</i> 1998	
	Serer	Ser	23	Niger-Congo	Atlantic-Serer	Rando <i>et al.</i> 1998	
	Tuareg	Tug	23	Afro-Asiatic	Berber	Watson <i>et al.</i> 1996	
	Mixed	Mal	26	Niger-Congo	Mixed	González <i>et al.</i> 2006 <sup>a</sup>	
	Bambara	Bab	71	Niger-Congo	Manding-East	Ely <i>et al.</i> 2006, González <i>et al.</i> 2006 <sup>a</sup>	
	Peul	Pe	15	Niger-Congo	Atlantic-Fulani	González <i>et al.</i> 2006 <sup>a</sup>	
Guinea-Bissau	Malinke	Mwk	92	Niger-Congo	Manding-West	Ely <i>et al.</i> 2006, González <i>et al.</i> 2006 <sup>a</sup>	
	Felupe-Djola	EJA	50	Niger-Congo	Atlantic-Bak	Rosa <i>et al.</i> 2004	
	Bijagós	BJG	22	Niger-Congo	Atlantic-Bijagó	Rosa <i>et al.</i> 2004	
	Balanta	BLE	62	Niger-Congo	Atlantic-Bak	Rosa <i>et al.</i> 2004	
	Papel	PBO	77	Niger-Congo	Atlantic-Bak	Rosa <i>et al.</i> 2004	
	Fulbe	FUL	77	Niger-Congo	Atlantic-Fulani	Rosa <i>et al.</i> 2004	
	Mandenka	MNK	58	Niger-Congo	Manding-West	Rosa <i>et al.</i> 2004	
	Nalú	NAJ	26	Niger-Congo	Atlantic-Nalu	Rosa <i>et al.</i> 2004	
	Sierra Leone	Limba	Lim	67	Niger-Congo	Atlantic-Limba	Jackson <i>et al.</i> 2005
		Loko	Lko	32	Niger-Congo	Manding Mende-Loko	Jackson <i>et al.</i> 2005
Temne		Tmn	121	Niger-Congo	Atlantic-Temne	Jackson <i>et al.</i> 2005	
Burkina-Faso	Mende	Mde	59	Niger-Congo	Manding Mende-Loko	Jackson <i>et al.</i> 2005	
	Fulani Banfora	FBa	50	Niger-Congo	Atlantic-Fulani	Cerny <i>et al.</i> 2006	
	Fulani Tindangou	FTi	47	Niger-Congo	Atlantic-Fulani	Cerny <i>et al.</i> 2006	
<i>Central Africa</i>							
Niger/Nigeria	Yoruba	Yor	33	Niger-Congo	Yoruboid	Vigilant <i>et al.</i> 1990, Watson <i>et al.</i> 1996	
	Fulbe	Fni	60	Niger-Congo	Atlantic-Fulani	Watson <i>et al.</i> 1996	
	Hausa	Hau	20	Afro-Asiatic	Chadic	Watson <i>et al.</i> 1996	
Chad	Mandara	Mad	37	Afro-Asiatic	Chadic	Destro-Bisol <i>et al.</i> 2004*	
	Uldeme	Oul	28	Afro-Asiatic	Chadic	Destro-Bisol <i>et al.</i> 2004*	
	Podokwo	Po	39	Afro-Asiatic	Chadic	Destro-Bisol <i>et al.</i> 2004*	
	Fulani Bongor	FBor	49	Niger-Congo	Atlantic-Fulani	Cerny <i>et al.</i> 2006	
North Cameroon	Kotoko	Kot	18	Afro-Asiatic	Chadic	Cerny <i>et al.</i> 2004	
	Tupuri	Tup	25	Niger-Congo	Adamawa	Destro-Bisol <i>et al.</i> 2004*	
	Daba	Dab	20	Afro-Asiatic	Chadic	Destro-Bisol <i>et al.</i> 2004*	
	Fali	Fal	41	Niger-Congo	Adamawa	Destro-Bisol <i>et al.</i> 2004*	
	Tali	Ta	20	Niger-Congo	Adamawa	Destro-Bisol <i>et al.</i> 2004*	
	Fulbe	Fca	34	Niger-Congo	Atlantic-Fulani	Destro-Bisol <i>et al.</i> 2004*	
	Hide	Hid	23	Afro-Asiatic	Chadic	Cerny <i>et al.</i> 2004	
South Cameroon	Fulani Tcheboua	FTc	40	Niger-Congo	Atlantic-Fulani	Cerny <i>et al.</i> 2006	
	Mafa	Maf	32	Afro-Asiatic	Chadic	Cerny <i>et al.</i> 2004	
	Bakaka	Bak	50	Niger-Congo	Bantu	Destro-Bisol <i>et al.</i> 2004*	
	Bamilike	Bam	48	Niger-Congo	Bantu	Destro-Bisol <i>et al.</i> 2004*	
	Masa	Mas	31	Afro-Asiatic	Chadic	Cerny <i>et al.</i> 2004	
	Bassa	Bis	46	Niger-Congo	Bantu	Destro-Bisol <i>et al.</i> 2004*	
	Ewondo	Ewo	53	Niger-Congo	Bantu	Destro-Bisol <i>et al.</i> 2004*	
Central African Republic	Biaka Pygmies	Bia	17	Niger-Congo	Bantu	Vigilant <i>et al.</i> 1990, Watson <i>et al.</i> 1996	
<i>East Africa</i>							
Ethiopia	Tigras	Tig	53	Afro-Asiatic	Semitic	Kivisild <i>et al.</i> 2004	
	Oromo/Afar	Oro	49	Afro-Asiatic	Cushitic	Kivisild <i>et al.</i> 2004	
	Amhara	Amh	120	Afro-Asiatic	Semitic	Kivisild <i>et al.</i> 2004	
	Gurage	Gur	21	Afro-Asiatic	Semitic	Kivisild <i>et al.</i> 2004	
Kenya	Turkana	Tur	37	Nilo-Saharan	Nilotic	Watson <i>et al.</i> 1996, 1997	
	Kikuyu	Kik	24	Niger-Congo	Bantoid	Watson <i>et al.</i> 1996, 1997	
	Nairobians	Nai	84	Mixed	Mixed	Brandstätter <i>et al.</i> 2004	
Somalia	Somali	Som	27	Afro-Asiatic	Mixed	Watson <i>et al.</i> 1996, 1997	
<i>South Africa</i>							
Mozambique	Mozambicans	Moz	109	Mixed	Mixed	Pereira <i>et al.</i> 2001	
	Bantu	MoB	307	Niger-Congo	Bantu	Salas <i>et al.</i> 2002	
South Africa/ Botswana	IKung	Ku	62	Khoisan	Northern	Vigilant <i>et al.</i> 1990; Watson <i>et al.</i> 1996, 1997; Chen <i>et al.</i> 2000	
	Khwe	Khw	31	Khoisan	Central	Chen <i>et al.</i> 2000	

“\*\*” Data in Coia *et al.* 2005, “a” – not included in the statistical analysis. The linguistic classification follows pertinent information from [www.ethnologue.com](http://www.ethnologue.com).

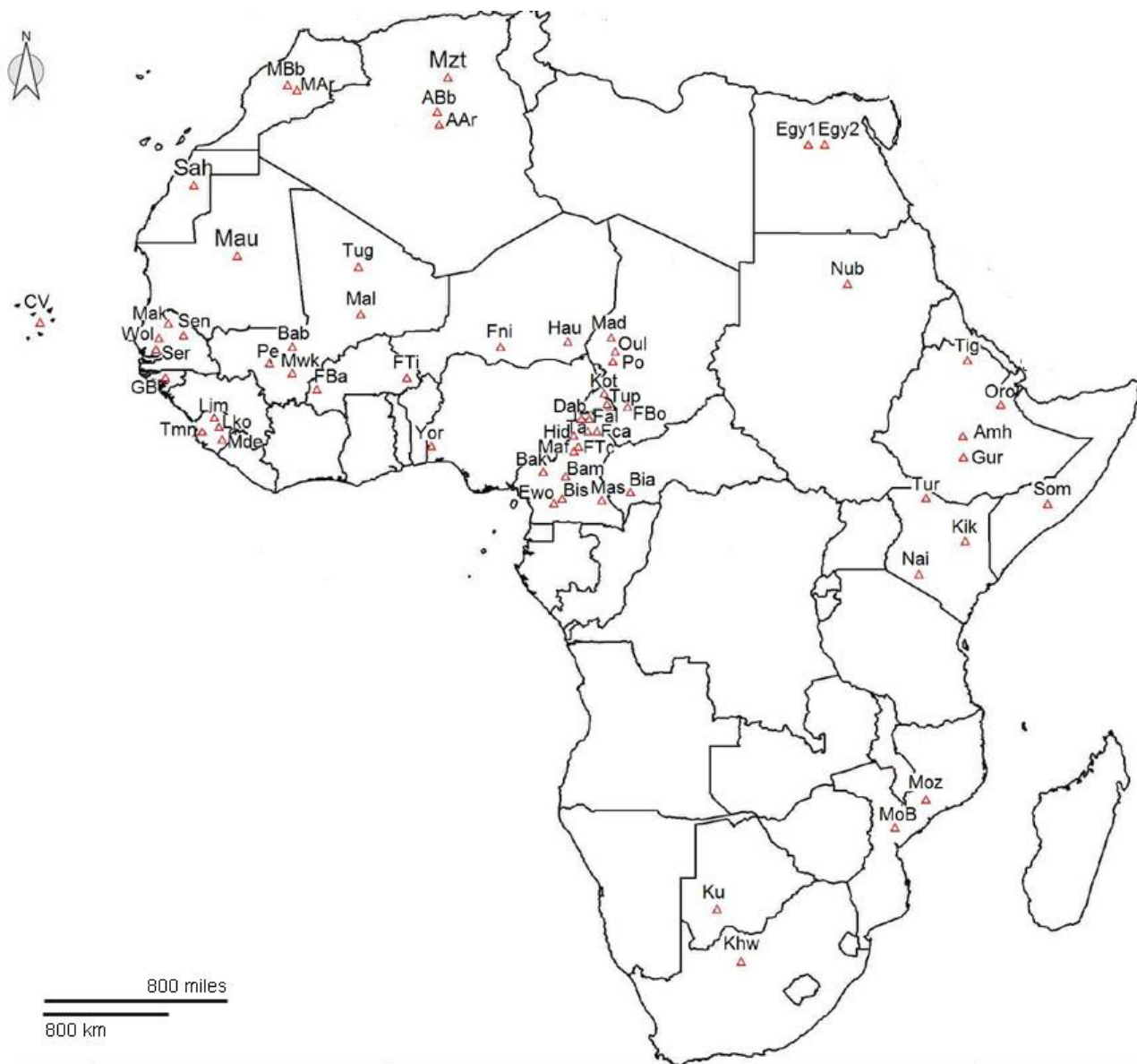


Figure S1 – Geographic distribution of the African datasets used for comparison of mtDNA genetic profiles (details in Table S2).

Table S3 – Absolute frequencies of mtDNA haplogroups in several African datasets, including Guinea-Bissau ethnic groups, and respective diversity indexes (H, sd)

Haplogroup	Northwest Africa								Northeast Africa			West Africa			East Africa										South Africa						
	Mauritanians	Saharawis	Arabe Morocco	Berbers Morocco	Arabe Algeria	Berbers Algeria	Mozabiles	Egyptians	Nile Valley-Egyptians	Nubians	Cape Verdeans	Senegal Mandinka	Senegalese	Wolof	Serer	Tuareg	Bambara	Malinke	Felupe-Djola	Blagos	Balania	Papel	Fulbe	Mandenka	Nalu	Limba	Loko	Temne	Mende	Fulani Bambara	Fulani Tindangou
L0a	0	0	1	0	1	1	0	5	5	15	2	2	0	0	0	1	0	0	2	2	7	3	1	3	1	1	1	6	1	0	0
L1b	5	1	18	10	1	2	0	1	5	4	23	22	4	11	5	2	5	7	2	4	8	5	13	5	1	3	7	14	16	16	12
L1c	2	0	3	2	0	0	0	2	0	1	20	2	2	1	1	0	0	2	2	0	4	3	5	3	2	3	4	8	1	0	0
L1*	0	0	0	0	0	0	0	1	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
L2a	3	2	17	13	8	0	5	9	4	34	60	13	9	10	5	9	3	16	6	3	9	14	17	8	4	15	8	20	11	3	1
L2b	0	0	3	3	1	0	0	4	1	0	12	3	5	7	4	0	2	5	5	1	4	10	0	4	5	3	1	1	1	6	3
L2c	1	5	3	2	2	0	0	0	0	0	47	43	5	3	1	1	3	9	9	3	7	13	12	13	4	12	1	17	4	5	6
L2d	0	2						0	0	1	2	0	0	3	0	0	1	2	0	2	1	2	1	1	1	1	1	7	3	0	3
L2*	0	0	4*	2*	3*	0	0	0	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L3b	1	0	15	5	5	0	3	2	1	2	31	9	10	2	2	1	1	5	7	3	3	6	8	3	1	11	3	12	6	15*	8
L3d	0	0	6	1	3	0	1	2	0	1	21	10	5	2	1	1	0	3	1	2	8	9	4	7	4	4	4	13	4	8	8
L3e	0	0	17	7	8	4	2	3	1	1	43	5	2	4	2	2	2	7	7	1	7	3	2	4	3	8	0	10	3	0	0
L3f	0	0	6	1	0	0	0	5	1	11	2	0	4	0	0	1	2	3	3	1	0	2	1	2	0	3	0	7	3	0	1
L3h	0	0	0	0	0	0	0	1	0	1	2	0	0	0	0	1	0	0	4	1	1	3	1	3	0	0	0	0	0	0	0
L3*	1	1	3	2	2	0	0	3	4	21	8	0	2	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	3	0	0
M1	0	1	15	9	0	2	4	14	12	9	0	0	0	0	0	0	0	2	1	2	0	0	0	0	0	0	0	0	0	0	0
U6	6	2	26	18	1	11	24	5	1	1	9	0	0	2	0	1	0	0	0	0	4	2	2	0	0	2	2	2	2	0	0
Eurasian	11	11	213	193	20	44	46	135	69	35	10	1	2	2	1	3	0	2	0	0	1	9	0	0	3	0	3	1	4	4	5
H	0.8069	0.7733	0.6138	0.4724	0.8222	0.4995	0.6272	0.4973	0.5689	0.8526	0.8811	0.7818	0.8963	0.8599	0.8893	0.8379	0.8889	0.8760	0.9045	0.9264	0.9090	0.8995	0.8759	0.9014	0.9015	0.8756	0.8690	0.9043	0.8743	0.8343	0.8585
sd	0.0481	0.0708	0.0295	0.0374	0.0376	0.0675	0.0422	0.0440	0.0546	0.0138	0.0073	0.0274	0.0186	0.0251	0.0370	0.0687	0.0420	0.0221	0.0160	0.0287	0.0113	0.0140	0.0160	0.0192	0.0270	0.0184	0.0318	0.0084	0.0258	0.0298	0.0229
N	30	25	350	268	55	64	85	192	107	148	292	110	50	48	23	19	61	22	50	22	62	77	77	58	26	67	32	121	59	50	47

Haplogroup	Central Africa															East Africa					South Africa												
	Yoruba	Fulbe	Hausa	Mandara	Udeme	Podokwo	Fulani Bongor	Kotoko	Tupuri	Daba	Fai	Tali	Cameroon Fulbe	Hdi	Fulani Tchibouba	Mafa	Bakaka	Bamillike	Masa	Bassa	Ewondo	Tigris	Oromo/Afar	Amhara	Gurage	Turkana	Kikuyu	Nairobi	Somalia	Mozambicans	Mozambique Bantu	Ikung	Khwe
L0a	1	0	0	0	2	2	0	3	1	1	5	0	0	2	0	5	10	7	3	2	8	3	3	11	0	7	3	15	1	16	88	1	3
L1b	6	11	1	4	3	2	14	1	2	1	0	1	4	0	8	1	3	1	0	3	3	0	3	1	1	0	0	2	0	1	4	0	1
L1c	2	0	1	4	1	0	0	0	0	2	1	1	2	3	2	0	7	3	2	11	11	0	0	0	0	0	2	0	5	17	0	0	
L1*	0	0	1	0	0	0	2	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	1	0	11	3	11	0	10	13	52	12	
L2a	7	9	5	10	6	5	3	2	4	2	14	5	3	4	0	3	5	14	5	7	8	3	7	18	2	4	3	8	7	47	90	0	3
L2b	1	0	0	1	2	0	3	1	1	0	0	0	2	1	3	6	1	2	1	0	0	1	0	4	0	0	0	0	2	4	2	2	2
L2c	0	2	2	1	0	1	2	0	0	0	0	0	0	0	4	0	0	0	0	0	1	0	0	0	0	0	0	0	1	2	0	0	
L2d	0	0	1	2	0	1	0	0	0	0	1	0	0	1	1	0	0	2	2	2	4	0	0	0	0	0	1	0	0	3	0	0	
L2*	0	0	0	0	0	0	0	0	0	0	1	1	5	0	0	1	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	
L3b	1	11	2	4	3	5	17*	0	1	4	5	2	4	1	7	9*	1	2	5	1	3	0	0	0	0	0	7	0	4	8	5	1	
L3d	5	7	2	0	0	2	0	0	1	1	0	0	3	7	0	3	1	0	0	0	0	3	3	0	0	0	0	2	21	0	0	0	
L3e	7	8	4	3	3	10	3	3	6	6	4	4	5	3	5	13	5	5	6	6	0	0	0	0	0	3	10	0	19	45	2	9	
L3f	2	5	1	1	1	2	0	7	3	0	2	3	0	1	2	2	2	5	7	5	3	3	4	6	1	3	2	5	3	2	8	0	0
L3h	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
L3*	0	0	0	7	5	8	1	1	7	3	5	3	6	1	1	7	3	0	7	6	9	9	23	6	10	8	16	8	0	4	0	0	
M1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	2	1	1	0	4	3	0	0	0	0	
U6	0	2	0	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23	12	30	7	0	2	1	0	0	0	0	0	
Eurasian	1	5	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	9	7	19	3	0	0	1	5	0	0	0	0	
H	0.8693	0.8774	0.9000	0.8679	0.9021	0.8691	0.8537	0.8170	0.8467	0.8632	0.8415	0.8789	0.9037	0.9091	0.8910	0.8851	0.8539	0.8706	0.8796	0.8754	0.8868	0.7569	0.8648	0.8469	0.8095	0.8048	0.8732	0.8873	0.7977	0.7560	0.8015	0.2924	0.7656
sd	0.0279	0.0142	0.0428	0.0303	0.0279	0.0293	0.0281	0.0698	0.0403	0.0486	0.0395	0.0398	0.0181	0.0336	0.0226	0.0233	0.0237	0.0298	0.0266	0.0219	0.0170	0.0453	0.0218	0.0140	0.0549	0.0324	0.0406	0.0134	0.0421	0.0316	0.0126	0.0734	0.0509
N	33	60	20	37	28	39	49	18	25	20	41	20	34	23	40	32	50	48	31	46	53	53	49	120	21	37	24	84	27	109	307	62	31

List of population datasets in Table S2 (units with N<20 not included). L1\*, L2\* and L3\* include lineages not ascertained in the mentioned clades. Guinea-Bissau Eurasian cluster refers only to U5b while in other populational units may include other non-L subsets. “\*” - the original publication does not allow distinguishing the clades.

Table S4 – Coalescence time estimates for the Guinea-Bissau mtDNA haplogroup variation (in ky)

Haplogroups	N	TMRCA	sd
L0a1	19	6.4	2.6
L1b	38	43.5	19.3
L1b1	33	23.8	8.8
L1b1 (x np114A)	21	15.4	7.7
L1b1 (x np274)	24	13.5	6.7
L1c	19	108.3	21.9
L2a*	15	36.3	13.0
L2a1	46	36.0	9.6
L2a1-β1	30	30.3	8.0
L2b	29	39.0	20.9
L2b1	21	8.6	4.0
L2c	61	20.8	5.1
L2d	7	121.1	33.1
L3b	32	36.6	12.6
L3b (x L3b1)	18	17.9	9.8
L3b1	14	40.4	16.3
L3d	35	42.7	10.8
L3e2a	10	8.1	4.0
L3e2b	6	10.1	5.8
L3e4	11	11.0	5.2
L3f1	9	49.3	16.2
L3h	13	14.1	8.4

Calculations based on the network in Figures 20 and 22, according to Forster *et al.* <sup>(1996)</sup> and Saillard *et al.* <sup>(2000)</sup>. “x” denotes lineages not considered for the calculations, e.g. L1b1 (x np274) considers all L1b1 mtDNAs except the lineages defined by np16274).

Table S5 - Exact matches of mtDNA haplotypes between Guinea-Bissau and other populations (based on HVS-I nps 16024-16400)

haplogrou	HVS-I motif (minus 16000 nps)	Guinea-Bissau ID <sup>a</sup>	N	Guinea-Bissau										Northwest Africa	Mauritanians	Saharawis	Moroccan Arabs	Morocco Berbers	Mozabites	Northeast Africa	Egyptians - Central & South <sup>1</sup>	Nubians, South Egypt, North Sudan <sup>1</sup>	Nubians, South Egypt, North Sudan <sup>1</sup>	Sudanese, Southern Sudan tribals <sup>1</sup>	West Africa	Cape Verdeans	Senegal Mandenka	Senegalese	Wolof	Serer	Tuareg	Mali (mixed)	Bambara	
				50	22	62	77	58	77	26	94	25	350																					60
L0a1	129 148 168 172 187 188G 189 223 230 311 320	GB4	1	2	3	1	2	2	1												6	2	2			2	2						1	
L1b1	069 093 126 187 189 223 264 270 278 293 311	GB7				1																												
L1b1	093 126 187 189 223 264 270 278 293 311	GB8				5	1																									1		
L1b	111 126 187 189 223 239 270 278 311	GB11			1																											1		
L1b1	114A 126 187 189 223 264 270 274 278 293 311	GB13			1																						3*							
L1b1	114A 126 187 189 223 264 270 278 293 311	GB14			1	1	1																				4							
L1b1	114G 126 187 189 223 264 270 274 278 293 311	GB15			1				1																	1								
L1b1	114G 126 187 189 223 264 270 274 278 311	GB17		2																														
L1b1	126 145 187 189 223 264 270 278 293 311	GB19							1			1														1	3	1		1				
L1b1	126 187 189 213 223 260 264 270 278 293 311	GB20					1																											
L1b1	126 187 189 223 256 264 270 278 293 311	GB21	1		1																						2					1		
L1b1	126 187 189 223 264 270 278 293 311	GB23			1	2	2	2	1			5		12	1						1		1			1	6		2	2	2	1	2	
L1b	126 187 189 223 264 270 278 311	GB24			1		1					1		1															2	1			2	
L1c3	093 129 (183C) 189 215 223 278 294 311 360	GB30	1																														1	
L1c	093 129 187 189 192 223 256 278 311 368	GB32				1																											1	
L1c3	093 129 189 209 215 223 278 289 294 311 360	GB33						1	1																									1
L1c	129 163 187 189 223 278 293 294 311 360	GB36							1																								2	
L2a-α2	(183C) 189 223 278 294 390	GB39					1														2										1			
L2a-α3	189 192 223 278 294 390	GB44				1																									1			
L2a-α2	189 223 278 294 362 390	GB46	1																															
L2a-α1	223 278 294 390	GB50					1							1																			1	
L2a1-β1	092 223 278 294 309 390	GB53						1																										
L2a1-β3	189 192 209 223 278 294 309 390	GB56								1																								1
L2a1-β3	(183C) 189 192 223 278 294 309 390	GB57					1					1	1	3												2		2	1					
L2a1-β3	189 192 223 278 294 309 390	GB58		2	1				1					2																		1	2	
L2a1-β2	189 223 264 278 294 309 390	GB59				1																												
L2a1-β2	189 223 278 294 309 390	GB62	2	1								2		4												2		1	1		1	1		
.2a1d-(β1)	213 223 278 294 309 390	GB68					1																				1			1				
.2a1d-(β1)	223 278 286 294 309 390	GB70					1		1																								2	
L2a1-β1	223 278 290 294 309 390	GB71				1		1		1																1	2							





Table S5 (continued)

haplogrou	HVS-I motif (minus 16000 nps)	<b>Guinea-Bissau</b>		<b>West Africa</b>							<b>Northwest Africa</b>					<b>Northeast Africa</b>					<b>West Africa</b>						
		ID <sup>a</sup>	N	50	22	62	77	58	77	26	94	25	350	60	85	27	68	14	79	292	110	50	48	23	23	26	71
L3b	124 223 278 362	GB137	1		1	1	1		1		2									10		5	1			1	3
L3b	124 278 362	GB138						1	1													1					
L3b1	223 278 362	GB141	2																		1						3
L3d	111 124 223	GB147				3				4										3		1					
L3d3	124 (183C) 189 223 278 304 311	GB148							1																		1
L3d	124 148 223 293 311 362	GB150								2										2							
L3d	124 223	GB152	1					1				1		1						3	2	1	1				
L3d	124 223 288	GB153								1										1							
L3d	124 223 311 399	GB156						1	2																1		
L3d1	124 223 319	GB157						1																			
L3d	124 223 399	GB159			1		1	1																			
L3e2a	223 320	GB162	2	1	1	1	1	1		1										6	1				1	1	2
L3e2b	172 (183C) 189 223 320	GB166							1					2						2			1			3	1
L3e2b	172 189 223 320	GB167	1																								
L3e4	051 223 264	GB170	2		3			1		1								1		17		1					
L3f1	209 223 292 295 311 362	GB177	1																								
L3f1	209 223 292 311	GB178					1													1					1		
L3h	129 223 256A 311 362	GB184	2	1	1	1	1	1												1							
M1	129 185 189 223 249 311	GB186	2		1							1		2													
U5b1b	189 192 270 320	GB188					8		1														1	1			
U6a	111 172 (183C) 189 219 278	GB189							1											1							
U6a	172 (183C) 189 219 278	GB191				1	1	2			3	1	1	10		1				4							
U6a	172 (183C) 189 219 278 390	GB192				1																			2		



















Table S5 (continued)

haplogrou	HVS-I motif (minus 16000)	ID <sup>a</sup>	Guinea-Bissau								Non-Africans	Palestinians <sup>3,4</sup>	Albanians <sup>1d</sup>	Sardinians <sup>3,4</sup>	Sicilians <sup>1d,4</sup>	Sicilians <sup>5</sup>	Siena <sup>1d</sup>	Romans <sup>4</sup>	Tuscanians <sup>6</sup>	Northern Portuguese <sup>7</sup>	Southern Portuguese <sup>7</sup>	Spain Leonese <sup>8</sup>	Spain Andalusians <sup>8</sup>	Spaniards <sup>9</sup>	English <sup>10</sup>	African American <sup>11</sup>			
			N	50	22	62	77	58	77	26																			
L3b	124 223 278 362	GB137	1		1	1	1		1																				3
L3b	124 278 362	GB138						1	1																				
L3b1	223 278 362	GB141	2																										
L3d	111 124 223	GB147				3																							
L3d3	124 (183C) 189 223 278 304 311	GB148							1																				
L3d	124 148 223 293 311 362	GB150																											
L3d	124 223	GB152	1					1																					
L3d	124 223 288	GB153																											
L3d	124 223 311 399	GB156						1	2																				
L3d1	124 223 319	GB157						1																					
L3d	124 223 399	GB159			1		1	1																					
L3e2a	223 320	GB162	2	1	1	1	1		1																				
L3e2b	172 (183C) 189 223 320	GB166							1		1																		
L3e2b	172 189 223 320	GB167	1																										
L3e4	051 223 264	GB170	2		3		1		1																				
L3f1	209 223 292 295 311 362	GB177	1																										
L3f1	209 223 292 311	GB178					1																						
L3h	129 223 256A 311 362	GB184	2	1	1	1	1	1	1																				
M1	129 185 189 223 249 311	GB186	2		1																								
U5b1b	189 192 270 320	GB188					8		1																				
U6a	111 172 (183C) 189 219 278	GB189																											
U6a	172 (183C) 189 219 278	GB191					1	1	2				1	1	1					2									
U6a	172 (183C) 189 219 278 390	GB192					1																						

Haplotype ID as in Table S1; "1d" - local database; "\*" 114A=114G; "\*\*\*" 16390 not determined in the original publications.

Original references 1 - Kringset *et al.* 1999, 2 - Mestpalu *et al.* 2004, 3 - Di Rienzo & Wilson 1991, 4 - Richardset *et al.* 2000, 5 - Cali *et al.* 2000, 6 - Francalacci *et al.* 1996, 7 - Pereira *et al.* 2000, 8 - Larrugaet *et al.* 2001, 9 - Crespiello *et al.* 2000, 10 - Helgasson *et al.* 2000, 11 - Budowle *et al.* 1999



Table S6 – Matrix of F<sub>ST</sub> distances of mtDNA haplogroup profile among African populations

	Population units																				
	Mar	MBb	Aar	Abb	Mzt	Sah	Mau	Egy1	Egy2	Nub	Amh	Tig	Gur	Oro	Som	Tur	Kik	Nai	Fni	Hau	Tug
Mar	0.0000																				
MBb	0.00953**	0.00000																			
Aar	0.05809**	0.12741***	0.00000																		
Abb	0.00901	0.00776	0.10551***	0.00000																	
Mzt	0.03373**	0.06672***	0.07638***	0.02201	0.00000																
Sah	0.03859**	0.09351**	0.01797	0.07735**	0.04975*	0.00000															
Mau	0.05784**	0.12953***	0.03007*	0.08847**	0.03298	0.01219	0.00000														
Egy1	0.01118**	0.00204	0.11519***	0.01949**	0.07456***	0.08733**	0.13240***	0.00000													
Egy2	0.00755	0.01253**	0.08274***	0.02579**	0.06657***	0.05490**	0.09303***	0.00132	0.00000												
Nub	0.13132**	0.19920**	0.03576**	0.17085**	0.12743**	0.05753**	0.05422***	0.17262***	0.12801***	0.00000											
Amh	0.17567**	0.25403**	0.07761**	0.19093**	0.11309**	0.08584***	0.04740**	0.23549***	0.18947***	0.03895***	0.00000										
Tig	0.21490**	0.30755***	0.13659***	0.22502**	0.11610***	0.12893**	0.06975**	0.29610***	0.24836***	0.10994**	0.01434	0.00000									
Gur	0.21264**	0.31773***	0.10963***	0.25061***	0.13762***	0.10871**	0.05362**	0.30406***	0.24591***	0.06802**	-0.00903	-0.01281	0.00000								
Oro	0.18126**	0.27670***	0.07380***	0.21014***	0.12130**	0.08302***	0.03635**	0.25858***	0.20400***	0.03547***	-0.00984	0.01290	-0.01656	0.00000							
Som	0.20494**	0.30454***	0.09359***	0.27566***	0.20207***	0.12058**	0.11457**	0.26953***	0.22032***	0.03963**	0.09268***	0.16458***	0.13391***	0.07979***	0.00000						
Tur	0.30803**	0.41152***	0.16939***	0.36160***	0.29896***	0.20150***	0.18463***	0.37732***	0.31793***	0.09118***	0.12282***	0.18955***	0.16616***	0.11627***	0.07452**	0.00000					
Kik	0.27352**	0.38696***	0.11584***	0.32943***	0.24419***	0.16018**	0.13301**	0.35706***	0.29796***	0.05950**	0.06282***	0.11731***	0.08768**	0.05042**	0.03182	0.00327	0.00000				
Nai	0.24963**	0.34359***	0.09444***	0.28231***	0.22484***	0.14139**	0.12266**	0.31226***	0.24900***	0.04678**	0.06080**	0.12288**	0.07900***	0.05630***	0.10142***	0.05252***	0.02197	0.00000			
Fni	0.20653**	0.30881**	0.05596**	0.25592***	0.19241**	0.11312**	0.05729***	0.29168***	0.23343***	0.07631**	0.09322***	0.15065***	0.11287**	0.07336***	0.10598***	0.13755***	0.08651***	0.06998***	0.00000		
Hau	0.25712**	0.37527***	0.05905**	0.33108***	0.24307***	0.12455***	0.10974**	0.35375***	0.29288***	0.06313**	0.09146**	0.16721***	0.12249***	0.07725***	0.09245***	0.09013**	0.10927***	0.05238***	0.04359**	0.00096	0.00000
Tug	0.20783**	0.31436***	0.04840*	0.28154***	0.19338***	0.10077**	0.07310**	0.29471***	0.24011**	0.02482	0.06471**	0.14785***	0.10779**	0.05406**	0.04379	0.12155***	0.06659**	0.07478***	0.02817	-0.00411	0.00000
Yor	0.24736**	0.35880**	0.07324***	0.30824***	0.23846***	0.14515**	0.10800**	0.33833***	0.27660***	0.07727**	0.10074***	0.17369**	0.13036**	0.08070**	0.10671***	0.13421***	0.07383**	0.06568***	0.00267	-0.01514	0.01505
Hid	0.23298**	0.34727***	0.05163**	0.29767***	0.22092***	0.12115**	0.10152**	0.32033***	0.26259***	0.05277**	0.07248***	0.14505**	0.10697**	0.06159**	0.08229**	0.10185**	0.03552*	0.04332**	0.02621*	-0.02236	0.02099
Maf	0.26564**	0.37726***	0.09011**	0.32196***	0.24957**	0.16084**	0.13759**	0.34762***	0.28750***	0.08590**	0.09116**	0.15999**	0.13170**	0.08202***	0.11276***	0.10407**	0.05294**	0.04332**	0.04257**	0.02541	0.06832**
Kot	0.29875**	0.41518***	0.13925**	0.37030**	0.29186***	0.19511**	0.16993**	0.38402***	0.32509***	0.08855**	0.10884**	0.17699**	0.14615**	0.09245***	0.03385	0.04615	-0.02301	0.05514**	0.10611**	0.08248**	0.09346**
Mas	0.26456**	0.37836***	0.08700**	0.32685***	0.24893***	0.15704**	0.13431**	0.34936***	0.29145***	0.07041**	0.09626**	0.16165**	0.13137**	0.08275**	0.04590**	0.06804**	0.00602	0.03773**	0.04616**	0.01336	0.04586**
FBo	0.22468**	0.32972***	0.08765***	0.28121***	0.22237***	0.13080**	0.08487**	0.31288**	0.24896**	0.10176**	0.11917**	0.17702**	0.13466**	0.09394**	0.14619**	0.15371**	0.11761**	0.08752**	0.00433	0.04534**	0.07853**
Dab	0.27161**	0.39177***	0.07917**	0.34236**	0.26256**	0.16974**	0.13444**	0.36966**	0.30661**	0.09142**	0.10091**	0.16722**	0.11535**	0.08952**	0.14698**	0.14718**	0.06766**	0.02361	0.00104	0.07302**	
Fal	0.27758**	0.38353***	0.09050**	0.33813**	0.25639**	0.16048**	0.13744**	0.35611**	0.29556**	0.03902**	0.07295**	0.16016**	0.11167**	0.06839**	0.07913**	0.09434**	0.04729**	0.02787**	0.05580**	0.00511	0.00518
Fca	0.18828**	0.29291**	0.04093**	0.24687**	0.18609**	0.08456**	0.06451**	0.27304**	0.20805**	0.04363**	0.06185**	0.12325**	0.06591**	0.05054*	0.10840**	0.13372**	0.07370**	0.03224**	0.03032**	0.02650	0.05720**
Mad	0.26585**	0.37379**	0.08786**	0.32857**	0.24967**	0.14036**	0.10570**	0.35080**	0.28644**	0.04662**	0.06871**	0.14732**	0.08081**	0.05612**	0.09476**	0.12733**	0.06647**	0.03694**	0.03896**	0.00637	0.02012
Po	0.26349**	0.37449**	0.07881**	0.31580**	0.24471**	0.15131**	0.12771**	0.35068**	0.28745**	0.07124**	0.07415**	0.13764**	0.08198**	0.06190**	0.11944**	0.12532**	0.04415**	0.01657	0.03606**	0.00917	0.06004**
Ta	0.26506**	0.38323**	0.07380**	0.34052**	0.25359**	0.14796**	0.11886**	0.35874**	0.29682**	0.04378**	0.06949**	0.14684**	0.08630**	0.05418**	0.05389**	0.09520**	0.01291	0.02177	0.03324**	-0.01077	0.01296
Tup	0.27949**	0.39354**	0.09875**	0.34466**	0.26979**	0.16735**	0.13916**	0.36827**	0.30360**	0.06217**	0.07013**	0.14462**	0.07471**	0.05778**	0.10361**	0.11778**	0.03019	0.01767	0.06422**	0.03433	0.06498**
Oul	0.25088**	0.36416**	0.07589**	0.31089**	0.22336**	0.13176**	0.08589**	0.34140**	0.27566**	0.03704**	0.03426**	0.10126**	0.04598**	0.02453	0.08584**	0.10168**	0.03465**	0.01137	0.02466**	0.00669	0.01667
Bak	0.27393**	0.37809**	0.10337**	0.31632**	0.25846**	0.16876**	0.13864**	0.35035**	0.28718**	0.08135**	0.09222**	0.15837**	0.11881**	0.08661**	0.13214**	0.11405**	0.04798**	0.01923**	0.07472**	0.03749**	0.08181**
Bam	0.26527**	0.36897**	0.08599**	0.31776**	0.24386**	0.14793**	0.12472**	0.33947**	0.27988**	0.04006**	0.06972**	0.15109**	0.11203**	0.06128**	0.05538**	0.07992**	0.02773	0.03265**	0.05369**	0.00418	0.00575
Bis	0.26643**	0.37059**	0.10297**	0.32493**	0.24943**	0.15035**	0.13436**	0.34364**	0.28214**	0.06500**	0.08256**	0.14795**	0.09565**	0.06924**	0.09176**	0.10126**	0.04147**	0.03490**	0.05736**	0.02947	0.06380**
Ewo	0.26037**	0.36235**	0.09463**	0.30516**	0.24080**	0.14131**	0.11170**	0.33426**	0.27118**	0.05872**	0.07721**	0.14434**	0.10229**	0.06857**	0.09820**	0.09602**	0.04545**	0.02251**	0.06251**	0.02401	0.05512**
FTc	0.23078**	0.34007**	0.08411**	0.28697**	0.22435**	0.11866**	0.09107**	0.31958**	0.25644**	0.10312**	0.11150**	0.16376**	0.12434**	0.08686**	0.13139**	0.13956**	0.09394**	0.07543**	0.01029	0.03336**	0.08677**
CV	0.21361**	0.28631**	0.06488**	0.24700**	0.19325**	0.10264**	0.09153**	0.26951**	0.22834**	0.07112**	0.08583**	0.14226**	0.10830**	0.07366**	0.09942**	0.12920**	0.07424**	0.06227**	0.02020**	-0.01483	0.02020
Sen	0.23038**	0.33468**	0.06765**	0.28663**	0.21417**	0.11124**	0.09232**	0.31013**	0.25280**	0.06222**	0.08222**	0.14804**	0.10661**	0.06686**	0.07647**	0.11088**	0.06404**	0.05712**	0.01245	0.00321	0.02791
Ser	0.23589**	0.34924**	0.07113**	0.30937**	0.23036**	0.12019**	0.07642**	0.32996**	0.26523**	0.06036**	0.08187**	0.15690**	0.10579**	0.06635**	0.09910**	0.13353**	0.08046**	0.06175**	0.00750	0.00356	0.01325
Wol	0.24242**	0.34512**	0.09053**	0.29470**	0.22559**	0.12778**	0.08505**	0.32651**	0.26525**	0.08077**	0.09037**	0.15452**	0.11338**	0.07637**	0.11507**	0.14726**	0.09612**	0.08560**	0.03063**	0.02767	0.03236**
Mak	0.29233**	0.38211**	0.14968**	0.33567**	0.28039**	0.13584**	0.15146**	0.36432**	0.30904**	0.14899**	0.16390**	0.22196**	0.18866**	0.14786**	0.18440**	0.19493**	0.15940**	0.13900**	0.08000**	0.06505**	0.10922**
Bab	0.26163**	0.37970**	0.09411**	0.33731**	0.25769**	0.11857**	0.09329**	0.36124**	0.29333**	0.08155**	0.10347**	0.17044**	0.12264**	0.08050**	0.08882**	0.11548**	0.05927**	0.06943**	0.01806	0.00905	0.03288
Mwk	0.23904**	0.33912**	0.06920**	0.29006**	0.22294**	0.10752**	0.09613**	0.31717**	0.25986**	0.06294**	0.09197**	0.16203**	0.12140**	0.07772**	0.07777**	0.12162**	0.06806**	0.06888**	0.02146**	-0.01494	0.00162
EJA	0.24941**	0.35474**	0.08000**	0.29738**	0.23004**	0.11339**	0.11213**	0.32860**	0.26820**	0.08266**	0.09781**	0.15635**	0.12267**	0.08778**	0.08925**	0.10455**	0.05542**	0.05289**	0.02281**	-0.00257	0.04437**
BJG	0.24634**	0.36411**	0.07855**	0.31758**	0.23512**	0.10894**	0.08640**	0.34020**	0.27095**	0.06228**	0.08222**	0.15040**	0.08618**	0.06240**	0.08181**	0.09230**	0.05377**	0.04137**	-0.00562	-0.00459	0.02336
PBE	0.24456**	0.34568**	0.07840**	0.28686**	0.22505**	0.11572**	0.09824**	0.31973**	0.25655**	0.07006**	0.08713**	0.15147**									

Table S6 (continued)

	Population units																						
	Yor	Hid	Maf	Kot	Mas	FBo	Dab	Fal	Fca	Mad	Po	Ta	Tup	Oul	Bak	Bam	Bis	Ewo	FTc	CV	Sen		
Mar																							
MBb																							
Aar																							
Abb																							
Mzt																							
Sah																							
Mau																							
Egy1																							
Egy2																							
Nub																							
Amh																							
Tig																							
Gur																							
Oro																							
Som																							
Tur																							
Kik																							
Nai																							
Fni																							
Hau																							
Tug																							
Yor	0.00000																						
Hid	-0.00705	0.00000																					
Maf	0.02494	-0.00174	0.00000																				
Kot	0.08632**	0.04882*	0.05221*	0.00000																			
Mas	0.04327**	0.00386	0.02314	0.00051	0.00000																		
FBo	0.02934**	0.06153**	0.05104**	0.13038***	0.08174***	0.00000																	
Dab	0.02095	-0.00911	0.03013	0.08607**	0.02109	0.06030**	0.00000																
Fal	0.04368**	0.01714	0.05252**	0.07753**	0.02371	0.10037***	0.03647*	0.00000															
Fca	0.03892**	0.02640	0.05058**	0.09282***	0.04388**	0.04045**	0.01139	0.04552**	0.00000														
Mad	0.02969**	0.01853	0.06349***	0.09052***	0.03829**	0.06561***	0.01910	0.00271	0.00688	0.00000													
Po	0.02978**	0.00478	0.03311**	0.06207**	0.02580	0.06849***	-0.02367	0.02735	0.01237	0.01631	0.00000												
Ta	0.01729	-0.00561	0.04152**	0.02445	-0.00952	0.07596**	-0.00700	-0.00715	0.00930	-0.01611	-0.01437	0.00000											
Tup	0.04326**	0.02000	0.04948**	0.03982	0.03731*	0.09259***	0.00137	0.03232	0.01814	0.01107	-0.01770	-0.02213	0.00000										
Oul	0.01883	0.00649	0.02601	0.05673**	0.02249	0.04856**	0.00314	-0.00143	0.00279	0.06018***	-0.01747	-0.00406	-0.01791	-0.00848	0.00000								
Bak	0.03959**	0.00005	0.03893**	0.05415**	0.03658**	0.10414***	0.00189	0.04776**	0.03908**	0.04243**	0.01363	0.01933	0.01487	0.01957	0.00000								
Bam	0.02477*	-0.00297	0.02589*	0.03791*	0.00609	0.09069***	0.04107**	-0.00882	0.04462**	0.01232	0.03248**	-0.00558	0.03215**	0.00343	0.02986*	0.00000							
Bis	0.04315**	0.00680	0.06662***	0.05766**	0.02780**	0.09578***	0.02209	0.03910**	0.02317*	0.00701	0.03107**	-0.00116	0.02214	0.01430	0.01458	0.02306*	0.00000						
Ewo	0.03998**	0.00069	0.04779**	0.05904**	0.02146	0.08786***	0.01802	0.02533**	0.02359**	0.01170	0.02997**	0.00953	0.03339**	0.00959	0.00197	0.01044	-0.00944	0.00000					
FTc	0.03171**	0.03630**	0.02925**	0.09974***	0.05356***	-0.00929	0.04084**	0.09690***	0.03067**	0.06018***	0.05185***	0.06074**	0.07787***	0.04468**	0.08239***	0.07802***	0.07092***	0.06480***	0.00000				
CV	0.01600*	0.00274	0.04208***	0.09293***	0.03811**	0.04834***	0.02509*	0.03581***	0.03741***	0.02145**	0.03074**	0.01744	0.05007**	0.01768*	0.05450***	0.03127***	0.04528***	0.03884***	0.03348***	0.00000			
Sen	0.03059**	0.01621	0.02312*	0.07587**	0.01886	0.02374**	0.03351**	0.03223**	0.02859**	0.01797	0.03716**	0.01665	0.05417**	0.01035	0.07534***	0.02871**	0.05099***	0.04464***	0.00921	0.00880	0.00000		
Ser	-0.00196	0.01719	0.02398	0.08879**	0.04607**	0.00940	0.03788	0.03845**	0.01380	0.00463	0.03971**	0.01758	0.04032	-0.00676	0.05881**	0.02692	0.04380**	0.04035**	0.01565	0.00920	-0.00180		
Wol	0.01760	0.04081**	0.04025**	0.10380***	0.07188***	0.02972**	0.07079**	0.06671***	0.03965**	0.03387**	0.06620***	0.04660**	0.06459**	0.01617	0.08149***	0.04934***	0.07049***	0.06678***	0.03575**	0.02812***	0.02201**		
Mak	0.09615***	0.09912***	0.12084***	0.17345***	0.13388***	0.07718***	0.13140***	0.13934***	0.11550***	0.10733***	0.12232***	0.12514***	0.14618***	0.10583***	0.14697***	0.12636***	0.13671***	0.12203***	0.05625***	0.04535***	0.06208***		
Bab	0.01425	0.03461	0.04558**	0.05579**	0.03747*	0.01753	0.05703**	0.06262**	0.02824	0.02642	0.04682**	0.02075	0.04629*	0.01554	0.06683**	0.04079**	0.05090**	0.04894**	0.01112	0.01134	0.00922		
Mwk	0.01251	0.01070	0.04154**	0.07901**	0.03164**	0.04671***	0.04560**	0.02752**	0.03813**	0.01549	0.04204**	0.01070	0.05159**	0.01434	0.06641***	0.01882**	0.04874***	0.04351***	0.03776**	-0.00192	0.00236		
EJA	0.03183**	0.01350	0.02657**	0.06580**	0.02262	0.04906***	0.02781	0.05029***	0.04057**	0.03989**	0.03267**	0.02488	0.05125**	0.02343*	0.05464***	0.04005**	0.05670***	0.04491***	0.02412**	0.00608	0.00441		
BJG	0.00686	0.01407	0.01485	0.06355**	0.02447	-0.00053	0.03483	0.03740**	0.02938**	0.02449	0.03444**	0.02637	0.05359**	0.00687	0.05483**	0.02489	0.05243**	0.03370**	-0.00805	0.00235	-0.01125		
PBO	0.00117	-0.00413	0.00990	0.07360***	0.03499**	0.02995**	0.03237**	0.04121**	0.03640**	0.03272**	0.03689**	0.03270**	0.05483**	0.01857	0.03524**	0.01981**	0.04175***	0.02423**	0.01571	0.00864	0.01464*		
BLE	0.02876**	0.01525	0.02305**	0.08848***	0.04400**	0.04529***	0.06063***	0.04696***	0.05019***	0.03750**	0.05680***	0.04235**	0.07273***	0.02458**	0.07572***	0.03140**	0.06223***	0.05019***	0.02589**	0.00768*	0.00143		
FUL	0.02737**	0.03352**	0.07107***	0.10989***	0.05800***	0.03156**	0.06566**	0.04864***	0.03742**	0.02652**	0.06544***	0.04039**	0.08140***	0.02903**	0.08566***	0.04335***	0.05899***	0.04954***	0.03043**	0.01058**	0.01095		
MNK	0.02742**	0.01354	0.03199**	0.08477***	0.04720**	0.04607***	0.05725**	0.05922***	0.05380***	0.04519**	0.05380***	0.04588**	0.07211***	0.03465**	0.06752***	0.03975**	0.05855***	0.04631***	0.02056**	0.00545	0.01111		
NAJ	0.01612	-0.00813	-0.00160	0.08872***	0.04051**	0.04927**	0.04376**	0.05286**	0.04208**	0.03974**	0.04740**	0.03937**	0.06389**	0.02736	0.05558**	0.02703**	0.05084**	0.04064**	0.02097	0.00428	0.00550		
Mde	0.01421	0.04276**	0.05756***	0.08749**	0.05213**	0.00909	0.05395**	0.05036**	0.02683**	0.01821	0.04706**	0.02698	0.05424**	0.01157	0.07429***	0.04157**	0.05292***	0.04824***	0.01657	0.02525***	0.01598		
Lko	0.00778	0.02813	0.05367**	0.11406***	0.05630**	0.02286	0.06362**	0.04118**	0.04420**	0.01514	0.06906***	0.03911**	0.08356***	0.01569	0.07893***	0.02839**	0.04401**	0.03612**	0.02803**	0.02225**	0.01359		
Lim	0.02957**	0.00846	0.04205**	0.08752***	0.02427**	0.05232***	0.03205*	0.02914**	0.04263**	0.02515**	0.03731**	0.01635	0.06071**	0.02300*	0.06834***	0.02805**	0.05523***	0.04458***	0.03430**	-0.00344	-0.00333		
Tmn	0.01471	0.00551	0.03273**	0.06980***	0.02410**	0.03078**	0.03519**	0.03591***	0.03123**	0.02350**	0.03631***	0.01979	0.05547***	0.01981**	0.05293***	0.02232**	0.03765***	0.02755**	0.01421*	0.00411	0.00364		
FBa	0.05076**	0.08583***	0.06690***	0.14694***	0.10461***	-0.01215	0.09722***	0.12327**	0.05799***	0.08120***	0.09968***	0.10284***	0.11901***	0.06480***	0.12909***	0.10785***	0.11531***	0.10525***	-0.00235	0.05771***	0.03036**		
FTI	0.05386**	0.07371***	0.06176***	0.13584***	0.08810***	-0.00849	0.08688***	0.11790***	0.04515**	0.08088***	0.08855***	0.09439***	0.11551***	0.06874***	0.12256***	0.10121***	0.10623***	0.09492***	-0.01258	0.05165***	0.02461**		
MoB	0.05419**	0.02329	0.04998**	0.08454**	0.05276**	0.13221***	0.07442**	0.02184**	0.10309***	0.06969***	0.07296***	0.05468**	0.08403***	0.04763**	0.04220***	0.01006	0.07425***	0.04113***	0.12724***	0.06689***	0.08266**		
Moz	0.05641**	0.03987**	0.09096**	0.12060**	0.06849**	0.15208***	0.09021**	0.01061	0.11614***	0.05682**	0.08565***	0.04111**	0.09195***	0.05169**	0.08062***	0.01695	0.08550***	0.06872***	0.15703***	0.06029***	0.08472***		
Ku	0.45659***	0.46387***	0.44500***	0.52295***	0.45021***	0.41009***																	

Table S6 (continued)

	Population units																							
	Ser	Wol	Mak	Bab	Mwk	EJA	BJG	BLE	PBO	FUL	MNK	NAJ	Mde	Lko	Lim	Tmn	FBa	FTi	MoB	Moz	Ku	Khw		
Mar																								
MBb																								
Aar																								
Abb																								
Mzt																								
Sah																								
Mau																								
Egy1																								
Egy2																								
Nub																								
Amh																								
Tig																								
Gur																								
Oro																								
Som																								
Tur																								
Kik																								
Nai																								
Fni																								
Hau																								
Tug																								
Yor																								
Hid																								
Maf																								
Kot																								
Mas																								
FBa																								
Dab																								
Fal																								
Fca																								
Mad																								
Po																								
Ta																								
Tup																								
Oul																								
Bak																								
Bam																								
Bis																								
Ewo																								
FTc																								
CV																								
Sen																								
Ser	0.00000																							
Wol	-0.02919	0.00000																						
Mak	0.06888**	0.07710***	0.00000																					
Bab	-0.02248	-0.01155	0.02689	0.00000																				
Mwk	-0.00824	0.00879	0.04712***	-0.01037	0.00000																			
EJA	0.01368	0.03095**	0.04376***	0.00105	0.00513	0.00000																		
BJG	-0.01182	0.00715	0.02231	-0.02276	-0.00454	-0.01179	0.00000																	
BLE	0.00261	0.01873*	0.04843***	0.00335	0.00693	0.00846	-0.01667	0.00000																
PBO	0.00482	0.01653*	0.04072***	0.00866	0.00170	0.00354	-0.00843	0.00255	0.00000															
FUL	0.00525	0.02424*	0.03939***	0.00342	0.00259	0.02458**	-0.00548	0.01631**	0.01398*	0.00000														
MNK	0.01753	0.02963**	0.01863*	0.00333	0.00504	-0.00058	-0.01283	-0.00283	-0.00863	0.01283	0.00000													
NAJ	-0.00091	0.01115	0.05228**	0.01055	0.00064	0.00174	0.00049	-0.00860	-0.01396	0.02828**	-0.00996	0.00000												
Mde	-0.00793	0.01054	0.05949***	-0.01548	0.01202	0.03405**	-0.01204	0.01483*	0.02710**	0.00655	0.02651**	0.03860**	0.00000											
Lko	-0.00773	0.01213	0.07755***	0.01067	0.01209	0.04485**	-0.00427	0.00816	0.01638	0.00376	0.02347**	0.02465	-0.00534	0.00000										
Lim	0.01413	0.03623**	0.04569***	0.01255	-0.00628	-0.00149	-0.00225	0.01341	0.00454	0.00765	0.00530	0.00622	0.02951**	0.02838*	0.00000									
Tmn	0.01044	0.02981**	0.03986***	0.00047	0.00124	0.00915	-0.01365	-0.00186	0.00473	0.00374	-0.00113	0.00536	0.00874	0.00699	0.00157	0.00000								
FBa	0.00878	0.02306*	0.06143***	0.01025	0.05001***	0.05337***	0.00098	0.03669**	0.04255***	0.03224**	0.04430***	0.04991**	0.01351	0.02776*	0.06095***	0.03854***	0.00000							
FTi	0.02563	0.04232**	0.05268***	0.01681	0.04911***	0.04622***	-0.00101	0.03196**	0.03769***	0.02701**	0.03378**	0.04397**	0.01771	0.03357*	0.05051***	0.02630**	-0.01116	0.00000						
MoB	0.07589***	0.09511***	0.15575***	0.09517***	0.06398***	0.07880***	0.05949**	0.04291***	0.07145***	0.08491***	0.07565***	0.06587***	0.09031***	0.07108***	0.06858***	0.06137***	0.15143***	0.14938***	0.00000					
Moz	0.07062**	0.09067***	0.17053***	0.09817***	0.04894***	0.08753***	0.08260**	0.06742***	0.07885***	0.08168***	0.09187***	0.08010***	0.09472***	0.07358***	0.06105***	0.07190***	0.17283***	0.17750***	0.02174***	0.00000				
Ku	0.47144***	0.43328***	0.42708***	0.48865***	0.40918***	0.40401***	0.45818***	0.39230***	0.38307***	0.39689***	0.40178***	0.45769***	0.41353***	0.46003***	0.39990***	0.35695***	0.43088***	0.43365***	0.35950***	0.39581***	0.00000			
Khw	0.11466***	0.12738***	0.19527***	0.12050***	0.11403***	0.10041***	0.11369***	0.09823***	0.12267***	0.14148***	0.12151***	0.10770***	0.13585***	0.14910***	0.11659***	0.11253***	0.16771***	0.16935***	0.11012***	0.11092***	0.23083***	0.00000		

Note: Population codes as in Table S2. Significance levels: \* - P<0.05, \*\* - P<0.01; \*\*\* - P<0.001

Table S7a – Analysis of Molecular Variance (AMOVA) of mtDNA haplogroups in African populations (1023 permutations)

Criteria	Ethnic clusters	Among groups				Among populations within groups				Within populations				
		%	Va	F <sub>CT</sub>	P	%	Vb	F <sub>SC</sub>	P	%	Vc	F <sub>ST</sub>	P	
Geography	African continent	<ul style="list-style-type: none"> <li>▪ Northwest</li> <li>▪ Northeast</li> <li>▪ West</li> <li>▪ Central</li> <li>▪ East</li> <li>▪ South</li> </ul>	11.97	0.05544	0.11970	0.00000± 0.00000	4.79	0.02219	0.05442	0.00000± 0.00000	83.24	0.38552	0.16760	0.00000± 0.00000
	Sub-Sahara	<ul style="list-style-type: none"> <li>▪ West</li> <li>▪ Central</li> <li>▪ East</li> <li>▪ South</li> </ul>	4.01	0.01855	0.04007	0.00000± 0.00000	4.74	0.02194	0.04935	0.00000± 0.00000	91.26	0.42257	0.08744	0.00000± 0.00000
	West Africa	<ul style="list-style-type: none"> <li>▪ Senegal</li> <li>▪ Mali</li> <li>▪ Guinea-Bissau</li> <li>▪ Sierra-Leone</li> <li>▪ Burkina-Faso</li> </ul>	0.70	0.00311	0.00696	0.06452± 0.00816	1.42	0.00633	0.01427	0.00000± 0.00000	97.89	0.43754	0.02113	0.00000± 0.00000
Linguistics	African continent	<ul style="list-style-type: none"> <li>▪ Afro-Asiatic</li> <li>▪ Nilo-Saharan</li> <li>▪ Niger-Congo</li> <li>▪ Khoisan</li> </ul>	9.77	0.04645	0.09767	0.00000± 0.00000	8.36	0.03973	0.09259	0.00000± 0.00000	81.88	0.38936	0.18122	0.00000± 0.00000
	African continent	<ul style="list-style-type: none"> <li>▪ AA Semitic</li> <li>▪ AA Berber</li> <li>▪ AA Chadic</li> <li>▪ Nilo-Saharan</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic-Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé</li> <li>▪ Khoisan</li> </ul>	11.14	0.05139	0.11142	0.00000± 0.00000	4.69	0.02162	0.05274	0.00000± 0.00000	84.17	0.38823	0.15828	0.00000± 0.00000
	Sub-Sahara	<ul style="list-style-type: none"> <li>▪ AA Semitic</li> <li>▪ AA Chadic</li> <li>▪ Nilo-Saharan</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé</li> <li>(excluding Khoisan)</li> </ul>	4.51	0.02082	0.04514	0.00000± 0.00000	2.49	0.01147	0.02604	0.00000± 0.00000	93.00	0.42886	0.07001	0.00000± 0.00000
	Sub-Sahara Niger-Congo	<ul style="list-style-type: none"> <li>▪ NC Bantu</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé</li> </ul>	3.41	0.01559	0.03412	0.00000± 0.00000	2.42	0.01104	0.02501	0.00000± 0.00000	94.17	0.43035	0.05827	0.00000± 0.00000
	West Africa	<ul style="list-style-type: none"> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé</li> </ul>	0.68	0.00304	0.00680	0.02542± 0.00489	1.57	0.00704	0.01584	0.00000± 0.00000	97.75	0.43701	0.02254	0.00000± 0.00000

Geographic and linguistic assignment according to information in Table S2.



Table S7b – Analysis of Molecular Variance (AMOVA) of mtDNA haplogroups in Guinea-Bissau ethnic groups (1023 permutations)

Criteria	Ethnic clusters	Among populations				Among populations within groups				Within populations			
		%	Va	F <sub>CT</sub>	P	%	Vb	F <sub>SC</sub>	P	%	Vc	F <sub>ST</sub>	P
Geography	<ul style="list-style-type: none"> <li>▪ Fula, Mandinga</li> <li>▪ Felupe-Djola, Papel, Balanta</li> <li>▪ Bijagós</li> <li>▪ Nálu</li> </ul>	-0.44	-0.00196	-0.00435	0.78592± 0.01115	0.74	0.00332	0.00733	0.05279± 0.00628	99.70	0.44942	0.00300	0.11828± 0.01036
	<ul style="list-style-type: none"> <li>▪ Fula, Mandinga</li> <li>▪ Felupe-Djola, Papel, Balanta, Nálu</li> <li>▪ Bijagós</li> </ul>	-0.03	-0.00013	-0.00029	0.55230± 0.01714	0.44	0.00199	0.00440	0.143099± 0.01145	99.59	0.44942	0.00411	0.13001± 0.00945
Linguistic	<ul style="list-style-type: none"> <li>▪ Bijagós</li> <li>▪ Felupe-Djola, Balanta, Papel</li> <li>▪ Nálu</li> <li>▪ Fula, Mandinga</li> </ul>	0.02	0.00010	0.00023	0.45552± 0.01517	0.41	0.00184	0.00407	0.17302± 0.01258	99.57	0.44942	0.00430	0.11926± 0.00942
	<ul style="list-style-type: none"> <li>▪ Bijagós</li> <li>▪ Felupe-Djola, Balanta</li> <li>▪ Papel</li> <li>▪ Fula</li> <li>▪ Nálu, Mandinga</li> </ul>	0.30	0.00135	0.00300	0.45259± 0.01264	0.15	0.00066	0.00146	0.35191± 0.01537	99.55	0.44942	0.00446	0.11241± 0.00891
	<ul style="list-style-type: none"> <li>▪ Bijagós</li> <li>▪ Felupe-Djola, Balanta, Papel, Nálu, Fula</li> <li>▪ Mandinga</li> </ul>	-1.32	-0.00590	-0.01317	0.94624± 0.00699	0.99	0.00442	0.00973	0.01369± 0.00309	100.33	0.44942	-0.00331	0.11730± 0.01278
Religion*	<ul style="list-style-type: none"> <li>▪ Felupe-Djola, Papel, Nálu, Bijagós</li> <li>▪ Fula, Mandinga</li> </ul>	0.37	0.00165	0.00366	0.28152± 0.01063	0.23	0.00102	0.00227	0.30596± 0.01463	99.41	0.44840	0.00592	0.14272± 0.01117

\* mostly Animists versus mostly Muslims. Negative values mean that one of the units is more similar to one in other cluster.

Table S8a – Statistical indices calculated from the mtDNA nucleotidic sequences (by haplogroup)

Haplogroup	L0a	L1b	L1c	L2a	L2b	L2c	L2d	L3b	L3d	L3e	L3f	L3h	M1	U5	U6
N	19	38	19	61	29	61	7	32	35	27	9	13	4	10	8
n hp	6	18	14	40	9	29	6	16	18	12	7	5	2	2	3
n trans	6	18	27	35	12	29	16	14	19	16	11	5	1	2	2
n transv	2	1	1	2	0	1	1	1	0	0	0	0	0	0	0
n subst	8	19	28	37	12	30	17	15	19	16	11	5	1	2	2
polim sites	8	18	28	36	12	29	17	15	19	16	11	5	1	2	2
Mean pairwise difference	1.0567	2.7408	8.8072	3.8681	1.8487	2.1114	9.6884	3.1240	3.7427	3.3413	4.3514	1.1586	0.5065	0.4105	0.6893
sd	0.7320	1.4859	4.2507	1.9696	1.0918	1.1935	5.0609	1.6635	1.9345	1.7691	2.3731	0.7979	0.5234	0.4093	0.5802
Nucl diversity	0.0028	0.0073	0.0234	0.0103	0.0049	0.0056	0.0257	0.0083	0.0099	0.0089	0.0115	0.0031	0.0013	0.0011	0.0018
sd	0.0022	0.0044	0.0126	0.0058	0.0032	0.0035	0.0154	0.0049	0.0057	0.0052	0.0071	0.0024	0.0017	0.0012	0.0018
Gene diversity	0.6023	0.9260	0.9708	0.9787	0.6404	0.8311	0.9524	0.9415	0.9378	0.8519	0.9444	0.6923	0.5000	0.2000	0.6071
sd	0.1242	0.0255	0.0239	0.0077	0.0934	0.0490	0.0955	0.0202	0.0237	0.0476	0.0702	0.1187	0.2652	0.1541	0.1640
Tajima D	-1.8824	-1.2756	-0.0920	-1.7255	-1.3846	-2.1705	1.0809	-0.6811	-0.7935	-0.8409	0.0096	-1.0688	-0.6124	-1.4009	-0.4479
P (sim<obs)	0.0150	0.0870	0.5360	0.0180	0.0720	0.0030	0.8970	0.2660	0.2290	0.2390	4.0687	0.1690	0.3680	0.7049	0.3500
Fu's Fs	-2.2625	-10.6175	-3.2076	-26.0725	-2.7564	-27.1834	-0.1804	-7.7759	-8.3836	-3.3959	-1.7403	-1.4540	0.1719	0.5862	-0.4776
P	0.0370	0.0000	0.0890	0.0000	0.0480	0.0000	0.3740	0.0010	0.0020	0.0530	0.1280	0.0650	0.3560	0.4340	0.1250
Harpending's Raggedness index	0.0627	0.0311	0.0190	0.0265	0.1675	0.0205	0.0522	0.0384	0.0249	0.0604	0.0340	0.0746	0.2500	0.7200	0.2411
P	0.8600	0.6200	0.4300	0.2700	0.2000	0.9300	0.9700	0.5600	0.5200	0.2500	0.9400	0.7800	0.9100	0.6300	0.3000

Abbreviations as follows: "n hp" – number haplotypes, "n trans" – number of transitions, "n transv" – number of transversions, "n subst" – number of substitutions, "polim" – polymorphic, "sd" – standard deviation. Note: Calculated with a Kimura-2P parameter ( $\gamma$  0.26; Meyer *et al.* 1999).

Table S8b – Statistical indices calculated from the mtDNA nucleotidic sequences (by ethnic group)

Ethnic group	Felupe-Djola	Bijagós	Balanta	Papel	Fulbe	Mandenka	Nalú
N	50	22	62	77	77	58	26
n hp	38	15	50	50	51	50	20
n trans	47	35	53	56	57	61	41
n transv	4	3	7	5	5	4	2
n subst	51	38	6	61	62	65	43
polim sites	49	37	54	58	59	64	43
Mean pairwise difference	8.0743	10.0005	10.6808	8.8267	8.9865	9.4867	9.0886
sd	3.8122	4.7542	4.9293	4.1150	4.1841	4.4162	4.3219
Nucl diversity	0.0214	0.0265	0.0283	0.0234	0.0238	0.0252	0.0241
sd	0.0112	0.0141	0.0145	0.0121	0.0123	0.0130	0.0128
Gene diversity	0.9902	0.9567	0.9921	0.9802	0.9764	0.9915	0.9692
sd	0.0056	0.0276	0.0046	0.0068	0.0083	0.0064	0.0220
Tajima D	-1.1713	-0.5317	-0.6401	-1.1065	-1.1236	-1.3601	-1.0881
P (sim<obs)	0.1236	0.3191	0.2793	0.0138	0.1336	0.0833	0.1468
Fu's Fs	-24.9468	-2.5533	-24.6740	-24.8486	-24.8353	-24.8188	-7.4231
P	0.0000	0.1430	0.0000	0.0000	0.0000	0.0000	0.0070
Harpending's Raggedness index	0.0046	0.0079	0.0039	0.0044	0.0056	0.0034	0.0091
P	0.9400	0.9700	0.9200	0.8800	0.8300	0.9700	0.8600

Abbreviations as follows: "n hp" – number haplotypes, "n trans" – number of transitions, "n transv" – number of transversions, "n subst" – number of substitutions, "polim" – polymorphic, "sd" –standard deviation. Note: Calculated with a Kimura-2P parameter (gamma 0.26; Meyer *et al.* 1999).

Table S9 –Y-chromosome SNP-defined haplogroups and associated extended Y-STR haplotypes among Guinea-Bissau ethnic groups

Haplotype	Haplogroup	Y-STR marker											Ethnic group							
		DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS385	DYS460	Felupe-Djola	Bijagos	Balanta	Papel	Fulbe	Mandenka	Nalu
H1	A1	14	12	30	23	10	11	14	14	8	12	15,16			1					
H2	A1	14	13	30	23	11	11	14	14	8	13	16,16		1						
H3	A1	14	13	31	22	11	11	14	14	8	12	17,17				1				
H4	A1	14	13	31	22	11	11	14	14	8	14	17,19					1			
H5	A1	15	13	32	23	11	11	13	14	8	12	17,17					1			
H6	A1	15	13	32	24	11	11	14	14	8	12	17,17					1			
H7	A1	15	14	31	23	11	11	14	14	8	12	15,17					1			
H8	B	15	13	29	24	10	14	12	14	10	12	14,15								1
H9	DE	13	12	28	21	11	11	13	14	8	11	15,16								1
H10	E1*	15	12	29	22	11	10	13	17	10	11	15,16					1			
H11	E1*	15	12	29	22	11	11	13	17	10	11	15,15				1				
H12	E1*	15	12	29	22	11	11	13	17	10	12	14,14					1			
H13	E1*	15	12	30	21	10	10	13	17	10	13	14,15						1		
H14	E1*	15	12	30	21	10	11	13	17	10	12	15,16								
H15	E1*	15	13	29	22	11	11	14	17	10	12	14,14		1						
H16	E1*	15	13	31	24	10	10	14	17	10	12	16,17				1				
H17	E1*	15	14	29	22	10	10	14	17	10	12	14,17		1						
H18	E1*	15	14	31	22	10	11	14	17	10	13	14,16						1		
H19	E1*	16	12	29	22	10	11	13	17	8	11	15,15		1						
H20	E1*	16	12	30	22	9	11	14	16	10	13	16,16							1	
H21	E1*	16	14	29	23	11	10	13	17	10	12	15,16		1						
H22	E2	14	12	28	24	10	11	13	13	11	11	19,19						1		
H23	E3*	13	13	30	22	10	11	14	14	10	12	14,17					1			
H24	E3*	13	14	31	21	10	11	14	15	10	11	16,17				1				
H25	E3*	14	14	32	21	10	10	14	14	10	11	17,17		1						
H26	E3a	13	13	30	21	10	11	14	14	11	14	15,16							1	
H27	E3a	13	13	31	23	10	11	13	14	10	11	16,17								1
H28	E3a	13	13	31	24	10	11	14	14	10	12	16,17							1	
H29	E3a	14	13	30	21	10	11	14	15	11	13	18,18			1					
H30	E3a	14	13	30	21	10	13	14	14	11	11	16,19								
H31	E3a	15	12	29	21	10	11	13	13	11	12	15,16					1			
H32	E3a	15	12	29	22	10	11	13	17	10	12	14,15		1						
H33	E3a	15	12	29	22	10	11	13	17	10	13	13,17								
H34	E3a	15	12	29	22	10	11	13	17	10	13	15,17				1				
H35	E3a	15	12	29	22	11	10	13	16	8	11	15,15							1	
H36	E3a	15	12	29	22	11	11	13	17	10	12	14,14							1	
H37	E3a	15	12	30	21	10	11	14	13	11	11	15,16					1			
H38	E3a	15	12	30	21	10	11	14	13	11	12	15,16							1	
H39	E3a	15	12	30	21	10	11	14	13	11	13	15,16					1			
H40	E3a	15	12	30	21	10	11	14	13	11	13	16,16						1		



Table S9 (continued)

Haplotype	Haplogroup	Y-STR marker											Ethnic group								
		DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS385	DYS460	Felupe-Djola	Bijagós	Balanta	Papel	Fulbe	Mandenka	Nalú	
H81	E3a	15	14	32	21	10	11	13	14	11	11	16,17									
H82	E3a	15	14	32	21	10	11	13	14	11	12	16,17									
H83	E3a	15	14	32	21	10	11	14	14	11	12	16,16									
H84	E3a	16	11	28	21	10	11	13	14	11	12	16,17									
H85	E3a	16	12	28	21	10	11	14	14	11	12	16,16									
H86	E3a	16	12	30	21	10	11	14	14	11	13	15,16				1				1	
H87	E3a	16	12	30	21	10	11	14	14	11	13	16,16								1	
H88	E3a	16	12	30	21	11	11	14	14	11	12	17,17				1					
H89	E3a	16	13	29	21	10	11	14	14	11	11	15,18								1	
H90	E3a	16	13	29	21	10	11	14	14	11	11	16,18		1							
H91	E3a	16	13	29	21	10	11	14	14	11	12	15,18								1	
H92	E3a	16	13	29	21	10	11	14	14	11	12	16,16									
H93	E3a	16	13	30	21	10	11	14	14	12	12	16,16				1					1
H94	E3a	16	13	30	21	11	11	14	14	11	11	17,17		1							
H95	E3a	16	13	30	21	11	11	14	14	11	12	17,18								1	
H96	E3a	16	13	30	22	10	11	14	14	12	12	16,16		1							
H97	E3a	16	13	30	22	10	11	14	15	11	11	15,17								1	
H98	E3a	16	13	31	21	10	11	14	14	10	12	16,16									
H99	E3a	16	13	31	21	10	11	14	14	11	10	16,16		1							
H100	E3a	16	13	31	21	10	11	14	14	12	11	16,16			1						
H101	E3a	16	13	31	21	10	11	15	14	10	12	16,17		1							
H102	E3a	16	13	31	22	10	11	14	14	11	12	14,17									1
H103	E3a	16	13	32	21	10	11	13	14	11	12	17,17		1							
H104	E3a	16	13	32	21	10	11	14	14	11	10	16,16								1	
H105	E3a	16	13	32	21	11	11	13	14	11	11	17,18				1					
H106	E3a	16	13	32	21	12	11	14	14	11	13	15,17								1	
H107	E3a	16	14	31	21	10	11	13	14	11	12	15,17								1	
H108	E3a	16	14	31	21	10	11	14	14	10	12	16,18									1
H109	E3a	16	14	31	21	10	11	14	14	11	11	17,18		1							
H110	E3a	16	14	31	21	10	11	14	14	11	12	16,16			1						
H111	E3a	16	14	31	21	10	12	14	15	12	11	18,19			1						
H112	E3a	16	14	31	22	10	11	13	14	11	13	16,16								1	
H113	E3a	16	14	31	22	10	11	13	14	11	14	16,17									
H114	E3a	16	14	31	22	10	11	13	15	11	12	16,17				1					
H115	E3a	16	14	31	22	10	11	14	14	11	11	15,17					1				
H116	E3a	16	14	31	22	10	11	14	15	11	11	16,18					1				
H117	E3a	16	14	32	21	10	11	13	14	11	12	15,15								1	
H118	E3a	16	14	32	21	10	11	14	14	11	11	15,17					1				
H119	E3a	16	14	32	21	10	11	14	15	11	12	17,2				1					
H120	E3a	16	14	32	21	10	11	14	15	11	13	20,21				1					



Table S9 (continued)

Haplotype	Haplogroup	Y-STR marker											Ethnic group							
		DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS385	DYS460	Felupe-Djola	Bijagós	Balanta	Papel	Fulbe	Mandenka	Nalú
H161	E3b1	13	13	31	24	10	12	12	15	10	12	15,15	11							
H162	E3b1	13	13	31	24	11	11	13	14	10	10	15,16	10					1		
H163	E3b1	13	13	32	25	11	11	13	15	10	12	15,18	11							
H164	E3b1	14	13	30	23	10	11	14	14	10	10	13,14	12		1					
H165	R1b	14	13	29	24	10	13	13	15	12	12	11,14			1					
H166	R1b	15	14	31	25	10	13	13	14	12	13	13,15					1			
H167	E1*		12	29	22	10	10	14		12	11	15,16		1						
H168	E1*		12	29	22	11		13			11	15,16		1						
H169	E1*		12	30	22	11		13			12	13,16				1				
H170	E1*		13	30	22	10		13			12	13,16				1				
H171	E1*		14		22	10		13			13	15,17								
H172	E3a		11	30	21	10		14			11	16,17			1				1	
H173	E3a	16	11		21	10		13	14											
H174	E3a		12	29	21	10		14			12	16,19								
H175	E3a		12	29	21	10		14			14	16,17		1		1				
H176	E3a		12	29	21	11		14			12	17,18							1	
H177	E3a		12	29	21	11		14	14		13	16,17		1						
H178	E3a		12	29	22	11		13			11	15,16		1						
H179	E3a		12	29	22	11		13			12	14,14				1				
H180	E3a		12	30	21	10		14			12	16,16						1		
H181	E3a		12	30	21	10		14			13	14,17			1					
H182	E3a		12	30	21	10		14			13	15,16						1		
H183	E3a		13	28	21	10		13			12	16,17						1		
H184	E3a		13	29	21	10		13			12	16,16						1		
H185	E3a		13	29	21	10		14			12	16,18				1				
H186	E3a		13	29	21	10		14			12									1
H187	E3a		13	30	20	10		14			11	15,16						1		
H188	E3a		13	30	21	9		13			13	15,15						1		
H189	E3a		13	30	21	9		13			13	15,16						1		
H190	E3a		13	30	21	10		13			12	16,16								1
H191	E3a		13	30	21	10		14			12	14,14				1				
H192	E3a		13	30	21	10		14			12	15,16						1		
H193	E3a		13	30	21	10		14			12	17,17						1		
H194	E3a		13	30	21	10		14			12	17,18		1						
H195	E3a		13	30	21	10		14			13	16,18						1		
H196	E3a		13	30	21	10		15			11	16,19							1	
H197	E3a		13	30	21	10		15			12	15,15							1	
H198	E3a		13	30	21	10		15			12	16,16			1					1
H199	E3a		13	30	21	11		12			11	17,17							1	
H200	E3a		13	30	21	11		13			11	16,17				1				





Table S10 – African datasets used for comparison of Y-chromosome genetic profiles

Geographic region/Ethnic group	Abbreviation	N	Linguistic		Reference	
			Family	Sublevel		
<i>Northwest Africa</i>						
West Sahara	Saharawis	Sah	29	Afro-Asiatic	Semitic	Bosch <i>et al.</i> 2001
Morocco	Arabs	MAR	93	Afro-Asiatic	Semitic	Scozzari <i>et al.</i> 1999, 2000; Bosch <i>et al.</i> 2001
	Berbers	MBb	167	Afro-Asiatic	Berber	Scozzari <i>et al.</i> 2001, Bosch <i>et al.</i> 2001
Algeria	Algerians	Alg	32	Afro-Asiatic	Mixed	Semino <i>et al.</i> 2004
	Arabs	AAR	35	Afro-Asiatic	Semitic	Semino <i>et al.</i> 2004
Tunisia	Tunisians	Tun1	146	Afro-Asiatic	Semitic	Arredi <i>et al.</i> 2004
	Tunisians	Tun2	28	Afro-Asiatic	Semitic	Wood <i>et al.</i> 2005
<i>Northeast Africa</i>						
Egypt	Egyptians	Egy1	73	Afro-Asiatic	Mixed	Arredi <i>et al.</i> 2004
	Egyptians	Egy2	92	Afro-Asiatic	Mixed	Wood <i>et al.</i> 2005
Sudan	Sudanese	Sud	40	Mixed	Mixed	Underhill <i>et al.</i> 2000
<i>West Africa</i>						
Cape Verde	Cape Verdeans	CV	201	Creole	Portuguese-based	Gonçalves <i>et al.</i> 2005
Senegal/Gambia	Senegalese	Mak	39	Niger-Congo	Manding	Wood <i>et al.</i> 2005
	Senegalese	Se	139	Niger-Congo	Mixed	Semino <i>et al.</i> 2002
	Wolof	Wo	34	Niger-Congo	Atlantic-Wolof	Wood <i>et al.</i> 2005
Mali	Mali	Mal	44	Niger-Congo	Mixed	Underhill <i>et al.</i> 2000
	Dogon	Do	55	Niger-Congo	Dogon	Wood <i>et al.</i> 2005
Guinea-Bissau	Felupe-Djola	EJA	50	Niger-Congo	Atlantic-Bak	Present study
	Bijagós	BJG	21	Niger-Congo	Atlantic-Bijagó	Present study
	Balanta	BLE	26	Niger-Congo	Atlantic-Bak	Present study
	Papel	PBO	64	Niger-Congo	Atlantic-Bak	Present study
	Fulbe	FUL	59	Niger-Congo	Atlantic-Fulani	Present study
	Mandenka	MNK	45	Niger-Congo	Manding West	Present study
	Nalú	NAJ	16	Niger-Congo	Atlantic-Nalu	Present study
Burkina-Faso	Fulbe	FBF	20	Niger-Congo	Fulani	Scozzari <i>et al.</i> 1997, 1999
	Mossi	Mo	49	Niger-Congo	Gur	Scozzari <i>et al.</i> 1997, 1999
	Rimaibe	Ri	37	Niger-Congo	Mande	Scozzari <i>et al.</i> 1997, 1999
Ghana	Ewe	Ewe	30	Niger-Congo	Kwa	Wood <i>et al.</i> 2005
	Ga	Ga	29	Niger-Congo	Kwa	Wood <i>et al.</i> 2005
	Fante	Fan	32	Niger-Congo	Kwa	Wood <i>et al.</i> 2005
<i>Central Africa</i>						
Chad	Mandara	Mad	28	Afro-Asiatic	Chadic	Wood <i>et al.</i> 2005
	Ouldeme	Ou1	52	Afro-Asiatic	Chadic	Scozzari <i>et al.</i> 1997, 1999; Wood <i>et al.</i> 2005
	Podokwo	Po	19	Afro-Asiatic	Chadic	Wood <i>et al.</i> 2005
North Cameroon	Daba	Dab	18	Afro-Asiatic	Chadic	Scozzari <i>et al.</i> 1997, 1999
	Fali	Fal	39	Niger-Congo	Adamawa	Scozzari <i>et al.</i> 1997, 1999
	Tali	Ta	15	Niger-Congo	Adamawa	Scozzari <i>et al.</i> 1997, 1999
	Fulbe	Fca	17	Niger-Congo	Fulani	Scozzari <i>et al.</i> 1997, 1999
South Cameroon	Bakaka	Bak	29	Niger-Congo	Bantu	Scozzari <i>et al.</i> 1999, Wood <i>et al.</i> 2005
	Bamileke	Bam	48	Niger-Congo	Bantu	Scozzari <i>et al.</i> 1997, 1999
	Bassa	Bis	11	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Ewondo	Ewo	29	Niger-Congo	Bantu	Scozzari <i>et al.</i> 1997, 1999
	Ngoumba	Ng	31	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Bakola	Bko	33	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
CAR	Biaka Pygmies	Bik	51	Niger-Congo	Bantu	Underhill <i>et al.</i> 2000, Wood <i>et al.</i> 2005
DRC	Mbuti Pygmies	Mb	59	Niger-Congo	Bantu	Underhill <i>et al.</i> 2000, Wood <i>et al.</i> 2005
	Nande	Nad	18	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Herna	Hen	18	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
<i>East Africa</i>						
Ethiopia	Ethiopians	Eth	88	Afro-Asiatic	Mixed	Underhill <i>et al.</i> 2000
	Oromo	Or	87	Afro-Asiatic	Cushitic	Semino <i>et al.</i> 2002, Wood <i>et al.</i> 2005
	Amhara	Am	66	Afro-Asiatic	Semitic	Semino <i>et al.</i> 2002, Wood <i>et al.</i> 2005
Uganda	Ganda	Gan	26	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
Kenya	Kikui & Kamba	K&K	42	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Maasai	Maa	26	Nilo-Saharan	Nilotic	Wood <i>et al.</i> 2005
<i>South Africa</i>						
Namibia	Herero	Her	24	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Ambo	Am	22	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	IKung Sekele	Ku	96	Khoisan	Northern	Scozzari <i>et al.</i> 1997, 1999; Wood <i>et al.</i> 2005
	Tsumkwe San		29	Khoisan	Central	Wood <i>et al.</i> 2005
	Dama	CKhoisan	18	Khoisan	Central	Wood <i>et al.</i> 2005
	Nama		11	Khoisan	Central	Wood <i>et al.</i> 2005
South Africa	Sotho-Tswana	ST	28	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Zulu	Zu	29	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Xhosa	Xh	80	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Shona	Sh	49	Niger-Congo	Bantu	Wood <i>et al.</i> 2005
	Khoisan	Khoi	39	Khoisan	Mixed	Underhill <i>et al.</i> 2000

Linguistic classification follows pertinent information from [www.ethnologue.com](http://www.ethnologue.com).

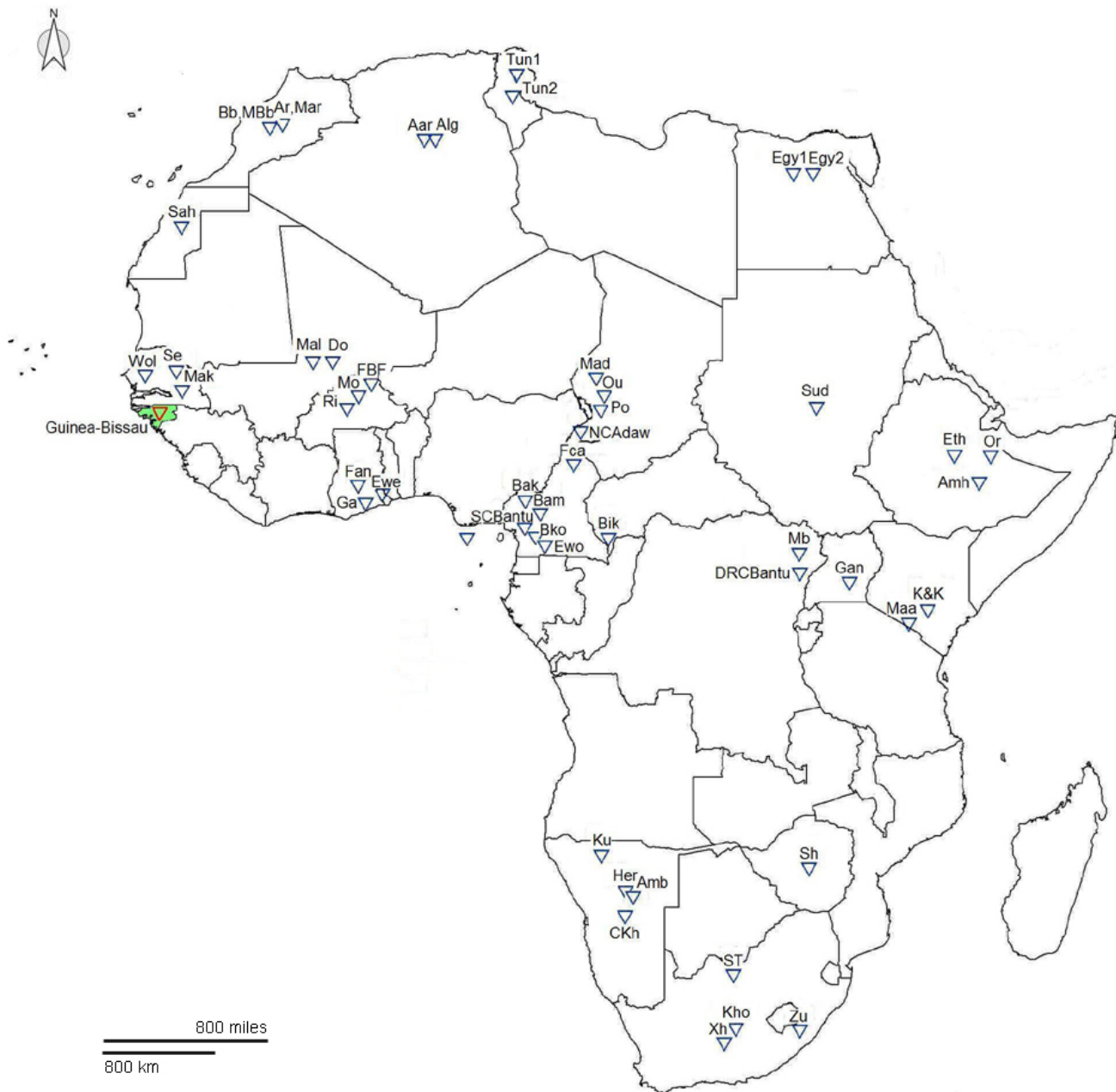


Figure S3 – Geographic location of the African datasets used for comparison of Y-chromosome genetic profiles (details in Table S10).

Table S11 – Absolute frequencies of Y-chromosome haplogroups for several African populations, including Guinea-Bissau ethnic groups, and respective diversity indexes (H, sd)

Haplogroup	North Africa								Northeast Africa				West Africa																			
	Saharawis	Arabs Morocco <sup>1</sup>	Arabs Morocco <sup>2</sup>	Berbers Morocco <sup>1</sup>	Berbers Morocco <sup>2</sup>	Algerians	Arabs Algeria	Tunisians <sup>1</sup>	Tunisians <sup>2</sup>	Egyptians <sup>1</sup>	Egyptians <sup>2</sup>	Sudanese	Cape Verdeans	Mandinka	Senegalese	Wolof	Mali	Dogon	Felupe-Djola	Bijagos	Balanta	Papel	Fulbe	Mandenka	Naiu	Fulbe Burkina-Faso	Mossi	Rimaibe	Ewe	Ga	Fante	
A1-M31	0	0	0	2	0	0	0	0	0	0	0	0	1	2	0	0	1	1	1	0	1	5	1	0	0	0	0	0	0	0	0	
A2-M14	0	0	0	0	0	0	0	0	0	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
A3-M32	0	0	0	0	0	0	0	0	0	1	3	18	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
B1-M146	0	0	0	0	0	0	0	0	0	0	0	0	0	1	nd	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0		
B2-M182	0	0	0	0	0	0	0	0	0	0	2	6	0	0	0	4	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
E*-SRY4064	0	0	0	0	0	0	0	0	0	0	1	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
E1-M33	1	0	0	1	2	0	1	1	0	0	1	1	10	1	7	4	15	25	17	1	3	13	4	4	2	2	0	2	0	1	1	
E2-M75	0	0	0	0	0	0	0	0	0	0	0	2	0	1	4	1	0	1	1	0	0	0	1	0	0	0	2	10	0	0	0	
E3*-PN2	0	0	0	0	0	0	0	0	0	0	0	0	1	0	4	1	0	0	1	0	1	1	0	0	0	0	1	1	1	0	1	
E3a*-M2	1	0	3	3	7	0	0	2	0	0	2	0	32	31	112	23	9	21	29	16	19	44	43	37	12	18	33	21	22	18	14	
E3a7-M191	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	3	0	0	0	0	1	2	0	0	11	3	7	10	13	
E3b*-M35	0	0	1	0	8	1	1	5	0	2	8	0	2	0	7	2	0	0	1	0	0	1	2	2	0	0	1	0	0	0	1	
E3b1-M78	0	21	5	7	6	2	4	8	4	17	29	7	35*	2	1	0	0	0	0	3	1	0	6	0	1	0	0	0	0	0	0	
E3b2-M81	22	16	23	44	67	17	14	56	10	7	0	2	6	1	1	2	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E3b3-M123	0	0	0	0	0	1	0	2	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J-12f2.1	0	10	1	4	3	7	10	53	13	15	22	0	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R-M207	0	1	3	0	2	nd	0	10	1	9	7	0	46	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	2
Other	5	1	8	3	8	4	5	9	0	16	17	4	39	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
H	0.4064	0.6811	0.6860	0.5139	0.5606	0.6694	0.7445	0.7129	0.6587	0.8349	0.8067	0.7474	0.8545	0.3698	0.3462	0.5348	0.7622	0.6505	0.5576	0.4143	0.4646	0.4871	0.4594	0.3192	0.4417	0.1895	0.5034	0.6111	0.4207	0.5123	0.6573	
sd	0.1012	0.0331	0.0632	0.0721	0.0555	0.0720	0.0450	0.0236	0.0524	0.0165	0.0215	0.0542	0.0097	0.0981	0.0517	0.0980	0.0325	0.0394	0.0495	0.1241	0.1160	0.0632	0.0774	0.0864	0.1446	0.1081	0.0707	0.0667	0.0874	0.0626	0.0528	
N	29	49	44	64	103	32	35	146	28	73	92	40	201	39	139	34	44	55	50	21	26	64	59	45	16	20	49	37	30	29	32	

Haplogroup	Central Africa								East Africa				South Africa																			
	Mandara	Podokwo	Ouldemé	NCAdaw (Fali, Tali)	Fulbe Cameroon	Bakala	Barnilake	SCBantu (Bassa, Ngoumba)	Ewondo	Bakola	Biaka Pygmies	Mbuti Pygmies	DRCBantu (Nande, Herza)	Ethiopians	Oromo	Amhara	Ganda	Kikuu & Kamba	Maasai	Herero	Ambo	!Kung Sekete	Ckhoisan (Tsumkwe San, Dama, Nama)	Sotho-Tswana	Zulu	Xhosa	Shona	Khoisan				
A1-M31	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
A2-M14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	13	15	0	0	0	0	0	0	6		
A3-M32	4	0	0	0	2	0	0	0	0	0	0	1	1	12	9	10	2	1	7	0	0	25	11	2	1	4	0	11				
B1-M146	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0		
B2-M182	1	0	5	7	0	0	0	7	3	5	24	32	0	9	0	1	0	1	2	0	0	7	12	5	5	4	5	11				
E*-SRY4064	0	0	0	0	0	0	0	0	0	0	0	0	0	16	0	0	0	0	0	0	1	0	1	1	1	3	0	0	0	0		
E1-M33	0	0	0	3	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
E2-M75	2	0	4	0	0	0	0	2	0	1	0	5	7	15	3	0	4	1	0	0	1	4	2	1	6	22	1	0	0	0		
E3*-PN2	0	0	0	0	1	0	0	0	0	0	0	0	0	0	13	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
E3a*-M2	3	0	5	18	0	17	19	18	19	7	7	4	9	3	0	0	8	14	3	9	11	21	7	10	10	27	25	7				
E3a7-M191	1	0	0	16	0	12	27	14	6	17	20	18	14	0	0	0	12	16	1	8	7	15	5	6	6	16	18	0	0	0		
E3b*-M35	0	0	0	0	0	0	0	0	0	0	0	5	6	16	7	0	8	9	0	1	11	2	2	0	4	0	4	0	0	0		
E3b1-M78	0	1	1	0	0	0	0	0	0	0	0	0	20	30	17	0	1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	
E3b2-M81	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
E3b3-M123	0	0	0	0	0	0	0	0	0	0	0	0	2	4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
J-12f2.1	0	0	0	0	0	0	0	0	0	0	0	0	0	6	22	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
R-M207	17	18	37	10	2	0	0	1	1	0	0	0	1	0	0	0	0	0	4	1	0	2	0	0	0	0	0	0	0	0	0	0
Other	0	0	0	0	3	0	0	0	0	0	0	0	4	5	2	0	0	0	2	0	0	0	1	0	0	0	0	0	0	0	0	0
H	0.6138	0.1053	0.4781	0.7610	0.6985	0.5025	0.5363	0.6911	0.5345	0.6780	0.6180	0.6243	0.7492	0.8584	0.8121	0.7907	0.6892	0.7224	0.7938	0.7428	0.6710	0.8300	0.8427	0.8095	0.7906	0.7712	0.6063	0.7949				
sd	0.0956	0.0920	0.0791	0.0263	0.1024	0.0397	0.0376	0.0402	0.0906	0.0658	0.0331	0.0462	0.0384	0.0144	0.0255	0.0267	0.0576	0.0373	0.0462	0.0515	0.0768	0.0147	0.0222	0.0485	0.0387	0.0243	0.0411	0.0262				
N	28	19	52	54	17	29	48	42	29	33	51	59	36	88	87	66	26	42	26	24	22	96	58	28	29	80	49	39				

List of population datasets in Table S10 (units with N<20 not included). “1” and “2” denote distinct sets of the same population; “\*” – only marker M35 was tested, limiting further resolution.

Table S12 – Coalescence time estimates for the Guinea-Bissau Y-chromosome haplogroup variation

Haplogroup	N	TMRCA (ky)	sd (ky)
A1-M91	7	9.8	2.9
E1*-M33	12	18.7	3.6
E3a*-M2	125	20.5	4.7
E3b*-M35	6	16.9	5.9
E3b1-M78	11	11.5	3.1

Time to the MRCA is calculated as in Zhivotovsky *et al.* <sup>(2004)</sup>, and standard error as in Thomas *et al.* <sup>(1998)</sup>.

Table S13 – Molecular diversity index ( $R_{ST}$ ) and TMRCA for haplogroup E3a\*-M2, by ethnic group

Ethnic group	N	$R_{ST}$	sd	TMRCA (ky)	sd (ky)
Felupe-Djola	19	0.4429	0.2608	21.0	5.7
Bijagós	14	0.3861	0.2426	14.2	3.4
Balanta	16	0.5166	0.2895	29.0	6.9
Papel	22	0.4386	0.2532	16.1	4.7
Fulbe	21	0.3581	0.2114	13.8	4.0
Mandenka	23	0.5208	0.2979	23.5	4.4
Nalú	9	0.4229	0.2434	14.9	5.4

Indexes calculated for 10 Y-STRs, except DYS385.  $R_{ST}$  is according to Reynolds *et al.* <sup>(1983)</sup> and TMRCA as in Zhivotovsky *et al.* <sup>(2004)</sup> and standard error as in Thomas *et al.* <sup>(1998)</sup>.

Table S14a – Exact matches of Y-chromosome haplotypes between Guinea-Bissau and other populations (for 10 Y-STR loci, except DYS437)

Haplogroup	Y-STR marker										ID*	Guinea-Bissau										Populations						
	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS438	DYS439	DYS385		Felupe-Djola	Bijagós	Balanta	Fulbe	Mandenka	Papel	Nalú	Mozambique <sup>1</sup>	Mozambique <sup>a</sup>	Angola <sup>a</sup>	Xhosa <sup>a</sup>	Spain <sup>2</sup>	Eurasia <sup>a</sup>	Afro UK <sup>a</sup>	Other <sup>a</sup>		
E3a	15	12	29	22	10	11	13	10	12	14,15	H32	1													1			
E3a	15	12	31	21	10	11	13	11	11	15,18	H41			1											1			
E3a	15	13	30	21	10	11	13	11	11	16,17	H46			1				1	1	1				1				
E3a	15	13	30	21	10	11	13	11	12	16,16	H48				1			1										
E3a	15	13	30	21	10	11	13	10	11	14,15	H49	1						1										
E3a	15	13	30	21	10	11	14	11	12	17,18	H54		1												1			
E3a	15	13	30	22	10	11	14	10	11	13,15	H58					1								1				
E3a	15	13	31	21	10	11	13	11	12	17,18	H60				1					1					1			
E3a	15	13	31	21	10	11	13	11	13	16,18	H61			1											1			
E3a	15	13	31	21	10	11	14	11	12	17,18	H67							1	1									
E3a	15	13	31	21	11	11	13	11	11	16,18	H69									1								
E3a	15	14	31	21	10	11	14	11	12	16,17	H76														1			
E3a	15	14	32	21	10	11	13	11	11	16,17	H81		1												1			
E3a	16	13	30	21	11	11	14	11	12	17,18	H95														1			
E3a	17	13	30	21	10	11	13	11	12	17,18	H125	1													3			
E3a	17	13	30	21	10	11	14	11	11	17,17	H127								1	1								
E3a	17	13	30	21	10	11	14	11	11	16,17	H128						1								1			
E3a	17	14	31	21	10	11	15	11	12	18,20	H140		1												1			
E3b1	13	13	30	24	10	11	13	10	12	16,17	H155											1		6				
E3b1	13	13	30	24	10	11	13	10	12	17,17	H156														1			
R1b	14	13	29	24	10	13	13	12	12	11,14	H165		1												71			

\* - Haplotype code as in supplementary data; 1- Alves *et al.* 2003; 2 – Zarrabeita *et al.* 2003; a - populational samples in YHRD database.

Table S14b – Haplotype exact matches between Guinea-Bissau Y chromosomes and other populations (for 8 Y-STR loci, except DYS437, DYS438 and DYS439)

Haplogroup	Y-STR marker								ID*	Guinea-Bissau							Populations			YHRD nr. pop <sup>a</sup>	
	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385		Felupe-Djola	Bijagos	Balanta	Fulbe	Mandenka	Papel	Nalú	Mozambique <sup>1</sup>	Spain <sup>2</sup>	Guinea Equatorial <sup>3</sup>		Arab <sup>4</sup>
E1*	15	12	29	22	11	11	13	14,14	H12						1						3
E3a	13	13	31	23	10	11	13	16,17	H27											1	3
E3a	15	12	29	22	10	11	13	14,15	H32	1											7
E3a	15	12	29	22	10	11	13	13,17	H33		1										7
E3a	15	12	31	21	10	11	13	15,18	H41			1									2
E3a	15	13	29	21	10	11	14	16,17	H42						1						1
E3a	15	13	29	21	10	11	14	16,18	H43						1						5
E3a	15	13	30	21	10	11	13	16,17	H46			1									12
E3a	15	13	30	21	10	11	13	15,16	H47				1			1					7
E3a	15	13	30	21	10	11	13	16,16	H48				1						2		7
E3a	15	13	30	21	10	11	13	14,15	H49	1											3
E3a	15	13	30	21	10	11	14	14,14	H50						1						1
E3a	15	13	30	21	10	11	14	16,17	H51,H52				1	1							7
E3a	15	13	30	21	10	11	14	17,17	H53				1								5
E3a	15	13	30	21	10	11	14	17,18	H54		1										6
E3a	15	13	30	21	11	11	14	16,17	H57						1						1
E3a	15	13	30	22	10	11	14	13,15	H58				1			1			1		4
E3a	15	13	30	22	10	11	14	15,16	H59				1								3
E3a	15	13	31	21	10	11	13	17,18	H60				1								4
E3a	15	13	31	21	10	11	13	16,18	H61												5
E3a	15	13	31	21	10	11	14	15,16	H62,H65				1	1	1						1
E3a	15	13	31	21	10	11	14	16,16	H63,H66				1	1					1		5
E3a	15	13	31	21	10	11	14	16,17	H64				1			1		1	1		8
E3a	15	13	31	21	10	11	14	17,18	H67					1							4
E3a	15	13	31	21	11	11	13	16,18	H69				1			2					6
E3a	15	13	32	21	10	11	14	15,16	H73				1								1
E3a	15	14	31	21	10	11	14	16,17	H76					1							2
E3a	15	14	31	21	11	11	13	16,18	H77					1							4
E3a	15	14	32	21	10	11	13	16,17	H81,H82												4
E3a	15	14	32	21	10	11	14	16,16	H83		1										1
E3a	16	11	28	21	10	11	13	16,17	H84					1					1		1
E3a	16	12	30	21	10	11	14	15,16	H87					1							2
E3a	16	13	30	21	10	11	14	16,16	H93					1							3

Table S14b (continued)

Haplogroup	Y-STR marker								ID*	Guinea-Bissau							Populations			
	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS385		Felupe-Djola	Bijagós	Balanta	Fulbe	Mandenka	Papel	Nalú	Mozambique <sup>1</sup>	Spain <sup>2</sup>	Guinea Equatorial <sup>3</sup>	Arab <sup>4</sup>
E3a	16	13	30	21	11	11	14	17,17	H94	1										1
E3a	16	13	30	21	11	11	14	17,18	H95			1								3
E3a	16	13	31	21	10	11	15	16,17	H101	1							1			2
E3a	16	14	31	21	10	11	14	17,18	H109	1										3
E3a	16	14	31	22	10	11	13	16,16	H112				1							1
E3a	16	14	31	22	10	11	13	16,17	H113,H114		1			1						1
E3a	16	14	31	22	10	11	14	16,18	H117		1									1
E3a	17	13	30	21	10	11	13	17,18	H125,H126	2										6
E3a	17	13	30	21	10	11	14	17,17	H127					2			1			18
E3a	17	13	30	21	10	11	14	16,17	H128					1						5
E3a	17	13	30	21	11	11	13	17,17	H129					1						1
E3a	17	13	31	21	10	11	14	16,16	H132		1									1
E3a	17	14	31	21	10	11	15	18,2	H140		1									1
E3a	17	14	31	22	10	11	13	15,17	H143			1								2
E3b*	13	13	30	23	10	13	13	16,16	H153				1				1			1
E3b1	13	13	30	24	10	11	13	16,17	H155			2								40
E3b1	13	13	30	24	10	11	13	17,17	H156			1			8					15
E3b1	13	13	30	24	10	11	14	16,16	H157			1								1
R1b	14	13	29	24	10	13	13	11,14	H165		1									71
R1b	15	14	31	25	10	13	13	13,15	H166			1								7

Note: This table summarizes the 8-STR matches in addition to the described in Table S14a.

“\*” – Haplotype code as in Table S9; 1- Alves *et al.* 2003; 2 – Zarrabeita *et al.* 2003; 3 - Arroyo-Pardo *et al.* 2005; 4 - Quintana-Murci *et al.* 2004; a - populational samples in YHRD database (nr pop – number of matching populations, details on request from the author).







Table S15 (continued)

Population units																				
BJG	BLE	PBO	FUL	MNK	NAJ	Mo	Ri	FBF	Ewe	Ga	Fan	Her	Am	Ku	cKhoisan	ST	Zu	Xh	Sh	Khoi
0.00000																				
-0.01714	0.00000																			
0.03024	-0.01171	0.00000																		
-0.02608	-0.01524	0.02078	0.00000																	
0.01077	-0.00491	0.02850	0.00699	0.00000																
-0.03417	-0.04187	-0.01373	-0.02791	-0.01654	0.00000															
0.05319*	0.04427	0.07685**	0.04515*	0.05444*	0.03888	0.00000														
0.08499*	0.06649*	0.08589**	0.07549**	0.11835**	0.06359	0.05353*	0.00000													
0.02361	0.00604	0.04011	0.02080	-0.01837	-0.00484	0.09347*	0.14441**	0.00000												
0.05372	0.04397	0.08052**	0.04538*	0.04590	0.04281	-0.02059	0.08273*	0.09100*	0.00000											
0.11538*	0.09910*	0.11894**	0.10189**	0.12574**	0.09664*	-0.00640	0.08948*	0.18117*	-0.00449	0.00000										
0.18125**	0.16616***	0.18224***	0.17637***	0.22256***	0.16390**	0.05593*	0.11493**	0.26770***	0.07514*	0.00383	0.00000									
0.17694**	0.17031**	0.19039***	0.18273***	0.24414***	0.16504**	0.08233*	0.11275**	0.28264***	0.10614*	0.03990	-0.01381	0.00000								
0.11934**	0.10784**	0.12945**	0.11192**	0.15891**	0.10382*	0.00714	0.05611*	0.21152**	0.02737	-0.01902	-0.02521	-0.01164	0.00000							
0.21160***	0.19069***	0.21363***	0.21370***	0.25640***	0.19636***	0.16774***	0.14399***	0.27950***	0.19070***	0.15152***	0.10437***	0.08765***	0.09127***	0.00000						
0.27057***	0.25332***	0.27742***	0.28375***	0.33751***	0.25372***	0.24544***	0.19904***	0.34914***	0.26786***	0.22656***	0.16860***	0.13365***	0.15619***	0.02215*	0.00000					
0.15046**	0.13296**	0.15648***	0.15155***	0.20781***	0.13019**	0.08139**	0.08036**	0.23977***	0.10777**	0.06226*	0.02858	0.01930	0.00769	0.03173*	0.06432**	0.00000				
0.17267***	0.15676***	0.17945***	0.17413***	0.23607***	0.15235**	0.10028**	0.04354	0.26242***	0.13423**	0.08836*	0.05428*	0.04335	0.02543	0.05875**	0.08173***	-0.01193	0.00000			
0.17058***	0.15661***	0.17578***	0.16716***	0.21200***	0.15530***	0.10331***	0.03425*	0.23733***	0.13559***	0.09857**	0.06835**	0.06064*	0.03644	0.06991***	0.11446***	0.02431	-0.00910	0.00000		
0.15171**	0.13948**	0.16030***	0.14474***	0.17744***	0.13677**	0.03184	0.09846**	0.21876***	0.04526	-0.00585	-0.00634	0.01985	-0.01745	0.12048***	0.17890***	0.01988	0.04137	0.07105***	0.00000	
0.27713***	0.25406***	0.28095***	0.28845***	0.35085***	0.25709***	0.26562***	0.21656***	0.36356***	0.29049***	0.25721***	0.20740***	0.17413***	0.18588***	0.02544*	0.00504	0.06528**	0.09883***	0.13658***	0.20659***	0.00000

Note: Population codes as in Table S10. Significance levels: \* - P<0.05, \*\* - P<0.01; \*\*\* - P<0.001.

Table S16a– Analysis of Molecular Variance (AMOVA) of Y-chromosome haplogroups in African populations (1023 permutations)

Criteria	Ethnic clusters	Among groups				Among populations within groups				Within populations				
		%	Va	F <sub>CT</sub>	P	%	Vb	F <sub>SC</sub>	P	%	Vc	F <sub>ST</sub>	P	
Geography	African continent	<ul style="list-style-type: none"> <li>▪ Northwest</li> <li>▪ Northeast</li> <li>▪ East</li> <li>▪ Central</li> <li>▪ West</li> <li>▪ South</li> </ul>	15.34	0.06941	0.15336	0.00000+ -0.00000	12.5 9	0.05700	0.14874	0.00000+ -0.00000	72.07	0.32620	0.27928	0.00000+ -0.00000
	Sub-Sahara	<ul style="list-style-type: none"> <li>▪ West</li> <li>▪ Central</li> <li>▪ East</li> <li>▪ South</li> </ul>	9.11	0.03864	0.09109	0.00000+ -0.00000	14.9 6	0.06345	0.16458	0.00000+ -0.00000	75.93	0.32205	0.24068	0.00000+ -0.00000
	West Africa	<ul style="list-style-type: none"> <li>▪ Cape Verde</li> <li>▪ Senegal/Gambia</li> <li>▪ Mali</li> <li>▪ Guinea-Bissau</li> <li>▪ Ghana</li> <li>▪ Burkina-Faso</li> </ul>	16.62	0.05712	0.16620	0.00000+ -0.00000	2.44	0.00839	0.02929	0.00000+ -0.00000	80.94	0.27816	0.19062	0.00000+ -0.00000
	West Africa	<ul style="list-style-type: none"> <li>▪ Senegal/Gambia</li> <li>▪ Mali</li> <li>▪ Guinea-Bissau</li> <li>▪ Ghana</li> <li>▪ Burkina-Faso</li> </ul>	10.31	0.02862	0.10314	0.00000+ -0.00000	3.38	0.00936	0.03763	0.00098+ -0.00098	86.31	0.23948	0.13690	0.00000+ -0.00000
Linguistics		<ul style="list-style-type: none"> <li>▪ Afro-Asiatic</li> <li>▪ Niger-Congo</li> <li>▪ Khoisan</li> </ul>	17.65	0.08322	0.17647	0.00000+ -0.00000	14.8 9	0.07024	0.18087	0.00000+ -0.00000	67.46	0.31812	0.32542	0.00000+ -0.00000
	African continent	<ul style="list-style-type: none"> <li>▪ AA Semitic</li> <li>▪ AA Berber</li> <li>▪ AA Chadic</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic-Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mande</li> <li>▪ Khoisan</li> </ul>	21.92	0.09557	0.21923	0.00000+ -0.00000	8.45	0.03685	0.10825	0.00000+ -0.00000	69.63	0.30353	0.30375	0.00000+ -0.00000

Table S16a (continued)

Criteria	Ethnic clusters	Among groups				Among populations within groups				Within populations			
		%	Va	F <sub>CT</sub>	P	%	Vb	F <sub>SC</sub>	P	%	Vc	F <sub>ST</sub>	P
	<ul style="list-style-type: none"> <li>▪ AA Semitic</li> <li>▪ AA Chadic</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé (excluding Khoisan)</li> </ul>	23.30	0.10101	0.23299	0.00000+ -0.00000	8.97	0.03891	0.11700	0.00000+ -0.00000	67.73	0.29361	0.32274	0.00000+ -0.00000
	<ul style="list-style-type: none"> <li>▪ AA Semitic</li> <li>▪ AA Chadic</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé</li> <li>▪ Khoisan</li> </ul>	16.15	0.06491	0.16150	0.00000+ -0.00000	9.16	0.03683	0.10928	0.00000+ -0.00000	74.69	0.30017	0.25313	0.00000+ -0.00000
Sub-Saharan	<ul style="list-style-type: none"> <li>▪ AA Semitic</li> <li>▪ AA Chadic</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé (excluding Khoisan)</li> </ul>	16.79	0.06583	0.16787	0.00000+ -0.00000	10.0 7	0.03951	0.12107	0.00000+ -0.00000	73.14	0.28683	0.26862	0.00000+ -0.00000
	<ul style="list-style-type: none"> <li>▪ AA Semitic</li> <li>▪ AA Chadic</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mandé (excluding Khoisan and Pygmies)</li> </ul>	16.45	0.06236	0.16446	0.00000+ -0.00000	8.67	0.03288	0.10379	0.00000+ -0.00000	74.88	0.28392	0.25118	0.00000+ -0.00000
Sub-Saharan Niger-Congo	<ul style="list-style-type: none"> <li>▪ NC Adamawa</li> <li>▪ NC Atlantic-Bak</li> <li>▪ NC Mandé</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Kwa</li> <li>▪ NC Bantu</li> <li>▪ NC Atlantic</li> </ul>	7.30	0.02494	0.07304	0.00000+ -0.00000	7.04	0.02404	0.07593	0.00000+ -0.00000	85.66	0.29252	0.14343	0.00000+ -0.00000

Table S16a (continued)

Criteria	Ethnic clusters	Among groups				Among populations within groups				Within populations			
		%	V <sub>a</sub>	F <sub>CT</sub>	P	%	V <sub>b</sub>	F <sub>SC</sub>	P	%	V <sub>c</sub>	F <sub>ST</sub>	P
	<ul style="list-style-type: none"> <li>▪ NC Atlantic-Bak</li> <li>▪ NC Mande</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Kwa</li> <li>▪ NC Atlantic</li> </ul>	2.51	0.00681	0.02510	0.10264+ -0.01130	7.25	0.01964	0.07432	0.00000+ -0.00000	90.24	0.24469	0.09755	0.00000+ -0.00000
	<ul style="list-style-type: none"> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Atlantic</li> <li>▪ NC Mande</li> </ul>	0.02	0.00006	0.00021	0.38416+ -0.01415	9.39	0.02536	0.09390	0.00000+ -0.00000	90.59	0.24469	0.09409	0.00000+ -0.00000
West Africa	<ul style="list-style-type: none"> <li>▪ NC Atlantic-Bak</li> <li>▪ NC Mande</li> <li>▪ NC Atlantic Fulani</li> <li>▪ NC Kwa</li> <li>▪ NC Atlantic-Dogon</li> <li>▪ NC Atlantic (other)</li> </ul>	7.05	0.01921	0.07052	0.00098+ -0.00098	3.10	0.00844	0.03334	0.00196+ -0.00136	89.85	0.24469	0.10151	0.00000+ -0.00000

Geographic and linguistic assignment according to information in Table S10.

Table S16b – Analysis of Molecular Variance (AMOVA) of Y chromosome haplogroups in Guinea-Bissau ethnic groups (1023 permutations)

Criteria	Ethnic clusters	Among populations				Among populations within groups				Within populations			
		%	Va	F <sub>CT</sub>	P	%	Vb	F <sub>SC</sub>	P	%	Vc	F <sub>ST</sub>	P
Geography	<ul style="list-style-type: none"> <li>▪ Fula, Mandinga</li> <li>▪ Felupe-Djola, Papel, Balanta</li> <li>▪ Bijagós</li> <li>▪ Nálu</li> </ul>	1.64	0.00394	0.01640	0.40078+- 0.01505	1.15	0.00277	0.01172	0.14076+- 0.00895	97.21	0.23349	0.02793	0.02346+- 0.00640
	<ul style="list-style-type: none"> <li>▪ Fula, Mandinga</li> <li>▪ Felupe-Djola, Papel, Balanta, Nálu</li> <li>▪ Bijagós</li> </ul>	2.07	0.00498	0.02069	0.24536+- 0.01415	0.98	0.00236	0.01002	0.14956+- 0.01141	96.95	0.23349	0.03050	0.01466+- 0.00368
Linguistic	<ul style="list-style-type: none"> <li>▪ Bijagós</li> <li>▪ Felupe-Djola, Balanta, Papel</li> <li>▪ Nálu</li> <li>▪ Fula, Mandinga</li> </ul>	1.64	0.00394	0.01640	0.39589+- 0.01520	1.15	0.00277	0.01172	1.00000+- 0.00000	97.21	0.23349	0.02793	0.02542+- 0.00468
	<ul style="list-style-type: none"> <li>▪ Bijagós</li> <li>▪ Felupe-Djola, Balanta</li> <li>▪ Papel</li> <li>▪ Fula</li> <li>▪ Nálu, Mandinga</li> </ul>	-0.09	0.00021	-0.00087	0.65200+- 0.01632	2.45	0.00585	0.02445	0.99413+- 0.00295	97.64	0.23349	0.02360	0.02639+- 0.00461
Religion*	<ul style="list-style-type: none"> <li>▪ Bijagós</li> <li>▪ Felupe-Djola, Balanta, Papel, Nálu, Fula</li> <li>▪ Mandinga</li> </ul>	-0.07	0.00017	-0.00070	0.45064+- 0.01714	2.40	0.00573	0.02397	0.99804+- 0.00196	97.67	0.23349	0.02329	0.02444+- 0.00475
	<ul style="list-style-type: none"> <li>▪ Felupe-Djola, Papel, Nálu, Bijagós</li> <li>▪ Fula, Mandinga</li> </ul>	1.54	0.00373	0.01541	0.28543+- 0.01182	2.04	0.00495	0.02074	0.08309+- 0.00843	96.42	0.23361	0.03583	0.01760+- 0.00338

\* mostly Animists versus mostly Muslims.

Appendix 1 – Primers information and PCR conditions for mtDNA HVS-I, HVS-II and coding region sites

Haplogroup	Polimorphism		Primer forward (5'-3')		Primer reverse (5'-3')	Annealing (°C)	Size (bp)
	HVSI	15907	atacaccagtctgtaaaccgg	16547 <sup>1</sup>	ggaacgtgtgggctattaggctt	52	640
	HVSII	029 <sup>2</sup>	ggctatcaccctattaaccac	408 <sup>2</sup>	ctgtaaaagtgcataccgcca	52	379
L2c	325CT	029 <sup>2</sup>	ggctatcaccctattaaccac	408 <sup>2</sup>	ctgtaaaagtgcataccgcca	52	379
L3h	1719GA	1615 <sup>3</sup>	acacaaagcaccacactacacttagga	1894 <sup>3</sup>	cttggtctccttgcaaagt	52	279
L1b	2352TC	2245	aactgaactcctcacaccaattgga	2528	ctggtagctagagggtgatg	52	283
L0-L1	2758GA	2706 <sup>*</sup>	cccgtgaagaggcgggcata	3006	tgcctgatccaacatcgag	52	300
L0-L1	3594CT	3388 <sup>3</sup>	ctaggctatatacaactacgc	3717 <sup>3</sup>	ggctactgctcgagtg	52	329
L2d	3693GA	3388 <sup>3</sup>	ctaggctatatacaactacgc	3717 <sup>3</sup>	ggctactgctcgagtg	52	330
L2b	4158AG	4057	tcccctgaactctacacaac	4251	ggcaatgctggagattgtaatg	52	194
L3f1	4218TC	4057	tcccctgaactctacacaac	4251	ggcaatgctggagattgtaatg	52	195
L1c?	4685AG	4308 <sup>3</sup>	ggagctaaacccccctta	4739 <sup>3</sup>	ggtagtattggtatggttc	52	431
L3e4	5584AG	5424 <sup>4</sup>	taacaacgtaaaaaataaaatgaca	5660	ctagtaagggttggttaa	52	236
U5b	5656AG	5548 <sup>5</sup>	agccctcagtaagtgcata	5677	ctagtaagggttggttaa	54	129
L1c	7055AG	6890 <sup>3</sup>	aagcaatatgaaatgatctg	7131 <sup>3</sup>	cgtaggttggtcta	52	241
L3d	8618TC	8551	ttcattgccccacaaatcc	8806	ggacgggtgaaatgagtgag	56	257
L3b	10086AG	9911 <sup>6</sup>	cgaagccgcccgtgatactgg	10107 <sup>6</sup>	gtagtaaggctaggaggag	60	196
L1c	10321TC	10284	ccatgagccctacaacaact	10484	gtaaatgaggggcatttggtta	60	200
N	10398AG	10284	ccatgagccctacaacaact	10484	gtaaatgaggggcatttggtta	60	200
M	10400CT	10284	ccatgagccctacaacaact	10484	gtaaatgaggggcatttggtta	60	200
L0-L1	10810TC	10672	gccatactagtcttggccgc	10959	attaggagggggttgtag	56	287
L3f1	11440GA	11158 <sup>4</sup>	cacccgatgaggcaaccagc	11502 <sup>4</sup>	agtgtagggcgtattaccatagc	52	344
L0a-f	11641AG	11295	tactctcactgccagaa	12017 <sup>4</sup>	tgagtgagcccattgtgttg	52	722
U	12308AG	12104 <sup>3</sup>	ctcaacccgacatcattacc	12338 <sup>3</sup>	attactttattggagtgacacaaaatt	52	234
U5b1	12618GA	12541	gccacaacccaacaacc	12818	cgggcgtatcatcaactgatgag	52	277
L0-L1,L3e1	13105AG	12744	cctattccaactgttcatcg	13154	agcagaaaatagcccactaa	52	410
L2a3	13803AG	13583 <sup>3</sup>	cctccctgacaagcgctatagc	13843 <sup>3</sup>	ctagggtctgtagagcttagg	60	260
L2c	13958GC	13899	tttctccaacatactcggattc	14347 <sup>4</sup>	tgatggggtggtgtgtgg	56	448
U5b	14182TC	13899	tttctccaacatactcggattc	14347 <sup>4</sup>	tgatggggtggtgtgtgg	56	448
L3f1	14766TC	14701	caatgatatgaaaaaccatc	14799	taattaatttattaggggg	52	98
L3e2a	14869GA	14701	caatgatatgaaaaaccatc	15161 <sup>4</sup>	atattggcctcacggaggacat	52	460

1 – Torroni *et al.* 1993, 2 - Bandelt *et al.* 2001, 3 - Torroni *et al.* 1996, 4 - Hofmann *et al.* 1997, 5 - Finnila *et al.* 2000, 6 - Torroni *et al.* 1992. The remaining primers are in use by the EBC group, original reference not specified. \* Allele-specific primer.

Note: Some of the primers were originally described for determining other haplogroup-defining mutations than the above mentioned. In the present analysis only the nucleotide position of interest was considered for their use.



## Appendix 2 – Primers information and PCR condition for NRY markers

SNP Marker	Region/STS	Polimorphism	Primer forward (5'-3')	Primer reverse (5'-3')	Annealing (°C)	Size (bp)
YAP <sup>1</sup>	DYS287	Alu <sup>-</sup> , Alu <sup>+</sup>	caggggagataaagaata	aagccactattagacaacct	57	599(Alu+)/308(Alu-)
92R7 <sup>2</sup>	92R7	504GA	gacccgctgtagacctgact	gcctatctactcagtgattct	60	722
SRY10831 <sup>3</sup>	SRY	135AG	ccacaacctctttcatc	aataaaaatcccgtaaaata	57	536
PN2 <sup>3</sup>	DYS287	153CT	gatgcaaatgagaagaact	ctaaaaactggaggagaaaa	54	536
M2 <sup>4</sup>	DYS271	168AG	aggcactggcagaatgaag	aatggaaaaacagctcccc	60	209
M9 <sup>5</sup>	G10.35a	68CG	gcagcatataaaacttcagg	gctgagcaaagttaggtttt	57	340
M10 <sup>5</sup>	G10.10	156TC	gcattgctataagttacctgc	taataaaaattgggtcacc	52	343
M13 <sup>5</sup>	G10.06	157GC	tcctaacctgggtcttctc	tgagccatgattttatccaac	52	233
M14 <sup>5</sup>	G10.07	180TC	agacggtagatcagttctctg	tagataaaagcacattgacacc	58	287
M31 <sup>6</sup>	G10.66b	71GC	gaaccagacaatacgaatagaag	tttagcggcttatctcattacc	50	486
M32 <sup>6</sup>	G10.68a	166TC	ttgaaaaatacagtggaac	caagtgtttaaggatacaga	48	370
M33 <sup>6</sup>	G10.68b	180AC	ttgaaaaatacagtggaac	caagtgtttaaggatacaga	51	370
M35 <sup>6</sup>	G10.72a	168GC	taagcctaaagagcagtcagag	agagggagcaatgaggaca	59	513
M40 <sup>3</sup>	SRY	258GA	gcatttgtacccttctcaac	tgccaagactacgagatttc	54	612
M44 <sup>6</sup>	G10.87	263GC	ctggcacctctgataatttgag	tgtgattctatgtgttgaggac	59	389
M60 <sup>6</sup>	B9.34	242, +1bp	gcactggcgttcatcatct	atgttcattatggttcaggagg	54	388
M75 <sup>6</sup>	B9.51	296GA	gctaacaggagaaataaatacagac	tattgaacagaggcattgtga	58	355
M78 <sup>6</sup>	B9.60a	197CT	ctcaggcattatTTTTTggt	atagtgtccttcaccttcctt	50	301
M81 <sup>6</sup>	B9.58a	147CT	acttaatttatagttcaatccctca	ttcatggagatgtctgtatctgg	51	422
M89 <sup>6</sup>	B9.94	347CT	agaagcagattgatgtcccact	tccagttaggagatcccctca	57	527
M91 <sup>6</sup>	B9.87a	368, 9T-8T	gagcttgacttttaggacgg	aaacttaaggcacttctggc	59	495
M116 <sup>6</sup>	G3.25a	176AC, AT	aagtatgactatgaagtacgaagaaa	attcagttagattttacaatgagca	55	429
M123 <sup>6</sup>	G3.27b	161GA	tggtaaactctactagttgccttt	cagcgaattagattttctg	53	393
M130	RPS4YC711	41CT	tatctccttctattgacag	ccacaagggggaaaaaacac	53	205
M155 <sup>6</sup>	G10.57c	251GA	tctctaactctgtgaccac	ggaaaaactaaactctaatctct	52	327
M168 <sup>6</sup>	DFFRY Ex01B site a	371CT	agtttgaggtagaataactgttgct	aatctcataggtctctgactgttc	62	473
M173 <sup>6</sup>	DBY Ex08	191AC	aagaaatggtgaactgaagttgat	aggtgtatctggcatccgta	53	417
M174 <sup>6</sup>	DffryEx38	219TC	acatctcagatcgtgttggt	aaaaagccatgcaattacctg	54	348
M191 <sup>6</sup>	DBY exon 2	342TG	ttgatttgcacgtgtggt	gccaggataattttgtatttctc	59	429

1 - Hammer and Horai 1995; 2 - Mathias *et al.* 1994; 3 - Hammer *et al.* 1998; 4 - Seielstad *et al.* 1994; 5 - Underhill *et al.* 1997; 6 - Underhill *et al.* 2001

### Appendix 3 – Information on Y-STR loci included on PowerPlex® Y System

Locus	Label	Location	GenBank® Accession	Repeat motif	Size range (bp)	Repeat number alleles
DYS391 <sup>1</sup>	FL	Yq	G09613	TCTA	90–118	6, 8–13
DYS389I/II <sup>1</sup>	FL	Yq	AF140635	[TCTG][TCTA] Complex	148–168, 256–296	10–15, 24–34
DYS439 <sup>2</sup>	FL	Yq	AC002992	GATA	203–231	8–15 <sup>a</sup>
DYS393 <sup>1</sup>	TMR	Yp	G09601	AGAT	104–136	8–16
DYS390 <sup>1</sup>	TMR	Yq	AC011289	[TCTG][TCTA] Complex	191–227	18–27
DYS385a/b <sup>1</sup>	TMR	Yq	Z93950	GAAA	243–315	7–25
DYS438 <sup>2</sup>	JOE	Yq	AC002531	TTTTTC	101–121	8–12
DYS437 <sup>2</sup>	JOE	Yq	AC002992	[TCTA][TCTG] Complex	183–199	13–17
DYS19 <sup>1</sup>	JOE	Yp	X77751	TAGA Complex	232–268	10–19
DYS392 <sup>1</sup>	JOE	Yq	G09867	TAT	294–327	7–18

1-Kayser *et al.* 1997, 2-Ayub *et al.* 2000; TMR = carboxy-tetramethylrhodamine, FL = fluorescein, JOE = 6-carboxy-4',5'-dichloro-2',7'-dimethoxyfluorescein; a- follows the original nomenclature as in Ayub *et al.* 2000.

## Appendix 4 – Typing methodology for mtDNA polymorphisms (and RFLP specifications)

Haplogroup	Polymorphism	Sequencing	RFLP/Site*	Uncutted state (-)	Cutted state (+)
HVSI		✓			
HVSII		✓			
L2c	325CT		✓ -323 <i>Hae</i> III	379	84+295
L3h	1719GA		✓ -1715 <i>Dde</i> I	23+30+227	23+30+48+179
L1b	2352TC		✓ +2348 <i>Mbo</i> I	283	103+180
L0-L1	2758GA		✓ -2759 <i>Rsa</i> I	145+156	54+91+156
L0-L1	3594CT		✓ +3594 <i>Hpa</i> I	330	123+207
L2d	3693GA		✓ -3693 <i>Mbo</i> I	59+271	24+34+271
L2b	4158AG		✓ +4158 <i>Alu</i> I	194	93+101
L3f1	4218TC	✓			
L1c?	4685AG		✓ -4686 <i>Alu</i> I	4+9+47+48+107+101+115	4+9+47+48+53+54+101+115
L3e4	5584AG		✓ -5585 <i>Alu</i> I	161+75	161+60+15
U5b	5656AG		✓ +5656 <i>Nhe</i> I	129	108+21
L1c	7055AG		✓ -7056 <i>Alu</i> I	241	166+75
L3d	8618TC		✓ -8615 <i>Mbo</i> I	64+113+80	40+24+113+80
L3b	10086AG		✓ +10084 <i>Taq</i> I	196	174+22
L1c	10321TC		✓ +10321 <i>Alu</i> I	200	37+163
N	10398AG		✓ -10394 <i>Dde</i> I	73+127	72+38+90
M	10400CT		✓ +10398 <i>Alu</i> I	200	115+85
L0	10810TC		✓ +10806 <i>Hinf</i> I	158+129	135+23+129
L3f1	11440GA		✓ +11438 <i>Mbo</i> I	344	280+64
L0a-f	11641AG		✓ +11641 <i>Hae</i> III	722	348+374
U	12308AG		✓ +12308 <i>Hinf</i> I	67+168	67+138+30
U5b1	12618GA	✓			
L0-L1,L3e1	13105AG	✓			
L2a3	13803AG		✓ +13804 <i>Hae</i> III	260	222+38
L2c	13958GC		✓ -13958 <i>Hae</i> III	448	60+388
U5b	14182TC	✓			
L3f1	14766CT		✓ +14766 <i>Mse</i> I	82+4+12	65+17+4+12
L3e2a	14869GA		✓ -14868 <i>Mbo</i> I	358+102	190+168+102

\* The restriction state is defined in relation to CRS

## Appendix 5 – Typing methodology for NRY markers (and RFLP specifications)

SNP Marker	Sequencing	RFLP	Enzyme	Ancestral state	Derived state
YAP				150	455
92R7		✓	<i>HindIII</i>	56+137	193
SRY10831	✓				
PN2	✓				
M2		✓	<i>NlaIII</i>	102+65+42	65+144
M9		✓	<i>HinfI</i>	182+93+66	284+93
M10	✓				
M13		✓	<i>MboI</i>	156+ 77	233
M14	✓				
M31		✓	<i>BtsI</i>	393+73+20	393+93
M32	✓				
M33		✓	<i>MseI</i>	10+59+118+183	10+118+242
M35		✓	<i>BsrI</i>	169+344	513
M40		✓	<i>BsrBI</i>	362+ 147	509
M44	✓				
M60	✓				
M75	✓				
M78	✓				
M81		✓	<i>HpyCH4IV</i>	276+146	422
M89	✓				
M91	✓				
M116	✓				
M123	✓				
M130		✓	<i>BseL1</i>	154+46+4+1	201+4
M155	✓				
M168		✓	<i>Hinf I</i>	234+106+ 81+52	234+187+52
M173	✓				
M174		✓	<i>BseNI</i>	348	218+130
M191	✓				

## MtDNA Profile of West Africa Guineans: Towards a Better Understanding of the Senegambia Region

Alexandra Rosa<sup>1,2</sup>, António Brehm<sup>2,\*</sup>, Toomas Kivisild<sup>1</sup>, Ene Metspalu<sup>1</sup> and Richard Villems<sup>1</sup>

<sup>1</sup>Department of Evolutionary Biology, Estonian Biocenter, Tartu University, Riia 23, 51010 Tartu, Estonia

<sup>2</sup>Human Genetics Laboratory, Center of Macaronesian Studies, University of Madeira, Campus of Penteada, 9000-390 Funchal, Portugal

### Summary

The matrilineal genetic composition of 372 samples from the Republic of Guiné-Bissau (West African coast) was studied using RFLPs and partial sequencing of the mtDNA control and coding region. The majority of the mtDNA lineages of Guineans (94%) belong to West African specific sub-clusters of L0-L3 haplogroups. A new L3 sub-cluster (L3h) that is found in both eastern and western Africa is present at moderately low frequencies in Guinean populations. A non-random distribution of haplogroups U5 in the Fula group, the U6 among the "Brame" linguistic family and M1 in the Balanta-Djola group, suggests a correlation between the genetic and linguistic affiliation of Guinean populations. The presence of M1 in Balanta populations supports the earlier suggestion of their Sudanese origin. Haplogroups U5 and U6, on the other hand, were found to be restricted to populations that are thought to represent the descendants of a southern expansion of Berbers. Particular haplotypes, found almost exclusively in East-African populations, were found in some ethnic groups with an oral tradition claiming Sudanese origin.

### Introduction

Unveiling the history of human settlement in the West Coast of Africa is a complex task. It is the result of a continuous complex network of migrations, invasions and admixture of peoples from different origins. Fossil evidence suggests a modern human presence in NW Africa around 40000 years before present (YBP) (Alimen, 1987). A pre-Neolithic Capsian culture evolved later locally or through a diffusion from the Near East (Camps-Faber, 1989). Around 9000 YBP, when the Sahara went through a period of maximum humidity (Aumassip *et al.* 1988), several Neolithic cultures flourished in the area, bringing together people of sub-Saharan and North African origin (Dutour *et al.* 1988). The domestication and spread of several African-specific plants probably started in western Sahel after 4000 YBP. The first phase of largely east and southward oriented Bantu migrations, originating from the cen-

tral Gulf of Guinea region, is a likely outcome of these cultural developments (Fage, 1995).

The Ghana Empire, between Niger and Senegal, is the oldest known occidental African Kingdom (Fage, 1995) which was followed in the 14<sup>th</sup>–16<sup>th</sup> centuries by other empires (Mali, Songhai). The admixture of Berbers with native populations of this area dates back at least to the 9<sup>th</sup> century A.D., after the arrival of pastoral Peuls or Fulbe (here designated as Fula). In 1086 Ommiades conquered North-Western Africa and pushed the populations from South Morocco and Mauritania to the Senegal region (Moreira, 1964). When the Europeans arrived in Senegambia in the 15<sup>th</sup> century they met most of the presently known ethnic groups settled in the region (Teixeira da Mota, 1954). The Fula arrived again two centuries later, coming from the Futa Toro and Sahel regions, dominating the whole area. The Mandinga (Mandenka) were the last to arrive in this region (Carreira & Quintino, 1964).

Present day Guinean ethnic groups are disseminated all over the territory. The Balanta are the biggest group, and in the first quarter of the 20<sup>th</sup> century spread over

territories occupied earlier by other ethnic groups. The origin of the Balantas is uncertain. Some see language affinities with the Sudanese from whom they could have separated 2000 years ago with the first spread of kushites migrations (Quintino, 1964). According to Stuhlmann (1910), the group derives from a Bantu branch, which separated in the Pleistocene near the Nile, following camite invasions. The Bijagós inhabit the Archipelago of the same name and some scholars see strong cultural resemblances to Egyptians (Quintino, 1964), but others relate them to the Senegalese Djola. The latter are a rather heterogeneous group, and include the Beafada which have an oral tradition of coming from Mali (Lopes, 1999). A mass arrival of Fula took place in the beginning of the 19<sup>th</sup> century. The origin of this ethnic group is unknown, but tradition relates them to Hiksos and Nubians. They show the typical phonetic "glottal catch" which characterizes the whole group.

Here we analyze the mtDNA lineages present in the major ethnic groups of Senegambia, covering a broad number of recognized groups underrepresented in previous studies (Graven *et al.* 1995; Watson *et al.* 1997; Rando *et al.* 1998), and compare them within the broader context of African mtDNA variability (Graven *et al.* 1995; Watson *et al.* 1997; Rando *et al.* 1998, 1999; Krings *et al.* 1999; Chen *et al.* 2000; Pereira *et al.* 2001; Brehm *et al.* 2002; Salas *et al.* 2002). Because mtDNA haplogroups show distinct geographic patterns in Africa, their frequency and diversity patterns in West Africa can be informative with respect to the origin of the different ethnic groups from Guiné-Bissau. The presence of Y-chromosomes of Eurasian affiliation among populations from Cameroon at a high frequency, as reported recently (Cruciani *et al.* 2002), raises the intriguing question of back migrations from Eurasia to Africa, here supported by the presence of particular Eurasian mtDNA lineages among Guineans.

### Material and Methods

#### Sampling

A total of 372 blood samples were collected from unrelated Guinean males whose maternal ancestors were known to belong exclusively to a specific ethnic group. The samples were collected either in military camps

with the permission of the Guiné-Bissau Chairman of the Joint Chiefs of Staff, or in the villages around the country with the help of the Ministry of Health. Every participant gave his consent in an individual interview after a detailed explanation of the project. Sample sizes and origins (along with additional information) are specified in Table 1 and 2. Due to the complex history involving the major ethnic groups in Guiné-Bissau, they do not all follow a clear present-day settlement pattern (see Figure 1).

Populations of low sample size were pooled according to their linguistic affinities. The linguistic clustering presented in Table 1 is based on anthropological or linguistic classifications following Almeida (1939), Barros (1947), Carreira (1962, 1983), Almada (1964), Carreira & Quintino (1964), Hair (1967), Quintino (1967, 1969), Diallo (1972) and Lopes (1999). Some groups were left unpooled: the Balanta, for whom a Sudanese origin has been suggested, and the Bijagós because of their particular geographical location.

#### HVS-I and HVS-II Sequencing

The leukocyte fraction of whole blood was used for DNA extraction by standard methods and the mtDNA hypervariable segment I (HVS-I) of the control region was amplified and sequenced. Sequencing products were separated on a MegaBACE 1000 automatic sequencer, following the manufacturer's specifications and aligned using Wisconsin Package GCG Version 10.0. All sequences were read between nucleotide positions (nps) 16024 and 16400. Additional information regarding polymorphic sites 185, 186, 189, 195, 236, 297 and 322 in HVS-II was obtained by directly sequencing all samples that could not be unambiguously classified on the basis of HVS-I information alone.

#### RFLP Testing

In case of ambiguity in defining mtDNA haplogroups on the basis of the HVS-I haplotype, additional data was gathered from restriction fragment length polymorphisms (RFLPs) of diagnostic sites. All restriction digests were made according to the manufacturer's instructions (Fermentas and New England BioLabs). The following polymorphic restriction sites were screened: 322HaeIII,

\*Corresponding author: António Brehm, phone +351291705383, fax +351291705393, e-mail: brehm@uma.pt

**Table 1** Population data of the Guinean samples ethnic distribution

Code	Ethnic group	Language group WA	Religion	Closest language group	Synonyms
BLE	Balanta <sup>a</sup>	Bak-Balanta-Ganja	A,M,C	Tenda	Ballante, Balant
BDA	Baiote	Bak-Diola-Bayot	A	Diola	Bayotte
BAB	Banhu	Eastern Senegal-Banyun	A,M	Tenda	Bainouk, Banyuk, Elomay
BIF	Beafada	Easter Senegal-Tenda	M	Badyara	Biafada, Bidyola, Biafar
BJG	Bijagó	Bijagó	A		Bidyogo, Bijougot
BRA	Brame				
CCJ	Cassanga	Eastern Senegal-Nun	A	Banhu-Felupe	Kasanga, I-Hadja
EJA	Djola <sup>b</sup>	Bak-Diola-HerEjamat	A,C	Diola-Wolof	Fulup, Floup, Ejamat, Ediamat
FUL	Fula	Fulani-West Central	M	Fula-Wolof	Fulbe, Futa Jallon
FUF	Futa-Fula	Fulani-West Central	M	Fula-Wolof	Fulbe, Futa Jallon
FUC	Fula-Preto	Fulani-Western	M	Fula-Wolof	Peul, Peull
FUC	Fula-Forro	Fulani-Western	M	Fula-Wolof	Peul, Peull
FUT	Fula-Toranca	Fulani	M	Fula-Wolof	Peul, Peull
JAD	Jancanca	Mandenkan	M	Mandinka	Jahanque, Jahanka, Diakanke
LAN	Landoma				
MAN	Mancanha	Bak-Manjaku-Papel	A,C	Manjaku-Papel	Mankanya, Mankanha
MNK	Mandinga	Mandekan	M	Kalenge, Jahanka	Mandingue, Mandenka
MFV	Manjaco	Bak, Manjaku-Papel	A,C,M	Mancanha, Papel	Mandyak, Manjiak
MSW	Mansonca	Sua	M		Kunante, Mansoanka
NAJ	Nalú	Mbulungish-Nalu	A,M	Susu	Nalou
SUD	Sussu	Susu-Yalunka	M,A,C	Yalunka	Susu, Sose, Soso
PBO	Papel	Bak, Manjaku-Papel	A,C	Mankanya, Mandyak	Pepel, Oium

A-Animist, M-Muslim, C-Christian; Population codes and language groups follow terminology from [www.sil.org/ethnologue/](http://www.sil.org/ethnologue/); <sup>a</sup>includes the so-called Balanta-Mané (Balanta islamized by Mandinga); <sup>b</sup>includes Felupes;

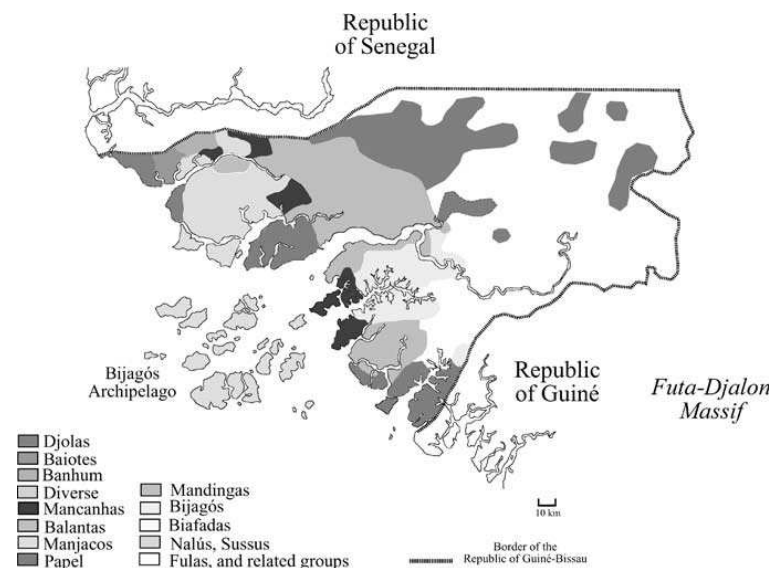
1715Ddel, 2349MboI, 2758RsaI, 3592HpaI, 3693MboI, 4157AluI, 4685AluI, 5584AluI, 5656NheI, 7055AluI, 8616MboI, 10084TaqI, 10321AluI, 10394Ddel, 10397AluI, 10806HinfI, 11439MboI, 11641HaeIII, 12308HinfI, 13803HaeIII, 13957HaeIII, 14766MseI and 14868MboI. The following coding region sites were ascertained by sequencing: 2758, 4218, 12618, 13105 and 14182. Primers and PCR conditions used in all analyses are available as Complementary Material at [www.ahg.com](http://www.ahg.com).

### Haplogroup characterization

The HVS-I sequence types were classified following the nomenclature of African and European mtDNA haplogroups (Quintana-Murci et al. 1999; Macaulay et al. 1999; Rando et al. 1999; Alves-Silva et al. 2000; Chen et al. 2000; Richards et al. 2000; Richards & Macaulay 2001; Bandelt et al. 2001; Torroni et al. 1997, 2001; Mishmar et al. 2003; Salas et al. 2002). Here, and in what follows, the nucleotide position (np) number relative to

the revised CRS (Anderson et al. 1981; Andrews et al. 1999) is used to designate haplotype-defining mutations. Character state change is specified only for transversions and insertions/deletions. Based on the previous knowledge of African complete sequences paraphyletic clade L1 is split into two monophyletic units L0, capturing previously defined L1a and L1d lineages, and L1 clade that includes L1b and L1c clades (Mishmar et al. 2003). The sub-clades of L0a (pro L1a) and L1b are defined as in Salas et al. (2002).

Haplogroup L2 is divided into L2a (characterized by 16294 and 13803), L2b (16114A, 16129, 16213 and 4158), L2c (322 and 13958), and L2d (16399 and 3693) sub-clades. Mutations 16278, 16362 and 10086 characterize haplogroup L3b; haplogroup L3d is defined by 8618 and shares with L3b a transition at np 13105. According to Bandelt et al. (2001) L3e (defined by 2352) is subdivided into L3e1 (16327), L3e2 (16320), L3e3 (16265T) and L3e4 (16264 and 5584) clades. L3e2 is further subdivided into L3e2\* (14869) and L3e2b (16172 and 16189). As in Salas et al. (2002), L3f captures all L3\*



**Figure 1** Geographic distribution of ethnic groups in Guiné-Bissau. The boundaries may not correspond entirely to the precise distribution of the groups involved since overlapping areas do exist.

lineages with a mutation at 16209. L3f1 is further defined by a T at np 16292 (and 14766). Here we further define a new sub-cluster, L3h characterized by a loss of the Ddel site at np 1715 (mutation at np 1719) and the HVS I motif 16129, 16256A and 16362. Following Finniliä et al. (2000) U5b is characterized by 5656 and 12618 over 14182. Haplogroup U6 (Rando et al. 1998) is defined by 16172 and 16219. Haplogroup M1 is characterized by 16129, 16189, 16249 and 10400 mutations (Quintana-Murci et al. 1999).

### Genetic Analysis and Population Comparisons

Median networks of HVS-I haplotypes (Bandelt et al. 1995, 2000) were drawn for each haplogroup separately, using the Network 3.1 program (Arne Röhl, [www.fluxus-engineering.com/sharenet.htm](http://www.fluxus-engineering.com/sharenet.htm)). Haplogroup frequencies, molecular diversity indexes ( $F_{ST}$ ) and genetic diversity ( $H - Nei$ , 1987) for haplotypes and haplogroups and analysis of molecular variance (AMOVA) were calculated using Arlequin

v2.0 (Schneider et al. 2000). Comparisons between populations were assessed by subjecting the (relative) frequency vectors of the haplogroups to a principal component analysis (PCA).

A local database with more than 19000 individuals taken from literature and our unpublished data from worldwide populations was employed to search for exact matches of Guiné-Bissau haplotypes, ignoring length variation in the C stretch of the HVS-I.

Coalescence times were estimated by means of the  $\rho$  statistic, assuming that a transition within 16090-16365 corresponds to 20180 years (Forster et al. 1996).

## Results and Discussion

### Haplogroup Profiles

The 372 Guinean samples clustered to 192 different haplotypes of all major West African mtDNA haplogroups (for the complete list see Complementary Material). Three predominant haplotypes (GB4, GB85

Table 2 Haplogroup relative frequencies and diversity index (H) in Guiné-Bissau ethnic groups and several other African populations. Superscripts (a-g) in Guinean ethnic groups refer to codes used in Figure 3 (PCA)

Table with 20 columns representing different populations (e.g., Senegal, Cabo Verde, Nile, Guiné-Bissau, Morocco, Algeria) and 16 rows representing haplogroups (L0a, L1b, L1c, L1\*, L2a, L2b, L2c, L2\*, L3a, L3b, L3c, L3\*, M1, U6, Eurasian, H, s-d, N). Each cell contains a numerical value representing the frequency or diversity index for that haplogroup in that population.

Pereira et al. (2001); 2Salas et al. (2002); 3unpublished; 4Watson et al. (1997); 5unpublished; 6Kriings et al. (1999); 7Chen et al. (2000); 8Vigilant et al. (1991); 9Brehm et al. (2002); 10Graven et al. (1995); 11Rando et al. (1998); 12unpublished (i - Turkana, Kikuyu and Somalia; ii - Senegalese, Wolof and Serer; iii - Songhai, Tuareg, Yoruba, Hausa, Fulbe and Kamur). The so-called Eurasian haplogroup refers to all non-Ls, M1 or U6 sequences. N, total sample number; H, Nei's (1987) diversity index; sd, H standard deviation.

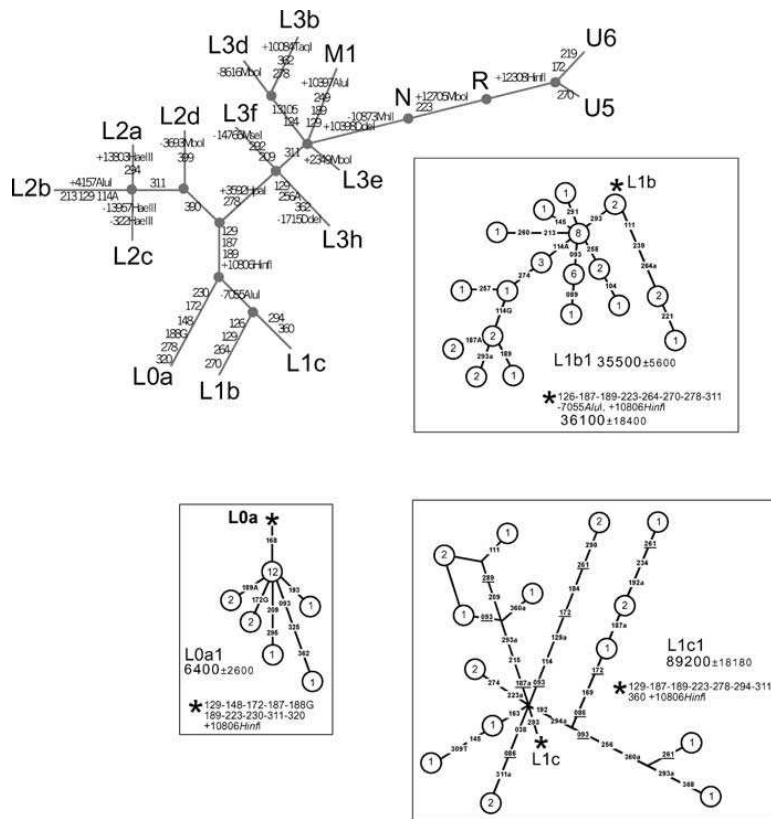
and GB117) captured 13% of the Guinean mtDNA variation, occurring at a frequency higher than 3% each. Most sequences (94%) could be classified as belonging to sub-Saharan African L0a1, L1b, L1c1, L2a, L2b, L2c, L2d1, L3b, L3d, L3e, L3f1 and L3h haplogroups and subhaplogroups. Unexpectedly for a West African population, 22 (5.9%) of the samples clustered to haplogroups M1 (1.1%), U5 (2.7%) and U6 (2.2%, Table 2; Graven et al. 1995; Watson et al. 1997; Rando et al. 1998; Salas et al. 2002). M1 and U6 are found in North and East Africa, Arabia, and the Middle East, whereas U5 has been sampled at appreciable frequencies only in Europe (Passarino et al. 1998; Quintana-Murci et al. 1999; Richards et al. 2000). The haplogroup profile for each ethnic group separately can be found in the Complementary Material.

L Lineages

Haplogroup L0 was represented in Guineans only by its daughter group L0a1 showing marginal frequencies ranging from 1% to 5% (Table 2), in contrast to its frequency in East African populations (e.g. 25% in Mozambique: Watson et al. 1997; Pereira et al. 2001; Salas et al. 2002). Interestingly, only the Balanta, a group claiming Sudanese origin, showed an increased frequency of this clade (11%). Haplogroup L0a has a Paleolithic time depth in East African populations (33,000 year old, Salas et al. 2002). The relatively young coalescent date of L0a1 in Guineans (6400±2600 years, assuming a single founder) suggests that only a small subset of L0a reached Guinea during the Holocene. The founder haplotype of L0a in Guineans, GB4 (see Table 4 in Complementary Material), has an exact match in East Africa, the Middle East and in Cape Verde and Senegal Mandenka populations, indicating that its spread is not strictly restricted to Guineans. The lack of the L0a2 clade, associated with the 9bp deletion in CoII/tRNA15s intergenic region, and widespread in Bantu speaking populations all over Africa (Soodyall et al. 1996), suggests that L0a has at least two distinct phylogeographic patterns in Central and West Africa. We cannot discard the possibility of a Bantu migration to West Africa, as the founder group could have a distinct composition from those who participated in the southwards migration(s).

Haplogroup L1b is restricted mostly to West African populations (Graven et al. 1995; Watson et al. 1997; Salas et al. 2002) and is represented by two different branches in Guineans. Its major cluster (Figure 2) L1b1 is associated with a transition at np 16293 and includes a frequent sub-clade defined by the combined presence of a transversion to A at np 16114 and a transition at np 16274 that has also been observed in Senegalese Mandenka (Graven et al. 1995) and Wolof (Rando et al. 1998). L1b1 presents a TMRCA of about 36000 years (Figure 2), predating the diversity of L0a1 in Guineans. The matches in this cluster have a West African distribution well represented in Mandenka (haplotypes GB8 and GB20) and their frequency is highest in the Fulani-western and Senegal-eastern language groups (Table 2). GB23 and GB24 are widespread in Africa and are found in nearly all West African populations considered here (Salas et al. 2002). Another West African specific clade, L1c, is present at a relatively low frequency (0-8%) yet with high haplotype diversity in the Guiné-Bissau sample.

Haplogroups L2a-L2c are frequent in Senegambia (Table 2) and reveal signatures of a recent expansion from a limited number of founder haplotypes that are shared between populations of different linguistic affiliation. In contrast, haplotypes belonging to haplogroup L2d are represented by single individuals and do not show a common founder sequence (Figure 2). Fifteen out of 42 L2a haplotypes sampled in Guinea Bissau had matches elsewhere: West Africa (Cabo Verde, Brehm et al. 2002; Wolofs & Senegalese, Rando et al. 1998; Mandenka, Graven et al. 1995) but can also be found in East, South and North Africa. The geographic distribution of L2b and L2c haplotypes is largely restricted to West Africa. Not surprisingly most of the haplotype matches are with Cabo Verdeans, Wolof and Senegalese. L2c is the haplogroup that shows a higher extent of shared lineages: Cape Verde, Senegal Mandenka, mixed Senegalese and São Tomé. The last case is likely due to a recent gene flow from the Cape Verde Islands (Brehm et al. 2002). However, several L2 haplotypes observed in Guineans appeared as unspecific to other West African populations but shared matches with East and North Africans. This was the case for the Balanta (BLE) haplotype GB44

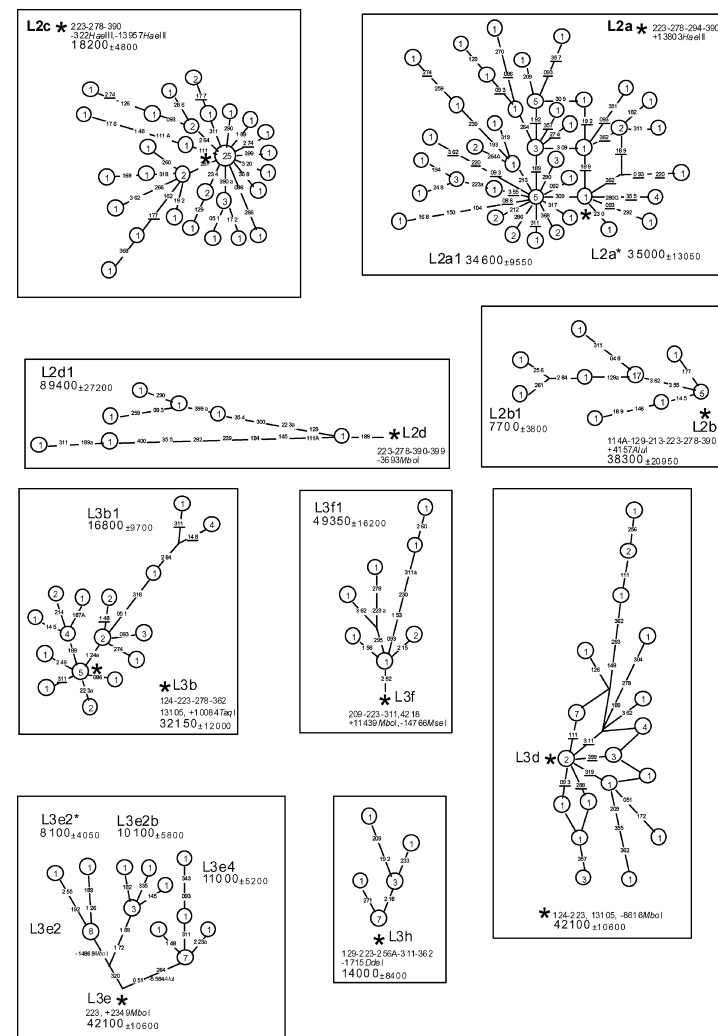


**Figure 2** MtDNA phylogeny of all Guinean haplogroups and skeletons of various L0, L1, L2 and L3 sub-haplogroups based on HVS-I sequences and coding-region RFLPs. The number of individuals assigned to the haplotypes is shown within the circles. The numbers over the lines represent the HVS-I (-16000 bp) and coding region mutations, with respective restriction sites. Transversions are represented with suffixes (length variation in the C-stretch is disregarded). Recurrent mutations are underlined and *a* refers to the mutation loss relative to root. The star indicates the putative root of the haplogroup. Coalescence estimates  $\pm$  sd (in ya) are shown for haplogroups or sub-haplogroups.

matching only with Sudanese (Watson *et al.* 1999), and GB59 matching with Moroccan sequences. Interestingly, haplotype GB83 (L2b) found in the Mansona (MSW) group had an exact match only with Ethiopians (our unpublished data). Also the Fula haplotype GB39 has not been reported in West Africa but appears in East Africa: Lake Turkana (Watson *et al.* 1997), Nubia,

Southern Sudan, Ethiopia and Saudi Arabia (our unpublished data).

Haplogroups L3b, L3d, and L3e are rare or absent in indigenous populations of North and South Africa but well represented in our sample. GB127 and GB134 are particular links of Guinean groups to North-west African Mozabites, Moroccans and Senegalese.



**Figure 2** Continued.

Particularly, GB136 from Fula-related people has been found so far in Hausa and again in Nubians and Sudanese. Apart from Mozambique (6%) the majority of L3d lineages are West African (7% in mixed Senegalese

to 12% in Niger/Nigeria) with an estimated age of 42100 ( $\pm$  10600, Salas *et al.* 2002). L3f is more frequent in Southeast Africa, ranging from 8% in Kenya/Sudan to 2% in Mozambique. The coalescence time of this



haplogroup in West Africa was calculated as 39400 ya ( $\pm 10400$ , Salas *et al.* 2002), within the error range of the estimate based on Guinean samples ( $49350 \pm 16200$  ya). Haplotype GB178 in Fula shared an exact match with sequences from a wide range of East-African populations (Somalia, Egypt) and even Saudi Arabia. Haplogroup L3h is found in Ethiopia, Cape Verde and Niger/Nigeria at marginal frequencies ( $\sim 1\%$ ) but reaches its highest known frequency in the Ejamat from Guinea (8%). Its coalescent time estimate ( $14000 \pm 8400$  ya) in Guineans is consistent with its late Pleistocene/early Holocene spread around Africa.

No significant differences between Guinean ethnic groups pooled by their linguistic affiliation were observed in haplogroup frequencies. As for their geographic neighbours (Table 2), haplogroups L1b, L1c, L2b, L2c, L2d, L3b, L3d, and L3e cover most of the mtDNA variation (64–85%). The Guiné-Bissau sample shows an overall genetic diversity of 0.901 (sd.005) that is significantly higher than among other samples from West Africa (Table 2).

### M1 and U6 Lineages

Haplogroup M1 has been characterized as an East African remnant of the major Asian haplogroup M (Quintana-Murci *et al.* 1999). It has been found mostly in Ethiopian populations (17%), its characteristic HVS-I motif being also well represented in Egyptian and Sudanese populations along the Nile Valley (7–8%, Krings *et al.* 1999). HVS-I haplotypes matching the East African M1 clade have also been identified in Northwest Africans (Plaza *et al.* 2003, unpublished data) where their frequency can reach 12.8% in Algerians and 4% among Moroccan and Algerian Arabs and Berbers. M1 is generally absent from autochthonous West African populations but was found among Balanta, Baiote, and Djola groups speaking Niger Congo Atlantic Bak languages. The Guinean M1 haplotypes matched exactly one West Saharan (Rando *et al.* 1998), 2 Mozabites (Côrte-Real *et al.* 1996), 2 Iranian and one Saudi Arabian sequence (unpublished data). This lineage derives from a particular cluster defined by a mutation at position 16185, which is also found in Ethiopia, Morocco and North African populations (Plaza *et al.* 2003, our unpublished results).

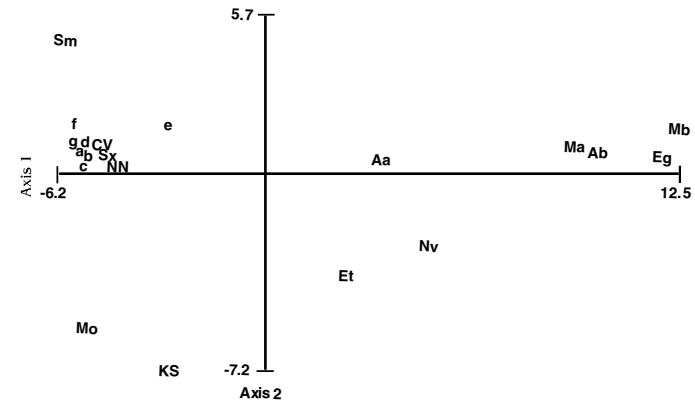
Haplogroup U6 is rather frequent in NW Africa, among Algerian Berbers, Moroccans and Mauritani-ans (Côrte-Real *et al.* 1996; Rando *et al.* 1998; Plaza *et al.* 2003), but is rare or absent in western sub-Saharan Africans. Three different U6 haplotypes were observed in Fula, Mandenka and Manjaco groups. These haplotypes match with sequences of a wide geographic range: North and West Africa (Cabo Verde, Tuareg, Mozabites, Moroccan Arabs and Berbers), East Africa (Nile Valley, Egypt and Ethiopia), the Middle East (Iran) and Mediterranean Europe (Sicily and Portugal, <http://www.ahg.com/>), suggesting that their spread might be related to the southern expansions of the Berber groups to whom the Fulani languages relate.

### European Lineages: U5

Ten individuals out of 372 samples, all related to Fulbe groups, carried mtDNA variants typical of western Eurasia, particularly Europe. Within these mtDNAs belonging to haplogroup U5 nine Fulanis share one particular HVS-I haplotype. Both haplotypes are only one mutational step away from a common node widespread in Europe. Although U5 is one of the most frequent mtDNA variants among western Eurasians (about 460 sequences in our mtDNA HVS-I database) no exact matches to the two Guinean haplotypes were found, as would be expected in the case of recent admixture. On the other hand, the Fulani U5 haplotype appears in a data set of West Africans (Wolof and Serer, Rando *et al.* 1998) and in Moroccans (unpublished data), pointing to the existence of a common African founder lineage of haplogroup U5. Again, as in haplogroup U6 the linguistic correlation suggests that the spread of the haplotype in Senegambia might be related to the movement of Berber populations. More data from North and West African populations is needed to better characterize the source and the time of the spread of this founder lineage.

### AMOVA and Principal Component Analysis

Analysis of molecular variance (AMOVA) in African populations attributed 15.6% to differences between groups, 3% to variation between populations within groups, and 81.6% to differences within populations



**Figure 3** PCA of African populations based on data from Table 2 but excluding the Kung/Khwe. Population codes are as follows: Mb (Morocco Berbers), Ab (Algerian Berbers), Ma (Morocco Arabs), Aa (Algerian Arabs), Eg (Egypt), Nv (Nile Valley), Et (Ethiopia), KS (Kenya/Sudan), Mo (Mozambique), Sm (Senegal Mandenka), Sx (Senegal mixed), NN (Niger/Nigeria), CV (Cabo Verde). Guinean ethnic groups were grouped (from a to g) as in Table 2. Axis 1 extracted 70.2% and axis 2, 12.5% of the total variation.

(overall  $F_{ST} = 0.184$ ,  $P < 0.0001$ ). A hierarchical structuring of populations into groups based on religion beliefs (Muslims vs. Animists) and geography (interior vs. littoral) gave similar values (data not shown).

A principal component (PC) analysis distinguished North Africans from sub-Saharan (Figure 3). The difference revealed by the first component is likely due to the presence of Eurasian mtDNA lineages among the North Africans and a relatively higher frequency of haplogroups L2a, L2c, L2d and U6 in Northwest Africa. The second component reflects L2/L0 frequencies. Moroccan Berbers and Arabs and Algerian Berbers are plotted close to Egyptians, supporting a common origin, while Algerian Arabs are placed apart. The Nile Valley sample occupies an intermediate position between Ethiopia and North Africans. The populations from Mozambique appear isolated and well differentiated from Kenya and Sudan. All the West Africans form a distinct and more compact cluster. Nevertheless the isolation of Senegalese Mandenka (Sm) and the Fula from Guiné (e) should be noted. As a whole, Guinean groups

are closer to West and then East Africa (see Axis 1, Figure 3).

### Final Remarks

Roughly 87% of the mtDNA lineages found in the Guinean populations are common in other West African populations. Not surprisingly, the highest number of matches was with Cape Verde followed by other populations from the area (Mandenka, Wolof, Fulbe), but also with Morocco. The notable L haplotype sharing with North Africans testifies to the absence of a real barrier between this region and typical sub-Saharan populations. On the other hand, some Guinean groups (Fula and Balanta for instance) present haplotypes otherwise observed to date in East-African and Middle East populations.

It is interesting to note that the Bantu-associated markers L0a 9bp del CoII/tRNA<sup>Lys</sup> (Soodyall *et al.* 1996), L3b motif 16124–16223–16278 (Watson *et al.* 1997), L3e1 particularly L3e1a characterized by mutation 16185 (Bandelt *et al.* 2001) or the 16192 L2a1

subclade (Pereira *et al.* 2001), were not found in our sample. This suggests that either Bantu migrations contributed very little to the gene pool of Guineans, despite the evidence of a Bantu migration starting from Cameroon and spreading towards Ghana, Nigeria, Burkina Faso and Mauritania, or that they had a distinct gene pool from that associated with the southwards migrants. The lack of Bantu branches of the Niger–Congo linguistic family, among a plethora of languages spoken in Guiné-Bissau, is more in agreement with the first hypothesis.

The finding of haplogroup M1 lineages of East African origin, albeit at low frequencies (3–5%) in Guinean groups with linguistic affinities to the Bak superfamily including Balanta, Baiote and Ejamat languages, supports the earlier suggestion of a Sudanese origin of the Balanta population and their spread to western Africa with kushitic migrants approximately 2000 years ago. Obviously, thereafter they were assimilated within the local population, acquiring their language. In particular the 16185 mutation might suggest a route through North Africa. The U6 presence in the Guinean pool, although at a low frequency, is not surprising, as these particular lineages have already been reported for this region. It seems plausible that the U5 lineages observed in the Fula arrived in Guiné via Sahel from North Africa before the slave trade. None of the typical European haplogroups (H, J, and T) were found in the present-day population of Guinea, whereas they exist at a fairly high frequency in North Africa in contrast to the U5 frequency (only 4.5%). This makes it less likely that the presence of U5 in Guiné, in particular, and in Northwest Africa in general, is due to recent admixture with the European population. A possible ancient migration from Asia to Africa was proposed by Cruciani *et al.* (2002) to explain the presence of some unusual Y-chromosome lineages identified in West Africa. Haplogroup R1 (defined by M173 mutation), without further branch defining mutations (M269 and M17) specific to Europeans, accounted for ~40% of the Y-chromosomes in North-Cameroon, while not yet having been sampled elsewhere in Africa. More data from Central and Western Africa are needed to cast light on the origin of such idiosyncratic mtDNA and Y chromosome lineages. Thus, our U5 sequences from the Guinean Fulbe people corroborate Cruciani's hypothesis of a prehistoric migration

from Eurasia to West Sub-Saharan Africa, testified by their present day restricted and localised distribution.

### Acknowledgments

The authors are grateful for the precious help of the Chairman of the Joint Chiefs of Staff and the Ministry of Health from Guiné-Bissau. ICCTI (Lisbon, Portugal) and the Regional Government of Madeira provided financial support to AB. AR is a recipient of a Ph.D. scholarship from Fundação para a Ciência e Tecnologia (FCT, Lisbon) reference SFRH/BD/12173/2003. TK was supported by the Estonian basic research grant 4769. We are also grateful to the contributions from two anonymous reviewers to an early version of the manuscript.

### Electronically Available Data

HVS-I and HVS-II haplotypes and their distribution among ethnic groups from Guiné-Bissau are available as Supplementary Material at the web site <http://www.ahg.com/>. A list of the PCR primers and conditions used to amplify all pertinent mtDNA regions are also included in the Supplementary Material web site.

### References

- Alimen, H. (1987) Evolution du climat et des civilisations depuis 40000 ans du nord au sud du Sahara occidental. (Premières conceptions confrontées aux données récentes.) Bull. L'Assoc. Franç. *L'Étude Quaternaire* **4**, 215–227.
- Alves-Silva, J., Santos, M., Guimarães, P., Ferreira, A. C., Bandelt, H.-J., Pena, S. D. & Prado, V. F. (2000) The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* **67**, 444–461.
- Almada, A. A. (1964) Tratado breve dos rios da Guiné do Cabo Verde, 2nd ed. Lisbon, 156p.
- Almeida, A. (1939) Sobre a etno-economia da Guiné Portuguesa. BGC, Lisboa Vol 15.
- Anderson, S., Bankier, A. T., Barrel, B. G., De Bruijn, M. H. L., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, E., Schreier, P. H., Smith, A. J. H., Staden, R. & Young, I. G. (1981) Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465.
- Andrews, R. M., Kubacka, I., Hinnery, P. E., Lightowlers, R. N., Turnbull, D. M. & Hoewll, N. (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nature Genetics* **23**, 147.
- Aumassip, G., Ferhat, N., Heddouche, A. & Vernet, R. (1988) Le milieu saharien aux temps pré historiques. In: *Milieux,*

- hommes et techniques du Sahara pré historique. Problèmes actuels* (eds. Aumassip G. *et al.*), pp. 9–29. L'Harmattan (ed.), Paris.
- Bandelt, H.-J., Forster, P., Sykes, B. C. & Richards, M. B. (1995) Mitochondrial portraits of human populations using median networks. *Genetics* **141**, 743–753.
- Bandelt, H.-J., Macaulay, V. & Richards, M. (2000) Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol Phylogeny Evol* **16**, 8–28.
- Bandelt, H.-J., Alves-Silva, J., Guimarães, P., Santos, M., Brehm, A., Pereira, L., Coppa, A., Larruga, J. M., Rengo, C., Scozzari, R., Torroni, A., Prata, M. J., Amorim, A., Prado, V. F. & Pena, S. D. J. (2001) Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. *Ann Hum Genet* **65**, 549–563.
- Barros, A. (1947) A invasão fula da circunscrição de Bafatá. Queda dos Beafadas e Mandingas, tribos “Gabungabé”. *BCCP (Bissau)* **15**, 737–743.
- Brehm, A., Pereira, L., Bandelt, H.-J., Prata, M. J. & Amorim, A. (2002) Mitochondrial portrait of the Cabo Verde Archipelago: the Senegambian outpost of Atlantic slave trade. *Ann Hum Genet* **66**, 49–60.
- Camps-Fabre, H. (1989) Capsien et Natoufien au Proche-Orient. In: Colloque L'homme maghrebien et son environnement depuis 100.000 ans. *Tiav. Du CAPMO (Algerie)* 71–104.
- Carreira, A. (1962) O fundamento dos etnónimos na Guiné Portuguesa. *Revista Garcia da Horta* **10**, 305–323.
- Carreira, A. & Quintino F. R. (1964) Antroponímia da Guiné Portuguesa. JIU, Lisboa, Vol. 49 and 52.
- Carreira A. (1983) Migrações nas Ilhas de Cabo Verde. Instituto Caboverdeano do Livro. 2nd. Ed., Lisboa.
- Chen, Y.-S., Olckers, A., Schurr, T. G., Kogelnik, A. M., Huoponen, K. & Wallace, D.C. (2000) mtDNA variation in the South African Kung and Khwe and their genetic relationships to other African populations. *Am J Hum Genet* **66**, 1362–1383.
- Côrte-Real, H. B. S. M., Macaulay, V. A., Richards, M. B., Hariti, G., Issad, M. S. & Cambon-Tomsen, A. *et al.* (1996) Genetic diversity in the Iberian Peninsula determined from mitochondrial sequence analysis. *Ann Hum Genet* **60**, 331–350.
- 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. & Underhill, P. (2002) A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* **70**, 1197–1214.
- Diallo, T. (1972) Les institutions politiques du Fouta Djallon. Initiation et Études Africaines, IFAN, Dakar, vol 28.

- Dutour, O., Vernet, R. & Aumassip, G. (1988) Le peuplement pré historique du Sahara. In: *Milieux, hommes et techniques du Sahara pré historique. Problèmes actuels* (eds. Aumassip, G. *et al.*), pp. 39–52. L'Harmattan (ed.), Paris.
- Page, J. (1995) A history of Africa. Routledge (ed.), London.
- Finnilä, S., Hänninen, I. E., Ala-Kokko, L. & Majamaa, K. (2000) Phylogenetic Network of the mtDNA Haplogroup U in Northern Finland Based on Sequence Analysis of the Complete Coding Region by Conformation-Sensitive Gel Electrophoresis. *Am J Hum Genet* **66**, 1017–1026.
- Forster, P., Harding, R., Torroni, A. & Bandelt, H.-J. (1996) Origin and evolution of Native American mtDNA variation: a reappraisal. *Am J Hum Genet* **59**, 935–945.
- Graven, L., Passarino, G., Semino, O., Boursot, P., Santachiara-Benerecetti, S., Langaney, A. & Excoffier, L. (1995) Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample. *Mol Biol Evol* **12**, 334–345.
- Hair, P. E. (1967) Ethnolinguistic continuity on the Guinea Coast. *JAH*, Vol VIII, 247–268.
- Krings, M., Salem, A. E., Bauer, K., Geisert, H., Malek, A. K., Chaix, L., Simon, C., Welsby, D., Di Rienzo, A., Utermann, G., Sajantila, A., Pääbo, S. & Stoneking, M. (1999) mtDNA analysis of Nile River Valley populations: a genetic corridor or a barrier to migration? *Am J Hum Genet* **64**, 1166–1176.
- Lopes, C. (1999) Kaabunké, espaço, território e poder na Guiné-Bissau, Gâmbia e Casamance pré-coloniais. CND, Lisboa.
- Macaulay, V., Richards, M., Hickey, E., Vega, E., Cruciani, F., Guida, V., Scozzari, R., Bonnè-Tamir, B., Sykes, B. & Torroni, A. (1999) The emerging tree of west Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs. *Am J Hum Genet* **64**, 232–249.
- Mishmar, D., Ruiz-Pesini, E., Golik, P., Macaulay, V., Clark, A. G., Hosseini, S., Brandon, M., Easley, K., Chen, E., Brown, M. D., Sukernik, R. I., Olckers, A. & Wallace, D.C. (2003) Natural selection shaped regional mtDNA variation in humans. *Proc Natl Acad Sci USA* **100**, 171–176.
- Moreira, J. M. (1964) Os Fulas da Guiné Portuguesa na panorâmica geral do mundo Fula. Boletim Cultural da Guiné Portuguesa XIX, 417–432.
- Nei, M. (1987) Molecular evolutionary genetics. Columbia University Press, New York.
- Passarino, G., Semino, O., Quintana-Murci, L., Excoffier, L., Hammer, M. & Santachiara-Benerecetti, A. S. (1998) Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. *Am J Hum Genet* **62**, 420–434.
- Pereira, L., Macaulay, V., Torroni, A., Scozzari, R., Prata, M. J. & Amorim, A. (2001) Prehistoric and historic traces

- in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* **65**, 439–458.
- Plaza, S., Calafell, F., Helal, A., Bouzerna, N., Lefranc, G., Bertranpetit, J., Comas, D. (2003) Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean. *Ann Hum Genet* **67**, 312–328.
- Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K. & Santachiara-Benerecetti, A. S. (1999) Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature Genetics* **23**, 437–441.
- Quintino, F. (1964) Sobrevivências da Cultura etiópica no ocidente africano. *BCG (Guiné)* **19**, 5–35.
- Quintino, F. (1967) Os povos da Guiné. *BCG (Guiné)* **22**, 5–40.
- Quintino, F. (1969) Os povos da Guiné. *BCG (Guiné)* **24**, 861–915.
- Rando, J. C., Pinto, F., González, A. M., Hernández, M., Larruga, J. M., Cabrera, V. M. & Bandelt, H.-J. (1998) Mitochondrial DNA analysis of North-west African populations reveals genetic exchanges with European, Near-Eastern, and Sub-Saharan populations. *Ann Hum Genet* **62**, 531–550.
- Rando, J. C., Cabrera, V. M., Larruga, J. M., Hernández, M., González, A. M., Pinto, F. & Bandelt, H.-J. (1999) Phylogeographic patterns of mtDNA reflecting the colonization of the Canary Islands. *Ann Hum Genet* **63**, 413–428.
- Richards, M., Macaulay, V., Hickey, E., Vega, E., Sykes, B., Guida, V., Rengo, C., Sellitto, D., Cruciani, F., Kivisild, T., Villems, R., Thomas, M., Rychkov, S., Rychkov, O., Rychkov, Y., Gölge, M., Dimitrov, D., Hill, E., Bradley, D., Romano, V., Cali, F., Vona, G., Demaine, A., Papiha, S., Triantaphyllidis, C., Stefanescu, G., Hatina, J., Belledi, M., Di Rienzo, A., Novelletto, A., Oppenheim, A., Norby, S., Al-Zaheri, N., Santachiara-Benerecetti, S., Torroni, A. & Bandelt, H.-J. (2000) Tracing European founder lineages in the Near Eastern mtDNA pool. *Am J Hum Genet* **67**, 1251–1276.
- Richards, M. & Macaulay, V. (2001) The mitochondrial gene tree comes of age. *Am J Hum Genet* **68**, 1315–1320.
- Salas, A., Richards, M., De la Fè, T., Lareu, M.-V., Sobrino, B., Sánchez-Diz, P., Macaulay, V. & Carracedo, A. (2002) The making of the African mtDNA landscape. *Am J Hum Genet* **71**, 1082–1111.
- Schneider, S. Roessli, D. & Excoffier, L. (2000) Arlequin ver. 2.000: A software for population genetics data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland.
- Soodyall, H., Vigilant, L., Hill, A. V., Stoneking, M. & Jenkins, T. (1996) mtDNA control-region sequence variation suggests multiple independent origins of an “Asian-specific” 9-bp deletion in sub-Saharan Africans. *Am J Hum Genet* **5**, 595–608.
- Stuhlmann, F. (1910) Handwerk und Industrie in Ostafrika, kulturgeschichtliche, Betrachtungen, Friedrichsen. Hamburg, 163p.
- Teixeira da Mota, A. (1954) Guiné Portuguesa. Agência Geral do Ultramar (ed.), Lisboa.
- Torroni, A., Huoponen, K., Francalacci, P., Petrozzi, M., Morelli, L., Scozzari, R., Obinu, D., Savontaus M. L. & Wallace D. C. (1997) Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**, 1835–1850.
- Torroni, A., Rengo, R., Guida, V., Cruciani, F., Sellito, D., Coppa, A., Calderon, F. L., Simionati, B., Valle, G., Richards, M., Macaulay, V. & Scozzari, R. (2001) Do the four clades of the mtDNA haplogroup L2 evolve at different rates? *Am J Hum Genet* **69**, 1348–1356.
- Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A.C. (1991) African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507.
- Watson, E., Forster, P., Richards, M. & Bandelt, H.-J. (1997) Mitochondrial footprints of human expansions in Africa. *Am J Hum Genet* **61**, 691–704.
- Wisconsin Package Version 10.0, Genetics Computer Group (GCG), Madison, Wisc. (1999).

Received: 2 July 2003

Accepted: 12 November 2003



## Announcement of Population Data

## Population data on 11 Y-chromosome STRs from Guiné-Bissau

Alexandra Rosa<sup>a,b,\*</sup>, Carolina Ornelas<sup>a,b</sup>, António Brehm<sup>a</sup>, Richard Villems<sup>b</sup><sup>a</sup> Department of Biology, Universidade da Madeira, Campus Universitário da Penteada, 9000-390 Funchal, Portugal<sup>b</sup> Department of Evolutionary Biology, Estonian Biocentre, University of Tartu, Riia 23, 51010 Tartu, Estonia

Received 12 January 2005

Available online 10 May 2005

## Abstract

The forensic value of Y-STR markers in Guiné-Bissau was accessed by typing of 215 males. Allele and haplotype frequencies, determined for loci DYS19, DYS389-I, DYS389-II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439 and the duplicated locus DYS385, are within the limits of variation found in other populations south of the Sahara. The level of discrimination achieved in Guineans is higher than for European or other African populations with comparable data. The haplotype diversity of 0.9995 is reduced to 0.9981 when the minimal haplotype is considered thus revealing the importance of increasing the number of typed loci.

© 2005 Published by Elsevier Ireland Ltd.

Keywords: Y-chromosome; Short tandem repeats; Powerplex® Y-System; Guiné-Bissau

**Population:** A total of 215 unrelated healthy males from Guiné-Bissau population whose ancestors were known to inhabit the same region for the last three generations.

**DNA extraction:** Chelex method [1] from leukocitary blood fraction.

**PCR:** A multiplex reaction for 11 markers was performed with Powerplex® Y-System (Promega) following the manufacturer's instructions. For the samples with individually typed Y-STRs (GB155 to GB207) published primers and conditions were used (DYS19, DYS389I/II, DYS390, DYS391, DYS392, DYS393 [1,2]; DYS385 [3]; DYS439 [4]).

**Typing:** Amplified PCR fragments were analyzed in ABI PRISM™ 310 Genetic Analyser along with Genescan 2.1 analysis software (AB Applied Biosystems). Typing followed the ISFG guidelines for Y-STR analysis [5]. The allele nomenclature system used is the proposed in [6,7] with the exception of the DYS389 locus [8]. Guidelines for

the presentation of population data, specified by Lincoln and Carracedo [9], have been considered.

**Results:** Described in Tables 1a–2. To note that the sample size is not the same for all markers, varying from  $N = 163$  to 215 (Table 1).

**Quality control:** Proficiency testing of the GEP-ISFG.

**Data analysis:** Frequency and diversity indexes [10] were calculated with Arlequin ver. 2.000 [11] for both loci (D) and haplotypes (H). The same software performed AMOVA tests for selected populations and locus-by-locus and an exact test of population differentiation (not considering DYS385). The YHRD database (www.yhrd.org) was consulted in the search for exact matches (both extended and minimal haplotypes).

**Other remarks:** The allele frequencies and range of the studied loci in Guiné-Bissau population (Table 1a and b) fit into the determined by other studies on sub-Saharan African populations [12–18]. The most outstanding differences are found when comparing with Non-Africans [13,14,19–25] or even North-Africans [26,27]. To note is the high prevalence in Guineans of alleles 15 for DYS19 (42%), 21 for DYS390 (67%), 11 for DYS392 (88%), 14 to DYS 437 (72%) and 11

Table 1a

Allele frequencies and gene diversity for ten Y-STR markers in Guiné-Bissau population

Allele	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439
8									0.0667	
9					0.0279					
10					0.7767	0.0736			0.2364	0.0421
11		0.014			0.1907	0.8773			0.6485	0.2804
12		0.1907			0.0047	0.0368	0.0093		0.0485	0.4533
13	0.092	0.5628				0.0086		0.0361		0.2009
14	0.0491	0.2279					0.586	0.7169		0.0234
15	0.4172	0.0047					0.0512	0.1205		
16	0.2638							0.012		
17	0.1779							0.1145		
20				0.0186						
21				0.6744						
22				0.214						
23				0.0419						
24				0.0512						
28			0.0143							
29			0.1667							
30			0.3714							
31			0.3095							
32			0.1333							
33			0.0048							
<i>N</i>	163	215	210	215	215	163	215	166	165	214
<i>D</i>	0.7182	0.5975	0.7239	0.497	0.3612	0.2248	0.5314	0.4598	0.52	0.6764
S.E.	0.0199	0.0249	0.0139	0.0339	0.0349	0.0422	0.0212	0.0431	0.0359	0.0178

*N*, sample size; *D*, gene diversity; S.E., standard error of *D*.

for DYS438 (65%) which in other populations are marginal or, if the most common allele, retain a lower fraction of the variation. For DYS393 a percentage of 59% of allele 14 is so far the highest reported, even for a West African population. A broader range for DYS389I (11–15 versus 12–14 for Europe) or a more limited one, as for DYS438 (10–12 versus 9–13 in Europe) are examples of other distinct features. The haplotype distribution of DYS385 ranges from alleles 13 to 21 where the most frequent haplotypes 15–16, 16–16 and 16–17 (~16%) are either absent or weakly represented outside of Africa. All loci show a unimodal distribution, including DYS392, which is bimodal in most non-Africans. As for the diversity indexes (*D*), DYS19 and DYS389II exhibit the highest diversity in this study ( $D = 0.7182$  and  $0.7239$ ), not to consider DYS385 ( $H = 0.9031$ ). Together with DYS393 these loci held higher gene/haplotype diversity than most Europeans, thus more informative for African populations. On the other hand, DYS391 seems to have a lower *D* than Europeans (but still higher than Asians) while DYS392 exhibits the lowest diversity. Data of both loci supports a limited utility in forensic casework in several sub-Saharan populations as previously suggested [18].

The eleven Y-STRs are fully typed for 161 individuals resulting in 154 distinct haplotypes ( $H = 0.9995 \pm 0.0008$ ), with the highest frequency of two individuals (Table 2). GB155 to GB207 were not taken into account for the calculation, as their 11 Y-STR pattern might be similar to the previous. The discriminatory power achieved is higher

Table 1b  
Haplotype frequency and diversity for DYS385 in Guiné-Bissau population

Haplotype	DYS385
13,14	0.0049
13,15	0.0049
13,16	0.0099
13,17	0.0049
14,14	0.0345
14,15	0.0148
14,16	0.0099
14,17	0.0345
15,15	0.0443
15,16	0.1576
15,17	0.0493
15,18	0.0197
16,16	0.1675
16,17	0.1478
16,18	0.0591
16,19	0.0197
16,20	0.0049
17,17	0.0985
17,18	0.069
17,20	0.0049
18,18	0.0197
18,19	0.0049
18,20	0.0099
20,21	0.0049
<i>N</i>	203
<i>H</i>	0.9031
S.E.	0.0084

*N*, sample size; *H*, haplotype diversity; S.E., standard error of *H*.

\* Corresponding author. Tel.: +351 966110876; fax: +351 291705390.

E-mail address: arosa@uma.pt (A. Rosa).





STRs set. For the African populations in the database, the highest number of matching lineages were with Angola, Mozambique and West Africa.

In order to evaluate the discriminatory power of an extended haplotype, *H* was determined for sets of ten markers (minimum haplotype plus one). The additional marker causes a variation in haplotype diversity as follows: DYS437 ( $H = 0.9982 \pm 0.0010$ , 143 haplotypes), DYS438 ( $H = 0.9986 \pm 0.0009$ , 146 haplotypes) and DYS439 ( $H = 0.9994 \pm 0.0008$ , 153 haplotypes). The level of discrimination obtained by additional typing of DYS439 confirms its usefulness for forensic purposes [12,24].

#### Acknowledgments

The authors are grateful for permissions to collect blood samples by the Ministry of Health of the Republic of Guiné-Bissau. AMI-Assistencia Médica Internacional gave local support. This work has been possible thanks to the technical help of Siiri Rootsi and Juri Parik from the Department of Evolutionary Biology, EBC, Estonia and Ana Teresa Fernandes and Rita Goncalves from the Human Genetics Laboratory, University of Madeira. AR is beneficiary of the fellowship grant SFRH/BD/12173/2003 from FCT, Fundacao para a Ciencia e Tecnologia. AB received a grant from the Regional Government of Madeira (Portugal).

#### References

- [1] M.V. Lareu, C.P. Phillips, A. Carracedo, P.J. Lincoln, S. Court, J.A. Thomson, Investigation of the STR locus HUMTH01 using PCR and two electrophoresis formats: UK and Galician Caucasian population surveys and usefulness in paternity investigations, *Forensic Sci. Int.* 66 (1994) 41–52.
- [2] M. Kayser, P. de Knijff, P. Dieltjes, M. Krawczak, M. Nagy, T. Zerjal, A. Pandya, C. Tyler-Smith, L. Roewer, Applications of microsatellite-based Y-chromosome haplotyping, *Electrophoresis* 18 (1997) 1602–1607.
- [3] P.M. Schneider, S. Meuser, W. Waiyawuth, Y. Seo, C. Rittner, Tandem repeat structure of the duplicated Y-chromosomal STR locus DYS385 and frequency studies in the German and three Asian populations, *Forensic Sci. Int.* 97 (1998) 61–70.
- [4] Q. Ayub, A. Mohyuddin, R. Qamar, K. Mazhar, T. Zerjal, S.Q. Mehdi, C. Tyler-Smith, Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information, *Nucleic Acids Res.* 28 (2000) e8.
- [5] P. Gill, C. Brenner, B. Brinkmann, B. Budowle, A. Carracedo, M.A. Jobling, K.P. de, M. Kayser, M. Krawczak, W.R. Mayr, N. Morling, B. Olaisen, V. Pascali, M. Prinz, L. Roewer, P.M. Schneider, A. Sajantila, C. Tyler-Smith, DNA Commission of the International Society of Forensic Genetics; recommendations on forensic analysis using Y-chromosome STRs, *Forensic Sci. Int.* 124 (2001) 5–10.
- [6] M. Kayser, A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, C. Schmitt, L. Roewer, Evaluation of Y-chromosomal STRs: a multicenter study, *Int. J. Legal Med.* 110 (1997) 125–129.
- [7] P. de Knijff, M. Kayser, A. Caglia, D. Corach, N. Fretwell, C. Gehrig, G. Graziosi, F. Heidorn, S. Herrmann, B. Herzog, M. Hidding, K. Honda, M. Jobling, M. Krawczak, K. Leim, S. Meuser, E. Meyer, W. Oesterreich, A. Pandya, W. Parson, G. Penacino, A. Perez-Lezaun, A. Piccinini, M. Prinz, L. Roewer, Chromosome-Y microsatellites: population genetic and evolutionary aspects, *Int. J. Legal Med.* 110 (1997) 134–149.
- [8] L. Roewer, M. Kayser, K.P. de, K. Anslinger, A. Betz, A. Caglia, D. Corach, S. Furedi, L. Henke, M. Hidding, H.J. Kargel, R. Lessig, M. Nagy, V.L. Pascali, W. Parson, B. Rolf, C. Schmitt, R. Szibor, J. Teifel-Greding, M. Krawczak, A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males, *Forensic Sci. Int.* 114 (2000) 31–43.
- [9] P. Lincoln, A. Carracedo, Publication of population data of human polymorphisms, *Forensic Sci. Int.* 110 (2000) 3–5.
- [10] M. Nei, *Molecular Evolutionary Genetics*, Columbia University Press, New York, 1987.
- [11] S. Schneider, D. Roessli, L. Excoffier, Arlequin ver. 2.000: A Software for Population Genetics Data Analysis, Genetics and Biometry Laboratory, University of Geneva, Switzerland, 2000.
- [12] L. Gusmao, C. Alves, A. Amorim, Molecular characteristics of four human Y-specific microsatellites (DYS434, DYD437, DYS438, DYS439) for population and forensic studies, *Ann. Hum. Genet.* 65 (2001) 285–291.
- [13] M. Kayser, M. Krawczak, L. Excoffier, P. Dieltjes, D. Corach, V. Pascali, C. Gehrig, L.F. Bernini, J. Jespersen, E. Bakker, L. Roewer, K.P. de, An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations, *Am. J. Hum. Genet.* 68 (2001) 990–1018.
- [14] M. Seielstad, E. Bekele, M. Ibrahim, A. Toure, M. Traore, A view of modern human origins from Y chromosome microsatellite variation, *Genome Res.* 9 (1999) 558–567.
- [15] M.J. Trovada, C. Alves, L. Gusmao, A. Abade, A. Amorim, M.J. Prata, Evidence for population sub-structuring in Sao Tome e Principe as inferred from Y-chromosome STR analysis, *Ann. Hum. Genet.* 65 (2001) 271–283.
- [16] S. Alvarez, A.M. Lopez-Parra, M.S. Mesa, V. Jimenez, S.J. Herrera, F. Bandres, A. Arraztio, J.M. Rubio, E. Arroyo-Pardo, Three Y-chromosome STR frequencies in a population from equatorial Guinea (Central Africa), *J. Forensic Sci.* 47 (2002) 224–225.
- [17] L. Pereira, L. Gusmao, C. Alves, A. Amorim, M.J. Prata, Bantu and European Y-lineages in Sub-Saharan Africa, *Ann. Hum. Genet.* 66 (2002) 369–378.
- [18] N. Leat, M. Benjeddou, S. Davison, Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa, *Forensic Sci. Int.* 144 (2004) 73–75.
- [19] L. Roewer, M. Krawczak, S. Willuweit, M. Nagy, C. Alves, A. Amorim, K. Anslinger, C. Augustin, A. Betz, E. Bosch, A. Caglia, A. Carracedo, D. Corach, A.F. Dekairelle, T. Dobosz, B.M. Dupuy, S. Furedi, C. Gehrig, L. Gusmao, J. Henke, L. Henke, M. Hidding, C. Hohoff, B. Hoste, M.A. Jobling, H.J. Kargel, K.P. de, R. Lessig, E. Liebeherr, M. Lorente, B. Martinez-Jarreta, P. Nieves, M. Nowak, W. Parson, V.L. Pascali, G. Penacino, R. Ploski, B. Rolf, A. Sala, U. Schmidt, C. Schmitt, P.M. Schneider, R. Szibor, J. Teifel-Greding, M. Kayser, Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes, *Forensic Sci. Int.* 118 (2001) 106–113.
- [20] M. Kayser, S. Brauer, H. Schadlich, M. Prinz, M.A. Batzer, P.A. Zimmerman, B.A. Boatman, M. Stoneking, Y chromosome STR haplotypes and the genetic structure of U.S. populations of African, European, and Hispanic ancestry, *Genome Res.* 13 (2003) 624–634.
- [21] L. Gusmao, C. Alves, S. Belez, A. Amorim, Forensic evaluation and population data on the new Y-STRs DYS434, DYS437, DYS438, DYS439 and GATA A10, *Int. J. Legal Med.* 116 (2002) 139–147.
- [22] R. Uchihii, T. Yamamoto, K. Usuda, T. Yoshimoto, M. Tanaka, S. Tokunaga, R. Kurihara, K. Tokunaga, Y. Katsumata, Haplotype analysis with 14 Y-STR loci using 2 multiplex amplification and typing systems in 2 regional populations in Japan, *Int. J. Legal Med.* 117 (2003) 34–38.
- [23] B. Quintans, S. Beleza, M. Brion, P. Sanchez-Diz, M. Lareu, A. Carracedo, Population data of Galicia (NW Spain) on the new Y-STRs DYS437, DYS438, DYS439, GATA A10, GATA A7.1, GATA A7.2, GATA C4 and GATA H4, *Forensic Sci. Int.* 131 (2003) 220–224.
- [24] E. Bosch, A.C. Lee, F. Calafell, E. Arroyo, P. Henneman, K.P. de, M.A. Jobling, High resolution Y-chromosome typing: 19 STRs amplified in three multiplex reactions, *Forensic Sci. Int.* 125 (2002) 42–51.
- [25] P. Grignani, G. Peloso, P. Fattorini, C. Previdere, Highly informative Y-chromosomal haplotypes by the addition of three new STRs DYS437, DYS438 and DYS439, *Int. J. Legal Med.* 114 (2000) 125–129.
- [26] E. Bosch, F. Calafell, A. Perez-Lezaun, D. Comas, H. Izaabel, O. Akhayat, A. Sefiani, G. Hariti, J.M. Dugoujon, J. Bertranpetit, Y chromosome STR haplotypes in four populations from northwest Africa, *Int. J. Legal Med.* 114 (2000) 36–40.
- [27] L. Quintana-Murci, A. Bigham, H. Rouba, A. Barakat, K. McElreavey, M. Hammer, Y-chromosomal STR haplotypes in Berber and Arabic-speaking populations from Morocco, *Forensic Sci. Int.* 140 (2004) 113–115.
- [28] B. Arredi, E.S. Poloni, S. Paracchini, T. Zerjal, D.M. Fathallah, M. Makrelouf, V.L. Pascali, A. Novelletto, C. Tyler-Smith, A Predominantly Neolithic Origin for Y-Chromosomal DNA Variation in North Africa, *Am. J. Hum. Genet.* 75 (2004) 338–345.
- [29] E. Arroyo-Pardo, L. Gusmao, A.M. López-Parra, C. Baeza, M.S. Mesa, A. Amorim, Genetic variability of 16 Y-chromosome STRs in a sample from Equatorial Guinea (Central Africa), *Forensic Sci. Int.* 149 (2005) 109–113.
- [30] C. Alves, L. Gusmao, J. Barbosa, A. Amorim, Evaluating the informative power of Y-STRs: a comparative study using European and new African haplotype data, *Forensic Sci. Int.* 134 (2003) 126–133.
- [31] M.T. Zarrabeitia, J.A. Riancho, L. Gusmao, M.V. Lareu, C. Sanudo, A. Amorim, A. Carracedo, Spanish population data and forensic usefulness of a novel Y-STR set (DYS437, DYS438, DYS439, DYS460, DYS461, GATA A10, GATA C4, GATA H4), *Int J Legal Med* 117 (2003) 306–311.

This Provisional PDF corresponds to the article as it appeared upon acceptance. Copyedited and fully formatted PDF and full text (HTML) versions will be made available soon.

### Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective

*BMC Evolutionary Biology* 2007, 7:124 doi:10.1186/1471-2148-7-124

Alexandra Rosa (parosa@igc.gulbenkian.pt)  
 Carolina Ornelas (nyjoana@yahoo.com)  
 Mark A Jobling (maj4@leicester.ac.uk)  
 Antonio Brehm (brehm@uma.pt)  
 Richard Villems (rvillems@ebc.ee)

ISSN 1471-2148

**Article type** Research article

**Submission date** 27 November 2006

**Acceptance date** 27 July 2007

**Publication date** 27 July 2007

**Article URL** <http://www.biomedcentral.com/1471-2148/7/124>

Like all articles in BMC journals, this peer-reviewed article was published immediately upon acceptance. It can be downloaded, printed and distributed freely for any purposes (see copyright notice below).

Articles in BMC journals are listed in PubMed and archived at PubMed Central.

For information about publishing your research in BMC journals or any BioMed Central journal, go to

<http://www.biomedcentral.com/info/authors/>

© 2007 Rosa 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.

### Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective

Alexandra Rosa<sup>1,2,§</sup>, Carolina Ornelas<sup>1,2</sup>, Mark A Jobling<sup>3</sup>, António Brehm<sup>2</sup>, Richard Villems<sup>1</sup>

<sup>1</sup>Department of Evolutionary Biology, Estonian Biocentre, Riia 23, 51010 Tartu, Estonia

<sup>2</sup>Human Genetics Laboratory, University of Madeira, Campus of Penteada, 9000-390, Funchal, Portugal

<sup>3</sup>Department of Genetics, University of Leicester, Leicester, LE1 7RH, United Kingdom

<sup>§</sup>Corresponding author

Email addresses:

AR: arosa@uma.pt

CO: nyjoana@yahoo.com

MJ: maj4@leicester.ac.uk

AB: brehm@uma.pt

RV: rvillems@ebc.ee

## Abstract

### Background

The geographic and ethnolinguistic differentiation of many African Y-chromosomal lineages provides an opportunity to evaluate human migration episodes and admixture processes, in a pan-continental context. The analysis of the paternal genetic structure of Equatorial West Africans carried out to date leaves their origins and relationships unclear, and raises questions about the existence of major demographic phenomena analogous to the large-scale Bantu expansions. To address this, we have analysed the variation of 31 binary and 11 microsatellite markers on the non-recombining portion of the Y chromosome in Guinea-Bissau samples of diverse ethnic affiliations, some not studied before.

### Results

The Guinea-Bissau Y chromosome pool is characterized by low haplogroup diversity ( $D=0.470$ ,  $sd\ 0.033$ ), with the predominant haplogroup E3a\*-M2 shared among the ethnic clusters and reaching a maximum of 82.2% in the Mandenka people. The Felupe-Djola and Papel groups exhibit the highest diversity of lineages and harbor the deep-rooting haplogroups A-M91, E2-M75 and E3\*-PN2, typical of Sahel's more central and eastern areas. Their genetic distinction from other groups is statistically significant ( $P=0.01$ ) though not attributable to linguistic, geographic or religious criteria. Non sub-Saharan influences were associated with the presence of haplogroup R1b-P25 and particular lineages of E3b1-M78.

### Conclusions

The predominance and high diversity of haplogroup E3a\*-M2 suggests a demographic expansion in the equatorial western fringe, possibly supported by a local agricultural center. The paternal pool of the Mandenka and Balanta displays evidence of a particularly marked population growth among the Guineans, possibly reflecting the demographic effects of the agriculturalist lifestyle and their putative relationship to the people that introduced early cultivation practices into West Africa. The paternal background of the Felupe-Djola and Papel ethnic groups suggests a better conserved ancestral pool deriving from East Africa, from where they have supposedly migrated in recent times. Despite the overall homogeneity in a multiethnic sample, which contrasts with their social structure, minor clusters suggest the imprints of multiple peoples at different timescales: traces of ancestral inhabitants in haplogroups A-M91 and B-M60, today typical of hunter-gatherers; North African influence in E3b1-M78 Y chromosomes, probably due to trans-Saharan contacts; and R1b-P25 lineages reflecting European admixture via the North Atlantic slave trade.

## Background

Many genetic studies of sub-Saharan Y chromosome variation have paid special attention to the large-scale Bantu expansions, and the particular pool of the "relic" Central African Pygmies and the South African Khoisan [1-7], while little is known about the events that have shaped the paternal structure of Equatorial West Africans. Although anthropological evidence is scarce, the earliest traces of West Atlantic occupation by modern humans dates back 40 ky [8,9]. Later climatic changes, when around 9 kya the Sahara was at its wettest [10], created conditions for both the massive displacement of



people and the spread of agriculture, reaching previously uninhabited areas and promoting admixture with isolated populations [11-14]. Although the farming practices in Sahel could have started earlier than 6 kya [15,16], firm archaeological evidence points to the domestication of local sorghum, millet and yams ~4 kya [17]. Together with the introduction of iron-smelting techniques ~ 2.7 kya, agriculture led ultimately to the large-scale Bantu migrations from the Gulf of Guinea to the south of the continent [18]. From the perspective of Y chromosome genetic variation, such movements are believed to have erased much of the pre-existing diversity, replacing it with the now dominant haplogroup E3a-M2 lineages [4,19,20].

The inhabitants of the Guinea-Bissau area have certainly been under the influence of several demographic events since prehistorical times, as a result of migratory movements, trade networks and consecutive invasions. The first recorded influx of ethnically defined groups is the arrival of Fulbe people in the 8<sup>th</sup> century AD, from a Central African epicenter [21]. First contact with the North African Berbers dates back to at least the 9<sup>th</sup> century, and was repeated in the 11<sup>th</sup> century when, pushed by the Omniades, these people came to occupy the vicinity of Senegal [22]. The economic shift in the Sahel allowed more centralized states to form (namely the "Black Kingdoms" in the period between the 8<sup>th</sup> and 16<sup>th</sup> centuries, [23]), linked by a trading corridor reaching from Mauritania to Niger [18]. In the following centuries pastoral Fulbe arrived again slowly but *en masse*, together with the Mandenka, and became the most prevalent people in Guinea-Bissau territory. Oral tradition also states that the Djola people – Felupe-Djola, Baiote and possibly Beafada – came from Sudan in the 15<sup>th</sup>-16<sup>th</sup> centuries [24]. As for the Balanta, Sudanese or Bantu affinities may argue for their cultural and phenotypic aspects. Though research on the background of the Nalú is less

advanced, Teixeira da Mota [25] considers them to be the autochthonous people of the region. The same author identifies Bijagós as a separated branch of Djola or relatives of Papel and Nalú. The main ethnic groups now present in Guinea-Bissau (Figure 1; see Additional file 1) were already settled in the region in the 15<sup>th</sup> century, at the time of arrival of the Portuguese. With the establishment of the Atlantic slave trade the region experienced an input of Europeans, in their vast majority males, whose genetic imprint is undetermined. Many of the ethnic barriers were brought down, in particular the endogamic practices, promoting an intense cultural contact and higher levels of admixture between groups than before.

The present study intends to characterize the paternal genetic pool of Guineans, focusing on their ethnic affiliation, by the use of binary markers and microsatellites on the non-recombining region of the Y chromosome (NRY). Our sample (n=282) extends significantly the Y-chromosomal coverage of West African populations (Senegal [5], Gambia/Senegal Wolof and Mandenka [7], Mali [2] and Dogon [7], Burkina-Faso [1,26], Ghana Ewe, Ga and Fante [7]) both in size and number of surveyed ethnic groups. The unique features of the Y chromosome system, namely its haploid and non-recombining nature and paternal inheritance, provide an opportunity to evaluate the temporal and spatial aspects of population movements, in the light of the available non-genetic evidence.

## Results and discussion

### Y chromosome haplogroup variation

The fairly homogeneous paternal structure of Guinea-Bissau ( $D=0.470$ ,  $sd\ 0.033$ ), is not surprising given the general landscape of sub-Saharan low Y-chromosomal haplogroup diversity [2,27] and its reported east-to-west decline along a Central African corridor [28]. Responsibility for the low diversity is attributed to the highly frequent E3a\*-M2 and E1\*-M33 lineages (72.0% and 15.6%, respectively) that are shared among all ethnic clusters (Figure 2). In our dataset the Mandenka harbor the highest frequency (82.2%) of the E3a\*-M2 paragroup, fitting the context of its closest neighbors (~80% in Senegalese [5] and Gambia/Senegal Mandinka [7]). The lack of diversity of West African Y chromosomes together with the predominance of E3a\*-M2 lineages (assuming a frequency peak only equivalent to that in Central Africa; Figure 3) reinforces its link to agricultural expansion [3,4,19,20] and hint at the existence of a large local center of cultivation [14-16,18,29]. We hypothesized that the newly adopted lifestyle created conditions for major demographic growth, obscuring earlier patterns of lineages. Alternatively, a moderate farming expansion may have occurred on a background of reduced diversity, following the 5.5 kya savanna retreat [30] or the malarial epidemic episodes which were an outcome of pastoral habits [31].

The lifestyle transition in West Africa was most likely promoted by people other than the Bantu, as no relevant westwards migrations of these people are reported and none or few Bantu languages are found in the area today. In fact, the West African center may date earlier than that documented for Central Africa and may have acted as a western source of knowledge [14-16]. Based on the high frequency and microsatellite diversity of E3a\*-M2 in the Mandenka and Balanta (Figure 2; see Additional file 2), we suggest that these people may have experienced a particular benefit from food production. If so, this might associate their ancestors with the people who implemented

the farming habits in the Guinea-Bissau area. The Mandenka are physically and culturally descendants of the Mande, protagonists of agricultural population expansions in Niger/Mali/Burkina-Faso region [18] and rulers of the West African Black Empires, based on trade and agriculture. For the Balanta, the cultural and physical affinities with Bantu suggest a common origin at the end of the Pleistocene [24], so it may be that different peoples jointly learnt the agricultural techniques. The E3a7-M191 lineages of one Fulbe and two Mandenka individuals of Guinea-Bissau are undoubtedly representatives of a Central African lineage that followed a trajectory to the west [2,3,5,32].

Haplogroup E1\*-M33, of probable local radiation (5-7% in Senegal and Burkina-Faso [2,3,5,7], 40.4% in Mali and 52.9% in Fulbe of Cameroon [1,26]), is surprisingly frequent in Felupe-Djola and Papel (34.0% and 20.3%). Both ethnic groups exhibit the highest haplogroup diversity ( $0.5 < D < 0.6$ ) and the deepest-rooting phylogenetic types in our dataset – haplogroups A-M91, E2-M75 and E3\*-PN2 – some with occasional occurrences in Fulbe and Balanta (Figure 2). These minor imprints may represent movements from Sahel's more central and eastern parts, seen, for example, in the typically Ethiopian/Sudanese E3\*-PN2 lineages that have reached Senegambia [2,3,5]. The Djola's oral tradition claims an arrival from Sudan in the 15<sup>th</sup>-16<sup>th</sup> centuries which is supported by their carrying the lowest fraction of E3a\* in our dataset (58.0%). At the same time, the relatively short time of residence and/or the genetic isolation on cultural grounds has not contributed to a greater homogeneity among the peoples. The Papel, curiously also affiliated to the Bak-speakers, may either represent a legacy left by earlier inhabitants of the Guinean delta, survivors of an ancient pool through demographic

reductions and expansions, or later arrivals who have preserved a more discrete genetic identity.

Of greater prevalence in the East quadrant of Africa and among South African Khoisan (~12% and 15%, respectively; [2,5]) the paragroup E3b\*-M35 is common to Felupe-Djola and Papel (~2%) but is also found among Fulbe and Mandenka (~4%). Its presence at ~2% in Guinea-Bissau and ~5% in Senegal may also indicate loose relationships to the North, where it is widespread at rather low frequencies (2-4%, [1,26,33-35]). A similar scenario of Eastern prevalence and North African spread traces the African distribution of E3b1-M78 (~26% in Sudan and Ethiopia and 19% in NW-African Arabs), not to mention the ~7% in the Near Eastern and European people [1,5,26,33-35]. In Guinea-Bissau this haplogroup attains the highest frequency so far reported for West Africa (~4%).

The remainder of binary marker variation falls into haplogroups A, B and R, each detected at marginal frequencies (0.4-3.9%). Clades A-M91 and B-M60, the most divergent of the haplogroups of the Y chromosome tree, are associated with the earliest modern human diversification and are putative markers of the first pan-African dispersals of hunter-gatherers [2,3,7,20,36]. However, the Guinea-Bissau A-M91 lineages do not belong to the widespread A3-M32 but to the A1-M31 subcluster, with reported marginal presence in Mali (2.0% [2,7]), Gambia/Senegal Mandinka (5.1% [7]) and North African Berbers (3.1% [1,33-35]). Any association of Balanta to the Sudanese-speakers is traceable only in the A3b2-M13 and E3\* Y chromosomes. The B-M60 variant observed in almost all sub-Saharan collections [28] was only found in Nalú. One other Nalú individual belongs to the rare and deep-rooting DE\* paragroup described in five Nigerians [37] and thus representing a coalescent "missing link",

8

paraphyletic to haplogroups D and E. The two Western European R1b-P25 lineages in Fulbe and Bijagós are best explained by recent European influence, at the time of the slave trade. A partial introduction through North African pastoral immigrants can not be rejected, where the 3-12% of R1b-P25 are due to the geographic proximity and the long reported contacts with Europe and Middle-East [33]. The European source seems nevertheless more likely: firstly, Y chromosome signatures of European presence have a reported great expression in the nearby Cape Verdians [38] and secondly, highly frequent North African haplogroups that would have been equally carried by the migrants (e.g. E3b2-M81) are absent in Guineans. The M173 and P25 derived states in both our samples rule out a relationship to the R1\*-M173 lineage previously found in Cameroon, Oman, Egypt and Rwanda, and adduced to support the "Back-to-Africa" theory [3,28].

Pairwise  $F_{ST}$  analysis of haplogroup frequencies reveals the Felupe-Djola as the only group statistically significantly different from others, namely Bijagós ( $F_{ST}=0.095$ ,  $P=0.027$ ), Fulbe ( $F_{ST}=0.081$ ,  $P=0.004$ ) and Mandenka ( $F_{ST}=0.107$ ,  $P=0.004$ ). The exact test of population differentiation reveals similar information, further distinguishing Papel from Bijagós ( $P=0.01$ ), Fulbe ( $P=0.003$ ) and Mandenka ( $P=0.04$ ). These results are in agreement with principal components analysis (PCA; see below) and the interpretation of the greater distinctiveness of the paternal pool of Felupe-Djola and Papel among other Guineans.

#### PCA and AMOVA analysis

A PCA of Guinean and other African populations Y chromosome haplogroup frequencies is depicted in Figure 4a [see Additional files 3 and 4 for population details].

9

The 1<sup>st</sup> PC clearly separates the Afro-Asiatic speakers from other linguistic families, independently of their geographic location. Consistent with geographical grouping, North and West Africans cluster in independent and tighter groups. The coordinates of North Africans are attributable to haplogroups E3b2-M81 and J-12f2 while West Africans' Y chromosomes cluster largely due to E3a\*-M2, and E1-M33 to a lesser extent. Central and South African people are more dispersed in the plot, many lying closer to the Eastern populations (due to the presence of R-M207, A3-M32 and B2-M182 lineages) while others lie closer to the Western cluster. A linguistic correlation is hypothesized to underlie the genetic proximity of Bantu-speakers occupying different quadrants of the continent, driven by the E3a7-M191. Guinea-Bissau groups are included in the western cluster of populations, in close vicinity to Gambia/Senegal Wolof and Mandinka [7] and Senegalese [5] with which they share numerous population groups. It is noteworthy that the Guinea-Bissau Fulbe show a distinct pool from other Fulbe people, namely the ones in Burkina-Faso and Cameroon, and are integrated within the Guinea-Bissau variation. A PCA of Guinea-Bissau ethnic groups is illustrated in Figure 4b, less biased by the major influence of haplogroup E3a\*-M2 and where the influence of minor clusters is emphasized. The Felupe-Djola and Papel have distinctive positions, largely a result of the high frequency of haplogroup E1-M33. The Bijagós, inhabitants of the archipelago, are placed apart in closer relation to the mainland Fulbe. The position of Mandenka is clearly defined by its E3a\*-M2 composition.

The AMOVA yielded no statistically significant results for ethnic group distinction on any of the defined criteria, with ~97% of the variance occurring at the within-population level ( $P<0.05$ ; see Additional file 5). These results suggest that in spite of obvious sociocultural differences among groups, marked by the supposedly strict

10

admixture barriers, their Y chromosome gene pool remains largely shared, because of common origin or common history of genetic admixture without language shift.

#### Microsatellite haplotypes within haplogroups

Y-chromosomal microsatellites provide further haplotype resolution, and are of particular use when, as in this case, some haplogroups are very prevalent. The E3a\*-M2 microsatellite profiles of Mandenka and Balanta are the most diverse among our data ( $R_{ST}$  average gene diversity, see Additional file 2) and attest to an earlier origin or more pronounced expansion. Since the corresponding parameter in Fulbe is less diverse we consider this to signal either a genetic bottleneck or their more recent expansion and late arrival in the West. The data are consistent with the less diverse E3a-M2 profile in Central and South Africans (data not shown). Haplotypes within E3b1-M78 are supposed to represent distinct clusters of local genetic drift [39]. The rare DYS439 allele 10 of a so-called E3b1- $\beta$  cluster particularly widespread among Moroccan Arabs defines a contribution to the Guinean Fulbe and Bijagós from North West Africans who have crossed the Sahara. The hypothesis of much later European contribution is valid though the remaining variability is absent (except for two R1b-P25 chromosomes) and none of the Guinean haplotypes carry the A7.1 allele with size 9, characteristic of Europe [39]. Microsatellite networks for paragroup E3a\*-M2 and haplogroup E3b1-M78 are not informative due to multiple reticulations and the absence of a clear haplotype sub-structure particularly associated to ethnic groups [see Additional file 6]. Further refinement awaits the finding of new markers especially within paragroup E3a\*-M2. The microsatellite profile of the DE\* individual is one mutational step away from the allelic

11

state described for Nigerians (DYS390\*21, DYS388 not tested; [37]), therefore suggesting a common ancestry but not elucidating the phylogenetics.

The Fulbe E3a\*-M2 extended haplotypes find exact matches in Equatorial Guinea, Mozambique, Angola and Xhosa (H61, H48, H69 [40,41]; see Additional file 7) supporting their broad distribution. The Mandenka share E3a\*-M2 variants with all other groups in Guinea-Bissau and do not match types outside Central-West Africa (except H67 in Mozambique), a sign of localized expansion and increased influence over their ethnic neighbors. The Felupe-Djola, Balanta and Papel each share one microsatellite haplotype (H49, H46 and H127, respectively) with Mozambique and Angola. Several E3a\* eight-loci profiles matched Europeans (H29, H38, H44, H30, H152, H153 and H55), most likely descendants of incoming slaves. Three Fulbe E3b1-M78 haplotypes (H155 and H156) were found to match Spanish haplotypes [42] and samples in Central Portugal, Macedonia, Romania and Poland (YHRD database [43]). Both profiles present the A7.1 allele 12, quite frequent in Equatorial Guinea [44]. The R1b-P25 H165 has a 10-loci haplotype found in 68 worldwide populations, of which 53 are European (nine matches in Portugal, YHRD). The picture for H166-R1b is quite different since on a 7-loci basis it matches four Europeans and two individuals from the Reunion Islands, known to have a European-permeable culture.

#### MtDNA haplogroup variation

Comparisons between mtDNA and Y-chromosomal diversity are hindered because of the very different mutational properties of their SNPs and Y microsatellites, and because of SNP ascertainment bias on the Y chromosome. Therefore, caution is needed when interpreting the results.

12

The maternal inheritance of Guineans is markedly West African, with haplogroups coalescing at distinct timeframes, from the initial occupation of the area to the later inputs of people [45]. Of relevance for comparison with the paternal counterpart are the signatures of recent expansion in haplogroups frequent in Senegambia, namely haplogroups L2a-L2c, the latter displaying an almost starlike phylogeny and being particularly frequent in the Mandenka ([45]; Tajima's D and Fu's Fs, our unpublished data). An intriguing increased frequency of L0a1 in the Balanta might parallel A1-M31 and A3b2-M13 Y chromosomes in representing East African traces. Although the founder L0a1 haplotype is shared in an east-to-west corridor, the emerging lineages are exclusive of Guineans, indicating a rapid spread and local expansion after arrival. These may therefore reflect the arrival of their ancestors in the Holocene (at about 7 kya, [45]). Moreover, the exact matches found between Balanta and North Africans in haplogroups L2a, L2b and L3b may represent evidence for their contact and long residence in the territory. L3e4 lineages, thought to signal the western expansion of food-production and iron-smelting, show a moderate frequency of 8% in the Balanta. The absence of mtDNA Bantu-markers [46-49] suggests either that Bantu people contributed very little to the maternal gene pool of Guineans, or that they had a different pool from that associated with the southwards migrations [45].

The widespread L3e2b is mainly a Felupe-Djola and Papel cluster with probable links to their homeland mirrored in exact matches with East and Central African haplotypes. Lineages within L3h, coalescing at the late Pleistocene/early Holocene in Guineans [45], exhibit one of the highest found frequencies among the Felupe-Djola (8%). Their increased frequency of West African mtDNA haplogroups L2b and L3d and

13

Y chromosome E1\*-M33 could be due to amplification in small founder groups, as these are absent in East Africa.

The mtDNA haplotypes in Guinean Fulbe exhibit a wide range of matches supported by their wide distribution and massive movements in recent history (e.g. [21]). The high frequency of L1b is otherwise a constant in the Fulbe "world" [50]. Conversely to what is seen on the paternal side, this is the only group that retains statistically significant differences in mtDNA lineages from its ethnic neighbors. As for the Y chromosome, the mtDNA pool of Bijagós shows higher affinity to that of Fulbe, making less likely any connections to the Djola, Papel or Nalú [25].

The North African mtDNA haplogroups demonstrates partial diffusion to Sahel, namely U6 found in Fulbe and Mandenka and M1b present in Guinea-Bissau Atlantic Bak-speakers ([51,52]; previously referred to as M1 in [45]). The U5b1b lineages in Fulbe and Papel are representatives of a link between the Scandinavian Saami and the North African Berbers, emphasizing the great importance of post-glacial expansions [53]. These lineages have most likely crossed the strait of Gibraltar and developed into local clusters, one of which is in West Africa. They do not seem to result from recent gene flow given that the North African Euroasiatic haplogroups H, J and T are absent in our sample.

#### Conclusions

The analysis of our data provides further evidence for the homogeneity of the Y chromosome gene pool of sub-Saharan West Africans, due to the high frequency of haplogroup E3a-M2. Its frequency and diversity in West Africa are among the highest

14

found, suggesting an early local origin and expansion in the last 20-30 ky. Hypothesizing on the existence of an important local agricultural centre, this could have supported a demographic expansion, on an E3a-M2 background, that almost erased the pre-existing Y chromosome diversity. Its pattern of diversity within Mandenka and Balanta hints at a more marked populational growth, these people possibly related to the local diffusion of agricultural expertise. The Papel and Felupe-Djola people retain traces of their East African relatives, to which the short timescale of residence in Guinea-Bissau and higher isolation from major influences have contributed. In the near absence of archaeological data, the signatures of North, Central and East Africans, traceable in less frequent extant paternal haplogroups, fit well with the linguistic and historical evidence regarding the origin and admixture processes of particular ethnic groups. Minor influences of North and East Africa, in particular, are corroborated by mtDNA data.

#### Methods

##### Sampling procedure

A total of 282 Guinea-Bissau unrelated healthy males were analyzed in this survey for the Y chromosome biallelic markers. The sample constitutes a subset of that typed for mtDNA [45] and therefore follows similar selection criteria and DNA extraction procedures. The present data are published as a Cape Verde source population [38] but here samples are described by ethnic affiliation. In the aforementioned article the authors were alerted to slight inconsistencies in Figure 2, which are corrected in the present work (see Figure 2), such as the missing 44 haplogroup E1\*-M33 individuals.

15

Note that discrepancies were not due to sample mistyping but to typographical errors in the original table.

In order to have a manageable number of units with reasonable sample size, many of the Guinean ethnic groups were clustered: Felupe-Djola includes the homonymous group, Baiote, Cassanga and Beafada; Papel includes Papel, Manjaco and Mancanha; Fulbe clusters Fulbe, Futa-Fulbe, Fulbe-Preto and Fulbe-Forro; Mandenka joins Mandenka, Mansonca, and Sussu; Balanta, Bijagós and Nalú were considered independently. The clustering is not without controversy, but follows pertinent information related either to history, anthropology or linguistics [24,54-59].

#### Typing of Y chromosome Binary and Microsatellite Polymorphisms

The hierarchical selection of the following 31 Y chromosome binary markers according to the Y Chromosome Consortium phylogeny [60,61] allowed the inclusion of each Y chromosome into specific haplogroups: YAP [62], 92R7 [63], SRY4064, SRY10831 [64], P25 [65], PN2 [62], M2, M9, M10, M13, M14, M31, M32, M33, M35, M44, M60, M75, M78, M81, M89, M91, M116, M123, M130, M155, M168, M173, M174 and M191 [2,20]. The typing details of restriction fragment length polymorphisms (RFLPs) and direct sequencing analysis are available from the authors. The Wisconsin Package Version 10.0 [66] was employed to align DNA sequences. The nomenclature and phylogenetic relationship of lineages followed the guidelines proposed by the YCC [60], referred to in the text by the (sub)haplogroup and the terminal mutation.

The microsatellite variation, previously determined for a subset of 215 individuals [67] and newly typed for five samples, was associated to the haplogroups. Typing methodology of microsatellites DYS19, DYS389I, DYS389II, DYS390, DYS391,

16

DYS392, DYS393, DYS385, DYS437, DYS438 and DYS439 is published elsewhere [67]. An additional GATA microsatellite A7.1 (DYS460 [68]) was tested for E3b1-M35 chromosomes.

#### Data analysis

A graphical representation of the haplogroup phylogeny and distribution among ethnic clusters was built in netViz 6.5 [69]. Arlequin program ver. 2.000 [70] was used for the summary statistics on both haplogroup and microsatellite haplotype frequencies for each population unit: diversity indexes [71];  $F_{ST}$  and  $R_{ST}$  calculation; exact test of population differentiation (DYS385 omitted from the analysis) [72]; AMOVA tests [73] with hierarchical clustering of the ethnic groups on geographical, linguistic and religious criteria. PCAs were performed with the software MSVP Version 3.13m [74] for haplogroup frequencies of our data and a wide selection of African populations (units as in Figure 4; see Additional files 3 and 4), to generate a more complete picture of the African Y-haplogroup variation and the phylogeographic relationships.

Haplotype networks of microsatellite data were drawn using the Network 4.1.1.2 program [75]. Information on seven microsatellites (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392 and DYS393) was sequentially submitted to reduced-median and median joining algorithms [76,77]. Singletons were excluded from the analysis and the threshold level of 2 was set, with weighted STR loci [78]. The YHRD database and published sources were consulted for exact matches of eight and ten microsatellites (minimal and extended haplotypes, respectively).

17

#### Abbreviations

kya, (kilo) thousand years ago; mtDNA, mitochondrial DNA; nps, nucleotide positions; PCA, Principal Components Analysis; AMOVA, Analysis of Molecular Variance.

#### Authors' contributions

AR conceived the study design and together with CO carried out the molecular genetic typing. AR performed the statistical analysis and interpreted the data to draft the manuscript. MJ, AB and RV have been involved in drafting and revising the manuscript, whose final version was read and approved by all.

#### Acknowledgments

The authors are grateful for permissions to collect blood samples by the Chairman of the Joint Chiefs of Staff and the Ministry of Health of the Republic of Guinea-Bissau. AMI - Assistência Médica Internacional gave local support. The laboratory work has been possible thanks to the technical help of Siiri Roots and Jüri Parik from the Department of Evolutionary Biology, EBC, Estonia and Ana Teresa Fernandes and Rita Gonçalves from the Human Genetics Laboratory, University of Madeira. Alexandra Rosa is beneficiary of the fellowship grant SFRH/BD/12173/2003 from FCT, Fundação para a Ciência e Tecnologia. MAJ is supported by a Wellcome Trust Senior Fellowship in Basic Biomedical Science (057559). António Brehm received a grant from the Regional Government of Madeira (Portugal).

18

#### References

1. Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D, Arredi B, Destro-Bisol G, De Stefano G, Rickards O, Martinez-Labarga C, Modiano D, Biondi G, Moral P, Olckers A, Wallace DC, Novelletto A: **Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations.** *Am J Hum Genet* 1999, **65**:829-846.
2. Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ: **Y chromosome sequence variation and the history of human populations.** *Nat Genet* 2000, **26**:358-361.
3. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli-Sforza LL, Scozzari R, Underhill PA: **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.
4. Pereira L, Gusmao L, Alves C, Amorim A, Prata MJ: **Bantu and European Y-lineages in sub-Saharan Africa.** *Ann Hum Genet* 2002, **66**:369-378.

19

5. Semino O, Santachiara-Benerecetti AS, Falaschi F, Cavalli-Sforza LL, Underhill PA: **Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny.** *Am J Hum Genet* 2002, **70**:265-268.
6. Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL: **African Y chromosome and mtDNA divergence provides insight into the history of click languages.** *Curr Biol* 2003, **13**:464-473.
7. 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.
8. Aïmen H: **Evolution du climat et des civilisations depuis 40000 ans du nord au sud du Sahara occidental, Premières conceptions confrontées aux données récentes.** *Bull L'Assoc Franç L'Étude Quaternaire* 1987, **4**:215-227.
9. Mercader J, Martí R: **The Middle Stone Age occupation of Atlantic central Africa: new evidence from Equatorial Guinea and Cameroon.** In *Under the Canopy*. Edited by Mercader J. New Brunswick: Rutgers University Press; 2003:93-118.
10. Aumassip G, Ferhat N, Heddouche A, Vernet R: **Le milieu saharien aux temps préhistoriques.** In *Milieus, hommes et techniques du Sahara préhistorique. Problèmes actuels*. Edited by Aumassip G et al. Paris: L'Harmattan; 1994:9-29.

20

11. Camps G: *Les Civilisations Préhistorique de l'Afrique du Nord et du Sahara*. Paris, Doin: 1974.
12. Hassan FA: **Archaeological Explorations of the Siwa Oasis Region, Egypt.** *Curr Anthropol* 1978, **19**:146-148.
13. Dutour O, Vernet R, Aumassip G: **Le peuplement préhistorique du Sahara.** In *Milieus, hommes et techniques du Sahara préhistorique. Problèmes actuels*. Edited by Aumassip G et al. Paris: L'Harmattan; 1988:39-52.
14. Clark JD: **Africa: From the appearance of *Homo sapiens sapiens* to the beginnings of food production.** In *Volume I - Prehistory and the Beginnings of Civilization*. Edited by De Laet SJ et al. New York: Routledge; 1994:191-206.
15. Atherton JH: **Excavations at Kamabai and Yagala Rock Shelters, Sierra Leone.** *West Afr J Archaeol* 1972, **2**:39-74.
16. Calvocoressi D, David N: **A new survey of radiocarbon and thermoluminescence dates for West Africa.** *J Afr Hist* 1979, **20**:1-29.
17. Stahl AB: **Reinvestigation of Kintampo 6 Rock Shelter, Ghana: Implications for the Nature of Culture Change.** *Afr Archaeol Rev* 1985, **3**:117-150.
18. Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes*. Princeton: Princeton University Press; 1994.
19. Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer MF, Santachiara-Benerecetti AS: **Different genetic components in the Ethiopian**

21

- population, identified by mtDNA and Y-chromosome polymorphisms.** *Am J Hum Genet* 1998, **62**:420-434.
20. Underhill PA, Passarino G, Lin AA, Shen P, Mirazon Lahr M, Foley R, 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.
  21. Carreira A, Meireles M: **Notas sobre os movimentos migratórios da população natural da Guiné-Portuguesa.** *Bol Cult Guiné Port* 1959, **XIV**:7-20.
  22. Moreira JM: **Os Fulas da Guiné Portuguesa na panorâmica geral do mundo Fula: os Fulas segundo os nossos cronistas.** *Bol Cult Guiné Port* 1964, **XIX**:417-432.
  23. Fage J: *A history of Africa*. London: Routledge; 1995.
  24. Quintino F: **Os povos da Guiné: estrutura social.** *Bol Cult Guiné Port* 1969, **XXIV**:861-915.
  25. Teixeira da Mota A: *Guiné Portuguesa*. Lisboa: Agência Geral do Ultramar; 1954.
  26. Scozzari R, Cruciani F, Santolamazza P, Sellitto D, Cole DE, Rubin LA, Labuda D, Marini E, Succi V, Vona G, Torroni A: **mtDNA and Y chromosome-specific polymorphisms in modern Ojibwa: implications about the origin of their gene pool.** *Am J Hum Genet* 1997, **60**:241-244.

22

27. Shen P, Wang F, Underhill PA, Franco C, Yang WH, Roxas A, Sung R, Lin AA, Hyman RW, Vollrath D, Davis RW, Cavalli-Sforza LL, Oefner PJ: **Population genetic implications from sequence variation in four Y chromosome genes.** *Proc Natl Acad Sci U S A* 2000, **97**:7354-7359.
28. Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, Herrera RJ: **The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations.** *Am J Hum Genet* 2004, **74**:532-544.
29. Jobling MA, Hurles ME, Tyler-Smith C: *Human Evolutionary Genetics - Origins, Peoples & Disease*. New York: Garland Publishing; 2004.
30. **Review and Atlas of Palaeovegetation: Preliminary land ecosystem maps of the world since the Last Glacial Maximum**  
[\[http://www.esd.ornl.gov/ern/gen/adams1.html\]](http://www.esd.ornl.gov/ern/gen/adams1.html)
31. Kwiatkowski DP: **How malaria has affected the human genome and what human genetics can teach us about malaria.** *Am J Hum Genet* 2005, **77**:171-192.
32. Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, Scozzari R, Cruciani F, Destro-Bisol G, Spedini G, Chambers GK, Herrera RJ, Yong KK, Gresham D, Tourné I, Feldman MW, Kalaydjieva L: **The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time.** *Am J Hum Genet* 2004, **74**:50-61.

23

33. Bosch E, Calafell F, Comas D, Oefner PJ, Underhill PA, Bertranpetit J: **High-resolution analysis of human Y-chromosome variation shows a sharp discontinuity and limited gene flow between northwestern Africa and the Iberian Peninsula.** *Am J Hum Genet* 2001, **68**:1019-1029.
34. Scozzari R, Cruciani F, Pangrazio A, Santolamazza P, Vona G, Moral P, Latini V, Varesi L, Memmi MM, Romano V, De Leo G, Gennarelli M, Jaruzelska J, Villemis R, Parik J, Macaulay V, Torroni A: **Human Y-chromosome variation in the western Mediterranean area: implications for the peopling of the region.** *Hum Immunol* 2001, **62**:871-884.
35. Arredi B, Poloni ES, Paracchini S, Zerjal T, Fathallah DM, Makrelouf M, Pascali VL, Novelletto A, Tyler-Smith C: **A predominantly Neolithic origin for Y-chromosomal DNA variation in North Africa.** *Am J Hum Genet* 2004, **75**:338-345.
36. Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, De Benedictis G, Francalacci P, Kouvatsi A, Limborska S, Marcikiae M, Mika A, Mika B, Primorac D, Santachiara-Benerecetti AS, Cavalli-Sforza LL, Underhill PA: **The genetic legacy of Paleolithic *Homo sapiens sapiens* in extant Europeans: a Y chromosome perspective.** *Science* 2000, **290**:1155-1159.
37. Weale ME, Shah T, Jones AL, Greenhalgh J, Wilson JF, Nymadawa P, Zeitlin D, Connell BA, Bradman N, Thomas MG: **Rare deep-rooting Y chromosome lineages in humans: lessons for phylogeography.** *Genetics* 2003, **165**:229-234.

38. Goncalves R, Rosa A, Freitas A, Fernandes A, Kivisild T, Villemis R, Brehm A: **Y-chromosome lineages in Cabo Verde Islands witness the diverse geographic origin of its first male settlers.** *Hum Genet* 2003, **113**:467-472.
39. Cruciani F, La Fratta R, Santolamazza P, Sellitto D, Pascone R, Moral P, Watson E, Guida V, Colomb EB, Zaharova B, Lavinha J, Vona G, Aman R, Cali F, Akar N, Richards M, Torroni A, Novelletto A, Scozzari R: **Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa.** *Am J Hum Genet* 2004, **74**:1014-1022.
40. Alves C, Gusmão L, Barbosa J, Amorim A: **Evaluating the informative power of Y-STRs: a comparative study using European and new African haplotype data.** *Forensic Sci Int* 2003, **134**:126-133.
41. Leat N, Benjeddou M, Davison S: **Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa.** *Forensic Sci Int* 2004, **144**:73-75.
42. Zarrabeitia MT, Riancho JA, Gusmão L, Lareu MV, Sanudo C, Amorim A, Carracedo A: **Spanish population data and forensic usefulness of a novel Y-STR set (DYS437, DYS438, DYS439, DYS460, DYS461, GATA A10, GATA C4, GATA H4).** *Int J Legal Med* 2003, **117**:306-311.
43. **YHRD - Y Chromosome Haplotype Reference Database** [[www.yhrd.org](http://www.yhrd.org)]

44. Arroyo-Pardo E, Gusmão L, Lopez-Parra AM, Baeza C, Mesa MS, Amorim A: **Genetic variability of 16 Y-chromosome STRs in a sample from Equatorial Guinea (Central Africa).** *Forensic Sci Int* 2005, **20**:109-113.
45. Rosa A, Brehm A, Kivisild T, Metspalu E, Villemis R: **MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region.** *Ann Hum Genet* 2004, **68**:340-352.
46. Soodyall H, Vigilant L, Hill AV, Stoneking M, Jenkins T: **mtDNA control-region sequence variation suggests multiple independent origins of an "Asian-specific" 9-bp deletion in sub-Saharan Africans.** *Am J Hum Genet* 1996, **58**:595-608.
47. Watson E, Forster P, Richards M, Bandelt H-J: **Mitochondrial footprints of human expansions in Africa.** *Am J Hum Genet* 1997, **61**:691-704.
48. Bandelt H-J, Alves-Silva J, Guimaraes PE, Santos MS, Brehm A, Pereira L, Coppa A, Larruga JM, Rengo C, Scozzari R, Torroni A, Prata MJ, Amorim A, Prado VF, Pena SD: **Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade.** *Ann Hum Genet* 2001, **65**:549-563.
49. Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A: **Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade.** *Ann Hum Genet* 2001, **65**:439-458.

50. Cerny V, Hajek M, Bromova M, Cmejla R, Diallo I, Brdicka R: **MtDNA of Fulani nomads and their genetic relationships to neighboring sedentary populations.** *Hum Biol* 2006, **78**:9-27.
51. Plaza S, Calafell F, Helal A, Bouzerna N, Lefranc G, Bertranpetit J, Comas D: **Joining the pillars of Hercules: mtDNA sequences show multidirectional gene flow in the western Mediterranean.** *Ann Hum Genet* 2003, **67**:312-328.
52. Olivieri A, Achilli A, Pala M, Battaglia V, Fornarino S, Al-Zahery N, Scozzari R, Cruciani F, Behar DM, Dugoujon JM, Coudray C, Santachiara-Benerecetti AS, Semino O, Bandelt H-J, Torroni A: **The mtDNA legacy of the Levantine early Upper Palaeolithic in Africa.** *Science* 2006, **314**:1767-1770.
53. Achilli A, Rengo C, Battaglia V, Pala M, Olivieri A, Fornarino S, Magri C, Scozzari R, Babudri N, Santachiara-Benerecetti AS, Bandelt H-J, Semino O, Torroni A: **Saami and Berbers - an unexpected mitochondrial DNA link.** *Am J Hum Genet* 2005, **76**:883-886.
54. Almada AA: *Tratado breve dos rios da Guiné do Cabo Verde.* Lisboa: Editorial LIAM; 1964.
55. Carreira A, Quintino FR: *Antroponímia da Guiné Portuguesa.* In *Memórias da Junta de Investigação do Ultramar.* Lisboa: Junta de Investigações do Ultramar; 1964.
56. Hair PE: **Ethnolinguistic continuity on the Guinea Coast.** *J Afr Hist* 1967, **VIII**:247-268.

57. Quintino F: **Os povos da Guiné**. *Bol Cult Guiné Port* 1967, **XXII**:5-40.
58. Diallo T: **Les institutions politiques du Fouta Djallon**. *Initiat Études Afr* 1972, **28**.
59. Lopes C: *Kaabunké, espaço, território e poder na Guiné-Bissau, Gâmbia e Casamance pré-coloniais*. Lisboa: Comissão Nacional para as Comemorações dos Descobrimentos Portugueses; 1999.
60. Y Chromosome Consortium: **A nomenclature system for the tree of human Y-chromosomal binary haplogroups**. *Genome Res* 2002, **12**:339-348.
61. Jobling MA, Tyler-Smith C: **The human Y chromosome: an evolutionary marker comes of age**. *Nat Rev Genet* 2003, **4**:598-612.
62. Hammer MF, Horai S: **Y chromosomal DNA variation and the peopling of Japan**. *Am J Hum Genet* 1995, **56**:951-962.
63. Mathias N, Bayes M, Tyler-Smith C: **Highly informative compound haplotypes for the human Y chromosome**. *Hum Mol Genet* 1994, **3**:115-123.
64. Whitfield LS, Sulston JE, Goodfellow PN: **Sequence variation of the human Y chromosome**. *Nature* 1995, **378**:379-380.
65. Hammer MF, Redd AJ, Wood ET, Bonner MR, Jarjanazi H, Karafet T, Santachiara-Benerecetti AS, Oppenheim A, Jobling MA, Jenkins T, Ostrer H, Bonne-Tamir B: **Jewish and Middle Eastern non-Jewish populations share a**

28

- common pool of Y-chromosome biallelic haplotypes**. *Proc Natl Acad Sci U S A* 2000, **97**:6769-6774.
66. **Wisconsin Package Version 10.0**. Genetics Computer Group. 2005.
67. Rosa A, Ornelas C, Brehm A, Villems R: **Population data on 11 Y-chromosome STRs from Guiné-Bissau**. *Forensic Sci Int* 2006, **157**:210-217.
68. White PS, Tatum OL, Deaven LL, Longmire JL: **New, male-specific microsatellite markers from the human Y chromosome**. *Genomics* 1999, **57**:433-437.
69. **NetViz 6.5 - NetViz LCC Corporation**. 2002. [[www.netviz.com](http://www.netviz.com)].
70. **Arlequin version 2.000: a software for population genetics data analysis**. Schneider S, Roessli D, Excoffier L. 2000. [<http://anthropologie.unige.ch/arlequin>].
71. Nei M: *Molecular Evolutionary Genetics*. New York: Columbia University Press; 1987.
72. Raymond C, Rousset F: **An exact test for population differentiation**. *Evolution* 1995, **49**:1280-1283.
73. Excoffier L, Smouse PE, Quattro JM: **Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data**. *Genetics* 1992, **131**:479-491.

29

74. **MVSP - A multi-variate statistical package for Windows ver 3.13m**. Kovach WL, Kovach Computing Services. 2004. [[www.kovcomp.co.uk/mvsp/index.html](http://www.kovcomp.co.uk/mvsp/index.html)].
75. **Network 4.1.1.2 - Fluxus Technology Ltd**. 2004. [[www.fluxus-engineering.com](http://www.fluxus-engineering.com)].
76. Bandelt H-J, Forster P, Sykes BC, Richards MB: **Mitochondrial portraits of human populations using median networks**. *Genetics* 1995, **141**:743-753.
77. Bandelt H-J, Forster P, Röhl A: **Median-joining networks for inferring intraspecific phylogenies**. *Mol Biol Evol* 1999, **16**:37-48.
78. Helgason A, Sigurdardottir S, Nicholson J, Sykes B, Hill EW, Bradley DG, Bosnes V, Gulcher JR, Ward R, Stefansson K: **Estimating Scandinavian and Gaelic ancestry in the male settlers of Iceland**. *Am J Hum Genet* 2000, **67**:697-717.
79. Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti AS, Soodyall H, Zegura SL: **Hierarchical patterns of global human Y-chromosome diversity**. *Mol Biol Evol* 2001, **18**:1189-1203.
80. Semino O, Magri C, Benuzzi G, Lin AA, Al-Zahery N, Battaglia V, Maccioni L, Triantaphyllidis C, Shen P, Oefner PJ, Zhivotovsky LA, King R, Torroni A, Cavalli-Sforza LL, Underhill PA, Santachiara-Benerecetti AS: **Origin, diffusion, and differentiation of Y-chromosome haplogroups E and J: inferences on the Neolithization of Europe and later migratory events in the Mediterranean area**. *Am J Hum Genet* 2004, **74**:1023-1034.

30

## Figure legends

**Figure 1 - Geographic location of Guinea-Bissau and present-day settlement pattern of the ethnic groups considered in this study.**

**Figure 2 - Y chromosome haplogroup diversity in Guinea-Bissau.** Absolute numbers are shown for the total sample and ethnical clusters. Haplogroup nomenclature and defining mutations assayed in this study, shown along the branches of the phylogeny, are as proposed by the YCC [60]. The bold link indicates the root, determined by comparisons with primates [2,79].

**Figure 3 – African spatial distribution of haplogroup E3a-M2.** Frequency scale (in percentage) is shown on the left. Data according to population datasets described in Additional files 3 and 4.

**Figure 4 – Principal Component Analysis for a) several African populations and b) Guinea-Bissau ethnic clusters, based on haplogroup frequencies.**

**a)** The 1<sup>st</sup> PC captures 42.6% of the variance and 16.9% are under the responsibility of the 2<sup>nd</sup> PC. For details on populational datasets see Additional file 2. The codes in italic refer to the following populations: Morocco Arabs: *Ar* [1,34], *Mar* [33]; Morocco Berbers: *Bb* [33], *MBb* [34]; Algeria: *Alg* [80], *Aar*-Algerian Arabs [35]; Tunisia- *Tun1* [35], *Tun2* [7]; West Sahara: *Sah*-Saharawis [33]; Egypt: *Egy1* [35], *Egy2* [7]; Sudan: *Sud* [2]; Ethiopia: *Eth* [2], *Or*-Oromo, *Amh*-Amhara [5,7]; Kenya: *K&K*-Kikuu & Kamba, *Maa*-Maasai [7]; Uganda: *Gan*-Ganda [7]; North Cameroon: *Po*-Podokwo, *Mad*-Mandara [7], *Ou*-Ouldeme, *Daba* [1,7,26], *NCA*adaw-Fali, *Tali* [1,26], *Fca*-Fulbe [1,26]; South Cameroon:

31

*SCBantu*-Bassa, Ngoumba [7], *Bak-Bakaka* [1,7], *Bam-Bamileke* [1,26], *Ewo-Ewondo* [1,26], *Bko-Bakola* Pygmies [7]; CAR: *Bik-Biaka* Pygmies [2,7]; DRC: *DRCBantu*-Nande, Hema [7]; *Mb-Mbuti* Pygmies [2,7]; Guinea-Bissau: *EJA*-Felupe-Djola, *BJG*-Bijagós, *BLE*- Balanta, *PBO*-Papel, *FUL*-Fulbe, *MNK*-Mandenka, *NAJ*-Nalú (Present study); Burkina Faso: *Mo*-Mossi [1,26], *Ri*-Rimaibe [1,26], *FBF*-Fulbe [1,26]; Gambia/Senegal: *Wo*-Wolof [7], *Mak*-Mandinka [7]; Mali: *Mal* [2], *Do*-Dogon [7]; Ghana: *Ewe*, *Ga*, *Fan*-Fante [7]; Senegal: *Se* [5]; Namibia: *Her*-Herero, *Amb*-Ambo [7], *Ku*-!Kung, Sekele [1,7,26], *CKh*-Tsumkwe San, Dama, Nama [7]; South Africa: *ST*-Sotho-Tswana, *Zu*-Zulu, *Xh*-Xhosa, *Sh*-Shona [7], *Kho*-Khoisan [2].

b) The PCA captures 87.0% of the variance with 74.0% and 13.0% attributed to the 1<sup>st</sup> and 2<sup>nd</sup> PC, respectively. The 1<sup>st</sup> PC reflects an axial proportion of E3a\* vs. E1\* where Papel and Felupe-Djola retain the higher proportions of the later. E3a\* is again a main influence in the 2<sup>nd</sup> axis against that of R1b and E3b1, placing Mandenka apart from Bijagós and Fulbe.

### List of additional files

#### Additional file 1

**File format:** EPS

**Title:** Population data on the surveyed ethnic groups of Guinea-Bissau.

**Description:** The table provides information on the linguistic and religious affiliations of the Guinea-Bissau ethnic groups.

#### Additional file 2

32

**File format:** EPS

**Title:** Diversity indices and TMRCA estimates.

**Description:** The table gives diversity indices for Guinea-Bissau Y chromosome haplogroups: a) coalescence time estimates for haplogroups and b) molecular diversity index ( $R_{ST}$ ) and TMRCA for haplogroup E3a\*-M2, by ethnic group.

#### Additional file 3

**File format:** EPS

**Title:** Comparative African data.

**Description:** The table summarizes previously published Y chromosome datasets on African populations, here considered for comparative purposes.

#### Additional file 4

**File format:** EPS

**Title:** Geographical distribution of African samples.

**Description:** The figure displays the geographical distribution of the African Y chromosome samples considered for comparative purposes [see Additional file 3].

#### Additional file 5

**File format:** EPS

**Title:** Analysis of Molecular Variance (AMOVA) in Guinea-Bissau

**Description:** The table summarizes the results of an AMOVA analysis (1023 permutations) for the Y chromosome variation among Guinean ethnic groups, clustered according to geographical, linguistic and religious criteria.

33

#### Additional file 6

**File format:** EPS

**Title:** Microsatellite haplotype networks.

**Description:** The networks describe the variability of 7 microsatellite loci in Y chromosome haplogroups, among ethnic groups. a) haplogroup E3a\*-M2 (N=75, singletons excluded); b) haplogroup E3b1-M78 (N=11), "\*" denoting the E3b1-β haplotypes. Node sizes are proportional to the number of individuals.

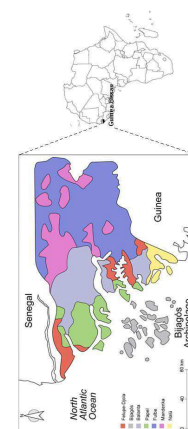
#### Additional file 7

**File format:** PDF

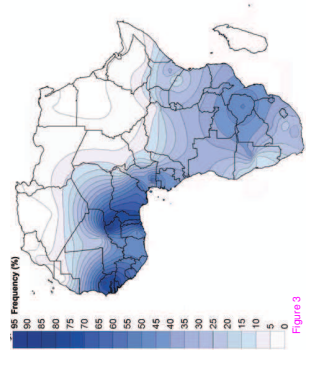
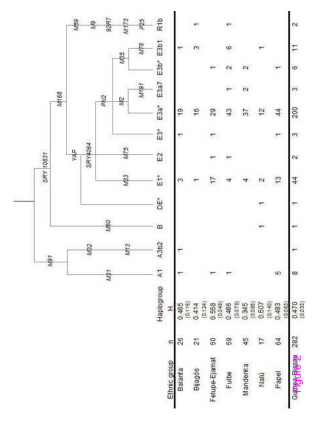
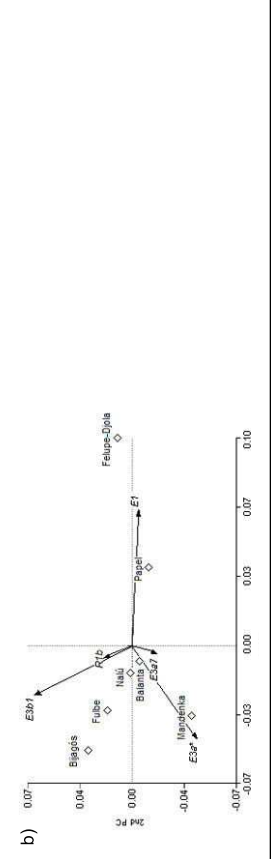
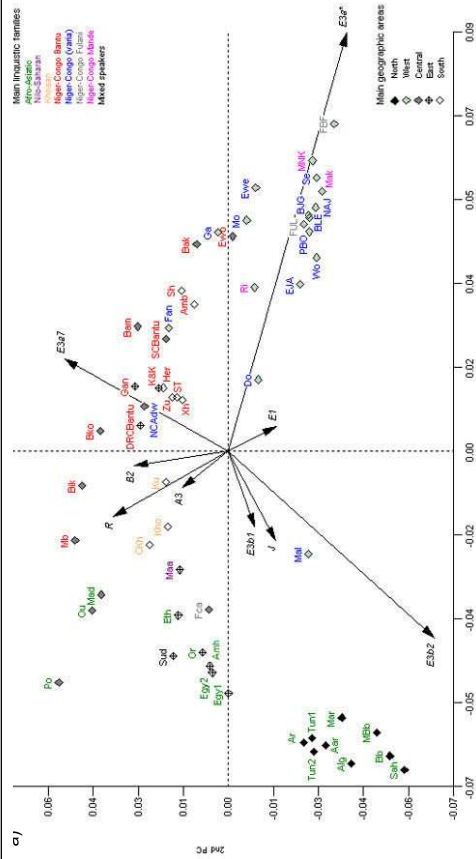
**Title:** Haplotypes in Guinean samples.

**Description:** List of the Y chromosome SNP-defined haplogroups and corresponding microsatellite haplotypes found in the Guinean sample set, by ethnic group.

34







**Additional files provided with this submission:**

Additional file 1: additional file 1.eps, 1484K  
[http://www.biomedcentral.com/imedia/2117185668144073/supp1\\_eps](http://www.biomedcentral.com/imedia/2117185668144073/supp1_eps)  
 Additional file 2: additional file 2.eps, 2959K  
[http://www.biomedcentral.com/imedia/1043891793153012/supp2\\_eps](http://www.biomedcentral.com/imedia/1043891793153012/supp2_eps)  
 Additional file 3: additional file 3.eps, 6207K  
[http://www.biomedcentral.com/imedia/1026045838144073/supp3\\_eps](http://www.biomedcentral.com/imedia/1026045838144073/supp3_eps)  
 Additional file 4: additional file 4.eps, 4014K  
[http://www.biomedcentral.com/imedia/937807491530137/supp4\\_eps](http://www.biomedcentral.com/imedia/937807491530137/supp4_eps)  
 Additional file 5: additional file 5.eps, 3811K  
[http://www.biomedcentral.com/imedia/844784223144073/supp5\\_eps](http://www.biomedcentral.com/imedia/844784223144073/supp5_eps)  
 Additional file 6: additional file 6.eps, 2231K  
[http://www.biomedcentral.com/imedia/1258096816144073/supp6\\_eps](http://www.biomedcentral.com/imedia/1258096816144073/supp6_eps)  
 Additional file 7: additional file 7.pdf, 41K  
[http://www.biomedcentral.com/imedia/6829754801530128/supp7\\_pdf](http://www.biomedcentral.com/imedia/6829754801530128/supp7_pdf)

Figure 3