GENES, PEOPLES AND LANGUAGES IN CENTRAL AFRICA

Gemma Berniell Lee

TESI DOCTORAL UPF/2010

DIRECTOR DE LA TESI

Dr. David Comas

Departament de Ciències Experimentals i de la Salud

Dedication

To my grandfather, to my mother, to Alex.

Acknowledgements

I would like to thank Dr. Jaume Bertranpetit for having given me the opportunity to carry out this PhD project under his supervision and become a member of his internationally renowed scientific research team. I would also like to thank Dr. David Comas for his valuable help and support (and faith!) throughout the development of the thesis. Likewise, I would like to thank all those people who were there for me while I was doing my PhD (all my lab colleagues and scientists) and those who are with me now.

Abstract

The present thesis titled " Genes, peoples and languages in Central Africa" examines the genetic diversity patterns in populations from west central Africa, more specifically, in Bantu and Pygmy populations from Gabon and Cameroon, two key areas in the understanding of the Bantu expansion. More than 800 samples have been analysed at the Y chromosome level in order to genetically characterise these populations and establish the genetic relationship between them. The results have shown that the Bantu expansion largely homogenised the gene pool of Bantu populations, erasing the pre-Bantu diversity, while it diversified that of Pygmy groups, introducing Bantu lineages into their gene pool. Furthermore, gene flow of paternal lineages seems to have taken place mainly in one direction; from Bantus to Pygmies. These results contrast with those found in studies of maternal (mtDNA) lineages in these areas, where considerable gene flow from Pygmy to Bantu populations have been observed, suggesting possible sex-biased admixtures rates between Bantu and Pygmy populations. An interesting finding, is the significant presence of a non-African lineage in these sub-Saharan populations.

Resumen

La presente tesis, titulada “Genes, peoples and languages in Central Africa”, examina los patrones de diversidad genética en poblaciones del oeste de Africa central, más específicamente, poblaciones Bantús y Pigmeas de Gabon y Camerún, dos zonas vitales para la comprensión de la expansión Bantú. Se han analizado más de 800 muestras a nivel del cromosoma Y con el fin de caracterizar genéticamente a estas poblaciones, y establecer la relación genética entre ellas. Los resultados han demostrado que la expansión Bantú homogeneizó el acervo genético de las poblaciones Bantús, eliminando la diversidad pre-Bantú, mientras que diversificó aquel de las poblaciones Pigmeas, introduciendo linajes Bantus. Además, se ha visto que el flujo de linajes paternos parece haber tenido una única dirección: de Bantus a Pigmeos. Estos resultados contrastan con aquellos obtenidos para linajes maternos (DNA mitocondrial) en estas zonas, donde se ha observado un considerable flujo genético de Pigmeos a Bantus, sugiriendo un posible sesgo sexual en la tasa de mestizaje entre poblaciones Bantus y Pigmeas. Un hallazgo interesante es la presencia de un linaje no-africano en estas poblaciones de África subsahariana.

Preface

Among the several demographic processes that have shaped the genetic landscape of Africa, the Bantu expansion is undoubtedly one of the most influential. Good proof of this is the current, practically exclusive, widespread distribution of Bantu populations throughout sub-Saharan Africa, an area that prior to the Bantu expansion was home to hundreds of hunter-gatherer populations. Although many genetic studies have been carried out on populations from sub-Saharan Africa few have focused on Central Africa, an area of vital importance for the understanding of this vast population movement.

The present thesis is based on the genetic analysis of paternal (Y chromosome) lineages in 22 Bantu and 3 Pygmy populations from the west central African regions of Gabon and Cameroon with the aims of determining the genetic structure of Pygmy and Bantu populations, the extent and symmetry of gene flow between them, and a possible existance of ancestral paternal lineages prior to the Bantu expansion.

Contents

1. INTRODUCTION

1.1. Human genetic variation

a) Before and after the human genome project

Until the sequencing of the human genome became a reality, back in the year 2003, our knowledge on the genetic variation present in human populations was fairly limited, being restricted to the analysis of only a few DNA polymorphisms (e.g. Short Tandem Repeats (STRs) and Variable Number of Tandem Repeats (VNTRs)) and the information from uniparental genomes (Y chromosome and mitochondrial DNA (mtDNA)). With this remarkable initiative, came the discovery of hundreds of Single Nucleotide Polymorphisms (SNPs), genetic changes at a specific site interspersed throughout the genome´s three billion nucleotides. Since then, and with the current sequencing and typing techniques available, many large-scale genome-wide studies have provided, and continue to provide, new insights into the nature of diseases, drug response and human evolution. The International HapMap Project for example, which analysed blocks of DNA containing numerous SNPs inherited initially in 270 individuals from four geographical populations (HapMap I), and subsequently extended to eleven populations (HapMap 3), paved the path towards a "catalogue" of human genetic variation that has contributed and is still contributing to our understanding of the structure of human genetic variation and its effect on human health.

Large arrays of genome-wide association studies using SNPs have been conducted for scores of human diseases and have identified

hundreds of genes and genetic variants that contribute to common diseases such as cancer (Stacey et al. 2006), cholesterol and triglycerides (Kathiresan et al. 2008), Alzheimer´s (Yucesoy et al. 2006), Parkinson´s (Edwards et al. 2010) and benign breast disease (Jorgensen et al. 2009), as well as to the response to therapeutic drugs (Seo et al. 2005; Goodman et al. 2008; Varenhorst et al. 2009), opening new frontiers in the treatment of human disease. In the same way, multiple genome-wide genomic studies using a variety of unlinked loci, have produced detailed descriptions of genetic diversity patterns of human populations (both within and among populations), both at the global level (Auton et al. 2009) and for individual geographic areas (Rosenberg et al. 2002; Lao et al. 2008), providing insights into human history and evolution, the evolutionary forces acting on human populations and the relationships among these.

Taking into account that projects such as the first comprehensive map of copy number variation (CNV) in the human genome (based on the study of 270 individuals from four HapMap populations), the first high-resolution sequence map of human structural variation, and the 1,000 Genomes Project have already started to offer important insights into genetic population structure and human disease, there is no doubt that more refined fine-scaled maps of human genetic diversity that will further our understanding of the nature of the human species are soon to come…

b) Types of human genetic variation

Genetic variation in the human genome can occur at any genetic level (e.g. at the DNA sequence level, at the chromosome or protein level etc.) and takes many forms, ranging from large, microscopically visible chromosome anomalies, to changes in a single nucleotide (Redon et al. 2006).

Until quite recently, it was thought SNPs, variations of a single nucleotide, were the most frequent form of genomic variation in the human genome, constituting up to 90% of the total variation (Sachidanandam et al. 2001; The International HapMap Consortium 2005). However, the discovery of the widespread presence of Copy Number Variations (CNVs), structural alterations such as duplications, inversions and rearrangements of large (1 kilobase (kb) in length and over) segments of DNA at a variable copy number throughout the human genome (Iafrate et al. 2004; McCarroll et al. 2006; Repping et al. 2006), has shown that this kind of polymorphism is also a major source of human genetic diversity (Freeman et al. 2006), to such an extent, that it is currently thought that CNVs affect more nucleotides than do SNPs (Redon et al. 2006).

The most common types of polymorphisms used in evolutionary studies are:

- *SNPs*

Also known as biallelic markers, because they usually only have two possible alleles (frequencies >1%) (http://www.promega.com/geneticidproc/ussymp9proc/content/10.p

df), SNPs occur approximately every 1-2 kilobases (kb), with more than 6 million SNPs having being been identified throughout the genome (http://www.ncbi.nlm.nih.gov/SNP/snp summary.cgi). SNPs are thought to have occurred only once in human history (unique events) and are characterized by having a low mutation rate ($2x10^{-8}$ per generation (Nachman and Crowell 2000)); two properties that make them especially useful for studying early demographic events.

- *Insertions/deletions*

Insertions/deletions (indels) represent around 8% of all human polymorphisms (Weber et al. 2002), and can be as small as one nucleotide in length up to several million nucleotides long (Lupski et al. 1996). Their stability - unlikely to undergo reverse or convergent mutation - makes them useful markers to distinguish chromosomal lineages that are identical by descent (Tishkoff and Verrelli 2003).

- *Alu insertions*

Alu insertions are the most common type of retrotransposable elements in the human genome. There are approximately one million copies (Batzer et al. 2002), comprising around 10% of the genome (Smit 1996). Given their length, around 300 base pairs (bp), they are classified as repetitive DNA elements belonging to the Short Interspersed Nuclear Elements (SINEs) type. The fact that the ancestral state of Alu insertions is known (absence or presence of Alu insertion) (Batzer and Deininger 1991; Batzer et al. 1994), makes them highly useful for the reconstruction of human demographic history and for the detection of migration and

populational events (Bamshad et al. 2003; Hammer et al. 1994; Jorde et al. 2000; Nasidze et al 2001; Romualdi et al. 2002; Sherry et al. 1997; Stoneking et al. 1997).

- *Satellite DNA*

There are two main families of satellite DNA or tandemly repeated DNA; microsatellites or STRs, and minisatellites or VNTRs. Microsatellites are less than 10 bp long, mostly 2-6 bp in length (Nakamura et al. 1987; Charlesworth et al. 1994; Chambers and Mac Avoy 2000), and minisatellites are longer tandem repeats, from 10-100 bp in length. Whilst STRs are widely used in evolutionary studies, minisatellites are minisatellites have only been used in some studies. It has been shown that strand slippage during replication and mutations via a complex meiotic recombination pathway give rise to a different number of repeats in STRs (Tishkoff 2003), leading to variation in length between individuals that acts as alleles. STRs have a much higher mutation rate (2.0×10^{-3} per generation (Weber and Wong 1993)) than SNPs, making them ideal markers to investigate demographic events that have taken place more recently.

c) Forces shaping human genetic variation

The magnitude and patterns of genetic variation vary across the genome, across individuals, across populations and across species, given that their precise genealogy is not the same. The more closely linked the regions are, the individuals are, the populations are and the species are, the more similar their genealogy is.

Genealogy is greatly influenced by natural on-going processes such as mutation, migration, recombination, drift and natural selection, and by characteristics of the species and historical events – all of which can increase/decrease/change genetic diversity e.g. selection can rapidly increase the frequency of a genetic variant or nearby regions (hitch-hiking effect) (Maynard Smith and Haigh 1974) or reduce local genetic diversity. Because these events can have an effect at the genomic, individual, populational and species level, patterns of genetic diversity also vary from region to region, individual to individual, population to population and from species to species. Furthermore, all these processes are not constant in time or space, so patterns of genetic diversity can also vary over a range of time and specific locations. Therefore, the patterns of genetic diversity we observe in a given genomic region, individual, population or species are the result of many demographic and evolutionary events acting on different timescales. In the human species, we have clear examples of how events that took place in the past have had a major impact on modern genetic diversity (e.g. colonisations, population expansions etc. (Jobling et al. 2004)). The Out-of-Africa origin for modern humans, for example, is probably one of the most representative examples of this fact.

Because of the differences in genealogy mentioned above, the analysis of different kinds of genetic data may yield different patterns of genetic diversity as a reflection of the environmental and biological events acting at one moment in time, with values becoming increasingly similar the more closely linked the regions are (Weir et al. 2005).

Genetic polymorphisms are created at random by mutation and can appear both in coding (functional) and non-coding (non-functional) regions. Whilst functional polymorphisms can lead to major biological alterations causing phenotypic variation or disease (Yamaguchi-Kabata et al. 2008), in principle, non-functional polymorphisms do not cause any alterations that impair gene function, with neutral loci being less likely to be subjected to selective forces such as natural selection than non-neutral loci. For this reason, functional polymorphisms are especially useful as genetic markers for medical diagnosis and genome mapping studies and neutral markers, such as the non-recombining portions of the genome, are especially efficient at detecting population substructure.

d) Patterns of human genetic variation

Humans are around 98.8% identical to chimpanzees at the nucleotide level (Britten et al. 2002; Fujiyama et al. 2002; Olson et al. 2003) and even more similar to each other. However, the small fraction of genetic material remaining that differs among people confers individuality, and is responsible for differences in appearance, susceptibility to diseases, and response to drugs and environmental factors.

A wide range of studies have shown that not all human populations show the same level of genetic variation, with 85 – 90% of the genetic variation being due to differences between individuals within populations (Lewontin 1972; Barbujani et al. 1997; Jorde et al. 2000; Watkins et al. 2003; International HapMap Consortium 2005; Rosenberg et al. 2005), and the remaining 10-15 % being due to differences among populations. However, it has also been established that individuals from the same population tend to be more closely related and more genetically similar (Powell and Taylor 1978), and that it is possible to accurately assign humans to populations (Rosenberg et al. 2002, 2005; Bamshad et al. 2003; Turakulov and Easteal 2003; Lao et al. 2006) even if they are of admixed origin (Bamshad et al. 2003). Furthermore, estimates of genetic differentiation e.g. the F_{ST} statistic (that varies from 0 (undifferentiated) to 1 (population-specific) (Weir and Hill 2002)), have shown that human populations show geographic clustering across all major continents.

The influence of geography on patterns of genetic variation was first identified in the first half of the 20th century (Wright et al.

1943). This evidence came from the correlation between the frequency of human blood group polymorphisms and geographic location (Mourant et al. 1976; Lewontin et al. 1972). Subsequent studies showed a strong correlation between genetic differentiation and geographic distance; the larger the geographic distance between populations, the larger the genetic distance between them (both at the continental and global scale (Cavalli-Sforza 1994; Cavalli-Sforza and Feldman 2003). This correlation was later re-confirmed by several subsequent studies (Ramachandran et al. 2005; Manica et al. 2005; Linz et al. 2007; Relethford 2004; Conrad et al. 2006) that have shown that geographic distance explains at least 75% of the variance between human populations (Ramachandran et al. 2005; Manica et al. 2005; Linz et al. 2007).

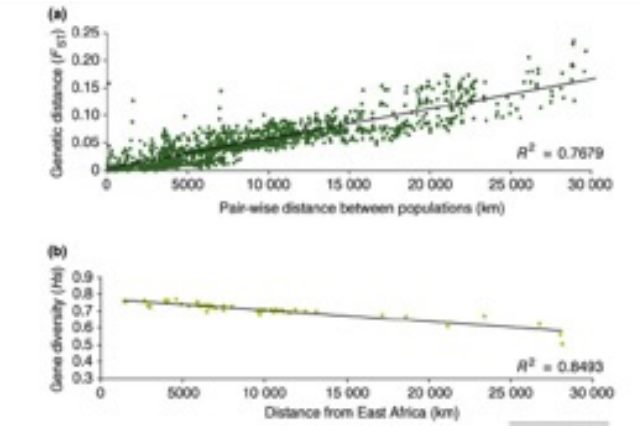


Figure 1. Effect of geography on genetic diversity patterns (from Lawson Handley et al. 2007). (a) F_{ST} distances between populations in the HGDP-CEPH cell line plotted against geographic distance. (b) Gene diversity (Hs) within the HGDP-CEPH populations plotted against geographic distance from East Africa.

The geographic distribution of human genetic diversity has been studied in depth with a wide variety of genetic polymorphisms in diverse human populations; classical protein polymorphisms (Corbo et al. 1999; Cavalli-Sforza et al. 1994), autosomal STRs (Deka et al. 1995; Bowcock et al. 1994; Rosenberg et al. 2002; Jorde et al. 2000; Jorde et al. 1997), autosomal SNPs (Shriver et al. 2005; Chen et al. 1995); mtDNA haplotypes (Seielstad et al. 1998; Jorde et al. 2000; Torroni et al. 2001; Chen et al. 2000), Y chromosome STRs and SNPs (Forster et al. 2000; Underhill et al. 2000; Hammer et al. 2001; Seielstad et al. 1998; Jorde et al. 2000; Kayser et al. 2001), non-STR, non-repetitive element indels (Weber et al. 2002), Alu insertion polymorphisms (Watkins et al. 2001, 2003; Hedges and Batzer 2005) and a few LINE-1 (L1) insertion polymorphisms (Sheen et al. 2000). Although these studies have offered slightly different estimates of genetic differentiation across human populations depending on the type of marker and population analysed - mtDNA and Y chromosome markers show higher F_{ST} values than autosomal markers: 0.24-0.27 (Seiestad et al. 1998; Jorde et al. 2000) and 0.23-0.64 (Seiestad et al. 1998; Jorde et al 2000; Kayser et al. 2001; Romualdi et al. 2002), respectively, versus 0.09-0.16 (Barbujani et al. 1997; Jorde et al. 2000; Deka et al. 1995; Bowcock et al. 1994), the degree of agreement across these studies is remarkable.

e) The Out-of-Africa model

Many studies have shown that, in general, genetic diversity is highest in Africa. Studies with mtDNA (Cann et al. 1987; Vigilant et al.1991; Ingman et al. 2000), Y chromosome (Seielstad et al. 1999; Hammer et al. 2001; Underhill et al. 2001) and autosomal markers (Watkins et al. 2001; Cavalli-Sforza et al.1994; Deka et al. 1995; Weber et al. 2002; Calafell et al. 1998) have shown that African populations are more diverse than non-African populations and some studies have even shown that allele diversity outside Africa is a subset of that found within Africa (Armour et al. 1996; Calafell et al. 1998; Kivisild et al. 1999; Yu et al. 2002), pointing towards an African origin for the human species. Although there are a few exceptions to this pattern of higher genetic diversity in Africa (Mountain and Cavalli-Sforza 1994; Rogers et al. 1995; Eller et al. 1999) these are mainly due to ascertainment bias, with the Recent African Origin model (RAO) being nowadays widely accepted. Also known as the Out-of-Africa model, the RAO claims that anatomically modern humans originated in East Africa around 200,000 years ago and then migrated out of Africa, spreading across the rest of the globe within the last 100,000 years (Campbell and Tishkoff 2008; Walter et al. 2000), replacing all non-African archaic humans (Manica et al. 2007)(a south-western African origin, near the coastal border of Namibia and Angola, has also been hypothesised (Tishkoff et al. 2009)). According to the model, modern humans lived in Africa a lot longer than in any other geographic region and maintained relatively large effective population sizes, resulting in high levels of within population

genetic diversity within Africa (Campbell and Tishkoff 2008; Reed and Tishkoff 2006). Also, the exit from Africa seems to have been associated to severe or repeated bottlenecks, whereby drift vastly reduced the genetic diversity of the populations leaving Africa, to such an extent, that 85% of genetic diversity within human populations decreases with geographic distance from East Africa (Ramachandran et al. 2005; Prugnolle et al. 2005; Linz et al. 2007).

1.2. Mitochondrial DNA

a) Evolution

It is thought that mtDNA originated from the circular genomes of endosymbiotic bacteria that were engulfed by proto-eukaryotic cells around 1.5 billion years ago. The mitochondrial genome is a double-stranded circular molecule made up of 16,569 bp that with time has lost most of its genes, currently only coding for 13 sub-units of the oxidative phosphorylation system within the mitochondria (2 ribosomal RNAs and 22 transfer RNAs (Wallace et al. 1999). Hence, most of the functional proteins of the mitochondria are coded by nuclear genes. MtDNA is highly abundant in the cell, with each mitochondrion containing around 2-10 copies of mtDNA, and each somatic cell containing ~ 1000 mitochondria.

b) Structure

The entire sequence of mtDNA was first published at the beginning of the 1980´s (Anderson et al. 1981) and was since known as the Cambridge Reference Sequence (CRS), being later reviewed by Andrews et al. 1999. The base content of its two strands is biased, with the purine-rich and the pyrimidine-rich strands named "heavy" and "light", respectively (Chinnery et al. 2006). In addition to its coding region, the mtDNA molecule only contains a small fragment of non-coding DNA known as the Control Region (CR) or D-loop, 1125 bp long. Like its name denotes, the CR is involved in the regulation of the transcription and replication of the molecule; containing the origin of replication, together with the main regulatory elements involved in these events. Already presenting a much higher mutation rate than the coding region, the CR contains two 350 bp long hypervariable regions, hypervariable region I (HVR I) and hypervariable region II (HVR II), that are even more variable.

c) Inheritance

MtDNA is inherited via the maternal line and, as is characteristic of a haploid system, in principle, it does not undergo recombination (Horai et al. 1995; Jorde et al. 1998), because despite sperm mitochondria entering the oocyte during fertilization, they are somehow degraded. Although the possibility of recombination taking place in mtDNA has been proposed by several authors (Eyre-Walker et al. 1999; Awadalla et al. 1999), which has lead to certain

controversy, paternal inheritance of mtDNA has only been truly observed once (Schwartz and Vissing 2002), with the rest of the cases having been ruled out due to insufficient evidence.

d) Selection

It is generally assumed that the mtDNA molecule is selectively neutral. However, some studies showing an excess of non-synonymous substitutions in humans, pointing towards purifying selection acting on mtDNA (Nachman et al. 1996) have raised certain controversy. Although other association studies have also suggested a possible link between specific mtDNA types and certain diseases, there is no conclusive evidence on this subject for the moment, with the general acceptance that even if purifying selection has been acting on human mtDNA, it does not seem to have modified the topology of its phylogeny.

1.3. Y Chromosome DNA

a) Evolution

The Y chromosome is the second smallest chromosome, with an average size of 60 megabases (Mb). It was originally thought that the Y chromosome was a profoundly degenerated X chromosome, with very few genes on it encoding male-specific features (Ohno et al. 1960). It was later discovered that, in fact, both the X and the Y chromosome originated from the same ancestral pair of identical autosomes 300 million years ago (mya) (Jegalian and Page 1999; Lahn and Page 1999) - shortly after the divergence of the

evolutionary lines leading to mammals and birds - and that they differentiated into their distinctive X and Y forms when the appearance of the Sex determining Region Y (SRY), that codes for the testes determining factor and multiple inversions on the Y chromosome, repressed X-Y recombination (Lahn and Page, 1999). Each inversion drove the sex chromosomes farther apart as they evolved, increasing the amount of DNA that could no longer align and recombine. Whereas the X chromosome continued to benefit from crossing over in females, maintaining stable gene function, it is thought the Y chromosome developed mechanisms to overcome the degradation of genes by the lack of X-Y crossing over. One of these mechanisms that has been thought to preserve gene function over evolutionary time is gene conversion (Rozen et al. 2003).

b) Structure

The Y chromosome is composed of a large non-recombining region NRY (NRY) that comprises 95% of its length, and two pseudo-autosomal regions that flank the NRY on either side (pseudo-autosomal regions 1 and 2 (2.6 Mb and 0.32 Mb in length respectively)), where X-Y crossing over is normal and frequent in male meiosis (Simmler et al. 1985; Cooke et al. 1985; Freije et al. 1992). Following the discovery of multiple processes occurring within the Y chromosome e.g. non-homologous recombination or gene conversion events (Skaletsky et al. 2003), the NRY region was recently renamed MSY or male-specific region (Rozen et al. 2003; Skaletsky et al. 2003). The MSY contains a heterochromatic region of variable size (can comprise up to 40 Mb) consisting of at least six

different types of sequence classes that form long homogeneous tandem arrays (Skaletsky et al. 2003) and an euchromatic region of around 23 Mb (8 Mb on the short arm (Yp) and 14.5 Mb on the long arm (Yq)), consisting of a mosaic of complex and interrelated sequences (Skaletsky et al. 2003). The euchromatic region contains 156 transcription units that include 78 protein coding genes that encode 27 distinct proteins/protein families, 12 of which are expressed ubiquitously throughout the body, 11 of which are mostly or exclusively expressed in the testis, and 4 of which are expressed in other tissues (Skaletsky et al. 2003). Euchromatic sequences can be classified into three groups: X - transposed, X - degenerate and ampliconic.

- X-transposed

Comprising 10-15% of the MSY (3.4 Mb), as their name indicates, these sequences are the result of a large X-Y transposition that took place around 3-4 million years ago after the divergence of the human and chimpanzee lineages (Rozen et al. 2003), still exhibiting 99% identity to their X chromosome counterparts. Of the three groups of euchromatic sequences, X-transposed sequences contain the lowest density of genes, as well as the highest density of interspersed repeated elements (Skaletsky et al. 2003).

- X-degenerate

Comprising 20% of the MSY, these sequences are surviving remnants of ancient autosomes from which the modern X and Y

chromosomes co-evolved (Lahn and Page et al. 1999), exhibiting between 60 - 96% identity to their X-linked homologues.

- Ampliconic

Comprising about 30% of the MSY (10.2 Mb), these sequences (also known as the segmentally duplicated portion of NRY (Hurles and Jobling 2003)), that are located in seven segments scattered across the long and proximal short arms of the Y chromosome include large regions that show 99.9% identity that is maintained by frequent gene conversion (Skaletsky et al. 2003). The most prominent feature are eight massive palindromes located in the ampliconic regions of the long arm of the Y chromosome (Yq) that comprise around one quarter of the NRY euchromatin, six of which carry testes genes. Of the three groups of euchromatic sequences, ampliconic sequences contain the highest density of genes.

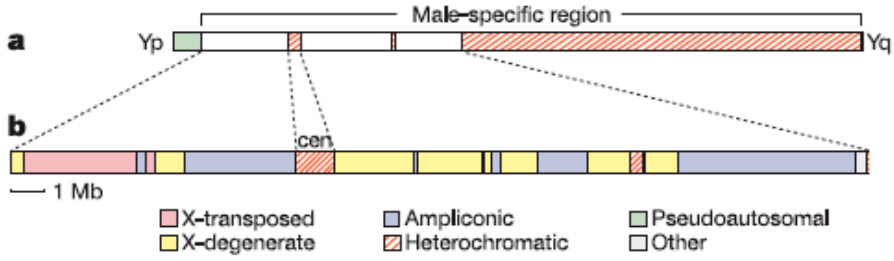


Figure 2. MSY region of the Y chromosome (from Skaletsky et al. 2003). (a) Schematic representation of the Y chromosome, including the pseudoautosomal and heterochromatic regions. (b) Enlarged view of a 24-Mb portion of the MSY.

c) Inheritance

Like its female counterpart that is only inherited via the maternal line, the Y chromosome is only inherited via the paternal line. As a haploid system, in principle, the Y chromosome does not undergo recombination. Although the possibility of recombination taking place within the Y chromosome has been questioned due to the observation of Y-Y gene conversion events (Skaletsky et al. 2003), this is not considered to be the case when phylogenetic population inferences are made.

d) Selection

Like for mtDNA, it is generally assumed that the Y chromosome is also selectively neutral. However, because it contains many Y-specific genes that are important for male fitness and it escapes recombination - meaning that any selection would affect the entire chromosome- the possibility that the Y chromosome could be susceptible to selection has been suggested. Many studies have tried to clarify this point, searching for associations between Y haplotypes and male illnesses. Whilst some studies have failed to find such association (Jobling and Tyler-Smith 2004), others have detected signatures of selection acting on the Y chromosome (Jobling et al 1998; Yen et al. 1998; Repping et al. 2003). Although further research needs to be carried out, at this moment, it is widely considered that even if selection is acting on the male-specific chromosome, its effects are not strong enough to modify the general topology of its phylogenetic tree.

1.4. Applications of mitochondrial and Y chromosome DNA

Both mtDNA and the Y Chromosome have a series of features that make them extremely useful for studies on recent human evolution (Jobling and Tyler-Smith 1995; Underhill et al. 2000, 2003; Hammer et al. 2001; Hammer and Zegura 2002), medical genetics (Jobling and Tyler-Smith 2000), DNA forensics (Jobling et al. 1997) and genealogical reconstructions (Jobling et al. 2001).

The fact that they are maternally and paternally inherited, respectively, and that they do not recombine, means that the mtDNA of all the maternally-linked members and the Y chromosome of all the paternally-linked members of a family/group is exactly identical, making it possible to trace modern mtDNA and Y chromosome lineages back to ancestral females and males. Because both mtDNA and the Y chromosome only change by accumulating mutations over time, it can be said, that these genomes reflect the history of female and male lineages in a much simpler way than autosomes. Furthermore, the lack of recombination on these markers results in only one copy of the mtDNA and Y chromosome gene pool being passed on from generation to generation, with their effective population size (N_e) being one-fourth that of diploid autosomes and one-third that of the X Chromosome (Takahata 1993). This makes them more susceptible to drift, which increases the levels of variation between populations, especially small populations, making them good reflectors of population sub-structuring.

All these characteristics are extremely useful in the field of forensic genetics. mtDNA, for example, is especially useful in cases involving evidence based on samples of poor nuclear DNA content e.g. degraded DNA, teeth, skeletal remains etc. (Just et al. 2009) and the Y chromosome is useful in sexual assault cases, where male profiles can be successfully identified in a sample containing multiple donors (both female - male(s) and male-male(s)), and for paternity cases.

1.5. mtDNA and Y Chromosome phylogenies

To our good fortune, both mtDNA and the Y Chromosome contain many types of polymorphisms, such as SNPs, that allow researchers to examine variation on different time scales. The mtDNA genome, on one hand, contains many SNPs in the coding region and a high level of variation at HVR I/II. The Y chromosome, on the other, contains hundreds of well-characterized SNPs and many STRs (Jobling and Tyler Smith 2003; Kayser et al. 2004). MtDNA and Y chromosome lineages defined by SNPs are referred to as *haplogroups,* and mtDNA and Y lineages defined by HVR I/II variation and Y- chromosome STRs, respectively, are referred to as *haplotypes*.

MtDNA has a mutation rate 5-10 times higher than nuclear DNA (Parsons et al.1997) – an estimated $2\text{-}3x10^{-7}$ per nucleotide per generation (Horai 1995; Jorde et al. 1998). However, the estimates of the mutation rate of the coding region vary significantly depending on the type of study carried out (Ward et al. 1991; Torroni and Wallace 1994; Forster et al. 1996; Parsons et al. 1997;

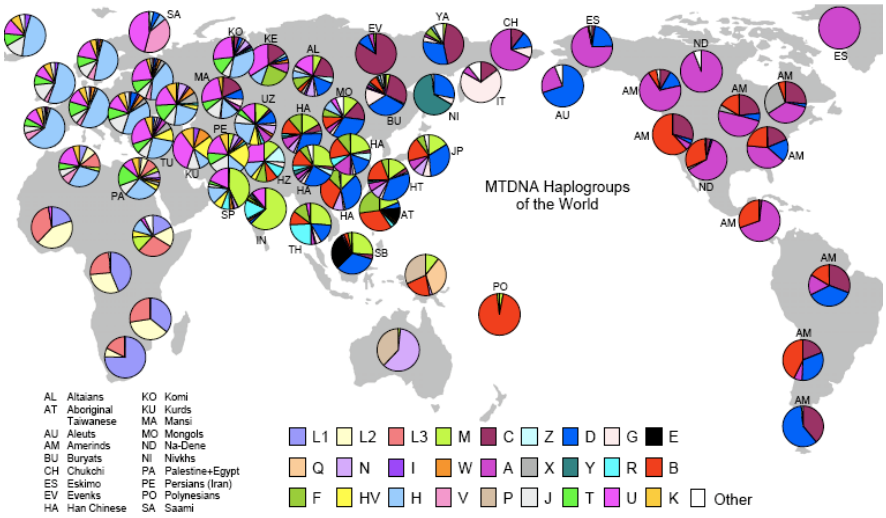
Siguroardóttir et al. 2000; Heyer et al. 2001), showing heterogeneous mutation rates across nucleotide positions, with some positions evolving at a faster rate than others (Soares et al. 2009).

Y-chromosome SNPs, like autosomal SNPs, have an average mutation rate of $2x10^{-8}$ (Jobling and Tyler-Smith 2003), whilst Y-STRs have been shown to have much higher mutation rates. Although the precise value of Y-chromosome STR mutation rates varies significantly depending on the estimation methods employed and the STR marker analysed, the following values have been estimated per generation; 2×10^{-3} (Heyer at al. 1997), $2.8x\ 10^{-3}$ (Kayser et al. 2000), 2.6×10^{-4} (Forster et al. 2000) and 6.9×10^{-4} (Zhivotovsky et al. 2004).

The slow evolving characteristic of SNPs enables the characterisation of deeply-rooted ancestral female and male lineages whilst the fast evolving characteristic of mtDNA HVR I/II and Y-chromosome STRs allow the study of more recently acquired variation. In this way, in evolutionary studies, the use of both types of markers is a valuable combination that increases resolution (de Knijff 2000), allowing the examination of the diversity within haplogroups. Furthermore, the examination of the evolutionary relationship between mtDNA and Y-chromosome SNPs has enabled the construction of robust highly resolved mtDNA and Y Chromosome Phylogenies (Richards and Macaulay 2001; Underhill et al. 2000), whose lineages are geographically structured and can be examined in an approach commonly known as "phylogeography" (Avise et al. 1987; Avise 2000).

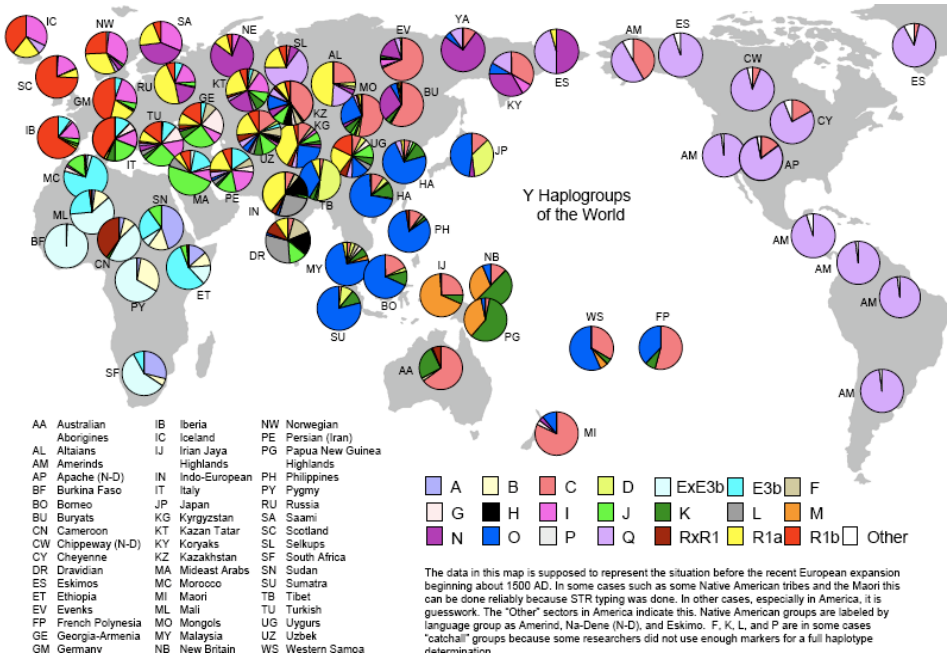
Since the discovery of the first mtDNA and Y-chromosome SNPs, technological advances (e.g. Polymerase Chain Reaction (PCR), Denaturing High Pressure Liquid Chromatography (DHPLC)) have led to the discovery of new SNPs resulting in the continuous up-dating and refinement of the mtDNA and Y Chromosome phylogenies. Furthermore, the possibility of screening populations from a wide-range of locations across continents for many mtDNA and Y-chromosome SNPs has increased our knowledge on the geographic distribution of their different mtDNA/Y Chromosome branches (haplogroups).

Studies have shown that the Time to the Most Recent Common Ancestor (TMCRCA) for the mtDNA and the Y chromosome phylogenies is not more ancient than 100-200 kilo years ago (kya) (Watson et al. 1997; Ingman et al. 2000; Hammer et al. 1998; Underhill et al. 2001; Underhill 2003;), and that their root is in Africa. The deepest branches of both mtDNA (L and its sub-groups) and Y chromosome (A and B) phylogenies are restricted to Africa and the rest are found in other parts of the world, completely separating African populations from non-African populations, in agreement with the Out-of-Africa hypothesis that suggests that modern diversity arose in Africa.



Geographic distribution of mtDNA lineages (from www.mcdonald.cam.ac.uk/genetics/links.html)

Archaeological evidence has suggested that there were at least two different migration events Out-of-Africa; a southern coastal route around the northern edge of the Indian Ocean reaching Australia 50-60 kya (Bowler et al. 2003), and a northern route, slightly later on, into Eurasia via the Levantine corridor 40,000 kya (Clark and Lindly 1989). The Americas was settled later on around 14,000 years ago. The geographic distribution of mtDNA and Y Chromosone lineages is in agreement with these findings, where apart from the ancestral set of African haplogroups, the rest of the non-African haplogroups cluster in three other geographic regions; Southeastern Asia/Australia, Central and Western Asia/Europe and the Americas and the Pacific, reflecting the migratory movements humans carried out after leaving Africa.



Geographic distribution of Y Chr lineages (from www.mcdonald.cam.ac.uk/genetics/links.html)

The current geographic clustering of both Y chromosome and mtDNA lineages has been/is still being further influenced by sex-specific behaviours such as patrilocality and matrilocality, which enhance local differentiation. The comparison of male and female patterns of migration allow us to better understand these sex-specific behaviours and how these and other demographic events have contributed to the current distribution of mtDNA and Y Chromosome gene pools observed.

1.6. Africa and the Bantu Expansion

a) The Bantu Expansion

Since the Out-of-Africa event, Africa has witnessed many important migrations, colonisations and population movements that have marked its history and shaped its genetic diversity patterns. The Bantu expansion, triggered by the transition from food collection to food production, is undoubtedly one of them.

The Bantu expansion of languages, one of the greatest population movements in recent African history (Diamond 1997), took place around 5,000 ya, when farmers speaking Bantu languages left their homeland situated in the present day border between Cameroon and Nigeria (Newman 1995) to initiate an unstoppable expansion throughout sub-Saharan Africa. During the expansion, that may have been driven by the advent of agriculture, and to some extent, to the emergence of iron techniques (Newman 1995), Bantu agriculturalists replaced neighbouring hunter-gatherer populations, namely Pygmy and Khoisan groups, forcing them to mix with them or to move to remote isolated areas (Cavalli-Sforza 1986).

Although the exact mode and tempo of the Bantu expansion are still somewhat controversial, data suggest that the expansion involved several dispersals with a relatively small number of people, rather than one single vast movement (Van Bakel 1981; Vansina 1995), following two main routes or waves; an eastern wave, leading north of the main African forest block to the Great Lakes region of East Africa (in what is nowadays Uganda) and then on to South Africa, and a western wave, leading southwards and penetrating the

equatorial forest in two directions, one following the coast line, and the other following the many major ridge lines running north-south into the interior (Oslisly 1995), with conversion events between both waves taking place in Central Africa (Newman 1995).

The Bantu expansion resulted in the simultaneous spread of Bantu languages and Bantu genes across the whole of sub-Equatorial Africa (Diamond and Bellwood 2003), with this area mainly being inhabited by Bantu groups (with a only few Pygmy and Khoisan populations in Central and Southern Africa, respectively) all speakers of Bantu languages. Although the demographic intensity of the Bantu expansion is still uncertain, its effects can be observed in the current genetic diversity patterns of sub-Saharan Africa and the distribution of Bantu languages.

b) Distribution of Bantu languages in Africa

It has been estimated that there are over 6,000 languages in the world (Grimes 2000) - a third of which are African languages (www.ethnologue.com). African languages belong to four main linguistic families; Afro-asiatic, Khoisan, Niger-Congo and Nilo-Saharan (Heine and Nurse 2000). The Niger-Congo family is the largest phyla of the four (Williamson and Blench 2000), both in terms of the amount of languages it contains and in its geographical distribution, with Niger-Congo languages being spoken throughout sub-Saharan Africa. Around one third of the Niger-Congo phyla are Bantu languages, a group of over 500 languages spoken in almost half of all sub-Saharan countries (Ruhlen 1991) - from Cameroon all the way to South Africa.

This large distribution of the Bantu languages is explained by the Bantu expansion of languages.

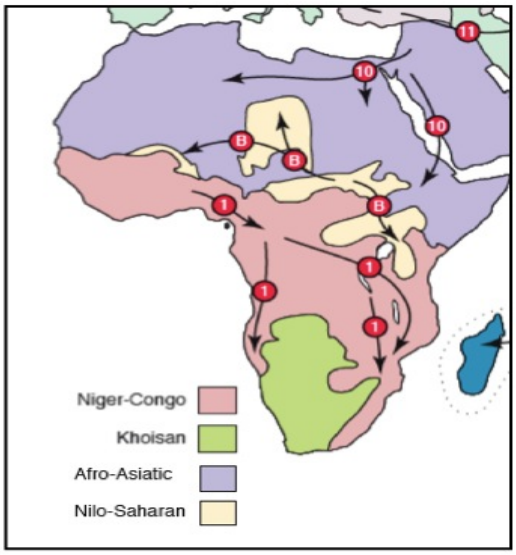


Figure 3. Language families in Africa (modified from Diamond and Bellwood 2003). (1), (B) and (10) represent the expansion routes suggested for Bantu languages, Nilo-Saharan and Afro-Asiatic languages respectively.

Studies on the distribution of Y chromosome genetic variation and linguistics in sub-Saharan Africa have shown an association between Bantu languages and Bantu genes, with Y chromosomal SNP haplogroups being significantly correlated with the distribution of Bantu languages (Wood et al. 2005). Given that my thesis is mostly based on the analysis of the Y chromosome in African populations involved in the Bantu expansion, the following part of my thesis will only make reference to this uniparentally inherited marker.

c) Distribution of Y lineages in sub-Saharan Africa

The Bantu expansion erased (at least partly) the ancient Y-chromosome diversity found in the areas it surpassed, vastly changing the African genetic landscape. However, some traces of ancient paternal lineages are still observed in sub-Saharan populations, mainly among hunter-gatherers. Sub-Saharan populations are characterized by the presence of four main Y chromosome haplogroups; A, B, E and R.

Haplogroups A and *B* are the deepest clades of the Y-chromosome phylogeny, being practically exclusive to Africa, and showing a wide distribution across all sub-Saharan Africa at moderate-to-low frequencies. Both haplogroups have been found in Khoisan-speaking populations from South Africa (Underhill et al. 2001; Cruciani et al. 2002; Wood et al. 2005), in the Fulbe population from West Central Africa (Cruciani et al. 2002), in populations from Sudan and Ethiopia, and the Sandawe population from East Africa (Underhill et al. 2001; Cruciani et al. 2002; Tishkoff et al. 2007). Haplogroup B has also been found in Biaka and Mbuti pygmy populations (Underhill et al. 2001), in various populations from East Africa (Knight et al. 2003; Luis et al. 2004; Tishkoff et al. 2007), and in Yoruba, Mossi and Bamileke from West Central Africa (Cruciani et al. 2002; Tishkoff et al. 2007). Outside Africa, haplogroup A has only been found in one individual from Sardinia that has been considered to be the result of a recent event (Underhill et al. 2001) and haplogroup B has not been found in any non-African populations.

Haplogroup E is thought to have originated in Africa around 50,000 years ago, being present throughout the whole of the African continent (Wood et al. 2005; Cruciani et al. 2002). Haplogroups E1a and E2 are quite rare, being mainly present in East Africa (Luis et al. 2004). Although haplogroup E2 is found at low frequencies across the entire African continent (Cruciani et al. 2002), its sub-branch E2b has been found to be more frequent among Bantu populations (Cruciani et al. 2002: Underhill et al. 2000). Haplogroup E1b1 is highly widespread and very frequent across sub-Saharan Africa. Its sub-branch E1b1a has been proposed as a marker of the Bantu expansion (Wood et al. 2005; Underhill et al. 2001; Scozzari et al. 1999; Passarino et al. 1998) and is present in most sub-Saharan populations, including hunter-gatherers (Underhill et al. 2001; Cruciani et al. 2002; Beleza et al. 2005; Tishkoff et al. 2007). Sub-haplogroup E1b1b is restricted to East and North Africa, although it has also been observed outside Africa in Europe and Western Asia (Semino et al. 2000; Underhill et al. 2001; Cruciani et al. 2004; Wood et al. 2005; Tishkoff et al. 2007).

Haplogroup R is not very frequent in Africa. Rather, it is mainly found in Asia and Europe, and is thought to have originated in Asia (Underhill et al. 2001). However, its sub-branch R1b1* has been observed in Northern and Southern Cameroon at a high frequency (Cruciani et al. 2002). Its presence in these areas has been thought to be due to a back migration from Asia to sub-Saharan Africa (Cruciani et al 2002).

2. AIMS

2.1.Technological Innovation

From a technological point of view, our aim was to create a Y-chromosome SNP multiplex capable of distinguishing samples from a wide range of geographic areas in a single reaction. Our multiplex would improve the current methodologies for Y chromosome typing, which, when used for multiple SNPs, can result extremely laborious and highly time and DNA consuming. The wide haplogroup coverage of our multiplex would eliminate the need for the hierarchical typing system widely used in Y chromosome studies. With the newly created multiplex, we aimed to carry out population genetic studies of several populations in a fast, secure and effective manner.

2.2. Population genetic inferences

a) West Central Africa

Although much is known in terms of linguistics and archaeology for this mass population movement, less is known in terms of genetics. Many studies have attempted to provide insights into the nature of the Bantu expansion from a genetic perspective using both mtDNA (Bandelt et al. 2001; Torroni et al. 2001; Ingman et al. 2000; Salas et al. 2002; Pereira et al. 2002; Plaza et al. 2004) and the Y chromosome (Cruciani et al. 2002; Coia et al. 2004; Luis et al. 2004; Semino et al. 2002; Knight et al. 2003; Underhill et al. 2001;

Beleza et al. 2005; Kayser et al. 2001; Thomas et al. 2000; Hammer et al. 2001, Destro-Bisol et al. 2004, Wood et al. 2005; Côrte-Real et al. 2000) and have postulated specific sets of markers as genetic footprints of the Bantu expansion, such as L0a, L2a, L3b and L3e for the mtDNA (Pereira et al. 2001; Salas et al. 2002; Plaza et al. 2004; Beleza. et al. 2005) and E1b1a (previously named E3a), E2 and B2a for the Y chromosome (Underhill et al. 2000; Cruciani et al. 2002; Beleza et al. 2005). However, most of these studies have focused on the Eastern migration wave of the Bantu expansion, with only a few analysing populations affected by the Western route. Taking into account that both routes varied significantly in geographical constraints, time-scale, and most probably, in the size of the migrating population, it is not surprising that differences between the two waves have been found. Data on the Western route is, therefore, crucial for the full understanding of this expansion.

An area of the Western route that has proved to be especially interesting is West Central Africa. Lying in the vicinity of the Bantu cradle, this area homes hundreds of different Bantu-speaking agriculturalists and small Pygmy hunter-gatherer groups; Bantu groups being spread across the whole of the area, western pygmies being present in the Western Congo basin, Cameroon, Gabon and the Central African Republic, and Eastern pygmies being present in the North-Eastern area of the Republic of Congo.

Pygmies lead a significantly different lifestyle to Bantus. Although they also mainly speak Niger-Congo languages that they have adopted from their neighbours, they are hunter-gatherers, physically

different (much shorter), monogamous, and practice patrilocality. Y chromosome studies have shown that these populations were greatly assimilated by their Bantu neighbours e.g. presence of haplogroup E1b1a in Pygmy populations (Wood et al. 2005), where as mtDNA studies have shown a lesser effect of Bantu populations, with the presence of ancestral mtDNA maternal lineages in Pygmy gene pools (Batini et al. 2007; Quintana-Murci et al. 2008). These results point towards the effect of sex-biased rates of admixture triggered by socio-cultural factors influencing the process of the Bantu expansion. Furthermore, the influence of Bantu people on the gene pool of pygmy populations has shown to vary among pygmy groups, with Western pygmies showing higher admixture levels with Bantu agriculturalists than Eastern pygmies (Thomas et al. 2000; Beleza et al. 2005; Pereira et al. 2001; Pereira et al. 2002) for the Y chromosome, and mtDNA marker of the bantu expansion haplogroup L1c, being found at extremely high frequencies in Western Pygmy populations (Batini et al. 2007; Quintana-Murci et al. 2008) but completely absent in Eastern Pygmies.

Taking all this data into consideration, in this thesis I have concentrated on studying the genetic diversity in Bantu and Western pygmy populations from the Central African regions of Gabon and Cameroon from the Y chromosome perspective in order to:

* Characterise the genetic structure of Bantu and pygmy populations from West Central Africa
* Measure the extent and symmetry of the gene flow between

Pygmy and non-pygmy populations

* Further explore the influence of Bantu populations on the Y chromosome gene pool of pygmy populations, tackling the possibility of ancestral paternal lineages prior to the Bantu expansion that may have been missed in other areas sampled

* Trace the spread of the Eurasian haplogroup R1b1* with the aim of refining the current hypothesis of a back migration to Africa as the origin of this haplogroup (Cruciani et al. 2002).

b) The Island of La Réunion

The island of La Réunion, a small island located in the Indian Ocean, is nowadays a multi-ethnic population – a true reflection of the population movements it experienced during its colonization back in the 17th century. The constant influx of foreign labourers from various continents throughout the colonization period resulted in people of very diverse origin living side-by-side, giving rise to the current melting pot of people living on this island. Although historical records have provided insights into the potential ethnic origin of these foreign labourers, there is little genetic data on how the different ethnic groups contributed to the current gene pool of La Réunion population, and what demographic processes acted on these during the colonization event.

Taking into consideration the lack of genetic data on this population, in this thesis I have concentrated on studying the genetic diversity in an admixed sample of the "general" population of La Réunion both from a mtDNA and Y chromosome DNA perspective in order to:

∗ Provide data on the genetic composition and diversity patterns of La Réunion Island

∗ Help understand how the colonization system implemented by the French shaped the genetic diversity currently observed

∗ Measure the contribution of the different incoming ethnic groups to the gene pool of this population

∗ Examine possible female/male biasing during the population settlement of the island

We also aimed to use the Y-chromosome SNP multiplex in the study of populations from Cuba in order to investigate the demographic processes that shaped the current Cuban population.

2.3. Collaboration in forensic genetic studies

Throughout the development of the thesis, on-going collaborations in forensic projects took place in order to improve and standardize the analysis of Y chromosome, mtDNA and autosomal markers and be up-to-date with innovations in this field that can also be applied to population genetics.

3. RESULTS

3.1.Creation of a 35 Y-chromosome SNP multiplex

Article 1: Berniell-Lee G, Sandoval K, Mendizabal I, Bosch E, Comas D (2007) SNPlexing the human Y-Chromosome: a single assay system for major haplogroup screening. Electrophoresis. 28:3201-6.

Berniell-Lee G, Sandoval K, Mendizabal I, Bosch E, Comas D. SNPlexing the human Y-chromosome: a single-assay system for major haplogroup screening. Electrophoresis. 2007; 28(18): 3021-6.

3.2. Population genetic inferences

Article 1: Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, van der Veen L, Hombert JM, Quintana-Murci L, Comas D (2009) Genetic and demographic implications of the Bantu expansion: insights from human paternal lineages. Mol Biol Evol. 26:1581-9.

Berniell-Lee G, Calafell F, Bosch E, Heyer E, Sica L, Mouguiama-Daouda P, et al. Genetic and demographic implications of the Bantu Expansion: insights from human paternal lineages. Mol Biol Evol. 2009; 26(7): 1581-9.

Article 2: Berniell-Lee G, Plaza S, Bosch E, Calafell F, Jourdan E, Césari M, Lefranc G, Comas D (2008) Admixture and sexual bias in the population settlement of La Réunion Island (Indian Ocean). Am J Phys Anthropol. 136:100-7.

Berniell-Lee G, Plaza S, Bosch E, Calafell F, Jourdan E, Césari M, et al. Admixture and sexual bias in the population settlement of La Réunion Island (Indian Ocean). Am J Phys Anthropol. 2008; 136(1): 100-7.

Annexe I. Population genetic study of Cuban populations

Article 1. Mendizabal I, Sandoval K, **Berniell-Lee G**, Calafell F, Salas A, Martínez-Fuentes A, Comas D (2008) Genetic origin, admixture and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol. 8: 213-223.

Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martínez-Fuentes A, et al. Genetic origin, admixture and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol. 2008; 8: 213.

Annexe II: Collaboration in forensic studies

Article 1: Crespillo M, Paredes MR, Prieto L, Montesino M, Salas A, Albarran C, Alvarez-Iglesias V, Amorin A, **Berniell-Lee G**, Brehm A, Carril JC, Corach D, Cuevas N, Di Lonardo AM, Doutremepuich C, Espinheira RM, Espinoza M, Gómez F, González A, Hernández A, Hidalgo M, Jimenez M, Leite FP, López AM, López-Soto M, Lorente JA, Pagano S, Palacio AM, Pestano JJ, Pinheiro MF, Raimondi E, Ramón MM, Tovar F, Vidal-Rioja L, Vide MC, Whittle MR, Yunis JJ, Garcia-Hirschfel J (2006) Results of the 2003-004 GEP-ISFG collaborative study on mitochondrial DNA: focus on the mtDNA profile of a mixed semen-salive stain. Forensic Sci Int. 2006 160:157-67.

Crespillo M, Paredes MR, Prieto L, Montesino M, Salas A, Albarran C, et al. Results of the 2003-2004 GEP-ISFG collaborative study on mitochondrial DNA: focus on the mtDNA profile of a mixed semen-saliva stain. Forensic Sci Int. 2006; 160(2-3): 157-67.

4. DISCUSSION

4.1.Technological advancements

Recent technological advancements in high-throughput DNA sequencing and SNP typing have offered, and continue to offer, more efficient and robust automated methods to characterize diversity patterns in human populations and identify candidate loci for disease. Techniques such as matrix-assisted laser desorption/ionisation-time of flight mass spectrometry (MALDI-TOF MS) or microarrays are extremely useful for the typing of SNPs, not only because they enable the simultaneous typing of thousands of these markers in one go, but also, because they are extremely fast and only require tiny quantities of DNA. Up until now, techniques of this type have proved highly successful for the typing of autosomal SNPs, but have not been attempted with mtDNA and Y chromosome markers, because studies with these genomes generally require the typing of a few selected SNPs and in a hierarchical manner. The aim of the first part of this thesis was to make use of the vast advantages these new technologies have to offer and provide a much more efficient alternative for the typing of Y chromosome SNPs than the allele-specific probes and single-base extension methods currently used. The high-throughput technology chosen for this purpose was the SNPlex system based on an Oligo Ligation Assay coupled to Polymerase Chain Reaction (OLA/PCR). The result was a highly robust 35 Y-chromosome SNP multiplex capable of distinguishing up to 31 different haplogroups and sub-haplogroups on the Y chromosome phylogeny (YCC 2002),

correctly assigning samples of different geographic origin to their corresponding haplogroup.

The 35 Y-chromosome multiplex created included 6 SNPs more than the largest multiplex to date for the Y chromosome (Brion et al. 2005), and showed no incompatibilities between alleles of the different SNPs. Furthermore, the lack of inconsistencies corroborated the accuracy of the current Y chromosome phylogeny consensually used for population genetic and forensic studies.

All in all, SNPlex proved to be a highly accurate, fast, and robust technique to type SNPs that can now also be applied to Y SNPs, allowing many markers in hundreds of samples to be typed in a short period of time. Applying SNPlex technology to Y chromosome studies, known to be highly tedious and time-consuming, could result highly advantageous and open a new door for large scale population studies. Although the pass rate and average call rates were slightly lower for the 35 Y-chromosome multiplex than for other autosomal markers (Tobler et al. 2005), probably due to the repetitive nature of the Y chromosome, one must take into consideration that this has been the first attempt to SNPlex Y chromosome SNPs and that, although there is plenty of room for improvement, this study has created a valuable typing tool for Y chromosome SNPs in population genetic studies.

4.2. Insights into the Bantu Expansion

With the Western wave of the expansion being very poorly characterised from a genetic point of view compared to the Eastern wave, West Central Africa represents a crucial area for the full understanding of the dynamics of this large population movement. Not only is West Central Africa one of the first stepping-stones of the expansion, it is also an area where hundreds of Bantu farmers and small hunter-gatherer pygmy groups co-exist and have been doing so for thousands of years. In order to provide insights into the genetic composition of this area, 22 Bantu-speaking groups and 3 Pygmy groups from Cameroon and Gabon were extensively analysed at the Y chromosome level.

The genetic characterization of Bantu and Pygmy populations showed that the Y chromosome gene pool of both groups show the effects of the Bantu expansion, which largely homogenised the gene pool of Bantu populations, having erased the pre-Bantu diversity, and diversified that of Pygmy groups, incorporating Bantu lineages into their gene pool. This can be seen in the fact that Bantu populations show very low levels of haplogroup diversity, being largely characterized by two main lineages (E1b1a and B2a) that have been identified as signatures of the Bantu expansion (Underhill et al. 2001), and Pygmies showing much higher haplogroup diversity, presenting ancient pre-Bantu lineages, such as B2b and A, that originated early during the history of modern humans (Cruciani et al. 2002; Knight et al. 2003), plus the lineages brought in by Bantus through gene flow from these farmers to the

hunter-gatherer populations. Most of the paternal lineages in West Central Africa, therefore, have a recent origin brought about by the Bantu Expansion with only Pygmy populations having retained some traces of ancient lineages previous to the Bantu expansion. Furthermore, the flow of paternal lineages among Bantu and Pygmy populations in West Central Africa has shown to be mainly unidirectional, from Bantu-to-Pygmy but not Pygmy-to-Bantu. This can be seen in the strong Bantu influence that Pygmies present, with these populations showing high frequencies of haplogroups related to the Bantu expansion, such as haplogroup E1b1a (previously known as E3a), and Bantu populations showing an extremely weak Pygmy influence, presenting exclusively Bantu lineages, with extremely low frequencies of Pygmy-related haplogroups, such as B2a.

These findings are surprising when compared to mtDNA data on this area of Africa, where several studies have shown high frequencies of ancient maternal lineages, considerable gene flow from Pygmy populations to Bantus, and a substantial common and deep ancestry between Bantu agriculturalists and Pygmies (Batini et al. 2007; Quintana-Murci et al. 2008). However, they are in agreement with a previous suggestion of the Bantu expansion having had a stronger influence on the Y chromosome pool of sub-Saharan populations than on the mtDNA pool (Destro-Bisol et al. 2004; Wood et al. 2005). These discrepancies between mtDNA and Y chromosome lineages in Pygmies could be due to sex-biased admixture rates between Bantu populations and Pygmy groups explained by socio-cultural factors such as patrilocality, whereby

Bantu Y lineages could have been introduced into the Pygmy gene pool through extramarital unions between Pygmy females and Bantu farmers due to their low bride price, fertility or similar conditions (Destro-Bisol et al. 2004).

One of the intriguing points of the analysis of the genetic variation in Bantu and Pygmy populations from Gabon and Cameroon, is the significant presence of paternal lineages belonging to Eurasian haplogroup R1b1*. This haplogroup, that has been hypothesised to represent footprints of a back migration into sub-Saharan Africa (Cruciani et al. 2002), has also been found in Cameroon, Oman, Egypt, Rwanda (Luis et al. 2004) and Sudan (Hassan et al. 2008).

According to the known Y-Chromosome phylogeny (Underhill et al. 2000), the presence of this haplogroup in sub-Saharan populations is somewhat strange, given that its origin has been placed in Eurasia and not Africa. The STR analysis of the diversity within this haplogroup seems to point towards its origin being related to a possible expansion prior to the Bantu expansion. However, given that its frequency is especially high in northern Cameroon, very close to the putative origin of the Bantu expansion, it could also simply represent the expansion of a sub set of the diversity within haplogroup R1b1* to West Central Africa through this population movement.

The fact that no traces of a non-African lineage have been found for the mtDNA further supports a sexually biased demographic expansion in West Central Africa, further highlighting the need to carry out more studies with Central African populations in order to

clarify this issue. Nigeria, Niger, Chad and Sudan, for example, could be interesting populations to sample in the near future.

Therefore, the current data point towards the following conclusions:

a) The Bantu expansion involved more male subjects than female ones
b) A Bantu-to-pygmy flow of paternal lineages
c) Pygmy populations retained some ancestral pre-Bantu lineages
d) A presence of Eurasian haplogroup R1b1* in Bantu populations being related to a pre-Bantu migration or to
e) A possible pre-Bantu migration having given rise to haplogroup R1b1* in Bantu populations and a possible pre-Bantu migration origin for this haplogroup

However, further population studies with mtDNA and Y chromosome markers in sub-Saharan Africa, and specifically in other areas of Central Africa, are needed to further elucidate these points.

4.3. Insights into the genetic composition of La Réunion

La Réunion is a relatively ‘‘young’’ population, given that it remained devoid of inhabitants until colonised by French in the 17th century. According to historical records, no ‘‘autochthonous’’ population originally lived on the island and, thus, the current Reunionese gene pool should be derived uniquely from the populations employed during the colonisation process.
A mixed ancestry of Asian and African origin has already been suggested in La Réunion by the presence and transmission of typically Asian and African gamma-immunoglobulin (Gm) alleles and immunoglobulin (IgG) haplotypes (Dugoujon et al., unpublished data).

In order to provide data on the genetic diversity patterns in La Réunion island, and help understand how the these were shaped by the colonisation system implemented by the French, a randomly selected sample of 41 individuals of “unknown” ethnic origin - considered to be representative of the general population - were typed both at the mtDNA and Y chromosome DNA level.

In agreement with the known colonization history of La Réunion, the genetic characterisation of this sample set showed a genetic pool composed of both European lineages, most probably derived from the European colonisers, and lineages from India, Africa and East Asia, most likely derived from the slaves and/or workers imported to work on the plantations. An interesting finding, however, was the

asymmetrical distribution of these lineages in the female and male gene pools - whereas most of the Y chromosome lineages were of European/Middle Eastern origin, with only a few lineages of East Asian origin and no lineages from Africa, most of the mitochondrial lineages were of Indian and East Asian origin, with only a few of European/Middle Eastern and African origin. These results point towards a clear sexual bias in the population settlement of La Réunion Island, in terms of both the influence of the European colonisers and the influence of the different incoming ethnic groups. The mtDNA pool from La Réunion shows a much lower European impact than the Y chromosome pool (19% vs 85%), and shows a much wider contribution from lineages of the different ethnic groups (Indian, East Asian and African) compared to male lineages, that only show a contribution from lineages from one incoming ethnic group (East Asian). These patterns suggest that the inter-ethnic crosses taking place at the time of colonisation, were predominantly between male European colonisers, and female imported workers (primarily of Indian origin). This kind of male-biased European admixture was probably due to the social organization of the French colonisation, characterised by an asymmetrical import of men versus women for their work on the sugar cane and coffee plantations. Furthermore, restrictions in the movement of male workers, together with the harsh living conditions that made it hard to survive, could have also reduced the reproductive events between male slaves and the few female slaves, leading to more reproductive events between the females slaves and the European colonisers and further exaggerating the sexual bias. In

addition, estimates of the genetic variation for both Y chromosome and mtDNA gene pools are in agreement with these findings, where the Y chromosome pool is highly diverse, and the mtDNA pool is highly homogeneous, characterised by a lack of diversity in its two most frequent haplogroups (M2 and E1). This reduction in diversity could be due to founder effects within the mtDNA pool caused by several reasons, for example, kinship-structured immigration or the sexually biased import of slaves, consequently leading to strong genetic drift due to a reduced population size.

Although the sources of the lineages found in this sample set are consistent with a European colonisation and the "putative" origins of the slaves/workers taken to La Réunion described by historical records, the fact that some of these lineages have also been found in other populations geographically close to La Réunion - also colonised by the French - suggests that these lineages might have been brought into La Réunion via these populations instead. A clear example of this possible scenario are the African lineages, which are also found in Malagasy individuals, and East Asian lineages, which are also found in Malagasy and Southeast Asian individuals. These lineages could be derived from Malagasy individuals (imported into the island to work on the sugar and coffee plantations) who have been found to have a dual Indonesian and African ancestry, result of the peopling of Madagascar by Indonesians via Southern India and by Africans via East Africa. This is also true for some of the European lineages (e.g. those belonging to Haplogroup R1b), that have been found in several

populations within Island Southeast Asia and Madagascar, and may have been brought into La Réunion through slaves from these areas that gained a European component when also colonised by Europeans.

Therefore, the current data point towards the following conclusions:
(a) A mixed ancestry of lineages in La Réunion (European, Asian, African and Indian) – in agreement with previous findings of a mixed ancestry of Asian and African (Dugoujon et al., unpublished data).
(b) A sexual bias in the population settlement of La Réunion – both in terms of the contribution from European lineages (very strong European component in the Y chromosome pool (85%) vs a week component in the mtDNA pool (18%)) and in terms of the contribution from lineages from the incoming slave/worker populations (a strong and wide contribution from several ethnic groups in the mtDNA pool (Indian 44%, East Asian 27%, and African 10%) and a weak contribution from only one ethnic group in the Y chromosome pool (East Asian 15%)).
(c) More possible origins for some of the lineages other than those documented by historical records, resulting from the import of slaves/workers from other nearby/colonized locations e.g. malagasy lineages.

Although our results have shown that the several historical and demographic episodes taking place in La Réunion have, without a doubt, shaped its genetic diversity patterns, especially, its

colonisation process, where the French implemented a system that led to people of European Indian, East Asian and African origin living side-by-side and intermingling, in order to ascertain whether the multi-ethnic character observed in our sample set is representative of La Réunion population as a whole, further larger-scale studies need to be carried out. Recently, the detailed analysis of mtDNA lineages in several Reunionese groups (Dubut et al. 2009) has reinforced the conclusions drawn in the present thesis and highlighted the importance of ethno-historical data to reconstruct the history of admixed populations. Furthermore, it would be interesting to establish the patterns and levels of genetic diversity for the Y-chromosome in the established ethnic groups currently described in La Réunion.

The present thesis has shown several examples of discrepancy between maternal (mtDNA) and paternal (Y chromosome) lineages as a result of a sexual bias in human populations. These examples, which include the sexual bias in farmer Bantu-speakers and Pygmies, the Reunionese, and the Cuban population (Mendizabal et al. 2008, see Annexe I), provide new knowledge to the growing evidence of sexual bias in the ethnogenesis of human populations (Batista dos Santos et al. 1999; Carvajal-Carmona et al. 2000; Carvalho-Silva et al. 2001; Kayser et al. 2003; Wooding et al. 2004; Quintana-Murci et al. 2010).

5. BIBLIOGRAPHY

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, *et al*. Sequence and organization of the human mitochondrial genome. Nature. 1981; 290(5806):457-65.

Auton A, Bryc K, Boyko AR, Lohmueller KE, Novembre J, Reynolds A, *et al*. Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. 2009; 19(5): 795-803.

Armour JA, Anttinen T, May CA, Vega EE, Sajantila A, Kidd JR, *et al*. Minisatellite diversity supports a recent African origin for modern humans. Nat Genet. 1996; 13(2): 154-60.

Awadalla P, Eyre-Walker A, Smith JM. Linkage disequilibrium and recombination in hominid mitochondrial DNA. Science. 1999; 286(5449): 2524-5.

Avise JC, Arnold J, Ball RM, Bermingham E, Lamb T, Neigel JE, *et al*. Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics. Ann Rev Ecol and Syst. 1987; 18: 489–522.

Avise JC (2000). Phylogeography: The History and Formation of Species. President and Fellows of Harvard College.

Bamshad MJ, Wooding S, Watkins WS, Ostler CT, Batzer MA, Jorde LB. Human population genetic structure and inference of group membership. Am J Hum Genet. 2003; 72(3): 578-89.

Bandelt HJ, Alves-Silva J, Guimarães PE, Santos MS, Brehm A, Pereira L, *et al*. Phylogeography of the human mitochondrial haplogroup L3e: a snapshot of African prehistory and Atlantic slave trade. Ann Hum Genet. 2001; 65(Pt 6): 549-63.

Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL. An apportionment of human DNA diversity. Proc Natl Acad Sci U S A. 1997; 94(9):4516-9.

Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, *et a*l. Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa. Mol Phylogenet Evol. 2007; 43(2): 635-44.

Batista dos Santos SE, Rodrigues JD, Ribeiro-dos-Santos AK, Zago MA. Differential contribution of indigenous men and women to the formation of an urban population in the Amazon region as revealed by mtDNA and Y-DNA. Am J Phys Anthropol. 1999; 109(2): 175-80.

Batzer MA, Deininger PL. A human-specific subfamily of Alu sequences. Genomics. 1991; 9(3): 481-7.

Batzer MA, Stoneking M, Alegria-Hartman M, Bazan H, Kass DH, Shaikh TH, *et al*. African origin of human-specific polymorphic Alu insertions. Proc Natl Acad Sci USA. 1994; 91(25): 12288-92.

Batzer MA, Deininger PL. Alu repeats and human genomic diversity. Nat Rev Genet. 2002; 3(5): 370-9.

Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A. The genetic legacy of western Bantu migrations. Hum Genet. 2005; 117(4): 366-75.

Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. High resolution of human evolutionary trees with polymorphic microsatellites. Nature. 1994; 368(6470): 455-7.

Bowler JM, Johnston H, Olley JM, Prescott JR, Roberts RG, Shawcross W, *et al*. New ages for human occupation and climatic change at Lake Mungo, Australia. Nature. 2003; 421(6925): 837-40.

Brión M, Sanchez JJ, Balogh K, Thacker C, Blanco-Verea A, Børsting C, *et al*. Introduction of an single nucleodite polymorphism-based "Major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages. Electrophoresis. 2005; 26(23): 4411-20.

Britten RJ. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. Proc Natl Acad Sci USA.

2002; 99(21):13633-5.

Calafell F, Shuster A, Speed WC, Kidd JR, Kidd KK. Short tandem repeat polymorphism evolution in humans. Eur J Hum Genet. 1998; 6(1): 38-49.

Campbell MC, Tishkoff SA. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annu Rev Genomics Hum Genet. 2008; 9: 403-33.

Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987; 325(6099): 31-6.

Carvajal-Carmona LG, Soto ID, Pineda N, Ortíz-Barrientos D, Duque C, Ospina-Duque J, *et a*l. Strong Amerind/white sex bias and a possible Sephardic contribution among the founders of a population in northwest Colombia. Am J Hum Genet. 2000; 67(5):1287-95.

Carvalho-Silva DR, Santos FR, Rocha J, Pena SD. The phylogeography of Brazilian Y-chromosome lineages. Am J Hum Genet. 2001;68(1): 281-6.

Cavalli-Sforza LL (1986). African Pygmies: an evaluation of the state of research. In Cavalli-Sforza LL (ed.): African Pygmies, pp 361-426. Orlando Academic Press, Orlando.

Cavalli-Sforza LL, Menozzi P, Piazza A (1994). The History and Geography of Human Genes. Princeton University Press, Princeton.

Cavalli-Sforza LL, Feldman MW. The application of molecular genetic approaches to the study of human evolution. Nat Genet. 2003; 33 Suppl: 266-75.

Chambers GK, MacAvoy ES. Microsatellites: consensus and controversy. Comp Biochem Physiol B Biochem Mol Biol. 2000; 126(4): 455-76.

Charlesworth B, Sniegowski P, Stephan W. The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 1994; 371(6494): 215-20.

Chen J, Sokal RR, Ruhlen M. Worldwide analysis of genetic and linguistic relationships of human populations. Hum Biol. 1995; 67: 595-612.

Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace DC. mtDNA variation in the South African Kung and Khwe-and their genetic relationships to other African populations. Am J Hum Genet. 2000; 66(4): 1362-83.

Chinnery PF, Vissing J. Treating MNGIE: is reducing blood nucleosides the first cure for a mitochondrial disorder? Neurology. 2006 Oct 24;67(8):1330-2.

Clark GA, Lindly J. The case of continuity: observations on the biocultural transition in Europe and Western Asia, in Mellars & Stringer. 1989: 626-76.

Coia V, Caglià A, Arredi B, Donati F, Santos FR, Pandya A, *et al*. Binary and microsatellite polymorphisms of the Y-chromosome in the Mbenzele pygmies from the Central African Republic. Am J Hum Biol. 2004; 16(1): 57-67.

Cooke HJ, Brown WR, Rappold GA. Hypervariable telomeric sequences from the human sex chromosomes are pseudoautosomal. Nature. 1985; 317(6039): 687-92.

Conrad DF, Jakobsson M, Coop G, Wen X, Wall JD, Rosenberg NA, *et al*. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet. 2006; 38(11): 1251-60.

Corbo RM, Scacchi R, Rickards O, Martinez-Labarga C, De Stefano GF. An investigation of human apolipoproteins B and E polymorphisms in two African populations from Ethiopia and Benin. Am J Hum Biol. 1999; 11(3): 297-304.

Corte-Real F, Andrade L, Carvalho M, Anjos MJ, Gamero J, Vieira DN, *et al*. Comparative analysis of STR data for Portuguese spoken countries. Prog Forensic Genet. 2000; 8: 212-214.

Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, *et al*. A Back Migration from Asia to Sub-Saharan Africa Is Supported by High-Resolution Analysis of Human Y-Chromosome Haplotypes. Am. J. Hum. Genet. 2002; 70: 1197–1214.

Deka R, Shriver MD, Yu LM, Ferrell RE, Chakraborty R. Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. Electrophoresis. 1995; 16(9): 1659-64.

de Knijff P. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am J Hum Genet. 2000; 67(5): 1055-61.

Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglià A, *et al*. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. Mol Biol Evol. 2004; 21(9): 1673-82.

Diamond J (1997). Guns, Germs, and Steel: The Fates of Human Societies. W.W. Norton & Company.

Diamond J, Bellwood P. Farmers and Their Languages: The First Expansions. Science. 2003; 300 (5619); 597 – 603.

Dubut V, Murail P, Pech N, Thionville MD, Cartault F. Inter- and extra-Indian admixture and genetic diversity in reunion island

revealed by analysis of mitochondrial DNA. Ann Hum Genet. 2009; 73(Pt 3): 314-34.

Edwards TL, Scott WK, Almonte C, Burt A, Powell EH, Beecham GW, *et al*. Genome-wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for Parkinson disease. Ann Hum Genet. 2010; 74(2):97-109.

Eller E. Population substructure and isolation by distance in three continental regions. Am J Phys Anthropol. 1999; 108(2): 147-59.

Eyre-Walker A, Smith NH, Smith JM. How clonal are human mitochondria? Proc Biol Sci. 1999; 266(1418): 477-83.

Forster P, Harding R, Torroni A, Bandelt HJ. Origin and evolution of Native American mtDNA variation: a reappraisal. Am J Hum Genet. 1996; 59(4): 935-45.

Forster P, Röhl A, Lünnemann P, Brinkmann C, Zerjal T, Tyler-Smith C, *et al*. A short tandem repeat-based phylogeny for the human Y chromosome. Am J Hum Genet. 2000 ;67(1): 182-96.

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, *et al*. Copy number variation: new insights in genome diversity. Genome Res. 2006; 16(8): 949-61.

Freije D, Helms C, Watson MS, Donis-Keller H. Identification of a

second pseudoautosomal region near the Xq and Yq telomeres. Science. 1992; 258(5089): 1784-7.

Fujiyama A, Watanabe H, Toyoda A, Taylor TD, Itoh T, Tsai SF, *et al*. Construction and analysis of a human-chimpanzee comparative clone map. Science. 2002; 295(5552): 131-4.

Goodman T, Ferro A, Sharma P. Pharmacogenetics of aspirin resistance: a comprehensive systematic review. Br J Clin Pharmacol. 2008; 66(2): 222-32.

Grimes BF, Grimes JE (2000). Ethnologue: languages of the world, 14th ed. Summer Institute of Linguistics, Dallas.

Hammer MF. A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. Mol Biol Evol. 1994; 11(5): 749-61.

Hammer MF, Karafet T, Rasanayagam A, Wood ET, Altheide TK, Jenkins T, *et al*. Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol. 1998; 15(4): 427-41.

Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, *et al*. Hierarchical patterns of global human Y-chromosome diversity.Mol Biol Evol. 2001; 18(7): 1189-203.

Hammer MF and Zegura SL. The human Y chromosome haplogroup tree: nomenclature and phylogeography of its major divisions. Annu. Rev. Anthropol. 2002; 31:303– 321

Hassan HY, Underhill PA, Cavalli-Sforza LL, Ibrahim ME. Y-chromosome variation among Sudanese: restricted gene flow, concordance with language, geography, and history. Am J Phys Anthropol. 2008; 137(3): 316-23.

Hedges DJ, Batzer MA. From the margins of the genome: mobile elements shape primate evolution. Bioessays. 2005; 27(8): 785-94.

Heine B and Nurse D (2000). African languages: an introduction. Cambridge (UK): New York: Cambridge University Press.

Heyer E, Puymirat J, Dieltjes P, Bakker E, de Knijff P. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet. 1997; 6(5): 799-803.

Heyer E, Zietkiewicz E, Rochowski A, Yotova V, Puymirat J, Labuda D. Phylogenetic and familial estimates of mitochondrial substitution rates: study of control region mutations in deep-rooting pedigrees. Am J Hum Genet. 2001; 69(5):1113-26.

Horai S, Hayasaka K, Kondo R, Tsugane K, Takahata N. Recent African origin of modern humans revealed by complete sequences of hominoid mitochondrial DNAs. Proc Natl Acad Sci U S A. 1995;

92(2): 532-6.

Horai S. Evolution and the origins of man: clues from complete sequences of hominoid mitochondrial DNA. Southeast Asian J Trop Med Public Health. 1995; 26 Suppl 1: 146-54.

Hurles ME, Jobling MA. A singular chromosome. Nat Genet. 2003; 34(3): 246-7.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, *et al*. Detection of large-scale variation in the human genome. Nat Genet. 2004; 36(9): 949-51.

Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000; 408(6813): 708-13.

International HapMap Consortium. A haplotype map of the human genome. Nature. 2005; 437(7063): 1299-320.

Jegalian K, Page DC. A proposed path by which genes common to mammalian X and Y chromosomes evolve to become X inactivated. Nature. 1998; 394(6695): 776-80.

Jobling MA, Tyler-Smith C. Fathers and sons: the Y chromosome and human evolution. Trends Genet. 1995; 11(11): 449-56.

Jobling MA, Pandya A, Tyler-Smith C. The Y chromosome in forensic analysis and paternity testing. Int J Legal Med. 1997; 110(3):118-24.

Jobling MA, Williams GA, Schiebel GA, Pandya GA, McElreavey GA, Salas GA, *et al*. A selective difference between human Y-chromosomal DNA haplotypes. Curr Biol. 1998; 8(25): 1391-4.

Jobling MA, Tyler-Smith C. New uses for new haplotypes the human Y chromosome, disease and selection. Trends Genet. 2000; 16(8): 356-62.

Jobling MA. In the name of the father: surnames and genetics. Trends Genet. 2001; 17(6): 353-7.

Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003; 4(8): 598-612.

Jobling MA, Hurles ME, Tyler-Smith C (2004). Human evolutionary genetics. New York: Garland Science.

Jorde LB, Rogers AR, Bamshad M, Watkins WS, Krakowiak P, Sung S, *et al*. Microsatellite diversity and the demographic history of modern humans. Proc Natl Acad Sci USA. 1997; 94(7): 3100-3.

Jorde LB, Bamshad M, Rogers AR. Using mitochondrial and

nuclear DNA markers to reconstruct human evolution. Bioessays. 1998; 20(2):126-36.

Jorde LB, Watkins WS, Bamshad MJ, Dixon ME, Ricker CE, Seielstad MT, *et al*. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y-chromosome data. Am J Hum Genet. 2000; 66(3): 979-88.

Jorgensen TJ, Helzlsouer KJ, Clipp SC, Bolton JH, Crum RM, Visvanathan K. DNA repair gene variants associated with benign breast disease in high cancer risk women. Cancer Epidemiol Biomarkers Prev. 2009;18(1): 346-50.

Just RS, Leney MD, Barritt SM, Los CW, Smith BC, Holland TD, *et al*. The use of mitochondrial DNA single nucleotide polymorphisms to assist in the resolution of three challenging forensic cases. J Forensic Sci. 2009; 54(4): 887-91.

Kathiresan S, Musunuru K, Orho-Melander M. Defining the spectrum of alleles that contribute to blood lipid concentrations in humans. Curr Opin Lipidol. 2008; 19(2): 122-7.

Kayser M, Roewer L, Hedman M, Henke L, Henke J, Brauer S, *et al*. 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. 2000;

66(5):1580-8.

Kayser M, Krawczak M, Excoffier L, Dieltjes P, Corach D, Pascali V, *et al*. An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. Am J Hum Genet. 2001; 68(4): 990-1018.

Kayser M, Brauer S, Weiss G, Schiefenhövel W, Underhill P, Shen P, *et al*. Reduced Y-chromosome, but not mitochondrial DNA, diversity in human populations from West New Guinea. Am J Hum Genet. 2003; 72(2): 281-302.

Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, *et al*. A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet. 2004; 74(6): 1183-97.

Kivisild T, Bamshad MJ, Kaldma K, Metspalu M, Metspalu E, Reidla M, *et al*. Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages. Curr Biol. 1999; 9(22): 1331-4.

Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, *et al.* African Y chromosome and mtDNA divergence provides insight into the history of click languages. Curr Biol. 2003; 13(6): 464-73.

Lahn BT, Page DC. Four evolutionary strata on the human X

chromosome. Science. 1999; 286(5441): 964-7.

Lao O, van Duijn K, Kersbergen P, de Knijff P, Kayser M. Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. Am J Hum Genet. 2006; 78(4): 680-90.

Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, *et a*l. Correlation between genetic and geographic structure in Europe. Curr Biol. 2008; 18(16): 1241-8.

Lawson Handley LJ, Manica A, Goudet J, Balloux F. Going the distance: Human population genetics in a clinal world. Trends in Genetics. 2007; 23(9): 432-439.

Lewontin R. The Apportionment of Human Diversity. Evolutionary Biology. 1972; 6:391–398.

Linz B, Balloux F, Moodley Y, Manica A, Liu H, Roumagnac P, *et al*. An African origin for the intimate association between humans and Helicobacter pylori. Nature. 2007; 445(7130): 915-8.

Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioğlu C, Roseman C, *et al*. The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations.Am J Hum Genet. 2004; 74(3): 532-44.

Lupski JR, Roth JR, Weinstock GM. Chromosomal duplications in bacteria, fruit flies, and humans. Am J Hum Genet. 1996; 58(1): 21-27.

Manica A, Prugnolle F, Balloux F. Geography is a better determinant of human genetic differentiation than ethnicity. Hum Genet. 2005; 118(3-4): 366-71.

Manica A, Amos W, Balloux F, Hanihara T. The effect of ancient population bottlenecks on human phenotypic variation. Nature. 2007; 448(7151): 346-8.

Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Gen. Res. Camb. 1974; 23: 23-35.

Mendizabal I, Sandoval K, Berniell-Lee G, Calafell F, Salas A, Martínez-Fuentes A, *et al*. Genetic origin, admixture, and asymmetry in maternal and paternal human lineages in Cuba. BMC Evol Biol. 2008; 8: 213.

McCarroll SA, Hadnott TN, Perry GH, Sabeti PC, Zody MC, Barrett JC, *et a*l. International HapMap Consortium. Common deletion polymorphisms in the human genome. Nat Genet. 2006; 38(1): 86-92.

Mountain JL, Cavalli-Sforza LL. Inference of human evolution

through cladistic analysis of nuclear DNA restriction polymorphisms. Proc Natl Acad Sci USA. 1994; 91(14): 6515-9.

Mourant AE, Tills D, Domaniewska-Sobczak K. Sunshine and the geographical distribution of the alleles of the Gc system of plasma proteins. Hum Genet. 1976; 33(3): 307-14.

Nachman MW, Brown WM, Stoneking M, Aquadro CF. Nonneutral mitochondrial DNA variation in humans and chimpanzees. Genetics.1996; 142(3): 953-63.

Nachman MW, Crowell SL. Estimate of the mutation rate per nucleotide in humans. Genetics. 2000; 156(1): 297-304.

Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, *et al.* Variable number of tandem repeat (VNTR) markers for human gene mapping. Science. 1987; 235(4796): 1616-22.

Nasidze I, Risch GM, Robichaux M, Sherry ST, Batzer MA, Stoneking M. Alu insertion polymorphisms and the genetic structure of human populations from the Caucasus. Eur J Hum Genet. 2001; 9(4): 267-72.

Newman JL (1995). The peopling of Africa. Yale University Press, New Haven, CT.

Ohno S, Kaplan WD, Kinosita R. On the sex chromatin of Gallus domesticus. Exp Cell Res. 1960;19:180-183.

Olson MV, Varki A. Sequencing the chimpanzee genome: insights into human evolution and disease. Nat Rev Genet. 2003; 4(1): 20-8.

Oslisly R. The middle Ogooué Valley, Gabon: Cultural changes and palaeoclimatic implications of the last four millenia. Azania. 1995; 39–40: 324–331.

Parsons TJ, Muniec DS, Sullivan K, Woodyatt N, Alliston-Greiner R, Wilson MR, *et al*. A high observed substitution rate in the human mitochondrial DNA control region. Nat Genet. 1997; 15(4): 363-8.

Passarino G, Semino O, Quintana-Murci L, Excoffier L, Hammer M, Santachiara-Benerecetti AS. Different genetic components in the Ethiopian population, identified by mtDNA and Y-chromosome polymorphisms. Am J Hum Genet 1998; 62: 420–434.

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(Pt 5): 439-58.

Pereira L, Gusmão L, Alves C, Amorim A, Prata MJ. Bantu and European Y-lineages in Sub-Saharan Africa. Ann Hum Genet. 2002; 66(Pt 5-6): 369-78.

Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, Carracedo A, *et al*. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. Hum Genet. 2004; 115(5): 439-47.

Powell JR and Taylor CE. Are human races "substantially" different genetically? Am. Nat. 1978; 112: 1139–1142.

Prugnolle F, Manica A, Balloux F. Geography predicts neutral genetic diversity of human populations. Curr Biol. 2005; 15(5): 159-60.

Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, *et al*. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. Proc Natl Acad Sci USA. 2008; 105(5): 1596-601.

Quintana-Murci L, Harmant C, Quach H, Balanovsky O, Zaporozhchenko V, Bormans C, *et al*. Strong maternal Khoisan contribution to the South African coloured population: a case of gender-biased admixture. Am J Hum Genet. 2010; 86(4): 611-20.

Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proc Natl Acad Sci USA. 2005; 102(44): 15942-7.

Reed FA, Tishkoff SA. African human diversity, origins and migrations. Curr Opin Genet Dev. 2006; 16(6): 597-605.

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, *et al*. Global variation in copy number in the human genome. Nature. 2006; 444(7118): 444-54.

Relethford JH. Global patterns of isolation by distance based on genetic and morphological data. Hum Biol. 2004; 76(4): 499-513.

Repping S, Skaletsky H, Brown L, van Daalen SK, Korver CM, Pyntikova T, *et al*. Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. Nat Genet. 2003; 35(3): 247-51.

Repping S, van Daalen SK, Brown LG, Korver CM, Lange J, Marszalek JD, *et al*. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. Nat Genet. 2006; 38(4):463-7.

Richards M, Macaulay V. The mitochondrial gene tree comes of age. Am J Hum Genet. 2001; 68(6):1315-20.

Rogers AR, Jorde LB. Genetic evidence on modern human origins. Hum Biol. 1995; 67(1): 1-36.

Romualdi C, Balding D, Nasidze IS, Risch G, Robichaux M, Sherry

ST, *et al*. Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. Genome Res. 2002; 12(4): 602-12.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, *et al*. Genetic structure of human populations. Science. 2002; 298(5602): 2381-5.

Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. 2005; 1(6): e70.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, *et al*. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. Nature. 2003; 423(6942): 873-6.

Ruhlen M (1991). A Guide to the World's Languages, Vol. 1: Classification (revised edition). London: Edward Arnold.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, *et al*. International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature. 2001; 409(6822): 928-33.

Salas A, Richards M, De la Fe T, Lareu MV, Sobrino B, Sánchez-

Diz P, *et al*. The making of the African mtDNA landscape. Am J Hum Genet. 2002; 71(5): 1082-111.

Schwartz M, Vissing J. Paternal inheritance of mitochondrial DNA. N Engl J Med. 2002; 347(8): 576-80.

Scozzari R, Cruciani F, Santolamazza P, Malaspina P, Torroni A, Sellitto D, *et al*. Combined use of biallelic and microsatellite Y-chromosome polymorphisms to infer affinities among African populations. Am J Hum Genet. 1999; 65(3): 829-46.

Seielstad MT, Minch E, Cavalli-Sforza LL. Genetic evidence for a higher female migration rate in humans. Nat Genet. 1998; 20(3): 278-80.

Semino O, Passarino G, Oefner PJ, Lin AA, Arbuzova S, Beckman LE, *et al*. The genetic legacy of Paleolithic Homo sapiens sapiens in extant Europeans: a Y chromosome perspective. Science. 2000; 290(5494): 1155-9.

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(1): 265-8.

Seo T, Ishitsu T, Ueda N, Nakada N, Yurube K, Ueda K, *et al*. ABCB1 polymorphisms influence the response to antiepileptic

drugs in Japanese epilepsy patients. Pharmacogenomics. 2006; 7(4): 551-61.

Sheen FM, Sherry ST, Risch GM, Robichaux M, Nasidze I, Stoneking M, *et al*. Reading between the LINEs: human genomic variation induced by LINE-1 retrotransposition. Genome Res. 2000; 10(10): 1496-508.

Sherry ST, Harpending HC, Batzer MA, Stoneking M. Alu evolution in human populations: using the coalescent to estimate effective population size. Genetics. 1997; 147(4): 1977-82.

Shriver MD, Mei R, Parra EJ, Sonpar V, Halder I, Tishkoff SA, *et al*. Large-scale SNP analysis reveals clustered and continuous patterns of human genetic variation.Hum Genomics. 2005; 2(2): 81-9.

Sigurğardóttir S, Helgason A, Gulcher JR, Stefansson K, Donnelly P. The mutation rate in the human mtDNA control region. Am J Hum Genet. 2000; 66(5): 1599-609.

Simmler MC, Rouyer F, Vergnaud G, Nyström-Lahti M, Ngo KY, de la Freije D, *et al*. Identification of a second pseudoautosomal region near the Xq and Yq telomeres. Science. 1992; 258(5089): 1784-7.

Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, *et al*.

Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 2009; 84(6): 740-59.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, *et al*. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003; 423(6942): 825-37.

Smith JM, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974; 23(1): 23-35.

Smit AF. The origin of interspersed repeats in the human genome. Curr Opin Genet Dev. 1996; 6(6): 743-8.

Stacey SN, Sulem P, Johannsson OT, Helgason A, Gudmundsson J, Kostic JP, *et al*. The BARD1 Cys557Ser variant and breast cancer risk in Iceland. PLoS Med. 2006; 3(7): e217.

Stoneking M, Fontius JJ, Clifford SL, Soodyall H, Arcot SS, Saha N, *et al*. Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res. 1997; 7(11): 1061-71.

Takahata N. Allelic genealogy and human evolution. Mol Biol Evol. 1993; 10(1): 2-22.

Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, le Roux M, *et al*. Y chromosomes traveling south: the cohen modal haplotype and the origins of the Lemba--the "Black Jews of Southern Africa". Am J Hum Genet. 2000; 66(2): 674-86.

Tishkoff SA, Verrelli BC. Patterns of human genetic diversity: implications for human evolutionary history and disease. Annu Rev Genomics Hum Genet. 2003; 4: 293-340.

Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, *et al*. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. Mol Biol Evol. 2007; 24(10): 2180-95.

Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, *et al*. The genetic structure and history of Africans and African Americans. Science. 2009; 324(5930): 1035-44.

Tobler AR, Short S, Andersen MR, Paner TM, Briggs JC, Lambert SM, *et al.* The SNPlex genotyping system: a flexible and scalable platform for SNP genotyping. J Biomol Tech. 2005; 16(4):398-406.

Torroni A, Wallace DC. Mitochondrial DNA variation in human populations and implications for detection of mitochondrial DNA mutations of pathological significance. J Bioenerg Biomembr. 1994; 26(3):261-71.

Torroni A, Bandelt HJ, Macaulay V, Richards M, Cruciani F, Rengo C, *et al*. A signal, from human mtDNA, of postglacial recolonization in Europe. Am J Hum Genet. 2001; 69(4): 844-52.

Turakulov R, Easteal S. Number of SNPS loci needed to detect population structure. Hum Hered. 2003;55(1): 37-45.

Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, *et al*. Y chromosome sequence variation and the history of human populations. Nat Genet. 2000; 26(3): 358-61.

Underhill PA, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, *et al*. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet. 2001; 65(Pt 1): 43-62.

Underhill PA. Inferring human history: clues from Y-chromosome haplotypes. Cold Spring Harb Symp Quant Biol. 2003; 68: 487-93.

Varenhorst C, James S, Erlinge D, Brandt JT, Braun OO, Man M, *et al*. Genetic variation of CYP2C19 affects both pharmacokinetic and pharmacodynamic responses to clopidogrel but not prasugrel in aspirin-treated patients with coronary artery disease. Eur Heart J. 2009; 30(14):1744-52.

Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC.

African populations and the evolution of human mitochondrial DNA. Science. 1991;253(5027):1503-1507.

Vansina J. New linguistic evidence and the "Bantu expansion." J Afr Hist . 1995; 36:173-195.

Van Bakel MA. The “bantu” expansion: demographic models. Curr. Anthropol. 1981; 22:688-691.

Wallace DC, Brown MD, Lott MT. Mitochondrial DNA variation in human evolution and disease. Gene. 1999; 238(1): 211-30.

Walter RC, Buffler RT, Bruggemann JH, Guillaume MM, Berhe SM, Negassi B, *et al*. Early human occupation of the Red Sea coast of Eritrea during the last interglacial. Nature. 2000; 405(6782): 65-9.

Ward RH, Frazier BL, Dewjager K, Paabo S. Extensive Mitochondrial diversity within a single Amerindian tribe. Proc Natl Acad Sci USA. 1991; 88: 8720–8724.

Watkins WS, Ricker CE, Bamshad MJ, Carroll ML, Nguyen SV, Batzer MA, *et al*. Patterns of ancestral human diversity: an analysis of Alu-insertion and restriction-site polymorphisms. Am J Hum Genet. 2001; 68(3): 738-52.

Watkins WS, Rogers AR, Ostler CT, Wooding S, Bamshad MJ, Brassington AM, *et al*. Genetic variation among world populations: inferences from 100 Alu insertion polymorphisms. Genome Res. 2003; 13(7): 1607-18.

Watson E, Forster P, Richards M, Bandelt HJ. Mitochondrial footprints of human expansions in Africa. Am J Hum Genet. 1997; 61(3): 691-704.

Weber JL, Wong C. Mutation of human short tandem repeats. Hum Mol Genet. 1993; 2(8):1123-8.

Weber JL, David D, Heil J, Fan Y, Zhao C, Marth G. Human diallelic insertion/deletion polymorphisms. Am J Hum Genet. 2002; 71(4): 854-62.

Weir BS, Hill WG. Estimating F-statistics. Annu Rev Genet. 2002; 36: 721-50.

Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG. Measures of human population structure show heterogeneity among genomic regions. Genome Res. 2005; 15(11): 1468-76.

Williamson K, Blench R (2000). 'Niger-Congo', in Heine, Bernd and Nurse, Derek (eds) African Languages - An Introduction. Cambridge: Cambridge University press, pp. 11–42.

Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, *et al*. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. Eur J Hum Genet. 2005; 13(7): 867-76.

Wooding S, Ostler C, Prasad BV, Watkins WS, Sung S, Bamshad M, *et al*. Directional migration in the Hindu castes: inferences from mitochondrial, autosomal and Y-chromosomal data. Hum Genet. 2004; 115(3): 221-9.

Wright S. Isolation by Distance. Genetics.1943; 28(2): 114-38.

Yamaguchi-Kabata Y, Shimada MK, Hayakawa Y, Minoshima S, Chakraborty R, Gojobori T, *et al*. Distribution and effects of nonsense polymorphisms in human genes. PLoS One. 2008; 3(10): e3393.

Yen PH. A long-range restriction map of deletion interval 6 of the human Y chromosome: a region frequently deleted in azoospermic males. Genomics. 1998; 54(1): 5-12.

YCC: A nomenclature system for the tree of Y chromosomal binary haplogroups. Genome Res. 2002; 12: 339–348.

Yu N, Fu YX, Li WH. DNA polymorphism in a worldwide sample of human X chromosomes. Mol Biol Evol. 2002; 19(12):2131-41.

Yucesoy B, Peila R, White LR, Wu KM, Johnson VJ, Kashon ML, *et al*. Association of interleukin-1 gene polymorphisms with dementia in a community-based sample: the Honolulu-Asia Aging Study. Neurobiol Aging. 2006; 27(2): 211-7.

Zhivotovsky LA, Underhill PA, Cinnioğlu C, Kayser M, Morar B, Kivisild T, *et al*. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet. 2004; 74(1):50-61.