Margarida Cardoso Reis Sá Coelho

Studying the evolutionary history of single markers and the demographic history of African populations:

The evolution of lactase persistence;

Genetic structure of Bantu-speaking populations; and the microevolutionary impact of the Atlantic slave trade

U.PORTO
FC FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

2010

Dissertação apresentada à Faculdade de Ciências da Universidade do Porto para a obtenção do grau de Doutor em Biologia

Thesis presented to the Faculty of Sciences, University of Porto for the Doctor degree in Biology

Orientador científico/Supervisor: Jorge Rocha

Agradecimentos/ Acknowledgments

A concretização desta tese não teria sido possível sem a colaboração e apoio de muitas pessoas às quais não poderia deixar de agradecer.

À Fundação para a Ciência e a Tecnologia agradeço a concessão da bolsa de doutoramento – SFRH/BD/22651/2005.

Ao Professor Jorge Rocha, agradeço a oportunidade de realizar os diversos trabalhos que compõem esta tese, nesta fascinante área de investigação. Agradeço também a ajuda e orientação ao longo destes anos bem como todos os conhecimentos que me transmitiu.

Ao Professor Sobrinho Simões, agradeço o caloroso acolhimento no IPATIMUP, a atenção dedicada ao meu trabalho, o entusiasmo e a confiança transmitidos.

Da Universidade Pedagógica de Moçambique, agradeço ao Professor António Prista e ao Doutor Sílvio Saranga, por terem tornado possível a minha viagem por Moçambique, da qual guardo preciosas recordações e experiências que enriqueceram a maneira de olhar para o meu próprio trabalho.

From Stanford University, I would like to thank to Joanna Mountain, Chris Gignoux, Brenna Henn and Matt Jobin, for the warm welcome during my short visit to their research group.

Aos colegas do CIBIO, agradeço a ajuda e colaboração, em particular ao Fernando Sequeira, pela paciência e entusiasmo.

A todos os co-autores das publicações apresentadas nesta tese agradeço a importante ajuda e colaboração, em especial à Isabel, no decurso dos trabalhos incluídos no Artigo 4. O “bom astral” e verdadeiro espírito de equipa no decorrer do trabalho “dos DIPSTRs” foram sem dúvida uma “armadura” face às dificuldades que surgiram. Incontornavelmente, as frases “É muito difícil levar a vida honradamente” e “Deixem-me dizer estas coisinhas” vão ficar ligadas a este trabalho. Ah, e obrigada também pela Figura I.4!

Gostaria também de agradecer a todos os meus colegas do IPATIMUP por terem contribuído, de diversas formas para a realização desta tese. Em particular:

Aos colegas da informática, pela ajuda na resolução dos quebra-cabeças "*in silico*";

Aos colegas e amigos mais próximos com quem tive o privilégio de partilhar saberes e experiências, assim como as alegrias e as tormentas do dia-a-dia: Mafalda, Cirnes, Rui, Alexandra, Filipe, Vânia, Sandra Martins, Sandra Beleza, Susana Seixas, João, Zélia, Patrícia;
e, de uma forma muito especial, à Rita, à Sofia, à Joana Rebelo, à Joana Campos e à Isabel (parte 2!)- foi um enorme privilégio poder contar com o vosso apoio e amizade a toda a prova. Obrigada por colorirem os meus dias com a vossa alegria e autenticidade.

Aos meus pais, ao Filipe, à Ana Alexandra, ao Ricardo, à Olinda, à Sara e ao Tiago. Foi entre vocês que encontrei a calma, a confiança e a força necessárias para levar este trabalho a bom porto.

Table of contents

Summary

The major contributions provided by genetic tools to our understanding of human evolutionary history can be divided in two main areas. One area comprises the identification and characterization of genes that have been targeted by selection, providing insights into the processes by which human species have dealt with different environments and lifestyles. The other area is related with the reconstruction of the demographic history of human populations, which mainly relies on the combined study of different types of markers with distinct characteristics and modes of transmission.

This thesis is divided in three parts. In the first part we present a study on the evolutionary history of lactase persistence, a genetic trait that allows lactose to be digested throughout life. This trait is thought to provide a selective advantage in dairying populations due to the added nutritional benefits of drinking milk. In the second and third parts, we focused on two episodes of the African demographic history: one related with the study of populations from Mozambique and Angola in the context of the large scale migration of Bantu-speaking populations and, the other, centred on the analysis of the population of the small island of São Tomé.

Our investigation on the lactase persistence was triggered by the finding that the -13910*T allele was a robust marker for this trait in individuals of Finnish origin. In fact, at the onset of our work, most studies on the evolutionary history of lactase persistence had been mainly based on the interpretation of correlations between phenotypic frequencies and environmental variables. In a first stage we intended to address two major key points related to the evolution of the -13910*T lactase persistence associated allele: a) to test the role of selection in shaping its present day distribution, and b) to estimate the age of the allele. To this purpose, we studied the microsatellite-defined haplotype variation closely linked to the -13910*T allele in several African and European populations to perform a formal neutrality test. Based on this test we conclude that the -13910*T allele is too recent to have reached its current frequencies without the influence of positive selection. The linked microsatellite variation was also used to calculate absolute ages of the allele, indicating that the -13910*T allele originated only after the separation between European and African populations and may be as recent as 12,500-7,500 years. This estimate associates the origin of the -13910*T allele with cattle domestication in Eurasia. Taken together, our results support the hypothesis that the -13910*T allele arose in Eurasia and reached its present distribution in a relatively short time due to the selective advantage of lifelong unrestricted use of milk. By applying a phylogeographic interpretation of the distribution of the haplotypes defined by the -13910*C/T and -22018*G/A polymorphisms

we were also able to predict an independent origin for lactase persistence mutations, in Europe and in the majority of African populations. In accordance with our prediction, several new sequence variants associated with lactase persistence were subsequently identified by others, in populations from Africa and Middle East: -14010*G/C, -13915*T/G and -13907*C/T. Given that adult milk consumption and lactase persistence seem to have been spread along with pastoralism, the distinct mutations constitute a tool to study the migratory events that led to the dispersion of pastoralist populations. In our study of the Bantu expansions, the identification of one of these mutations in Southwest Angola allowed us to explore alternative scenarios about the presence of the pastoralism in this region (see below).

The analysis of the levels and patterns of genetic variation in Angola and Mozambique allowed us to retrieve important insights into the demographic processes underlying the spread of the Bantu-speaking populations at both regional and continental geographic levels.

In the study of southwestern Angola we analysed the patterns of Y-chromosome, mtDNA and lactase persistence genetic variation in four representative West-Savanna Bantu-speaking groups (Ovimbundu, Ganguela, Nyaneka-Nkhumbi and Kuvale), relying on different combinations of agricultural and pastoral lifestyles. The results obtained with the Y-chromosome and mtDNA data indicated that, in spite of their peripheral location, the studied populations retained a clear genetic link to West-Central African populations from areas that are adjacent to the original homeland of the Bantu expansions. We observed also unequivocal signs of admixture with local Khoisan peoples, especially among the pastoralist Herero-speaking Kuvale, where the frequencies of Y-chromosome and mtDNA Khoisan lineages reached 12% and 22% values, respectively. These results, along with historical and archaeological data, highlight the relevance of the contacts between Bantu and Khoisan speakers in southwestern Angola. The evidence that Khoisan ethnic groups with pastoral subsistence had a significant presence in southwestern Africa before the arrival of Bantu-speaking populations supports the notion that Bantu-Khoisan interactions likely involved cattle herders from the two groups. Moreover, we found that the Kuvale were additionally characterized by the presence of the -14010*C lactase persistence allele, which likely originated in non-Bantu pastoralists from East Africa. The observation of Khoisan lineages along with the -14010*C allele in Bantu-speaking populations from southwestern Angola provides unique insights into the presence of pastoralism in this region. According to our interpretation, the link between East and Southwest African pastoral scenes was established indirectly, through migrations of Khoe herders across southern Africa.

We have further developed a set of 14 UEPSTR markers, consisting of a Unique Event Polymorphism (UEP) closely linked to a Short Tandem Repeat (STR), to analyse several

populations from Mozambique and Angola. The sample from Mozambique was countrywide and included 17 population groups representing most of the country's ethnolinguistic diversity. In Angola we analyzed the Ovimbundu and the Kuvale groups. We found that sampled genetic variation is not primarily structured between southwestern and southeastern regions of Africa, and a high genetic homogeneity between most Bantu populations was observed. Only the Kuvale from Angola and the Chopi from Mozambique were found to be major outliers. We have also inferred basic parameters of the Bantu expansions, by taking the Angolan and Mozambican populations as representative of the Southwest and Southeast branches of Bantu expansions, respectively. Estimates of migration rates based on STR multiloci showed that gene flow has played an important role in maintaining a close relationship between Bantu populations living in those regions (N_em~5). The study of uniparental markers additionally revealed that this gene flow seems to have been undertaken mainly by females (mtDNA: N_em >10; crY: N_em~0). Moreover, Y-chromosome and mtDNA data have shown that, while both male and female populations underwent significant size growth after the split between the western and eastern branches of Bantu expansions, males had lower population sizes than females throughout the Bantu dispersals. The close genetic relationship between most sampled Bantu populations is consistent with high degrees of interaction between peoples living in savanna areas located to the south of the rainforest. The genetic evidence accumulated so far is becoming increasingly difficult to reconcile with widely accepted models postulating an early split between eastern and western Bantu-speaking populations.

In São Tomé, our major aim was to study the patterns of population structure within the island without relying on predefined ethnic, anatomical, or geographical population categories. To this purpose, we analyzed 15 unlinked autosomal microsatellite loci in individuals from 14 localities across the island and used a Bayesian clustering approach to sort individuals into genetic clusters. The genetic variation in additional phylogeographic informative markers across inferred clusters was subsequently analysed in order to address the major factors that shaped the observed genetic structure. We found that, despite the fact that maximum distance between any two sampled sites was less than 50 km, São Tomé presents an unusual level of genetic structure. The genetic structuring observed was mainly caused by the grouping of Angolar Creole-speakers in a separate cluster carrying a distinctive imprint of genetic drift. The overall patterns of genetic variation suggest the occurrence of a kin-structured founder event, possibly resulted from the flight of a patrilineal clan of rebel slaves who established secondary contacts with the rest of the island mostly through restricted, female mediate gene flow.

Resumo

As principais contribuições das ferramentas genéticas para o conhecimento da história evolutiva humana podem ser divididas em duas grandes áreas. Uma das áreas compreende a identificação e caracterização de genes-alvo de selecção, fornecendo pistas sobre o modo como a espécie humana se tem adaptado a diferentes ambientes e estilos de vida. A outra área está relacionada com a reconstrução da história demográfica das populações humanas, e baseia-se principalmente na análise combinada de vários tipos de marcadores, com diferentes características e modos de transmissão.

Esta tese está dividida em três partes. Na primeira parte apresentamos um estudo sobre a história evolutiva da persistência da lactase, uma característica genética que se caracteriza pela capacidade de digerir a lactose durante toda a vida. Os benefícios nutricionais do leite parecem conferir uma vantagem selectiva aos indivíduos com persistência da lactase pertencentes a comunidades pastoris com elevada dependência deste alimento. Na segunda e terceira partes, focamo-nos em dois episódios da história demográfica de África: um relacionado com o estudo de populações de Moçambique e Angola, no contexto da migração em grande escala de populações de agricultores de língua Bantu, e outro, centrado na população da pequena ilha de São Tomé.

A nossa investigação acerca da persistência da lactase foi motivada pela identificação de um alelo (-13910*T) com forte associação com esta característica em indivíduos de origem finlandesa. Quando iniciámos o trabalho, a maioria dos estudos sobre a história evolutiva da persistência da lactase eram baseados na interpretação de correlações entre frequências fenotípicas e variáveis ambientais. Numa primeira fase, abordámos duas questões centrais sobre o alelo -13910*T, associado à persistência da lactase: a) testar a influência da selecção na sua actual distribuição geográfica, e b) estimar a idade do alelo. Para tal, caracterizámos a variação genética de haplótipos definidos por microssatélites na proximidade do alelo -13910*T, em várias populações de África e Europa de modo a realizar um teste formal de neutralidade selectiva. Com base neste teste concluímos que o alelo -13910*T é demasiado recente para que as suas frequências actuais possam ter sido atingidas sem favorecimento selectivo. A diversidade dos microssatélites associados ao alelo -13910*T foi também usada para calcular a idade absoluta do alelo. As nossas estimativas de idade indicam que o alelo -13910*T ter-se-á originado após a separação entre as populações africanas e europeias e situam a sua idade mínima no intervalo entre 12.500-7.500 anos. Esta estimativa associa a origem do alelo -13910*T com o início da domesticação de gado na Eurásia. No seu conjunto, a evidência recolhida apoia a hipótese de que o alelo -13910*T surgiu na Eurásia e atingiu a sua actual

distribuição num período de tempo relativamente curto devido à vantagem selectiva conferida pelo uso de leite ao longo de toda a vida. A interpretação filogeográfica da distribuição dos haplótipos definidos pelos polimorfismos -13910*C/T e -22018*G/A permitiu-nos concluir que uma origem independente para a persistência de lactase, na Europa e na maioria das populações de África, teria que ter ocorrido. De acordo com a nossa hipótese, novas variantes associadas à persistência da lactase foram posteriormente identificadas noutros estudos, em populações de África e do Médio Oriente: -14010*G/C, -13915*T/G e -13907*C/T. Uma vez que o consumo de leite na idade adulta e a persistência da lactase parecem estar associados à dispersão da pastorícia, as várias mutações identificadas constituem uma ferramenta para estudar os eventos migratórios que levaram à dispersão das diferentes populações de pastores nómadas. No âmbito do nosso estudo sobre as expansões Bantu, a identificação de uma destas mutações no sudoeste de Angola, permitiu-nos explorar cenários alternativos sobre a presença da pastorícia nesta região (ver abaixo).

A análise dos níveis e padrões de variação genética nas populações de Angola e Moçambique permitiu-nos fazer inferências acerca dos processos demográficos subjacentes à expansão das populações de língua Bantu tanto a nível regional como continental.

Para o estudo da população do sudoeste de Angola analisámos os padrões de variação genética ao nível do cromossoma Y (CRY), do ADN mitocondrial (ADNmt) e das mutações associadas à persistência da lactase. Estudamos quatro grupos representativos das línguas Bantu "West-Savanna" (Ovimbundu, Ganguela, Nyaneka-Nkhumbi e Kuvale), com diferentes intensidades de utilização da agricultura e da pastorícia. A análise dos dados do CRY e do ADNmt permitiu-nos concluir que, apesar da sua localização periférica, as populações estudadas retêm uma clara ligação genética a populações de África Ocidental e Central da região onde si iniciaram as expansões Bantu. Detectámos também sinais inequívocos de miscigenação com povos Khoisan locais, especialmente entre os pastores Kuvale de língua Herero, que apresentam frequências de linhagens Khoisan de 12% e 22% para o CRY e ADNmt, respectivamente. Estes resultados, juntamente com dados históricos e arqueológicos, ilustram a importância dos contactos entre grupos Bantu e Khoisan no sudoeste de Angola. Vários indícios apontam para que a presença de grupos étnicos Khoisan com elevada dependência da actividade pastoril no sudoeste de África seja anterior à chegada das populações de língua Bantu. Este cenário sugere que os contactos entre grupos Khoisan e Bantu tenham sido levados a cabo entre populações de pastores destes dois grupos étnicos. Verificámos também que os pastores Kuvale são ainda caracterizados pela presença do alelo -14010*C, associado à persistência da lactase, com provável origem em pastores não-Bantu da África Oriental. A observação simultânea de linhagens Khoisan e do alelo -14010*C em populações de

língua Bantu do sudoeste de Angola constitui um dado importante na questão da presença da pastorícia nesta região. De acordo com a nossa interpretação, o elo entre as cenas pastoris do oriente e sudoeste de África, terá sido estabelecido indirectamente, através da migração de pastores Khoe por toda a África Austral.

Desenvolvemos também um conjunto de 14 marcadores UEPSTR, cada um formado por um polimorfismo de ocorrência única (UEP) ligado a um microssatélite (STR), para a caracterização de diversas populações de Moçambique e Angola. A amostra de Moçambique incluiu 17 grupos, representando grande parte da diversidade etnolinguística deste país. Relativamente a Angola analisámos os grupos Kuvale e Ovimbundu. Verificámos que a variação genética não se encontra primariamente estruturada entre o sudoeste e sudeste de África, observando-se uma grande homogeneidade entre a maioria das populações analisadas. Os Kuvale de Angola e os Chopi de Moçambique constituem os dois principais "outliers". Estimámos também parâmetros básicos das expansões Bantu, usando as populações de Angola e Moçambique para representar, respectivamente, o ramo sudoeste e sudeste das expansões Bantu. Estimativas das taxas de migração baseadas na análise dos marcadores STRs mostraram que o fluxo de genes tem desempenhado um papel importante na manutenção da proximidade genética entre as populações Bantu que vivem nessas regiões (Nem ~ 5). O estudo de marcadores uniparentais revelou adicionalmente que as migrações entre as duas regiões foram empreendidas principalmente por elementos do sexo feminino (ADNmt: Nem> 10; CRY: Nem ~ 0). A análise dos dados do CRY e do ADNmt mostrou também que, após a separação entre os ramos ocidental e oriental das Expansões Bantu, tanto a fracção feminina como a masculina das populações sofreram um crescimento significativo, e que a fracção masculina apresentou menores tamanhos populacionais ao longo do tempo. A proximidade genética entre a maioria das populações de língua Bantu amostradas sugere um alto grau de interacção entre os povos que vivem nas áreas de savana localizadas para sul da floresta tropical. As evidências genéticas acumuladas até ao momento tornam-se cada vez mais difíceis de conciliar com os modelos amplamente aceites que defendem uma divergência inicial dos grupos localizados ao longo dos ramos oeste e este das expansões Bantu.

No estudo de São Tomé, o nosso principal objectivo era analisar o padrão de estrutura populacional da ilha sem ter em conta critérios pré-definidos de ordem étnica, anatómica ou geográfica. Para tal, analisámos 15 microssatélites autossómicos independentes em indivíduos de 14 localidades da ilha e usámos uma abordagem bayesiana para o agrupamento dos indivíduos em grupos genéticos. Estudamos também a variação genética dos grupos inferidos ao nível de marcadores filogeográficos informativos de modo a tentar perceber os principais factores na origem da estrutura genética detectada. Observamos que, apesar do facto de a

distância máxima entre dois quaisquer locais de amostragem ser inferior a 50 km, São Tomé apresenta uma estrutura genética bem marcada. A estrutura genética de São Tomé é causada pelo agrupamento de indivíduos de língua Angolar, apresentando um claro sinal de deriva genética. A análise global dos padrões de variação sugere a ocorrência de um evento fundador, levado a cabo por indivíduos com uma relação de parentesco próxima. Este evento fundador poderá ter sido resultado da fuga de um clã patrilineal de escravos que terá estabelecido, posteriormente, contactos com o resto da ilha, principalmente mediados por elementos do sexo feminino.

GENERAL INTRODUCTION

1. The study of human evolutionary history through the use of genetic tools

The present state of knowledge about the human evolutionary history results from contributions from different research areas such as paleontology, archaeology, history, physical anthropology, linguistics and, more recently, genetics (Jobling et al. 2004).

The major contributions provided by genetic tools can be divided in two main areas. One of these areas comprises the characterization of functionally relevant genes. It includes the determination of the time and place of origin of specific mutations and the study of the factors that have influenced its current frequency and geographic distribution. The identification of regions of the genome that have been targeted by positive selection is particularly important to provide insights into the processes by which human species have adapted to different environments and lifestyles (Sabeti et al. 2006). The other area, is related with the reconstruction of the demographic history of human populations, through the combined study of different types of markers with distinct characteristics and modes of transmission (Garrigan and Hammer 2006). Since the aim is to capture the general trends of genetic diversity, the specific properties of each locus are not so relevant.

1.1 Study of functionally relevant genes: the importance of genes underlying traits with adaptive relevance

The identification of molecular signatures of positive natural selection has constituted an important aspect of the study of genes with adaptive relevance. Several tests have been developed to detect signatures of positive selection in the human genome. Each signature is informative about selective processes that took place at specific windows of evolutionary time (Sabeti et al. 2006). For example, the comparison of functionally significant differences between species can give information about the selection episodes that occurred millions of years ago and that may have been important in the process of the differentiation of the human lineage (Nielsen et al. 2005). Studies based on these genetic signatures suggest that genes related to immune response, reproduction (e.g. spermatogenesis) and sensory perception (e.g. olfaction) may have played an important role in human evolution (Nielsen et al. 2005). On the other hand, genetic signatures like the observation of relatively large differences in allele frequencies between populations or the occurrence of unusually long haplotypes, allow to explore more recent selective episodes, helping to better understand the way human species have adapted to different environments and lifestyles (Sabeti et al. 2006). The loci that have been associated with these kind of signatures often involve traits related with human response to

regional and temporal changes in the climate (e.g. skin pigmentation, energy metabolism), pathogens (e.g. resistance to malaria, immune function) and diet (e.g. ability to digest milk) (Balaresque et al. 2007).

Until recently, inferences regarding positive selection were made almost exclusively through candidate gene studies (Akey 2009). Such studies were applied, for example, to lactase persistence (Ingram et al. 2009), associated with the ability to digest lactose throughout adulthood, and to a number of traits associated with reduced susceptibility to malaria, like the Duffy blood group (Hamblin and Di Rienzo 2000, Hamblin et al. 2002, Seixas et al. 2002) and the haemoglobin genes (Allison 1954, Ohashi et al. 2004).

The increasing availability of large-scale genotyping data is allowing a systematic survey of the genome for "signatures" of positive selection without an a priori hypothesis about which genes may be under selection (Akey 2009). Indeed, recent genome-wide studies have identified new potentially selected genomic regions and individual genes, and re-evaluating previously proposed candidates (e.g. Akey et al. 2004, Voight et al. 2006, Williamson et al. 2007, Barreiro et al. 2008, Pickrell et al. 2009). Despite the advances provided by genome wide scans of selection, they just represent the beginning of a longer process in the study of new positively selected loci (Akey 2009). In fact, the regions identified to be under selection are typically large, containing several genes and thousands of polymorphisms. It is thus necessary to undertake more detailed examinations to identify and characterize causal variants and to retrieve biological insights about their function. Considerable caution is also needed when interpreting genome-wide scans of selection given the likely detection of false positive cases and the low power associated with currently used approaches (Akey 2009). Recent studies have incorporated some improvements in the approaches applied in genome-wide scans of selection, by combining population genetics data with environmental information (Coop et al. 2010, Hancock et al. 2010) or by using multiple signals of selection simultaneously (Grossman et al. 2010).

Another important aspect of the genome-wide scans for selection is related with the populations that are studied. The occurrence of geographically restricted selective pressures led to regional-specificity of some adaptive traits. Consequently, the study of population panels that do not represent the worldwide genetic diversity may hinder our understanding of human adaptation. The fact that past genome wide scans for selection (reviewed in Akey 2009), have relied mainly on the study of panels with low number of representative populations (e.g. HapMap, Perlegen), highlights the need for increasing the set of populations included in future analyses.

Researchers are becoming increasingly aware of the high level of complexity underlying several adaptive traits (Akey 2009). Several studies have shown that genetic variants presenting signs of recent positive selection are not restricted to single-base-pair substitutions, but also include genomic rearrangements (Stefansson et al. 2005) and copy-number variants (Perry et al. 2007). Additionally, it has been observed that positive selection is not restricted to the protein-coding regions of the human genome, as adaptive regulatory variants have also been identified (Tournamille et al. 1995, Troelsen et al. 2003). Other related issues include the study of parallel (Ralph and Coop 2010) and polygenic (Pritchard et al. 2010) adaptation, epistatic selection (Williams et al. 2005) or selection acting at a post-transcriptional levels (Quach et al. 2009). Overall, it would be important to devise new approaches in order to incorporate these findings in the study of selective events.

There is sometimes the belief that, in contemporary societies, selective pressures caused by environmental changes are buffered through culture. This could be correct in some cases, like for the temperatures changes to which humans learnt how to counteract (Laland et al. 2010). However, there is increasing evidence showing that culture itself could be a source of selection in humans (Durham 1991, Laland et al. 2010, Stearns et al. 2010). In fact, genes under culturally derived selection have been associated with surprisingly high selection coefficients (Laland et al. 2010). The genetic adaptations to cultural changes associated with transitions in subsistence patterns provide good examples of gene-culture co-evolution. For example, the transition from hunter-gathering to agriculture and pastoralism changed many aspects of the human environment such as diet, lifestyle, population density or pathogen load, giving rise to new selective pressures (Hancock et al. 2010, Richerson et al. 2010, Scheinfeldt et al. 2010). Many genes have been identified as having been subject to selection in response to shifts in subsistence patterns, including the lactase gene (LCT), the amylase gene (AMY1) or the β-globin gene (HBB). While the first two examples are related with genetic adaptations to dietary specializations, the third is associated with protection against epidemic diseases (Laland et al. 2010). Several variants in the proximity of LCT have been associated with high lactase activity during all life. This trait allows pastoralist populations to respond to the selective environment created by dairy farming and is present at relatively high frequencies in these populations (Tishkoff et al. 2007). On the other hand, an improved digestion of starchy food (e.g. roots, tubers), associated with higher AMY1 copy number, has been found in populations with high-starch diets (Perry et al. 2007). The β-globin S (HBB*S) “sickle-cell” allele, which confers protection against malaria, underwent a rapid increase in some agriculturalist populations when they start clearing the forest to grow crops, creating the conditions to the breeding of malaria-carrying mosquitoes (reviewed in Durham 1991).

1.2 Study of demographic processes underlying human evolutionary history: the importance of the African continent

The research on genetic variation of present human populations is giving important contributions to our knowledge about the human demographic history, allowing the reconstruction of past events, such as, early and recent human migrations, population expansions and bottlenecks, and admixture between isolated groups (Jobling et al. 2004). In some cases, the information collected from genetic data played a decisive role in long-standing archaeological debates, like the discussion on the origin of modern humans (Stringer 2002). DNA evidence and fossil skeletal remains indicate that anatomically moderns humans arose in Africa ~200,000 years ago and that ~60,000 years ago a subset of these African populations may have started to disperse from northeastern Africa across the world (Mellars 2006). Several recent studies have also shown that worldwide patterns of genetic diversity are consistent with a model of colonization of the world through serial founder effects starting at a single African origin (Ramachandran et al. 2005, DeGiorgio et al. 2009, Deshpande et al. 2009).

The position of Africa as the most likely place of origin of all modern humans makes the study of African populations particularly important in the context of population history. However, the interest of this continent is not limited to the modern human origins debates. African populations are themselves an important topic. In fact, with its highly diverse and dynamic geographical, physical and human features (Figures 1 and 2), Africa harbors a great variety of genetic patterns that remain to be explored (Reed and Tishkoff 2006).

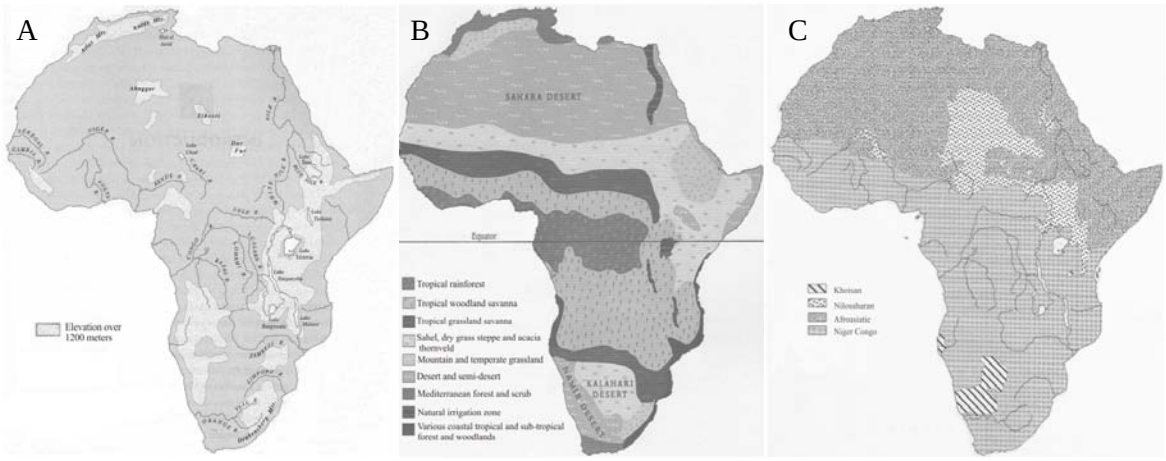


Figure 1 Maps showing different aspects of Africa. (A) Rivers and major elevations; (B) vegetation; (C) language distributions. (A) and (C) were retrieved from Newman (1995) and (B) from Reader (1997).

African landscapes range from the driest deserts to the most humid forests. The African dwellers are divided in more than 2,000 ethnic and linguistic groups, with contrasting social practices and subsistence patterns that include various modes of agriculture, pastoralism and some of the last remnants of the world's hunter-gatherers (Reed and Tishkoff 2006). Extensive population movements have occurred through and from Africa both in ancient and historical times, leading to contact and admixture between groups that had been separated for thousands years and, in some cases, resulting in the peopling of uninhabited areas (Newman 1995). One of the most significant migration events in recent African history was the expansion of Bantu-speaking agriculturalists ~5,000 years ago from its homeland around Cameroon/Nigeria, which extensively reshaped the human landscape of the continent south of the Sahara (Newman 1995) (Figure 2). Africa is also the place of origin of millions of individuals that were caught in the Atlantic slave trade. The study of the populations that are known to have been involved in these "migrations" is providing important clues to understand the history of a multitude of populations arising from the African Diaspora (Klein 1999) (Figure 2).

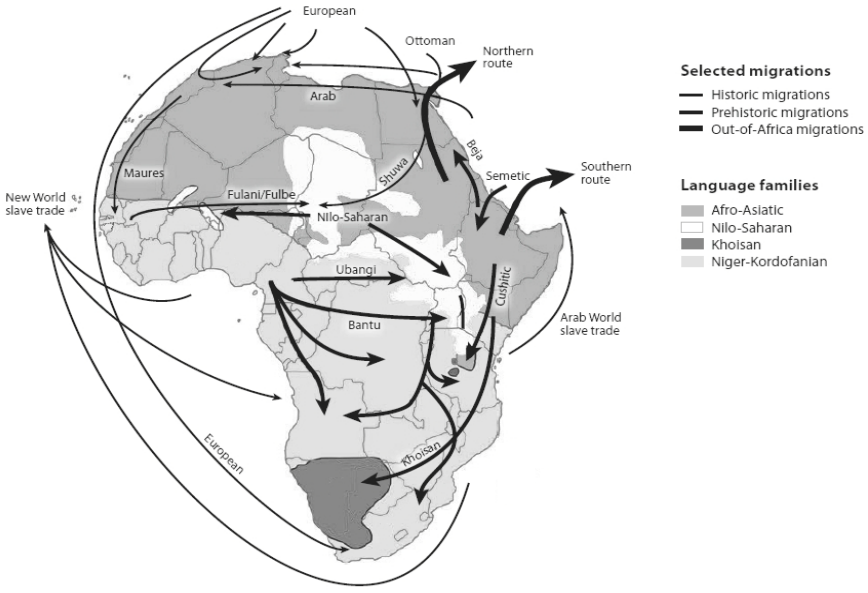


Figure 2 Map depicting the routes of selected migration events within and out of Africa. Each language family distribution is identified by a different grade. The diagram was adapted from Campbell and Tishkoff (2008).

The study of African populations also presents a great potential for the identification of genes underlying the resistance/susceptibility to a series of diseases currently affecting human populations. Given the long-term impact of infectious diseases in Africa (e.g. malaria, tuberculosis), the identification of genetic determined resistance to infection in African populations, may yield insights into devising strategies to struggle against some diseases. In addition, the high levels of genetic diversity and low level of linkage disequilibrium observed in African populations, make them highly informative for the identification of the genetic factors underlying some complex diseases (Campbell and Tishkoff 2008).

Two major issues concerning genetic studies in Africa can be pointed out: one related with the type of markers commonly used and the other with sample coverage.

In what concerns the type of markers used, a significant part of our present understanding of African genetic variation is based on the study of mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY) (e.g. Cruciani et al. 2002, Salas et al. 2002). Because of their uniparental patterns of inheritance and lower effective population size, mtDNA and NRY haplotypes provide complementary information about female- and male-specific aspects of genetic variation and are especially sensitive to the effects of drift. MtDNA and NRY markers tend to be highly geographically structured and, due to lack of recombination, haplotype phylogenies can be easily reconstructed, providing a temporal framework for mutation accumulation, which can be related to the geographic distribution of different lineages. Several NRY and mtDNA haplogroups are particularly informative because their origins appear to be geographically and temporally distinct from each other. However, despite the high informativeness of the uniparentally inherited mtDNA and NRY in the study of sex-specific aspects of human history, each of these fragments behave evolutionary as single locus, which, due to the stochasticity of evolutionary processes, is insufficiently robust to generate meaningful estimates of relevant population history parameters. To benefit from the evolutionary information from other regions of the genome, multilocus approaches based on several independently evolving genetic systems are clearly needed.

Recently, several studies have tried to uncover the African evolutionary history by analysing multi-locus patterns of genetic variation throughout several African populations (Campbell and Tishkoff 2008). These studies were based on either unique event polymorphisms (UEPs) or short tandem repeat (STR) data. Different empirical and theoretical studies have shown that the combination of UEPs and STRs in compound autosomal haplotype systems (UEPSTRs) may counterbalance the limitations of each marker type and maximize their specific advantages (Hey et al. 2004, Ramakrishnan and Mountain 2004, Payseur and Cutter 2006). For example, simulation data showed that autosomal UEPSTRs provide improved estimates of

population divergence times and may be especially useful for characterizing population size changes (Ramakrishnan and Mountain 2004; Payseur and Cutter 2006). Mountain et al. (2002) have developed a general approach to identify and genotype multiple UEPSTR loci to be used in human evolutionary genetics. However the use of multilocus UEPSTR marker sets to address specific population history questions is still not widespread.

In what concerns sample coverage, inferences about human population history typically relied on few African populations that were assumed to be representative of the whole continental diversity. For example, the Human Genome Diversity Panel (HGDP) of populations (Cann et al. 2002), widely used in worldwide genetic studies, includes only eight populations from Africa (Mozabite, Mandenka, Yoruba, Biaka Pygmies, Mbuti Pygmies, Kenyan and South African Bantus and San), and do not reflect the full extent of African human diversity (Campbell and Tishkoff 2008). A series of recent studies of African populations have been based on this panel (e.g. Ramachandran et al. 2005, Ray et al. 2005, Rosenberg et al. 2005). While the limited number of populations used did not challenge the validity of general conclusions about the origins and global distribution of human genetic variability, insufficient sampling has certainly hampered our perception of how human diversity was shaped within Africa (Slatkin 2005). In an effort to characterize the genetic variation and the relationships among African populations, a recent study genotyped a panel of 1,327 polymorphic markers in an extended sample of 113 geographically diverse populations in Africa (Tishkoff et al. 2009). This work constitutes a landmark in the study of the African evolutionary history, especially as concerns to East African populations. In spite of the major advance provided by this study, it is important to note that regions like the Sahel, the Atlantic West Africa, Namibia, Angola and the central corridor comprising the Democratic Republic of Congo, Central Zimbabwe and the Zambia, remain sparsely sampled. Additional sampling of these and other areas will be important in order to achieve a more complete picture of the patterns of genetic diversity present in Africa.

1.3 Integrating adaptive and demographic histories

Studies about selection and demography have been normally regarded as separated fields. However, it is important to integrate both areas in order to achieve a more accurate picture of the human evolutionary history.

On the one hand, knowledge of the demographic history of human populations is essential to study positively selected loci by providing the general expectations under neutral evolution. Recent studies have shown that geography and population history have played an important role in the present distribution of selected alleles (Coop et al. 2009, Pickrell et al. 2009). Additionally, it has been observed that the pattern of geographic distribution of a selected allele, coupled in the framework of a known demographic history, could give insights into the strength of positive selection (Coop et al. 2009, November and Di Rienzo 2009).

On the other hand, loci under selection can be very informative in the context of the demographic studies. When selective pressures occur only in geographic restricted areas or affect populations with specific lifestyles, selected mutations may became frequent in some specific regions while remaining virtually inexistent in others. Such a geographic segregation makes these selected mutations particularly informative about recent migration events. The S allele of β-globin gene is one of such "migration markers". The high level of geographic segregation observed for the haplotypes associated with β-globin S allele (HBB*S) – called Benin, Bantu, Arab-Indian and Senegal- makes them particularly useful for assessing the regional ancestry of lineages bearing these variants and dispersed worldwide (Nagel and Ranney 1990) (Figure 3A). At another scale, the -13910*T and the Duffy-null (FY*O) alleles, can be very informative about intercontinental migrations. The -13910*T allele was found to be especially frequent in populations from Europe and Middle East (Ingram et al. 2009). Its presence in Africa has been explained as the result of recent gene flow into this continent (Figure 3B). For example, the presence of this mutation in North African Berber populations seems to represent the genetic signature of a past migration of pastoralists from the Middle East (Myles et al. 2005). On the other hand, the FY*O allele is almost fixed in the majority of sub-Saharan African populations and is virtually absent in Europe (Cavalli-Sforza et al. 1994) (Figure 3C), constituting a highly informative marker for evaluating admixture with sub-Saharan individuals.

Figure 3 Examples of mutations/haplotypes presenting regional specificity. (A) Distribution of the four major HBB*S linked haplotypes; (B) Worldwide distribution of the -13910*T lactase persistence allele. The darker the colour, the higher the frequency of the -13910*T allele (figure modified from Ingram et al. 2009); (C) Distribution of the FY*O allele in Africa (figure retrieved from Cavalli-Sforza et al. 1994).

2. Work outline

In this work we use genetic tools to address several aspects of the human evolutionary history- including adaptive and demographic episodes.

In the first part we focus on the evolutionary history of lactase persistence, a genetically determined trait, characterized by the ability to digest lactose during all life. This trait has long been recognized as an example of dietary adaptation in human species. It is thought to provide a selective advantage due to the added nutritional benefits of drinking milk in dairying populations (Simoons 1970, McCracken 1971). The co-evolution of dairy farming and lactase persistence constitutes one of the best examples of the influence that culturally derived selection can have on human genes (Durham 1991). Although the genetic inheritance of lactase persistence was already recognized in the early 1970s (Sahi et al. 1973), the molecular mechanisms behind lactase expression remained unknown until recently. The identification of the -13910*T allele as the putative variant responsible for the lactase persistence in northern Europeans (Enattah et al. 2002) created the possibility to further study several functional and evolutionary aspects related to this trait. In this part of the work we studied the evolutionary history of lactase persistence based on the analysis of the haplotypic variation associated with the -13910*T allele. Our major goal was to formally test the role of positive selection in the present distribution of the -13910*T allele and to estimate its age. We also developed a filogeographic hypothesis about the possibility that multiple mutations associated with lactase persistence had occurred in different geographic regions. The results are described in *Article 1* and include work partially carried out in my master's thesis (Coelho 2005). Posterior developments related with the identification of several mutations associated with lactase persistence in Africa (Tishkoff et al. 2007), allowed us to explore alternative scenarios about the presence of the pastoralism in Southwest Africa (see below).

In the second and third parts of this thesis, two episodes of African demographic history are studied: the Bantu expansions and the Atlantic Slave trade.

The dispersal of Bantu-speaking agriculturalists stands as one of the most impressive examples of long-range human migrations. These expansions have had a major influence on biological and cultural diversity in sub-Saharan Africa. It is generally accepted that Bantu expansions started 4,000-5,000 years ago in the area between Cameroon and Nigeria (Newman 1995). However there is still no consensus about many aspects of the history of Bantu populations, including the major dispersal routes followed by Bantu speakers and the nature of the interactions between spreading populations. We focused on populations from Southwest Angola and Mozambique, located, respectively, at the southwestern and southeastern of the two

principal branches of Bantu expansions. In addition we developed a new set of compound haplotype systems (UEPSTRs), each consisting of a rapidly evolving short tandem repeat (STR) closely linked to a unique event polymorphism (UEP), to use a multilocus approach in the characterization of our sample of Bantu-speaking populations. Our analysis of the levels and patterns of genetic variation in Bantu-speaking populations allowed us to discuss the demographic scenarios that better explain the observed genetic diversity. The genetic differentiation between groups located in southwest Angola and Mozambique was also used to examine the implications of the spread of Bantu speakers at more regional contexts. Our results are presented in *Articles 2 and 3*. In *Article 2*, we used mtDNA and NRY uniparental markers to study several Bantu-speaking groups from the southwest of Angola and further combined our results with published data from Southeast Africa and other regions of Africa. Given that our samples from Southwest Angola were collected in a region where the extinct click language Kwadi was reported to be spoken, this study was also important to characterize the contacts between advancing farmers and retreating hunter-gatherers in this region of Africa. Additionally, the presence in this region of Bantu-speaking groups depending on cattle raising to different degrees constituted an opportunity to investigate the relationships between the southwestern Angolan and other African pastoral scenes. This issue was further investigated by screening different lactase persistence associated alleles in different ethnic groups from this region. In *Article 3*, we studied the genetic structure of 19 Bantu-speaking groups from Mozambique and Angola using a multilocus approach based on the 14 newly developed compound UEPSTRs systems. This study allowed us to additionally compare the ability of UEP-only, STR-only and joint UEPSTR datasets to document genetic variation both at the intercontinental level and among the African Bantu-speaking populations.

The shipping of African slaves during the Atlantic slave trade represents one of the largest intercontinental “migrations” in human history (Klein 1999), leading to new forms of social encounter (Curtin 1998). The study of recent populations emerged from the slave trade may provide unique opportunities for studying the evolutionary determinants that modelled human cultural and biological variation at a relatively small temporal scale, in different geographic settings. The peopling of the island of São Tomé, in the Gulf of Guinea, is intimately related with the Atlantic slave trade. São Tomé played a crucial role as a slave entrepôt and became one of the first examples of the plantation complex that spread into the tropical New World (Curtin 1998). Moreover, the diversity of contributions to the peopling of São Tomé has promoted intense cultural interactions that resulted in the emergence of distinct autochthonous creoles (Forro and Angolar) that are still widely spoken. Driven by this combination of interesting features, we undertook a fine scale analysis of the genetic structure

of São Tomé in order to understand how different evolutionary factors have shaped current genetic variability in the island (*Article 4*). In this study we paid special attention to the sampling approach and used a study design that inverted the sequence by which the relationships between genetic and cultural variation are usually investigated.

References

Akey, J. M., M. A. Eberle, M. J. Rieder, C. S. Carlson, M. D. Shriver, D. A. Nickerson, and L. Kruglyak. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol* 2:e286.

Akey, J. M. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res* 19(5):711-22.

Allison, A. C. 1954. Protection afforded by sickle-cell trait against subtertian malareal infection. *Br Med J* 6;1(4857):290-4.

Balaresque, P. L., S. J. Ballereau, and M. A. Jobling. 2007. Challenges in human genetic diversity: demographic history and adaptation. *Hum Mol Genet* 16 Spec No. 2:R134-9.

Barreiro, L. B., G. Laval, H. Quach, E. Patin, and L. Quintana-Murci. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet* 40:340-5.

Campbell, M. C., and S. A. Tishkoff. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* 9:403-33.

Cann, H. M., C. de Toma, L. Cazes, M. F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. 2002. A human genome diversity cell line panel. *Science* 296:261-2.

Cavalli-Sforza, L. L., P. Menozzi and A. Piazza. 1994. *The history and geography of human genes*: Princeton University Press, Princeton, New Jersey.

Coelho, M. 2005. Human lactase persistence: Evaluation of concordance between the breath hydrogen test and molecular genotyping; Analysis of evolutionary history using a microsatellite approach, Master's thesis. Faculdade de Ciências, University of Porto.

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, D. Absher, R. M. Myers, L. L. Cavalli-Sforza, M. W. Feldman, and J. K. Pritchard. 2009. The role of geography in human adaptation. *PLoS Genet* 5:e1000500.

Coop, G., D. Witonsky , A. Di Rienzo, and J. K. Pritchard. 2010. Using environmental correlations to identify loci underlying local adaptation. *Genetics* [Epub ahead of print].

Cruciani, F., P. Santolamazza, P. Shen, V. Macaulay, P. Moral, A. Olckers, D. Modiano, S. Holmes, G. Destro-Bisol, V. Coia, D. C. Wallace, P. J. Oefner, A. Torroni, L. L. Cavalli-Sforza, R. Scozzari, and P. A. Underhill. 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 70:1197-214.

Curtin, P. 1998. The rise and fall of the plantation complex. Cambridge: Cambridge University Press.

DeGiorgio, M., M. Jakobsson, and N. A. Rosenberg. 2009. Explaining worldwide patterns of human genetic variation using a coalescent-based serial founder model of migration outward from Africa. *Proc Natl Acad Sci U S A* 106:16057-62.

Deshpande, O., S. Batzoglou, M. W. Feldman, and L. L. Cavalli-Sforza. 2009. A serial founder effect model for human settlement out of Africa. *Proc Biol Sci* 276:291-300.

Durham, W. H. 1991. "Cultural Mediation: The Evolution of Adult Lactose Absorption," in *Coevolution: Genes, Culture and Human diversity*. Stanford: Stanford University Press.

Enattah, N. S., T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Jarvela. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30:233-7.

Garrigan, D., and M. F. Hammer. 2006. Reconstructing human origins in the genomic era. *Nat Rev Genet* 7:669-80.

Grossman, S. R., I. Shylakhter, E. K. Karlsson, E. H. Byrne, S. Morales, G. Frieden, E. Hostetter, E. Angelino, M. Garber, O. Zuk, E. S. Lander, S. F. Schaffner, and P. C. Sabeti. 2010. A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* 12;327(5967):883-6.

Hamblin, M. T., and A. Di Rienzo. 2000. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am J Hum Genet* 66(5):1669-79.

Hamblin, M. T., E. E. Thompson, and A. Di Rienzo. 2002. Complex signatures of natural selection at the Duffy blood group locus. *Am J Hum Genet* 70(2):369-83.

Hancock, A. M., D. B. Witonsky, E. Ehler, G. Alkorta-Aranburu, C. Beall, A. Gebremedhin, R. Sukernik, G. Utermann, J. Pritchard, G. Coop, and A. Di Rienzo. 2010. Human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proc Natl Acad Sci U S A* 11;107 Suppl 2:8924-30.

Hey, J., Y.J. Won, A. Sivasundar, R. Nielsen, and J. A. Markert. 2004. Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Mol Ecol* 13: 909-919.

Ingram, C. J., C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow. 2009. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet* 124:579-91.

Jobling, M. A., M. E. Hurles, and C. Tyler-Smith. 2004. *Human Evolutionary Genetics: origins people & disease*: Garland Science.

Klein, H. S. 1999. The Atlantic slave trade. Cambridge: Cambridge University Press.

Laland, K. N., J. Odling-Smee, and S. Myles. 2010. How culture shaped the human genome: bringing genetics and the human sciences together. *Nat Rev Genet* 11:137-48.

McCracken, R. D. 1971. Lactase deficiency: an example of dietary evolution. *Curr Anthropol* 12:479–517.

Mellars, P. 2006. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science* 313:796-800.

Mountain, J.L., A. Knight, M. Jobin, C. Gignoux,.A. Miller, A. A. Lin, and P. A. Underhill. 2002. SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome Res* 12:1766-1772.

Myles, S., N. Bouzekri, E. Haverfield, M. Cherkaoui, J. M. Dugoujon, and R. Ward. 2005. Genetic evidence in support of a shared Eurasian-North African dairying origin. *Hum Genet* 117:34-42.

Nagel, R. L., and H. M. Ranney. 1990. Genetic epidemiology of structural mutations of the beta-globin gene. *Semin Hematol* 27:342-59.

Newman, J. L. 1995. *The peopling of Africa: a geographic interpretation*: Yale University Press New Haven and London.

Nielsen, R., C. Bustamante, A. G. Clark, S. Glanowski, T. B. Sackton, M. J. Hubisz, A. Fledel-Alon, D. M. Tanenbaum, D. Civello, T. J. White, J. Sninsky, M. D. Adams, and M.

Cargill. 2005. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLoS Biol* 3:e170.

Novembre, J., and A. Di Rienzo. 2009. Spatial patterns of variation due to natural selection in humans. *Nat Rev Genet* 10:745-55.

Ohashi, J, I. Naka, J. Patarapotikul, H. Hananantachai, G. Brittenham, S. Looareesuwan, A. G. Clark, and K. Tokunaga. 2004. Extended linkage disequilibrium surrounding the hemoglobin E variant due to malarial selection. *Am J Hum Genet* 74(6): 1198–1208.

Payseur, B.A., and A.D. Cutter. 2006. Integrating patterns of polymorphism at SNPs and STRs. *Trends Genet* 22:424-429.

Perry, G. H., N. J. Dominy, K. G. Claw, A. S. Lee, H. Fiegler, R. Redon, J. Werner, F. A. Villanea, J. L. Mountain, R. Misra, N. P. Carter, C. Lee, and A. C. Stone. 2007. Diet and the evolution of human amylase gene copy number variation. *Nat Genet* 39:1256-60.

Pickrell, J. K., G. Coop, J. Novembre, S. Kudaravalli, J. Z. Li, D. Absher, B. S. Srinivasan, G. S. Barsh, R. M. Myers, M. W. Feldman, and J. K. Pritchard. 2009. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res* 19:826-37.

Pritchard, J. K., J. K. Pickrell, and G. Coop. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol* 20:208-215.

Quach, H., L. B. Barreiro, G. Laval, N. Zidane, E. Patin, K. K. Kidd, J. R. Kidd, C. Bouchier, M. Veuille, C. Antoniewski, and L. Quintana-Murci. 2009. Signatures of purifying and local positive selection in human miRNAs. *Am J Hum Genet* 84:316-27.

Ralph, P. L. and G. Coop. 2010. Parallel Adaptation: One or many waves ofadvance of an advantageous allele? *Genetics* [Epub ahead of print].

Ramakrishnan, U., and J.L. Mountain 2004. Precision and accuracy of divergence time estimates from STR and SNPSTR variation. *Mol Biol Evol* 21:1960-1971.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 102:15942-7.

Ray, N., M. Currat, P. Berthier, and L. Excoffier. 2005. Recovering the geographic origin of early modern humans by realistic and spatially explicit simulations. *Genome Res* 15:1161-7.

Reader, J. 1997. *Africa- A biography of the continent*. New York: Vintage Books.

Reed, F. A., and S. A. Tishkoff. 2006. African human diversity, origins and migrations. *Curr Opin Genet Dev* 16:597-605.

Richerson, P. J, R. Boyd, and J. Henrich. 2010. Gene-culture coevolution in the age of genomics. *Proc Natl Acad Sci U S A* 11;107 Suppl 2:8985-92.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. W. Feldman. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genet* 1:e70.

Sabeti, P. C., S. F. Schaffner, B. Fry, J. Lohmueller, P. Varilly, O. Shamovsky, A. Palma, T. S. Mikkelsen, D. Altshuler, and E. S. Lander. 2006. Positive natural selection in the human lineage. *Science* 312:1614-20.

Sahi, T., M. Isokoski, J. Jussila, K. Launiala, and K. Pyorala. 1973. Recessive inheritance of adult-type lactose malabsorption. *Lancet* 2:823-6.

Salas, A., M. Richards, T. De la Fe, M. V. Lareu, B. Sobrino, P. Sanchez-Diz, V. Macaulay, and A. Carracedo. 2002. The making of the African mtDNA landscape. *Am J Hum Genet* 71:1082-111.

Scheinfeldt, L. B., S. Soi, and S. A. Tishkoff. 2010. Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. *Proc Natl Acad Sci U S A* 11;107 Suppl 2:8931-8.

Seixas, S., N. Ferrand, and J. Rocha. 2002. Microsatellite variation and evolution of the human Duffy blood group polymorphism. *Mol Biol Evol* 19(10):1802-6.

Simoons, F. J. 1970. Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis. *Am J Dig Dis* 15:695-710.

Slatkin, M. 2005. Seeing ghosts: the effect of unsampled populations on migration rates estimated for sampled populations. *Mol Ecol* 14:67-73.

Stearns, S. C., S. G. Byars, D. R. Govindaraju, and D. Ewbank. 2010. Measuring selection in contemporary human populations. *Nat Rev Genet* 11(9):611-22.

Stefansson, H., A. Helgason, G. Thorleifsson, V. Steinthorsdottir, G. Masson, J. Barnard, A. Baker, A. Jonasdottir, A. Ingason, V. G. Gudnadottir, N. Desnica, A. Hicks, A. Gylfason, D. F. Gudbjartsson, G. M. Jonsdottir, J. Sainz, K. Agnarsson, B. Birgisdottir, S. Ghosh, A. Olafsdottir, J. B. Cazier, K. Kristjansson, M. L. Frigge, T. E. Thorgeirsson, J. R. Gulcher, A. Kong, and K. Stefansson. 2005. A common inversion under selection in Europeans. *Nat Genet* 37:129-37.

Stringer, C. 2002. Modern human origins: progress and prospects. *Philos Trans R Soc Lond B Biol Sci* 357:563-79.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas. 2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39:31-40.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, and S. M. Williams. 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035-44.

Tournamille, C., Y. Colin, J. P. Cartron, and C. Le Van Kim. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nat Genet* 10:224-8.

Troelsen, J. T., J. Olsen, J. Moller, and H. Sjostrom. 2003. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology* 125:1686-94.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.

Williams, T. N., T. W. Mwangi, S. Wambua, T. E. Peto, D. J. Weatherall, S. Gupta, M. Recker, B. S. Penman, S. Uyoga, A. Macharia, J. K. Mwacharo, R. W. Snow, and K. Marsh. 2005. Negative epistasis between the malaria-protective effects of alpha+-thalassemia and the sickle cell trait. *Nat Genet* 37(11):1253-7.

Williamson, S. H., M. J. Hubisz, A. G. Clark, B. A. Payseur, C. D. Bustamante, and R. Nielsen. 2007. Localizing recent adaptive evolution in the human genome. *PLoS Genet* 3:e90.

PART 1

The evolutionary history of lactase persistence

1.1. Introduction

1.1.1 Geographic distribution of lactase persistence and evolutionary hypotheses

The ability to digest lactose in adulthood (lactase persistence) is an autosomal dominant trait characterized by the maintenance of high levels of the lactase enzyme activity beyond the suckling phase (Flatz 1987). Lactase persistence (LP) is the derived trait, since in most mammals and in the majority of humans, lactase activity declines during childhood as part of the normal developmental regulation of the enzyme, a condition called lactase restriction (Flatz 1987).

The diagnosis of the lactase activity phenotype may be done through the use of physiological tests. These tests are based either on the direct determination of lactase activity from intestinal biopsies or on the quantification of metabolites that are formed when lactose is not digested (e.g. the hydrogen gas). Studies of the prevalence of the LP phenotype worldwide have shown that the frequency of this trait is highly variable in human populations and appears to be positively correlated with the importance of the milk in diet (Figures I.1) (Ingram et al. 2009).

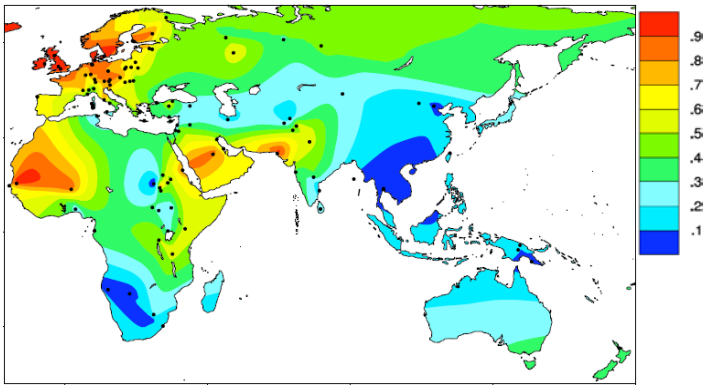


Figure I.1 Interpolated map of Old World lactase persistence frequencies. Dots represent the collection locations. Colours and colour keys show the frequencies of the LP phenotype estimated by surface interpolation (figure retrieved from Itan et al. 2010).

In Europe, lactase persistence is most frequent in the north- where there is a long history of milk dependent cattle pastoralism- and gradually decreases towards the south and east of the continent (Flatz 1987, Swallow 2003) (Fig.I.1). The association between milk dependent cattle pastoralism and lactase persistence in Europe is supported by archaeological data, suggesting that the earliest centres of European cattle pastoralism were located in the north part of the continent (Beja-Pereira et al. 2003). The importance of this region in the context of the pastoralism was also confirmed by the observation that the cattle from Northern Europe present the highest levels of genetic diversity in milk proteins, likely due to the selection for increased milk yield in this region (Beja-Pereira et al. 2003).

Outside Europe, lactase persistence is usually rare. However, the trait is present at high frequencies in nomadic pastoralist communities from Africa and Middle East (like the Fulbe from Senegal, Beja from Sudan, Bedouins from Saudi Arabia, Maasai from Tanzania or the Rendille from Kenya) (Itan et al. 2010). There are, however, pastoralist groups who present low frequencies of lactase persistence (Figure I.2A), like the Nilotic tribes of southern Sudan (Swallow 2003). In these cases, there is a cultural adjustment to the inability to digest lactose through the reduction of fresh milk drinking and increased consumption of fermented milk products with low lactose content (Durham 1991, Holden and Mace 2002). Consequently, the prevalence of lactase persistence is more strongly correlated with fresh milk consumption than with the degree of pastoralism dependence (Figure I.2B).

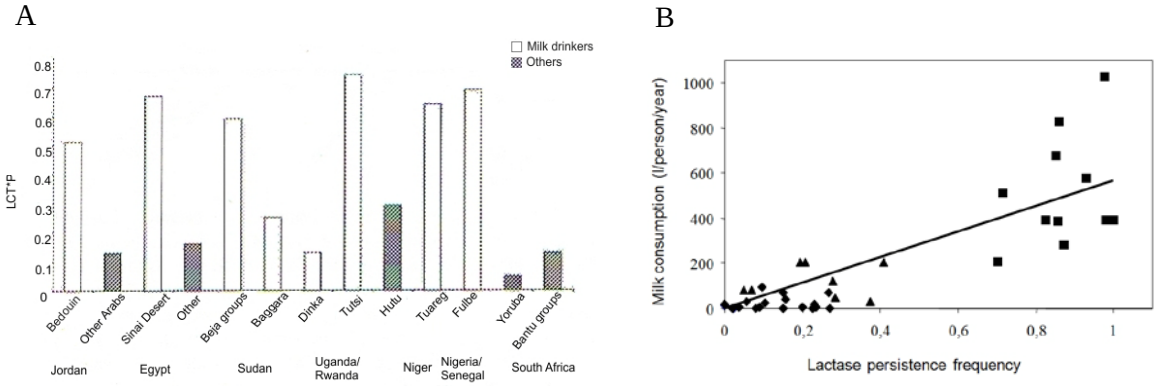


Figure I.2 A. Frequency distribution of the lactase persistence allele (LCT*P) in different Arab and African groups. Milk-drinking pastoralists (white) are compared with non-milk consumers (black) from neighbouring communities (Swallow 2003). B. Correlation between the frequencies of LP and the milk consumption. The squares and the triangles represent pastoral populations with LP frequencies higher than 60% and lower than 40%, respectively. Non-pastoral populations are represented as diamonds (data was compiled by Durham 1991).

The major hypothesis that has been raised to explain the correlation between pastoral tradition, milk consumption and lactase persistence is the "Culture-historical hypothesis", whereby the nutritional benefits of the milk during adult life are thought to confer a selective advantage in groups that adopted dairying (Simoons 1970, McCracken 1971). The "Culture-historical hypothesis" emphasizes the role of culture in the evolution of lactase persistence. The new selective pressures arising in dairying populations conferred an higher fitness to those individuals who were lactase persistent and this would lead, in turn, to the increase of the frequency of this trait in those populations (Durham 1991). The selective advantage of the lactase persistence trait would derive from what Simoons (1970) calls "the general nutritional advantage" of fresh milk consumption, of special relevance in pastoral nomads with high dependence upon fresh milk (Durham 1991). Additional hypotheses stressing other benefits of the milk have been proposed (Swallow 2003). The improvement of the calcium assimilation provided by milk lactose at high-latitude regions constitutes the core argument of the so-called "Calcium-absorption hypothesis" (Flatz and Rotthauwe 1973). According to this hypothesis, the consumption of milk during lifetime would be advantageous in regions with low incidence of ultraviolet B radiation, where milk lactose compensate for low vitamin D photosynthesis. Indeed, it has been observed that lactose of fresh milk act physiologically like a vitamin D supplement, facilitating the absorption of calcium from the small intestine (Durham 1991). This hypothesis would explain, for example, the high frequencies of lactase persistence observed in northwestern Europe but could not be applied to other milk-dependent pastoralist groups, showing high prevalence of lactase persistence regardless of their homeland latitude (Durham 1991). Another hypothesis is focused on the value of milk as a source of water in arid environments, important to guarantee adequate hydration and the electrolyte balance (Flatz 1987, Holden and Mace 2002). This hypothesis is supported by the high lactase persistence frequencies found in pastoralist groups in arid regions from Middle East and North Africa (Holden and Mace 2002). Contrasting with the previous hypotheses, which implicate a selective advantage for milk consumption, the so called "Reverse cause argument" suggests that the increase of lactase persistence in some human populations is unrelated to milk use and that dairying was adopted precisely by those populations that could tolerate lactose (reviewed in Aoki 2001). Basically, this proposal differs from the previous ones in the temporal priority given to cultural or genetic change.

1.1.2. The molecular basis of lactase persistence

Although the genetic inheritance of lactase persistence was already recognized in the early 1970s (Sahi et al. 1973), the molecular mechanisms behind the developmental regulation of the lactase expression remained unknown until recently. Initial efforts to identify the cause of the persistence/restriction variation led to the identification of several DNA polymorphisms within the lactase gene (LCT) and in neighbouring promoter regions, but none showed the appropriate phenotype-genotype correlation (reviewed in Swallow 2003). It was only in 2002 that Enattah et al. (2002) identified a genetic polymorphism located 13,910 base pairs 5' of the initiation codon of LCT (-13910*C/T) that was completely associated with lactase persistence in individuals of Finnish origin. A similar, but less strong association, was also found with a G/A single nucleotide polymorphism (-22018*G/A), about 8kb further upstream. These two single nucleotide polymorphisms (SNPs) are located within introns of the neighbouring minichromosome maintenance-6 gene (MCM6) and illustrate the complexity of the regulation of LCT by showing how apparently silent DNA variants in non-coding regions of the genome may play unexpected functional roles many kilobases away.

Several studies have assessed functional differences between the two alleles (C and T) at the -13910 position seem to support the notion that this SNP is causative of the lactase persistence/restriction variation. *In vitro* studies have shown that the -13910*T allele has an increased enhancer effect in activating lactase promoter activity (Olds and Sibley 2003, Troelsen et al. 2003). The analysis of intestinal biopsy samples detected that the -13910*T allele was associated with higher lactase mRNA expression levels (Kuokkanen et al. 2003, Rasinperä et al. 2005). Functional *in vitro* studies have further shown that several nuclear transcription factors (e.g. Oct-1) present a higher affinity to the sequences harbouring the T allele of the -13910*C/T SNP, directing increased lactase promoter activity (Troelsen et al. 2003, Lewinsky et al. 2005) . It became then clear that the primary mechanism for both the lactase persistence and restriction was the regulation of gene transcription. Based on these results it has been proposed that a decreased availability of transcription factors after weaning would explain the postweaning down-regulation of the lactase expression. According to this scenario, the stronger enhancer effect provided by the -13910*T variant would compensate for these changes, enabling the high expression of lactase throughout adulthood (Troelsen et al. 2003).

1.1.3 The -13910*T allele as a tool to study the evolutionary history of lactase persistence

The finding of the -13910*T allele as a robust marker for lactase persistence in Eurasian populations added a new dimension to the studies of the evolutionary history of lactase persistence which, until then, were essentially based on the interpretation of correlations between phenotypic frequencies and environmental variables (Durham 1991). Some studies based on the haplotype analysis of SNPs surrounding the -13910*T allele were undertaken. It was shown that the -13910*T allele lies in a SNP-defined haplotype extending over at least 1 Mb in Europeans, suggesting a recent origin for the polymorphism (Poulter et al. 2003). The observation of an unusually extended haplotype associated with the -13910*T allele in Northern Europe was consistent with a model whereby the high frequencies of lactase persistence might have been reached in a short time frame due to the action of positive selection. However an effective estimate of the age of the -13910*T allele and a formal test for the role of positive selection in the present distribution of the allele were still missing.

Moreover, it was also reported that the -13910*T allele has a low predictive value outside Eurasia (Mulcare et al. 2004). With the exception of the Fulbe and Hausa from Cameroon (11.2% and 13.9%, respectively), the -13910*T allele was found to be very rare and could not account for the frequency of the lactase persistence phenotype throughout Africa (Mulcare et al. 2004). This discrepancy could be explained by the fact that the allele was not causal of lactase persistence but simply a highly associated marker, which postdates the causal mutation. Alternatively, the lack of concordance between phenotypic and genotypic results could be indicative of genetic heterogeneity underlying lactase persistence. In this case, different lactase persistence mutations would have been originated in different regions of the world and consequently the predictive value of each allele would be dependent on the population studied.

When we started this work, several key aspects of the evolutionary history of lactase persistence were still to be addressed. It was important to pursue the characterization of the -13910*T allele in different populations in order to: a) determine the age of lactase persistence candidate mutation; b) assess the effective role of positive selection in the actual distribution of the -13910*T allele; and c) understand whether lactase persistence arose multiple times in human populations or was caused by a single mutation. In order to address these questions about the evolutionary history of human lactase persistence, two studies based on the haplotypic characterization of the -13910*T allele were independently undertook. One of the studies was performed by Bersaglieri et al. (2004) and was based on the typing of a panel of 100 SNPs in

northern European-derived populations. The other study was conducted by us and was based on the analysis of the haplotypic variation at four microsatellite loci close to the -13910*T allele in several populations from Europe and Africa. The results from our study are presented in the *Article 1*. A more detailed discussion and posterior results are presented as a commentary.

References

Aoki, K. 2001. Theoretical and empirical aspects of gene-culture coevolution. *Theor Popul Biol* 59:253-61.

Beja-Pereira, A., G. Luikart, P. R. England, D. G. Bradley, O. C. Jann, G. Bertorelle, A. T. Chamberlain, T. P. Nunes, S. Metodiev, N. Ferrand, and G. Erhardt. 2003. Gene-culture coevolution between cattle milk protein genes and human lactase genes. *Nat Genet* 35:311-3.

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 74:1111-20.

Durham, W. H. 1991. "Cultural Mediation: The Evolution of Adult Lactose Absorption," in *Coevolution: Genes, Culture and Human diversity*. Stanford: Stanford University Press.

Enattah, N. S., T. Sahi, E. Savilahti, J. D. Terwilliger, L. Peltonen, and I. Järvelä. 2002. Identification of a variant associated with adult-type hypolactasia. *Nat Genet* 30:233-7.

Flatz, G., and H. W. Rotthauwe. 1973. Lactose nutrition and natural selection. *Lancet* 2:76-7.

Flatz, G. 1987. Genetics of lactose digestion in humans. *Adv Hum Genet* 16:1-77.

Holden, C., and R. Mace. 2002. "Pastoralism and the evolution of lactase persistence," in *Human Biology of Pastoral Populations*. Cambridge: Cambridge University Press.

Ingram, C. J., C. A. Mulcare, Y. Itan, M. G. Thomas, and D. M. Swallow. 2009. Lactose digestion and the evolutionary genetics of lactase persistence. *Hum Genet* 124:579-91.

Itan, Y., B. L. Jones, C. J. Ingram, D. M. Swallow, and M. G. Thomas. 2010. A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol Biol* 10:36.

Kuokkanen, M., N. S. Enattah, A. Oksanen, E. Savilahti, A. Orpana, and I. Järvelä. 2003. Transcriptional regulation of the lactase-phlorizin hydrolase gene by polymorphisms associated with adult-type hypolactasia. *Gut* 52:647-52.

Lewinsky, R. H., T. G. Jensen, J. Moller, A. Stensballe, J. Olsen, and J. T. Troelsen. 2005. T-13910 DNA variant associated with lactase persistence interacts with Oct-1 and stimulates lactase promoter activity in vitro. *Hum Mol Genet* 14:3945-53.

McCracken, R. D. 1971. Lactase deficiency: an example of dietary evolution. *Curr Anthropol* 12:479–517.

Mulcare, C. A., M. E. Weale, A. L. Jones, B. Connell, D. Zeitlyn, A. Tarekegn, D. M. Swallow, N. Bradman, and M. G. Thomas. 2004. The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet* 74:1102-10.

Olds, L. C., and E. Sibley. 2003. Lactase persistence DNA variant enhances lactase promoter activity in vitro: functional role as a cis regulatory element. *Hum Mol Genet* 12:2333-40.

Poulter, M., E. Hollox, C. B. Harvey, C. Mulcare, K. Peuhkuri, K. Kajander, M. Sarner, R. Korpela, and D. M. Swallow. 2003. The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. *Ann Hum Genet* 67:298-311.

Rasinperä, H., M. Kuokkanen, K. L. Kolho, H. Lindahl, N. S. Enattah, E. Savilahti, A. Orpana, and I. Järvelä. 2005. Transcriptional downregulation of the lactase (LCT) gene during childhood. *Gut* 54:1660-1.

Sahi, T., M. Isokoski, J. Jussila, K. Launiala, and K. Pyorala. 1973. Recessive inheritance of adult-type lactose malabsorption. *Lancet* 2:823-6.

Simoons, F. J. 1970. Primary adult lactose intolerance and the milking habit: a problem in biologic and cultural interrelations. II. A culture historical hypothesis. *Am J Dig Dis* 15:695-710.

Swallow, D. M. 2003. Genetics of lactase persistence and lactose intolerance. *Annu Rev Genet* 37:197-219.

Troelsen, J. T., J. Olsen, J. Moller, and H. Sjostrom. 2003. An upstream polymorphism associated with lactase persistence has increased enhancer activity. *Gastroenterology* 125:1686-94.

1.2. Results and Discussion

Article 1

Coelho, M., D. Luiselli, G. Bertorelle, A. I. Lopes, S. Seixas, G. Destro-Bisol, and J. Rocha. 2005. Microsatellite variation and evolution of human lactase persistence. *Hum Genet* 117:329-39.

Hum Genet (2005) 117: 329–339
DOI 10.1007/s00439-005-1322-z

ORIGINAL INVESTIGATION

Margarida Coelho · Donata Luiselli · Giorgio Bertorelle
Ana Isabel Lopes · Susana Seixas
Giovanni Destro-Bisol · Jorge Rocha

Microsatellite variation and evolution of human lactase persistence

Received: 19 January 2005 / Accepted: 8 April 2005 / Published online: 1 June 2005

Abstract The levels of haplotype diversity within the lineages defined by two single-nucleotide polymorphisms (SNPs) (−13910 C/T and −22018 G/A) associated with human lactase persistence were assessed with four fast-evolving microsatellite loci in 794 chromosomes from Portugal, Italy, Fulbe from Cameroon, São Tomé and Mozambique. Age estimates based on the intraallelic microsatellite variation indicate that the −13910*T allele, which is more tightly associated with lactase persistence, originated in Eurasia before the Neolithic and after the emergence of modern humans outside Africa. We detected significant departures from neutrality for the −13910*T variant in geographically and evolutionary distant populations from southern Europe (Portuguese and Italians) and Africa (Fulbe) by using a neutrality test based on the congruence between the frequency of the allele and the levels of intraallelic variability measured by the number of mutations in adjacent microsatellites. This result supports the role of selection in the evolution of lactase persistence, ruling out possible confounding effects from recombination suppression and population history. Reevaluation of the available evidence on variation of the −13910 and −22018 loci indicates that lactase persistence probably originated from different mutations in Europe and most of Africa, even if 13910*T is not the causal allele, suggesting that selective pressure could have promoted the convergent evolution of the trait. Our study shows that a limited number of microsatellite loci may provide sufficient resolution to reconstruct key aspects of the evolutionary history of lactase persistence, providing an alternative to approaches based on large numbers of SNPs.

Electronic supplementary material Supplementary material is available for this article at http://dx.doi.org/10.1007/s00439-005-1322-z

M. Coelho · S. Seixas · J. Rocha
Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

M. Coelho · J. Rocha (✉)
Departamento de Zoologia Antropologia, Faculdade de Ciências, Universidade do Porto, Porto, Portugal
E-mail: jrocha@ipatimup.pt
Tel.: +351-22557-0700
Fax: +351-22557-0799

D. Luiselli
Dipartimento di Biologia Evoluzionistica Sperimentale, Universita di Bologna, Bologna, Italia

G. Bertorelle
Sezione di Biologia Evolutiva, Dipartimento di Biologia, Universitá di Ferrara, Ferrara, Italia

A. I. Lopes
Unidade de Gastrenterologia Pediátrica, Hospital de Santa Maria, Lisbon, Portugal

G. Destro-Bisol
Dipartimento di Biologia Animale e dell' Uomo, Universitá "La Sapienza", Rome, Italy

G. Destro-Bisol
Istituto Italiano di Antropologia, Rome, Italy

Introduction

The ability to digest lactose in adults is an autosomal dominant hereditary condition caused by the persistence of lactase activity in the small intestine after weaning. The frequency of lactase persistence, as evaluated by different physiological tests, varies widely in human populations and is well correlated with the distribution of dairy farming (reviewed in Swallow 2003). In the majority of populations, the ancestral mammalian developmental pattern prevails, and most people have a marked decline in lactase levels after infancy (lactase restriction), which may limit their use of large amounts of fresh milk in adulthood. In Europe, the highest frequencies of lactase persistence are observed in northwestern populations, where milk-dependent cattle pastoralism was developed very early (Midgley 1992),

and there is a decrease in prevalence towards the south and east. In Africa, both north and south of the Sahara, lactase persistence is typically much more frequent among pastoralists than in neighboring non-pastoralist communities.

There are different views on the microevolutionary forces underlying the present-day distribution of lactase persistence. Several studies have proposed that the match between the geographic distribution of lactase persistence and dairy farming could be the result of the recent selective pressure associated with the added nutritional benefit of high milk consumption in populations that shifted their subsistence patterns to become crucially dependent on milk (Simoons 1970; McCracken 1971; Kretchmer 1972; Flatz 1987; Holden and Mace 1997). A quite different view is sustained by Nei and Saitou (1986), who questioned the role of selection based on the assumption that the origin of lactase persistence predated the geographical dispersion of modern humans and on the lack of a sufficiently long period of time for selection to act since the introduction of dairying. According to this interpretation, the differences in lactase persistence among human populations arose by genetic drift and preceded the major changes in subsistence patterns associated with the Neolithic. In this case, the correlation between lactase persistence and milk-based pastoralism could be either entirely fortuitous or caused by the adoption of milk drinking habits only by those populations with the ability to digest lactose (Bayless 1971; McCracken 1971; Aoki 2001).

Recently, the T allele of a C/T polymorphism in a potential regulatory site located 13,910 bp upstream the lactase gene was found to be completely associated with lactase persistence in Northern Europeans (Enattah et al. 2002). A significant, although less strong association, was also observed with a second −22,018-bp G → A mutation (Enattah et al. 2002). Besides creating a new tool for assessing lactose-digesting capability through single-nucleotide polymorphism (SNP) genotyping, these findings provided a basis for studying the major forces that shaped the distribution of lactase persistence, namely through the analysis of the levels of mutational heterogeneity and haplotype diversity associated with this trait.

In the Eurasian populations studied so far, the frequencies of the −13910*T allele are concordant with the expectations from physiological tests (Swallow 2003). However, with the exception the Fulbe and Hausa from Cameroon, the −13910*T allele was found to be rare in many African pastoralist communities where high frequencies of lactase persistence were previously found by using physiological tests (Mulcare et al. 2004). It is still to be demonstrated whether the observed discrepancy is caused by the fact that the allele is not causative or that lactase persistence had separate mutational origins in Africa and in Eurasia. The latter hypothesis clearly favors a prominent role of selection, since it would be unlikely that different persistence mutations might have risen in frequency and spread throughout human dairying societies without being driven by an adaptive advantage.

Haplotype diversity studies in populations of Northern European ancestry have shown that most −13910*T and −22018*A alleles lie in an extended SNP-defined haplotype that was found to be unusually long for its frequency, indicating that there was not enough time for recombination to break it down (Poulter et al. 2003; Bersaglieri et al. 2004). This finding is consistent with the hypothesis that the current distribution of lactase persistence in Northern Europeans was caused by recent positive selection, but the possible confounding effects of allele-specific recombination suppression and/or population history on the extent of linkage disequilibrium need to be ruled out (Hollox 2004).

Here we present an analysis of the genetic variation and the evolutionary history of lactase persistence based on a microsatellite approach. By applying a neutrality test based on the intraallelic accumulation of mutations, which is not influenced by recombination suppression, we provide evidence of selection acting on the −13.910 kb*T allele in four ethnically diverse populations from Europe (Portuguese, Italians, and Finnish) and Africa (Fulbe), whose heterogeneity makes the role of population history an unlikely confounding factor. Furthermore, we reevaluate the available evidence on variation of the −13910 and −22018 SNPs and conclude that lactase persistence probably originated from different mutations in Europe and most Africa even if the −13910*T is not the causal allele. Our study shows that a battery of four microsatellite loci that can be easily typed in large samples is able to capture the information necessary to reconstruct the evolution of lactase persistence in human populations, providing an alternative to approaches based on large numbers of SNPs.

Materials and methods

Populations

DNA samples were obtained upon informed consent from Central Italy ($n = 67$ individuals; 37 from Tocco da Causaria and 30 from Rome), Northern Portugal ($n = 90$), the Fulbe ethnic group from Cameroon ($n = 51$), São Tomé Island in the Gulf of Guinea ($n = 142$; from different locations in the Island), and Mozambique ($n = 47$; from speakers of the Ronga Bantu language from Maputo).

The Fulbe sample was obtained in the province of the Extreme Nord in Cameroon, in the villages of Marua, Meme, and Mora. This population descends from nomadic herders that moved from Nigeria to the Cameroon from the eighteenth century onwards and progressively abandoned sheep farming to become settled agriculturists (Spedini et al. 1999). The samples from Mozambique and São Tomé are from populations that have neither traditions of pastoralism nor dairy practices, but provide useful information on the distri-

bution of background microsatellite haplotype variability associated with the lactase gene. Mozambique lies at the southeastern edge of the Bantu expansion and might have been a contact zone between Bantu-speaking farmers and more ancestral Khoisan (Salas et al. 2002). São Tomé started to be peopled by the end of the fifteenth century with slaves imported by Portuguese colonists from the adjacent coasts of the Gulf of Guinea and the Congo–Angola area. As a consequence of this settlement pattern this insular population has retained the high levels of genetic diversity that are generally observed in the African mainland (Tomás et al. 2002).

SNP and microsatellite typing

Haplotype diversity was assessed through the analysis of the two SNPs associated with lactase persistence (−13910 C/T and −22018 G/A) and four linked microsatellites: D2S3010 [a $(TATC)_n$ repeat], D2S3013 [a $(TA)_n$ repeat], D2S3015 [a $(CAAAA)_n$ repeat] and D2S3016 [a $(TG)_n$ repeat] (Fig. 1).

The SNPs were typed by PCR-restriction fragment length polymorphism (PCR-RFLP) methods. The −13910 C/T polymorphism was amplified within a 125-bp fragment with primers 5′-GCAGGGCTCAAAGAACAATC-3′ (forward) and 5′-TGTACTAGTAGGCCTCTGCGCT-3′ (reverse). The −13910*T allele introduces a *Bsm*FI restriction site that originates digestion product sizes of 80 and 45 bp. The −22018 G/A locus was amplified within a 271-bp product with primers 5′-CTCAGTGATCCTCCCACCTC-3′ and 5′-CCCCTACCCTATCAGTAAAGGC-3′. Digestion with *Hin*6I generates 196- and 75-bp fragments in the presence of the −22018*G allele. PCR reactions contained 0.5 μM of each primer, 0.2 mM of each deoxynucleotide triphosphate (dNTP), 10 mM Tris–HCl (pH 8.8), 50 mM KCl, 0.08% Nonidet, 1.5 mM $MgCl_2$ (1.0 mM for the −22018 G/A locus) and 1 U *Taq* polymerase. Samples were denatured for 5 min at 94°C, followed by 35 cycles of 94°C for 1 min, 58°C for 1 min, and 72°C for 1 min, followed by a 20-min extension at 72°C. Digestions (1 U/μl) were performed at 50°C for 1 h and DNA fragments were visualized by silver staining after non-denaturing electrophoresis separation in 9% polyacrylamide gels.

Microsatellites were typed by PCR amplification in two duplex reactions followed by separation of amplification products in an ABI 310 DNA sequencer. Fragment analysis and weight determination were performed with the GeneScan software. The first duplex reaction included the primers for D2S3013 (5′-GAGAATATAGTCATAAACTATGTT-3′ and 5′-ATTTTGGATTATATATGCTTTCTTG-3′, labeled with FAM fluorescence) and D2S3015 (5′-CCTGTAGTCCCAGCTAATTTC-3′ and 5′-CAGAGAAGTTTTGTTTGTGGA-3′, labeled with TET fluorescence) at 0.5 and 0.075 μM concentrations, respectively. The second duplex reaction included the primers for D2S3010 (5′-TTAGGCCTCTCTTCGAATGAT-3′ and 5′- GATTTAGGTGGAGACACAC-3′, labeled with FAM fluorescence) and D2S3016 (5′-GAGAAAAATTAGGTGTGAAACCA-3′ and 5′-CCCTTTAGCTGCCTGAACTG-3′, labeled with TET fluorescence) at 0.5 and 0.075 μM concentrations, respectively. All other reagents were as above. In both duplex reactions, samples were denatured for 5 min at 94°C, followed by 35 cycles of 94°C for 1 min, 55°C for 1 min, and 72°C for 1 min, followed by a 20-min extension at 72°C.

Haplotype determination

Haplotypes of sampled individuals were reconstructed from the combined genotype data for all the populations by the statistical inference method implemented in the software package PHASE, version 2.0.2 (Stephens et al. 2001; Stephens and Donnelly 2003), which provides the probabilities of the most likely pairs of haplotypes for each individual. Haplotypes were inferred with two alternative approaches. In the first approach we determined three-locus haplotypes consisting of the two −13910 and −22018 SNPs and each microsatellite marker. In the second approach we inferred the full six-locus haplotypes, combining the two SNPs and the four microsatellites. Individual haplotype phases were assigned by choosing the most probable haplotype pair that was compatible with the individual multi-locus genotypes. The haplotype frequencies were calculated by direct counting after resolution of each individual haplotype phase.

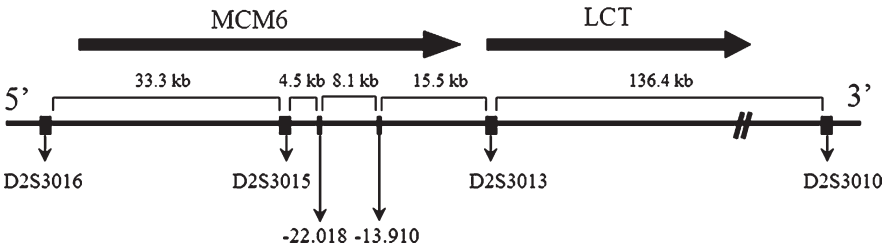


Fig. 1 Schematic representation of the genetic interval including the lactase locus (*LCT*) and the neighbor gene for the human homologue of a yeast gene involved in the cell cycle (MCM6), with the relative locations of the two SNPs (−13.910 kb and −22.018 kb) and the four microsatellites (D2S3010, D2S3013, D2S3015, D2S3016) used to characterize the haplotype diversity associated with the lactase restriction/persistence polymorphism. Distances are as in BAC clone RP11-34L23 (GenBank accession no. ACO118937) 75x23mm (300×300 DPI)

Age estimates

To estimate the time to the most recent common ancestor (TMRCA) of the −13910*T allele associated with lactase persistence, we used two different methods based on the intraallelic accumulation of microsatellite diversity, assuming a stepwise mutation model and using a 25-year generation time.

In the first method, an unbiased estimator of the TMRCA was calculated, assuming no recombination, by the average squared difference in repeat number between each sampled −13910*T haplotype and the root haplotype (Stumpf and Goldstein 2001). The root of the −13910*T clade was obtained by combining together the modal allele lengths at each microsatellite locus in the pooled sample from all the populations. The TMRCA central estimates and confidence intervals were calculated using the program Ytime (Behar et al. 2003).

The second method is based on the simulation of the overtime decay in the frequency of the allele originally associated with −13910*T in each microsatellite locus (Seixas et al. 2001). Unlike the previous method, this approach allows for recombination to be taken into account. The modal allele length at each microsatellite locus in the pooled sample was considered to be the ancestral and the combined TMRCA was calculated as the weighted average of the single locus estimates, with the weight of each microsatellite locus determined by the sum of its corresponding mutation and recombination rates. Recombination rates (r) were calculated using the general relation 1 cM = 1 Mb, according to the approximate estimates provided by Kong et al. (2002) for the region encompassing the four microsatellite loci. Confidence intervals were calculated assuming a rapid population growth according to Goldstein et al. (1999).

For each age estimation method we used two sets of microsatellite mutation rates (μ). The first set was derived indirectly from the parameter $\theta = 4Ne\mu$ assuming mutation-drift equilibrium and using the unbiased θ estimator proposed by Xu and Fu (2004), based on the sample homozygosity under the single-step stepwise mutation model. We assumed $Ne = 10{,}000$ (Takahata 1993) and estimated homozygosities from the microsatellite allele frequency distributions in São Tomé, which are less likely to have been distorted by a possible increase in the frequency of tolerance-associated chromosomes due to selection (see below). The second set of mutation rates was derived from the average 0.001 value obtained from observed mutations in pedigrees (Weber and Wong 1993). Locus specific mutation rates were calculated by apportioning this average according to the ratios of the locus-specific estimates calculated by the indirect approach.

Neutrality tests

To assess the role of natural selection in shaping the distribution of the −13910*T allele, we used the test developed by Slatkin and Bertorelle (2001), which evaluates whether the observed frequency of an allele is consistent with its levels of variability under a given demographic pattern, assuming neutrality. We used the test modality that measures the intraallelic variability by the minimum number of mutations (S_0) observed at linked microsatellite marker loci (Slatkin and Bertorelle 2001; Slatkin 2002).

The tests were performed by considering the simultaneous combination of all four microsatellites with the −13910*T allele. The minimum number of mutations necessary to generate the observed haplotypes (S_0) was inferred by using median-joining networks (Bandelt et al. 1999) calculated with the program NETWORK 4.0.0.0 (http://www.fluxus-engineering.com). All tests were performed under a number of different demographic models (see below) with the two sets of mutation rates used for calculating the TMRCA of the −13910*T allele.

Results

Haplotype diversity

The frequencies of the core haplotypes defined by the −13910 C/T and −22018 G/A SNPs, and the expected prevalence of lactase persistence in different populations are shown in Table 1. Estimates from northern Portugal, Italy and the Fulbe are within the frequency ranges previously reported on the basis of physiological tests (Flatz 1987; Swallow 2003). The estimates from São Tomé and Mozambique are within the range observed for the majority of African non-pastoralist populations (Flatz 1987; Swallow 2003). It is likely that the occurrence of the −13910*T allele in these two populations is due to recent admixture with Europeans. In São Tomé, for example, the frequency of the −13910*T allele is very close to that expected from a previously calculated 11% level of admixture with the Portuguese colonists (Tomás et al. 2002).

The C–A haplotype is rare in all samples. Since, as previously shown (Poulter et al. 2003; Swallow 2003), the −13910 and −22018 polymorphisms were originated according to a C–G → C–A → T–A phylogenetic sequence, the low frequency of this intermediate haplotype indicates that the −22018 G → A mutation might have occurred only shortly before the −13910 C → T mutation. It is the occasional occurrence of C–A chromosomes that may lead to the wrong identification of lactase persistence on the basis of −22018 genotyping.

Microsatellite allele frequency distributions within the common C–G and T–A −13910/−22018 SNP core haplotypes in a pooled sample combining the data from all populations are shown in Fig. 2. Equivalent distributions for each sample are shown in Fig. S1 of the Electronic supplementary material (ESM). The data presented were retrieved from an inferred distribution of

Table 1 Frequencies of the haplotypes defined by the −13910 and −22018 SNPs and predicted prevalence of lactase persistence in the different populations

Haplotype[a]		Populations				
−13910	−22018	Portugal ($n=90$)	Italy[b] ($n=67$)	Fulbe ($n=51$)	São Tomé ($n=142$)	Mozambique ($n=47$)
C	G	0.62	0.87	0.79	0.94	0.99
C	A	0.01	–	–	0.02	–
T	A	0.37	0.13	0.21	0.04	0.01
Predicted frequency of lactase persistence[c]		0.62	0.24	0.38	0.08	0.02

[a]The haplotype frequencies were retrieved from the inferred distribution of six locus SNP/microsatellite haplotypes presented in Table S1 of the Electronic supplementary material

[b]Samples from Tocco da Causaria and from Rome were pooled since they are not significantly different ($P=0.61$), using the exact test of population differentiation of Raymond and Rousset (1995) implemented in the Arlequin 2.1 software (Schneider et al. 2000)

[c]Frequency of −13910 CT + TT genotypes assuming Hardy–Weinberg equilibrium.

six-locus full haplotypes combining the two −13910 and −22018 SNPs, and the four microsatellite markers, which had a 70% proportion of phase callings with confidence values ≥75% (Table S1, ESM). An alternative inference approach based on the determination of three-locus haplotypes consisting of the two SNPs and each microsatellite yielded higher fractions of haplotypes with phase calling probabilities ≥75%, ranging from 95% for locus D2S3010 to 99% for loci D2S3015 and D2S3016 (data not shown). However, we found no significant differences between the microsatellite allele frequency distributions within the core SNP haplotypes obtained by the two approaches, using an exact test of population differentiation (Raymond and Rousset 1995). Moreover, no significant differences were found between the confidence of phase callings involving C–G and T–A SNP core haplotypes in each of the two approaches.

A clear reduction in microsatellite variation was found within the T–A haplotype (Fig. 2). This decrease in the −13910*T intraallelic diversity is not uniform across all microsatellite markers and the higher variability accumulated in D2S3010 and D2S3013 suggests that these loci have higher mutation rates than D2S3015 and D2S3016 (Fig. 2). This is also confirmed by the decreased heterozygosities observed for the D2S3015 and D2S3016 markers among C–G haplotypes (Fig. 2).

A median-joining network relating the compound SNP-microsatellite haplotypes in the pooled sampled is shown in Fig. 3. The network has two main branches that reflect the bimodality of the D2S3013 microsatellite allele frequency distributions within C–G core haplotypes (Figs. 2, 3). In contrast with the high variability associated with C–G chromosomes, T–A haplotypes are tightly clustered within one of the two main branches irrespectively of their geographic location, as expected from a unique, relatively recent origin. Within the T–A clade, the microsatellite configuration 10–21–4–2 (D2S3010–D2S3013–D2S3015–D2S3016) is found in all populations, except Mozambique, and is likely to represent the ancestral chromosome since it is the most frequent haplotype and combines the modal allele from each individual locus (Fig. 3, inset; Table S1, ESM). Alone it represents 25% of all sampled T–A chromosomes, 52% together with its four one-step neighbors (11–21–4–2; 10–22–4–2, 10–20–4–2 and 9–21–4–2).

An additional feature of the microsatellite allele frequency distributions is the apparent lack of recombinant T–A haplotypes within the 61.4-kb region encompassing the D2S3013, D2S3015 and D2S3016 loci (Fig. 1). This is indicated by the complete absence of diversity in D2S3015 and D2S3016 and by the observation of a clear unimodal distribution at the D2S3013 locus, which suggests the occurrence of a stepwise accumulation of mutations in an ancestral T–A haplotype carrying the D2S3013*21 allele (Fig. 2). If recombination had played a major role in the generation of D2S3013 diversity, the striking bimodality observed within C–G haplotypes would be at least partially reflected among the T–A chromosomes and these would not cluster just in one side of the haplotype network (Figs. 2, 3). Due to a less clear difference between the shape of the D2S3010 microsatellite allele frequency distributions among C–G and T–A haplotypes, it is more difficult to evaluate the role of recombination in the regeneration of diversity in this locus. Taken together, these observations highlight the usefulness of using faster evolving markers to subtype haplotypes that could be otherwise homogeneous if defined only by SNPs.

TMRCA of the −13910*T allele

The TMRCAs of the −13910*T allele calculated in different samples are presented in Table 2. Calculations for the Finnish sample were performed with data taken from Enattah et al. (2002) and do not include locus D2S3010.

Since age calculations can be typically affected by the misestimation of mutation and recombination rates

(Reich and Goldstein 1999), we considered different combinations of these parameters to assess the sensitivity of estimates to variation in their values. In the pooled sample, estimates that do not take recombination into account (m_1 and m_2) are within 45,000–30,000 or 17,500–11,750 year ranges, depending on the use of indirect or direct estimates of microsatellite mutation rates, respectively. In general the average square distance method (Δ) leads to higher age estimates than calculations based on the decrease in frequency of the modal microsatellite allele (p). This discrepancy is particularly noticeable in the sample from São Tomé and may be due to the presence of low frequency microsatellite alleles that are several mutational steps away from the modal haplotype and which may not be taken into account in the "p method".

If we assume that both recombination and mutation contribute to the intraallelic diversity, the observed haplotype homogeneity and lack of T–A recombinant chromosomes can be only explained by a more recent TMRCA of the −13910*T allele. Accordingly, calculations based on the decrease in frequency of the modal microsatellite allele that do not assume recombination suppression lead to the lowest TMRCA estimates: 12,300 and 7,450 years in the pooled sample, under assumption sets m_3 and m_4, respectively (Table 2).

Age estimates in the Fulbe and Finnish samples were found to be lower than in Portugal and Italy. These differences reflect the interpopulation variation in the levels of intraallelic diversity that may be caused by the combined effects of sampling variance and specific demographic histories of each population, like differ-

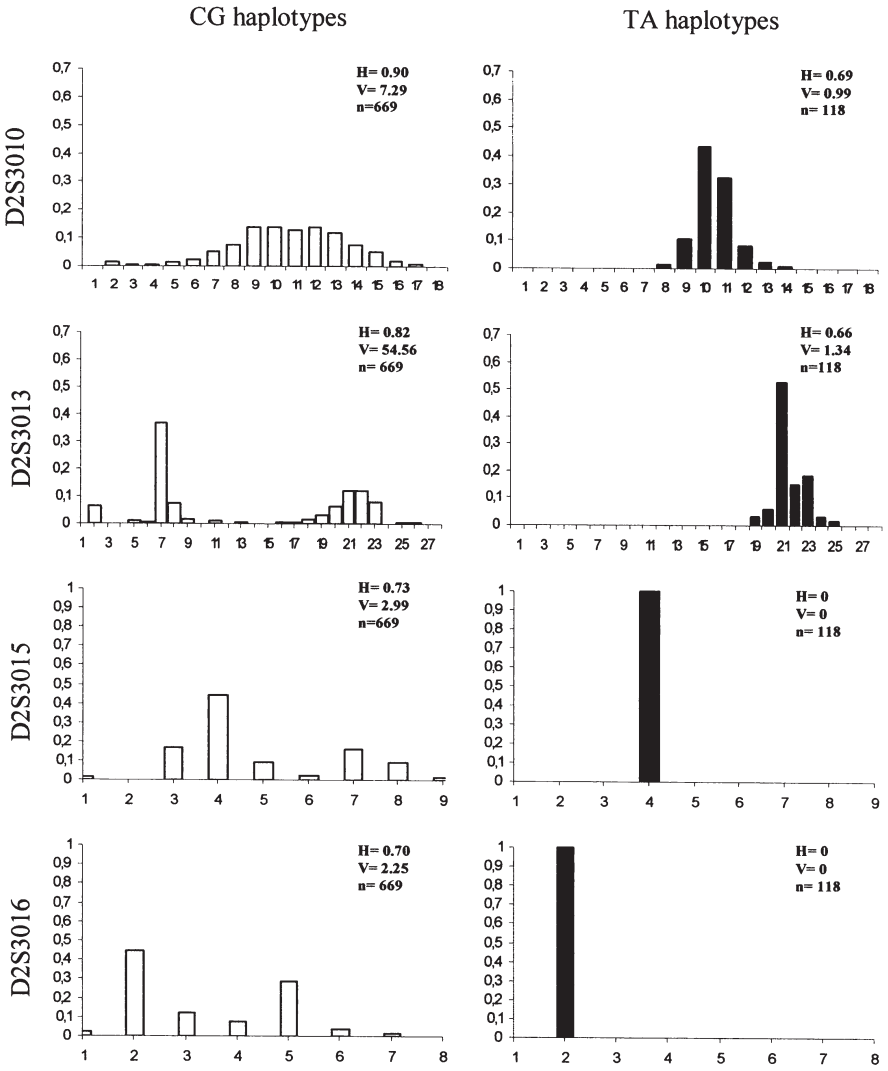


Fig. 2 Microsatellite allele frequency distributions within C-G and T-A −13910/−22018 SNP haplotypes in a pooled sample from São Tomé, Mozambique, Portugal, Italy and Cameroonian Fulbe. The estimated sizes of allele 1 in each microsatellite are: 188 bp for D2S3010; 133 bp for D2S3013; 190 bp for D2S3015 and 148 bp for D2S3016. *H* heterozygosity; *V* variance in repeat number; *n* number of chromosomes 148×177mm (300×300 DPI)

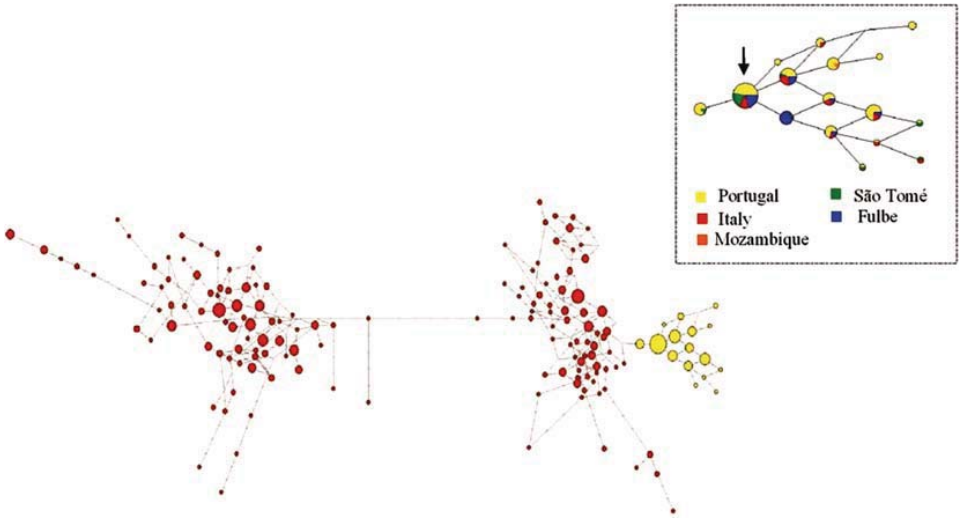


Fig. 3 Median-joining network (Bandelt et al. 1999) representing the compound SNP-microsatellite haplotype variation in a pooled sample from São Tomé, Mozambique, Portugal, Italy and Cameroonian Fulbe. C-G haplotypes are shown in red and T-A haplotypes are shown in yellow. The distribution of haplotype variation within T-A chromosomes is shown in the *inset*. Haplotypes are represented by *circles*, with areas proportional to the number of individuals harboring the haplotype. The putative ancestral 10-21-4-2 (D2S3010-D2S3013-D2S3015-D2S3016) haplotype is indicated with an *arrow*. Networks were calculated with the program NETWORK 4.0.0.0., using the same weight for SNP and microsatellite loci and the 'frequency >1' option, which selects only the haplotypes that occur more than once in the data set 350×192mm (72×72 DPI)

ences in the long term population size or in levels of drift during the dispersion of the trait.

Neutrality tests

Table 3 presents the results of the neutrality tests for the −13910*T allele in different samples using the method of Slatkin and Bertorelle (2001). The data consist of the full haplotypes combining all four microsatellite markers linked to the −13910*T allele. For illustrative purposes we present the outcomes obtained with different combinations of two global demographic models (D1 and D2) and the two sets of microsatellite mutation rates used for age calculation (see Table 2). The first demographic model (D1) is based on the analysis of Pritchard et al. (1999) and assumes a constant exponential growth rate of 0.008 starting 900 generations ago from an initial population of 10^3. The second model (D2) is a variation of the scenarios simulated by Kruglyak (1999) and assumes that the effective population size increased exponentially from 10^4 to 5×10^9, also starting 900 generations ago. A smaller long term population size and lower genetic diversity is expected under demographic model D1.

Neutrality is rejected at the 0.001 level for the Portuguese, Finnish and Fulbe samples under all sets of assumptions (Table 3). In the Italian sample, neutrality cannot be rejected at the same significance level under the most conservative assumption, which combines demographic model D1 with the set of lower mutation rates (m_1), decreasing the expected levels of intraallelic diversity under neutrality. In São Tomé, where a much lower frequency of the −13910*T allele is found, neutrality is rejected only with the assumptions involving the set of higher mutation rates (m_2).

Similar patterns and conclusions were obtained with a variety of other reported demographic models, differing in the rate of exponential growth, time of onset of population growth, and effective population sizes before expansion (Rogers and Harpending 1992; Marjoram and Donnelly 1994; Wall and Przeworski 2000; Slatkin and Bertorelle 2001; Pluzhnikov et al. 2002) (results not shown).

Discussion

The distribution of the ability to digest lactose in human populations is generally claimed to be an example of genetic adaptation to recent modifications in human dietary habits. However, the support of population genetics for a causative link between dairy farming and lactase persistence has been mostly based on geographic correlations and considerable controversy still exists about the relative roles that selection and population history might have played in the origin, evolution, and spread of this trait. As a contribution to the understanding of this topical issue, we have characterized the patterns of haplotype variation within lineages defined by SNPs −13910 C/T and −22018 G/A, using four microsatellite loci encompassing 198 kb around the lactase gene (*LCT*) in 794 chromosomes from five ethnically diverse populations with different genetic backgrounds

Table 2 Estimates of the time (years) to the most recent common ancestral of the −13910*T allele

Population	Age estimation method					
	Average square distance, Δ		Decrease in frequency of modal allele, *p*			
	m_1[a]	m_2[b]	m_1	m_2	m_3[c]	m_4[d]
Portugal ($n_i = 66$)[e]	48,370 (9,910–127,870)[f]	18,930 (3,870–53,620)	38,940 (23,690–75,500)	15,250 (11,690–39,440)	15,560 (10,000–28,440)	9,370 (5,940–17,190)
Italy ($n_i = 17$)	52,220 (10,990–142,000)	20,440 (4,310–57,000)	34,625 (13,440–140,000)	13,560 (5,250–54,440)	13,560 (5,750–42,190)	8,315 (3,440–27,690)
Finland[g] ($n_i = 33$)	23,640 (0–88,125)	9,250 (0–34,000)	20,750 (9,625–39,875)	8,125 (3,750–15,625)	nd[h]	nd[h]
Fulbe ($n_i = 21$)	20,025 (1,475–60,075)	7,840 (575–26,125)	23,815 (9,315–54,130)	9,310 (3,625–28,250)	10,125 (4,000–26,250)	6,060 (2,440–16,810)
São Tomé ($n_i = 13$)	46,750 (9,625–131,810)	18,290 (3,750–54,440)	17,560 (3,940–51,690)	6,875 (1,560–20,250)	7,375 (1,750–19,000)	4,400 (1,000–11,940)
Pooled[i] ($n_i = 117$)	44,610 (10,040–110,000)	17,460 (4,125–48,300)	30,000 (21,125–43,690)	11,750 (8,250–17,125)	12,300 (8,940–17,125)	7,440 (5,375–10,440)

[a]Assuming suppression of recombination and microsatellite indirect estimation of mutation rates: μ_1(D2S3010) = 0.0009; μ_2(D2S3013) = 0.0005; μ_3(D2S3015) = 0.000095; μ_4(D2S3016) = 0.00011

[b]Assuming suppression of recombination and microsatellite mutation rates calculated from a 0.001 direct average estimate: μ_1(D2S3010) = 0.0023; μ_2(D2S3013) = 0.0013; μ_3(D2S3015) = 0.0002; μ_4(D2S3016) = 0.0003

[c]Mutation rates as in m_1 and assuming the following recombination rates between the −13910 site and each microsatellite locus: r_1(D2S3010) = 0.0015; r_2(D2S3013) = 0.00016; r_3(D2S3015) = 0.00013; r_4(D2S3016) = 0.00046

[d]Mutation rates as in m_2 and recombination rates as in m_3

[e]n_i represents the number of chromosomes bearing the −13910*T allele

[f]95% confidence intervals are given *in parentheses*; confidence intervals for the Δ method were calculated assuming no population growth

[g]Based on the data from Enattah et al. (2002), not including locus D2S3010

[h]Not done, due to unavailable distribution of microsatellite allele frequencies in the general population

[i]Excluding Finland

and subsistence patterns. Our study infers the intraallelic genealogy of lactase persistence based on the use of mutations in fast evolving markers, yields new results concerning the role of selection in the evolution of this trait in populations outside northern Europe and presents a phylogeographic interpretation suggesting that lactase persistence most probably arose from different mutations in Europe and most of Africa.

The assessment of intraallelic diversity indicates that, even assuming recombination suppression, the TMRCA of the −13910*T variant, which is more tightly associated with lactase persistence, is unlikely to be much older than ~45,000 years, although it may be as recent as 12,500–7,500 years if more realistic sets of assumptions that take recombination into account are applied. Since these estimates refer only to the age since intraallelic diversity began to accumulate due to an increase in frequency or population expansion, the actual age of the −13910 C → T mutation may be older than inferred by its TMRCA (Slatkin and Rannala 2000). However, when the TMRCA estimates are taken together with the observation of low frequencies of −13910*T in most sub-Saharan African populations (Mulcare et al. 2004), it is reasonable to conclude that the mutation originated in Eurasia before the Neolithic and after the emergence of modern humans outside Africa 100,000–50,000 years ago (Klein 2000; Relethford 2001).

Our analysis of neutrality, using a wide range of demographic models, provides strong support to the notion that lactase persistence underwent a rapid increase in frequency, due to a selective advantage. Given the low probability values for finding the observed intraallelic diversities under neutrality (Table 3), this conclusion is unlikely to have been affected by ascertainment bias and seems to be sufficiently robust to remain unaltered by further corrections for multiple tests.

The major implications of these results lie in the assumptions underlying the neutrality test and in the composition of the population dataset. Differently from a previous assessment of selection in the *LCT* gene based on the observation of an extended haplotype in northern European-derived populations with a panel of 100 SNPs covering 3.2 Mb (Bersaglieri et al. 2004), our approach does not rely upon recombination and measures intraallelic diversity by the number of mutations in fast evolving linked microsatellites. Therefore, our test is exempt from the possible confounding effects of allele-specific recombination rates, which may produce unusually long *LCT* haplotypes at high frequency, even in the absence of selection (Hollox 2004). The robustness of our conclusions is further strengthened by the fact that evidence for selection is not confined to a single, relatively homogeneous, northern-European population, but can be found in samples from southern Europe (Portugal and Central Italy) and in the Cameroonian Fulbe, which lie in the periphery of the distribution of the −13910*T allele and are separated from the major regions of high frequency of lactase persistence in Europe by a belt of agricultural northern-African popula-

Table 3 Probabilities of finding a number of mutations $\leq S_0$ in the microsatellite loci linked to the −13910*T allele

Population	n_i[a]	n[b]	S_0[c]	D1[d]		D2[e]	
				m_1[f]	m_2[f]	m_1	m_2
Portugal	66	180	17	1.40×10^{-14}	6.09×10^{-52}	2.31×10^{-30}	1.54×10^{-95}
Italy	17	134	8	3.56×10^{-3}	1.97×10^{-11}	1.42×10^{-6}	5.30×10^{-21}
Finland[g]	33	57	3	2.80×10^{-6}	1.57×10^{-17}	7.45×10^{-13}	2.56×10^{-32}
Fulbe	21	102	7	2.66×10^{-5}	2.79×10^{-17}	2.21×10^{-10}	2.90×10^{-30}
Sao Tomé	13	284	11	0.380	9.85×10^{-5}	0.0139	5.50×10^{-11}

[a]Number of chromosomes bearing the −13910*T allele
[b]Total number of chromosomes in the sample
[c]Minimum number of mutations in linked microsatellite loci
[d]Assuming a constant exponential growth rate of 0.008, starting 900 generations ago from an initial population of 10^3
[e]Assuming exponential growth from 10^4 to 5×10^9, starting 900 generations ago
[f]Sets of mutation rates as defined in Table 2
[g]Based on the data from Enattah et al. (2002), not including locus D2S3010

tions where lactase restriction predominates (Flatz 1987). This indicates that it is unlikely that the present observations are simply caused by shared population history.

While the current distribution of lactase persistence in Eurasia and the African Fulbe seems to be due to the dispersion of a single mutation, it is still unclear what is the significance of the recent finding that −13910*T allele is absent from most African populations in which high frequencies of lactase persistence have been previously found with physiological tests (Mulcare et al. 2004). This evidence in itself is not necessarily in contrast with a unique origin of lactase persistence in Africans and Europeans if the −13910*T is not a causative factor but a mutation associated with this trait. However, consideration of the temporal relationships between the −22018 G → A and −13910 C → T mutations and their correspondent levels of association with lactase persistence suggests that, even if the −13910*T is not the causative allele, the trait is likely to be due to an independent mutation in Europe and in most Africa (Fig. 4). Assuming that −13910*T is not the causal allele, its association with lactase persistence in Europe, but not in many African populations, can only be explained if the −13910 C → T mutation postdates both the true causal mutation and the separation between the two continental groups (Fig. 4). On the other hand, the −22018 G → A mutation is not completely associated with lactase persistence in Europe, which would imply that the true causative mutation must have occurred after this mutation. Therefore, the −22018 G → A and −13910 C → T mutations would provide an upper and lower temporal bound, respectively, to the causal mutation. With this scenario, if Africans and European shared the same causal mutation, a high level of geographic segregation between related haplotypes should have occurred and a strong association between the −22018*A allele and lactase persistence would be expected in African populations with high frequencies of the trait, in spite of the absence of the −13910*T allele (Fig. 4). However, this is contradicted by the low frequency of −22018*A allele in Africa (Bersaglieri et al. 2004) and an independent origin of lactase persistence in Europe and Africa due to separate mutations in the same or in different regulatory elements remains the most plausible explanation.

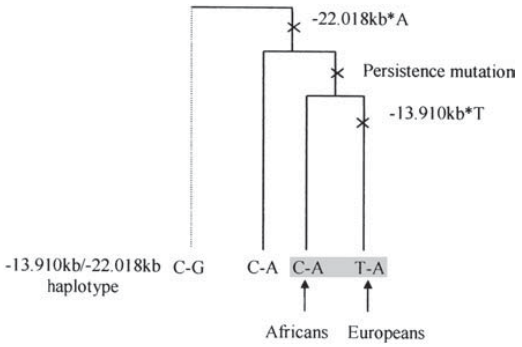


Fig. 4 Phylogeograhic implications of the hypothesis that European populations and African populations with high frequencies of lactase persistence share a common causal mutation different from −13910 C → T. Lineages associated with lactase persistence are *shaded*80×55mm (300×300 DPI)

Our finding that a battery of four microsatellite loci that can be easily typed in large samples is able to capture the information necessary to reconstruct the evolution of lactase persistence in human populations, highlights the significance of using these faster evolving markers to increase the efficiency and resolution of phylogeographic studies and to provide an alternative to approaches based on large numbers of SNPs. This approach might prove useful in future studies of genetic determinants of lactase persistence in African populations. Assuming that the trait is not caused by a *trans*-acting mutation, a possible research direction could be to use microsatellite markers to subtype SNP-defined haplotypes and to identify sub-haplotypes with low microsatellite diversity that could have undergone recent positive selection and show high frequency differences across African populations with diverse milk-drinking traditions. If association with lactase persistence is confirmed through concordance with physiological tests,

a further search for candidate mutations could then be restricted to the selected subhaplotypes.

Acknowledgements We thank Luís Pedro Resende and Cinzia Battaggia for assistance in typing the Portuguese and Italian samples, respectively. We are also grateful to Gabriella Spedini for the Fulbe DNA samples and to António Prista and Sílvio Saranga for the Mozambique samples. We thank Eduardo Tarazona-Santos and Nuno Ferrand for comments on the manuscript. This research was supported by the Sociedade Portuguesa de Gastrenterologia and by the Fundação para a Ciência e a Tecnologia (grants POCTI/42510/ANT/2001 and POCTI/BIA-BDE/56654/2004). D.L. and G.D.B. were supported by the M.I.U.R. (grant numbers 2003054059 and 2005058414).

References

Aoki K (2001) Theoretical and empirical aspects of gene-culture coevolution. Theor Popul Biol 59:253–261

Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16:37–48

Bayless TM (1971) Junior, why didn't you drink your milk?. Gastroenterology 60:479–480

Behar DM, Thomas MG, Skorecki K, Hammer MF, Bulygina E, Rosengarten D, Jones AL, Held K, Moses V, Goldstein D, Bradman N, Weale ME (2003) Multiple origins of Ashkenazi Levites: Y chromosome evidence for both Near Eastern and European ancestries. Am J Hum Genet 73:768–779

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE and Hirschhorn JN (2004) Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet 74:1111–1120

Enattah NS, Sahi T, Savilahti E, Terwilliger JS, Peltonen L, Järvelä I (2002) Identification of a variant associated with adult-type hypolactasia. Nat Genet 30:233–237

Flatz G (1987) Genetics of lactose digestion in humans. Adv Hum Genet 16:1–77

Goldstein DB, Reich DE, Bradman N, Usher S, Seligsohn U, Peretz H (1999) Age estimates of two common mutations causing factor XI deficiency: recent genetic drift is not necessary for elevated disease incidence among Ashkenazi Jews. Am J Hum Genet 64:1071–1075

Holden C, Mace R (1997) Phylogenetic analysis of the evolution of lactose digestion in adults. Hum Biol 69:605–628

Hollox E (2004) Genetics of lactase persistence-fresh lessons in the history of milk drinking. Eur J Hum Genet 13:267–269

Klein RG (2000) Archeology and the evolution of human behavior. Evol Anthropol 9:17–36

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Kretchmer N (1972) Lactose and lactase. Sci Am 227:70–78

Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. Nat Genet 22:139–144

Marjoram P, Donnelly P (1994) Pairwise comparisons of mitochondrial DNA sequences in subdivided populations and implications for early human evolution. Genetics 136:673–683

McCracken RD (1971) Lactase deficiency: an example of dietary evolution. Curr Anthropol 12:479–517

Midgley MS (1992) TRB culture: the first farmers of the North European plain. Edinburgh University Press, Edinburgh

Mulcare CA, Weale ME, Jones AL, Connell B, Zeitlyn D, Tarekegn A, Swallow DM, Bradman N, Thomas MG (2004) The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (*LCT*) (C−13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. Am J Hum Genet 74:1102–1110

Nei M, Saitou N (1986) Genetic relationship of human populations and ethnic differences in relation to drugs and food. In: Kalow W, Goedde HW, Agarwal DP (eds) Ethnic differences in reactions to drugs and other xenobiotics. Alan L Riss, New York, pp 21–37

Pluzhnikov A, Di Rienzo A, Hudson R (2002) Inferences about human demography based on multilocus analyses of noncoding sequences. Genetics 161:1209–1218

Poulter M, Hollox E, Harvey CB, Mulcare C, Peuhkuri K, Kajander K, Sarner M, Korpela R, Swallow DM (2003) The causal element for the lactase persistence/non-persistence polymorphism is located in a 1 Mb region of linkage disequilibrium in Europeans. Ann Hum Genet 67:298–311

Pritchard JK, Seielstad MT, Pérez-Lezaun A, Feldman MW (1999) Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol 16:1791–1798

Raymond M, Rousset F (1995) An exact test for population differentiation. Evolution 49:1280–1281

Reich DE, Goldstein DB (1999) Estimating the age of mutations using the variation at linked markers. In: Goldstein DB, Schlötterer C (eds) Microsatellites: evolution and applications. Oxford Unviversity Press, Oxford, pp 129–138

Relethford JH (2001) Genetics and the search for modern human origins. Wiley, New York

Rogers AR, Harpending H (1992) Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9:552–569

Salas A, Richards M, De La Fe T, Lareu MV, Sobrino B, Sánchez-Diz P, Macaulay V, Carracedo A (2002) The making of the African mtDNA landscape. Am J Hum Genet 71:1082–1111

Schneider S, Roessli D, Excoffier L (2000) Arlequim, ver. 2.000: a software for population genetics data analysis. University of Geneva, Geneva

Seixas S, Garcia O, Trovoada MJ, Santos MT, Amorim A, Rocha J (2001) Patterns of haplotype diversity within the serpin gene cluster at 14q32.1: insights into the natural history of the alpha1-antitrypsin polymorphism. Hum Genet 108:20–30

Simoons FJ (1970) Primary adult lactose intolerance and the milking habit: a problem in biological and cultural interrelations. II. A culture historical hypothesis. Am J Dig Dis 15:695–710

Slatkin M (2002) The age of alleles. In: Slatkin M, Veuille M (eds) Modern developments in theoretical population genetics: the legacy of Gustave Malécot. Oxford University Press, New York, pp 233–260

Slatkin M, Bertorelle G (2001) The use of intraallelic variability for testing neutrality and estimating population growth rate. Genetics 158:865–874

Slatkin M, Rannala B (2000) Estimating allele age. Annu Rev Genomics Hum Genet 1:225–249

Spedini G, Destro-Bisol G, Mondovi S, Kaptué L, Taglioli L, Paoli G (1999) The peopling of sub-Saharian Africa: the case study of Cameroon. Am J Phys Anthropol 110:143–162

Stephens M, Donnelly P (2003) A comparison of Bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73:1162–1169

Stephens M, Smith N, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. Am J Hum Genet 68:978–989

Stumpf MPH, Goldstein DB (2001) Genealogical and evolutionary inference with the human Y chromosome. Science 291:1738–1742

Swallow DM (2003) Genetics of lactase persistence and lactose intolerance. Annu Rev Genet 37:197–219

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

Tomás G, Seco L, Seixas S, Faustino P, Lavinha J, Rocha J (2002) The peopling of São Tomé (Gulf of Guinea): origins of slave settlers and admixture with the Portuguese. Hum Biol 74:397–411

Wall JD, Przeworski M (2000) When did the human population size start increasing?. Genetics 155:1865–1874

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Xu H, Fu Y (2004) Estimating effective population size or mutation rate with microsatellites. Genetics 166:555–563

Supplementary figure 1

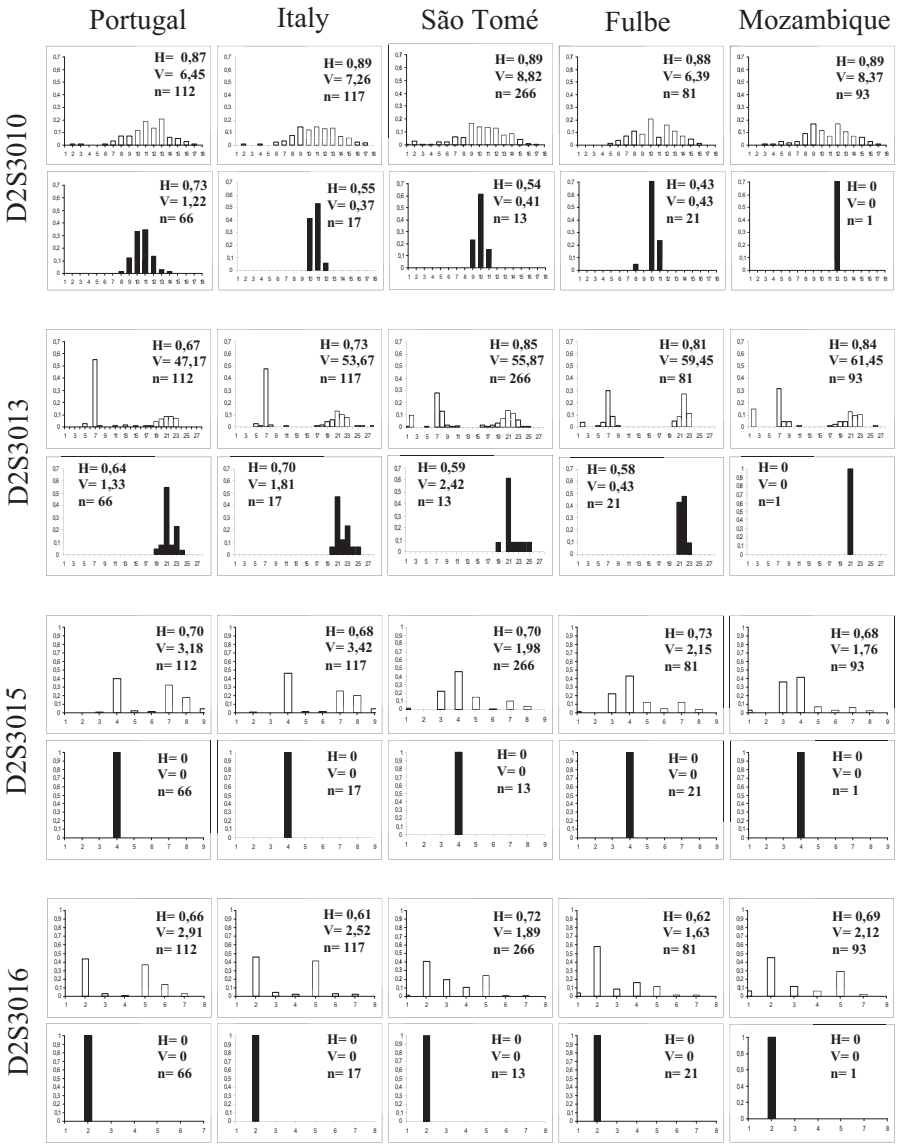


Fig S1: Microsatellite allele frequency distributions within C-G (white bars) and T-A (black bars) haplotypes defined by –13.910kb/-22.018kb SNPs in different populations. The estimated sizes of allele 1 in each microsatellite are: 188 bp for D2S3010; 133 bp for D2S3013; 190 bp for D2S3015 and 148 for D2S3016. H= heterozygosity; V= variance in repeat number; n= number of chromosomes.

1.2.1 Comments

1.2.1.1 Implications for the evolutionary history of lactase persistence

Our study shows that a limited number of microsatellite loci (STR) may provide sufficient resolution to reconstruct key aspects of the evolutionary history of lactase persistence, providing a useful alternative to approaches based on large numbers of single nucleotide polymorphisms (SNPs).

Our results support the notion that natural selection has been an important factor driving the -13910*T lactase persistence-associated allele to high frequencies in populations from southern Europe (Portuguese and Italians) and Africa (Fulbe). By using an approach based on the microsatellite diversity linked with the -13910*T allele in several geographically and genetically distinct populations, we were able to rule out possible confounding effects from recombination suppression and population history. Contrary to what happens with SNPs, the diversity generated by STRs is not influenced by recombination suppression. The STR high mutation rates allow also for a high resolution of the estimates of evolutionary parameters of recent events (Gaspar et al. 2004, Payseur and Cutter 2006). Additionally, the analysis of a set of heterogeneous populations, rules out confounding factors related with population history that studies based on a single population could not discard.

Our estimates of the time to the most recent common ancestor (TMRCA) showed that the -13910*T allele may be as recent as 12,500-7,500 years, indicating that that the origin of the -13910*T mutation is unlikely to substantially pre-date pastoralism or the differentiation of Europeans and Africans. In view of these results, the presence of this allele in some African populations (like the Fulbe from Cameroon), seems to have been the result of recent back-migrations from Eurasian populations, probably originated in Middle East. Moreover the results on dating are compatible with the “Culture-historical hypothesis” associating the increase in frequency of lactase persistence with the emergence of dairying as a component of the domestication of animals. On the other hand, the results obtained are clearly at odds with the so-called “Reverse cause argument”, which would require the mutation to be older in order to reach its actual distribution without a selective “push”.

By applying a phylogeographic interpretation of the distribution of the haplotypes defined by the -13910*C/T and -22018*G/A polymorphisms we were able to predict that even if the -13910*T allele was not the causative allele, an independent origin for lactase persistence, due to separate mutations, in Europe and in the majority of African populations, had to be claimed.

1.2.1.2 Recent developments in the study of lactase persistence evolution

Since our work has been published, the main developments in the study of the evolutionary history of lactase persistence have been done in the following topics: a) genome-wide scans for selection; b) antiquity of the -13910*T allele; c) geographic explicit studies about the dispersion of the -13910*T allele in Europe; and d) detection of genetic convergence underlying the lactase persistent trait.

a) Genome-wide studies trying to detect molecular signatures of selection have found that the haplotypes carrying the -13910*T allele present one of the clearest signals of positive selection ever found in the human genome (e.g. Voight et al. 2006, Sabeti et al. 2007), emphasizing the role of positive selective in driving the -13910*T lactase persistence allele to high frequencies in a short period of time.

b) Our age estimates for the -13910*T allele (7,500-12,500 years) were confirmed by subsequent studies on LCT haplotype variation (Mulcare 2006, Enattah et al. 2007). Moreover, studies of ancient DNA extracted from Neolithic individuals from Central Europe (Burger et al. 2007) and Scandinavia (Malmström et al. 2010) showed that the -13910*T allele was either absent or present at very low frequencies 6,000-8,000 years ago. The finding that the milking of ruminant animals was already practised in the northwestern Anatolia in the sixth and seventh millennia BC (Evershed et al. 2008) is also compatible with the idea that lactase persistence and dairying have co-evolved.

c) Two recent studies used forward computer simulations in order to address important questions regarding the mode and direction of spread of the -13910*T allele and the precise nature of the selective advantage conferred by lactase persistence (Itan et al. 2009, Gerbault et al. 2009). Both studies highlight the importance of the combined effects of the demographic expansion of farmers during the Neolithic and of selective pressures to explain the present distribution of the -13910*T allele. They differ however on the nature of the selective pressure in question. Itan et al. (2009) stress the importance of positive selection in all groups that adopted dairying cultures. On the other hand, Gerbault et al. (2009) found that in southern Europe, genetic drift alone was able to explain the frequencies of the -13910*T allele, while in the north-western Europe a specific selective factor should have been present. Such a scenario of increased selective intensity in high-latitude regions would be compatible with the “Calcium-absorption hypothesis” (Gerbault et al. 2009).

Itan et al. (2009) tried to infer the most probable place of origin of the co-evolution between lactase persistence and dairying, proposing that it lies in a region between the central Balkans and Central Europe (Figure I.3). The fact that the co-evolution between lactase persistence and dairying had occurred along the wave of advance of the Neolithic demographic expansion is, according to these same authors, central to explain the actual distribution of lactase persistence in Europe and, in particular, its highest frequencies in the Northwest of the continent (Itan et al. 2009).

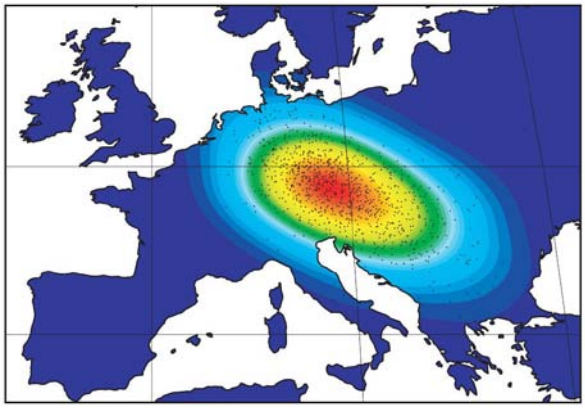


Figure I.3 Approximate posterior density of the region of origin for the co-evolution between the lactase persistence and dairying. Points represent regression-adjusted latitude and longitude coordinates from simulations accepted at the 0.5% tolerance level (retrieved from Itan et al. 2009).

d) As we stated in *Article 1*, besides -13910*T allele, additional mutations associated with lactase persistence had to be present in Africa in order to explain its present lactase persistence frequencies. Our prevision was fully confirmed by subsequent studies that showed that new sequence variants (-14010*G/C, -13915*T/G and -13907*C/G) in close proximity to the -13910*C/T SNP, were associated with lactase persistence in different populations from East Africa and Middle East (Ingram et al. 2007, Tishkoff et al. 2007, Enattah et al. 2008) (Figure I.4).

Tishkoff et al. (2007) found that the -14010*C allele is particularly frequent in Nilo-Saharan pastoralist populations like the Maasai (57.8% and 44.7% in Kenya and Tanzania, respectively) or the Datog (62.5%, Tanzania), as well as in Afro-Asiatic agro-pastoralist from

Tanzania like the Iraqw (58%) (Figure I.4 A). The oldest estimates of the -14010*C allele, ~6,000-7,000 years, were found in these populations, suggesting that the -14010*C allele is associated with the emergence of the pastoral tradition of the Great Lakes in East Africa, centred on cattle herding (Newman 1995).

The geographic distribution and allele age estimates for the -13915*G variant, suggest that this allele may have originated in the Arabian Peninsula around 4,000 years ago and then spread into northern regions of Middle East and different regions of Africa (Figure I.4 B) (Ingram et al. 2007, Enattah et al. 2008). The highest frequencies reported for this allele are found among Saudi Arabian and Jordan Bedouins (~50% and 40%, respectively) (Ingram et al. 2007). This allele is present also at relatively high frequencies in East Africa north to Tanzania, only in Afro-Asiatic-speaking populations, including the Beja pastoralists where it is observed at frequencies between 9.1% and 24.4% (Ingram et al. 2007, Tishkoff et al. 2007). The rise of the -13915*G allele seems to be associated with the domestication of the Arabian camel (Enattah et al. 2008).

The -13907*G is likely to be the variant with more restricted geographic distribution and it was found in Afro-Asiatic populations from Sudan, Kenya and Ethiopia, like the Beja and Afar pastoralists (~20% frequency) (Ingram et al. 2007, Tishkoff et al. 2007) (Figure I.4 C). The distribution of this allele may be related to the emergence of the kingdom of Aksum in the highlands of northern Ethiopia, around the first century AD, where agriculture and cattle herding seem to have had an important role in the kingdom's political economy (Curtin et al. 1995, Tishkoff et al. 2007).

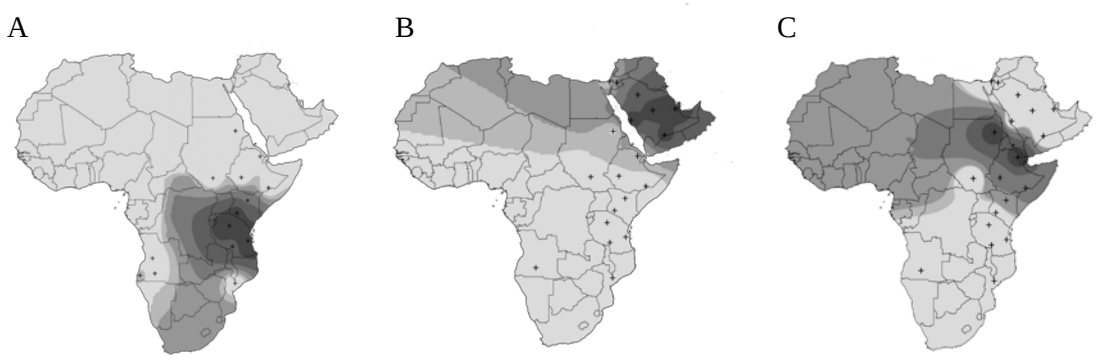


Figure I.4 Frequencies of the alleles -14010*C (A), -13915*G (B) and -13907*G (C), underlying LP in populations from East Africa and the Arabian Peninsula. The darker the colour, the higher the frequency of the allele. The sampled locations are marked with a cross. In the remaining places, the allelic frequencies were deducted assuming that there was a linear decrease of the frequency as far as the distance to the area of highest frequency decreased.

Taken together, these results show that in the last 2,000-12,000 years at least four causal variants associated with lactase persistence (-14010*C, -13915*G, -13910*T and -13907*G) have evolved independently, reaching high frequencies in diverse human groups with long histories of pastoralism and milk drinking. Additional studies have shown that it is likely that other alleles associated with lactase persistence are still to be uncovered especially in populations from western, southern and parts of eastern Africa, eastern Europe, and parts of western, central and southern Asia (e.g. Ingram et al. 2009, Itan et al. 2010, Xu et al. 2010). This convergent evolution of lactase persistence is a strong evidence that this trait has been subject to positive selection, since the observed distribution of multiple lactase persistence associated mutations would hardly be generated by chance. At the same time, these results illustrate how cultural processes- as the dairying farming- can provide strong selective pressures affecting the rate of change of allele frequencies.

The identification of several lactase persistence-associated alleles evolving independently in different human populations as a result of local adaptation to a specific subsistence pattern, highlights the need for fine-scale geographic sampling when searching for new genetic variants under selection (Tishkoff et al. 2007). Indeed, genome wide scans of selection have been mainly based on population panels that are not representative of the worldwide genetic diversity. For example, a scan for selection signals in the HapMap data set (where individuals of northern and western European origin are present), detected a selection signal in Europeans in the lactase region (Voight et al. 2006). On the contrary, another study based on the HGDP-HGDP panel of samples (where there is a lack of North European groups), failed to identify the lactase gene as a candidate for recent positive selection in Europe (Lopéz Herráez et al. 2009). Both studies, failed to identify any signal of selection for lactase persistence in Africa, reflecting the shortage of African samples in these panels.

Both selection and demographic processes seem to have been important factors shaping the actual distribution of the lactase persistence- associated alleles worldwide. The expansion of Neolithic farmers may have been crucial to the distribution of the -13910*T allele in Europe (Itan et al. 2009, Gerbault et al. 2009). On the other hand, the migratory movements of pastoralist communities seem to have been important factors in the distribution of the several lactase persistence- associated alleles present in Africa (Tishkoff et al. 2007, Ingram et al. 2009). Given that adult milk consumption and lactase persistence are assumed to have spread along with pastoralism, the mutations that are associated with different occurrences of lactase persistence in different geographic locations, constitute unforeseen tools to reconstruct the migratory events that led to the dispersion of pastoralist populations.

In order to further explore the distribution of the lactase persistence-associated alleles, we were able to additionally investigate the presence of those mutations in a set of populations from Cabo Verde, Israel (Palestinians), Pakistan, Cameroon (Fulbe), Angola and Mozambique (Figure I. 5).

In the population from Cabo Verde we found the -13910*T allele at a frequency of 10% (Figure I.5). It is likely that the occurrence of the -13910*T allele in this population is due to recent admixture with Europeans, mainly from Portugal. This is in agreement with studies based on other genetic markers reporting high levels of European admixture in the Cape Verdean population (Parra et al. 1995, Gonçalves et al. 2003).

In Israel we found both the -13910*T and -13915*G alleles at 5% frequency and, in Pakistan, we only detect the presence of the Eurasian -13910*T allele at a frequency of 16%. These observations seem to indicate that the -13915*G allele presents a low dispersion to the north and east of the Arabian Peninsula (Figure I.5).

In the Fulbe population we found that the -13910*T and the -13915*G alleles are present at 10% and 3% frequencies, respectively (Figure I. 5). The Fulbe population, along with the Hausa group from Cameroon (with a 13,9% frequency of the -13910*T allele) (Mulcare et al. 2004), represent exceptions to the general rule of absence of the -13910*T allele south of the Sahara. Previous studies testing for the presence of the -13910*T allele (the unique mutation identified till then) in North African populations, found that this mutation was present in the Mozabites from Algeria (21.7%) (Bersaglieri et al. 2004) and in three Berber populations from Morocco and Algeria (15%) (Myles et al. 2005). Myles et al. (2005) propose that the observation of the -13910*T allele in North Africa could be explained by the Neolithic spread of ovicaprid herders from Middle East, speaking some form of Berber. The admixture between Fulbe and Afro-Asiatic Berber nomads (Curtin et al. 1995) could constitute the reason for the presence of the -13910*T allele in the Fulbe population. However, it is not known whether the introduction of both the -13915*G and -13910*T alleles in northern and western Africa occurred simultaneously. It could be that the presence of the -13915*G observed in these regions resulted from a subsequent migratory movement linked with the Arab expansion. It is interesting to note that in the populations from Cameroon studied by Ingram et al. (2007), the Shuwa Arabs (Semitic) present only the -13915*G at a frequency of 6.3%, while the Mambila (Niger-Congo, non-pastoralists) did not present any of the mutations. Other cases of populations presenting more than one lactase persistence- associated allele have been reported. For example, Ingram et al. (2009) reported in Somali camel-herders from Ethiopia, the occurrence of the -13910*T, -13915*G, -13907*G and -14010*C mutations. The authors interpret the

simultaneous presence of these mutations as a mark of past contact between migratory milk-drinking peoples through shared cultural practices.

In *Article 2* (see Part II) we looked for LP alleles in several groups from Southwest Angola and Mozambique. We found that the lactase persistence variant -14010*C was present at a frequency of 6% in the Kuvale people (Figure I.5), one of the most cattle-exclusive pastoral Bantu-speaking peoples (Estermann 1961). Since the -14010*C allele is especially frequent among Nilo-Saharan and Afro-Asiatic-speaking pastoral populations from Kenya and Tanzania (Tishkoff et al. 2007), our observation provides genetic evidence for a link between the relatively isolated southwestern Africa pastoral scene and the major cattle herding centers of East Africa. Recently, the -14010*C variant has been reported to occur at 13% frequency in the Xhosa population from South Africa (Figure I.5), which is well known for extensive cultural and genetic interactions with Khoisan pastoralist groups from whom it borrowed a number of click words (Torniainen et al. 2009). On the other hand, we found no lactase persistence variants in Bantu communities from southern Mozambique (Figure I.5), that are somewhat related to the Xhosa but did not interacted as extensively with the Khoisan pastoralists. In the comments to *Articles* 2 and 3 (Part II) these results are discussed with further detail.

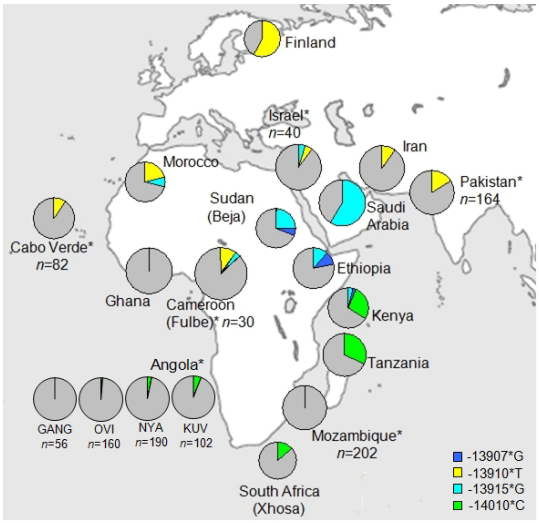


Figure I.5 Distribution of LP allele frequencies in selected populations. Our data is marked with an asterisk. The population from Angola is divided in the major ethnolinguistic groups sampled. The remaining data are from Ingram et al. (2007), Imtiaz et al. (2007), Tishkoff et al. (2007), Enattah et al. (2008) and Torniainen et al. (2009). *n*= number of haplotypes studied. Samples from Israel and Pakistan are from the Human Genome Diversity Panel (Cann et al. 2002). GANG = Ganguela; OVI = Ovimbundu; NYA = Nyaneka-Nkhumbi; KUV = Kuvale.

References

Bersaglieri, T., P. C. Sabeti, N. Patterson, T. Vanderploeg, S. F. Schaffner, J. A. Drake, M. Rhodes, D. E. Reich, and J. N. Hirschhorn. 2004. Genetic signatures of strong recent positive selection at the lactase gene. *Am J Hum Genet* 74:1111-20.

Burger, J., M. Kirchner, B. Bramanti, W. Haak, and M. G. Thomas. 2007. Absence of the lactase-persistence-associated allele in early Neolithic Europeans. *Proc Natl Acad Sci U S A* 104:3736-41.

Cann, H. M., C. de Toma, L. Cazes , M. F. Legrand, V. Morel, L. Piouffre, J. Bodmer, W. F. Bodmer, B. Bonne-Tamir, A. Cambon-Thomsen, Z. Chen, J. Chu, C. Carcassi, L. Contu, R. Du, L. Excoffier, G. B. Ferrara, J. S. Friedlaender, H. Groot, D. Gurwitz, T. Jenkins, R. J. Herrera, X. Huang, J. Kidd, K. K. Kidd, A. Langaney, A. A. Lin, S. Q. Mehdi, P. Parham, A. Piazza, M. P. Pistillo, Y. Qian, Q. Shu, J. Xu, S. Zhu, J. L. Weber, H. T. Greely, M. W. Feldman, G. Thomas, J. Dausset, and L. L. Cavalli-Sforza. 2002. A human genome diversity cell line panel. *Science* 12;296(5566):261-2.

Curtin, P., S. Feierman, L. Thompson, and J. Vansina. 1995. *African history: from earliest times to independence*. London: Longman.

Enattah, N. S., A. Trudeau, V. Pimenoff, L. Maiuri, S. Auricchio, L. Greco, M. Rossi, M. Lentze, J. K. Seo, S. Rahgozar, I. Khalil, M. Alifrangis, S. Natah, L. Groop, N. Shaat, A. Kozlov, G. Verschubskaya, D. Comas, K. Bulayeva, S. Q. Mehdi, J. D. Terwilliger, T. Sahi, E. Savilahti, M. Perola, A. Sajantila, I. Järvelä, and L. Peltonen. 2007. Evidence of still-ongoing convergence evolution of the lactase persistence T-13910 alleles in humans. *Am J Hum Genet* 81:615-25.

Enattah, N. S., T. G. Jensen, M. Nielsen, R. Lewinski, M. Kuokkanen, H. Rasinperä, H. El-Shanti, J. K. Seo, M. Alifrangis, I. F. Khalil, A. Natah, A. Ali, S. Natah, D. Comas, S. Q. Mehdi, L. Groop, E. M. Vestergaard, F. Imtiaz, M. S. Rashed, B. Meyer, J. Troelsen, and L. Peltonen. 2008. Independent introduction of two lactase-persistence alleles into human populations reflects different history of adaptation to milk culture. *Am J Hum Genet* 82:57-72.

Estermann, C. 1961. *Etnografia do sudoeste de Angola: o grupo étnico Herero*. Volume 3. Lisboa, Junta de investigações do Ultramar.

Evershed, R. P., S. Payne, A. G. Sherratt, M. S. Copley, J. Coolidge, D. Urem-Kotsu, K. Kotsakis, M. Ozdogan, A. E. Ozdogan, O. Nieuwenhuyse, P. M. Akkermans, D. Bailey, R. R. Andeescu, S. Campbell, S. Farid, I. Hodder, N. Yalman, M. Ozbasaran, E. Bicakci, Y. Garfinkel, T. Levy, and M. M. Burton. 2008. Earliest date for milk use in the Near East and southeastern Europe linked to cattle herding. *Nature* 455:528-31.

Gaspar, P., S. Seixas, and J. Rocha. 2004. Genetic variation in a compound short tandem repeat/Alu haplotype system at the SB19.3 locus: properties and interpretation. *Hum Biol* 76:277-87.

Gerbault, P., C. Moret, M. Currat, and A. Sanchez-Mazas. 2009. Impact of selection and demography on the diffusion of lactase persistence. *PLoS One* 4:e6369.

Gonçalves, R., A. Rosa, A. Freitas, A. Fernandes, T. Kivisild, R. Villems, and A. Brehm. 2003. Y-chromosome lineages in Cabo Verde Islands witness the diverse geographic origin of its first male settlers. *Hum Genet* 113:467-72.

Imtiaz, F., E. Savilahti, A. Sarnesto, D. Trabzuni, K. Al-Kahtani, I. Kagevi, M. S. Rashed, B. F. Meyer, and I. Järvelä. 2007. The T/G 13915 variant upstream of the lactase gene (LCT)

is the founder allele of lactase persistence in an urban Saudi population. *J Med Genet* 44:e89.

Ingram, C. J., M. F. Elamin, C. A. Mulcare, M. E. Weale, A. Tarekegn, T. O. Raga, E. Bekele, F. M. Elamin, M. G. Thomas, N. Bradman, and D. M. Swallow. 2007. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum Genet* 120:779-88.

Ingram, C. J., T. O. Raga, A. Tarekegn, S. L. Browning, M. F. Elamin, E. Bekele, M. G. Thomas, M. E. Weale, N. Bradman, and D. M. Swallow. 2009. Multiple Rare Variants as a Cause of a Common Phenotype: Several Different Lactase Persistence Associated Alleles in a Single Ethnic Group. *J Mol Evol* 69(6):579-88.

Itan, Y., A. Powell, M. A. Beaumont, J. Burger, and M. G. Thomas. 2009. The origins of lactase persistence in Europe. *PLoS Comput Biol* 5:e1000491.

Itan, Y., B. Jones, C. Ingram, D. Swallow and, M. G. Thomas. 2010. A worldwide correlation of lactase persistence phenotype and genotypes. *BMC Evol Biol* 2010, 10:36.

López Herráez, D., M. Bauchet, K. Tang, C. Theunert, I. Pugach, J. Li, M. R. Nandineni, A. Gross, M. Scholz, and M. Stoneking. 2009. Genetic variation and recent positive selection in worldwide human populations: evidence from nearly 1 million SNPs. *PLoS One* 4:e7888.

Malmström, H., A. Linderholm, K. Lidén, J. Storå, P. Molnar, G. Holmlund, M. Jakobsson, and A. Götherström. 2010. High frequency of lactose intolerance in a prehistoric hunter-gatherer population in northern Europe. *BMC Evol Biol* 10:89.

Mulcare, C. A., M. E. Weale, A. L. Jones, B. Connell, D. Zeitlyn, A. Tarekegn, D. M. Swallow, N. Bradman, and M. G. Thomas. 2004. The T allele of a single-nucleotide polymorphism 13.9 kb upstream of the lactase gene (LCT) (C-13.9kbT) does not predict or cause the lactase-persistence phenotype in Africans. *Am J Hum Genet* 74:1102-10.

Mulcare, C. A. 2006. The evolution of the lactase persistence phenotype. Ph.D. thesis, University of London.

Myles, S., N. Bouzekri, E. Haverfield, M. Cherkaoui, J. M. Dugoujon, and R. Ward. 2005. Genetic evidence in support of a shared Eurasian-North African dairying origin. *Hum Genet* 117:34-42.

Newman, J. L. 1995. *The peopling of Africa: a geographic interpretation*: Yale University Press New Haven and London.

Parra, E. J., J. C. Ribeiro, J. L. Caeiro, and A. Riveiro. 1995. Genetic structure of the population of Cabo Verde (west Africa): evidence of substantial European admixture. *Am J Phys Anthropol* 97:381-9.

Payseur, B. A., and A. D. Cutter. 2006. Integrating patterns of polymorphism at SNPs and STRs. *Trends Genet* 22:424-9.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, X. Xie, E. H. Byrne, S. A. McCarroll, R. Gaudet, S. F. Schaffner, E. S. Lander, International HapMap Consortium, K. A. Frazer, D. G. Ballinger, D. R. Cox, D. A. Hinds, L. L. Stuve, R. A. Gibbs, J. W. Belmont, et al. 2007. Genome-wide detection and characterization of positive selection in human populations. *Nature* 449:913-8.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas.

2007. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39:31-40.

Torniainen, S., M. I. Parker, V. Holmberg, E. Lahtela, C. Dandara, and I. Järvelä. 2009. Screening of variants for lactase persistence/non-persistence in populations from South Africa and Ghana. *BMC Genet* 10:31.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard. 2006. A map of recent positive selection in the human genome. *PLoS Biol* 4:e72.

Xu, L., H. Sun, X. Zhang, J. Wang, D. Sun, F. Chen, J. Bai1, and S. Fu1. 2010. The -22018A allele matches the lactase persistence phenotype in Northern chinese populations. *Scand J Gastroenterol* 45(2): 168-174.

PART 2

On the edge of Bantu Expansions:

Genetic studies in Southwest Angola and Mozambique

2.1. Introduction

Among the complex series of demographic events that have occurred in Africa, the dispersal of Bantu-speaking agriculturalists stands as one of the most impressive examples of long-range human migrations.

It is generally accepted that Bantu expansions started in the adjacent grasslands of Nigeria-Cameroon around 5,000 years ago and progressed throughout Central, Eastern and Southern Africa (Curtin et al 1995, Newman 1995). The common and recent ancestry of the different Bantu-speaking populations is reflected in the similarity at the lexical and morphological level of the several Bantu languages. For example when referring to "people", the Duala from Cameroon say *bato*, individuals in Kongo, on the Atlantic coast, say *bantu* and the Swahili, on the Indian Ocean coast, say *watu* (Curtin et al. 1995).

The Bantu expansions may have been triggered by developments related with the mastering of the tropical agriculture (Curtin et al. 1995). Indeed, a new type of economy has arisen based on root (e.g. the yams) and tree (e.g. oil palms) crops supplemented by some trapping, gathering and fishing (Curtin et al. 1995, Newman 1995). As some Bantu-speaking populations moved southwards, away from the rainforest, they acquired knowledge about cultivation of grain crops (e.g. millets and sorghum) and cattle herding that became fundamental in their adaptation to drier environments (Curtin et al. 1995). Bantu expansions seem to be also associated with the diffusion of metallurgy further south from both the Great Lakes region and Gabon-Congo centres, around 800 BC (Curtin et al. 1995). The knowledge of the iron making seems to have enabled Bantu-speaking populations to clear larger fields and to build digging sticks, important tasks to the practice of woodland and savannah agriculture (Newman 1995).

Despite the contribution from different areas of research, like linguistics, archaeology and genetics, there is still no consensus about many aspects of the history of Bantu populations, including the major dispersal routes followed by Bantu speakers and the nature of the interactions between spreading populations. Current views about Bantu expansions based on archaeological and linguistic data can be divided into two main models. According to the most widely accepted dispersion model, the Bantu expansions involved an early population split into two major routes leading to the separation of east and west Bantu primary language branches (Newman 1995), one following an eastern path, first circumventing the rainforest to the area of the Great Lakes, and then proceeding to Southeast Africa; the other, moving south, through the rainforest into the arid steppes of Southwest Africa (Figure II.1A). The results of Holden (2002) support this hypothesis by showing a clear East-West divergence of Bantu languages (Figure II.1C). The alternative model challenges the early split between western and eastern branches of Bantu languages, proposing a single passage through the rainforest, followed by a later spatial divergence in subequatorial Africa (Figure II.1B) (Ehret 1998). In accordance with this

hypothesis, Rexová et al. (2006) found that all Bantu languages located south and east of the rainforest areas of Congo-Kinshasa formed a monophyletic clade (Figure II.1D). In particular, southwestern populations were not shown to cluster with populations up north (Figure II.1D).

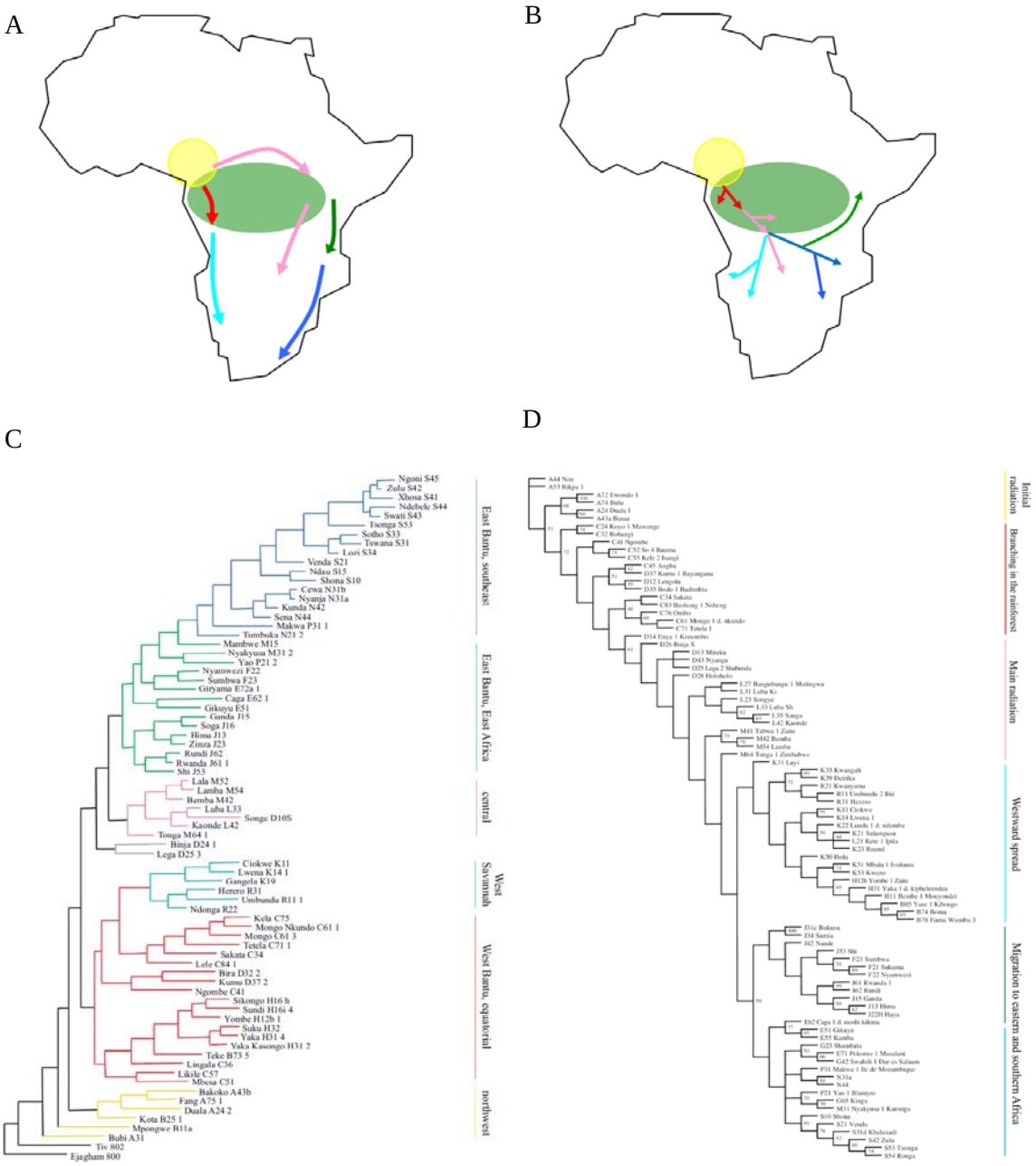


Figure II.1 (A) and (B) Maps showing possible directions of dispersion of Bantu-speaking populations in accordance with models proposed by Newman (1995) and Ehret (1998), respectively. (C) and (D) Maximum parsimony trees relating Bantu languages, retrieved from Holden (2002) and Rexová et al. (2006), respectively.

Other important aspect of the Bantu expansions is the degree of admixture between Bantu newcomers and local indigenous populations, like Pygmies, Khoisan and Nilo-Saharans. Cross-cultural linguistic comparisons have given important insights about the social, cultural and economic interactions that the Bantu populations have established along their dispersals (e.g. Ehret 1998). However, the linguistic data do not give information about the biological significance of such interactions. Genetic data have great potential for investigating this and other aspects of the complex population history of Bantu expansions.

At the beginning of our study, most of the available genetic information had been gathered in phylogeographic studies of Y-chromosome (NRY) and mitochondrial DNA (mtDNA) variation. These studies found evidence for a high similarity between Bantu groups and identified several mtDNA haplogroups and NRY lineages likely to be associated with the Bantu migrations (Thomas et al. 2000, Underhill et al. 2001, Cruciani et al. 2002, Salas et al. 2002). Other haplogroups, characteristic of specific population groups like the Pygmies, Khoisan or East African non-Bantus, have been used to evaluate the degree of admixture between different Bantu population and those groups (e.g. Soodyall and Jenkins 1993, Underhill et al. 2001, Salas et al. 2002). Although NRY and mtDNA uniparentally inherited markers are highly informative, the analysis of additional independently evolving genetic systems is clearly needed in order to deepen our understanding of the demographic history of Bantu expansions.

Apart from the paucity of genetic markers that are used to analyze genetic variation, current studies on Bantu expansions were also hampered by poor sampling within the vast region encompassing sub-equatorial Africa. For example, no genetic information was available from southern Angola, crucial for understanding the dramatic push of Bantu peoples towards the arid steppes and deserts of Southwest Africa. The extension of the number of sampled regions will be crucial to discriminate between contrasting population-history models.

In this work we focused on populations from Angola and Mozambique, encompassing areas that have been poorly sampled in previous studies of African genetic diversity. The assumption of one of the two models for the Bantu expansions previously described has implications on the degrees of proximity expected between these two populations. According to models favouring an early split between eastern and western Bantu groups (Newman 1995; Holden 2002), these populations lie on opposite edges of the two more ancient Bantu dispersal routes and should be maximally divergent. In the alternative model, favouring a single passage of ancestral Bantu peoples through the rainforest and a subsequent radiation south of the equatorial forest (Ehret 1998; Rexová et al. 2006), the genetic divergence between Mozambican and Angolan populations is expected to be reduced and the opportunities for gene flow increased.

2.1.1 Southwestern Angola

The region of Angola remained a persistent gap in studies of African genetic variation. Although new genetic data has become available from Kimbundu and Bakongo speakers from northern Angola and Cabinda (Plaza et al. 2004, Beleza et al. 2005), no information existed on the broad area encompassing the dry woodlands to the south of the Cuanza river. The study of populations inhabiting the region south of the Cuanza River is particularly important in the context of the study of Bantu expansions to understand the migration of the "West Savannah" Bantu-speaking peoples out of the rain forest into the arid steppes and deserts of Southwest Africa. In fact, the fertility of the soils has constituted an important factor in the expansion of human agricultural societies. It is known that the pastoralism, by promoting the conversion of low quality plant resources into portable, high quality foods as meat and milk, constitutes an important factor allowing the exploration and the creation of settlements in regions that given their aridity, could not be otherwise occupied (Leonard and Crawford 2002).

Bantu groups in southwestern Angola form a broad cultural and economic cluster relying on cattle raising to various degrees (Redinha 1971). Among them, the Herero are the most cattle-dependent economy in Bantu Africa (Newman 1995). The origin of this group is uncertain. Oral traditions point to a homeland location in the upper Zambezi or even in the region of the Great Lakes (Estermann 1961). However, linguistic evidence does not support this hypothesis given that Herero language presents a closer relationship with the other "West Savannah" languages than with Bantu languages from Central and East Africa (Holden 2002, Rexová et al. 2006) (Figure II.1 C and D).

The relative isolation of Southwest Africa from the major East African pastoral centres represents an important challenge for the identification of the processes that led to the emergence of a cattle-herding zone in the southwestern periphery of the Bantu expansions. The archaeological record in eastern and southern Africa documents the spread of herding between these two regions between c.200 BC onwards to the sixth century AD (Curtin et al. 1995). Linguistic analyses indicate that the transmission of cattle culture from eastern Africa was unrelated with the Bantu dispersals and the mediation from Cushitic (Blench 2009) or Eastern Sudanic (Ehret 1998) language groups has been proposed. Other kinds of evidences linking the pastoral communities from the East and Southwest Africa have been described, like shared cultural features, such as mat huts, sandals and butter-making equipment (Figure II.2 A) (Blench 2009). Khoe speakers seem to have played an important role in the spread of herding towards western regions of southern Africa. There are evidences showing that around 200 BC, in the region of the Middle Zambezi valley, this Khoisan group, who had hitherto practiced

hunter-gathering, start to herd cattle and sheep, and spread southerly (Figure II.2 B) (Curtin et al. 1995, Newman 1995). It is possible that the Khoe had acquired their pastoral culture following contact with pastoralist communities originating from East Africa (Blench 2009).

The adoption of pastoralism by some Bantu groups from southwestern Angola, like the Herero, is still an unclear issue. The cultural and geographical proximity between Bantu pastoralists from southern Angola and the Khoe pastoralists poses intriguing questions about possible interactions between these two groups. The genetic analysis of groups settled in southwestern Angola is therefore important to understand to what extent Bantu groups have interacted with autochthonous groups from southern Angola and the role played by Khoe herders in the adoption of pastoralism by Bantu speakers.

A

B

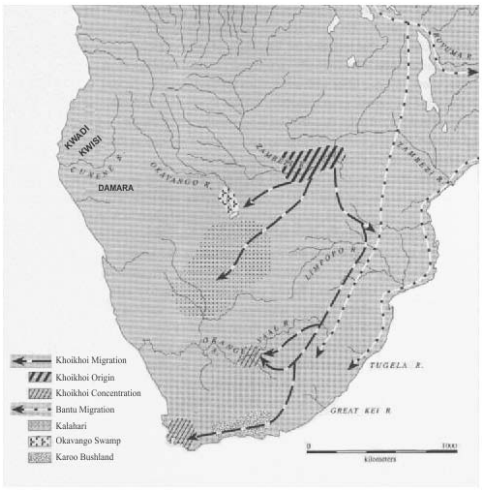


Figure II.2 (A) *Eholo*, leather bag used for butter production by Bantu-speaking pastoralists along the Southern Angola/Namibia borderland. This artefact, which is found among cattle producers from the Horn of Africa, Ethiopia and all the way to Egypt, represents one of the ethnographic evidences supporting a link between the pastoral scenes from the East and Southwest Africa (Blench 2009). (B) Map depicting the advance of Bantu and Khoikhoi (Khoe) groups across the Zambezi River (retrieved from Newman 1995).

In order to better understand the peopling of the western edge of the Bantu expansions we studied the NRY, mtDNA and lactase persistence genetic variation in "West Savannah" Bantu-speaking groups from southwestern Angola. We analysed the data in the context of regional and continental genetic diversity by assessing the differentiation between these groups and their levels of admixture with Khoisan-speaking populations, and by examining the relationship between southwestern Angola and other areas of Africa. Furthermore, we combined our dataset with published data on NRY and mtDNA variation from Southeast Africa to infer key demographic parameters underlying the history of Bantu expansions.

The results of this study are presented in the *Article 2*. A more detailed discussion of some aspects of the results is presented afterwards as a commentary.

2.1.2 Mozambique

Mozambique occupies a strip over a wide latitudinal cline in southeastern Africa, connecting the hinterland with the Indian Ocean coast. It is crossed by different rivers, which can be viewed both as geographic barriers and "highways" linking the interior and the coast. For example, the Zambezi River almost links the eastern and western coasts of Africa. On the other hand, in Mozambique, the S allele of the β-globin gene is almost restricted to regions located to the north of this river (Lehmann and Huntsman 1974), suggesting that the Zambezi River may be an important barrier to gene flow and/or an ecological transition. It is interesting to note that Curtin et al. (1995), in their definition of southern Africa, consider the Zambezi River as its northeast limit. Moreover, several ecoregions can be found across the country as the Eastern Miombo woodlands in the north, southern Zambezian-Inhambane coastal forest mosaic in the coast or the Maputaland coastal forest mosaic in the south (http://www.worldwildlife.org).

According to Ehret (1998), populations presently inhabiting Mozambique, trace their origins back to the so-called Mashariki Bantu group who settled along a broad front in the immediate west of the western rift. The author divided the Mashariki group into two subgroups of communities (the Kaskazi and the Kusi) reflecting the historical geography of the early Mashariki settlement (the two words were taken from the Swahili for "northwind" and "southwind", respectively) (Ehret 1998). In Mozambique, the Kaskasi settled in the far northern regions and include the Swahili, Yao and Makonde speaking-populations. The Kusi comprised the remaining Bantu speakers in the country who had moved to the south (for example, the Nyanja, Makua, Tewe and the Chopi).

The Mashariki Bantu arrived in the African Great Lakes region from the west around 1,000 BC, and there they contacted with several communities inhabiting the African Great Lakes region at that time (Ehret 1998). From the interactions with these communities, Mashariki Bantu learnt new cultural practices and technologies, like grain cultivation and iron working knowledge, which enabled them to undertake a series of expansions all across the eastern and southeastern parts of Africa (Ehret 1998). Around the early first millennium AD (~c.300 BC to the sixth century AD), Bantu-speaking farmers began to move into southern Africa from the region of Great Lakes and became the major population of the region (Curtin et al. 1995, Ehret 1998).

During the process of territorial expansion, different groups adapted to different ecological niches, becoming isolated and giving rise to the several ethnic and linguistic groups presently observed. Additionally, each group may have had different degrees of interaction with pre-existent populations and with individuals originated from different geographic regions. Consequently, each particular community of Bantu-speakers created its own local synthesis on the basis of the common heritage (Curtin et al. 1995). In the northern part of Mozambique, populations like the Yao and Swahili, seemed to have had a much more steady contact with other populations of East Africa up to the north (Curtin et al. 1995, Newman 1995). In the nineteenth century, Yao traders had brought ivory from the woodlands east of Lake Malawi to Kilwa to exchange it for various manufactured items (Newman 1995), linking the interior of East Africa with the coast. Other linguistic group, paradigmatic because its multiple cultural influences, is the Swahili Bantu language, spoken by coastal groups in the north of Mozambique. Along with the mastering of technologies characteristic of Bantu speakers (such as iron working, grain and yam cultivation), the Swahili specialized in the maritime coastal way of living by learning, for example, how to construct wells and built boats for the ocean (Curtin et al. 1995). Influences from southwestern Asia were clearly significant to Swahili identity and such a double heritage characterizes its culture (Curtin et al. 1995, Newman 1995). Other group, the Tsonga Bantu speakers, including the Ronga and the Shangaan (Lewis 2009), lived in small, scattered farmsteads in the southern region of Mozambique (Curtin et al. 1995, Newman 1995). They have adapted to an environment where tse tse flies made impossible for them to breed cattle in most parts of their territory, but where it was possible to have fowl and goats. Their economy relied also on fishing, shellfish collecting and grain cultivation (Curtin et al. 1995, Newman 1995). In the southern part of Mozambique, Bantu communities seem to have had important contacts with the pastoral Khoe. This can be observed in the Bantu herding vocabulary in Southeast Africa where, for example, a new term for cattle, *-kòmò- almost completely displaced the older Mashariki Bantu *-gòmbè (Ehret 1998). Despite that, cattle

raising must have been a relatively uncommon activity among the southeastern African Bantu communities until more recently (Ehret 1998).

A total of 43 languages are listed for Mozambique (Lewis 2009). All these languages, except the Portuguese, belong to the Narrow Bantu, Central branch of the Niger-Congo family, reflecting, on the one hand, the remarkably rapid linguistic differentiation undertaken since the arrival of the Bantu-speakers from the north and, at the other hand, the proximity between all the languages.

The geographic and cultural heterogeneity of Mozambique makes this country particularly interesting to study how genetic diversity is distributed throughout all these different micro-environments. The study of the patterns of genetic diversity at regional scale is being the focus of an increasing number of studies (Biswas et al. 2009). The knowledge of regional patterns of variation may be highly informative about the human population history, the way humans adapt to different micro-environments, interact and differentiate from each other. Additionally, this information can give insights about the way genetic diversity could be modelled in response to different cultural, social, climatic, and physical features of the environment.

Previous studies about the genetic diversity of Mozambican populations have mainly analysed the uniparentally transmitted mtDNA and NRY data (Pereira et al. 2001, Pereira et al. 2002, Salas et al. 2002). The data point to a high genetic homogeneity of the different populations analysed, likely reflecting their common origin, high levels of gene flow, or both. Notwithstanding, virtually no Y-data is available from Mozambique north to the Save river. When analyzed in a broad African context, mtDNA data indicate that the Mozambican populations are closely related to Western-Central populations, in accordance with its putative origin in the area comprised between present-day Nigeria and Cameroon. It presents however an eastern African component higher than western Bantu populations, which seems to be the result of an en route assimilation of eastern African non-Bantu lineages (Pereira et al. 2001, Salas et al. 2002). It is also interesting to note that, despite the total replacement of Khoisan languages in Mozambique, a small level of Khoisan assimilation was observed at both the NRY and mtDNA levels (~9% and 6%, respectively) (Pereira et al. 2001, Pereira et al. 2002, Salas et al. 2002). Contrasting with mtDNA data, which did not present neither European, east or south Asian lineages (Pereira et al. 2001, Salas et al. 2002), the analysis of the NRY data showed that 6% of the lineages analysed were of European origin (Pereira et al. 2001).

Thus, the knowledge of the patterns of genetic diversity of the different populations from Mozambique is far from being accomplished. Further studies based on genetic markers exploring the full information contained in the genome beyond the mtDNA and NRY components as well as a more comprehensive sampling of Mozambican populations are necessary for a full understanding of the evolutionary history of this country. We developed a battery of 14 independently evolving non-recombining autosomal UEPSTRs with widespread chromosomal locations to characterize a countrywide population sample from Mozambique. The newly developed UEPSTRs were also used to study the split between western and eastern Bantu dispersal routes by comparing the Bantu groups sampled in Mozambique with representative Bantu-speaking populations from southwestern Angola

The results of this study are presented in the *Article 3*. A more detailed discussion of some aspects of the results is presented afterwards as a commentary.

References

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

Biswas S., L. B. Scheinfeldt, and J. M. Akey. 2009. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am J Hum Genet* 84(5): 641–650.

Blench, R. 2009. Was there and interchange between Cushitic pastoralists and Khoisan speakers in the prehistory of Southern Africa and how can this be detected? *Sprache und Geschichte in Afrika*, 20. (in press).

Cruciani, F., P. Santolamazza, P. Shen, V. Macaulay, P. Moral, A. Olckers, D. Modiano, S. Holmes, G. Destro-Bisol, V. Coia, D. C. Wallace, P. J. Oefner, A. Torroni, L. L. Cavalli-Sforza, R. Scozzari, and P. A. Underhill. 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet* 70:1197-214.

Curtin, P., S. Feierman, L. Thompson, and J. Vansina. 1995. *African history: from earliest times to independence*. London: Longman.

Ehret, C. 1998. *An African Classical Age: Eastern & Southern Africa in World History,1000B.C. to A.D.400*. Charlottesville: University Press of Virginia.

Estermann, C. 1961. *Etnografia do Sudoeste de Angola: O Grupo Étnico Herero*. Vol. 3. Lisbon: Junta de Investigações do Ultramar.

Holden, C. J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc Biol Sci* 269:793-9.

Lehmann, H., and R. G. Huntsman. 1974. Man's haemoglobins including the haemoglobinopathies and their investigation. Amsterdan. Oxford. North-Holland Publishing, p. 308.

Leonard, W. R., and M. H. Crawford. 2002. "The biological diversity of herding populations: an introduction," in *Human biology of pastoral populations*, ed. W. Leonard and M. H. Crawford, 1-9. Cambridge: Cambridge University Press.

Lewis, M. P., (ed). 2009. *Ethnologue: Languages of the World*. Sixteenth edition. Tex.: SIL International. Dallas. Online version: http://www.ethnologue.com.

Newman, J. L. 1995. *The peopling of Africa: a geographic interpretation*: Yale University Press New Haven and London.

Pereira, L., V. Macaulay, A. Torroni, R. Scozzari, M. J. Prata, and A. Amorim. 2001. Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* 65:439-58.

Pereira, L., L. Gusmão, C. Alves, A. Amorim, and M. J. Prata. 2002. Bantu and European Y-lineages in Sub-Saharan Africa. *Ann Hum Genet* 66:369-78.

Plaza, S., A. Salas, F. Calafell, F. Corte-Real, J. Bertranpetit, A. Carracedo, and D. Comas. 2004. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. *Hum Genet* 115:439-47.

Redinha, J. 1971. *Distribuição Étnica de Angola*. Luanda: Cita.

Rexová, K., Y. Bastin, and D. Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften* 93:189-94.

Salas, A., M. Richards, T. De la Fe, M. V. Lareu, B. Sobrino, P. Sanchez-Diz, V. Macaulay, and A. Carracedo. 2002. The making of the African mtDNA landscape. *Am J Hum Genet* 71:1082-111.

Soodyall, H., and T. Jenkins. 1993. Mitochondrial DNA polymorphisms in Negroid populations from Namibia: new light on the origins of the Dama, Herero and Ambo. *Ann Hum Biol* 20:477-85.

Thomas, M. G., T. Parfitt, D. A. Weiss, K. Skorecki, J. F. Wilson, M. le Roux, N. Bradman, and D. B. Goldstein. 2000. Y chromosomes traveling south: the cohen modal haplotype and the origins of the Lemba--the "Black Jews of Southern Africa". *Am J Hum Genet* 66:674-86.

Underhill, P. A., G. Passarino, A. A. Lin, P. Shen, M. Mirazon Lahr, R. A. Foley, P. J. Oefner, and L. L. Cavalli-Sforza. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet* 65:43-62.

2.2. Results and Discussion

Article 2

Coelho, M., F. Sequeira, D. Luiselli, S. Beleza, and J. Rocha. 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC Evol Biol* 9:8.

BMC Evolutionary Biology

BioMed Central

Research article

Open Access

On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola

Margarida Coelho[1,2], Fernando Sequeira[3], Donata Luiselli[4], Sandra Beleza[1] and Jorge Rocha*[1,2]

Address: [1]IPATIMUP, Instituto de Patologia e Imunologia Molecular da Universidade do Porto, R Dr Roberto Frias s/n, 4200-465 Porto, Portugal, [2]Departamento de Zoologia e Antropologia, Faculdade de Ciências da Universidade do Porto, Praça Gomes Teixeira, 4099-002 Porto, Portugal, [3]CIBIO, Centro de Investigação em Biodiversidade e Recursos Genéticos, Campus Agrário de Vairão, 4485-661 Vairão, Portugal and [4]Dipartimento di Biologia Evoluzionistica Sperimentale, Universita di Bologna, Via Selmi, 3 Bologna, Italy

Email: Margarida Coelho - mcoelho@ipatimup.pt; Fernando Sequeira - fsequeira@mail.icav.up.pt; Donata Luiselli - donata.luiselli@unibo.it; Sandra Beleza - sbeleza@ipatimup.pt; Jorge Rocha* - jrocha@ipatimup.pt

* Corresponding author

Published: 21 April 2009

BMC Evolutionary Biology 2009, **9**:80 doi:10.1186/1471-2148-9-80

Received: 28 July 2008
Accepted: 21 April 2009

This article is available from: http://www.biomedcentral.com/1471-2148/9/80

Abstract

Background: Current information about the expansion of Bantu-speaking peoples is hampered by the scarcity of genetic data from well identified populations from southern Africa. Here, we fill an important gap in the analysis of the western edge of the Bantu migrations by studying for the first time the patterns of Y-chromosome, mtDNA and lactase persistence genetic variation in four representative groups living around the Namib Desert in southwestern Angola (Ovimbundu, Ganguela, Nyaneka-Nkumbi and Kuvale). We assessed the differentiation between these populations and their levels of admixture with Khoe-San groups, and examined their relationship with other sub-Saharan populations. We further combined our dataset with previously published data on Y-chromosome and mtDNA variation to explore a general isolation with migration model and infer the demographic parameters underlying current genetic diversity in Bantu populations.

Results: Correspondence analysis, lineage sharing patterns and admixture estimates indicate that the gene pool from southwestern Angola is predominantly derived from West-Central Africa. The pastoralist Herero-speaking Kuvale people were additionally characterized by relatively high frequencies of Y-chromosome (12%) and mtDNA (22%) Khoe-San lineages, as well as by the presence of the -14010C lactase persistence mutation (6%), which likely originated in non-Bantu pastoralists from East Africa. Inferred demographic parameters show that both male and female populations underwent significant size growth after the split between the western and eastern branches of Bantu expansions occurring 4000 years ago. However, males had lower population sizes and migration rates than females throughout the Bantu dispersals.

Conclusion: Genetic variation in southwestern Angola essentially results from the encounter of an offshoot of West-Central Africa with autochthonous Khoisan-speaking peoples from the south. Interactions between the Bantus and the Khoe-San likely involved cattle herders from the two groups sharing common aspects of their social organization. The presence of the -14010C mutation in southwestern Angola provides a link between the East and Southwest African pastoral scenes that might have been established indirectly, through migrations of Khoe herders across southern Africa. Differences in patterns of mtDNA and Y-chromosome intrapopulation diversity and interpopulation differentiation may be explained by contrasting demographic histories underlying the current female and male genetic variation.

BMC Evolutionary Biology 2009, **9**:80 http://www.biomedcentral.com/1471-2148/9/80

Background

Among the complex series of demographic events that shaped the patterns of human genetic variation in Africa, the massive dispersal of Bantu-speakers stands as one of the most impressive examples of human migration. Both linguistic and archeological evidences suggest that the spread of Bantu languages started about 4000 years ago in the adjacent grasslands of Cameroon-Nigeria and involved large movements of farmers carrying an agricultural tradition especially well-suited to the climate conditions prevailing in subequatorial Africa [1,2]. According to a widely accepted dispersion model, one major population movement involved the expansion of ancestors of East Bantu speakers along the northern fringe of the African rain forest into the interlacustrine areas surrounding Uganda [1-3]. Another important movement is thought to be linked to the early penetration of ancestors of West Bantu speakers into the wet coastal areas of the central African forest, beyond the Cameroon plateau [2]. More recent major expansions would include the migrations of West and East Bantu speakers into the dry territories located beyond the southern borders of the rain forests, which eventually culminated with the diffusion of Bantu languages across southern Africa [1,2]. However, this basic representation of the major trends of Bantu dispersals has not remained unchallenged [3-5], and many specific details of the migration dynamics leading to the emergence of widespread Bantu-speaking communities are still poorly understood [3].

Genetic data have great potential for unraveling the complex population history underlying Bantu expansions, but there are a number of difficulties related to sampling coverage and parameter estimation that need to be overcome. So far, most of the available genetic information has been gathered in phylogeographic studies of Y-chromosome and mitochondrial DNA (mtDNA) variation. These studies identified several mtDNA haplogroups likely to be associated with the Bantu migrations that trace their ancestries to different geographic regions of Africa [6]. In contrast, the great majority of Bantu Y-chromosome lineages were found to belong to a single widespread haplogroup (E3a), which seems to have overrun most preexisting diversity [7-9]. Recently, a few studies have begun to address more detailed aspects of regional mtDNA variation by increasing both the resolution of sequence data and the density of population sampling [10,11]. However, in spite of this progress, current understanding of Bantu expansions is still hampered by lack of sampling of crucial regions in subequatorial Africa. The area of Angola, in particular, has remained persistently underrepresented in most studies of African genetic variation [12]. Although new genetic data has become available from Kimbundu and Bakongo speakers from northern Angola and Cabinda [13,14], no information exists on the broad area encompassing the dry woodlands to the south of the Cuanza river, which is critical for understanding the push of West Savanna Bantu-speaking peoples out of the rain forest into the arid steppes of southwestern Africa.

Being exposed to the effects of the Benguela current, southern Angola provided a new environment characterized by increasing levels of aridity that challenged the progression of the agricultural lifestyle that had predominated in the well irrigated lands of the Congo basin [1,2]. Faced with this environmental shift, some groups, like the Ovimbundu, settled the high grounds of the Bié plateau where they could find areas of relatively fertile soil and higher rainfall [2]. In the coastal areas, and further to the south, settlements had to be limited to major river valleys and subsistence economies became increasingly dependent on cattle raising. The Herero, the Ovambo and the Nyaneka-Nkhumbi are examples of such Bantu groups in southwestern Angola and form a broad cultural and economic cluster relying on cattle raising to various degrees [15,16]. Among them, the Herero are the most exclusively pastoral of all Bantu peoples from southwestern Africa and penetrated well into the arid regions of the Namib Desert where they shared their mode of life with neighboring non-Bantu Khoe cattle herders [17]. This cultural and geographical proximity between Bantu and Khoisan-speaking groups poses intriguing questions about the development of the Southwest African pastoral scene and the nature of the interactions between the vanguard of West Bantu speakers and the non-Bantu peoples from the desert. For example, the role played by Khoe herders in the adoption of the present pastoral specialization of Bantu speakers is still not clear [18]. Moreover, the relative isolation of Southwest Africa from the major East African pastoral centers represents an important challenge for the identification of the migration routes that led to the emergence of a cattle-herding zone in the southwestern periphery.

Here we present an analysis of the western edge of the Bantu expansions based in the study of Y-chromosome, mtDNA and lactase persistence genetic variation in West Savanna Bantu-speaking groups from southwestern Angola. We analyzed the data in the context of regional and continental genetic diversity by assessing the differentiation between these populations and their levels of admixture with Khoisan-speaking groups, and by examining the relationship between southwestern Angola and other areas of Africa. Furthermore, we combined our dataset with published data on Y-chromosome and mtDNA variation from Southeast Africa to explore a general isolation with migration (IM) model [19,20] and infer key demographic parameters underlying the history of Bantu expansions.

BMC Evolutionary Biology 2009, **9**:80 http://www.biomedcentral.com/1471-2148/9/80

Methods

Population Samples

Buccal swabs were collected from 54 Kuvale, 153 Nyaneka-Nkhumbi, 21 Guanguela, 96 Ovimbundu, and 45 Bantu-speaking individuals with other ethnic affiliations. Individuals were grouped according to self-identified ethnicity, and only samples from unrelated individuals were included in the study. Samples were collected after informed consent in donors' hometowns and villages located in the administrative province of Namibe, southwestern Angola (Figure 1A). Besides including the original core area of Herero-speaking peoples, the province is presently inhabited by different ethnic groups due to relatively recent migrations from surrounding areas.

The groups included in our sample represent West-Savanna Bantu-speaking populations [22] from the southwestern edge of the Bantu expansions (Figure 1A) and rely on different combinations of agricultural and pastoral lifeways. The Ovimbundu form the largest ethnolinguistic cluster in Angola, making up 35% of the total population. Their original core area was located in the Bié plateau, but they underwent a series of southward expansions that considerably enlarged their territory and are presently one of the major groups inhabiting Namibe [15,16] (Figure 1A). Traditionally, most Ovimbundu groups practiced mixed farming and kept livestock. However, cattle raising was not crucial for subsistence and few families owned large herds [15]. The Nyaneka-Nkhumbi-speaking groups, who have also spread to present-day Namibe, originally settled the area located West from the middle Cunene River, including the Huíla plateau in the eastern limit of the Namibe province (Figure 1A). These peoples are agro-pastoralists that depend in part on cultivation, but keep large cattle herds and use dairying products as an important source of subsistence [16,23]. The Kuvale people dwell in the arid lowland areas of the Namibe province and are one of the most representative Herero-speaking groups from Angola [17] (Figure 1A). Like other groups included in the Herero cultural division, they are semi-nomadic cattle herders and rank among the most exclusively pastoral peoples of southwestern Africa.

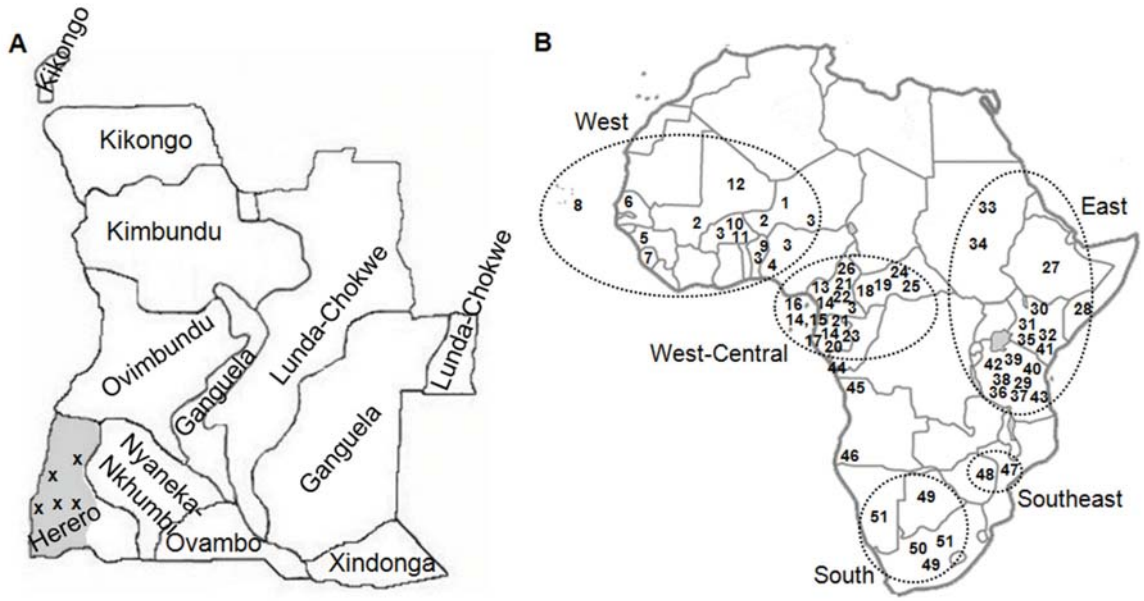


Figure 1
Major ethnolingusitic groups from Angola and population samples used in this study. A) Map of Angola depicting the core areas of the country's major ethnolinguistic groups and sampled locations in the Namibe province (modified from [21]). The area encompassing the sampled locations is shaded. B) Map of Africa with the approximate locations of the population groups used in the present analysis. Populations are coded with numbers. The correspondence between numbers and populations is given in Additional files 4 and 5. Major geographic regions are encircled.

The Ganguela-speaking peoples originally settled southeastern Angola, which is well removed from Namibe (Figure 1A). However, during the Angolan civil war many Ganguela families flew to neighboring countries and to other regions of Angola, including Namibe. The Ganguela originally included a number of scattered farming communities that were split by the southern expansion of the Chokwe peoples in the 19th century. The populations that remained in the western side of the Chokwe penetration progressively adopted cattle and became mixed farmers [16].

Laboratory methods

mtDNA

We have sequenced both hypervariable segments I (HVS-I; positions 16024–16400) and II (HVS-II; positions 73–340) of the mtDNA control region. MtDNA sequencing was performed as described previously [14]. All HVS-I and HVS-II sequences are shown in Additional file 1. To assign mtDNA sequences to previously defined haplogroups, we initially followed established criteria based on HVS-I sequence variation [6,24] updated as recently discussed [25]. Occasional ambiguities in these assignments were resolved by additional typing of a selected set of four diagnostic restriction fragment length polymorphisms (RFLPs): 3592 *Hpa*I (absent in L3), 2349 *Mbo*I (present in L3e), 10084 *Taq*I (present in L3b) and 8616 *Mbo*I (absent in L3d). After this initial assignment step, we used the available information on HVS-I and HVS-II variation provided by published complete mtDNA sequences to refine and/or rename the classifications according to the most recently updated mtDNA phylogeny [26] (see Additional file 1). For the sake of comparison we refer to the HVS-I-based nomenclature throughout the article.

Y-chromosome

To characterize the nonrecombining portion of the Y-chromosome (NRY) we genotyped 9 unique event polymorphisms (UEPs; M2, M35, M60, M91, M112, M150, M213, YAP, SRY4064) and 11 short tandem repeats (STRs; DYS19, DYS389I, DYS389II, DYS385, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439). The DYS385 locus consists of a duplicated tetranucleotide STR region and was omitted from some analyses. Except for YAP, UEPs were typed by direct sequencing of PCR products. Primer sequences and protocols are provided upon request. Short tandem repeats were typed with the Promega Powerplex Y System. All Y-chromosome combined haplotypes, defined by UEP and STRs, are shown in Additional file 2. For the sake of comparison NRY haplogroups based on the UEP variation were named according to the Y-chromosome Consortium guidelines [27,28], but we also provide haplogroup names according to a most recent update [29], which in our dataset essentially involves the renaming of haplogroup E3a as E1b1a (see Additional file 2).

Lactase persistence

Lactase persistence was screened by direct sequencing of a 359 bp PCR fragment located within intron 13 of the MCM6 gene, which contains all single nucleotide polymorphisms (SNPs) that have been so far associated with lactase persistence in human populations: G/C -14010; T/G -13915; C/T -13910; and C/G -13907 [30-32] (see Additional file 3 for typing details). In addition to the southwestern Angolan sample, we further typed the lactase persistence-associated SNPs in a total of 111 Bantu speaking individuals belonging to 11 different population groups from Mozambique: 3 Chopi, 4 Chwabo, 19 Makhwa, 15 Makonde, 15 Ndau, 11 Nyanja, 15 Ronga, 2 Sena, 15 Shangaan, 1 Shona and 11 Tswa.

Data analyses

Summary statistics for mtDNA and Y-chromosome haplotype variation, and Tajima's *D* and Fu's *Fs* tests were calculated and performed with the ARLEQUIN 3.11 software package [33]. Analyses of molecular variance (AMOVA) to evaluate the apportionment of genetic variation were also performed using ARLEQUIN 3.11. Correspondence analysis based on mtDNA and Y-chromosome haplogroup frequencies was performed using the POPSTR program [34].

Population cross-comparisons for the mtDNA data were restricted to the 16090–16365 HVS-I sequence range and were based in an assembled dataset comprising approximately 5400 mtDNA profiles from 73 populations (Figure 1B and Additional file 4). For the NRY data, cross-population comparisons were based either on UEP-defined haplogroups or on higher resolution haplotypes defined by a subset of 7 STR loci (DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393) common to all samples assembled in a database of about 5072 haplotypes from 72 populations (see Additional file 5). Networks of NRY haplotypes and mtDNA sequences were constructed using the NETWORK 4.5 software [35]. For NRY haplotypes, the reduced-median [36] and median-joining [37] algorithms were applied sequentially and differential microsatellite weighting was used to resolve extensive reticulation at microsatellite loci. Weights for each microsatellite were inversely proportional to the ratio of the variance displayed by each marker within the respective haplogroups and the average variance value across loci in those haplogroups. For mtDNA sequences, the median-joining algorithm [37] was used without further weighting. Ages of mtDNA and NRY lineages were estimated with the ρ (rho) statistic [38] using NETWORK 4.5, assuming 25 years per generation, a mtDNA control region mutation rate of $\mu = 7.55 \times 10^{-6}$ per nucleotide per generation, based in a recent Bayesian estimate [39], and

BMC Evolutionary Biology 2009, **9**:80 http://www.biomedcentral.com/1471-2148/9/80

the following NRY-STR per generation mutation rates [40]: $\mu_{DYS19} = 0.0017$; $\mu_{DYS389I} = 0.0019$; $\mu_{DYS389II} = 0.0023$; $\mu_{DYS390} = 0.0023$; $\mu_{DYS391} = 0.0035$; $\mu_{DYS392} = 0.0006$; $\mu_{DYS393} = 0.0007$.

Admixture proportions were estimated with the ADMIX2.0 program [41]. MtDNA-based estimates were calculated from haplogroup frequencies without taking into account molecular distances between haplogroups. NRY-based estimates were calculated from the frequency of haplotypes defined by STR loci DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, and DYS393, not taking into account molecular distances between haplotypes.

We have also attempted to infer the key demographic parameters of Bantu expansions by analyzing our NRY and mtDNA data from southern Angola together with additional data from Southeast Africa (see Additional files 4 and 5) within the framework of a general isolation with migration (IM) model [19,20], using the IMa program [42]. The IM model describes the historical demographic properties of two related populations that may have varied in size after diverging from a single ancestral population, with bidirectional migration occurring at constant rate following the initial split [19,20]. Thus, we reasoned that this framework could be applied to populations located in the two opposite edges of the Bantu expansions in order to analyze the split between the eastern and western streams of Bantu dispersals after a common origin in an area likely to be located in West-Central Africa (see Additional file 6). The model has six parameters whose posterior probability distributions can be estimated by using the Markov chain Monte Carlo (MCMC) approach implemented in the IMa computer program [42]: effective population sizes for both extant (N_1 and N_2) and ancestral (N_A) populations, time since divergence (t) and migration rates in both directions (m_1 and m_2). These demographic terms are obtained by conversion from estimated basic parameters that are scaled by the mutation rate (per locus per generation): $\theta_A = 4N_A\mu$; $\theta_1 = 4N_1\mu$; $\theta_2 = 4N_2\mu$; $t = t\mu$; $m_1 = m1/\mu$; $m_2 = m2/\mu$. The mtDNA dataset consisted of 724 HVS-I 276 bp-long sequences ranging from positions 16090 to 16365, including 358 sequences from southwestern Angola, collected in the present study, and 366 assorted sequences from different ethnic groups from Mozambique and Zimbabwe (see Additional file 4). The NRY data consisted of 348 haplotypes defined by 7 STR loci totaling 236 Y chromosomes from the present Angolan sample and 112 chromosomes from Mozambique (see Additional file 5). Parameter conversions were done by using the aforementioned mtDNA control region and NRY STR mutation rates. After preliminary runs to determine plausible uniform prior ranges, the IMa program was run for at least 10 million steps after 100000 steps of burn-in with 8 Metropolis-coupled chains, with geometric heating. For each dataset at least two independent replicates were performed using the same running options and a different random seed to assess convergence of the parameter estimates. MtDNA sequences were assumed to mutate under the Hasegawa-Kishino-Yano (HKY) finite sites mutation model [43]. Mutation in NRY STRs was modeled by the stepwise mutation model (SMM). The mode of each marginal posterior distribution generated by the program was considered a point estimate of the corresponding parameter value. Reported parameter estimates are means from replicate runs.

Results

mtDNA

Haplotype and sequence diversity

Summary statistics for mtDNA HVS-I sequence diversity are presented in Table 1. The estimated levels of sequence variation in the whole sample from Namibe ($\theta_k = 85$; $\pi = 0.025$; H = 0.986) are within the range found in other Sub-Saharan regions [24]. However, diversity indices sug-

Table 1: MtDNA HVS-I sequence diversity in populations from southwestern Angola

Population	N	k (k/N)	H (SD)	θk (95% CI)	θs (SD)	π (SD)	Tajima's D (P)	Fu's Fs (P)
Kuvale	54	23 (0.42)	0.937 (0.017)	14.61 (8.43–25.05)	11.19 (3.37)	0.026 (0.013)	-0.45 (0.37)	-2.04 (0.29)
Ganguela	21	16 (0.76)	0.962 (0.030)	28.92 (12.08–73.02)	10.28 (3.76)	0.025 (0.013)	-0.37 (0.38)	-3.76 (0.06)
Nyaneka-Nkhumbi	153	73 (0.48)	0.982 (0.003)	54.11 (38.93–75.04)	12.67 (3.18)	0.024 (0.012)	-0.93 (0.16)	-24.44 (0.00)
Ovimbundu	92	61 (0.66)	0.987 (0.004)	77.83 (50.90–120.09)	13.74 (3.70)	0.024 (0.012)	-1.10 (0.13)	-24.61 (0.00)
Total	365	142 (0.39)	0.986 (0.002)	84.92 (67.85–106.00)	15.29 (3.33)	0.025 (0.013)	-1.12 (0.10)	-24.01 (0.00)

N, number of sequences; k, number of different haplotypes; H, haplotype diversity; θ, mutation drift statistic calculated from the number of different haplotypes (θk) and number of segregating sites (θs); π, nucleotide diversity. The total sample includes 45 additional sequences from other groups.

gest that subsistence economy and population history had a measurable impact in the demography of the peoples from southern Angola. The semi-nomadic Herero-speaking Kuvale pastoralists display the lowest value of θ_k (= 14.61), and fail to reveal any significant signal of population expansion, as reflected by the nonsignificance of both Tajima's D and Fu's Fs tests (Table 1). By contrast, the Ovimbundu and the Nyaneka-Nkhumbi agro-pastoralist groups exhibit higher levels of haplotype diversity and have significantly negative Fs values that are consistent with a history of demographic expansion. In order to confirm that the differences in Fs values were not influenced by sample size, we recalculated the Fs values in random subsamples of the Nyaneka-Nkhumbi and the Ovimbundu that matched the lower sample sizes of the Ganguela (N = 21) and the Kuvale (N = 54). For N = 21, Fs values remained significant for the Ovimbundu (Fs = -6.57; P = 0.01) but became nonsignificant for the Nyaneka-Nkhumbi (Fs = -3.97; P = 0.06). With N = 54, Fs values remained significant both for the Ovimbundu (Fs = -24.42; P = 0.00) and for Nyaneka-Nkhumbi (Fs = -11.40; P = 0.008), suggesting that the absence of traces of population expansion in the Kuvale is not attributable their lower sample size.

Haplogroup composition

To investigate the relationship between Namibe and other African populations, a correspondence analysis based on mtDNA haplogroup frequencies was performed using contextual samples from different Sub-Saharan regions (Figure 1B; Additional file 4). The southern African Khoe-San samples distort the correspondence analysis (data not shown) due to their genetic uniqueness (high frequencies of haplogroups L0d/k) and were excluded from further analysis in order to achieve better resolution of the genetic relationships among populations.

Figure 2A shows the results of the correspondence analysis after exclusion of the Khoe-San samples. To facilitate graphical display, populations from Southeast Africa, which are known to be genetically homogeneous [6], were pooled into a single group. Samples from the northern Angolan regions of Cabinda and Luanda (here designated as Angola) were also combined. The pooled sample from

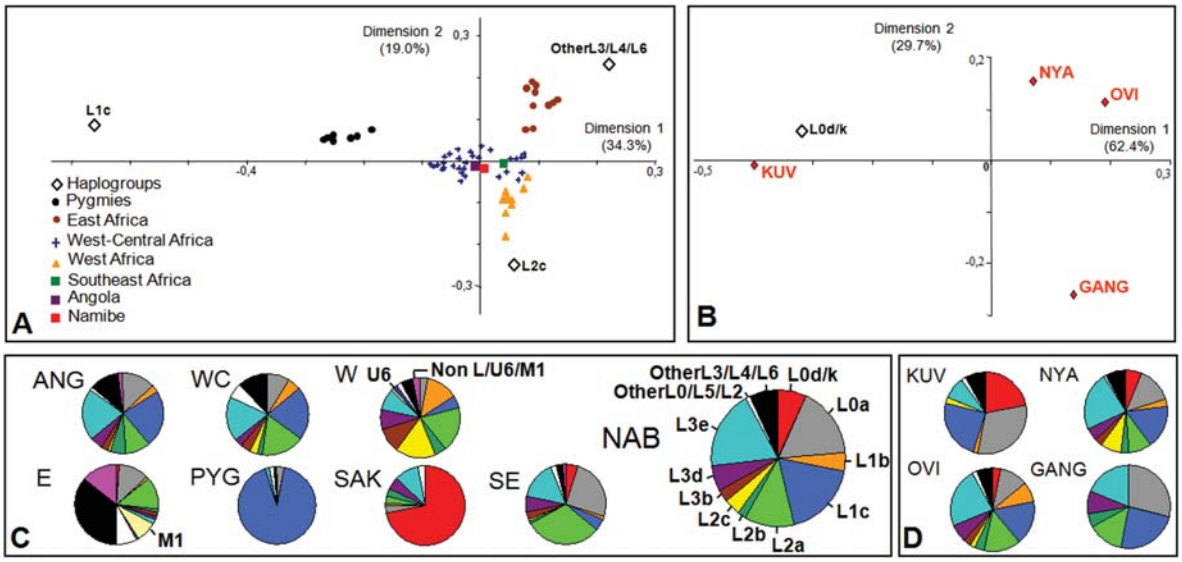


Figure 2

MtDNA haplogroup variation in southwestern Angola and other African populations. A) and B) Correspondence analysis plots based on haplogroup frequency profiles from several African populations (A) and different ethnolinguistic groups from Namibe (B). Percentages in parentheses indicate the total fraction of the genetic variation that was captured by each dimension. Geographic regions were defined as in Figure 1B. Populations from Mozambique and Zimbabwe were pooled into a single Southeast Africa group. Angola includes samples from Cabinda and Luanda. Namibe includes all groups sampled in this study. C) MtDNA haplogroup frequencies in the Namibe province and in other African population groups. D) MtDNA haplogroup frequencies in the four population groups sampled in the Namibe province. Haplogroup frequencies were broken down as in [25]. ANG = Angola (Luanda + Cabinda); WC = West-Central Africa; W = West Africa; E = East Africa; PYG = Pygmies; SAK = South Africa Khoe-San; SE = Southeast Africa; NAB = Namibe; KUV = Kuvale; OVI = Ovimbundu; NYA = Nyaneka-Nkhumbi; GANG = Ganguela.

BMC Evolutionary Biology 2009, **9**:80 http://www.biomedcentral.com/1471-2148/9/80

Namibe lies in a cluster including most West-Central African populations, Angola and Southeast Africa. Pygmies display unusually high frequencies of haplogoup L1c and are clear genetic outliers. Despite some genetic continuity between West-Central and West Africa, the geographic regions of West-Central, West and East Africa seem to be well correlated with mtDNA variation, as populations from different regions tend to cluster around different coordinates. Furthermore, when populations were classified by four major geographic regions (West, East, Southeast and West-Central Africa) and two additional outlier ethnic groups (Pygmies and southern African Khoisan-speakers), AMOVA analysis showed that 14.2% of the variability lies between groups, 4.8% among populations within groups and 81% within populations (all P values <0.01).

Most mtDNA haplotypes that are commonly found in Sub-Saharan populations were also observed in Namibe (see Additional file 1; Figure 2C). The most frequent (>5%) haplogroups were L0d (6%), L0a1 (9%), L0a2 (8%), L1c1 (8%), L1c2 (7%), L2a1 (10%), L3e1 (9%), L3e2 (7%) and L3f (6%). Haplogroup L1c1a, which is typical of Pygmy populations from Central Africa [10,11], was virtually absent from our sample (see Additional files 1; Figure 2C). The relatively high frequency of the typical Khoe-San L0d haplogroup contrasts with previous findings from northern Angola [13,14] but compares with observations in Bantu groups from Southeast Africa [6,44]. However, this haplogroup is not evenly distributed in the Namibe samples and reaches much higher frequencies in the Kuvale (22%; Figure 2D) than in other groups. When correspondence analysis is focused on the four populations from Namibe (Figure 2B) the genetic peculiarity of the Kuvale caused by the high frequency of L0d becomes obvious.

Patterns of lineage sharing

In order to analyze the likely origin of mtDNA sequences from southwestern Angola, we used the comparative mtDNA African dataset (see Figure 1B and Additional file 4) to study the patterns of HVS-I lineage sharing between Namibe and other Sub-Saharan populations. Although restriction of population cross-comparisons to the HVS-I control region increases the number of available samples, it is important to note that some matches may involve sequences that are phylogenetically unrelated. However, these cases are expected to seriously bias the conclusions only if convergence episodes are non-randomly distributed across lineages.

We evaluated the patterns of haplotype sharing using matching scores based both on individual sequences and on haplotypes. Matching scores based on individual sequences count the number of individuals from Namibe with at least one match in a given African region. Matching scores based on haplotypes, count the number of different haplotypes from Namibe with at least one match in a given African region. We found that 76% of all individual sequences and 53% of the different haplotypes sampled in Namibe match at least one sequence from elsewhere. Figure 3 shows the distribution of shared lineages across different Sub-Saharan populations. As much as 96% of all individual mtDNAs from Namibe that were found to be shared with other African populations had matches with West-Central Africa, or with regions up north in Angola (Cabinda and Luanda), which lie close to plausible migration paths between West-Central and Southwest Africa (Figure 3A). Lineage sharing with either West or East Africa is significantly lower and represents 37% and 60%, respectively, of all shared sequences from Namibe. Although lineage sharing between Namibe and the southeastern Africa is high (Figure 3), ~97% of the observed matches were found to be also shared with West-Central Africa. The link between southwestern Angola and West-Central Africa is not restricted to haplogroups that are thought to have originated in this region, like L1c or L3e [6,24]. Even shared sequences belonging to haplogroups that may trace their phylogenetic origin back to regions outside West-Central Africa were found to match sequences that occur around this area (see Additional file 7). Consideration of shared haplotypes instead of individual sequences replicates these overall sharing trends (Figure 3B).

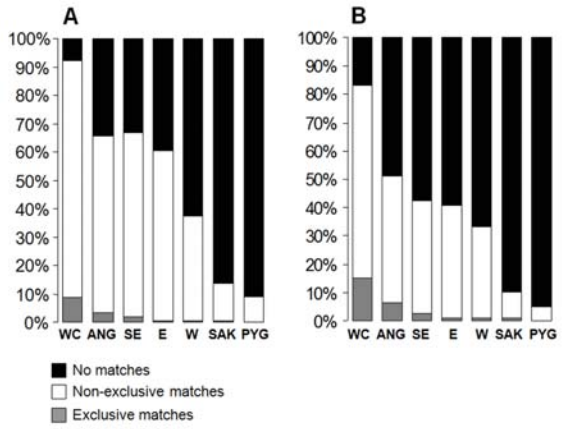


Figure 3
Patterns of mtDNA lineage sharing. Lineage sharing between individual mtDNAs (A) and haplotypes (B) from southwestern Angola and from other population groups in Africa. Only mtDNAs and haplotypes found to be shared between Namibe and at least one other African population were included in the calculations. Abbreviations are the same as Figure 2.

Lineage sharing with Pygmies and southern African Khoisan-speaking peoples is low (Figure 3). Even the sequences that belong to the typical Khoe-San L0d haplogroup did not match any Khoisan-speaking population from the database. However, network analysis clearly shows that the Angolan L0d lineages are phylogenetically related to other typical Khoe-San sequences from southern Africa (see Additional file 8). We have attempted to calculate the age of the unmatched L0d lineages by estimating the average number of mutational changes to their closest southern African ancestor, using the ρ statistic (data not shown). Estimated ages were found to vary between 4816 (± 4816) and 17308 (± 9667) years.

Admixture analysis

Although patterns of lineage sharing suggest that a substantial fraction of the mtDNA pool from southwestern Angola may have derived from West-Central Africa, important contributions from other regions cannot be firmly ruled out due to the relatively low proportion of lineages that are shared exclusively with each of several potential source areas (Figure 3). To complement the study of matching patterns, we used an explicit model of admixture [41] in which the southwestern Angolan population was considered to be a hybrid containing variable contributions from five different parental regions and populations (Table 2): West Africa, West-Central Africa, East Africa, Pygmies and the southern Africa Khoe-San (see Additional file 4). This model oversimplifies the complex demographic scenarios underlying the Bantu migrations by assuming that the Namibe pool was instantaneously created by combining different proportions of parental populations. However, the location of Namibe in a border of the Bantu expansion range, where different populations could have merged, seems to justify the use of this admixture model as an exploratory tool to assess the relative contribution of different putative source regions.

Despite being associated with high standard deviations, estimates of the admixture proportions are consistent with a major (0.74 ± 0.15) contribution of West-Central Africa to the southwestern Angola mtDNA pool (Table 2). Contributions from West Africa (0.04 ± 0.07), East Africa (0.04 ± 0.06) and the Pygmies (0.05 ± 0.05) seem to have been residual and significantly lower than that from southern African Khoisan-speaking peoples (0.13 ± 0.02). Moreover, it is interesting to note that the calculated contribution from the Khoe-San in each population group shows a stepwise increase that appears to be correlated with the degree of dependence on animal husbandry of the different groups: Ganguela (0.03 ± 0.02) < Ovimbundu (0.07 ± 0.03) < Nyaneka-Nkhumbi (0.12 ± 0.03) < Kuvale (0.33 ± 0.08). The relatively low standard deviations associated with these estimates reflects the high levels of differentiation of the Khoe-San, showing that the use of an admixture model is more adequate when parental populations have remained isolated for a long time.

Y-chromosome

Haplotype diversity

In accordance with the trend observed for mtDNA (Table 1), NRY diversity in STR-defined haplotypes loci was found to be lower among the Kuvale (θ_k = 24.6) than in the Nyaneka-Humbi (θ_k = 60.3) and the Ovimbundu (θ_k = 61.1), revealing a consistent pattern of population size reduction and genetic drift in the Kuvale group.

Haplogroup composition

Figure 4A displays the results of a correspondence analysis based on SNP-defined NRY lineages, summarizing the genetic relationships between Namibe and other sub-Saharan populations. Populations where the majority of NRY lineages belong to haplogroup E3a-M2 -including Namibe, northern Angola (Cabinda), West-Central Africa Bantu groups, West Africa and the Pygmies- are clustered together (Figures 4A and 4C). West-Central African populations located in the upper-right quadrant of the correspondence analysis plot are non-Bantu populations from Northern Cameroon with high frequencies of C, FR haplogroups. South African and East African Khoisan-speak-

Table 2: Estimated admixture proportions of mtDNA lineages from southwestern Angola

	Parental Population				
Hybrid population	West-Central Africa	West Africa	East Africa	Pygmies	South Africa Khoe-San
Kuvale	0.52 (0.16)		0.16 (0.15)		0.33 (0.08)
Ganguela	0.77 (0.49)	0.05 (0.22)	0.05 (0.21)	0.10 (0.18)	0.03 (0.02)
Nyaneka-Nkhumbi	0.81 (0.09)	0.07 (0.09)			0.12 (0.09)
Ovimbundu	0.87 (0.12)	0.06 (0.11)			0.07 (0.03)
Total	0.74 (0.15)	0.04 (0.07)	0.04 (0.06)	0.05 (0.05)	0.13 (0.02)

Estimates and their standard deviations (shown between parentheses) were computed from 1000 bootstrap replications. Empty cells correspond to unsupported parental populations that were not used in admixture calculations.

BMC Evolutionary Biology 2009, 9:80 http://www.biomedcentral.com/1471-2148/9/80

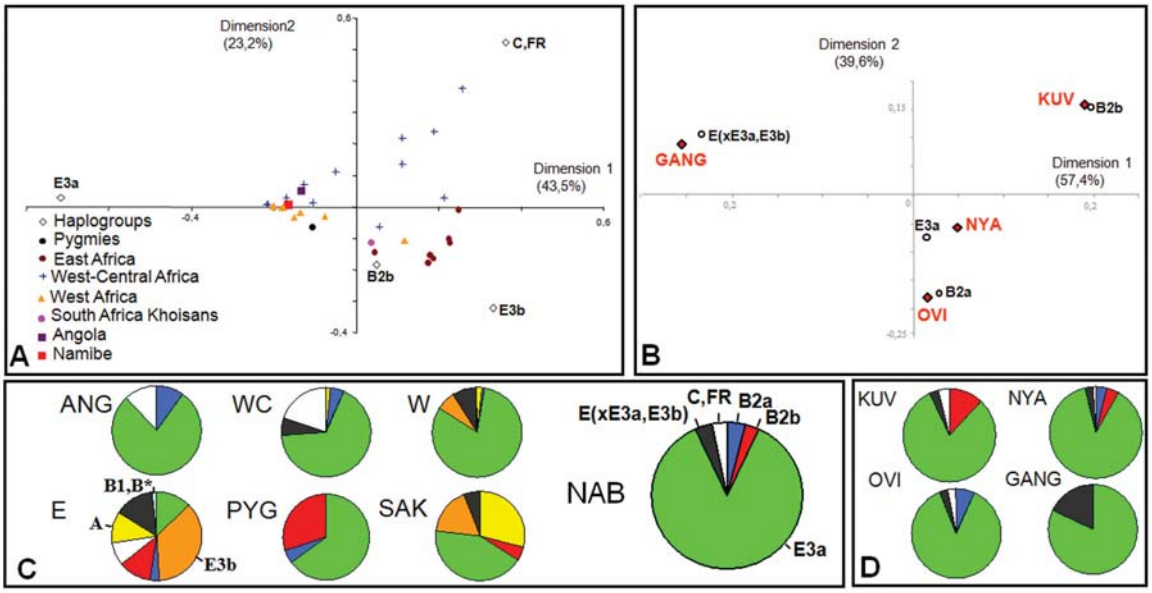


Figure 4

Y-chromosome haplogroup variation in southwestern Angola and other African populations. A) and B) Correspondence analysis plots based on haplogroup frequency profiles from several African populations (A) and different ethnolinguistic groups from Namibe (B). Percentages in parentheses indicate the total fraction of the genetic variation that was captured by each dimension. Geographic regions were defined as in Figure 1B. Angola refers to a sample from Cabinda. Namibe includes all groups sampled in this study. C) Y-chromosome haplogroup frequencies in the Namibe province and in other African population groups. D) Y-chromosome haplogroup frequencies in the four population groups sampled in the Namibe province. Abbreviations are the same as Figure 2.

ing groups lie close to each other in the lower-right quadrant, which groups populations where B2b-M112 and E3b-M35 haplogroups are common, including all samples from East Africa (Figures 4A and 4C). A single West African pooled sample from Mali (coded as 12 in Figure 1B; see Additional file 5) is also located in the lower-right quadrant due to a high frequency of E3b-M35. However, due to paraphyly of the E3b-M35 clade [45], the sharing of E3b-M35 may not indicate a close genetic relationship between East Africa populations and the West Africa sample. When populations were classified into the same major groups defined for mtDNA (West Africa, East Africa, Southeast Africa, West-Central Africa, Pygmies and southern African Khoisan-speakers), AMOVA analysis showed that 18.2% of the variability lies between groups, 18.8% among populations within groups and 62.3% within populations (all P values <0.01). This apportionment reflects the high levels of genetic heterogeneity observed in populations from the same broad geographic area.

Although the majority (~80%) of Y-chromosome lineages in southwestern Angola belong to haplogroup E3a-M2 (Figure 4C; see Additional file 2) the distribution of the remaining lineages is not uniform across the Namibe samples (Figure 4D). The small sample from the Ganguela has E(xE3a, E3b) lineages that are less common in the other groups, while the Kuvale and the the Nyaneka-Nkhumbi carry the B2b-M112 haplogroup, which is known to be frequent both in Pygmies and Khoisan-speakers [8,9,46]. Figure 4B emphasizes the influence of the E(xE3a, E3b) and B2b minor haplogroups in separating the Ganguela and Kuvale from the Nyaneka-Nkhumbi and the Ovimbundu. This local pattern is remarkably congruent with that obtained with mtDNA (Figure 2B).

Patterns of lineage sharing

To study the patterns of lineage sharing in the Y-chromosome, we used the populations from the comparative NRY African dataset with reported haplotype data for a common set of seven STR loci (see Additional file 5). Due to the high level of convergent evolution among NRY haplotypes based on this limited subset of STRs, the possibility of phylogenetically unrelated matches cannot be completely ruled out, as for mtDNA.

Approximately 75% of the total number of Y chromosomes and 52% of different haplotypes from southwestern Angola had at least one match with another African population. As observed for mtDNA, most Y-chromosome matches involved West-Central Africa (92% of shared Y chromosomes; 81% of shared haplotypes; Figure 5). Levels of Y-chromosome sharing with Southeast Africa were also high (72% of Y chromosomes; 42% of haplotypes), but ~97% of these individual matches (85% of different haplotypes) were shared with West-Central Africa.

Haplotypes within haplogroup B2b remained unmatched, but phylogenetic relationships inferred by network analysis suggest that these haplotypes are more likely to have been derived from Khoe-San populations than from Pygmies (see Additional file 9). Estimated ages of the unmatched B2b lineages based in the average number of mutational changes to the closest southern African ancestor were found to vary between 19445 (± 13749) and 29168 (± 20624) years.

Admixture analysis

As for mtDNA, we used an explicit admixture model to infer the relative contributions of different African regions to the sampled southwestern Angola Y-chromosome pool. However, admixture calculations based on the frequencies of SNP-defined haplogroups lead to estimates that were associated with high standard deviations and often exceeded the 0–100% range under different combinations of parental populations. As these implausible results were likely to be due to the high similarity of haplogroup frequency profiles of West and West-Central Africa, both dominated by the E3a-M2 haplogroup, we performed a higher resolution analysis using the frequencies of haplotypes defined by a common set of 7 STR loci (see Additional file 5). At this level of resolution, the E3a-M2 haplotype subset defined by alleles 15-21-10-11-13 at loci DYS19, DYS390, DYS391, DYS392 and DYS393, which has been considered a founder lineage of Bantu expansions [7], has very different frequencies in West-Central (~0.24) and West Africa (~0.08).

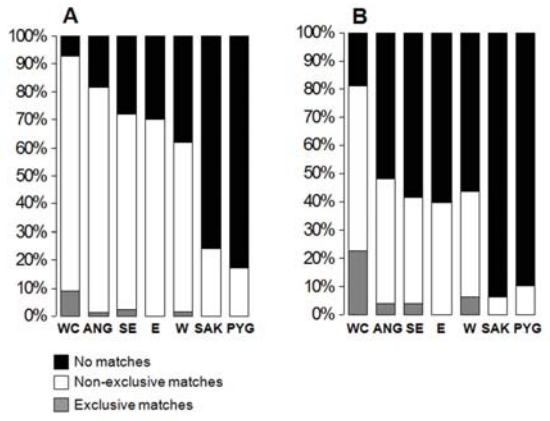


Figure 5
Patterns of Y-chromosome lineage sharing. Lineage sharing between individual Y chromosomes A) and haplotypes B) from southwestern Angola and from other population groups in Africa. Only Y chromosomes or haplotypes that were found to be shared between southwestern Angola and at least one other African population were included in the calculations. Haplotypes were defined by STR loci DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393. Abbreviations are the same as Figure 2.

To perform the STR-based admixture reanalysis, we excluded all haplotypes that did not reach a minimal 0.02 frequency threshold in at least one source region or in the whole Namibe sample. West-Central Africa and West Africa were the only supported source regions in most calculations (Table 3), with West-Central Africa providing the major admixture contribution (0.88 ± 0.19), as observed for mtDNA. Only the Nyaneka-Nkhumbi showed a signal for a possible Khoe-San contribution (0.12 ± 0.17), but we note the large standard deviation associated with this calculation.

Lactase persistence

Given the well known association between lactose tolerance and pastoralism [47], we reasoned that the study of lactase persistence mutations in Namibe might be informative for exploring historical links between southwestern Bantu cattle herders and other pastoral communities elsewhere in Africa. To this end, we screened our sample for all SNPs that are currently known to be associated with lactase persistence in human populations. We found that the -14010C allele, which is most frequent in Nilo-Saharan and Afro-Asiatic populations from Kenya and Tanzania (32–42%, [32]), was present at lower frequencies in the Ovimbundu (1%), the Nyaneka-Nkhumbi (3%) and the Kuvale (6%). By contrast, we could not find any lactase persistence-associated allele in an additional sample of 111 individuals from several ethnolinguistic groups in Mozambique.

Estimation of demographic parameters

Table 4 displays the estimated terms for the basic demographic parameters of the IM model using NRY and mtDNA datasets from Southwest Angola and Southeast Africa, assumed to represent the endpoints of the western and eastern branches of the Bantu migrations, respectively (see Additional file 6). Independent runs based on the NRY dataset converged on the approximate marginal posterior probability distributions for all parameters of the

BMC Evolutionary Biology 2009, **9**:80 http://www.biomedcentral.com/1471-2148/9/80

Table 3: Estimated admixture proportions of Y-chromosome lineages from southwestern Angola

	Parental Population		
Hybrid population	West-Central Africa	West Africa	South Africa Khoe-San
Kuvale	0.96 (0.39)	0.04 (0.39)	
Nyaneka-Nkhumbi	0.68 (0.29)	0.20 (0.28)	0.12 (0.17)
Ovimbundu	0.99 (0.30)	0.01 (0.30)	
Total	0.88 (0.19)	0.12 (0.19)	

Estimates and their standard deviations (shown between parentheses) were computed from 1000 bootstrap replications. East Africans, and Pygmies were always unsupported as parental populations and are not shown. The Ganguela were omitted due to low sample size. Empty cells correspond to unsupported parental populations that were not used in admixture calculations.

model (see Additional file 10). However, no convergence was found for the migration rate parameter from Southwest to Southeast Africa (m_2, $2N_2m_2$) under reasonable computation times using the mtDNA dataset (see Additional file 11). Therefore, we chose to present separately the outcome of individual runs with different probability density peaks for this parameter (Table 4).

Estimated current population sizes (N_1 and N_2) are 4 to 8 fold higher than ancestral population sizes (N_A) showing that the dispersal of Bantu speaking groups involved both significant size growth and geographic expansions. Population growth could actually have been even more marked, if bottlenecks occurring at the formation of daughter populations caused initial size reductions [48]. There are important differences in the estimates based in the mtDNA and NRY datasets. The Y-chromosome-based estimate for the current Southeast Africa population size is about one half that from the Southwest (N_1~10000, N_2~5000; Table 4), while current population sizes estimated from the mtDNA data are similar in both edge populations (N_1 and N_2 ~30000; Table 4). Current population sizes based on Y-chromosome estimates lie on the lower range of reported African-specific population size, while current population size estimates from the mtDNA are in the upper range of reported values from African populations [49]. Ancestral population sizes inferred from the Y-chromosome (N_A~1200) and mtDNA (N_A~7100) are also different.

With regard to migration, all Y-chromosome runs consistently yielded values close to zero (see Additional file 10), corresponding to the first bin of the surveyed parameter space (Table 4). In contrast, population migration rates inferred from mtDNA are high ($2N_1m_1$ and $2N_2m_2 > 15$; Table 4), pointing to extensive female-mediated gene flow

Table 4: Estimates of demographic parameters in the Southwest and Southeast edges of the Bantu expansions

	N1	N2	NA	m1	m2	2N1m1	2N2m2	t (years)
Y chrom								
	10020	5510	1195	0	0	0	0	1950
	(6684–21557)	(3313–12057)	(647–2372)	(0–4 × 10^{-3})	(0–6 × 10^{-3})	(0–80.2)	(0–66.1)	(1388–2940)
mtDNA								
A	33212	31885	7090	2.6 × 10^{-4}	5.5 × 10^{-4}	17	35	25410
	(22680–44905)	(23592–46315)	(3441–16129)	(9.0 × 10^{-5}-4.5 × 10^{-3})	(2.8 × 10^{-4}-4.7 × 10^{-3})	(6–297)	(18–298)	(14612–39135)
B	35700	30558	7173	2.4 × 10^{-4}	3.0 × 10^{-3}	17	184	24978
	(24256–46066)	(23055–46315)	(3607–17041)	(5.0 × 10^{-5}-3.9 × 10^{-3})	(4.5 × 10^{-4}-4.8 × 10^{-3})	(4–275)	(21–296)	(13749–38319)
Joint								
A	6894	6150	1394	2 × 10^{-3}	7 × 10^{-3}	28	86	4133
	(4631–9571)	(4403–8619)	(343–2246)	(4.7 × 10^{-4}-1.6 × 10^{-2})	(1.4 × 10^{-3}-1.6 × 10^{-2})	(6–215)	(17–191)	(3071–13384)
B	7558	6274	1372	1 × 10^{-3}	9 × 10^{-3}	15	113	3981
	(5417–9929)	(4884–8578)	(933–2433)	(4.5 × 10^{-4}-1.1 × 10^{-2})	(4.5 × 10^{-4}-1.5 × 10^{-2})	(7–211)	(6–185)	(3147–6332)

N_1-Current effective population size in the Southwest edge; N_2-Current population size in the Southeast edge; N_A-Ancestral effective population size; m1-Probability of migration from Southeast to Southwest Africa, per gene copy per generation; m_2-Probability of migration from Southwest to Southeast Africa, per gene copy per generation; $2N_1m_1$-Effective number of genes migrating into Southwest Africa, per generation; $2N_2m_2$-Effective number of genes migrating into Southeast Africa, per generation; t-time since divergence from a common ancestor. 95% credibility intervals are given in parentheses. A and B show the outcome of runs with different probability density peaks for m_2.

between the west and east branches of Bantu expansions. In all cases, $2N_2m_2$ estimates (migration from Southwest into Southeast) were found to be consistently higher than $2N_1m_1$ (migration from Southeast into Southwest), but this observation must be regarded with caution, since $2N_2m_2$ estimates from independent runs failed to converge to a single maximum (see Additional file 11). Although the credibility intervals obtained in different runs were quite similar, migration rate distributions were typically two-peaked. While a fraction of the runs yielded a major peak for lower $2N_2m_2$ values (~35), in other cases the pattern reversed and the major peak corresponded to unusually high 2_2Nm_2values (~180) (Table 4; Additional file 11).

In order to overcome the limitations of single locus estimates, we have also generated inferences based on the combined mtDNA and NRY datasets (Table 4; Additional file 12). Joint estimates of population sizes support a 5-fold growth after population splitting ($N_1$and N_2 ~7000; N_A~1300; Table 4). Divergence time estimates were remarkably consistent with the archeological data (t = 4000 years; Table 4), while migration rates from the western to the eastern branch remained difficult to resolve, reflecting the uncertainty associated with mtDNA dataset (Table 4; Additional file 10).

Discussion

Southwest Angola in the African context

It is generally accepted that the mtDNA pool of Bantu speaking populations comprises a diverse set of lineages that trace their phylogeographical ancestry into three major sub-continental regions: West Africa, East Africa and West-Central Africa [6,11,24]. In spite of the growing knowledge about the ultimate regional sources of Bantu mtDNA lineages, the understanding of the major demographic processes that led to the assemblage and distribution of these diverse regional contributions among the different areas of the Bantu-speaking universe is still far from being complete. Our analysis shows that haplogroups currently associated with the Bantu mtDNA pool from southwest Angola reflect the combination of different regional contributions generally observed in most Bantu-speaking populations [6,11,24]. However, both the patterns of lineage sharing and admixture estimates from different potential source populations strongly suggest that the bulk (~75%) of mtDNA variation in southwestern Angola can be traced back just to West-Central Africa, in areas that are adjacent to the original heartland of Bantu expansions [2]. The only additional region with a significant (~13%) genetic contribution to Southwest Angola was southern Africa, indicating that most extant mtDNA variation from southwestern Angola may have simply resulted from the encounter of an offshoot of West-Central Africa with autochthonous Khoisan-speaking peoples from the south.

It is, therefore, likely that the occupation of Southwest Africa has been preceded by a period of assemblage of diverse mtDNA contributions up north, in West-Central Africa, followed by subsequent migrations from specific dispersal centers into the southwest. According to linguistic and archeological evidences, a likely dispersal center to Angola would have been located in savanna areas just south of the equatorial forest, around the Tshikapa site, where premetallurgical Bantu speakers originating in Cameroon/Gabon might have acquired iron technology and livestock from eastern Bantu peoples, before proceeding to the southwest [21,50]. The location of this center on the southern savanna edge of West-Central Africa would explain the lack of Pygmy L1c1a lineages in Angola, in contrast with the areas closer to Cameroon and Gabon where gene flow from Pygmies was more important [11].

In contrast with the collection of diverse haplogroups that is generally found in the maternal pool, the NRY haplogroup composition is highly homogeneous in most potential source areas of Bantu dispersions, due to the predominance of haplogroup E3a-M2 in West and West-Central Africa [8,9,51]. However, STR-defined haplotypes yielded sufficient resolution to allow discrimination between Y-chromosome contributions from West and West-Central Africa and reveal a link between southwestern Angola and West-Central Africa that is remarkably congruent with the results from the mtDNA dataset.

Southwest Angola in the local and regional contexts

Despite the substantial differences between their levels of haplogroup variation, both NRY and mtDNA data concurred in showing that the populations sampled in Namibe are clustered together with other Bantu groups from elsewhere. Within the local context of southwestern Angola, the divergence of the Herero-speaking Kuvale from other population groups was found to be associated with the lack of signals of demographic expansions (Table 1), suggesting that this differentiation was shaped by increased genetic drift. Evidence for reduced levels of mtDNA diversity that are likely to have been caused by recent bottlenecks were previously described in Herero populations from Namibia and Botswana, and seem to be a pervasive feature of these groups [52,53]. However, it is difficult to know to what extent the present diversity patterns reflect the traditional semi-nomadic pastoral way of life of the Herero or were caused by population size reductions ensuing recent conflicts with colonial rulers [54-56]. Moreover, it is not clear whether genetic drift was sufficient to generate the divergence among Herero-speaking groups that is evidenced by comparisons between the

Kuvale from Angola and the Herero from Namibia, which are believed to be the most representative population of the group [17,56]. In fact, while our Kuvale sample is essentially composed of mtDNA haplogroups that are commonly found in other Bantu populations from Angola (Figure 2 and Additional file 1), earlier data indicates that the Herero from Namibia display an unusually high (~50%) frequency of haplogroup L3d [57,58], which was not found in the Kuvale and is known to be much less common in most Bantu populations [6]. This pattern may imply that the broad Herero cultural division encompasses a very heterogeneous set of population groups with no obvious common origin.

A further feature of the genetic composition of the Kuvale that is not paralleled by the Herero from Namibia is their substantial levels of assimilation of Khoe-San lineages (Table 2). In fact, while Khoe-San lineages were absent in sampled Y chromosomes from the Namibian Herero [59] and may represent at most 8% of their mtDNA pool [57,58], typical Khoe-San mtDNA L0d and NRY B2b haplogroups reached 22% and 12% frequencies, respectively, in the Kuvale (Figures 2 and 4). Other sampled Bantu groups from southern Angola that are not as cattle dependent as the Kuvale exhibit much lower levels of Khoe-San lineage assimilation (Figure 2, Table 2 and Additional file 1), suggesting that most gene flow occurred between the herding Khoe peoples and the herding Bantu, probably due to the similarity of their social organization. Given the lack of shared haplotypes between Namibe L0d and B2b haplotypes and the sequences available in databases of Khoisan-speaking populations, we have estimated the ages of the introgressed lineages in order to assess the time depth underlying their present differentiation. In spite of their large uncertainty, coalescent estimates pointed to ages ranging from 4816 (± 4816) to 29168 (± 20624) years, which consistently pre-date the expected arrival time of Bantu-speaking populations to southwestern Africa [1,2]. Thus, it is likely that the divergence of these lineages occurred prior to the recent Bantu expansion. In this context, it is tempting to speculate that the unmatched Angolan L0d and B2b lineages may represent a legacy of the original speakers of Kwadi, an extinct click language remotely related with Central Khoisan that is known to have been spoken in the geographical area presently occupied by the Kuvale [60-62].

The presence in Namibe of the -14010C lactase persistence mutation, which had only been previously found in Kenya and Tanzania [32], raises intriguing questions about the relationships between the East and Southwest African pastoral scenes. The simplest explanation for this observation would be the occurrence of a direct link between the two regions, leading to the introduction of the -14010C mutation in southwestern Angola, most likely by incoming East Bantu migrants originating in East Africa (Figure 6A). However, it is difficult to explain how the -14010C mutation could have spread from a putative Kenyan/Tanzanian center of origin into the remote areas of southwestern Angola without reaching neighboring regions in Mozambique, where no lactase persistence variants could be found. We thus favor an alternative hypothesis that takes into account the association between the frequency of the -14010C variant, the levels of Khoe-San lineage assimilation and the degree of dependence on pastoralism observed in the populations from southwestern Angola (Figure 6). According to this interpretation, the -14010C allele could have been brought to southern Africa by migrant Khoe cattle herders that had previously made contact with Nilotic or Cushitic pastoralists from East Africa. Subsequent interactions in southwestern Angola would have transferred the mutation to Bantu herders concomitantly with mtDNA and Y-chromosome lineages that are specific of the Khoe-San. Under this framework, it is conceivable that the first Bantu speakers arriving in southern Angola acquired cattle only after contact with the Khoe people. Three major lines of evidence support the Khoe-San mediated transfer of the -14010C allele. First, archeological, linguistic and ethnographic data suggest that herding Khoe peoples expanded into southern Africa 2000 years ago from areas around the upper Zambezi River, where they may have acquired pottery and livestock from East African pastoralists spreading as far south as Central Zambia [1,2,63]. Second, a recent analysis of NRY lineages from southern and eastern Africa has defined a new haplogroup (E3b1f), shared by Khoe-San and non-Bantu East Africans, whose distribution is consistent with a Bantu-unrelated demic diffusion of pastoralism from East into southern Africa, which may have started 2000 years ago [45]. Third, and as noted before [32], early reports on lactose tolerance based in physiological tests indicate that Khoisan-speaking people may have moderate levels of lactase persistence [64]. Future studies of lactase persistence based on haplotype resolution of flanking regions may shed light on the levels of genetic differentiation between variants that are presently shared by the compared populations.

The demography of Bantu expansions

We have attempted to infer basic demographic properties of Bantu expansions using the framework of the IM model by assuming that populations located in the southwestern and southeastern edges of sub-equatorial Africa encompass the deepest branches of Bantu divergence after a common origin in West-Central Africa.

A major advantage of the IM class of models is the ability to disentangle the effects of evolutionary factors that are typically confounded in summary statistics based in equi-

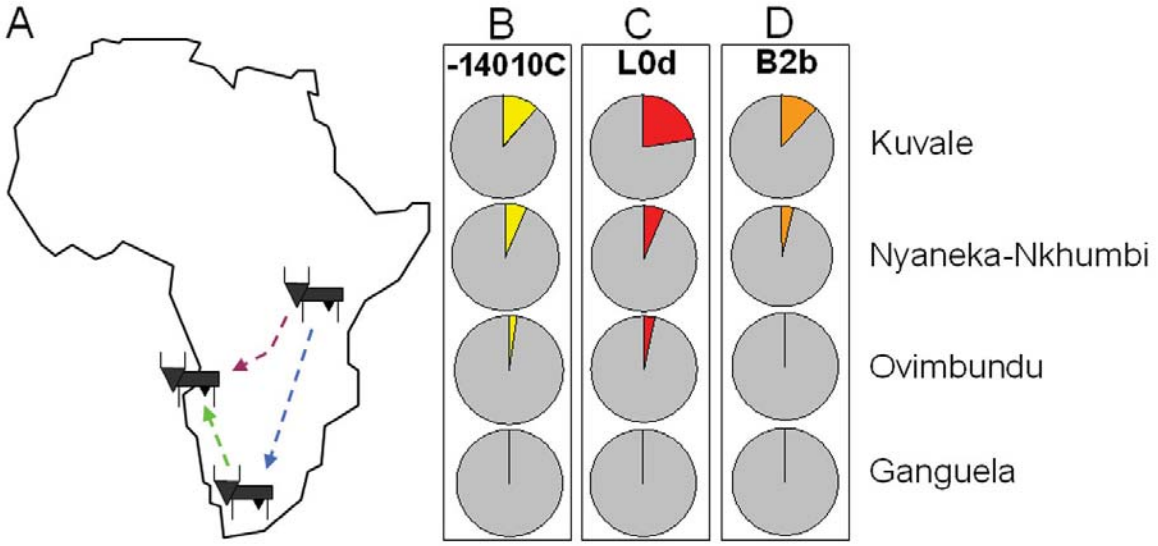


Figure 6
Possible trajectories of the lactase persistence -14010C mutation from East to Southwest Africa. A) Major hypotheses about the migration of the -14010C mutation: a direct migratory link between East and Southwest Africa (violet arrow); a Khoe mediated link, with a first contact between East African pastoralists and the herding Khoe (blue arrow) followed by subsequent transfer to Southwest Bantu pastoralists through Bantu-Khoe interactions (green arrow). B-D) carrier frequencies of the -14010C mutation (B), and typical Khoe-San mtDNA (C) and NRY lineages (D) in major ethnolinguistic groups sampled in the Namibe province.

librium models [19,20,48]. However, like other model-based approaches, the IM framework relies on a number of simplifying assumptions that may be violated by empirical datasets. There are at least two assumptions that may influence the validity of parameter estimates in the context of the Bantu expansions. First, the model does not take into account the effects of gene flow from third party populations, whereas Bantu-speakers did undergo regional interactions with local non-Bantu groups that may distort the interpretation of inferred parameter values. Moreover, gene exchange involving unsampled demes lying between the two edge populations may affect inferences on the true patterns of migration, including the degree of asymmetrical gene flow [65]. A second limiting assumption is that the ancestral population is assumed to be unstructured and to have persisted in isolation for a long time before population splitting [66]. However the patterns of mtDNA variation suggest that the Bantu expansions might have been preceded by complex female-mediated population dynamics involving lineage assemblage in West-Central Africa and the formation of an admixed ancestral population that had no time to achieve panmixy before the expansion. It is possible that the implausibly high divergence time inferred from our mtDNA dataset (t~25000 years; Table 4) was influenced by this kind of older population structure, reflecting lack of panmixy in the ancestral population. In contrast, the more consistent divergence time estimate inferred from the Y-chromosome data (t~2000 years; Table 4) may be related to the erasure of previous ancestral variation that seems to have caused the current predominance of haplogroup E3a-M2, leading to a better fitting of the Y-chromosome data to the model. A further limitation lies in the lack of a geographical specific framework accounting for the spatial expansion of Bantu-speaking peoples.

In spite of these caveats, several consistent results could be found, showing that the analyses presented here do provide informative parameter estimates that may be contrasted in the future with other inferential frameworks and empirical datasets. Our joint estimation of the time of split between the two edges of Bantu migrations (t~4000 years; Table 4), inferred from clearly resolved posterior density peaks, is remarkably consistent with archeology-based estimates for the onset of the dispersion of Bantu speaking peoples across Africa [1,2]. On the other hand,

comparisons between estimates based on the NRY and mtDNA data reveal clearly contrasting patterns between the historic demographic parameters of male and females that may account for key present-day properties of Bantu genetic variation.

Previous comparative studies on Y-chromosome and mtDNA variation in Africa provided evidence for sex biased demographic patterns, including the observation of different levels of correlation between genetic, linguistic and geographic variation [59], as well as the finding that interpopulation differentiation measured by Fst estimators is higher for the Y-chromosome than for the mtDNA in food-producing societies [67]. The latter pattern was interpreted as the result of higher migration rates and/or effective sizes in females than in males. More recently, a resequencing study of mtDNA and Y-chromosome stretches, performed in the same set of sub-Saharan African populations, has found that signals of population expansion in food-producing populations, including one combined Bantu sample, were limited to the mtDNA, while Y-chromosome data better fit models of population stationarity [68]. We found evidence for a demographic expansion both using the Y-chromosome and the mtDNA datasets (Table 4). However, since we used a different inferential framework, a different set of populations and distinct types of genetic information, it is difficult to evaluate the causes of this discrepancy. In any case, our inferences based on the IM model seem to confirm and extend the previous trends by showing that expanding Bantu females most likely had both greater population sizes (N) and higher migration rates (m) (Table 4).

As previously proposed [67,68], it is likely that cultural practices like polygyny, leading to a lower male effective size, and patrilocality, leading to a higher female migration rate, were the major driving forces underlying the observed patterns of genetic variation in current Bantu speaking populations. However, it is important to stress that differences in migration rates among Bantu populations do not necessarily imply differences in the ability to advance and settle new territory. Thus, the higher mobility of females does not mean that males advanced slower than females during the range expansion of Bantu populations, but simply that females were more likely to migrate across the different settlements that were progressively established as the Bantu dispersions unfolded.

Conclusion

Based on patterns of lineage sharing and admixture estimates, our analysis provides evidence that most genetic variation from southwestern Angola is likely to have derived from West-Central Africa. The differences in the amount of haplogroup variation between the mtDNA and Y-chromosome data suggest that the push of Bantu peoples out of the rain forests was preceded by the assemblage of diverse mtDNA contributions in West-Central Africa, a process that was not paralleled by the Y-chromosome, in which lineage extinction must have prevailed. Estimates of demographic parameters have shown that contrasting patterns of female and male genetic variation were a pervasive feature of Bantu expansions, characterized by lower male than female effective sizes and migration rates. Local interactions between the western vanguard of the Bantu migrations and Khoisan-speaking peoples from the arid regions of the South were essentially mediated by Bantu pastoral peoples like the Herero-speaking Kuvale, who share aspects of their social organization with Khoe cattle herders from adjacent areas. We hypothesize that the East African lactase persistence -14010C mutation has been carried to southern Africa by Khoe herders who contacted East African pastoralists and subsequently transferred the mutation to Bantu cattle herders in the course of genetic interactions in the Southwest.

Authors' contributions

JR, SB and MC conceived the study. DL carried out the molecular typing together with MC. MC and FS performed the statistical analyses. JR and SB carried out the fieldwork in southwestern Angola. MC, SB and JR have been involved in interpreting the data and drafting the manuscript. All authors revised critically the manuscript and have given final approval of the version to be published.

Additional material

Additional File 1

MtDNA sequence data from southwestern Angola.
MtDNA sequence data from southwestern Angola. *The table displays HVS-I and HVS-II sequence data and haplogroup classifications in population groups sampled in the Namibe province.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S1.xls]

Additional File 2

NRY haplotype data from southwestern Angola.
NRY haplotype data from southwestern Angola. *The table displays NRY haplotypes defined by UEP and STRs in population groups sampled in the Namibe province.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S2.xls]

Additional File 3

Typing procedure for lactase persistence mutations.

Typing procedure for lactase persistence mutations. *The file provides details on the genotyping method for the following polymorphisms associated with lactase persistence: G/C -14010; T/G -13915; C/T -13910 and C/G -13907.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S3.pdf]

Additional File 4

MtDNA comparative African data.

MtDNA comparative African data. *The table summarizes previously published mtDNA HVS-I datasets on African populations, here considered for comparative purposes.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S4.pdf]

Additional File 5

NRY comparative African data.

NRY comparative African data. *The table summarizes previously published NRY datasets on African populations, here considered for comparative purposes.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S5.pdf]

Additional File 6

The IM framework and the splitting of the western and eastern streams of Bantu migrations.

The IM framework and the splitting of the western and eastern streams of Bantu migrations. *The scheme presents the basic parameters of the IM model in the context of the Bantu expansion. N_A = population effective size of the ancestral population; N_1 = current population size in the Southwest edge; N_2 = current population size in the Southeast edge; m_1 = migration rate from the eastern into the western stream; m_2 = migration rate from the western into the eastern stream. Note that the migration parameters are identified by the destination of migrants as time goes forward.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S6.pdf]

Additional File 7

Patterns of mtDNA lineage sharing by haplogroup.

Patterns of mtDNA lineage sharing by haplogroup. *The figure shows the fractions of lineage sharing between southwestern Angola and other African regions for the most common mtDNA haplogroups (see Figure 3).*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S7.pdf]

Additional File 8

Median-joining network derived from African HVS-I mtDNA sequences belonging to haplogroup L0d.

Median-joining network derived from African HVS-I mtDNA sequences belonging to haplogroup L0d. *The figure shows the phylogenetic relationships between the mtDNA L0d sequences from southwestern Angola and from other African populations. Each circle represents a different haplotype. The area of the circles is proportional to the frequency of the haplotype in the populations. The branch lengths are proportional to the number of mutations separating two sequences.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S8.pdf]

Additional File 9

Median-joining network derived from African Y-chromosome STR-haplotypes belonging to haplogroup B2b

Median-joining network derived from African Y-chromosome STR-haplotypes belonging to haplogroup B2b. *The figure shows the phylogenetic relationships between Y-chromosome B2b haplotypes from southwestern Angola and from other African populations. Haplotypes were defined with a common set of 5 STR loci: DYS19, DYS389I, DYS389II, DYS390, and DYS392. The area of the circles is proportional to the frequency of the haplotype in the populations. The branch lengths are proportional to the number of mutations separating two haplotypes.*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S9.pdf]

Additional File 10

Probability densities for the basic demographic parameters of the IM model.

Probability densities for the basic demographic parameters of the IM model. *The figure provides marginal posterior probability densities for independent runs of the program IMa using the Y-chromosome STR haplotype dataset (L = likelihood).*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S10.pdf]

Additional File 11

Probability densities for the basic demographic parameters of the IM model.

Probability densities for the basic demographic parameters of the IM model. *The figure provides marginal posterior probability densities for independent runs of the program IMa using the mtDNA HVS-I sequence dataset (L = likelihood).*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S11.pdf]

Additional File 12

Probability densities for the basic demographic parameters of the IM model.

Probability densities for the basic demographic parameters of the IM model. *The figure provides marginal posterior probability densities for independent runs of the program IMa using the joint mtDNA and Y-chromosome datasets (L = likelihood).*

Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2148-9-80-S12.pdf]

Acknowledgements

We are grateful to all sample donors, to the Governor of the Namibe Province, Dr. Álvaro de Boavida Neto, and to Dr. Pedro Viyayauca, Chairman of Namibe's Provincial Health Department, for permission to collect samples and José Pimentel for logistic support during field work in Angola. This work was partially financed by the following research grants from Fundação para a Ciência e a Tecnologia (FCT): PPCDT/BIA-BDE/56654/2004 and PTDC/BIA-BDE/68999/2006. MC, FS and SB are supported by FCT grants SFRH/BD/22651/2005, SFRH/BPD/27134/2006, SFRH/BPD/21887/2005, respectively. We also like to acknowledge António Prista and all colleagues of the Human Biological Variability in Mozambique project for the Mozambican samples, and Nuno Ferrand for comments on the manuscript.

References

1. Curtin P, Feierman S, Thompson L, Vansina J: *African History: From Earliest Times to Independence* London, Longman; 1995.
2. Newman JL: *The peopling of Africa: A Geographical Interpretation* New Haven, Yale University Press; 1995.
3. Ehret C: *An African Classical Age: Eastern & Southern Africa in World History, 1000B.C. to A.D.400* Charlottesville, University Press of Virginia; 1998.
4. Blench R: *Archeology, Language, and the African Past* Lanham, Altamira Press; 2006.
5. Rexová K, Bastin Y, Frynta D: **Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data.** *Naturwissenschaften* 2006, **93:**189-194.
6. Salas A, Richards M, De La Fe T, Lareu MV, Sobrino B, Sanchez-Diz P, Macaulay V, Carracedo A: **The making of the African mtDNA landscape.** *Am J Hum Genet* 2002, **71:**1082-1111.
7. Thomas MG, Parfitt T, Weiss DA, Skorecki K, Wilson JF, Le Roux M, Bradman N, Goldstein DA: **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:**674-686.
8. Underhill PA, Passarino G, Lin AA, Shen P, Mirazón-Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL: **The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations.** *Ann Hum Genet* 2001, **65:**43-62.
9. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro-Bisol G, Coia V, Wallace DC, Oefener PJ, Torroni A, Cavalli-Sforza LL, Scozarri R, Underhill PA: **A back migration from Asia to Sub-Saharan Africa is supported by high-resolution analysis of Human Y-chromosome haplotypes.** *Am J Hum Genet* 2002, **70:**1197-1214.
10. Batini C, Coia V, Battaggia C, Rocha J, Pilkington MM, Spedini G, Comas D, Destro-Bisol G, Calafell F: **Phylogeography of the human mitochondrial L1c haplogroup: genetic signatures of the prehistory of Central Africa.** *Mol Phylogenet Evol* 2007, **43:**635-644.
11. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, Balanovsky O, Kidd KK, Kidd JR, Veen L van der, Hombert JM, Gessain A, Verdu P, Froment A, Bahuchet S, Heyer E, Dausset J, Salas A, Behar DM: **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:**1596-1601.
12. Reed FA, Tishkoff SA: **African human diversity, origins and migrations.** *Curr Opin Genet Dev* 2006, **16:**597-605.
13. Plaza S, Salas A, Calafell F, Côrte-Real F, Bertranpetit J, Carracedo A, Comas D: **Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola.** *Hum Genet* 2004, **115:**439-447.
14. Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A: **The genetic legacy of western Bantu migrations.** *Hum Genet* 2005, **117:**366-375.
15. Estermann C: *Etnografia de Angola (Sudoeste e Centro): Colectânea de Artigos Dispersos Volume 2*. Lisbon, Instituto de Investigação Científica Tropical; 1983.
16. Redinha J: *Distribuição Étnica de Angola* Luanda, Cita; 1971.
17. Estermann C: *Etnografia do Sudoeste de Angola: O Grupo Étnico Herero Volume 3*. Lisbon, Junta de Investigações do Ultramar; 1961.
18. Gibson GD: **Foreword.** In *The Herero of Western Botswana: Aspects of Change in a Group of Bantu-Speaking Cattle Herders* Edited by: Vivelo FR. St. Paul, West Publishing; 1977.
19. Nielsen R, Wakeley J: **Distinguishing migration from isolation: a Markov chain Monte Carlo approach.** *Genetics* 2001, **158:**885-896.
20. Hey J, Nielsen R: **Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*.** *Genetics* 2004, **167:**747-760.
21. Redinha J: *Carta Étnica da Província de Angola* Luanda, Olisipo; 1973.
22. Holden CJ: **Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum parsimony analysis.** *Proc Biol Sci* 2002, **269:**793-799.
23. Estermann C: *Etnografia do Sudoeste de Angola: O Grupo Étnico Nhaneca-Humbe Volume 2*. Lisbon, Junta de Investigações do Ultramar; 1961.
24. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A: **The African Diaspora: mitochondrial DNA and the Atlantic slave trade.** *Am J Hum Genet* 2004, **74:**454-465.
25. Cerný V, Salas A, Hájek M, Zaloudková M, Brdicka R: **A bidirectional corridor in the Sahel-Sudan belt and the distinctive features of the Chad Basin populations: a history revealed by the mitochondrial DNA genome.** *Ann Hum Genet* 2007, **71:**433-452.
26. Behar DM, Villems R, Soodyall H, Blue-Smith J, Pereira L, Metspalu E, Scozzari R, Makkan H, Tzur S, Comas D, Bertranpetit J, Quintana-Murci L, Tyler-Smith C, Wells RS, Rosset S, Genographic Consortium: **The dawn of Human matrilineal diversity.** *Am J Hum Genet* 2008, **82:**1130-1140.
27. Consortium YC: **A nomenclature system for the tree of human Y-chromosomal binary haplogroups.** *Genome Res* 2002, **12:**339-348.
28. Jobling MA, Tyler-Smith C: **The human Y chromosome: an evolutionary marker comes of age.** *Nat Rev Genet* 2003, **4:**598-612.
29. Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF: **New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree.** *Genome Res* 2008, **18:**830-838.
30. Enattah NS, Sahi T, Savilahti E, Terwilliger JS, Peltonen L, Järvelä I: **Identification of a variant associated with adult-type hypolactasia.** *Nat Genet* 2002, **30:**233-237.
31. Ingram CJ, Elamin MF, Mulcare CA, Weale ME, Tarekegn A, Raga TO, Bekele E, Elamin FM, Thomas MG, Bradman N, Swallow DM: **A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence?** *Hum Genet* 2007, **120:**779-788.
32. Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Barbbitt CC, Silverman JS, Powell K, Mortensen HM, Hirbo JB, Osman M, Ibrahim M, Omar SA, Lema G, Nyambo TB, Ghori J, Bumpstead S, Pritchard JK, Wray GA, Deloukas P: **Convergent adaptation of human lactase persistence in Africa and Europe.** *Nat Genet* 2007, **39:**31-40.
33. Excoffier L, Laval G, Schneider S: **Arlequin ver. 3.0: An integrated software package for population genetics data analysis.** *Evolutionary Bioinformatics Online* 2005, **1:**47-50.
34. **Henry C Harpending Homepage** [http://harpending.humanevo.utah.edu]
35. **Network 4.5** [http://www.fluxus-engineering.com]
36. Bandelt HJ, Forster P, Sykes BC, Richards MB: **Mitochondrial portraits of human populations using median networks.** *Genetics* 1995, **141:**743-753.
37. Bandelt HJ, Forster P, Röhl A: **Median-joining networks for inferring intraspecific phylogenies.** *Mol Biol Evol* 1999, **16:**37-48.
38. Forster P, Harding R, Torroni A, Bandelt HJ: **Origin and evolution of native American mtDNA variation: a reappraisal.** *Am J Hum Genet* 1996, **59:**935-945.
39. Endicott P, Ho SY: **A Bayesian evaluation of human mitochondrial substitution rates.** *Am J Human Genet* 2008, **82:**895-902.
40. Gusmão L, Sánchez-Diz P, Calafell F, Martín P, Alonso CA, Alvarez-Fernández F, Alves C, Borjas-Fajardo L, Bozzo WR, Bravo ML, Builes JJ, Capilla J, Carvalho M, Castillo C, Catanesi CI, Corach D, Di Lonardo AM, Espinheira R, Fagundes de Carvalho E, Farfán MJ, Figueiredo HP, Gomes I, Lojo MM, Marino M, Pinheiro MF, Pontes ML, Prieto V, Ramos-Luis E, Riancho JA, Souza Góes AC, Santapa OA, Sumita DR, Vallejo G, Vidal Rioja L, Vide MC, Vieira da Silva CI, Whittle MR, Zabala W, Zarrabeitia MT, Alonso A, Carracedo A, Amorim A: **Mutation rates at Y chromosome specific microsatellites.** *Hum Mutat* 2005, **26:**520-528.

41. Dupanloup I, Bertorelle G: **Inferring admixture proportions from molecular data: extension to any array of parental populations.** *Mol Biol Evol* 2001, **18:**672-675.
42. **Hey Lab Software and Data** [http://lifesci.rutgers.edu/~heylab/HeylabSoftware.htm#IM]
43. Hasegawa M, Kishino H, Yano T: **Dating of the human-ape splitting by a molecular clock of mitochondrial DNA.** *J Mol Evol* 1985, **22:**160-174.
44. 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 Atlantic slave trade.** *Ann Hum Genet* 2001, **65:**439-458.
45. Henn BM, Gignoux C, Lin AA, Oefner PJ, Shen P, Scozzari R, Cruciani F, Tishkoff SA, Mountain JL, Underhill PA: **Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa.** *Proc Natl Acad Sci USA* 2008, **105:**10693-10698.
46. Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL: **History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation.** *Mol Biol Evol* 2007, **24:**2180-2195.
47. Swallow DM: **Genetics of lactase persistence and lactose intolerance.** *Annu Rev Genet* 2003, **37:**197-219.
48. Hey J: **On the number of New World founders: a population genetic portrait of the peopling of the Americas.** *PLOS Biology* 2005, **3:**e193.
49. 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.
50. Philipson DW: **The spread of the Bantu language.** *Sc Am* 1977, **4:**106-114.
51. Rosa A, Ornelas C, Jobling MA, Brehm A, Villems R: **Y chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective.** *BMC Evolutionary Biology* 2007, **7:**124.
52. Harpending H, Sherry ST, Rogers AR, Stoneking M: **The genetic structure of ancient human populations.** *Curr Anthrop* 1993, **34:**483-496.
53. Excoffier L, Schneider S: **Why hunter-gatherer populations do not show sign of Pleistocene demographic expansions.** *Proc Natl Acad Sci USA* 1999, **96:**10597-10602.
54. Vivelo FR: *The Herero of Western Botswana: Aspects of Change in a Group of Bantu-Speaking Cattle Herders* St. Paul, West Publishing; 1977.
55. Carvalho RD: *Vou lá visitar pastores* Lisbon, Cotovia; 2000.
56. Pennington RL: **Economic stratification and health among the Herero of Botswana.** In *Human Biology of Pastoral Populations* Edited by: Leonard W, Crawford MH. Cambridge, Cambridge University Press; 2002.
57. Soodyall H, Jenkins T: **Mitochondrial DNA polymorphisms in Negroid populations from Namibia: new light on the origins of the Dama, Herero and Ambo.** *Ann Hum Biol* 1993, **20:**477-485.
58. Chen YS, Olckers A, Schurr TG, Kogelnik AM, Huoponen K, Wallace D: **mtDNA Variation in the South African Kung and Khwe-and their genetic relationship to other African populations.** *Am J Hum Genet* 2000, **66:**1362-1383.
59. Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF: **Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes.** *Eur J Hum Genet* 2005, **13:**867-876.
60. de Almeida A: *Os Bosquímanos de Angola* Lisbon, Instituto de Investigação Científica Tropical; 1994.
61. Westphal EOJ: **The linguistic pre-history of southern Africa: Bush, Kwadi, Hottentot, and Bantu linguistic relationships.** *Africa* 1963, **33:**237-265.
62. Güldemann T: **Reconstruction through "de-construction": the marking of person, gender and number in the Khoe family and Kwadi.** *Diachronica* 2004, **21:**251-306.
63. Blench R: **Was there and interchange between Cushitic pastoralists and Khoisan speakers in the prehistory of Southern Africa and how can this be detected?** *Sprache und Geschichte in Afrika* in press.
64. Casimir MJ: **On milk-drinking San and the "myth of the primitive isolate".** *Curr Anthrop* 1990, **31:**551-554.
65. Garrigan D, Kingan SB, Pilkington M, Wilder J, Cox MP, Soodyall H, Strassman B, Destro-Bisol G, de Knijff P, Novelletto A, Friedlaender J, Hammer MF: **Inferring human population sizes, divergence times and rates of gene flow from mithocondrial, X and Y chromosome resequencing data.** *Genetics* 2007, **177:**2195-2207.
66. Won Y-J, Sivasundar A, Wang Y, Hey : **On the origin of Lake Malawi cichlid species: a population genetic analysis of divergence.** *Proc Natl Acad Sci* 2005, **102:**6581-6586.
67. Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglià A, Tofanelli S, Spedini G, Capelli C: **Variation of female and male lineages in Sub-Saharan populations: the importance of sociocultural factors.** *Mol Biol Evol* 2004, **21:**1673-1682.
68. Pilkington MM, Wilder JA, Mendez FL, Cox MP, Woerner A, Angui T, Kingan S, Mobasher Z, Batini C, Destro-Bisol G, Soodyall H, Strassmann BI, Hammer MF: **Contrasting signatures of population growth for mitochondrial DNA and Y chromosomes among human populations in Africa.** *Mol Biol Evol* 2008, **25:**517-525.

Additional File 3

Typing procedure for lactase persistence mutations

Lactase persistence mutations genotyping protocol

Lactase persistence mutations G/C -14010, T/G -13915, C/T -13910 and C/G -13907 were genotyped by direct sequencing. A 359bp fragment containing all mentioned mutations and located in the intron 13 of the *MCM6* gene was amplified using primers 5'-GCAGGGCTCAAAGAACAATC-3' (forward) and 5'-TGTTGCATGTTTTTAATCTTTGG-3' (reverse). PCR reactions contained 0,5µM of each primer, 0,2mM of each deoxynucleotide triphosphate (dNTP), 750 mM Tris-HCl (pH 8.8 at 25°C), 200 mM $(NH_4)_2SO_4$, 0.1% (v/v) Tween 20, 1,5mM $MgCl_2$ and 1 U Taq polymerase. The PCR profile consisted of: 94ºC for 5 min, 35 cycles of 94ºC for 1 min, 58ºC for 1 min and 72ºC for 1min, followed by a 20-min extension at 72ºC.

Sequencing reactions were carried out using the ABI Big Dye v3.1 Ready Reaction Kit and using the protocol specified by the manufacturer (Applied Biosystems, Inc. Foster City, CA). Products were run on an ABI PRISM 3130xl sequencer and analyzed in the ABI PRISM 3130xl Genetic Analyzer software (Applied Biosystems, Inc. Foster City, CA). The resulting chromatograms were inspected for the presence/absence of the lactase mutations using MEGA4.0 software [1].

References:

1. Tamura K, Dudley J, Nei M, Kumar S: **MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0**. *Mol Biol Evol* 2007, **24**:1596-1599.

Additional File 4
MtDNA comparative African data

African populations used for mtDNA comparisons.

Geographic area/Contextual samples; Pop name	Place of origin	Linguistic Affiliation	n	Code in Figure1	Reference
West Africa					
Tuareg[1,2,3]	Niger	Afro-Asiatic	23	1	[1]
Songhai[1,2,3]	Mali, Niger	Nilo-Saharan	10	2	[1]
Fulbe[1,2,3]	Benin, Burkina Faso, Níger, Nigeria	Niger Congo	60	3	[1]
Yoruba[1,2,3]	Nigeria	Niger-Congo	33	4	[1];[2]
Guinea-Bissau (various) [1,2,3]	Guinea-Bissau	Niger-Congo	372	5	[3]
Serer[1,2,3]	Senegal	Niger-Congo	23	6	[4]
Wolof[1,2,3]	Senegal	Niger-Congo	48	6	[4]
Senegalese (various) [1,2,3]	Senegal	Niger-Congo	50	6	[4]
Mandenka[1,2,3]	Senegal	Niger-Congo	110	6	[5]
Sierra Leone (various) [1,2,3]	Sierra Leone	Niger-Congo	276	7	[6]
Cabo Verde[1,2,3]	CaboVerde	Creole	292	8	[7]
West-Central Africa					
Bamileke[1,2,3]	Cameroon	Niger-Congo	48	13	[8]
Fali[1,3]	Cameroon	Niger-Congo	41	13	[9]
Fulbe[1,2,3]	Cameroon	Niger-Congo	34	3	[9]
Tali[1,3]	Cameroon	Niger-Congo	20	13	[9]
Tupuri[1,3]	Cameroon	Niger-Congo	25	13	[9]
Daba[1,3]	Cameroon	Afro-Asiatic	20	13	[9]
Mandara[1,3]	Cameroon	Afro-Asiatic	37	13	[9]
Podokwo[1,3]	Cameroon	Afro-Asiatic	39	13	[9]
Uldeme[1,3]	Cameroon	Afro-Asiatic	28	13	[9]
Bassa[1,2,3]	Cameroon	Niger-Congo (Bantu)	46	13	[9]
Bakaka[1,2,3]	Cameroon	Niger-Congo (Bantu)	50	13	[9]
Ngumba[1,2,3]	Cameroon	Niger-Congo (Bantu)	88	13	[10]
Fang[1,2,3]	Cameroon, Equatorial Guinea, Gabon	Niger-Congo (Bantu)	116	14	[10];[11]
Ewondo[1,2,3]	Cameroon	Niger-Congo (Bantu)	78	13	[8];[10]
Bubi[1,2,3]	Equatorial Guinea	Niger-Congo (Bantu)	45	16	[12]
Akele[1,2,3]	Gabon	Niger-Congo (Bantu)	48	17	[10]
Ateke[1,2,3]	Gabon	Niger-Congo (Bantu)	54	17	[10]
Benga[1,2,3]	Gabon	Niger-Congo (Bantu)	50	17	[10]
Dumba[1,2,3]	Gabon	Niger-Congo (Bantu)	47	17	[10]
Eshira[1,2,3]	Gabon	Niger-Congo (Bantu)	40	17	[10]
Eviya[1,2,3]	Gabon	Niger-Congo (Bantu)	38	17	[10]
Galoa[1,2,3]	Gabon	Niger-Congo (Bantu)	51	17	[10]
Kota[1,2,3]	Gabon	Niger-Congo (Bantu)	56	17	[10]
Makina[1,2,3]	Gabon	Niger-Congo (Bantu)	45	17	[10]
Mitsogo[1,2,3]	Gabon	Niger-Congo (Bantu)	64	17	[10]
Ndumu[1,2,3]	Gabon	Niger-Congo (Bantu)	39	17	[10]
Nzebi[1,2,3]	Gabon	Niger-Congo (Bantu)	63	17	[10]
Obamba[1,2,3]	Gabon	Niger-Congo (Bantu)	47	17	[10]
Orungu[1,2,3]	Gabon	Niger-Congo (Bantu)	20	17	[10]
Punu[1,2,3]	Gabon	Niger-Congo (Bantu)	52	17	[10]
Shake[1,2,3]	Gabon	Niger-Congo (Bantu)	51	17	[10]
Pygmies					
Babongo Pygmies[1,2,3]	Gabon	Niger-Congo	45	20	[10]

Baka Pygmies[1,2,3]	Cameroon, Gabon	Niger-Congo	127	21	[10]
Bakola Pygmies[1,2,3]	Cameroon	Niger-Congo	118	22	[10];[13]
Bakoya Pygmies[1,2,3]	Gabon	Niger-Congo	31	23	[10]
Biaka Pygmies[1,2,3]	Central African Republic	Niger-Congo	73	24	[2];[10]
Mbenzele Pygmies[1,2,3]	Central African Republic	Niger-Congo	57	25	[8]
Tikar Pygmies[1,2,3]	Cameroon	Niger-Congo	35	26	[10]
East Africa					
Etiopia[1,2,3]	Ethiopia	Afro-Asiatic	270	27	[14]
Somali[1,2,3]	Somalia	Afro-Asiatic	27	28	[1]
Burunge[1,2,3]	Tanzania	Afro-Asiatic	38	29	[13]
Turkana[1,2,3]	Kenya	Nilo-Saharan	37	30	[1]
Nairobi[2]	Kenya	Niger-Congo	100	31	[15]
Luo[1,2,3]	Kenya	Nilo-Saharan	53	32	Luiselli et al, unpublished
Nubian[1,2,3]	Nubia	Nilo-Saharan	80	33	[16]
Sudanese[1,2,3]	Sudan	Nilo-Saharan	76	34	[16]
Kikuyu[2]	Kenya	Niger-Congo (Bantu)	24	35	[1]
Turu[2]	Tanzania	Niger-Congo (Bantu)	29	36	[13]
Iraqw[1,2,3]	Tanzania	Afro-Asiatic	12	37	[17]
Datog[1,2,3]	Tanzania	Nilo-Saharan	57	38	[13]; [17]
Sukuma[2]	Tanzania	Niger-Congo (Bantu)	32	39	[13];[17]
Hadza[1,2]	Tanzania	Khoisan	145	42	[2];[13];[17]
Sandawe[1,2]	Tanzania	Khoisan	82	43	[13]
Southwest Africa					
Cabinda (Kikongo) [1,2]	Angola	Niger-Congo (Bantu)	109	44	[18]
Luanda (Kimbundu)[1,2]	Angola	Niger-Congo (Bantu)	44	45	[19]
Namibe (various) [1,2,3,4]	**Angola**	**Niger-Congo (Bantu)**	**365**	**46**	**Present study**
Southeast Africa					
Shona[1,2,4]	Zimbabwe	Niger-Congo (Bantu)	61	48	Luiselli et al, unpublished
Mozambique (various) [1,2]	Mozambique	Niger-Congo (Bantu)	109	47	[20]
Mozambique (various) [1,2,4]	Mozambique	Niger-Congo (Bantu)	307	47	[21]
Southern Africa Khoisans					
!Kung[1,2,3]	Botswana, South Africa	Khoisan	68	49	[2];[22]
Khwe[1,2,3]	South Africa	Khoisan	31	50	[22]
!Xun/Khwe[1,2,3]	South Africa	Khoisan	18	50	[13]

[1]Population sample used in haplogroup frequency analyses (pie charts and PC analysis);
[2]Population sample used in Lineage sharing analysis;
[3]Population sample used in admixture analyses;
[4]Population sample used in IM model and estimation of demographic parameters.

References:

1. Watson E, Bauer K, Aman R, Weiss G, von Haeseler A, Pääbo S: **mtDNA sequence diversity in Africa**. *Am J Hum Genet* 1996, **59**:437-444.
2. Vigilant L, Stoneking M, Harpending H, Hawkes K, Wilson AC: **African populations and the evolution of human mitochondrial DNA.** *Science* 1991, **253**:1503-1507.
3. Rosa A, Brehm A, Kivisild T, Metspalu E, Villems R: **MtDNA profile of West Africa Guineans: towards a better understanding of the Senegambia region.** *Ann Hum Genet* 2004, **68**: 340-352.
4. Rando JC, Pinto F, González AM, Hernández M, Larruga JM, Cabrera VM, Bandelt HJ: **Mitochondrial DNA analysis of northwest African populations reveals genetic exchanges with European, near-eastern, and sub-Saharan populations.** *Ann Hum Genet* 1998, **62**:531-550.

5. Graven L, Passarino G, Semino O, Boursot P, Santachiara-Benerecetti S, Langaney A, Excoffier L: **Evolutionary correlation between control region sequence and restriction polymorphisms in the mitochondrial genome of a large Senegalese Mandenka sample**. *Mol Biol Evol* 1995, **12**: 334-345.
6. Jackson BA, Wilson JL, Kirbah S, Sidney SS, Rosenberger J, Bassie L, Alie JA, McLean DC, Garvey WT, Ely B: **Mitochondrial DNA genetic diversity among four ethnic groups in Sierra Leone.** *Am J Phys Anthropol* 2005, **128**: 156-163.
7. Brehm A, Pereira L, Bandelt HJ, Prata MJ, Amorim A: **Mitochondrial portrait of the Cabo Verde archipelago: the Senegambian outpost of Atlantic slave trade.** *Ann Hum Genet* 2002, **66**: 49–60.
8. Destro-Bisol G, Coia V, Boschi I, Verginelli F, Cagliá A, Pascali V, Spedini G, Calafell F: **The analysis of variation of mtDNA hypervariable region 1 suggests that Eastern and Western Pygmies diverged before the Bantu expansion.** *Am Nat* 2004, **163**: 212-226.
9. Coia V, Destro-Bisol G, Verginelli F, Battaggia C, Boschi I, Cruciani F, Spedini G, Comas D, Calafell F: **Brief communication: mtDNA variation in North Cameroon: Lack of Asian lineages and implications for back migration from Asia to sub-Saharan Africa.** *Am J Phys Anthropol* 2005, **128**: 678-681.
10. Quintana-Murci L, Quach H, Harmant C, Luca F, Massonnet B, Patin E, Sica L, Mouguiama-Daouda P, Comas D, Tzur S, Balanovsky O, Kidd KK, Kidd JR, van der Veen L, Hombert JM, Gessain A, Verdu P, Froment A, Bahuchet S, Heyer E, Dausset J, Salas A, Behar DM: **Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers.** *Proc Natl Acad Sci U S A* 2008, **105**: 1596-1601.
11. Pinto F, Gonzalez AM, Hernandez M, Larruga JM, Cabrera VM: **Genetic relationship between the Canary Islanders and their African and Spanish ancestors inferred from mitochondrial DNA sequences.** *Ann Hum Genet* 1996, **60**: 321-330.
12. Mateu E, Comas D, Calafell F, Pérez-Lezaun A, Abade A, Bertranpetit J: **A tale of two islands: Population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea.** *Ann Hum Genet* 1997, **61**: 507-518.
13. Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL: **History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation.** *Mol Biol Evol* 2007, **24**: 2180-2195.
14. Kivisild T, Reidla M, Metspalu E, Rosa A, Brehm A, Pennarun E, Parik J, Geberhiwot T, Usanga E, Villems R: **Ethiopian mitochondrial DNA heritage: Tracking gene flow across and around the gate of tears.** *Am J Hum Genet* 2004, **75**: 752-770.
15. Brandstätter A, Peterson CT, Irwin JA, Mpoke S, Koech DK, Parson W, Parsons TJ: **Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database.** *Int J Legal Med* 2004, **118**: 294-306.
16. Krings M, Salem AE, Bauer K, Geisert H, Malek AK, Chaix L, Simon C, Welsby D, Di Rienzo A, Utermann G, Sajantila A, Pääbo S, Stoneking M: **mtDNA analysis of Nile River Valley populations: A genetic corridor or a barrier to migration?** *Am J Hum Genet* 1999, **64**: 1166-1176.
17. Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL: **African Y chromosome and mtDNA divergence provides insight into the history of click languages.** *Curr Biol* 2003, **13**: 464-473.
18. Beleza S, Gusmão L, Amorim A, Carracedo A, Salas A: **The genetic legacy of western Bantu migrations.** *Hum Genet* 2005, **117**: 366-375.
19. Plaza S, Salas A, Calafell F, Corte-Real F, Bertranpetit J, Carracedo A, Comas D: **Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola.** *Hum Genet* 2004 **115**: 439-447.
20. Pereira L, Macaulay V, Torroni A, Scozzari R, Prata MJ, Amorim A: **Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade**. *Ann Hum Genet* 2001, **65**: 439-458.
21. Salas A, Richards M, Lareu MV, Scozzari R, Coppa A, Torroni A, Macaulay V, Carracedo A: **The making of the African mtDNA landscape.** *Am J Hum Genet* 2002, **71**: 1082-1111.
22. 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**: 1362-1383.

Additional File 5
NRY comparative African data

African populations used for Y-chromosome comparisons

Geographic area/Contextual samples; Pop name	Place of origin	Linguistic Affiliation	n	Code in Figure1	Reference
West Africa					
Tuareg[2,3]	Niger	Afro-Asiatic	9	1	[1]
Songhai[2,3]	Niger	Nilo-Saharan	5	2	[1]
Fon[1]	Benin	Niger-Congo	100	9	[2]
Rimaibe[1]	Burkina Faso	Niger Congo	37	10	[3]
Fulbe[1]	Burkina Faso	Niger-Congo	20	3	[3]
Fulbe[2,3]	Niger	Niger-Congo	6	3	[1]
Mossi[1]	Burkina Faso	Niger-Congo	49	11	[3]
Guinea Bissau (various) [2,3]	Guinea Bissau	Niger-Congo	203	5	[4],[5]
Guinea Bissau (various) [1]	Guinea Bissau	Niger-Congo	232	5	[6]
Mali (various) [1]	Mali	-[5]	44	12	[7]
Yoruba[1,2,3]	Nigeria	Niger-Congo	13	4	[8]
Senegal (various) [1]	Senegal	Niger-Congo	139	6	[9]
West-Central Africa					
Bamileke[2,3]	Cameroon	Niger-Congo	53	13	[10]
Bamileke[1]	Cameroon	Niger-Congo	133	13	[2],[3]
Bakaka[1]	Cameroon	Niger-Congo (Bantu)	12	13	[3]
Cameroon (various) [1]	Cameroon	Niger-Congo (Bantu)	14	13	[2]
Daba[1]	Cameroon	Afro-Asiatic	18	13	[3]
Ewondo[2,3]	Cameroon	Niger-Congo (Bantu)	20	13	[10]
Ewondo[1]	Cameroon	Niger-Congo (Bantu)	29	13	[3]
Tali[1]	Cameroon	Niger-Congo	15	13	[3]
Fali[1]	Cameroon	Niger-Congo	39	13	[3]
Fulbe[1]	Cameroon	Niger-Congo	17	3	[3]
Mixed Adamawa[1]	Cameroon	Niger-Congo	18	13	[3]
Mixed Chadic[1]	Cameroon	Afro-Asiatic	15	13	[3]
Mixed Nilo-Saharan[1]	Cameroon	Nilo-Saharan	9	13	[3]
Ouldeme[1]	Cameroon	Afro-Asiatic	21	13	[3]
Bangui[2,3]	Central African Republic	Niger-Congo	122	18	[11]
Lissongo[2,3]	Central African Republic	Niger-Congo (Bantu)	4	19	[1]
Bubi[2,3]	Equatorial Guinea	Niger-Congo (Bantu)	133	16	[12]
Fang[2,3]	Equatorial Guinea	Niger-Congo (Bantu)	110	14	[12]
Equatorial Guinea (various) [2,3]	Equatorial Guinea	Niger Congo	101	15	[13]
Pygmies					
Biaka Pygmies[1]	Central African Republic	Niger-Congo	20	24	[3]
Biaka Pygmies[2,3]	Central African Republic	Niger-Congo	8	24	[8]
Pygmies[2,3]	Central African Republic	Niger-Congo	20	24	[1]
East Africa					
Ethiopian Nilo-Saharan[2,3]	Ethiopia	Nilo-Saharan	40	27	[1]
Ethiopian Afro-Asiatic[2,3]	Ethiopia	Afro-Asiatic	44	27	[1]
Ethiopians[1]	Ethiopia	Afro-Asiatic	126	27	[9]

Ethiopia (various) [1]	Ethiopia	-[5]	88	27	[7]
Somalia (various) [2,3]	Somalia	-[5]	201	28	[15]
Burunge[1,2,3]	Tanzania	Afro-Asiatic	24	29	[8]
Sudan (various) [1]	Sudan	-[5]	40	34	[7]
Turu[2]	Tanzania	Niger-Congo (Bantu)	20	36	[8]
Iraqw[1]	Tanzania	Afro-Asiatic	6	37	[16]
Datog[1,2,3]	Tanzania	Nilo-Saharan	35	38	[8]
Sukuma[2]	Tanzania	Niger-Congo (Bantu)	30	39	[8]
Mbugwe[2]	Tanzania	Niger-Congo (Bantu)	14	40	[8]
Maasai[1]	Kenya	Nilo-Saharan	26	41	[14]
Hadzabe[1,2]	Tanzania	Khoisan	54	42	[8]
Hadzabe[1]	Tanzania	Khoisan	23	42	[16]
Sandawe[1,2]	Tanzania	Khoisan	67	43	[8]
Southwest Africa					
Cabinda (various) [1,2]	Angola	Niger-Congo (Bantu)	74	44	[17]
Luanda (various) [2]	Angola	Niger-Congo (Bantu)	50	45	[5]
Namibe (various) [1,2,3,4]	**Angola**	**Niger-Congo (Bantu)**	**236**	**46**	**Present study**
Southeast Africa					
Mozambique (various) [2,4]	Mozambique	Niger-Congo (Bantu)	112	47	[18]
South African					
Omega San[2,3]	Namibia	Khoisan	15	51	[1]
Sekele San[2,3]	South Africa	Khoisan	14	51	[1]
San[2,3]	South Africa	Khoisan	8	51	[8]
!Kung[1]	South Africa	Khoisan	64	49	[3]
Kwe[1]	South Africa	Khoisan	26	50	[3]

[1]Population sample used in haplogroup frequency analyses (pie charts and PC analysis);
[2]Population sample used in lineage sharing analysis;
[3]Population sample used in admixture analyses;
[4]Population sample used in IM model and estimation of demographic parameters.
[5]Linguistic Affiliation was not mentioned in the reference

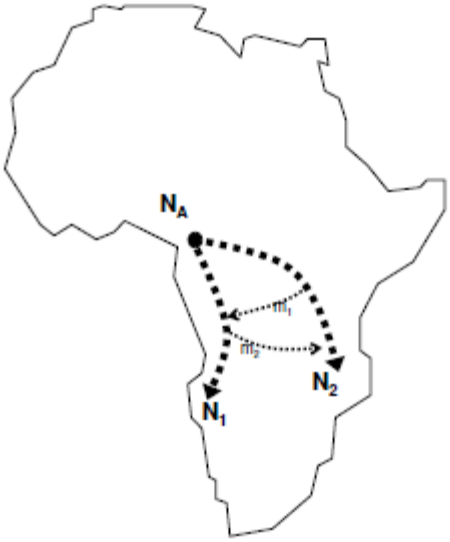
References:

1. Pritchard JK, Seielstad M, Perez-Lezaun A, Feldman M: **Population Growth of Human Y Chromosomes: A Study of Y Chromosome Microsatellites.** *Mol Biol Evol* 1999, **16**: 1791-1798.
2. Luis JR, Rowold DJ, Regueiro M, Caeiro B, Cinnioglu C, Roseman C, Underhill PA, Cavalli-Sforza LL, Herrera RJ: **The Levant versus the Horn of Africa: evidence for bidirectional corridors of human migrations.** *Am J Hum Genet* 2004, **74**: 532-544.
3. Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, Modiano D, Holmes S, Destro–Bisol G, Coia V, Wallace DC, Oefner PJ, Torroni A, Cavalli–Sforza LL, Scozzari R, Underhill PA: **A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes.** *Am J Hum Genet* 2002, **70**:1197–1214.
4. Rosa A, Ornelas C, Jobling MA, Brehm A, Villems R: **Y-chromosomal diversity in the population of Guinea-Bissau: a multiethnic perspective.** *BMC Evol Biol* 2007, **7**: 124.
5. Côrte-Real F, Carvalho M, Andrade L, Anjos MJ, Pestoni C, Lareu MV, Carracedo A, Vieira DN, Vide MC: **Chromosome Y STRs analysis and evolutionary aspects for Portuguese spoken countries.** In *Progress in forensic Genetics* 8; *Amsterdam*. Edited by Sensabaugh, G F P, Lincoln J, Olaisen, B: Elsevier Science; 2000: 272-274.
6. Gonçalves R, Rosa A, Freitas A, Fernandes A, Kivisild T, Villems R, Brehm A: **Y-chromosome lineages in Cabo Verde Islands witness the diverse geographic origin of its first male settlers.** *Hum Genet* 2003, **113**: 467-472.

7. Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, Kauffman E, Bonne-Tamir B, Bertranpetit J, Francalacci P, Ibrahim M, Jenkins T, Kidd JR, Mehdi SQ, Seielstad MT, Wells RS, Piazza A, Davis RW, Feldman MW, Cavalli-Sforza LL, Oefner PJ: **Y chromosome sequence variation and the history of human populations.** *Nat Genet* 2000, **26**: 358–361.
8. Tishkoff SA, Gonder MK, Henn BM, Mortensen H, Knight A, Gignoux C, Fernandopulle N, Lema G, Nyambo TB, Ramakrishnan U, Reed FA, Mountain JL: **History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation.** *Mol Biol Evol* 2007, **24**: 2180-2195.
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**:265–268.
10. Caglià A, Tofanelli S, Coia V, Boschi I, Pescarmona M, Spedini G, Pascali V, Paoli G, Destro-Bisol G: **A study of Y-chromosome microsatellite variation in sub-Saharan Africa: a comparison between F(ST) and R(ST) genetic distances.** *Hum Biol* 2003, **75**: 313-330.
11. Lecerf M, Filali M, Grésenguet G, Ndjoyi-Mbiguino A, Le Goff J, de Mazancourt P, Bélec L: **Allele frequencies and haplotypes of eight Y-short tandem repeats in Bantu population living in Central Africa.** *Forensic Sci Int* 2007, **171**: 212-215.
12. Barrot C, Sánchez C, Xifró A, Ortega M, Mas J, Huguet E, Corbella J, Gené M: **Data for Y-chromosome haplotypes in Fang and Bubi populations from Bioko (Equatorial Guinea).** *Forensic Sci Int* 2007, **168**: e10-12.
13. Arroyo-Pardo E, Gusmão L, López-Parra AM, Baeza C, Mesa MS, Amorim A: **Genetic variability of 16 Y-chromosome STRs in a sample from Equatorial Guinea (Central Africa).** *Forensic Sci Int* 2005, **149**: 109-113.
14. Wood ET, Stover DA, Ehret C, Destro-Bisol G, Spedini G, McLeod H, Louie L, Bamshad M, Strassmann BI, Soodyall H, Hammer MF: **Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes.** *Eur J Hum Genet* 2005, **13**: 867-876.
15. Hallenberg C, Simonsen B, Sanchez J, Morling N: **Y-chromosome STR haplotypes in Somalis.** *Forensic Sci Int* 2005, **151**: 317-321.
16. Knight A, Underhill PA, Mortensen HM, Zhivotovsky LA, Lin AA, Henn BM, Louis D, Ruhlen M, Mountain JL: **African Y chromosome and mtDNA divergence provides insight into the history of click languages.** *Curr Biol* 2003, **13**: 464-473.
17. Beleza S: **Phylogenetic and demographic history of two human populations revealed by the analysis of two non-recombining segments of the genome: Y-chromosome and mitochondrial DNA.** *PhD thesis*. Santiago Compostela University, 2005
18. Alves C, Gusmão L, Barbosa J, Amorim A: **Evaluating the informative power of Y-STRs: a comparative study using European and new African haplotype data.** *Forensic Sci Int* 2003, **134**: 126-133.

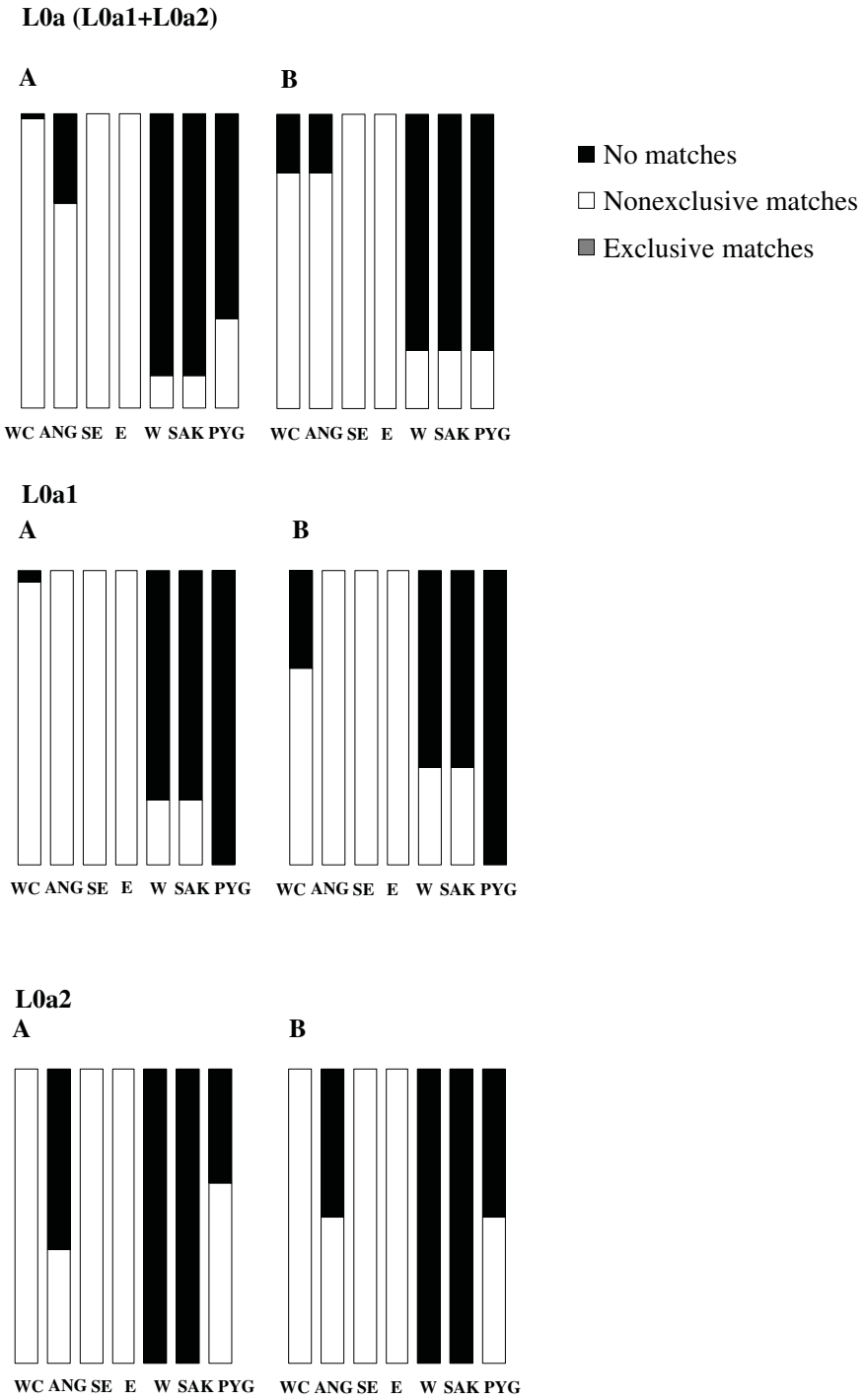
Additional File 6

The IM framework and the splitting of the western and eastern streams of Bantu migrations

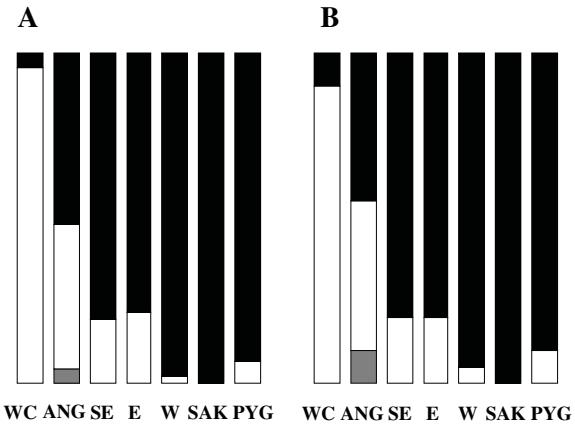


Additional File 7

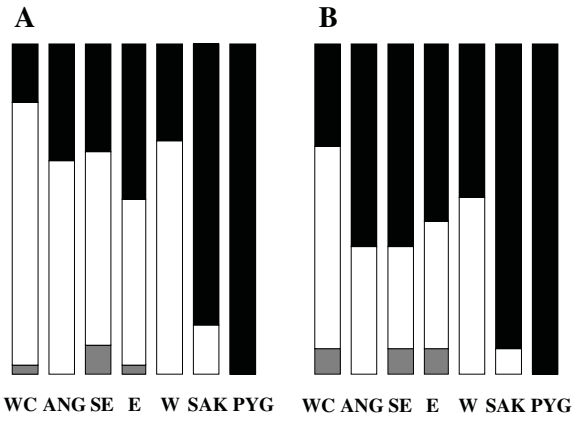
Patterns of mtDNA lineage sharing by haplogroup



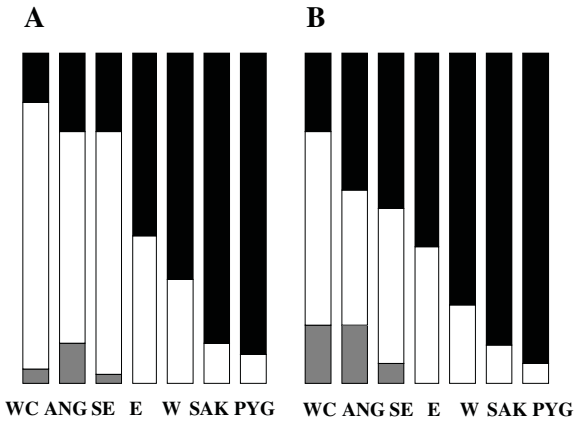
L1c

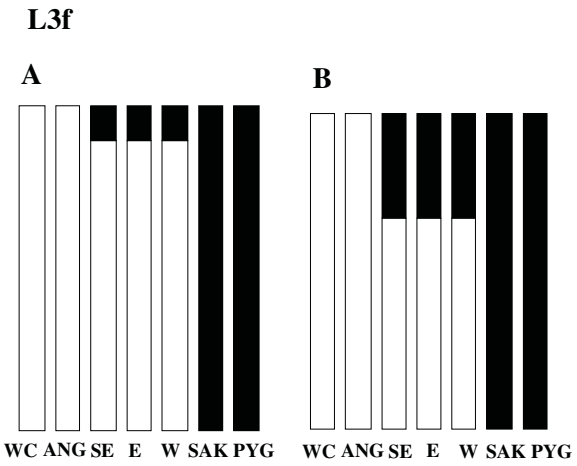

A
B
WC ANG SE E W SAK PYG
WC ANG SE E W SAK PYG

L2a


A
B
WC ANG SE E W SAK PYG
WC ANG SE E W SAK PYG

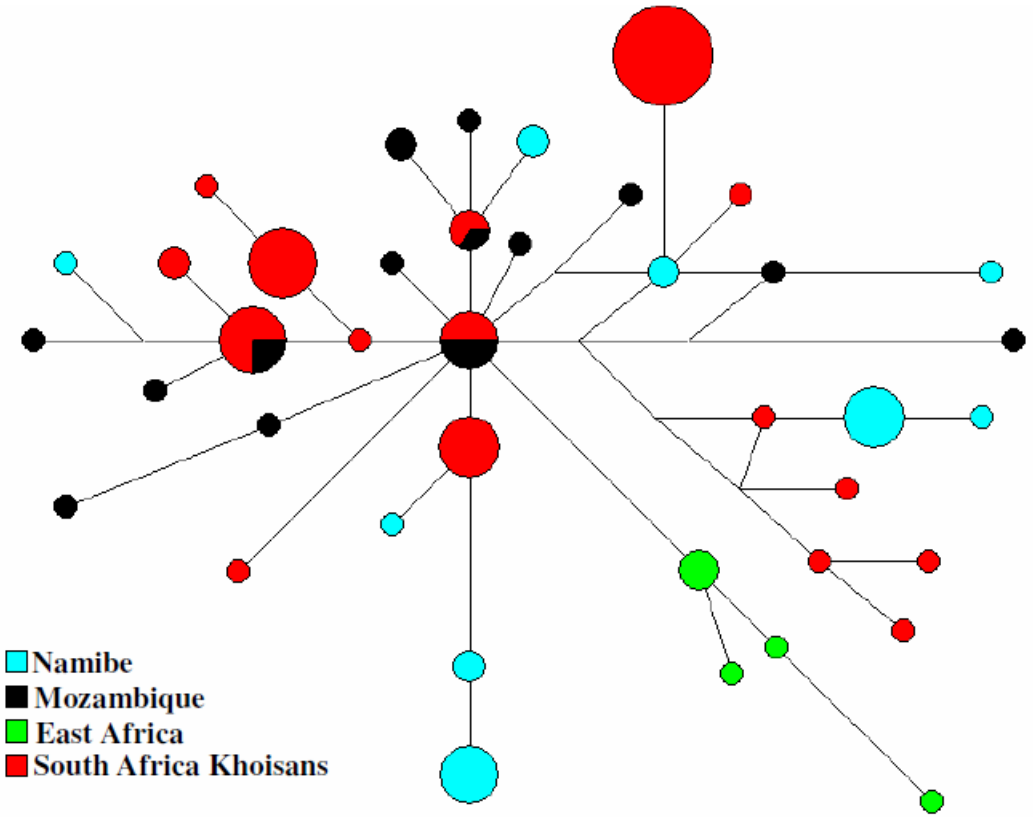
L3e


A
B
WC ANG SE E W SAK PYG
WC ANG SE E W SAK PYG


L3f
A
B
WC ANG SE E W SAK PYG
WC ANG SE E W SAK PYG

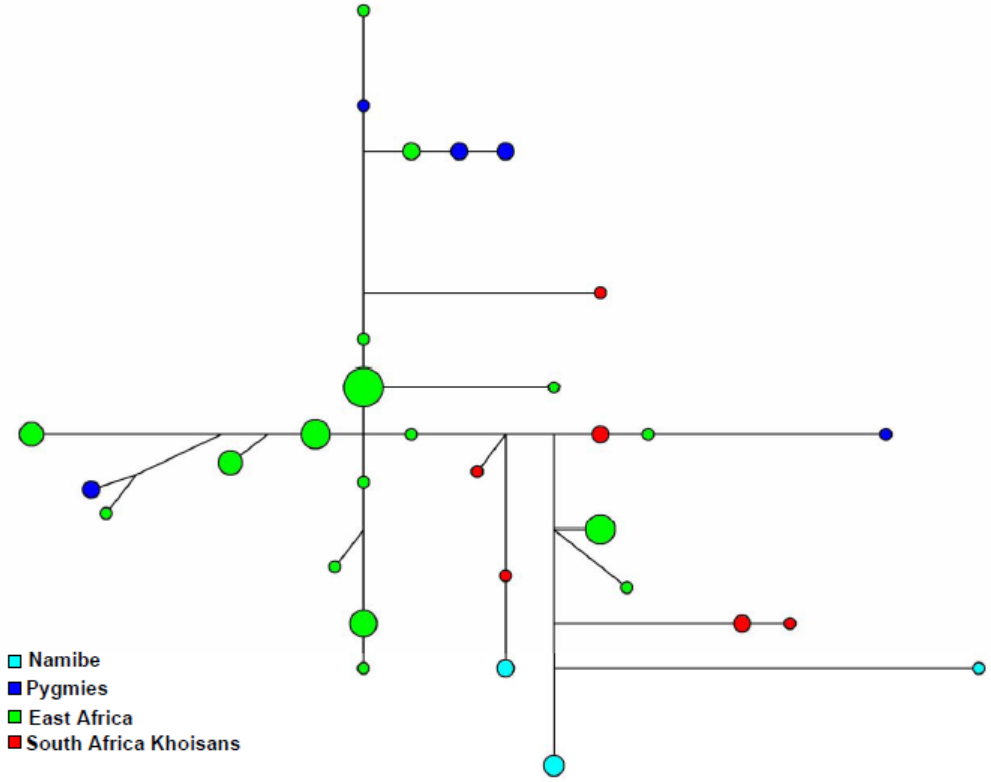
Additional File 8

Median-joining network derived from African HVS-I mtDNA sequences belonging to haplogroup L0d



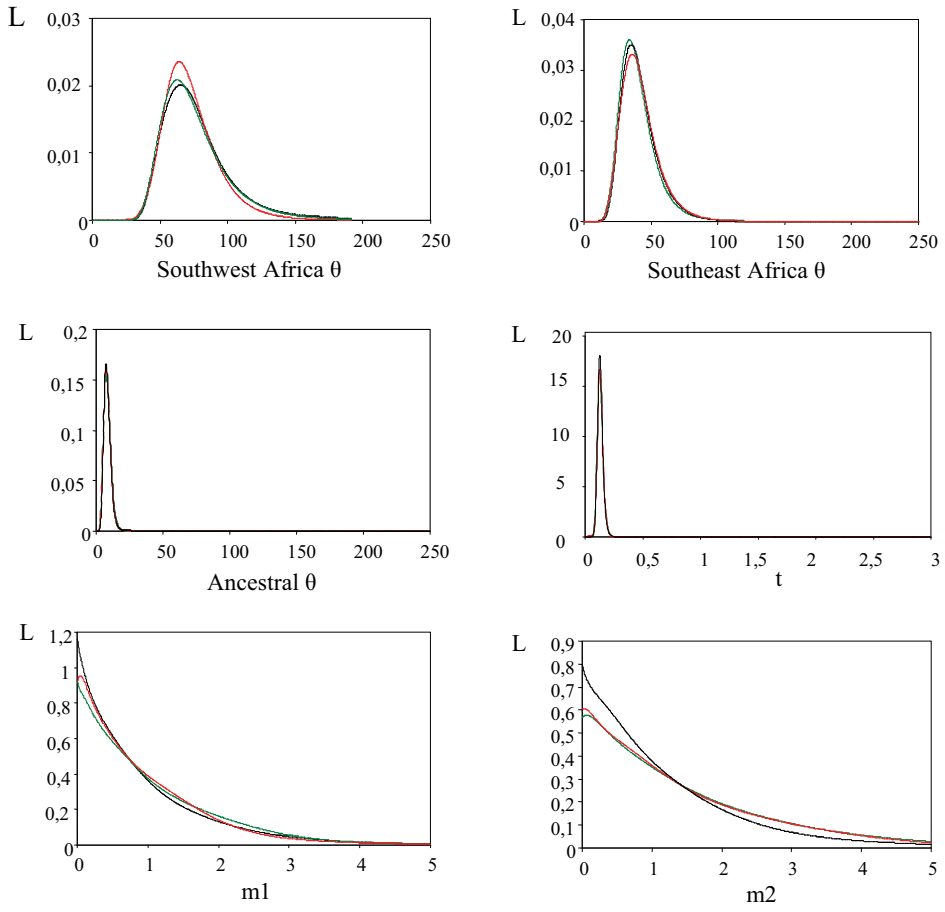
Additional File 9

Median-joining network derived from African Y-chromosome STR-haplotypes belonging to haplogroup B2b



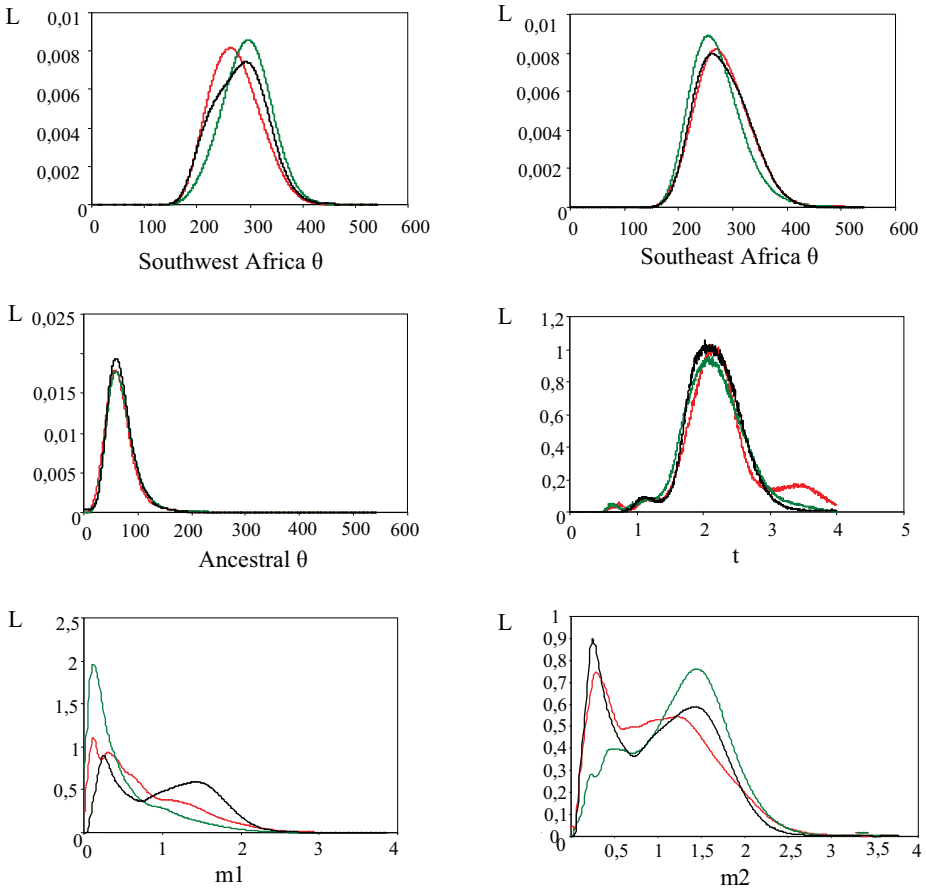
Additional File 10

Probability densities for the basic demographic parameters of the IM model (Y- chromosome STR haplotype dataset)



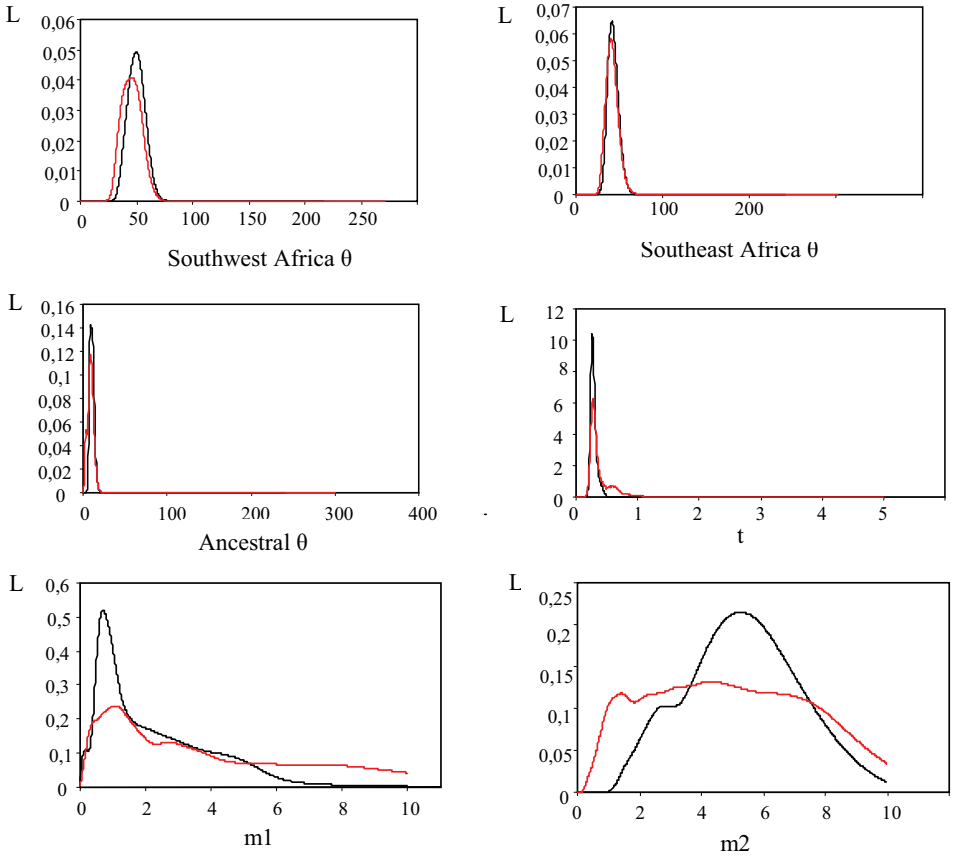
Additional File 11

Probability densities for the basic demographic parameters of the IM model (mtDNA HVS-I sequence dataset)



Additional File 12

Probability densities for the basic demographic parameters of the IM model (joint mtDNA and Y- chromosome datasets)



Article 3

Alves, I, M. Coelho, C. Gignoux, A. Damasceno, A. Prista and J. Rocha. 2010. Genetic homogeneity across Bantu-speaking groups from Mozambique and Angola challenges early split scenarios between East and West Bantu populations. *Hum Biol* (in press).

Genetic homogeneity across Bantu-speaking groups from Mozambique and Angola challenges early split scenarios between East and West Bantu populations

Isabel Alves[1]; Margarida Coelho[1,2]; Christopher Gignoux[3]; Albertino Damasceno[4]; António Prista[5]; Jorge Rocha[1,2]

1. IPATIMUP, Instituto de Patologia e Imunologia Molecular da Universidade do Porto, Portugal.
2. Departamento de Biologia, Faculdade de Ciências, Universidade do Porto, Portugal
3. Institute for Human Genetics, University of California at San Francisco, USA.
4. Faculdade de Medicina, Universidade Eduardo Mondlane, Moçambique.
5. Faculdade de Educação Física e Desporto, Universidade Pedagógica, Moçambique

Correspondence: Jorge Rocha
IPATIMUP
R. Dr. Roberto Frias s/n
4200-465 Porto
Portugal
Email: jrocha@ipatimup.pt

KEY WORDS: UEPSTRs, Bantu expansions, Angola, Mozambique

Abstract

The large scale spread of Bantu-speaking populations remains one of the most debated questions in African population history. In this work we studied the genetic structure of 19 Bantu-speaking groups from Mozambique and Angola using a multilocus approach based on 14 newly developed compound haplotype systems (UEPSTRs), each consisting of a rapidly evolving short tandem repeat (STR) closely linked to a unique event polymorphism (UEP). We compared the ability of UEPs, STRs and UEPSTRs to document genetic variation at the intercontinental level and among the African Bantu populations, and found that UEPSTR systems clearly provided more resolution than UEPs or STRs alone. The observed patterns of genetic variation revealed high levels of genetic homogeneity between major populations from Angola and Mozambique, with two main outliers: the Kuvale from Angola and the Chopi from Mozambique. Within Mozambique, two Kaskazi–speaking populations from the far north (Yao and Mwani) and two Nyasa-speaking groups from the Zambezi River basin (Nyungwe and Sena) could be differentiated from the remaining groups, but no further population structure was observed across the country. The close genetic relationship between most sampled Bantu populations is consistent with high degrees of interaction between peoples living in savanna areas located to the south of the rainforest. Our results highlight the role of gene flow during the Bantu expansions and show that the genetic evidence accumulated so far is becoming increasingly difficult to reconcile with widely accepted models postulating an early split between eastern and western Bantu populations.

Introduction

The dispersal of Bantu-speaking agriculturalists across sub-Saharan Africa provides one of the most striking examples of long-range human migrations. Although it is generally accepted that Bantu expansions started 4,000-5,000 years ago in the area between Cameroon and Nigeria (Newman 1995), there is still no consensus about many aspects of the history of Bantu populations, including the major dispersal routes followed by Bantu speakers and the nature of the interactions between spreading populations. Briefly, current views about Bantu expansions based on archeological and linguistic data can be divided into two main models. According to the most widely accepted model, the Bantu dispersals involved an early population split into two major routes leading to the separation of East and West Bantu primary language branches (Newman 1995; Holden 2002): one following an eastern path, first circumventing the rainforest to the area of the Great Lakes, and then proceeding to Southeast Africa; the other moving south, through the rainforest, into the arid steppes of Southwest Africa. The alternative model challenges the early split between West and East Bantu daughter communities, proposing a single passage through the rain forest, followed by a later spatial divergence in subequatorial Africa (Ehret 1998; 2001; Rexová et al. 2006). According to this view, most Bantu languages spoken to the south and east of the rainforest should be included in a single "Savanna-Bantu" group (Ehret 2001; Rexová et al. 2006).

So far, most studies on the genetic structure of Bantu-speaking populations have focused almost only on Y chromosome and mitochondrial DNA variation (e.g, Salas et al. 2002; Pereira et al. 2002). Although these two uniparentally inherited markers are highly informative, they represent a relatively small fraction of the total genome, providing only limited insights into the history of the human populations. To benefit from the evolutionary information from other regions of the genome, multilocus approaches based on several independently evolving genetic systems are clearly needed.

Unique event polymorphisms (UEPs) and short tandem repeats (STRs) are the two major types of molecular markers currently used to generate multilocus genotype data in large scale surveys of human genetic variation (Rosenberg et al. 2002; Li et al. 2008). Because of their contrasting mutational properties, the two marker types provide complementary information about different aspects of population history (Harpending et al. 1996; de Knijff 2000). UEPs, including single nucleotide polymorphisms (SNPs) and biallelic deletion/insertion polymorphisms (DIPs), are unlikely to experience recurrent mutation and are effective in recording ancient demographic events. However, individual UEP loci have low resolution and very large marker sets are required to study recent population divergence (Rosenberg et al.

2003). Moreover, UEPs are usually affected by ascertainment bias, which may distort the patterns of genetic variation (Mountain and Cavalli-Sforza 1994; Rogers and Jorde 1996). Faster evolving STRs are generally less affected by ascertainment bias and offer more power than UEPs to assess the evolutionary history of closely related populations (Rogers and Jorde 1996; Rosenberg et al. 2003). However, repeat size homoplasy due to convergent mutation obscures the causes of allele sharing and may lead to underestimation of population structure, especially in cases of ancient population divergence (Flint et al. 1999; Estoup et al. 2002).

Different empirical and theoretical studies have shown that the combination of UEPs and STRs in compound autosomal haplotype systems (UEPSTRs) may counterbalance the limitations of each marker type and maximize their specific advantages (Ramakrishnan and Mountain 2004; Hey et al. 2004; Payseur and Cutter 2006). Moreover, Mountain et al. (2002) have developed a general approach to identify and genotype multiple UEPSTR loci to be used in human evolutionary genetics. However, to our knowledge, the use of multilocus UEPSTR marker sets to address specific population history questions is still not widespread.

Here, we use a newly developed set of 14 UEPSTR systems to provide novel insights into the history of Bantu-speaking populations by focusing on the levels and patterns of genetic variation in 19 Angolan and Mozambican groups, encompassing areas that have been poorly sampled in previous studies of African genetic diversity. We confirmed the usefulness of the compound UEPSTR approach to assess genetic variation and evaluated the degree of genetic differentiation between Angola and Mozambique in the context of the split between western and eastern Bantu dispersal routes. In addition, we explored a countrywide population sample from Mozambique to examine the genetic implications of the spread of Bantu speakers across Southeast Africa. Our observation of low levels of genetic divergence across the studied groups is difficult to reconcile with an early split between East and West Bantu-speaking groups, favoring dispersal models that emphasize the close relationship between populations from western and eastern savanna areas located southerly to the African rainforest.

Material and Methods

Finding and ascertainment of UEPTR haplotype systems

Based on the criteria proposed by Mountain et al. (2002), we developed a battery of 14 independently evolving non-recombining autosomal UEPSTRs with widespread chromosomal locations that could be typed with rapid cost-efficient protocols. We focused our efforts on systems consisting of one DIP tightly linked to one STR (hereafter called DIPSTRs), because allele specific PCR required for determination of UEPSTR haplotypic phase (see below) has a higher success rate for DIPs than for SNPs. Taking advantage of publicly available databases, we extracted lists of STRs and DIPs by scanning the STR database from the National Cancer Institute (http://grid.abcc.ncifcrf.gov/str.php; Collins et al. 2003) and the short biallelic indel database from the Marshfield Center for Medical Genetics (http://www.marshfieldclinic.org/mgs/; Weber et al. 2002), respectively. To ensure variation, only STR sequences with more than 6 repeats were chosen from the database. We next searched for DIPs located within 500bp upstream or downstream of each STR, to minimize recombination between DIP and STRs (Mountain et al. 2002; Ramakrishnan and Mountain 2004), and to generate DIPSTR amplicon sizes easy to type that were suitable for multiplex analysis. Selected STRs were further confirmed to be polymorphic using a set of 10 European and 10 African individuals. The chromosomal location, STR repeat motif, indel sequence and DIP derived alleles for each DIPSTR are shown in Table 1.

Genohaplotyping and structure characterization of DIPSTRs

Genohaplotyping of DIPSTRs, which comprises the empirical determination of an individual's pair of haplotypes at each DIPSTR, was carried out in a similar way to that primarily introduced by Mountain et al. (2002), using allele specific PCR, in three PCR multiplex reaction systems. To increase genotyping efficiency, we implemented the Universal Tail approach proposed by Schuelke (2000), which allowed for the same fluorescent primers to be used in a large number of systems by adding FAM and VIC labels to universal tails at the 5' end of indel allele-specific unlabeled primers. Deletion and insertion alleles at DIPs were identified by FAM and VIC fluorescence, respectively, while STR alleles were identified by the fragment length. Genohaplotypes were determined on an ABI3130 automated sequencer. Detailed protocols are available upon request. To characterize the structure of each DIPSTR system and accurately determine the number of STR repeats corresponding to different allele specific PCR fragments, we sequenced a set of diverse STR alleles that were cloned into the

pCR4 plasmid vector using the TOPO TA cloning kit (Invitrogen). Sequencing was performed using the Big Dye Terminator v3.1 cycle sequencing kit (Applied Biosystems).

Table 1: Characteristics of the DIPSTR systems

DIPSTR identification[a]	Chromosome	Position[b]	DIP[c]	DIP derived allele[d]	Repeat motif[e]
MID1777	3	12,204,516	[TC]	DEL	(TG)n
MID2078	3	116,101,902	[CATAT]	DEL	(AT)n(AC)m
MID1590	4	88,278,825	[AAT]	INS	(GT)n
MID999	5	110,089,776	[AA]	DEL	(CT)nTT(CA)m
MID1013	5	126,870,938	[CCAG]	DEL	(GT)n
MID473	6	95,295,850	[TTACATTT]	INS	(AGGA)n
MID1739	6	79,080,937	[GTCAGG]	INS	(TG)n
MID592	6	56,123,516	[AT]	DEL	(CA)n
MID1073	7	29,210,403	[ACAA]	DEL	(TC)n(TG)m
MID2500	10	124,527,156	[AGA]	DEL	(AC)n
MID2170	11	91,479,705	[ATC]	INS	(TG)n
MID2563	14	96,894,098	[GATG]	DEL	(TAGA)n
MID681	14	39,838,079	[GTCA]	INS	(GT)n
MID1827	15	51,163,404	[GT]	DEL	(ATT)n

a. According to Indel identification at the Marshfield Center for Medical Genetics;
b. Chromosomal position according to UCSC genome browser (hg 18; Build 36.1 assembly);
c. Indel polymorphic sequence;
d. INS=insertion; DEL= deletion, according to the Marshfield Center database;
e. Short tandem repeat structure recovered from the cloning procedure (see Material and Methods).

Population samples

Buccal swabs were collected from 286 Mozambican, 104 Angolan and 50 Portuguese individuals, after informed consent. The Portuguese sample was included in the study to contrast patterns of genetic variation between African and European populations. Since Angola and Mozambique were Portuguese colonies until 1975, this sample also provides an appropriate reference to detect the European genetic impact in the African populations. The Angolan sample includes 50 Ovimbundu and 54 Herero-speaking Kuvale individuals, representing Bantu-speaking populations from the southwestern edge of the Bantu language distribution, belonging to the West-Savanna sub-branch of Savanna Bantu, according to the classification proposed by Ehret (1998, 2001; see also Tishkoff et al. 2009). The Ovimbundu form the largest ethnolinguistic group from Angola, making up 35% of the country's total population. The Kuvale, like other Herero-speaking groups, are semi-nomadic cattle-herders, ranking among the most exclusively pastoral peoples of southwestern Africa. All samples from Angola were collected in the southwestern province of Namibe (Figure 1A), as previously described (Coelho et al. 2009), with the approval and collaboration of the local Provincial Health Department. The sample from Mozambique is countrywide and includes 17 population groups, representing most of the country's ethnolinguistic diversity (Figure 1B). All languages belong to the narrow East, or Mashariki, cluster of East-Savanna Bantu, according to Ehret's classification. Following this classification, Mozambican populations can be further subdivided on the basis of linguistic criteria. The Yao (n = 20) and the Mwani (n = 11), from northern Mozambique (Figure 1B), speak languages included in the Kaskazi, or northern, cluster of East Bantu, which extends to regions as far north as Tanzania and Kenya. Within Kaskazi, Yao belongs to the Rufiji-Ruvuma subgroup, while Mwani, a language close to Swahili, would be placed among the North-East Coastal subgroup. All other Mozambican populations speak languages from the southern, or Kusi, East Bantu cluster, which may be additionally subdivided into the following groups: i) Makua (Makua, n = 20 ; Makua-Meeto, n = 20; Lomwe, n = 19; Marenje, n = 10 and Chwabo, n = 10); ii) Nyasa (Nyanja, n = 19; Nyungwe, n = 21 and Sena, n = 20); iii) Shona (Tewe, n = 14 and Ndau, n = 20); and iv) South-East Bantu (Shangaan, n = 19; Tswa, n = 22; Ronga, n = 15, Gitonga, n = 14; Chopi, n = 12) (Figure 1B). Mozambican samples were collected with the approval and collaboration of Pedagogic and Eduardo Mondlane Universities of Mozambique.

Cryptic relatedness between pairs of individuals was inferred from the genetic data using the RELPAIR program (Epstein et al. 2000). Based on this analysis, we excluded from the

study six individuals belonging to pairs inferred to be related at a level equal or closer to first cousins: 1 Portuguese, 1 Kuvale, 1 Makua and 3 Nyungwe.

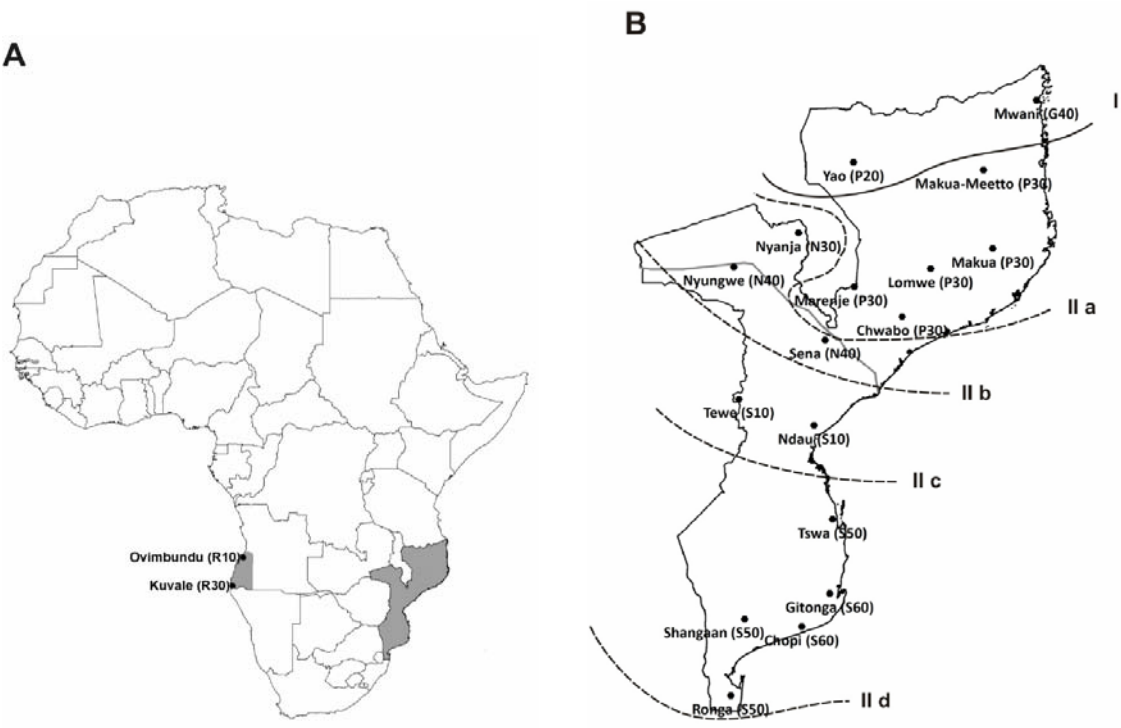


Figure 1. Geographic location of the populations analyzed in the study. A) Map of Africa with sampled regions shown in grey. B) Map of Mozambique with the geographic distribution of sampled ethnolinguistic groups. The solid line (I) separates Kaskazi-speaking populations (Yao and Mwani) from Kusi-speaking populations (all others). Dashed lines delimitate four Kusi-speaking sub-groups: IIa, Makua; IIb, Nyasa; IIc, Shona; IId, Southeast Bantu (see text for more details. Guthries's alphanumeric codes (Guthrie 1967-1971) are shown between parentheses. The grey line between regions IIa and IIb represents the Zambezi River.

Data analyses

Within-population variability

Allele frequencies per locus per population and expected heterozygosity (*H*) were calculated using the ARLEQUIN 3.11 software package (Excoffier et al. 2005). The 95% confidence intervals (CIs) of *H* were constructed by the bootstrap method implemented in the software Genetix 4.0 (Belkir et al. 1998). STR heterozygosity within the deletion and insertion

DIP backgrounds was calculated using the REJSTATS software, included in the REJECTOR package (Jobin and Mountain 2008).

Apportionment of genetic variation and pairwise distances

Hierarchical analysis of molecular variance (AMOVA) was performed using ARLEQUIN 3.11. When considering only DIPs or the combined DIPSTR data, the analyses were performed using conventional F-statistics, without taking into account the molecular divergence between alleles. For the DIPSTR data this means that STR allele sizes were ignored and all distances between different DIPSTR haplotypes were considered identical. When using the STR data alone, we used both conventional F-statistics and the Rst genetic distance (Slatkin 1995), which quantifies the divergence between STR alleles taking into account the probability of recurrent mutation.

Genetic structure was additionally explored through the spatial analysis of molecular variance implemented in the SAMOVA v.1.0 software (Dupanloup et al. 2002), which defines groups of populations that are geographically homogeneous and maximally differentiated from one another, without requiring the *a priori* definition of the number of populations to include in a given group.

Pairwise genetic differentiation among populations was evaluated through Fst genetic distances (Reynolds 1983) calculated using ARLEQUIN and visualized in a multi-dimensional scaling (MDS) plot using the STATISTICA 7.0 software (Stat. Soft, Inc. 2004). The significance of the stress value was evaluated according to Sturrock and Rocha (2000). Mantel tests, also implemented in ARLEQUIN, were performed to assess the correlation among genetic, linguistic and geographic pairwise distances in Mozambique. Geographic distances between populations were calculated using Great Circle Distances (www.gb3pi.org.uk/great.html) based on latitude and longitude data for the sample sites. Linguistic distances (d) were calculated in arbitrary units as follows: d = 0 for populations sharing the same language; d = 1 for populations speaking different languages sharing the same Guthrie code (Guthrie M.1967-1971); d = 2 for populations speaking languages with different Guthrie codes sharing the same subgroup within the Kusi or Kaskazi clusters; d = 3 for populations with languages from different subgroups within Kusi or Kaskazi; d = 4 for populations speaking languages from Kaskazi and Kusi clusters.

Genetic clustering of individuals

We investigated the presence of genetic clusters using the Bayesian approach implemented in the STRUCTURE software package, version 2.3.3 (Pritchard et al. 2000; Falush et al. 2003), assuming that individuals may have mixed ancestry (admixture model) and a model of correlated allele frequencies (F model). We performed 20 independent runs for a number of clusters (K) ranging from 1 to 3, with 600,000 iterations after a burn-in of length 200,000. To identify groups of runs with similar clustering patterns (modes) for each K value, structure outputs were processed using the Greedy algorithm implemented in CLUMPP (Jakobsson and Rosenberg 2007). Individual genotype membership proportions were averaged across runs within the same mode and subsequently plotted using DISTRUCT1.1 (Rosenberg 2004).

Inference of population-history parameters

To infer basic population-history parameters underlying the split between Bantu populations dispersing along western and eastern routes in Africa, we analyzed the STR-only data using the approximate Bayesian computation (ABC) approach (Beaumont et al. 2002) implemented in the software REJECTOR (Jobin and Mountain 2008). In the ABC framework, gene genealogies are generated through coalescent simulations under specified models of population history whose parameter values are drawn from prior distributions. Parameter values that produce simulated summary statistics sufficiently close to those observed in the experimental data are retained to construct posterior distributions. We explored 4 models of population split between Mozambique and Angola, assumed to represent the eastern and western edges of Bantu dispersals: in model 1 (no migration and no growth) two populations, with constant effective sizes (N_{Ang} and N_{Moz}) are assumed to have split t generations ago, remaining isolated from each other, and from third party populations; model 2 (growth without migration) is similar to model 1 but the two populations are allowed to growth exponentially, with rates r_{Ang} and r_{Moz}, after divergence; model 3 (migration without growth) is similar to model 1, but allows for bidirectional migration between the two populations at constant rates, m_{AngMoz} and m_{MozAng}; model 4 (migration and growth) is the most parameter rich and adds population growth to model 3. Genetic datasets were simulated by drawing the parameters of each model from the following uniform prior ranges: N_{Ang} and N_{Moz} (5,000-20,000 individuals); t (1-2,000 generations); m_{AngMoz} and m_{MozAng} (0.005-0.000005, fraction of individuals from one population migrating to the other per generation); r_{Ang} and r_{Moz} (0-0.05, exponential growth rate per generation). The Kuvale sample was excluded from the analyses because the

documented admixture between this population and Khoisan peoples (Coelho et al. 2009) violates the assumption of no third party migration, common to both population-history models. Based on tests of inferential power performed by Jobin and Mountain (2008), the following 5 summary statistics were calculated: variance in STR repeat number (*V*; averaged across loci) and expected heterozygosity (*H*; averaged across loci), to describe the genetic diversity within Mozambique and Angola; and $\delta\mu^2$ (Goldstein et al. 1995), Dsw (Shriver et al. 1995) and Nei's minimum genetic distance (Nei 1987), to assess the extent of differentiation between the two populations. The presented results are based on 10^6 iterations of REJECTOR and only simulations generating summary statistics within 1% of the observed value were accepted (tolerance level of 0.01).

In currently available versions of REJECTOR, STR mutation is modeled under a strict stepwise mutation model (SMM) with fixed mutation rates that cannot be estimated in the ABC inference procedure. To calculate per locus mutation rates, we used the homozygosity-based estimator proposed by Xu and Fu (2004), which is relatively robust to deviations from pure SMM and departures from mutation-drift equilibrium (Xu et al. 2005). Interestingly, our estimate of the average mutation rate (2.27×10^{-4}) is close to other calculations ($2,4\text{-}2.45 \times 10^{-4}$) based on ABC methods (Verdu et al. 2009, Wegmann et al. 2009).

Results

Properties of markers with different mutation rates

We compared the properties of DIPs (STR blind), STRs (DIP blind) and joint DIPSTR datasets by considering three major aspects of the data providing information about ascertainment bias, homoplasy levels and power to detect population structure.

Patterns of within-population diversity

Figure 2A presents estimates of average heterozygosity at the DIP-only, the STR-only and the joint DIPSTR datasets in 19 Bantu-speaking populations from Angola and Mozambique and one European sample from Portugal. In accordance with the well-known trend for higher heterozygosity in sub-Saharan African populations (Garrigan and Hammer 2006), the STR-

only dataset revealed greater genetic diversity in Africa than in Europe. On the contrary, no excess African heterozygosity was detected with the DIP set, suggesting that DIPs may be affected by ascertainment bias (Mountain and Cavalli-Sforza 1994). However, the DIP bias is not reflected in the joint DIPSTR dataset, which clearly showed higher heterozygosity levels in Africa (Figure 2A). Figure 2B shows that, as expected, average STR heterozygosity values on the background of ancestral DIP alleles are significantly higher than on the background derived alleles ($P < 0.001$; one tail paired Wilcoxon signed rank test).

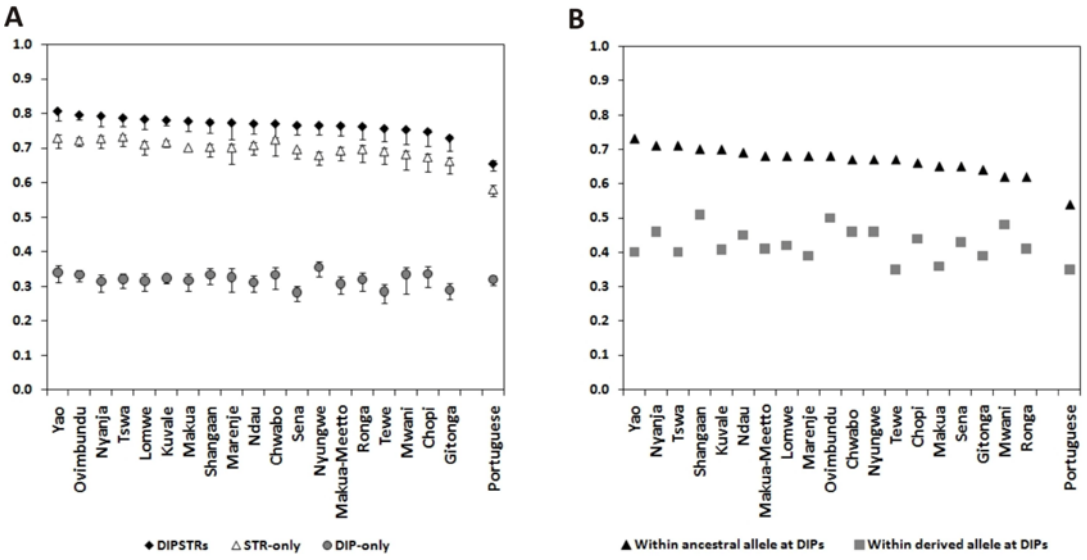


Figure 2. A) Population average heterozygosities (*H*) and 95% CIs calculated with the DIP-only (grey circles), the STR-only (white triangles) and the joint DIPSTR (black diamonds) datasets. Populations are ranked by declining heterozygosity values at the DIPSTR dataset. B) STR average heterozygosity values calculated on the background of ancestral (black triangles) and derived (grey squares) DIP alleles. Populations are ordered according to decreasing values of diversity within ancestral DIP alleles.

In order to informally evaluate the degree of overlapping between STR allele distributions on the background of different DIP alleles, we calculated:

$$H_{excess} = \frac{H_{DIPSTR} - H_{STR}}{H_{DIPSTR}} \qquad (1)$$

which is indirectly related to the amount of homoplasy in STRs, and where H_{DIPSTR} and H_{STR} are heterozygosity estimates at DIPSTR and STR datasets, respectively. The H_{DIPSTR} values were found to be significantly higher than H_{STR}, reflecting the ability of DIPs to split allele size overlaps caused by homoplastic mutations at linked STRs ($P < 0.001$; one tail paired Wilcoxon signed rank test; Figure 2A). However, only a fraction of the STR homoplasy is revealed by DIPSTRs, since convergent STR size mutations within each DIP allelic class, or in different repeat tracts of imperfect STRs, will remain undetected. Figures 3A and 3B display H_{excess} values at the 14 DIPSTR loci, in a pooled African sample including all Bantu-speaking populations and in the Portuguese population, respectively. Figures 3C and 3D illustrate the difference in extent of allele size overlap at two STR loci (MID1073 and MID1013, Table 1) with clearly distinct H_{excess} values.

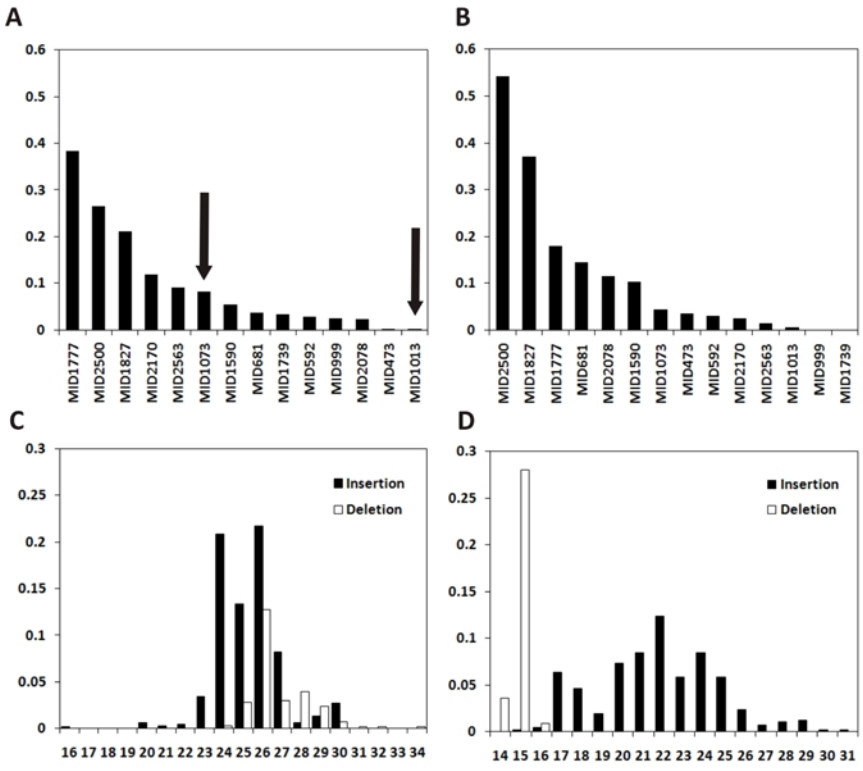


Figure 3. A) and B) Levels of STR allele size overlap measured by H_{excess} in Bantu-speaking populations (A) and in the Portuguese population (B). Arrows in panel A denote the two STRs presented in panels C and D. C) and D) Distributions of STR allele frequencies at DIPSTRs MID1073 (C) and MID1013 (D) in the African sample. STR allele sizes, in number of repeats, are shown on the *x* axes. STR alleles linked to a deletion or insertion are shown as white and black bars, respectively.

Levels of intercontinental differentiation

To assess the ability of different marker sets to describe population divergence at the intercontinental level, we performed an AMOVA analysis dividing the sampled populations into one group including all 19 Bantu-speaking populations from Angola and Mozambique, and one group including the Portuguese population (Table 2). The among-group proportion of genetic variance calculated with the DIP-only data (27.1%; $P < 0.001$) was very high when compared with previous studies (Barbujani et al. 1997), even considering that only African and European population groups were contrasted (Tishkoff and Kidd 2004). This inflation was essentially due to loci MID473, MID1739 and MID1073, in which absolute differences in deletion/insertion allele frequencies in the African and the European samples were found to be very high (0.60, 0.63 and 0.59, respectively). After removal of the three loci, the DIP-only estimates (13.79%; $P < 0.001$) were clearly more concordant with previous results (Barbujani et al. 1997; Tishkoff and Kidd 2004) and with estimates based on the STR (11.93%-12.6%; $P < 0.001$) and the DIPSTR datasets (12.10%; $P < 0.001$). These results provide further indication that the DIP-only set may be affected by ascertainment bias, not only obscuring the excess African diversity but also favoring the differentiation between African and European populations.

Table 2. Analysis of molecular variance (AMOVA) in European and African samples

			Variance components in percentage		
	No. populations	No. groups[a]	Among groups	Among populations within groups	Within groups
DIPs	20	2	27.10**	0.25	72.64
STRs (Fst)[b]	20	2	12.19**	0.72**	87.09
STRs (Rst)[c]	20	2	17.46**	0.75	81.79
DIPSTRs	20	2	12.50**	0.91**	86.59

*P<0.05; **P<0.001

a. The European sample was treated as one group and all the remaining 19 populations from Mozambique and Angola were included in another group.
b. Using the Fst distance, without accounting for the molecular divergence among STR alleles.
c. Using the Rst distance to account for the molecular divergence among STR alleles.

Individual clustering

We evaluated the power of the different marker sets to detect population structure by using the individual clustering algorithm implemented in STRUCTURE (Figure 4). For K = 2 (Figure 4A), all independent runs at each dataset were included in the same mode. Using the DIP-only dataset, the Portuguese individuals were grouped into a single "White" cluster, with an average proportion of genotype membership of 93% (standard deviation; SD = 3%; Figure 4A). The African samples were less clearly grouped, with only 65% (SD = 22%) of genotypes belonging to the alternative "Black" cluster (Figure 4A). The STR and DIPSTR datasets clearly improved the assignment of African samples to the "Black" cluster, whose individuals had average proportions of genotype membership of 96% (SD = 8%) and 97% (SD = 5%), for STRs and DIPSTRs, respectively. These clustering patterns indicate that, despite the longstanding Portuguese colonization of Angola and Mozambique, the signals of admixture between Europeans and Africans are negligible. For K = 3, independent runs at each dataset were grouped in different modes. Figure 4B displays the averaged clustering patterns across runs belonging to the mode with the highest average probability of observing the data (Ln(P)). In all datasets, African samples were split into two clusters (Figure 4B, in black and grey). At the DIP-only dataset, the mode with highest overall likelihood (Ln(P) = -6,214; SD = 42.92; Figure 4B) included 16 out of 20 runs, and presented a fairly symmetric clustering pattern within African populations, with individual membership proportions in the "Grey" and "Black" clusters of 46% (SD = 9% and 8%, respectively). Two additional modes included 3 and 1 runs and gave qualitatively similar results (not shown). The mode with highest overall likelihood (Ln(P) = -19,206; SD = 55.3) for the STR-only data included 5 out of 20 runs and also presented symmetric clustering patterns within Africans, with individual membership proportions in the "Grey" and "Black" clusters of 48% (SD = 17%) and 49% (SD = 17%), respectively (Figure 4B). Average membership proportions remained practically unchanged in an additional mode, including the remaining 15 runs, although SD values were considerably lower (8%). Using the DIPSTR data, the mode with highest average probability of observing the data (Ln(P) = -22,667; SD = 55.3) contained 2 out of 20 runs. In this mode (Figure 4B) two populations could be distinguished from the other African populations: the Chopi from Mozambique, presenting higher membership proportions in the "Black" cluster (73%; SD = 14% vs 49%; SD = 24%); and the Kuvale from Angola, showing increased proportions of membership in the "Grey" cluster (76%, SD = 18% vs 44%, SD = 22% Figure 4B). In a second mode (not shown), including 11 runs with lower average likelihood (Ln(P) = -22,680; SD = 18.3), the Kuvale and the Chopi had less asymmetric clustering patterns, but still

displayed increased membership proportions in the "Grey" (64%; SD = 11%) and the "Black" clusters (63%; SD = 8%), respectively. Finally, in a third mode, including 7 runs (Ln(P) = -22677; SD = 8.7), clustering patterns of Chopi and Kuvale could not be differentiated from the other African populations (not shown). Taken together, the results for K = 3 using the DIPSTR data, provide signals of an increased differentiation of the Chopi and the Kuvale, and indicate that compound DIPSTR haplotypes are more informative than DIPs and STRs alone to detect population structure. In the following, we focus on the use of DIPSTR data to characterize the patterns of genetic variation in the sampled African populations.

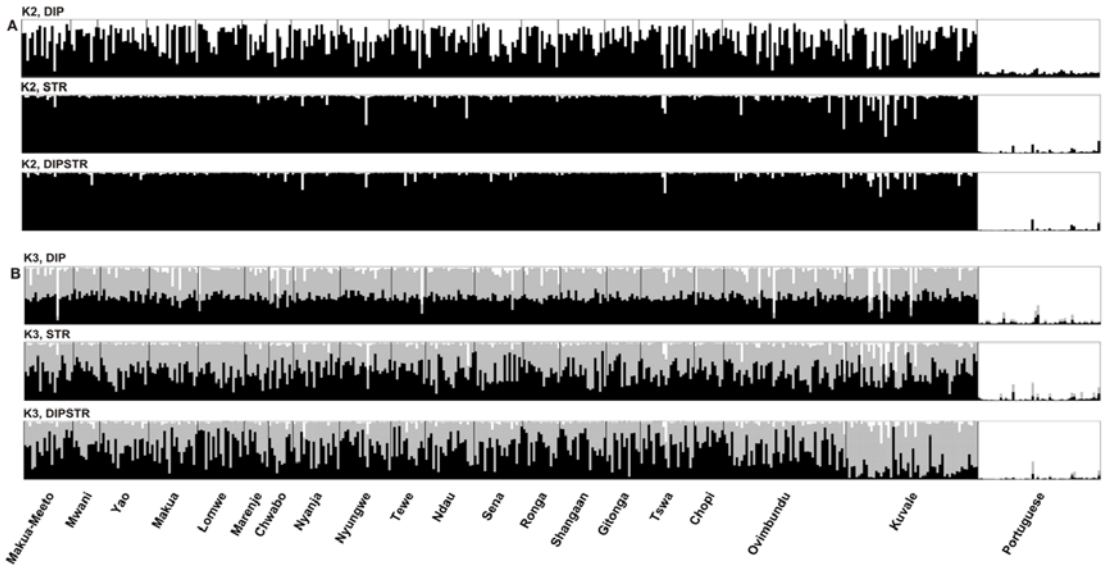


Figure 4. A**)** Distributions of individual genotype membership proportions in each cluster at K = 2 based on the DIP-only, the STR-only and the DIPSTR datasets in 434 individuals from 19 Bantu-speaking groups and one European (Portuguese) population. B) Proportions of genotype membership at K = 3. Each individual is represented by a vertical line partitioned in K segments representing the individual's estimated genotype proportions. Thin black lines separate individuals from different populations.

Genetic structure of Bantu-speaking populations

Apportionment of genetic variation

In order to examine how the genetic diversity was distributed in the 19 Bantu-speaking populations, several AMOVA analyses were performed including all populations and also considering different groupings based on language or geography (Table 3). Significant

fractions of the total variance were found when all Bantu-speaking populations were considered as a single group (0.91%; P < 0.001); among the two Angolan samples (1.28%; P < 0.001); and among the 17 Mozambican populations (0.60%; P < 0.001). The level of genetic differentiation between Angola and Mozambique (0.50%; P < 0.05) was found to be smaller than between populations within each country (0.69%; P < 0,001), suggesting that the sampled genetic variation is not primarily structured across these two regions of subequatorial Africa.

Table 3. Analysis of molecular variance (AMOVA) in African populations using DIPSTR systems

Groups	No. populations	No. groups	Among groups (%) DIPSTRs	Among populations within groups (%) DIPSTRs	Within populations (%) DIPSTRs
All African populations	19	1		0.91**	99.09
All Mozambican populations	17	1		0.60*	99.40
All Angolan populations	2	1		1.28**	98.72
Mozambique and Angola	19	2	0.50*	0.69**	98.80
Mozambique					
Northern and Southern Zambezi	17	2	-0.04	0.62*	99.42
Northern, Central and Southern provinces[a]	17	3	0.04	0.58*	99.38
Kaskazi and Kusi	17	2	0.58*	0.48*	98.94
Makua, Nyassa, Shona and South-East Bantu	15	4	0.12	0.37	99.51
G40, P20, P30, N30, N40, S10, S50, S60[b]	17	8	0.31*	0.33	99.36

*P<0.05; **P<0.001.

a. The northern group includes Cabo Delgado, Niassa and Nampula provinces; the central includes Zambezia, Tete, Manica and Sofala, and the southern one consists of Gaza, Inhambane and Maputo.
b. Guthrie codes (Guthrie M.1967-1971).

To ascertain how genetic variation was partitioned in Mozambique, the 17 populations from the country were further grouped according to different geographic and linguistic criteria. No significant among-group differences were found across the major Zambezi River (Figure 1B), nor across a North/Centre/South geographic segmentation. When populations were grouped according to linguistic criteria, significant differences were found among groups speaking languages from different Guthrie's codes (0.31%; P < 0.05) and among the Kaskazi (Yao and Mwani) and Kusi (all other populations) major linguistic clusters of Mashariki Bantu

(0.58%; P < 0.05). However, applying a Bonferroni correction threshold of 0.05/8=0.00625 to the 8 comparisons involving Mozambican samples (Table 3), the Guthrie code (P = 0.038) and the Kaskazi/Kusi divisions (P = 0.008) did not remain statistically significant.

We next performed a SAMOVA analysis (Dupanloup et al. 2002) in order to additionally identify maximally differentiated groups without using grouping criteria defined *a priori*. Table 4 presents the results of this analysis for a number of groups (K) comprised between 2 and 6. For K > 6, the among-group component of the total variance became artificially inflated, increasing monotonically until K = 19 because of the reduction of the proportion of variance due to differences between populations within each group (see Dupanloup et al. 2002). The genetic pattern revealed by SAMOVA suggests that most Bantu populations form a single, fairly homogeneous cluster from which a series of populations were sequentially separated as K increased. Consistent with the results from the STRUCTURE analysis (Figure 4B), the Chopi and the Kuvale were separated from each other and from the remaining populations at the first significant partition (K = 3; Table 4), suggesting that these two groups are the most extreme outliers. The Mwani and the Nyungwe were separated from the major cluster at K = 4 and K = 5, respectively. At K = 6, the Mwani are joined to the Yao forming a Kaskazi-speaking group, while the Nyungwe are joined to the Sena, forming a Nyasa Kusi-speaking group (see Figure 1B). The Ovimbundu are also separated at K = 6, but remain isolated from the Kuvale, in spite of the linguistic and geographic proximity of these two groups.

Table 4. Spatial analysis of molecular variance (SAMOVA) in African populations using DIPSTR systems

K	Among groups	Among populations within groups	Groupings
2	1.39	0.80**	Chopi/ all other populations
3	1.07*	0.56**	Chopi/ Kuvale/ all other populations
4	1.06**	0.51**	Chopi/ Kuvale/ Mwani/ all other populations
5	1.00**	0.46**	Chopi/ Kuvale/ Mwani/ Nyungwe/ all other populations
6	1.00**	0.17	Chopi/ Kuvale/ Mwani + Yao/ Nyungwe + Sena/ Ovimbundo/ all other populations

*P<0.05; **P<0.001

Pairwise genetic distances

MDS analysis of pairwise Fst genetic distances among all Bantu-speaking populations showed a genetic pattern very similar to the one revealed by SAMOVA (Figure 5). In accordance with the results from STRUCTURE and SAMOVA the Kuvale and Chopi have marginal positions in the MDS plot. The Mwani and the Nyungwe are also separated from the other groups, consistent with SAMOVA at K = 4 and K = 5, respectively. Moreover, the Yao are the nearest genetic neighbors of the Mwani, and the Sena are the nearest neighbors of the Nyungwe, in agreement with the Mwani+Yao and Nyungwe+Sena groupings at K = 6 (see Table 4). The Ovimbundu occupy a relatively marginal position within the major central cluster, consistent with the separation of this group by SAMOVA only at K = 6.

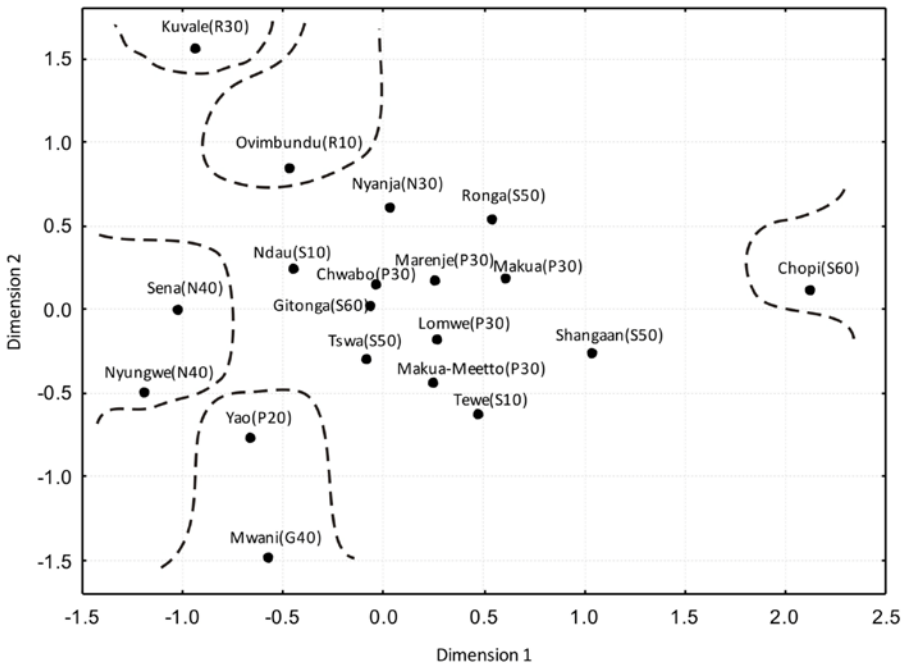


Figure 5. MDS plot based on Fst genetic distances calculated for the joint DIPSTR dataset. Mozambican and Angolan populations are represented by filled and open circles, respectively. Dashed lines delimit population groups isolated by SAMOVA at K = 6. (Stress value = 0.191, under the 1% cutoff value of 0,269 defined by Sturrock and Rocha, 2000).

Relationships between language, geography and genetics

We have performed Mantel tests to compare matrices of genetic, geographic and linguistic distances and assess the influence of geography and language in the shaping of genetic diversity within Mozambique (Table 5). Significant correlations were found between geography and language, reflecting the clear latitudinal segmentation of linguistic groupings across the country, where related languages are, on average, spoken in neighboring regions (Figure 1B). Genetic distance was also significantly correlated with geographical and language distance. However, these correlations were weak and probably represent by-products of the association between language and geography, since they did not remain significant when partial correlations were used.

Table 5. Correlation and partial correlation coefficients between genetic, geographic and linguistic distances in the Mozambican populations[a]

		r value[b]
Correlations	Genetics and Geography	0.27*
	Genetics and Linguistics	0.29*
	Geography and Linguistics	0.45**
Partial correlations	Genetics and Geography, Linguistics held constant	0.16
	Genetics and Linguistics, Geography held constant	0.19

*P<0.05; **P<0.01.
a. Genetic data is based on the DIPSTR set.
b. Correlation or partial correlation coefficient values.

Inferring population history parameters

Figures 6 and 7 display posterior distributions of the parameters describing the 4 models of population history explored with the ABC framework. In spite of using only a limited number of 5 summary statistics and performing 10^6 iterations, we were unable to retrieve more than 50 accepted simulations when all summary statistics were required to simultaneously fall within a 0.01 tolerance level. Moreover, expanding the tolerance level to values of 0.05 or 0.10, did not improve acceptance rates sufficiently for combinations of different subsets of statistics to be

used (not shown). Therefore, we chose to present parameter estimates using the summary statistics separately, based on numbers of acceptances ranging from 500 to ~20,000. A major limitation of not using summary statistics jointly is that no formal choice of the model best fitting the data is possible, since this evaluation involves the comparison of acceptance rates under each model using all summary statistics, for a given tolerance (Pritchard et al. 1999; Cornuet et al. 2008; Verdu et al. 2009). However, our approach may be useful to empirically assess the informative power of each statistic and the degree of redundancy in the data. We found that virtually no information about population size or population growth was captured by the three genetic distances ($\delta\mu^2$, Dsw and Nei's minimum genetic distance). Conversely, *H* and *V* did not convey information on time of divergence and migration. For this reason, posterior distributions for population size and population growth are displayed only for *H* and *V*, while distributions for time of divergence and migration are displayed only for genetic distances (Figures 6 and 7). In models 2 (Figure 6B) and 4 (Figure 7B) the number of acceptances for *V* was less than 70 and the corresponding distributions are not shown. When more than one summary statistic was used to infer a parameter, in some cases all distributions from different statistics were informative and centered around neighboring modes (see, e.g., effective size in Angola, Figures 6A and 7A; divergence time in Figures 6A and 7B). In other cases, some statistics were clearly informative, while the others yielded flat distributions that remained close to the priors (e.g., divergence time in Figure 6B and migration rates in Figure 7A and 7B). We found no instances of different summary statistics providing very different, clearly peaked distributions for the same parameter, suggesting that the loss of computational efficiency associated with joining different statistics would not be compensated by an increase in information. This, of course, does not mean that the used statistics are sufficient, and that other summaries of the data could not improve the estimations.

Estimated divergence times remained congruent across models of population-history, and were in agreement with the notion that Bantu-speaking populations became differentiated only within the last 5,000 years (modal *t* values ~2600-5100 years, Figures 6A; 6B, for Nei's distance and 7B; assuming 25 years per generation). However, the 95% credibility intervals (CI) were generally quite large, ranging from 225 to 47,000 years. In Angola, the posterior distributions for the growth rate showed non-zero, clearly peaked probability distributions in models 2 (growth without migration; Figure 6B) and 4 (migration and growth Figure 7B), with similar 0.002-0.003 modal *r* values (95% CIs ranging from 0.0003-0.0451). The estimates for the Angolan effective population size were close to 11,000 (modal N_{Ang} values ~10,200-12,400; 95% CIs ranging from 5,700-18900), in models 1 and 3 with constant population size (Figures 6A and 7A). In models allowing for exponential growth (model 2, Figure 6B; model

4, Figure 7B) modal N_{Ang} values tended to be larger (modal N_{Ang} values ~15,000-18,000 individuals; 95% CIs ranging from 9,500-19,700). Finally, the posterior distributions for the levels of gene flow had similar non-zero 4-6 x 10^{-4} modes both for m_{AngMoz} and m_{MozAng} in models 3 and 4 (Figures 7A, for Nei's distance and 7B for $\delta\mu^2$) with 95% CIs ranging from 4 x 10^{-5}-5 x 10^{-3}. When estimated population sizes for Angola are taken into account, these values correspond to approximately 4-10 individuals (Nm) migrating from Angola to Mozambique, each generation. Several distributions remained always flat and were not distinguishable from the priors, suggesting that, in these cases, the data did not contain information for inference. These included population size and population growth in Mozambique (Figures 6 and 7), as well as time of divergence in model 3 (Figure 7A).

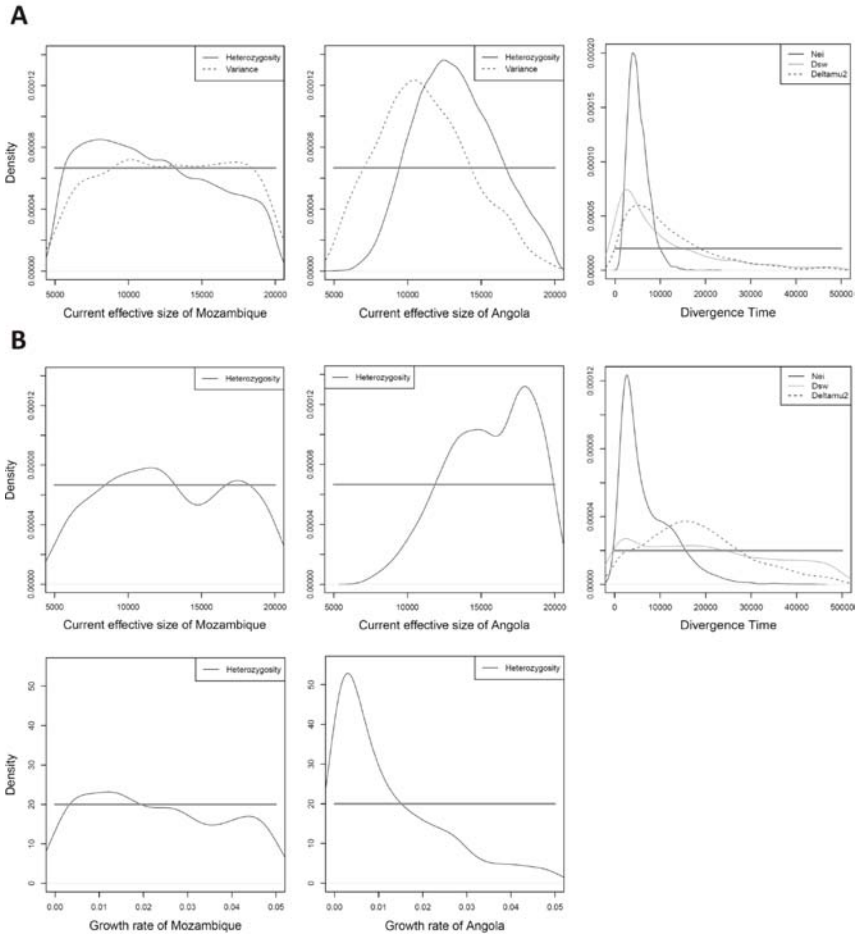


Figure 6. Prior and approximated posterior distributions of population history parameters specifying population-history models 1 (no migration and no growth; panel A) and 2 (growth without migration; panel B). Straight lines represent the uniform distributions and the curves represent Kernel density plots of the approximated posterior distributions obtained with different summary statistics using the ABC analysis implemented in REJECTOR (Jobin and Mountain 2008). For details see text.

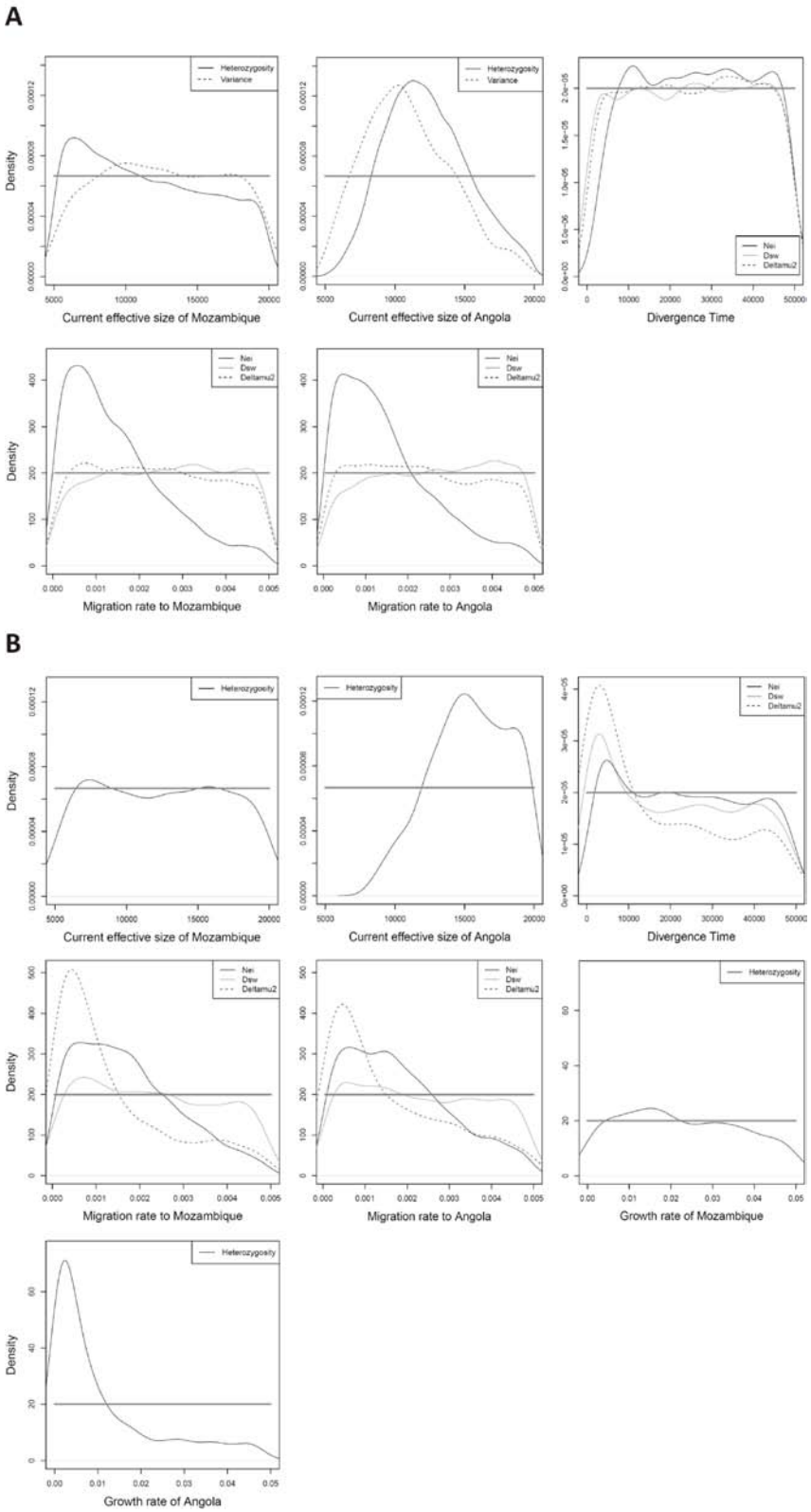


Figure 7. Prior and approximated posterior distributions of population history parameters specifying population-history models 3 (migration without growth; panel A) and 4 (migration and growth; panel B). Straight lines represent the uniform distributions and the curves represent Kernel density plots of the approximated posterior distributions obtained with different summary statistics using the ABC analysis implemented in REJECTOR (Jobin and Mountain 2008). For details see text.

Discussion

Utility of DIPSTR markers

Owing to their contrasting mutational properties, STRs and UEPs offer diverse levels of resolution to document different aspects of population-history events (Payseur and Cutter 2006). However, the validity of most comparisons of genetic patterns at these two marker types is generally hampered by the lack of shared genealogical history among the loci being compared (Payseur and Jing 2009). Since DIPs and STRs share the same genealogical background in fully linked DIPSTRs, our marker set provides an additional opportunity to evaluate the impact of mutational differences without the confounding factor of variation in genealogical history. We did not find strong correlations between measures of population variation and structure at UEPs and STRs and confirmed that STRs are substantially more informative than UEPs to infer population structure. In addition, we could empirically evaluate the specific advantages of using UEPs and STRs in combination.

The most striking differences in the genetic patterns observed at DIPs and STRs were the absence of excess African heterozygosity and the high proportion of variation between Africans and Europeans at the DIP-only dataset (Figure 2A and Table 2). The same trends were recently observed by Romero et al. (2009), who suggested that a large set of DIPs ascertained through the Marshfield Center for Medical Genetics (Weber et al. 2002), was affected by an extreme ascertainment bias. We have selected our DIP markers from the Marshfield database without any further ascertainment criteria besides proximity to STRs and easiness to type. Thus, it is likely that the properties of our 14 DIP dataset reflect original biases in the database. However, we found that these biases could be largely mitigated when DIPs were combined with STRs, as the joint DIPSTR systems displayed a pronounced excess African heterozygosity and showed levels of differentiation between African and European populations that were clearly congruent with those estimated from other datasets (see Figure 2A and Table 2; Barbujani et al. 1997). On the other hand, DIPs provided a noticeable increase in the resolution of linked STRs. For example, heterozygosity was significantly higher at DIPSTRs than at STRs due to the ability of DIPs to distinguish between STR alleles that are identical by state but not identical by descent (see Figures 2A and 3). Moreover, DIPSTRs offered more power than STRs to detect population structure, as shown by the ability to distinguish outlier Bantu populations using STRUCTURE (see Figure 4B).

Taken together, our results highlight the advantages of combining UEPs and STRs in compound UEPSTR systems to explore different aspects of human population history. Although recent advances in genotyping technology have made possible the use of very large

numbers of SNPs, these methods are still too expensive to be applied to most populations (Need and Goldstein 2009). Thus, it is likely that population studies focusing on regional and local aspects of human genetic variation will still depend on carefully chosen marker sets. Since UEPSTRs are relatively simple to develop, a more widespread use of these systems is expected to improve cost/benefit ratios in studies using a limited number of loci to address specific patterns of human genetic variation.

Implications for the history of the Bantu expansions

Patterns of genetic variation

In this study we used a newly developed suite of UEPSTR markers to provide insights into the history of the Bantu expansions by analyzing, for the first time, multilocus genotype data in Bantu-speaking populations from Angola and Mozambique. According to models favoring an early split between Eastern and Western Bantu groups (Newman 1995; Holden 2002) these populations lie in opposite edges of the two more ancient Bantu dispersal routes and should be maximally divergent. However, our study did not reveal any clear-cut separation between Angola and Mozambique, and showed only very small amounts of genetic differentiation among the sampled groups. In fact, with the exception of few outliers, like the Kuvale from Angola and the Chopi from Mozambique, the general picture emerging from our analyses (individual clustering patterns, AMOVA, SAMOVA and MDS) is that most groups may be included into a single cluster.

In view of this broad genetic homogeneity, the outlier groups suggest that local factors, like drift and differential admixture with non-Bantu populations, may have played a more important role in producing genetic differentiation than large scale continental-wide events. In fact, it is likely that the genetic distinctiveness of the Kuvale resulted from increased drift and local admixture with Khoikhoi-herders from southwest Africa, as shown by recent studies of on mtDNA variation (Coelho et al. 2009). Unlike the Kuvale, the Chopi were previously found to be unremarkable at the mtDNA level and did not reveal special signs of increased admixture with Khoisan groups (Salas et al. 2002). However, this group is ethnographically notable for using large xylophone ensembles resembling Javanese and Balinese xylophone orchestras, a peculiar cultural trait that has sometimes been taken as evidence of an Indonesian influence via Madagascar (Hogan 2006). In the future, it will be worth to investigate more thoroughly if a genetic link underlies the cultural connection between the Chopi and Madagascar.

Within Mozambique, two sets of spatially contiguous populations were allocated to different linguistic groups by SAMOVA (Table 4; K = 6). The first set includes the Kaskazi-speaking Yao and Mwani, from far northern Mozambique (Figure 1B) that are linguistically closer to Bantu peoples from Tanzania and Kenya than to the other Mozambican populations, who speak languages grouped in the Kusi cluster (Ehret 1998). Taking into account the location of the Yao and Mwani at the southern edge of the Kaskazi distribution, further studies on their relationship with their northern and southern neighbors might help to better understand the occupation of East and Southeast Africa. The second set of populations includes the Nyungwe and the Sena from the Zambezi basin, who speak languages belonging to Nyasa, one of the major Kusi-speaking groups that scattered across southeastern Africa, together with Makua, Shona and South-East Bantu (Figure 1B; Ehret 1998). However, this cluster does not include the Nyanja, who also speak a Nyasa language, and the robustness of the Nyasa group remains to be more fully investigated. In spite of these two groupings, language does not seem to be a good overall predictor of the genetic diversity in Mozambique, as no significant partial correlations were found between genetic distance and our rough measure of linguistic differentiation (Table 5). However, it will be useful to further evaluate if more refined linguistic distances will lead to different results. Like language, geographic distance was also found to be a poor predictor of genetic differences between Mozambican groups (Table 5). Moreover, additional analyses based on comparisons of individuals and populations, using the methods implemented in the software GENEPOP (Rousset 2008), failed to reveal a consistent pattern of isolation by distance (not shown). Taken together, these observations suggest that present-day southeastern Bantu peoples descend from closely related ancestors that spread recently across a broad range of territories, with no time for restricted gene flow to generate isolation by distance or genetic differentiation across language barriers (see, eg., Lansing et al. 2007). The linguistic and archeological evidence suggesting that Bantu peoples settled Mozambique between 2,000 to 1,600 years ago (Ehret 1998) provides additional support to this interpretation.

Population-history parameters

A major limitation of our ABC approach was the inability to obtain sufficient numbers of accepted simulations under reasonable computation times using multiple summary statistics simultaneously. This is probably related to the relatively inefficient rejection-sampling method implemented in REJECTOR, which can only handle a small number of summary statistics. When many summary statistics were used, either the number of accepted simulations became

too low, or the tolerance had to be increased to levels that compromised the approximation to the parameter. Due to these limitations, demographic parameters were estimated using summary statistics separately, and no formal test was performed to assess which of the simulated scenarios better explained the data.

Recently, a number of methods have been developed to increase the efficiency of ABC. Major improvements include modifications of the original rejection-sampling algorithms (Beaumont et al. 2002), sounder criteria for choosing informative summary statistics (Wegmann et al. 2009, Joyce and Marjoram 2008), and effective ways to compare alternative scenarios (Fagundes et al. 2007). These recent developments have been successfully applied to infer the branching history and migration dynamics of Pygmy and agricultural populations (Wegmann et al. 2009, Verdu et al. 2009, Patin et al. 2009), but ABC methods are still not commonly applied to the history of most African populations. The study of the Bantu expansions, in particular, is especially challenging since the underlying demographic events are very recent and difficult to resolve. In this context, our contribution is intended to provide a first approximation that must be improved in future inferential studies using more efficient ABC frameworks and increased numbers of loci to address more complex demographic models. To this end, it will be crucial to extend the number of sampled regions, as contrasting population-history models may be easier to discriminate on the basis of their different expectations for the relationships among populations scattered across many Bantu-speaking regions.

With these caveats in mind, we obtained a set of estimates that may be useful to contrast with other inferential frameworks and empirical datasets. Our point estimates of divergence times between the Angolan and Mozambican populations (~2600-5100 years) are in good agreement with the age of diversification of Bantu-speaking communities (Newman 1995; Ehret 1998), but lack precision to discriminate between early or late split scenarios. The well peaked posterior distributions with nonzero rates of migration and exponential growth, suggest that more complex models allowing for continuous gene flow (Figures 7A and 7B) and/or population growth (Figures 6B and 7B) are more realistic representations of the history of the sampled Bantu populations. The signal for exponential growth in Angola (r ~0.002-0.003) is in agreement with previous findings based on mtDNA and Y-chromosome variation (Salas et al. 2002; Coelho et al. 2009), falling within the range of early ABC estimates of population growth in southern African populations (Pritchard et al. 1999). Finally, the migration rate calculations (Nm ~4-10) between Southeast and Southwest Africa are congruent with previous estimates based on mtDNA data ($N_em > 5$; Castrì et al. 2009; $N_em > 10$; Coelho et al. 2009),

indicating that extensive gene flow occurred across the savanna areas located to the south of the equatorial rainforest.

Conclusion

The genetic homogeneity revealed in the present study is difficult to reconcile with an early split between eastern and western Bantu populations, suggesting that the spreading of Bantu languages is better portrayed as a gradual unfolding of interconnected populations than a series of successive bifurcations involving small sized groups. Alternative models that place the ancestors of subequatorial Bantu peoples at the southern outskirts of the rainforest, shortening divergence times and increasing opportunities for gene flow, seem to better fit the observed genetic patterns.

Acknowledgements

We are grateful to all sample donors, to the Governor of the Province of Namibe, and to Dr. Pedro Viyayauca, Chairman of Namibe's Provincial Health Department, for support during sample collection in Angola. The work was financed by grant PTDC/BIA-BDE/68999/2006 from Fundação para a Ciência e a Tecnologia (FCT, Portugal) and by projects "Variabilidade Biológica Humana em Moçambique" and "STEPS" at the Pedagogic University and the University Eduardo Mondlane of Mozambique, respectively. Margarida Coelho was supported by FCT grant SFRH/BD/22651/2005. We thank Sandra Beleza and Nuno Ferrand for comments and suggestions, and Joanna Mountain for support in developing UEPSTRs and for critically reading the manuscript.

Literature Cited

Barbujani, G., A. Magagni, E. Minch and L.L. Cavalli-Sforza. 1997. An apportionment of human DNA diversity. *Proc. Natl. Acad. Sci. USA.* 94:4516-4519.

Beaumont, M.A., W. Zhang, D.J. Balding. 2002. Approximate Bayesian computation in population genetics. *Genetics*, 162, 2025–2035.

Belkhir, K., P. Borsa, J. Goudet et al. 1998. GENETIX, logiciels pour Windows pour la genétique des populations. Laboratoire Génome et Populations, Université de Montpellier II, Montpellier.

Castrì, L., S. Tofanelli, P. Garagnani et al. 2009. mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. *Am. J. Phys. Anthropol.* 140: 302-311.

Coelho M., F. Sequeira, D. Luiselli et al. 2009. On the edge of Bantu expansions: mtDNA, Y chromosome and lactase persistence genetic variation in southwestern Angola. *BMC. Evol. Biol.* 9: 80.

Collins, J.R., R.M. Stephens, B. Gold et al. 2003. An exhaustive DNA micro-satellite map of the human genome using high performance computing. *Genomics.* 82:10-19.

Cornuet, J.M., F. Santos, M..A. Beaumont et al. 2008. Inferring population history with DIY ABC: a user-friendly approach to Approximate Bayesian Computation. *Bioinformatics* 24: 2713-2719.

de Knijff, P. 2000. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. *Am. J. Hum. Genet.* 67:1055-1061.

Dupanloup, I., S. Schneider, L. Excoffier 2002. A simulated annealing approach to define the genetic structure of populations. *Mol. Ecol.*11: 2571-2581.

Ehret, C. 1998. *An African Classical Age: Eastern & Southern Africa in World History,1000B.C. to A.D.400.* Charlottesville, University Press of Virginia.

Ehret, C. 2001. Bantu expansion: re-envisioning a central problem of early African history. *Int. J. Afr. Hist. Stud*. 34:5–41.

Epstein, M. P., W. L. Duren, M. Boehnke. 2000. Improved inference of relationship for pairs of individuals. *Am. J. Hum. Genet.* 67: 1219-1231.

Estoup, A., P. Jarne, J.M. Cornuet. 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.* 11:1591-1604.

Excoffier, L., G. Laval, S. Schneider. 2005. Arlequin (version 3.0): An integrated software package for population genetics data analysis *Evol. Bioinform. Online.* 1: 47-50.

Fagundes, N. J., N. Ray, M. Beaumont et al. 2007. Statistical evaluation of alternative models of human evolution. *Proc. Natl. Acad. Sci. USA.* 104: 17614-17619.

Falush, D., M. Stephens, J.K. Pritchard. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* 164: 1567-1587.

Flint, J., J. Bond, D.C. Rees et al. 1999. Minisatellite mutation processes reduce Fst estimates. *Hum. Genet.* 105:567-576.

Garrigan, D., M.F. Hammer. 2006. Reconstructing human origins in the genomic era. *Nat. Rev. Genet.* 7: 669–680.

Goldstein, D.B., A. Ruiz-Linares, L.L. Cavalli-Sforza, et al. 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics.* 139:463-471.

Guthrie M.1967-1971. Comparative Bantu: an introduction to the comparative linguistics and prehistory of Bantu languages, vols 1-4. Farnborough, UK: Greg International.

Harpending, H.C., J.H. Relethford, S.T. Sherry. 1996. Methods and models for understanding human diversity. In *Molecular Biology and Evolution*, AJ Boyce and C.G.N. Mascie-Taylor, eds. Cambridge, England: Cambridge University Press, 283-299.

Hey, J., Y.J. Won, A. Sivasundar et al. 2004. Using nuclear haplotypes with microsatellites to study gene flow between recently separated Cichlid species. *Mol. Ecol.* 13: 909-919.

Hogan, B. 2006. Locating The Chopi Xylophone Ensembles of Southern Mozambique. *Pac. Rev. Ethnomusicol*. 11 (Winter 2006).

Holden, C.J. 2002. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proc. R. Soc. Lond. B* 269:793-799.

Jakobsson, M., N.A. Rosenberg. 2007. CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. *Bioinformatics.* 23:1801-1806.

Jobin, M.J., and J.L. Mountain. 2008. REJECTOR: software for population history inference from genetic data via a rejection algorithm. *Bioinformatics* 24:2936-2937.

Joyce P., and P. Marjoram. 2008. Approximately sufficient statistics and Bayesian computation. *Stat. Appl. Genet. Mol. Biol.* 7: 26.

Lansing, J. S., M. P. Cox, S. S. Downey et al. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc. Natl. Acad. Sci. USA.* 41: 16022-16026.

Li, J.Z., D.M. Absher, H. Tang et al. 2008. Worldwide human relationships inferred from genome-wide patterns of variation. *Science.* 319:1100–1104.

Mountain, J.L., and L.L. Cavalli-Sforza. 1994. Inference of human evolution through cladistic analysis of nuclear DNA restriction polymorphisms. *Proc. Natl. Acad. Sci. USA.* 91:6515–6519.

Mountain, J.L., A. Knight, M. Jobin et al. 2002. SNPSTRs: empirically derived, rapidly typed, autosomal haplotypes for inference of population history and mutational processes. *Genome. Res.* 12:1766-1772.

Need, A.C. and D.B. Goldstein 2009. Next generation disparities in human genomics: concerns and remedies. *Trends Genet.* 25: 489-494.

Nei, M. 1987. Molecular Evolutionary Genetics. New York, Columbia University Press.

Newman, J.L.: *The peopling of Africa: A Geographical Interpretation* New Haven, Yale University Press; 1995.

Patin, E., G. Laval, L. B. Barreiro et al. 2009. Inferring the demographic history of African farmers and Pygmy hunter-gatherers using a multilocus resequencing data set. *Plos Genet.* 5: e1000448.

Payseur, B.A., and A.D. Cutter. 2006. Integrating patterns of polymorphism at SNPs and STRs. *Trends. Genet.* 22:424-429.

Payseur, B.A., and P. Jing. 2009. A genomewide comparison of population structure at STRPs and nearby SNPs in humans. *Mol. Biol. Evol.* 26:1369-1377.

Pereira, L., L. Gusmão, C. Alves et al. 2002. Bantu and European Y-lineaes in sub-Saharian Africa. *Ann. Hum. Genet.* 66: 369-378.

Pritchard, J.K., M. T. Seielstad, A. Perez-Lezaun et al. 1999. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Mol. Biol. Evol.* 16: 1791-1798.

Pritchard, J.K., M. Stephens, P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics.* 155:945-959.

Ramakrishnan, U., and J.L. Mountain 2004. Precision and accuracy of divergence time estimates from STR and SNPSTR variation. *Mol. Biol. Evol.* 21:1960-1971.

Rexová, K., Y. Bastin, D. Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften.* 93:189-94.

Reynolds, J., B.S. Weir, and C.C. Cockerham. 1983. Estimation for the coancestry coefficient: basis for a short-term genetic distance. *Genetics*. 105:767-779.

Rogers, A.R., and L.B. Jorde. 1996. Ascertainment bias in estimates of average heterozygosity. *Am. J. Hum. Genet*. 58: 1033–1041.

Romero, I.G., A. Manica, J. Goudet et al. 2008. How accurate is the current picture of human genetic variation? *Heredity.* 102**:**120-126.

Rosenberg, N.A. 2004. DISTRUCT: a program for the graphical display of population structure. *Mol. Ecol. Notes*. 4: 137-138.

Rosenberg, N.A., J.K. Pritchard, J.L. Weber et al. 2002. Genetic structure of human populations. *Science.* 298: 2381–2385.

Rosenberg, N.A., L.M. Li, R. Ward et al. 2003. Informativeness of genetic markers for inference of ancestry. *Am. J. Hum. Genet.* 73: 1402-1422.

Rousset, F., 2008. Genepop'007: a complete reimplementation of the Genepop software for Windows and Linux. *Mol. Ecol. Resources* 8: 103-106.

Salas, A., M. Richards, T. De la Fe et al. 2002. The making of the African mtDNA landscape. *Am. J. Hum. Genet.* 71:1082-1111.

Schuelke, M. 2000. An economic method for the fluorescent labeling of PCR fragments. *Nat. Biotechnol.* 18: 233-234.

Shriver, M.D., L. Jin, E. Boerwinkle et al. 1995. A novel measure of genetic distance for highly polymorphic tandem repeat loci. *Mol. Biol. Evol.* 12:914-920.

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

Sturrock, K., J. Rocha. 2000. A multidimensional scaling stress evaluation table. *Field. Methods.* 12:49–60.

Tishkoff, S.A., and Kidd K.K. 2004. Implications of biogeography of human populations for 'race' and medicine. *Nat. Genet.* 36:S21-27.

Tishkoff, S.A., F.A. Reed, F.R. Friedlaender et al. 2009. The genetic structure and history of Africans and African Americans. *Science.* 324: 1035-1044.

Verdu, P., F. Austerlitz, A. Estoup et al. 2009. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr. Biol.* 19: 312-318.

Weber, J.L., D. David, J. Heil et al. 2002. Human diallelic insertion/deletion polymorphisms. *Am. J. Hum. Genet.* 71: 854-862.

Wegmannn, D., C. Leuenberger, L. Excoffier. 2009. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *Genetics* 182: 1207-1218.

Xu, H., and Y.X. Fu. 2004. Estimating effective population size or mutation rate with microsatellites. *Genetics* 166: 555-563.

Xu, H., R. Chakraborty, Y.X. Fu. 2005. Mutation rate variation at human dinucleotide microssatellites. *Genetics* 170: 305-312.

2.2.1 Comments

2.2.1.1 Interactions between Bantu-speaking people and hunter-gatherer populations

The spread of Bantu-speaking peoples from West-Central Africa around ~5,000 years ago profoundly altered the genetic landscape of the African sub-Saharan region (Campbell and Tishkoff 2008). The kind of interactions that were established between the spreading Bantu agriculturalists and hunter-gatherers are still not well known (Diamond and Bellwood 2003).

Results from Tishkoff et al. (2009) suggest a common ancestry of contemporary hunter-gatherers, including Khoisan, Hadza, Sandawe and Pygmies. The divergence between the major hunter-gatherers groups in Africa is likely to have occurred much earlier than the expansion of Bantu agriculturalists. A first split, around 35,000 years ago, may have led to the separation of the ancestors of Sandawe and Southern African Khoisan groups (Tishkoff et al. 2007a, Behar et al. 2008). A later split, around 20,000 years ago, seems to have led to the divergence between Western and Eastern Pygmy groups (Destro-Bisol et al. 2004, Patin et al. 2009).

Genetic studies are providing important insights into the biological impact of the interactions between Bantu agriculturalists and hunter-gatherers (e.g. Verdu et al. 2009, de Filippo et al. 2010). These interactions have been shown to be quite heterogeneous, leading to different degrees of cultural and/or genetic exchanges (e.g. Verdu et al. 2009). It seems that sociocultural boundaries between the Bantu and hunter-gatherer communities could have been important factors driving long-term interaction. For example, strong discrimination against Pygmy populations may have prevented marriages between Pygmies and Bantu populations (Destro-Bisol et al. 2004). Also, subsistence strategies seem to have been important factors determining the kind of interactions established between Bantu and non-Bantu populations (de Filippo et al. 2010). In southern Africa, where the environment was less suitable to agriculture, Bantu populations may have diversified their subsistence resources to include in a larger extent food gathering, hunting, fishing and sometimes, where the conditions were favourable, also pastoralism (Newman 1995). In several situations, Bantu agriculturalists may have competed directly with resident Khoisans (Newman 1995) and a great variety of responses has been observed. Some groups, like the Kuvale preserved their Bantu language and became pastoralists (Newman 1995). Others, like the Bergdama, may have been absorbed by the Khoe and abandoned their language (Curtin et al. 1995).

2.2.1.1.1 Interactions between Bantu and Pygmy populations

Pygmies are represented by two main groups: "Eastern Pygmies" from the Democratic Republic of Congo, and by "Western Pygmies", from Cameroon, Congo, Gabon and Central African Republic (Destro-Bisol et al. 2004). These hunter-gatherers coexist with neighbouring Bantu agriculturalist populations and speak Bantu languages. Recent analyses of multilocus autosomal genetic variation in several African populations suggest the occurrence of an asymmetrical gene flow between Bantu and Pygmy populations, with higher levels of Bantu introgression occurring from Bantu into Pygmies populations than the opposite (Patin et al. 2009, Verdu et al. 2009, Tishkoff et al. 2009). A heterogeneous level of gene flow from Bantu to Pygmy populations has been also observed (Verdu et al. 2009, Tishkoff et al. 2009). For example, by using STRUCTURE, Tishkoff et al. (2009) found that Bantu ancestry among Pygmy groups could range from 0.13 in the Eastern Mbuti Pygmies to 0.54 in the Western Bedzan Pygmies.

The analysis of uniparental genetic markers provided evidence of sex-biased rates of admixture between both populations: maternal lineages (e.g. L1c1a) were introduced mainly into Bantu from Pygmies populations (Quintana-Murci et al. 2008), while substantial paternal Bantu lineages (e.g. E3a) have been introduced into Pygmies from Bantu populations (Berniell-Lee et al. 2009). Intermarriage practices between Pygmies and Bantu populations seem to explain the observed patterns of genetic admixture. Indeed, following ethnographic reports, marriages between a Bantu woman and a Pygmy man seem to be more socially constrained than marriages between a Pygmy woman and a Bantu man. Given that these populations are often patrilocal, the introgression from Bantu into Pygmy populations is sometimes the result of the return of the Pygmy woman and her children to her original community after divorce or by the presence in the Pygmy communities of illegitimate children from Bantu men and Pygmy women (Verdu et al. 2009).

In our sample from southern Angola and in the sample from Mozambique previously studied by Salas et al. (2002), the haplogroup L1c1a, typical of "Western Pygmy" populations, was found to be virtually absent. In contrast, this haplogroup is found at high frequencies in Bantu populations from Cameroon and Gabon (Quintana-Murci et al. 2008). This difference adds support to the model of Bantu dispersion whereby Bantu languages located to the south and east of rainforest would have had origin on a dispersal centre located in savannah areas just south of the equatorial forest (Ehret 1998, Rexová et al. 2006). In fact, if populations from southwestern Angola were directly derived from rainforest Bantu groups in Cameroon and Gabon, a heavier Pygmy component would be expected in these populations.

2.2.1.1.2 Interactions between Bantu and Khoisan-speaking populations

2.2.1.1.2.1 Inference based on haplogroups characteristic of Khoisan or Bantu populations

In southern regions of Africa the assimilation of male and female- Khoisan characteristic lineages varies widely across geographic regions and ethnic groups. In eastern Zambia it was found that the Bisa and Kunda groups, present low levels of assimilation of putative hunter-gatherer lineages in both the paternal and maternal line (~3%) (de Filippo et al. 2010). In the Shona from the north of Zimbabwe the L0d/L0k haplogroups are also present at low frequencies (1.69% each) (Castri et al. 2009). In Mozambique, Khoisan characteristic lineages are found at relatively low frequencies in both the maternal (~6%) and paternal (~9%) lineages (Pereira et al. 2001, Pereira et al. 2002, Salas et al. 2002,). However, it was observed that some Mozambican populations, such as the Ronga, Tswa and Ndau present higher levels of the L0d-Khoisan characteristic haplogroup (19, 16, and 16%, respectively) (Salas et al. 2002). At the southern African Xhosa and Zulu, a high frequency of the L0d- female Khoisan characteristic haplogroup- is observed (~25% and ~50%, respectively) (Soodyall and Jenkins 1993). In the male counterpart, a lower Khoisan introgression seems to have occurred, as reflected by the presence of the low frequency of the A3b1-M51 haplogroup (5% and 3 % in Xhosa and Zulu, respectively) (Wood et al. 2005).

L0d/k haplogroups were not found in previous studies from northern Angola (Plaza et al. 2004, Beleza et al. 2005). In our analysis of the genetic pool of southwestern Angola we found signs of admixture with local Khoisan, especially among the Kuvale, where typical Khoisan mtDNA L0d and NRY B2b haplotypes reached as much as 22% and 12%, respectively. In addition, our UEPSTRs multilocus analysis of populations from Mozambique and Southwest Angola shows that the Kuvale stands out as an outlier group among several Bantu populations. This observation highlights the importance of the interactions established between Bantu and non-Bantu groups in shaping the general pattern of genetic diversity among Bantu populations.

2.2.1.1.2.2 Novel insights provided by the study of the -14010*C lactase persistence associated allele

Another interesting aspect of our study, was the finding of the -14010*C lactase persistence allele at a relatively low frequency (6%) in the Kuvale people (Figure II.3 A and B). Since the -14010*C variant is especially frequent among Nilo-Saharan and Afro-Asiatic-speaking pastoral populations from Kenya and Tanzania (Tishkoff et al. 2007b), our observation provides genetic evidence for a link between the relatively isolated southwestern Africa pastoral scene and the major cattle herding centers of East Africa. Direct links between Southwest and East Africa were favorite topics of early Anthropology scholars studying southwestern Africa (Blench 2009). Estermann (1961), for example, claimed that the Kuvale, together with other Bantu herding peoples from southern Angola, had a "Chamitic" provenance, implying that they could trace at least part of their origins to non-Bantu populations from the region of the Great Lakes. However, on the basis of our mtDNA and NRY data, we failed to find any consistent genetic affinity between the peoples from southwestern Angola and non-Bantu East Africans (Figure II.4 A, B, C, and D). Thus, the transference of the -14010*C allele between East and Southwest Africa does not appear to be a consequence of a direct migration between both regions but was otherwise mediated by other groups. The presence in southern Africa of an ancient pastoral tradition, represented by the Khoisan speaking Khoe was the basis for our alternative hypothesis. Linguistic and archeological evidences indicate that Khoe may have acquired their pastoral culture as a result of contacts with migrating herders from East Africa in the northeast Zambia and Zimbabwe (Blench 2009). A genetic imprint of such contacts seems to be attested by the observation of a new Y- chromosome haplogroup (E3b1f) with high frequencies in pastoral communities from both East Africa Great Lakes and southern African Khoe (Henn et al. 2008) (Figure II.3C).

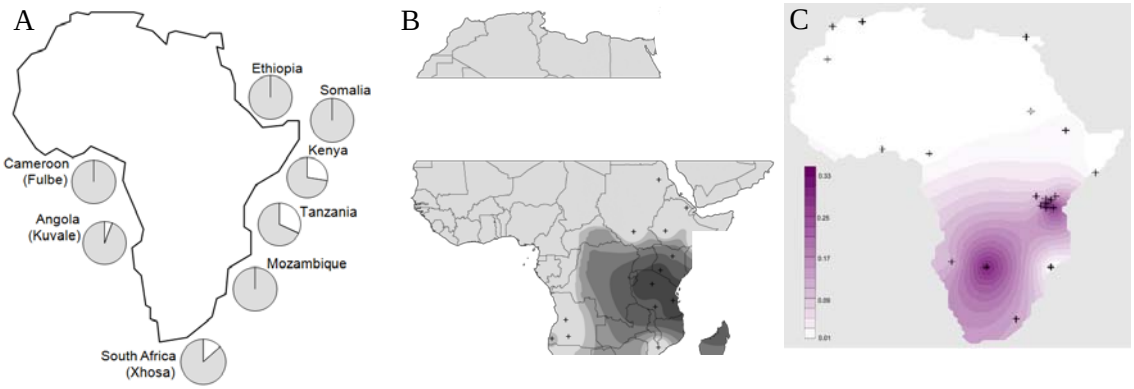


Figure II.3 (A) Frequencies of the -14010*C lactase persistence associated allele in several populations from Africa (in white). Data from Cameroon, Angola and Mozambique were obtained in the present work. The remaining data are from Ingram et al. (2007), Ingram et al. (2009) and Tishkoff et al. (2007b); (B) Interpolation graphic of the data shown in panel A. The darker the color, the higher the frequency of the allele; (C) Frequencies of the Y-chromosome haplogroup E3b1f (figure retrieved from Henn et al. 2008). In (B) and (C), the sampled locations are marked with a cross. In the remaining places, the allelic frequencies were deducted assuming that there was a linear decrease of the frequency as far as the distance to the area of highest frequency decreased.

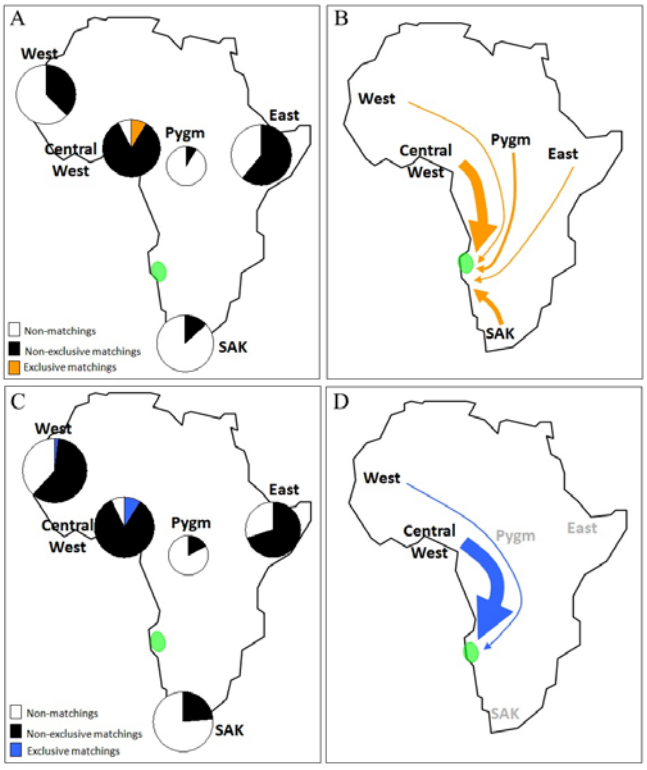


Figure II.4 (A) and (C) Patterns of sequence sharing between southwestern Angola and other Sub-Saharan regions/contextual populations, by using mtDNA and NRY data, respectively. (B) and (D) Estimated genetic contribution of different parental groups to the southwestern Angola population using mtDNA and NRY data, respectively. The thickness of the arrows is proportional to the admixture proportions of the group to southwestern Angola population.

In this setting, it is possible that the -14010*C lactase persistence associated allele was transferred to the Khoe together with domestic animals and pastoral habits. This transfer would explain early observations of moderate frequencies of the lactase persistence trait among the Khoe (Casimir 1990). Subsequent migrations of Khoe herders, would led to the transference of the -14010*C allele from the Khoe to southern Bantu pastoral peoples in well defined contact zones like southwestern Angola. Recently, the -14010*C allele was reported to occur at 13% frequency in the Xhosa population from South Africa (Torniainen et al. 2009) (Figure II.3A), which is well known for extensive cultural and genetic interactions with the Gonaqua Khoe, from whom it borrowed a number of click words. On the other hand, we found no lactase persistence variants in Bantu communities from southern Mozambique that are somewhat related to the Xhosa but did not interacted as extensively with the Khoe (Figure II. 3A). Taken together, these observations add further support to our hypothesis by showing that the only Southern Bantu groups with lactase persistence are the ones that made cultural and genetic contact with the Khoe. However, further investigation on the connection between lactase persistence and pastoralism in southern Africa is clearly needed, especially on the poorly sampled Khoe from whom lactase persistence data still relies on physiological tests performed more than twenty years ago.

2.2.1.1.2.3 Implications for the search of the extinct Kwadi gene pool

The fact that our Angolan samples were collected in regions encompassing the area previously inhabited by speakers of the extinct Kwadi click language provided an opportunity to explore possible interactions between Bantu and Kwadi populations (Figure II.5). The Kwadi language lies on a separate branch of the Khoe family, which groups click languages spoken by pastoral Khoisan peoples, like the Nama from neighboring Namibia (Figure II.5) (Güldemann 2008). The last records of the Kwadi language were made by the Professor António de Almeida in 1955, when this group was already reduced to about fifty people. According to de Almeida's reports and pictures, the Kwadi were physically very distinct from other Khoisan and resembled Herero groups from southern Angola, like the Himba and the Kuvale (de Almeida 1994). The cultural and physical resemblance between the Kuvale and their neighboring Kwadi raises the possibility that the later were an offshoot of the former who adopted the Kwadi speech in the course of more active interactions with Khoe herders originally speaking this language.

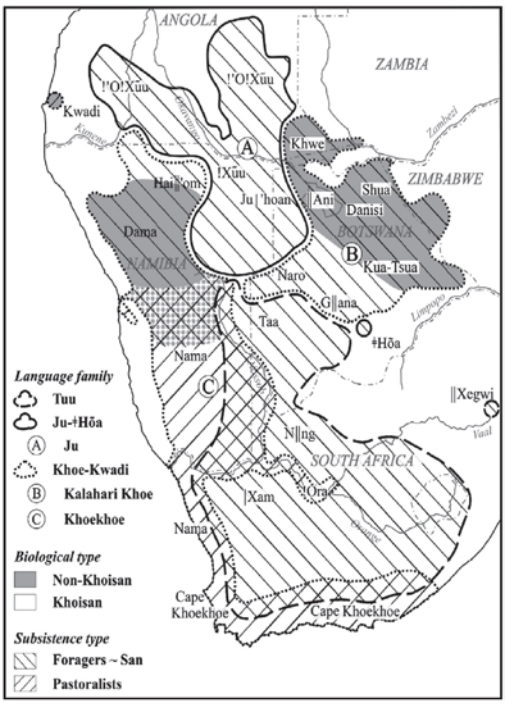


Figure II.5 Map displaying the past distribution of southern African non-Bantu groups and their basic linguistic, biological and cultural classifications (retrieved from Güldemann 2008)

2.2.1.2 Recent developments in the study of the Bantu expansions

The study of Bantu expansions is quite challenging since the underlying demographic events are very recent and complex. Recent studies involving a more comprehensive sample coverage and the analysis of a high number of markers are increasingly providing the tools to address the demographic history of the Bantu expansions (e.g. Tishkoff et al. 2009).

Several studies have applied dense sampling strategies in order to analyse regional aspects of the Bantu expansions like the interactions between Pygmy hunter-gatherers and Bantu-speaking farmers (Quintana-Murci et al. 2008, Berniell-Lee et al. 2009, Patin et al. 2009, Verdu et al. 2009), or the genetic differentiation between several populations inhabiting the Cross River region of Nigeria (Veeramah et al. 2010).

Multilocus approaches have received a remarkable boost with the recent publication of Sarah Tishkoff's study on 2432 individuals from 113 populations using a panel of 1327 polymorphic markers (Tishkoff et al. 2009). This study included a set of populations belonging to the Bantoid branch of the Niger-Kordofanian family, located in Cameroon, Democratic Republic of Congo, Rwanda, Tanzania, Kenya and South Africa. By considering the geographic data along with clustering analysis it was possible to distinguish five major groups of clusters, including one corresponding to the Niger-Kordofonian language family (Figure II.6A, in orange). Moreover, it was possible to detect a slight substructure among Niger-Kordofanian speakers. This substructure was mainly due to the identification of specific genetic components in non-Bantu Niger-Kordofanians from West Africa, West Central African Bantus, and East Bantu populations (Tanzania and Kenya) speaking Kaskazi languages (Figure II.6B) (Tishkoff et al. 2009). Interestingly, the only Kusi groups from South Africa included in Tishkoff's study (the Venda and the Xhosa), did not reveal high levels of the unique genetic component present in Kaskazi populations (Figure II.6B). The separation between the Kaskazi-speaking Yao and Mwani, from the Kusi-speaking groups in Mozambique, reported in *Article 3*, seems to be in agreement with these findings. Taken together, these patterns seem to support the idea that the occupation of East and Southeast Africa involved separate movements of the two related groups of populations and not a single eastern path as implied by the early split model of the Bantu expansions (see *Article 3*).

In other study, Patin et al. (2009) resequenced 24 independent noncoding regions, in Western and Eastern Pygmies and in several agricultural populations (Figure II.7 A). Through the use of the STRUCTURE program, they were able to separate the Mozambican population from other populations, including Bantu-speaking populations like the Ngumba (Cameroon), Akele (Gabon) and Chagga (Tanzania) (Patin et al. 2009) (Figure II.7 B). Again, these results indicate no close relationship between southeast (a.k.a. Kusi) and east (a.k.a Kaskazi) Bantu populations.

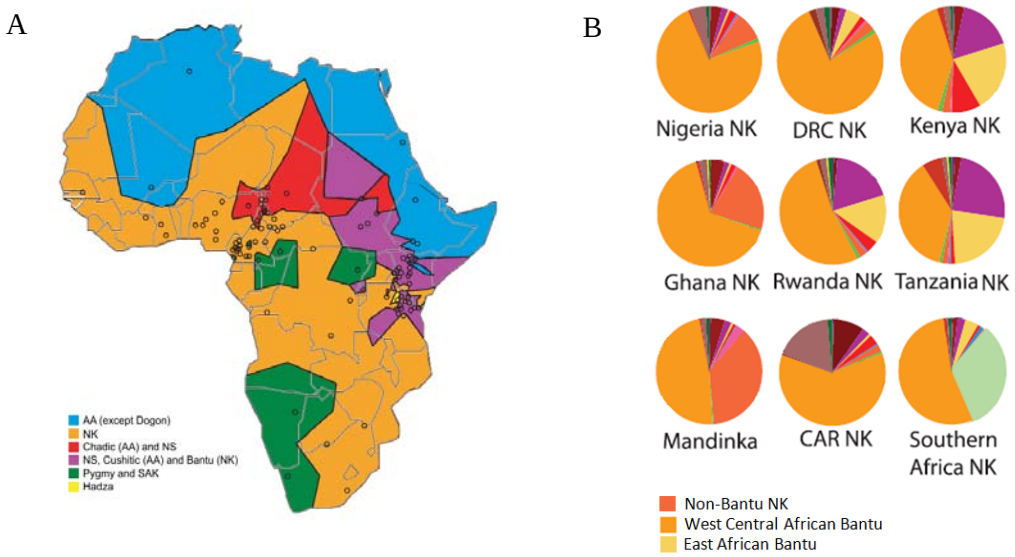


Figure II.6 (A) Geographic discontinuities among African populations using the TESS program (Chen et al. 2007). Circles indicate location of the populations analysed; (B) Inferred proportion of ancestral clusters using STRUCTURE. Dark, medium and light orange corresponds to non-Bantu Niger Kordofanian, West Central African Bantu and East African Bantu ancestries, respectively. Both figures were retrieved from Tishkoff et al. 2009. AA = Afroasiatic; NK = Niger-Kordofanian; NS = Nilo-Saharan; SAK = southern African Khoisan; DRC = Democratic Republic of Congo; CAR = Central African Republic.

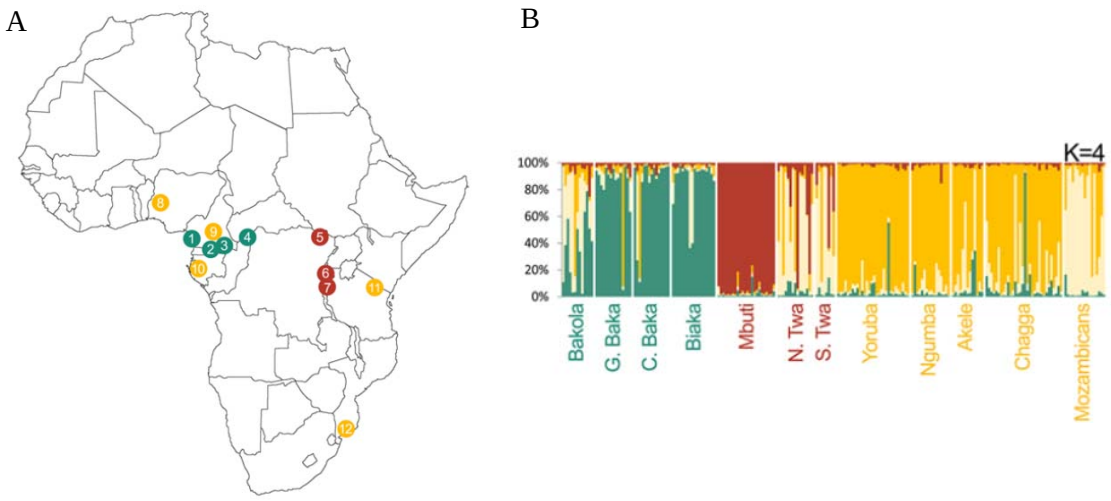


Figure II.7 (A) Geographic location of the populations studies by Patin et al. 2009. Green, red and yellow dots represent Western Pygmy, Eastern Pygmy and agricultural populations, respectively. 1. Bakola from Cameroon, 2. Baka from Gabon, 3. Baka from Cameroon, 4. Biaka from the Central Africa Republic, 5. Mbuti from the Democratic republic of Congo, 6. Twa from northern Rwanda, 7. Twa from southern Rwanda, 8. Yoruba from Nigeria, 9. Ngumba from Cameroon, 10. Akele from Gabon, 11. Chagga from Tanzania, 12. Mozambicans from Mozambique (retrieved from Patin et al. 2009).

In the future it will be interesting to further investigate the implications of the Bantu peopling of east and southeast Africa in the more general debate about the alternate models of Bantu dispersals. In this context it will be crucial to extend the number of sampled regions, as contrasting population-history models may be easier to discriminate on the basis of their different expectations for the relationships among populations scattered across many Bantu-speaking regions. It is important to note that the absence of equivalent sets of markers and populations in the different studies make the comparison of the several results very difficult. Therefore it will be also important to define a minimum subset of highly informative markers to be used in future works in order to have a more complete picture of genetic diversity and population history of Bantu populations.

References

Behar. D. M., R. Villems, H. Soodyall, J. Blue-Smith, L. Pereira, E. Metspalu, R. Scozzari, H. Makkan, S. Tzur, D. Comas, J. Bertranpetit, L. Quintana-Murci, C. Tyler-Smith, R.S. Wells, and S. Rosset. 2008. The dawn of human matrilineal diversity. *Am J Hum Genet* 82(5):1130-40.

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

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

Blench, R. 2009. Was there and interchange between Cushitic pastoralists and Khoisan speakers in the prehistory of Southern Africa and how can this be detected? *Sprache und Geschichte in Afrika*, 20. (in press).

Campbell, M. C., and S. A. Tishkoff. 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* 9:403-33.

Casimir, M.J. 1990. On milk drinking San and the "myth of the primitive isolate". *Curr. Anthrop* 31: 551-554.

Castri, L., S. Tofanelli, P. Garagnani, C. Bini, X. Fosella, S. Pelotti, G. Paoli, D. Pettener, and D. Luiselli. 2009. mtDNA variability in two Bantu-speaking populations (Shona and Hutu) from Eastern Africa: implications for peopling and migration patterns in sub-Saharan Africa. *Am J Phys Anthropol* 140:302-11.

Chen, C., E. Durand, F. Forbes, and O. François. 2007. Bayesian clustering algorithms ascertaining spatial population structure: A new computer program and a comparison study. *Mol Ecol Notes* 7:747-756.

Curtin, P., S. Feierman, L. Thompson, and J. Vansina. 1995. *African history: from earliest times to independence*. London: Longman.

de Almeida A. 1994. Bushmen and other non-Bantu peoples of Angola: three lectures. In the Almeida A. *Os bosquímanos de Angola*. Lisboa, Ministério do Planeamento e da Administração do Território, Secretaria de Estado da Ciência e Tecnologia, Instituto de Investigação Científica Tropical. (Originally published in 1965).

de Filippo, C., P. Heyn, L. Barham, M. Stoneking, and B. Pakendorf. 2010. Genetic perspectives on forager-farmer interaction in the Luangwa Valley of Zambia. *Am J Phys Anthropol*. 141(3):382-94.

Destro-Bisol G., V. Coia, I. Boschi, F. Verginelli, A. Cagliá, V. Pascali, G. Spedini, and F. Calafell. 2004. The analysis of variation of mtDNA hypervariable region 1 suggests that Eastern and Western Pygmies diverged before the Bantu expansion. *Am Nat.* 163(2):212-26.

Diamond J., and P. Bellwood. 2003. Farmers and their language: the first expansions. *Science*. 300: 597-603.

Ehret, C. 1998. *An African Classical Age: Eastern & Southern Africa in World History,1000B.C. to A.D.400* Charlottesville, University Press of Virginia.

Estermann, C. 1961. *Etnografia do sudoeste de Angola: o grupo étnico Herero*. Volume 3.

Lisboa, Junta de investigações do Ultramar.

Güldemann, T. 2008. A linguist's view: Khoe-Kwadi speakers as the earliest food-producers of southern Africa. *Southern African Humanities* 20: 93-132.

Henn, B. M., C. Gignoux, A. A. Lin, P. J. Oefner, P. Shen, R. Scozzari, F. Cruciani, S. A. Tishkoff, J. L. Mountain, and P. A. Underhill. 2008. Y-chromosomal evidence of a pastoralist migration through Tanzania to southern Africa. *Proc Natl Acad Sci U S A* 105:10693-8.

Ingram, C. J., M. F. Elamin, C. A. Mulcare, M. E. Weale, A. Tarekegn, T. O. Raga, E. Bekele, F. M. Elamin, M. G. Thomas, N. Bradman, and D. M. Swallow. 2007. A novel polymorphism associated with lactose tolerance in Africa: multiple causes for lactase persistence? *Hum Genet* 120:779-88.

Ingram, C. J., T. O. Raga, A. Tarekegn, S. L. Browning, M. F. Elamin, E. Bekele, M. G. Thomas, M. E. Weale, N. Bradman, and D. M. Swallow. 2009. Multiple rare variants as a cause of a common phenotype: several different lactase persistence associated alleles in a single ethnic group. *J Mol Evol* 69(6):579-88.

Newman, J. L. 1995. *The peopling of Africa: a geographic interpretation*: Yale University Press New Haven and London.

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

Pereira, L., L. Gusmão, C. Alves, A. Amorim, and M. J. Prata. 2002. Bantu and European Y-lineages in Sub-Saharan Africa. *Ann Hum Genet* 66:369-78.

Pereira, L., V. Macaulay, A. Torroni, R. Scozzari, M. J. Prata, and A. Amorim. 2001. Prehistoric and historic traces in the mtDNA of Mozambique: insights into the Bantu expansions and the slave trade. *Ann Hum Genet* 65:439-58.

Plaza, S., A. Salas, F. Calafell, F. Corte-Real, J. Bertranpetit, A. Carracedo, and D. Comas. 2004. Insights into the western Bantu dispersal: mtDNA lineage analysis in Angola. *Hum Genet* 115:439-47.

Quintana-Murci, L., H. Quach, C. Harmant, F. Luca, B. Massonnet, E. Patin, L. Sica, P. Mouguiama-Daouda, D. Comas, S. Tzur, O. Balanovsky, K. K. Kidd, J. R. Kidd, L. van der Veen, J. M. Hombert, A. Gessain, P. Verdu, A. Froment, S. Bahuchet, E. Heyer, J. Dausset, A. Salas, and D. M. Behar. 2008. Maternal traces of deep common ancestry and asymmetric gene flow between Pygmy hunter-gatherers and Bantu-speaking farmers. *Proc Natl Acad Sci U S A* 105:1596-601.

Rexová, K., Y. Bastin, and D. Frynta. 2006. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften* 93:189-94.

Salas, A., M. Richards, T. De la Fe, M. V. Lareu, B. Sobrino, P. Sanchez-Diz, V. Macaulay, and A. Carracedo. 2002. The making of the African mtDNA landscape. *Am J Hum Genet* 71:1082-111.

Soodyall, H., and T. Jenkins. 1993. Mitochondrial DNA polymorphisms in Negroid populations from Namibia: new light on the origins of the Dama, Herero and Ambo. *Ann Hum Biol* 20:477-85.

Tishkoff, S. A., M. K. Gonder, B. M. Henn, H. Mortensen, A. Knight, C. Gignoux, N. Fernandopulle, G. Lema, T. B. Nyambo, U. Ramakrishnan, F. A. Reed, and J. L.

Mountain. 2007a. History of click-speaking populations of Africa inferred from mtDNA and Y chromosome genetic variation. *Mol Biol Evol* 24:2180-95.

Tishkoff, S. A., F. A. Reed, A. Ranciaro, B. F. Voight, C. C. Babbitt, J. S. Silverman, K. Powell, H. M. Mortensen, J. B. Hirbo, M. Osman, M. Ibrahim, S. A. Omar, G. Lema, T. B. Nyambo, J. Ghori, S. Bumpstead, J. K. Pritchard, G. A. Wray, and P. Deloukas. 2007b. Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39:31-40.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, and S. M. Williams. 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035-44.

Torniainen, S., M. I. Parker, V. Holmberg, E. Lahtela, C. Dandara, and I. Jarvela. 2009. Screening of variants for lactase persistence/non-persistence in populations from South Africa and Ghana. *BMC Genet* 10:31.

Veeramah, K., B. Connell, N. Pour, A. Powell, C. Plaster, D. Zeitlyn, N. Mendell, M. Weale, N. Bradman, and M. Thomas. 2010. Little genetic differentiation as assessed by uniparental markers in the presence of substantial language variation in peoples of the Cross River region of Nigeria. *BMC Evol Biol*. 2010; 10: 92.

Verdu, P., F. Austerlitz, A. Estoup, R. Vitalis, M. Georges, S. Thery, A. Froment, S. Le Bomin, A. Gessain, J. M. Hombert, L. Van der Veen, L. Quintana-Murci, S. Bahuchet, and E. Heyer. 2009. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol* 19:312-8.

Wood, E. T., D. A. Stover, C. Ehret, G. Destro-Bisol, G. Spedini, H. McLeod, L. Louie, M. Bamshad, B. I. Strassmann, H. Soodyall, and M. F. Hammer. 2005. Contrasting patterns of Y chromosome and mtDNA variation in Africa: evidence for sex-biased demographic processes. *Eur J Hum Genet* 13(7):867-76.

PART 3

Human microevolution and the Atlantic slave trade: A case study from São Tomé

3.1. Introduction

The slave trade in Africa was already a common feature long before European exploitation of other continents, driving the forced migration of large number of individuals within and outside the continent (Klein 1999). Indeed, the slave trade was an integral part of African societies and states, like the Wolof, the nomadic Tuareg or the Arab traders (Thomas 1997, Klein 1999). The Atlantic slave trade undertaken by Europeans from the final of the fifteenth to the nineteenth centuries should then be seen as an extension of the internal African markets (Klein 1999). Its developments would make the shipping of African slaves across the Atlantic ocean one of the largest migrations in human history (Curtin et al. 1995).

The coercive population relocations launched by the Atlantic slave trade have led to new forms of social encounter with important implications in the recent evolution of human cultural and biological variation. Indeed, the merging of populations with different cultural and biological backgrounds often led to cultural syncretism, creoles languages and miscegenation. Among the specific features of this bio-cultural process are: its association with networks of maritime trade; the concentration of large scale migrations in a relatively short time; the occurrence of episodes of true colonization whereby new immigrants clear outnumbered native populations; and the displacement of large numbers of individuals to new disease environments (Curtin 1998).

3.1.1 The origins of the Atlantic slave trade

Portuguese voyages in the fifteenth century along the West African coast opened the doors to all the European commercial networks that were later to evolve in this region. The primary interest driving the Portuguese exploration of the West African coast was the gold, with slaves and products like pepper and ivory constituting only secondary concerns (Klein 1999). During the early decades of the trade, the purchased slaves were mainly sent to Europe to be used as domestic servants and as a convenient merchandise with which to buy gold from African traders on the coast of modern Ghana (Klein 1999). By the end of the fifteenth century, a major alteration in the use of the slaves occurred. With the introduction of sugar production to eastern Atlantic islands, like São Tomé, and later to the New World, slaves became a crucial factor in agricultural production, and the European major commercial focus shifted from gold trade to slavery (Klein 1999). The whole process of sugar production required large concentrations of workers and massive labour migration to sugar plantations was always needed. This labour was mainly supplied by slaves brought directly from Africa and, it was the

trading of these slaves through the Atlantic ocean, that gave rise to the Atlantic slave trade (Klein 1999). With time, several economic cycles were taking place in European colonies spread through the Atlantic. Sugar plantations gave place to other industries like gold and diamond mining as well as coffee and cacao plantations. Even though all these industries were also depended on African slave labour, the high interest of Europe for sugar constituted the incentive leading the Atlantic trade to the dimensions it reached (Klein 1999).

3.1.2 The origins of African slaves

Over nine million slaves were shipped across Atlantic between 1451 and 1870 (Curtin 1969). Following Curtin's figures, less than 5% landed in what is now US, 42% were sold to plantation owners on the sugar producing islands of Caribbean, 38% were shipped to Brazil and between 10 and 20% died en route (Curtin 1969).

The geographic sources of the slave trade shifted from one part of the African coast to another, depending on the African political and military conditions or on the development of new trade routes from the interior (Bohannan and Curtin 1995). Military initiatives often made available a high number of prisoners taken in warfare but constituted only point situations in the slave trade. A steady nature of slave exports were often related with political initiatives like village raiding of unprotected peasant groups, and persons condemned to slavery by their respective communities for various reasons (Klein 1999). The increasing American demand for slaves of the eighteenth and nineteenth centuries and the consequent rise in the slave prices triggered the search for slaves in places previously not explored, like the more interior regions of Africa and the Southeast Africa (Klein 1999).

On the basis of the historical records of the slaves shipped from specific geographic points, 4 major regions of the African coast could be distinguished as sources of slaves in the Atlantic trade (Figure III.1):

A- The region in the Atlantic West Africa extending from the Senegambia area to present-day Ivory Coast.

B- The region comprised between current-day Ghana and Gabon, including the so- called Gold Coast (actual Ghana) and the coasts defined by the Bight of Benin, from Togo to Western Nigeria (also known as Slave Coast), and the Bight of Biafra (from the Niger river delta of eastern Nigeria to the Cameroon).

C- The region from southern Gabon to Angola, constituting the most important source of African slaves for America from the sixteenth century until the late nineteenth century. Estimates indicate that between 1700 and the 1860s, 3.8 million of slaves left this region (Klein 1999).

D- The Southeast Africa (Mozambique), entering in the Atlantic slave trade in a steady fashion only in the nineteenth century, representing at that time the third largest supplier of slaves to America (only behind West Central Africa and the area of the bight of Benin) (Klein 1999).

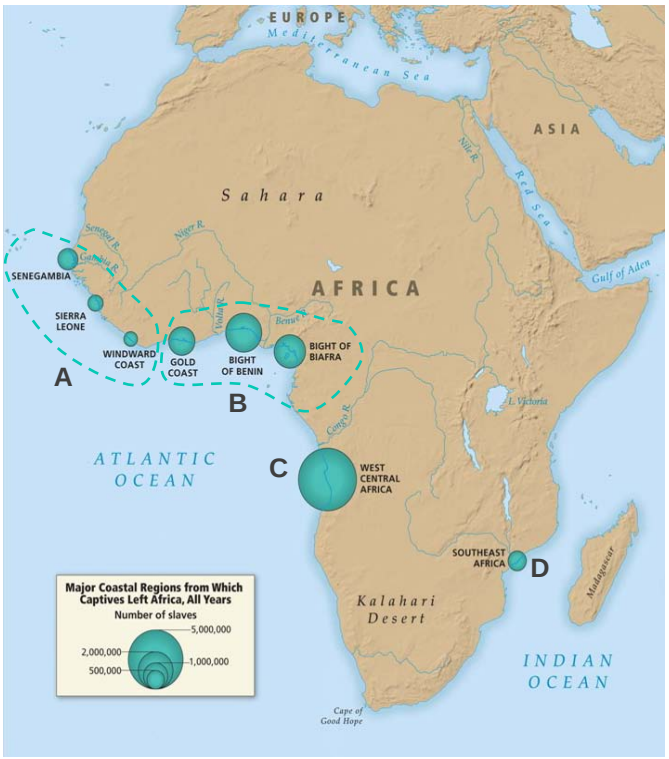


Figure III.1 Major coastal regions from which slaves left Africa during the Atlantic slave trade (adapted from http://www.slavevoyages.org/)

The figures taken from historic records are very informative about the origin of the African slaves in broad geographic lines. However they tend to refer to coastal areas from where the slaves were shipped, rather than their places of origin. Thus, when slaves were originated far from the coastal shipping points, it become not possible to precise their real place of origin (Curtin 1969). The inland capture of slaves is well described in regions where rivers were important for transporting slaves, like the Senegal and Gambia Rivers, the Niger Delta and the Congo Basin. In these cases, most of the slave exports originated in the far interior, beyond the shores and the head of navigation of rivers (Curtin et al. 1995, Thomas 1997, Klein 1999). Other approaches have been applied in order to circumvent the limitations related with the historic records about the place of origin of the African slaves.

Archaeological studies have been undertaken with the aim of identifying first generation captives in burial assemblages located in places of destination of African slaves (Handler 1994). One of the features analysed is the occurrence of intentionally modified teeth. However, some of these features are not regional-specific enough to allow precise identification of the places of origin (Handler 1994). Another technique that has been used in these studies is the isotope analysis of skeletons and teeth. Originally developed in geological and environmental sciences, isotope analyses are now being applied in archaeology to reconstruct the diets and movements of people in the past (Bentley 2006). The combination of carbon, nitrogen, oxygen, and strontium isotope in the teeth and skeleton provides a tool to identify their place of origin, by allowing the reconstruction of the diets and of physic characteristics of the place where they have originated. However, these analyses are hampered by the fact that similar combinations of isotope ratios could be found in geographically distant places.

Genetic information has a high potential to address several unanswered questions about the Atlantic slave trade. There are still some technical difficulties related with the genetic analysis of archaeological remains. Nevertheless, the major added value of the genetic data is that it allows the inference of the most likely geographic origin of the ancestors of present-day populations that emerged during the Atlantic slave trade. These analyses are based on the occurrence of alleles/haplotypes that are relatively specific to some restricted geographic areas of Africa (Nagel and Ranney 1990, Salas et al. 2005).

3.1.3 The plantation complex and the Atlantic slave trade

It was in the eastern Mediterranean, during the Crusades, that Europeans learnt about sugar cane cultivation and processing (Figure III.2). On the basis of eastern Mediterranean models of sugar production, Portuguese and Spanish established their own plantations in the Atlantic islands like the Canary Islands, and Madeira and São Tomé archipelagos (Curtin et al. 1995, Klein 1999).

Figure III.2: Migration of sugar cultivation from Asia into the Atlantic (retrieved from http://www.slavevoyages.org/)

Only in São Tomé was the plantation model characterized by massive slave work imported from the African mainland, with large-scale plantations aimed to supply the distant European markets. The political control of the production was mainly located in Europe (Curtin et al. 1995). The labourer populations of the plantation colonies were not self-sustained by natural increase and a steady supply of slaves was required. As Vansina wrote "In this way, the plantation complex consumed people, just as others industries consume raw materials" (Curtin et al. 1995). This form of production constitutes a paradigm of what is called the "Plantation complex" or the "Atlantic system", that served as model for the plantations that were later to be found in New World (Curtin 1998).

3.1.4 Populations emerging in the context of the Atlantic slave trade as models of human microevolution: the case-study of São Tomé

The impact that the entry of Europeans and Africans had in their arriving places was variable and dependent on the presence or absence of pre-existent populations, the strategies of colonization of each European nation, the availability of products to explore, and the suitability of the soils to different plantations (Klein 1999). When addressing the dynamics underlying the emergence of new populations during the Atlantic slave trade, one may follow two basic approaches. One option is to undertake general comparative studies involving different geographic settings and historical times. The alternative is to carry out monographic approaches focused on the thorough characterization of representative cases.

In this work we undertook a fine scale analysis of the genetic structure of the small island of São Tomé (832 km^2), located in the heart of the Gulf of Guinea, 300 km from the nearest African seashore. The island constituted an important point in the Atlantic slave trade both as slave entrepôt for the assemblage and redistribution of slaves bound for Lisbon, Elmina (in present-day Ghana) and the Americas, as well as the place of destination of slaves that were required to work in the island mainly in agricultural plantations. Due to the rapid development of several sugar cane plantations by the beginning of the sixteenth century, São Tomé has been one of the first examples of the typical plantation complex that spread into the tropical New World (Curtin 1998).

With a present population of about 150,000 individuals, S. Tomé was uninhabited when Portuguese sailors arrived there in the early 1470s. The peopling of the island was made by Europeans, mainly Portuguese, and Africans arrived from different regions of continental mainland (de Almeida 1962). Different economic cycles triggered several waves of African migrations to São Tomé and the present features of the island population are the result of the interactions that have emerged from this complex settlement. The diversity of contributions to the peopling of São Tomé has promoted intense cultural interactions that resulted in the emergence of two distinct autochthonous creoles (São-Tomense and Angolar) that are still widely spoken, alongside with the official Portuguese language (Hagemeijer 2009). Both creoles languages derive significant portions (> 80%) of their lexicon from Portuguese and are likely to descent from a single proto-creole that originally had important syntactic affinities with Edo from Nigeria and later incorporated contributions from western Bantu languages, like Kikongo and Kimbundu, spoken in the Kongo-Angolan area (Hagemeijer 2009) (Table1). Angolar is particularly notorious for its high proportion (10–20%) of Bantu words, claimed to be derived from the Kimbundu (Maurer 1992, Lorenzino 1998) (Table1).

TableIII.1 Examples illustrating the influence of Portuguese and African lexicon in the formation of the creoles from São Tomé (Source: Hagemeijer 2009)

	São-Tomense	**Angolar**	**Etymology**	**English**
Portuguese (Ptg.) influence	[kaso]	[kaθo][1]	Ptg. cão	dog
Edo influence	[obo]	[obo]	Edo *ógo*	forest
Kikongo (Kk.) influence	[bobo]	[bobo]	Kk. booba	ripe
Specific Kimbundu (Kb.) influence in Angolar	[piʃi]	[kikie]	Ptg. Peixe Kb. kikele	Fish

[1] the sound θ is pronounced "th" as in "thin".
[2]the sound ʃ is pronounced "sh" as in "fish".

Previous genetic studies in São Tomé have mostly focused in the analysis of uniparental markers through the characterization of mtDNA sequence diversity (Mateu et al. 1997, Trovoada et al. 2003), Y-chromosome short tandem repeat (STR) (Trovoada et al. 2001) and Y-Unique Event Polimorphisms (UEP) (Gonçalves et al. 2007, Trovoada et al. 2007) variation. These studies have shown that the population of São Tomé had retained the global high levels of genetic diversity that are generally observed in the continental mainland, as expected from a settlement pattern based on massive importation of slaves from different African regions. Tomás et al. (2002) estimated the relative contributions of major geographic sources of African slaves to São Tomé by studying different β-globin cluster haplotypes bearing the β-globin S mutation (β*S). The high levels of geographic segregation of the different β*S-haplotypes in different major areas of slave recruitment, allowed an approximate assessment of the relative contribution of each of those areas to S. Tomé. The following β*S -haplotypes distribution was found: 36.4% Bantu, 52.3% Benin, 4.5% Cameroon, 4.5% Senegal, and 2.3% atypical. The high percentage of the Benin haplotype, which occurs in populations from present-day Ghana to northern Gabon, showed that the impact of Central-West Africa to the present population of São Tomé was higher than previously thought (Tomás et al. 2002). Genetic markers have also been used to evaluate the extent of African-European admixture in present-day São Tomé. A study on autosomal markers informative for European admixture estimated the impact of European in present population of São Tomé to be low (10.7%±0.9%) (Tomás et al. 2002).

Studies using uniparentally inherited genetic markers have additionally shown that European males and females did not contribute equally to the genetic pool of Santomeans. Indeed, while the contribution of putative European descent mtDNA lineages to the Santomean maternal pool seems to have been virtually nil (Mateu et al. 1997, Trovoada 2003), 12.5% to 23.9% of the male lineages found in present genetic pool of S. Tomé were found to have an West Eurasia/European origin (Gonçalves et al. 2007, Trovoada et al. 2007). The studies on uniparental markers were based on different sampling approaches. Mateu et al. (1997) characterized São Tomé's mtDNA sequence variation by treating the island as a single sampling unit. Trovoada et al. (2001, 2003) have studied the Y- chromosome microsatellites and mtDNA variation in preconceived ethno-linguistic entities from São Tomé by characterizing self-identified individuals from the São-Tomense and Angolar linguistic communities and from an additional group, the Tonga, descendents of contracted labourers that started arriving on the island in the mid-nineteenth century. Using this approach, they found signs of a slight genetic microdifferentiation and a reduction of mtDNA and Y-chromosome diversity in self-reported Angolares relative to the other population groups.

Given the complexity of the settlement history of São Tomé, it is clear that, as with most studies of population genetics, a major problem in any analysis of its current patterns of genetic variation is that the choice of a sampling strategy may strongly influence the results. Here, we attempt to describe and interpret the genetic structure of the island of São Tomé without relying on predefined ethnic, anatomical or geographical categories. To this end, we inverted the sequence by which the relations between genetic and cultural variation are usually investigated. We first sorted individuals into genetic clusters based solely on multilocus microsatellite genotypes and then compared the distribution of additional phylogeographic informative markers across the inferred genetic clusters. The results of this study are presented in the *Article 4*. A more detailed discussion of some aspects of the results is presented afterwards as a commentary.

References

Bentley, R. A. 2006. Strontium isotopes from the earth to the archaeological skeleton: a review. *J. Arch. Meth. Theo* 13:135-187.

Bohannan, P., and P. Curtin. 1995. *Africa and Africans*: Waveland Press.

Curtin, P. 1969. *The Atlantic Slave Trade*: a census. Madison: University of Wisconsin Press.

Curtin, P., S.Feierman, L. Thompson, and J. Vansina. 1995. *African history: from earliest times to independence*. London: Longman.

Curtin, P. 1998. *The rise and fall of the plantation complex*. Cambridge: Cambridge University Press.

de Almeida, A. 1962. *Da origem dos Angolares habitantes da ilha de S. Tomé*. Separata das "Memórias". Academia de Ciências de Lisboa, Tomo VIII.

Gonçalves, R., H. Spinola, and A. Brehm. 2007. Y-chromosome lineages in São Tomé e Principe islands: evidence of European influence. *Am J Hum Biol* 19:422-8.

Hagemeijer, Tjerk. 2009. As línguas de S. Tomé e Príncipe. Revista de Crioulos de Base Lexical Portuguesa e Espanhola, Vol. 1.1.

Handler, J. S. 1994. Determining African birth from skeletal remains: a note on tooth mutilation. *Hist Arch* 28:113-119.

Klein, H. S. 1999. *The Atlantic slave trade*. Cambridge: Cambridge University Press.

Lorenzino, G. 1998. *Angolar Creole Portuguese*: Newcastle: Lincom Europa.

Mateu, E., D. Comas, F. Calafell, A. Perez-Lezaun, A. Abade, and J. Bertranpetit. 1997. A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea. *Ann Hum Genet* 61:507-18.

Maurer, P. 1992. L'apport lexical bantou en Angolar. *Afrikanische Arbeitspapiere* 29:163–74.

Nagel, R. L., and H. M. Ranney. 1990. Genetic epidemiology of structural mutations of the beta-globin gene. *Semin Hematol* 27:342-59.

Salas, A., A. Carracedo, M. Richards, and V. Macaulay. 2005. Charting the ancestry of African Americans. *Am J Hum Genet* 77:676-80.

Thomas, H. 1997. The Slave Trade: *The story of the Atlantic Slave Trade*: 1440-1870. New York: Simon & Schster.

Tomás, G., L. Seco, S. Seixas, P. Faustino, J. Lavinha, and J. Rocha. 2002. The peopling of São Tomé (Gulf of Guinea): origins of slave settlers and admixture with the Portuguese. *Hum Biol* 74:397-411.

Trovoada, M. J., C. Alves, L. Gusmão, A. Abade, A. Amorim, and M. J. Prata. 2001. Evidence for population sub-structuring in São Tomé e Príncipe as inferred from Y-chromosome STR analysis. *Ann Hum Genet* 65:271-83.

Trovoada, M. J., L. Pereira, L. Gusmão, A. Abade, A. Amorim, and M. J. Prata. 2003. Pattern of mtDNA variation in three populations from São Tomé e Príncipe. *Ann Hum Genet* 68:40–54.

Trovoada, M. J., L. Tavares, L. Gusmão, C. Alves, A. Abade, A. Amorim, M. J. Prata. 2007. Dissecting the genetic history of São Tomé e Príncipe: a new window from Y-chromosome biallelic markers. *Ann Hum Genet* 71(Pt 1):77-85.

3.2. Results and Discussion

Article 4

Coelho, M., C. Alves, V. Coia, D. Luiselli, A. Useli, T. Hagemeijer, A. Amorim, G. Destro-Bisol, and J. Rocha. 2008. Human microevolution and the Atlantic slave trade: a case study from São Tome. *Curr Anthropol* 49:134-143.

Human Microevolution and the Atlantic Slave Trade

A Case Study from São Tomé

Margarida Coelho, **Cíntia Alves Valentina Coia**, **Donata Luiselli**, **Antonella Useli**, **Tjerk Hagemeijer**, **António Amorim**, **Giovanni Destro-Bisol**, and **Jorge Rocha**

Instituto de Patologia e Imunologia Molecular da Universidade do Porto (IPATIMUP), R. Dr. Roberto Frias s/n, 4200-465 Porto, Portugal (jrocha@ipatimup.pt) (Coelho, Alves, Rocha)/Dipartimento di Biologia Animale e dell' Uomo, Università "La Sapienza," Roma, Italy (Coia, Destro-Bisol)/ Dipartimento di Biologia Evoluzionistica Sperimentale, Università di Bologna, Italy (Hagemeijer). 15 VII 07

Populations derived from the Atlantic slaving process provide unique opportunities for studying key evolutionary determinants of current patterns of human cultural and biological variation. Examination of the genetic patterning of the small plantation island of São Tomé (Gulf of Guinea) using a study design that avoids the use of preconceived ethno-linguistic labels to define genetic sampling units reveals that, despite the fact that maximum distance between any two sampled sites is less than 50 km, the island has an unusual level of genetic structure that is mainly caused by the grouping of Angolar Creole-speakers in a separate cluster carrying a distinctive imprint of genetic drift. This pattern may have been shaped by a kin-structured founder effect associated with the flight of a patrilineal clan of rebel slaves who established a remarkably successful maroon community in the vicinity of the plantation complex. The observation that population-discontinuous jumps may occur even under social conditions of massive coercive amalgamation provides an illustration of the way in which human clusters emerge and eventually shape the genetic background of human populations.

The worldwide maritime revolution of the fifteenth and early sixteenth centuries, together with the Atlantic slave trade, triggered an unprecedented wave of population relocations whose implications can only be properly assessed by multidisciplinary approaches. However, most studies of the new patterns of maritime trade have been based on the contributions of the historical and social sciences; much less effort has been dedicated to addressing their biological impact on human populations. And yet, human populations derived from the Atlantic slaving process are natural laboratories that provide a unique opportunity for integrative studies aiming at the identification of key evolutionary determinants of current patterns of human cultural and biological variation. Here we present a detailed analysis of the genetic structure of São Tomé, the major island of the São Tomé e Príncipe archipelago, located in the Gulf of Guinea (1° N, 7° E), which, because of its limited dimensions (1,000 km^2), small population size (130,000), and relatively recent complex peopling process, may be considered an excellent model for assessing the microevolutionary impact and biocultural implications of the slave trade.

 0011-3204/2008/4901-0007$10.00. DOI: 10.1086/524762

Historical Outline

The previously uninhabited archipelago of São Tomé e Príncipe, discovered in 1471–72 by Portuguese sailors, was peopled mainly by slaves imported from different regions of the African mainland, quickly becoming an important stepping-stone in the Atlantic slave trade and one of the first examples of the plantation complex that was later to be found in the New World (Tenreiro 1961; Curtin 1998). The role of São Tomé e Príncipe as both a plantation complex and a slave entrepôt has given rise to an intricate settlement pattern that may be conveniently divided into three major periods (Tenreiro 1961). The first period, ranging from the beginning of permanent settlement by the Portuguese in 1493 to the end of the sixteenth century, was largely dominated by sugarcane production. According to historical records, most slaves brought to São Tomé during the earliest stages of colonization originated from the Niger delta as a result of trade with the ancient kingdom of Benin (Nigeria). The shift from the homestead society that characterized the first two decades to a quickly growing plantation economy on São Tomé (Garfield 1992) is roughly correlated with the expansion of slave recruitment first to the Congo and slightly later to Angola. Slave transshipment also became less confined to the limits of the Gulf of Guinea and was increasingly integrated into the intercontinental trade networks with the Americas (Klein 1999).

By the end of the sixteenth century the sugar economy was already declining as a result of competition from the emergent Brazilian sugar industry and consecutive raids on plantations organized by escaped slaves (Tenreiro 1961). This decline marks the beginning of a second period in the history of São Tomé, which encompassed most of the seventeenth and eighteenth centuries. During this period the archipelago became essentially a provisioning stop for the slave ships crossing the Atlantic (Tenreiro 1961).

The third period in the history of São Tomé began in the early nineteenth century. This phase, which may be considered one of recolonization, was characterized by an unprecedented economic and demographic boom associated with the introduction of coffee and, more important, cacao plantations (Tenreiro 1961). With the abolition of the Atlantic slave trade, most plantation labor power was to be provided by indentured la-

borers (*serviçais*) on five-year contracts, mostly recruited in the Portuguese colonies of Angola, Mozambique, and Cape Verde.

Portuguese settlers have had a profound political and cultural influence on São Tomé e Príncipe's society during almost five centuries of colonization but remained a small fraction of the total population of the archipelago. Despite several measures implemented by the crown to encourage miscegenation (Tenreiro 1961), the impact of European admixture in the present-day population has been estimated to be as low as 11% (Tomás et al. 2002).

Linguistic Diversity

The diversity of contributions to the settlement of São Tomé e Príncipe has led to a surprising level of linguistic variation. Alongside the official Portuguese language, three autochthonous creole languages are spoken in the archipelago: São-Tomense (Lungwa Santome) and Angolar (Lunga Ngola), both spoken on São Tomé, and Principense (Lung'ie), spoken on Príncipe. The three creoles derive significant portions (> 80%) of their lexicon from Portuguese and are likely to descend from a single proto-creole of the Gulf of Guinea that originally had important syntactic affinities with Edo and other typologically similar languages (Ferraz 1979; Maurer 1992; Lorenzino 1998; Hagemeijer 1999). Additional contributions from western Bantu languages may have been incorporated slightly later and had more impact on the lexicon and phonology than on syntax (Hagemeijer 1999).

Contemporary São-Tomense, which exhibits lexical affinities with both Kikongo (spoken around the Congo River) and Edo, is spoken by the descendents of the slaves that were freed in the initial stages of the history of São Tomé (Forros or "sons of the Island"). Príncipense is currently spoken by a minority on the island of Príncipe and has remained less influenced by subsequent inputs from Bantu-speakers (Hagemeijer 1999). Finally, Angolar is notorious for its high proportion (10–20%) of Bantu words, claimed to be derived from the Kimbundu spoken in Angola (Maurer 1992; Lorenzino 1998). Its speakers, commonly called Angolares, are generally considered a separate ethno-linguistic group concentrated in fishing communities on the northwestern and southwestern shores of São Tomé, but their origin and settlement history remain controversial (Seibert 1998).

Aims and Study Design

Previous genetic studies in São Tomé have mostly focused on the analysis of uniparental markers using various sampling approaches. Mateu et al. (1997) characterized mtDNA sequence variation by treating the island as a single sampling unit. Trovoada et al. (2001, 2003) have studied the Y-chromosome microsatellites and mtDNA among self-reported donors from preconceived ethno-linguistic groups. Here we attempt to describe and interpret the genetic structure of São Tomé without relying on predefined ethnic, anatomical, or geographical categories. Our major goal was to understand how different evolutionary factors had shaped the current genetic variation in São Tomé by addressing the following questions: Is there any detectable level of population structure in São Tomé? How many different genetic clusters, if any, may account for the observed structure? How did they originate, and how do they relate to the available linguistic and historical information on the peopling of the island?

To this end, we coupled a comprehensive sampling strategy with a two-tiered approach based on the sorting of sampled individuals into genetic clusters using multilocus microsatellite genotypes and an examination of variation in additional phylogeographic informative markers across inferred genetic clusters. [Details of this methodology may be found in the electronic edition of this issue on the journal's web site.] We found São Tomé to be characterized by a remarkable level of genetic structuring best captured by two clusters that are more closely related to language differentiation than to geographic distance. These results demonstrate that methods appropriate for genetic cluster analysis in widely separated regions may be useful in sorting individuals at the microgeographical level. More generally, our observations show that analyses based on genetically defined demic boundaries are a useful alternative to study designs relying on the use of preconceived groups as starting points for population-genetics comparisons.

Results

Autosomal Microsatellite Variation and Inference of Population Structure

We used the unsupervised algorithm implemented in STRUCTURE (Pritchard, Stephens, and Donnelly 2000; Falush, Stephens, and Pritchard 2003) to group into genetic clusters 394 unrelated individuals sampled in 14 localities across the island of São Tomé (fig. 1) and typed for 15 autosomal microsatellite markers. Clustering solutions of highest likelihood were found for numbers of clusters (K) equal to 2 and 3, indicating that the population of São Tomé is not genetically uniform and that a significant signal of structure can be detected even with a relatively small number of loci. The sample division corresponding to $K = 2$ was found to produce significantly higher-likelihood solutions than $K = 3$ (Wilcoxon two-sample test, two-sided $p < 0.0002$). Moreover, the posterior probability that the proper number of clusters is 2, calculated according to Pritchard, Stephens, and Donnelly (2000), was essentially 1. However, since current methods for inferring K provide only a rough guide for determining the real number of clusters (Pritchard, Stephens, and Donnelly 2000), we also explored cluster partitions with $K = 3$.

Representative estimates of the population structure are shown in figure 2, which displays the proportions of ancestry in inferred clusters for each individual at $K = 2$ and $K = 3$. At $K = 2$ (fig. 2, *a*), 70% of the sampled individuals (277/394) had fractions of ancestry in one of the two clusters (A and B) that were higher than 70%. When individuals from each location were assigned to the cluster in which they had

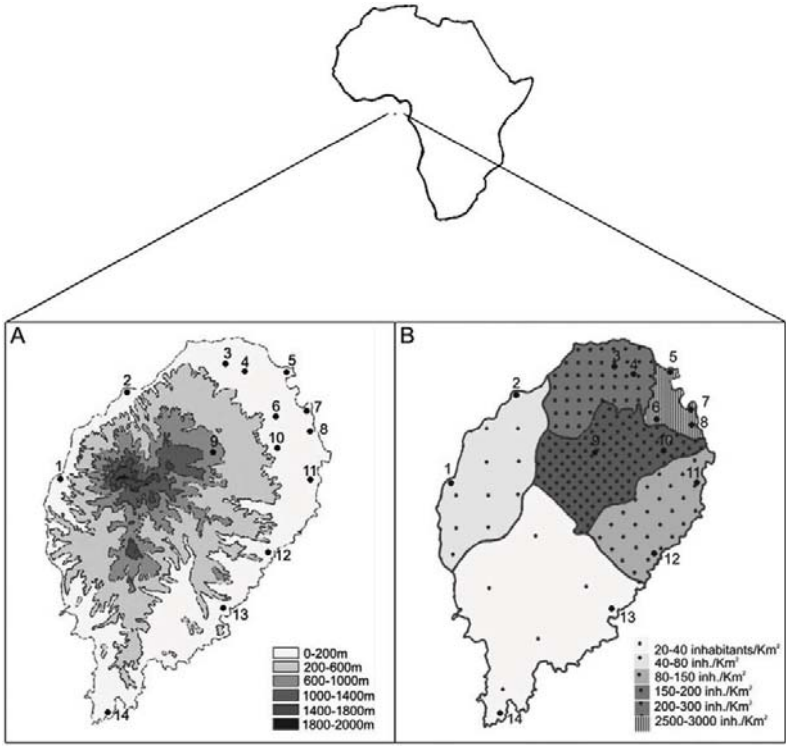


Figure 1. São Tomé: geographical position in the Gulf of Guinea, topography (*A*), demography (*B*), and location of sampling sites (adapted from Pinto et al. 2003). *1*, Santa Catarina; *2*, Neves; *3*, Guadalupe; *4*, Agostinho Neto; *5*, Praia Gamboa; *6*, Madalena; *7*, Cidade de São Tomé; *8*, Pantufo; *9*, Monte Café; *10*, Trindade; *11*, Santana; *12*, Ribeira Afonso; *13*, São João dos Angolares; *14*, Porto Alegre. Locations 4, 9, and 14 are plantations.

the highest proportion of ancestry, we found a highly significant association between cluster assignment and sampling location ($\chi^2 = 111.1$, $p < 0.001$, 13 df).

At K = 3 (fig. 2, *b*), cluster B split into two additional clusters (B1 and B2) that are not associated with geographic origin ($\chi^2 = 5.49$, $0.975 > p > 0.95$, 13 df). We interpret the splitting pattern as indicating that cluster B is more heterogeneous than cluster A, but it is less clear whether subclusters B1 and B2 may be considered meaningful genetic entities or just reflect random inference of population structure. When additional STRUCTURE runs were performed using only the individuals with the largest fractions of their genomes in cluster B, the best clustering solutions were invariably found at K = 1, indicating no further structure. While additional clusters may be found in the future by increasing both the sample sizes and the number of loci, we will focus on the interpretation of the major split between clusters A and B.

Figure 3, *a*, displays the geographic distribution of the average per-individual proportions of ancestry. Cluster A strongly predominates in villages such as São João dos Angolares and Ribeira Afonso, in the southeast, and Santa Catarina, in the northwest, where Angolar is known to be the predominant autochthonous language (Maurer 1995). In contrast, cluster B is linguistically more heterogeneous, including individuals from locations where São-Tomense is the predominant autochthonous language (such as Madalena, Cidade de São Tomé, and Trindade), together with descendents of *serviçais* who live on plantations (Agostinho Neto, Monte Café, and Porto Alegre) and speak or originally spoke languages from Mozambique, Angola, and Cape Verde. The fact that individuals from São João dos Angolares and Santa Catarina show the highest proportions of ancestry in cluster A, in spite of the relatively great distance between these villages, indicates that the genetic differentiation of cluster A is more closely related to language than to geographic distance.

Figure 3, *b*, is a synthetic map of the first principal com-

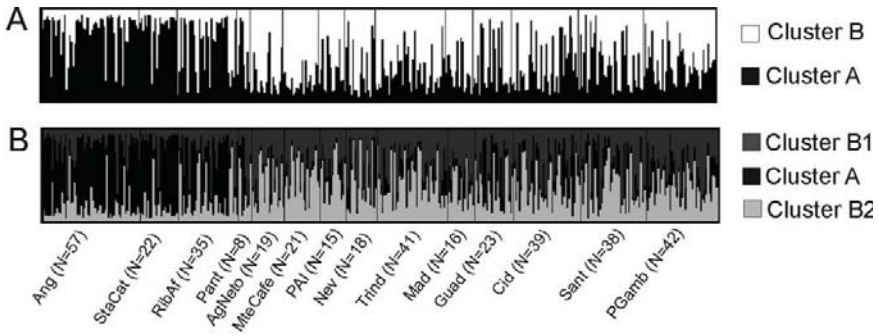


Figure 2. Distribution of individual proportions of ancestry in 14 sampled locations from the island of São Tomé assuming K = 2 (A) and K = 3 (B). Each individual is represented by a vertical line partitioned into K segments representing the individual's estimated membership fraction in each cluster. Analysis was performed without prior knowledge of sample provenance. *Ang*, São João dos Angolares; *StaCat*, Santa Catarina; *RibAf*, Ribeira Afonso; *Pant*, Pantufo; *AgNeto*, Agostinho Neto; *MteCafé*, Monte Café; *PA1*, Porto Alegre; *Nev*, Neves; *Trind*, Trindade; *Mad*, Madalena; *Guad*, Guadalupe; *Cid*, Cidade de São Tomé; *Sant*, Santana; *PGamb*, Praia Gamboa.

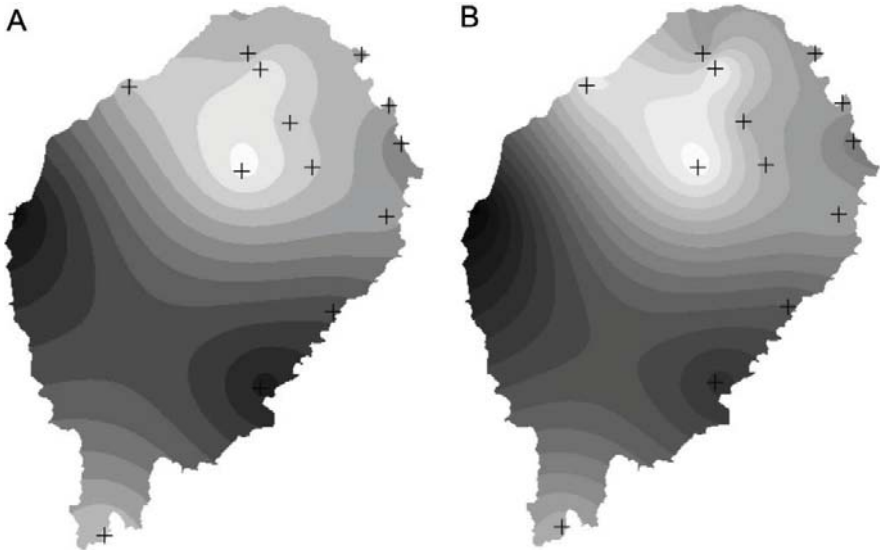


Figure 3. *A*, geographic distribution of average per-individual proportions of ancestry in clusters A and B. *Black*, highest average ancestry proportions in cluster A; *white*, highest average ancestry proportions in cluster B. *B*, synthetic map of first principal component resulting from the allele frequencies at autosomal microsatellite loci, extracting 23% of the total variance. Values are ranked according to color tonalities ranging from black to white. *Crosses*, sampling locations.

ponent derived from the distribution of autosomal microsatellite allele frequencies. The close parallel with the spatial distribution of genetic clusters confirms that the geographic patterning of genetic variation in São Tomé is mainly determined by the dichotomy between Angolares and non-Angolares. Additional confirmation is provided by the pattern of pairwise F_{st} genetic distances calculated between sampling locations. The significance of the divergence between Angolares (cluster A) and non-Angolares (cluster B) is further attested by the finding that, when compared with 11 additional African-derived samples using pairwise genetic distances, the two clusters were never grouped together in any bootstrap replication, showing a level of differentiation ($F_{st} = 0.03$) that is clearly above the average ($F_{st} = 0.01$).

Assuming that each cluster has undergone independent drift away from a common ancestral population, it is possible to explore the model of correlated allele frequencies implemented in STRUCTURE (*F* model) to calculate *F*, an analog of F_{st} that measures the genetic distance between each cluster and the ancestral population—which is expected to be inversely proportional to the size of the cluster (Falush, Stephens, and Pritchard 2003). In our data set, the *F* value for cluster A (0.060; 0.033–0.113) was found to be substantially higher than for cluster B (0.001; 0.000–0.017), consistent with increased drift in the Angolar group, in spite of the lack of large differences in per-locus heterozygosities in clusters A (0.77) and B (0.82). Although the model assumption may not hold, since the two clusters may not coalesce into a real ancestral population, this result can be interpreted as an informal indication that, in contrast to cluster B, cluster A is a genetic outlier relative to a single population amalgam.

Genetic Variation across Inferred Clusters

β-globin and Duffy blood group loci. To evaluate the relative contributions of different African areas of slave recruitment, we compared the distribution of the β-globin C allele (HBB*C) and the haplotypes associated with β-globin S allele (HBB*S) across the major groups defined by STRUCTURE, using 45 HBB*S and 25 HBB*C carriers. We found no significant differences between clusters A and B (table 1; $\chi^2 = 0$, 26, $p = 0.61$, 1 df) in the distribution of the haplotypes HBB*S-Benin, which predominates in Central-West Africa, and HBB*S-Bantu, which is particularly common in the Congo-Angola area, suggesting that the two clusters share HBB lineages with widespread origins on the African mainland. Consistent with this observation, no significant differences were found between the distributions of pooled lineages with a putative Central-West African origin (HBB*S-Benin+HBB*C) and the HBB*S-Bantu haplotype ($\chi^2 = 0.60$, $p = 0.44$, 1 df).

In order to evaluate the relative impact of European admixture, we additionally estimated the frequencies of the Duffy blood group FY*O allele in a subsample of 156 individuals previously assigned to the two clusters. The frequency of FY*O, which is virtually fixed in most sub-Saharan African populations (Cavalli-Sforza, Menozzi, and Piazza 1994), was found to be significantly lower in cluster B (0.86) than in cluster A (0.94) (table 1; $\chi^2 = 5.99$, $p = 0.01$, 1 df), indicating that the latter remained more closed to gene flow from European settlers.

Table 1. Distribution of HBB Types and FY Alleles among Genetic Clusters in São Tomé

	Cluster		
	A	B	Total
HBB type			
A-S_{Be}[a]	7	15	22
A-S_{Ba}[b]	9	11	23
A-C	16	9	25
Total	32	38	70
FY allele			
FY*O	160	122	282
Non-FY*O	10	20	30
Total	170	142	312

[a]Heterozygotes for the HBB*S-Benin haplotype.
[b]Heterozygotes for the HBB*S-Bantu haplotype.

MtDNA variation. MtDNA sequence variation was studied in a subset of 82 unrelated individuals who had more than 70% of their ancestry proportions in one of the two major clusters. All individuals belonged to known African haplogroups, with no European lineages detected. As in many African American communities, the global mtDNA profile from São Tomé consists of haplogroups that are characteristic of West Africa (L1b, L2c, L3b, and L3e4) and West-Central Africa (L1c and L3e except L3e4), reflecting the major contributions of these broad geographic regions to the settlement of the island. Interestingly, haplotype H30 (L3e1*), which is shared only by two African Americans outside São Tomé, and haplotypes H8 and H9 (L1c1), which are no more than one mutational step away from sequences that are typical of Western Pygmies (Batini et al. 2007), are virtually restricted to cluster A and may have originated from the poorly sampled area of the Congo Basin between northern Angola and southern Cameroon. Significant levels of differentiation between the two clusters were found only when molecular divergence between mtDNA sequences was not taken into account ($\Phi^{st} = 0.019$, $p = 0.07$; $F_{st} = 0.032$; $p = 0.001$). Compared with cluster B, cluster A has a sample configuration characterized by a relatively small number of different haplotypes, few rare haplotypes, and more haplotypes with intermediate frequencies (fig. 4, table 2). This pattern is likely to have been caused by genetic drift leading to an increase in the frequency of a small number of highly divergent haplotypes.

Y-chromosome variation. We typed 26 biallelic markers and 11 microsatellite loci in a subset of 47 unrelated males who

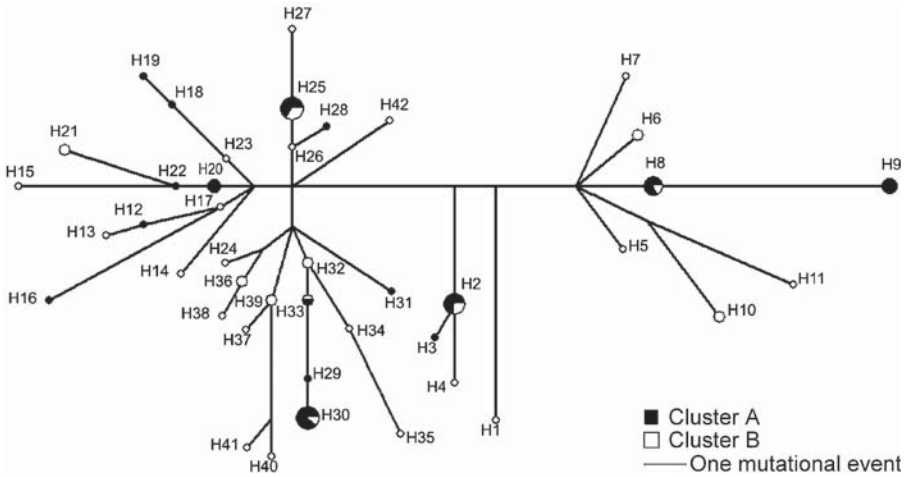


Figure 4. Network representing mtDNA HVR-I sequence variation in genetic clusters from São Tomé. Haplotypes are represented by circles with areas proportional to frequency.

had more than 70% of their ancestry proportions in one of the two major genetic clusters. In contrast with the mtDNA data, several Y-chromosome lineages clustered in haplogroups that are likely to have a European origin (G, J, K, and R1), indicating that gene flow from Europeans was essentially mediated by males. Moreover, consistent with the data from the Duffy blood group, this gene flow is likely to have had its greatest impact in cluster B, where the frequency of European lineages (27% [7/22]) is higher than in cluster A (8% [2/25]).

The levels of differentiation between the two clusters for Y-chromosome microsatellite haplotypes ($R_{st} = 0.158$; $p = 0.001$; $F_{st} = 0.231$; $p = 0.001$) were found to be substantially higher than for mtDNA haplotypes. The major aspect of this differentiation is a clear reduction in Y-chromosome diversity in cluster A (table 3, fig. 5), where a single modal haplotype (H9; haplogroup E3a [xE3a7]) represents as much as 60% (15/25) of the sampled lineages, suggesting the occurrence of a remarkable founder effect. In contrast with cluster A, cluster B has a much more heterogeneous lineage distribution, with several low-frequency divergent haplotypes.

The modal H9 haplotype in cluster A has a 15-21-11-11-13 stretch defined by microsatellites DYS19, DYS390, DYS391, DYS392, and DYS393 that is just one mutational step from a core 15-21-10-11-13 haplotype, which is generally considered to be a major founder lineage of the Bantu expansion (Thomas et al. 2000). This 15-21-11-11-13 stretch is relatively common in populations from West-Central Africa such as Cabinda (13%) and Angola (10%) but may have a much wider distribution in Africa, since it was also found with lower frequencies in Mozambique (1%) and Guinea-Bissau (2%) (Pereira et al. 2002; Beleza 2005; Rosa et al. 2006).

Discussion

As a contribution to the understanding of population structure at the microgeographical level, we studied the genetic patterns of the small plantation island of São Tomé using a Bayesian clustering approach that circumvents the usual difficulties associated with the need to rely on cultural, geographical, or anatomical categories for prior definition of sample units. We found evidence for a strongly uneven distribution of genetic variation that is best captured by two

Table 2. MtDNA HVR-I Sequence Diversity in Two Genetic Clusters from São Tomé

Cluster	N	K (K/N)	H (SD)	S (S/L)	Θ_k (95%CI)	Θ_{hom} (SD)	Θ_s (SD)	Θ_π (SD)
A	42	16 (0.38)	0.911 (0.021)	36 (0.1)	8.96 (4.75–16.59)	8.83 (2.48)	8.37 (2.72)	8.62 (4.51)
B	40	31 (0.78)	0.987 (0.008)	48 (0.13)	61.40 (31.36–125.13)	75.10 (51.07)	11.28 (3.59)	8.47 (4.45)
Pooled	82	42 (0.52)	0.961 (0.010)	57 (0.15)	33.84 (21.81–52.44)	22.66 (6.19)	11.45 (3.20)	8.63 (4.64)

Note: N, number of sequences; K, number of different haplotypes; H, haplotype diversity; S, number of segregating sites; L, sequence length (376 bp); Θ, mutation drift statistics calculated from K (Θ_k [Ewens 1972]), 1-H (Θ_{hom} [Kimura and Crow 1964]), S (Θ_s [Watterson 1975]), and the mean number of pairwise differences (Θ_π [Tajima 1983]).

Table 3. Y-chromosome Microsatellite Haplotype Diversity in Two Genetic Clusters from São Tomé

Cluster	N	K (K/N)	H (SD)	Θ_k (95%CI)	Θ_{hom} (SD)	Average Locus Diversity (SD)
A	25	10 (0.40)	0.600 (0.116)	5.68 (2.56–1.26)	1.13 (0.56)	0.218 (0.139)
B	22	18 (0.82)	0.978 (0.021)	43.88 (17.65–117.36)	43.37 (45.29)	0.526 (0.295)
Pooled	47	26 (0.55)	0.881 (0.044)	23.14 (13.12–40.88)	6.20 (2.82)	0.418 (0.230)

Note: Abbreviations are as in table 2. The DYS385 locus was omitted from this analysis.

clusters. One of the clusters (A) carries a clear imprint of genetic drift and predominates in villages where Angolar Creole is the major autochthonous language. The other cluster (B) remained open to more diverse genetic contributions and is linguistically much more heterogeneous. How do these patterns relate to the available historical and linguistic data? There are two major perspectives on the origin of the Angolares that lead to different expectations about the patterning of genetic diversity on the island. The first perspective corresponds to the long-held popular belief that the Angolares are a reasonably well-delimited group founded by a small number of survivors of the wreck of a slave ship from Angola just off the southeastern shore of São Tomé around 1540–50 (Seibert 1998; Caldeira 2004). The second perspective challenges this tradition on linguistic and historical grounds and considers the Angolares a conglomerate of groups with diverse origins that was formed by the assimilation of successive waves of slave runaways (Ferraz 1974; Seibert 1998; Caldeira 2004).

Our results are clearly at odds with the second hypothesis and point to the strong impact of a founder group in the shaping of the present genetic diversity of the Angolares. Otherwise, it would have been very difficult to detect any clear clustering pattern using our conservative approach, based on a limited set of 15 randomly chosen microsatellites. In fact, the STRUCTURE program typically requires a minimum of 20 to 60 loci to detect population subdivision on a continental scale, and therefore more markers would normally be expected to be required under the demographic conditions of coercive amalgamation found in São Tomé (Bamshad et al. 2003).

Whether the Angolar ancestors were survivors of a shipwreck is currently impossible to determine because of the limitations of the historical record. What the observed levels of genetic differentiation do suggest is that that the founder group probably consisted of fugitives from a specific region of Africa who had little initial contact with the remaining population. Moreover, it is possible that the initial Angolar fugitives were members of an extended kinship group and did not simply constitute a random sample of their original homeland population. Such a kin-structured founder event would have significantly enhanced genetic differentiation and increased the chances of cluster detection (Smouse, Vitzthum, and Neel 1981; Fix 1999). For example, using discriminant function analyses, Crawford et al. (2002) have shown that in

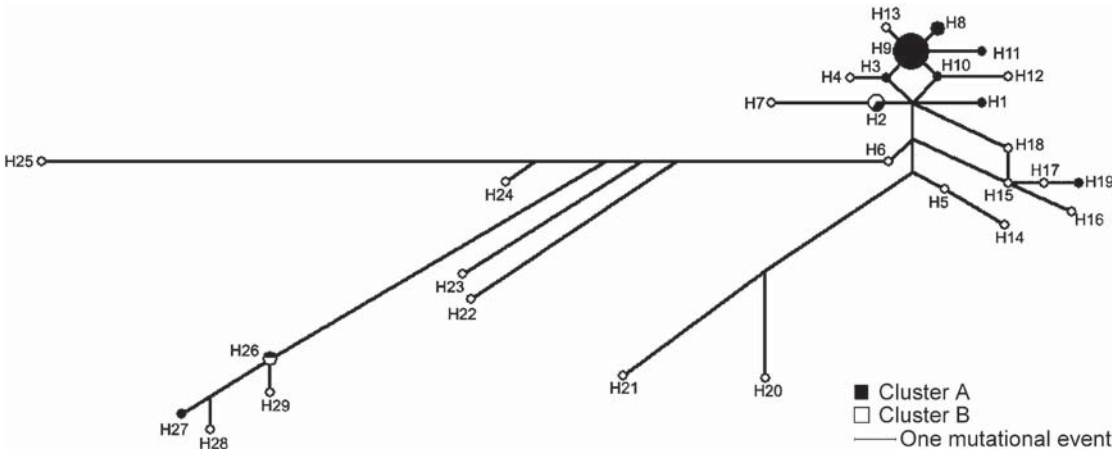


Figure 5. Network representing Y-chromosome haplotype variation in genetic clusters from São Tomé. Haplotypes are represented by circles with areas proportional to frequency.

Altai and Evenki populations of Siberia, individuals could be accurately assigned to clans with a relatively low number of autosomal minisatellite loci. In addition, it is likely that kinship ties between members of the founding clan were mainly patrilineal, since the Angolares show the sexually asymmetric apportionment of genetic diversity that is typically observed in many patrilineal descent groups with patrilocal residence (Seielstad, Minch, and Cavalli-Sforza 1998; Oota et al. 2001; Hage and Marck 2003; Chaix et al. 2007). In fact, while the Y-chromosomes of Angolares combine high levels of differentiation from non-Angolares ($F_{st} = 0.231$) with a pattern of low within-population diversity ($H = 0.60$), the mtDNA data show higher levels of within-population genetic diversity ($H = 0.911$) and lower levels of among-population differentiation ($F_{st} = 0.032$ [see table 2 and fig. 5]). The coupling of strong genetic drift with the absence of important subsequent male immigration appears to be the most likely explanation for the Y-chromosome pattern. By contrast, the levels of mtDNA diversity reflect a pattern of female-mediated introgression that is consistent with the likely shortage of available wives and with historical references to repeated episodes of abduction of women from plantations (Caldeira 2004). This pattern of restricted female migration is likely to explain the widespread origin of some mtDNA and hemoglobin lineages currently observed in the Angolares, as well as the emergence of Angolar Creole. In fact it seems implausible that the unique Bantu features found in Angolar could have derived from an ancestral proto-Creole simply through isolation (Hagemeijer 1999). It is more likely that these specific Bantu features are remnants of the original language spoken by fugitive founders who adopted a relexified version of the autochthonous creole, given the advantage of language understanding in dealing with the hostility of the surrounding plantation complex.

In summary, our results suggest that the current structure of the Angolar group has been strongly shaped by a kin-structured founder event resulting from the flight of a patrilineal clan of rebel slaves who established secondary contacts with the rest of the island mostly through restricted, female-mediated gene flow. In spite of this gene flow, the group has retained high levels of genetic differentiation that are more typical of a centralized band than a loose federation of refugees.

The impact of a founder effect on the genetic structure of the Angolares is further substantiated by consideration of the available demographic data. Today, Angolares are thought to amount to about 10,000 people (Seibert 1998), which roughly corresponds to the combined populations of São João dos Angolares (4,082), Ribeira Afonso (4,955), and Santa Catarina (2,199), representing approximately 8% of the 131,633 inhabitants of São Tomé (see fig. 1). Since these populations are not much smaller than those from other sampling locations, it is likely that drift effects were essentially caused by the initial small effective size of a founder group that underwent population growth as it settled on the outskirts of the plantation complex. The remarkable success of this flight is further attested by the fact that Angolares retained their autonomy long enough to arrive at a first understanding with the plantation authorities by 1693 and preserve their heritage down to the present, unlike many other runaways in São Tomé and most insular maroon communities emerging from the slave trade (Seibert 1998; Caldeira 2004; Curtin 1998).

Taken together, our results show that the microcosm of São Tomé captures many important features of the slave trade and illustrates the intricate way in which the economic and political order of an early plantation complex shaped the genetic structure and cultural heritage of the people who were forced to live in it. This study demonstrates that strong genetic microdifferentiation not only may occur under social conditions forcing massive amalgamation but also can be an inevitable consequence of freedom preservation and survival under such adverse conditions. Therefore, it may be only apparently paradoxical that a clustering method which is especially effective in detecting human subdivision on a continental scale has proved useful in capturing population structure on a small island like São Tomé.

In a recent analysis of population structure on a global level, Rosenberg et al. (2005) have shown that clusters and clines are not necessarily mutually exclusive representations of human genetic patterning and that the ability to detect clusters is due to small jumps in an otherwise clinal surface. Our focus on the small-scale study of genetic patterning in São Tomé provides an illustration of how such jumps and their consequences may arise. In this sense, local studies of recent populations emerging from the slave trade may prove to be useful windows into fundamental questions of human microevolution.

Acknowledgments

We thank Juliana Ramos of the Ministry of Health of the Democratic Republic of São Tomé e Príncipe, all the sample donors, and Nuno Ferrand for their help and collaboration during fieldwork in São Tomé. We are also grateful to Nuno Ferrand, Raquel Godinho, Sandra Beleza, and José Alberto Gonçalves for comments on the manuscript and advice in data analysis. This research was supported by Fundação para a Ciência e a Tecnologia (grants POCTI/42510/ANT/2001 to J. R. and SFRH/BD/22651/2005 to M. C.). D. L. was supported by MIUR (COFIN Grant 2003054059) and by RFO (ex 60% 2004, Università di Bologna), Project 491.

References Cited

Bamshad, M. J., S. Wooding, W. S. Watkins, C. T. Ostler, M. A. Batzer, and L. B. Jorde. 2003. Human population genetic structure and inference of group membership. *American Journal of Human Genetics* 72:578–89.

Batini, C., V. Coia, C. Battaggia, J. Rocha, M. M. Pilkington, G. Spedini, D. Comas, G. Destro-Bisol, and F. Calafell.

2007. Phylogeography of the human mitochondrial L1c haplogroup: Genetic signatures of the prehistory of central Africa. *Molecular Phylogenetics and Evolution* 43:635–44.

Beleza, S. 2005. Phylogenetic and demographic history of two human populations revealed by the analysis of two non-recombining segments of the genome: Y-chromosome and mitochondrial DNA. Ph.D. diss., University of Santiago de Compostela.

Caldeira, A. M. 2004. Rebelião e outras formas de resistência à escravatura na ilha de São Tomé (séculos XVI–XVIII). *Africana Studia* 7:101–36

Cavalli-Sforza, L. L., P. Menozzi, and A. Piazza. 1994. *The history and geography of human genes*. Princeton: Princeton University Press.

Chaix, R., L. Quintana-Murci, T. Hegay, M. Hammer, Z. Mobasher, F. Austerlitz, and E. Heyer. 2007. From social to genetic structures in Central Asia. *Current Biology* 17:43–48.

Crawford, M. H., J. McComb, M. Schanfield, and J. Mitchell. 2002. Genetic structure of pastoral populations of Siberia: The Evenki of Central Siberia and the Kizhi of Gorno Altai. In *Human biology of pastoral populations*, ed. W. Leonard and M. H. Crawford, 10–49. Cambridge: Cambridge University Press.

Curtin, P. D. 1998. *The rise and fall of the plantation complex*. Cambridge: Cambridge University Press.

Ewens, W. J. 1972. The sampling theory of selectively neutral alleles. *Theoretical Population Biology* 3:87–112.

Falush, D., M. Stephens, and J. K. Pritchard. 2003. Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequency. *Genetics* 164: 1567–87.

Ferraz, L. I. 1974. A linguistic appraisal of Angolar. In *In Memoriam António Jorge Dias*, 177–86. Lisbon: Junta de Investigações Científicas do Ultramar.

———. 1979. *The Creole of São Tomé*. Johannesburg: Witwatersrand University Press.

Fix, A. G. 1999. *Migration and colonization in human microevolution*. Cambridge: Cambridge University Press.

Garfield, R. 1992. *A history of São Tomé Island, 1470–1655: The key to Guinea*. San Francisco: Mellen Research University Press.

Hage, P., and J. Marck. 2003. Matrilineality and the Melanesian origin of Polynesian Y chromosomes. *Current Anthropology* 44:S121–27.

Hagemeijer, T. 1999. As ilhas de Babel: A crioulização no Golfo da Guiné. *Camões* 6:74–88.

Kimura, M., and J. F. Crow. 1964. The number of alleles that can be mantained in a finite population. *Genetics* 40: 725–38.

Klein, H. S. 1999. *The Atlantic slave trade*. Cambridge: Cambridge University Press.

Lorenzino, G. 1998. *Angolar Creole Portuguese*. Newcastle: Lincom Europa.

Mateu, E., D. Comas, F. Calafell, A. Pérez-Lezaun, A. Abade, and J. Bertranpetit. 1997. A tale of two islands: Population history and mitochondrial DNA sequence variation of Bioko and S. Tomé, Gulf of Guinea. *Annals of Human Genetics* 61:507–18.

Maurer, P. 1992. L'apport lexical bantou en Angolar. *Afrikanische Arbeitspapiere* 29:163–74.

———. 1995. *L'Angolar: Un créole afro-portugais parlé à São Tomé*. Hamburg: Helmut Buske Verlag.

Oota, H., W. Settheetham-Ishida, D. Tiwawech, T. Ishida, and M. Stoneking. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence. *Nature Genetics* 29:20–21.

Pereira, L., L. Gusmão, C. Alves, A. Amorim, and M. J. Prata. 2002. Bantu and European lineages in sub-Saharan Africa. *Annals of Human Genetics* 66:369–78.

Pinto, J., M. J. Donnelly, C. A. Sousa, J. Malta-Vacas, V. Gil, C. Ferreira, V. Petrarca, V. E. do Rosário, and J. D. Charlwood. 2003. An island within an island: Genetic differentiation of *Anopheles gambiae* in São Tomé, West Africa, and its relevance to malaria vector control. *Heredity* 91:407–14.

Pritchard, J. K., M. Stephens, and P. Donnelly. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–59.

Rosa, A., C. Ornelas, A. Brehm, and R. Villems. 2006. Population data on 11 Y-chromosome STRs from Guiné-Bissau. *Forensic Science International* 157:210–17.

Rosenberg, N. A., S. Mahajan, S. Ramachandran, C. Zhao, J. K. Pritchard, and M. Feldman. 2005. Clines, clusters, and the effect of study design on the inference of human population structure. *PLoS Genetics* 1:70.

Seibert, G. 1998. *A questão das origens dos angolares de São Tomé*. Lisbon: CEsA Brief Papers 5.

Seielstad, M. T., E. Minch, and L. L. Cavalli-Sforza. 1998. Genetic evidence for a higher female migration rate in humans. *Nature Genetics* 20:278–80.

Smouse, P. E., V. J. Vitzthum, and J. V. Neel. 1981. The impact of random and lineal fission on the genetic divergence of small human groups: A case study among the Yanomama. *Genetics* 98:179–97.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105:437–60.

Tenreiro, F. 1961. *A ilha de São Tomé*. Lisbon: Memórias da Junta de Investigaçãoes Científicas do Ultramar.

Thomas, M. G., T. Parfitt, D. A. Weiss, K. Skorecki, J. F. Wilson, M. le Roux, N. Bradman, and D. B. Goldstein. 2000. Y chromosome travelling south: The Cohen modal haplotype and the origins of the Lemba, the "black Jews of southern Africa." *American Journal of Human Genetics* 66:674–86.

Tomás, G., L. Seco, S. Seixas, P. Faustino, J. Lavinha, and J. Rocha. 2002. The peopling of São Tomé (Gulf of Guinea): Origins of slave settlers and admixture with the Portuguese. *Human Biology* 74:397–411.

Trovoada, M. J., C. Alves, L. Gusmão, A. Abade, A. Amorim, and M. J. Prata. 2001. Evidence for population sub-structuring in São Tomé e Príncipe as inferred from Y-chromo-

some STR analysis. *Annals of Human Genetics* 65:271–83.

Trovoada, M. J., L. Pereira, L. Gusmão, A. Abade, A. Amorim, and M. J. Prata. 2003. Pattern of mtDNA variation in three populations from São Tomé e Príncipe. *Annals of Human Genetics* 68:40–54.

Watterson, G. 1975. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology* 7:256–76.

 DOI: 10.1086/524762

Supplement A from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

(Current Anthropology, vol. 49, no. 1, p. 134)

Methodology

General Approach

As with most studies of population genetics, a major problem in any analysis of the current patterns of genetic variation in São Tomé is that the choice of a sampling strategy may strongly influence the results. Treating a sampling area a priori as a uniform genetic landscape may of course risk obscuring any underlying genetic structuring. At the same time, sampling strategies based on the identification of individuals according to their supposed membership in social or cultural groups may be used, in principle, to assess the level of genetic differentiation among culturally constructed taxonomies. However, it remains to be shown whether preconceived ethnic labels capture the most important features of the genetic composition of human groups and therefore may be used as adequate proxies for delimiting biologically meaningful categories (Juengst 1998; MacEachern 2000).

An alternative approach to grouping genetic data on the basis of predefined cultural labels is to consider individuals as the starting units of analysis and investigate with what degree of accuracy each individual can be assigned to a cultural group on the basis of multilocus genotype data (McComb et al. 1996; Crawford et al. 2002). This type of approach has recently been extended to allow the identification of groups a posteriori, without the need to specify any prior hypothesis about the genetic origin of individuals or provide information on sampling locations (Pritchard, Stephens, and Donnelly 2000; Falush, Stephens, and Prichard 2003).

To study the patterns of population structure in São Tomé without relying on preconceived population categories, we followed a two-tiered approach similar to that described by Wilson et al. (2001). We inferred the major clustering patterns of the island by using individual multilocus genotypes and compared the distribution of additional markers across the inferred clusters to analyze the major factors that shaped the observed genetic structure. Individual clustering was based on 15 randomly chosen unlinked autosomal microsatellite loci, while additional markers (Duffy blood group, mtDNA, Y-chromosome, and β-globin haplotypes) were specifically selected to address questions such as the genetic contributions of different African regions, the amount of European admixture, the asymmetry of male- and female-mediated gene flow, and the relative levels of differential genetic drift. The method relies on the expectation that genetic information from unlinked loci is more closely associated within the postulated clusters than would be expected in a general unstructured population. This expectation is justified by the fact that it is the association among different loci that provided the cumulative information allowing individuals to be assigned to different clusters (Edwards 2003).

Sampling

We sampled buccal swabs from 394 apparently unrelated individuals from 14 localities, encompassing 11 villages that contain more than 80% of the 131,600 inhabitants of São Tomé, as well as 3 plantations (*roças*) on which many descendents of contract laborers still reside (see fig. 1; table A1). The distribution of sampling locations reflects São Tomé's highly uneven settlement pattern. Because of the mountainous topography of the island, most settlements are concentrated in the northeast and along the coast. The heavily forested southwest is virtually uninhabited, and settlements are scarce in the mountainous center and the southeast (fig. 1). To avoid the inclusion of closely related individuals in the sample, volunteers were requested to gather in each location at a particular sampling point (school or health center) and asked to acknowledge any known family relationship with one another. Cases of acknowledged relationship included parent/offspring, full siblings, half siblings, avuncular relatives, and first cousins, and none of these pairs were used in this study.

Suppl. A from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

Table A1
Population Sizes of Sampled Locations in São Tomé

Sampled Location	Number (%) of Residents[a]	Sample Size
São João dos Angolares	4,082 (3)	57
Ribeira Afonso	4,955 (4)	35
Santana	8,303 (6)	38
Pantufo	1,929 (2)	8
Cidade de São Tomé	49,957 (38)	39
Praia Gamboa[b]	n.a.[c]	42
Guadalupe	5,329 (4)	23
Neves	8,497 (6)	18
Santa Catarina	2,199 (2)	22
Madalena	2,165 (2)	16
Trindade	19,234 (15)	41
Agostinho Neto[d]	n.a.	19
Monte Café[d]	n.a.	21
Porto Alegre[d]	n.a.	15

Sources: National Institute of Statistics (São Tomé e Príncipe), unpublished results from the 2001 census.
[a] Percentage of the total population of the island (131,633) in parentheses.
[b] A suburb of the capital, Cidade de São Tomé.
[c] Not available.
[d] Plantation.

Laboratory Analyses

Autosomal Microsatellite Markers

Individuals were genotyped for 15 unlinked autosomal microsatellites (D3S1358, TH01, D21S11, D18S51, Penta E, D5S818, D13S317, D7S820, D16S539, CSF1PO, Penta D, vWA, D8S1179, TPOX, and FGA) using the Powerplex 16 System (Promega) according to the manufacturer's instructions.

β-globin (HBB) and Duffy Blood Group (FY) Loci

Hemoglobin β^*S (HBB*S) and β^*C (HBB*C) mutations were detected as described by Modiano et al. (2001). β-globin S haplotypes were determined according to Tomás et al. (2002). The Duffy blood group O allele (FY*O) was identified as in Tournamille et al. (1995).

Mitochondrial DNA Variation

The mtDNA hypervariable region-1 (HVR-1) between nucleotide positions 16024 to 16399 was sequenced according to Vigilant et al. (1991), with minor modifications. For population comparisons only nucleotide positions 16090–16365 were considered, since this is the stretch shared by most sequences available from the literature. To resolve occasional ambiguities in haplogroup assignments, a selected set of six diagnostic restriction fragment length polymorphism (RFLP) sites was additionally analyzed: 2349 *Mbo*I (haplogroup L3e), 3592 *Hpa*I (paragroup L1 and haplogroup L2), 10084 *Taq*I (L3b), 10397 *Alu*I (M), 12308 *Hinf*I (U), and 13957 *Hae*III (L2c). Haplogroup assignment followed the classification from Salas et al. (2002, 2004).

Y-chromosome Variation

Y-chromosome variation was characterized by the joint analysis of fast-evolving microsatellite loci and slow-evolving biallelic markers. We followed a hierarchical approach based on the updated phylogeny provided by Jobling and Tyler-Smith (2003) to type a set of 26 biallelic Y-chromosome markers (M2, M9, M22, M26, M34, M35, M42, M60, M62, M70, M78, M81, M96, M123, M170, M172, M173, M191, M201, M213, YAP, Tat, 12f2 deletion, 92R7, P25, and SRY1831.1/2) using RFLP analysis, direct sequencing, or the SNaPshot minisequencing procedure (Applied Biosystems) according to previously described methods (Seielstad et al.

1994; Hammer and Horai 1995; Shen et al. 2000; Brion et al. 2005). Samples were additionally typed for 11 Y-chromosome microsatellite loci (DYS19, DYS389I, DYS389II, DYS385, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, and DYS439) using the Powerplex Y System (Promega). The DYS385 locus consists of a duplicated tetranucleotide short tandem repeat region and was not used for phylogenetic reconstructions.

Clustering Analysis

Individual Clustering

Genetic clusters of individuals were inferred without prior knowledge of geographical or ethno-linguistic affiliations by using the Bayesian approach implemented in the STRUCTURE program, version 2 (Pritchard, Stephens, and Donnelly 2000; Falush, Stephens, and Pritchard 2003), assuming that individuals may have mixed ancestry and therefore that fractions of ancestry in each cluster could be estimated for each individual (admixture model). The program uses the genotypes of sampled individuals (X) to identify clusters with distinctive allele frequencies and assigns individuals to a number of clusters (K) that is initially defined by the user and can be varied across independent runs. We performed ten independent runs for each K between 1 (no structure) and 14 (the total number of sampled locations) with 10^6 iterations after a burn-in of length 30,000. To determine the number of clusters most appropriate for interpreting the data, we calculated the posterior probability Pr(X/K) according to Pritchard, Stephens, and Donnelly (2000) and compared the various ln P(X/K) values produced for each K using the Wilcoxon two-sample test, as proposed by Rosenberg et al. (2001). All STRUCTURE runs were performed assuming a model of allele frequency correlations (*F* model) which supposes that alleles in different clusters have correlated frequencies due to shared ancestry. The choice of the correlated frequencies model in São Tomé, which is essentially an African-derived population, is justified by the observation that large allele frequency correlations are usually found across most human populations, even when samples from different continents are compared (Rosenberg et al. 2005). Moreover, we made a preliminary comparison of the microsatellite allele frequency distributions across the 14 sampling locations from São Tomé and found a mean Pearson correlation coefficient of 0.81 (0.64–0.92) based on 175 alleles at 15 loci, which indicates a high level of correlation in our data. Although the *F* model enhances the power of STRUCTURE to detect subtle population subdivision, a minimum of 20–60 loci is typically required to distinguish between clusters corresponding to major world continental regions (Pritchard, Stephens, and Donnelly 2000; Bamshad et al. 2003; Falush, Stephens, and Pritchard 2003; Ramachandran et al. 2004). Since more markers are expected to be required in less divergent human groups or in admixed populations like São Tomé, our approach using only 15 microsatellite loci is conservative.

Population Clustering

We further analyzed the genetic structure of São Tomé by exploring the relationships between the 14 sampled locations with both principal component and pairwise genetic distance analyses. Pairwise F_{st} genetic distances (Reynolds, Weir, and Cockerham 1983) were used to compute neighbor joining networks with the PHYLIP 3.5c software package (Felsenstein 1993). The reliability of network nodes was assessed with the bootstrap option implemented in the PHYLIP SEQBOOT program using 10,000 replicate data sets. Principal-component factor scores were calculated with POPSTR (Henry Harpending, personal communication) and displayed as interpolated synthetic maps (Cavalli-Sforza, Menozzi, and Piazza 1994) drawn with SURFER 6.0 (Golden Software 1996).

Analysis of Genetic Variation across Inferred Clusters

β-globin Locus

Because of their high levels of geographic segregation in major areas of slave recruitment, haplotypes associated with β-globin S allele (HBB*S) are particularly useful for assessing the origins of lineages with diverse regional African ancestries: the Benin haplotype predominates in Central-West Africa, from present-day Ghana to northern Gabon, the Bantu haplotype is particularly common in the Congo-Angola area, and the Senegal haplotype is restricted to Atlantic West Africa, where the HBB*S mutation is less frequent than in the other regions (Nagel and Ranney 1990). The β-globin C allele (HBB*C) also has a remarkable geographic specificity, being virtually confined to Central-West Africa, with a peak in the region of Burkina Faso (Cavalli Sforza, Menozzi, and Piazza 1994).

Suppl. A from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

To study the distributions of the HBB*S haplotypes and the HBB*C allele across major groups defined by STRUCTURE, we selected a total of 45 HBB*S and 25 HBB*C carriers that were previously identified in a global sample of 613 unrelated individuals not included in the initial STRUCTURE runs. HBB*S and HBB*C carriers were subsequently typed for the 15 microsatellites, and their proportion of ancestry was inferred in additional STRUCTURE runs together with the initial set of 394 individuals. The carriers were finally assigned to the major cluster in which they had the highest proportion of ancestry. By using this procedure, we tried to reduce the chances that the ascertainment of an enriched collection of the relatively infrequent HBB*S and HBB*C alleles could bias the initial clustering analysis because of a possible association between HBB alleles and specific clusters.

Duffy Blood Group

Because the Duffy blood group FY*O allele is almost fixed in the majority of sub-Saharan African populations and virtually absent in Europe, it is one of the most informative ancestral markers for evaluating the amounts of admixture between Africans and Europeans. To assess the relative amount of European admixture across genetic clusters, we simply compared the FY*O allele frequencies in a subsample of 156 individuals previously assigned to different clusters on the basis of their microsatellite multilocus genotypes.

MtDNA and Y-chromosome Haplotypes

Because of their uniparental patterns of inheritance and lower effective population size, mtDNA and Y-chromosome haplotypes provide complementary information about female- and male-specific aspects of genetic variation and are especially sensitive to the effects of drift. We characterized mtDNA and Y-chromosome variation in subsets of individuals with posterior probabilities of cluster assignment higher than 70% in order to assess the contributions of different African regions and study male- and female-specific levels of gene flow and genetic drift. Summary statistics for mtDNA and Y-chromosome variation within and among subpopulation clusters were estimated with the ARLEQUIN software (version 2.1; Schneider, Roessli, and Excoffier 2000). Intercluster differentiation was assessed either taking or not taking into account the molecular differences between haplotypes by calculating F_{st} and Φ_{st} or R_{st}, respectively. For Φ_{st} estimation, divergence between mtDNA sequences was measured by the number of nucleotide differences. R_{st} was calculated using distances between Y-chromosome microsatellite data estimated by the sum of the squared number of microsatellite repeat differences between haplotypes. The significance of population subdivisions was evaluated by permutation. Networks for Y-chromosome haplotypes and mtDNA sequences were constructed using NETWORK 4.0 (http://www.fluxus-engineering.com), applying the median-joining and reduced-median algorithms, respectively (Bandelt et al. 1995; Bandelt, Forster, and Röhl 1999). MtDNA sequence networks were weighted according to the relative mutation rates of the different nucleotide positions estimated by Meyer, Weiss, and van Haeseler (1999). For Y-chromosome haplotype networks, each microsatellite was weighted according to its variance in repeat number. Biallelic markers were given a weight ten times higher than the highest microsatellite weight.

Additional References

Alves, C., L. Gusmão, A. Damascene, B. Scares, and A. Amorim. 2004. Contribution for an African autosomic STR data base (AmpF/STR Identifiler and Powerplex 16 System) and a report on genotypic variations. *Forensic Science International* 139:201–5.

Bandelt, H-J., P. Forster, and A. Röhl. 1999. Median-joining networks for inferring intraspecific phytogenies. *Molecular Biology and Evolution* 16:37–48.

Bandelt, H-J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141:743–53.

Beleza, S., C. Alves, F. Reis, A. Amorim, A. Carracedo, and L. Gusmão. 2004. 17 STR data (AmpF/STR Identifiler and Powerplex 16 System) from Cabinda (Angola). *Forensic Science International* 141:193–96.

Brion, M., B. Sobrino, A. Blanco-Verea, M. V. Lareu, and A. Carracedo. 2005. Hierarchical analysis of 30 Y-chromosome SNPs in European populations. *International Journal of Legal Medicine* 119:10–15.

Côrte-Real, F., L. Andrade, M. Carvalho, M. J. Anjos, J. Gamero, D. N. Vieira, A. Carracedo, and M. C. Vide. 2000. Comparative analysis of STR data for Portuguese spoken countries. *Progress in Forensic Genetics* 8: 212–14.

Edwards, A. W. F. 2003. Human genetic diversity: Lewontin's fallacy. *Bioessays* 29:798–801.
Felsenstein, J. 1993. *PHYLIP-phylogeny inference package, version 3.5c*. Department of Genetics, University of Washington, Seattle.
Golden Software. 1996. Surfer for Windows (Win32): Surface Mapping System, version 6.04. Golden: Golden Software, Inc.
Hammer, M. F., and S. Horai. 1995. Y-chromosomal DNA variation and the peopling of Japan. *American Journal of Human Genetics* 56:951–62.
Jobling, M. A., and C. Tyler-Smith. 2003. The human Y-chromosome: An evolutionary marker comes of age. *Nature Reviews Genetics* 4:598–612.
Juengst, E. T. 1998. Group identity and human diversity: Keeping biology straight from culture. *American Journal of Human Genetics* 63:673–77.
Lane, A. B., H. Soodyall, S. Arndt, M. E. Ratshikhopa, E. Jonker, C. Freeman, L. Young, B. Morar, and L. Toffie. 2002. Genetic structure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies. *American Journal of Physical Anthropology* 119:175–85
MacEachern, S. 2000. Genes, tribes, and African history. *Current Anthropology* 41:357–84.
McComb, J., M. H. Crawford, L. Osipova, T. Karaphet, O. Posukh, and M. Schanfield. 1996. DNA interpopulational variation in Siberian indigenous populations: The Mountain Altai. *American Journal of Human Biology* 8:599–608.
Meyer, S., G. Weiss, and A. von Haeseler. 1999. Pattern of nucleotide substitution and rate heterogeneity in hypervariable regions I and II of human mtDNA. *Genetics* 152:1103–10.
Modiano, D., G. Luoni, B. S. Sirima, J. Simporé, F. Verra, A. Konaté, E. Rastrelli, A. Olivieri, C. Calissano, G. M. Paganotti, L. D'Urbano, I. Sanou, A. Sawadogo, G. Modiano, and M. Coluzzi. 2001. Haemoglobin C protects against *Plasmodium falcinarum malaria. Nature* 414:305–8.
Nagel, R. L., and H. M. Ranney. 1990. Genetic epidemiology of structural mutations of the β-globin gene. *Seminars in Hematology* 27:342–59.
Ramachandran, S., N. A. Rosenberg, L. Zhivotovsky, and M. Feldman. 2004. Robustness of the inference of human population structure: A comparison of X-chromosomal and autosomal microsatellites. *Human Genomics* 1:87–97.
Reynolds, J., B. S. Weir, and C. C. Cockerham. 1983. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics* 105:767–79.
Rosenberg, N. A., T. Burke, K. Elo, M. W. Feldman, P. J. Freidlin, M. A. M. Groenen, J. Hillel, A. Mäki-Tanila, M. Tixier-Boichard, A. Vignal, K. Wimmers, and S. Weigend. 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* 159:699–713.
Salas, A., A. Carracedo, M. Richards, and V. Macaulay. 2005. Charting the ancestry of African Americans. *American Journal of Human Genetics* 74:454–65.
Salas, A., M. Richards, T. De La Fe, M. V. Lareu, B. Sobrino, P. Sanchez-Diz, V. Macaulay, and A. Carracedo. 2002. The making of the African mtDNA landscape. *American Journal of Human Genetics* 71:1082–1111.
Salas, A., M. Richards, M. V. Lareu, R. Scozzari, A. Coppa, A. Torroni, V. Macaulay, and A. Carracedo. 2004. The African diaspora: Mitochondrial DNA and the Atlantic slave trade. *American Journal of Human Genetics* 74:454–65.
Schneider, S., D. Roessli, and L. Excoffier. 2000. *Arlequin, version 2.000: A software for population genetics data analysis*. Geneva: University of Geneva.
Seielstad, M. T., J. M. Hebert, A. A. Lin, P. A. Underbill, M. Ibrahim, D. Vollrath, and L. L. Cavalli-Sforza. 1994. Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. *Human Molecular Genetics* 3:2159–61.
Shen, P., F. Wang, P. A. Underhill, C. Franco, W-H. Yang, A. Roxas, R. Sung, A. A. Lin, R. W. Hyman, D. Vollrath, R. W. Davis, L. L. Cavalli-Sforza, and P. Oefner. 2000. Population genetics implications from sequence variation at four Y chromosome genes. *Proceedings of the National Academy of Sciences, USA* 97: 7354–59.
Sun, G., S. T. McGarvey, R. Bayoumi, C. J. Mulligan, R. Barrantes, S. Raskin, Y. Zhong, J. Akey, R. Chakraborty, and R. Deka. 2003. Global genetic variation at nine short tandem repeat loci and implications on forensic genetics. *European Journal of Human Genetics* 11:39–49.
Tofanelli, S., I. Boschi, S. Bertoneri, V. Coia, L. Taglioli, M. G. Franceschi, G. Destro-Bisol, V. Pascali, and G.

Paoli. 2003. Variation at 16 STR loci in Rwandans (Hutu) and implications on profile frequency estimation in Bantu-speakers. *International Journal of Legal Medicine* 117:121–26.

Tournamille, C., Y. Colin, J. P. Cartron, and C. Le Van Kim. 1995. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy-negative individuals. *Nature Genetics* 10:224–28.

Vigilant, L., M. Stoneking, H. Harpending, K. Hawkes, and A. C. Wilson. 1991. African populations and the evolution of human mitochondrial DNA. *Science* 253:1503–7.

Wilson, J. F., M. E. Weale, A. C. Smith, F. Gratrix, B. Fletcher, M. G. Thomas, N. Bradman, and D. B. Goldstein. 2001. Population genetic structure of variable drug response. *Nature Genetics* 29:265–69.

 DOI: 10.1086/524762

Supplement B from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

(Current Anthropology, vol. 49, no. 1, p. 134)

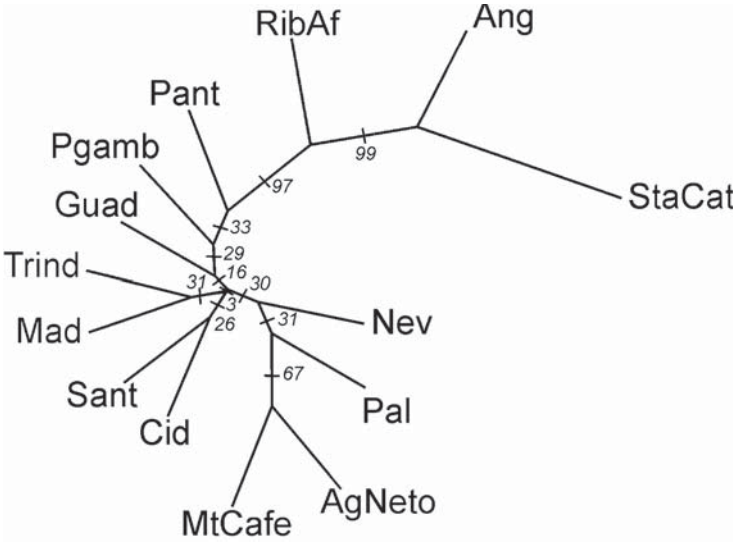


Figure B1. Neighbor-joining network estimated from F_{st} genetic distances between 14 sampling locations from São Tomé using 15 autosomal microsatellite loci. Location abbreviations are as in figure 2. Bootstrap percentage values shown on branches are based on 10,000 replications.

Suppl. B from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

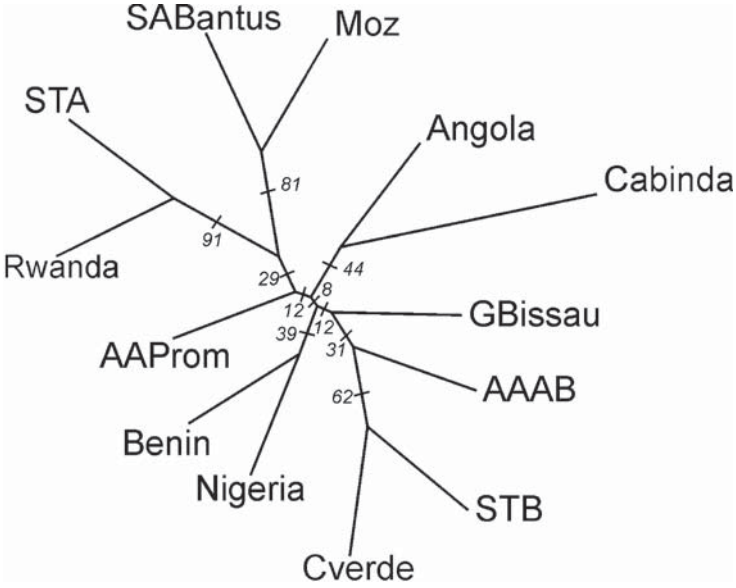


Figure B2. Network based on F_{st} genetic distances estimated with a subset of 9 shared microsatellite loci (D3S1358, D5S818, D7S820, D8S1179, D13S317, D18S51, D21S11, vWA, and FGA) between the two genetic clusters from São Tomé (STA = cluster A; STB = cluster B) and 11 additional African-derived samples: Nigeria (N = 46) and Benin (N = 51) (Sun et al. 2003); Rwanda (N = 52; Tofanelli et al. 2003); Mozambique (= Moz; N = 142; Alves el al. 2004); Angola (N = 102; Côrte-Real et al. 2000); Cabinda (N = 110; Beleza et al. 2004); South African Bantus (= SABantus; N = 166; Lane et al. 2002); Guinea-Bissau (= GBissau; N = 100; Côrte-Real et al. 2000); Cape Verde (= Cverde; N = 107; Côrte-Real et al. 2000); the Afro-American panel from Promega (= AAProm; N = 218) and the Afro-American panel from Applied Biosystems (= AAAB; N = 357). Bootstrap percentage values shown on branches are based on 10,000 replications.

 DOI: 10.1086/524762

Supplement C from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

(Current Anthropology, vol. 49, no. 1, p. 134)

Table C1
mtDNA HVR-I Sequence Diversity in Two Genetic Clusters from São Tomé

			Cluster	
Haplotype ID	HVR-I Haplotypes	Haplogroup	A	B
H1	129 148 168 172 187 188G 189 223 230 278 293 311 320	L0a1a	–	1
H2	126 187 189 223 264 270 278 293 311	L1b1	6	2
H3	126 187 189 223 264 270 278 293 300 311	L1b1	1	–
H4	104 126 187 189 223 256 264 270 278 293 311	L1b1	–	1
H5	86 129 187 189 223 241 278 293 294 311 360	L1c1	–	1
H6	129 163 187 189 209 223 278 293 294 311 360	L1c1	–	2
H7	93 129 187 189 223 263 278 293 294 311 360 368	L1c1	–	1
H8	129 187 189 274 278 293 294 311 360	L1c1a	5	1
H9	51 129 187 189 214 234 249 258 274 278 293 294 311	L1c1a1	4	–
H10	129 187 189 223 265C 286A 292 294 311 360	L1c2	–	2
H11	129 163 187 189 265C 278 286G 294 311 320 360	L1c2	–	1
H12	189 192 223 278 294 390	L2a*	1	–
H13	189 192 223 278 294 309 390	L2a1	1	–
H14	111 223 278 294 309 390	L2a1	–	1
H15	93 223 278 286 294 309 390	L2a1a	–	1
H16	114A 129 192 213 223 278 355 362 390	L2b1	1	–
H17	192 223 278 390	L2c*	–	1
H18	81 175 223 278 320 390	L2c*	1	–
H19	81 93 175 223 278 320 390	L2c*	1	–
H20	223 264 278 390	L2c2	3	–
H21	84 93 220 223 264 278 311 390	L2c2	–	2
H22	93 223 264 278 390	L2c2	1	–
H23	223 278 320 390	L2c*	–	1
H24	51 223 355	L3*	–	1
H25	124 223 278 362	L3b*	6	3
H26	124 223 278	L3b*	–	1
H27	124 189 214 223 278 362	L3b*	–	1
H28	124 148 223 278	L3b*	1	–
H29	172 223 399	L3e1*	1	–
H30	172 223 327 399	L3e1*	8	1
H31	176 223 256 327	L3e1*	1	–
H32	223 320	L3e2*	–	2
H33	223 320 399	L3e2*	1	1
H34	172 189 223 320	L3e2b	–	1
H35	172 189 256 311 320	L3e2b	–	1
H36	51 223 264	L3e4	–	2
H37	209 223 264 311	L3e4	–	1
H38	51 223 264 299	L3e4	–	1
H39	209 223 311	L3f*	–	2
H40	129 209 223 256 292 295 311	L3f1	–	1
H41	129 209 223 288 292 295 311	L3f1	–	1
H42	172 219 278	U6	–	1
			42	40

Note: Variant positions from the Cambridge Reference Sequence are shown for a 16024–16399 stretch (minus 16000).

Suppl. C from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

Table C2
Geographic Distribution of Matches to HVR–I Sequences from São Tomé

No. Observed					Perfect Matches[a]							Imperfect Matches[b]						
Haplotype ID	Haplotypes	Cluster A	Cluster B	São Tomé[c]	North Africa	West Africa	West Central Africa	East Africa	South East Africa	South Africa	America	North Africa	West Africa	West Central Africa	East Africa	South East Africa	South Africa	America
H1	L0a1a	–	1	T, ST	1	–	5	5	10	–	1	–	–	–	–	–	–	–
H2	L1b1	6	2	F, ST	3	13	10	2	1	–	3	–	–	–	–	–	–	–
H3	L1b1	1	–	–	–	–	–	–	–	–	–	3	13	10	2	1	–	3
H4	L1b1	–	1	–	–	1	–	–	–	–	–	–	1	–	–	–	–	1
H5	L1c1	–	1	–	–	–	–	–	–	–	–	–	–	1	–	1	–	1
H6	L1c1	–	2	–	–	–	3	3	1	–	1	–	–	–	–	–	–	–
H7	L1c1	–	1	T, F	–	–	1	–	–	–	–	–	–	–	–	–	–	–
H8	L1c1a	5	1	–	–	1	–	–	–	–	–	–	–	3	–	2	–	1
H9	L1c1a1	4	–	A, ST	–	–	2	–	–	–	–	–	–	–	–	–	–	–
H10	L1c2	–	2	A, ST	–	–	–	–	–	–	–	–	–	–	–	–	–	–
H11	L1c2	–	1	A	–	–	–	–	–	–	–	–	–	–	–	–	–	1
H12	L2a*	1	–	ST	–	1	2	1	3	–	–	–	–	–	–	–	–	–
H13	L2a1	–	1	T, F, ST	2	9	7	4	2	1	3	–	–	–	–	–	–	–
H14	L2a1	–	1	ST	–	1	–	–	–	–	–	–	1	–	–	–	–	–
H15	L2a1a	–	1	–	–	–	–	–	–	–	–	–	3	–	–	9	–	1
H16	L2b1	1	–	–	–	4	–	–	–	–	–	–	–	–	–	–	–	2
H17	L2c*	–	1	–	1	2	–	–	–	–	–	–	–	–	–	–	–	–
H18	L2c*	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
H19	L2c*	1	–	A, F, ST	–	–	–	–	–	–	–	–	–	–	–	–	–	–
H20	L2c2	3	–	A, ST	–	4	–	–	–	–	–	–	–	–	–	–	–	–
H21	L2c2	–	2	T	–	1	–	–	–	–	–	–	1	–	–	–	–	–
H22	L2c2	1	–	A, ST	–	3	–	–	–	–	–	–	–	–	–	–	–	–
H23	L2c*	–	1	–	–	2	1	–	–	–	–	–	–	–	–	–	–	–
H24	L3*	–	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
H25	L3b*	6	3	F	1	12	7	1	1	–	2	–	–	–	–	–	–	–
H26	L3b*	–	1	–	–	3	2	–	1	–	–	–	–	–	–	–	–	–
H27	L3b*	–	1	–	–	3	–	–	–	–	–	–	–	–	–	–	–	–
H28	L3b*	1	–	–	–	–	–	–	–	–	–	–	3	2	–	1	–	–
H29	L3el*	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2
H30	L3el*	8	1	A, ST	–	–	–	–	–	–	2	–	–	–	–	7	–	–
H31	L3el*	1	–	–	–	–	1	–	–	–	–	–	–	–	–	2	–	2
H32	L3e2*	–	2	–	–	13	4	–	–	–	2	–	–	–	–	–	–	–
H33	L3e2*	1	1	–	–	13	4	–	–	–	2	–	–	–	–	–	–	–
H34	L3e2b	–	1	–	–	5	3	–	1	2	3	–	–	–	–	–	–	–
H35	L3e2b	–	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
H36	L3e4	–	2	T	–	5	1	1	1	–	1	–	–	–	–	–	–	–
H37	L3e4	–	1	ST	–	–	–	–	–	–	–	–	–	–	–	–	–	–
H38	L3e4	–	1	T, F	–	1	–	–	–	–	–	–	1	–	–	–	–	–
H39	L3f*	–	2	–	1	1	7	1	2	–	2	–	–	–	–	–	–	–
H40	L3f1	–	1	–	–	–	–	–	–	–	–	–	–	–	1	1	–	2
H41	L3f1	–	1	ST	–	1	–	–	–	–	–	–	–	–	–	–	–	–
H42	U6	–	1	–	3	1	1	–	–	–	–	–	–	–	–	–	–	–

Note: Database sequences were assembled from previous compilations described in Salas et al. (2002, 2004, 2005) and Batini et al. (2007).

[a] Number of population groups with at least one match in each geographic area. North Africa: Tunisia, Morocco and Mauritania. West Africa: Cape-Verde, Guinea-Bissau, Senegal, Sierra Leone, Mali, Niger, Nigeria. West-Central Africa: Cameroon, Equatorial Guinea, Congo, Cabinda, Angola. East Africa: Sudan, Kenya, and Tanzania. South-East Africa: Mozambique. South Africa: South Africa Khoisan; Afro-American: North, Central, and South America.

[b] Number of different populations with at least one sequence differing at a single position in each geographic area. Imperfect matches were searched when perfect matches either were not found or were restricted to just one population outside São Tomé.

[c] Matches to previous samples from São Tomé, including nonidentified individuals (ST) (Mateu et al 1997) and self-reported Tonga (T), Forro (F), and Angolares (A) (Trovoada et al. 2003).

Suppl. C from Coelho et al., "Human Microevolution and the Atlantic Slave Trade"

Table C3
Y-chromosome Haplotypes in Two Genetic Clusters from São Tomé

		Microsatellites											Cluster	
Haplogroup	Haplotype	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	DYS437	DYS438	DYS439	DYS385	A	B
E3a(XE3a7)	H1	15	12	31	21	10	11	13	14	11	12	17, 17	1	–
	H2	15	13	31	21	10	11	13	14	11	13	17, 17	1	2
	H3	15	13	31	21	10	11	13	14	11	11	16, 17	1	–
	H4	15	13	32	21	10	11	13	14	11	11	16, 17	–	1
	H5	15	14	31	22	10	11	13	14	11	12	15, 18	–	1
	H6	15	13	30	21	10	11	12	14	11	12	14, 17	–	1
	H7	14	13	30	21	10	10	13	14	11	13	16, 18	–	1
	H8	15	13	31	21	11	11	12	14	11	11	16, 17	2	–
	H9	15	13	31	21	11	11	13	14	11	11	16, 17	15	–
	H10	15	13	31	21	11	11	13	14	11	12	16, 17	1	–
	H11	15	13	29	21	11	11	13	14	11	11	16, 17	1	–
	H12	15	13	32	21	11	11	13	14	11	13	16, 16	–	1
	H13	16	13	31	21	11	11	13	14	11	11	16, 17	–	1
	H14	16	14	31	22	11	11	13	14	11	12	16, 17	–	1
E3a7	H15	16	13	30	21	10	11	14	14	11	12	17, 18	–	1
	H16	16	13	30	21	10	11	15	14	12	12	18, 18	–	1
	H17	17	13	30	21	10	11	14	14	11	12	17, 17	–	1
	H18	16	13	31	21	10	11	14	14	11	12	17, 17	–	1
	H19	17	14	31	21	10	11	14	14	11	12	16, 17	1	–
E3b1	H20	13	13	30	24	10	11	13	14	10	12	15, 18	–	1
E3b3*	H21	15	13	31	24	10	11	12	13	10	11	13, 18	–	1
B	H22	15	13	29	24	11	14	12	14	10	12	14, 15	–	1
G	H23	15	12	28	22	10	11	13	16	10	11	13, 15	–	1
J*	H24	14	13	31	22	10	11	12	14	10	12	14, 18	–	1
K2	H25	14	12	30	24	11	15	13	14	9	10	14, 18	–	1
R1*	H26	14	13	29	24	10	13	13	15	12	12	11, 14	1	1
	H27	14	14	30	24	11	13	13	15	12	11	11, 14	1	–
	H28	14	13	30	24	11	13	13	15	12	11	11, 14	–	1
R1b	H29	14	13	29	24	10	13	13	15	12	12	12, 13	–	1
													25	22

Note: Haplogroup nomenclature according to Jobling and Tyler-Smith (2003).

3.2.1 Comments

3.2.1.1 Implications for study designs in human populations

To avoid the inherent circularity of traditional population genetics study-designs, we attempted to interpret the genetic structure of the island of São Tomé without relying on predefined label categories. By coupling a transect sampling strategy with a Bayesian clustering approach, each individual could be treated as a basic sampling unit, so that genetic groups could be delimited without relying on non-genetic criteria. Other critical point in the genetic study of populations is the choice of the type and number of genetic markers. By using a set of 15 autosomal markers, we avoided the problems associated with single-locus studies, whereby the history of a locus could be erroneously taken as the history of a population.

The geographic patterning of genetic variation in São Tomé was found to be mainly determined by the dichotomy between Angolares and non-Angolares. The occurrence of a kin-structured founder event seems to have significantly enhanced genetic differentiation and increased the chances of cluster detection.

By analyzing the distribution of phylogeographic informative markers only among individuals presenting a fraction of ancestry higher than 70% in one of the two clusters, we removed possible effects of recent admixture and identified the "core lineages" that shaped the general patterns of genetic variation observed. For example, we were able to find a single modal Y-chromosome haplotype representing as much as 60% of the lineages sampled in Angolares, supporting the occurrence of a remarkable founder effect. Past genetic studies in São Tomé (e.g. Trovoada et al 2003, Trovoada et al. 2007), based on self-reported identity, have found weaker signs of genetic microdifferentiation of the Angolares. In fact, if individuals with loose genetic relationships to the "original" Angolares declare themselves as belonging to this group, the genetic signal of the initial founder event will be diluted.

3.2.1.2 Implications for the genesis of the Angolar language

Admitting that the Angolar group derived from a founder population, two major alternative scenarios may be proposed to account for the present levels of differentiation. According to one of the scenarios, the founder group consisted of a small group of fugitives from a specific region of Africa who had little initial contact with the remaining population (Figure III.3A.1,2). This scenario is fully compatible with the long held popular belief that the Angolares are survivors of the wreckage of a slave ship, occurring just off the southeast shore of São Tomé around 1540-50 (Castelo-Branco 1971, Seibert 1998, Caldeira 2004). According

to this view, the survivors escaped to the mountainous interior and launched several raids on plantations with the purpose of capturing women. A scenario of women abduction could have provided the basis for restricted female-mediated gene flow, leading to the observed differences in the levels of between-cluster mtDNA and Y-chromosome divergence (Φst=0.019/Fst=0.032 versus Rst=0.158/ Fst=0.231). According to the alternative scenario, Angolares could have resulted from the split and subsequent isolation of a small group of escaped slaves that separated from the major population living in plantations (FigureIII.3B.1,2). This is compatible with available historic references to runaway slaves (bush negroes), slave uprisings, punishment expeditions (bush wars) and, arguably, refugee villages, called mocambos (Caldeira 2004). The word "Mocambo" is likely from Kimbundo or Kikongo origin and was later widely used to designate maroon societies in Brazil (Caldeira 2004). It is possible that these splits were not just random fissions of unrelated individuals, but kin-structured founder events.

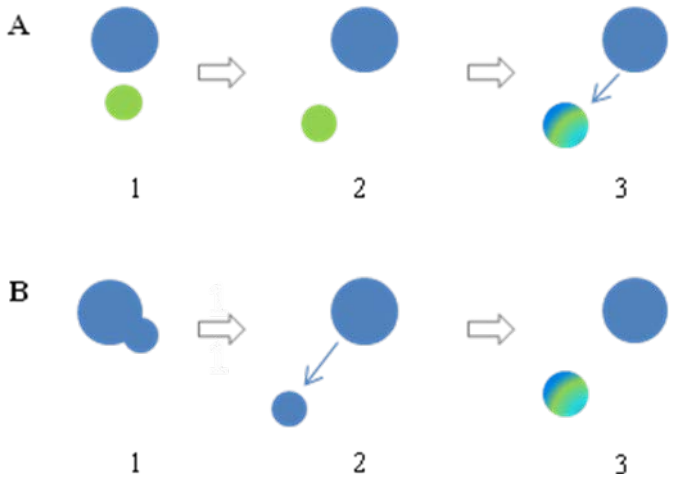


Figure III.3: Possible scenarios for the origin of the Angolar group based on the occurrence of a founder event. (A) It is considered that the Angolares (in green) had little initial contact with the remaining population of the island (in blue) (1 and 2). Later introgression has attenuated initial differences (3). (B) The Angolar group resulted from a split of a small group of individuals from the major population of the island (1 and 2). Subsequent isolation has led to its present differentiation (represented by the green color) (3). 1, 2 and 3 refer to consecutive periods of time.

The two scenarios bear clearly different genetic and linguistic implications. Under the first scenario (FigureIII.3.A), genetic divergence between Angolares and non-Angolares are the result of the retention of the initial differences, while genetic resemblances are caused by gene flow in more recent secondary contacts. Following this view, Bantu words specific to the Angolar Creole would be remnants of the original language, or languages, spoken by founders that adopted a relexified version of the autochthonous Creole in subsequent contacts with non-Angolares. Under the alternative scenario (Figure III.3.B), they are the similarities, and not the differences, that are ancestral. In this case, Angolar founders would have been exposed to the formative phases of an ancestral proto-Creole and linguistic differences, like genetic differences, would have been caused by subsequent isolation. This late scenario is however less likely given the recent history of the island (around five centuries) and also because it is implausible that the unique Bantu features found in Angolares could have derived from an ancestral proto-Creole simply through isolation (Hagemeijer 1999).

3.2.1.3 The relevance of founder events during human evolution

Overall, our data suggest that the pattern of genetic variation observed in Angolares group has been shaped by a founder event likely undertaken by a kin-structured group of fugitive slaves. Subsequent female-mediated introgression seems also to have occurred. Other historical examples of populations that emerged or were greatly influenced by the arrival of a relatively low number of individuals to a remote place, include the formation of the Garifunas in the Caribbean islands (Salas et al. 2005), or the population of Norfolk Island, located off the eastern coast of Australia (Macgregor et al. 2009). The Garifuna are believed to be the descendents of shipwrecked African slaves that have reached the island of St. Vicent and have admixed with Native Americans. The population of Norfolk islands have originated from a small number of “Bounty” mutineer founders, with British ancestry, and Tahitian women. These results, obtained through the analysis of populations evolving at a small geographic scale over a short period of time, may provide important insights into general aspects of human evolution and differentiation. In particular, a better understanding of the genetic impact of founder events is especially important as recent studies have shown that founder events strongly shaped the present configuration of worldwide patterns of genetic diversity (e.g. Ramachandran et al. 2005).

3.2.1.4 Analysing the associations between languages and genetics

Our results also show that the genetic differentiation observed in S. Tomé is clearly more related to language than to geographic distance. Several studies have identified associations between language and genes in human populations at both large (e.g. Belle and Barbujani 2007, Cavalli-Sforza 2000, Scheinfeldt et al. 2010) and fine scales (e.g. Friedlaender et al. 2007, Lansing et al. 2007). The conceptual link between genetics and linguistics resides in the fact that both, genes and languages, can be vertically transmitted between generations and are expected to be similarly affected by the same kind of evolutionary factors. In the *The Descent of Man*, Darwin already noted that: "The formation of different languages and of distinct species, and the proofs that both have been developed through a gradual process, are curiously parallel" (Darwin 1882). Moreover, when intermarriage tends to occur essentially within linguistic groups, language differences may contribute to genetic isolation and be a causal factor of human microevolution. However, a correspondence between patterns of genetic and linguistic diversity is not always verified. Such lack of correspondence could arise when genetic exchange occurs with little or no linguistic exchange, or in cases where horizontal transmission of language is not accompanied by a parallel genetic change. In their study of African populations, Tishkoff et al. (2009) have found several examples of genetically close populations that do not share the same linguistic affiliation (Figure III.4).

It is interesting to note that even in cases where language and genetics are unpaired, significant insights about the evolutionary dynamics can be retrieved. For example, the Fulbe population, who speak a west African Niger-Kordofanian language are genetically close to the Chadic and Central-Sudanic speaking populations, suggesting the occurrence of gene flow into this population without a clear repercussion in the language (Tishkoff et al. 2009) (Figure III.4). Also, the genetic divergence between Pygmies and the remaining speakers of Niger-Kordofanian speaking-populations languages, gives support to the idea that Pygmies have lost their indigenous languages and adopted the language of their neighbouring populations, with whom they have interact (Tishkoff et al. 2009) (Figure III.4). Thus, the correlation between genetic and linguistic variation is dependent on the populations being studied and should be more a hypothesis to test and interpret than a assumption that could be made *a priori* (Barbujani 1997).

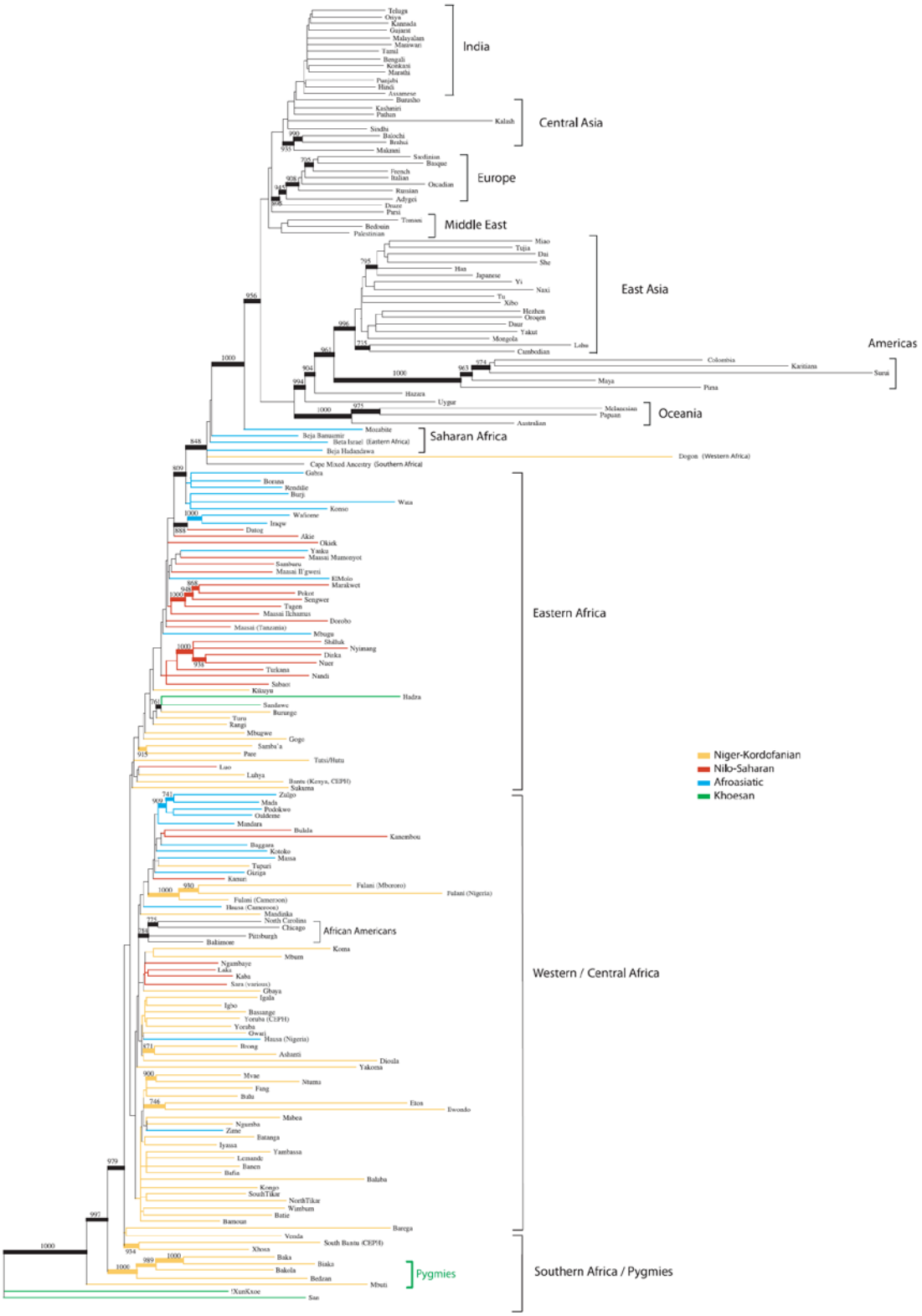


Figure III.4 Neighbor-joining tree from pairwise D^2 genetic distances between populations. African populations are colour-coded according to language family classification (retrieved from Tishkoff et al. 2009).

3.2.1.5 Searching for biogeographic ancestry

By comparing the mtDNA and Y chromosome haplogroup variation between S.Tomé and a reference database from several populations from Africa, we found that the observed haplogroups could be traced back mainly to regions of West Africa and West-Central Africa, reflecting the major contributions of these broad geographic settlement to the island. Given the high susceptibility of the mitochondrial DNA (mtDNA) and the non-recombining portion of the Y chromosome (NRY) to genetic drift, these markers display marked frequency differences in major regions of the African continent. This feature could be used to trace the relative ancestry of populations derived from the slave trade to broad geographic regions of Africa. However, the usefulness of ancestry estimation is not restricted to the study of populations or groups. It has been observed that an increasing number of African American individuals, descendant from enslaved Africans, are seeking more information on their Old World ancestries. The lack of personalized historical information about individual's geographic ancestry, make the genetic information an helpful resource (Shriver and Kittles 2004). MtDNA and NRY haplotypes have been the markers of choice to the majority of the commercial companies providing direct-to-consumer genetic ancestry tests. Nevertheless, the results obtained through such genetic systems are often misunderstood and misrepresented to costumers (Royal et al. 2010). It is important to have in mind that the information obtained from uniparental markers represents only a small fraction (less than 1%) of individuals genetic inheritance (Reed and Tishkoff 2006). There are also important confounding factors related with recent population movements that reshuffled past lineage distributions and with the fact that many lineages are spread throughout broad geographic areas (Reed and Tishkoff 2006). Another major difficulty concerning the identification of the origin of lineages with known African ancestry is the low quality of reference databases. In fact, the major areas of slave recruitment in Africa- mainly the region between Gabon and Angola- are largely unrepresented in genetic studies, making the conclusions retrieved highly prone to error (Reed and Tishkoff 2006). Our work about the genetic diversity of the Southwest Angola (see Part 2 of this thesis) will fill to some extent this gap.

References

Barbujani, G. 1997. DNA variation and language affinities. *Am J Hum Genet* 61:1011-4.

Belle, E. M., and G. Barbujani. 2007. Worldwide analysis of multiple microsatellites: language diversity has a detectable influence on DNA diversity. *Am J Phys Anthropol* 133:1137-46.

Caldeira, A.M. 2004. Rebelião e outras formas de resistência à escravatura na ilha de São Tomé (séculos XVI - XVIII). *Africana Studia* 7: 101-136.

Castelo-Branco, F. 1971. Subsídios para o estudo dos «angolares» de S.Tomé. *Studia* 33: 149-159.

Cavalli-Sforza, L. L. 2000. *Genes, peoples and languages*. New York: North Point Press.

Darwin, C. R. 1882. The descent of man, and selection in relation to sex. London: John Murray. 2nd edition, fifteenth thousand http://darwin-online.org.uk/

Friedlaender, J. S., F. R. Friedlaender, J. A. Hodgson, M. Stoltz, G. Koki, G. Horvat, S. Zhadanov, T. G. Schurr, and D. A. Merriwether. 2007. Melanesian mtDNA complexity. *PLoS One* 2:e248.

Hagemeijer, T. 1999. As ilhas de Babel: a crioulização no Golfo da Guiné. Camões 6: 74-88.

Lansing, J. S., M. P. Cox, S. S. Downey, B. M. Gabler, B. Hallmark, T. M. Karafet, P. Norquest, J. W. Schoenfelder, H. Sudoyo, J. C. Watkins, and M. F. Hammer. 2007. Coevolution of languages and genes on the island of Sumba, eastern Indonesia. *Proc Natl Acad Sci U S A* 104:16022-6.

Macgregor, S., C. Bellis, R. A. Lea, H. Cox, T. Dyer, J. Blangero, P. M. Visscher, and L. R. Griffiths. 2009. Legacy of mutiny on the Bounty: founder effect and admixture on Norfolk Island. *Eur J Hum Genet* 18(1):67-72.

Ramachandran, S., O. Deshpande, C. C. Roseman, N. A. Rosenberg, M. W. Feldman, and L. L. Cavalli-Sforza. 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci U S A* 102:15942-7.

Reed, F. A., and S. A. Tishkoff. 2006. African human diversity, origins and migrations. *Curr Opin Genet Dev* 16:597-605.

Royal C. D., J. Novembre, S. M. Fullerton, D. B. Goldstein, J. C. Long, M. J. Bamshad, and A. G. Clark. 2010. Inferring genetic ancestry: opportunities, challenges, and implications. *Am J Hum Genet* 14;86(5):661-73.

Salas, A., M. Richards, M. V. Lareu, B. Sobrino, S. Silva, M. Matamoros, V. Macaulay, and A. Carracedo. 2005. Shipwrecks and founder effects: divergent demographic histories reflected in Caribbean mtDNA. *Am J Phys Anthropol* 128:855-60.

Scheinfeldt, L. B., S. Soi, and S. A. Tishkoff. 2010. Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. *Proc Natl Acad Sci U S A*. 11;107 Suppl 2:8931-8.

Seibert, G. 1998. A questão das origens dos angolares de São Tomé. CEsA Brief Papers n.º 5. Lisbon.

Shriver, M. D., and R. A. Kittles. 2004. Genetic ancestry and the search for personalized genetic histories. *Nat Rev Genet* 5:611-8.

Tishkoff, S. A., F. A. Reed, F. R. Friedlaender, C. Ehret, A. Ranciaro, A. Froment, J. B. Hirbo, A. A. Awomoyi, J. M. Bodo, O. Doumbo, M. Ibrahim, A. T. Juma, M. J. Kotze, G. Lema, J. H. Moore, H. Mortensen, T. B. Nyambo, S. A. Omar, K. Powell, G. S. Pretorius, M. W. Smith, M. A. Thera, C. Wambebe, J. L. Weber, and S. M. Williams. 2009. The genetic structure and history of Africans and African Americans. *Science* 324:1035-44.

Trovoada, M. J., L. Pereira, L. Gusmão, A. Abade, A. Amorim, and M. J. Prata. 2003. Pattern of mtDNA variation in three populations from São Tomé e Príncipe. *Ann Hum Genet* 68:40–54.

Trovoada, M. J., L. Tavares, L. Gusmão, C. Alves, A. Abade, A. Amorim, M. J. Prata. 2007. Dissecting the genetic history of São Tomé e Príncipe: a new window from Y-chromosome biallelic markers. *Ann Hum Genet* 71(Pt 1):77-85.

CONCLUDING REMARKS

The studies included in this work fall within two main areas of research on human evolutionary history: the analysis of functionally relevant genes and the reconstruction of human demographic history. The study of the evolutionary history of the lactase persistence may be included in the former area while the demographic studies on Bantu expansions and of the peopling of São Tomé belong to the latter area.

By studying the genetic variation associated with the -13910*T lactase persistence-associated allele, we provided support to the notion that lactase persistence underwent a rapid increase in frequency due to a selective advantage (*Article 1*). The use of microsatellite markers and the analysis of geographically and ethnically distinct populations allowed us to exclude the possible confounding effects of recombination suppression and shared population history. We have also shown that the low predictive value of the -13910*T allele for lactase persistence in most African populations had to be due to separate mutations, in Europe and in Africa. Our predictions were fully confirmed by posterior studies showing that several new sequence variants (-14010*G/C, -13915*T/G and -13907*C/G), in very close proximity to the -13910*C/T polymorphism, were associated with the lactase persistence in populations from East Africa and Middle East. The observation that at least four causal mutations associated with lactase persistence have evolved independently in geographically distinct populations, illustrates how the sharing of a selective pressure- such as adult milk consumption- may lead to convergent evolution. At the same time these results call attention into the limitations associated with genome-wide scans for natural selection that are based only in a limited set of populations. In fact, several important signs of selection might have been bypassed due to lack of representative samples in the commonly used population panels.

On the other hand, these observations provide interesting information about the spread of pastoralist populations, illustrating that the separation between adaptive and demographic studies is somewhat artificial. We found that the -14010*C allele, associated with lactase persistence in East Africa, was also present in southwestern Angola, reaching its maximum frequency among the pastoralist Herero-speaking Kuvale population (*Article 2*). Given that the Kuvale rank among the most exclusively pastoral peoples of southwestern Africa, our observation provides a genetic evidence for a link between the relatively isolated southwestern Africa pastoral scene and the major cattle herding centers of East Africa. These observations coupled with the results on mtDNA and Y-chromosome variation showing high levels of admixture between the Kuvale and the Khoisan, led us to suggest that the link between the eastern and southwestern African pastoral scenes was established indirectly, through migrations of Khoisan herders across southern Africa. Taken together, our results indicate that Khoisan peoples, far from being isolated remnants from an ancient past, have played a major

role in the cultural and genetic transitions that shaped the current patterns of human diversity in southern Africa.

The choice of the genetic markers is an important aspect in studies of demographic history. Due to its intrinsic characteristics, like mutation rate and mode of transmission, and depending on the type of questions to be addressed, the use of each marker is associated with specific benefits and drawbacks. It has been shown that the combination of markers with different properties may offer insights that could not be achieved using each marker type in isolation. Our newly developed battery of UEPSTRs provides an illustration of the advantages of combining different type of markers to explore the evolutionary history of human populations (*Article 3*). However, it is important to note that, independently of progresses in marker developing, to make full use of the data generated by different laboratories it is crucial to have increasingly comparable datasets. This should be achieved by defining a minimum subset of highly informative markers to be used in future works about other African populations.

Another aspect of the recent advances in understanding human genetic diversity is related with data analysis. Datasets based on multiple, independently evolving genetic systems are particularly well suited to simulation-based inferential frameworks that are aimed to distinguish between alternative models of population history and to estimate key microevolutionary parameters under a given model. Recent applications of rejection algorithms and Approximate Bayesian Computation to infer the branching history of Pygmy and agricultural populations provide good examples of the usefulness of new computational methods to address population history in Africa (Patin el al. 2009, Verdu et al. 2009). Our application of model based methods using full likelihood (*Article 2*) or approximate methods (*Article 3*), provides a further contribution to improve inferational framework studies addressing the Bantu expansions.

To disentangle the spatial-temporal processes that gave rise to the present human genetic diversity, it is important to address both deep-time and more fine-scale questions, combining continent-wide studies with more detailed pictures provided by regional or local case studies. Our study of the small island of São Tomé located in the Gulf of Guinea, demonstrates that monographic studies of recently emerged populations may give important insights into general aspects of human microevolution (*Article 4*). Given the small size of the island and the social conditions forcing massive amalgamation, the strong genetic differentiation in São Tomé is particularly intriguing, providing an illustration of the way genetic clusters may arise and how the human genetic diversity can be shaped by such unexpected factors as people's desire to get free.

References

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

Verdu, P., F. Austerlitz, A. Estoup, R. Vitalis, M. Georges, S. Thery, A. Froment, S. Le Bomin, A. Gessain, J. M. Hombert, L. Van der Veen, L. Quintana-Murci, S. Bahuchet, and E. Heyer. 2009. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol* 19:312-8.