Slavery, Language and DNA: A tale of São Tomé and Príncipe

João Tiago Brochado Almeida
Mestrado em Biodiversidade, Genética e Evolução
Departamento de Biologia
2017

Orientador
Jorge Macedo Rocha, Professor Associado, FCUP

Coorientador
Magdalena Gayà Vidal, Investigador Auxiliar, CIBIO-InBIO

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Acknowledgements

This work would not have been a possibility without the help, support and dedication of the following people whom I thank immensely:

First, to my supervisor Prof. Jorge Macedo Rocha for the time dedicated to this project and to teaching and guiding me. Thank you for allowing me to do this work alongside you and specially thank you for all the enthusiasm, wisdom and will that you passed on to me during this Masters.

My deepest thanks to Magdalena Gayà Vidal for teaching me the so much needed practices for this work as well as for having the time and patience to answer and to discuss my recurrent doubts and opinions.

My thanks to the other members of the HUMANEVOL group, Armando Semo, Sandra Oliveira and Anne Fehn for the spontaneous help and care.

I would like to thank to CTM's technicians Jolita Dilytė and Sandra Afonso for the help and the advices during the laboratory work.

I am thankful to all the people who were willing to help, in particular my friends and colleagues.

And finally, I want to thank and dedicate this work to my parents and sister for all the support and care.

Abstract

The archipelago of São Tomé and Príncipe, located in Gulf of Guinea, was found uninhabited in 1470's and played a significant role in the Atlantic slave trade. European settlers, mainly from Portugal, and slaves recruited from the Bight of Benin to Angola were the main sources of peopling of the islands. Because of this multiplicity of source populations, the archipelago is notable for its human diversity, which is reflected in the three creole languages presently spoken in its islands. Lungwa Santome and Lunga Ngola, both spoken in São Tomé by Forros and Angolares respectively, display a greater influence of Bantu languages found in Congo-Angola. In contrast Lung'ie, the creole spoken in Príncipe, retained more characteristics of Edoid languages from Niger-Congo.

Here, we used for the first time a Whole-Exome sequencing approach to evaluate the levels of genetic differentiation of the major groups found in São Tomé and Príncipe and to measure the relative contribution of different populations to the current genetic makeup of the archipelago.

By analyzing 102,772 single nucleotide polymorphisms in a sample of 24 individuals belonging to different creole-speaking communities we estimated an influence of the area of the Gulf of Guinea of ~60% and that of the region of Congo-Angola of ~40%, both in Lungwa Santome-speaking Forros and in inhabitants of Príncipe descending from Lung'ie speakers. Angolares displayed a remarkable level of genetic differentiation, displaying an overrepresentation of a minor genetic component that was exclusively found in Bantu-speaking populations.

Our estimates show that there is no apparent correlation between genes and languages in Forros and Príncipe inhabitants. In contrast, Angolares represent a remarkable case of gene-language correlation, the strong influence of their lexicon is congruent with a Bantu overrepresentation in their genes.

We found contributions from European colonizers to be relatively low in Forros (13%) but much higher than in Príncipe (3%) and virtually absent in Angolares.

Our results confirm the previously observed signal of Angolares as an extremely differentiated group within the archipelago. Their genetic distinctiveness is not paralleled by Príncipe. Limits on the resolution capacity of our polymorphisms, as well as a high degree of genetic distinctiveness in Angolares makes it difficult to pinpoint a region of origin for them, even though we detect a major Bantu influence.

Finally, we explored the functional consequences of an extreme genetic differentiation by identifying coding SNPs that have an excess of differentiation between Forros and Angolares, recognizing functional differences in relevant processes that are related to fructose 2,6−bisphosphate metabolism, negative regulation of myoblast fusion, skeletal muscle cell proliferation, muscle atrophy and the Major Histocompatibility Complex.

Keywords

São Tomé and Príncipe; Slave trade; Gulf of Guinea creoles; Expanded exome; Population structure; Admixture; Niger-Congo; Western Bantu; Gene Ontology.

Sumário

O arquipélago de São Tomé e Príncipe, localizado no Golfo da Guiné, foi descoberto desabitado nos anos de 1470 e teve um papel significativo no comércio transatlântico de escravos. Colonos Europeus, principalmente de Portugal, e escravos recrutados desde a Baia de Benim até Angola foram as principais fontes de povoamento das ilhas. Foi devida a esta multiplicidade de populações de origem que o arquipélago se tornou notável pela sua diversidade humana que se encontra refletida em três línguas crioulas caraterísticas de grupos distintivos. Lungwa Santome e Lunga Ngola, ambas faladas em São Tomé por Forros e Angolares respetivamente, exibem uma maior influência de línguas Bantu encontradas no Congo-Angola. Contrariamente, Lung'ie, o crioulo falado no Príncipe, retém mais caraterísticas gramaticais de línguas Edoides do Níger-Congo.

Realizamos pela primeira vez um estudo de alta-definição explorando uma abordagem de sequenciação do exoma completo para avaliar os níveis de diferenciação genética dos grupos principais encontrados em São Tomé e Príncipe e para medir a contribuição relativa das diferentes populações para a composição genética do arquipélago.

Ao analisar 102.772 marcadores genéticos num tamanho amostral de 24 indivíduos pertencentes a diferentes comunidades crioulas estimamos uma influência de ~60% da área do Golfo da Guiné e de ~40% da região do Congo-Angola, tanto nos Forros falantes de Lungwa Santome como nos habitantes de Príncipe descendentes de falantes Lung'ie. Os Angolares mostram um nível notável de diferenciação genética, apresentando uma sobre-representação de um componente genético encontrado exclusivamente em populações de língua Banto.

As nossas estimativas mostram que não existe uma correlação aparente entre genes e linguagem nem nos Forros nem nos habitantes de Príncipe. Contrariamente, os Angolares representam um caso notável de correlação genes-linguagem, as fortes influências no seu léxico são congruentes com uma sobre-representação Banto nos seus genes.

Encontramos relativamente baixas contribuições de colonizadores Europeus nos Forros (13%), mas ainda assim mais altas que em Príncipe (3%) e praticamente inexistentes nos Angolares.

Os nossos resultados confirmam os sinais observados anteriormente identificando os Angolares como um grupo extremamente diferenciado no arquipélago. Este tipo de distinção genética não encontra um paralelo em Príncipe. Limites na capacidade de

resolução dos nossos polimorfismos, assim como um grande grau de distinção genética nos Angolares tornou difícil apontar uma região de origem para estes, mesmo detetando uma influência Banto grande.

Por ultimo, exploramos as consequências funcionais de uma diferenciação genética extrema através da identificação de polimorfismos codificantes que apresentam um excesso de diferenciação entre Forros e Angolares reconhecendo diferenças na função de processos relevantes á metabolização da frutose 2,6-bifosfato, á regulação negativa da fusão de mioblastos, á proliferação de células musculares esqueléticas, á atrofia muscular e ao Complexo Principal de Histocompatibilidade.

Palavras-chave

São Tomé e Príncipe; Comércio de escravos; Crioulos do Golfo da Guiné; Exoma expandido; Estrutura populacional; Miscigenação; Níger-Congo; Bantos Ocidentais; Ontologia Genética.

Index

List of Tables and Figures

List of Abbreviations

Abbreviation	Description
DNA	Deoxyribonucleic Acid
gDNA	Genomic DNA
mtDNA	Mitochondrial DNA
ESN	Esan from Nigeria
F_{ST}	Fixation index that accounts for deviations in the subdivided groups relative to the total population
GATK	Genome Analysis Tool Kit
GGC	Gulf of Guinea Creole
GO	Gene Ontology
IBS	Iberian Peninsula populations from Spain
LWK	Luhya from Webuye, Kenya
MDS	Multidimensional Scaling
MSL	Mende from Sierra Leone
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
RNA	Ribonucleic Acid
miRNA	Micro RNA
SNP	Single-Nucleotide Polymorphism
UTR	Untranslated Region
VQSLOD	Variant Quality Score log-odds
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

Introduction

Historical background

The Democratic Republic of São Tomé and Príncipe (Figure 1), located in Central Africa, in the Gulf of Guinea, near the Equator and west of Gabon, is composed of two of the four main islands that constitute the volcanic Cameroon line, the other two being Bioko (previously known as Fernando Pó) and Annobón (also known as Pagalu). The archipelago has a total area of 964 km^2 and a populational size of 197,541 (estimated in June 2016) (Henriques, 2000, CIA.gov, 2017).

The islands were discovered in the early 1470's by Portuguese captains Fernão do Pó, João de Santarém and Pêro Escobar (or Pedro Escobar), that were hired by an entrepreneur from Lisbon named Fernão Gomes (Tenreiro, 1961, Caldeira, 1999). Unlike Bioko, that was inhabited by the Bubi, or Annobón, which was too small, São Tomé and Príncipe were uninhabited at the time of their discovery and offered good settling conditions (Thomas, 1997, Henriques, 2000).

Figure 1 – São Tomé and Príncipe islands and main cities. *[Retrieved from Google Maps. Accessed 18 May 2017]*

The official colonization of São Tomé started fourteen years after its discover, with the nomination of Governor Álvaro Caminha, in a letter by the King of Portugal stating that the land would be used for sugarcane plantation and slave trade with the continental land near the delta of the Niger river (Thomas, 1997, Tenreiro, 1961). Príncipe's settlement begun in similar circumstances, only seven years later than São Tomé. In Príncipe, however, the plantation was less dominant, with the population focusing more on the trade with the kingdom of Benin, with which the island had exclusive trade privileges during the ruling of Governor António Carneiro (1500 - 1545) (Caldeira, 2005, Hagemeijer, 2011). From that time on, and until the end of the 16th century, São Tomé and Príncipe economies were dominated by the sugarcane production, with their settlement being made by Europeans, mainly Portuguese, and slaves traded with the ancient kingdom of Benin or Edo in the bight of Benin and Niger Delta, nowadays Nigeria (Thomas, 1997).

In the beginning of the 16th century an expansion of the slave recruitment areas heading to West Central Africa started to occur. These were Bantu-speaking regions, encompassing the kingdom of Kongo and Ndongo, located in parts of nowadays Gabon, the Republic of the Congo, the Democratic Republic of the Congo, Cabinda and North Angola (Caldeira, 1999, Klein, 1999, Hagemeijer, 2009). This expansion of the slave recruitment area made the Portuguese dominant in the Niger-Congo-speaking area, with the traffic being made all over the Gulf of Guinea, from the Kingdom of Dahomey, in the Slave Coast, nowadays Ghana and Togo, to the Kingdom of Ndongo, in Angola (Curtin, 1969, Thomas, 1997). Due to the increased economic status in São Tomé, a decree was issued by the King, encouraging miscegenation in the population. This statement arrived at the same time as the execution of two orders to manumit slaves, in 1515 and 1517, freeing the slaves offered to the first settlers (Caldeira, 1999). This manumission ("Alforria" in Portuguese) is what is said to be in the origin of the name Forro, to designate the descendants of freed slaves in the island of São Tomé. In spite of all the measures encouraging miscegenation between Portuguese and Africans, individuals with mixed descendance were still the minority of the population (Tenreiro, 1961, Caldeira, 1999, Henriques, 2000).

With the end of the 16th century came the decline in the production of sugarcane caused by the Brazilian competition. It is said that this decline was aggravated by a series of external French and Dutch attacks as well as robberies and attacks directed by groups of escaped slaves who lived in the inaccessible parts of the island, in the central forest and in the south. One of these groups was named "Angolares", after the legend that they descended from a group of slaves originating in Angola, who survived a shipwreck that

hypothetically occurred around 1540-1550, in the southwestern tip of São Tomé. Generally considered a separate ethno-linguistic group, this population stayed isolated for a long time and ended up settling in São João dos Angolares and Santa Catarina (Cunha Matos, 1916, Seibert, 1998). Others refute this hypothesis saying that it is more likely that the Angolares were formed by several waves of slaves that escaped from the plantations and gathered in the areas of difficult access of the island. Numerous records that blamed the Angolares for raiding the major cities and plantations, report not only damaged machinery and robbery of provisions, but also women kidnaping (Henriques, 2000, Caldeira, 1999).

After the decline and cease of sugarcane plantation, slavery in the Gulf of Guinea and provision supply to transatlantic slave ships became the main sustenance for São Tomé and Príncipe until the second half of the 18th century. However, in 1637 the Elmina fortress in Ghana (São Jorge da Mina) was taken by the Dutch leading to a big loss in the number of slaves imported to São Tomé and Príncipe from the Slave Coast and Gold Coast. This prevented Portugal from using export markets in what is now Ghana, in a time that the sale of slaves increased after the decline in gold production and the wars related to the rise of Asante and Akwanu in 1680. From 1721 onwards, after the reconstruction of the Fort of São João Baptista de Ajudá in Benin, the trades occurred East of what is now Benin, but mainly in the West coast of Central Africa (Klein, 1999, Tenreiro, 1961, Curtin, 1969).

At the beginning of the 19th century a demographic and economic increase occurred in São Tomé and Príncipe due to the newly introduced cultures of coffee and cacao. This boom, along with the abolition of slavery in Portuguese territories between 1869 and 1878, the measures against the Atlantic slave trade by other European nations and the refusal of the freed slaves to work in plantations, forced the hiring of laborers in other Portuguese colonies, such as Cape Verde, Angola and Mozambique. The descendants of this newly arrived contract laborers are named in the archipelago as Tongas (Caldeira, 1999, Henriques, 2000, Miers and Roberts, 1988).

A massive contribution of immigrants coming to the island of Príncipe from Cape Verde took effect after 1900, following a major outbreak of sleeping sickness that almost depopulated the entire island (Maurer, 2009, Nascimento, 2003). This lead to an underrepresentation of the original Príncipe's population and made it more difficult to localize the descendants of earlier settlers.

Linguistic diversity

Besides the official Portuguese language, three creoles with an African substrate that are lexically related to Portuguese can be found in São Tomé and Príncipe: Forro (also known as Lungwa Santome or São-Tomense), Angolar (a.k.a. Lunga Ngola) and Lung'ie (or Principense). These creoles are likely to descend from an early proto-creole that originated in the first years of colonization, retaining the grammatical characteristics of Edoid, a group of non-Bantu idioms of Nigeria that belong to the Niger-Congo language family (Hagemeijer, 2009, Hagemeijer, 2011). According to this view, Forro can be thought as a modification of this proto-creole of the Gulf of Guinea, with later contributions from western Bantu languages (Ferraz, 1979), reflecting the expansion of the original areas of slave recruitment into the region of Congo-Angola (Kikongo and Kimbundu); Lung'ie has a more visible Edoid linguistic influence that might have resulted from the retention of the original proto-creole and/or increased interactions between the rulers of the Príncipe island and the kingdom of Benin (see above). Presently it is an endangered language spoken only by 5-16% of the population on the island (Maurer, 2009, Hagemeijer, In press). Angolar, despite the extensive phonology and syntax similarities with Forro, is the language with a greater Bantu influence in its lexicon, particularly from Kimbundu, which is spoken around the city of Luanda, in Angola (Lorenzino, 1998). Table 1 summarizes the quantitative differences within the African component of the lexicon of the three creoles. The quantified contributions seem to agree with the historical data and the influence of the main areas of slave recruitment at different times.

Table 1 – Percentages of Edo and Bantu-derived words in the African lexicon of the three Gulf of Guinea creoles (GGCs) spoken in São Tomé and Príncipe (Rocha and Hagemeijer, 2012).

	Edo	**Bantu**
Lung'ie	76%[1]	24%
Santome	37%	63%
Ngola	4%	96%

[1] Based on Maurer (2009) this value goes up to 93%

Previews genetic studies

Genetic studies in São Tomé and Príncipe, focused on mtDNA (Coelho et al., 2008, Mateu et al., 1997, Trovoada et al., 2004), Y-chromosome (Coelho et al., 2008, Trovoada et al., 2001), or limited collections of autosomal data (Coelho et al., 2008, Tomás et al., 2002), which have shown high levels of diversity, similar to what can be found in the continent. This is expected given the massive importation of slaves coming from all the coast of the Gulf of Guinea and extending to Angola (Mateu et al., 1997).

Based on a limited set of 15 autosomal microsatellite markers, Coelho et al. (2008) found that the Angolares formed a genetically homogeneous group, displaying a remarkable level of differentiation from other groups of São Tomé. Other analyses based on the mtDNA (Coelho et al., 2008, Trovoada et al., 2004), and especially the Y-chromosome, showed that the Angolares might have experience a strong founder effect (Coelho et al., 2008).

In Príncipe the available genetic studies are not as developed as in São Tomé, in part due to the difficulties in identifying members of the Lung'ie-speaking community, or their descendants, given the high levels of admixture with individuals from Cape Verde in the island. More recent studies on Y-chromosome and mtDNA suggest that unrelated individuals with lineages descending from presumptive Lung'ie-speaking paternal grandfathers and maternal grandmothers, respectively, do not cluster in a distinct group when compared with other individuals of the archipelago, aside from Angolares (Ferreira, 2012, Machado, 2013).

Regarding admixture with European settlers, no contribution of European females was detected in São Tomé on the basis of mtDNA (Mateu et al., 1997, Coelho et al., 2008). When examining the Y-chromosome variation, however, traces of European lineages could be traced (Trovoada et al., 2001, Coelho et al., 2008). In addition, by using a battery of eight autosomal ancestry informative markers, a 10.7% average, overall European contribution was estimated for the island of São Tomé. However, this value dropped to 6.5% after removing from the sample pool all the individuals with at least one Portuguese or Cape Verdean ancestor in the two previous generations (Tomás et al., 2002).

To estimate the demographic contributions of different African regions to the peopling of São Tomé, Tomás et al. (2002) analyzed the distribution of β-globin S haplotypes that are known to predominate in different areas of slave recruitment. The authors estimated that, contrary to what might be expected from historical registries, the current population

of the Island received higher inputs from the Gulf of Guinea (~52%) than from the Congo-Angola area (36%).

However, in spite of these works, there are still important outstanding questions about the genetic history of São Tomé and Príncipe archipelago that can be better addressed using genome-wide approaches based in more powerful arrays of genetic markers. Before listing these questions, we will provide in the following section a brief summary of the methodological progresses involving the use of exome sequencing strategies to explore genome-wide variation.

Exome sequencing

With the extensive use of sequencing methods in current laboratories, it is now possible to aim for the whole human genome. However, given the cost and information load involved in such studies, it is still difficult to obtain the substantial amounts of data that are needed for populational studies (Warr et al., 2015). Therefore, individual laboratories preferably focus on retrieving data from known loci, few chromosomes or through the use of chip arrays. The issue with these kind of data, on the other hand, is that, as it used specific variants that were pinpointed *a priori*, it would be impossible to identify previously undocumented variation, introducing an ascertainment bias that leads to the underestimation of diversity in populations that were not used in variant-discover. This problem is well known and for some data it is possible to attempt to correct the ascertainment bias created by the method (Albrechtsen et al., 2010, Lachance and Tishkoff, 2013).

The expanded exome sequencing strategies aims for retrieving variation from about 62 Mb of genomic content (about 2% of the genome), including exons, untranslated regions (UTRs), and miRNA (Chilamakuri et al., 2014). The portion of the genome responsible for coding proteins and functional elements, even though it is not the most prone to accumulate variations, it is ideal for phenome, disease related and selection studies (Warr et al., 2015). Moreover, exome sequencing provides high coverage for private and rare variants that can be missed by low-coverage whole-genome data (The 1000 Genomes Project Consortium, 2012). This means that whole-exome sequencing (WES) is a viable solution for the study of the substantial amounts of variation that is needed for population genetics studies - at least while the improvement of coverage in the whole-genome sequencing (WGS) is in development and until the price for WGS inevitably reaches the cost of capture and sequencing needed for WES (Warr et al.,

2015). But possibly the most important feature of the WES is that, by not having an *a priori* characterization, it is not biased towards the over representation of certain populations, as it happens, e.g., with SNP Chips.

The WES techniques have been used mainly for genomic medicine (Teer and Mullikin, 2010, Bamshad et al., 2011), and human population studies (Kidd et al., 2014, Kim et al., 2014, Bustamante et al., 2005, Tennessen et al., 2010, Yi et al., 2010), being capable to find population structure and signs of selection in a feasible way. The present marketed WES kits are specialized for human sequencing, however, it is possible to extrapolate the use of this method to the study of animals (Warr et al., 2015) and even microbiome (Kidd et al., 2014).

Aims

The present work expects to contribute to a better understanding of the peopling history of São Tomé and Príncipe, by undertaking for the first time a high-resolution study of its genetic diversity, exploring a WES approach.

To achieve this goal, we focused on the following questions:

1) Can non-Bantu and Bantu-speaking Niger-Congo populations from major areas of slave recruitment be distinguished genetically?
2) What was the contribution of European colonizers and different African areas of slave recruitment to the genetic composition of the peoples that speak the three creole languages of São Tomé and Príncipe? Is there a link between these contributions and the linguistic characteristics of the creole languages of São Tomé?
3) How far is the previously identified genetic distinctiveness of Angolares reproducible with the genome-wide data and how does it compare with new samples from the island of Príncipe?
4) What is the origin of the Angolares?
5) What are the functional consequences of extreme genetic differentiation?

The answers to these questions will not only be important for understanding the emergence of creole societies that were formed during the Atlantic slave trade but will also assist in evaluating the phenotypic consequences of genetic differentiation between recently formed populations.

Methods

Sampling

The DNA samples were collected from cheek scraps, along with the registration of donor's demographic and ethnographic information. This information encompasses name, sex, age, birth locality (both from the individuals and their parents), language (individual wise and up to the grandparents) and information on relatedness.

The sequenced samples were selected according to the following criteria: the sampled individuals and their parents had to be born in the same city; sampled individuals could not be closely related, that is, if the data is available, they could not have a common ancestor as far as two previous generations; the four grandparents of individuals from Príncipe must have been born in the island and be acknowledged descendants from Lung'ie speakers.

We were able to isolate 24 samples in total: 16 from the island of São Tomé and the remaining eight from Príncipe. The samples from São Tomé were divided into two ethnic groups, eight samples from Angolares and eight samples of the Forro group. The Angolares samples were collected in São João dos Angolares (4) and Santa Catarina (4); the samples from Forros were collected in Guadalupe (3), Trindade (3) and Madalena (2). We chose these localities based on the historical distribution of these linguistic groups while trying to avoid areas with highly admixed individuals.

Whole Exome Sequencing

Capture of the expanded exome

We preformed the library preparation and the expanded exome enrichment using Nextera® Rapid Capture Enrichment kit by Illumina, Inc. and following the protocol version #15037436 v01 (January 2016).

The previously extracted genomic DNA (gDNA) samples were calibrated to achieve a concentration of 5 ng/µl in a volume of 10 µl which was attained through several quantifications with the fluorometric method Quant-iT™ PicoGreen®.

About 50 ng of gDNA of each sample were fragmented and tagged in a transposon-based method. To verify the success of the previous procedure, the length of the

fragmented DNA on half of the sample was calculated using an Agilent High Sensitivity DNA Chip in the Agilent Technologies 2200 TapeStation and found to have a median size of 262 bp (ranging from 179 bp to 321 bp).

The addition of the indexes and the adapters needed for cluster generation and sequencing was done through a 10-cycle PCR amplification program. The indexing was followed by a fragment purification process using Sample Purification Beads to select for size and eliminating unwanted products such as unligated adapters and adapter dimers. After purification, another analysis on the TapeStation was made to confirm the fragment selection, this time with an Agilent DNA 1000 Chip, showing a median size of 221 bp (ranging from 179 bp to 323 bp), along with this analysis we verified the quantity of DNA by using the Qubit™ 2.0 Fluorometer as dictated by the respective High Sensitivity protocol.

A concentration process was done prior to the hybridization to obtain the 500 ng in 40 µl required for the next step, calculated using the High Sensitivity fluorometric method for Qubit™ 2.0 Fluorometer. This concentration was performed using Amicon Ultra-0,5 centrifugal filter unit (0.5 ml, 30 kDa) with a centrifuge setup of 14000 G for 7 minutes.

In the hybridization process, there were added Expanded Exome Oligos and Enrichment Hybridization Buffer to target the regions of interest of the DNA library mix. To capture the probes hybridized to the target regions we used a process that required Streptavidin Magnetic Beads followed by two washes and an elution. This process of hybridization and cleaning was performed a second time with an increased incubation time to ensure high specificity.

A last step of library validation included quantification through fluorometry and a quality assessment. Two quantifications were done, one using Qubit™ 2.0 Fluorometer with High Sensitivity settings and another using qPCR along with the KAPA Library Quantification Kit for Illumina® platforms by Kapa Biosystems. The quality was measured through Agilent Technologies 2200 TapeStation with an Agilent High Sensitivity DNA Chip.

The pool of indexed samples was sequenced in two lanes using Illumina's HiSeq 1500 System in rapid run mode.

Read mapping and SNP calling

We did the quality control check of the raw sequence data with FastQC (v0.10.1) (Andrews, 2010) and applied a filter for Phred Quality Score of 30 (Q30) using Sickle (v1.33) (Joshi and Fass, 2011) in pair-end mode.

Quality accepted reads were aligned with the reference genome b37, available on the Genome Analysis Tool Kit (GATK) bundle, using the -mem option of Burrows-Wheeler Aligner (BWA) software (v0.7.15) (Li, 2013).

File conversion, sorting, indexing and merging was done with SAMtools (v1.3.1) (Li et al., 2009, Li, 2011). PCR duplicate reads were flagged with *MarkDuplicates*, a tool from Picard toolkit (v2.8.0) (http://broadinstitute.github.io/picard). Table 2 summarizes statistics for sequencing and mapping of the raw data.

The variant discovery workflow was done according to Genome Analysis Tool Kit (GATK) Best-Practices recommendations for exome sequencing (DePristo et al., 2011, Van der Auwera et al., 2013) and using corresponding software (v3.7) (McKenna et al., 2010). In the appropriate steps of the pipeline we selected the options that allowed an analysis only of the intervals for which the enrichment kit was optimized with an additional 100 bp of padding for each interval. All additional files needed to run the pipeline, which include dbsnp build 138, HapMap 3 genotypes, OMNI 2.5 genotypes for 1000 Genomes samples, and a set of known indels, were collected from the GATK bundle. First, a local realignment of reads around indels was performed using the GATK tools *RealignerTargetCreator* and *IndelRealigner*. Then, a detection of systematic errors in base quality scores (BQSR) was evaluated with GATK *BaseRecalibrator* and the recalibrated bases printed with GATK *PrintReads*.

Table 2 – Summary statistics for exome sequencing data from the populations in São Tomé and Príncipe.

	Total reads	**Unmapped reads**	**Mapped reads**	**%**	**Duplicates**	**Properly Paired**	**%**
Forros	20,334,599.0	745,465.9	19,589,133.1	95.1	1,323,027.6	19,191,186.0	93.1
Príncipe	41,045,322.9	236,285.8	40,809,037.1	99.4	2,825,011.5	40,403,632.0	98.5
Angolares	27,949,913.4	672,894.7	27,277,018.8	97.2	1,933,935.0	26,904,306.7	95.8

The simultaneous calling for SNPs and indels was achieved with GATK *HaplotypeCaller* per sample. We followed a guide recommendation that suggests the use of at least 30 samples in the next step. We gathered data from the 24 samples from this study and 13 additional samples from another study that used the same exome sequencing technic and pipeline (Gayà-Vidal, unpublished) and pass them all together to the joint genotyping tool, GATK *GenotypeGVCFs*. This resulted in 538,960 variant positions with 12.34 ± 5.27 mean coverage across our 24 samples. To filter this raw callset, the variant quality score log-odds (VQSLOD) was evaluated with GATK *VariantRecalibrator* and a threshold for the sensitivity to access true variation was applied with GATK *ApplyRecalibration* with a level of 99,0% for both SNPs and insertions/deletions (Indels), this lowed the number of variants that pass this filter to 497,473 with 11.97 ± 5.08 depth.

Callset Refinement

Variants from the dataset previously collected were further filtered relative to quality and data type in the following way: (i) Biallelic SNPs with genotype coverage between 3 and 100 and genotype quality at least 20 were retained using VCFtools (v0.1.13) (Danecek et al., 2011); (ii) Y-chromosome and pseudo-autosomal region (chrX:60,001-2,699,520 and chrX:154,931,044-155,260,560) were filtered out; (iii) Heterozygotic calls for male individuals in the X-chromosome were set to missing data; (iv) Hardy-Weinberg equilibrium for autosomes and female X-chromosome was evaluated with VCFtools and sites with a *p*-value below 0.0001 were excluded; (v) Sites with more than 15% missing data were also excluded. Statistics relative to the filtered dataset is presented in Table 3.

Phasing and genotype imputation was inferred with BEAGLE (v4.1) (Browning and Browning, 2007, Browning and Browning, 2016). To achieve this we added, temporarily, 8 individuals from European and African populations available in 1000 Genome database (The 1000 Genomes Project Consortium et al., 2015) and the African expanded exome data from our research group used in GATK *GenotypeGVCFs* to our dataset and run this cohort in the tool. The resulting set had phased data for 24 samples across 102,772 SNPs with no missing information.

Datasets

Relatedness statistics calculated on VCFtools (v0.1.13) (Manichaikul et al., 2010, Yang et al., 2010) showed that a pair of individuals from Príncipe had a kinship coefficient of 0.235, which led to the elimination of one sample, the one with the most missing data before imputation. The resulting dataset had, at most, 3rd-degree related individuals (highest value for kinship coefficient = 0.032) from São Tomé and Príncipe with data for 23 individuals across 102,772 SNPs.

The second dataset was a merge of the previous dataset with the populations Mende from Sierra Leone (MSL, n=85), Esan from Nigeria (ESN, n=99), Luhya from Kenya (LWK, n=97) and Iberian Peninsula populations from Spain (IBS, n=107) from 1000 Genome Project Database (The 1000 Genomes Project Consortium et al., 2015) and an aforementioned set from Gayà-Vidal (Unpublished), referred in this study as "Bantu", which includes the Nambya from Zimbabwe (n=1), Ronga (n=1) and Makhuwa (n=1) from Mozambique and Ganguela (n=5), Nyaneka (n=5), Ovibumdu (n=5), Himba (n=7) and Kuvale (n=7), all from Angola. The SNPs from 1000 Genome Project were filtered so that A-T and G-C positions were eliminated prior to the merge. All datasets were joined in a way that the resulting set consists only of common variants. This was then filtered for number of alleles per position to include biallelic SNPs only. The final dataset has information for 64.285 SNPs across 443 individuals with, at most, a 3rd-degree relationship.

Table 3 – Exome variant and statistics before imputation

	Forros	Príncipe	Angolares	All
Size	7*	8	9*	24
Variant dataset (Biallelic SNPs)				
T_i/T_v [1]				1.983
Mean Depth	7.05	14.65	11.02	11.98
Refined dataset				
T_i/T_v [1]				2.599
Mean Depth	16.81	31.01	28.05	27.75
Number of Sites	70,333	81,762	79,919	77,738
Singletons	1243	1228	844	1088
Missing (%)	0.145	0.005	0.027	0.054

* One individual from Forros was later identified as part of Angolares and thus had the population changed.

[1] Ratio of the number of transitions to the number of transversions

Statistical analysis

Population differentiation and admixture

Multidimensional Scaling (MDS) was made using PLINK (v1.07) (Purcell et al., 2007) which uses a clustering process based on pairwise identity-by-state, using the options --cluster and --mds-plot for two dimensions. Data was then plotted using R (v3.3.2) (R Core Team, 2016, Wickham, 2009).

Principal component analysis (PCA) was carried out with the *smartpca* tool from EIGENSOFT software (v6.1.4) (Patterson et al., 2006). The eigenvectors were plotted using *ggplot2* package from R (v3.3.2). The value for the variance explained for each principal component (PC) was calculated given the percentage of the eigenvalue of that component relative to the sum of the eigenvalues for all possible components.

Population-based F_{ST} pairwise values were calculated with EIGENSOFT (v6.1.4) which uses the Hudson estimator (Hudson et al., 1992, Keinan et al., 2007), which is not dependent on sample size (Bhatia et al., 2013). Neighbor-joining trees (Saitou and Nei, 1987) from F_{ST} data were drawn with SplitsTree4 (v4.14.5) (Huson and Bryant, 2006).

Model-based structure analyses were carried out with the program ADMIXTURE (Alexander et al., 2009), using the option –cv, for cross-value calculation, for *K* ranging from 2 to 5 in 10 iterations. The best run for each K was identified with pong (v1.4.7) (Behr et al., 2016) followed by respective visualization. We have also used a -supervised option, which allows to estimate the contributions of pre-defined parental populations into hypothesized hybrid groups.

Three-population tests (Reich et al., 2009) were done with *threepop* using the program TreeMix (v1.13) (Pickrell and Pritchard, 2012) which allows to detect evidence of admixture through the use of allele frequencies in modern populations. We accepted as indication of population admixture for f_3 statistic values with statistical significance (Z-score) of -1.64 or lower (corresponding to a *p*-value for one-tailed test of 0.05), as in Yang et al. (2014). We also tested for admixture using inferred trees from allelic frequencies using the option *treemix* in the program TreeMix (v1.13). Migrations were added successively from 1 to 4 with the option -m and the outgroup was set to be the Iberian Peninsula population.

Locus Differentiation

In an attempt to analyze the functional consequences of extreme interpopulation differentiation, that is, between the sub-populations of the population of São Tomé, we explored the levels of the per-locus differentiation between the Angolares and Forros using per-site F_{ST} values calculated with the Weir and Cockerham estimator (Weir and Cockerham, 1984). Since this method depends on the ratio of sample size, we randomly selected 10 individuals of each population that exceed that number.

We began our analysis by selecting the SNPs with most extreme differentiation, with F_{ST}>0.25 in pairwise comparisons involving the Angolares and Forros, Esan, and Iberian Peninsula. Then we eliminated all SNPs with F_{ST}>0.5 in comparisons with Esan and Iberian Peninsula, in order to target possible ancestral informative markers. By using these criteria, we obtained a list of 347 SNPs that was compared with a second list containing all polymorphisms in the tested populations.

The two lists were uploaded to Ensembl BioMarts (Kinsella et al., 2011) and the information for the Protein Identification Code was retrieved. The two Protein ID lists were tested for significant over-representation of functional annotations with the gene ontology (GO) term enrichment tool FatiGO (Al-Shahrour et al., 2004) in Babelomics 5 (Alonso et al., 2015). The information for Molecular Function (activities, such as catalytic or binding activities, that occur at the molecular level), Biological Process (series of events accomplished by one or more ordered assemblies of molecular functions) and Cellular Component for the data was considered to be over-represented in the first list when adjusted *p*-value from the Fisher's exact test (Fisher, 1935) was less than 0.05 after correction for multiple testing with the FDR Benjamini-Hochberg controlling procedure (Benjamini and Hochberg, 1995). The list of significantly over-represented terms was converted, from the Protein ID's associated with each GO term, to Gene Name and accompanied with a description based on information retreated from the databases Ensembl BioMarts and OMIM (https://omim.org).

REVIGO (Supek et al., 2011) relies on semantic similarity measures and it was used to summarize the list of GO terms using the clustering settings of 0.4 similarity (terms with 40% semantic similarities or more are merged in a more comprehensive term), adjusted *p*-values, the GO term database *Homo Sapiens* and the semantic similarity measure *simRel* (Schlicker et al., 2006).

Results and Discussion

Genetic Diversity

Based on a total of 102,772 SNPs we found that the mean values of the inbreeding coefficient *F* (Table 4) were higher for Angolares (0.0491) followed by the Príncipe sample (0.0164) and the Forros (-0.0405). These values show that Angolares are less diverse than the other two populations. Singletons followed an inverse correlation in contrast with heterozygosity, Angolares have the lowest number of singletons (844), then Príncipe (1228) and Forros (1243).

Interpopulation differentiation

Individual clustering

Figure 2 displays a PCA plot of the genetic relationships among the individuals from the three population samples from São Tomé and Príncipe. The Angolares are clearly isolated from other two groups in the first PC, confirming previous findings about their genetic distinctiveness (Coelho et al., 2008). Apart from the separation of two outlier individuals in the second PC, there is no clear differentiation between Príncipe and the Forro-speakers from São Tomé. Interestingly one individual initially classified as “Forro” was found to have a genetic composition indistinguishable from the Angolar samples. For this reason, his population label was changed in further analyses.

To assess the levels of intrapopulation differentiation in a wider context, the 23 samples from São Tomé and Príncipe were merged with the comparative dataset and their genetic relationships were assessed by Multidimensional Scaling (MDS). The MDS plot (Figure 3) shows: i) a clear separation between European and African populations along

Table 4 – Exome inbreeding coefficient and number of heterozygotes for populations on the study area before imputation.

	Forros		**Príncipe**		**Angolares**	
Heterozygotes	13,548	(19.26%)	15,034	(18.39%)	14,172	(17.73%)
F (Inbreeding Coefficient) [1]		-0.0405		0.0164		0.0491
Singletons	1243	(1.77%)	1228	(1.50%)	844	(1.06%)

[1] Mean values of the inbreeding coefficient *F* calculated as one minus the observed heterozygotes divided by the expected heterozygotes at Hardy–Weinberg equilibrium.

the first axis; ii) a close relationship between the non-Bantu Niger-Congo speakers Mende and Esan groups, which are nevertheless distinguishable from each other; iii) a clear separation between Bantu and non-Bantu speakers (Esan and Mende), within Niger-Congo; iv) a clear difference between the Luhya Bantu from Kenya and most Bantu speakers from Angola; vi) a tendency for samples from São Tomé to lie between the Bantu from Angola and the Esan from Nigeria, reflecting the historically acknowledged importance of areas of slave recruitment in the Gulf of Guinea and Congo-Angola to the peopling of São Tomé (Curtin, 1969, Thomas, 1997, Caldeira, 1999, Klein, 1999, Hagemeijer, 2009); vii) a tendency for some Forro individuals to be "pulled" towards the European populations.

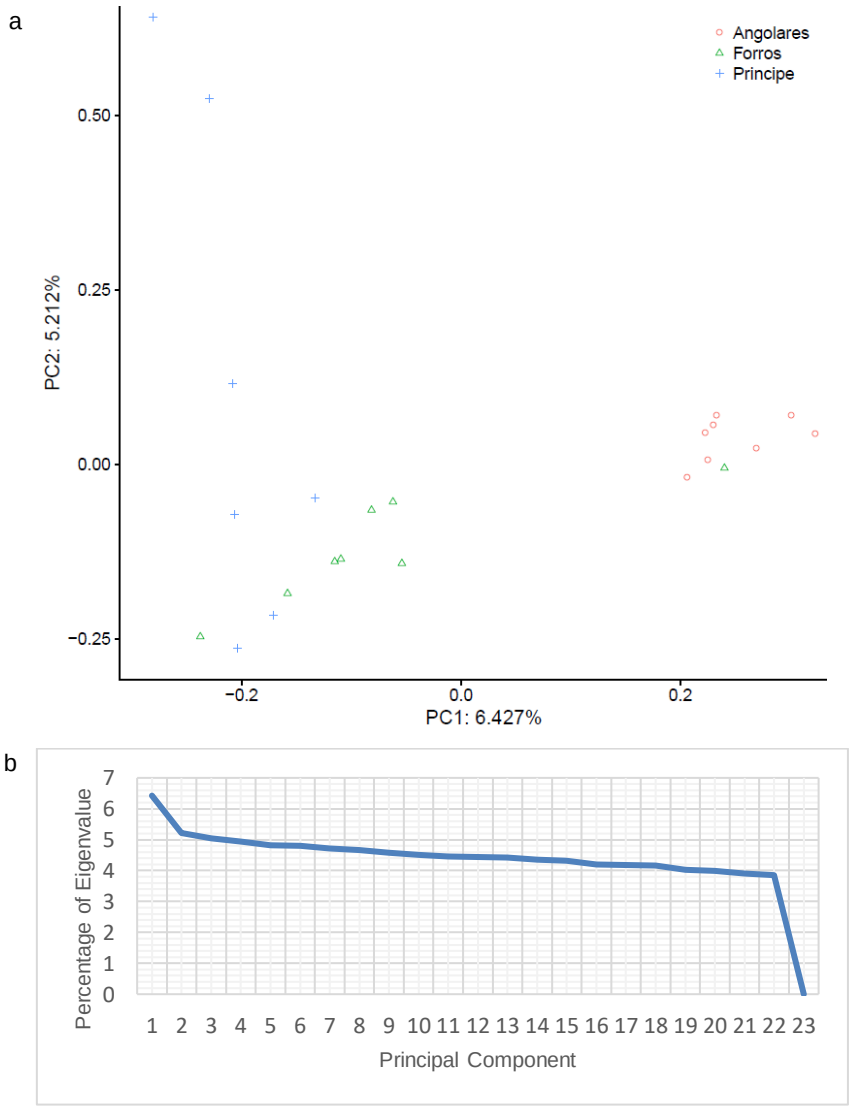

Figure 2 – a) PCA for the 23 São Tomé and Príncipe samples. The principal component variance explained, showed in the axis label, was calculated as percentage of the eigenvalue of that component relative to the sum of the eigenvalues for all possible components; b) Plot of the variance in eigenvalues across all principal components.

Since the MDS relationships could be affected by the high stress associated with axis fit, we performed a PCA with the same samples that is displayed in Figure 3. The results of the two first PCs are essentially identical to those obtained with MDS (Figure 4a). A third PC isolates the Angolares from all other samples (Figure 4b).

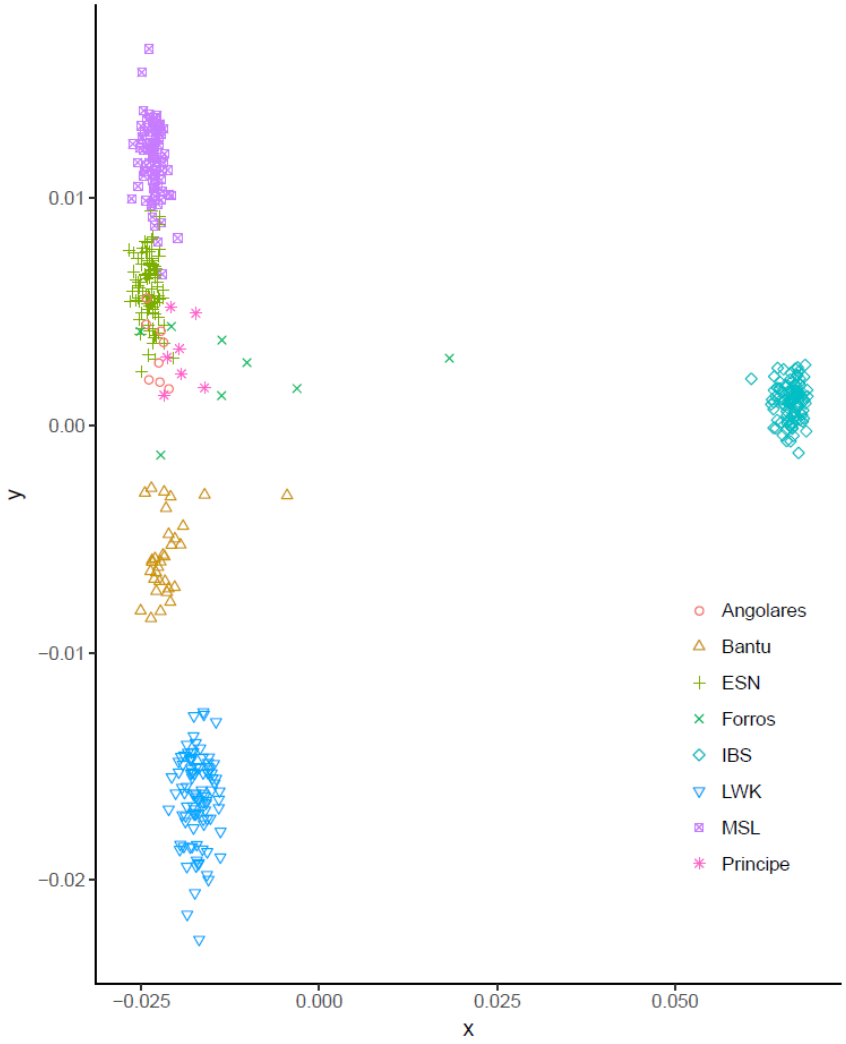

Figure 3 – Multidimensional Scaling plot for the second dataset. Population codes: IBS = Iberian Peninsula, LWK = Luhya, MSL = Mende, ESN = Esan.

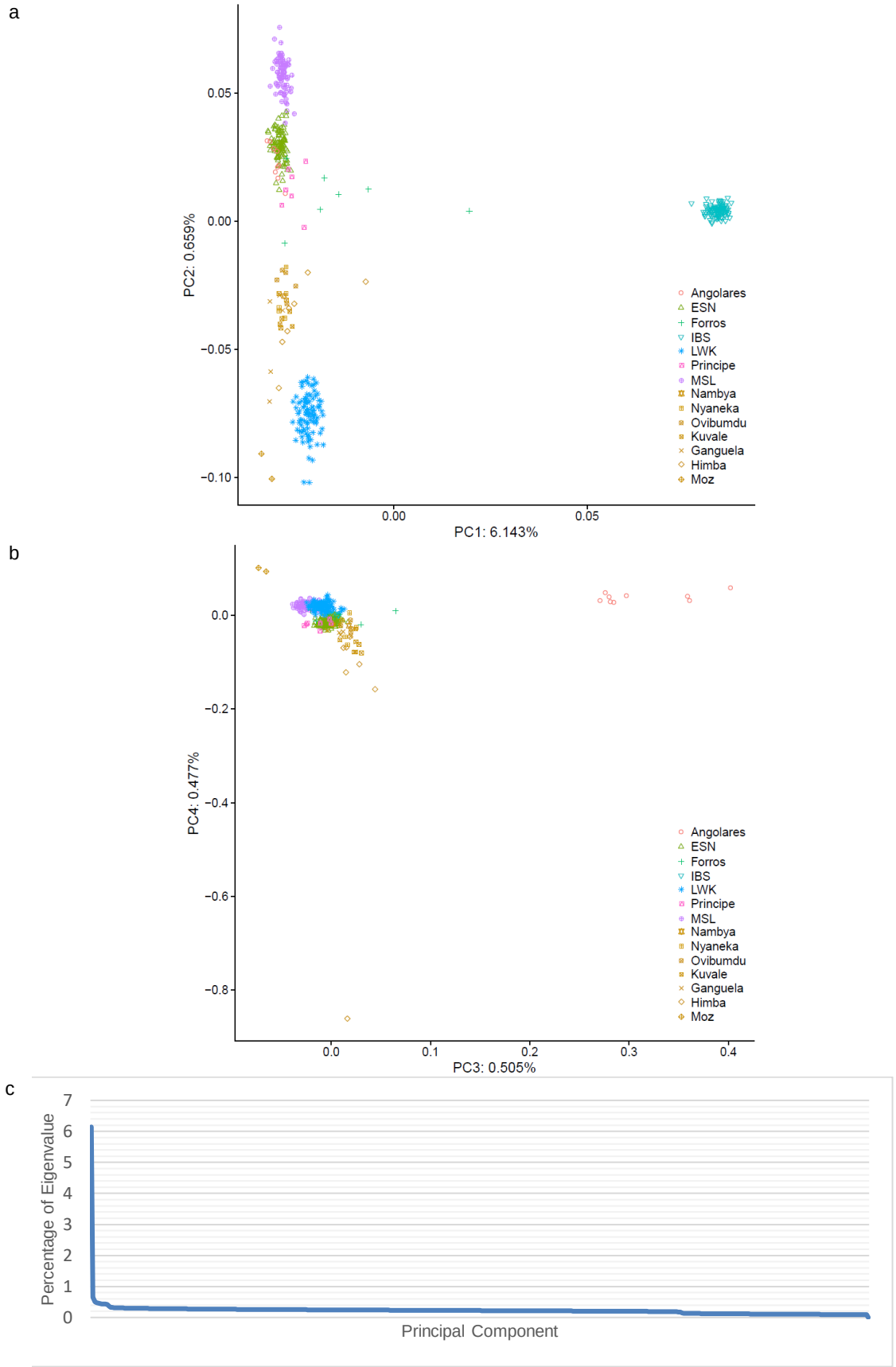

Figure 4 – a-b) PCA for all samples in the second dataset. Same color represents the same population with the Bantu (Beige) divided in distinct groups according to shape. Principal component variance explained, showed in the axis label, was calculated as percentage of the eigenvalue of that component relative to the sum of the eigenvalues for all possible components; c) Plot of the variance in eigenvalues across all principal components. Population codes: IBS = Iberian Peninsula, LWK = Luhya, MSL = Mende, ESN = Esan, Moz = Mozambique.

Finally, we carried out a model-based analysis of population structure using the ADMIXTURE program (Figure 5; Table 5). For K=2, corresponding to the lowest cross-validation error, European and African populations are clearly separated, as in the first axes of MDS and PCA (Figures 3 and 4). In accordance with these analyses, the Forros show non-negligible proportions of the European genetic component (13.2%), which is much lower in Príncipe (3.6%) and virtually absent in the Angolares. At K=3, the Bantu-speaking Luhya from Kenya are separated from the remaining populations, although the Luhya component (green) is also present in Bantu speakers from Angola and Mozambique (57.1%) and in all the groups from São Tomé and Príncipe (27.7%). At K=4, the two non-Bantu Niger-Congo-speaking populations are clearly differentiated: the Mende from Sierra Leone display a unique genetic component (violet), while the Esan from Nigeria remain more similar (but still different) to other Bantu-speakers. At K=5 the Angolares become individualized and display a genetic component (orange) that is partially shared with the Bantu from Angola and Mozambique.

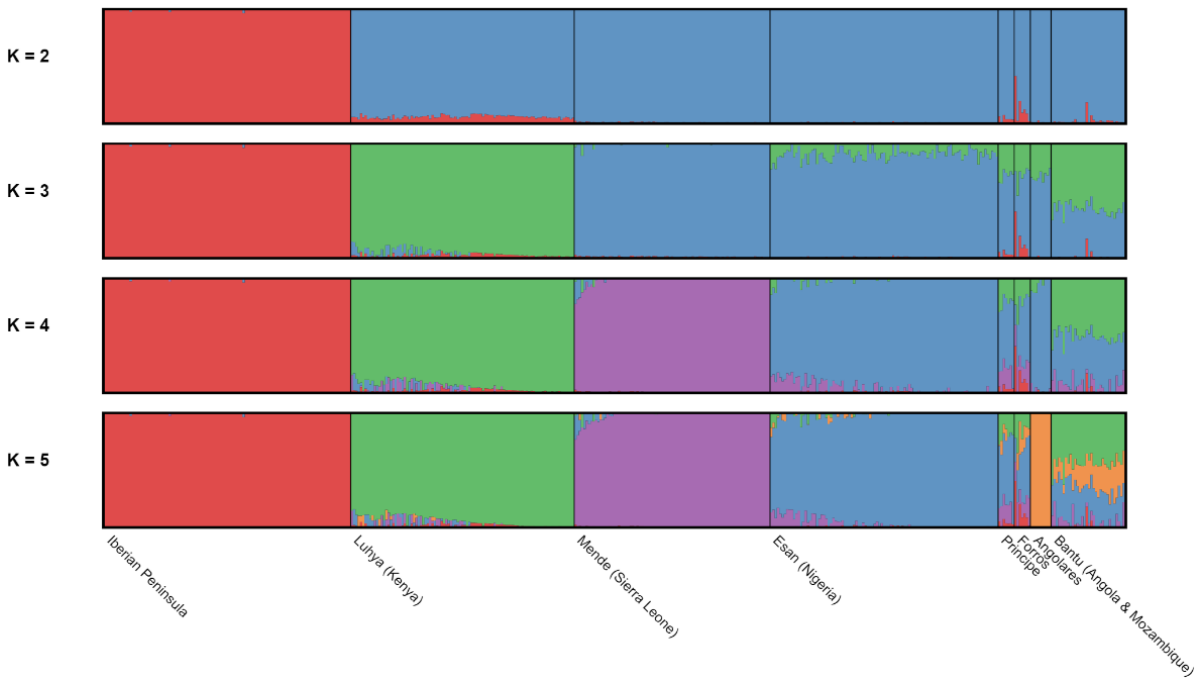

Figure 5 – Unsupervised inferred populational structure according to ADMIXTURE for the dataset encompassing all populations (443 individuals) for K (number of clusters) from 2 to 5. Individuals are grouped by population. Each individual is represented by a vertical bar. The proportion of the bar in each of k colors corresponds to the average posterior likelihood that the individual is assigned to the cluster indicated by that color.

Table 5 – Percentage of the estimated contribution, represented with the same colors, as in the ADMIXTURE for the populations in K=5.

Iberian Peninsula	99,90	0,03	0,03	0,03	0,03
Luhya (Kenya)	1,00	1,60	95,10	1,60	0,70
Mende (Sierra Leone)	0,17	1,30	0,17	98,20	0,17
Esan (Nigeria)	0,10	95,50	0,80	3,00	0,60
Príncipe	3,12	59,12	18,42	14,82	4,52
Forros	12,80	40,30	17,20	16,80	12,90
Angolares	0,00	0,00	0,00	0,00	100,00
Bantu (Angola + Mozambique)	0,78	28,68	43,88	5,38	21,28

Taken together, the results displayed in Figures 2 to 5 suggest that the sample from Príncipe and the Forros from São Tomé have a genetic composition that reflects the contributions of major areas of slave recruitment in the African mainland, while favoring a higher demographic impact from the Gulf of Guinea (non-Bantu) than the Congo-Angola (Bantu) regions. An exogenous contribution from Europeans is also visible and shows that the highest levels of European/African admixture occurred among the Forros. However, it is important to note that the way Príncipe's samples were selected, requiring all grand-parents to be born in the island, may have biased downward the amount of European miscegenation in this population. Finally, the Angolares are clear outliers and provide a remarkable example of gene-language correlation. According to the ADMIXTURE analysis, they seem to be more related to the African mainland Bantu peoples than to any other population from São Tomé (Figure 5; k=5). This could be interpreted as evidence that the Angolares represent a native population that settled São Tomé in pre-colonial times, as suggested by some Santomean historians (Esboço (1975) as cited in Seibert (1998)). However, since the results of MDS, PCA and ADMIXTURE analyses are mere descriptions of the patterns of genetic structure, there can be different explanations for this observation. An alternative explanation is that the Angolares were a group of escaped slaves that experienced a bottleneck effect leading to an overrepresentation of a genetic component of Bantu origin (Coelho et al., 2008, Trovoada et al., 2004, Trovoada et al., 2007). This hypothesis seems to fit better the available linguistic data on the derivation of the Angolar language from a common proto-creole, as well as the historical records attesting that São Tomé was uninhabited when found by Portuguese sailors (Thomas, 1997, Henriques, 2000).

Population clustering

We have also analyzed the relationships between the different samples from São Tomé and Príncipe using a more classical population approach based on F_{ST} genetic distances (Table 6).

As expected, Angolares show extreme levels of differentiation, and have average genetic distances to Forros and Príncipe (0.026) that are not very different from their genetic distances to other African populations, like the Luhya Bantu (0.029), the Angola and Mozambique Bantu (0.027), or the non-Bantu Esan (0.026) and Mende (0.028).

The results from pairwise comparisons are graphically displayed in the Neighbor-Joining trees from Figure 6. Of note are: i) the high differentiation of the Angolares when compared not only with the other two groups of the archipelago, but also with all other African populations; ii) the higher proximity of the Forros to the Europeans than any other African population; and iii) a slight link between Angolares and Forros.

These patterns are broadly consistent with the individual clustering analysis.

Table 6 – F_{ST} for all populations on the dataset according to Hudson estimator. Values differ in tone from red, lower, to green, higher. Population codes: IBS = Iberian Peninsula, LWK = Luhya, MSL = Mende, ESN = Esan.

	LWK	Príncipe	Forros	Angolares	MSL	ESN	IBS
Bantu	0,006	0,007	0,005	0,027	0,008	0,006	0,140
IBS	0,134	0,138	0,111	0,166	0,147	0,146	
ESN	0,009	0,005	0,004	0,026	0,006		
MSL	0,010	0,008	0,006	0,028			
Angolares	0,029	0,028	0,023				
Forros	0,005	0,006					
Príncipe	0,009						

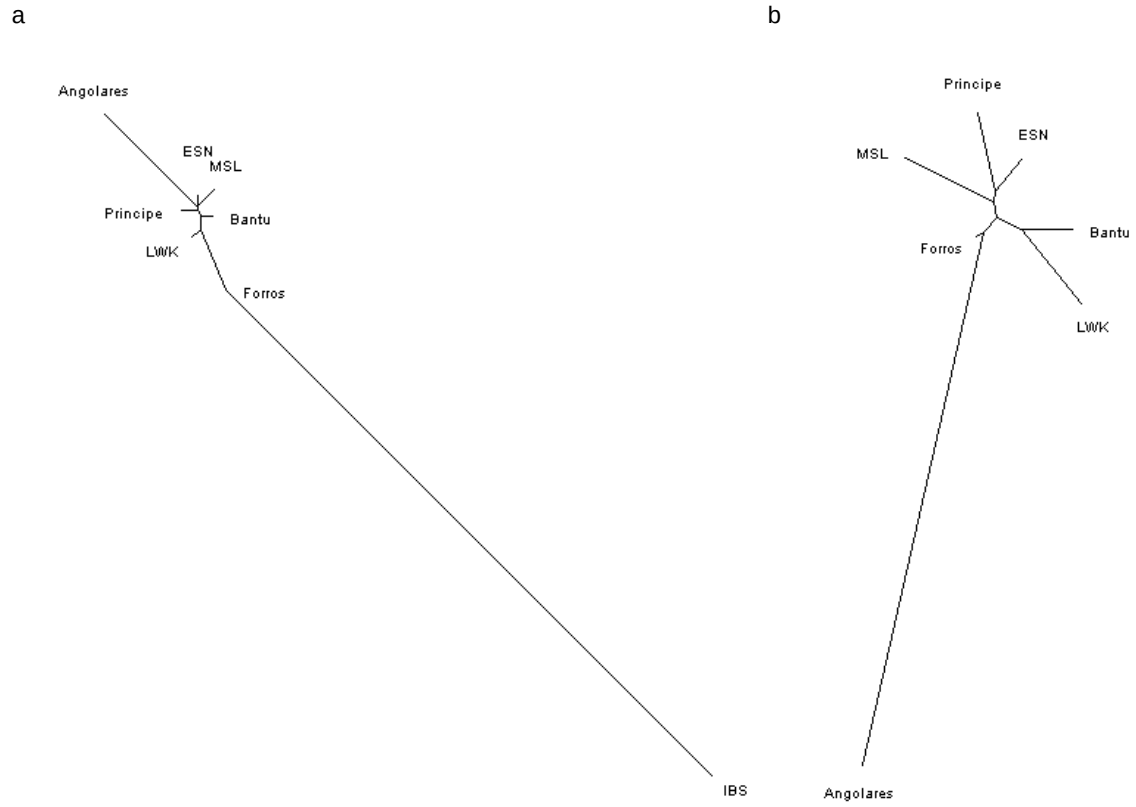

Figure 6 – Neighbour-Joining Equal Angle tree based on F_{ST} calculation for: a) all populations; b) African populations, on the second dataset. Population codes: IBS = Iberian Peninsula, LWK = Luhya, MSL = Mende, ESN = Esan.

Admixture

Since the results from population structure analyses indicate that the genetic composition of São Tomé and Príncipe might have resulted from contributions involving different areas of Africa and Europe, we attempted to characterize and quantify these contributions more formerly, in the context of an admixture framework.

TreeMix analysis

We started this analysis by using the approach implemented in the program TreeMix (Pickrell and Pritchard, 2012), which infers population splitting and mixing from genome-wide allele frequency data. The program initially provides a maximum-likelihood tree of populations assuming no migration. In subsequent steps, populations with a poor fit to the tree model are identified, and migration events involving these populations are modeled.

Figure 7a displays the initial tree of the studied populations, which is remarkably similar to the NJ network shown in Figure 6a. According to this tree the major separation is between Africans and Europeans, and the branching order for the African populations is:

Forros, Luhya, Bantu, Angolares, Príncipe and lastly Esan and Mende. It is interesting to note that, in spite of their high levels of differentiation, the Angolares are still closely related with Príncipe and the non-Bantu Niger-Congo populations (Figure 7a). Figures 7b-e show the sequential addition of four migration events with an estimation of the migration weight in Table 7. The first event (Figure 7b) results in gene flow from the European to Forros (12.56%) as expected from the analyses performed in the above sections. When this migration is taken into account, the branching pattern is altered to: Luhya, Bantu, Forros, Mende/Esan and finally Príncipe/Angolares. The second migration event reveals additional gene flow (3.90%) from the Europeans to Príncipe (Figure 7c). With this migration the branching order becomes: Luhya, Bantu, Forros/Angolares, Príncipe and Esan/Mende. This pattern is remarkably similar to the NJ tree that includes only the African populations (Figure 6b). Interestingly, when the contributions of Europeans are taken into account (or Europeans are removed from comparisons), the two populations from São Tomé (Forros and Angolares) have a recognizable relationship and become slightly closer to each other than to Príncipe, which is nearer the Niger-Congo non-Bantu-speaking groups (Figure 7c). The trees with three and four migrations maintain their topology and display additional migration events from Europeans to Luhya (Figure 7d) and from Angolares to Forros (Figure 7e), respectively. The migration from Europeans to Luhya is probably an indirect signal of the known presence of Eurasian genes in the peoples of East Africa, like the Maasai, with whom the Luhya are known to have admixed with (Busby et al., 2016). The migration from Angolares to Forros reflects the geographic proximity of the two populations in the island of São Tomé, and may explain the link observed in Figure 6b.

These results identify the genetic contribution from Europeans as the most important admixture event shaping the inferred relationships between the groups of São Tomé and Príncipe. Unexpectedly, none of these populations could be modeled as hybrids of parental populations from the African mainland, as suggested by the historical records or by previous studies of hemoglobin haplotypes (Tomás et al., 2002).

Three-population tests

We have also used the three-population test (Reich et al., 2009) that is designed to evaluate if a given population could have resulted from admixture between two hypothetical parental populations. Significant scores for admixture evidence based on the three-population test are presented in Table 8. According to this test, Forros could have resulted from admixture between Europeans and any of the African populations included in our dataset. The Luhya, for the reasons discussed above, are also

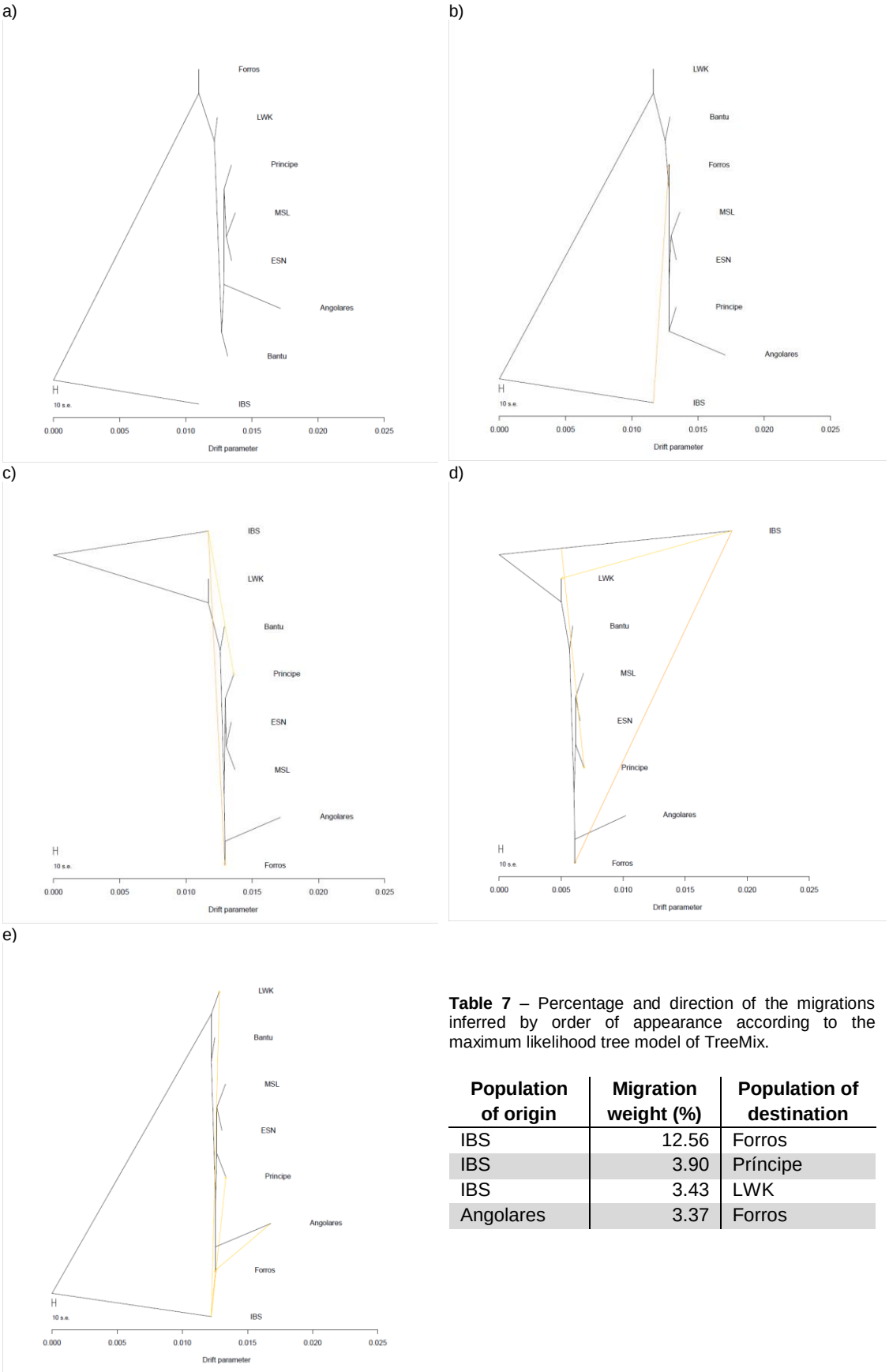

Table 7 – Percentage and direction of the migrations inferred by order of appearance according to the maximum likelihood tree model of TreeMix.

Population of origin	Migration weight (%)	Population of destination
IBS	12.56	Forros
IBS	3.90	Príncipe
IBS	3.43	LWK
Angolares	3.37	Forros

Figure 7 – Maximum likelihood trees and migration events inferred in TreeMix for the populations in the second dataset. **a**) Base tree without migration; **b** to **e**) Trees with 1 to 4 events of migration (modeled as arrows and colored according to their weight); The scale bar shows ten times the average standard error of the estimated entries in the sample covariance matrix. Population codes: IBS = Iberian Peninsula, LWK = Luhya, MSL = Mende, ESN = Esan.

Table 8 – Statistically significant results for admixture according to the f_3 test calculated for all populations on the dataset with the most populations.

Admixed Population	Parental Population 1	Parental Population 2	f_3	Standard error	Z-score
Forros	Esan	Iberian Peninsula	-3.108E-03	7.616E-05	-40.808
Forros	Mende	Iberian Peninsula	-2.970E-03	7.813E-05	-38.008
Forros	Iberian Peninsula	Bantu	-2.615E-03	8.215E-05	-31.833
Forros	Angolares	Iberian Peninsula	-3.247E-03	1.032E-04	-31.465
Forros	Príncipe	Iberian Peninsula	-2.237E-03	1.018E-04	-21.982
Forros	Luhya	Iberian Peninsula	-1.705E-03	7.956E-05	-21.428
Luhya	Esan	Iberian Peninsula	-5.876E-04	3.990E-05	-14.729
Luhya	Mende	Iberian Peninsula	-4.175E-04	4.292E-05	-9.729
Luhya	Iberian Peninsula	Bantu	-3.960E-04	4.303E-05	-9.202
Luhya	Angolares	Iberian Peninsula	-5.658E-04	8.198E-05	-6.901
Príncipe	Esan	Iberian Peninsula	-2.145E-04	9.203E-05	-2.331

outstanding for being modeled as hybrids between African and Europeans. The Príncipe sample could be described essentially as an Esan population with European admixture. However, despite this result, the Forros and Angolares could not be described as resulting from admixture involving specific African populations.

The relative importance of European/African admixture events detected by both TreeMix and three-population tests, combined with the lack of signals of admixture involving only African populations, may reflect a lack of power to detect admixture when the parental populations are genetically close. Therefore, we decided to further analyze this type of admixture by using the “supervised” option of the program ADMIXTURE, in which the parental and hybrid populations are pre-defined. The results of this analysis are presented in the next section.

Supervised Admixture analysis

Figure 8 shows the results of a “supervised” ADMIXTURE analysis in which Esan, Iberian Peninsula and Bantu (Angola and Mozambique) are considered parental populations, while Forros and Príncipe are the hybrid populations. In these conditions the Esan and Bantu are proxies for the Gulf of Guinea and Congo-Angola areas, respectively, while the Iberian Peninsula represents the contribution of European genes, mostly mediated by the Portuguese colonizers. The estimated contribution of the three parental populations for the Forro population was (Table 9): 13% (Europe), 54% (Gulf of

Guinea), 33% (Congo- Angola). In Príncipe, these contributions were: 3% (Europe); 59% (Gulf of Guinea), 38% (Congo-Angola). If the European contribution is not accounted for, the respective contributions of Gulf of Guinea and Congo-Angola for the results were 62%:38% in Forros, and 61%:39% in Príncipe. These values are practically the same and suggest that the Gulf of Guinea area had an impact around 1.6 times higher than the Congo-Angola region in both the Forros and Príncipe inhabitants. This conclusion is remarkably consistent with the results previously obtained with hemoglobin haplotypes, suggesting a predominance of Gulf of Guinea (~60%) contributions for a general sample from the island of São Tomé (Tomás et al., 2002).

The supervised approach could not be applied to the Angolares, which, due to their high levels of genetic divergence, have a genetic composition that is inconsistent with the choice of parental populations.

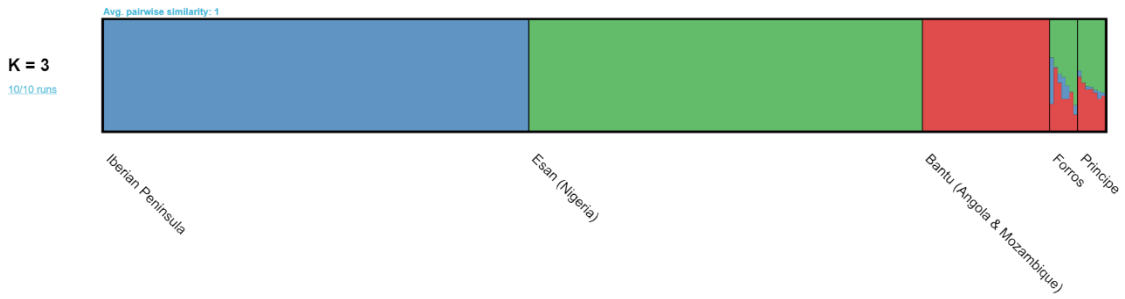

Figure 8 – Inferred ancestral components according to ADMIXTURE for the populations Forros and Príncipe with Iberian Peninsula, Esan and Bantu as fixed (supervised) populations.

Table 9 – Estimated contribution of three parental populations as inferred by ADMIXTURE for the populations of Forros and Príncipe.

	Iberian Peninsula	Esan	Bantu (Angola and Mozambique)
Forros	13%	54%	33%
Príncipe	3%	59%	38%

Locus Differentiation

As Angolares and Forros showed high levels of genetic differentiation that seem to be due to increased genetic drift, we explored this differentiation to understand the function consequences of random demographic processes.

With this aim, we identified the coordinates on the human genome of 1096 SNPs with high per-site differentiation (F_{ST}>0.25) between Angolares and Forros (Figure 9). After eliminating possible ancestral informative markers (SNPs with per-site F_{ST}>0.5 for Esan/Iberian Peninsula), the remaining 347 positions were identified as part of 235 Biological Processes (Table 10). It was possible to group various Biological Processes (Table 10) as part of pathways that manage fructose 2,6−bisphosphate metabolism (orange), negative regulation of myoblast fusion (Table 10; red), skeletal muscle cell proliferation (blue), muscle atrophy (violet) or Major Histocompatibility Complex mediated processes (green).

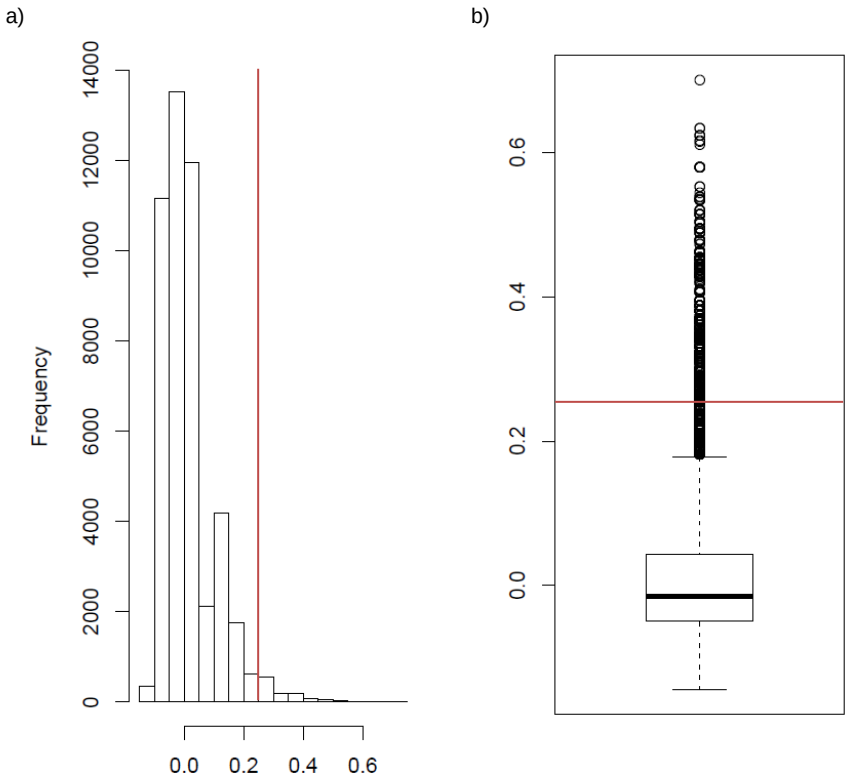

Figure 9 – Bar-plot (a) and Box-plot (b) displaying the F_{ST} for each position calculated between the populations of Angolares and Forros. The red line represents F_{ST}=0.25.

Table 10 – Top 20 of 235 biological processes, ordered by Adjusted *p*-value, for positions with F_{ST}>0.25 (target) between Angolares and Forros with the respective percentage of presence in the target or background lists. Processes colored the same participate in similar proceedings thus making a "supercluster", e.g., fructose 2,6−bisphosphate metabolism (orange), negative regulation of myoblast fusion (red), skeletal muscle cell proliferation (blue), muscle atrophy (violet) or Major Histocompatibility Complex (green) related. The color association was defined by REVIGO using the settings tiny similarity, *p*-values, GO term database *Homo Sapiens* and semantic similarity *SimRel*.

Biological Process	Presence (%)		Adjusted *p*-value
	Target list	Background list	
Fructose 2,6-bisphosphate metabolic process	0.65	0.03	2.75E-07
Negative regulation of myoblast fusion	0.53	0.01	3.81E-07
Negative regulation of syncytium formation by plasma membrane fusion	0.53	0.02	7.77E-07
Regulation of satellite cell proliferation	0.53	0.02	9.66E-07
Regulation of skeletal muscle cell proliferation	0.53	0.02	1.04E-06
Chorismate metabolic process	0.53	0.02	1.04E-06
Phosphatidylinositol phosphorylation	1.00	0.13	1.16E-06
Skeletal muscle atrophy	0.59	0.03	1.61E-06
Regulation of syncytium formation by plasma membrane fusion	0.65	0.04	1.63E-06
Syncytium formation by plasma membrane fusion	0.76	0.07	1.76E-06
Skeletal muscle satellite cell proliferation	0.53	0.02	1.76E-06
Striated muscle atrophy	0.59	0.03	1.76E-06
Muscle atrophy	0.59	0.03	1.76E-06
Syncytium formation	0.76	0.07	1.76E-06
Skeletal muscle cell proliferation	0.53	0.03	2.49E-06
Positive regulation of antigen processing and presentation of peptide antigen via MHC class I	0.41	0.01	2.49E-06
Antigen processing and presentation of peptide antigen via MHC class Ib	0.41	0.01	2.49E-06
Antigen processing and presentation of endogenous peptide antigen via MHC class Ib	0.41	0.01	2.49E-06
Antigen processing and presentation of endogenous peptide antigen via MHC class Ib via ER pathway, TAP-dependent	0.41	0.01	2.49E-06
Antigen processing and presentation of endogenous peptide antigen via MHC class Ib via ER pathway	0.41	0.01	2.49E-06

The genes with highest contributions for the Biological Processes showed on Table 10 are represented in Table 11 along with a brief functional description.

The extreme differentiation of the two populations present in São Tomé allowed us to identify metabolic divergences and target genes that may be important in shaping patterns of health and disease. A global view over the function of the genes most present in the Biological Processes with lower *p*-value (Table 10) shows various genes that can be part of the network of processes related with carcinogenesis.

Table 11 – List of coding genes that produce the proteins responsible for the Biological Processes on Top 20 (Table 9) along with a brief functional description. Genes appear in decreasing order of presence, that is, the number of different terms on the Top 20 Biological Processes for which each gene is directly responsible.

Gene Name	Description and Function	Presence
CFLAR	CASP8 and FADD like, apoptosis regulator	12
ABCB5	ATP binding cassette subfamily B member 5, participant in transmembrane transport of structurally diverse molecules	5
TAP2	ATP binding cassette subfamily B member, translocates peptides from the cytosol to awaiting MHC class I molecules in the endoplasmic reticulum	5
AL669918.1	Paralog of TAP2, includes ATPase activity and MHC protein binding	5
ADAM9	ADAM metallopeptidase domain 9, combined with ADAM10 and ADAM17 catalyzes the alpha-secretase activity displayed by a human glioblastoma cell line toward amyloid precursor protein	3
TRIM63	Tripartite motif involved in oncogenesis, signal transduction, peroxisome biogenesis, viral infection, development, transcriptional repression and ubiquitination	3
ADAM12	ADAM metallopeptidase domain 12, participant in muscle regeneration	2
PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3, regulates the steady-state concentration of fructose-2,6-bisphosphate, a potent activator of phosphofructokinase, a key regulatory enzyme of glycolysis	1
INPP1	Inositol polyphosphate-1-phosphatase, plays a key role in intracellular signaling. Reduced activity may have a role in cardiac hypertrophy	1
PLEC	Pectin, acts as a crosslinking element of the cytoskeleton providing mechanical strength to cells and tissues	1
PIK3C2G	phosphatidylinositol-4-phosphate 3-kinase catalytic subunit type 2 gamma, regulates diverse cellular responses, such as cell proliferation, oncogenic transformation, cell migration, intracellular protein trafficking, and cell survival	1
PIP5K1C	phosphatidylinositol-4-phosphate 5-kinase type 1 gamma, catalyzes the synthesis of phosphatidylinositol 4,5-bisphosphate, essential molecule in various cellular processes. Overexpression disrupted focal adhesion plaques and caused cell rounding	1

Conclusions

Our study of the genetic diversity of São Tomé and Príncipe shows that the use of high resolution WES is an invaluable approach to provide answers to outstanding questions about the history and population structure of the archipelago, even with the use of a relatively small sample size of only 24 individuals.

In the following we summarize the answers to those questions, previously outlined in the introductory section.

1. Can non-Bantu and Bantu-speaking Niger-Congo populations from major areas of slave recruitment be distinguished genetically?

Yes. As shown in in Figures 3, 4a and 5, populations like the Esan (Nigeria) and Mende (Sierra Leone), who speak non-Bantu Niger-Congo languages, can be clearly distinguished from each other and from different Bantu-speaking populations that are widespread across Africa (Angola, Mozambique, Kenya). This differentiation shows that our 102,772 discovered genetic markers can be used to evaluate the contribution of important regions of slave recruitment to the demographic history of the Atlantic Slave trade.

2. What was the contribution of European colonizers and different African areas of slave recruitment to the genetic composition of the peoples that speak the three creole languages of São Tomé and Príncipe? Is there a link between these contributions and the linguistic characteristics of the creole languages of São Tomé?

Based on our study, and focusing just on the African contribution, we estimate (Table 9) that the relative influence of the area of the Gulf of Guinea was nearly 1.5 times higher than that of the region of Congo-Angola both in the Forros from São Tomé and the inhabitants of Príncipe (~60% Gulf of Guinea; ~40% Congo-Angola). Contributions from European colonizers are relatively low in Forros (13%) but much higher than Príncipe (3%). These partitions are very different in Angolares, who have an overrepresentation of a minor genetic component that was exclusively found in Bantu-speaking populations (mostly from Angola) and has no detectable contribution from Europeans (Figure 5; K=5). These estimates show that there is no apparent correlation between genes and languages in Forros and Príncipe inhabitants, since both have similar contributions from the Gulf of Guinea and Congo-Angola, while displaying clear differences in the proportions of Edo and Bantu words that are found in their lexica (Table 1). On the contrary, Angolares represent a remarkable case of gene-language

correlation: their genetic differentiation is in pace with their linguistic peculiarity; the strong influence of Bantu in their language is congruent with the overrepresentation of a unique Bantu component in their genes.

3. How far is the previously identified genetic distinctiveness of Angolares reproducible with the genome-wide data and how does it compare with new samples from the island of Príncipe?

Our results confirm and amplify the signal of a strong differentiation of Angolares that was previously observed with a limited set of only 15 microsatellites (Figure 2; Figure 4b; Figure 5; Figure 6) (Coelho et al., 2008). These results were neither an artifact nor a spurious observation that could be explained by ascertainment bias. Angolares are an extremely differentiated group within the archipelago. Their genetic distinctiveness is not paralleled by Príncipe.

4. What is the origin of the Angolares?

We don't know. Their high degree of genetic distinctiveness, together with limits on the resolution capacity of our polymorphisms, makes it difficult to pinpoint a region of origin. Their unique genetic component is found in Bantu-speakers but is also found in other populations from the archipelago (Figure 5). TreeMix analysis (Figure 7e) detect a migration from Angolares to Forros that may explain the sharing of this component. This could have occurred after the formation of the Angolares through a founder effect from a group of escaped slaves without primary contact with other São Tomé and Príncipe inhabitants. Alternatively, Angolares could be a subset of the Forros that experienced a profound founder effect and subsequent drift. In the moment, all that genetic data can do is to tentatively exclude a Gulf of Guinea component from the gene pool of Angolares, which ultimately may derive from the Congo-Angola area. Further details must be sought by taking into account other sources of evidence, like history and linguistics.

5. What are the functional consequences of extreme genetic differentiation?

Levels of genetic differentiation are not the same in all markers, even when they are shaped by random (non-selective) events. Like in non-coding regions, the distribution of functional relevant variants may be influenced by extreme demographic events leading to biochemical, metabolic and phenotypic differences across populations. Our study on the identification of coding SNPs that have an excess of differentiation between Forros and Angolares recognized five groups of biological processes that have functionally relevant differences between the two populations. We conclude that random demographic processes can rapidly create meaningful functional differentiation.

References

AL-SHAHROUR, F., DÍAZ-URIARTE, R. & DOPAZO, J. 2004. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics,* 20**,** 578-580.

ALBRECHTSEN, A., NIELSEN, F. C. & NIELSEN, R. 2010. Ascertainment biases in SNP chips affect measures of population divergence. *Mol Biol Evol,* 27**,** 2534-47.

ALEXANDER, D. H., NOVEMBRE, J. & LANGE, K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research,* 19**,** 1655-64.

ALONSO, R., SALAVERT, F., GARCIA-GARCIA, F., CARBONELL-CABALLERO, J., BLEDA, M., GARCIA-ALONSO, L., SANCHIS-JUAN, A., PEREZ-GIL, D., MARIN-GARCIA, P., SANCHEZ, R., CUBUK, C., HIDALGO, M. R., AMADOZ, A., HERNANSAIZ-BALLESTEROS, R. D., ALEMAN, A., TARRAGA, J., MONTANER, D., MEDINA, I. & DOPAZO, J. 2015. Babelomics 5.0: functional interpretation for new generations of genomic data. *Nucleic Acids Research,* 43**,** W117-21.

ANDREWS, S. 2010. FastQC: A quality control tool for high throughput sequence data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

BAMSHAD, M. J., NG, S. B., BIGHAM, A. W., TABOR, H. K., EMOND, M. J., NICKERSON, D. A. & SHENDURE, J. 2011. Exome sequencing as a tool for Mendelian disease gene discovery. *Nat Rev Genet,* 12**,** 745-55.

BEHR, A. A., LIU, K. Z., LIU-FANG, G., NAKKA, P. & RAMACHANDRAN, S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics,* 32**,** 2817-2823.

BENJAMINI, Y. & HOCHBERG, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological),* 57**,** 289-300.

BHATIA, G., PATTERSON, N., SANKARARAMAN, S. & PRICE, A. L. 2013. Estimating and interpreting FST: the impact of rare variants. *Genome Research,* 23**,** 1514-21.

BROWNING, B. L. & BROWNING, S. R. 2016. Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics,* 98**,** 116-26.

BROWNING, S. R. & BROWNING, B. L. 2007. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American Journal of Human Genetics,* 81**,** 1084-97.

BUSBY, G. B., BAND, G., SI LE, Q., JALLOW, M., BOUGAMA, E., MANGANO, V. D., AMENGA-ETEGO, L. N., ENIMIL, A., APINJOH, T., NDILA, C. M., MANJURANO, A., NYIRONGO, V., DOUMBA, O., ROCKETT, K. A., KWIATKOWSKI, D. P., SPENCER, C. C. & MALARIA GENOMIC EPIDEMIOLOGY, N. 2016. Admixture into and within sub-Saharan Africa. *eLife,* 5.

BUSTAMANTE, C. D., FLEDEL-ALON, A., WILLIAMSON, S., NIELSEN, R., HUBISZ, M. T., GLANOWSKI, S., TANENBAUM, D. M., WHITE, T. J., SNINSKY, J. J., HERNANDEZ, R. D., CIVELLO, D., ADAMS, M. D., CARGILL, M. & CLARK, A. G. 2005. Natural selection on protein-coding genes in the human genome. *Nature,* 437**,** 1153-7.

CALDEIRA, A. 2005. *Centro de História de Além-Mar, Enciclopédia Virtual da Expansão Portuguesa. Topónimos: Ilha do Príncipe* [Online]. http://www.fcsh.unl.pt/cham/eve//content.php?printconceito=148. [Accessed 12 May 2017].

CALDEIRA, A. M. 1999. *Mulheres, sexualidade e casamento em São Tomé e Príncipe (séculos XV-XVII),* Lisbon, Edições Cosmos.

CHILAMAKURI, C. S. R., LORENZ, S., MADOUI, M.-A., VODÁK, D., SUN, J., HOVIG, E., MYKLEBOST, O. & MEZA-ZEPEDA, L. A. 2014. Performance comparison of four exome capture systems for deep sequencing. *BMC Genomics,* 15.

CIA.GOV. 2017. *The World Factbook — Central Intelligence Agency* [Online]. https://www.cia.gov/library/publications/resources/the-world-factbook/geos/tp.html. [Accessed 17 May 2017].

COELHO, M., ALVES, C., COIA, V., LUISELLI, D., USELI, A., HAGEMEIJER, T., AMORIM, A., DESTRO-BISOL, G. & ROCHA, J. 2008. Human Microevolution and the Atlantic Slave Trade: A Case Study from São Tomé. *Current Anthropology,* 49**,** 134-143.

CUNHA MATOS, R. 1916. *Corografia Historica Das Ilhas De S. Tomé, Príncipe, Anno Bom, E Fernando Pó,* São Tomé.

CURTIN, P. D. 1969. *The Atlantic slave trade: a census,* Madison, University of Wisconsin Press.

DANECEK, P., AUTON, A., ABECASIS, G., ALBERS, C. A., BANKS, E., DEPRISTO, M. A., HANDSAKER, R. E., LUNTER, G., MARTH, G. T., SHERRY, S. T., MCVEAN, G., DURBIN, R. & GENOMES PROJECT ANALYSIS, G. 2011. The variant call format and VCFtools. *Bioinformatics,* 27**,** 2156-8.

DEPRISTO, M. A., BANKS, E., POPLIN, R., GARIMELLA, K. V., MAGUIRE, J. R., HARTL, C., PHILIPPAKIS, A. A., DEL ANGEL, G., RIVAS, M. A., HANNA, M.,

MCKENNA, A., FENNELL, T. J., KERNYTSKY, A. M., SIVACHENKO, A. Y., CIBULSKIS, K., GABRIEL, S. B., ALTSHULER, D. & DALY, M. J. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics,* 43**,** 491-498.

FERRAZ, L. I. 1979. *The Creole of São Tomé,* Johannesburg, Witwatersrand University Press.

FERREIRA, M. 2012. Looking for the genetic footprints of a vanishing language: the Lung'le in the Island of Príncipe (Gulf of Guinea). Unpublished manuscript, Faculty of Science, Porto University.

FISHER, R. A. 1935. The Logic of Inductive Inference. *Journal of the Royal Statistical Society,* 98**,** 39-82.

GAYÀ-VIDAL, M. [Population structure and admixture patterns in Southern Angola based on Expanded Exome sequences]. Unpublished raw data.

HAGEMEIJER, T. 2009. As Línguas de S. Tomé e Príncipe. *Revista de Crioulos de Base Lexical Portuguesa e Espanhola,* 1**,** 1-27.

HAGEMEIJER, T. 2011. The Gulf of Guinea Creoles: Genetic and typological relations. *Journal of Pidgin and Creole Languages,* 26**,** 111-154.

HAGEMEIJER, T. In press. From Creoles to Portuguese: Language shift in São Tomé and Príncipe. *In:* LAURA ÁLVAREZ LÓPEZ, P. G., JUANITO AVELAR (ed.) *The Portuguese language continuum in Africa and Brazil.* Amsterdam/Philadelphia: John Benjamins.

HENRIQUES, I. C. 2000. *São Tomé e Príncipe. A invenção de uma sociedade,* Lisbon, Vega editora.

HUDSON, R. R., SLATKIN, M. & MADDISON, W. P. 1992. Estimation of Levels of Gene Flow from DNA Sequence Data. *Genetics,* 132**,** 583-589.

HUSON, D. H. & BRYANT, D. 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution,* 23**,** 254-67.

JOSHI, N. A. & FASS, J. N. 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33) [Software]. Available at: https://github.com/najoshi/sickle.

KEINAN, A., MULLIKIN, J. C., PATTERSON, N. & REICH, D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *Nat Genet,* 39**,** 1251-1255.

KIDD, J. M., SHARPTON, T. J., BOBO, D., NORMAN, P. J., MARTIN, A. R., CARPENTER, M. L., SIKORA, M., GIGNOUX, C. R., NEMAT-GORGANI, N., ADAMS, A., GUADALUPE, M., GUO, X., FENG, Q., LI, Y., LIU, X., PARHAM, P., HOAL, E. G., FELDMAN, M. W., POLLARD, K. S., WALL, J. D., BUSTAMANTE,

C. D. & HENN, B. M. 2014. Exome capture from saliva produces high quality genomic and metagenomic data. *BMC Genomics,* 15**,** 262.

KIM, H. L., RATAN, A., PERRY, G. H., MONTENEGRO, A., MILLER, W. & SCHUSTER, S. C. 2014. Khoisan hunter-gatherers have been the largest population throughout most of modern-human demographic history. *Nature Communications,* 5**,** 5692.

KINSELLA, R. J., KAHARI, A., HAIDER, S., ZAMORA, J., PROCTOR, G., SPUDICH, G., ALMEIDA-KING, J., STAINES, D., DERWENT, P., KERHORNOU, A., KERSEY, P. & FLICEK, P. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database (Oxford),* 2011**,** bar030.

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

LACHANCE, J. & TISHKOFF, S. A. 2013. SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. *Bioessays,* 35**,** 780-6.

LI, H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics,* 27**,** 2987-93.

LI, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. [arXiv:1303.3997v2] Available at: http://arxiv.org/abs/1303.3997

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & SUBGROUP, G. P. D. P. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LORENZINO, G. 1998. *Angolar Creole Portuguese,* Newcastle, Lincom Europa.

MACHADO, M. 2013. The Genetic Structure of the Island of Príncipe and its implications for the Origins and Development of Human Creole Societies in the Gulf of Guinea. Unpublished manuscript, Faculty of Science, Porto University.

MANICHAIKUL, A., MYCHALECKYJ, J. C., RICH, S. S., DALY, K., SALE, M. & CHEN, W. M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics,* 26**,** 2867-73.

MATEU, E., COMAS, D., CALAFELL, F., PÉREZ-LEZAUN, A., ABADE, A. & BERTRANPETIT, J. 1997. A tale of two islands: population history and mitochondrial DNA sequence variation of Bioko and São Tomé, Gulf of Guinea. *Annals of Human Genetics,* 61**,** 507-518.

MAURER, P. 2009. *Principense: Grammar, texts, and vocabulary of the Afro-Portuguese creole of the island of Príncipe, Gulf of Guinea,* London, Battlebridge.

MCKENNA, A., HANNA, M., BANKS, E., SIVACHENKO, A., CIBULSKIS, K., KERNYTSKY, A., GARIMELLA, K., ALTSHULER, D., GABRIEL, S., DALY, M. & DEPRISTO, M. A. 2010. The Genome Analysis Toolkit: a MapReduce framework

for analyzing next-generation DNA sequencing data. *Genome Research,* 20**,** 1297-1303.

MIERS, S. & ROBERTS, R. 1988. *The End of slavery in Africa,* Madison, Wisconsin, University of Wisconsin Press.

NASCIMENTO, A. 2003. Os São-Tomenses e as Mutações Sociais na sua História Recente. *AFRICANA STUDIA, Faculdade de Letras da Universidade do Porto Edition* 6**,** 7-44.

Online Mendelian Inheritance in Man, OMIM®. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2017. Available at: https://omim.org/

PATTERSON, N., PRICE, A. L. & REICH, D. 2006. Population structure and eigenanalysis. *PLoS Genet,* 2**,** e190.

PICKRELL, J. K. & PRITCHARD, J. K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. *PLoS Genetics,* 8**,** e1002967.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. & SHAM, P. C. 2007. PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics,* 81**,** 559-575.

R CORE TEAM 2016. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: https://www.R-project.org/.

REICH, D., THANGARAJ, K., PATTERSON, N., PRICE, A. L. & SINGH, L. 2009. Reconstructing Indian population history. *Nature,* 461**,** 489-494.

ROCHA, J. & HAGEMEIJER, T. 2012. *RE: Línguas e genes na ilha de S. Tomé. In Colóquio S. Tomé e Príncipe numa perspectiva interdisciplinar.*

SAITOU, N. M. & NEI, M. 1987. The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees. *Molecular Biology and Evolution,* 4**,** 406-25.

SCHLICKER, A., DOMINGUES, F. S., RAHNENFUHRER, J. & LENGAUER, T. 2006. A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics,* 7**,** 302.

SEIBERT, G. 1998. *A questão da origem dos angolares de São Tomé,* Lisbon, Brief Papers nº5/98, CEsA

SUPEK, F., BOSNJAK, M., SKUNCA, N. & SMUC, T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS One,* 6**,** e21800.

TEER, J. K. & MULLIKIN, J. C. 2010. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet,* 19**,** R145-R151.

TENNESSEN, J. A., MADEOY, J. & AKEY, J. M. 2010. Signatures of positive selection apparent in a small sample of human exomes. *Genome Res,* 20**,** 1327-34.

TENREIRO, F. 1961. *A ilha de São Tomé,* Lisbon, Junta de Investigações Científicas do Ultramar.

THE 1000 GENOMES PROJECT CONSORTIUM 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature,* 491**,** 56-65.

THE 1000 GENOMES PROJECT CONSORTIUM, AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A. & ABECASIS, G. R. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

THOMAS, H. 1997. *The slave trade: the story of the Atlantic slave trade, 1440-1870*, Simon & Schuster.

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. *Human Biology,* 74**,** 397-411.

TROVOADA, M. J., ALVES, C., GUSMÃO, L., ABADE, A., AMORIM, A. & PRATA, M. J. 2001. Evidence for population sub-structuring in São Tomé e Príncipe as inferred from Y-chromosome STR analysis. *Annals of Human Genetics,* 65**,** 271-283.

TROVOADA, M. J., PEREIRA, L., GUSMÃO, L., ABADE, A., AMORIM, A. & PRATA, M. J. 2004. Pattern of mtDNA variation in three populations from São Tomé e Príncipe. *Annals of Human Genetics,* 68**,** 40-54.

TROVOADA, M. J., TAVARES, L., GUSMAO, L., ALVES, C., ABADE, A., AMORIM, A. & PRATA, M. J. 2007. Dissecting the genetic history of Sao Tome e Principe: a new window from Y-chromosome biallelic markers. *Annals of Human Genetics,* 71**,** 77-85.

VAN DER AUWERA, G. A., CARNEIRO, M. O., HARTL, C., POPLIN, R., DEL ANGEL, G., LEVY-MOONSHINE, A., JORDAN, T., SHAKIR, K., ROAZEN, D., THIBAULT, J., BANKS, E., GARIMELLA, K. V., ALTSHULER, D., GABRIEL, S. & DEPRISTO, M. A. 2013. From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. *Current Protocols in Bioinformatics,* 43**,** 11 10 1-33.

WARR, A., ROBERT, C., HUME, D., ARCHIBALD, A., DEEB, N. & WATSON, M. 2015. Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda),* 5**,** 1543-50.

WEIR, B. S. & COCKERHAM, C. C. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution,* 38**,** 1358-1370.

WICKHAM, H. 2009. *ggplot2: Elegant Graphics for Data Analysis*, Springer-Verlag New York.

YANG, J., BENYAMIN, B., MCEVOY, B. P., GORDON, S., HENDERS, A. K., NYHOLT, D. R., MADDEN, P. A., HEATH, A. C., MARTIN, N. G., MONTGOMERY, G. W., GODDARD, M. E. & VISSCHER, P. M. 2010. Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics,* 42**,** 565-9.

YANG, X., AL-BUSTAN, S., FENG, Q., GUO, W., MA, Z., MARAFIE, M., JACOB, S., AL-MULLA, F. & XU, S. 2014. The influence of admixture and consanguinity on population genetic diversity in Middle East. *Journal of Human Genetics,* 59**,** 615-622.

YI, X., LIANG, Y., HUERTA-SANCHEZ, E., JIN, X., CUO, Z. X. P., POOL, J. E., XU, X., JIANG, H., VINCKENBOSCH, N., KORNELIUSSEN, T. S., ZHENG, H., LIU, T., HE, W., LI, K., LUO, R., NIE, X., WU, H., ZHAO, M., CAO, H., ZOU, J., SHAN, Y., LI, S., YANG, Q., ASAN, NI, P., TIAN, G., XU, J., LIU, X., JIANG, T., WU, R., ZHOU, G., TANG, M., QIN, J., WANG, T., FENG, S., LI, G., HUASANG, LUOSANG, J., WANG, W., CHEN, F., WANG, Y., ZHENG, X., LI, Z., BIANBA, Z., YANG, G., WANG, X., TANG, S., GAO, G., CHEN, Y., LUO, Z., GUSANG, L., CAO, Z., ZHANG, Q., OUYANG, W., REN, X., LIANG, H., ZHENG, H., HUANG, Y., LI, J., BOLUND, L., KRISTIANSEN, K., LI, Y., ZHANG, Y., ZHANG, X., LI, R., LI, S., YANG, H., NIELSEN, R., WANG, J. & WANG, J. 2010. Sequencing of 50 Human Exomes Reveals Adaptation to High Altitude. *Science,* 329**,** 75-78.

Appendix

"TreeMap" view of REVIGO for the 235 biological processes identified as differentiated between Angolares and Forros. Each rectangle is a single cluster representative as it is in the *Homo Sapiens* database. The representatives are joined into "superclusters" of related terms, visualized with different colors using the settings of tiny similarity and model of semantic similarity *SimRel*. The size of the rectangles reflects the *p*-value of the GO term as calculated with FatiGO.