

PhD

3rd
CYCLE

FCUP
FCUL
2025

U.PORTO

Genetic admixture histories on the island of São
Tomé during and after the Trans-Atlantic Slave Trade

Marta Ciccarella

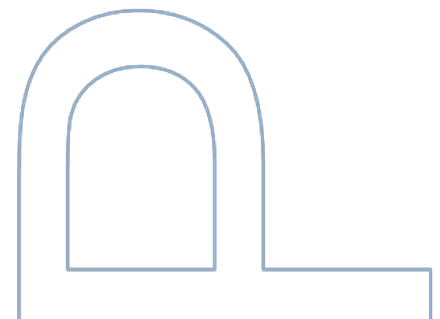
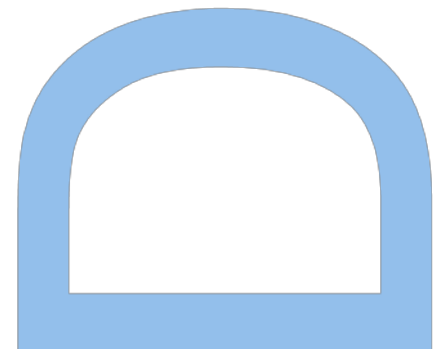
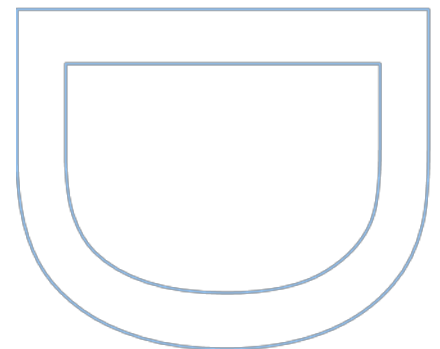
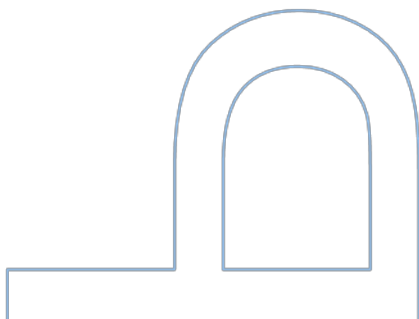
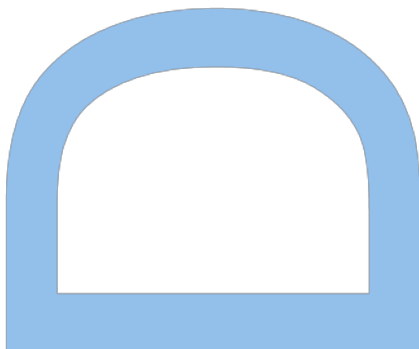
FC



Genetic admixture histories on the island of São Tomé during and after the Trans- Atlantic Slave Trade

Marta Ciccarella

Biodiversity, Genetics and Evolution
Faculty of Sciences of the University of Porto and
Faculty of Sciences of the University of Lisbon
2025



Genetic admixture histories on the island of São Tomé during and after the Trans-Atlantic Slave Trade

Marta Ciccarella

Biodiversity, Genetics and Evolution
Department of Biology
2025

Supervisor

Jorge Rocha, Associate Professor, Faculty of Sciences University of Porto

Co-supervisor

Paul Verdu, Director of Research, UMR7206 EA, CNRS-MNHN-UPC

Dedicated to my sisters

Acknowledgements

The work presented in this dissertation would have not been possible without the support of several people and institutions. I am deeply grateful to all of them. While I cannot name everyone, the following is a brief and non-exhaustive list of those to whom I owe special thanks.

Firstly, I would like to express my sincere gratitude to my supervisors, Jorge Rocha and Paul Verdu, for their constant availability, patience and guidance throughout this journey. Jorge, thank you for inspiring me with your outstanding scientific curiosity. Paul, thank you for teaching me the highest standards of scientific rigour and professional ethics. I am deeply grateful to you both for the opportunity to discuss science and everything that goes with it.

I would also like to take this opportunity to express my sincere gratitude to all the São Toméan and Cabo Verdean participants in this study.

I would like to thank the institutions where I conducted my research: the Research Centre for Biodiversity and Genetic Resources (CIBIO-InBio) in Portugal and the National Museum of Natural History (MNHN) at the Musée de l'Homme in France.

Thanks to the Foundation for Science and Technology (FCT) and the Faculty of Sciences of the University of Porto for their financial support in the form of a PhD grant and a research grant, respectively. I would also like to thank the UMR 7206 Eco-Anthropology (EA), the MNHN and the Centre National de la Recherche Scientifique (CNRS) for the excellent working conditions and material resources they provided me with at the Musée de l'Homme, and for funding all my research trips. Thanks to the Agence Nationale de la Recherche (ANR) for funding part of this study and providing material resources.

Special thanks go to Taoues Lahrem, the administrative manager of the UMR 7206 EA, and Dora Bastos, the postgraduate manager at CIBIO-InBio, as well as all the administrative staff, for their continuous support. Without them, research would simply not be possible.

Thanks to my colleagues from the Human Evolutionary Genetics group at CIBIO: to Bérénice Alard, who welcomed me to CIBIO even before my PhD began; to Anne-Maria Fehn, who gave me a warm welcome in Portugal and even took me to the health centre; to Magdalena Gayà-Vidal, for our chats about science and life over coffee at Teresa's bar; to Beatriz Amorim, for her kindness and lovely company in the office; and to Armando Semo, for always being available and having a big smile.

Thanks to my PhD colleagues of the BIODIV Doctoral Programme who welcomed me warmly to CIBIO and made me feel at home in Portugal since the beginning. Special thanks go to Ninda Baptista, for the great times spent in libraries, demonstrations and clubs; to Adam Marques and Costanza Piccoli, for all the listening at work and the excursions outside.

Thanks to my colleagues from the Anthropological Genetics group (Agene), from whom I learnt a lot about science and how it works behind the scenes. Special thanks go to Bruno Toupance for his trust and support from the very beginning; to Romain Laurent, for his patience and dedication in guiding students like me, who began the PhD with limited coding experience. Thanks also to Céline Bon for her enthusiasm for research, which motivate me to continue; to Nina Marchi for her positive attitude and advice on what to do next; to Frédéric Austerlitz for the inclusive gatherings on the terrace since my first days at the Musée de l'Homme; and to Evelyne Heyer, Raphaëlle Chaix and Marie-Claude Marsolier-Kergoat.

I would also like to thank the researchers of the UMR 7206 EA, and particularly Priscille Touraille, who first welcomed me to the Musée de l'Homme and introduced me to Paul. I would also like to express my deepest gratitude to Frank Alvarez-Pereyre for guiding me through the complexities of interdisciplinary research since the early stages of my PhD.

I would like to sincerely thank all my PhD colleagues at the Musée de l'Homme, where I learnt that doing a PhD can be a joyful experience. Special thanks go to Margaux Bieuville for listening to me, for always having the good advice, and for convincing me that boxing and sport in general are worth it. Thanks also to Johanne Adam, Léa Guyon and Juliette Sauvage for supporting each other throughout our PhDs, for making coming to work a joy every day, and for sharing congresses and holidays together.

Finally, I wouldn't have made it through without the love and support of my friends and family. Special thanks go to my friend Lorenzo, who convinced me that I could do anything I wanted and then helped me to do it. Thanks to my beloved partner, Spyros, who has always stood by my side and brought me joy. Thanks to Erica and Geri, my sisters and my soul mates, to whom this work is dedicated. And thanks to my parents, Mariuccia and Giuseppe, my greatest supporters. I would be nothing without your love.

Resumo

A deslocação forçada de milhões de africanos durante o tráfico transatlântico de escravos pôs em contacto populações anteriormente isoladas, dando origem a padrões complexos de miscigenação genética. A antropogenética procura reconstruir os prazos e a extensão destes episódios de miscigenação a partir dos padrões de diversidade genética observados nas populações atuais, descendentes de africanos escravizados. Também procura examinar a forma como os diferentes contextos históricos e sociais influenciaram quem se reproduziu com quem, e deixou descendentes nas sociedades coloniais de ambos os lados do Atlântico.

Este estudo centra-se em duas dessas sociedades: uma na ilha em São Tomé, no Golfo da Guiné, e outra em Cabo Verde, um arquipélago situado ao largo da costa do Senegal. Ambos os territórios eram desabitados antes da colonização portuguesa no final do século XV e foram povoados na sequência da migração forçada de africanos escravizados. Embora as duas sociedades tenham sido colónias sob o domínio português, São Tomé e Cabo Verde experimentaram diferentes formas de gestão colonial e organização social, que moldaram as respetivas histórias genéticas. Além disso, as migrações de mão de obra de Cabo Verde para São Tomé no século XX, após a abolição da escravatura, puseram em contacto as duas populações.

O trabalho procura reconstruir o impacto genético das migrações históricas para São Tomé, através de uma análise de dados recém-gerados sobre o genoma de 96 indivíduos nascidos na ilha, juntamente com amostras representativas do arquipélago de Cabo Verde e de outras populações africanas, europeias e americanas. Os nossos resultados são consistentes com os registos históricos e revelam como cinco séculos de estruturas sociais em mutação, relacionadas com a escravatura, a resistência ao trabalho forçado e o trabalho contratado, deixaram marcas mensuráveis na diversidade genética dos atuais são-tomenses. Em última análise, o estudo revela a importância de interpretar as descobertas genéticas no contexto histórico e social único de cada uma das populações formadas no contexto do tráfico de pessoas escravizadas originárias do continente africano.

Palavras-chave: mistura genética; comércio transatlântico de escravos; ancestralidade genética.

Abstract

The forced displacement of millions of Africans during the Trans-Atlantic Slave Trade brought previously isolated populations into contact, resulting in complex patterns of genetic admixture. Anthropological geneticists seek to reconstruct the timing and extent of these admixture events from patterns of genetic diversity observed in present-day populations descended from enslaved Africans. They also examine how different historical and social contexts influenced who mated with whom and left descendants in colonial societies on both sides of the Atlantic.

This study focuses on two such populations: those currently living in São Tomé, an island in the Gulf of Guinea, and Cabo Verde, an archipelago off the coast of Senegal. Both locations were uninhabited before Portuguese colonization in the late 15th century and were settled through the forced migration of enslaved Africans. Although both were early peopling colonies under Portuguese rule, São Tomé and Cabo Verde experienced different forms of colonial exploitation and social organization, which shaped their genetic histories. Additionally, labour-driven migrations from Cabo Verde to São Tomé in the 20th century, after the abolition of slavery, brought into contact these two distinct populations descended from enslaved Africans.

By analysing newly-generated genome-wide data from 96 individuals born in São Tomé, alongside extensive samples from Cabo Verde and other African, European and American populations, I reconstructed the genetic impact of historical migrations to the island. Our genetic findings align with historical records and reveal how five centuries of shifting social structures, related to slavery, marronage, and indentured labour, have left measurable traces in the genetic diversity of present-day São Toméans. This study ultimately demonstrates the importance of interpreting genetic findings within the evolving historical and social contexts unique to each population descended from enslaved Africans.

Keywords: Genetic admixture; Trans-Atlantic Slave Trade; Genetic ancestry.

Table of Contents

List of Figures	xi
List of Abbreviations	xii
INTRODUCTION	1
Aim of the Research	9
Content of the Thesis	11
CHAPTER 1	15
1 Genetic and historical perspectives on the Trans-Atlantic Slave Trade	17
1.1 The Trans-Atlantic Slave Trade from an historical perspective	19
1.2 What is the genetic impact of the TAST on both sides of the Atlantic?	22
1.3 Where in Africa did the genetic ancestors of African-descendant populations come from?	24
1.4 How much female and male genetic ancestry came from Africa?	26
1.5 Who mated with whom in admixed populations?	29
1.6 When did admixture occur?	31
1.7 Final remarks and future directions	33
CHAPTER 2	35
2 Nested admixture during and after the Trans-Atlantic Slave Trade on the island of São Tomé	37
2.1 Introduction	39
2.1.1 São Tomé and the emergence of the slave trade and slave societies	41
2.1.2 Cacao plantations in São Tomé: from slavery to indentured labour	42
2.1.3 Social stratification and creole languages in São Tomé	43
2.1.4 Previous genetic studies	43
2.1.5 The aim of this study	44

2.2	Results	45
2.2.1	São Tomé genetic diversity in the global context.....	45
2.2.2	Haplotype-based population structure within São Tomé	52
2.2.3	Haplotypic ancestry of genetic clusters.....	54
2.2.4	Long IBD-tracts shared among São Toméans and Cabo Verdeans.....	57
2.2.5	Runs of Homozygosity	57
2.2.6	Dating admixture events for each São Toméan cluster separately	59
2.3	Discussion.....	61
2.3.1	A population descended from freed slaves.....	63
2.3.2	One of the first communities of runaway slaves	64
2.3.3	Cabo Verdean migration to São Tomé	65
2.3.4	Limited European genetic contribution.....	66
2.3.5	Nested admixture between recently admixed populations.....	67
2.3.6	The genetic contributions from Mozambique and Angola.....	68
2.3.7	Final methodological considerations.....	69
2.4	Conclusions and perspectives	71
CHAPTER 3		73
3	Sex-specific demographic processes shaping genetic diversity on the X chromosome and the autosomes in São Tomé and Cabo Verde.....	75
3.1	Introduction	77
3.1.1	Ancestry-related mating patterns	78
3.1.2	Sex-specific demography.....	79
3.1.3	Enslaved and non-enslaved communities	80
3.1.4	IBD sharing patterns on the X and autosomal chromosomes.....	81
3.1.5	Gender-specific social dynamics in São Tomé and Cabo Verde.....	82
3.1.6	The genetic diversity of putative source populations	83
3.1.7	Insights from uniparental lineages in São Tomé.....	85
3.1.8	Aim of this study.....	85

3.2	Results	87
3.2.1	Genetic differentiation of individual samples based on autosomes and X chromosomes	87
3.2.2	ADMIXTURE Analysis	93
3.2.3	IBD-sharing	96
3.2.4	Runs of Homozygosity	98
3.3	Discussion.....	102
3.3.1	Methodological discussion	102
3.3.2	São Tomé and Cabo Verde	105
3.3.3	Founder effects in the archipelago of Cabo Verde	107
3.3.4	Proposing complex models of genetic admixture	109
3.4	Conclusions and Perspectives.....	115
CONCLUSIONS AND PERSPECTIVES.....		117
	The genetic impact of historical migrations in São Tomé	119
	The evolution of genetic structures	122
	Mating patterns related to genetic ancestry	124
	Complex models of genetic admixture	125
	Nested admixture	126
	Genetic ancestry labels.....	127
	Communicating results to a wider audience	128
	Final considerations on São Tomé and Cabo Verde	130
REFERENCES		133
ATTACHMENTS		149
SUPPLEMENTARY INFORMATION ON GENETIC DATASETS		151
SUPPLEMENTARY INFORMATION FOR CHAPTER 2		155
	Supplementary Materials and Methods	157
	Supplementary Figures.....	165

Supplementary Tables	176
SUPPLEMENTARY INFORMATION FOR CHAPTER 3	189
Supplementary Materials and Methods	191
Supplementary Figures.....	197
Supplementary Tables	200

List of Figures

Figure 0.1. The Cantino Planisphere	6
Figure 1.1. Volume and direction of the Trans-Atlantic Slave Trade	19
Figure 2.1 Major historical migrations to the island of São Tomé.....	40
Figure 2.2. Map of São Tomé showing the position of 14 sampling sites.....	45
Figure 2.3. The genetic differentiation of the worldwide population samples	46
Figure 2.4. Unsupervised ADMIXTURE analysis	47
Figure 2.5 Genetic clustering across São Tomé.....	50
Figure 2.6 Genetic clustering in the worldwide context	51
Figure 2.7. Genetic clustering and population structure in São Tomé.....	53
Figure 2.8. Haplotypic ancestry of São Toméans and other populations	56
Figure 2.9. IBD tracts sharing and ROH analysis	58
Figure 2.10. Admixing sources and timing of admixture events	60
Figure 2.11. Nested admixture in São Tomé.	62
Figure 3.1. ASD-MDS analysis of worldwide population samples	88
Figure 3.2. ASD-MDS analysis of a subset of population samples	90
Figure 3.3. ASD-MDS analysis using two autosomal chromosomes.....	92
Figure 3.4. ADMIXTURE analysis on the chromosome 20 and the X chromosome ...	95
Figure 3.5. IBD sharing on the autosomal chromosome 7 and the X chromosome. ...	97
Figure 3.6. Distribution of ROH Length by Chromosome Size	99
Figure 3.7. ROH by length classes on the X chromosome and chromosome 7	101
Figure 3.8. Admixture scenarios in São Tomé including gene flow from Europe	110
Figure 3.9. Demographic scenarios for the <i>Angolares</i> in São Tomé	112
Figure 3.10. Simplified admixture model of Cabo Verdean migration to São Tomé..	114

List of Abbreviations

FCUP	Faculty of Sciences of the University of Porto
UP	University of Porto
UMR7206 EA	UMR7206 Eco-Anthropologie
MNHN	Museum National d'Histoire Naturelle
CNRS	Centre National de la Recherche Scientifique
UPC	Université Paris Cite

ABC	Approximate Bayesian Computation
ASD	Allele Sharing Dissimilarities
EM	Expectation-Maximization
IBD	Identical-By-Descent
LD	Linkage Disequilibrium
MDS	Multi-Dimensional Scaling
mtDNA	Mitochondrial DNA
Ne	Effective Population Size
NRY	Non-Recombining Y Chromosome
PCA	Principal Component Analysis
ROH	Runs Of Homozygosity
SNP	Single Nucleotide Polymorphism

TAST	Trans-Atlantic Slave Trade
------	----------------------------

INTRODUCTION

The Trans-Atlantic Slave Trade (TAST) was one of the largest forced migrations in recent human history, leaving a profound impact on the genetic diversity of populations on both sides of the Atlantic. Historians have provided invaluable insights into the timing and scale of this migration (Curtin 1990; Berlin 1998; Eltis and Richardson 2010; Lovejoy 2012). However, historical accounts and records are often inaccurate, missing, or falsified (Geggus 1989a; Eltis and Richardson 2008; Morgan 2021). Since the mid-20th century, population geneticists have investigated new data and developed new tools to address specific questions of genetic admixture and population demography that can complement historical reconstruction (Fortes-Lima and Verdu 2020).

The migrations triggered by the TAST and the social contexts in which individuals and communities interacted have likely shaped the distribution of genetic variation in populations descended from enslaved Africans. However, we must be cautious about claiming that the history of the TAST is recorded in the genomes of these populations. There is no direct causal link between slavery and genetic markers: slavery is not 'written' in the genes. The observed patterns of genetic diversity within the genomes of contemporary populations can be regarded as a biological archive, documenting who mated with whom in the past and left descendants, which is then analysed through definitions and tools specific to population genetics.

The successive waves of forced and deliberate migration from Africa and Europe during the TAST brought previously isolated populations into contact. This allowed individuals from different source populations to mate. Such process, known as genetic admixture, has profoundly shaped the genetic diversity of populations descended from enslaved Africans on both sides of the Atlantic (Beleza et al. 2012; Moreno-Estrada et al. 2013; Bryc et al. 2015; Baharian et al. 2016; Ongaro et al. 2019; Micheletti et al. 2020; Almeida et al. 2021; Laurent et al. 2023). Even after the formal end of the TAST and the abolition of slavery, the demography and genetic diversity of these populations continued to evolve under the influence of new waves of migration, driven by economic opportunity or other labour systems such as indentured servitude (Carreira 1976; Galenson 1984; Nascimento 2008; Allen 2014; Baharian et al. 2016; Kiesow 2017; Laurent et al. 2023).

In addition, hierarchies of race, class, and gender strongly influenced mating patterns during admixture processes (Wade 2009; Mas-Sandoval et al. 2023). Previous genetic studies have suggested that in several populations descended from enslaved Africans,

female and male contributions from source populations were not equivalent, and that mating might have occurred preferentially between individuals with certain genetic ancestries. These processes, known as ancestry-related sex bias and assortative mating, have been shown to have profoundly affected the genome-wide genetic diversity of at least certain populations descended from enslaved Africans on both sides of the Atlantic (Goldberg et al. 2014; Zaitlen et al. 2017; Goldberg et al. 2020; Micheletti et al. 2020; Kim et al. 2021; Ongaro et al. 2021; Korunes et al. 2022; Mas-Sandoval et al. 2023).

These processes, resulting in mating patterns that are opposite to one another, may have alternated between the two over time. Discrimination based on social stratification, or even segregation laws, discouraged or prohibited unions across racial and social boundaries (Seibert 2015; Valente and Berry 2020; Ruef 2022). Conversely, in certain contexts, ideologies of “*branqueamento*”, literally meaning “whitening”, promoted unions in line with the eugenic aim of diminishing African ancestry in the overall population (Caldeira 2007). To thoroughly understand the signature that these processes left in the genomes of admixed populations, in form of patterns of genetic diversity, we thus require a careful interpretation of genetic findings within the evolving historical and social contexts unique to each population during and after the TAST.

Studying the legacy of the TAST from a genetic perspective means engaging with various disciplines, and in particular with History. A productive dialogue between historians and geneticists requires a clear understanding of the objectives and limits of each discipline. One of the main differences between historical and genetic investigation is that genetic studies based on modern genomes focus on the historical and social events that have shaped the current distribution of genetic diversity. For this reason, such studies are not able to address events such as the migration of individuals who left few descendants in the current population, or migrations followed by major demographic shifts or high mortality rates that radically altered the population’s genetic diversity. However, recent advances in the analysis of ancient genomes, namely genetic material retrieved from individuals who lived in the past, offer new opportunities to explore these particular questions (Fortes-Lima and Verdu 2020).

Many of the questions posed by historians may not align with those of geneticists, nor can they always be answered using genetic data. Acknowledging which questions can actually be addressed by genetic research is the first step in advancing this interdisciplinary dialogue. As a geneticist, studying the TAST's historical record and historiography can be challenging, yet this information is extremely important for both study design and the interpretation of genetic results. A critical understanding of the primary historical sources, the state-of-the-art work of historians, and the evolving narratives associated with the TAST, is essential to provide a context for anthropological genetics studies.

Anthropological geneticists and historians are eventually interested in similar questions and sometimes face the same limitations, especially the difficulty of interpreting ethnic, political, and geographic categories used in the past with current classifications (Lovejoy 2010; Fortes-Lima and Verdu 2020; Jackson 2021). Studies that focus on large inter-continental datasets, including genetic and/or historical data, provide valuable insights into large-scale patterns associated with the TAST. However, they often risk overlooking the complexity of historical migrations and genetic admixture histories, which can lead to biased or over-simplified general interpretations. On the other hand, at more local scales, researchers can better tailor data, categories, and methods to specific research questions, allowing for a deeper understanding of the unique histories of African-descended populations. In turn, these detailed endeavour at a local scale may obscure the diversity of TAST histories and their equally diverse consequences on extant genomic diversity. Generalization in these cases also represents a very delicate challenge.

During my PhD, I focused on populations descended from enslaved Africans that are currently living in São Tomé, the largest island of the archipelago of São Tomé e Príncipe in the Gulf of Guinea, and Cabo Verde, another archipelago off the coasts of Senegal composed of nine main inhabited islands today. These African Atlantic islands played a crucial role in the establishment of the slave trade with the African coasts, even before it reached the other side of the Atlantic, and, in the case of São Tomé, in the emergence of an economic system based on the exploitation of enslaved labour, the plantation economy, which later became central to the global commercial and political landscape of all European colonial empires in the Americas.

São Tomé e Príncipe and Cabo Verde represent among the earliest instances of interaction between Europeans and Africans in stratified colonial societies (Seibert 2014). The two archipelagos share many similarities in their historical trajectories: they both were uninhabited prior to Portuguese settlement in the late 15th century; they both became the first European peopling colonies in sub-Saharan Africa and played central roles in the Trans-Atlantic Slave Trade; and they both were governed as slave societies under Portuguese colonial rule until independence from Portugal in 1975. Despite these parallels, the demographic and colonial histories of São Tomé and Cabo Verde have diverged significantly over time. First, they were involved in different slave trade routes, with Cabo Verde trading primarily with the nearby Senegambia region and São Tomé with the Gulf of Guinea and the Congo-Angola region (Caldeira 2013). Secondly, Cabo Verde never developed large-scale plantations, while São Tomé's history and society were heavily influenced by plantation economies. Finally, after the abolition of slavery, many Cabo Verdeans went to São Tomé as indentured workers in plantations (Nascimento 2003).



Figure 0.1. The Cantino Planisphere.

This 1502 map, attributed to an anonymous Portuguese cartographer, already depicts São Tomé and Cabo Verde as emerging commercial hubs in the African Atlantic. Biblioteca Estense Universitaria, Modena, Italy.

The island of São Tomé stands out as one of the earliest examples of slave societies in the colonial era, with its 16th-century sugar plantations heavily dependent on enslaved labour (Curtin 1990; Berlin 1998; Almeida Mendes 2008; Caldeira 2011). In contrast to a “society with slaves”, a “slave society” is characterized not only by a high proportion of slaves in the population, but also by the centrality of slavery to its politics, economy, and social organization (Finley 1980). However, slavery was not a static institution, but a dynamic system that adapted to the fluctuations of plantation economies (Berlin 1998). Although São Tomé’s plantation economy began to decline in the 17th century, the institution of slavery persisted, with enslaved people still comprising 40% of the island’s population at the beginning of the 19th century (Lucas 2015). In the mid 19th century, the emergence of coffee and cacao plantations re-established forced labour at the centre of São Tomé’s economy, with indentured servitude emerging after the abolition of slavery in 1875 (Kiesow 2017).

The history of São Tomé includes some of the most relevant aspects of the colonial era: the island was a major hub for the slave trade with African, European, and American ports; it hosted one of the first plantation-based economies dependent on slavery; it saw the emergence of one of the first communities of runaway slaves, whose descendants would later be known as the *Angolares* (Seibert 1998); and it transitioned to indentured servitude after the abolition of slavery. São Tomé’s population also experienced unique historical and demographic events, as it maintained one of the highest proportions of Africans throughout colonization, and went through a period of economic decline when the Portuguese largely abandoned the island, leaving the *Forros*, literally meaning freed slaves, as the major social group for over two centuries (Hagemeijer 2011).

In São Tomé and Cabo Verde, linguistic studies have revealed a remarkable diversity of creole languages (Baptista 2015; Cardoso et al. 2015). Creoles are defined as new language varieties that emerged from the contact between different European and non-European languages (Mufwene 2015). In the case of São Tomé, linguists hypothesise that a proto-Gulf of Guinea Creole language is the ancestor of the various creoles spoken in the region (Hagemeijer 2011). Among these, the Forro is the second most spoken language after Portuguese in São Tomé e Príncipe, and its prevalence declined significantly between 2001 and 2012, when two major censuses were conducted. Notably, the most recent census also included data on the Angolar creole, primarily

spoken by the *Angolares*, and Cabo Verdean Kriolu, spoken mainly in rural areas by descendants of indentured workers (Instituto Nacional de Estatística 2012).

From a genetic perspective, this small-scale island setting provides unique opportunities to investigate complex histories of admixture in the colonial context. The availability of historical, linguistic and demographic data can facilitate the interpretation of genetic findings and provide insight into the relationship between cultural and genetic diversity. In line with previous studies on São Tomé and Cabo Verde (Almeida et al. 2021; Laurent et al. 2023), it would be of great interest to investigate whether migration patterns reconstructed from genetic data reflect the African and European components of present-day creoles, and to what extent the observed genetic structure aligns with linguistic and anthropological categories.

The joint analysis of genetic samples from São Tomé and Cabo Verde in this study presents at least two unique opportunities. First, we aim to assess the impact of Cabo Verdean migration to São Tomé after the abolition of slavery, which represented a major yet very recent (on evolutionary time scales), demographic shift on the island. To our knowledge, this is the first time that admixture between two distinct populations descended from enslaved Africans is explicitly explored, and in particular in Africa. Second, studying the genetic diversity of these two populations allows for a direct comparison of the impacts of their shared, yet distinct, colonization experiences on genetic diversity. The different origins of the enslaved Africans who were forcibly brought to the islands, and the different slave societies that developed in São Tomé and Cabo Verde, have likely influenced admixture and demographic histories, which in turn may have left different signatures on the patterns of genetic diversity of these populations.

Aim of the Research

The main aim of my PhD is to reconstruct the complex processes of genetic admixture that have shaped current genetic diversity in the island of São Tomé. This research is based on a new genotyping dataset containing an unprecedented number of genome-wide markers and samples from the island, combined with an extensive reference dataset of 70 population samples from Africa, Europe and the Americas. This study aims to explore how historical migrations during and after the TAST, along with shifting social structures in the colonial context, influenced admixture patterns at a micro-scale in Africa. Specifically, this research addresses the following questions:

1. Where did the genetic ancestors of São Tomé's population come from?
 - a. Do African and European genetic contributions reflect the major slave trade routes and further migrations after the abolition of slavery?
 - b. What is the relative impact of slavery and post-slavery migrations on the genetic diversity of the São Toméan population?
2. Is there a genetic structure within São Tomé's population?
 - a. Is this structure explained by geography or by differences in genetic ancestry?
 - b. Does it reflect linguistic-anthropological communities, such as the *Angolares*, *Forros*, and Cabo Verdeans?
3. When did the admixture occur in São Tomé?
 - a. How did historical events, such as the rise and fall of plantation economies, influence the timing and process of genetic admixture on the island?
 - b. Did demographic dynamics associated with changing historical and social contexts shape genetic diversity and admixture patterns over time?
4. How many females and males from Africa and Europe contributed to the current genetic diversity in São Tomé?
 - a. To what extent sex-biased migrations from Africa and Europe translate into ancestry-related sex bias in the admixture process?
 - b. Are there any differences in ancestry-related sex bias across different linguistic-anthropological communities?
5. Who admixed with whom in the admixed population?

- a. Did race, class, and gender hierarchies influence the mating patterns and reproductive success of particular individuals or groups?
 - b. Did individuals mate primarily within linguistic-anthropological communities?
 - c. To what extent individuals mate between different linguistic-anthropological communities?
6. What is the impact of sex-specific demography and mating patterns on São Tomé's genetic diversity?
 - a. Have different mating behaviours and family structures in enslaved and non-enslaved communities influenced patterns of genetic diversity in São Tomé?
 - b. Did demographic dynamics often characteristic of enslaved communities, such as restricted reproduction, leave signatures in the island's genetic diversity, or not?
7. What are the differences in genome-wide genetic diversity patterns between São Tomé and Cabo Verde?
 - a. How did the differing origins of enslaved Africans and the distinct slave societies that developed in São Tomé and Cabo Verde shape the sex-specific demographic histories and mating behaviours of their populations?
 - b. In what ways might these contrasting colonial experiences have resulted in distinct genetic diversity patterns between the two archipelagos?

Content of the Thesis

In **Chapter 1**, I review the key questions that anthropological genetics has attempted to address in the context of the TAST over the past two decades. I highlight both the similarities and discrepancies between the demographic events of the TAST, as reconstructed from historical records, and the genetic findings based on observed patterns of genetic diversity in present-day populations descended from enslaved Africans.

The chapter begins with a historical overview of the TAST, focusing on the key events reconstructed from historical data, while also discussing the limitations of these reconstructions, as acknowledged by many historians. In particular, despite recent efforts to compile large-scale databases, the structure of such databases is highly dependent on classification methods. In addition, it is important to note that most TAST databases primarily record ship voyages and do not record the detailed origins of enslaved Africans before departure nor their destinations after arrival. There are also significant gaps in the data, particularly regarding the sex ratio of the slave trade, which appears to have varied over time and across regions, rather than reaching a steady ratio of two males to one female as some studies have assumed.

The chapter then shifts to genetic perspectives, where I examine the central questions that genetic studies of populations descended from enslaved Africans are able to address. A key question is the genetic impact of the TAST on both sides of the Atlantic. I first discuss the feasibility of comparing results from different studies, each based on different data sets and driven by different research questions. Then, I consider the potential for tracing the African genetic origins of these populations, stressing that genetic ancestry reflects genetic similarity, and the results of genetic studies are dependent on the specific samples used. I also report studies that try to disentangle the relative genetic contributions of African females and males. While the demographic patterns of TAST suggest that more males than females were taken from Africa, many studies of African-descended populations show a higher genetic contribution from African females. This discrepancy is often interpreted as evidence of ancestry-related sex bias in the admixture process. However, these patterns may be further influenced by other non-random mating behaviours and sex-specific migration patterns. Finally, demographic features typical of enslaved communities, such as family structures and

restricted reproductive practices, may also have played a role in shaping these genetic patterns.

Overall, this chapter lays the foundations for the questions that will be explored in the following chapters. It provides an overview of the state-of-the-art work in the field of population genetics and aims to provide a comprehensive framework for future research in this area, which I aim to publish as a review paper.

In **Chapter 2**, I investigate the genetic diversity of the island of São Tomé, focusing on an extensive dataset of 2 million autosomal SNPs genotyped in 96 individual samples collected on the island. I developed a genotyping quality control pipeline to curate genomic data of samples from São Tomé, Cabo Verde, and other African, European and American populations included in available genetic datasets. The final merged dataset included African and European populations that, based on historical records and previous genetic studies, may have contributed to admixture processes in São Tomé and Cabo Verde. Using this dataset, I aimed to investigate the genetic ancestry of São Tomé's inhabitants, identify potential source populations, and explore the admixture and demographic processes that have shaped the island's genetic diversity.

My analysis revealed significant genetic differentiation between individual samples from São Tomé. To better understand this genetic structure, I investigated whether individuals in São Tomé could be grouped on the basis of genetic diversity, which would allow reconstructing the admixture and demographic history of each cluster separately. The most effective approach I found was to use haplotypes shared between individuals, as haplotype-based methods offer higher resolution in detecting fine-scale population structure and recent ancestry than approaches based on allele frequencies only.

I thus modelled the haplotypic ancestry of each São Toméan cluster as a mixture of the possible source populations included in the reference dataset. The results are consistent with historical migrations driven by slavery and indentured servitude, revealing significant differences in ancestry proportions between groups. Analysis based on recent shared haplotypes highlighted that, in addition to differences in admixture histories, São Tomé's genetic diversity has also been shaped by demographic events such as drift during periods of isolation.

Finally, I interpreted the observed genetic clusters in relation to known historical and anthropological communities, including the *Angolares*, the *Forros* and the Cabo Verdeans, using multiple lines of genetic evidence and insights from previous genetic and linguistic studies. I propose that shifts in the historical and social contexts in São Tomé altered the interaction between these communities, leading to gene flow between recently admixed groups that had previously remained isolated on the island. However, disentangling these complex layers of admixture, which form in the process that I call “nested admixture”, remains challenging with current methods and represents an interesting direction for future research.

This research not only highlights the unique admixture and demographic histories that shaped genetic diversity in São Tomé, but also offer broader insights into complex admixture processes. Part of this work has been published as a scientific manuscript entitled “Nested admixture during and after the Trans-Atlantic Slave Trade on the island of São Tomé” in preprint on BioRxiv (Ciccarella et al. 2024) and it is currently *in press* in the Molecular Biology and Evolution journal.

In **Chapter 3**, I analyse the genetic diversity on the X chromosome in São Tomé and Cabo Verde, and compare patterns of genetic diversity on the X chromosome to those on the autosomes to investigate sex-specific demography and mating patterns. The X chromosome spends two third of its time in females and one third in males, and this difference compared to the autosomes allows us to investigate sex-specific dynamics during admixture processes, as well as sex-specific demographic histories, including differences in female and male migration patterns and effective population sizes.

I curated and analysed for the first time several tens of thousands of X chromosome markers in São Tomé, as well as in Cabo Verde and other African population samples that were included in previously published datasets (Patin et al. 2017; Semo et al. 2020; Laurent et al. 2023). Until now, only autosomal data from these populations had been analysed. These African populations can be considered proxies for the source populations involved in the admixture process that occurred in São Tomé, as identified in Chapter 2. This allowed to highlight the importance of considering sex-specific dynamics in source populations too, particularly to explain the patterns of genetic diversity observed in the admixed population that actually formed in the source populations prior to admixture.

In short, I examined genetic diversity on the X chromosome, applying various approaches and methods to compare these patterns with those observed on the autosomes. I interpreted these differences in terms of ancestry-related mating patterns and sex-specific demographic processes during admixture, and proposed several competing demographic models a priori consistent with observed genetic patterns, that will need to be formally tested in future studies. In particular, I proposed to consider specific scenarios, based on genetic findings and historical accounts, as well as specific statistics, such as haplotype sharing, IBD, and ROH patterns, to explain the observed genetic diversity.

I intend to publish these results in a dedicated part of a future article, and I hope that this Chapter may lay the ground for future research developing new methodological frameworks capable of considering such complex demographic scenarios and haplotype-based summary statistics.

CHAPTER 1

1 Genetic and historical perspectives on the Trans-Atlantic Slave Trade

In this chapter, I review the key questions that anthropological genetics has sought to answer in relation to the Trans-Atlantic Slave Trade (TAST) over the past two decades. First, I provide a historical overview of the TAST, emphasising key events derived from historical records, while acknowledging the limitations of these reconstructions. In particular, I note that large-scale databases heavily rely on classification methods, and may lack of detailed information on the specific origins and destinations of enslaved Africans. There are significant gaps in the data, including inconsistencies in the sex ratio of the trade, which varied across time and space rather than being the commonly assumed ratio of two males to one female. The chapter then turns to genetic perspectives, examining the questions that genetic studies can address, such as the impact of the historical migrations associated to the TAST on genetic diversity on both sides of the Atlantic. I assess the comparability of studies based on different datasets and research aims, and explore the challenges of tracing in particular African genetic ancestries. I highlight that genetic ancestry inference reflects patterns of genetic similarity shaped by sampling strategies. I also review studies investigating sex-specific contributions from source populations, where genetic evidence often indicates a greater contribution from African females despite the overall male-biased nature of the trade. This discrepancy is often attributed to ancestry-related sex bias admixture, but may further result from other non-random mating patterns and sex-specific demography. I also open for considerations on how demographic features typical of enslaved communities, such as family structures and constrained reproductive practices, may have further shaped these patterns.

Overall, this chapter lays the ground for the investigations presented in the subsequent chapters and provide a comprehensive framework for future research. My aim is to refine the chapter and publish it as a review paper that will be of interest to other researchers in the field, as well as to a more general audience interested in these questions.

1.1 The Trans-Atlantic Slave Trade from an historical perspective

Our understanding of the TAST has significantly improved in recent decades thanks to the contributions of many historians, particularly from North America (Curtin 1990; Berlin 1998; Brooks 2003; Eltis and Richardson 2010; Lovejoy 2012). One of the most comprehensive studies up to date is the project coordinated by David Eltis and published in 2008 as the Trans-Atlantic Slave Trade Database, which is periodically updated (The Slave Voyages Consortium 2021). In addition to detailed, freely accessible data, the authors of the TAST database provide comprehensive summary maps (

Figure 1.1) and documentation about the methods used to build the dataset. Despite the vast amount of data that they were able to collect, including ports of embarkation and disembarkation of enslaved Africans, and the sex and age-specific demography associated with the slave trade, they warn that in some cases data had to be imputed for completeness and that important gaps remain, particularly with regard to the slave trade under Portuguese and Brazilian control (Eltis and Richardson 2008).

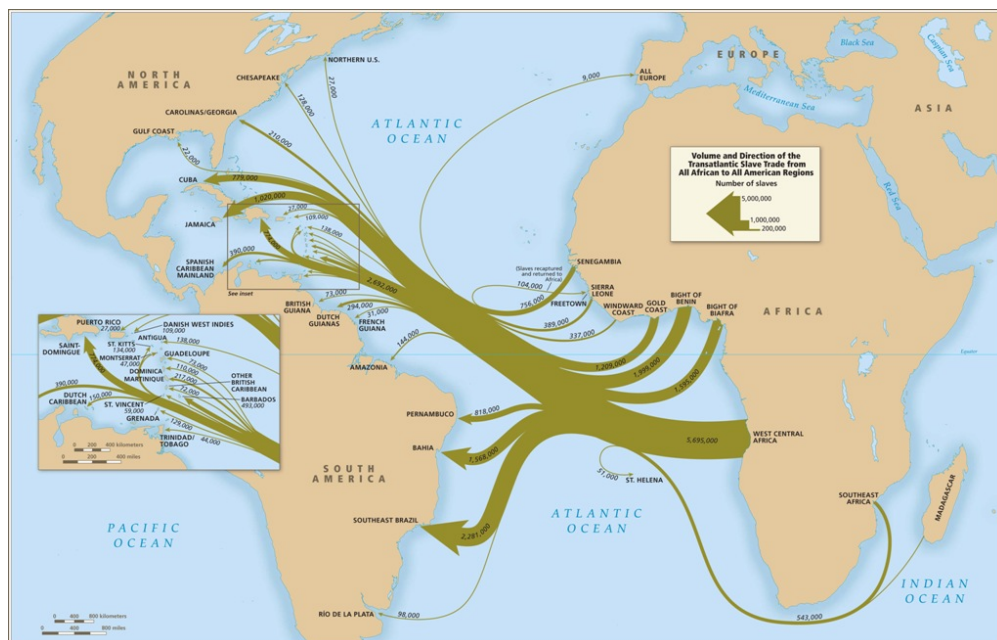


Figure 1.1. Volume and direction of the Trans-Atlantic Slave Trade.

The major regions of embarkation and disembarkation of enslaved Africans are reported in Africa, Europe and the Americas. From Eltis and Richardson (Eltis and Richardson 2010).

Initiated by the Portuguese in the 15th century, the TAST evolved significantly over time, reflecting shifts in European colonial power and their various coastal contacts in Africa (Caldeira 2013). Initially dominated by the Portuguese and Spanish, the trade expanded under Dutch, English, and French influence in the 18th century, reaching then its peak in numbers of displaced humans (Curtin 1969). The major ports of embarkation in Africa also changed over time, depending on European access to African coasts and their relationship with African kingdoms, from early trading regions such as “Senegambia”, to the Gulf of Guinea, including the “Gold Coast”, the “Bight of Benin”, and the “Bight of Biafra”, and finally to “West Central Africa”, especially parts of present-day Congo and Angola, which accounted for almost half of all embarkations between the 16th and the 19th century (Caldeira 2013).

Destinations in the Americas also varied according to colonial demand, with Brazil and the Caribbean receiving the majority of enslaved Africans, over 80% of the total slave voyages included in the TAST database, while relatively less enslaved Africans disembarked in Spanish America and North America. Finally, no more than 2% of slave voyages went to Europe and another 2% to other ports within Africa (Figure 1) (The Slave Voyages Consortium 2021). However, the scale of the intra-African slave trade remains underestimated, and it primarily involved the archipelagos of Cabo Verde and São Tomé e Príncipe, where some enslaved Africans were retained to meet local labour needs, while some others were resold and transported to other destinations (Caldeira 2013).

Some historians have cautioned about using the TAST database as a migration database, since it is based on records of slave trade voyages derived from ship registers and commercial documents, which do not directly indicate the actual origins of enslaved Africans, nor their final destinations after disembarkation. Moreover, for over thirty years, since Curtin’s first categorization (Curtin 1969), critics have raised attention to the regional organisation of the data in historical studies about the TAST. The African regions used in the TAST database do not represent the inland African routes of slave supply, but rather the main coastal ports through which European colonial powers traded slaves. The categories “Senegambia”, “Gold Coast”, etc., which make sense mainly from a European point of view, are fundamental to the structure of the database. At this regard, Lovejoy points out that if the database were rearranged on the basis of historical African regional boundaries, its structure would be very different (Lovejoy 2010).

As well as providing estimates of the total number of enslaved Africans shipped across the Atlantic, the TAST database breaks down the data by sex and age. In the 18th century, the ratio of males to females increased steadily, eventually reaching about two males for every female, according to the authors (Eltis and Richardson 2010). In contrast, the overall sex ratio was more balanced in the earlier period, at least from the second half of the 17th century for which there is relatively reliable historical data. From the 18th century onwards, the TAST saw an increase in the number of males and children traded (The Slave Voyages Consortium 2021). The age used to define ‘children’ could vary based on local legal definitions or societal norms of the time, making it challenging to generalize. Nevertheless, estimates suggest that at the peak of the TAST, children accounted for approximately a quarter of the enslaved Africans (Eltis and Richardson 2010).

Overall sex ratios from the TAST database should be treated with caution, as information on the sex of enslaved Africans is reported for less than 10% of the slave voyages (The Slave Voyages Consortium 2021). Importantly, the sex is not reported for most of the numerous voyages directed to Brazil and Spanish America until the 19th century. Female enslaved Africans were generally more numerous before the 18th century than after, and in some slave voyages, they actually outnumbered males (Morgan 2021).

The sex-specific demographic composition of enslaved Africans varied considerably among exporting regions as well as among importing regions, likely depending on which colonial power was in charge of trading, when the slave voyages were organized, and where the enslaved Africans came from (Geggus 1989a; Eltis and Richardson 2010). In the mid-17th century, for example, sex ratios were particularly low among slave voyages organized by British, French and Dutch traders, including departures from West Central Africa, which accounted for nearly half of all departures and had some of the highest reported sex ratios of the TAST. By the 18th century, French traders in West Central Africa actually reached the ratio of two males to one female, while British traders in the same region did not (Geggus 1989b). Because the sex ratio of enslaved Africans varied widely and remain largely unknown, it is important not to assume that it was universally or consistently male biased.

1.2 What is the genetic impact of the TAST on both sides of the Atlantic?

A comprehensive understanding of the genetic impact of the TAST in populations descended from enslaved Africans is challenging because each genetic study is shaped by specific research questions that make comparisons between studies difficult. Importantly, the reported proportions of African genetic ancestry among different studies are significantly influenced by sampling strategies (Fortes-Lima and Verdu 2020).

Studies that focus exclusively on self-identified African descendants are more likely to observe higher proportions of African ancestry than studies that consider a more heterogeneous sample of the population. In one such study of self-identified African Americans in the United States, the average African ancestry was estimated to be 75% (Patin et al. 2017). In contrast, the Brazilian sample of the EPIGEN project, which included individuals who self-identified as “Black”, “Mixed”, and “White”, found diverse African contributions ranging from 50.8% in the city of Salvador, including most of the self-identified “Black” and “Mixed” individual samples, to less than 16% in Pelotas and Bambui, mostly including self-identified “White” individuals (Kehdy et al. 2015). In addition, the authors caution against interpreting the Salvador sample as representative of the entire population, as individuals living in poor environments were oversampled (Lima-Costa et al. 2015). Another study on maroon communities in French Guiana and Suriname, specifically designed to target these isolated populations, found some of the highest levels of African ancestry in the Americas at 98% (Fortes-Lima et al. 2017). In the same research, a sample of African-Brazilians from Rio de Janeiro in Brazil showed 71% African ancestry. These examples underscore the difficulty of assessing the overall genetic impact of the TAST across the Americas, as the results depend heavily on the specific population samples studied and the questions driving the research.

To gain a broader understanding of the effects of European colonization and forced migrations associated with TAST, particularly in the Americas, researchers have recently studied multiple enslaved African populations simultaneously using samples collected through different sampling strategies, but using consistent methodologies that allow for more comparable results across populations (Ongaro et al. 2019; Gouveia et al. 2020; Micheletti et al. 2020). Despite these efforts, general conclusions about broader patterns of the TAST must be approached with caution due to the inherent limitations of the

datasets used and the diversity of complex admixture processes that, over time, shaped genetic diversity patterns across the Americas. Indeed, the genetic ancestry observed in present-day populations reflects not only the historical migrations and forced displacements of the TAST but also the diverse and complex processes of genetic admixture and/or social barriers to reproduction across socio-historical communities that occurred over centuries, shaped by local social, cultural, and historical contexts.

More generally, genetic studies in recent decades have focused on the impacts of African forced migrations to the Americas, with considerably less attention paid to the genetic impacts of the TAST on African populations themselves. The slave trade likely had a significant impact on the genetic diversity of contemporary African populations, both through the loss of individuals to deportation and through genetic interactions that resulted from the contact of Africans and Europeans within Africa. Notably, the African Atlantic islands of Cabo Verde and São Tomé offer some of the earliest examples of slave societies of the colonial era, and the micro-scale genetic studies on these populations have shed new light on the formation of complex genetic admixture patterns in the context of the TAST (Tomas et al. 2002; Coelho et al. 2008; Beleza et al. 2012; Verdu et al. 2017; Almeida et al. 2021; Hamid et al. 2021; Korunes et al. 2022; Hamid et al. 2023; Laurent et al. 2023). Future research should examine these and other African populations to provide a more comprehensive picture of the multiple impacts of the TAST on both sides of the Atlantic.

1.3 Where in Africa did the genetic ancestors of African-descendant populations come from?

One of the main questions that genetics has sought to answer in the context of TAST is where in Africa the genetic ancestors of enslaved African-descendant populations came from: a question that is often difficult or even impossible to address using historical records alone.

Researchers aim to infer genetic ancestry by comparing some measure of genetic similarity between a target population of interest, which is assumed to be admixed *a priori*, and a reference panel that includes samples from populations that are assumed to be possible sources (Mathieson and Scally 2020; Coop 2022; Lewis et al. 2022). Therefore, these inferences depend heavily on the quality and representativeness of the genetic datasets employed. Recent advances, such as the H3Africa initiative, have significantly expanded available data for African populations, uncovering novel variants and enhancing our understanding of African genetic diversity (Ramsay et al. 2021). However, access to genetic datasets is often restricted to specific research purposes, and technical limitations, such as differences in sequencing and genotyping technologies, complicate data integration into joint analyses. More generally, there are inherent limitations in capturing the full diversity of African populations, and the genetic diversity of current populations has been shaped by the past 500 years of historical, political and social changes in Africa, thus it may not fully represent the genetic diversity of past populations involved in the TAST.

Another critical challenge in genetic ancestry inference studies is sample classification. When building the reference panel, researchers can either retain the original labels assigned by the studies from which the samples were drawn or re-cluster individuals based on genetic criteria. Retaining original labels introduces the difficulty of harmonizing population labels of different samples, which may be based on different descriptors - such as ethnicity, language, geography, self-identification, and/or genetic criteria - and reassigning them into broader categories for analysis. Conversely, clustering by genetic similarity can avoid some of these issues, but introduces others: assigning labels to genetically defined groups can lead to challenges in interpreting results, particularly in terms of what these genetic clusters represent anthropologically. Transparent methods of classification and standardized approaches are essential to avoid potential

misinterpretation of genetic findings, particularly when presenting complex genetic results to broader audiences (National Academies of Sciences, Engineering, and Medicine 2023).

When genetic data are explicitly compared with historical records, such as those in the TAST database, discrepancies or unexpected patterns may emerge. Some authors found an over-representation of West Central African ancestry in African descendant populations across the Americas compared to their expectations based on the number of slaves that have been deported from this region (Gouveia et al. 2020; Micheletti et al. 2020). However, this comparison is complicated by incomplete historical records, biases in genetic sampling and categorization of African populations (Jackson 2021), and the complexities of admixture histories influenced by forced migrations, internal redistribution, and forced interbreeding during and after slavery, as acknowledged by several authors (Ongaro et al. 2019; Fortes-Lima and Verdu 2020; Micheletti et al. 2020; Micheletti et al. 2021). These factors highlight the need for careful, multidisciplinary approaches that integrate genetic data with historical and anthropological perspectives to effectively reconstruct the demographic history of the African diaspora and the complex admixture processes that shaped the genetic diversity of current populations descended from enslaved Africans.

1.4 How much female and male genetic ancestry came from Africa?

Several genetic studies have suggested a discrepancy between the sex-specific demography of enslaved Africans reconstructed from historical records and genetic findings about the female and male ancestors of admixed populations. Although overall more males than females were most probably deported from Africa, genetic studies suggest a bias toward European males reproducing with African and/or Native American females in several admixed populations across the Americas and in Africa (Lind et al. 2007; Bryc et al. 2010; Beleza et al. 2012; Moreno-Estrada et al. 2013; Bryc et al. 2015; Homburger et al. 2015; Kehdy et al. 2015; Fortes-Lima et al. 2017; Ongaro et al. 2019; Micheletti et al. 2020; Korunes et al. 2022). We refer to this event, when females and males of the source populations contribute differently to the admixed population, as ancestry-related sex bias.

To investigate ancestry-related sex bias, early studies analysed markers on maternally inherited mitochondrial DNA (mtDNA) and paternally inherited Y chromosomes (Parra et al. 1998; Alves-Silva et al. 2000; Parra et al. 2001; Lind et al. 2007; Stefflova et al. 2009; Perego et al. 2012). However, interpreting and comparing results based on the mtDNA and the Y chromosome is challenging. These uniparental markers experience different selective pressures and follow different inheritance patterns: mtDNA is transmitted to both female and male offspring, whereas the Y chromosome is transmitted only to male offspring. In addition, demographic events, such as bottlenecks, can strongly influence their genetic diversity and lead to sudden lineage losses. Indeed, they each represent single loci and therefore each represent a single genealogical lineage in the entire genealogy of each sampled individual.

More recently, advances in genome-wide next-generation sequencing, combined with an improved theoretical understanding of complex genetic admixture processes, have made it possible to analyse and compare the genetic diversity on the autosomes and the X chromosome, which is typically present in two copies in females and one copy in males (Ramachandran et al. 2004; Hammer et al. 2008; Segurel et al. 2008; Bustamante and Ramachandran 2009; Keinan et al. 2009; Casto et al. 2010; Heyer et al. 2012; Goldberg and Rosenberg 2015; Cai et al. 2023; Cotter et al. 2024). Compared to uniparental markers, the autosomes and X chromosome offer higher resolution for inferring

admixture patterns due to the larger number of markers available and the effects of recombination, which rearranges genetic material from the female and male parental genomes during reproduction, providing a detailed record of ancestry segments across generations that is highly informative for retracing admixture events over time.

Many genetic studies on populations descending from enslaved Africans include some analysis of the mtDNA and sex chromosomes to look for patterns that may result from ancestry-related sex bias (Parra et al. 1998; Lind et al. 2007; Coelho et al. 2008; Bryc et al. 2010; Beleza et al. 2012; Moreno-Estrada et al. 2013; Bryc et al. 2015; Homburger et al. 2015; Baharian et al. 2016; Mathias et al. 2016; Fortes-Lima et al. 2018; Ongaro et al. 2019; Micheletti et al. 2020; Almeida et al. 2021; Korunes et al. 2022). In most cases, if markers on the mtDNA or X chromosome show greater genetic similarity to a source population compared to markers on the Y chromosome or autosomal DNA, it is interpreted as evidence that this population contributed more females than males to the gene pool of the admixed population. Sometimes, such genetic signal is also used to conclude that the other source populations contributed with more males than females. However, this is a specific case of ancestry-related sex bias, and different scenarios may produce the same patterns of genetic diversity.

Mechanistic admixture models permit to quantitatively examine how admixture processes shape ancestry levels within populations. Verdu and Rosenberg (Verdu and Rosenberg 2011) built a mechanistic model of admixture that accounts for changing admixture contributions over time, allowing to study how ancestry distributions evolve in an admixed population. Goldberg et al. (Goldberg et al. 2014) and Goldberg and Rosenberg (Goldberg and Rosenberg 2015) expanded on this model to include varying contributions from females and males of different source populations, exploring how sex-specific contributions from source populations affect autosomal and X-chromosomal ancestry. Interestingly, they found that ancestry-related sex bias leaves a signature also on the variance of ancestry segments on the autosomes, which are not inherited differentially in females and males. Moreover, they estimated sex-specific contributions from African and European sources in African-Americans, and they found that differences in X-chromosomal and autosomal admixture could also arise from a male bias from both source populations, with a larger such male bias from Europeans (Goldberg and Rosenberg 2015).

There are two main approaches to the analysis of autosomal and X chromosome diversity. Genotype-based methods, implemented in STRUCTURE-like algorithms such as ADMIXTURE (Alexander et al. 2009), provide Global Ancestry Inferences (GAI) that estimate the overall proportions of ancestry in an individual's genome. In contrast, haplotype-based methods, implemented in chromosome painting algorithms such as RFMIX (Maples et al. 2013) and ChromoPainter (Lawson et al. 2012), assign specific chromosome segments to different ancestries, generating also Local Ancestry Inferences (LAI) with finer resolution. A recent study applied LAI approaches for the first time to investigate ancestry-related sex bias across the Americas, and found that both GAI and LAI approaches yield consistent results (Ongaro et al. 2019; Ongaro et al. 2021). Higher European ancestry is observed on the autosomes, and higher Native American ancestry on the X chromosome, suggesting strong European male and Native American female bias. Conversely, African ancestry shows more variability, with only some populations presenting slightly higher African ancestry on the X chromosome. Variability in ancestry-related sex bias has also been highlighted in another study (Micheletti et al. 2020); in particular, the greater African female genetic contributions in Latin America compared to British colonized regions in the Americas may reflect differences in colonial and demographic history that may be explained by scenarios other than ancestry-related sex bias.

1.5 Who mated with whom in admixed populations?

Admixture in human populations is influenced by a number of sociocultural factors that make mate choice far from random. Ancestry-related sex bias, in which individuals preferentially mate with those of different genetic ancestry, and ancestry-related assortative mating, in which individuals preferentially mate with those of similar genetic ancestry, are both forms of non-random mating that shape genetic diversity and admixture patterns (Goldberg and Rosenberg 2015; Norris et al. 2019; Goldberg et al. 2020; Mas-Sandoval et al. 2023). Because ancestry cannot be determined without genetic testing, these mating behaviours are related to social and cultural dynamics. In particular, socio-cultural barriers may strongly limit reproduction between human groups, leading to the accumulation of genetic diversity within groups that eventually form subpopulations. In recently admixed populations, this population structure may also be associated to different proportions of genetic ancestry.

Particularly in the colonial context, social stratification was deeply rooted in racial ideologies, meaning that individuals were assigned to social groups largely on the basis of their presumed genetic ancestry. Some phenotypes became key markers in social categorization, based on the assumption that they reliably reflected genetic ancestry. However, genetic studies suggest that the link between phenotypes, such as skin colour, and genetic ancestry can rapidly decouple (Parra et al. 2003; Beleza et al. 2013; Ruiz-Linares et al. 2014; Kim et al. 2021). In other words, phenotypes associated with ancestry in the early stages of admixture eventually become uncorrelated with overall genetic ancestry, highlighting the limited genetic basis for social hierarchies constructed around presumed genetic ancestry. A recent study tested this hypothesis using a mechanistic model that considers genetic admixture, phenotypic traits, and mating behaviour together, and found that this is particularly true when phenotypes are determined by a few genetic loci, and that assortative mating slows a bit this rapid decoupling (Kim et al. 2021).

Ancestry plays a pivotal role in shaping mating patterns in modern admixed populations, regardless of the specific social and cultural dynamics at play (Sebro et al. 2017; Zaitlen et al. 2017). In contemporary Latin American populations, individuals are more likely to choose partners with similar ancestry profiles than would occur by chance, with the specific ancestries driving mating choices varying across different populations, reflecting

the influence of distinct social and cultural dynamics on ongoing processes of assortative mating (Risch et al. 2009; Zou et al. 2015). While previous studies relied on estimates of global ancestry fractions between mating pairs, recent advances allow researchers to detect signals of ancestry-related assortative mating without specific paired-data. Evidence of assortative mating can be inferred from the analysis of local ancestry tracts in homologous chromosomes within individual genomes (Norris et al. 2019). However, these methods are primarily effective for identifying assortative mating events that occurred within the past few generations.

Few studies have developed admixture models that incorporate non-random mating, particularly ancestry-related sex bias and assortative mating (Goldberg et al. 2020; Kim et al. 2021; Mas-Sandoval et al. 2023). These processes may evolve over time or coexist within the same admixture history. For example, norms favouring mating of females of one group with males of another group, initially associated with different genetic ancestry, may drive sex bias. In parallel, sociocultural barriers that reinforce mating of individuals within the same group who share similar genetic ancestry may contribute to assortative mating. Such dynamics are deeply linked to historical contexts, including policies of social segregation designed to maintain social barriers that reinforced assortative mating, and ideologies such as *branqueamento* in Latin America, which blurred and paradoxically reinforced race, class, and gender hierarchies through highly gendered admixture processes (Wade 2009). A recent study modelled these dynamics and incorporated both ancestry-related assortative mating and sex bias parameters into genetic simulations under different admixture scenarios. Using neural networks to estimate non-random mating parameters for different populations across the Americas, the study revealed that in Latin Americans mating probabilities were correlated to the proportion of Amerindian ancestry between males and females, while in African Americans to African ancestry (Mas-Sandoval et al. 2023).

1.6 When did admixture occur?

Over the past decade, genetic studies have uncovered remarkably diverse and complex admixture dynamics across populations descended from enslaved Africans on both sides of the Atlantic, revealing how different socio-historical contexts influenced the timing and extent of admixture (Moreno-Estrada et al. 2013; Bryc et al. 2015; Baharian et al. 2016; Fortes-Lima et al. 2018; Ongaro et al. 2019; Micheletti et al. 2020; Korunes et al. 2022; Laurent et al. 2023).

Admixture events do not always coincide with major periods of the TAST. For example, in Cabo Verde, early admixture occurred during the 16th century, aligning with the onset of European colonial expansion, while admixture during the TAST was limited (Laurent et al. 2023). Conversely, in many Caribbean islands, substantial admixture coincide with the plantation economy (Moreno-Estrada et al. 2013; Fortes-Lima et al. 2018). In some regions across the Americas, such as parts of Central and South America and the Caribbean, as in Cabo Verde, admixture intensified after slavery was abolished, driven by changing inter-community relationships and socio-political conditions (Fortes-Lima et al. 2018; Ongaro et al. 2019; Gouveia et al. 2020; Micheletti et al. 2020). In North America, the distribution of African ancestry in African Americans has been shaped not only by the TAST but also by more recent migrations, including movements during the Great Migration of the 20th century (Baharian et al. 2016). These examples highlight the need to consider local social and historical contexts to fully capture the complexity of admixture histories.

Dating admixture events based on genetic data is particularly challenging due to several factors. Most methods depend heavily on the accuracy of Local Ancestry Inferences (LAI), which identify genomic segments from different source populations. Most LAI methods rely on accurate genetic phasing, as phasing errors can fragment long ancestry tracts into smaller, misassigned segments. Additionally, the choice of source populations is critical: if these populations fail to capture the full diversity of source populations, certain segments may be misassigned or unclassified. LAI performs best when source populations are highly differentiated, such as African and non-African populations, where tools like RFMix are highly effective (Maples et al. 2013). However, for populations with closely related sources, methods like Globetrotter (Hellenthal et al. 2014), or recently developed tools like FLARE (Browning et al. 2023) and Gnomix (Hilmarsson et al. 2021),

may yield better results. Admixture dating methods also face challenges due to their reliance on simplified models, which often assume one or two discrete admixture pulses, potentially overlooking patterns of continuous or fluctuating gene flow (Gravel 2012; Hellenthal et al. 2014). However, recent advances have addressed some of these limitations. For instance, novel frameworks can model continuous and multiple pulses of admixture, providing insights into complex demographic processes (Fortes-Lima et al. 2021; Laurent et al. 2023).

Ancestry-related sex bias and assortative mating significantly affect the decay of Linkage Disequilibrium in admixed populations, which is critical for accurately dating admixture events (Zaitlen et al. 2017). In admixed populations, the length of continuous ancestry tracts is often used to infer the time since admixture, as these tracts become progressively shorter at each generation due to recombination. However, in the presence of ancestry-related assortative mating, haplotypes remain longer intact, leading to the underestimation of the number of generations since admixture. This discrepancy can result in timing estimates that conflict with historical records (Korunes et al. 2022). Moreover, tract length patterns can also reveal the degree of non-random mating associated with genetic ancestry (Mas-Sandoval et al. 2023). Incorporating assortative mating into demographic models improves the alignment of admixture timing estimates with documented history, though these estimates still depend on assumptions about the underlying population structure and demography over time.

1.7 Final remarks and future directions

The increasing availability of genome-wide data and novel statistical and bioinformatic methods to infer admixture have enabled population geneticists to shed new light over the past 500 years of history of admixture of populations descended from enslaved Africans. Genetic studies have traced the genetic ancestries of these populations to various regions in Africa and highlighted the diversity of admixture dynamics across the Americas and in Africa. Importantly, the patterns of genetic diversity of each population descended from enslaved Africans have been shaped by specific ranges of historical and social factors influencing genetic admixture processes both during and following the TAST (Moreno-Estrada et al. 2013; Bryc et al. 2015; Homburger et al. 2015; Baharian et al. 2016; Fortes-Lima et al. 2017; Ongaro et al. 2019; Gouveia et al. 2020; Micheletti et al. 2020; Korunes et al. 2022; Laurent et al. 2023; Mas-Sandoval et al. 2023).

Ancestry-related sex bias has emerged as a key focus in understanding the genetic histories of populations descended from enslaved Africans. Most studies find evidence of this bias by analysing patterns of genetic diversity in genomic regions that follow different rules of inheritance. These studies often reveal higher levels of European male and African and/or Native American female genetic ancestry, which are presented in contrast to the demography of the TAST reconstructed from historical records, reporting that overall more African males than females were forcibly displaced during the TAST (Eltis and Richardson 2010).

While some largely attribute this pattern to the sexual exploitation of enslaved women during the TAST (Micheletti et al. 2020), this explanation, although plausible, risks oversimplifying the complexities of the colonial era and reinforcing stereotypical narratives. Other scenarios, beyond European male bias and African or Native American female bias, may also explain these genetic patterns (Heyer et al. 2012; Goldberg and Rosenberg 2015). Moreover, historical data on the sex of enslaved individuals remains limited and suggests considerable variability in sex ratios across regions and time periods (Geggus 1989a; Eltis and Richardson 2010; Morgan 2021). These findings underscore the importance of interpreting genetic diversity within broader historical and social contexts that acknowledge the multiple roles of African, Native American and admixed women, not only as subjects of exploitation, but also as central figures in enslaved communities and colonial societies (Morgan 2021).

Ancestry-related assortative mating, another form of non-random mating, has recently been studied mainly in Latin American populations. In colonial societies, mate choice was heavily influenced by rigid social constraints, and even after the abolition of slavery, social hierarchies based on race, class and gender continued to shape social norms and policies, significantly influencing patterns of mate choice (Wade 2009). Key studies have demonstrated that accounting for non-random mating, such as ancestry-related sex bias and assortative mating, is essential for accurately estimating both the extent and timing of admixture events (Zaitlen et al. 2017; Korunes et al. 2022). Moreover, recent findings emphasize the importance of incorporating changing social structures (Mas-Sandoval et al. 2023) and admixture dynamics (Fortes-Lima et al. 2021; Laurent et al. 2023) into admixture models.

Future research on the genetic diversity of populations descended from enslaved Africans could benefit from extending admixture models to include complex scenarios that better represent past admixture dynamics. Incorporating sex-biased migration patterns, and temporal shifts in ancestry-related sex bias and assortative mating could provide deeper insights into admixture histories. Furthermore, differences in effective population sizes between the sexes, driven by sex-specific demographic behaviours such as residence and descent rules and the transmission of reproductive success, also play a critical role in shaping patterns of genetic diversity on sex chromosomes (Heyer et al. 2012; Guyon et al. 2024).

Historical records and legal frameworks underscore the unique constraints faced by enslaved communities, including forced reproduction, unequal sex ratios, and high mortality rates. In many colonial contexts, the principle of *partus sequitur ventrem*, for which enslaved status was inherited through the maternal line, was a critical mechanism for maintaining slavery (Morgan 2018). This principle was included in legal codes such as the Code Noir for French colonies (Niort 2023), was also applied in British colonies (Newman 2020), and reaffirmed in post-independence Brazil after the abolition of the slave trade (Santos 2016). These matrilineal descent norms in enslaved communities, applied in otherwise patrilineal societies, varied across colonial powers in their adoption, enforcement and duration. Investigating whether these particular descent rules have left detectable genetic signatures could shed light on their impact on genome-wide patterns of genetic diversity, and offer insight into the extent and timing of application of such norms.

CHAPTER 2

2 Nested admixture during and after the Trans-Atlantic Slave Trade on the island of São Tomé

In this chapter, I investigate the histories of genetic admixture in São Tomé, an African island in the Atlantic Ocean first settled by the Portuguese in the 15th century. The successive plantation economies that developed on the island triggered several forced and deliberate migrations from Africa and Europe, and resulted in strong social stratification. I examined 2.5 million autosomal SNPs that were newly genotyped in 96 São Toméans, alongside an extensive sample from the archipelago of Cabo Verde, located off the coast of Senegal, which also experienced five centuries of Portuguese colonisation. Based on haplotype sharing, I identified five genetic clusters in São Tomé and reconstructed the admixture histories of each genetic cluster separately, considering 70 African and European population samples. Using Local Ancestry Inference methods and Identical-By-Descent tract sharing, I further identified gene flow between admixed groups that were previously isolated on the island. I emphasise the importance of considering this process, which I named “nested admixture”, since it creates layers of genetic ancestry in the admixed population that are difficult to disentangle using current methods. Source populations are not static entities, but are themselves the result of previous admixture events in dynamic and complex genetic admixture process. This study sheds new light on the impact of historical migrations on the current patterns of genetic ancestry observed in São Tomé, as well as on the influence of social structures on the formation and transformation of genetic structures.

This study is the subject of a scientific paper that is currently *in press* in the Molecular Biology and Evolution journal. A preprint version of the paper has also been previously published in BioRxiv (Ciccarella et al. 2024). This chapter provides further information and analyses that go beyond the scope and format of the scientific paper.

2.1 Introduction

The island of São Tomé provides unique opportunities to disentangle the impact of historical migrations and social stratification on genetic diversity patterns in the context of the Trans-Atlantic Slave Trade (TAST), throughout the expansion of one of the first Plantation Economies, and after the abolition of the TAST and that of slavery.

The archipelago of São Tomé e Príncipe, in the Gulf of Guinea, was uninhabited at the time the Portuguese arrived on the island of São Tomé around 1470 and settled the island from 1493 onwards (Caldeira 2011). In São Tomé, the colonizers established one of the first slave societies based on systematized plantations. This economic system, founded on social and political inequality and the exploitation of enslaved labour for large-scale production and export of sugar, was later deployed throughout European colonies in the Caribbeans and the Americas (Curtin 1990).

Two major plantation economies marked São Tomé's history and society (Figure 2.1). The first, based on sugarcane production and the exploitation of enslaved labour, emerged in the late 15th century and lasted until the 17th century, reaching its peak in the mid-16th century. The second, based on coffee and cacao production, initially relied on enslaved labour in the 19th century, but, following the abolition of slavery in the Portuguese Empire, it transitioned to a new and controversial form of indentured labour until the 20th century.

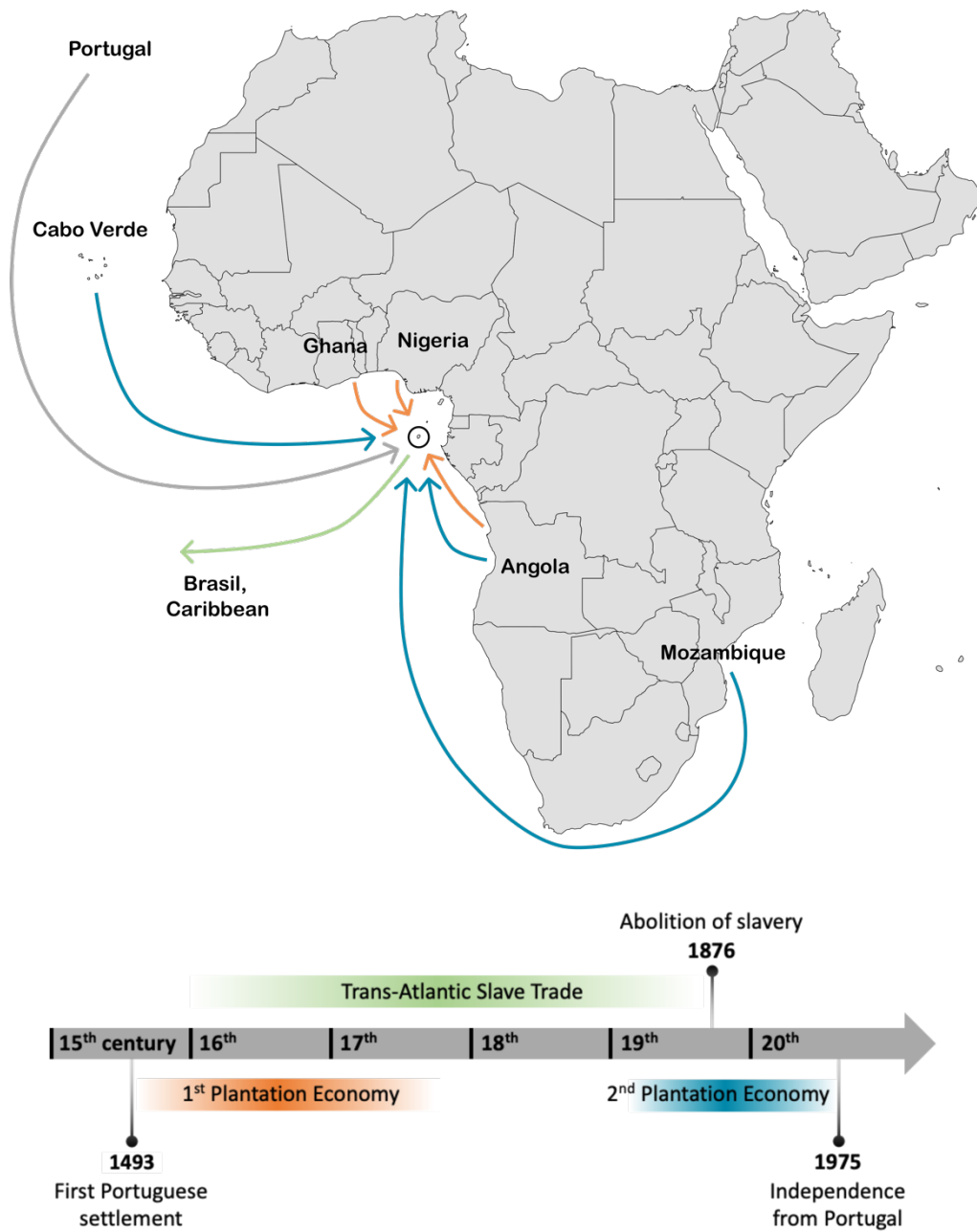


Figure 2.1 Major historical migrations to the island of São Tomé.

The map of Africa illustrates the most significant waves of forced and deliberate migration to São Tomé throughout five centuries of Portuguese colonisation. The colours of the arrows correspond to the major historical events associated with the settlement and demography of the island, as reported in the timeline below.

2.1.1 São Tomé and the emergence of the slave trade and slave societies

The colonization of the island of São Tomé in the late 15th century was crucial to the establishment of the slave trade along the African coasts, even before it reached the other side of the Atlantic. According to historical records, the Portuguese first established commercial relations with the Kingdom of Benin, in a region included into present-day Nigeria, and initially transported enslaved Africans from the Niger Delta to São Tomé. The exchange of European goods, enslaved people, and gold, namely the precursor of the triangular trade, was first established between São Tomé, the Kingdom of Benin, and Mina, another important Portuguese trading port on the so-called “Gold Coast” in present-day Ghana (Vogt 1973). As the demand for enslaved labour force increased, Portuguese and São Toméan merchants expanded their slave recruitment areas southwards, to the Kingdom of Kongo, encompassing the present-day regions of northern Angola, western DRC, and southern Republic of the Congo and Gabon (Almeida Mendes 2008).

The Portuguese introduced sugarcane cultivation to São Tomé at the very beginning of colonization, following earlier attempts on other Atlantic islands such as the Azores, Madeira, and Cabo Verde (Almeida Mendes 2008). The tropical environment of São Tomé proved particularly favourable for sugarcane cultivation, and the island quickly became one of the world’s leading producers of sugar in the 16th century – a process largely driven by the central role of São Tomé in the slave trade with the African coasts (Caldeira 2011). Slavery soon became central to São Tomé’s economic and social structure, turning the island into one of the first examples of “slave society” (Berlin 1998). This model was later replicated in the Americas, where large plantations also relied on enslaved labour and the TAST. Notably, Brazil emerged as a major slave society and one of the largest producers of sugar. The competition with the Brazilian market was indeed one of the main factors contributing to the economic decline of São Tomé’s plantation economy in the 17th century (Caldeira 2011).

São Tomé remained a slave society even after the decline of sugarcane production, as slavery was central to the island’s economy, but more dependent on the slave trade than plantations. From the 16th century until the abolition of the slave trade and slavery in the 19th century, São Tomé was a major hub of the TAST, with enslaved Africans imported

from various regions along the African coasts and exported to the Americas, particularly Brazil and the Caribbean (Caldeira 2013; Caldeira 2021).

2.1.2 Cacao plantations in São Tomé: from slavery to indentured labour

In the 19th century, coffee and cacao plantations were introduced in São Tomé e Príncipe, and the islands suddenly became the most important producers of cacao in Africa in the early 20th century (Kiesow 2017). The establishment of this second major plantation economy in São Tomé first relied on the exploitation of enslaved Africans from Angola (Clarence-Smith 1976). However, the 19th century marked a turning point, as slavery was gradually abolished in the Portuguese Empire. Following the official abolition of slavery in 1869, enslaved Africans working on São Tomé's plantations were replaced by indentured laborers, known as *serviçais*, who were recruited from other Portuguese colonies, including Angola, Mozambique, and Cabo Verde (Clarence-Smith 2000).

Although the working conditions on the *roças*, the colonial plantation estates shaping land ownership and social structure on the island, were generally better than those previously experienced by enslaved individuals, indentured workers were often not granted the means or freedom to return to their countries once their contracts had ended. This provoked significant debate in Europe regarding whether slavery had truly ended in the Portuguese Empire (Higgs 2013).

Historical records document the number and origins of these indentured labourers, which represented a substantial demographic shift on the island of São Tomé. Notably, over the course of the 20th century, around 80,000 individuals were recruited only from the archipelago of Cabo Verde (Carreira 1976). However, little is known about how many of them were actually repatriated at the end of their contracts. In this context, genetic data can provide new insights into the history of the *serviçais*, and the extent to which they remained in São Tomé and left descendants.

2.1.3 Social stratification and creole languages in São Tomé

Extant communities in São Tomé are associated with different histories related to marooning, slavery, and indentured servitude, which are also reflected in the linguistic landscape of the island.

According to historical records, runaway slaves escaped plantations since the beginning of the Portuguese colonisation of São Tomé. They took refuge in the island's forested interior, remaining isolated until the 19th century, when Portuguese colonists finally took possession of their land. Thus, some of these maroon communities began to establish social and economic links with the rest of the island's population, becoming known as the *Angolares* (Seibert 1998). Today, *Angolares* communities are mostly found in the north-eastern and west-southern coasts of the island, and they are known to speak the Angolar, one of the two autochthonous creole languages of the island (Lorenzino 1998; Hagemeijer 2011).

During the 17th and 18th centuries, freed enslaved Africans became the largest social group in São Tomé. This period was marked by a profound economic decline between the two major phases of plantation economies on the island, during which the Portuguese settlers and the Portuguese crown almost abandoned the colony (Seibert 2015). The descendants of the freed population of *Forros* are often associated with the Forro creole language, "Forro" literally meaning "freed slave" (Hagemeijer 2011).

Finally, the migration of contractual workers to São Tomé in the 19th and 20th century resulted in a significant demographic shift on the island. In particular, the many Cabo Verdeans that arrived in São Tomé, and their descendants, now constitute a third major social group on the island. According to the most recent census, 8% of São Toméans speak Cabo Verdean Kriolu (Instituto Nacional de Estatística 2012).

2.1.4 Previous genetic studies

Previous genetic studies in São Tomé identified substantial genetic structure and complex genetic admixture patterns on the island, even using a limited number of genetic markers (Tomas et al. 2002; Coelho et al. 2008) or samples (Almeida et al. 2021). Coelho et al. suggested that the observed genetic structure was mainly caused by a strong signal of genetic differentiation between *Angolares* and non-*Angolares* individuals, with increased genetic drift having occurred in the former group. More recently, Almeida et

al. generated exome sequence data in 25 individuals from the islands of São Tomé e Príncipe, and found similar genetic contributions from the Gulf of Guinea and Angola in both Angolar and Forro speakers in São Tomé.

2.1.5 The aim of this study

I aim to investigate the impact of successive historical migrations to São Tomé, associated to different periods of Portuguese colonization, as well as changing social structures on the island over time, on the current genetic diversity of São Toméans. The strong genetic structure and mosaic of African genetic diversity found by previous genetic studies in São Tomé raised a series of fundamental questions about the histories of admixture in the island:

- What impact have historical voluntary and forced migrations to São Tomé had on the island's genetic diversity?
- How are the various genetic ancestries distributed among present-day São Toméans?
- Are existing linguistic and anthropological communities associated with the genetic structure of the island?
- When and how did genetic admixture and isolation processes occur?

I investigated 96 unrelated individuals from 14 sampling sites across the main inhabited areas of the island of São Tomé. Individual samples were newly genotyped at 2.5 million genome-wide SNPs. Based on haplotype-sharing patterns within the island, I identified five clusters of genetically resembling individuals in São Tomé. I inferred shared haplotypic ancestries for the five São Toméan clusters using previously published genome-wide data from 70 African and European populations, including an extensive sample from the Cabo Verde archipelago (Auton et al. 2015; Gurdasani et al. 2015; Patin et al. 2017; Semo et al. 2020; Laurent et al. 2023). I was able to reconstruct recent gene flow between admixed groups using patterns of long Identical-by-Descent tracts shared among individuals from São Tomé and Cabo Verde. Finally, I propose a chronology of successive admixture and isolation events that may explain the observed genetic diversity in São Tomé.

2.2 Results

2.2.1 São Tomé genetic diversity in the global context

I investigated the genetic diversity in the island of São Tomé and in the worldwide context, by combining the genotype data of the 96 unrelated individuals from São Tomé (Figure 2.2) with a reference dataset of population samples from Africa, Europe, and the Americas, including other populations descended from enslaved Africans in Cabo Verde and the Americas, notably African-Americans in southwest USA (ASW) and the African-Caribbeans in Barbados (ACB) (Supplementary Table S 2.3) (Auton et al. 2015; Gurdasani et al. 2015; Patin et al. 2017; Semo et al. 2020; Laurent et al. 2023).

Figure 2.3 A shows the sampling locations of the 70 populations included in the reference dataset. I grouped these populations into 10 distinct geographical regions, focusing particularly on regions in Africa that were used as key ports of embarkation during the Trans-Atlantic Slave Trade (Eltis and Richardson 2010). Figure 2.2B shows the first two dimensions of a Multi-Dimensional Scaling analysis calculated on Allele-Sharing Dissimilarities (ASD) (Bowcock et al. 1994), between each pair of individuals in the final merged dataset including 411,121 autosomal SNPs. The first axis of variation is explained by major genetic differentiation between African and European populations. Instead, the second axis captures genetic differentiation within Africa, revealing a north-south gradient ranging from West Western African to South African populations.

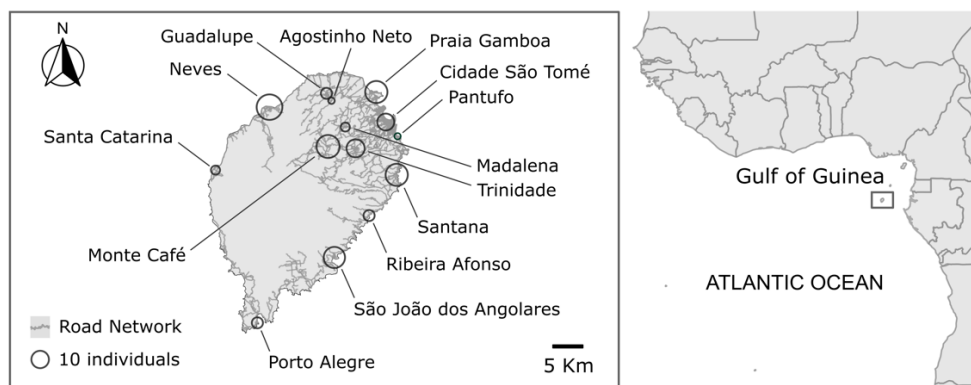


Figure 2.2. Map of São Tomé showing the position of 14 sampling sites.

Sampling sites are concentrated in the most populated areas where the road network (shown in grey) is denser. Each circle represents a sampling site, with its size proportional to the sample size. The map on the right shows the location of São Tomé in the Gulf of Guinea.

Populations descended from enslaved Africans exhibit genetic diversity patterns typical of admixed populations, falling between Africans and European population samples in the ASD-MDS analysis: a) Cabo Verdeans fall along a trajectory going from European to West Western African populations, reflecting the impact of slave recruitment from the neighbour areas of Senegambia and Guinea-Bissau in their current genetic profile (Beleza et al. 2012; Laurent et al. 2023); b) African-Americans (ASW) and African-Barbadians (ACB) cluster along a trajectory going from Central Western African to European populations, in agreement with a different history of recruitment of slaves mainly in the Gulf of Guinea (Martin et al. 2017; Laurent et al. 2023); c) São Toméans are highly dispersed across the two first dimensions of the ASD-MDS, displaying a relatively complex pattern. Some individuals cluster close to African populations, between East and West Southern Africans, while others cluster with Cabo Verdeans. The remaining São Toméans lie between African, Cabo Verdean and European populations.

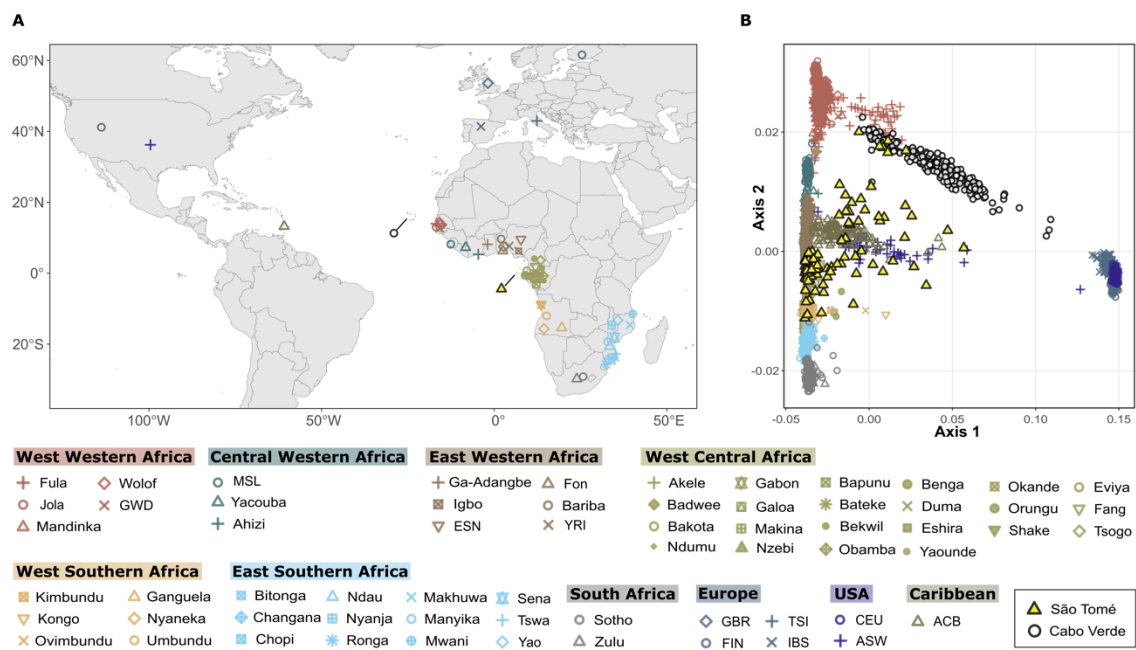


Figure 2.3. The genetic differentiation of the worldwide population samples.

(A) Geographical locations of the 70 population samples considered in the present study. (B) MDS projection of pairwise allele sharing dissimilarities (ASD) among São Toméans, Cabo Verdeans and other African, American and European populations. The first two axes of variation in the MDS were calculated among 3203 individuals using 411,121 autosomal SNPs. On the bottom, the color-coded legend sorts the populations into 10 distinct geographical regions in continental Africa, Europe and the Americas.

I further applied an unsupervised clustering algorithm, ADMIXTURE (Alexander et al. 2009), to visualize several axes of inter-individual genetic variation (Figure 2.4). The ADMIXTURE results reflect the patterns of genetic diversity observed in the ASD-MDS, while providing further information on genetic differentiation within Cabo Verde and São Tomé at higher levels of K .

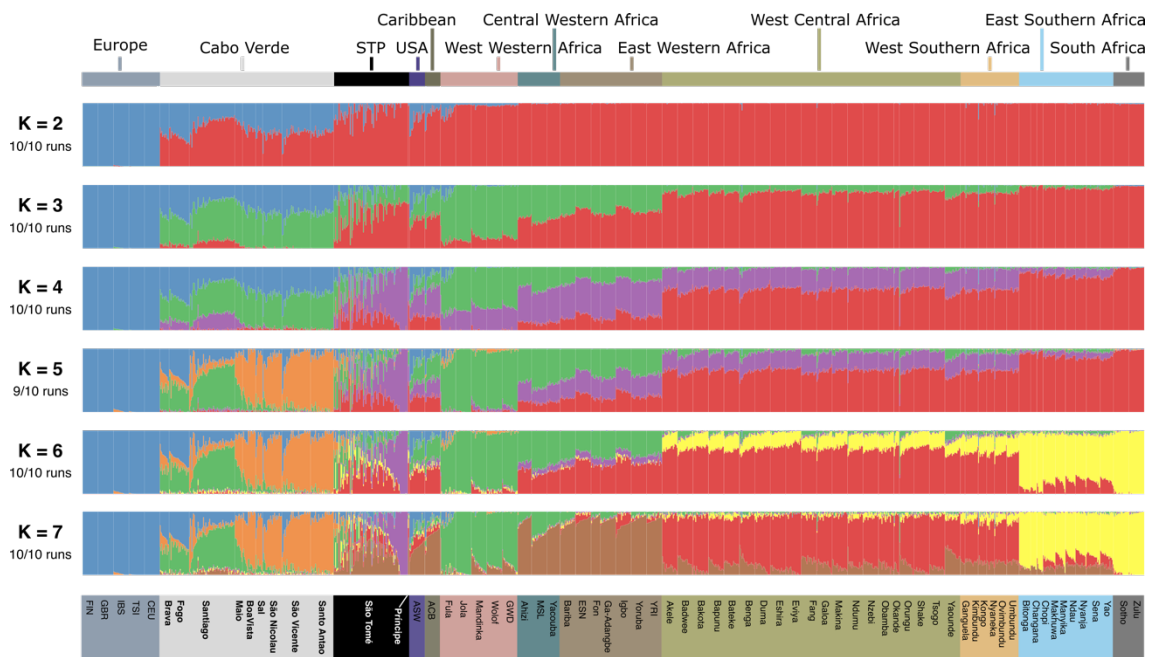


Figure 2.4. Unsupervised ADMIXTURE analysis.

Unsupervised ADMIXTURE analysis using 110,499 LD-pruned ($r^2 < 0.1$) autosomal SNPs from 1347 unrelated individuals, including 96 individuals that reported to be born in São Tomé, 2 individuals born in Príncipe, 225 individuals born on seven different islands in Cabo Verde, and a random sample of 20 individuals for each of the remaining population samples (Supplementary Table S 2.4). The proportion of independent ADMIXTURE runs closely resembling one-another is indicated on the left of each panel below each K value. Population samples are ordered by geographical regions according to the colour code indicated in Figure 2.

To better understand the ADMIXTURE results, it is recommended to interpret the genetic clustering at increasing levels of K (Lawson et al. 2012):

- At $K=2$, the clustering is explained by differences between African and European populations, maximizing the red and blue clusters, respectively. This clustering mirrors the results of the first two dimensions of the MDS analysis based on ASD.
- At $K=3$, the red cluster is partially replaced by the green cluster, which is present in the highest proportions in West Western Africans and Cabo Verdeans, without ever being close to 100% in any individual, thus indicating unresolved clustering at this value of K .
- At $K=4$, the new violet cluster is present in many Africans and enslaved African descendants, but is maximized close to 100% only in nine individuals born in São Tomé.
- At $K=5$, the new orange cluster is maximized in individuals born on the northern Cabo Verdean islands of Santo Antão, São Vicente, and São Nicolau. Interestingly, some individuals from São Tomé also present this orange cluster, as well as the green cluster, which is maximized in West Western African populations at this value of K .
- At $K=6$, the new yellow cluster is maximized in Sotho and Zulu populations from South Africa, and is also present in decreasing proportions in East Southern African, West Southern African, and West Central African populations.
- At $K=7$, West Western African populations maximize the green cluster; East Western African populations maximize the brown cluster; West Bantu-speaking populations from West Central and West Southern Africa present the highest proportions of the red cluster, while East Bantu-speaking groups from East Southern and South Africa maximize the yellow cluster.

Overall, genetic clusters at $K=7$ discriminate among different areas of slave recruitment in Africa, revealing genetic similarities between populations currently living in Africa and populations descended from enslaved Africans. African-Barbadians (ACB) and African-Americans (ASW) display relatively simple patterns with resemblance mainly to Central and East Western Africa. Cabo Verde and São Tomé exhibit more complex genetic structures that cannot be simply explained by the diverse African origins of their enslaved settlers.

In Cabo Verde, a) the southern islands of Fogo, Santiago, and Brava show high proportions of the green cluster found in West Western Africa. Instead, b) the northern islands of Santo Antão, São Vicente and São Nicolau maximize the orange cluster, which is found predominantly there. The orange cluster might represent the genetic differentiation of these populations due to high genetic drift (Beleza et al. 2012; Korunes et al. 2022; Laurent et al. 2023).

In São Tomé, a) some individual profiles are close to that of Cabo Verdeans, while b) others present varying proportions of clusters found in East Western, West Central and West Southern Africans. Finally, c) another group of individuals has a unique genetic pattern, represented by the violet cluster, which is not observed outside the island.

I examined the distribution of genetic clustering results across sampling sites in São Tomé (see Fig. 2.5), focusing specifically on the proportion of genetic diversity assigned to distinct clusters among all individuals collected at each site. While individuals from the same sampling site often exhibited different proportions of genetic clusters, multiple clusters were generally present across most locations. At some sites, certain clusters were more dominant than others. Notably, only two sampling sites, each with just one or two individuals, were characterised by a single genetic cluster: the violet cluster. This violet cluster tended to be more prevalent in coastal areas. In contrast, the red cluster was most prevalent in the interior Monte Café site, although it was also present at almost all other locations. Overall, the observed genetic clustering does not appear to correspond directly to the geographic location of the sample, suggesting that additional analyses are required to determine the factors underlying genetic diversity across the island.

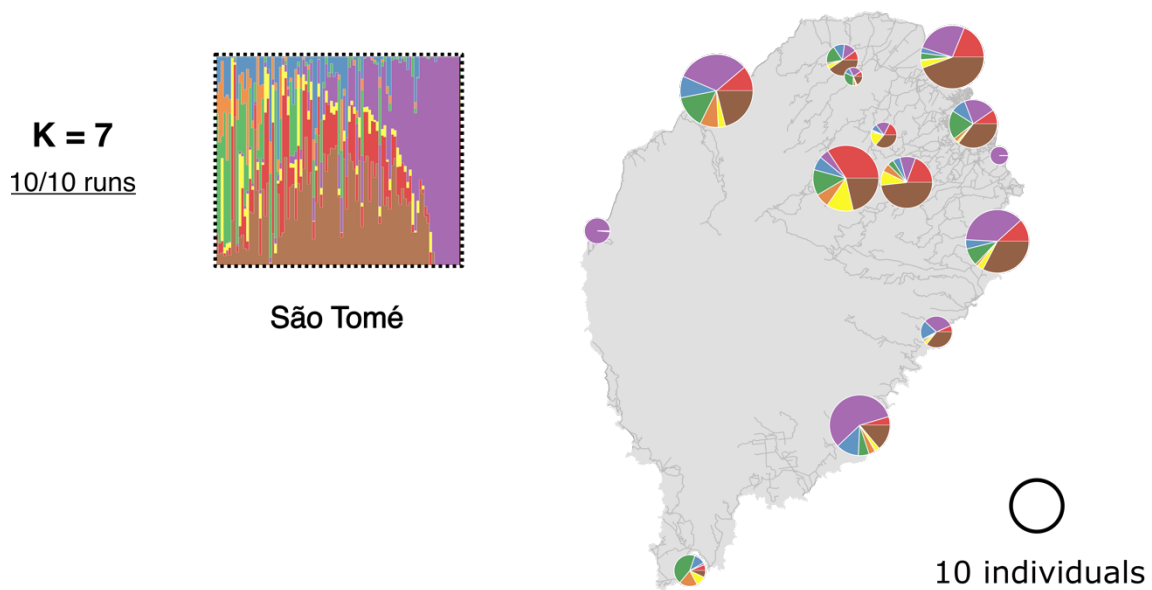


Figure 2.5 Genetic clustering across São Tomé.

The ADMIXTURE result at $K=7$ on the left is a zoom into Figure 2.4. The map on the right shows the geographic locations of 14 sampling sites, as in Figure 2.2, with pie charts representing the proportion of genetic diversity assigned to distinct genetic clusters across all individuals collected in each sampling site. Pie chart sizes are proportional to the number of individuals sampled at each site, and colors correspond to the seven genetic clusters inferred from ADMIXTURE analysis at $K=7$ using 110,499 LD-pruned ($r^2 < 0.1$) autosomal SNPs from 1347 unrelated individuals from Africa, Europe and the Americas (Figure 2.4).

I examined the positioning of individuals with the highest proportions of specific genetic clusters in the ASD-based MDS analysis. Specifically, I focused on the first ten individuals with the highest proportions of a genetic cluster. I only considered four genetic clusters due to difficulties in classifying individuals in distinct groups, as for instance many individuals with high proportions of the blue, orange and yellow clusters also exhibited high proportions of other clusters.

- Top Green individuals (those with the highest proportions of the green cluster) clustered with Cabo Verdeans or were positioned between Cabo Verdeans and other São Toméans. The ADMIXTURE results at $K=7$ show that the green cluster is maximised in West Western Africa and present at high proportions in individuals from the southern islands of Cabo Verde, as well as at varying levels in São Tomé.

- Top Brown individuals clustered near populations from eastern West Africa, where the brown cluster is maximised.
- Top Red individuals grouped closer to populations from South Western Africa, consistent with relatively high red cluster proportions in that region.
- Top Violet individuals were positioned between these two groups, and in general close to African populations.

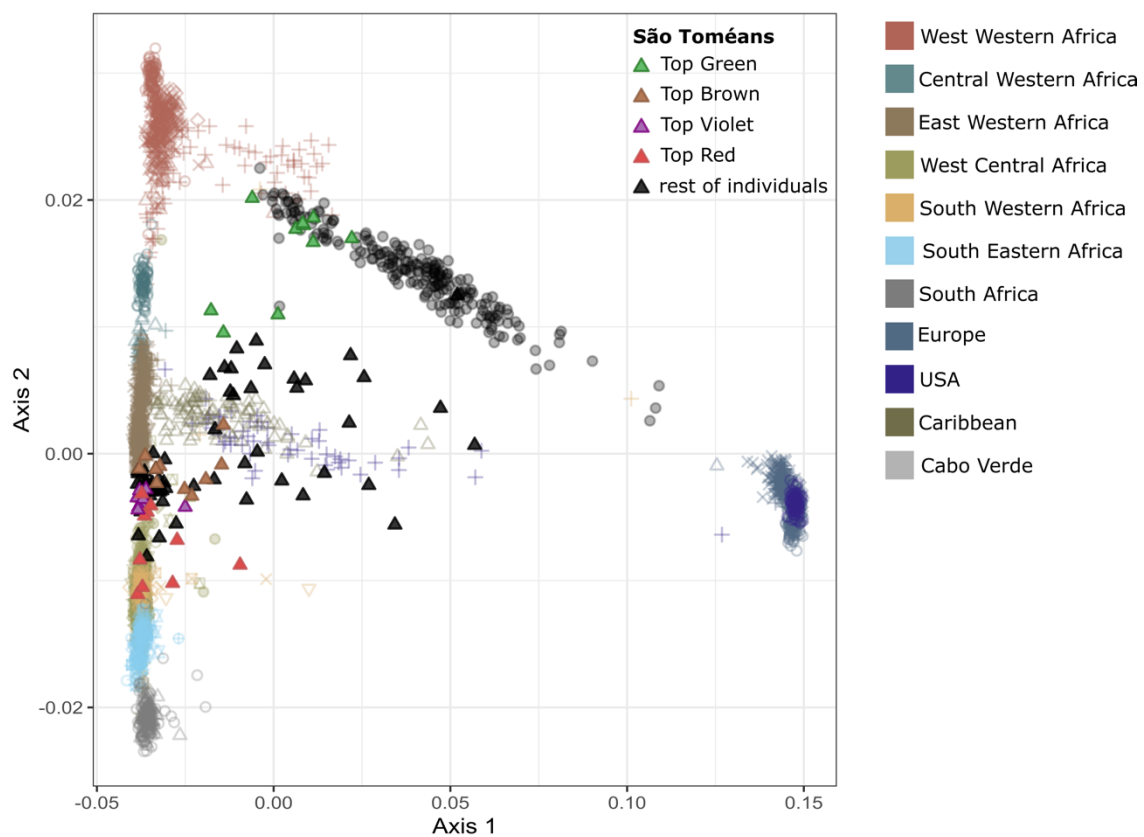


Figure 2.6 Genetic clustering in the worldwide context.

MDS analysis based on allele sharing distance (ASD), using the same dataset as in Figure X. São Toméan individuals are represented as triangles, coloured according to five categories. Specifically, the top 10 individuals with the highest proportions of the green, brown, red, and violet clusters, considering ADMIXTURE results at $K = 7$, are labelled as *Top Green*, *Top Brown*, *Top Red*, and *Top Violet*, respectively. All remaining individuals are shown in black.

2.2.2 Haplotype-based population structure within São Tomé

I analysed patterns of shared haplotypes between all pairs of individuals in São Tomé using the pipeline ChromoPainter2-fineSTRUCTURE (Lawson et al. 2012). I identified five different clusters of individuals (C1-C5) with high levels of inter-individual haplotypic resemblance between each other, further considering the number and length of shared haplotypes across the island (Figure 2.7).

Fig. 2.7A shows the dendrogram presenting the phylogenetic relationships between the five clusters, as well as their distribution across the sampling sites in São Tomé. Individuals sampled in the same locations are almost always assigned to more than one genetic cluster. C1 and C2 individuals are mostly found in coastal sampling sites, C3 individuals are mostly found in northern sampling sites, C4 individuals are found throughout the island, and C5 individuals are mostly found in the eastern part of the island.

Figure 2.7 B shows a PCA calculated on the same co-ancestry matrix used to infer the dendrogram. The first PC separate individuals of clusters C1 and C4, while the second PC separates individuals of cluster C3. Within this triangle, individuals of cluster C2 lay in between C1 and C3 along the second PC, and individuals of cluster C5 lay in between C4 and the rest of São Tomé individuals along the first PC.

Figure 2.7 C and D shown heatmaps representing the number and length of shared haplotypes, respectively, between each pair of São Toméans, which are ordered by genetic clusters. Individuals of cluster C1 mostly share haplotypes with each other, both in terms of total number and cumulative length, and copy relatively few haplotypes from other individuals on the island. On the contrary, individuals of cluster C2 copy relatively long haplotypes from C1 individuals, as well as several haplotypes from the rest of São Toméan individuals with varying cumulative lengths. Individuals of cluster C3 copy relatively more haplotypes from themselves than from the rest of the island, while individuals of clusters C4 and C5 differ markedly from other clusters, sharing the highest relative number of haplotypes.

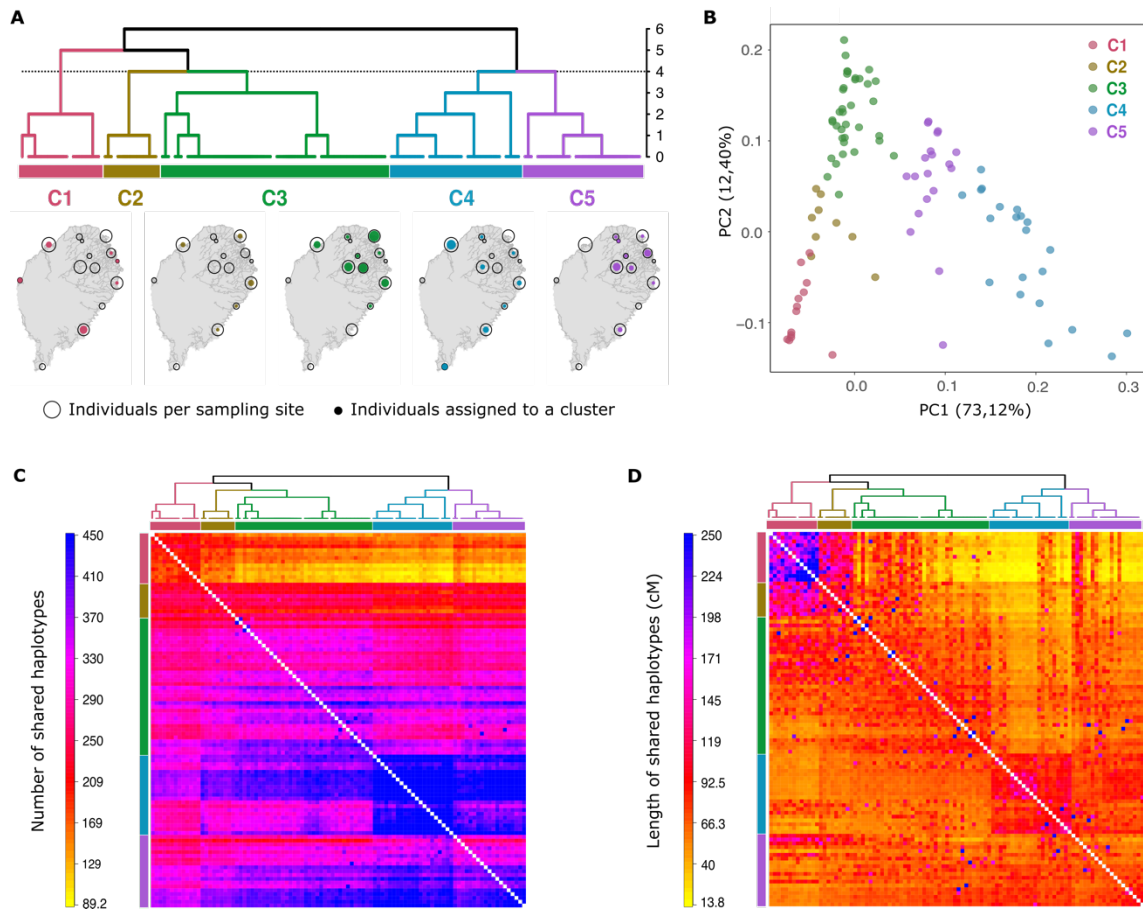


Figure 2.7. Genetic clustering and population structure in São Tomé.

(A) The 96 individuals from São Tomé grouped into five genetic clusters using fineSTRUCTURE, based on shared haplotype segments. The vertical axis of the dendrogram represents the distance or dissimilarity between clusters. The distribution of individuals across sampling sites is shown below. Black circles indicate the number of individuals sampled at each site, while coloured dots represent the number of individuals per cluster. (B) Principal Component Analysis (PCA) based on the co-ancestry matrix in (C), with individual labels color-coded by genetic clusters in (A). (C) Co-ancestry matrix showing the number of haplotypes each individual (row) copies from any other (column). (D) Similar to (C), but based on the cumulative length (in cM) of copied haplotypes.

2.2.3 Haplotypic ancestry of genetic clusters

I modelled the haplotypic ancestry of each São Toméan cluster as a mixture of the possible source populations included in the reference dataset using the Chromopainter2-SOURCEFIND pipeline (Chacón-Duque et al. 2018). I inferred a maximum of eight source populations per “Target” population sample using SOURCEFIND.

First, I considered each one of the admixed populations descended from enslaved-Africans as a separate “Target” population, and all other populations from continental Africa and Europe as “Donors” (Figure 2.8). In agreement with previous studies (Beleza et al. 2012; Patin et al. 2017; Ongaro et al. 2019; Micheletti et al. 2020; Laurent et al. 2023), I found that:

- Cabo Verdeans resulted from admixture between Europeans (from 27% in Santiago to 51% in Fogo) and Africans from West Western Africa (from 46% in Fogo to 69% in Santiago);
- African-Barbadians (ACB), and African-Americans (ASW) resulted from admixture between Europeans (15% in ACB, 24% in ASW) and Africans from Central and East Western Africa (49% in ASW, 82% in ACB), West Southern Africa (3% in ACB, 18% in ASW), and West Western Africa (9% in ASW only);
- São Toméans displayed a more complex ancestry profile, including contributions from Europe (9%), West Western Africa (16%), East Western Africa (37%), West Southern Africa (36%), as well as East Southern Africa (2%). Importantly, when Cabo Verdean populations were considered as possible “Donors” (Figure 2.8B), the haplotypic ancestry from the island of Santiago in Cabo Verde (15%) replaced the Mandinka GWD haplotypic ancestry and part of the Iberian IBS component, previously identified in São Tomé (Figure 2.8A).

In Figure 2.8C, each of the five São Toméan genetic clusters (C1 to C5), was considered as a separate “Target” population and Cabo Verdean individuals from seven different islands were also included as seven separate “Donor” populations. Interestingly, clusters corresponding to the most basal bifurcation of the fineSTRUCTURE dendrogram (C1-C3 vs. C4-C5, Figure 2.7A) showed different ancestry profiles:

- Clusters C1, C2 and C3 shared haplotypic ancestry with Kongo-speakers from Angola, in West Southern Africa (from 39% in C1 to 48% in C3), and the Ga-Adangbe from Ghana and the Igbo from Nigeria, in East Western Africa (from 49% in C1 to 51% in C2);
- Clusters C4 and C5 shared haplotypic ancestry with Cabo Verdeans, mostly from the island of Santiago (30% in C5, 49% in C4). Moreover, I found evidence of relatively low levels of shared haplotypic ancestry with the Sena and Manyika population from Mozambique, in East Southern Africa (2% and 8%, respectively).

Overall, these results suggests that the major division between the São Toméan clusters (C1-C3 vs. C4-C5) is due to variable contributions from different source populations, while further splits occurring within each major division may be related to subsequent episodes of *in situ* isolation and gene flow.

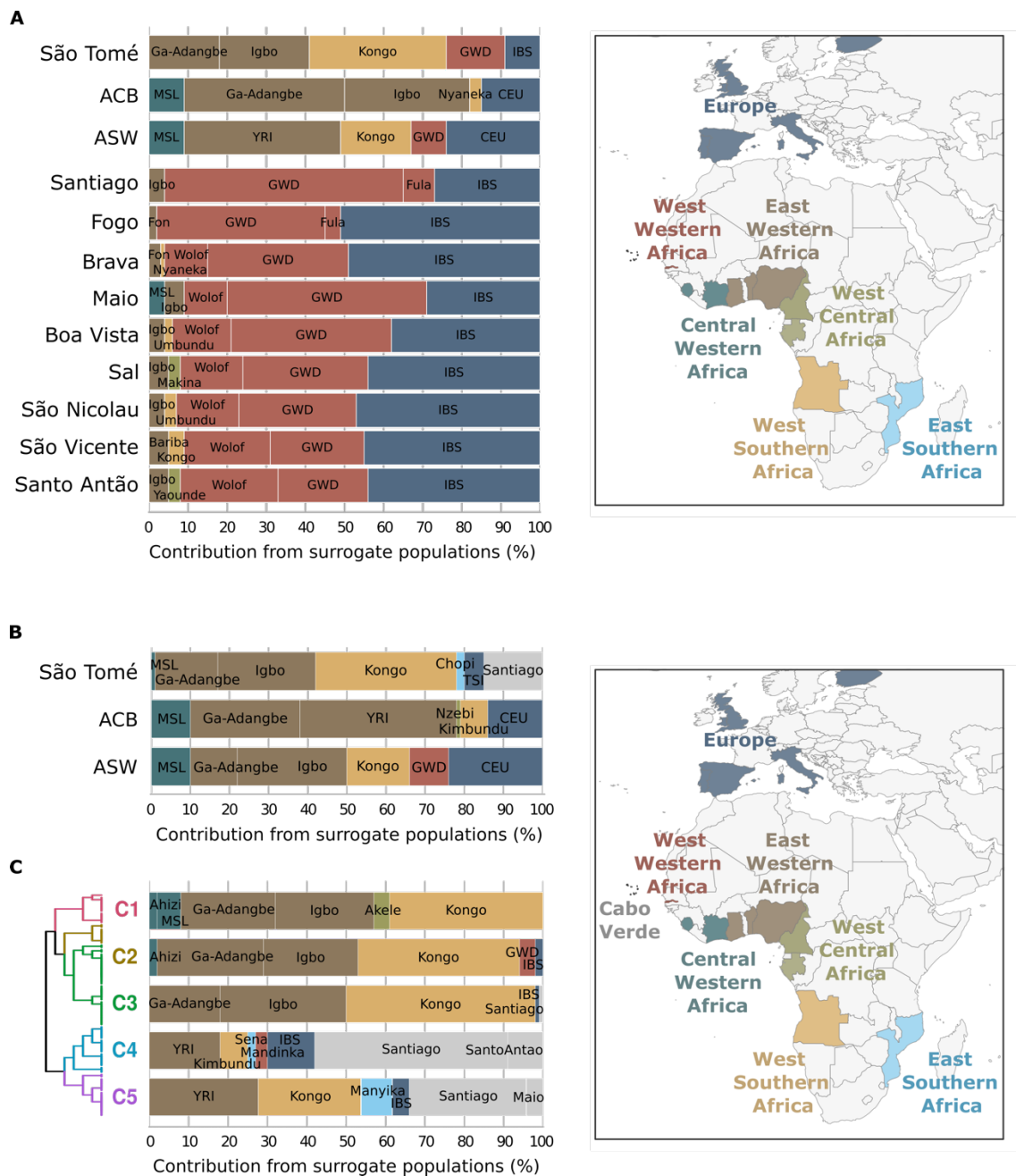


Figure 2.8. Haplotypic ancestry of São Toméans and other populations descended from enslaved-Africans.

On the left, each bar plot represents a distinct Target population, whose haplotypic ancestry is modelled using SOURCEFIND as a mosaic of Donor populations shown on the right. Three panels are presented: in (A) Cabo Verdeans are not included as Donors, while in (B) and (C) Cabo Verdeans are considered as potential sources of admixture. The target populations are São Toméans (São Tomé), Barbadians (ACB), African Americans (ASW), Cabo Verdeans (from Santiago to Santo Antão), and São Toméan genetic clusters (from C1 to C5). The Donor populations are color-coded by geographic regions in Africa and Europe, as indicated in the legend at the right.

2.2.4 Long IBD-tracts shared among São Toméans and Cabo Verdeans

I explored tracts that are identical by descent (IBD), shared among the five genetic clusters in São Tomé, and with each of the island populations of Cabo Verde (Figure 2.9A). In particular, I focused on IBD tracts longer than 18 cM, representing approximately the last eight generations of recombination between individuals (Baharian et al. 2016). I observed differences in IBD sharing patterns between and within each São Toméan group. Individuals of cluster C1 shared the highest total length of long-IBD tracts between each other, and with individuals of cluster C3. Individuals of cluster C2 shared relatively low levels of long IBD tracts with each other, while sharing long IBD tracts with both clusters C1 and C3. Individuals of clusters C4 and C5 were the only ones to share long IBD tracts with Cabo Verdean populations, in accordance with the SOURCEFIND results (Figure 2.8). However, cluster C5 showed higher values of total length of long IBD tracts shared with clusters C1 and C3 in São Tomé, when compared to cluster C4.

2.2.5 Runs of Homozygosity

I further investigated the presence of Runs of Homozygosity (ROH), which are stretches of DNA where an individual genome is homozygous across all sites. They arise when an individual inherits IBD segments from a recent common ancestor. ROH of different lengths are usually generated by different demographic processes: medium ROH can be generated by recent demographic events such as bottlenecks, while long ROH are usually due to recent parental relatedness (Pemberton et al. 2012; Szpiech et al. 2017; Szpiech et al. 2019).

I inferred and classed ROH in long, medium and short length classes using GARLIC (Szpiech et al. 2017). In Figure 2.9B, I reported the distribution of the sum of total length of ROH calculated on individual genomes, for each genetic cluster in São Tomé separately, and for other populations in Africa and Europe. Cluster C1 presented a total length of long ROH higher than any other population considered in the analysis. The sum of total length of long ROH in C2 individuals was intermediate between those of clusters C1 and C3 in São Tomé. Clusters C4 and C5 presented similar mean levels of medium and long ROH as the Santiago population in Cabo Verde.

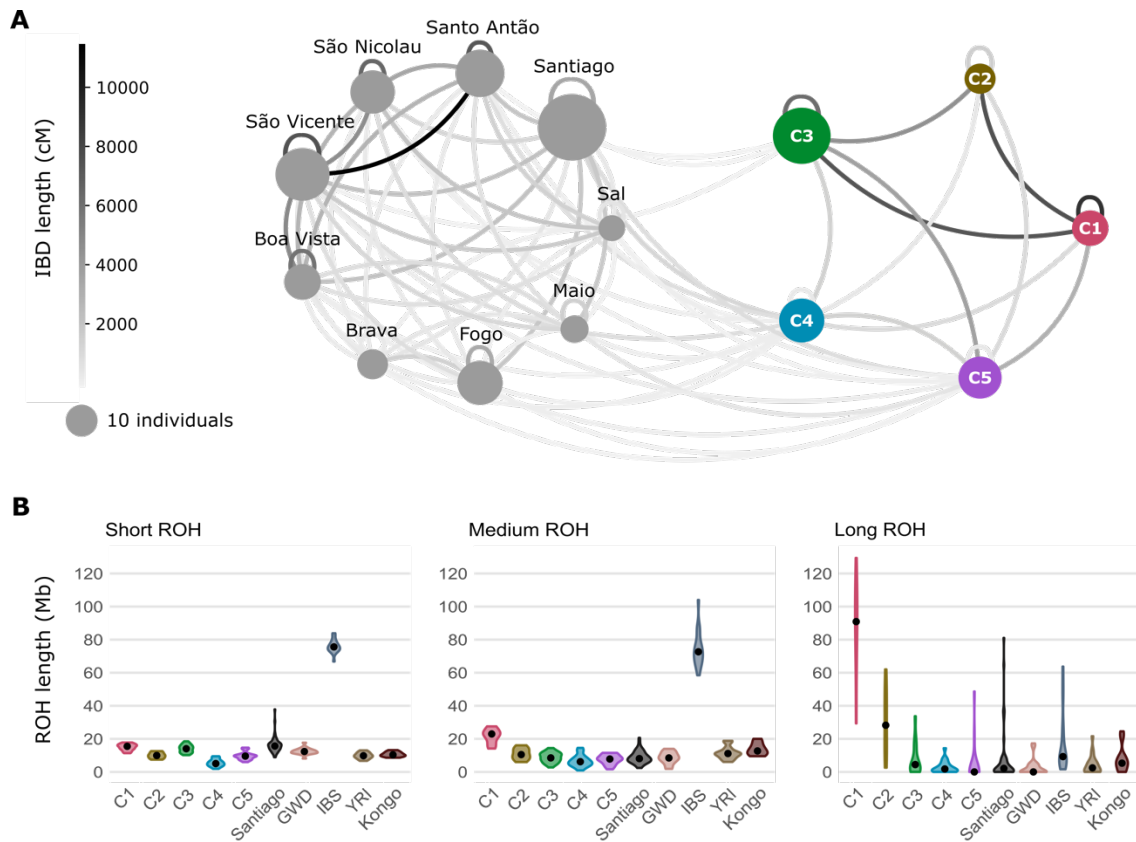


Figure 2.9. IBD tracts sharing and ROH analysis.

(A) Network showing the cumulative length of Identical-by-Descent (IBD) tracts longer than 18 cM shared within and between population samples from Cabo Verde (left) and São Tomé (right). The line colour indicates the total length of long IBD shared (cM) and the point size is proportional to sample size. (B) Violin plots of the total length of Runs of Homozygosity (ROH) per individual genome across clusters in São Tomé, the Santiago population in Cabo Verde, the Mandinka from Gambia (GWD), the Iberians from Spain (IBS), the Yoruba from Nigeria (YRI), and the Kongo from Angola.

2.2.6 Dating admixture events for each São Toméan cluster separately

I inferred the timings of major admixture events for each of the five São Toméan genetic clusters separately (Figure 2.10). I used the algorithm implemented in fastGLOBETROTTER (Wangkumhang et al. 2022), which analyse recombination distances between ancestry tracts. I specified a separate set of “Surrogate” populations for each São Toméan genetic cluster – the same populations that were identified by SOURCEFIND as the “Donors” most likely involved in the admixture events of each cluster (Figure 2.8C). The fastGLOBETROTTER algorithm utilizes a mixture model for the “Surrogate” populations and infers the relative contributions of each “Surrogate” population to two alternative mixtures, representing the first and second admixing sources for each admixture event.

Overall, fastGLOBETROTTER identified strong signals of admixture in all the five São Toméan clusters. It also dated each admixture event by considering a simplified scenario that assumes one major pulse of admixture - an important caveat that need to be kept in mind for the interpretation of the following results (See Materials and Methods and Discussion).

- For clusters C1, C2, and C3, the major admixture event primarily involved populations from East Western Africa, specifically the Igbo and Ga-Adangbe, on one side, and the Kongo population from West Southern Africa on the other. The dates of admixture events, considering a single pulse of admixture, ranged from 9 to 11 generations ago, and their relative Confidence Intervals were largely overlapping. For cluster C3, the single pulse of admixture inferred 11 generations before present would correspond to the year 1694 (CI: 1636-1769) when considering a generation time of 30 years. Moreover, admixture was estimated to approximately 9 generations before present for cluster C1, corresponding to the year 1763 (CI: 1654-1824), and similarly for the C2 cluster, around 1748 (CI: 1683-1864).
- Clusters C4 and C5 resulted from varying levels of admixture between Cabo Verdeans and African populations that were inferred as sources of admixture of other São Toméan clusters. In particular, the mixture of East Western Africans and West Southern Africans may represent the rest of the São Toméan populations. Admixture events were inferred at approximately 9 generations before present, around 1760 (CI: 1736-1795) for cluster C4, and 1751 (CI: 1714-1790) for cluster C5.

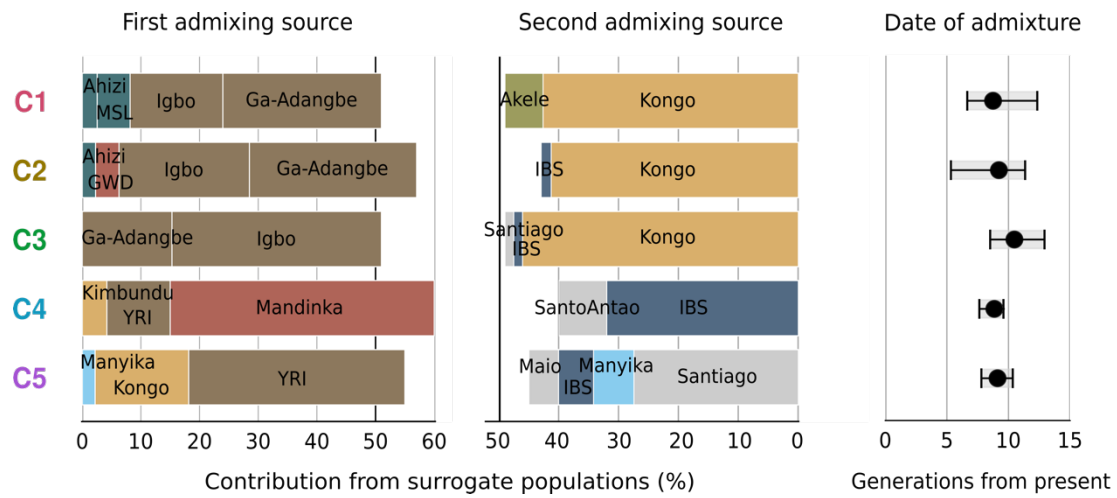


Figure 2.10. Admixing sources and timing of admixture events.

For each São Toméan cluster, the best-supported single-pulse admixture event from fastGLOBETROTTER is shown. The surrogate populations considered here are those inferred by SOURCEFIND as the best proxies for sources among the populations considered in this study (Figure 2.8C). (A, B) show the relative contributions of surrogates to the first and second sources, totalling 100%. (C) shows admixture dates in generations, with 95% confidence intervals from 100 bootstraps.

2.3 Discussion

In this work, I reconstructed complex genetic admixture processes on the island of São Tomé, the earliest European colony in Equatorial Africa and one of the first examples of Plantation Economic system that was later deployed in numerous other European colonies across the Americas (Curtin 1990). My investigation, based on unprecedented number of markers and samples from São Tomé, revealed that the genetic structure of this admixed population is more complex than that of several other populations descended from enslaved Africans on both sides of the Atlantic, due to an intricate pattern of admixture and isolation events during and after the TAST.

To account for genetic structure, I identified five genetic clusters in São Tomé, based on the hypothesis that individuals sharing the most haplotypes across the island may have experienced similar genetic histories. Examining the distribution of individuals belonging to the five genetic clusters across 14 sampling sites on the island, I found that individuals from the same site were frequently assigned to different genetic clusters (Figure 2.7A). This finding suggests that geography does not fully explain genetic clustering in São Tomé. Instead, community boundaries are likely to have played a role in maintaining reproductive isolation between distinct social groups living in close proximity on the island.

Genetic clustering in São Tomé can be partially explained by admixture events between different African populations. I reconstructed the genetic ancestry of each São Toméan genetic cluster separately, using 70 population samples from Africa and Europe as proxies for the sources of admixture. The major division in the clustering results separates groups sharing similar proportions of haplotypic ancestries with East Western and West Southern African populations (C1-C3), from groups with a clear contribution from Cabo Verde (C4 and C5) (Figure 2.8C). This differentiation is in line with the historically known recruitment of enslaved Africans from the Gulf of Guinea and Congo-Angola for the peopling of São Tomé from the 16th to the 18th centuries, which was followed by the arrival of Cabo Verdean immigrants for indentured labour in the 20th century (Figure 2.1).

Genetic diversity and ancestry patterns in São Tomé have also been shaped by isolation and admixture events that happened between recently admixed groups on the island. I propose an interpretation of this complex history of admixture based on different types

of genetic evidence (Figure 2.11). In the following paragraphs, I discuss the relationships between the observed genetic clusters and identifiable social groups in São Tomé, considering the familial anthropology information available for the sample, as well as the historical, demographic, and linguistic information available on the island.

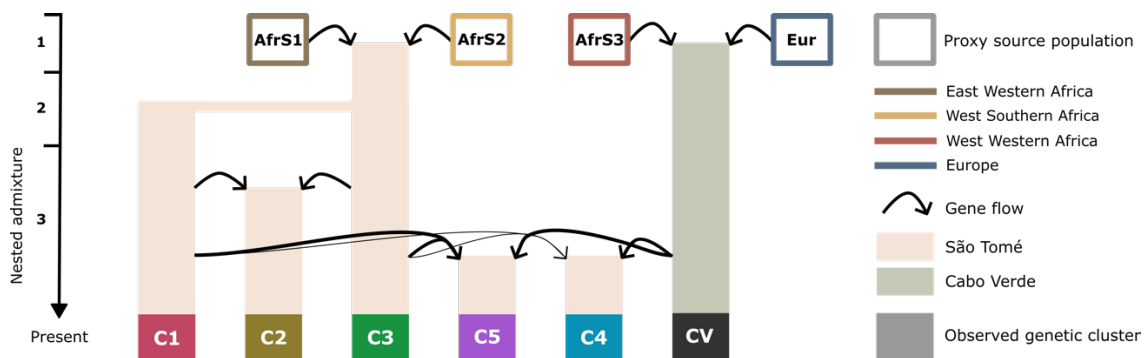


Figure 2.11. Nested admixture in São Tomé.

Demographic scenario explaining the observed genetic diversity in São Tomé and Cabo Verde. The rectangles at the bottom represent the five São Toméan genetic clusters (C1-C5), and the Cabo Verdean sample (CV). Bars represent genetic populations evolving over time, with colours marking the locations of admixture events. Proxies for source populations are shown at the top, coloured by African and European regions, and arrows indicate gene flow over different time periods. The timeline is divided into three epochs that could be linked to significant historical events; 1) the foundation of the São Toméan population in the 15th century; 2) the formation of isolated maroon communities during the first plantation economy from the 15th to the 17th century; 3) the changing historical and social contexts during two centuries of economic decline, through the second plantation economy in the 19th and 20th centuries, up to the present. In most cases, explaining the observed genetic diversity in São Tomé requires taking into account gene flow between admixed groups that were previously isolated - a process that I call “nested admixture”.

2.3.1 A population descended from freed slaves

Throughout the 17th and 18th centuries, a period of economic decline between the two plantation economies, the growing population of *Forros*, literally “freed slaves”, eventually became the dominant social group in São Tomé (Seibert 2015). They are historically associated to the Forro, a creole language with a syntax similar to Edo languages in Nigeria, and a lexicon mostly based on Portuguese, but also Edo and Kikongo, a Bantu language from Congo-Angola (Hagemeijer and Rocha 2019). The number of Forro creole speakers has considerably declined in recent decades. However, it remains the most widely spoken creole on the island (Gonçalves and Hagemeijer 2015), and it is spoken throughout São Tomé, particularly in the north-eastern areas and the capital city of São Tomé (Instituto Nacional de Estatística 2012).

Cluster C3 may represent the *Forros* of São Tomé. Most individuals belonging to the C3 cluster were sampled in the northeast of São Tomé (Figure 2.7A), and C3 individuals showed similar patterns of genetic diversity to those of Forro-speaking individuals analysed in previous studies (Almeida et al. 2021). In particular, they derive almost equal proportions of their genomic ancestry from populations currently living in Nigeria and Ghana in East Western Africa and Angola in West Southern Africa (Figure 2.5C). These haplotypic ancestries reflect the routes of the São Toméan slave trade with the Kingdom of Benin, the “Gold Coast”, and the Kingdom of Kongo in the 16th century.

While it is likely that the *Forros* result from admixture between Africans of different origins, with limited genetic contribution from Europeans, it is less clear when this admixture occurred. The Forro creole probably formed at the beginning of the colonization of the island, at the beginning of the 16th century, with Bantu features appearing slightly later than Edo features, according to linguistics. However, the mixing of East Western African and West Southern African genetic contributions in cluster C3 is inferred around 1694 (CI: 1636-1769) (Figure 2.10), substantially later than the mixing of linguistic features and the timing of arrival of slaves on the island. This is likely due to limitations of our dating approach. Indeed, fastGLOBETROTTER cannot always distinguish between discrete pulses and continuous admixture. The algorithm may infer a single admixture event instead of continuous if the period of recurring admixture is short, in which case the estimated date may fall within this period (Hellenthal et al. 2014). Moreover, the admixture model here assumes random mating, which is known to bias

admixture time estimates toward more recent dates (Zaitlen et al. 2017; Korunes et al. 2022).

2.3.2 One of the first communities of runaway slaves

Historical records report that enslaved Africans escaped plantations from the onset of Portuguese colonization in São Tomé. Maroon communities managed to control the mountainous interior of the island until occupation by colonial authorities in the late 19th century. The descendants of some of these communities thus became known as the *Angolares* (Seibert 2013). Today, the *Angolares* are mostly found in fishing communities scattered along the coast, with major settlements in the villages of São João dos Angolares and Santa Catarina.

Cluster C1 may represent the *Angolares* of São Tomé. C1 individuals were mostly sampled in coastal areas where *Angolares* are known to live (Figs 1 and 4A), and C1 individuals showed similar patterns of genetic diversity to those of Angolar-speaking individuals analysed in previous studies (Almeida et al. 2021).

Investigating the origins of the *Angolares* has been of interest to many disciplines. Colonial narratives have suggested that the *Angolares* descended from survivors of a shipwreck with enslaved Africans from Angola (Seibert 2013). However, historical, linguistic, and genetic studies have proved this hypothesis unrealistic (Hagemeijer and Rocha 2019; Almeida et al. 2021). Although the Angolar creole has a stronger lexical influence from the Bantu language Kimbundu, currently spoken in northern Angola, it shares many linguistic features with the Forro creole. It therefore seems likely that the two languages descend from the same proto-creole in the Gulf of Guinea, which subsequently accumulated differences in the amount and type of Bantu features incorporated into the ancestral creole (Hagemeijer 2011). My findings align with this hypothesis, as the *Forros* and *Angolares* of the C3 and C1 clusters, respectively, both derived the most of their genetic ancestry from the Gulf of Guinea (East Western Africa, 50% and 49%) and Bantu-speaking regions in Congo-Angola (West Southern Africa, 48% and 39%).

Despite this similarity, C1 individuals exhibited markedly distinct genetic patterns. They maximized the unique ADMIXTURE violet cluster (Figure 2.3), shared very long haplotypes (Figure 2.7D) and long IBD tracts (Figure 2.6A), and presented particularly

high levels of long ROH (Figure 2.6B). These features are typical of isolated populations, as may be explained by high levels of inbreeding and genetic drift due to low effective population size (Ceballos et al., 2018). Similar genetic signals have been observed in admixed population isolates from South America (Mooney et al. 2018), and in other populations descended from maroon communities, such as the Noir Marron from Suriname and French Guyana (Fortes-Lima et al. 2017).

Genetic drift can affect admixture dating, as it influences recombination distances between ancestry tracts (Pool and Nielsen 2009; Gravel 2012). For the C1 cluster, a single pulse of admixture is estimated to have occurred approximately nine generations ago, around the year 1763 (confidence interval: 1654–1824), slightly earlier than that inferred for the C3 cluster. However, the broad overlap in their confidence intervals suggests that both clusters likely share a similar founding admixture event.

The *Angolares*, who share a significant portion of their genetic ancestry with the *Forros*, thus likely became differentiated through isolation and inbreeding after running away from the plantations (Figure 2.11). According to the interpretation of previous studies (Almeida et al. 2021), additional contributions from runaway slaves from Angola may explain the impact of the Kimbundu language that distinguishes the Angolar and Forro creole languages.

2.3.3 Cabo Verdean migration to São Tomé

The establishment of coffee and cacao plantations in the 19th century marked a profound demographic shift in São Tomé. The number of indentured laborers represented approximately half of the population of the archipelago of São Tomé e Príncipe during the first half of the 20th century (Nascimento 2002).

Cluster C4 likely represents the Cabo Verdean *serviçais* and their descendants in São Tomé. The individuals of the C4 cluster have mostly been sampled in sites close to historical *roça* plantations, such as Agostinho Neto, Monte Café, São João dos Angolares, and Porto Alegre, which hosted numerous Cabo Verdean *serviçais* to work in plantations. Moreover, they include the five São Toméan individuals that reported in the interviews that one or both their parents were born in Cabo Verde. In terms of haplotypic ancestry, the C4 cluster is more similar to Cabo Verdeans than to the rest of the São Toméan samples (Figure 2.8).

My findings suggest that the Cabo Verdean *serviçais* who stayed in São Tomé after their contracts ended, and left descendants on the island, came from the islands of Santiago and Santo Antão, in the south and the north of the Cabo Verde archipelago, respectively (Figure 2.8 D). However, the C4 cluster also shows recent common ancestry with admixed populations from other islands in Cabo Verde (Figure 2.9A). Moreover, Cabo Verdeans from different islands share many long-IBD tracts with each other, and particularly with individuals from Santiago, the first island to be settled in the 15th century that played a key role in the founder effects shaping the genetic diversity of the other islands (Laurent et al. 2023). This may explain why SOURCEFIND, limited to using a maximum of eight surrogate populations, identified Santiago as a major contributor to Cabo Verdean haplotypic ancestry in São Tomé, potentially underestimating contributions from other islands (Figure 2.8).

The main admixture event of cluster C4 was inferred approximately 9 generations before present (circa 1760, CI: 1736-1795, assuming 30 years per generation; Figure 2.10) between West Western African and European populations. This likely reflects the founding of the admixed gene pool in Cabo Verde. However, this date is more recent than expected for the settlement of the archipelago (Laurent et al. 2023), which may reflect the limitations of the admixture model.

The accuracy of admixture dating depends on how well the model captures the complexity of the underlying demographic history and on the extent to which the surrogate populations represent the true sources. The simplified admixture model used here, which assumes a single admixture pulse, does not account for subsequent, albeit minor, admixture between Cabo Verdean and São Toméan individuals. In this model, the first admixture source is mostly formed by West Western African contributions, but also includes a mixture of East Western and West Southern African ancestry representing gene flow from local São Toméan sources. The second source is mostly formed by European ancestry, but also includes haplotypic ancestry from admixed Cabo Verdean genomes (Figure 2.10). These limitations should be taken into account when interpreting dating results.

2.3.4 Limited European genetic contribution

I found relatively little haplotypic ancestry from European source populations in São Tomé, compared to numerous other enslaved-African descendant populations in the

Americas and particularly compared to Cabo Verde (Ongaro et al. 2019; Fortes-Lima and Verdu 2020; Laurent et al. 2023).

Historically, the African population in São Tomé, including both enslaved and free individuals, far outnumbered the European presence on the island (Henriques 2000), and to a larger extent than in Cabo Verde (Albuquerque and Santos 1991). However, admixture was generally tolerated at the beginning of settlement of both colonies as part of Portugal's colonization efforts (Caldeira 2007). The European genetic relative contribution may thus have been diminishing over time, as a substantial part of the Portuguese settlers left the island in the 18th century due to the economic decline between the two main Plantation Economies (Lucas 2015).

Interestingly, almost half of the European ancestry tracts in São Tomé actually originate from Cabo Verdean admixed genomes. If Cabo Verdeans are not considered as sources of admixture, the European haplotypic ancestry in São Tomé is overestimated (compare Fig. 2.8A and B). When I considered the São Toméan clusters separately (Fig. 2.8C), I found a low but detectable level of European genetic contribution in all clusters except for the *Angolares* of cluster C1, where it was absent. Higher levels were found in clusters with Cabo Verdean haplotypic ancestry. This may suggest that, in the haplotype-sharing analyses that I used to define clusters in São Tomé (Figure 2.7), individuals with higher European descent may have clustered preferentially with those of Cabo Verdean descent.

2.3.5 Nested admixture between recently admixed populations

In the following paragraphs, I focus on the two remaining clusters, C2 and C5, which may result from relatively recent gene flow between the three social groups previously identified: the *Forros*, descendants of freed slaves, represented by cluster C3; the *Angolares*, descendants of runaway slaves, represented by cluster C1; and the Cabo Verdean *serviçais* and their descendants, represented by cluster C4.

2.3.5.1 Admixture between *Angolares* and *Forros*

Cluster C2 may result from admixture between the *Angolares* and the freed population of *Forros*, after the maroon communities begun to establish economic, social and cultural connections with the rest of the island population in the 19th century. Indeed, the nine individuals of the C2 cluster share numerous genetic features with both the C1 and the

C3 cluster, representing the *Angolares* and the *Forros*, respectively. In particular, they share many haplotypes and many long IBD-tracts with C1 and C3 (Figure 2.7 and 2.9A), while showing intermediate levels of medium and long ROH (Figure 2.9B).

Previous studies on populations descended from maroon communities have emphasized strong genetic isolation and differentiation, consistent with historical accounts of enslaved individuals escaping bondage and establishing refuge in remote, inaccessible areas (Fortes-Lima et al. 2017; Almeida et al. 2021). However, despite the relative isolation of these communities, admixture with surrounding populations may happen over time. My findings show that maroon communities also result from complex admixture processes, prior and after their foundation. This stresses the importance of reconsidering genetic expectations for descendants of “isolated” populations in the context of the TAST and admixture in general.

2.3.5.2 Admixture between Cabo Verdeans and São Toméans

Cluster C5 results from relatively high gene flow between Cabo Verdeans and the rest of the local São Toméan population. I found evidence of varying degrees of recent common ancestry between the C4 and C5 clusters, of Cabo Verdean descent, and the C3 and C1 clusters, representing the *Forros* and *Angolares*, respectively (Figure 2.6A).

The second plantation economy in São Tomé is marked by the emergence of new forms of discrimination and social stratification on the island (Bouchard 2023). After the abolition of slavery, the local São Toméan population refused to engage in plantation work, and immigrant indentured workers were largely confined to the *roças*. The low São Toméan genetic contribution to the haplotypic ancestry of cluster C4 aligns with the history of discrimination (Fig. 2.8). However, the haplotypic ancestry of cluster C5 suggests substantial admixture between Cabo Verdeans and São Toméans. This indicates that social stratification in the colonial *roça* plantations may have been permeable or that social boundaries between communities changed when the plantation system ended.

2.3.6 The genetic contributions from Mozambique and Angola

Enslaved Africans and contractual workers from Angola and Mozambique also worked in coffee and cacao plantations in São Tomé (Nascimento 2003; Nascimento 2023). However, I found little genetic contribution from Mozambique, primarily in the C4 and C5

clusters (2% and 8%, respectively; Figure 2.8). Identifying the impact of 19th and 20th century migrations from Angola is more challenging, as West Southern African ancestry in São Tomé also largely reflect TAST-related migrations in the 16th and 17th centuries.

2.3.7 Final methodological considerations

Genotype-based analysis was particularly useful for exploring patterns of genetic diversity and making assumptions fundamental to genetic ancestry inference (Figure 2.3, 2.4). Combining genetic findings from different analyses allowed to further understand the observed patterns. For instance, in Figure 2.6, I expected that ADMIXTURE results would somehow mirror patterns already observed in the ASD-MDS plot, since both the analyses were based on similar datasets including the same African regions. However, examining the Top Green individuals in the MDS plot suggested that the green ADMIXTURE cluster in São Tomé likely reflects genetic diversity similar to that of Cabo Verde rather than West Western Africa. Both analyses provided valuable insights into patterns of genetic diversity across the island and raised questions about the underlying histories of admixture.

Analysing shared haplotypes among São Toméans and other African and European populations allowed a fine-scale investigation of genetic diversity and ancestry patterns in São Tomé. Firstly, the haplotype-based analysis was particularly useful for identifying distinct clusters of individuals in São Tomé (Figure 2.7), which proved to be difficult considering genetic measures based on genotypes (Figure 2.5, 2.6). Secondly, I was able to reconstruct the haplotypic ancestry of each cluster separately, using the African and European populations included in the dataset as proxies for admixture events.

Genetic ancestry inference is highly dependent on the population samples selected as proxies. For instance, including or excluding Cabo Verdeans as source populations in São Tomé significantly affects the results and their interpretation (Figure 2.8A and B). Moreover, the population samples considered in this study may not be the most accurate proxies - an important issue that must be considered when drawing conclusions. My findings further prove the importance of accounting for genetic structure in the admixed population. In São Tomé, for example, when considering genetic clusters as separate target populations for admixture inference, I found that some of them resulted from substantially different admixture histories (Fig. 2.8C).

Inferring and dating admixture events requires accurate local ancestry assignments. Since different Local Ancestry Inference (LAI) methods are based on different assumptions and have specific limitations, it is crucial to choose the right method, taking into account the research questions and the available data.

During my PhD, I explored multiple LAI approaches on the São Toméan dataset. One of the most widely used methods, RFMix (Maples et al. 2013), proved to be not optimal for my specific research questions. RFMix is particularly effective when distinguishing ancestry tracts from highly divergent source populations, such as continental African and European proxies. However, as acknowledged by the authors, it is not best suited when sources are genetically relatively little differentiated. This was especially relevant in my study, which involved distinguishing between West Central African and West Southern African ancestry tracts. Moreover, several admixture modelling and dating methods, such as TRACTS (Gravel 2012), rely on the distribution of ancestry tract lengths. Although not reported in detail in this chapter, additional analyses using TRACTS based on RFMix results revealed that admixture inference was highly dependent on the number of source populations specified for local ancestry inference. In particular, using a larger number of source populations in RFMix resulted in shorter inferred ancestry tracts compared to models with fewer sources, heavily affecting admixture dating with TRACTS.

Instead, I found that the ChromoPainter2-SOURCEFIND-fastGLOBETROTTER pipeline was better suited to my research aims. First, I used ChromoPainter to “paint” the genomes of São Toméans using all the other population samples from Africa and Europe in the dataset, with the aim to minimise prior assumptions. Based on this chromosome painting, I then used SOURCEFIND to identify the subset of source populations that best explained the genetic diversity for each São Toméan genetic cluster (Figure 2.8). Finally, I used fastGLOBETROTTER to model each admixture event as involving two major sources per admixture event. This allowed to “split” the source populations identified by SOURCEFIND into two main groups, assuming a simple model with one pulse of admixture, and to infer the most likely number of generations since the main admixture event occurred (Figure 2.10).

2.4 Conclusions and perspectives

Historical migrations and social structures associated with different periods of Portuguese colonisation in São Tomé have likely influenced reproductive isolation and gene flow between populations and communities on the island. My results suggest that the three genetic clusters in São Tomé correspond to significant anthropological, historical and linguistic categories: the *Angolares*, the *Forros* and the Cabo Verdeans. These groups are associated with histories of marooning, slavery and indentured labour, respectively. Interestingly, two other clusters appear to have arisen from admixture among these groups. This reflects the gradual dismantling of a genetic structure previously established by successive historical migrations, demographic events and social factors.

To characterize the mosaic of genetic ancestries observed in São Tomé, I used the expression “nested admixture”, which refer to the gene flow between groups that are recently admixed (Figure 2.8). This emphasizes the importance of accounting for multiple layers of genetic ancestry resulting from successive admixture events, which is an inherently difficult task using current methods, as demonstrated in several previous studies based on both modern and ancient DNA (Baharian et al. 2016; Busby et al. 2016; Sarno et al. 2017; Wangkumhang and Hellenthal 2018). Recent advances have begun to address some challenges by developing methods that can infer sources without relying on modern populations as proxies, or disentangle contributions from closely related source populations (Salter-Townshend and Myers 2019; Browning et al. 2023). Future studies on São Tomé could attempt to infer and date admixture between the *Forros* and the *Angolares*, using, for example, clusters C1 and C3 as proxies of source populations for cluster C2. Similarly, it may be possible to disentangle the two main admixture processes - the founding of the Cabo Verdean population and subsequent admixture with São Toméans - for cluster C5.

A genealogical perspective on admixture could further help investigating the complex patterns of genetic ancestry observed in São Tomé. For instance, novel theoretical developments have investigated the relationship between the number of genetic and genealogical ancestors in the Afro-American population (Mooney et al. 2023; Agranat-Tamir et al. 2024), which would be of major interest to further understand how admixture occurred over time in the different São Toméan genetic groups here identified.

Finally, in several populations descended from enslaved Africans, female and male contributions from source populations were not equal, and mating might have occurred preferentially between individuals with certain genetic ancestry - both processes affecting genome-wide genetic diversity (Goldberg et al. 2014; Zaitlen et al. 2017; Goldberg et al. 2020; Micheletti et al. 2020; Kim et al. 2021; Ongaro et al. 2021; Korunes et al. 2022; Mas-Sandoval et al. 2023). It is therefore important, in future studies, to investigate the impact of ancestry-related sex bias and non-random mating on genetic diversity and admixture patterns in São Tomé.

CHAPTER 3

3 Sex-specific demographic processes shaping genetic diversity on the X chromosome and the autosomes in São Tomé and Cabo Verde

While autosomal genetic studies provide a broad view of population history, analysis of X chromosome data offers unique insights into gender and social dynamics affecting mating patterns and sex-specific demography. In this chapter, I introduce the latest research on sex-specific demographic inference, focusing on comparisons between the X chromosome and the autosomes. I further highlight the need for theoretical and analytical advances, particularly in disentangling the effects of sex-specific demographic processes and non-random mating patterns on the genetic diversity of recently admixed populations. I then present the first detailed study of X chromosome genetic diversity in São Tomé, along with several other populations in Africa, including seven islands in Cabo Verde, that may have contributed to the admixture processes. Using genotype- and haplotype-based methods, I found interesting differences in patterns of genetic diversity on the X chromosome and the autosomes in São Tomé and Cabo Verde. My findings suggest hypotheses that could be tested in future studies. In particular, I highlight the importance of accounting for sex-specific demographic histories in source populations prior to admixture, and of using specific haplotype-based statistics that highlight differences in genetic diversity on the X chromosome and the autosomes. This chapter not only enriches our understanding of the genetic diversity of the population of São Tomé, but also contributes to the broader question of how gender and social dynamics shape patterns of genetic diversity on the X chromosome and autosomes in recently admixed populations.

I plan to present these results in a dedicated section of a future article and hope that this chapter will help lay the foundation for future research aimed at developing new methodological frameworks capable of addressing complex demographic scenarios and using haplotype-based summary statistics.

3.1 Introduction

Genetic admixture processes in human populations occur within specific historical and social contexts in which individuals and communities interact. During the Trans-Atlantic Slave Trade (TAST), historical records document sex-biased migrations from both Europe and Africa, leading to distinct patterns of genetic ancestry on the autosomes and the X chromosome, which is inherited differently in females and males. Beyond migration, social constraints on reproduction and mating between individuals and groups, influenced by the complex interplay of racial, social and gender hierarchies in colonial societies, may have further shaped genetic diversity patterns on both X and autosomal chromosomes.

Sex-specific demographic processes in admixed populations impact the distribution of genetic ancestry tracts on the X chromosome and the autosomes, complicating the study of patterns resulting from non-random mating during the admixture process. In turn, non-random mating shape genetic diversity on both the X chromosome and the autosomes, making it difficult to investigate the effect of sex-specific demographic dynamics. This complex interplay of social and gender dynamics influencing genetic diversity and ancestry on the X chromosome and the autosomes highlights the need for a deeper understanding of these processes in recently admixed populations.

It is also particularly interesting to investigate sex-specific processes in populations that are considered as proxies of source populations in the admixture process. In fact, the patterns of genetic diversity in source populations may themselves be influenced by sex-specific demographic events that occurred prior to admixture. To understand patterns of genetic diversity in the admixed population is therefore important to know which patterns were already formed in source populations, and which formed after admixture.

In this introduction, I outline some expectations regarding ancestry-related mating patterns and sex-specific demography on the genetic diversity of the X and autosomal chromosomes, including a focus on investigating Identical-By-Descent shared haplotypes patterns on the X chromosome. I thus highlight the importance of examining these dynamics in São Tomé and its source populations. São Tomé indeed provides a unique context for understanding how social and gender dynamics contribute to shape genome-wide genetic diversity during complex admixture processes.

3.1.1 Ancestry-related mating patterns

The analysis of genetic ancestry on the autosomes and the X chromosome, which is inherited in two copies in females and one copy in males, permit to investigate: (1) the number of females and males from source populations who have contributed to the gene pool of admixed populations (Goldberg et al. 2014; Goldberg and Rosenberg 2015); and (2) the degree to which individuals preferentially mate with those who share similar genetic ancestry (Zaitlen et al. 2017; Goldberg et al. 2020; Kim et al. 2021; Mas-Sandoval et al. 2023).

Ancestry-related sex bias, where admixture involves a different number of females and males from the source populations, has often been used to explain the patterns of genetic diversity observed in populations descended from enslaved Africans during the Trans-Atlantic Slave Trade (Parra et al. 1998; Lind et al. 2007; Coelho et al. 2008; Bryc et al. 2010; Beleza et al. 2012; Moreno-Estrada et al. 2013; Bryc et al. 2015; Homburger et al. 2015; Baharian et al. 2016; Mathias et al. 2016; Fortes-Lima et al. 2018; Ongaro et al. 2019; Micheletti et al. 2020; Almeida et al. 2021).

The higher prevalence of African or Native American mitochondrial DNA (inherited through the maternal line) and European Y chromosomes (inherited through the paternal line), and/or a higher African or Native American ancestry on the X chromosome (present in two copies in females and one in males) compared to the autosomes, may indeed stem from a greater genetic contribution from African or Native American females and European males to the admixture process. However, this is not the only ancestry-related sex bias scenario that may explain these patterns (Goldberg and Rosenberg 2015). Other non-random mating processes and sex-specific demographic behaviours affect patterns of genetic diversity on these chromosomes (Heyer et al. 2012; Zaitlen et al. 2017; Goldberg et al. 2020).

Ancestry-related assortative mating, where individuals tend to mate with others with similar shared ancestry, can further affect genome-wide genetic diversity. For instance, it can affect the distribution of the length of ancestry tracts on the autosomes that are usually used to date admixture events (Zaitlen et al. 2017; Goldberg et al. 2020). This means that an earlier admixture process with assortative mating could be interpreted as a later migration pulse under random mating (Korunes et al. 2022; Mas-Sandoval et al. 2023). Moreover, the effect of ancestry-related assortative mating on the variance of

admixture is weaker in the X chromosome than on the autosomes, because in males the X chromosome is inherited only from the female individual of the mating pair, and is therefore independent of the male ancestry (Goldberg et al. 2020).

Recent admixture models have begun to incorporate both ancestry-related sex bias and assortative mating parameters (Mas-Sandoval et al. 2023). As the authors also suggest, forward-in-time simulations offer the flexibility to account for even greater complexity, and may allow in the future to make these parameters change over time and account for multiple pulses of migration and population growth. In addition, the inclusion of socio-economic variables in the sample would provide insights into the interplay of class, race and gender hierarchies in influencing mating patterns. Finally, including sex-specific demography would allow us to better understand the complexity of the patterns of genetic diversity of populations descended from enslaved Africans on both sides on the Atlantic.

3.1.2 Sex-specific demography

Genetic diversity on the X chromosome is theoretically expected to be $3/4$ that of the autosomes under the simple demographic model of a population of constant size with a balanced sex ratio. Under neutrality, deviations from this ratio can be explained by: (1) demographic events, such as changes in effective population size, affecting the X chromosome and the autosomes differently due to their different inheritance pattern (Pool and Nielsen 2007); and (2) sex-specific demographic processes (Heyer et al. 2012). A variety of social and cultural factors can differentially affect female and male demography, influencing in turn patterns of X chromosome and autosomal genetic diversity, with certain statistical measures more effective than others in distinguishing between these processes (Segurel et al. 2008; Verdu et al. 2013; Guyon 2024). Finally, we do not account for selection in this study, although it remains an important factor to consider, as the interplay of all evolutionary forces shape genetic diversity on the X chromosome differently than on the autosomes across human populations (Casto et al. 2010; Lucotte et al. 2016).

Sex-specific demography refers to differences in effective population size (N_e) and migration rates respective to females and males. In population genetics, N_e is a theoretical construct representing the number of individuals that effectively contribute to the genetic diversity of the next generation in a Wright-Fisher population at the drift-mutation equilibrium (Wright 1931; Hartl and Clark 2007). In an idealized population,

where mating is random, sex ratio is balanced, generations do not overlap, and there is no selection, migration, or recombination, N_e would equal the number of breeding individuals. However, real populations rarely or never meet these conditions, meaning that N_e often differs from the actual number of reproducing individuals and is typically (much) smaller than the census size. Since N_e determines the strength of genetic drift, smaller N_e leads to greater genetic drift, which in turn reduces genetic diversity. Differences in female and male N_e differentially affect genetic diversity on uniparental lineages, as well as on the X chromosome compared to the autosomes.

3.1.3 Enslaved and non-enslaved communities

In the context of slavery and colonization, several social factors may have differentially impacted female and male effective populations sizes, thus affecting patterns of genetic diversity on the X chromosome and autosomes. For example, sex-biased mortality rates due to sex-specific structures of labour on plantations could have disproportionately affected the N_e of one sex compared to the other. Moreover, social factors can differentially affect the extent of reproductive success of females and males (Austerlitz and Heyer 1998; Guez et al. 2023). For instance, polygyny, an alliance rule favouring unions between a single male and multiple females, increases male reproductive success compared to females. Social hierarchies based on race, class, and gender could have further influenced female and male demography in colonial societies. European settlers and their descendants could have had higher reproductive success than the rest of the population, and within enslaved populations in plantation systems, reproduction could be deliberately limited or promoted, to increase labour force, further shaping the genetic structure of enslaved and non-enslaved groups within the population.

Migration patterns have a significant impact on genetic diversity, particularly in structured populations where gene flow occurs primarily between subgroups rather than within a panmictic population. In such contexts, males and females may follow different migration rules shaped by cultural and social norms. For example, in patrilocal societies, females migrate to the community of their male partner, whereas in matrilineal societies, males migrate instead (Murdock 1949). These sex-specific migration patterns leave distinct genetic signatures. Patrilocality reduces Y chromosome diversity within groups while increasing genetic differentiation between them. Conversely, it increases mitochondrial DNA diversity within groups while decreasing differentiation between them (Oota et al.

2001). Similarly, patrilocality is expected to increase X chromosome diversity within groups and decrease it between groups relative to autosomes (Guyon 2024).

In human populations, patrilocality is often associated with patrilineal descent, where descendants of the male line remain in the same community. However, in slave societies, matrilineal descent may have played a role within enslaved communities due to the legal principle of *partus sequitur ventrem*, which dictated that children inherited their enslaved status from their mothers (Morgan 2018). This possibly created a complex demographic structure in which enslaved and non-enslaved populations followed opposing descent systems: one predominantly patrilineal and the other matrilineal. Investigating whether these contrasting patterns have left detectable signatures in genetic diversity could provide valuable insights into the demographic history of populations descended from enslaved Africans.

3.1.4 IBD sharing patterns on the X and autosomal chromosomes

High IBD (Identity-by-Descent) sharing within a population means that individuals share large segments of their genome due to recent common ancestry. Patterns of IBD sharing can provide insight into population structure, effective population size, and recent shared ancestry between individuals and groups (Browning et al. 2018; Nait Saada et al. 2020). In structured or isolated populations, IBD sharing is expected to be high due to reduced gene flow. The length and frequency of IBD segments also provide insights into demographic history, with longer segments indicating more recent shared ancestry and shorter segments reflecting older, more distant relationships (Baharian et al. 2016).

Comparing IBD sharing patterns on the X chromosome versus the autosomes reveals important differences in female and male demography due to their distinct inheritance patterns. The X chromosome has a lower effective population size ($3/4$ that of autosomes), leading to stronger genetic drift and higher IBD sharing in small populations. Additionally, because the X chromosome recombines only in females, its IBD segments tend to be longer than those found on autosomes. However, differences in male and female demography can also cause the X chromosome to exhibit different IBD patterns compared to autosomes. By comparing the effective population size estimated from the IBD-sharing on the X chromosome to that estimated from the autosomes it is possible to infer female and male effective population sizes, though these estimates are currently less robust for more recent generations (Cai et al. 2023).

In addition, non-random mating patterns, such as ancestry-related assortative mating, and inbreeding, when close relatives mate, promote the formation of long runs of homozygosity (ROH) by increasing the chances of inheriting from both parents identical haplotypes (IBD) (Pemberton et al. 2012; Ceballos et al. 2018). Different length classes of ROH can capture different demographic events, with shorter ROH often reflecting older demographic events and longer ROH indicating more recent inbreeding. In recently admixed populations, ROH patterns may partly reflect those of the source populations, especially for shorter ROH that predate admixture. In the case of sex-biased contributions from source populations, differences in ROH patterns between the autosomes and the X chromosome may reflect pre-existing ROH differences in source populations (Korunes et al. 2022).

Both IBD sharing and ROH are generally expected to be higher on the X chromosome due to its smaller effective population size and faster coalescence time. However, sex-specific social behaviours can affect this difference. For instance, it is expected to observe a greater difference in the case of matrilineal consanguinity, where related women are more likely to reproduce within the same lineage (Cotter et al. 2024).

Much of the behaviour of IBD and ROH patterns on the X chromosome under different sex-specific processes remains to be understood. Differences in IBD and ROH patterns between the X chromosome and the autosomes may arise from a variety of factors affecting female and male demography. Moreover, source populations may introduce different patterns of IBD and ROH, and if admixture is sex-biased, these patterns will differentially affect patterns of IBD and ROH on the X chromosome and the autosomes in admixed populations. Analysis of IBD sharing and ROH in recently admixed populations and their sources thus represents a promising approach to investigate mating patterns and sex-biased demography during admixture.

3.1.5 Gender-specific social dynamics in São Tomé and Cabo Verde

The island of São Tomé and the archipelago of Cabo Verde provide unique opportunities for exploring these dynamics. These African Atlantic islands served as sites of European settlement, exploitation of enslaved Africans, and admixture between these sources since the earliest stages of the TAST (Seibert 2014). São Tomé, in particular, hosted among the earliest instances of slave societies and maroon communities (Caldeira 2011). Following the end of the TAST and the abolition of slavery, waves of labour-driven

migration from Cabo Verde to São Tomé put into contact two distinct populations descended from enslaved Africans (Nascimento 2008). We aim to investigate whether major differences in peopling histories and the evolving social contexts have shaped distinct patterns of genetic diversity in São Tomé and Cabo Verde on the autosomes and on the X chromosome.

Based on historical records and previous genetic studies, the population of São Tomé likely result from admixture, since the 15th century, between African populations from different regions (Caldeira 2011), mostly populations similar to those living in present-day Ghana and Nigeria in East Western Africa and Congo and Angola in South Western Africa, with limited genetic contribution from European populations. Since the late 19th century, migrations from the archipelago of Cabo Verde represented a major demographic shift on the island (Nascimento 2003), and Cabo Verdeans admixed over time with the rest of the population of São Tomé. In Cabo Verde, on the other hand, admixture mainly happened between African populations, resembling to those currently living in Senegal and Gambia in West Western African, and European populations, mostly similar to those in the Iberian Peninsula. The peopling of the archipelago consisted of successive founder effects between islands and took several centuries, from the first settled island of Santiago peopled in the 15th century to the island of São Vicente peopled in the 18th century (Beleza et al. 2012; Laurent et al. 2023).

3.1.6 The genetic diversity of putative source populations

Several different African and European populations, including recently admixed populations, likely contributed to admixture processes in São Tomé. Understanding the genetic diversity of these putative source populations is essential to determine whether migrations to the island were sex-biased and how the incoming female and male individuals admixed on the island.

The genetic diversity of most populations in sub-Saharan Africa, and in particular populations from East Western Africa, West Central Africa, South Western Africa and South Eastern Africa included in this study, has been profoundly shaped by a major demographic event known as the “Bantu expansion”. This large-scale migration of Bantu-speaking agriculturalists from West Central Africa into the rest of Sub-Saharan Africa started around 5000 years ago, and introduced complex patterns of admixture between newly-arrived Bantu-speaking agriculturalists and pre-existing local hunter-gatherer

populations (Tishkoff et al. 2009; Verdu et al. 2009; Verdu et al. 2013; Li et al. 2014; Patin et al. 2014; Bostoen et al. 2015; Patin et al. 2017; Fehn et al. 2022; Fortes-Lima et al. 2024). Genetic studies have consistently highlighted strong sex-biased admixture, with agriculturalist males preferentially mating with hunter-gatherer females, leading to distinct genetic signatures across different genomic regions, especially on the X chromosome and uniparentally inherited markers (Destro-Bisol et al. 2004; Verdu et al. 2013; Patin et al. 2014; Oliveira et al. 2018; Oliveira et al. 2019).

In South Western Africa, matrilineal social structures have played an important role in shaping genetic diversity, particularly on the mitochondrial DNA (mtDNA) and the Y chromosome, as observed in southern Angola (Oliveira et al. 2018; Oliveira et al. 2019). Most of the Africans brought to São Tomé from South Western Africa were from northern Angola, where the kingdom of Kongo was located (Caldeira 2013). While whole-genome studies have been recently conducted in this region (Tallman et al. 2023), to my knowledge, sex-specific patterns of genetic variation remain largely unexplored, highlighting the need for further research.

A recent study in Cabo Verde examined the effects of ancestry-related sex bias and assortative mating on X chromosome and autosomal diversity (Korunes et al. 2022). The authors found that accounting for non-random mating in dating admixture produced older estimates that are more consistent with historical records. They also observed that ROH patterns can be difficult to interpret in admixed populations because of differences in contributions from source populations and demographic processes that happened across the islands after admixture. In addition, the study confirmed higher African ancestry on the X chromosome than on the autosomes, consistent with an excess of European male and African female genetic contributions.

While these findings provide valuable insights, the study clustered together individuals from the northern islands of Cabo Verde, including Santo Antão, São Vicente and São Nicolau, whereas Laurent et al. 2023 showed explicitly important genetic dissimilarities among these islands as well as varying admixture patterns and varying underlying admixture histories (Laurent et al. 2023). Thus, exploring in detail the X and autosome diversities among Santo Antão, São Vicente and São Nicolau remain to be conducted and will provide a more detailed perspective on the history of settlement of these islands and possibly more recent migrations to São Tomé.

3.1.7 Insights from uniparental lineages in São Tomé

In São Tomé, previous genetic studies have provided valuable insights into the genetic diversity of maternal and paternal lineages, particularly concerning the differentiation of three social groups: the *Angolares*, descendants of runaway slaves who formed isolated maroon communities; the *Forros*, descendants of freed African slaves who became one of the main social groups on the island; and the Cabo Verdeans and their descendants.

These studies revealed an absence of European maternal lineages in São Tomé (Trovoada et al. 2004), but a significant presence of European paternal lineages (Trovoada et al. 2001; Gonçalves et al. 2007; Trovoada et al. 2007; Coelho et al. 2008), suggesting that Portuguese males likely contributed more than females to the formation of the admixed population. Furthermore, the *Forros* and Cabo Verdeans showed higher proportions of paternal European lineages than the *Angolares* (Trovoada et al. 2007; Almeida et al. 2021). Previous studies on the Y chromosome also pointed out that the relative abundance of male European lineages in São Tomé was higher than expected from the reduced demographic impact of the Portuguese on the islands (Trovoada et al. 2007). A possible explanation for these patterns could be that Portuguese males left much more offspring than African males, implying a significant variance in the reproductive success of individuals based on race, class and gender hierarchies.

Uniparental lineages further suggest distinct male and female demographic patterns within São Tomé and across different social groups. The *Angolares*, for instance, show low Y chromosome diversity and limited male gene flow from other groups, whereas mitochondrial DNA data reveal greater within-population diversity and higher levels of female introgression (Coelho et al. 2008). Overall, the genetic diversity of uniparental lineages suggest that female mobility within the island was less restricted than male mobility, a characteristic often observed in patrilineal societies with patrilocal residence (Trovoada et al. 2004; Coelho et al. 2008; Almeida et al. 2021).

3.1.8 Aim of this study

This study aims at investigating ancestry-related mating patterns and sex-specific demographic processes affecting genetic diversity in São Tomé. Results from previous genetic studies on uniparental lineages in São Tomé provide a promising framework and raise new questions that need further investigation. By comparing genetic diversity on

the X chromosome and autosomes, we gain deeper insights into the island's complex admixture processes. In addition, considering the nested genetic structure and admixture previously observed in São Tomé in chapter 2, this study examines the social and gender dynamics at play on the island between different social communities: the *Angolares*, *Forros* and Cabo Verdeans.

In this chapter, I investigate X chromosome and autosomal markers in genotyping datasets from São Tomé, Cabo Verde and other worldwide regions, including Central, East and West Africa (Patin et al. 2017), Mozambique and Angola (Semo et al. 2020), and other African, European and American populations. For some of the African populations considered in this study, this is the first time that X chromosome genetic diversity is analysed at this resolution. Using both genotype-based and haplotype-based methods, I identify statistics that effectively highlight differences in patterns of genetic diversity between the autosomes and the X chromosome in these populations. I interpret these differences in terms of ancestry-related mating patterns and sex-specific demographic processes, proposing sex-specific admixture scenarios for further explicit testing. These results lay the foundation for future studies modelling complex sex-specific admixture scenarios using approaches such as Approximate Bayesian Computation (ABC) inference (e.g. Fortes-Lima et al. 2021), where the selection of appropriate summary statistics is critical.

3.2 Results

3.2.1 Genetic differentiation of individual samples based on autosomes and X chromosomes

We conducted Multi-Dimensional Scaling (MDS) analysis on individual pairwise Allele-Sharing Dissimilarities (ASD, Bowcock et al. 1994), using 541,734 autosomal SNPs or 12,861 X chromosome SNPs. The X chromosome data for males and females were analysed separately due to differences in ploidy.

The ASD-MDS analyses based on autosomal and X chromosome data, considering only diploid X chromosome markers from female samples, exhibited similar patterns of genetic differentiation in the first three axes of variation (Figure 3.1): 1) the first axis separated South American and African populations, 2) the second axis differentiated European and East Asian populations, and 3) the third axis separates South American and South Asian populations.

However, the ASD-MDS analyses based on different datasets resulted in different patterns of genetic differentiation. The analysis based on the X chromosome data from female samples resulted in more dispersed population clustering compared to the analysis based on the autosomal dataset, including both female and male samples, as well as more genetic markers. Moreover, the analysis based on haploid X chromosome markers in male samples only, resulted in patterns of genetic differentiation that were very different to those previously described, most probably due to the strong overrepresentation of males in West Central African and Central African population samples.

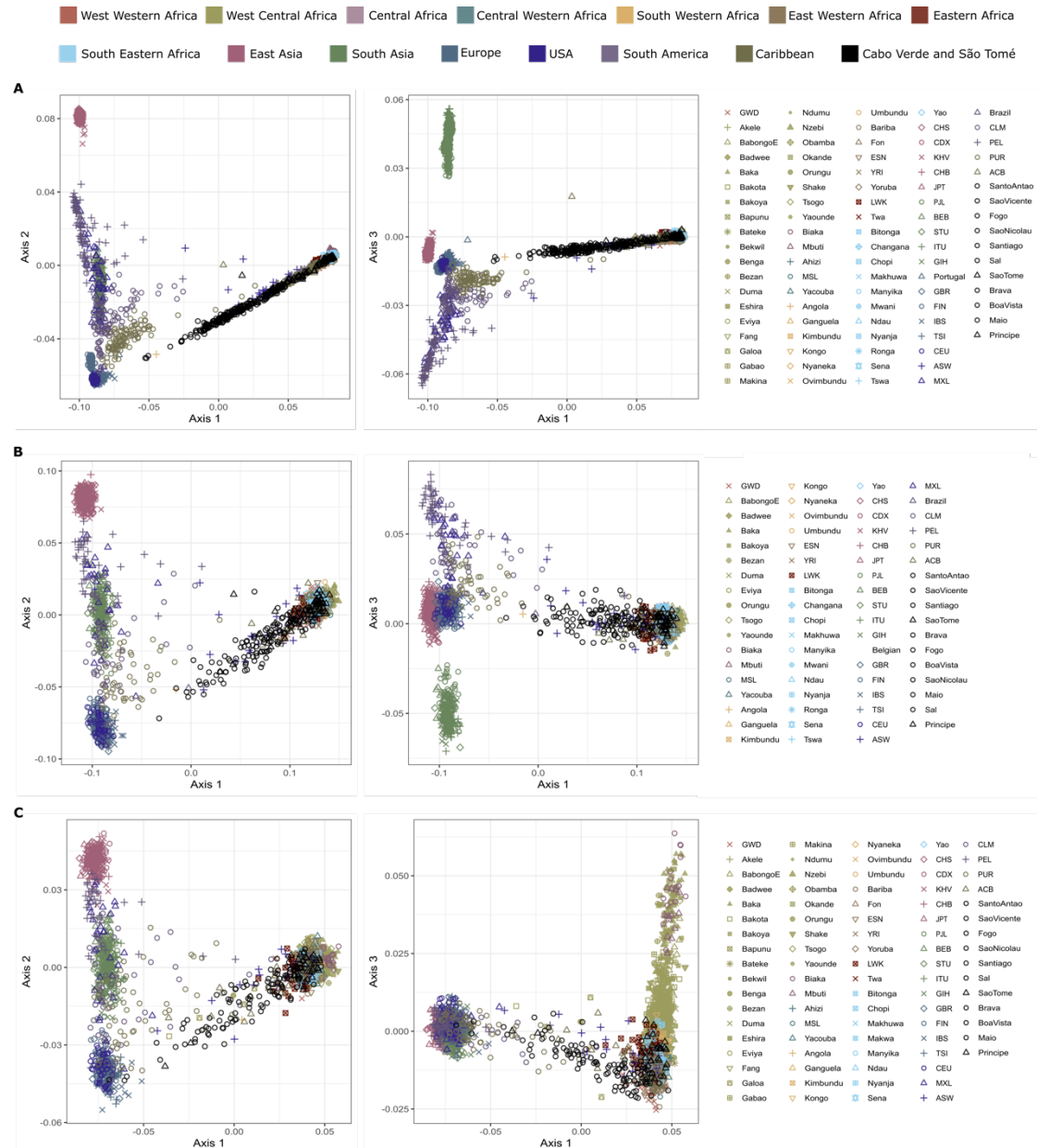


Figure 3.1. ASD-MDS analysis of worldwide population samples.

(A) ASD-MDS calculated on 541,734 Autosomal SNPs of 4127 individuals from 93 populations. (B) ASD-MDS calculated on 12,861 X Chromosome SNPs of 1666 female individuals from 72 populations. (C) ASD-MDS calculated on 12,861 X Chromosome SNPs of 2534 male individuals from 86 populations.

To better investigate the patterns of genetic differentiation in São Tomé and Cabo Verde, we analysed a subset of populations from Western, Central and Southern Africa, the Caribbean, the USA, and Europe (Figure 3.2). Indeed, the strong genetic differentiation between population some samples previously considered, such as those from South America and Asia, may have obscured finer patterns of genetic differentiation in other populations, which were possibly emerging at higher order dimensions.

Based on this subset of populations and considering autosomal data, the first ASD-MDS axis separates African and European populations, and the second axis captures variation between West Western African and East Southern African populations. Interestingly, the third axis highlights differences within Europeans, separating population samples from i) northern Europe, namely the Finnish in Finland (FIN), ii) central Europe, including the Utah residents with Northern and Western European ancestry (CEU) and the British in England and Scotland (GBR), and iii) southern Europe, including the Iberian populations in Spain (IBS), the Toscani in Italy (TSI), and an individual sample collected in São Tomé that self-reported to be born Portugal. Populations descended from enslaved Africans, including São Toméans, Cabo Verdeans, as well as individuals with African Ancestry in Southwest USA (ASW) and African Caribbeans in Barbados (ACB), lie between the African and European clusters.

Based on similar datasets and considering X chromosome data, the first axis of the ASD-MDS analyses also differentiate population samples from Africa and Europe. However, the second axis reflects differences between Cabo Verdeans and South Eastern African when considering female samples, and between Cabo Verdeans and West Central Africans when considering male samples. In both cases, the third axis captures genetic differentiation within the Cabo Verdean sample, and do not differentiate between European populations as observed for the autosomal dataset. Finally, individual samples from São Tomé and Cabo Verde do not form distinct clusters, but are widely dispersed along the second axis of variation of the ASD-MDS analyses for both female and male X chromosome datasets.

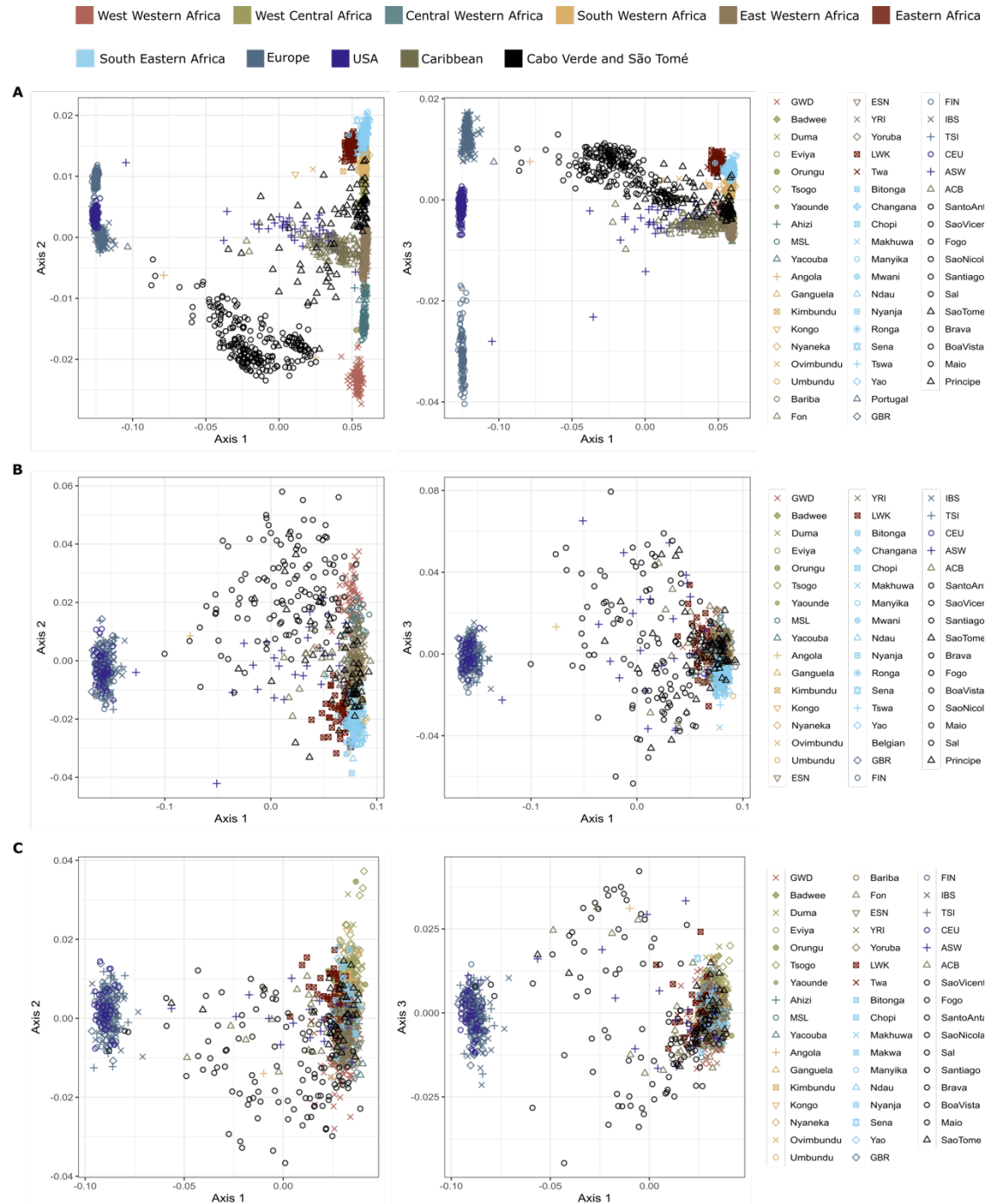


Figure 3.2. ASD-MDS analysis of a subset of population samples.

(A) ASD-MDS calculated on 541,734 autosomal SNPs of 1998 individuals from 56 populations. (B) ASD-MDS calculated on 12,861 X Chromosome SNPs of 922 female individuals from 51 populations. (C) ASD-MDS calculated on 12,861 X Chromosome SNPs of 1081 male individuals from 50 populations.

Overall, the differences in the results based on autosomal and X chromosome data may be due to the smaller number of markers considered on the X chromosome compared to those on the autosomes. In particular, individual samples form more distinguishable clusters in the ASD-MDS analyses when considering 541,734 Single Nucleotide Polymorphisms (SNPs) on the autosomes than when considering 12,861 SNPs on the X chromosome.

To explore this, I performed additional ASD-MDS analyses on autosomal chromosomes 7 and 20, which contained 29,494 and 13,380 SNPs respectively: the first autosomal chromosome is similar to the X chromosome in the number of total base pairs, the second in the number of SNPs genotyped in this study (Figure 3.3).

For both autosomal chromosomes, the first axis highlights differences between African and European populations, consistent with previous ASD-MDS analyses based on autosomal and X chromosome data. However, unlike analyses using all autosomal chromosomes, the second and third axes do not clearly differentiate European populations. Instead, the second axis captures differences between East Southern African and Cabo Verdean populations, mirroring patterns observed on diploid X chromosome makers when considering female samples. These findings demonstrate that the number and ploidy of markers has a significant impact on ASD-MDS results, leading to more dispersed clustering when fewer markers are considered in the analysis.

In the following analyses, I considered the following autosomal chromosomes for comparison with the X chromosome: 1) autosomal chromosome 20, which has a similar number of SNPs to the X chromosome in this study, for genotype-based analyses; and 2) autosomal chromosome 7, which has a similar size to the X chromosome in terms of base pairs, for haplotype-based analyses.

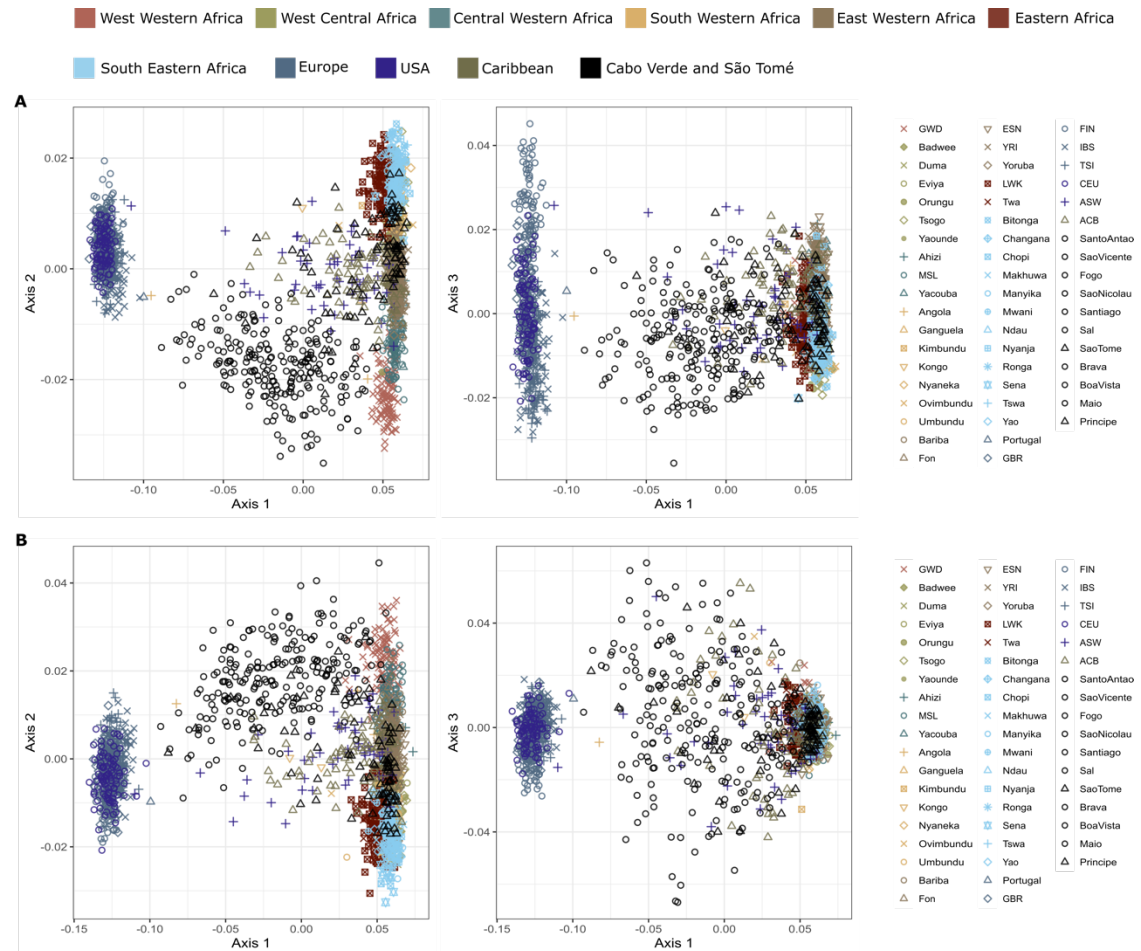


Figure 3.3. ASD-MDS analysis using two autosomal chromosomes.

(A) ASD-MDS calculated on 29,494 SNPs on autosomal chromosome 7 of 1998 individuals from 56 populations. (B) ASD-MDS calculated on 13,380 SNPs on autosomal chromosome 20 of 1998 individuals from 56 populations.

3.2.2 ADMIXTURE Analysis

A finer genetic differentiation within and between population samples may lie in further higher dimensions of ASD variation. However, it is difficult to visualize more than two or three dimensions at once using the ASD-MDS approach. Instead, STRUCTURE-like clustering algorithm (Pritchard et al. 2000; Falush et al. 2003), such as ADMIXTURE (Alexander et al. 2009), permit to easily visualize numerous higher-order axes of inter-individual genetic variation simultaneously. Moreover, a recent implementation of ADMIXTURE can deal with haplo-diploid data, permitting to consider both female and male samples for the X chromosome (Shringarpure et al. 2016).

We filtered the dataset prior to analysis, as clustering analyses are particularly sensitive to the composition of the dataset in terms of the individual samples and the linkage disequilibrium (LD) structure of the markers considered. To ensure the dataset was balanced overall, we considered a maximum of ten populations per region and up to twenty individuals per population. All individuals from São Tomé e Príncipe, Cabo Verde, the USA (African Americans) and the Caribbean (African Barbadians) were included. We thus retained 902 samples present in both the autosomal and X chromosome datasets and considered 5,421 LD-pruned SNPs on the autosomal chromosome 20 and 5,364 LD-pruned SNPs on the X chromosome.

At $K=2$, the ADMIXTURE algorithm differentiated between continental European populations, maximizing the red cluster, and continental African populations, maximizing the blue cluster, using the X chromosome and the autosomal chromosome 20 (Figure 3.4A). We thus focused on the variation in blue “continental African” and red “continental European” genetic cluster proportions among individuals descended from enslaved Africans, including São Toméans, Cabo Verdeans, African Americans and African Barbadians. In São Tomé, we categorized individuals into the five genetic clusters, previously identified using autosomal data (Figure 2.4). For each population separately, we compared the distribution of blue cluster proportions on autosomal chromosome 20 and the X chromosome, and we assessed the statistical significance of this difference using a Wilcoxon signed-rank test (Figure 3.4B).

The São Toméan genetic clusters C1, C2, and C3, which based on previous results represent the São Toméan population without Cabo Verdean genetic contributions, exhibit overall high proportions of the blue “continental African” cluster, with distributions

skewed toward 1. In clusters C1 (X chromosome: mean = 0.97, SD = ± 0.04 ; chromosome 20: mean = 0.99, SD = ± 0.01) and C2 (X chromosome: mean = 0.97, SD = ± 0.06 ; chromosome 20: mean = 0.98, SD = ± 0.03), the mean blue cluster proportion is slightly lower on the X chromosome than on chromosome 20. Conversely, in cluster C3, the proportion is slightly higher on the X chromosome (mean = 0.99, SD = ± 0.03) than on chromosome 20 (mean = 0.98, SD = ± 0.05). However, these differences are not statistically significant.

Clusters C4 and C5 in São Tomé, on the contrary, show overall lower proportions of blue “continental African” cluster. The mean of blue cluster proportion is higher on the X chromosome than on autosome 20. The distributions are not significantly different in C4 (chromosome 20: mean = 0.69, SD = ± 0.2 ; X chromosome: mean = 0.79, SD = ± 0.18 ; p-value = 0.09) but they are in C5 (chromosome 20: mean = 0.86, SD = ± 0.12 ; X chromosome: mean = 0.94, SD = ± 0.06 ; p-value = 0.02).

Across the Cabo Verdean islands, mean blue “continental African” cluster proportions are consistently higher on the X chromosome than on the chromosome 20. The magnitude of this difference varies between the islands of birth of individuals, and the difference is significant in most cases, except for the islands of Fogo (chromosome 20: mean = 0.48, SD = ± 0.12 ; chromosome 20: mean = 0.59, SD = ± 0.19 ; p-value = 0.07) and Boa Vista (chromosome 20: mean = 0.6, SD = ± 0.13 ; X chromosome: mean = 0.7, SD = ± 0.15 ; p-value = 0.06).

Among populations of African descent in the Americas, African Americans (ASW) show similar distributions of blue cluster proportions on the X chromosome (mean = 0.80, SD = ± 0.18) and autosomal chromosome 20 (mean = 0.80, SD = ± 0.18), with no significant difference (p-value = 0.8). In contrast, African Barbadians (ACB) present higher blue cluster proportions on the X chromosome (mean = 0.91, SD = ± 0.11) compared to chromosome 20 (mean = 0.88, SD = ± 0.12), a significant difference (p-value = 0.01).

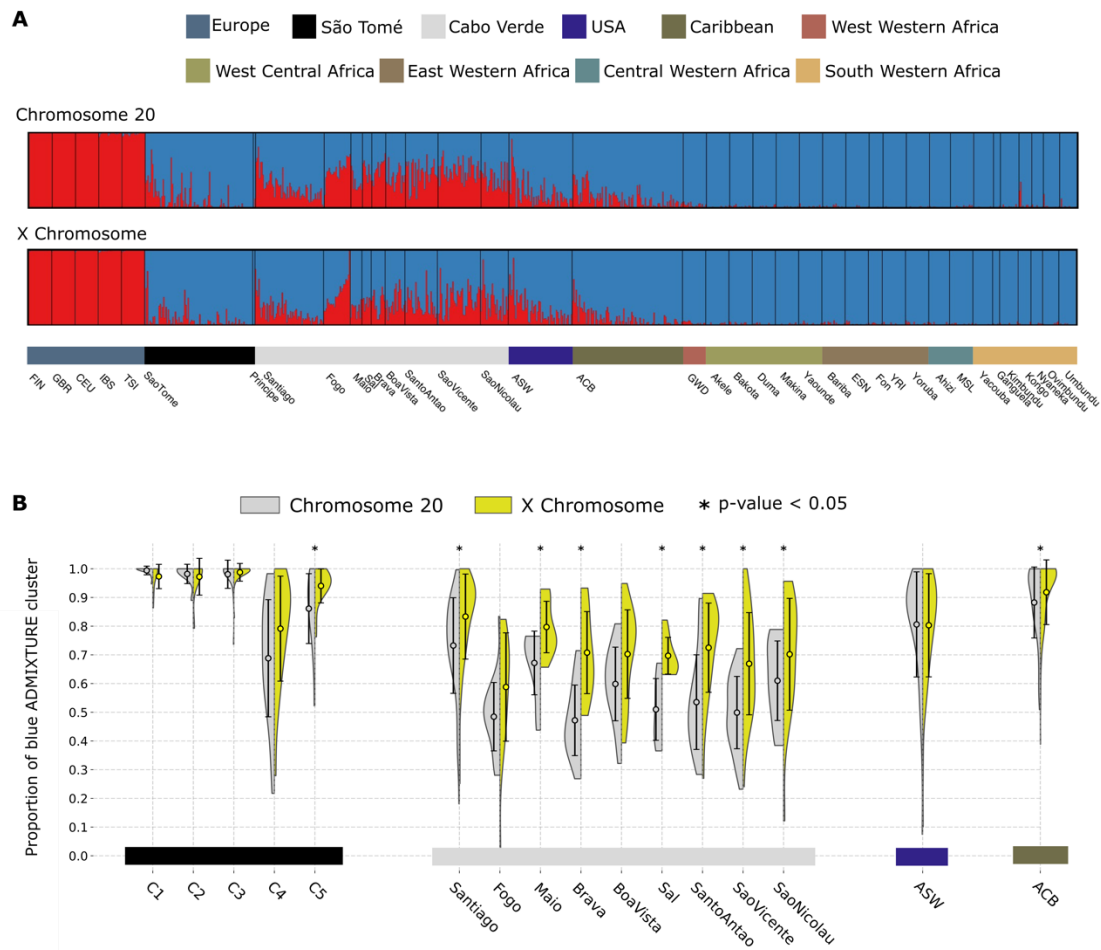


Figure 3.4. ADMIXTURE analysis on the autosomal chromosome 20 and the X chromosome.

(A) ADMIXTURE analysis at $K=2$ based on 5421 LD-pruned SNPs on the autosomal chromosome 20 and 5364 LD-pruned SNPs on the X chromosome from 902 individuals. Modal results and plots were generated using PONG (Behr et al. 2016), with each plot representing the consensus of 10 independent runs. Individuals are ordered by populations according to geographical regions, as indicated in the legend. Population codes: Finnish (FIN), English (GBR), Northern Europeans (CEU), Iberians (IBS), Italians (TSI), African Americans (ASW), and African Barbadians (ACB). (B) Violin plots representing the Kernel Density Distribution of blue “continental African” cluster proportion in ADMIXTURE results at $K=2$ for each population sample separately. The violin plots are truncated at the observed minimum and maximum values, and mean values are reported with standard deviation bars. Asterisks indicate p -values < 0.05 (Wilcoxon signed-rank test) for differences between chromosome 20 and the X chromosome in each population. The São Tomé population is divided into five genetically similar clusters (from C1 to C5), as defined in the previous study based on autosomal data (Figure 2.4).

3.2.3 IBD-sharing

Individual genomes may share contiguous stretches of DNA inherited from a common ancestor without recombination. These segments, called Identical-By-Descent (IBD), provide valuable information about recent common ancestry and population structure. We examined the IBD segments shared by pairs of individuals on the X chromosome and autosomal chromosome 7, which are similar in number of base pairs (chromosome 7: 159 Mb, chromosome X: 156 Mb). We calculated the cumulative length of long-IBD (>18 cM), reflecting shared ancestry within the last eight generations (Baharian et al. 2016), for pairs of individuals both within and between populations. This analysis included five genetic clusters from São Tomé and seven groups of individuals born in different islands of Cabo Verde (Figure 3.5). Notable differences in long-IBD sharing patterns were observed between chromosome 7 and the X chromosome across these populations. To assess the significance of these differences, I performed 10,000 permutations of the long-IBD sharing matrices. Significant differences were found within Cabo Verde, while within São Tomé were not statistically significant.

Between São Tomé and Cabo Verde, genetic cluster C3 showed no long-IBD sharing with Cabo Verde on chromosome 7 but did share long-IBD with several Cabo Verdean samples on the X chromosome. Conversely, clusters C4 and C5 shared long-IBD with more Cabo Verdean populations on chromosome 7 than on the X chromosome. Indeed, for certain population pairs, such as C4 and C5 with Santiago and Santo Antão, long-IBD sharing was higher on the X chromosome than on chromosome 7, while for others (e.g., C4-Brava, C5-São Vicente), it was absent. None of these differences, however, were statistically significant.

Within Cabo Verde, I observed several significant differences in long-IBD sharing between chromosome 7 and the X chromosome. In particular, the São Vicente–São Nicolau pair showed significantly higher long-IBD sharing on the X chromosome than on chromosome 7. I found a similar trend in São Vicente and Santo Antão, though it was not statistically significant. Additionally, long-IBD sharing on the X chromosome was higher between Santiago and other islands, but only the Santiago-Sal pair showed a significant difference. Notably, Sal exhibited significantly higher long-IBD sharing on the X chromosome than on chromosome 7 with São Nicolau, São Vicente, Santo Antão, Santiago, and Boa Vista.

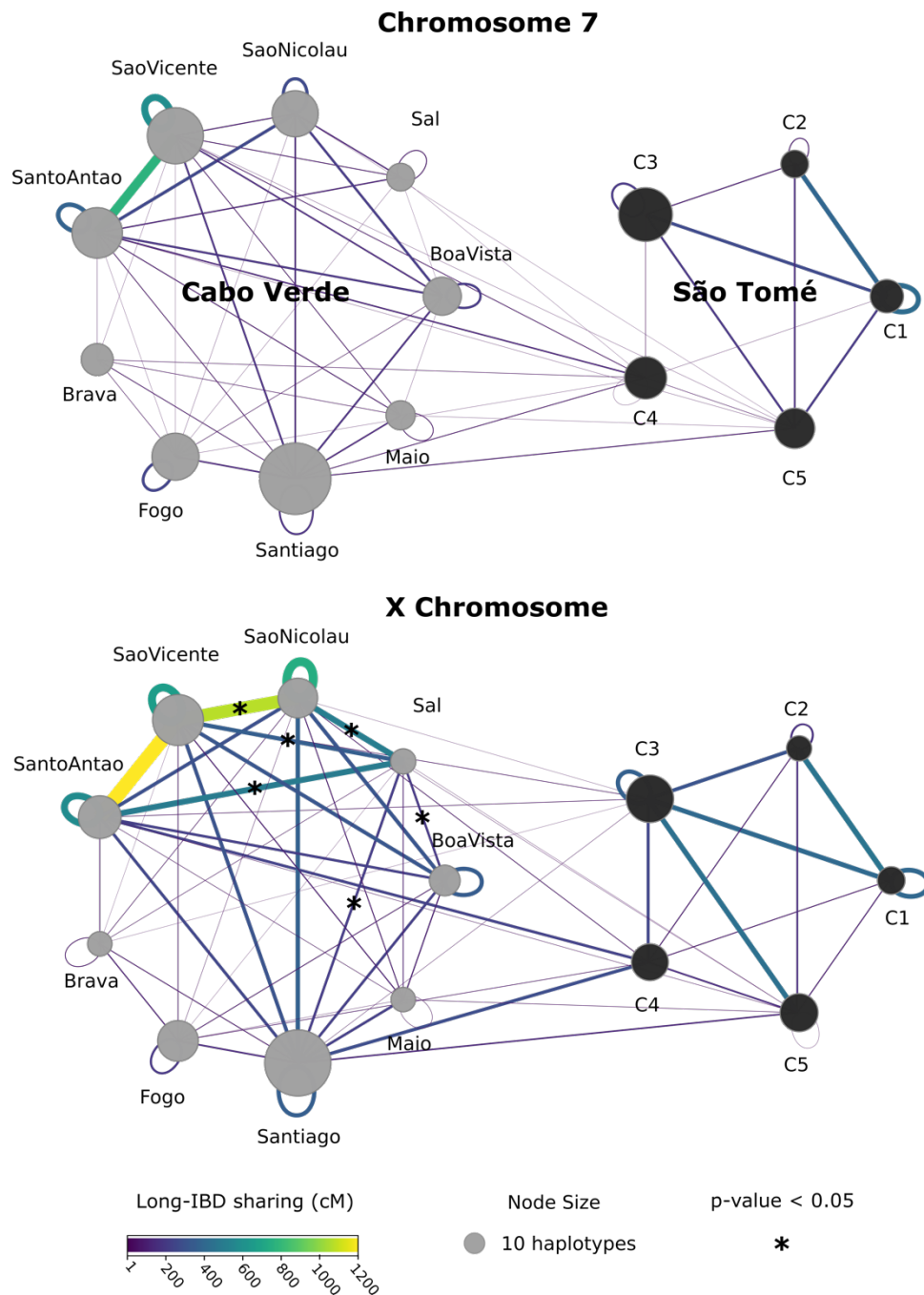


Figure 3.5. IBD sharing on the autosomal chromosome 7 and the X chromosome.

Network analysis of long-IBD (>18 cM) sharing patterns on chromosome 7 (up) and the X chromosome (bottom). Node size represents the sample size in number of haplotypes (chromosome 7: 2 haplotypes per individual; X chromosome: 2 haplotypes for females, 1 for males). Edge colour and thickness indicate cumulative long-IBD length (cM). Asterisks denote statistically significant differences (p-value < 0.05) in cumulative long-IBD sharing between chromosome 7 and the X chromosome (10,000 permutations). Population samples include individuals born in Cabo Verde on the left and São Tomé on the right of each graph.

3.2.4 Runs of Homozygosity

Within individual genomes, runs of homozygosity (ROH) occur when an individual inherits two identical IBD segments from both parents. ROH of different lengths are indicative of different demographic events: long ROH reflect recent inbreeding, medium ROH suggest population bottlenecks, and short ROH indicate more ancient population dynamics (Pemberton et al. 2012; Ceballos et al. 2018; Szpiech et al. 2019).

We calculated the total ROH length for each individual and plotted the median and interquartile ranges of total ROH length in each region for each chromosome, ordered by increasing chromosome size (Figure 3.6). While longer chromosomes tend to have higher ROH length, this pattern is not consistent across all regions and chromosomes. For instance, European populations show a different curve of median ROH length per chromosome size compared to African populations, as well as Central African compared to other African populations. For this reason, the relative difference in median ROH length between the X chromosome and an autosomal chromosome of similar size vary depending on the population considered. This variation is important to consider when comparing ROH length on the X chromosome to that of an autosomal chromosome.

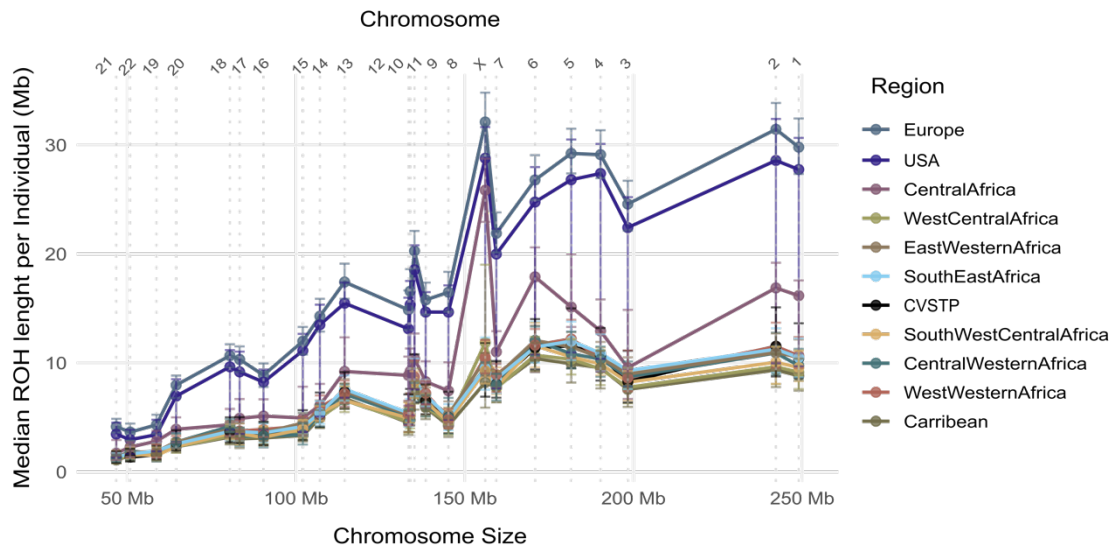


Figure 3.6. Distribution of ROH Length by Chromosome Size.

Scatterplot illustrating the median and interquartile range (25th and 75th percentiles) of the total ROH length per individual, calculated using GARLIC (Szpiech et al. 2017), and averaged by region. Autosomal and X chromosomes are analysed separately and ordered by chromosome size.

We analysed the distribution of total ROH length on the X chromosome (diploid females only) and chromosome 7 across genetic clusters in São Tomé and the islands of Cabo Verde (Figure 3.7). The distributions show differences in ROH length between population samples, with variations in the proportion of short (Amb), medium (Bmb) and long (Cmb) ROH categories, as calculated by GARLIC (Szpiech et al. 2017). In the case of the Brava and Maio populations in Cabo Verde, there was insufficient data to plot distributions on the X chromosome, i.e. at least two individual values, for some or all of the length classes. To statistically compare the total ROH length between the X chromosome and chromosome 7, we used the Wilcoxon rank sum test for each population sample.

Interestingly, cluster C1 in São Tomé show higher median long ROH length on the autosomal chromosome 7 than on the chromosome X, even though this difference is not significant. The opposite pattern is observed in all other clusters in São Tomé and all populations in Cabo Verde (except for Sal), meaning higher median long ROH length on the chromosome X than on the autosomal chromosome 7. These differences are significant only in the case of cluster C4 in São Tomé, and Santiago, São Nicolau, and São Vicente in Cabo Verde. Differences in short and medium length classes were not significant for any population sample using the Wilcoxon rank sum test.

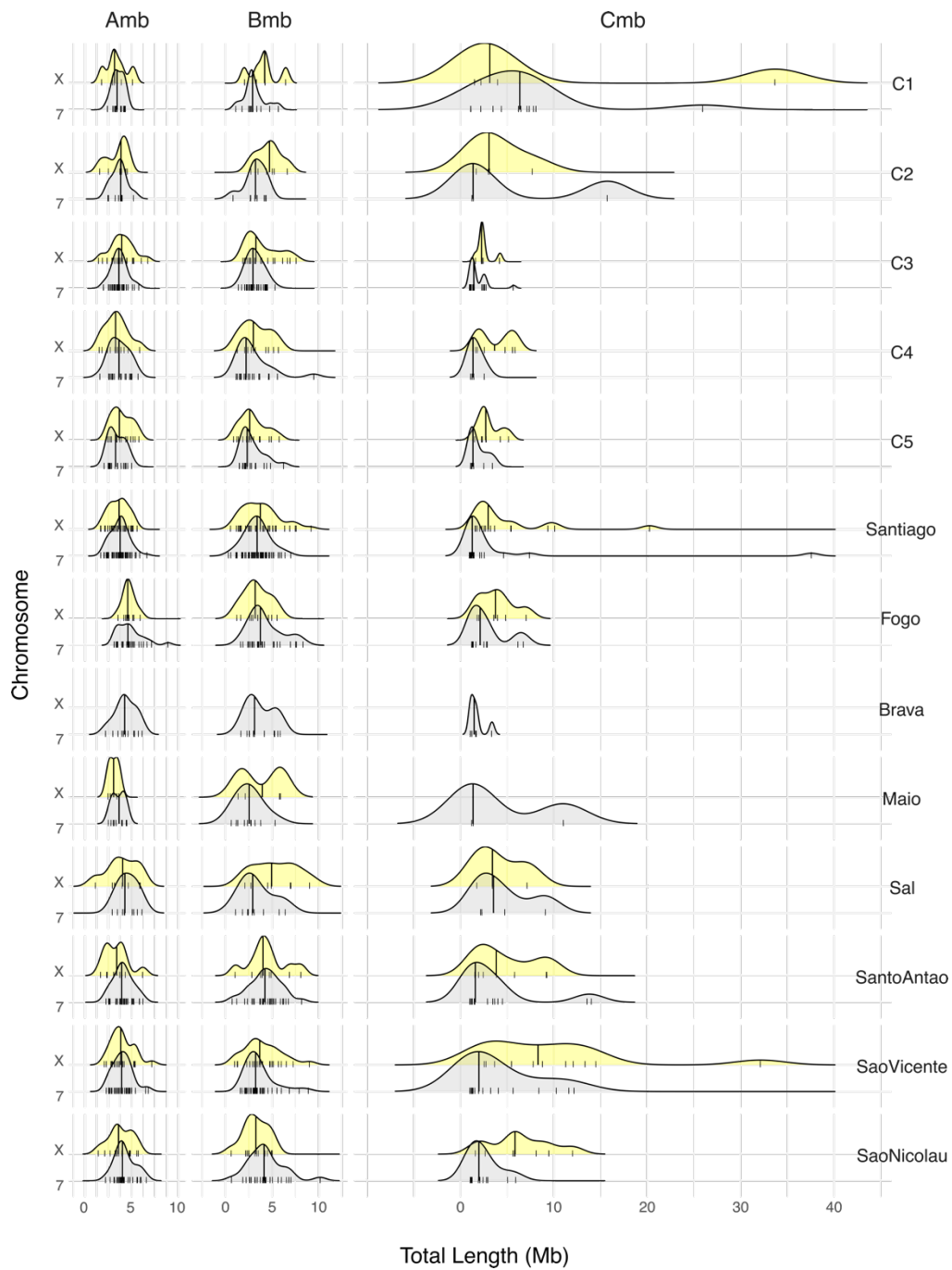


Figure 3.7. ROH by length classes on the X chromosome and chromosome 7.

The distribution of total runs of homozygosity (ROH) length per individual is shown for different genetic clusters in São Tomé (C1 to C5) and islands in Cabo Verde (Santiago to São Nicolau). ROH length is categorized into three classes: short (Amb), medium (Bmb), and long (Cmb), as determined for each population sample using GARLIC. The density ridgeline plots depict the variation across population samples and length classes, with median values indicated by quantile lines, and individual values represented as short vertical lines at the bottom of each distribution.

3.3 Discussion

In this section, I first focus on the methods based on genotypes and haplotypes to analyse and compare genetic diversity patterns on the X chromosome and the autosomes. I then turn to the island of São Tomé and the archipelago of Cabo Verde, providing a detailed discussion of the results and formulating hypotheses about social and gender dynamics that might have shaped these patterns. Finally, I propose complex models of genetic admixture including ancestry-related sex bias and sex-specific demography to explicitly test some of these hypotheses in future works.

3.3.1 Methodological discussion

Both genotype-based and haplotype-based methods provided valuable statistics to highlight differences between genetic diversity patterns on the X chromosome and the autosomes. However, such comparisons are inherently limited by the smaller size and overall lower number of markers available on the X chromosome compared to the complete set of autosomal chromosomes.

3.3.1.1 Genotype-based analyses

Exploring genetic differentiation based on individual genotypes is the first step to understand general patterns of genetic diversity and identify potential differences between autosomes and the X chromosome. Our study highlights that these differences partly stem from differences in the number of markers and individual samples considered in the analysis. When considering an autosomal chromosome with a number of markers closer to that of the X chromosome (Figure 3.3), we observed patterns of genetic diversity more similar to that of the X chromosome (Figure 3.2B and C), than when considering all autosomal markers compared to X-markers (Figure 3.2A). Therefore, in this study, we contrasted results obtained for the X chromosome with those based on all autosomes, or with those based on a single autosome with a comparable number of markers (chr. 20) or similar length (chr. 7), respectively.

The ADMIXTURE (Alexander et al. 2009) clustering algorithm is a valuable tool for comparing genetic differentiation on individual genomes between the X chromosome and the autosomes (Shringarpure et al. 2016). Comparing the proportion of genetic clusters maximized in African, European, and Native American populations between the X chromosome and the autosomes has been used to investigate ancestry-related sex

bias in many populations descended from enslaved Africans on both sides of the Atlantic (Moreno-Estrada et al. 2013; Homburger et al. 2015; Baharian et al. 2016; Mathias et al. 2016; Fortes-Lima et al. 2017; Fortes-Lima et al. 2018; Ongaro et al. 2019). However, considering all autosomal chromosomes or a single autosomal chromosome can affect clustering results, and thus affect the comparison between X chromosome and autosomes (Ongaro et al. 2019).

In this study, we analysed linkage disequilibrium (LD)-pruned data with ADMIXTURE on the X chromosome and chromosome 20, since the two chromosomes had a similar number of markers after LD-pruning. At $K=2$, African populations maximized the blue cluster and European populations the red. However, at higher values of K ($K=3$ to $K=10$), most clusters were not fully resolved, meaning that they not reached 100% assignment in any individual sample (see Supplementary Figure 3.1). This contrasts with the results obtained in the previous chapter using markers from all autosomal chromosomes (Figure 2.3). Despite the fact that a slightly different data set in terms of population composition has been considered in the two analyses, the discrepancy between clustering results when considering a single or all autosomal chromosomes is most probably due to differences in the number of genetic markers used in the analysis (Rosenberg et al. 2005).

The overrepresentation of individuals from African-descended populations in our dataset may further affect the accuracy of genetic clustering. One solution could be to perform an unsupervised ADMIXTURE analysis separately for each recently admixed population of interest. Alternatively, a supervised approach to ADMIXTURE analysis, where the clustering of the admixed individuals of interest is based on a prior clustering of individual genomes assumed to represent the source populations, may produce accurate results even when a limited number of markers are considered (Alexander and Lange 2011). This approach could thus be useful for comparing results based on the X chromosome and an autosomal chromosome. However, the assumptions required for supervision significantly influence ADMIXTURE results.

We observed interesting differences between unsupervised ADMIXTURE results based on the X chromosome and autosomal chromosome 20 at $K=2$ in several populations descending from enslaved Africans on both sides of the Atlantic, which may reflect sex-specific demographic histories and mating patterns during admixture processes. In

particular, in São Tomé, we observed considerable variation in the “continental African” genetic diversity on the X chromosome and the autosomal chromosome 20 across five clusters of genetically similar individuals, as identified in our previous study (Figure 2.4).

3.3.1.2 Haplotype-based analyses

Analysis of Identical-by-descent (IBD) tracts shared between individuals and runs of homozygosity (ROH) within the genome of each individual can inform on sex-specific demography and mating patterns (Korunes et al. 2022; Cai et al. 2023; Cotter et al. 2024). However, expectations regarding IBD-sharing and ROH patterns on both the autosomes and the X chromosome in recently admixed populations remain largely unexplored (Korunes et al. 2022). If the population experienced recent admixture, strong bottlenecks, and sex-specific demographic processes affecting its genetic diversity, the X chromosome could show different IBD-sharing and ROH patterns than expected under simplified demographic models. Future studies and better understanding of these dynamics for recently admixed populations may give more insight into sex-specific demography and mating patterns in populations descended from enslaved Africans.

In this study, patterns of long-IBD sharing within and between São Tomé and Cabo Verde may reflect different sex-specific demographic histories. However, we expect IBD sharing to be higher on the X chromosome than on the autosomes due to the smaller effective population size and faster coalescence time of the X chromosome (Cotter et al. 2024). This raises the question of whether the observed higher sharing on the X chromosome is driven by these genetic processes or by sex-specific demographic processes. The statistical test used here, a permutation test, does not allow for this distinction. Future analyses and a stronger theoretical framework will be needed to determine whether significant differences truly reflect sex-biased processes.

The same considerations apply to ROH patterns. Moreover, individual ROH levels vary across different chromosomes in ways that do not always correlate with chromosome size (Figure 3.6). Therefore, comparing ROH patterns on the X chromosome with those on an autosomal chromosome of similar size, as previously done in Cabo Verde (Korunes et al. 2022), may not be the best approach.

Nevertheless, I observed interesting patterns of long ROH on the X chromosome and the autosomal chromosome 7 in São Tomé and Cabo Verde (Figure 3.7). For instance,

in the case of cluster C1, the medium length of long ROH is higher on the autosomal chromosome than on the X chromosome. Although this difference is not significant using simple statistical tests, this pattern is not expected under neutrality, as the smaller effective population size of the X chromosome compared to the autosomes would result in higher ROH on the X chromosome. Further analysis is needed to assess which sex-specific demographic processes might explain this pattern.

3.3.2 São Tomé and Cabo Verde

The main aim of the previously discussed approaches, comparing genetic diversity patterns on the X chromosome and the autosomes, was to investigate ancestry-related sex bias and sex-specific demography in São Tomé and Cabo Verde. In the following paragraphs, I interpret genetic findings in light of the historical and anthropological information available on the islands.

3.3.2.1 Female and male genetic contributions from Africa and Europe in São Tomé

We found limited European haplotypic ancestry in São Tomé based on autosomal data (Ciccarella et al. 2024), and previous studies on uniparental lineages found relatively high levels of European lineages in the Y chromosome and virtually no European lineages in the mitochondrial DNA (Trovoada et al. 2004; Trovoada et al. 2007; Coelho et al. 2008; Almeida et al. 2021). These patterns suggest a greater European male than female contribution to admixture in São Tomé, which would also result in lower European genetic ancestry on the X chromosome compared to the autosomes. However, the patterns of genetic variation observed in São Tomé require more complex hypotheses, as I found distinct patterns of African and European genetic diversity on the X chromosome and the autosomes across the five genetic clusters identified on the island.

Focusing on the clusters without Cabo Verdean ancestry in São Tomé, specifically clusters C1, C2, and C3, only cluster C3 shows a relatively higher level of European genetic diversity on the autosomes compared to the X chromosome, which aligns with the previously proposed hypothesis. In contrast, clusters C1 and C2, which correspond to the *Angolares* of São Tomé according to the interpretations of the previous study, presented the opposite pattern, namely higher European genetic diversity on the X chromosome compared to autosomal chromosomes (Figure 3.4).

These patterns are not statistically significant and may have formed by chance or, particularly in the case of clusters C1 and C2, under the effect of genetic drift. The genetic diversity of cluster C1 and C2 resulted from strong genetic drift due to small effective population size, which may have amplified differences in genetic diversity between the autosomes and the X chromosome solely by chance. Also, the inherently smaller effective population size of the X chromosome compared to the autosomes could accentuate these differences (Pool and Nielsen 2007). Finally, historical accounts suggest a lack of women in *Angolares* communities, meaning a smaller female effective population size, which could have increased genetic drift on the X chromosome compared to the autosomes.

However, the patterns observed in clusters C1 and C2 may also be explained by sex-specific gene flows into *Angolares* communities. Historical records suggest that *Angolares* men mate with non-*Angolares* women living in plantations and neighbouring communities (Seibert 2004; Caldeira 2018). If these women carried non-African ancestry, their genetic contribution would be disproportionately represented on the X chromosome, lowering the level of African ancestry in this region of the genome and increasing the levels of long-IBD sharing between clusters C1 and C2 with cluster C3 on the X chromosome compared to the autosomes.

This latter scenario is also consistent with previous findings on uniparental markers and could explain the ROH patterns observed in cluster C1 in this study. The strong Y chromosome differentiation of *Angolares* groups compared to non-*Angolares*, together with low mitochondrial DNA differentiation between groups, suggests patrilocality and/or patrilineal descent in the *Angolares* (Coelho et al. 2008). Moreover, the small effective size of the *Angolares*, and the fact that gene flow from non-*Angolares* communities was female-mediated, may have led to inbreeding primarily in the male lineage. This could explain why individuals in cluster C1 have longer stretches of ROH on the autosomal chromosome 7 than on the X chromosome.

Clusters C4 and C5 in São Tomé, with Cabo Verdean descent, present genetic diversity patterns on the X chromosome and the autosomes more similar to those observed in Cabo Verde than to the rest of São Toméan clusters. The African genetic diversity is overall lower, but is always higher on the X chromosome than on the autosomes, as in Cabo Verdean populations. This suggests that more African females and European

males left descendants during the admixture process. According to our results on Cabo Verde and previous studies that explicitly tested this hypothesis (Korunes et al. 2022), ancestry-related sex bias occurred in Cabo Verde. The genetic diversity patterns observed in these particular São Toméan clusters most likely reflect patterns that formed in Cabo Verde prior to migration to São Tomé.

Moreover, our results suggest that more females than males migrated from Cabo Verde and left descendants in São Tomé – a hypothesis that need to be further investigated. We observed overall higher levels of long-IBD sharing between São Toméan clusters and Cabo Verdean populations when considering X chromosome markers compared to autosomal markers. Cluster C3, for instance, shared long IBD tracts with several Cabo Verdean populations on the X chromosome, but not on chromosome 7. Moreover, it shared more long-IBD tracts on the X chromosome than on chromosome 7 with cluster C5, resulting from gene flow between Cabo Verdeans and the local São Toméan population.

Finally, these differences in patterns of long-IBD sharing on the X chromosome and the autosomes further suggest that the five genetic clusters considered here, as defined in the previous study based on autosomal data, might not well represent the genome-wide genetic structure in São Tomé. A fineSTRUCTURE analysis based on X chromosome markers in São Tomé could reveal different clustering patterns, which would be interesting to investigate in future studies.

3.3.3 Founder effects in the archipelago of Cabo Verde

In Cabo Verde, we found a higher African genetic diversity on the X chromosome than on the autosomes in the seven populations from different islands of the archipelago: a significant difference in almost all populations. These findings suggest that more African females and European males might have contributed to the admixture process, as previously tested by other researchers under a model of constant admixture (Korunes et al. 2022). They also found considerable variation in the ratio of X to autosomal African genetic diversity across islands in Cabo Verde, as we observed in this study. This could mean that the extent of ancestry-related sex bias between the European and African source populations varied from island to island.

However, this could also result from the settlement history of Cabo Verde, involving several waves of migration and founder effects among islands that strongly impacted patterns of genetic diversity across the archipelago (Laurent et al. 2023). If the initial settlers of an island were already admixed, they would have carried pre-existing patterns of African and European genetic diversity. Thus, differences in patterns of genetic diversity on the X chromosome and the autosomes on that island would stem from ancestry-related sex bias that happened prior to the founding of its population.

Moreover, we observed higher long IBD sharing on the X chromosome compared to the autosomes in certain island pairs, suggesting female-biased migration between islands. Santiago, the first settled island of the archipelago, shares long-IBD segments with all other islands and exhibits higher long-IBD sharing on the X chromosome, although most comparisons with the autosomal chromosome 7 are not statistically significant. Sal, the last perennially settled island, shows significantly higher X chromosome IBD sharing with Santiago and several northern islands, possibly indicating increased female migration to Sal or strong founder effects that amplified pre-existing diversity patterns.

The results of this study are particularly relevant to investigate migrations in northern islands, which were considered as a single population in a previous study (Korunes et al. 2022). Santo Antão and São Vicente share high levels of recent genetic ancestry between each other, with significantly higher long IBD sharing on the X chromosome than on the autosomes. Historical records document substantial migration from Santo Antão to São Vicente (Albuquerque and Santos 1991), and our findings suggest that more females than males may have moved between these islands over time. Likewise, São Vicente and São Nicolau exhibited some of the highest levels of long IBD sharing on the X chromosome in the archipelago, which could also reflect a history of female-biased migration between these two islands.

Finally, we found significantly higher levels of long ROH on the X chromosome compared to autosomal chromosome 7 in Santiago, São Nicolau, and São Vicente. While higher ROH levels on the X chromosome are generally expected under neutrality, due to stronger effects of genetic drift and inbreeding on the X chromosome compared to autosomes, this pattern could also be shaped by sex-specific demographic processes. Specifically, a smaller effective population size of females compared to males and/or

mating between close relatives along the maternal lineage could contribute to the increase in ROH on the X chromosome (Cotter et al. 2024).

Future studies investigating ancestry-related sex bias in Cabo Verde should take into account the founder effects from one island to another, and if more females or males migrated among islands.

3.3.4 Proposing complex models of genetic admixture

I finally propose demographic admixture models incorporating sex-specific parameters, such as sex-biased gene flow and sex-specific demography, based on the hypotheses drawn on the previously discussed genetic findings. These complex demographic scenarios might be explicitly tested in future studies using simulations-based approaches (Fortes-Lima et al. 2021; Mas-Sandoval et al. 2023).

3.3.4.1 European genetic contribution to São Tomé

Our findings reveal different proportions of European genetic diversity on the X chromosome compared to the autosomes, suggesting the hypothesis of ancestry-related sex bias in São Tomé. To formally test this hypothesis, we need to consider a European source population in São Tomé (Figure 3.8) – a fundamental implementation to the demographic model of admixture presented in the previous chapter (Figure 2.11).

At first, we could test whether gene flow from this European source population more likely occurred at the founding of the São Toméan admixed population, or after the isolation event that gave rise to cluster C1. The latter scenario would more easily explain why we found virtually none genetic contribution from Europe in the *Angolares* of cluster C1 based on autosomal data. Finally, we could add sex-specific parameters to the admixture model to investigate whether gene flow from Africa or Europe was sex-biased and to what extent.

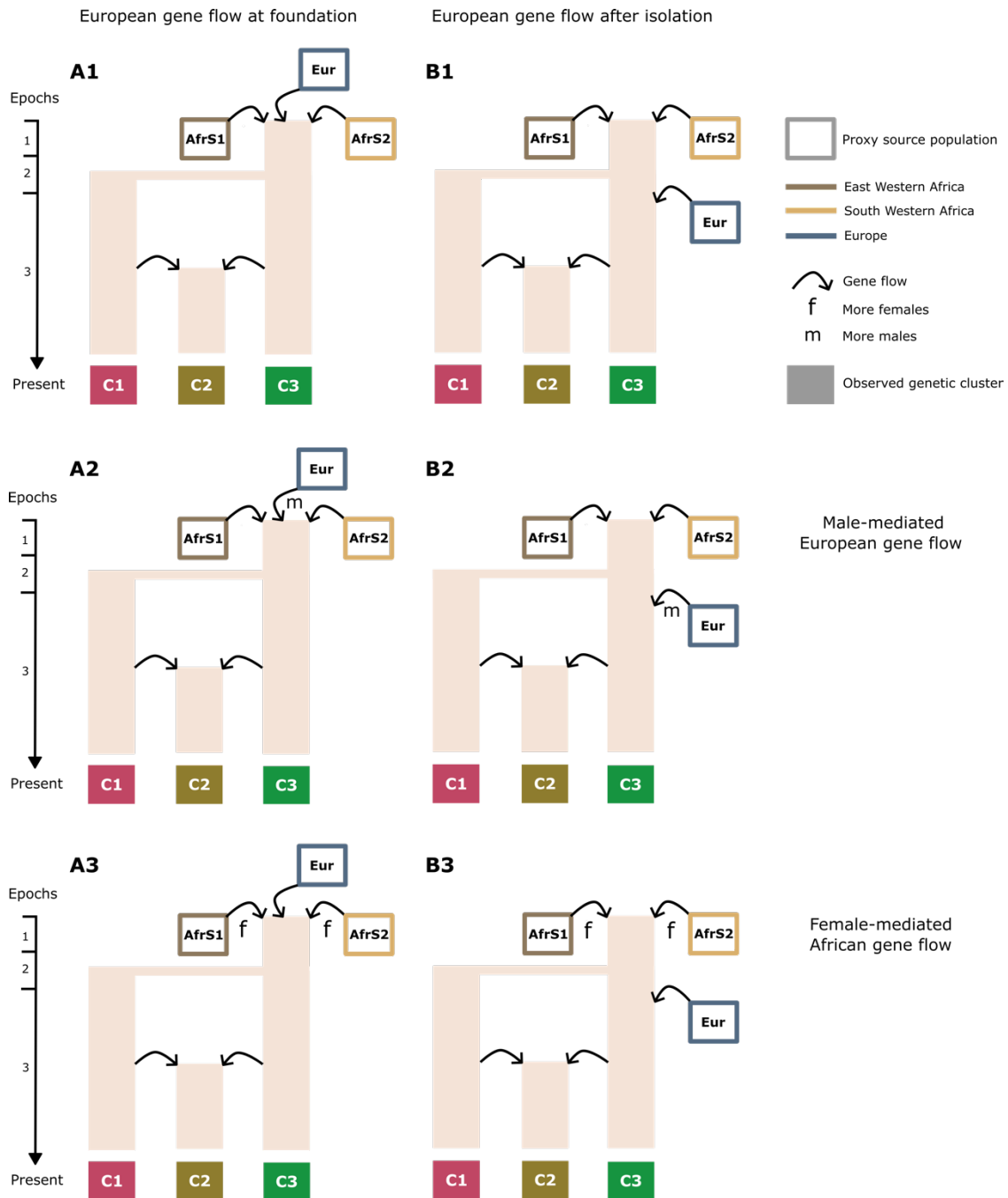


Figure 3.8. Demographic admixture scenarios for the C1, C2 and C3 clusters in São Tomé including gene flow from Europe.

Each model presents three alternative scenarios: Model A includes gene flow from Europe at the foundation of the admixed population in São Tomé, while Model B includes gene flow from Europe after the isolation event that give rise to cluster C1. Scenarios 1, 2, and 3 consider balanced sex-ratio from both Africa and Europe, male-mediated gene flow from Europe, and female-mediated gene flow from Africa, respectively. The extent of female and male gene flow can be estimated a posteriori on the scenario that best fit the observed genetic diversity.

3.3.4.2 The demographic history and nested admixture of the *Angolares*

Interestingly, the *Angolares* of São Tomé, represented by cluster C1 and C2, presented relatively higher levels of European genetic diversity on the X chromosome than on the autosomes. To investigate this further, we need to consider models in which the *Angolares* can receive some degree of gene flow of European ancestry tracts at some point in their demographic history.

If admixture with European source populations in São Tomé occurred before the isolation event that gave rise to the *Angolares* of cluster C1 (Scenario A in Figure 3.8), we could test whether a limited number of *Angolares* individuals contributing to reproduction, or an even smaller effective number of *Angolares* females than males, might be enough to explain the different genetic diversity patterns on the X chromosomes and the autosomes observed in clusters C1 and C2 compared to cluster C3 (Scenario A in Figure 3.9).

On the other hand, if admixture with European source populations occurred after isolation (Scenario B in Figure 3.8), we could test whether later ancestry-related sex bias might explain the observed patterns (Scenario B in Figure 3.9). For instance, to explain the introgression of European ancestry tracts particularly on the X chromosome after the isolation event, we could consider at least two alternative scenarios: 1) gene flow from European females, and 2) gene flow from São Toméan non-*Angolares* females carrying European ancestry tracts.

Finally, to explicitly test which of these scenarios best explain the observed genetic diversity of clusters C1 and C2, we would need to consider a model that incorporate both sex-specific effective population size and ancestry-related sex bias.

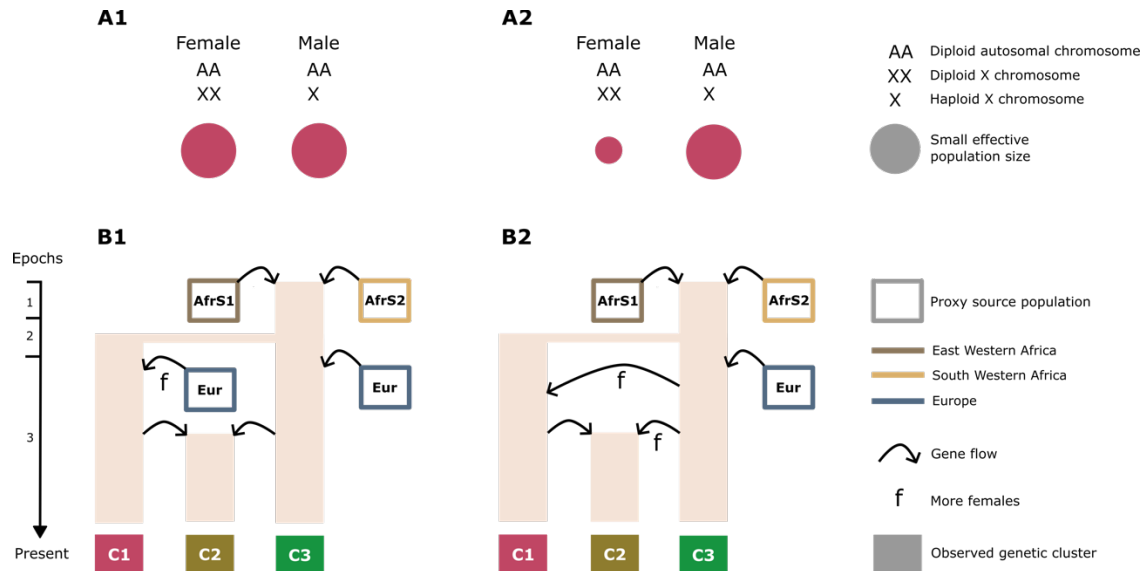


Figure 3.9. Alternative demographic scenarios for the *Angolares*, here represented by clusters C1 and C2 in São Tomé.

Model A focuses on sex-specific effective population sizes, while Model B explores sex-specific gene flow dynamics. In Model A, one scenario (A1) represents a limited number of *Angolares* females and males contributing to reproduction, whereas the other (A2) suggests an even smaller effective number of females compared to males. In Model B, ancestry-related sex biases are considered, with *Angolares* represented by cluster C1 and C2, and non-*Angolares* by cluster C3 in São Tomé. One scenario (B1) involves European females mating with *Angolares* males, while the other (B2) assumes that non-*Angolares* females with some European ancestry contributed to the *Angolares* gene pool through mating with *Angolares* males. A comprehensive model integrating both sex-specific effective population size (Model A) and gene flow (Model B) would provide a more complete understanding of the *Angolares* genetic diversity.

3.3.4.3 Ancestry-related sex bias in Cabo Verde and São Tomé

This study stresses the importance of considering sex-specific demographic histories in source populations to explain the genetic diversity of the admixed population. In particular, the higher African genetic diversity observed on the X chromosome compared to the autosomes in clusters C4 and C5, of Cabo Verdean descent in São Tomé, most probably reflect patterns of genetic diversity already formed in Cabo Verde. We thus need to consider ancestry-related sex bias in Cabo Verde prior to migration to São Tomé to address the following questions: Did more Cabo Verdean females or males migrated to São Tomé? Did more Cabo Verdean females or males admix with the local São Toméans?

I propose here a simplified model of admixture (Figure 3.10) accounting for: i) three source populations, including East Western Africa, South Western Africa, and Europe, contributing to the admixture event that founded the São Toméan population prior to admixture with Cabo Verdeans, here represented by clusters C1, C2, and C3; ii) two source populations, including West Western Africa and Europe, contributing more females and males, respectively, to the admixture event that founded the Cabo Verdean population, CV; iii) two source populations, deriving from the two distinct admixture events in i) and ii), including São Toméans and Cabo Verdeans contributing to the admixture event that resulted in clusters C4 and C5 in São Tomé.

Finally, to formally test whether more Cabo Verdean females or males contributed to the observed genetic diversity of these two clusters, we would need to consider three alternative scenarios: 1) same number of Cabo Verdean females and males migrating to São Tomé; 2) a greater number of Cabo Verdean females; and 3) a greater number of Cabo Verdean males.

Another aspect to take into account is that the source populations, as well as the admixed population, could be genetically structured. Thus, migrants contributing to a particular admixture process may not represent the entire source population. For example, social structures in Cabo Verde may also have influenced genetic structures within the archipelago. Due to the harsh working conditions of the indentured labour system, Cabo Verdeans from lower socioeconomic backgrounds, who may have had different levels of African and European ancestry to the rest of the Cabo Verdean population, may have been more likely to migrate to São Tomé. If the genetic structure in Cabo Verde was at

least partly explained by differences in levels of African and European genetic ancestry, future studies accounting for this genetic structure would be able to address the following question: Have Cabo Verdeans with higher or lower African ancestry, compared to the rest of the archipelago's population, migrated to São Tomé?

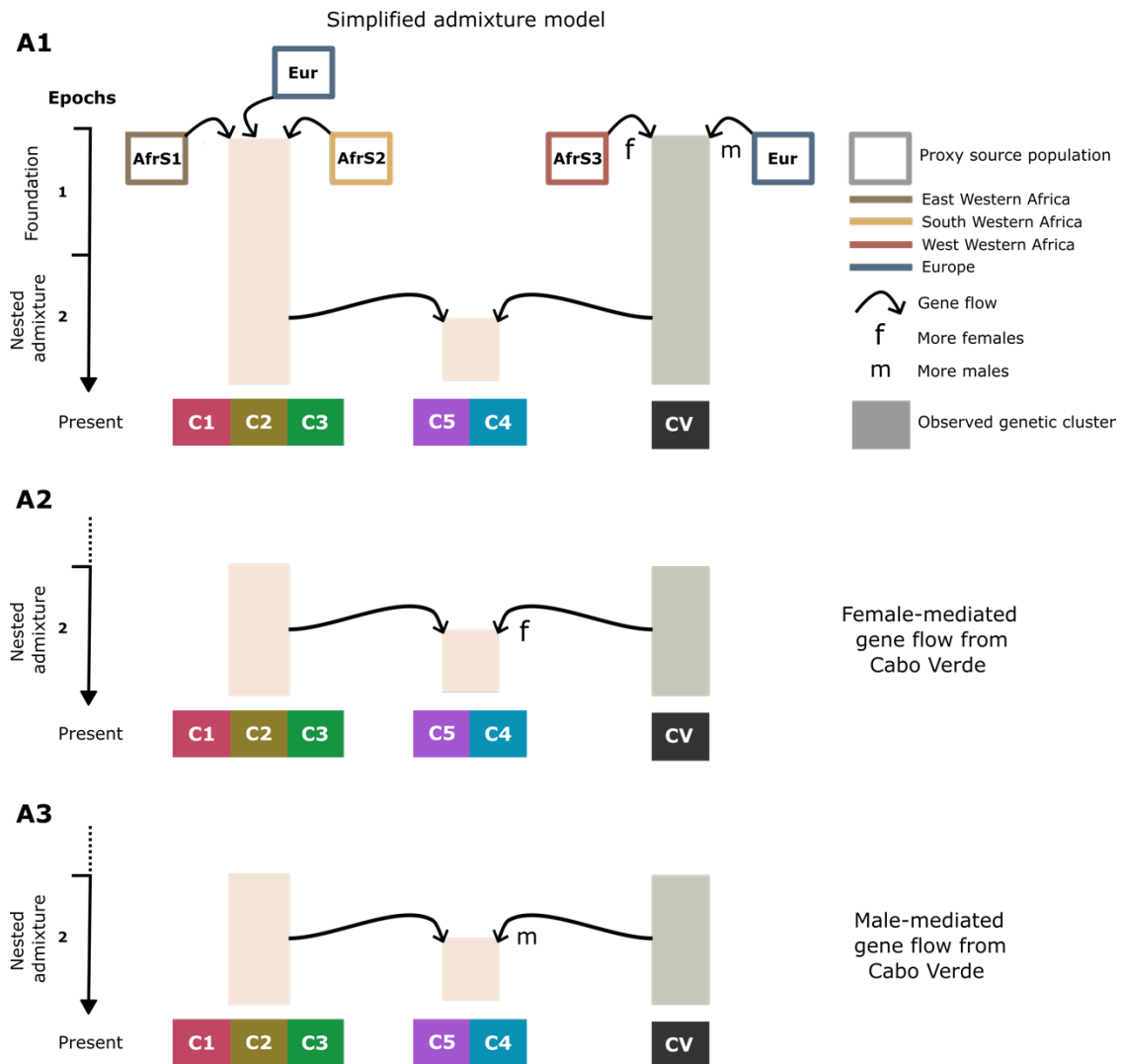


Figure 3.10. Simplified admixture model of Cabo Verdean migration to São Tomé.

The São Tomé group is represented as a single population comprising clusters C1, C2 and C3, while the Cabo Verdean group is represented by the CV sample. These groups contributed to the formation of the Cabo Verdean descent group in São Tomé, represented by the combined clusters C4 and C5. In the first epoch, the model assumes sex-biased contributions from African and European ancestry in the formation of the Cabo Verdean gene pool. In the second epoch, three migration scenarios are considered: (A1) sex-balanced migration from Cabo Verde to São Tomé, (A2) a higher proportion of female migrants, and (A3) a higher proportion of male migrants. Explicit tests are needed to determine which of these scenarios best explains the observed genetic diversity in clusters C4 and C5 in São Tomé.

3.4 Conclusions and Perspectives

Using a single autosomal chromosome with a comparable number of markers or overall size to the X chromosome seems to be good approach for comparing genetic diversity patterns on these two chromosomal regions. However, a single autosomal chromosome may not represent the genetic variation of all autosomes. Future studies may wish to compare results on the X chromosome with those on other autosomal chromosomes and develop statistical tests to assess whether the selection of a particular autosomal chromosome significantly affects the results. Nevertheless, based on this approach, we were able to highlight interesting differences in genetic diversity patterns on the X chromosome and autosomes in São Tomé and Cabo Verde, suggesting hypothesis of ancestry-related sex bias and sex-specific demography.

The results of genotype-based clustering analysis considered in this study are not well suited to disentangling contributions from different source populations. In São Tomé, the blue cluster proportions in individual genomes reflect contributions from different African sources, including African ancestry tracts from admixed Cabo Verdean genomes. In both São Tomé and Cabo Verde, recently admixed groups of African and European descent have themselves been sources of further admixture events. Future studies should focus on disentangling the genetic contributions of African source populations, including recently admixed groups, on the X chromosome as previously done on the autosomes in São Tomé and Cabo Verde using Local Ancestry Inference approaches.

One of the main highlights of this study is that a comprehensive understanding of genetic diversity in admixed populations requires a thorough examination of the genetic diversity and demographic history of their source populations. In São Tomé, the patterns of genetic diversity on the autosomes and the X chromosome of some genetic clusters likely reflect the patterns already formed in Cabo Verde prior to migration to São Tomé. Accounting for ancestry-related sex bias in Cabo Verde may thus enhance our understanding of Cabo Verdean migrations to São Tomé (Figure 3.10). More generally, source populations may exhibit patterns of genetic diversity that are themselves shaped by gender and social factors influencing sex-specific demography, further complicating the disentangling of sex-specific demographic processes that affected genetic diversity in the admixed population.

The observed patterns of genetic diversity in São Tomé and Cabo Verde reflect complex demographic histories shaped by multiple migration waves, founder effects, and sex-biased processes. Understanding these histories requires statistical frameworks capable of testing alternative demographic scenarios. Maximum-likelihood approaches struggle with highly complex admixture histories involving multiple pulses and recurrent admixture events, often rendering explicit likelihood calculations infeasible (Gravel 2012; Hellenthal et al. 2014). Approximate Bayesian Computation (ABC) and other simulations-based approaches, on the other hand, has been successfully used to reconstruct the genetic history of populations descended from enslaved Africans, allowing formal model comparison and parameter estimation (Fortes-Lima et al. 2021; Laurent et al. 2023; Mas-Sandoval et al. 2023).

CONCLUSIONS AND PERSPECTIVES

In this thesis, I reconstructed complex admixture histories of populations descended from enslaved Africans, particularly populations currently living on the island of São Tomé, in the Gulf of Guinea, and in the archipelago of Cabo Verde, off the coasts of Senegal. In chapter 1, I explored how different historical and social contexts influenced genetic admixture processes in populations descended from enslaved Africans on both sides of the Atlantic. I provided a critical review of some of the key questions addressed in previous genetic studies, and explored how genetic findings can be interpreted and analysed in relation to the unique histories and social dynamics of each population. In chapter 2 and 3, I focused on the admixed population of São Tomé, the African Atlantic island that became the site of one of the first slave societies based on plantation organization and economy. The fluctuation of plantation economies in São Tomé led to significant demographic shifts and social stratification associated with intensive slavery, marooning, and indentured labour over five centuries of Portuguese colonization, which make of this island a unique setting to investigate the impact of complex admixture histories on genetic diversity in TAST-related societies.

The genetic impact of historical migrations in São Tomé

Our results based on genetic data in São Tomé largely align with historical records on the successive waves of migration to the island. I identified genetic admixture events between populations of different origins in Africa and Europe that happened since the beginning of the colonization of the island and the slave trade until after the abolition of slavery.

São Toméans mostly result from admixture between East Western African and West Southern African source populations. This is in line with the historically known routes of recruitment of enslaved Africans during the first plantation economy from the 15th to the 18th century (Caldeira 2011). São Tomé became the world's leading producer of sugar cane due to the exploitation of slave labour force deported from the Kingdom of Benin and the "Gold Coast" in the Gulf of Guinea, where the Portuguese established commercial relations that were the precursors of the triangular trade (Vogt 1973); and, shortly afterwards, from the Kingdom of Kongo, a great empire that stretched from the north of Angola, to the west of the Democratic Republic of Congo, and to the south of present-day Gabon and Congo (Caldeira 2021).

The timing of this admixture process requires further investigation. Our results suggest that admixture between these two sources occurred in the 17th century (Figure 2.10), well after the arrival of enslaved individuals on the island, and the formation of Creole languages currently spoken on the island. Notably, these languages feature syntactic elements from Edo languages, spoken in present-day Nigeria, as well as lexical influences from Portuguese as well as Bantu languages, spoken in present-day Angola (Hagemeijer 2011; Hagemeijer and Rocha 2019). This indicates that genetic admixture between Africans of different origins occurred later than the initial arrival of slaves and the emergence of Creole languages. However, this finding may primarily reflect the limitations of the admixture models considered in this study, which assume a single pulse of admixture and random mating. More complex scenarios, considering recurrent admixture or assortative mating, could potentially place the initial genetic admixture earlier (Zaitlen et al. 2017; Korunes et al. 2022).

We found a relatively small European genetic contribution in São Tomé. This was much less than what observed in other populations descended from both sides of the Atlantic, such as Cabo Verdeans, African Americans in the US and African Caribbeans in Barbados, which we also analysed in this study.

This result is in line with the particular history of Portuguese colonisation and settlement of the island. According to demographic records, Europeans were always outnumbered by Africans, both free and enslaved, and almost completely left the island during periods of economic decline. However, historical records suggest that admixture with Africans was somewhat tolerated at the beginning of the colonization, in line with the Portuguese effort to populate the island (Caldeira 2007). In Chapter 2, we propose that the European genetic contribution may have occurred preferentially at the founding of the admixed populations. The signal of this admixture event may have diminished over time due to successive migrations that represented significant demographic shifts on the island. In Chapter 3, we present admixture models that can be tested in future studies to determine the extent to which Europeans contributed to the admixture process in São Tomé, when this admixture may have occurred, and whether gene flow was sex-biased (Figure 3.8).

We found a particularly high Cabo Verdean genetic contribution in São Tomé. Interestingly, almost half of the European haplotypic ancestry observed in São Tomé actually derive from Cabo Verdean admixed genomes (Figure 2.8).

The migrations associated with indentured servitude from Angola, Mozambique and Cabo Verde represented a major demographic shift on the island in the 19th and 20th century (Clarence-Smith 1976; Nascimento 2003; Nascimento 2008; Nascimento 2023). Our study reveals that these migrations have also resulted in a significant change of the genetic diversity currently present on the island. We observed a great genetic contribution associated with these migrations, particularly from the archipelago of Cabo Verde. We also identified a smaller genetic contribution from Mozambique. In the case of Angola, which was part of the slave routes during the first and second plantation economies in São Tomé, it was more difficult to identify which DNA segments are linked to the migrations of either era. Future studies based on refined Local Ancestry Inference could eventually permit to address this question by examining the length of genetic ancestry tracts from West Southern Africa in São Tomé.

The relative genetic impact of historical migrations from these three countries may reflect the specific characteristics of the São Toméan sample analysed in this study. Sampling in São Tomé was not explicitly targeted at specific subgroups and individuals from the main inhabited areas of the island were included in the sample. Nevertheless, as with any population sample, it may not fully represent the genetic diversity of the entire population. The observed genetic results, particularly the higher genetic contribution from Cabo Verde compared to Mozambique, could be explained by several factors, which are not mutually exclusive: (a) communities descended from Mozambicans may be underrepresented in the sample by chance; (b) the Cabo Verdean migration to São Tomé may have had a greater genetic impact on the island's population than Mozambican migration; and/or (c) communities descended from Mozambicans may be more socially or geographically isolated, potentially resulting in their limited genetic contribution to the rest of the island's population and underrepresentation in the current sample. If this is the case, further targeted sampling including known communities of Mozambican descent could help clarify the extent of genetic contributions from Mozambique.

Overall, these findings suggest that at least some of the indentured workers from Cabo Verde, Mozambique, and probably Angola, must have remained on the island after their contracts ended or left descendants there, thereby contributing to the genetic diversity of present-day São Toméans. This is an interesting result in itself, given that the difficulty faced by indentured workers in São Tomé in returning to their home countries at the end

of their contracts was at the centre of a debate in Europe about the persistence of slavery in the Portuguese Empire after its official abolition in 1869 (Higgs 2013).

By studying the genomes of a sample of São Toméans, we have been able to retrace the history of settlement of the island in great detail. The successive admixture events here discussed actually created a mosaic of genetic diversity in São Toméan genomes. This mosaic is not a random mix, but further reflects the complex social dynamics that influenced the interaction between populations and communities on the island during five centuries of Portuguese colonisation.

The evolution of genetic structures

We observed a striking genetic differentiation in the admixed population of São Tomé. We explored various methods of clustering individuals into distinct groups with the aim to analyse the demographic history of admixture of each group separately. Based on patterns of haplotype-sharing within the island, we identified five genetic clusters in São Tomé that reflect the current island's genetic structure.

Our findings provide valuable insights into the evolution of genetic structures during complex processes of genetic admixture in stratified societies. Historical migrations from Africa and Europe had different impacts across the island and gene flow between previously isolated communities have shaped the distribution of genetic diversity on the island over time.

Groups of individuals resulting from the same admixture event formed distinct genetic clusters in São Tomé due to reproductive isolation. Community boundaries limit mating between individuals from distinct social groups and eventually result in different patterns of genetic diversity accumulating in these groups over time. This process is particularly evident in the demographic history of one genetic cluster in São Tomé, which is genetically very differentiated from the rest of the island population due to strong genetic drift, despite having a similar composition of African genetic ancestry as the majority of the island population. This cluster likely represents the descendants of runaway slaves who formed maroon communities since the beginning of colonisation in the late 15th century, and remained relatively isolated until the 19th century, when some of these communities become known as the *Angolares* (Seibert 1998). We suggest that the *Angolares* result from the isolation of a subset of the founding populations of the

island, and that the founder effect, small effective population size and inbreeding might have caused the strong signals of rapid genetic differentiation from the rest of the island population.

Groups of individuals resulting from previous admixture events and successive isolation on the island eventually admixed with each other. Changes in the historical and social context can break down some community boundaries, creating new ones and resulting in new forms of social stratification. This can eventually lead to increased mating between previously isolated admixed groups. In São Tomé, for example, one genetic cluster presented intermediate genetic diversity patterns and recent shared genetic ancestry between the cluster representing the *Angolares* and that representing the *Forros*, the descendants of freed African slaves who became the dominant social group in São Tomé between the two plantation economies (Seibert 2015). These findings suggest that admixture occurred between the *Angolares* and the *Forros*. This could have resulted from the occasional or long-term contact between the *Angolares* and the rest of the São Toméan population during their period of relative isolation, or it could have occurred more intensely after the 19th century, when the *Angolares* had to establish commercial and social links with the other inhabitants of the island.

Another group of individuals seems to have been formed by relatively high levels of admixture between Cabo Verdean migrants and the local São Toméan population, which both derive from complex histories of admixture in Cabo Verde and São Tomé, respectively. Historically, contract workers were confined to the *roças*, the large colonial coffee and cocoa plantations, and had little contact with the rest of the São Toméan population, who refused to work on the plantations (Kiesow 2017). In line with this history of discrimination, one genetic cluster in São Tomé mainly result from Cabo Verdean genetic contributions, indicating that these Cabo Verdeans have remained relatively isolated from the rest of the island's population. Nevertheless, the gene flow between São Toméans and Cabo Verdeans observed in the former genetic cluster may be explained by a certain permeability of the social barriers to reproduction between indentured labourers and the rest of the island's population, or by the weakening of this social structure with the end of the plantation economy in São Tomé.

Overall, it is striking to observe the dismantling of a pre-existing genetic structure and the formation of a new one over only a few generations. In other words, the genetic

structure of the population is not fixed: it evolved in response to changes in the social boundaries between communities over time.

Mating patterns related to genetic ancestry

The joint analysis of markers on the X chromosome and the autosomes suggest that ancestry-related mating patterns affected genome-wide genetic diversity in São Tomé. Following previous genetic studies (Goldberg and Rosenberg 2015; Goldberg et al. 2020; Mas-Sandoval et al. 2023), we define these processes: a) ancestry-related sex bias, when female and male contributions from source populations are not equal during the admixture process, and b) ancestry-related assortative mating, when mating occurs preferentially between individuals with similar genetic ancestry in the admixed population. These two processes can be linked to the unique historical and social contexts of populations descended from enslaved Africans on both sides of the Atlantic, where laws and norms were intended to promote or restrict unions between specific pairs of individuals based on their presumed genetic ancestry.

The differences in genetic ancestry patterns on the X chromosome (inherited in two copies by females and one copy by males) and the autosomes (inherited in two copies by all individuals) allow us to infer mating patterns related to genetic ancestry. However, when comparing these two genomic regions, it is important to consider their inherent differences. In this study, I used for comparison a single autosomal chromosome with a similar number of markers or overall size to the X chromosome. Nevertheless, one autosomal chromosome may not represent the genetic variation of all autosomes, and statistical tests should be developed to assess the impact of selecting a particular chromosome on the results.

In chapter 3, I further underline the importance of accounting for ancestry-related mating patterns in source populations, as their genetic diversity may have also been affected by social and gender dynamics prior to admixture. In other words, genetic contributions to the admixture process may bear signals of genetic diversity resulting from ancestry-related mating patterns specific to source populations. One evident example in São Tomé are the patterns of African and European genetic ancestry observed in particular in two genetic clusters. The higher African genetic ancestry on the X chromosome compared to the autosomes in two clusters with Cabo Verdean descent in São Tomé likely reflects patterns already formed in Cabo Verde prior to migration to São Tomé.

Accounting for ancestry-related sex bias in Cabo Verde may thus enhance our understanding of Cabo Verdean migrations to São Tomé (Figure 3.10).

Apart from sex biased processes specific to admixture histories, the genetic diversity of the X chromosome can be further influenced by sex-specific demography, meaning differences in female and male respective migration and effective population size. These differences may be caused by a variety of social factors, including sex-biased mortality rates, the cultural transmission of reproductive success, rules of residence and descent, etc. (Austerlitz and Heyer 1998; Segurel et al. 2008; Heyer et al. 2012; Verdu et al. 2013; Guez et al. 2023). In the introduction of Chapter 3, I highlight that social and gender dynamics in stratified societies may influence both ancestry-related mating patterns and sex-specific demography, complicating the inference of the impact of such processes on the genetic diversity of the X chromosome and the autosomes. We thus need a better theoretical understanding of the joint impact of these processes on genome-wide genetic diversity.

Complex models of genetic admixture

In Chapters 2 and 3, I propose complex models of genetic admixture that could be tested in future studies to infer which scenarios best explain the observed genetic diversity. Such inference requires statistical frameworks capable of considering large numbers of parameters and high levels of complexities. Recent advances have made progress in this direction, particularly through simulation-based methods that can model multiple pulses of admixture and recurrent gene flow (e.g., Fortes-Lima et al. 2021; Laurent et al. 2023), or account for ancestry-related sex bias and assortative mating (Mas-Sandoval et al. 2023). Future studies should aim to model these processes jointly, and allow sex-specific parameters to vary over time and between communities, to more accurately reflect shifts in historical and social contexts.

The admixture histories reconstructed for São Tomé in this study highlight important aspects to consider in future complex admixture modeling approaches. For example, to understand the impact of social stratification on genetic patterns, we could model stratification within the admixed population, with different subgroups experiencing different admixture and demographic processes. Similarly, we could model stratification within the source populations, as migrants involved in admixture may have come from specific subgroups rather than be representative of the broader source population.

Furthermore, this study on São Tomé highlights the need to consider nested admixture events. Modelling source populations as both structured and admixed, rather than considering them as homogeneous and static entities, would allow to better understand the complexities of admixture histories.

However, we must be realistic about the challenges that come with increasing model complexity. Complex models can be computationally demanding and it may be harder to distinguish between competing scenarios or to obtain reliable parameter estimates. Starting with an extensive genome-wide analysis of genetic diversity, as I have done in this study, helps to identify the most relevant demographic processes and the most informative patterns of genetic diversity. For example, haplotype-based information and identity-by-descent (IBD) sharing were particularly useful for identifying recent gene flow between admixed groups in São Tomé and Cabo Verde. These insights are crucial for both model design and the selection of summary statistics, particularly when using methods such as Approximate Bayesian Computation (ABC), which are highly sensitive to these complex choices.

Nested admixture

In the current population genetics theoretical framework, we must position ourselves at a certain evolutionary scale to assume that the population of interest is admixed between two or more source populations. These source populations must have been isolated from each other for a relatively long time, and their previous history of admixture is not taken into account. This fundamental assumption is at the basis of what we currently define genetic ancestry. However, source populations are not static entities, but they are themselves resulting from admixture processes at some point in their evolutionary history. As all human populations are to some extent the result of successive, nested, admixture events, studying the effects of gene flow between admixed groups is essential to understanding genetic admixture processes more broadly.

In this study, I introduced the expression “nested admixture” to describe this well-known process, which however lacked a precise term to address it in genetic studies. This process creates layers of genetic ancestry in the resulting admixed groups, making it difficult to identify the exact source populations of certain genetic tracts using current methods. By naming this process, I aim to draw attention and encourage the

development of analytical frameworks that can consider source populations as resulting from complex admixture processes themselves.

Genetic ancestry labels

Language plays a crucial role in scientific investigation, influencing how researchers conceptualise ideas, conduct research, and communicate results. Genetic ancestry labels such as “African” or “European” imply that there are clear boundaries between otherwise homogeneous human groups. However, genetic studies demonstrate that human genetic diversity is better explained by clinal distributions shaped by complex genetic admixture processes over time and across space (Peter et al. 2020). Anthropological geneticists are beginning to acknowledge the limitations of using these categories, instead of more continuous definitions of genetic ancestry, and aim to develop new models and methods that account for this complexity.

Nested admixture in São Tomé highlight the limitations of using genetic ancestry labels. What do we define as “European” ancestry in São Tomé, given that most of the genetic tracts that share similar genetic diversity with European populations actually come from admixed Cabo Verdean genomes? In other words, at what point does a piece of DNA cease to be classified as originating from one place and instead be classified as originating from somewhere else?

Similarly, what do we define as “African” ancestry in São Tomé? One approach could be to exclude from Cabo Verdean genetic contributions the genetic ancestry tracts with “European” ancestry, including only those with “African” ancestry in Cabo Verdean admixed genomes. Alternatively, we could include all genetic contributions from populations currently living in Africa, and thus considering Cabo Verdean genetic contributions as “African” ancestry as a whole. However, ultimately, what would it mean “African” ancestry in São Tomé, since São Tomé is in Africa?

Another possibility is to differentiate “African” ancestry in São Tomé based on the historical context. On the one hand, we could specifically consider the “African” ancestry resulting from the TAST migrations to São Tomé, assuming that it is best represented by genetic tracts similar to those of contemporary West Central and West Southern African populations. On the other hand, genetic contributions from Cabo Verde and Mozambique could be considered as a distinct “African” ancestry resulting from

indentured labour migrations. However, historical records indicate that indentured labourers also came from regions such as Angola in West Southern Africa, and it is difficult to separate “West Southern African” genetic tracts resulting from TAST migrations from those resulting from more recent indentured labour migrations using current methods.

Finally, in São Tomé, we used genetic clustering to group genetically similar individuals, rather than relying on pre-labelled samples based on non-genetic information. This approach allowed us to minimise prior assumptions while investigating the genetic structure of the admixed population. I only interpreted the genetic diversity of each cluster after defining and analysing it. However, my interpretation of the observed genetic diversity, such as the hypothesis of nested admixture between “São Toméans” and “Cabo Verdeans” or between the “*Angolares*” and the “*Forros*”, which are again genetic ancestry labels, is highly influenced by the methodological framework we used. Notably, I decided to split the São Toméan sample in five genetic clusters based on haplotype-sharing patterns across the island. However, considering more or fewer genetic clusters, or using alternative clustering approaches, could lead to different interpretations of the island’s genetic structure.

Communicating results to a wider audience

Ultimately, what “African” and “European”, “São Toméan” and “Cabo Verdean”, “*Angolares*” and “*Forros*” mean cannot be determined by studies of human genetic variation. Instead, these categories are rooted in historical, social, and political contexts and are intimately linked to issues of collective identity. As population geneticists, we should pay attention, as the use of these labels in studies of genetic admixture may somehow interact with how individuals and groups perceive their identity.

Communicating results to a wider audience forcibly involves some extent of simplifying the assumptions behind these labels. This issue is even more pronounced in the case of personalised genetic tests offered by private companies, which frame individual genetic diversity within categories of genetic ancestry that are complex to understand even for experts in the field (Abel and Schroeder 2020). Furthermore, private companies generally do not provide transparent definitions of genetic ancestry and descriptions of the data and methods used to construct these categories, making impossible the

understanding and reproducibility of the results that is at the core of scientific investigation.

The use of population labels in genetic studies has been recently critically reviewed in the NASEM report (National Academies of Sciences, Engineering, and Medicine 2023). As researchers, we need to be explicit about what we mean by genetic ancestry in our studies. One possible solution might be to stop using “genetic ancestry” and instead prefer expressions that are more explicit about the underlying assumptions, such as “genetic similarity to a set of contemporary population samples that are assumed to be sources”, as some authors have provocatively suggested (Coop 2022).

In line with these considerations, in chapter 2 and 3 of this thesis, I introduced the population samples used as proxies for source populations before referring to the genetic contributions of these sources as the “genetic ancestry” of the admixed populations. I also opted for the expression “haplotypic ancestry” when inferring genetic ancestry from shared haplotypes to emphasise the method used for such inference. Finally, in the description of the results, I referred to genetic clusters in São Tomé as clusters C1 to C5, and in the discussion, I explicitly explained how and why I attempted to make correlations between genetic clusters and social groups in São Tomé based on historical, linguistic, and anthropological information available on the island.

I am aware that these measures may not be sufficient to avoid misinterpretations, and I am committed to paying attention as much as possible to how the scientific community, the general public and particularly the interested communities in São Tomé may interpret and use these genetic findings.

Final considerations on São Tomé and Cabo Verde

A key objective of this thesis was to investigate jointly and compare the impact of different admixture processes on the current genetic diversity of admixed populations from São Tomé and Cabo Verde. These African Atlantic islands were colonised by the same colonial power for five centuries, from the first Portuguese settlement in the late 15th century until independence from Portugal in 1975 (Seibert 2014). However, the two colonies' distinct historical and social contexts still resulted in profoundly different genetic admixture histories.

Genetic contributions from different African populations reflect the distinct routes of slave trade to São Tomé and Cabo Verde. In the case of São Tomé, the African haplotypic ancestry linked to the TAST primarily traces back to the nearby Gulf of Guinea and the Congo-Angola regions to the south. São Toméan merchants were actively involved in the slave trade with local African kingdoms in these regions. Some of the most important ports of embarkation of the TAST were located there, and amongst the largest overall numbers of enslaved Africans that reached the Americas were forcibly taken from these regions (Caldeira 2013). In contrast, Cabo Verdean African haplotypic ancestry traces almost entirely to the Senegambia region, located just offshore from the archipelago, which was a relatively important region for slave trade during the early period of the TAST (Eltis and Richardson 2010). Although historical records indicate that Cabo Verdeans engaged in slave trade with other African regions, the lack of genetic contributions from these areas in the present-day population suggests that individuals enslaved from these regions did not leave descendants on the islands (Laurent et al. 2023). Instead, they were probably shipped to other destinations where the demand for enslaved labour was higher. Indeed, both São Tomé and Cabo Verde functioned as both destinations and critical transit hubs within the broader network of the TAST.

The differing levels of European haplotypic ancestry in São Tomé and Cabo Verde are closely linked to the distinct types of slave societies that developed on the islands. Present-day São Toméans showed relatively low levels of European haplotypic ancestry, especially compared to Cabo Verdeans. This pattern is more similar to that observed in other populations in the Americas with a strong African genetic ancestry and a history of plantation-based slavery. In São Tomé, a plantation-based economy developed early, relying heavily on a large enslaved African population (Caldeira 2011).

The plantation system might have prevented social integration and admixed unions between Europeans and Africans. When the plantation economy collapsed, many Europeans abandoned the island, leaving behind a population that was likely largely descended from Africans with limited European genetic ancestry. In contrast, Cabo Verde did not develop a plantation economy to the same extent. Instead, the trade-based economy relied on a smaller number of enslaved individuals working mainly as domestic servants. Over time, admixed individuals formed creole elites that remained influential throughout Cabo Verde's history (Carreira 1972). The cultural transmission of reproductive success of European and admixed individuals in Cabo Verde could thus explain the relative higher European haplotypic ancestry found in present-day Cabo Verdean admixed genomes. Moreover, from the beginning of the colonization, enslaved Africans in São Tomé managed to escape the plantations and form maroon communities, and their descendants, the *Angolares*, exhibited little to no European genetic ancestry.

Admixture in Cabo Verde, and possibly in São Tomé, occurred mostly between African females and European males. Both São Tomé and Cabo Verde were remote colonies affected by tropical diseases, which discouraged permanent European settlement (Seibert 2014). Only a few European families settled on the islands, and the European settlers, who were mostly male, likely admixed with enslaved African females in line with the Portuguese colonial policy that encouraged admixture in order to populate the islands. In Cabo Verde, the significantly higher African ancestry on the X chromosome compared to the autosomes strongly supports this hypothesis of ancestry-related sex-bias during admixture. However, the situation in São Tomé is more complex. The limited overall European genetic contribution makes interpretations difficult, and admixture patterns vary across genetic clusters. São Toméan clusters with Cabo Verdean ancestry exhibited elevated African ancestry on the X chromosome, reflecting Cabo Verde's admixture dynamics. Another São Toméan cluster showed only a slight and non-significant increase in African ancestry on the X chromosome, and further analysis is required to clarify this pattern. Interestingly, the cluster representing the *Angolares* exhibited the opposite trend, namely higher African ancestry on the autosomes than on the X chromosome. This may be explained by admixture between non-*Angolares* females, with some European ancestry, and *Angolares* males, in line with historical accounts (Figure 3.9).

The Cabo Verdean migration to São Tomé marked a major demographic and social shift, with admixture mainly between Cabo Verdean females and São Toméan males. When the second plantation economy was established in the 19th century, many local São Toméans resisted working under the harsh conditions of the *roças*, leading colonial authorities to bring enslaved individuals and contracted laborers from other colonies, including Cabo Verde (Nascimento 2003). This contributed to a new form of social stratification, which is reflected in our genetic findings. We identified one genetic cluster in São Tomé with almost entirely Cabo Verdean ancestry, consistent with a history of discrimination that may have limited gene flow with local communities. However, we also observed another cluster that likely resulted from higher levels of admixture between Cabo Verdean immigrants and local São Toméan individuals. This suggests that, despite social barriers, mating occurred, perhaps due to more recent shifts in social and historical contexts. Interestingly, this admixture also appears to have been sex-biased, with Cabo Verdean females mating more frequently with São Toméan males; however, this process requires further investigation.

Finally, my findings highlight the importance of considering the unique historical and social contexts of populations descended from enslaved Africans when interpreting genetic data, as populations with relatively similar colonial experiences may have been shaped by very different admixture histories. In São Tomé and Cabo Verde, distinct slave trade routes brought different African populations to each island, and the development of a plantation-based slave society in São Tomé, with major fluctuations over time, had a profound impact on the genetic diversity of its admixed population.

REFERENCES

- Abel S, Schroeder H. 2020. From Country Marks to DNA Markers: The Genomic Turn in the Reconstruction of African Identities. *Current Anthropology* 61:S198–S209.
- Agranat-Tamir L, Mooney JA, Rosenberg NA. 2024. Counting the genetic ancestors from source populations in members of an admixed population. *Genetics* 226:iyae011.
- Albuquerque L de, Santos M. 1991. História geral de Cabo Verde. 1st ed. Lisboa: Centro de Estudos de História e Cartografia Antiga
- Alexander DH, Lange K. 2011. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics* 12:246.
- Alexander DH, Novembre J, Lange K. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research* 19:1655–1664.
- Allen RB. 2014. Slaves, Convicts, Abolitionism and the Global Origins of the Post-Emancipation Indentured Labor System. *Slavery & Abolition* 35:328–348.
- Almeida J, Fehn A-M, Ferreira M, Machado T, Hagemerijer T, Rocha J, Gayà-Vidal M. 2021. The Genes of Freedom: Genome-Wide Insights into Marronage, Admixture and Ethnogenesis in the Gulf of Guinea. *Genes* 12:833.
- Almeida Mendes A. 2008. Les réseaux de la traite ibérique dans l'Atlantique nord (1440-1640). *Annales. Histoire, Sciences Sociales* 63e année:739–768.
- Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, et al. 2010. Integrating common and rare genetic variation in diverse human populations. *Nature* 467:52–58.
- Alves-Silva J, da Silva Santos M, Guimarães PE, Ferreira AC, Bandelt HJ, Pena SD, Prado VF. 2000. The ancestry of Brazilian mtDNA lineages. *Am J Hum Genet* 67:444–461.
- Austerlitz F, Heyer E. 1998. Social transmission of reproductive behavior increases frequency of inherited disorders in a young-expanding population. *Proceedings of the National Academy of Sciences of the United States of America* 95.
- Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. *Nature* 526:68–74.
- Baharian S, Barakatt M, Gignoux CR, Shringarpure S, Errington J, Blot WJ, Bustamante C, Kenny EE, Williams SM, Aldrich MC, et al. 2016. The Great Migration and African-American genomic diversity. *PLOS Genetics* 12:e1006059.
- Baptista M. 2015. Continuum and variation in Creoles: Out of many voices, one language. *Journal of Pidgin and Creole Languages* 30:225–264.
- Behr AA, Liu KZ, Liu-Fang G, Nakka P, Ramachandran S. 2016. pong: fast analysis and visualization of latent clusters in population genetic data. *Bioinformatics* 32:2817–2823.

- Beleza S, Campos J, Lopes J, Araújo II, Almada AH, Silva AC e, Parra EJ, Rocha J. 2012. The Admixture Structure and Genetic Variation of the Archipelago of Cape Verde and Its Implications for Admixture Mapping Studies. *PLoS One* 7:e51103.
- Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, Campos J, Araújo II, Anderson TM, Vilhjálmsson BJ, et al. 2013. Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. Spritz RA, editor. *PLoS Genet* 9:e1003372.
- Berlin I. 1998. Many Thousands Gone: The First Two Centuries of Slavery in North America. Harvard University Press
- Blant A, Kwong M, Szpiech ZA, Pemberton TJ. 2017. Weighted likelihood inference of genomic autozygosity patterns in dense genotype data. *BMC Genomics* 18:928.
- Bostoen K, Clist B, Doumenge C, Grollemund R, Hombert J-M, Muluwa JK, Maley J. 2015. Middle to Late Holocene Paleoclimatic Change and the Early Bantu Expansion in the Rain Forests of Western Central Africa. *Current Anthropology* 56:354–384.
- Bouchard M-E. 2023. Scaling proximity to whiteness: Racial boundary-making on São Tomé Island. *Ethnography* 24:197–126.
- Bowcock AM, Ruiz-Linares A, Tomfohrde J, Minch E, Kidd JR, Cavalli-Sforza LL. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368:455–457.
- Brooks GE. 2003. Eurafricans in Western Africa: Commerce, Social Status, Gender, and Religious Observance from the Sixteenth to the Eighteenth Century. James Currey Publishers
- Browning SR, Browning BL, Daviglus ML, Durazo-Arvizu RA, Schneiderman N, Kaplan RC, Laurie CC. 2018. Ancestry-specific recent effective population size in the Americas. *PLOS Genetics* 14:e1007385.
- Browning SR, Waples RK, Browning BL. 2023. Fast, accurate local ancestry inference with FLARE. *The American Journal of Human Genetics* 110:326–335.
- Bryc K, Auton A, Nelson MR, Oksenberg JR, Hauser SL, Williams S, Froment A, Bodo J-M, Wambebe C, Tishkoff SA, et al. 2010. Genome-wide patterns of population structure and admixture in West Africans and African Americans. *Proc Natl Acad Sci U S A* 107:786–791.
- Bryc K, Durand EY, Macpherson JM, Reich D, Mountain JL. 2015. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *The American Journal of Human Genetics* 96:37–53.
- Busby GB, Band G, Si Le Q, Jallow M, Bougama E, Mangano VD, Amenga-Etego LN, Enimil A, Apinjoh T, Ndila CM, et al. 2016. Admixture into and within sub-Saharan Africa. Pickrell JK, editor. *eLife* 5:e15266.

- Bustamante CD, Ramachandran S. 2009. Evaluating signatures of sex-specific processes in the human genome. *Nat Genet* 41:8–10.
- Cai R, Browning BL, Browning SR. 2023. Identity-by-descent-based estimation of the X chromosome effective population size with application to sex-specific demographic history. *G3 Genes|Genomes|Genetics* 13:jkad165.
- Caldeira AM. 2007. Mestiçagem, estratégias de casamento e propriedade feminina no arquipélago de São Tomé e Príncipe nos séculos XVI, XVII e XVIII. *ARQUIPÉLAGO. História* 11–12:49–71.
- Caldeira AM. 2011. Learning the Ropes in the Tropics: Slavery and the Plantation System on the Island of São Tomé. *African Economic History* 39:35–71.
- Caldeira AM. 2013. Escravos e traficantes no império português: o comércio negreiro português no Atlântico durante os séculos XV a XIX. Lisboa: A Esfera dos Livros
- Caldeira AM. 2018. Do refúgio nos picos da ilha de São Tomé à absorção colonial: a questão dos angolares. *Biblos. Revista da Faculdade de Letras da Universidade de Coimbra* 4:123–147.
- Caldeira AM. 2021. The Island Trade Route of São Tomé in the 16th Century: Ships, Products, Capitals. *Rivista Istituto Storia Europa Mediterranea RiMe*:55–76.
- Cardoso HC, Hagemeijer T, Alexandre N. 2015. Crioulos de base lexical portuguesa. In: Iliescu M, Roegiest E, editors. *Manuel des anthologies, corpus et textes romans*. De Gruyter. p. 670–692.
- Carreira A. 1972. Cabo Verde: formação e extinção de uma sociedade escravocrata (1460-1878). Bissau: Centro de Estudos da Guiné Portuguesa
- Carreira A. 1976. Migrações nas ilhas de Cabo Verde. Lisboa: Universidade Nova
- Casto AM, Li JZ, Absher D, Myers R, Ramachandran S, Feldman MW. 2010. Characterization of X-Linked SNP genotypic variation in globally distributed human populations. *Genome Biology* 11:R10.
- Ceballos FC, Joshi PK, Clark DW, Ramsay M, Wilson JF. 2018. Runs of homozygosity: windows into population history and trait architecture. *Nature Reviews Genetics* 19:220–234.
- Chacón-Duque J-C, Adhikari K, Fuentes-Guajardo M, Mendoza-Revilla J, Acuña-Alonso V, Barquera R, Quinto-Sánchez M, Gómez-Valdés J, Everardo Martínez P, Villamil-Ramírez H, et al. 2018. Latin Americans show wide-spread Converso ancestry and imprint of local Native ancestry on physical appearance. *Nature Communications* 9:5388.
- Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. 2015. Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* 4:s13742-015-0047–0048.

- Ciccarella M, Laurent R, Szpiech ZA, Patin E, Dessarps-Freichay F, Utgé J, Lémée L, Semo A, Rocha J, Verdu P. 2024. Nested admixture during and after the Trans-Atlantic Slave Trade on the island of São Tomé. :2024.10.21.619344. Available from: <https://www.biorxiv.org/content/10.1101/2024.10.21.619344v1>
- Clarence-Smith WG. 1976. Slavery in Coastal Southern Angola, 1875-1913. *Journal of Southern African Studies* 2:214–223.
- Clarence-Smith WG. 2000. Cocoa and Chocolate, 1765-1914. 1st Edition. London: Taylor & Francis
- Coelho M, Coia CAV, 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.
- Coop G. 2022. Genetic similarity and genetic ancestry groups. Available from: <http://arxiv.org/abs/2207.11595>
- Cotter DJ, Severson AL, Kang JTL, Godrej HN, Carmi S, Rosenberg NA. 2024. Modeling the effects of consanguinity on autosomal and X-chromosomal runs of homozygosity and identity-by-descent sharing. *G3 Genes|Genomes|Genetics* 14:jkad264.
- Curtin PD. 1969. The Atlantic Slave Trade: A Census. London: University of Wisconsin Press
- Curtin PD. 1990. The Rise and Fall of the Plantation Complex: Essays in Atlantic History. Cambridge: Cambridge University Press
- Delaneau O, Zagury J-F, Robinson MR, Marchini JL, Dermitzakis ET. 2019. Accurate, scalable and integrative haplotype estimation. *Nature Communications* 10:5436.
- Destro-Bisol G, Donati F, Coia V, Boschi I, Verginelli F, Caglià A, Tofanelli S, Spedini G, Capelli C. 2004. Variation of female and male lineages in sub-Saharan populations: the importance of sociocultural factors. *Mol Biol Evol* 21:1673–1682.
- Eltis D, Richardson D. 2008. A New Assessment of the Transatlantic Slave Trade. In: *Extending the Frontiers: Essays on the New Transatlantic Slave Trade Database*. New Haven: Yale University Press.
- Eltis D, Richardson D. 2010. Atlas of the Transatlantic Slave Trade. New Haven: Yale University Press
- Falush D, Stephens M, Pritchard JK. 2003. Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* 164:1567–1587.
- Fehn A-M, Amorim B, Rocha J. 2022. The linguistic and genetic landscape of southern Africa. *J Anthropol Sci* 100:243–265.
- Finley MI. 1980. Ancient Slavery and Modern Ideology. New York: Viking Press

- Fortes-Lima C, Burgarella C, Hammarén R, Eriksson A, Vicente M, Jolly C, Semo A, Gunnink H, Pacchiarotti S, Mundeke L, et al. 2024. The genetic legacy of the expansion of Bantu-speaking peoples in Africa. *Nature* 625:540–547.
- Fortes-Lima C, Bybjerg-Grauholm J, Marin-Padrón LC, Gomez-Cabezas EJ, Bækvad-Hansen M, Hansen CS, Le P, Hougaard DM, Verdu P, Mors O, et al. 2018. Exploring Cuba's population structure and demographic history using genome-wide data. *Scientific Reports* 8:11422.
- Fortes-Lima C, Gessain A, Ruiz-Linares A, Bortolini M-C, Migot-Nabias F, Bellis G, Moreno-Mayar JV, Restrepo BN, Rojas W, Avendaño-Tamayo E, et al. 2017. Genome-wide Ancestry and Demographic History of African-Descendant Maroon Communities from French Guiana and Suriname. *The American Journal of Human Genetics* 101:725–736.
- Fortes-Lima C, Laurent R, Thouzeau V, Toupance B, Verdu P. 2021. Complex genetic admixture histories reconstructed with Approximate Bayesian Computation. *Molecular Ecology Resources* 21:1098.
- Fortes-Lima C, Verdu P. 2020. Anthropological genetics perspectives on the transatlantic slave trade. *Human Molecular Genetics* 30.
- Galenson DW. 1984. The Rise and Fall of Indentured Servitude in the Americas: An Economic Analysis. *The Journal of Economic History* 44:1–26.
- Geggus D. 1989a. Sex Ratio, Age and Ethnicity in the Atlantic Slave Trade: data from French shipping and plantation records. *J. Afr. Hist.* 30:23–44.
- Geggus D. 1989b. Sex Ratio and Ethnicity: A Reply to Paul E. Lovejoy. *The Journal of African History* 30:395–397.
- Goldberg A, Rastogi A, Rosenberg NA. 2020. Assortative mating by population of origin in a mechanistic model of admixture. *Theoretical Population Biology* 134:129–146.
- Goldberg A, Rosenberg NA. 2015. Beyond 2/3 and 1/3: The Complex Signatures of Sex-Biased Admixture on the X Chromosome. *Genetics* 201:263–279.
- Goldberg A, Verdu P, Rosenberg NA. 2014. Autosomal Admixture Levels Are Informative About Sex Bias in Admixed Populations. *Genetics* 198:1209–1229.
- Gonçalves R, Hagemeijer T. 2015. O português num contexto multilingue: o caso de São Tomé e Príncipe. *Revista Científica da UEM: Série Letras e Ciências Sociais* 1.
- Gonçalves R, Spínola H, Brehm A. 2007. Y-chromosome lineages in São Tomé e Príncipe islands: Evidence of European influence. *American Journal of Human Biology* 19:422–428.
- Gouveia MH, Borda V, Leal TP, Moreira RG, Bergen AW, Kehdy FSG, Alvim I, Aquino MM, Araujo GS, Araujo NM, et al. 2020. Origins, Admixture Dynamics, and

- Homogenization of the African Gene Pool in the Americas. *Molecular Biology and Evolution* 37:1647–1656.
- Gravel S. 2012. Population genetics models of local ancestry. *Genetics* 191:607–619.
- Guez J, Achaz G, Bienvenu F, Cury J, Toupance B, Heyer E, Jay F, Austerlitz F. 2023. Cultural transmission of reproductive success impacts genomic diversity, coalescent tree topologies, and demographic inferences. *Genetics* 223.
- Gurdasani D, Carstensen T, Tekola-Ayele F, Pagani L, Tachmazidou I, Hatzikotoulas K, Karthikeyan S, Iles L, Pollard MO, Choudhury A, et al. 2015. The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517:327–332.
- Guyon L. 2024. Inferring the history of human kinship systems from genetic data. Available from: <https://hal.science/tel-04889960>
- Guyon L, Guez J, Toupance B, Heyer E, Chaix R. 2024. Patrilineal segmentary systems provide a peaceful explanation for the post-Neolithic Y-chromosome bottleneck. *Nat Commun* 15:3243.
- Hagberg AA, Schult DA, Swart PJ. 2008. Exploring Network Structure, Dynamics, and Function using NetworkX. *Proceedings SciPy* [Internet]. Available from: <https://proceedings.scipy.org/articles/TCWV9851>
- Hagemeijer T. 2011. The Gulf of Guinea Creoles: Genetic and typological relations. *Journal of Pidgin and Creole Languages* 26:111–154.
- Hagemeijer T, Rocha J. 2019. Creole languages and genes: the case of São Tomé and Príncipe. *Faits de Langues* 49:167–182.
- Hamid I, Korunes KL, Beleza S, Goldberg A. 2021. Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. *Elife* 10:e63177.
- Hamid I, Korunes KL, Schrider DR, Goldberg A. 2023. Localizing Post-Admixture Adaptive Variants with Object Detection on Ancestry-Painted Chromosomes. *Molecular Biology and Evolution* 40:msad074.
- Hammer MF, Mendez FL, Cox MP, Woerner AE, Wall JD. 2008. Sex-biased evolutionary forces shape genomic patterns of human diversity. *PLoS Genet* 4:e1000202.
- Hartl DL, Clark AG. 2007. Principles of Population Genetics. 4th ed. Sunderland, Massachusetts: Sinauer Associates
- Hellenthal G, Busby GBJ, Band G, Wilson JF, Capelli C, Falush D, Myers S. 2014. A genetic atlas of human admixture history. *Science* 343:747–751.
- Henriques I de C. 2000. São Tomé e Príncipe: a invenção de uma sociedade. Lisboa: Vega
- Heyer E, Chaix R, Pavard S, Austerlitz F. 2012. Sex-specific demographic behaviours that shape human genomic variation. *Molecular Ecology* 21:597–612.

- Higgs C. 2013. *Chocolate islands: Cocoa, slavery, colonial Africa*. Ohio: Ohio University Press
- Hilmarsson H, Kumar AS, Rastogi R, Bustamante CD, Montserrat DM, Ioannidis AG. 2021. High Resolution Ancestry Deconvolution for Next Generation Genomic Data. :2021.09.19.460980.
- Hofmeister RJ, Ribeiro DM, Rubinacci S, Delaneau O. 2023. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat Genet* 55:1243–1249.
- Homburger JR, Moreno-Estrada A, Gignoux CR, Nelson D, Sanchez E, Ortiz-Tello P, Pons-Estel BA, Acevedo-Vasquez E, Miranda P, Langefeld CD, et al. 2015. Genomic Insights into the Ancestry and Demographic History of South America. *PLOS Genetics* 11:e1005602.
- Illumina Inc. 2010. Updated Cluster Generation Protocol for X Chromosome SNPs.
- Instituto Nacional de Estatística. 2012. Dados Distritais e Nacional Recenseamento 2012. *Instituto Nacional de Estatística* [Internet]. Available from: <https://www.ine.st/index.php/publicacao/documentos/category/71-dados-distritais-e-nacional-recenseamento-2012>
- Jackson FLC. 2021. So many Nigerians: why is Nigeria overrepresented as the ancestral genetic homeland of Legacy African North Americans? *The American Journal of Human Genetics* 108:202–208.
- Kehdy FSG, Gouveia MH, Machado M, Magalhães WCS, Horimoto AR, Horta BL, Moreira RG, Leal TP, Scliar MO, Soares-Souza GB, et al. 2015. Origin and dynamics of admixture in Brazilians and its effect on the pattern of deleterious mutations. *Proc Natl Acad Sci U S A* 112:8696–8701.
- Keinan A, Mullikin JC, Patterson N, Reich D. 2009. Accelerated genetic drift on chromosome X during the human dispersal out of Africa. *Nat Genet* 41:66–70.
- Kiesow S. 2017. Cocoa Culture on São Tomé and Príncipe: The Rise and Fall of Cocoa on the Islands in the Nineteenth and Twentieth Centuries. *Agricultural History* 91:55–77.
- Kim J, Edge MD, Goldberg A, Rosenberg NA. 2021. Skin deep: The decoupling of genetic admixture levels from phenotypes that differed between source populations. *American Journal of Physical Anthropology* 175:406–421.
- Korunes KL, Soares-Souza GB, Bobrek K, Tang H, Araújo II, Goldberg A, Beleza S. 2022. Sex-biased admixture and assortative mating shape genetic variation and influence demographic inference in admixed Cabo Verdeans. *G3 Genes[Genomes]Genetics* 12:jkac183.
- Laurent R, Szpiech ZA, da Costa SS, Thouzeau V, Fortes-Lima CA, Dessarps-Freichay F, Lémée L, Utgé J, Rosenberg NA, Baptista M, et al. 2023. A genetic and

- linguistic analysis of the admixture histories of the islands of Cabo Verde. *eLife* 12:e79827.
- Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics* 8:e1002453.
- Lewis ACF, Molina SJ, Appelbaum PS, Dauda B, Di Rienzo A, Fuentes A, Fullerton SM, Garrison NA, Ghosh N, Hammonds EM, et al. 2022. Getting genetic ancestry right for science and society. *Science* 376:250–252.
- Li S, Schlebusch C, Jakobsson M. 2014. Genetic variation reveals large-scale population expansion and migration during the expansion of Bantu-speaking peoples. *Proc Biol Sci* 281:20141448.
- Lima-Costa MF, Rodrigues LC, Barreto ML, Gouveia M, Horta BL, Mambrini J, Kehdy FSG, Pereira A, Rodrigues-Soares F, Victora CG, et al. 2015. Genomic ancestry and ethnoracial self-classification based on 5,871 community-dwelling Brazilians (The Epigen Initiative). *Sci Rep* 5:9812.
- Lind JM, Hutcheson-Dilks HB, Williams SM, Moore JH, Essex M, Ruiz-Pesini E, Wallace DC, Tishkoff SA, O'Brien SJ, Smith MW. 2007. Elevated male European and female African contributions to the genomes of African American individuals. *Hum Genet* 120:713–722.
- Lorenzino G. 1998. The Angolar Creole Portuguese of São Tomé: its grammar and sociolinguistic history. Munchen: LINCOM Europa
- Lovejoy PE. 2010. The Upper Guinea Coast and the Trans-Atlantic Slave Trade Database. *African Economic History* 38:1–27.
- Lovejoy PE. 2012. Transformations in Slavery: A History of Slavery in Africa. 3rd edition. Cambridge: Cambridge University Press
- Lucas PG. 2015. The demography of São Tomé and Príncipe (1758–1822): preliminary approaches to an insular slave society. *Anais De Historia De Alem-Mar* XVI:51–78.
- Lucotte EA, Laurent R, Heyer E, Ségurel L, Toupance B. 2016. Detection of Allelic Frequency Differences between the Sexes in Humans: A Signature of Sexually Antagonistic Selection. *Genome Biology and Evolution* 8:1489–1500.
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen W-M. 2010. Robust relationship inference in genome-wide association studies. *Bioinformatics* 26:2867–2873.
- Maples BK, Gravel S, Kenny EE, Bustamante CD. 2013. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet* 93:278–288.
- Martin AR, Gignoux CR, Walters RK, Wojcik GL, Neale BM, Gravel S, Daly MJ, Bustamante CD, Kenny EE. 2017. Human Demographic History Impacts Genetic

- Risk Prediction across Diverse Populations. *American Journal of Human Genetics* 100:635–649.
- Mas-Sandoval A, Mathieson S, Fumagalli M. 2023b. The genomic footprint of social stratification in admixing American populations. *eLife* 12:e84429.
- Mathias RA, Taub MA, Gignoux CR, Fu W, Musharoff S, O'Connor TD, Vergara C, Torgerson DG, Pino-Yanes M, Shringarpure SS, et al. 2016. A continuum of admixture in the Western Hemisphere revealed by the African Diaspora genome. *Nat Commun* 7:12522.
- Mathieson I, Scally A. 2020. What is ancestry? *PLOS Genetics* 16:e1008624.
- Micheletti S, Bryc K, Esselmann S, Freyman W, Moreno M, Poznik G, Shastri A, Beleza S, Mountain J, Agee M, et al. 2020. Genetic Consequences of the Transatlantic Slave Trade in the Americas. *The American Journal of Human Genetics* 107:265–277.
- Micheletti SJ, Esselmann SGA, Mountain JL. 2021. Response to Jackson. *The American Journal of Human Genetics* 108:209–210.
- Mooney JA, Agranat-Tamir L, Pritchard JK, Rosenberg NA. 2023. On the number of genealogical ancestors tracing to the source groups of an admixed population. *Genetics* 224:iyad079.
- Mooney JA, Huber CD, Service S, Sul JH, Marsden CD, Zhang Z, Sabatti C, Ruiz-Linares A, Bedoya G, Fears SC, et al. 2018. Understanding the Hidden Complexity of Latin American Population Isolates. *The American Journal of Human Genetics* 103:707–726.
- Moreno-Estrada A, Gravel S, Zakharia F, McCauley JL, Byrnes JK, Gignoux CR, Ortiz-Tello PA, Martínez RJ, Hedges DJ, Morris RW, et al. 2013. Reconstructing the Population Genetic History of the Caribbean. *PLOS Genetics* 9:e1003925.
- Morgan JL. 2018. Partus sequitur ventrem: Law, race, and reproduction in colonial slavery. *Small Axe* 22:1–17.
- Morgan JL. 2021. *Reckoning with Slavery: Gender, Kinship, and Capitalism in the Early Black Atlantic*. Durham, North Carolina: Duke University Press
- Mufwene SS. 2015. The Emergence of Creoles and Language Change. In: *The Routledge Handbook of Linguistic Anthropology*. Routledge.
- Murdock GP. 1949. *Social structure*. Oxford: Macmillan
- Nait Saada J, Kalantzis G, Shyr D, Cooper F, Robinson M, Gusev A, Palamara PF. 2020. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nat Commun* 11:6130.
- Nascimento A. 2002. *Poderes e Quotidiano nas Roças de S. Tomé e Príncipe. De finais de oitocentos a meados de novecentos*. Lousa: Tipografia Lousanense

- Nascimento A. 2003. O sul da diáspora : Cabo-Verdianos em plantações de S. Tomé e Príncipe e Moçambique. Praia: Edição da Presidência da República de Cabo Verde
- Nascimento A. 2008. Cape Verdeans in São Tomé and Príncipe. In: Batalha L, Carling J, editors. Transnational Archipelago. Perspectives on Cape Verdean Migration and Diaspora. Amsterdam University Press. p. 55–60. Available from: <https://www.jstor.org/stable/j.ctt46msd4.8>
- Nascimento A. 2023. Exile and contract: journeys of the mozambicans to S. Tomé and Príncipe (1940 to 1960). Centro de História da Universidade de Lisboa Available from: <http://hdl.handle.net/10451/58340>
- National Academies of Sciences, Engineering, and Medicine. 2023. Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field. Washington, D.C.: National Academies Press
- Newman BN. 2020. Blood Fictions, Maternal Inheritance, and the Legacies of Colonial Slavery. *Women's Studies Quarterly* 48:27–44.
- Niort J-F. 2023. Le code noir : idées reçues sur un texte symbolique. 2e éd. Paris: le Cavalier bleu éditions
- Norris ET, Rishishwar L, Wang L, Conley AB, Chande AT, Dabrowski AM, Valderrama-Aguirre A, Jordan IK. 2019. Assortative Mating on Ancestry-Variant Traits in Admixed Latin American Populations. *Front. Genet.* 10.
- Oliveira S, Fehn A, Aço T, Lages F, Gayà-Vidal M, Pakendorf B, Stoneking M, Rocha J. 2018. Matriclans shape populations: Insights from the Angolan Namib Desert into the maternal genetic history of southern Africa. *American J Phys Anthropol* 165:518–535.
- Oliveira S, Hübner A, Fehn A-M, Aço T, Lages F, Pakendorf B, Stoneking M, Rocha J. 2019. The role of matrilineality in shaping patterns of Y chromosome and mtDNA sequence variation in southwestern Angola. *Eur J Hum Genet* 27:475–483.
- Ongaro L, Molinaro L, Flores R, Marnetto D, Capodiferro MR, Alarcón-Riquelme ME, Moreno-Estrada A, Mabunda N, Ventura M, Tambets K, et al. 2021. Evaluating the Impact of Sex-Biased Genetic Admixture in the Americas through the Analysis of Haplotype Data. *Genes* 12:1580.
- Ongaro L, Scliar MO, Flores R, Raveane A, Marnetto D, Sarno S, Gneccchi-Ruscione GA, Alarcón-Riquelme ME, Patin E, Wangkumhang P, et al. 2019. The Genomic Impact of European Colonization of the Americas. *Current Biology* 29:3974–3986.e4.
- Oota H, Settheetham-Ishida W, Tiwawech D, Ishida T, Stoneking M. 2001. Human mtDNA and Y-chromosome variation is correlated with matrilineal versus patrilineal residence. *Nat Genet* 29:20–21.

- Parra EJ, Kittles RA, Argyropoulos G, Pfaff CL, Hiester K, Bonilla C, Sylvester N, Parrish-Gause D, Garvey WT, Jin L, et al. 2001. Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina. *Am J Phys Anthropol* 114:18–29.
- Parra EJ, Marcini A, Akey J, Martinson J, Batzer MA, Cooper R, Forrester T, Allison DB, Deka R, Ferrell RE, et al. 1998. Estimating African American Admixture Proportions by Use of Population-Specific Alleles. *The American Journal of Human Genetics* 63:1839–1851.
- Parra FC, Amado RC, Lambertucci JR, Rocha J, Antunes CM, Pena SDJ. 2003. Color and genomic ancestry in Brazilians. *Proceedings of the National Academy of Sciences* 100:177–182.
- Patin E, Lopez M, Grollemund R, Verdu P, Harmant C, Quach H, Laval G, Perry GH, Barreiro LB, Froment A, et al. 2017. Dispersals and genetic adaptation of Bantu-speaking populations in Africa and North America. *Science* 356:543–546.
- Patin E, Siddle KJ, Laval G, Quach H, Harmant C, Becker N, Froment A, Régnauld B, Lemée L, Gravel S, et al. 2014. The impact of agricultural emergence on the genetic history of African rainforest hunter-gatherers and agriculturalists. *Nat Commun* 5:3163.
- Pemberton TJ, Absher D, Feldman MW, Myers RM, Rosenberg NA, Li JZ. 2012. Genomic patterns of homozygosity in worldwide human populations. *American Journal of Human Genetics* 91:275–292.
- Perego UA, Lancioni H, Tribaldos M, Angerhofer N, Ekins JE, Olivieri A, Woodward SR, Pascale JM, Cooke R, Motta J, et al. 2012. Decrypting the Mitochondrial Gene Pool of Modern Panamanians. *PLOS ONE* 7:e38337.
- Peter BM, Petkova D, Novembre J. 2020. Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography. *Molecular Biology and Evolution* 37:943–951.
- Pool JE, Nielsen R. 2007. Population Size Changes Reshape Genomic Patterns of Diversity. *Evolution* 61:3001–3006.
- Pool JE, Nielsen R. 2009. Inference of Historical Changes in Migration Rate From the Lengths of Migrant Tracts. *Genetics* 181:711–719.
- Pritchard JK, Stephens M, Donnelly P. 2000. Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155:945–959.
- R Core Team. 2021. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing Available from: <https://www.R-project.org/>
- Ramachandran S, Rosenberg NA, Zhivotovsky LA, Feldman MW. 2004. Robustness of the inference of human population structure: A comparison of X-chromosomal and autosomal microsatellites. *Hum Genomics* 1:87–97.

- Ramsay M, Schlebusch C, Davies K. 2021. Evolutionary genomics in Africa. *Human Molecular Genetics* 30:R1.
- Risch N, Choudhry S, Via M, Basu A, Ronnie Sebro, Ronnie Sebro, Sebro R, Eng C, Beckman KB, Thyne S, et al. 2009. Ancestry-related assortative mating in Latino populations. *Genome Biology* 10:1–16.
- Rosenberg NA, Mahajan S, Ramachandran S, Zhao C, Pritchard JK, Feldman MW. 2005. Clines, Clusters, and the Effect of Study Design on the Inference of Human Population Structure. *PLOS Genetics* 1:e70.
- Ruef M. 2022. Racial Segregation under Slavery. *Soc Forces* 100:935–960.
- Ruiz-Linares A, Adhikari K, Acuña-Alonzo V, Quinto-Sanchez M, Jaramillo C, Arias W, Fuentes M, Pizarro M, Everardo P, Avila F de, et al. 2014. Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLOS Genetics* 10:e1004572.
- Salter-Townshend M, Myers S. 2019. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics* 212:869–889.
- Santos MS. 2016. “Slave Mothers”, Partus Sequitur Ventrem, and the Naturalization of Slave Reproduction in Nineteenth-Century Brazil. *Tempo* 22:467–487.
- Sarno S, Boattini A, Pagani L, Sazzini M, De Fanti S, Quagliariello A, Gneccchi Ruscone GA, Guichard E, Ciani G, Bortolini E, et al. 2017. Ancient and recent admixture layers in Sicily and Southern Italy trace multiple migration routes along the Mediterranean. *Sci Rep* 7:1984.
- Sebro R, Peloso GM, Dupuis J, Risch NJ. 2017. Structured mating: Patterns and implications. *PLOS Genetics* 13:e1006655.
- Segurel L, Martínez-Cruz B, Quintana-Murci L, Balaesque P, Georges M, Hegay T, Aldashev A, Nasyrova F, Jobling M, Heyer E, et al. 2008. Sex-Specific Genetic Structure and Social Organization in Central Asia: Insights from a Multi-Locus Study. *PLoS genetics* 4:e1000200.
- Seibert G. 1998. A questão da origem dos Angolares de São Tomé. *Instituto Superior de Economia e Gestão - CEsa Brief papers*; 5/1998.
- Seibert G. 2004. Os angolares da ilha de São Tome: Náufragos, Autóctones ou Quilombolas? *Revista do Programa de Pós-graduação em História da UnB*. 12:43–64.
- Seibert G. 2013. Tenreiro, Amador e os Angolares ou a reinvenção da história de São Tomé. In: Silva MC, Saraiva C, editors. *Antropologia, História, África e Academia*. Lisboa: Etnográfica Press. p. 171–185.
- Seibert G. 2014. Crioulização em Cabo Verde e São Tomé e Príncipe: divergências históricas e identitárias. *Afro-Ásia* 49:41–70.

- Seibert G. 2015. Colonialismo em São Tomé e Príncipe: hierarquização, classificação e segregação da vida social. *Anuário Antropológico* 40:99–120.
- Semo A, Gayà-Vidal M, Fortes-Lima C, Alard B, Oliveira S, Almeida J, Prista A, Damasceno A, Fehn A-M, Schlebusch C, et al. 2020. Along the Indian Ocean Coast: Genomic Variation in Mozambique Provides New Insights into the Bantu Expansion. *Molecular Biology and Evolution* 37:406–416.
- Shringarpure SS, Bustamante CD, Lange K, Alexander DH. 2016. Efficient analysis of large datasets and sex bias with ADMIXTURE. *BMC Bioinformatics* 17:218.
- Stefflova K, Dulik MC, Pai AA, Walker AH, Zeigler-Johnson CM, Gueye SM, Schurr TG, Rebbeck TR. 2009. Evaluation of Group Genetic Ancestry of Populations from Philadelphia and Dakar in the Context of Sex-Biased Admixture in the Americas. *PLOS ONE* 4:e7842.
- Szpiech ZA, Blant A, Pemberton TJ. 2017. GARLIC: Genomic Autozygosity Regions Likelihood-based Inference and Classification. *Bioinformatics* 33:2059–2062.
- Szpiech ZA, Mak ACY, White MJ, Hu D, Eng C, Burchard EG, Hernandez RD. 2019. Ancestry-Dependent Enrichment of Deleterious Homozygotes in Runs of Homozygosity. *American Journal of Human Genetics* 105:747–762.
- Tallman S, Sungo M das D, Saranga S, Beleza S. 2023. Whole genomes from Angola and Mozambique inform about the origins and dispersals of major African migrations. *Nat Commun* 14:7967.
- The Slave Voyages Consortium. 2021. Slave Voyages v2.2.13. Available from: <https://www.slavevoyages.org/>
- Tishkoff SA, Reed FA, Friedlaender FR, Ehret C, Ranciaro A, Froment A, Hirbo JB, Awomoyi AA, Bodo J-M, Doumbo O, et al. 2009. The Genetic Structure and History of Africans and African Americans. *Science* 324:1035–1044.
- Tomas 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 MJ, Alves C, Gusmão L, Abade A, Amorim A, Prata MJ. 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 MJ, Pereira L, Gusmão L, Abade A, Amorim A, Prata MJ. 2004. Pattern of mtDNA Variation in Three Populations from São Tomé e Príncipe. *Annals of Human Genetics* 68:40–54.
- Trovoada MJ, Tavares L, Gusmão L, Alves C, Abade A, Amorim A, Prata MJ. 2007. Dissecting the Genetic History of São Tomé e Príncipe: A New Window from Y-Chromosome Biallelic Markers. *Annals of Human Genetics* 71:77–85.

- Valente R da R, Berry BJL. 2020. Residential Segregation by Skin Color: Brazil Revisited. *Latin American Research Review* 55:207–226.
- Verdu P, Austerlitz F, Estoup A, Vitalis R, Georges M, Théry S, Froment A, Le Bomin S, Gessain A, Hombert J-M, et al. 2009. Origins and genetic diversity of pygmy hunter-gatherers from Western Central Africa. *Curr Biol* 19:312–318.
- Verdu P, Becker NSA, Froment A, Georges M, Grugni V, Quintana-Murci L, Hombert J-M, Van der Veen L, Le Bomin S, Bahuchet S, et al. 2013. Sociocultural behavior, sex-biased admixture, and effective population sizes in Central African Pygmies and non-Pygmies. *Mol Biol Evol* 30:918–937.
- Verdu P, Jewett EM, Pemberton TJ, Rosenberg NA, Baptista M. 2017. Parallel Trajectories of Genetic and Linguistic Admixture in a Genetically Admixed Creole Population. *Current Biology* 27:2529.
- Verdu P, Rosenberg NA. 2011. A General Mechanistic Model for Admixture Histories of Hybrid Populations. *Genetics* 189:1413.
- Vogt JL. 1973. The Early Sao Tome-Principe Slave Trade with Mina, 1500-1540. *The International Journal of African Historical Studies* 6:453–467.
- Wade P. 2009. Race and Sex in Latin America. London: Pluto Press
- Wangkumhang P, Greenfield M, Hellenthal G. 2022. An efficient method to identify, date, and describe admixture events using haplotype information. *Genome Research* 32:1553–1564.
- Wangkumhang P, Hellenthal G. 2018. Statistical methods for detecting admixture. *Current Opinion in Genetics & Development* 53:121–127.
- Wright S. 1931. Evolution in Mendelian Populations. *Genetics* 16:97–159.
- Zaitlen N, Huntsman S, Hu D, Spear M, Eng C, Oh SS, White MJ, Mak A, Davis A, Meade K, et al. 2017. The Effects of Migration and Assortative Mating on Admixture Linkage Disequilibrium. *Genetics* 205:375–383.
- Zhou Y, Browning SR, Browning BL. 2020. A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *The American Journal of Human Genetics* 106:426–437.
- Zou JY, Halperin E, Burchard E, Sankararaman S. 2015. Inferring parental genomic ancestries using pooled semi-Markov processes. *Bioinformatics* 31:i190-196.

ATTACHMENTS

SUPPLEMENTARY INFORMATION ON GENETIC DATASETS

Genome-wide genotyping data available for further studies

Generating data is a fundamental step in data analysis, and curating genetic markers from scratch allows for a thorough understanding of the datasets used for analysis. For this study, we genotyped new genome-wide genetic data from 100 individual samples collected in São Tomé and, for the first time, curated and analysed X chromosome markers from three other datasets for which only autosomal data had previously been analysed (see Supplementary Table S 1). I developed *ad hoc* Genotyping Quality Control pipelines to curate markers on the autosomes and X chromosome separately, using custom Illumina software and Python scripts.

In Chapter 2, I examined the 2.5 million newly generated autosomal markers from São Tomé, providing insights into the genetic structure and demographic history of the admixed population. In Chapter 3, I further examined 44,162 X chromosome markers newly genotyped in São Tomé. For the first time, X chromosome data were curated and analysed in São Tomé, Cabo Verde and other African populations, investigating the influence of gender dynamics on patterns of genetic diversity in admixed populations and their sources.

The newly generated autosomal datasets has already been deposited in the European Genome-phenome Archive (EGAD50000001348) and will be made available to the scientific community for further studies upon request, in line with ethical guidelines and informed consent from the study participants.

Study	Sampling Site	Autosomes			X chromosome		
Ciccarella et al. 2024	São Tomé	G	C	A	G	C	A
Laurent et al. 2023	Cabo Verde					C	A
Semo et al. 2020	Mozambique, Angola					C	A
Patin et al. 2017	Ivory Coast, Benin, Cameroon, Gabon, Angola, CAR, DCR					C	A

Supplementary Table S 1. Newly Generated and Curated Datasets of African Population Samples. This study includes 96 individual samples from São Tomé that were newly genotyped. Additionally, X chromosome markers from three previously published datasets were curated and analysed. Legend: G = Genotyped; C = Curated; A = Analysed for the first time in this study.

Sampling strategies

DNA sampling in São Tomé e Príncipe and Cabo Verde has been conducted by two different research groups on the two archipelagos, respectively. Sampling strategies are detailed in (Coelho et al. 2008) for São Tomé e Príncipe and in (Laurent et al. 2023) for Cabo Verde.

In São Tomé, the sampling reflected the island's uneven population distribution. Due to the island's mountainous geography and peopling history, settlements are mainly in the northeast and along the coast, with few in the centre and southeast, and almost none in the forested southwest. Samples were collected across 14 locations, including 11 villages and 3 plantations, representing over 80% of São Tomé's population.

Research sampling protocols followed the Declaration of Helsinki guidelines and written informed consent was obtained from all subjects involved in the sampling in São Tomé and Cabo Verde. Anthropological questionnaires, completed alongside DNA sampling, provide birthplace locations of each individual and their parents. DNA was collected with buccal swabs and extracted from saliva samples following standard protocols.

SUPPLEMENTARY INFORMATION FOR CHAPTER 2

Supplementary Materials and Methods

Genetic dataset

Genotyping and Quality Control

The 100 samples collected in São Tomé e Príncipe were newly genotyped using the HumanOmni2.5Million-BeadChip genotyping array read with iScan at the OMICS platform of the Institut Pasteur. 261 samples from Cabo Verde were previously genotyped on the same chip (albeit different versions) using the same technology at the same platform (Laurent et al. 2023).

I curated the raw genotypes of 361 samples from São Tomé e Príncipe and Cabo Verde from 5 separate batches produced with different versions of the same genotyping array (Supplementary Table S 2.1). I conducted a comprehensive Genotyping Quality Control in four phases using the Genotyping Module v1.9.4 from Illumina GenomeStudio and custom Python scripts. I finally retained 2,104,148 autosomal SNPs from 330 family unrelated individuals, including 233 individuals sampled in Cabo Verde and 97 individuals sampled in São Tomé e Príncipe (Supplementary Table S 2.2).

I assigned each individual sample from Cabo Verde and São Tomé e Príncipe to an island of birth based on individual birthplaces recorded in the family anthropology questionnaires. Out of the 233 individuals sampled in Cabo Verde, 225 were born on seven of the archipelago's islands, while seven were born outside of Cabo Verde, in particular two in São Tomé, three in Angola, one in Brazil, and one in Portugal. I excluded one individual due to missing birthplace information. Out of 97 individuals sampled in São Tomé e Príncipe, almost all individuals were born in São Tomé, apart from one individual born in Gabon, and two individuals born in the island of Príncipe.

Finally, I retained 96 individuals born in São Tomé, including the 94 individuals sampled in São Tomé and the two sampled in Cabo Verde. I will refer to the dataset containing 323 individuals born in Cabo Verde and São Tomé e Príncipe as CVSTP henceforth.

Merging with worldwide populations

The CVSTP dataset was merged with 2504 samples from 26 populations worldwide included in the 1000 Genomes project Phase 3 (Auton et al. 2015), with 1307 samples from 14 African populations included in the African Genome Variation Project

(EGAS00001000959) (Gurdasani et al. 2015), with 188 samples from 15 populations in Mozambique and Angola (E-MTAB-8450) (Semo et al. 2020), and with 1366 samples from 38 sub-Saharan African populations (EGAS00001002078) (Patin et al. 2017) (Supplementary Table S 2.3).

For the Semo et al. (2020) dataset, which was genotyped on a similar version of Illumina HumanOmni2.5 SNPs array as the São Tomé e Príncipe and Cabo Verde samples, I performed the same steps of Genotyping Quality Control from the raw genotyping data as for the CVSTP dataset. After each successive merging, I checked for genetic relatedness among samples up to the 2nd degree using KING (Manichaikul et al. 2010). I finally retained 411,121 SNPs from 5423 family unrelated individuals.

Population labels and geographical regions

I grouped the 5423 unrelated individuals from 107 populations of the final merged dataset into 16 distinct geographic regions, including ten regions within Africa (with CVSTP identified separately), three regions in the Americas, and one region in Europe, South Asia and East Asia.

When two cardinal points are used to define a region, such as ‘East Western Africa’, the second term indicates the wider region in Africa, including Western, Central or Southern Africa, while the first term specifies the relative position within that region (e.g. the eastern part of ‘Western Africa’).

Population genetics descriptions

Allele Sharing Dissimilarity

I investigated genetic diversity patterns between each pair of individuals based on successive subsets of populations in the merged dataset. I first computed a matrix of pairwise allele sharing dissimilarities (Bowcock et al. 1994), including all the individuals and all SNPs in the merged dataset using the ASD software (<https://github.com/szpiech/asd>). Then I explored three axes of variation of Multi-Dimensional Scaling projections of various, separate, subsets of this ASD pairwise-matrix using the *cmdscale* function in R (R Core Team 2021). In particular, I removed East Asian, South Asian, South American, Puerto Rican PUR, Mexican American MXL, Central African, East African, and four West-Central African hunter gatherers populations. Figure 2.3 B reports the first two dimensions of the ASD-MDS based on

411,121 autosomal SNPs of 3203 individuals from 77 populations, and the third dimension is presented in Supplementary Figure S 2.1.

Genetic clustering

I used ADMIXTURE version 1.3 (Alexander et al. 2009) to explore further genetic resemblances among individuals. ADMIXTURE analysis is sensitive to sample size heterogeneities, therefore I randomly resampled without replacement 20 individuals for each population, and I removed 7 populations with less than 5 individuals, except for the island's populations of interest in Cabo Verde and São Tomé e Príncipe, where I retained all individuals. This analysis was performed using 1347 individuals from 70 populations, referred to as the Working Dataset henceforth (Supplementary Table S 2.4).

I filtered the initial set of 411,121 autosomal SNPs for low Linkage Disequilibrium using the *--indep-pairwise* function in PLINK with a 50 SNP-window moving every 10 SNPs, and 0.1 r^2 cutoff. I thus run ADMIXTURE on 110,499 LD-pruned autosomal SNPs from the 1347 individuals in the Working Dataset. I performed 10 independent runs of ADMIXTURE for values of K ranging from 2 to 7 (Figure 2.4). I used PONG (Behr et al. 2016) to define ADMIXTURE “modes” with a greedy approach for similarity threshold 0.95. The alternative ADMIXTURE mode for $K=5$ is presented in Supplementary Figure S 2.2. I conducted an evaluation of the cross-validation error across 10 distinct runs for 15 values of K and found that $K=4$ yielded the lowest error (Supplementary Figure S 2.3). ADMIXTURE results for $K>7$ are reported in Supplementary Figure S 2.4.

Local-ancestry inferences

Phasing with Shapeit4

I phased individual genotypes using SHAPEIT4 version 4.2.2 (Delaneau et al. 2019), considering each autosomal chromosome separately with the HapMap Phase 3 Build GRCh38 genetic recombination map (Altshuler et al. 2010), and default parameters for autosomal data: minimum phasing window length of 2.5 Mb, and a total of 15 MCMC iterations, including 7 burn-in, 3 pruning, and 5 main iterations.

Chromosome painting for fineSTRUCTURE

I used the inferential algorithm implemented in ChromoPainter v2 (Lawson et al. 2012) to paint each São Toméan genome as a combination of fragments received from other

São Toméan individuals. I performed a first run of ChromoPainter to estimate nuisance parameters on four chromosomes using 10 EM iterations. I obtained $N_e = 792.022979835471$ and $\mu = 0.0020978990636338$, then used to run ChromoPainter on the whole dataset. I averaged the results for all individuals by chromosome.

Clustering São Tomé individuals with fineSTRUCTURE

I ran fineSTRUCTURE (Lawson et al. 2012) using the co-ancestry matrix based on the number of shared chunks between each pair of individuals previously computed with ChromoPainter. I performed 100,000 burn-in steps to allow the algorithm to reach a stable state. Subsequently, 100,000 further iterations were performed, retaining samples every 10,000th iteration. Following the MCMC sampling, a tree representing the inferred relationships among individuals was constructed. The best state observed during the MCMC sampling, reflecting the optimal population structure, served as the initial state for tree inference.

fineSTRUCTURE classified the 96 São Toméan individuals into 17 clusters. In order to increase the interpretability of subsequent analysis, and based on the haplotype sharing patterns of the co-ancestry matrices (Figure 2.7 C and D), I identified five genetic clusters by cutting the dendrogram at height 4. Supplementary Figure S 2.5 provide a more detailed representation of the fineSTRUCTURE dendrogram. The PCA in Figure 2.7 B has been calculated on the co-ancestry matrix with the *pcares* function in R. The third and fourth Principal Components are reported in Supplementary Figure S 2.6. The heatmaps in Figure 2.7 C and D are produced using the function provided by the authors in the *FinestructureDendrogram.R* script, and are capped at 450 and 250 cM, respectively, for visualization purposes. For a visual representation of how consistently pairs of individuals are grouped together across different iterations of the MCMC process, refer to the pairwise coincidence matrix in Supplementary Figure S 2.7.

Chromosome painting for SOURCEFIND and fastGLOBETROTTER

I used the phased haplotype information to reconstruct the chromosomes of each individual sample from Africa, Europe and the Americas in the Working Dataset as a series of genomic segments inherited from a set of Donor individuals using ChromoPainter v2 (Lawson et al. 2012). As with the previous ChromoPainter run on the São Toméan sample, I first estimated nuisance parameters using 10 E-M iterations, this time on a subset of the Working Dataset, following the author's recommendations. I

randomly sampled 3 individuals per population (1/10th of the entire dataset), and I selected 4 chromosomes: 1, 7, 14 and 21. I averaged the estimated values across chromosomes, weighted by chromosome size, over the 10 replicate analyses. Finally, I used the a posteriori estimated nuisance parameters to run the ChromoPainter algorithm on the entire Working Dataset.

I run ChromoPainter on 411,121 autosomal SNPs from 1347 individuals of the Working Dataset to prepare input files for three distinct SOURCEFIND analyses.

- For the first SOURCEFIND analysis, I run ChromoPainter setting each individual as both Donor and Recipient, except for the admixed individuals of interest Cabo Verde CV, São Tomé e Príncipe STP, African-Barbadians ACB, and African-Americans ASW, which were set as Recipient only. Importantly, each Donor population consisted of a maximum of 20 individuals in the Working Dataset. I obtained the averaged values of nuisance parameters $N_e = 228.3939$, and $\mu = 0.00076$. Finally, I combined painted chromosomes for each individual in each of CV, STP, ACB, and ASW populations, separately.
- For the second and third SOURCEFIND analysis, and for the fastGLOBETROTTER analysis, I run ChromoPainter setting all individuals as both Donor and Recipient, except for the admixed individuals of interest STP, ACB, and ASW, which were set as Recipient only. Importantly, the nine CV populations were set as both Donor and Recipient. I obtained the averaged values of nuisance parameters $N_e = 193.2902$, and $\mu = 0.00040$. Finally, I combined painted chromosomes for each individual in each of STP, ACB, and ASW populations, separately.
- Finally, I ran ChromoPainter setting all individuals as both Donor and Recipient to have a symmetric matrix for PCA computation (Supplementary Figure S 2.9), itself computed using the *eigen* function in R on the normalized matrix of chunks counts considering all individuals in the Working Dataset.

Estimating possible sources populations using SOURCEFIND

I applied the model-based method SOURCEFIND (Chacón-Duque et al. 2018) to estimate the shared haplotypic ancestry among populations based on chromosome painting. I modelled the copying vector of each “Target” admixed individual, obtained with the previous ChromoPainter analysis, as a weighted mixture of copying vectors from a set of “Surrogate” individuals. The Target and Surrogate individuals were allowed to

copy from the same set of Donors in the ChromoPainter analysis, in particular all the populations included in the Working Dataset except for the Target populations of interest. I ran three separate analyses to model different hypotheses of admixture.

- In the first analysis (Figure 2.8A), I used CV, STP, ACB, and ASW populations as Target, and all other populations in the dataset as Surrogates.
- In the second analysis (Figure 2.8B), I used STP, ACB, and ASW populations as Target, and all other populations in the dataset including CV populations as Surrogates, in order to estimate the relative contribution of CV populations to the haplotypic ancestry of each Target. To reduce the influence of differences in sample size among CV populations, CV populations were randomly subsampled to a maximum of 20 individuals per island.
- In the third analysis (Figure 2.8C), I used the five São Toméan genetic clusters as separate Targets, and the same set of Surrogate populations as the second analysis.

I conducted 20 independent SOURCEFIND runs for each analysis. Each Target individual was modelled as a mixture of Surrogate individuals only (“selfcopying” self.copy.ind), allowing for a maximum of 8 Surrogates (num.surrogates), with 4 expected number of Surrogates (exp.num.surrogates), and dividing each Target individual genome in 100 slots (num.slots) with possibly different ancestry. I considered 400,000 MCMC iterations (num.iterations), I discarded the first 100,000 as “burn-in” (num.burnin), and I sampled an MCMC iteration every 10,000 (num.thin). The final results consisted of 30 MCMC samples following the formula $M = (\text{num.iterations} - \text{num.burnin}) / \text{num.thin}$. For each Target population, I retained as the final result the inferred contributions to the mixture model of the MCMC iteration with the highest posterior probability over 20 independent runs (Supplementary Figure S 2.8).

Dating admixture events with fastGLOBETROTTER

I used fastGLOBETROTTER (Wangkumhang et al. 2022) to infer admixture using the Surrogate populations imputed by SOURCEFIND for each Target São Toméan cluster, and dated admixture events based on the LD decay patterns among haplotypes matching any pair of Surrogates. I used the same co-ancestry matrix used for the third SOURCEFIND analysis. Additionally, I prepared a copying vector file in which each Target genome is copying from each individual genome in the Donor population samples

for dating admixture events. I used all the populations of the reference panel including CV as Donors, and the populations imputed by SOURCEFIND for each São Toméan cluster as Surrogates.

I first ran fastGLOBETROTTER separately for each cluster setting “null.ind: 1”, inferring admixture proportions, dates, and sources over 5 iterations, and performing 100 bootstrap re-samples. In all bootstrap resamples for all Target São Toméan cluster, no date estimate was less or equal to 1, nor more or equal to 400. Therefore, the p-value for evidence of “any detectable admixture” was 0 in all cases.

In this study, I presented the output of a second fastGLOBETROTTER run, setting “null.ind: 0”. The results of the first and second run were consistent in terms of best guess of admixture dating. The “null individual analysis” is meant to detect signals in the LD decay curve due to strong genetic drift (Hellenthal et al. 2014). Moreover, fastGLOBETROTTER can detect if the left end of the co-ancestry curve is affected by long chunks, and eventually trim it prior to model fitting. Despite these corrections, the authors advise caution, as LD effects can still influence results (Wangkumhang et al. 2022).

Four São Toméan clusters better fitted a single admixture pulse, while cluster C5 fitted multiple pulses, since the additional goodness-of-fit (R^2) explained by adding a second date was 0.44, which is higher than the threshold of 0.35 suggested by Hellenthal et al. (2014). For all pairs of Surrogate populations of cluster C5, there is an increased probability to copy DNA segments separated by small cM distances. However, the 1-date curve fit the data as well as the 2-date curve starting from 2 cM distances. I thus decided to retain the 1-date results for C5, also because the 2-date results accounted for a pulse of admixture between Cabo Verdean populations 197 generations ago, which is highly improbable since Cabo Verdean populations have been founded in the late 15th century as per historical records (Albuquerque and Santos 1991). Confidence intervals for the date of admixture were obtained from 100 bootstrap replicates (Figure 2.10).

Long identical by descent (IBD) tracts

I used hapIBD (Zhou et al. 2020) to generate shared identity-by-descent (IBD) segments from the phased data of 96 São Toméan and 225 Cabo Verdean unrelated individuals. I used the same genetic distance map used for phasing, and I employed the seed-and-

extend algorithm implemented in hap-IBD with default parameters including a maximum non-IBS gap of 1,000 base pairs and a minimum extension length of 1 cM.

I found a total of 259,624 IBD segments shared between the 323 unrelated individuals from São Tomé e Príncipe. I filtered for IBD segments larger than 18 cM, taking into account the last 8 generations of recombination according to previous calculations (Baharian et al. 2016). By summing the length of all shared IBD segments greater than 18 cM, I calculated the cumulative IBD sharing between each pair of individuals. I plotted the resulting matrix as a heat map, capped at 10,000 cM (Supplementary Figure S 2.10). In Figure 6A, I built a network based on this matrix using the “networkx” package in Python (Hagberg et al. 2008).

Runs of Hemizyosity (ROH)

I called runs of homozygosity (ROH) with GARLIC (Szpiech et al. 2017), considering the five genetic clusters in São Tomé as separate populations, as well as 59 individuals from the island of Santiago, in Cabo Verde, 20 Gambian GWD individuals, 20 Iberian IBS individuals, 11 Kongo individuals, and 20 Yoruba YRI individuals. For each population separately, I ran GARLIC using the weighted logarithm of the odds (wLOD) (Blant et al. 2017), with a genotyping-error rate of 0.001, and using the same genetic distance map used for phasing, window sizes ranging from 30 to 90 SNPs in increments of 10 SNPs, 100 resampling to estimate allele frequencies, and all other GARLIC parameters set to default values. GARLIC calculates the ROH using window sizes for ROH detection and length boundaries for ROH class detection separately for each population, depending on the specific allele frequencies and distributions of ROH lengths, respectively (Supplementary Table S 2.5). I calculated the sum of ROH for each class length, for each individual (Figure 2.9B).

Supplementary Figures

Supplementary Figure S 2.1. Multi-Dimensional Scaling (MDS) analysis.

Supplementary Figure S 2.2. Alternative ADMIXTURE mode for $K=5$.

Supplementary Figure S 2.3. Cross-validation error of 10 independent ADMIXTURE runs for K from 2 to 15.

Supplementary Figure S 2.4. Unsupervised ADMIXTURE analysis for K from 2 to 15.

Supplementary Figure S 2.5. fineSTRUCTURE dendrogram of the São Toméan sample.

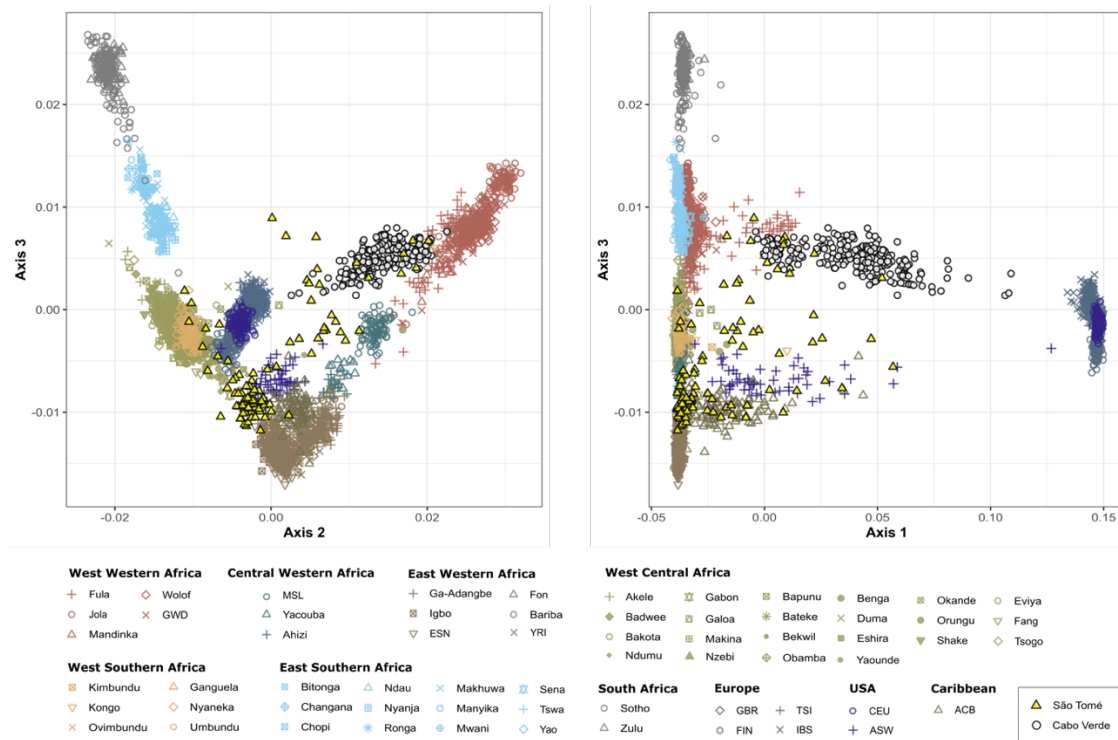
Supplementary Figure S 2.6. Principal Component Analysis (PCA) based on the co-ancestry matrix of the São Toméan sample.

Supplementary Figure S 2.7. Pairwise coincidence matrix of São Toméan individual samples.

Supplementary Figure S 2.8. MCMC iterations with the highest likelihood over 20 independent SOURCEFIND runs.

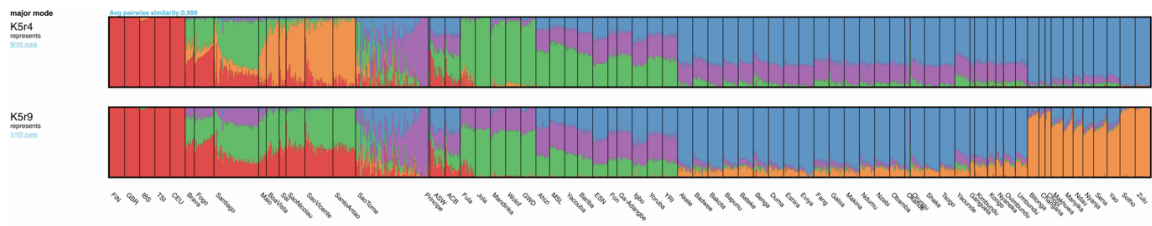
Supplementary Figure S 2.9. Haplotype-based PCA.

Supplementary Figure S 2.10. Heatmap of the cumulative length of long Identical by Descent (IBD) tracts.



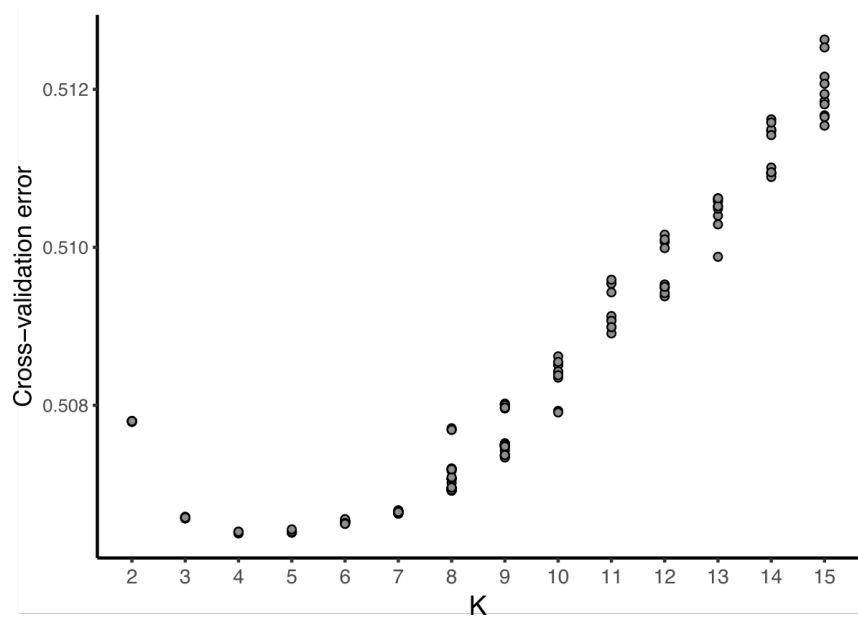
Supplementary Figure S 2.1. Multi-Dimensional Scaling (MDS) analysis.

This figure extends Figure 2.3 B by including the third axis of variation in the MDS projection of pairwise allele sharing dissimilarities (ASD, Bowcock et al. 1994). The MDS includes São Toméans, Cabo Verdeans, and various African, American, and European populations, with the projection based on 3203 individuals and 411,121 autosomal SNPs.



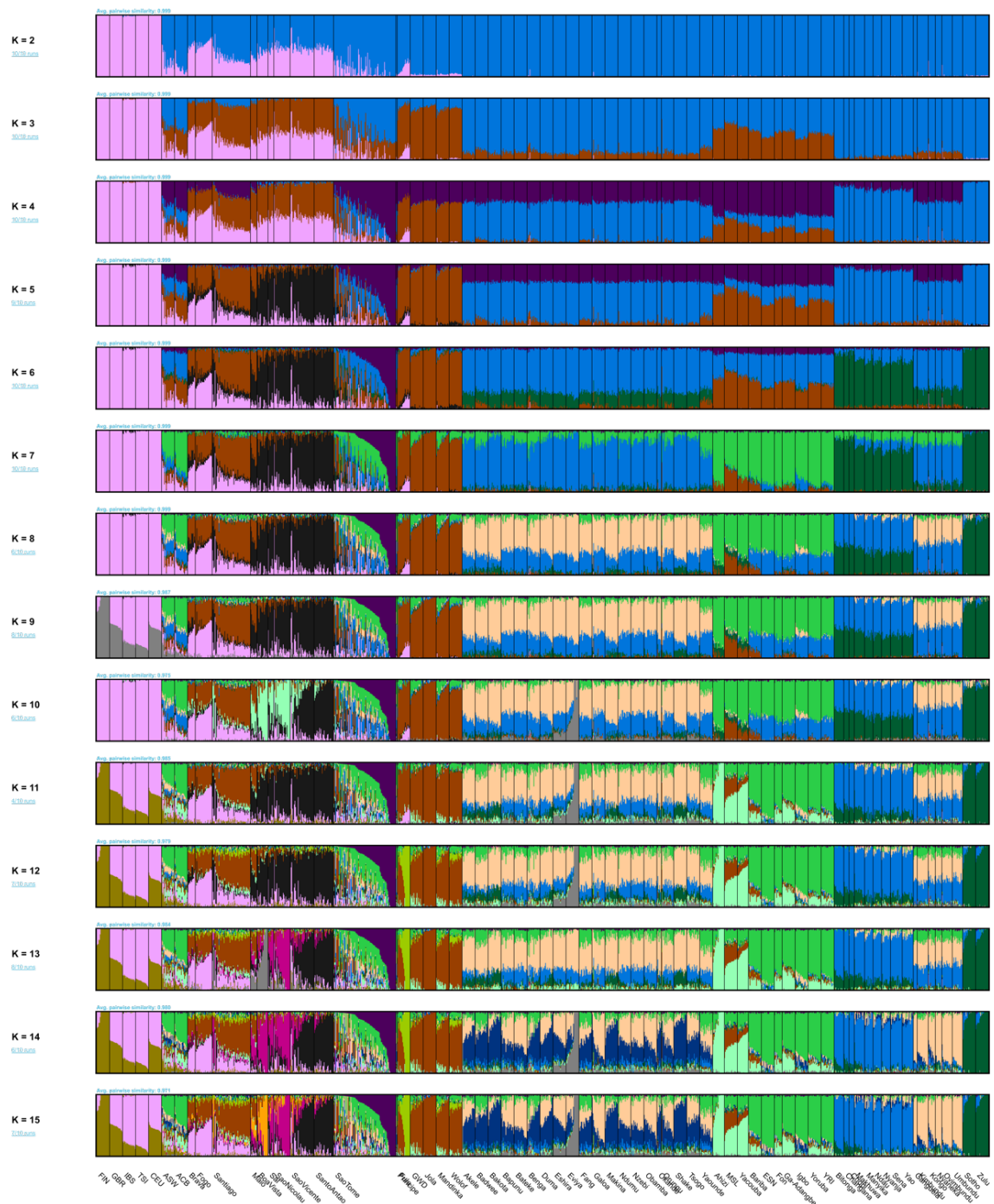
Supplementary Figure S 2.2. Alternative ADMIXTURE mode for $K=5$.

Alternative admixture mode for $K=5$, differing from the major mode reported in Figure 2.4 that represents 9 out of 10 independent ADMIXTURE runs. The average similarity between this alternative mode and the major mode is 0.814581, as calculated with PONG.



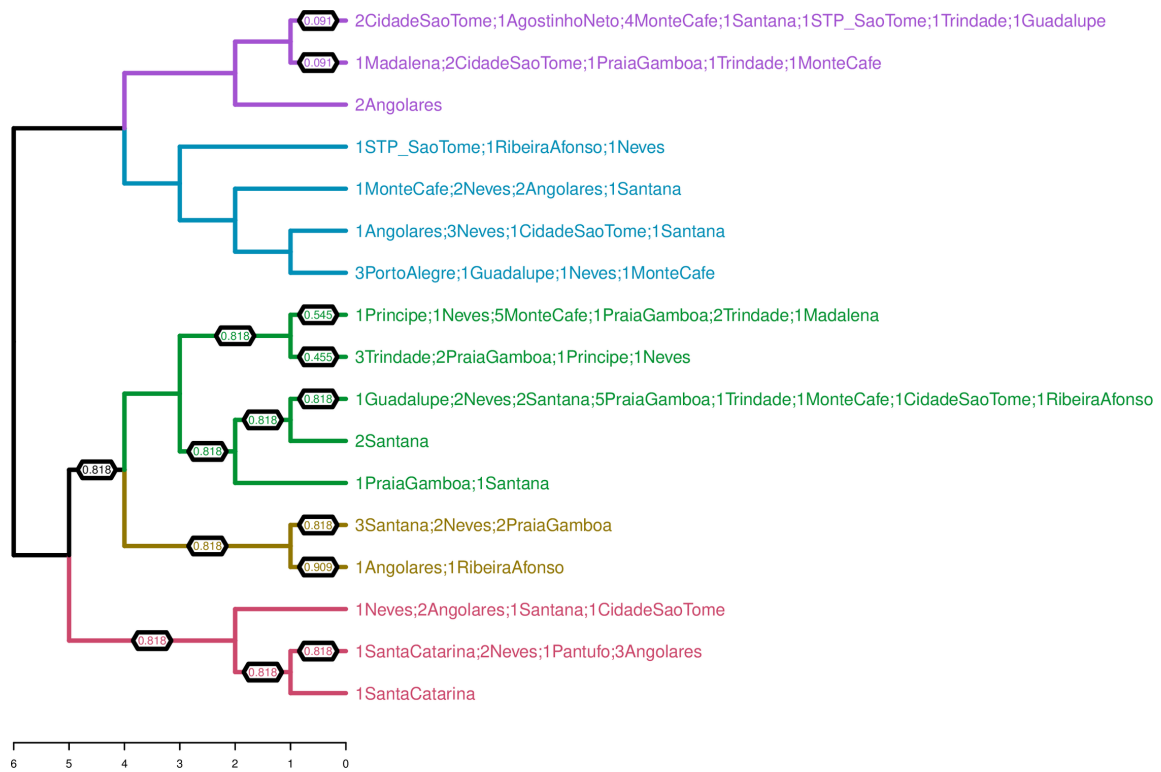
Supplementary Figure S 2.3. Cross-validation error of 10 independent ADMIXTURE runs for K from 2 to 15.

The cross-validation error for 10 independent ADMIXTURE runs begins to increase starting from $K=7$.



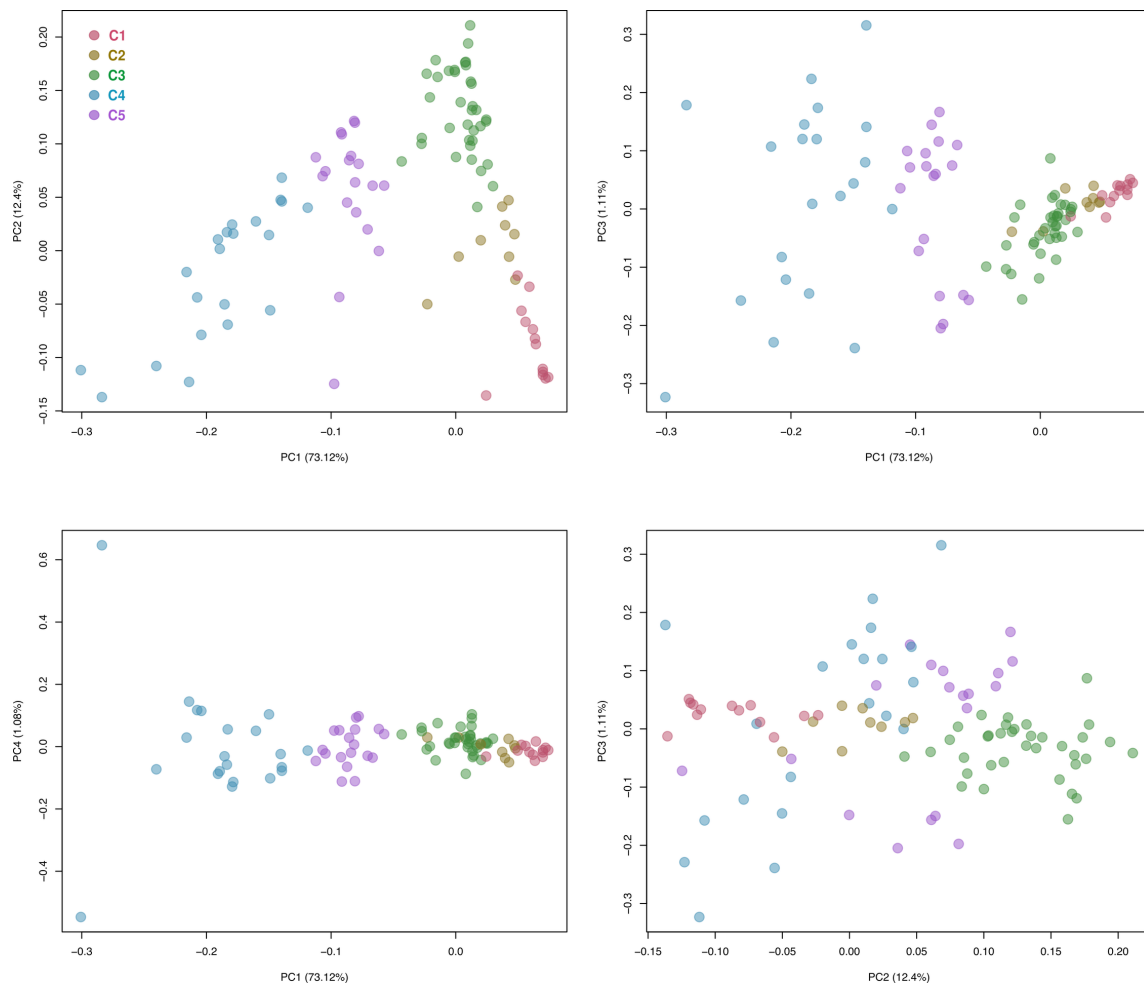
Supplementary Figure S 2.4. Unsupervised ADMIXTURE analysis for K from 2 to 15.

Unsupervised ADMIXTURE analysis as in Figure 2.4. Here, results for $K > 7$ are reported.



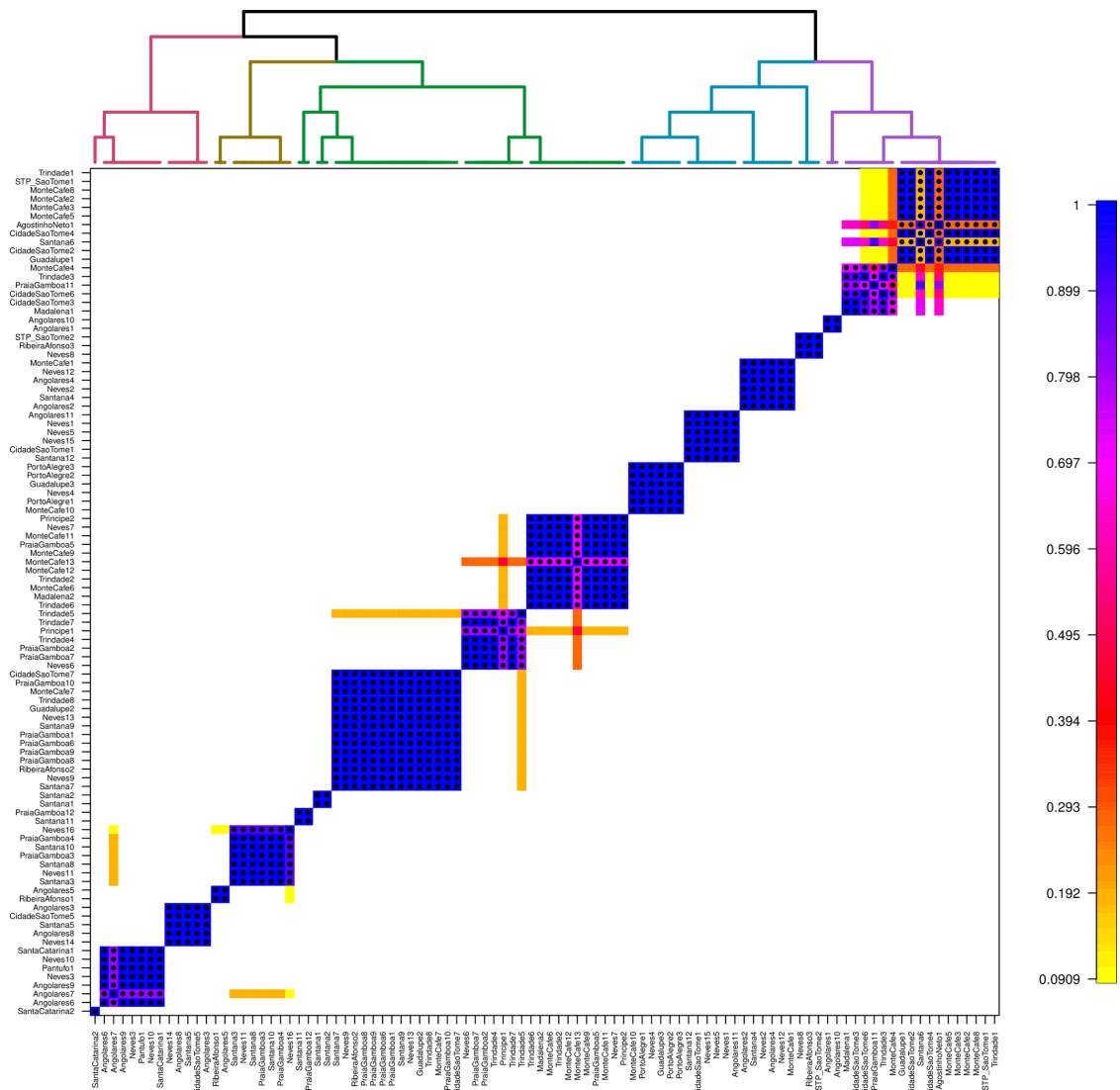
Supplementary Figure S 2.5. fineSTRUCTURE dendrogram of the São Toméan sample.

The numbers on the edges of the dendrogram give the proportion of MCMC iterations for which each population split is observed (only displayed when the proportion is below 1).



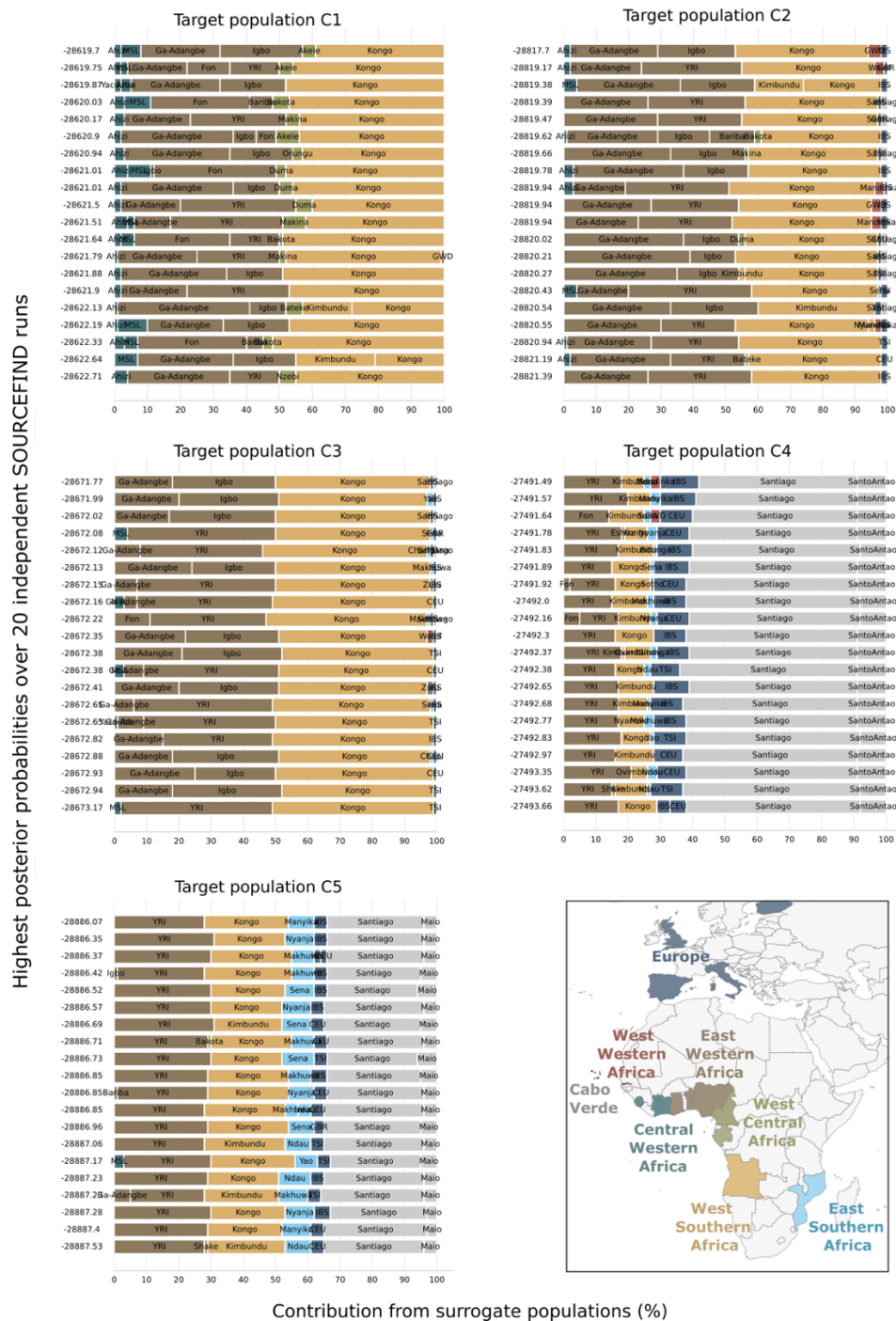
Supplementary Figure S 2.6. Principal Component Analysis (PCA) based on the co-ancestry matrix of the São Toméan sample.

This figure expands upon Figure 2.7 B by including additional principal components, specifically PC3 and PC4.



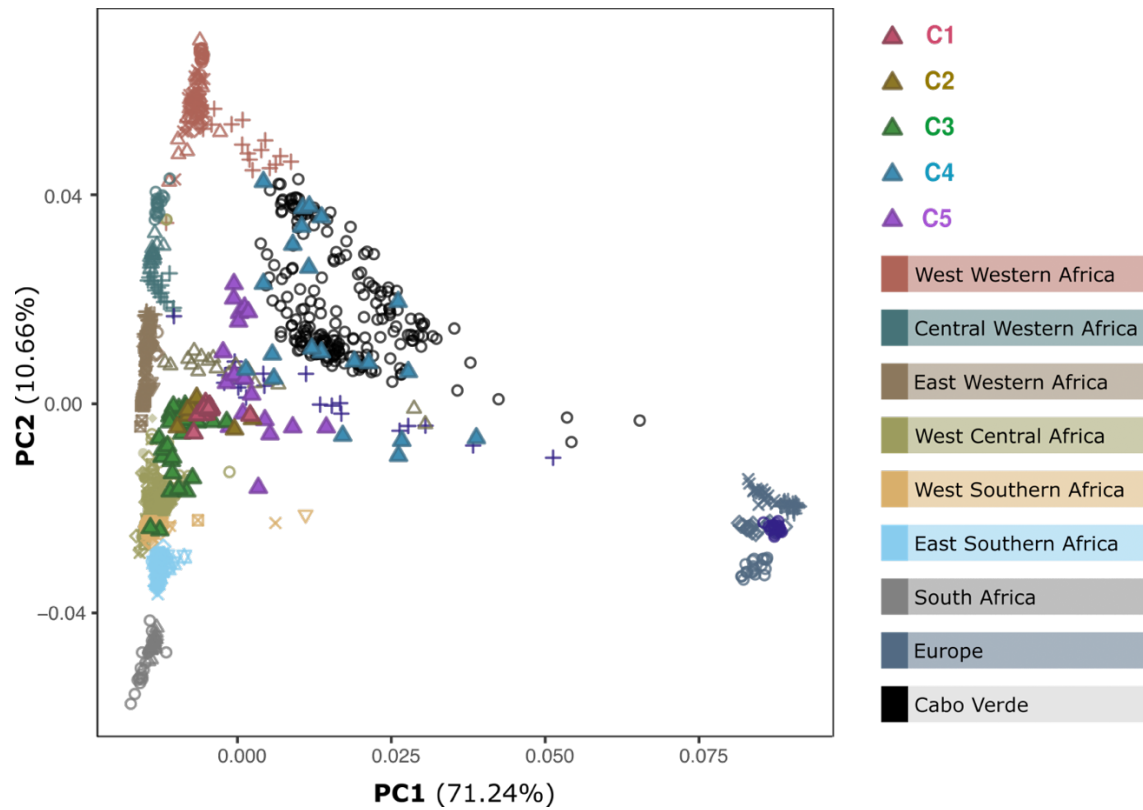
Supplementary Figure S 2.7. Pairwise coincidence matrix of São Toméan individual samples.

The coincidence matrix is used to summarize the results of fineSTRUCTURE Markov Chain Monte Carlo (MCMC) clustering process. It captures how consistently pairs of individuals are grouped together across different iterations of the MCMC process. The colouring represents the average pairwise coincidence across MCMC samples. If the value is close to 1, it means that the corresponding pair of individuals is almost always grouped together in the same cluster, indicating strong genetic similarity.



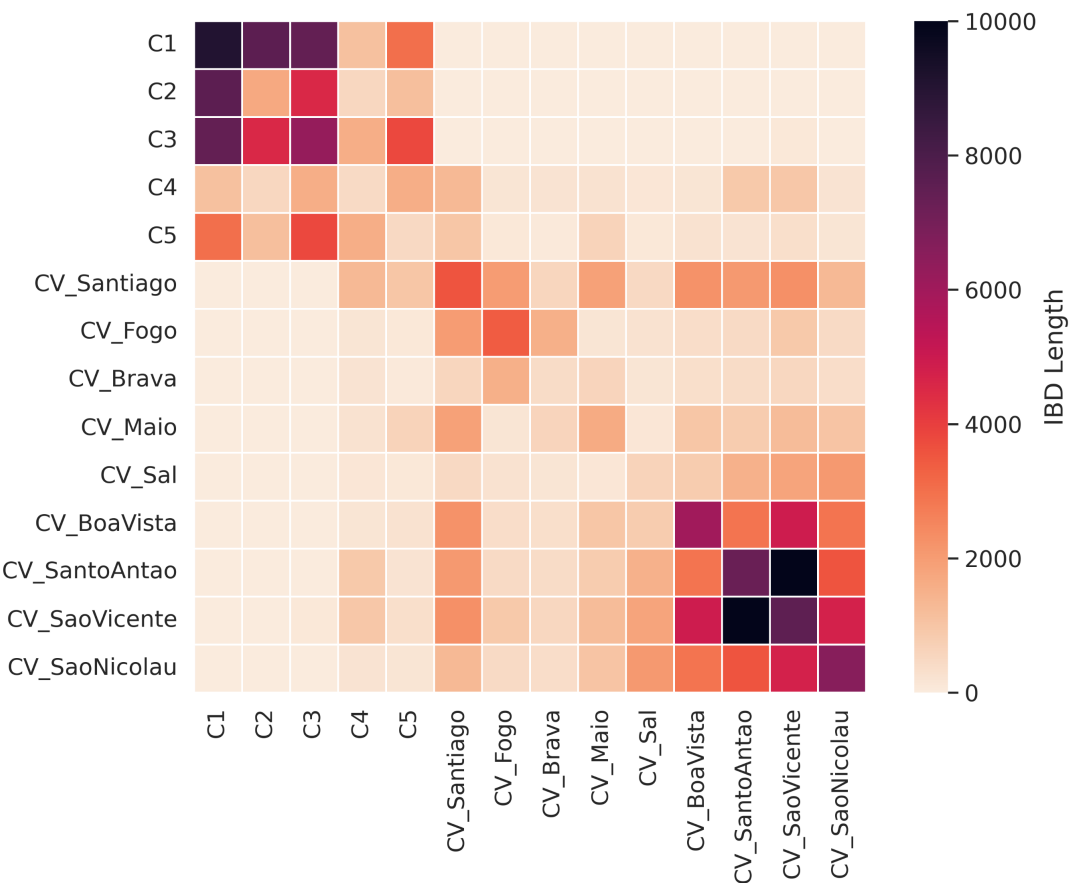
Supplementary Figure S 2.8. MCMC iterations with the highest likelihood over 20 independent SOURCEFIND runs.

The results are ordered by posterior probability and illustrate both the consistency and variability in the source estimates across different runs.



Supplementary Figure S 2.9. Haplotype-based PCA.

Principal Component Analysis (PCA) based on chromosome painting with Chromopainter2 using all the 1347 individuals of the Working Dataset as both Donors and Recipients. São Toméan samples are coloured according to genetic clusters (C1-C5).



Supplementary Figure S 2.10. Heatmap of the cumulative length of long Identical by Descent (IBD) tracts.

Alternative representation of the results shown in Figure 6A, displaying the cumulative length of IBD tracts longer than 18 cM, that are shared within and between samples from São Tomé (C1-C5) and Cabo Verde.

Supplementary Tables

Supplementary Table S 2.1. Genotyping details for the São Toméan and Cabo Verdean sample batches. Batch codes, sampling locations, genotyping years, and the Illumina manifest used for each batch.

Supplementary Table S 2.2. Overview of the Genotyping Quality Control (QC) Pipeline across four phases.

Supplementary Table S 2.3. Population datasets included in this study. The original publication source for each population dataset.

Supplementary Table S 2.4. Population datasets included in the Working Dataset. Populations and number of individual samples included in the analyses presented in this study.

Supplementary Table S 2.5. Length boundaries of ROH classes as calculated by GARLIC.

Supplementary Table S 2.1. Genotyping details for the São Toméan and Cabo Verdean sample batches.

<i>Code</i>	<i>Sampling place</i>	<i>Genotyping Year</i>	<i>Illumina Manifest</i>	<i>Number of Samples*</i>	<i>Number of Markers**</i>
CV1	Cabo Verde	2010-2011	HumanOmni2.5-8v1_C	48	2 372 616
CV2	Cabo Verde	2014	HumanOmni2-5 Exome-8-v1-0-B	42	2 565 665
CV3	Cabo Verde	2016-2017	InfiniumOmni2-5-8-v1-3_A1	128	2 369 989
CV4	Cabo Verde	2018	InfiniumOmni2-5-8v1-4_A1	80	2 381 001
ST1	São Tomé e Príncipe	2020	HumanOmni2.5-8v1_C	111	2 379 172

* including duplicated samples

** including markers on sex-chromosomes and mitochondrial DNA

Supplementary Table S 2.2. Overview of the Genotyping Quality Control (QC) Pipeline across four phases.

Phase 1) Genotype Calling

In Phase 1, we called genotypes using Illumina GenomeStudio Genotyping Module version 1.9.4 for each batch separately. We used GenCall version 7.0.0 with a low cutoff of 0.15. We removed markers with ambiguous positions on the human genome reference sequence, and we removed markers on sex chromosomes and mitochondrial DNA. We filtered autosomal SNPs that failed the 95% call rate and 0.2 cluster separation cutoffs, and samples that failed the 95% SNP call rate cutoff. Transversions (A/T and G/C markers) and markers with duplicated positions were also removed.

	CV1	
	Number of Markers	Number of Samples
<i>Raw data loaded in GenomeStudio</i>	2 379 855	48
<i>Removing markers with ambiguous positions</i>	-7 239	
<i>Removing markers on sex-chromosomes and mtDNA</i>	-58 442	
<i>Autosomal dataset before GQC</i>	2 314 174	48
<i>Updating SNP statistics</i>		
<i>Removing markers call rate < 90%</i>	-6 738	
<i>Recalculating sample statistics</i>		
<i>Removing samples call rate <95%</i>		
<i>Clustering SNPs and update statistics</i>		
<i>Removing markers call rate < 95%</i>	-8 654	
<i>Removing markers cluster sep. <0.2</i>	-2	
<i>Removing symmetric A/T and G/C markers</i>	-73 001	
<i>Removing markers with duplicated positions</i>	-3 141	
<i>Autosomal dataset after GQC Phase 1)</i>	2 222 638	48

CV2		CV3	
Number of Markers	Number of Samples	Number of Markers	Number of Samples
2 567 845	42	2 372 784	128
-2 180		-2 795	
-60 994		-58 087	
2 504 671	42	2 311 902	128
-5 042		-2 687	
	-1		-1
-14 349		-4 619	
-55		-218	
-97 389		-71 826	
-37 499		-6 911	
2 350 338	41	2 225 646	127

CV4		ST1	
Number of Markers	Number of Samples	Number of Markers	Number of Samples
2 382 209	80	2 380 364	111
-1 208		-1 192	
-58 566		-57 595	
2 322 435	80	2 321 577	111
-1 166		-1 162	
			-1
-4 493		-10 366	
-74		-118	
-73 133		-72 333	
-6 987		-6 991	
2 236 454	80	2 230 107	110

Phase 2) Batch Merging

In Phase 2, we only kept markers genotyped in all 5 batches using *ad hoc* Python scripts. To check the concordance rate of genotype calls across batches, we genotyped 32 individuals separately on two, three, or four different batches. We removed SNPs that were called differently on different batches for the same duplicated individual sample. We then removed the duplicated individual samples with the highest missing rate. Finally, we remapped markers positions based on rsIDs according to dbSNP Human Build 154 Release on GRCh38 genome assembly (Genome Reference Consortium Human Build 38). We removed markers whose rsIDs have multiple positions in the same build, or that do not have a position in this build. We finally retained a total of 2,104,297 autosomal markers mapped to the GRCh38 genome assembly (hg38) from 358 samples.

	Number of Markers	Number of Samples
<i>Autosomal markers with same rsID among batches</i>	2 111 666	406
<i>Removing markers that differ among replicated samples</i>	-6 878	
<i>Removing duplicated samples</i>		-48
<i>Removing markers without position in dbSNP</i>	-491	
<i>Autosomal dataset after GQC Phase 2)</i>	2 104 297	358

Phase 3) Genetic Relatedness

In Phase 3, we removed genetically related individuals up to the 2nd degree using KING version 2.2.7 (Manichaikul et al. 2010).

	Number of Markers	Number of Samples
<i>Removing related individuals up to 2nd degree</i>		-28
<i>Autosomal dataset after GQC Phase 3)</i>	2 104 297	330

Phase 4) Population Genetics QC

In Phase 4, we conducted quality control filters using PLINK version 1.9 (Chang et al. 2015). We removed markers with high missing data (10%), markers that deviate from Hardy-Weinberg equilibrium (HWE), and insertion/deletions. We thus retained 2,104,148 curated autosomal SNPs from 330 unrelated individuals, including 233 unrelated individuals sampled in Cabo Verde and 97 unrelated individuals sampled in São Tomé e Príncipe.

	Number of Markers	Number of Samples
<i>Removing markers with >10% missing data</i>		
<i>Removing markers in HW disequilibrium ($p\text{-value} < 2.63 \times 10^{-8}$)</i>	-149	
<i>Autosomal dataset after GQC Phase 4)</i>	2 104 148	330

Supplementary Table S 2.3. Population datasets included in this study.

The original publication source for each population dataset.

Dataset	Continent	State	Region	Population	Size
<i>This study</i>	AFR	CaboVerde	CVSTP	Santiago	59
<i>This study</i>	AFR	CaboVerde	CVSTP	SãoVicente	37
<i>This study</i>	AFR	CaboVerde	CVSTP	SantoAntão	30
<i>This study</i>	AFR	CaboVerde	CVSTP	Fogo	26
<i>This study</i>	AFR	CaboVerde	CVSTP	SãoNicolau	25
<i>This study</i>	AFR	CaboVerde	CVSTP	BoaVista	17
<i>This study</i>	AFR	CaboVerde	CVSTP	Brava	12
<i>This study</i>	AFR	CaboVerde	CVSTP	Maio	10
<i>This study</i>	AFR	CaboVerde	CVSTP	Sal	9
<i>This study</i>	AFR	CaboVerde	CVSTP	NA	1
<i>This study</i>	AFR	SãoToméPríncipe	CVSTP	SãoTomé	96
<i>This study</i>	AFR	SãoToméPríncipe	CVSTP	Príncipe	2
<i>This study</i>	EUR	Portugal	Europe	Portugal	1
<i>This study</i>	AMR	Brazil	SouthAmerica	Brazil	1
<i>This study</i>	AFR	Angola	WestSouthernAfrica	Angola	3
<i>This study</i>	AFR	Gabon	WestCentralAfrica	Gabon	1
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Sena	18
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Yao	17
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Makhuwa	16
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Bitonga	15
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Nyanja	14
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Manyika	13
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Ndau	13
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Changana	8
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Chopi	8
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Mwani	5
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Tswa	5
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Ronga	3
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Umbundu	16
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Nyaneka	10
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Ganguela	6
<i>Patin et al. 2017</i>	AFR	CAR	CentralAfrica	Biaka	15
<i>Patin et al. 2017</i>	AFR	DRC	CentralAfrica	Mbuti	13
<i>Patin et al. 2017</i>	AFR	IvoryCoast	CentralWesternAfrica	Ahizi	18

<i>Patin et al. 2017</i>	AFR	IvoryCoast	CentralWesternAfrica	Yacouba	17
<i>Patin et al. 2017</i>	AFR	Rwanda	EastAfrica	Twa	2
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Bariba	20
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Yoruba	20
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Fon	12
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Kimbundu	17
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Ovimbundu	16
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Kongo	11
<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Yaounde	39
<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Badwee	38
<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Bezan	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Baka	98
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Fang	61
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Nzebi	61
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Tsogo	60
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bapunu	51
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Shake	47
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bakota	46
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Galoa	46
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Obamba	45
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bateke	41
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Eshira	40
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Duma	39
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Makina	39
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Benga	38
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Akele	37
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Ndumu	37
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	BabongoE	27
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Eviya	22
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bakoya	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Orungu	19
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Okande	7
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bekwil	5
<i>Gurdasani et al. 2015</i>	AFR	Ethiopia	EastAfrica	Ethiopia	107
<i>Gurdasani et al. 2015</i>	AFR	Kenya	EastAfrica	Kalenjin	100
<i>Gurdasani et al. 2015</i>	AFR	Kenya	EastAfrica	Kikuyu	99
<i>Gurdasani et al. 2015</i>	AFR	Uganda	EastAfrica	Baganda	99
<i>Gurdasani et al. 2015</i>	AFR	Uganda	EastAfrica	Banyarwanda	96

<i>Gurdasani et al. 2015</i>	AFR	Uganda	EastAfrica	Barundi	93
<i>Gurdasani et al. 2015</i>	AFR	Ghana	EastWesternAfrica	Ga-Adangbe	100
<i>Gurdasani et al. 2015</i>	AFR	Nigeria	EastWesternAfrica	Igbo	99
<i>Gurdasani et al. 2015</i>	AFR	SouthAfrica	SouthAfrica	Zulu	100
<i>Gurdasani et al. 2015</i>	AFR	SouthAfrica	SouthAfrica	Sotho	86
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Mandinka	87
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Jola	79
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Wolof	78
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Fula	74
<i>Auton et al. 2015</i>	AMR	ACB	Carribean	ACB	95
<i>Auton et al. 2015</i>	AMR	PUR	Carribean	PUR	104
<i>Auton et al. 2015</i>	AFR	MSL	CentralWesternAfrica	MSL	85
<i>Auton et al. 2015</i>	AFR	LWK	EastAfrica	LWK	97
<i>Auton et al. 2015</i>	EAS	CDX	EastAsia	CDX	93
<i>Auton et al. 2015</i>	EAS	CHB	EastAsia	CHB	103
<i>Auton et al. 2015</i>	EAS	CHS	EastAsia	CHS	104
<i>Auton et al. 2015</i>	EAS	JPT	EastAsia	JPT	104
<i>Auton et al. 2015</i>	EAS	KHV	EastAsia	KHV	99
<i>Auton et al. 2015</i>	AFR	ESN	EastWesternAfrica	ESN	98
<i>Auton et al. 2015</i>	AFR	YRI	EastWesternAfrica	YRI	108
<i>Auton et al. 2015</i>	EUR	FIN	Europe	FIN	99
<i>Auton et al. 2015</i>	EUR	GBR	Europe	GBR	90
<i>Auton et al. 2015</i>	EUR	IBS	Europe	IBS	107
<i>Auton et al. 2015</i>	EUR	TSI	Europe	TSI	107
<i>Auton et al. 2015</i>	AMR	CLM	SouthAmerica	CLM	94
<i>Auton et al. 2015</i>	AMR	PEL	SouthAmerica	PEL	85
<i>Auton et al. 2015</i>	SAS	BEB	SouthAsia	BEB	86
<i>Auton et al. 2015</i>	SAS	GIH	SouthAsia	GIH	100
<i>Auton et al. 2015</i>	SAS	ITU	SouthAsia	ITU	102
<i>Auton et al. 2015</i>	SAS	PJL	SouthAsia	PJL	95
<i>Auton et al. 2015</i>	SAS	STU	SouthAsia	STU	99
<i>Auton et al. 2015</i>	AMR	ASW	USA	ASW	55
<i>Auton et al. 2015</i>	EUR	CEU	USA	CEU	99
<i>Auton et al. 2015</i>	AMR	MXL	USA	MXL	64
<i>Auton et al. 2015</i>	AFR	GWD	WestWesternAfrica	GWD	113

Supplementary Table S 2.4. Population datasets included in the Working Dataset.

Populations and number of individual samples included in the analyses presented in this study.

<i>Dataset</i>	<i>Continent</i>	<i>State</i>	<i>Region</i>	<i>Population</i>	<i>Size</i>
<i>This study</i>	AFR	CaboVerde	CVSTP	Santiago	59
<i>This study</i>	AFR	CaboVerde	CVSTP	SãoVicente	37
<i>This study</i>	AFR	CaboVerde	CVSTP	SantoAntao	30
<i>This study</i>	AFR	CaboVerde	CVSTP	Fogo	26
<i>This study</i>	AFR	CaboVerde	CVSTP	SãoNicolau	25
<i>This study</i>	AFR	CaboVerde	CVSTP	BoaVista	17
<i>This study</i>	AFR	CaboVerde	CVSTP	Brava	12
<i>This study</i>	AFR	CaboVerde	CVSTP	Maio	10
<i>This study</i>	AFR	CaboVerde	CVSTP	Sal	9
<i>This study</i>	AFR	SãoToméPríncipe	CVSTP	SãoTomé	96
<i>This study</i>	AFR	SãoToméPríncipe	CVSTP	Príncipe	2
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Sena	18
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Yao	17
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Makhuwa	16
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Bitonga	15
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Nyanja	14
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Manyika	13
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Ndau	13
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Changana	8
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Chopi	8
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Umbundu	16
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Nyaneka	10
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Ganguela	6
<i>Patin et al. 2017</i>	AFR	IvoryCoast	CentralWesternAfrica	Ahizi	18
<i>Patin et al. 2017</i>	AFR	IvoryCoast	CentralWesternAfrica	Yacouba	17
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Bariba	20
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Yoruba	20
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Fon	12
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Kimbundu	17
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Ovimbundu	16
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Kongo	11
<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Badwee	20
<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Yaounde	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Akele	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bakota	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bapunu	20

<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bateke	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Benga	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Duma	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Eshira	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Eviya	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Fang	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Galoa	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Makina	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Ndumu	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Nzebi	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Obamba	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Shake	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Tsogo	20
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Orungu	19
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Okande	7
<i>Gurdasani et al. 2015</i>	AFR	Ghana	EastWesternAfrica	Ga-Adangbe	20
<i>Gurdasani et al. 2015</i>	AFR	Nigeria	EastWesternAfrica	Igbo	20
<i>Gurdasani et al. 2015</i>	AFR	SouthAfrica	SouthAfrica	Sotho	20
<i>Gurdasani et al. 2015</i>	AFR	SouthAfrica	SouthAfrica	Zulu	20
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Fula	20
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Jola	20
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Mandinka	20
<i>Gurdasani et al. 2015</i>	AFR	Gambia	WestWesternAfrica	Wolof	20
<i>Auton et al. 2015</i>	AMR	Barbados	Carribean	ACB	20
<i>Auton et al. 2015</i>	AFR	SierraLeone	CentralWesternAfrica	MSL	20
<i>Auton et al. 2015</i>	AFR	Nigeria	EastWesternAfrica	ESN	20
<i>Auton et al. 2015</i>	AFR	Nigeria	EastWesternAfrica	YRI	20
<i>Auton et al. 2015</i>	EUR	Finland	Europe	FIN	20
<i>Auton et al. 2015</i>	EUR	England	Europe	GBR	20
<i>Auton et al. 2015</i>	EUR	Spain	Europe	IBS	20
<i>Auton et al. 2015</i>	EUR	Italy	Europe	TSI	20
<i>Auton et al. 2015</i>	AMR	SouthWestUSA	USA	ASW	20
<i>Auton et al. 2015</i>	EUR	Utah	USA	CEU	20
<i>Auton et al. 2015</i>	AFR	Gambia	WestWesternAfrica	GWD	20

Supplementary Table S 2.5. Length boundaries of ROH classes as calculated by GARLIC.

<i>Population</i>	<i>Short/Medium ROH Length Boundary (Mb)</i>	<i>Medium/Long ROH Length Boundary (Mb)</i>	<i>Window Size</i>	<i>LOD Score Cutoff</i>
C1	0.641322	1.85764	50	1.52488
C2	0.527860	1.47489	50	2.83498
C3	0.722295	2.01409	60	3.96535
C4	0.510380	1.32756	60	7.56954
C5	0.699838	2.42269	60	4.43695
Santiago	0.692757	2.00713	60	4.74461
GWD	0.689603	1.96557	60	4.19922
IBS	0.443377	1.45504	40	1.00673
Kongo	0.764603	2.11275	50	-48.5687
YRI	0.648575	2.02074	60	6.57242

SUPPLEMENTARY INFORMATION FOR CHAPTER 3

Supplementary Materials and Methods

Genetic Dataset

Genotyping Quality Control and Merging

We genotyped markers on the X chromosome, NRY chromosome, and, mitochondrial DNA for the first time across 1,922 individual samples from 49 populations, including Cabo Verde and São Tomé e Príncipe (This study), as well as populations from Angola and Mozambique (Semo et al. 2020), and other population samples from Central, East and West Africa (Patin et al. 2017). Additionally, we merged these data with markers on the X chromosome and NRY chromosome from 26 worldwide population samples included in the 1000 Genome Project (Auton et al. 2015). Specifically, we genotyped 361 DNA samples (261 samples from Cabo Verde and 100 samples from São Tomé e Príncipe) in 5 batches, separately, using different versions of the Illumina HumanOmni2.5Million-BeadChip genotyping array at the OMICS platform of the Institut Pasteur. We also curated for the first time markers on the X chromosome, Y Chromosome and mitochondrial DNA on the 200 samples from Angola and Mozambique (Semo et al. 2020) genotyped using a similar HumanOmni2.5Million-BeadChip, and markers on the X chromosome and Y Chromosome on the 1366 samples from Central, East and West Africa (Patin et al. 2017) genotyped using the Illumina InfiniumOmniExpress720K-BeadChip.

We conducted comprehensive Genotyping Quality Control during the merging of datasets, outlined in three phases (Supplementary Table S 3.1).

Phase 1) For each batch, we called genotypes using Illumina GenomeStudio Genotyping Module version 1.9.4. We used GenCall version 7.0.0 with a low cutoff of 0.15. We removed markers with ambiguous positions on the human genome reference sequence and markers on the autosomes. We filtered separately Autosomes, X chromosome, Y chromosome, and mitochondrial DNA SNPs that failed the 95% call rate and 0.2 cluster separation cutoffs, and samples that failed the 95% call rate cutoff. Symmetric A/T and G/C markers and markers with duplicated positions were also removed. For curating markers on the X chromosome, we first estimated the genetic sex of each individual sample in GenomeStudio, and we temporally excluded male samples for clustering

markers. This permit a better clustering by reducing the noise in the X chromosome data from the males (Illumina 2010).

Phase 2) We merged the 5 batches containing individual samples collected in Cabo Verde and São Tomé e Príncipe. We only kept markers genotyped in all 5 batches using *ad hoc* Python scripts. To check the concordance rate of genotype calls, we genotyped 32 individuals separately on two, three, or four different batches. We removed SNPs that were called differently on different batches for the same duplicated individual sample. We thus removed the duplicated individual samples with the highest missing rate. Finally, we remapped markers positions based on rsIDs according to dbSNP Human Build 154 Release on GRCh38 genome assembly (Genome Reference Consortium Human Build 38). We removed markers whose rsIDs have multiple positions in the same build, or that do not have a position in this build. After filtering, we removed genetically related individuals up to the 2nd degree with KING version 2.2.7 (Manichaikul et al. 2010) using autosomal markers. For autosomal data, genotyping QC is further detailed in the previous study (first paper). We finally retained 4,416 curated SNPs on the X chromosome, 808 on the NRY chromosome, and 123 on mitochondrial DNA from 328, 157, and 331 genetically unrelated individuals, respectively. We refer to this dataset as CVSTP.

Phase 3) We finally merged all the datasets together. Before merging, we converted genome coordinates on the on GRCh38 genome assembly and remapped markers positions based on rsIDs according to dbSNP Human Build 154 on GRCh38. We identified positions shared between each dataset and the previous merged dataset, following the order: CVSTP (first paper), 1000K (Auton et al. 2015), SEMO (Semo et al. 2020), PATIN (Patin et al. 2017). After each merging step, we removed genetically related individuals up to the 2nd degree with KING version 2.2.7 (Manichaikul et al. 2010) using the respective set of autosomal markers. The final datasets include 12,976 curated SNPs on the X chromosome from 4,208 individual samples, 226 curated SNPs on the NRY chromosome from 2,567 individual samples, and 52 curated SNPs on the mitochondrial DNA from 513 individual samples. We refer to these datasets in the format: ChromosomalType_dataset1_dataset2_etc., where the chromosomal type is X, NRY, or MT, and the datasets are listed in the order of successive merging.

X chromosome heterozygosity

For the X_CVSTP_KGP_SEMO_PATIN dataset, we further checked X chromosome heterozygosity (F) as implemented in PLINK. We first inferred the genetic sex of individual samples using the *--check-sex* function. Genetic sex was defined to be female if $F < 0.4$ and male if $F > 0.8$. We removed 8 individual samples with F values included in the interval 0.4-0.8 indicative of ambiguous sex. For the remaining samples, the genetically imputed sex was concordant with self-reported sex when available. We also removed all markers that were heterozygous in more than 5 male individual samples, for a total of 115 positions. We finally retained 12,861 SNPs on the X chromosome for 4,200 individual samples, including 2,534 samples assigned as male and 1,666 as female.

Genotype-based analyses

Allele Sharing Dissimilarity

We investigated genetic diversity patterns between each pair of individuals based on successive subsets of populations in our dataset. We first computed a matrix of pairwise Allele Sharing Dissimilarities (ASD) including all the individuals and all SNPs in the merged dataset using the ASD software (<https://github.com/szpiech/asd>). Then we explored three axes of variation of Multi-Dimensional Scaling projections of various subsets of this ASD pairwise-matrix using the *cmdscale* function in R (R Core Team 2021). We grouped population samples into 16 distinct regions across the globe, including ten regions within Africa, with CVSTP identified separately, three regions in the Americas, and one region each in Europe, South Asia and East Asia (Supplementary Table S 3.2). Figure 3.1 report the ASD-MDS based on 541,734 autosomal SNPs, 12,861 X chromosome SNPs on male samples, and 12,861 X chromosome SNPs on female samples, respectively. For the subset in Figure 3.2 and 3.3, we removed East Asian, South Asian, South American, Puerto Rican PUR, Mexican American MXL, Central African, East African, and four West-Central African hunter-gatherer populations.

ADMIXTURE

We used the software package ADMIXTURE version 1.3 (Alexander et al. 2009) to explore genetic differentiation among individual samples. ADMIXTURE analysis is sensitive to population size heterogeneities, therefore we randomly resampled without replacement 20 individuals for each population, and we removed populations with less

than 5 individuals, except for the admixed populations of interest, where we considered all the individuals: São Toméans, Cabo Verdeans, African Americans, and African Barbadians. We thus retained the same sample set of 902 female and male individuals for the autosomes and X chromosome data.

Following author's recommendations, we filtered SNPs for low Linkage Disequilibrium using the *--indep-pairwise* function in PLINK with a 50 SNP-window moving every 50 SNPs, and 0.2 r^2 cutoff. We thus run ADMIXTURE on 5421 LD-pruned SNPs on the autosomal chromosome 20 and 5364 LD-pruned SNPs on the X chromosome. We run ADMIXTURE using the '*--haploid="male:23"*' command, to account for haploidy on the X chromosome in male individuals (Shringarpure et al. 2016). We performed 10 independent runs of ADMIXTURE for values of K ranging from 2 to 10 (Supplementary Figure 3.1). We used PONG (Behr et al. 2016) to define ADMIXTURE 'modes' with a greedy approach for similarity threshold 0.95.

We used a Wilcoxon signed-rank test to assess the significance of differences in blue cluster proportions between the X chromosome and autosomal chromosome 7. This non-parametric test serves as an alternative to the paired t-test, accounting for paired data without assuming a normal distribution. It ranks the absolute differences between the two groups and evaluates whether the median difference is significantly different from zero.

Haplotype-based analyses

Phasing

We phased X chromosome and autosomal chromosomes separately without reference panel using the 'phase_common' algorithm in SHAPEIT 5 (Hofmeister et al. 2023). We used the genetic map on genome build h38 provided by the authors. For the X chromosome, we used the 'haploids' function for dealing with male samples.

IBD

We inferred genomic tracts that are Identical-by-Descent using hapIBD (Zhou et al. 2020). For the X chromosome, male samples were converted to pseudo-haploid data, and after IBD inference, we manually removed 755,802 IBD tracts on the second male haplotype, representing 48% of the total IBD segments, following the recommendations of the authors (Cai et al. 2023). We finally retain 794,143 IBD segments on the X chromosome.

We filtered IBD segments smaller than 2 cM and larger than 72 cM, as IBD inference on very short segments is less reliable, while very long segments often represent anomalies. We calculated the cumulative sum and total length of shared IBD segments for each pair of population samples. We represented the results in form of heatmaps

Among these IBD segments, we filtered those larger than 18 cM to account for recent common ancestry in the last 8 generations of recombination, according to previous calculations (Baharian et al. 2016). We represented the resulting long-IBD sharing patterns between Cabo Verdeans and São Toméan samples as networks using the “networkx” package in python (Figure 3.5 B).

We developed a permutation test to assess whether the observed difference in IBD sharing between chromosome 7 and the X chromosome was statistically significant for each population pair. A paired test was not appropriate in this case, as the same pairs of individuals do not necessarily share IBD on both chromosome 7 and the X chromosome. Furthermore, the number of haplotypes that can share IBD differ between the autosomes (2 per individual) and the X chromosome (2 per female, 1 per male).

The permutation test allowed us to construct a null distribution by randomly shuffling data labels. First, we calculated the observed difference in total IBD length shared between each population pair for the X chromosome and chromosome 7. To determine whether this observed difference could occur by chance, we generated a null distribution by randomly shuffling the chromosome labels, breaking the association between chromosome type and IBD lengths. For each permutation, we recalculated the total IBD length and calculated the difference between the X chromosome and chromosome 7. This process was repeated 10,000 times to create a distribution of permuted differences.

The p-value was then calculated as the proportion of permuted differences that were equal to or greater than the absolute value of the observed difference (ensuring a two-tailed test). A low p-value indicates that the observed difference is unlikely under the null hypothesis that there is no difference in the distribution of IBD between the X chromosome and chromosome 7. We rejected the null hypothesis if the p-value was less than 0.05, indicating a significant difference in IBD sharing patterns.

Runs of Homozygosity

Runs of homozygosity (ROH) were inferred using GARLIC (Szpiech et al. 2017) on phased data with hg38 genome build. The analysis was performed separately for the autosomes and for the X chromosome, focusing on 4,127 female and male samples on the autosomes and 1,666 female samples using 12,861 diploid markers on the X chromosome. For the male samples, ROH could not be observed as the X chromosome is haploid in males.

We categorized the ROHs into short, medium, and long length classes as defined by GARLIC. To identify the optimal parameters, we explored different window sizes with smoothness threshold. For the X chromosome, the search resulted in the window size optimal value of 30, and the Gaussian mixture model was used to classify the ROHs into three classes, with ROH size limits 558,684 and 1.52201e+06 base pairs. Window sizes were calculated separately on each autosomal chromosome using the same smoothness threshold of 0.5. For chromosome 6, the selected ROH size boundaries were 290998 and 1.06968e+06 base pairs. Future approaches may consider defining additional size classes and/or establishing similar boundaries for the X chromosome and autosomes to allow more meaningful comparisons between the two.

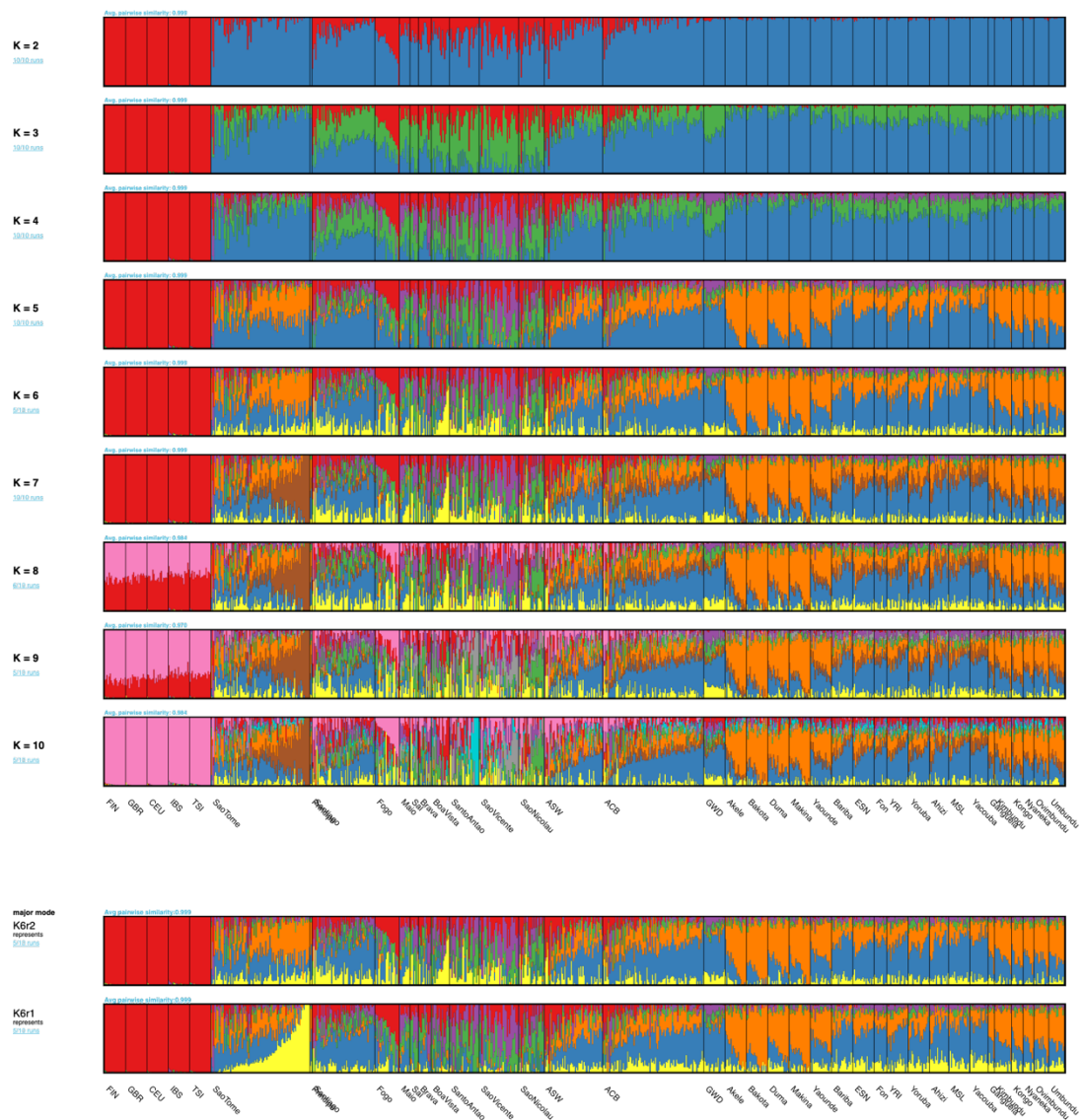
For plotting, we calculated the total ROH length of each class size for each individual on each chromosome and converted the values to Mega bases (Mb). We then visualized the distribution using the 'geom_density_ridges' function in the '**gggridges**' package in R (R Core Team 2021) applying quantile lines to indicate median values. Individual data points were represented as jittered vertical bars to capture variation within populations.

To test for significant differences in total ROH length between the X chromosome and chromosome 7, we performed Wilcoxon rank-sum tests separately for each population sample. Only populations with data for both chromosomes were included in the analysis. The test was implemented using the `wilcox.test()` function in R, with p-values computed without the exact method (`exact = FALSE`).

Supplementary Figures

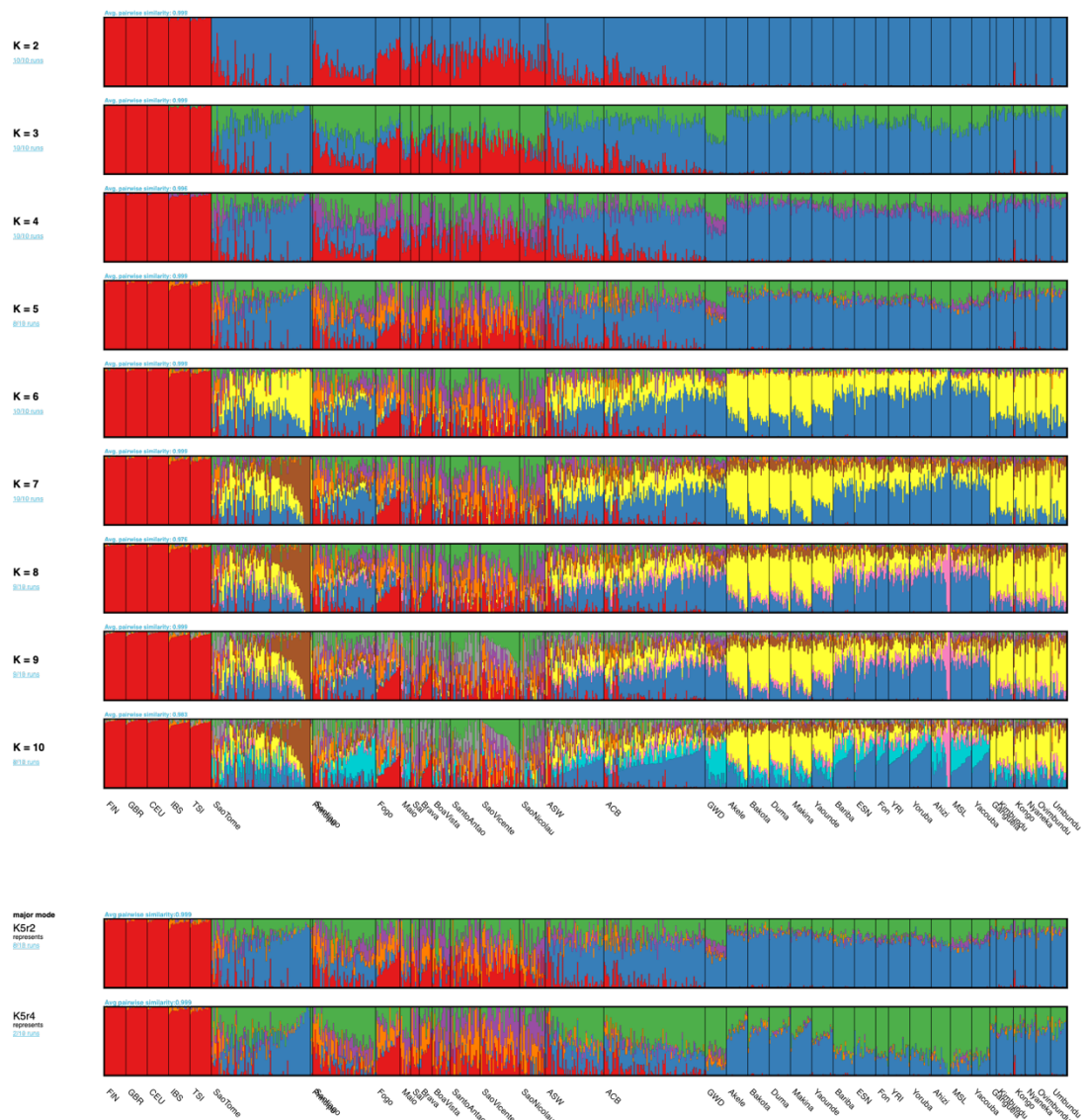
Supplementary Figure 3.1. Unsupervised ADMIXTURE analysis on X chromosome data for $K = 2$ to 10.

Supplementary Figure 3.2. Unsupervised ADMIXTURE analysis on autosomal data for $K = 2$ to 10.



Supplementary Figure 3.1. Unsupervised ADMIXTURE analysis on X chromosome data for K = 2 to 10.

In Figure 3.4, only results at $K = 2$ are presented. At the bottom, the Supplementary Figure Shows alternative clustering modes for $K = 6$, visualized using PONG.



Supplementary Figure 3.1. Unsupervised ADMIXTURE analysis on autosomal data for K = 2 to 10.

In Figure 3.4, only results at $K = 2$ are presented. At the bottom, the Supplementary Figure Shows alternative clustering modes for $K = 5$, visualized using PONG.

Supplementary Tables

Supplementary Table S 3.1. Overview of the Genotyping Quality Control (QC) Pipeline across three phases.

Supplementary Table S 3.2. Population datasets used in the study.

Supplementary Table S 3.3. Population descriptors for 1000 Genome Project Phase 3

Supplementary Table S 3.1. Overview of the Genotyping Quality Control (QC) Pipeline across three phases.

Phase 1) Genotype Calling

	CV1	
	Number of Markers	Number of Samples
<i>Raw data loaded in GenomeStudio</i>	2 379 855	48
<i>Removing markers with ambiguous positions</i>	-7 239	
<i>X chromosome dataset before GQC</i>	55208	48
<i>Estimate gender for samples</i>		
<i>Exclude male samples</i>		
<i>Cluster X chromosome markers</i>		
<i>Include male samples</i>		
<i>Update X chromosome SNP statistics</i>		
<i>Removing markers call rate < 90%</i>	-190	
<i>Recalculating sample statistics</i>		
<i>Removing samples call rate < 95%</i>		
<i>Clustering SNPs and update statistics</i>		
<i>Removing markers call rate < 95%</i>	-377	
<i>Removing markers cluster sep. < 0.2</i>	-1	
<i>Removing symmetric A/T and G/C markers</i>	-1499	
<i>Removing markers with duplicated positions</i>	-2377	
<i>X chromosome dataset after GQC Phase 1)</i>	50764	48

CV2		CV3	
Number of Markers	Number of Samples	Number of Markers	Number of Samples
2 567 845	42	2 372 784	128
-2 180		-2 795	
55939	42	53253	128
-755		-366	
	-2		-1
-543		-597	
0		-5	
-1822		-1319	
-2798		-2508	
50021	40	48458	127

CV4		ST1	
Number of Markers	Number of Samples	Number of Markers	Number of Samples
2 382 209	80	2 380 364	111
-1 208		-1 192	
53425	80	53505	111
-358		-394	
	-1		-3
-963		-923	
0		-1	
-1333		-1336	
-2515		-2529	
48256	80	48322	110

Phase 2) Batch Merging

	Number of Markers	Number of Samples
<i>X chromosome markers with the same rsID</i>	44308	402
<i>Removing markers differing between samples</i>	-117	
<i>Removing duplicated samples</i>		-46
<i>Removing markers without position in dbSNP</i>	-29	
<i>Removing related individuals up to 2nd degree</i>		-28
<i>X chromosome dataset after GQC Phase 2)</i>	44162	328

Phase 3) Final Merging

	Number of Markers	Number of Samples
<i>Auton et al. 2015</i>	whole genome	2504
<i>Semo et al. 2020</i>	58176	188
<i>Semo et al. 2020 – after GQC Phase 2)</i>	48370	162
<i>Patin et al. 2017</i>	17814	1366
<i>Patin et al. 2017 – after GQC Phase 2)</i>	16309	1304
<i>X chromosome final merged dataset</i>	12976	4208

Supplementary Table S 3.2. Population datasets used in the study.

<i>Dataset</i>	<i>Continent</i>	<i>State</i>	<i>Region of origin</i>	<i>Population</i>
<i>This study</i>	AFR	CaboVerde	CVSTP	Santiago
<i>This study</i>	AFR	CaboVerde	CVSTP	SãoVicente
<i>This study</i>	AFR	CaboVerde	CVSTP	SantoAntao
<i>This study</i>	AFR	CaboVerde	CVSTP	Fogo
<i>This study</i>	AFR	CaboVerde	CVSTP	SãoNicolau
<i>This study</i>	AFR	CaboVerde	CVSTP	BoaVista
<i>This study</i>	AFR	CaboVerde	CVSTP	Brava
<i>This study</i>	AFR	CaboVerde	CVSTP	Maio
<i>This study</i>	AFR	CaboVerde	CVSTP	Sal
<i>This study</i>	AFR	CaboVerde	CVSTP	NA
<i>This study</i>	AFR	SãoToméPríncipe	CVSTP	SãoTomé
<i>This study</i>	AFR	SãoToméPríncipe	CVSTP	Príncipe
<i>This study</i>	EUR	Portugal	Europe	Portugal
<i>This study</i>	AMR	Brazil	SouthAmerica	Brazil
<i>This study</i>	AFR	Angola	WestSouthernAfrica	Angola
<i>This study</i>	AFR	Gabon	WestCentralAfrica	Gabon
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Sena
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Yao
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Makhuwa
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Bitonga
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Nyanja
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Manyika
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Ndau
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Changana
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Chopi
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Mwani
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Tswa
<i>Semo et al. 2020</i>	AFR	Mozambique	EastSouthernAfrica	Ronga
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Umbundu
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Nyaneka
<i>Semo et al. 2020</i>	AFR	Angola	WestSouthernAfrica	Ganguela
<i>Patin et al. 2017</i>	AFR	CAR	CentralAfrica	Biaka
<i>Patin et al. 2017</i>	AFR	DRC	CentralAfrica	Mbuti
<i>Patin et al. 2017</i>	AFR	IvoryCoast	CentralWesternAfrica	Ahizi
<i>Patin et al. 2017</i>	AFR	IvoryCoast	CentralWesternAfrica	Yacouba
<i>Patin et al. 2017</i>	AFR	Rwanda	EasternAfrica	Twa
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Bariba
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Yoruba
<i>Patin et al. 2017</i>	AFR	Benin	EastWesternAfrica	Fon
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Kimbundu
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Ovimbundu
<i>Patin et al. 2017</i>	AFR	Angola	WestSouthernAfrica	Kongo

<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Yaounde
<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Badwee
<i>Patin et al. 2017</i>	AFR	Cameroon	WestCentralAfrica	Bezan
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Baka
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Fang
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Nzebi
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Tsogo
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bapunu
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Shake
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bakota
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Galoa
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Obamba
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bateke
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Eshira
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Duma
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Makina
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Benga
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Akele
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Ndumu
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	BabongoE
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Eviya
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bakoya
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Orungu
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Okande
<i>Patin et al. 2017</i>	AFR	Gabon	WestCentralAfrica	Bekwil
<i>Auton et al. 2015</i>	AMR	ACB	Carribean	ACB
<i>Auton et al. 2015</i>	AMR	PUR	Carribean	PUR
<i>Auton et al. 2015</i>	AFR	MSL	CentralWesternAfrica	MSL
<i>Auton et al. 2015</i>	AFR	LWK	EastAfrica	LWK
<i>Auton et al. 2015</i>	EAS	CDX	EastAsia	CDX
<i>Auton et al. 2015</i>	EAS	CHB	EastAsia	CHB
<i>Auton et al. 2015</i>	EAS	CHS	EastAsia	CHS
<i>Auton et al. 2015</i>	EAS	JPT	EastAsia	JPT
<i>Auton et al. 2015</i>	EAS	KHV	EastAsia	KHV
<i>Auton et al. 2015</i>	AFR	ESN	EastWesternAfrica	ESN
<i>Auton et al. 2015</i>	AFR	YRI	EastWesternAfrica	YRI
<i>Auton et al. 2015</i>	EUR	FIN	Europe	FIN
<i>Auton et al. 2015</i>	EUR	GBR	Europe	GBR
<i>Auton et al. 2015</i>	EUR	IBS	Europe	IBS
<i>Auton et al. 2015</i>	EUR	TSI	Europe	TSI
<i>Auton et al. 2015</i>	AMR	CLM	SouthAmerica	CLM
<i>Auton et al. 2015</i>	AMR	PEL	SouthAmerica	PEL
<i>Auton et al. 2015</i>	SAS	BEB	SouthAsia	BEB
<i>Auton et al. 2015</i>	SAS	GIH	SouthAsia	GIH
<i>Auton et al. 2015</i>	SAS	ITU	SouthAsia	ITU

<i>Auton et al. 2015</i>	SAS	PJL	SouthAsia	PJL
<i>Auton et al. 2015</i>	SAS	STU	SouthAsia	STU
<i>Auton et al. 2015</i>	AMR	ASW	USA	ASW
<i>Auton et al. 2015</i>	EUR	CEU	USA	CEU
<i>Auton et al. 2015</i>	AMR	MXL	USA	MXL
<i>Auton et al. 2015</i>	AFR	GWD	WestWesternAfrica	GWD

Supplementary Table S 3.3. Population descriptors for 1000 Genome Project Phase 3.

<i>ID</i>	<i>Population name</i>	<i>Population description</i>
<i>ACB</i>	African Caribbean	African Caribbean in Barbados
<i>ASW</i>	African Ancestry SW	African Ancestry in Southwest US
<i>BEB</i>	Bengali	Bengali in Bangladesh
<i>CDX</i>	Dai Chinese	Chinese Dai in Xishuangbanna, China
<i>CEU</i>	CEPH	Utah residents with Northern and Western European ancestry
<i>CHB</i>	Han Chinese	Han Chinese in Beijing, China
<i>CHS</i>	Southern Han Chinese	Han Chinese South
<i>CLM</i>	Colombian	Colombian in Medellin, Colombia
<i>ESN</i>	Esan	Esan in Nigeria
<i>FIN</i>	Finnish	Finnish in Finland
<i>GBR</i>	British	British in England and Scotland
<i>GIH</i>	Gujarati	Gujarati Indians in Houston, TX
<i>GWD</i>	Gambian Mandinka	Gambian in Western Division, The Gambia - Mandinka
<i>GWF</i>	Gambian Fula	Gambian in Western Division, The Gambia - Fula
<i>GWJ</i>	Gambian Jola	Gambian in Western Division, The Gambia - Jola
<i>GWW</i>	Gambian Wolof	Gambian in Western Division, The Gambia - Wolof
<i>IBS</i>	Iberian	Iberian populations in Spain
<i>ITU</i>	Telugu	Indian Telugu in the UK
<i>JPT</i>	Japanese	Japanese in Tokyo, Japan
<i>KHV</i>	Kinh Vietnamese	Kinh in Ho Chi Minh City, Vietnam
<i>LWK</i>	Luhya	Luhya in Webuye, Kenya
<i>Masai</i>	Masai	Masai in Kenya
<i>MSL</i>	Mende	Mende in Sierra Leone
<i>MXL</i>	Mexican Ancestry	Mexican Ancestry in Los Angeles, California
<i>PEL</i>	Peruvian	Peruvian in Lima, Peru
<i>PJL</i>	Punjabi	Punjabi in Lahore, Pakistan
<i>PUR</i>	Puerto Rican	Puerto Rican in Puerto Rico
<i>STU</i>	Tamil	Sri Lankan Tamil in the UK
<i>TSI</i>	Toscani	Toscani in Italy
<i>YRI</i>	Yoruba	Yoruba in Ibadan, Nigeria