Evolutionary Genomics of the Trans-Atlantic Slave Trade

by

Mannda Ndou

15000753

Research Dissertation for Masters in Biological Sciences: Zoology

School of Mathematical and Natural Sciences

Department of Biological Sciences

University of Venda

Thohoyandou, Limpopo

South Africa

Student: Mr. Mannda Ndou

Co-supervisor: Dr. Andrinajoro Rakotoarivelo

Supervisor: Prof. Yoshan Moodley

Abstract

The Trans-Atlantic Slave Trade is a remarkable large-scale migration in the human history. On several occasions between the 16th and 19th century, millions of African men, women, and children were purchased from African traders and some abducted for slavery by the Europeans, for forced labour in the European colonies established in the American continent. The barbarous nature of the slavery left significant genetic modifications in the ancestry of modern-day descendants of former slaves (African Americans in United States and in Barbados). This research uses differently inherited high coverage Whole Genome Sequences (WGS) from autosomal, X, Y, and low coverage mitochondrial chromosomes collectively to present a detailed genetic point of view of the African Americans, their genetic relations to Africans and their interactions with America's other residents: Europeans and Native Americans. The results show that African slaves were abducted from West Africa (dominantly from Nigerian populations). Gene flow patterns were observed among former African slaves, their European slave masters, and Native American populations, resulting in genetic diversity among modern-day African Americans that is greater than any other population currently inhabiting the Americas and even higher than their source populations in Africa. The gene flow pattern was unidirectional from Europeans to African Americans and Native Americans, but bidirectional between the African Americans and Native Americans.

Declaration

I, Mannda Ndou of University of Venda student number 15000753, declare that this research dissertation is my original work and has not been submitted for any degree at any other university or institution. The dissertation does not contain any other persons' writing unless specifically acknowledged and referenced accordingly.

Signed (Student):  Date: 11-February-2022

Acknowledgements

First and foremost, I thank God, the Almighty, for His grace and favour of life and good health throughout from the beginning to completion of this research.

I cannot express enough gratitude to the University of Venda and the Department of Zoology (now part of the Department of Biological Science), the Moodley Evolutionary Genetics Lab FF029 for allowing me to advance a Masters' degree with them. I equally acknowledge the National Research Foundation of the Republic of South Africa (NRF) and the Deutscher Akademischer Austauschdienst/German Academic Exchange Service (DAAD) for funding this degree, I could have not afforded to pay the tuition and accommodation fees for my studies.

This degree could have not been accomplished without the support and supervision of my supervisor Prof. Yoshan Moodley, he was always there to guide every step of my academic journey throughout this degree and even extended his hand financially to support me for research purpose before my bursary funds were disbursed by the university; he inspired scientific greatness in me.

I extend my gratitude to my co-supervisor Dr. Andrinajoro Rakotoarivelo for assisting with compiling the bioinformatic software that I needed for my research analyses on the local server and helping with errors I experienced with bioinformatic commands and pipelines. Acknowledgement is given also to the CHPC (Centre for High Performance Computing) server for providing a platform under our project title: Mammalian Evolutionary Genomics, project code: CBBI0911, for some heavy analyses that could not be handled with our local lab server. I also thank my fellow lab mates Thabelo Rambuda, Marcia Ngoveni, and Kagisho Dibakoane for being an active group and offering combinative support and strategies towards solutions to some common lab problems we faced.

I am extremely grateful to my mother Rendani Joyce Sidahela who continued to give me her unconditional love although she was always worried to when I will finish studying and get a job. At least, her worry should last a little longer until I secure a doctoral degree. She supported me as much as she could emotionally, with prayers, and financially although she was unemployed. Indeed, a mother's love cannot be compared to any.

Conclusively I am also grateful to my young brother (Takalani Ndou), grandmother (Mukatshelwa Khorombi), uncles (Rudzani Sidahela, Londani Sidahela, Ntungufhadzeni Sidahela), my girlfriend (Rabelani Nenungwi) and the Fountain of Life Internal Well Church for having faith in me and keeping me in their prayers; it takes a community to raise a scientist in the making.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1. Human evolutionary history

1.1.1. Out of Africa hypotheses of human evolution

Archaeological and human genomic studies revealed that Africa is the origin of worldwide human diversity (Templeton, 2002; Scheinfeldt and Tishkoff, 2013). There are at least three Out of Africa migrations that were responsible for human evolution at large. The first ancient people migrated out of Africa about 1.9 million years ago as *Homo erectus*, followed by the second exodus of the *Homo ergaster* species about 1.7 million years ago (Weidenreich, 1946; Wolpoff et al., 2000; Aguirre and Carbonell, 2001; Derricourt, 2005). The third Out of Africa exodus was of the modern humans which occurred later than 100 thousand years ago (Smith et al., 1989; Stoneking and Soodyall, 1996). The Out of Africa migrations were a hot topic in the beginning of evolutionary science with great controversy concerning the origin, migrations, and spread of the *Homo* genus throughout the Old World (Relethford, 2001). The controversy was centred in the three major hypotheses: Multiregional, Out of Africa with replacement, and the Out of Africa with assimilation hypotheses. The Multiregional hypothesis proposed that a subset of *Homo ergaster* species in Africa migrated to Asia and Europe about 1.7 million years ago. The Asian *Homo ergaster* evolved into Denisovan whereas the European *Homo ergaster* evolved into Neanderthal; with evolutionary time the *Homo ergaster* that remained in Africa, the Denisovan and the Neanderthal in their respective continents evolved into the modern African, Asian, and European human populations (Weidenreich, 1946; Wolpoff et al., 2000).

In contrast to the Multiregional hypothesis, the Out of Africa with replacement hypothesis proposed that a subset of modern humans that evolved in Africa, migrated to Asia and Europe between 50 and 100 thousand years ago and eliminated the Denisovan ancestry that was native to Asia and the Neanderthal ancestry that was native to Europe (Stoneking and Soodyall, 1996). In support to the Out of Africa with replacement hypothesis, the mitochondrial Deoxyribonucleic Acid (mtDNA) of Neanderthals is very divergent from modern humans including the modern Europeans that should be descendants of Neanderthal (Ovchinnikov et al., 2000; Krings et al., 2000). This scenario is similar to the ancient Australian Lake Mungo3 (LM3) fossil specimen whose mtDNA was divergent from mtDNA of modern Australians; but was proposed to be mtDNA lineage extinction of this fossil in modern Australians (Adcock et al., 2001). Similarly, Relethford (2001) postulated the unanticipated divergence of Neanderthal mtDNA from modern Europeans as extinction of Neanderthal mtDNA lineage, possibly from selection sweeps and genetic drift due to its historical low effective population size. This implies

that whatever hominids existed in Eurasia, their mtDNA was completely replaced by modern human mtDNA.

The Out of Africa with assimilation hypothesis integrated the contradicting Multiregional and Replacement hypotheses into one scientifically reasonable guess with regards to the high human diversity in Africa compared to Europe and Asia. It proposed that the subset of the modern humans that migrated from Africa to Asia and Europe about 60 thousand years ago did not replace Denisovan nor Neanderthal, in fact they had gene flow with one another and evolved into the admixed modern human populations we see today (Smith et al., 1989). Majority of scholars supported the Out of Africa with assimilation hypothesis to be the best possible explanation of the modern human diversity. Clarke (2000) and Templeton (2002) supported this hypothesis over the Out of Africa with replacement hypothesis using model-based statistical methods on mtDNA, Y chromosome DNA, X chromosome DNA, and few autosomal DNA regions. They showed that 'Out of Africa' migrations to Eurasia were characterized by gene flow, not replacement. And proposed that if the Out of Africa with replacement hypothesis was true, the integration of the old phylogenetic trees and the recent genome-wide analyses would show a complete 100 % African ancestry in the Eurasian populations instead of the observed 90 % African and 10 % Eurasian ancestries, which happened through a minimum of three Out of Africa migrations (Out of Africa again and again with assimilation).

1.1.2. Modern humans in and out of Africa

The first modern human populations evolved in Africa at least 200 thousand years ago and a subset successfully migrated throughout the European and Asian continents between 80 and 50 thousand years ago (Cann et al. 1987; Scheinfeldt et al., 2010; Henn et al., 2012). Some studies (Cavalli-Sfoza et al., 1994; Semino et al., 2002; Lovell et al., 2005) propose a late migration back into Africa from Eurasia to Ethiopia, which gave rise to the modern East Africans. But before the Out of Africa migrations, early anatomically modern Africans already showed noticeable population differentiation (Mellars, 2006; Henn et al., 2011; Mallick et al., 2016). The Africans that remained in Africa after the Out of Africa migration continued separating into distinct groups and evolved into more than 2000 ethnolinguistic groups that are categorised into four main language families: Niger-Congo (Niger-Congo A and Niger-Congo B (Bantu)), Afroasiatic, Nilo-Saharan, and the Khoisan which is the most ancient existing human population, but the smallest African family both in language and population size (Heine and Nurse, 2000; Fan et al., 2019; Boyeldieu et al., 2008; Sands, 1998). The Niger-Congo family alone with more than 1430 distinctive languages, exceeds any of the world's other language families (Fan et al., 2019). African populations maintained large population sizes with several population divergences and

continuous migrations within the continent but forming many different ancestral lineages that distinguished Saharan and Eastern Africa from Southern, central, and Western Africa, and making Africa the most genetically diverse continent in the world (Tishkoff et al., 2009 and Beltrame et al., 2016).

After the Out of Africa migrations the different geographic areas of human populations in different continents had distinct and unique environmental factors including extreme conditions that could potentially confer different adaptive pressures from those on the African continent of origin. Thousands of years of human population persistence and adaptation to their unique continental environmental pressures shaped each continent independently into the rich human 'rainbow-colours' of diversity we observe today (Stoneking and Soodyall, 1996; Clarke, 2000). The contemporary populations were not only influenced to vary by environmental and temperature shifts, but also the different food sources that come with the environment. These changes also created new infectious diseases and would also have selected certain genotypes and phenotypes driven by the response of the human immune system (Jobling et al., 2019). The further division of the African gene pool into the European, Asian, and American continents resulted in different traditions and cultures, influencing distinct genetic patterns and structure among the continents (Gannon and Pillai, 2015; Mendes et al., 2020). These continental dependant genetic signatures can be investigated from continent to population level differentiation using whole genomes, evolutionary software, and bioinformatic techniques.

1.1.3. The peopling of the Americas

The first anatomical modern humans from Africa inhabited Asia between 80 and 50 thousand years ago (Cann et al., 1987; Scheinfeldt et al., 2010; Henn et al., 2012). Before 36 thousand years ago, the Asian ancestry was already spread into the Mal'ta and Han ancestries (Mendes et al., 2020; Figure 1). The Mal'ta and Han ancestries had secondary contact about 36 thousand years ago which resulted in the formation of the ancient East Asian population (Raghavan et al., 2014; Mendes et al., 2020; Figure 1). This ancient East Asian population later split from East Asia into the Bering land bridge (Beringia) between East Asia and Northwest America but continued to have gene flow with the Asian population it split from until approximately 25 thousand years ago (Moreno-Mayer et al., 2018; Llamas et al., 2016). Between 4.5 and 15 thousand years ago, the Beringian population completely separated from the Asian lineage and remained in Beringia because of geographic barriers bordering America, this period was termed the Beringian standstill (Szathmary, 1993; Bonatto and Salzano, 1997; Tamm et al., 2007). The Beringian population later successfully entered North America and became the first Native

Americans who later distributed throughout the Americas. But based on ancient DNA (about 11,5 thousand years old) from United States' Sunriver and the 20 thousand years and older archaeological sites in Northeast Asia (Buvit et al., 2016; Potter et al., 2018), Moreno-Mayer et al. (2018) proposed that the Beringian population isolated from the Native American ancestor prior to the Beringian standstill. Nevertheless, the Native American ancestor migrated south through North America, with establishment along the migration route all the way to South America between 16 and 1.5 thousand years ago (Mendes et al., 2020; Figure 1).

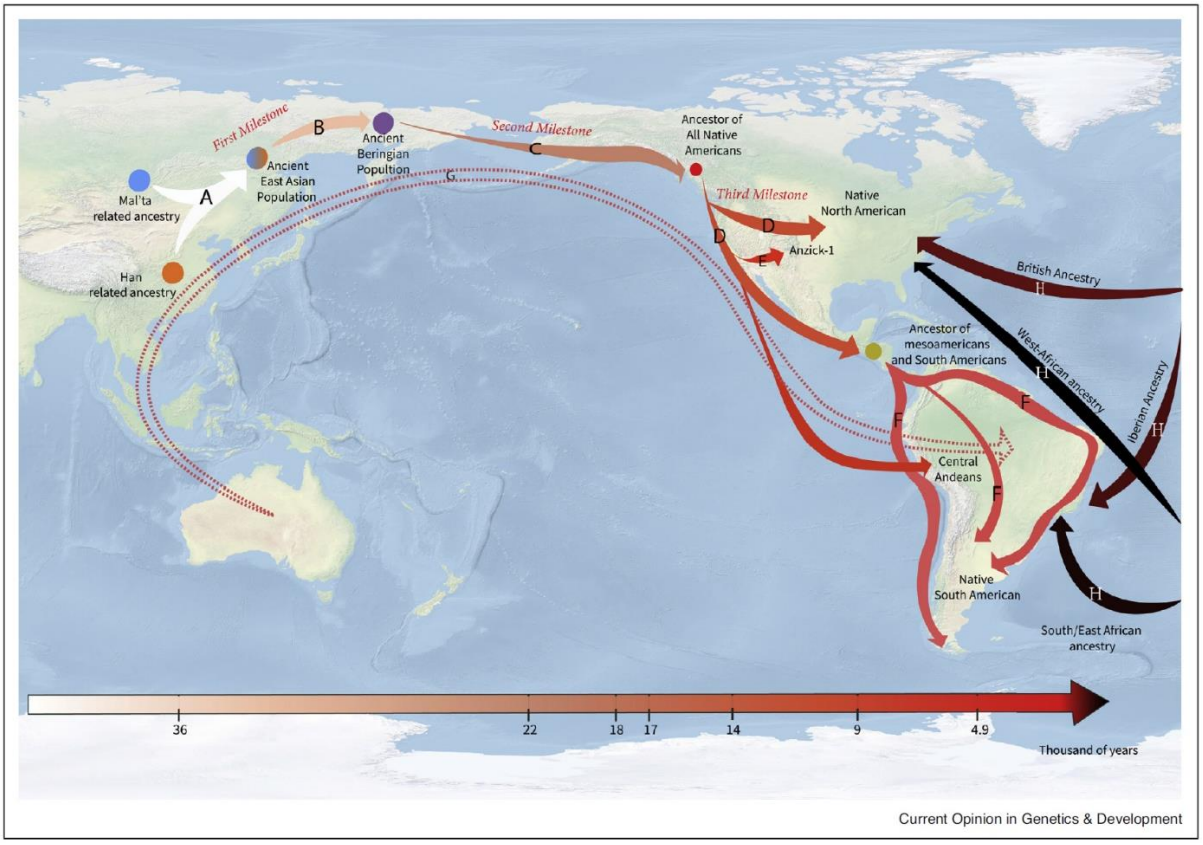

Figure 1. Time-scaled migration events of the Native American ancestry from Northeast Asia across the North and South America, and the subsequent peopling of America from Europe and Africa. Picture adopted from Mendes et al. (2020).

Europeans were the second immigrants into the American continent, long after the original Native Americans arrived from Asia. Historical and archaeological findings revealed that the German Norse wanderers (Germanic) were the first Europeans to inhabit North America as early as the 10th century (Price, 2015), but a better documented and popular European colonization of America started in 1492

by Spanish and Portuguese into the Caribbean islands of Puerto Rico and Cuba. About a century after the Spanish colonization, the British colonized North America, implementing the same colonial strategies and conduct (Haring, 1985). The European colonizers gained power over the Native American people and became a threat to their existence. They established colonies in the Native American homelands which became dependent on the Native American labour as slaves for functionality. By 1650, about 80 % of the Native American population size was lost due to combination of hard labour, harsh removal to indigenous homelands, and new diseases brought by the Europeans colonies (Reséndez, 2016; Ostler, 2019). The demand for production, economic development, and capital profit from the European American colonies increased to more than could be met by the Native American labour, and so they turned to Africa in search for more labourers. This quest prompted the main topic of this dissertation: The Trans-Atlantic Slave Trade.

1.2. Historical background of slavery and the Trans-Atlantic slave trade

The Trans-Atlantic Slave Trade is one of the recent Out of Africa mass migration in literature, and it defines the third major wave of human migrations into the Americas. The brutal nature of slave transportation, the American slavery period, and the substantial establishment of the modern African American society in the Americas stimulated the interest in investigating the Trans-Atlantic Slave Trade. Although African slavery started in the early 16th century, the first recorded African slaves disembarked in North America in 17th century (1619) at the British American colony of Jamestown (now in the United States, McCartney, 2003) from West and West Central African regions such as Nigeria, Sierra Leone, Gambia, Congo, Angola, and others (Bryc et al., 2010; Gates, 2014). Slave trade was not a new concept introduced by Europeans in Africa, but majority of the African slaves were already serving as slaves for the kings and other wealthy African leaders before the arrival of the Europeans in Africa (Lovejoy, 2011). The Atlantic slave trade was not a once-off event, but on several occasions between the 16th and 19th century Europeans voyaged from Europe and America to Africa abducting and purchasing huge numbers of African men and women from West and West Central African lords and traders to work the established European colonies in America (Gemery and Hogendorn, 1974; Fage, 1989).

The slave trade business was conducted through the so called Triangular Trans-Atlantic Slave Trade between Europe, Africa, and America (Oldfield, 2012). Cargo slave ships would leave Europe in a southerly direction to West Africa carrying large amount of goods such as glass, clothes, guns, liquor, and other European products. When the Europeans arrived in Africa, they kidnapped unsuspicious locals into waiting areas in their ships' slave barracks and then exchanged their goods for more

enslaved Africans from West African traders (Sylvester, 1998). From West Africa, the Europeans shipped the abducted and purchased African slaves to America through the Middle Passage, selling them to the proprietors of European colonies in South America, the Caribbean islands, and North America for labour in cash crop (sugar, rice, cotton, tobacco) plantations (Morgan, 2007). Not only where the slaves sold to European slave masters for money, they were also exchanged for goods produced in the plantations and after that the European traders sailed back to Europe, completing the triangular trade pathway.

The slave trade migrations were the largest distant human movements ever recorded in human history that lasted for over 400 years, with an average of six months transportation per migration from West and West Central Africa to European colonies in America (Clayborne, 2011). Some Africans were also enslaved from Southeast Africa and Madagascar (Figure 2). From the 17th century both the North American and Caribbean slave masters preferred a balanced male to female slave ratio. But this ratio changed to be dominated by males and children in the 19th century in North American colonies while remained balanced in Caribbean colonies (Eltis and Richardson, 1997; Nunn, 2008). This fluctuation in gender ratio in North America increased from century to century and was influenced by slave master's belief that men and children were more energetic than women for working in sugar plantations (Eltis and Engerman, 1993). In the series of the Trans-Atlantic slave trade, many enslaved Africans died during rebellions upon abduction, harsh environmental conditions, and disease outbreaks during the average six months transportation and upon arrival in the New World (Eltis, 1999). The slave transportation alone constituted 14% slave death rate (Clayborne, 2011), which increased with the adaptation and bad treatment in America. The overall slave trade summed to approximately 12.5 million Africans abducted and purchased from Africa, of which only about 10.7 million made it to America (Gates, 2014). Of the 10.7 million, between 500 and 650 thousand were sold to the British colonies of what was to become the United States, and the remaining millions sent to other European colonies throughout Central America, South America, and Caribbean islands (Thomas, 1999; Zakharia et al., 2009; Jin et al., 2012). The African slaves were the foundational labor force for the operation of European colonies in America, building of cities from badlands, mining, and the development of world economy at large (McMillan, 2002).

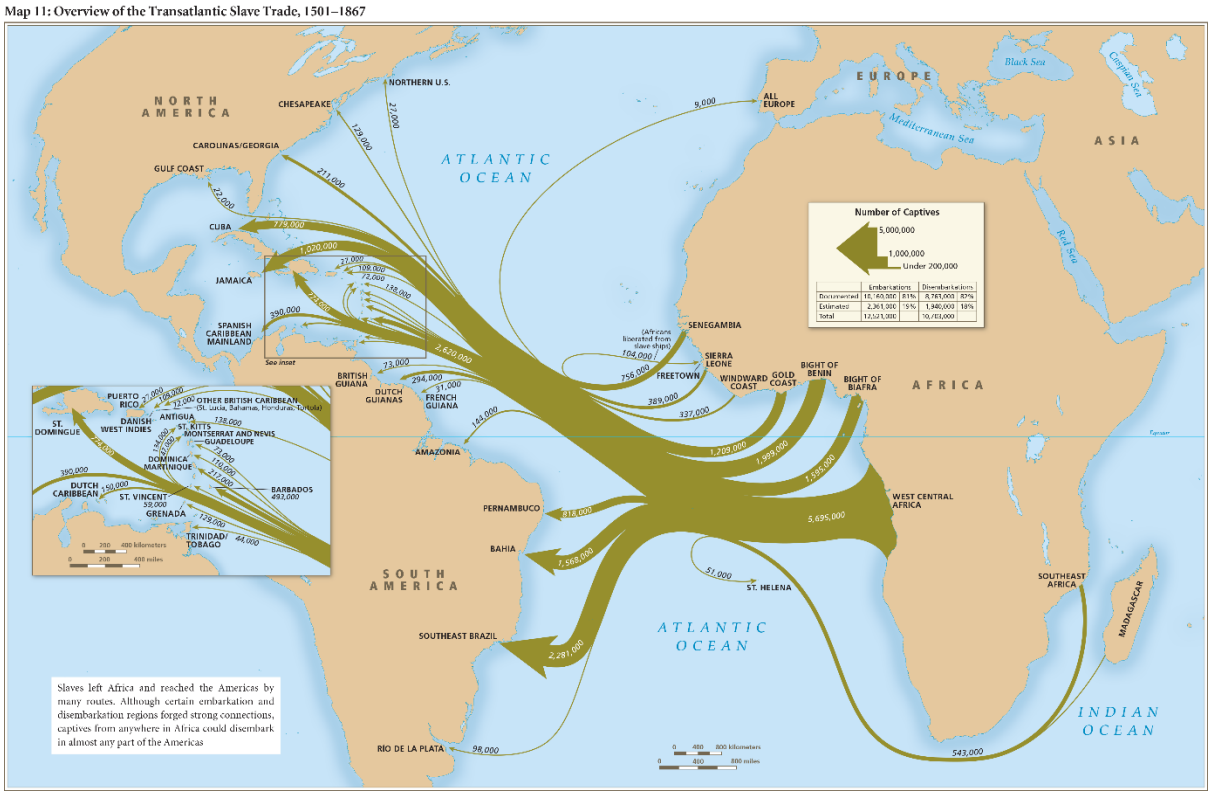

Figure 2. Trans-Atlantic Slave Trade routes from Africa to the Americas from 1501 to 1867. Picture adopted from Li (2020).

Archaeological research in the Americas has shed light upon the slave habitation and the development of the African American culture. However, the vast historic knowledge that linked the emergence of the slave populations in the Americas to their African origins is constrained and limited by the paucity of excavated archaeological materials and sites both in West Africa and in the New World (Simmonds, 1973; DeCorse, 1991). Moreover, non-biological archaeological materials are difficult to impossible to use as reference for identification of descendent slave populations in the African communities (De Corse, 1989). Tishkoff et al. (2009) unravelled this enigmatic concept using numerous nuclear microsatellites and INDEL (insertion/deletion) markers, which reflected the ancestry of the African Americans mainly to Niger-Kordofanian populations in West Africa. Nevertheless, the use of microsatellite genetic data did not provide a complete historic picture, as will be explained in the next section.

1.3. The rise of genomics over microsatellite markers

Many ideas and speculations about the nature of the Trans-Atlantic slave trade have been published in social sciences research platforms and historical books. However, “Nothing in biology makes sense except in the light of evolution” (Dobzhansky, 2013). The great Greek philosophers Anaximander (610-

546 BC) and Aristotle (384-322 BC) proposed a scientific thinking based on natural and testable explanations for the natural world and events that occur in it (Blundell, 2016).

Previous use of microsatellite genetic markers for inferring evolution of human populations was a good start in the early research years (Cavalli-Sforza et al., 1988; Templeton et al., 1992; Wallace et al., 1999). The sense and direction of human evolution was obtained, but many details were still lacking due to the paucity of genetic information in such small genetic markers. Microsatellite data may offer limited human genetic information, but never the complete correlation of all and different genes that may be offered by whole genome sequences. There are no other genetic markers, even highly polymorphic microsatellites, that offers the level of resolution offered by whole genome sequences (Pérez-Reche et al., 2020). The detailed genetic representation of African and African American history, thus far, was inferred by Tishkoff et al. (2009) from 1327 nuclear microsatellite loci and Insertion or Deletion mutations (INDELS) markers. Although this was a large microsatellite dataset, it pales in comparison to over 3 million Single Nucleotide Polymorphisms (SNPs) (Bentley et al., 2008; Wheeler et al., 2008), 111 million Single Nucleotide Variants (SNVs), and 14 million INDELS (Byrska-Bishop et al., 2021) obtainable from human whole genome sequences. The emergence of whole genome sequences, together with new bioinformatics software and statistical tests have brought about a new era in unravelling patterns of human evolutionary history and genetic diversity (Vitti et al., 2012). The worldwide genomic projects such as the 1000 Genomes (1000 Genomes Project Consortium, 2015), International HapMap (Gibbs et al., 2003), and the African Genome Variation Projects (AGVP) (Gurdasani et al., 2015) improve our understanding of how humans respond to different diseases relative to their geographic regions. However, these databases also provide an excellent genomic data for investigating human evolutionary biology in relation to their past and present geographic regions using high resolution whole genome data.

1.4. Hypotheses

This study investigates the genomic consequences of the Trans-Atlantic slave trade, with refence to the African slaves, based on three major hypotheses:

Hypothesis 1: The genetic diversity of Africa is basal to all modern humans; thus, a portion of African genes were already shared with European, Asian, and American continents since these populations derive from Africa through the Out of Africa migrations that happened long before the Trans-Atlantic slave trade (Cann et al., 1987; Scheinfeldt et al., 2010; Henn et al., 2012). However, non-African

populations generally have a lower genetic diversity than African populations due to subsampling and genetic drift (Prugnolle et al., 2005; Liu et al., 2006). Thus, the Trans-Atlantic Slave Trade (Gemery and Hogendorn, 1974; Fage, 1989), could be yet another subsampling of African genes and given that a large proportion of the African slaves died during the passage and due to harsh living conditions in the Americas, it is possible that genetic drift has also resulted in lower genetic diversity in the African slaves compared to their counterparts in West and West Central Africa.

Hypothesis 2: Based on the literature that African slaves were purchased and abducted from West and West Central African regions (Bryc et al., 2010; Gates, 2014), the genealogy of their modern-day descendant (African American) populations should evidently reflect their geographic origin in those African regions; leading to the second hypothesis that the genetic structure of present-day African Americans will reflect their origin(s) in Africa.

Hypothesis 3: African slaves worked in close contact with European slave masters on plantations and in residences (Berlin, 1980). Not only were slaves in close contact with the slave masters, but also with the enslaved local Native American populations (Fisher, 2017). It is thus conceivable that gene flow could have occurred between African slaves and both their European slave master and Native American counterparts.

This project used whole genome multi-sample SNPs from the 1000 Genomes Project from four West Africa, one East Africa, two African American, four Native American, one European American, and two European populations to help test the above three hypotheses, using bioinformatics and evolutionary genomic analyses.

2. Method

2.1. Sample datasets background

The International Genome Sample Resource (IGSR) website (https://www.internationalgenome.org) has been a reliable repository for world-wide diversity of human genomic data, freely available for use by researchers throughout the world. The website contains variants, alignments, and sequence genetic and genomic data types available in different formats (Bam, VCF, Cram, and Fastq), and

generated through different technologies for integrated variant call sets, exome, low coverage whole genome sequences, and PCR-free high coverage genomes. The release of more than 85 million variants in completion of the 1000 Genomes Project final phase (phase 3) on human reference GRCh37 (1000 Genomes Project Consortium, 2015), has prompted the release of the sequences, alignments, and variants based on the most recent and upgraded GRCh38 human assembly (Lowy-Gallego et al., 2019). The New York Genome Center (NYGC), funded by the National Human Genome Research Institute (NHGRI), sequenced a total number of 3202 samples (2504 unrelated and 698 related individuals) from the phase 3 dataset to high (30X) coverage. These were mapped to the new GRCh38 human reference assembly, resulting in high coverage recalibrated multi-sample genotypes, structural, and phased VCF files.

2.2. Downloading the sample datasets

Recalibrated 30X high coverage phased multi-sample chromosomal level VCF files were downloaded from the IGSR website. High coverage data were chosen for better reliability for analytical stages, including the accurate scoring of heterozygote SNPS, as opposed to low coverage data. Chromosome 1 to chromosome 22 VCF files were concatenated into an autosomal chromosomes VCF file using the Bcftools command: bcftools concat -O z –threads 30 chr1.vcf.gz …chr22.vcf.gz -o Autosomal_chromosomes.vcf.gz, through the Ubuntu terminal platform. Accounting for the varying modes of inheritance, which confer different levels of genetic drift (effective population size) to different genomic regions, Chromosome X and Y VCF files were left intact for comparative analyses with the autosomal and mitochondrial chromosomes datasets. There was no high coverage mitochondrial chromosome dataset in the website during download, hence the available low coverage phase 3 mitochondrial genotype dataset was downloaded (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/ALL.chrMT.phase3_callmom-v0_4.20130502.genotypes.vcf.gz) for parallel analyses with the Y chromosome dataset. This resulted in four VCF datasets (autosomal chromosomes, X chromosome, Y chromosome, and mitochondrial chromosome) for all sampled populations to analyse in parallel.

2.3. Sub-setting desired populations and markers based on modes of inheritance from downloaded datasets

A total of 14 populations and 728 high coverage human genomes were used in this study. These populations were all that was available in the high coverage multi-samples VCF files, representing the

populations that were involved during the African Slave Trade. These include: two slave descendant populations with African ancestry (African Americans) in the United States and Caribbean which were chosen to represent the African slaves that made it to the European colonies in the American continent during the slave trade; four West African populations (two from Nigeria, one from Gambia, and one from Sierra Leone) representing potential populations from where the slaves were abducted in West Africa; one East African population (Kenya) served as a control for the Trans-Atlantic slave trade as no slaves were speculated to be abducted from it; four American populations (Mexico, Peru, Colombia, and Puerto Rico) representing the Native American populations and speculated to have interacted with the slave and the slave master populations during slavery in America, one American population with European ancestry (United States) representing the Europeans that colonized America (slave masters), and two European populations (Britain and Spain) representing the slave masters place of origin before they colonized America. See Table 1 for a full breakdown of samples.

Table 1. Fourteen populations used to study the genomics of the Trans-Atlantic Slave Trade with sample sizes

Population code	Population full name	Representation based on the time of slavery	Continent	Number of high coverage genomes per population
1. ASW	African Ancestry in South West United States	West Africans that were abducted to European colonies in America as slaves	America	52
2. ACB	African Caribbean in Barbados			52
3. ESN	Esan in Nigeria	Potential West African populations of slave origin	Africa	52
4. YRI	Yoruba in Nigeria			52
5. GWD	Mandinka in Gambia			52
6. MSL	Mende in Sierra Leone			52
7. LWK	Luhya in Kenya	East African population where no slaves were thought to have been abducted		52

Population code	Population full name	Representation based on the time of slavery	Continent	Number of high coverage genomes per population
8. MXL	Mexican Ancestry in Los Angeles, California	Native Americans that possibly interacted with both the African slaves and the European slave masters	America	52
9. PEL	Peruvian in Lima			52
10. PUR	Puerto Rican in Puerto Rico			52
11. CLM	Colombian in Medellin			52
12. CEU	Utah residents with Northern and Western European ancestry	European slave masters in North America		52
13. IBS	Iberian populations in Spain	Origin of the Europeans that introduced slave colonies in America	Europe	52
14. GBR	British in England and Scotland			52

The sample list for these populations was subsetted from the overall unrelated samples of the 1000 genomes project sample list in an Excel spreadsheet. Unrelated samples were chosen for capturing as much variation and relationships within and among variable populations, and to infer the phylogeny of individuals and populations, not biased by family-based relatedness, but accounting for the wide scale variability of human populations. Related samples may be good for inferring most recent family ancestries but may offer biased parameters estimations for world-wide scale population studies (Wang, 2018).

Genetic markers constituting the human whole genome sequences (WGS) are inherited from one generation to another through different modes of inheritance. These genetic markers can be divided into four kinds of genetic datasets: autosomal chromosomes, X chromosome, Y chromosome, and the mitochondrial chromosome datasets. Autosomal chromosomes consist of the largest nucleotide base number and are inherited by offspring from both male and female parents. The X and Y chromosome sex-determining datasets are gender-based inherited from parents to offspring. The Y chromosome is

paternally inherited and can only be passed on to male offspring, whereas the X chromosomes can be inherited both paternally and maternally, but the greatest chance is of maternal inheritance as females possess two copies or a pair of X chromosomes opposed to one copy in the males. The last dataset, mitochondrial chromosome, is maternally transferred from female parent but inherited by both male and female offsprings. These varying modes of inheritance effect variable effective populations sizes of these markers (Jorde et al., 2000; Ellegren, 2009), and thus the level of genetic drift, qualifying them to be investigated separately.

Migration rates and gene flow patterns or direction in natural populations usually vary from one gender to another and can be reflected in patterns of diversity and structure of the four different genetic datasets above. Both genders have equally important genetic stories to portray. The male to female ratio of all studied populations was balanced based on the population with the least individuals of either gender, to infer fairly and evenly the phylogeny created by these genders; thus, getting rid of gender-based biased phylogeny and gene flow patterns. E.g., ASW population had the least 26 males, therefore all the populations were down sampled to 26 males and 26 females per population, explaining the 52 genome population sizes shown in Table 1 above. Notepad was used to create a text file (samples list) for the samples with balanced sex ratios, which was used to subset the four downloaded VCF datasets into autosomal chromosomes, X chromosome, Y chromosome, and mitochondrial chromosome datasets using Bcftools (Li et al., 2009, command: bcftools view -O z -S Samples_list.txt Dataset.vcf.gz -o Subsampled_dataset.vcf.gz). The Y chromosome dataset had half the total number of samples per population as it only represents the male gender lineage (only passed on and inherited by males).

2.4. General filtering of datasets

The three (autosomal chromosomes, X chromosomes, and Y chromosome) high coverage datasets and mitochondrial low coverage dataset were generally filtered with three Bcftools commands, and two Plink (Purcell et al., 2007) commands. The first filtering was for minor allele frequency (MAF). The MAF value defines the frequency of the less shared alleles among the individuals in the populations that are being compared. This is how many times the less common alleles were observed in the populations being studied. The frequency of this allele is computed across the corresponding allele position in the sequence of each sample in the multi-sample VCF file. This computation is an important factor for determining the genetic variability conferred by specific loci, and the overall diversity represented by the individual sample. A small MAF value is preferable for enclosing as much zoomed-

in deep variation among closely related individuals or populations belonging to the same species (Linck and Battey, 2019). Filtering using a large MAF value generalises and masks deep variation among individuals or populations, hence it is preferable for studying individuals or populations that are distantly related (usually belonging to different species). In the present case, I was interested in evolution within a species, so only loci with a MAF greater than 2% were kept in the datasets using the Bcftools command: bcftools view --threads 20 -O z -i 'MAF>0.02' Subsampled_dataset.vcf.gz -o Subsampled_dataset1.vcf.gz.

Further filtering was carried out to bring the datasets down to a manageable size that traditional analytical software can handle and reasonable for the completion of this project. Multi-sample whole genome sequences with millions of SNPs would either run unsuccessfully or take a long period of time (months) to run to completion depending on the computational power demand of the bioinformatic tool being used. INDELS (Insertion or deletion of nucleotide bases in a DNA sequence resulting from mutations) are important for understanding mechanisms behind many human diseases in clinical research but are not particularly informative about the evolutionary history. The second filtering was to retain only SNPs from the samples VCF files, discarding INDELS, using the command: bcftools view --types snps Subsampled_dataset1.vcf.gz -O z -o Subsampled_dataset2.vcf.gz --threads 20.

Multi-sample VCF files contain a certain degree of missing SNP data. The amount of missing data is highly dependent on the SNP coverage of the samples within the multi-sample file. Low coverage datasets contain a lot of missing SNP data compared to high coverage datasets. When there is a huge amount of missing data randomly occurring within the multi-sample sequence file, samples SNP data cannot be parallelly compared, hence they might portray a misleading topology or genetic structure. The human genome generally consists of biallelic sites at SNPs, but it is common to observe multiallelic sites throughout the genome. Most evolutionary software requires only biallelic SNPs to infer the genetic history and this also provides another opportunity to reduce the size of the multi-sample data set. The third filtering was to remove certain degree of missing data and multi-allelic sites from the data set. This was done using Bcftools command: bcftools view --threads 30 -e 'AC==0 || F_MISSING > 0.2' -m2 -M2 -O z -o Subsampled_dataset3.vcf.gz Subsampled_dataset2.vcf.gz. Only the loci of the individual genomes containing biallelic and SNPs with less than 20 percent missing data were retained by this command. The missing data filtering did not result in a large amount of data loss as the three (autosomal, X chromosome, and Y chromosome) datasets were of high coverage, the considerable number of SNPs lost was observed in the low coverage mitochondrial dataset.

The last general filtering was for linkage disequilibrium (LD) SNPs. Of the several linked alleles (SNPs), only the most informative ones were used. The command that was used for LD pruning of SNPs in Plink was: plink --vcf Subsampled_dataset3.vcf.gz --allow-extra-chr --indep-pairwise 50 10 0.8 --out Subsampled_dataset3; This command directs Plink to filter the dataset in independent-pairwise 50 base pairs windows cycle, with 10 base pairs step size between the windows, and a filtering threshold of 0.8. The filtering threshold value determines how much linked SNPs are pruned out of the dataset, the smaller the threshold value the more SNPs are pruned out (Purcell et al., 2007). The outputs from this command were Subsampled_dataset3.prune.in and Subsampled_dataset3.out. The '.prune.in' file contains information of the SNPs that meet the filtering threshold, whereas the '.prune.out' file entitles the information of SNPs that do not meet the filtering requirements. Secondly the original VCF file was recoded based on the '.prune.in' file that was created from the first Plink command using the command: plink --allow-extra-chr --extract Subsampled_dataset3.prune.in --make-bed --out final_ Subsampled_dataset --recode vcf-iid --vcf Subsampled_dataset3.vcf.gz. The number of SNPs remaining in the datasets was viewed using the command: bcftools stats -s – File.vcf.gz , and the remaining SNPs were recorded in Table 2 (see Section 3.1).

2.5. Preparing the Outgroup

An outgroup genome was necessary to root the phylogeny tree so that the correct branching pattern within the ingroups could be established relative to a reproductively isolated, but closely related species. An outgroup genome was also necessary to conduct ABBABABA tests for introgression (see Section 2.8). Two paired-end whole genome sequences (WGS) short reads of a male captive born chimpanzee (*Pan troglodytes*), from New Iberia (Louisiana, United States of America) were downloaded from ENA (European Nucleotide Archive) website (https://www.ebi.ac.uk/ena/browser/view/SAMN15033246). Their run accession number was SRR11892906. They were generated in ILLUMINA instrumental platform, and Illumina NovaSeq 6000 instrumental model. The male chimpanzee was chosen over the female chimpanzee to have an outgroup that can be subsetted and merged into the four human datasets, as the male genome consist of both the Y and the X sex chromosomes opposed to female genome that only consist of a pair of X sex chromosomes. The same human reference assembly (GRCh38) that was used by the 1000 genomes project to map the four downloaded datasets was downloaded from the IGSR website. The chimpanzee short reads were mapped to this assembly using BWA MEM (Li and Durbin, 2010), with 1000 genomes project command options: bwa mem -t 30 -B 4 -O 6 -E 1 -M $ GRCh38_full_analysis_set_plus_decoy_hla.fa $fastq_file(1) $fastq_file(2) | k8 bwa-postalt.js

GRCh38_full_analysis_set_plus_decoy_hla.fa.alt | samtools view -1 - > $Outgroup.bam. The same reference assembly and the mapping command options were replicated as in the 1000 genomes project pipeline to ensure the compatibility of the chimpanzee dataset with the 1000 genomes project human dataset. The bwa-postalt.js file specified in the K8 option was downloaded from GitHub (https://gitlab.citius.usc.es/github/bigbwa/-/blob/23996308dc30008b40993c2753f1f78129a4e71d/bwa-0.7.12/bwakit/bwa-postalt.js). After mapping successfully, the Samtools (Li et al., 2009) command: 'samtools flagstat' was used to view how much of the chimp's short reads was successfully mapped to the reference genome, and it was found that 99.08 % had mapped properly and 96.27 % was properly paired. The resultant bam file was sorted by coordinates using the command: samtools sort -@ 30 -O BAM --reference GRCh38_full_analysis_set_plus_decoy_hla.fa Outgroup.bam -o sorted_Ourtgroup.bam. The sorting by coordinates command arranges SNPs based on position, and this step is necessary and a requirement for calling SNPs from the bam file.

SNPs were called from the resultant sorted bam into a VCF file using the piped 1000 genomes project mpileup command: bcftools mpileup – threads 30 -E -a DP -a SP -a AD -P ILLUMINA -pm3 -F0.2 -C50 -d 700000 -f GRCh38_full_analysis_set_plus_decoy_hla.fa Sorted_Outgroup.bam | bcftools call – threads 30 -mv -O z GRCh38 -o Outgroup.vcf.gz. The autosomal chromosomes 1 to 22 were extracted from the resultant chimpanzee (outgroup) SNPs VCF file into a new autosomal VCF file, using the command: bcftools view Outgroup.vcf.gz -O z --regions chr1,chr2,chr3,….chr22 -o Autosomal_Chimp.vcf.gz. The same was done for the outgroup's X chromosome, Y chromosome, and mitochondrial chromosome, resulting in 4 outgroup VCF files that match the 4 human datatsets. The autosomal, X chromosome, Y chromosome, and mitochondrial chromosome outgroup VCF file were then merged to the respective human datasets using Bcftools command: bcftools merge -O v 1000_genomes_autosomal_individuals.vcf.gz Autosomal_Chimp.vcf.gz -o my_samples_and_outgroup.vcf. The merged datasets were again filtered with the general filtering commands used previously in the four human datasets (see Section 2.4). The outgroup datasets were merged with the human datasets for phylogeny reconstruction (see Section 2.7.3 and 2.7.4) and gene flow analyses (see Section 2.8).

2.6. Genomic diversity

Genomic diversity is the basis for individual, populations, to species variability throughout natural landscapes. It was important to be examined in this project to compare the populations variability among the African, African American, Native American, and European populations, and help answer hypothesis 1 which proposed that African populations will have higher genomic diversity than the American and European populations. Heterozygosity and nucleotide diversity are the two most common and effective parameters to measure genetic diversity in populations. The two measures were analysed in parallel to compare how they portray the genomic diversity from one population to another and were performed in the autosomal chromosomes' dataset that was already filtered for MAF, multiallelic SNPs, and missing data (see filtering Section 2.4). This dataset was chosen because of its huge SNPs size and that it is a diploid genetic marker, thus heterozygosity and nucleotide diversity can be computed on the alternating base pairs.

Heterozygosity is measured across the genome for each diploid genome individually. For simplicity with both the nucleotide diversity and heterozygosity analyses, the above mentioned filtered autosomal VCF dataset was subsetted into 14 population level VCF files using the command: bcftools view -O z -S ASW_sample_list.txt All_populations_file.vcf.gz > ASW.vcf.gz. The heterozygosity of each population VCF file was analysed using the VCFtools (Danecek et al., 2011) command: vcftools --gzvcf ASW.vcf.gz --het --out ASW. The individual heterozygosity was then viewed and computed for average heterozygosity of the population in Excel and then plotted in a graph.

Nucleotide diversity is the population statistics calculated per genomes loci across the entire VCF dataset. The results from this analysis are given on bases of loci pairwise differences without the information of individuals, therefore the already subsetted population level VCF files were handy for average population nucleotide diversity. The nucleotide diversity for each population was inferred per window fashion using the VCFtools command: vcftools --gzvcf ASW.vcf.gz --window-pi 100000 --out ASW. The nucleotide diversity for each population was viewed and averaged based on the windows into the nucleotide diversity of the population using Excel and compared among the populations in a bar graph plot.

2.7. Genomic Structure analyses

Genetic structure analyses were carried out to identify whether the different populations will fall into their correct evolutionary positions in the phylogeny of the African, European, and the American

continents. This was aimed on answering hypothesis 2, whether the African Americans will be closely positioned to the African populations of their origin, and not with either Native American or European populations.

2.7.1. Principal Component Analyses (PCA) on autosomal and X chromosome datasets

PCA method introduced by Jolliffe and Lovric (2011) does not infer population structure based on any genetic model. Hence, the PCA uses less computational power and time compared to model-based methods (e.g. Admixture and Iqtree) for reconstructing population structure. Although it would have been interesting to perform and compare the PCA from all four datasets, the Plink software used for PCA could only carry out the analyses on diploid genetic markers, not the haploid Y and mitochondrial chromosomes datasets. Hence, the PCA were performed on the diploid autosomal and the X chromosome datasets that were generally filtered (see section 2.4) and without an outgroup genome. Plink requires a Binary Variant Call Format (BCF) file to be converted to Plink format for the PCA analyses. Hence, the filtered VCF files were converted to BCF file using the command: bcftools view -O b input.prune.in.vcf.gz -o output_samples.prune.in.bcf.gz. The output files were indexed using the command: bcftools index output_samples.prune.in.bcf.gz. Each BCF file was converted to Plink format using the command: plink --noweb --bcf Autosomal_samples.prune.in.bcf.gz --keep-allele-order --vcf-idspace-to _ --const-fid --allow-extra-chr 0 --split-x b37 no-fail --make-bed --out Autosomal_samples.prune.in. The PCA analyses were performed on the resultant output files using the command: plink --bfile Autosomal_samples.prune.in –pca.

Two Plink output files (plink.eigenval and plink.eigenvec) were produced from the PCA run. The plink.eigenval file consist of the values of the magnitude of variability among individuals, and the plink.eignevec file consist of the X-Y plane positions of the individuals when plotting the PCA. A '.ped' file consisting of the populations and samples information is required when plotting the PCA results. The ped file was constructed in Notepad, with the sample names on the first column and their population codes on the corresponding second column. Header of the first column was 'Individual.ID' and the header of the second column was 'Populations'. The PCAs were plotted in R4.0.4 (R core team, 2013) using a custom script, given in Appendix A.

2.7.2. Admixture analyses

2.7.2.1. Admixture plots

The Admixture software (Alexander et al., 2009) was preferred over the NGSAdmix software (Skotte et al, 2013) because of the high coverage and diploid state of both the autosomal and X chromosome datasets. NGSAdmix software would have been preferred if the datasets were of low coverage for inferring admixture based on genotype likelihoods, but with high coverage SNPs datasets, it is more reliable to use the allele calls to infer admixture among populations. The Admixture software analyses the two alleles in an individual that are inherited from both the male and female parents and examines similarities and differences in the distribution patterns of genetic variants across the genome assuming Hardy-Weinberg equilibrium. Populations whose individuals share similar genetic variation patterns are grouped or mixed, and vice versa. The admixture analyses were performed on diploid datasets that recombine, that is on the autosomal and the X chromosome datasets, in Linux terminal platform. The recombination process exchanges the genetic information within an individual between the chromosome inherited from both parents, thus admixing the variation. Free recombination is an assumption of the Admixture model. The datasets that were used for these analyses were the VCF files after general filtering, without the outgroup. An outgroup is not required in the Admixture analyses as the outgroup is reproductively isolated from the ingroups and thus without the possibility of gene flow.

For Admixture plots presentations, the populations were grouped based on their geographic continents, the first populations in the VCF file were the African populations, followed by the slave descendant American and Caribbean populations, followed by the Native American populations and lastly the populations with European ancestry (American European, Spain and Britain). The Admixture software requires a VCF input file that is pruned for LD, emphasizing the use of a generally filtered autosomal VCF file. The next step was to run the actual admixture with a loop for several K values. But the Admixture looping requires the .map file of the input bed file to be present, which was created through a Plink command: plink –vcf Autosomal_chromosomes.prune.in.vcf –recode –out Autosomal_chromosomes.prune.

Admixture was forced to model population structure from K = 2 to K = 12, for the 14 populations used in this study. The maximum K was less than the number of populations because most populations were likely closely related and descendants of one another. The best K for the data set was chosen as that with the least cross validation (CV) error, representing the best possible population clusters in the dataset. The Admixture loop command used was: for i in {1..12}; do /media/moodley/seagate/admixture_linux-1.3.0/dist/admixture_linux-1.3.0/admixture --cv

Autosomal_chromosomes.prune.in.bed $i -B5 -j30 > log${i}.out; done. The CV value was calculated by default of 5 folds, with 5 replicates and using 30 threads for parallel processing.

After the Admixture analyses finished running the CV errors of all the K values were printed using the command: grep "CV" *out | awk '{print $3,$4}' | sed -e 's/(//;s/)//;s/://;s/K=//' > Autosomal_chromosomes.prune.in.cv.error and were plotted in Excel line graph. The K value with the smallest CV error was used as the best representation of the true population structure, and the maximum K value to be interpreted. The admixture plots were made in R (terminal option) through an R script that was downloaded from GitHub (https://github.com/speciationgenomics/scripts/raw/master/plotADMIXTURE.r), using the command: Rscript plotADMIXTURE.r -p Autosomal_chromosome.prune.in -i Samples.list -k 14 -l ESN,YRI,LWK,MSL,GWD,ACB,ASW,MXL,PEL,PUR,CLM,CEU,IBS,GBR. The sample list text file used as one of the inputs in the Admixture plots command was created in text editor from the original autosomal VCF file that was used to create the input bed file. The text file list contained the sample names on the first column and the population codes on the corresponding second column in the order of the VCF file samples. Sample names that were appearing on the Admixture plots were cropped out for better graphical representation of the plots.

2.7.2.2. Admixture pie charts plot

The Admixture analyses produces two output files, the P and Q files. The Q file is of interest for pie plots production as it contains the K proportion of each sampled populations involved in the analyses. The Q file of the autosomal chromosomes' dataset was opened in Excel, and the K value with the least CV error was identified. Pie charts were created using the CV error values descending from the K value with the least CV error. The autosomal chromosomes dataset was chosen for these charts over the X chromosome because it represents the population evolutionary history of both sexes using the most SNP data. The K proportion were averaged to population level and pie charts were drawn in Excel and modified to have same colours as representing the K values in admixture plots. A high-resolution blank world map was downloaded from Google. The downloaded picture was opened with Microsoft PowerPoint, and cropped to only include African, American, and European geographic regions represented by this study. The pie charts for each population were pasted onto the map relevantly to the geographic location of the populations. A PowerPoint slide with this map and pies was exported and saved as an SVG image file.

2.7.3. Phylogenetic analyses from Autosomal and X chromosome datasets

The IQ-TREE software (Nguyen et al., 2015) was used for inferring the phylogeny of the individuals within the autosomal and X chromosome datasets. Topological integrity of software that reconstruct phylogeny of different individuals and populations is sensitive to the amount of genetic data available. Incomplete evolutionary data and paucity of SNPs data in multi-samples analysed concurrently may result in a deceptive topology (Wolf et al., 2002; Vishnoi et al., 2010). The X and autosomal chromosomes genetic markers comprise of thousands and millions of SNPs data sufficient to infer a genetic topology consistent with the true structure of populations. Hence, the two SNPs rich datasets were used to serve as basis of populations structure. The autosomal and X chromosome input VCF datasets that were used for the phylogenetic tree are the ones that were merged and generally filtered as described in Section 2.4 with the outgroup genome. An outgroup genome was required to root the trees so than ingroups can be arranged in their correct evolution phylogeny.

The sample names of the two pruned VCF files of the two datasets were viewed using the command: bcftools query -l file.vcf.gz and random subset of 10 samples (5 males and 5 females) per population list was created using Notepad. Iqtree software requires best fit genetic models to reconstruct structure based on the available DNA sequences, hence requires more computational time and specs to search and assign the right model to the data. The heavy computational power required for phylogenetic reconstruction by Iqtree was not afforded trough local or remote server given the enormous amount of sample and SNP data. Thus, the number of samples for each population was reduced to ten, as a compromise between computational requirements and the need for an accurate and uncompromised phylogeny. Fewer samples in the phylogenetic tree are also advantageous for better topology visualisation. A balanced male to female ratio was made for unbiased gender related phylogeny. The two VCF datasets were then subsetted to this smaller samples size using the command: bcftools view -O z -S sample_list.txt input.vcf.gz > subset_output.vcf.gz. The same sample list was used to subset the two datasets so that the inferred phylogeny is of the same individuals through the two datasets. The sample names of the subsetted VCF files were renamed based on their population codes with numbers 1 to 10 respectively, using the command: bcftools reheader -O z -s new_names.txt input.vcf.gz > output.vcf.gz. The renamed sample list used in this command was also created through Notepad. Iqtree takes the Phylip format file as an input file. A Python vcf2phylip script (Ortiz, 2019) for converting a VCF file to Phylip format file was downloaded from Github. The autosomal and X chromosome renamed datasets were converted to Phylip format files using the command: python vcf2phylip.py -i Autosomal_tree_samples.prune.in.vcf -o sorted_OutGroup.bam. The -o option indicates the name of the outgroup as it appears in the input VCF file.

An account was created in Centre for Higher Performance Computing server (CHPC), and respective autosomal and X chromosome folders were created in the account. The autosomal and X chromosome Phylip format files were uploaded to the CHPC server account folders using the command: scp file.phy username@lengau.chpc.ac.za:path/to/chpc_account_folder. Iqtree was run on the CHPC server via a job submission script with the command: iqtree -s samples.prune.in.min4.phy -bb 1000 -nt AUTO, -bb option for bootstrapping value (Felsenstein, 1985; Hoang et al., 2018), with automatic best fit model finding for the genetic data (Kalyaanamoorthy et al., 2017). A thousand bootstraps were chosen for presenting the best possible topology. Iqtree was run on the CHPC server for speed and computational power, as the local lab server does not have enough computational power for such a large number of individuals and SNPs. The job submission script used is attached as a text file in the supplementary section (Appendix B). When the Iqtree analyses were completed, the outputs were downloaded to the local lab server from a clean Linux terminal window using the command: scp -r username@lengau.chpc.ac.za:path/to/file/on/chpc_server_account path/to/local_lab_server/folder. A clean Linux terminal is required for this download to be successful, otherwise the local computers will give connection errors. The two tree.file outputs of the two datasets were used to draw the phylogenetic trees using Figtree (Rambaut, 2016) software (http://tree.bio.ed.ac.uk/software/figtree/). The trees were modified for better visualisation in the same Figtree software and output as SVG files for higher graphical resolution.

2.7.4. Phylogenetic trees and network analyses for mitochondrial and Y chromosome datasets

Both the Y and mitochondrial chromosomes genetic markers are haploid, uniparentally inherited (only show one parental lineage of evolution) and have small number of SNPs compared to autosomal and X chromosome genetic markers. Hence, both the phylogenetic trees and networks were used to boost the interpretation of structure and geneflow patterns among populations. Haploid genetic markers do not recombine; therefore, hybrids cannot result from any of these individual genetic markers although their genetic information may be shared among individuals and clades of different ancestry. The phylogenetic trees supported by bootstrap values would assign the parental clades to such individuals that are shared between/among different clades. Both Y and mitochondrial chromosome phylogenetic trees were created using Iqtree software and similarly to the phylogenetic trees of the autosomal and the X chromosome datasets (see Section 2.7.3 for method). As the Y chromosome is only inherited by males, half the number of samples presented in the autosomal, X chromosome, and mitochondrial trees and network were presented in the Y chromosome tree and network, but 5 samples were added so that the number of samples per population in the Y chromosome tree and

network may be uniform with that of other datasets. Haplotype reticulate network through PopART software (Leigh and Bryant, 2015) were analysed in parallel with phylogenetic trees in representing the structure of these two haploid datasets, as networks show the proportion of genetic information shared between/among individuals or populations and describe more complex evolutionary events and processes imposed through the few genes shared by individuals. Reticulate scenarios such as hybridization, gene loss or duplication, and horizontal transfer of genes are better represented in phylogeny networks (Huson and Bryant, 2006).

Unlike the phylogenetic trees, haplotype reticulate networks were created without an outgroup genome but filtered with the same general filtering commands (Section 2.4). The outgroup genome was excluded from these analyses because networks do not need to be rooted based on evolution phylogeny as they cluster individuals or populations not based on evolution but the proportion of shared genetic information. PopART requires an alignment file and the text traits file as an input to create the haplotype reticulate network. The Phylip format files (alignment files) for the two datasets were created in the same procedure as the Phylip format files of the autosomal and the X chromosome datasets (Section 2.7.3). The traits text file was created in Notepad software. The traits file requires the information of the population names separated by comma at the first row. On the second row, the sample names must follow on the first column and grouped based on the population names on first row. Each sample name is followed by columns of zeroes and ones, separated by comma, that are equivalent and in order of the populations given in the first row, 1 indicating the population in which the sample belongs to. The PopART software was launched in a Windows operating platform and the alignment and traits files were of the two datasets were imported in separate runs. The networks were drawn using the median joining network option (Bandelt et al., 1999) with 5000 replicates for best network support, and then modified for better presentation before they were output as SVG files.

2.8. Patterson's D statistic and ABBABABA gene flow tests

The gene flow tests were carried out to detect if there was gene flow occurring among the African, European, and American populations during the American slavery period. But the tests were mainly aimed on resolving hypothesis 3, whether the slave masters had gene flow with the African Americans and/or Native American that they enslaved, and whether African Americans and Native Americans had gene flow between themselves.

The Patterson's D statistic ABBABABA gene flow tests (Green et al., 2010; Durand et al., 2011) were carried out using Dsuite software (Malinsky et al., 2021). Dsuite software was preferred as it is straightforward to use, is the most recent, and most accurate gene flow inference software (Malinsky et al., 2021). The ABBABABA gene flow test analyses the gene flow patterns between populations with respect to the ancestral (denoted by A) and derived (denoted by B) alleles. This test is structured as a 4 branched phylogeny set, whereby the first two branch tips (P1 and P2) represent the sister taxa individuals, populations, or species where gene flow is obvious based on the sharing of the most recent common ancestor. The third branch (P3) consists of the individual, population, or species of which gene flow is speculated to have occurred with any of the sister taxa components in P1 and P2. These three set branches are the test ingroups. The last branch (P4) consists of an outgroup component which is not closely related to any of the ingroups, and it is with certainty that there is no gene flow between the outgroup and any of the ingroups, otherwise the ABBABABA test will be flawed. The results from this test have two possible outcomes, the proportion of the ABBA gene flow to the BABA gene flow. The gene flow is always inferred between the derived alleles denoted by B; ABBA infers the proportion of gene flow between the P2 and P3 component, whereas BABA infers the proportion of gene flow between P1 and P3; if the proportion of ABBA gene flow is equal to the proportion of BABA gene flow, then there is no gene flow inferred from either of the sister taxa components with the P3 component. Roughly equal ABBA and BABA ratios are expected under a model of incomplete lineage sorting (genetic drift), however significant ABBA will have a positive D statistic and a Z score greater than positive 3, whereas a significant BABA will have a negative D statistic and a Z score less than negative 3, both of which implying that gene flow has skewed allele frequencies beyond what is expected due to random genetic drift. The populations chosen for these gene flow tests were based on four scenarios (sets), aimed at testing Hypothesis 3 of whether gene flow occurred between African slaves, their European slave masters, and possibly the local Native American populations.

2.8.1. Set 1: Gene flow between European Americans (slave masters) and African Americans (African slaves)

P1 and P2 must always be sister taxa relative to P3 or P4. Therefore, the first set consisted of either European British (GBR) or European Spanish (IBS) populations in P1, an American population with European ancestry (CEU) at P2 position, and the African American populations (either ASW or ACB) in P3, and the outgroup (chimpanzee) position in P4. This test was for determining whether Americans of European descent (P2) share more alleles with the descendants of former slaves (P3) compared to Europeans from Europe (P1) who have never been to the Americas. Therefore, the expectation if the hypothesis is true, would be a significant ABBA pattern.

2.8.2. Set 2: Gene flow between the European Americans (slave masters) and Native Americans

The second set consisted of either European British (GBR) or European Spanish (IBS) populations in P1, an American population with European ancestry (CEU) in P2 (as in Set 1), but with Native American populations (either Colombian (CLM), Peruvian (PEL), Puerto Rican (PUR), or Mexican (MXL)) in P3, and the outgroup in P4. This test was for determining whether the Americans with European ancestry (P2) share more alleles with the Native Americans (P3), compared to Europeans (P1) who have never been to America. Therefore, the expectation if the hypothesis is true, would be a significant ABBA pattern. Both Sets 1 and 2 may also highlight any differences in rates of gene flow between Europeans of British and Spanish descents with African or Native Americans.

2.8.3. Set 3: Gene flow between the Native Americans and the African Americans (African slaves)

The third set consisted of Native American populations (either Colombian (CLM), Peruvian (PEL), Puerto Rican (PUR), or Mexican (MXL)) in P1 and P2 positions, the African American populations (either ASW or ACB) in P3, and the outgroup in P4 position. This test was for determining whether Native American populations in either or both P1 and P2 had gene flow with the African slave populations during American slavery. Either ABBA or BABA patterns would be significant if the hypothesis is true.

2.8.4. Set 4: Gene flow between the African Americans, Native Americans, and British and Spanish Europeans

The fourth set consisted of Spanish (IBS) and British (GBR) populations in P1 and P2 positions, African American (ASW and ACB) and Native American populations (CLM, PEL, PUR, and MXL) in P3 position, and an outgroup in P4 position as usual. This set was designed as a control for results of Sets 1 and 2, to see if it is only European descendant Americans (CEU) who had gene flow with African/Native Americans or whether the people of British descent (GBR) were more likely to have gene flow with Native American and African American populations than people of Spanish descent (IBS). Either ABBA or BABA patterns would be significant if the hypothesis is true.

Another set with the African Americans (either ASW or ACB) in P1, either Nigerian (ESN and YRI), Kenyan (LWK), Gambian (GWD), or Mende in Sierra Leone (MSL) African populations in P2, and American population with European ancestry (CEU) in P3 was thought-out, but it could not work with

the assigned P1 and P2 populations because of the possibility of their most recent common ancestor being more ancient than the ancestor of the African Americans populations and the American population with the European ancestry. The Dsuite command used for the ABBABABA analyses was: ./Build/Dsuite Dtrios -c -n Mannda_Dsuite3 Renamed_Autosomal_tree_samples_OG.vcf Tree.txt (see Table 3 for results).

3. Results

3.1. Genomic data

The genomic dataset used in each analysis is summarised in Table 2 below. The statistics are simplified to give coverage depth of the dataset, number of individuals represented, and the number of SNPs. Additional genotype statistics (before filtering) for all four datasets such as the number of INDELS, number of MNPs (Multi Nucleotide Polymorphisms), and multi allelic sites are presented in the supplementary section (Appendix C), but it must be noted that they were filtered out through the general filtering commands (See filtering commands under the method Section 2.4). The total genotypes that were generally filtered out are 25385308 for autosomes, 1148678 for X chromosome, 46877 for Y chromosome, and 581 for mitochondrial dataset. The autosomal, X ,Y, and mitochondrial chromosome datasets had different genotype statistics, with the autosomal dataset having the highest and mitochondrial dataset having the lowest. The autosomal, X, and mitochondrial datasets had 728 individuals (samples) and the Y chromosome dataset had half the number of samples as it represents only the paternal route of inheritance. After filtering, the number of samples and SNPs per dataset were the same for respective analyses, except for the phylogenetic tree and network analyses where samples were subsetted to 10 individuals per populations, hence a lower number of SNPs in these datasets.

Table 2. Total number of SNP variants for the four subsetted human genomic datasets before and after filtering for the various analyses conducted in this study**.**

		The four dataset types			
		Autosomal chromosomes (30X coverage)	**X chromosome (30X coverage)**	**Y chromosome (30X coverage)**	**Mitochondrial chromosome (Low coverage)**
		Genotype statistics before filtering			
	Number of samples	728	728	364	728
	Number of SNPs	111860496	4468198	176147	3892
Analyses	**Genotype statistics after general filtering and analyses-based dataset subset**				
Genomic diversity	Samples number	728			
	Number of SNPs	11023827			
Phylogenetic trees	Samples number	141	141	141	141
	Number of SNPs	1561384	373548	9326	307
Phylogenetic Networks	Samples number			141	141
	Number of SNPs			9326	307
PCA	Samples number	728	728		
	Number of SNPs	11023827	401006		
Admixture and pie plots	Samples number	728	728		
	Number of SNPs	11023827	401006		
ABBABABA tests	Number of samples	728			
	Number of SNPs	11023827			

3.2. Genomic diversity

Genomic diversity among the African, European, and American populations shown in Appendix D of supplementary section was visualised and compared on the two bar graphs representing heterozygosity (Figure 3) and nucleotide diversity (Figure 4). In Figure 3, both the African American populations (ACB and ASW) showed greater percentages of heterozygous sites (7.92% and 7.90%) than all other populations. This is of similar magnitude to percentages of African populations: GWD (7.81%), ESN (7.85%), LWK (7.81%), MSL (7.87%), and YRI(7.88%) compared to noticeably lower

heterozygosity of Native American: PEL(6%), CLM(6.67%), MXL(6.39%), PUR(6.83%) and European descent populations CEU (6.34%), IBS(6.39%), GBR(6.33%). Within the least Native American and European observed heterozygosity, the Puerto Rican (PUR) populations has the highest percentage count of observed heterozygosity while the Peruvian (PEL) population bares the overall least.

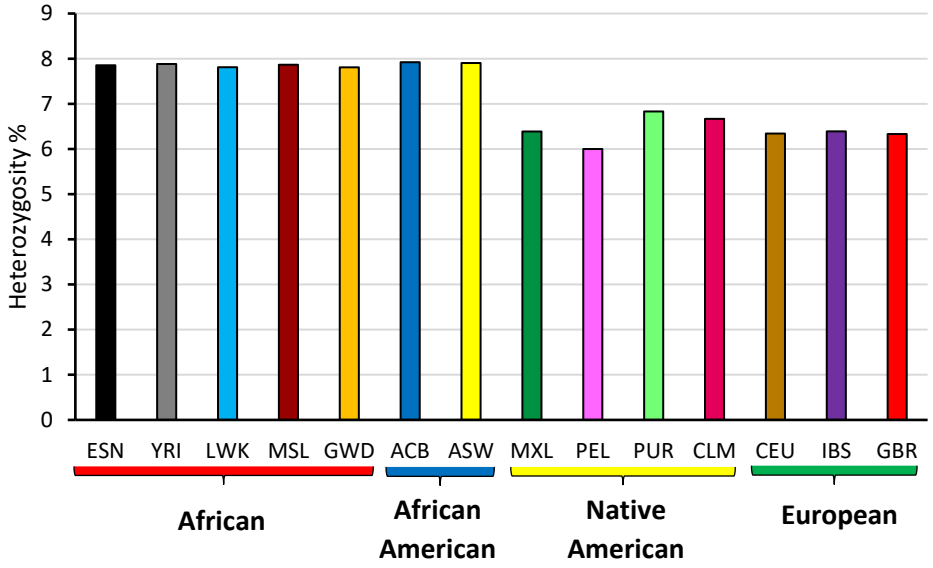

Figure 3. Bar graph comparing observed genome-wide heterozygosity among the European, American, and African populations. The population sample size is 52 individuals and the same for each population. Populations bars are colour coded, and these colours will represent the same populations throughout these results section, except for Admixture results that produced their own colours. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

Similarly to the Heterozygosity graph above (Figure 3), the African American populations (ACB and ASW) have the highest nucleotide diversity (7.92% and 7.90%) than the rest of the populations, but does not exceed by large the African populations: ESN (7.86%), YRI (7.87%), LWK (7.80%), GWD (7.84%), MSL(7.87%) compared to the Native American populations: PEL (6%), CLM (6.67%), MXL (6.44%), PUR(6.82%) and European descent: CEU(6.30%), IBS(6.37%), GBR(6.30%) populations. There is no definite difference between the nucleotide diversity of Native American and European

populations, because while those of European descent populations (CEU, IBS, GBR) are lower than three Native American populations (CLM, MXL, PUR) they exceed that of Peruvian (PEL) Native American population, but the Puerto Rican nucleotide diversity is the highest among them whiles the Peruvian's remains the least.

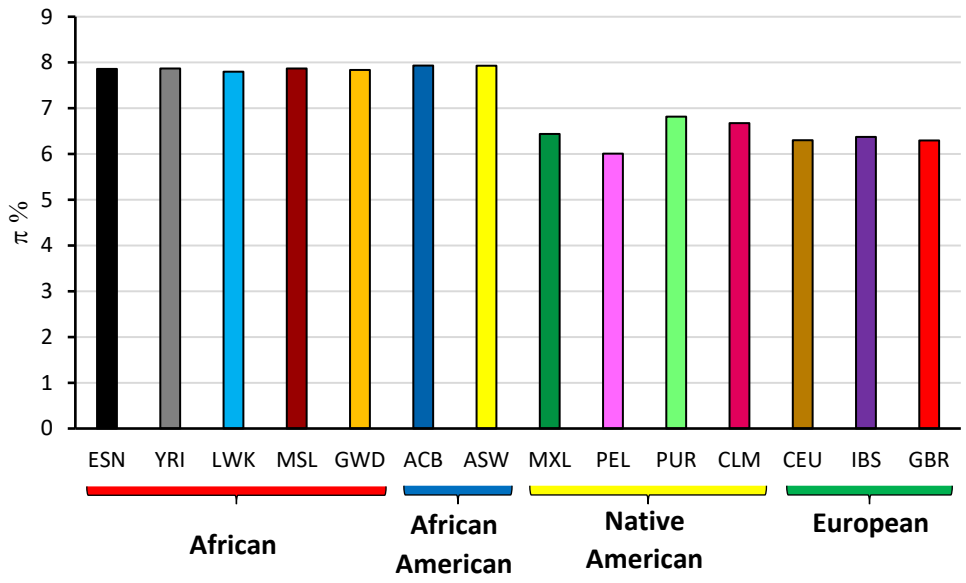

Figure 4. Bar graph comparing the nucleotide diversity (π) of the European, American, and African populations. Like with the heterozygosity graph, the population sample size is 52 individuals and the same for each population. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

3.3. Genomic structure

3.3.1. PCA plots for autosomes and X chromosome

PCA analyses showed that only the first four principal components of the autosomal dataset were informative about the genomic structure of the African, American, and European populations (Figure 5). The 1st component separated the African and the African American populations from the Native

American and European populations, with a proportion of African American individuals spreading from Nigerian populations to the American-European group and the 2nd component separated the Native American populations from the European populations (Figure 5A). The 3rd component (Figure 4B) showed structure within Africa, separating the Sierra Leone and Gambia (MSL and GWD) from the two Nigerian populations (ESN and YRI), with both the African American populations more closely related to the Nigerian populations, and the East African Kenyan population (LWK) separated from the rest of the West Africans. Although the 4th component (Figure 5C) also showed structure within Africa similarly to the 3rd component, the East African (LWK) population is grouped closely to the two (GWD and MSL) West African populations than the two (YRI and ESN) West African Nigerian populations. After the 4th component, there was no further structure revealed by the next components, therefore they are not shown.

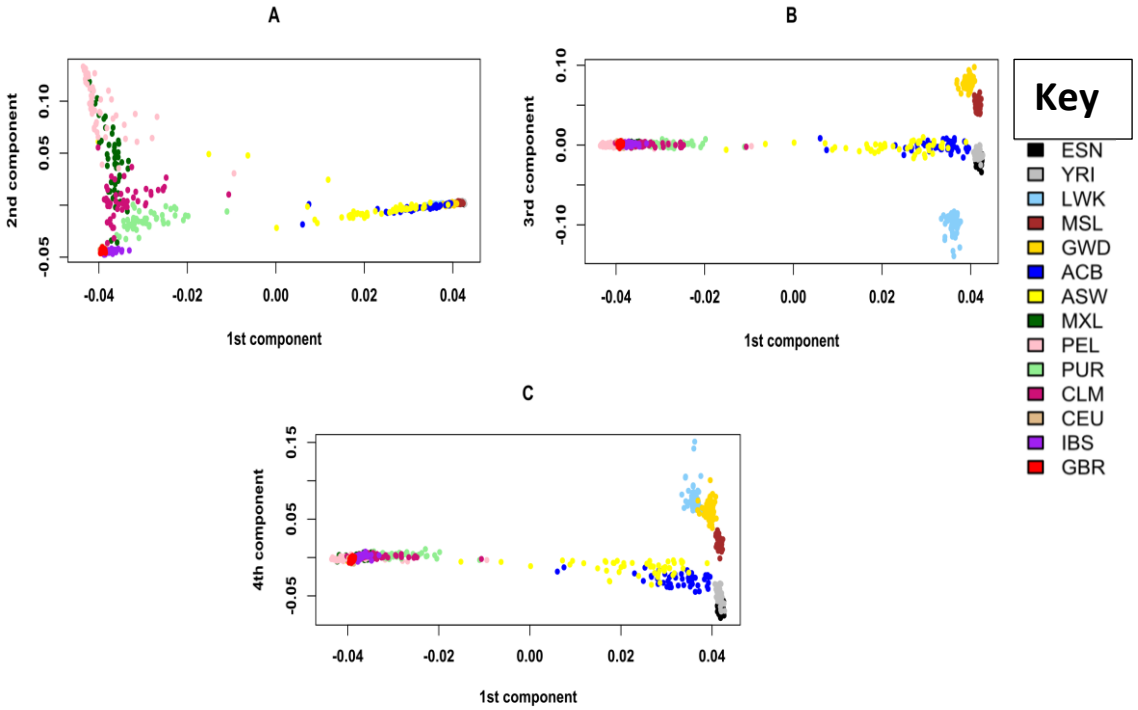

Figure 5. Principal Components analyses for autosomal chromosomes, comparing components 1 to 4. Panels A-C. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

The PCA plots of the X chromosome dataset (Figure 6) were less structured compared to the autosomal chromosomes' dataset PCA plots (Figure 5). Unexpectedly, Figure 6A appeared as if there were three separate alignments with individuals equally shared and stretching from Africa to out of Africa. Despite this observation, the 1st component separates the Africans and African Americans from the Native American and European populations, and the 2nd component divides European and Native Americans, although not as clearly as in the autosomal chromosomes PCAs. Like the 2nd component, the 3rd component revealed a little structure between Native Americans and European populations, whereas the African populations are mixed up together with both the African American populations. Neither the 4th (figure 6C) nor the next components revealed further structuring in the X chromosome data set.

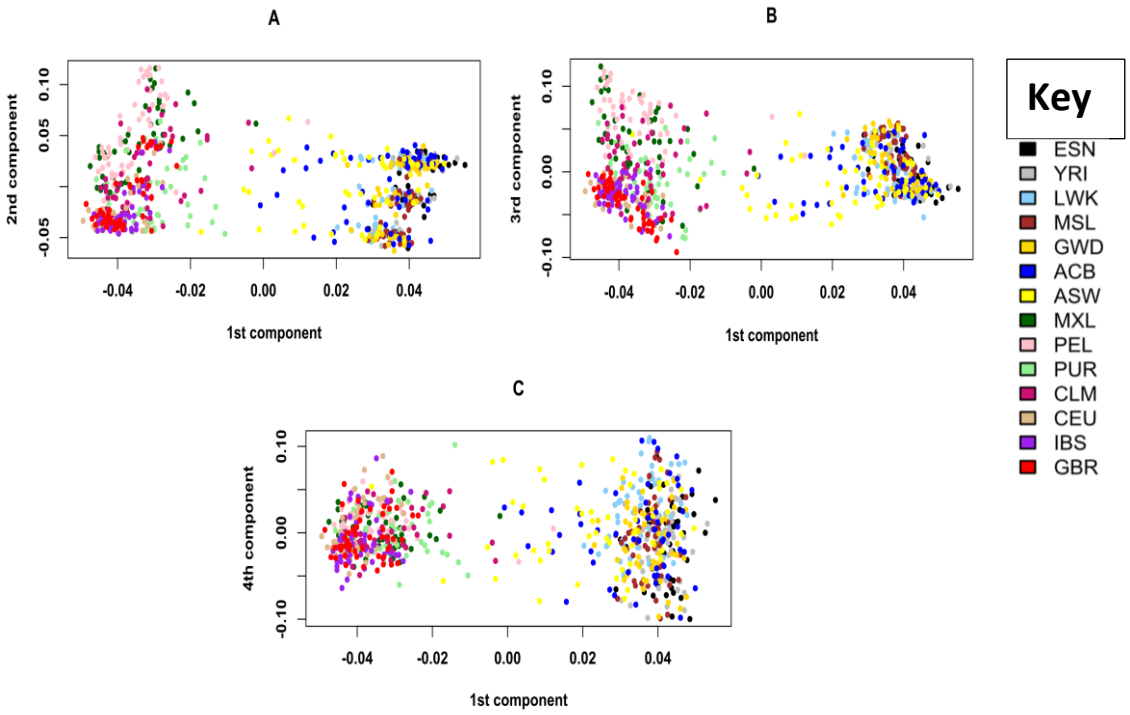

Figure 6. Principal Components analyses for the X chromosome. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

3.3.2. Admixture plots for autosomes and X chromosome

For the autosomal data set, the best K chosen based on the CV distribution across 12 values of K was seven (Figure 7A). The admixture plots reveal considerable structure within the autosomal data set (Figure 7B). K = 2 separates African from non-African populations, with some indication of admixture in LWK, GWD, MXL, PEL and IBS, but are most pronounced in ACB, ASW, PUR and CLM. K = 3 assigns each population to the three major groups (African, European, and Native American) and shows that the Native American ancestry is strongest in Mexico and in Peru. K = 4 adds greater variability within the African populations, with a clearly East African component emerging among the Kenyans. K = 5 introduces a new component (sky blue) that is strongest among Puerto Ricans, but also prominent in Mexicans, Colombians, and Spanish. K = 6 more clearly differentiates the African populations from one another, with pink colour for Kenya population, Red for MSL and GWD, and Blue for Nigerian populations. The African American populations appear highly admixed but dominantly with Nigerian genotypes. K = 7 differentiates Puerto Rico from other populations. Structure greater that K = 7 was difficult to interpret with no clear additional distinctions among populations.

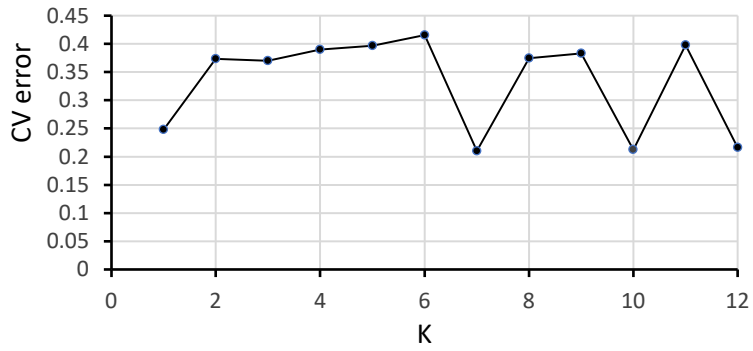

Figure 7A. Autosomal chromosomes admixture plot Cross Validation errors extracted from the 12 K values. K = 7 has the lowest CV error value.

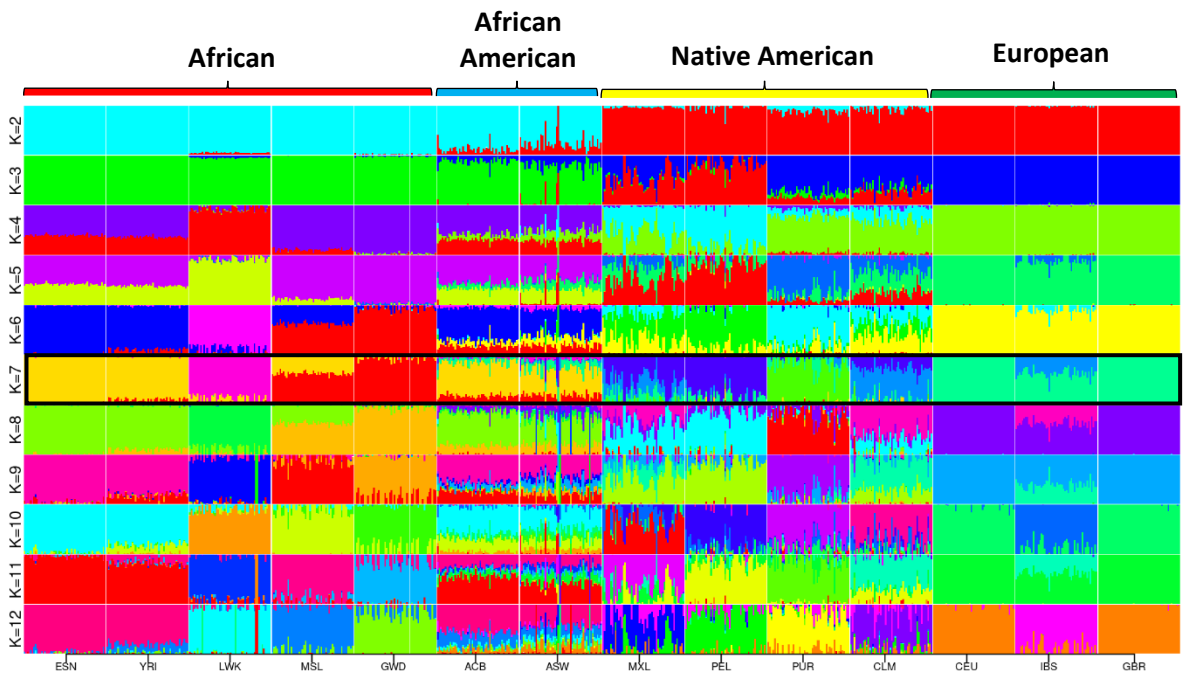

Figure 7B. Autosomal chromosomes Admixture analyses of the African, American, and European populations. The boxed K = 7 marks the admixture value with the lowest Cross Validation error, limiting the analyzation of this plot to this K value. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

When these admixture proportions were plotted onto a map of the world, more clear geographic patterns emerged (Figure 7C). The East and West African populations were clearly differentiated, with very little proportion of the West African admixture in East African populations. West African itself was also clearly divided into far western Sierra Leone and Gambia from Nigerian Esan and Yoruba. The two African American populations in the United States and the Barbados were the most admixed with mainly Nigerian admixture, smaller proportions of Gambian-Sierra Leone, European admixture, and even smaller amounts of Native American admixture. The African Americans from the USA had a notable East African ancestry component, unlike those of the Caribbean. The Mexican, Peruvian, and Colombian populations had a common dominant Native American ancestry (blue) which is also found at low frequency in African Americans from the USA and Puerto Ricans. The Puerto Rican population is dominated by its own (light green) ancestry which is presumably of Native American origin since it

is only present at low frequency in other Native American populations. Although European ancestry is common among Native Americans and African American populations, it is completely absent from Africa. A further European ancestral component (sky blue) is present in the Spanish population, as well as among Mexicans and Colombians and to a lesser extent in Puerto Ricans.

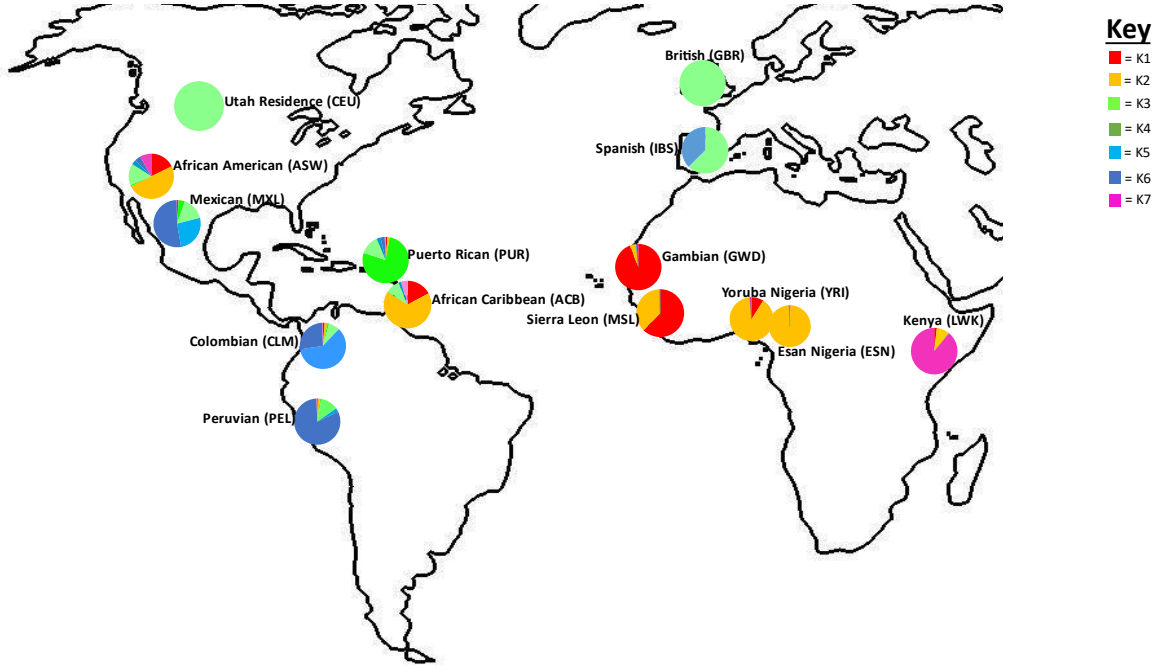

Figure 7C. Autosomal chromosomes mapped admixture proportions pie charts for African, European, and the American populations from K = 1 to K = 7.

Like with the autosomal dataset, the best K for the X chromosome dataset was chosen based on the CV distribution across 12 values, but it was K = 5 (Figure 8A). Although the autosomal chromosomes admixture plots (Figure 7B) showed more structure than the X chromosome admixture plot (Figure 8B), the X chromosome admixture plots nevertheless showed differentiation among the African, Americans, and European populations. K = 2 generally separated the African populations (sky blue) from the American-European populations (Red), with indication of admixture in LWK, MXL, PEL, PUR, CLM, but most pronounced in the African American slave descendant populations (ASW and ACB). K = 3 outlined the structure within Africa, differentiating East Africa (Red colour in LWK) from the rest of West Africa (light green). At K = 4, the Native American populations (MXL, PEL, PUR, CLM) are

differentiated from European ancestry populations (CEU, IBS, GBR), but with higher level of European admixture in the PUR and CLM populations than in MXL and PEL populations. K = 5 is as good as K = 4, the Native American admixture is better represented in the PEL and the PUR populations than the more European admixed PUR and CLM populations. There is also a noticeable but very small structure loss (less proportion of European green and more proportion of sky-blue) in the IBS population compared the other CEU and GBR populations with European ancestry.

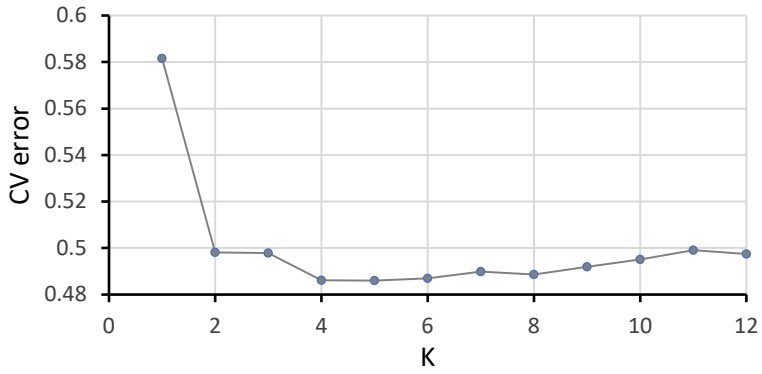

Figure 8A. Showing the autosomal chromosomes Admixture plot Cross Validation errors for 12 K values. K = 5 has the lowest CV error.

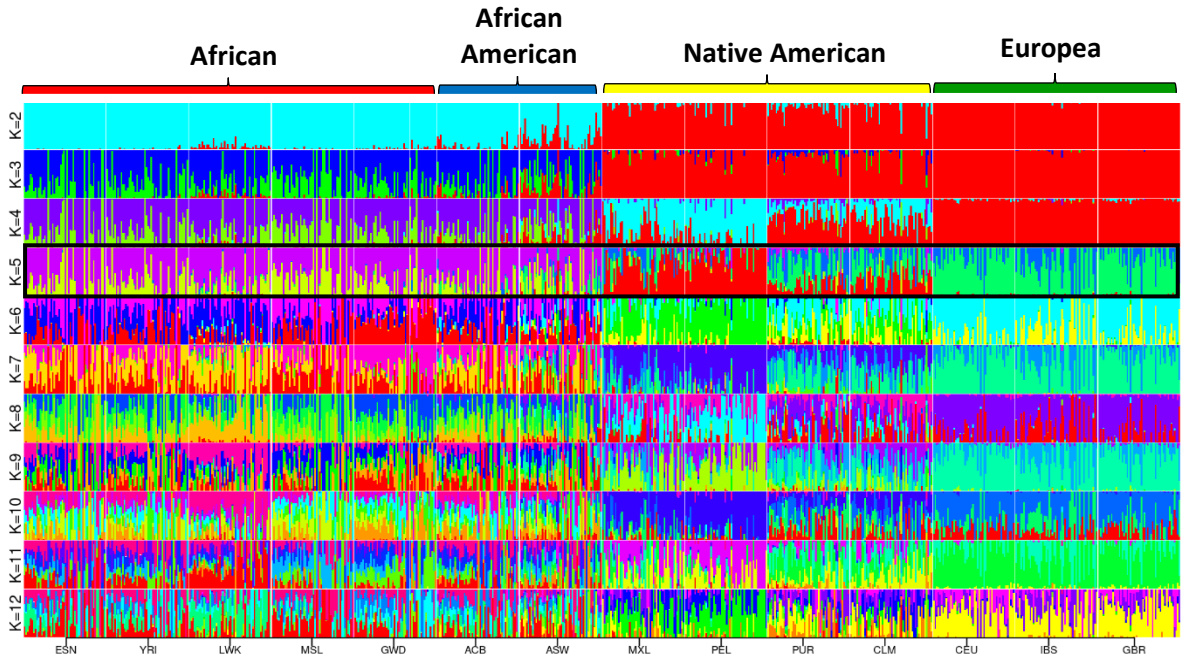

Figure 8B. Showing the X chromosome Admixture analyses of the 14 African, American, and European populations. The boxed K = 5 marks the admixture value with the lowest Cross Validation error read,

limiting the analyzation of this plot to this K value. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

3.3.3. Phylogenetic trees for autosomal and X chromosome data sets

Despite two African American individuals (ACB4 and ASW3) being basal to the whole tree, the phylogenetic structure among Africans, African Americans, Native Americans, and Europeans is clearly differentiated by the rest of the individuals within the autosomal chromosomes' dataset (Figure 9). The structure within Africa, separated the East African Kenyan population from the rest of West African populations, but there is one African American (ACB7) and Nigerian (YRI8) individual clustering within the East Africa Kenyan clade. Even within then West African region, the Nigerian, Gambian and Sierra Leone populations are also structurally differentiated, with two African Caribbean individuals clustering within the Nigerian clade. Most African American individuals and four Puerto Ricans were intermediate between the African and Native American clades, and one African American individual (ASW9) is clustered within the Native American clade. As much as slave descendant individuals are clustered within and between the Native American and African clades, there were no European individuals clustered either within the Native American or African clades, nor African Americans clustered within the European clades. However, three Puerto Rican (PUR3, PUR6, PUR7) and two Mexican (MXL2, MXL9) Native American genomes were intermediate between the Native American and European clades, and two Puerto Rican (PUR4, PUR10), two Colombian (CLM2, CLM9), and one Mexican (MXL1) Native American genomes were clustered within the European clade.

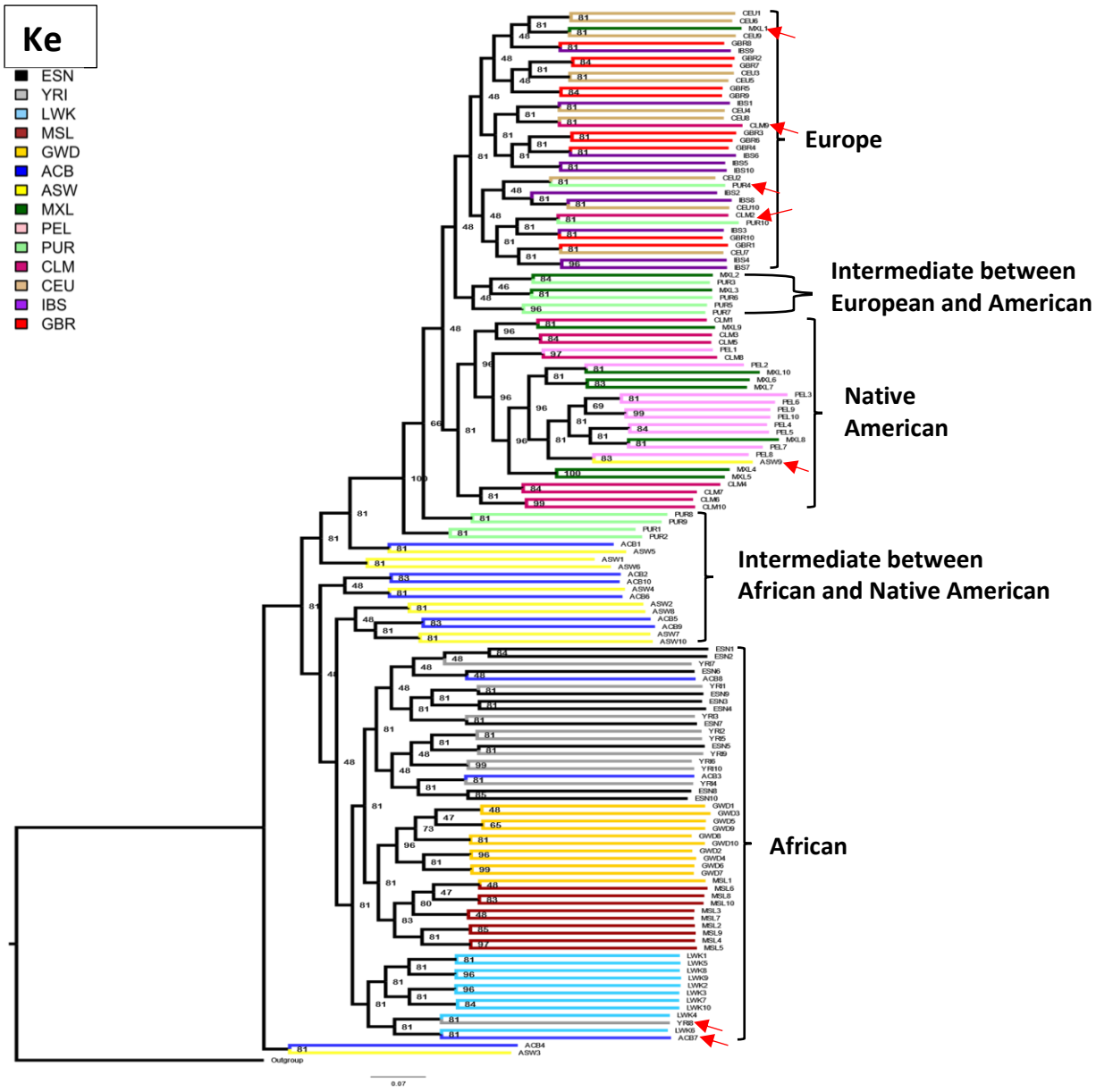

Figure 9. Rectangular phylogenetic tree of the autosomal chromosomes reconstructed from maximum likelihood and the best fit substitution model PMB+F+R5 with 1000 bootstraps. The scale bar underneath the tree represents the genetic distance (nucleotide per site) between the individuals and the red arrows indicate individuals in unexpected clades given their ancestry. The sample size is 10 individuals per population. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

The X chromosome phylogenetic tree (Figure 10) showed a grouping of some African individuals as basal to the Native American and European populations, but other African individuals as derived and phylogenetically more recent than the American and European clades. There is no structural differentiation between the East African and West African populations in both the basal and derived African populations, individuals from both the West and East Africa are mixed with one another. The African American individuals also conform to the loss of West and East African structure, mixed between the African and European clades. There is barely differentiation between European and Native American populations but are differentiated from the African populations, although some Native Americans are intermediate between the European and derived African clades. Majority of African American individuals are clustered within both the African and derived African clades, but one African American individual is sister taxa to a Peruvian individual in the Native American clade.

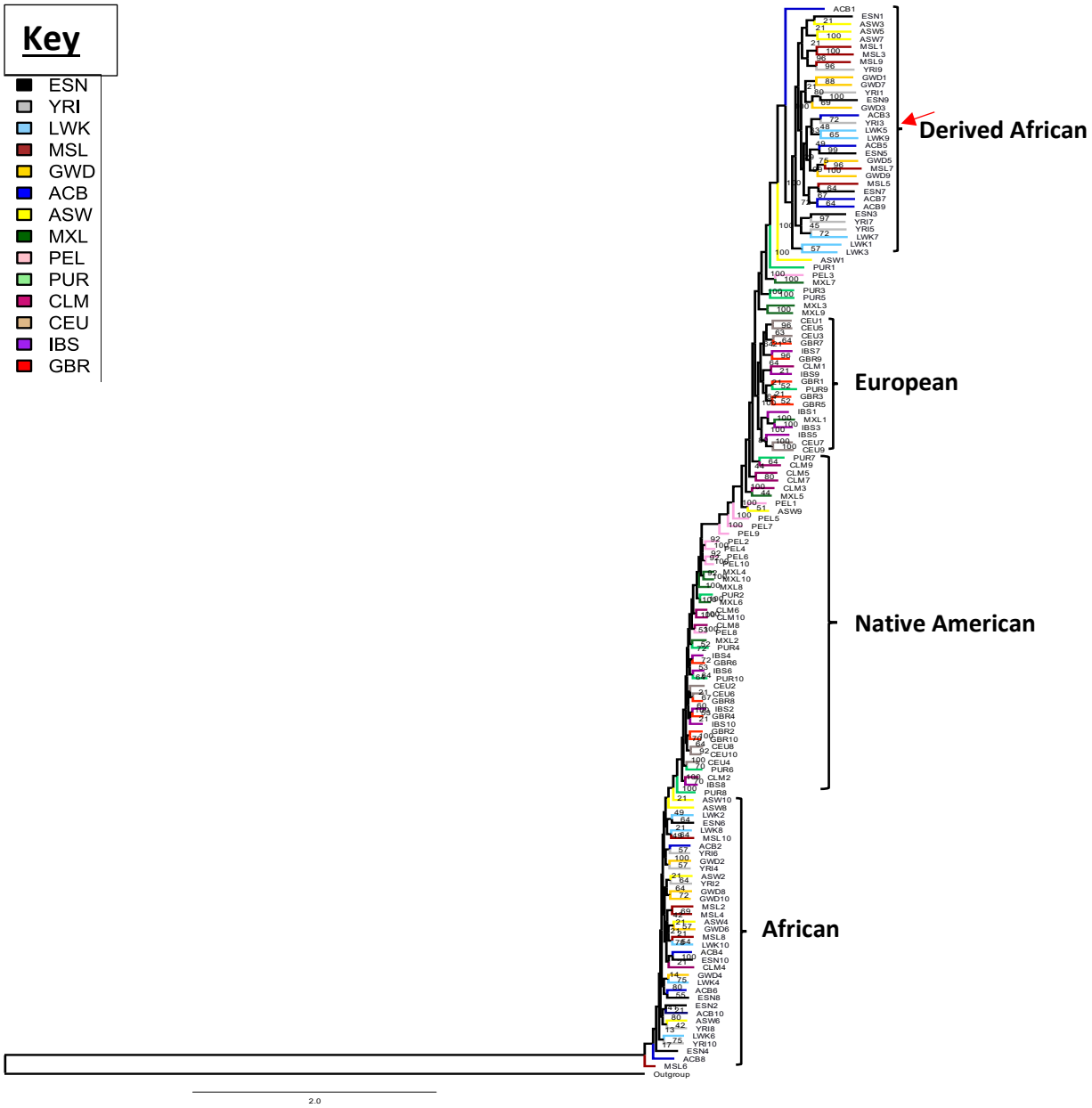

Figure 10. Rectangular phylogenetic tree of the X chromosome reconstructed from maximum likelihood and best fit substitution model PMB+F+R8 with 1000 bootstraps. The same 10 samples per populations used in the autosomal chromosomes tree were used for the X chromosome tree. The outgroup branch has been reduced for better visibility of the ingroups, hence its length is not on scale, but the rest of the branches of the ingroups are on scale. The scale bar underneath the tree represents the genetic distance (nucleotide per site) between the individuals and the red arrows indicate individuals in unexpected clades. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in

Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

3.3.4. Phylogenetic trees and networks for Y and mitochondrial chromosome datasets

Phylogenetic trees (Figure 11A and 11B) and networks (Figure 12A and 12B) were used to investigate the structure of the Y chromosome and mitochondrial chromosome datasets. Unlike the autosomal and X chromosomes that can be inherited from either male or female parent, the Y chromosome is uniparentally inherited from a male parent to male offsprings. Therefore, the structure inferred from Y chromosome phylogenetic tree (Figure 11A) and network (Figure 11B) depicts only the paternal ancestry in evolution. Both the phylogenetic tree and network show the same topology, but the phylogenetic tree supports the actual topology with bootstrap values for determining the true parental clades of individuals. Both the Y chromosome phylogenetic tree and network differentiated between the African and out of Africa structure with 100 % bootstraps, but the structure within Africa is not visible, the African individuals irrespective of their East or West African origin and majority of the African Americans from United States and Caribbean are mixed. The structure loss was also observed between the Native Americans and European individuals, they are as mixed as the African clade. Despite the structure differentiation of African and Native American/European clades three Native American individuals (PUR1, PUR5 and CLM10) are clustered within the African clade with the Kenyan and Nigerian individuals whereas four African American individuals (ASW3, ASW5, ACB3, and ACB1) are clustering within the Native American-European mixed clade.

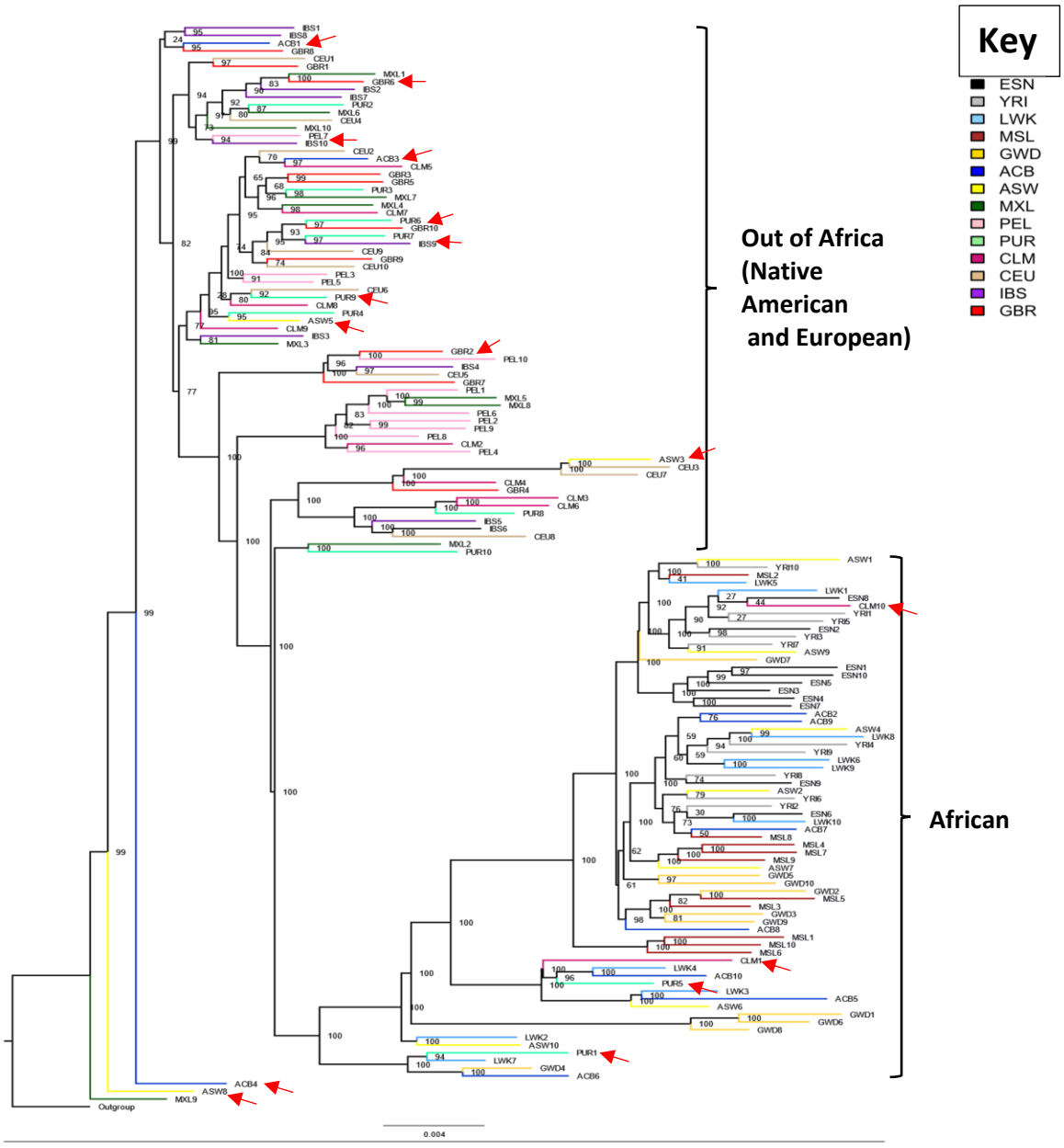

Figure 11A. Rectangular phylogenetic tree of the Y chromosome reconstructed from maximum likelihood and best fit substitution model TVMe+ASC+R6 with 1000 bootstraps. All the branches in scale. Half the number of exact samples per populations are shown compared to the mitochondrial tree because the Y chromosome is inherited only by males, but the remaining half was added of other available male samples. The scale bar underneath the tree represents the genetic distance (nucleotide per site) between the individuals and the red arrows indicate individuals in unexpected clades. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria); LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto

Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

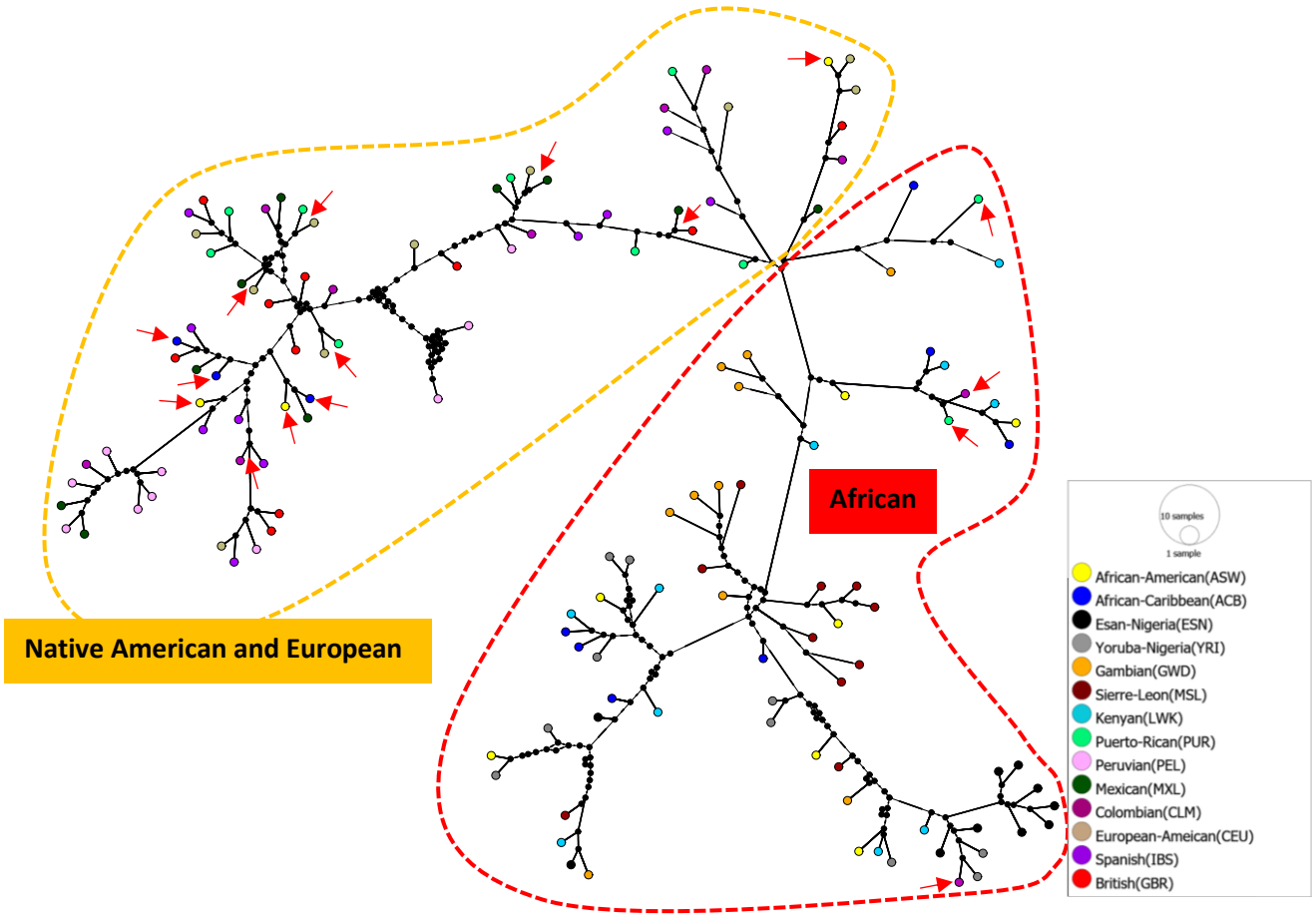

Figure 11B. Haplotype reticulate network of the Y chromosome reconstructed from Median-Joining method based on the topology of Y chromosome phylogenetic tree. The small black dots joining the branches represent interior node haplotypes that were not present in the samples, and cross bars in the on the branches indicate nucleotide differences. Dashed lines represent the topology of Y chromosome phylogenetic tree, and the red arrows indicate individuals in unexpected groups.

Like with the Y chromosome, the mtDNA is also uniparentally inherited, maternally transferred from parent to male/female offsprings. Although this DNA is represented in both the male and female offsprings, it depicts only the maternal ancestry of evolution as it can only be inherited from female parents. But unlike the Y chromosome phylogenetic tree and network, the mitochondrial phylogenetic tree and network (Figure 12A and 12B) showed more structure differentiation among the African, Native American, and European groups. Although African is differentiated from Native American there are six Native Americans (PUR3, PUR4, PUR7, PUR9, PEL5, MXL8) clustered within the African group

and one African American (ASW9) clustering within the Native American A group . Like with the Y chromosome, mitochondrial chromosome also showed no structure differentiation between the East and West African regions as the East Africa Kenyan individuals are clustered with the West African individuals. This lack of structure was also observed within American and European clades. Although one Native American clade (Native American B) appeared clustered with the European clade in the mitochondrial network (Figure 12B), the low bootstrap value (38) of this node (Figure 12A) does not support its close relatedness to the Europeans, hence it was labelled as a Native American clade. Therefore, the Native American and European clades were divided into A and B groups and the separated Native American clades circled with similar blue dotted lines.

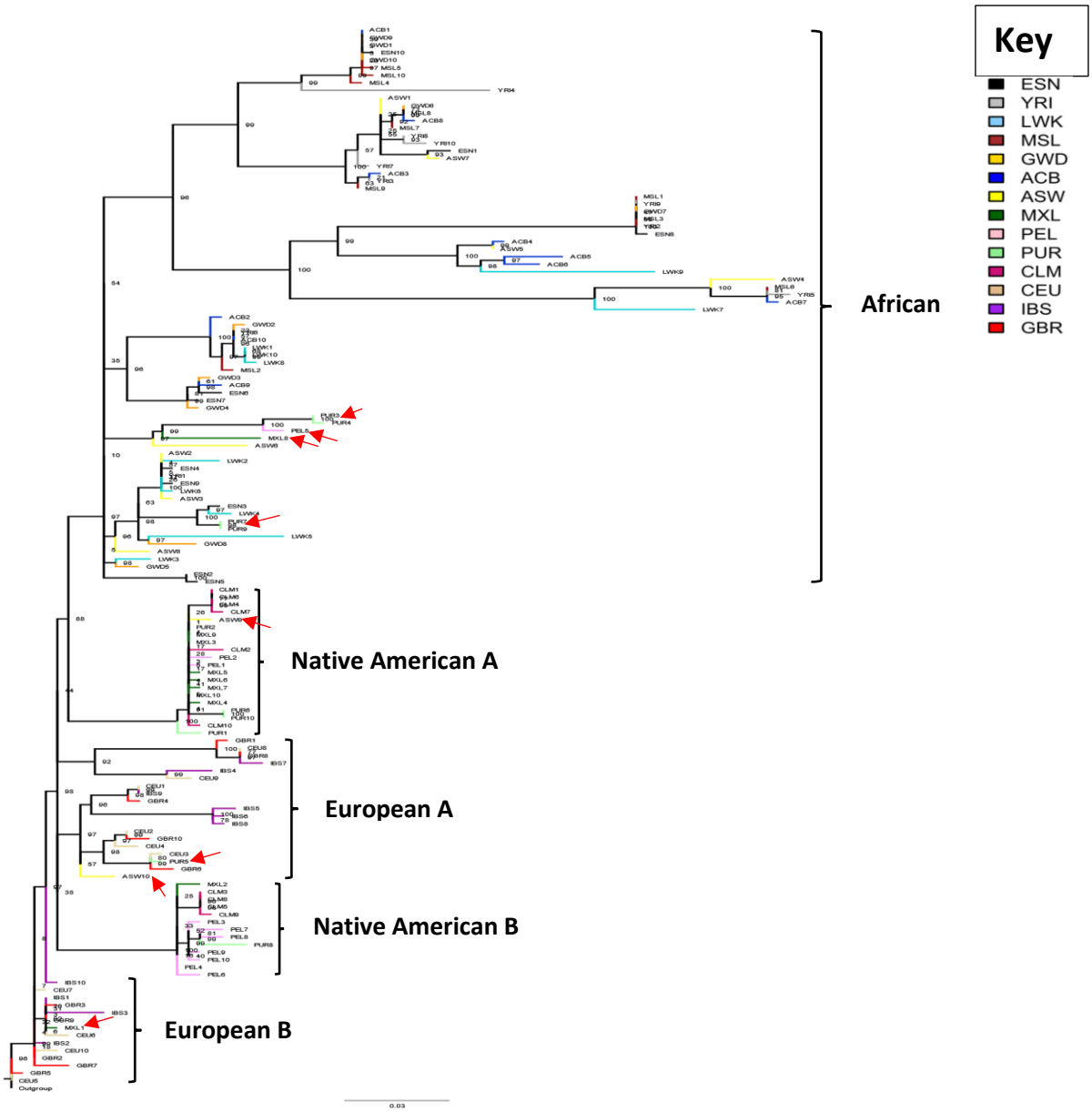

Figure 12A. Rectangular phylogenetic tree of the mitochondrial chromosome reconstructed from maximum likelihood and best fit substitution model K2P+R3 with 1000 bootstraps. The scale bar underneath the tree represents the genetic distance (nucleotide per site) between the individuals and the red arrows indicate individuals in unexpected clades. **Key:** ESN (Esan in Nigeria); YRI (Yoruba in Nigeria; LWK (Luhya in Kenya); MSL (Mende in Sierra Leone); GWD (Gambian Mandinka); ACB (African Caribbean in Barbados); ASW (African Ancestry in South West United States); MXL (Mexican Ancestry in Los Angeles, California); PEL (Peruvian in Lima); PUR (Puerto Rican in Puerto Rico); CLM (Colombian in Medellin); CEU (Utah residents with Northern and Western European ancestry); IBS (Iberian populations in Spain); GBR (British in England and Scotland).

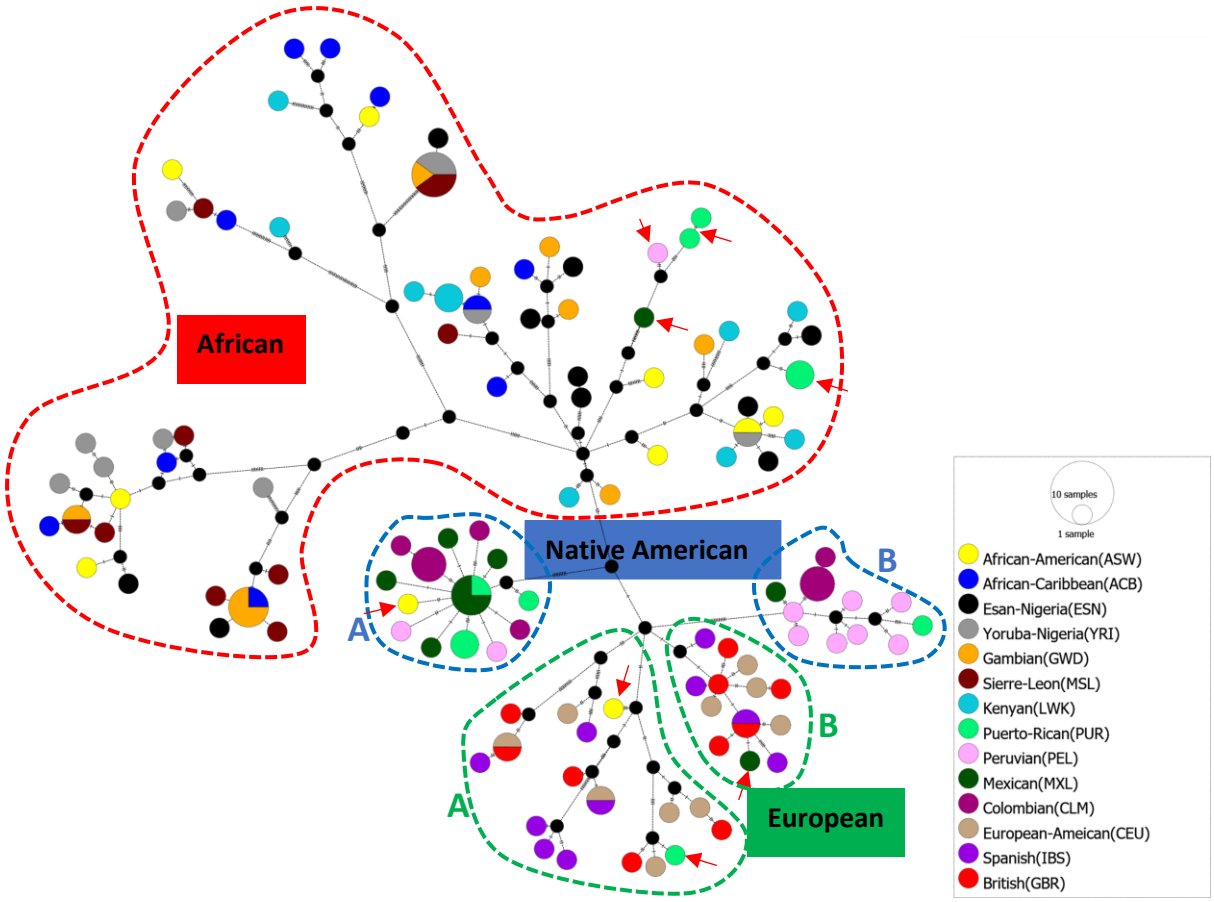

Figure 12B. Haplotype reticulate network of the mitochondrial chromosome reconstructed from median-joining method based on the topology of the mitochondrial chromosome tree. The colour coded dashed circles represent the African, Native American, and European groups and the red arrows indicate individuals in unexpected clades.

3.4. Gene flow

Significant gene flow tests statistics (Table 3) were collectively summarized based on the four sets, see the gene flow method Section 2.8 for sets explanations. For all four sets, tests were either significant or non-significant only for BABA gene flow. See supplementary section (Appendix E) for a complete set of both significant and non-significant gene flow tests.

As hypothesised (hypothesis 3), significant gene flow was observed in set 1 between the European Americans (CEU, descendant of slave masters) and both African American (ASW and ACB) populations. Not only did the European Americans (CEU) have gene flow with the African American populations, but they also had gene flow with the Native American populations (CLM, MXL, PEL, and PUR) as shown

in set 2. In the meantime, when both the Native Americans and African Americans were having genetic contact with the European slave masters, only two (CLM and MXL) of the four (PEL, CLM, PUR, and MXL) Native American populations had significant gene flow with the African American populations (ASW and ACB). In the set 4 , significant gene flow was observed between the British population (GBR) and both the Native Americans (PEL, CLM, PUR, and MXL) and African American (ASW and ACB) populations, compared to Spanish (IBS) with either.

Table 3. Autosomal chromosomes significant ABBABABA gene flow tests for Set 1, Set 2, Set 3, and Set 4. P1 and P2 are sister taxa populations, and P3 is a distant population suspected to have had gene flow with either of the sister taxa populations. D statistics determines the direction of gene flow, and the Z score determines the significance of gene flow; negative D statistic would indicate gene flow between P1 and P3 significant with Z score less than negative 3, hence more gene flow between P1 and P2 (BABA) counts than gene flow between P2 and P3 (BABA). Positive D statistic indicate gene flow between P2 and P3 significant with Z score greater than positive 3, hence more ABBA counts than BABA. The P value expresses the level of gene flow significance ranging from 0 to 1, P value less/equal to 0.05 is statistically significant. **Key:** CEU (Utah residents with Northern and Western European ancestry); GBR (British in England and Scotland); ACB (African Caribbean in Barbados); ASW (African Ancestry in Southwest United States); CLM (Colombian in Medellin); PUR (Puerto Rican in Puerto Rico); PEL (Peruvian in Lima); MXL (Mexican Ancestry in Los Angeles, California); IBS (Iberian populations in Spain).

P1	P2	P3	D statistic	Z-score	p-value	ABBA	BABA
Set 1: Gene flow between European American slave masters and African American slaves							
IBS	CEU	ACB	0.005402	4.25514	2.09E-05	153109	151464
IBS	CEU	ASW	0.006887	5.52759	3.25E-08	156473	154332
Set 2: Gene flow between the European American slave masters and Native Americans							
IBS	CEU	CLM	0.011237	8.62733	6.29E-18	174376	170501
IBS	CEU	MXL	0.012809	10.485	0	174747	170327
IBS	CEU	PEL	0.014298	12.4801	0	174474	169555
IBS	CEU	PUR	0.010174	7.76127	8.41E-15	173393	169900
Set 3: Gene flow between the Native Americans and the African Americans							
PUR	CLM	ACB	0.004595	4.10377	4.06E-05	164603	163097
PEL	MXL	ACB	0.007868	3.3229	0.000891	152838	150452

P1	P2	P3	D statistic	Z-score	p-value	ABBA	BABA
PUR	CLM	ASW	0.007272	6.10005	1.06E-09	168581	166147
PUR	MXL	ASW	0.009326	4.37268	1.23E-05	168172	165064
Set 4: Gene flow between the African Americans, Native Americans, and British or Spanish							
IBS	GBR	ACB	0.005043	4.30219	1.69E-05	153019	151483
IBS	GBR	ASW	0.006584	5.6326	1.78E-08	156391	154345
IBS	GBR	CLM	0.011337	9.71119	0	174350	170441
IBS	GBR	MXL	0.013144	11.4607	0	174770	170235
IBS	GBR	PEL	0.014588	13.5644	0	174502	169484
IBS	GBR	PUR	0.01037	8.78477	1.52E-18	173384	169825

4. Discussion

Both the heterozygosity and nucleotide diversity are comparable, with the African Americans having the highest genetic diversity than the rest of the populations, followed by African, Native American, and European populations. The PCA, admixture plots, phylogenetic trees, and networks of the four datasets differentiated African from non-African (American and European) populations, with the origin of African American populations in West Africa clearly reflected by the autosomes. Generally, the Native American and European genetic affinity was observed in the African American populations. Therefore, the African Americans had geneflow with both the Native Americans and the Europeans.

4.1. Genomic diversity (Heterozygosity and nucleotide diversity)

Considering Africa being the mother continent of origin of human diversity, the non-African populations are subsets of its diversity that evolved independently. Before the Atlantic slave trade, a subset of African genes had already migrated to Europe, Asia, and America during and after the Out of Africa migrations. Africans that remained in Africa during and after these migrations continue to exist in older clades than the European and American subsets in the other continents and preserved relatively greater effective population sizes up to this date (Reed and Tishkoff, 2006; Campbell and Tishkoff, 2008). The African continent has vast culturalization, and concurrently constitutes greater linguistic diversity, delineating nearly a third of the earth's languages (Rotimi, 2016). Africa has a lot of ancestries that approximate the complete worldwide ancestries combined (Shriner et al., 2014; Baker et al., 2017). Thus, and with reference to the Out of Africa hypotheses, the diversity of genes

should be greater in African populations and decrease as we move away from Africa to other continents (Rosenberg et al., 2002; Liu et al, 2006). Therefore, one would expect the African populations to have greater genetic diversity than any other populations outside Africa, and Native Americans to have the least diversity as they evolved later in time from the Asian subpopulation. The observed heterozygosity (Figure 3) and nucleotide diversity (Figure 4) fits this expectation. Both the heterozygosity and nucleotide diversity graphs show a similar trend among populations which was anticipated. However, genomic diversity of the African American populations (ACB and ASW) exceeds slightly (not substantially higher) that of African populations and is much higher than any European or Native American population. Considering that a large proportion of African genetic diversity is found in San and Pygmy people (the original inhabitants of southern and central Africa) compared to the African populations used in this study, the African genetic diversity could have exceeded that of African Americans. Unfortunately, the Trans-Atlantic slave trade was not cantered in San and Pygmy populations, hence they are not incomparable and incorporated in this study. But, it is possible that the African populations regarded as slave source populations in this study temporarily left Africa and then returned, suggesting bottlenecks that could explain their genetic diversity inferiority observed compared to African Americans.

The human variation that we observe throughout the world across the continents took thousands of years to be shaped by evolution, however the observed genomic diversity cannot be explained by selection because of the shorter evolutionary timeframe of the Trans-Atlantic Slave Trade to the current date. Despite the subsampling of African populations, high death rate during slave transportation, and harsh environmental conditions upon arrival in the American continent, heterozygosity and nucleotide diversity of African American populations was much higher than expected (Hypothesis 1). This could have been caused by introgression from European (Table 3, Set 1 and Set 4) and Native American (Table 3, Set 3) populations, supported evenly by the admixture plots (Figure 7B and 8B). Looking at the low to zero proportion of African ancestry in the Native American (PUR, CLM, PEL, MXL) and European American (CEU) populations (Figure 7C), it is possible that majority of the descendants of African-European/Native American mixes more likely became part of the African American populations and not the European/Native American populations. Therefore, African Americans could have a much higher than expected heterozygosity as was observed, even more than Africans in Africa.

The Puerto Rican population (PUR) showed greater genomic diversity (Figure 3 and 4) than any other Native American and European (Out of Africa) populations. Although introgression from other population may explain higher genetic diversity like observed in the African American populations, the

higher genomic diversity of Puerto Rico cannot only be based on introgression otherwise its diversity would have been comparable to that of Native American populations (Table 3). Therefore, it is an interesting observation of an island population bearing greater genomic diversity than multiple nearby mainland populations. According to the island biogeography concept (Wilson and MacArthur, 1967; MacArthur and Wislon, 2016), it would be expected that the mainland populations have greater genetic diversity as they are more likely connected by active migration (gene flow). The uniqueness of the Puerto Rican population was observed in the Admixture plots (Figure 7B and 7C) with its dominant light green that was observed in very small proportion in other Native American populations and with several of its individuals intermediately placed between the African-non-African and Native American-European clades in the autosomal phylogenetic tree (Figure 9). However, Puerto Rico only showed significant gene flow with Europeans (CEU/GBR) and not with the African American populations (ACB and ASW) (see Table 3), despite low proportions of African ancestry in Figure 7C. It is possible that D-statistics may not have detected gene flow between PUR and African American populations if African Americans were having more gene flow with Colombian and Mexican populations. Despite the separation of Puerto Rico from the American mainland, the diverse culture of this population and its mixed ancestries than other Native American populations could have influenced their elevated genomic diversity than any other Native American and European populations.

Both the heterozygosity and nucleotide diversity results do not support the original hypothesis 1 which stated that 'the Trans-Atlantic Slave Trade could be yet another subsampling of African genes, it is possible that genetic drift has lowered genomic diversity among both African American populations (African Americans in US and African Caribbean in Barbados) compared to their counterparts in West and West Central Africa'. However, it is clear from these results that gene flow between the African American populations and non-African populations in the New World has played a major role in shaping the high genomic diversity of the African Americans. The alternate hypothesis would be that gene flow between the African Americans, Native Americans, and Europeans on the American continent has offset any effect of genetic drift and maintained the genomic diversity of African American populations.

4.2. Genetic structure within and among the African, American, and European populations

4.2.1. Genetic Structure between West and East Africa

The African continent harbours diverse cultural activities and traditions, with numerous languages grouped under the four language phyla (Niger-Congo, Afroasiatic, Nilo-Saharan, Khoisan). Although

this continent was not fully represented at a population level by the studied samples, the autosomal chromosome dataset (Figure 5, 7B, 7C, and 9) nevertheless outlined the differentiation between West Africa and East Africa, even with the few populations included. This East and West African structure differentiation is strongly supported with 81% bootstraps on the first node of African group in the autosomal chromosomes tree (Figure 9), clearly denoted by grouping of the East African population separately from the West African populations (Nigerian, Sierra Leone, and the Gambia) in the 3rd component of the PCA (Figure 5B), and the East African pink colour in the admixture plots (Figure 7B). The West and East African structure is mainly differentiated by geography and language (Fan et al., 2019). From all the African countries (including the ones not sampled), Kenya in East African comprises of the most diverse language families; a combination of Nilo-Saharan, Afro-Asiatic, and Niger-Congo Bantu language families (Figure 13). This is in comparison with all the West African populations in this study that speak languages belonging to one common Niger-Congo A language family and were not surrounded by populations of different language families (Ehret and Posnansky, 1982; Tishkoff et al, 2009; see also Figure 13). The Luhya in Kenya population sampled in this study is a Bantu language (Niger-Congo B) speaking population, therefore the language variability could explain the West and East African genomes differentiation observed in the autosomal chromosomes' dataset.

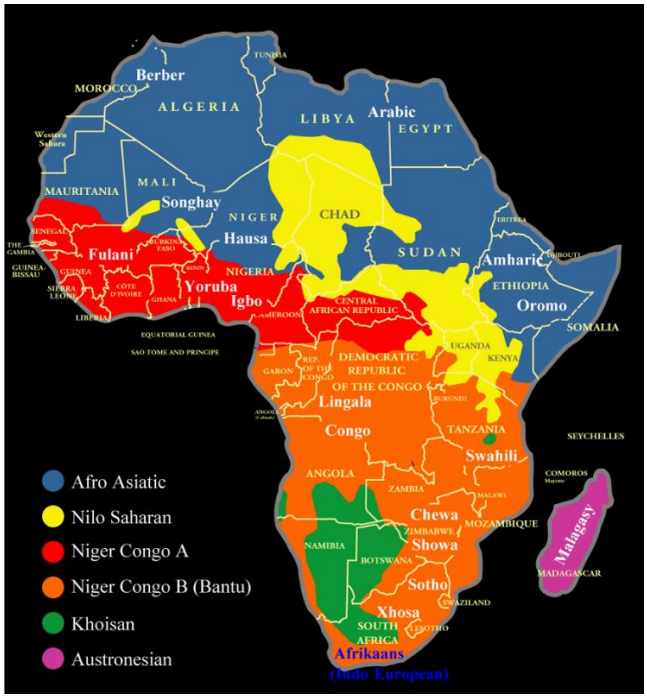

Figure 13. The map of colour-coded distribution of African language families and their major languages (Greenberg, 1963; Heine and Nurse, 2000; Dingemanse, 2004).

The differences in both language families and genomic variation between the East African population and the West African populations, could have resulted from the geographic distance between the Kenyan population and West Africa. The displacement between Luhya in Kenya and the Gambia, Sierra Leon, and Nigeria West African populations is a minimum of 3300 km. Although there is small proportion of the West African admixture in the Luhya population (Figure 7B/C) that might signal migration (gene flow) between the two regions, the distance between the West and East Africa is substantially long with numerous geographic barrier such as rivers, mountains, forests, etc. This distance possibly minimized migration between West Africa and East Africa, resulting in the two groups gradually changing independently and becoming variable due to isolation by distance (reduced long distance migration) (Wright, 1943). In addition to the isolation by distance, the environmental influences between the West Africa and East Africa are not the same, they may have contributed to influencing the gradual changes of populations between these two regions. Another reason for the Kenyan population to be differentiated from West African populations could be from the gene flow influence from presumably differentiated neighbouring ethnic groups that speak languages belonging to different language families to Luhya which are not neighbouring the West Africans. Despite the Niger Congo Bantu speaking populations, the Luhya is also neighbouring populations of the Afro-Asiatic and Nilo-Saharan language families which compares differently to the Niger Congo A language family speaking populations that are neighbouring the West African populations. These neighbouring populations could have variable genetic information that might increase the divergence of Luhya population from the Niger-Congo A speakers in West Africa. However, Afro-Asiatic and Nilo-Saharan populations neighbouring the Luhya were not sampled and it would be interesting to investigate their genomic relationship with the Luhya population to shed further light on this result. The West and East African differentiation could be from any of the discussed reasons, or a combination of some if not all.

Although the West and East African populations are differentiated, they are not monophyletic. There is significant West African admixture in the East African population (Figure 7B and 7C), one Nigerian (YRI8) and one African American (ACB7) individual in the Kenyan clade of the autosomal tree (Figure 9). This observation could not have resulted from gene flow between West and-East Africa, otherwise they would have not been differentiated in the first place or the “hybrid” YRI8 and ACB7 would have been in an intermediate position between the East and West African clades in the phylogenetic tree (Figure 9). This is likely the result of incomplete lineage sorting (ancestral polymorphisms) reflecting the time long before the abduction of slaves before the isolation of West and East Africa.

4.2.2. Genetic Structure within West Africa

The Gambian, Mende in Sierra Leone, and Nigerian populations all speak languages within the Niger-Congo A phylum, but differentiation was observed between the Nigerian populations and the other West African populations. Considering the common language phyla shared by these populations one would expect them to have very little differentiation, but with reference to the autosomal dataset (Figure 5B and 5C, Figure 7B, Figure 7C, and Figure 9), these West African populations (Nigerian and Gambian-Sierra Leone clades) were indeed differentiated with 81% bootstraps for the second node of the African clade (Figure 9). The West African (Nigerian and Gambian) differentiation observed in this study from 11023827 autosomal SNPs (Table 2) was also inferred through PCA and Fst analyses by Bhatia et al. (2011) from 309373 autosomal SNPs, but with high (500) sample size per population compared to 52 sample size per population used in this study. Though these West African populations are not geographically distantly isolated like the West to East African (Kenya) regions, there is still considerable distance between them that might influence differentiation. The Gambian and Mende in Sierra Leone populations are geographically closely orientated compared to the Esan and Yoruba Nigerian populations, therefore, migration (gene flow) is more plausible between them than to the far Nigerian regions. Thus, possibly explaining differentiation by the dominant Gambian admixture (red colour, Figure 7C) shared more intensively with Mende population than with the Nigerian populations. Although these West African populations speak languages belonging to common Niger Congo A language family, their languages, traditions, and culture practises are variable (Bulley et al., 2017) and such variability affect chances of gene flow (Fix, 1995).

4.2.3. The genetic origins of African Americans

Although a huge number of African slaves could have been sourced from West Central Africa, most of the West Central African slaves were taken to Southeast of South America (Salas et al., 2004; Li, 2020; Figure 2). Their genetic affinity may be less represented in the European colonies in North American, Central American, North of South American and Caribbean Islands that were constituted mostly of West African slaves. And previous historic and genomic studies proposed that the West African region was the source for majority of modern African Americans in United States (Salas et al., 2005; Tishkoff et al., 2009; Bryc et al., 2010; Lovejoy, 2011). It would have been interesting to investigate the proportion of West Central African genetic affinity in these sampled African American populations, but there were no samples representing the West Central Africa during the download and analyses of samples of this study, hence this study did not detect West Central African admixture and cannot make conclusions concerning the genetic proportion within African American populations relatively to West Central Africa. If the West Central African populations were sampled, they possible would be

differentiated from both Nigerian and far West populations (Gambia and Sierra Leone) based on gradual changes by geographic isolation and different language families, but some of the slaves possibly would have appear closely related to these West Central populations in the autosomal PCAs. The West Central populations could have been separated from the rest of the West Africans with higher bootstrap values in the autosomal phylogenetic tree, but sister taxa to Nigerian clade with possible African American individuals clustering within and intermediating their clades.

Both the African American populations are most closely affiliated to the Yoruba and Esan Nigerian populations in the autosomal PCA plots (Figure 5), admixture plots (Figure 8B), and phylogenetic tree (Figure 9). This attest to more than 70% preponderant Niger-Kordofanian ancestry that was found in African American populations (Tishkoff et al., 2009; Shriner et al., 2014; Campbell et al., 2014; The 1000 Genomes Project Consortium, 2015). Although this and two African American individuals (ACB8 and ACB3) in the Nigerian clade (Figure 9) may reflect a Nigerian origin of the African Americans, many of the African American individuals formed intermediate clades between the African and Native American populations with low bootstrap values support. The most plausible reason for these African Americans being African-Native American intermediates could be gene flow with the non-African populations, and based on the admixed genome hypothesis, admixed genomes will always occupy an intermediate position on the tree from a recombining genetic marker (Lavretsky et al., 2015). This is also evident in the 1st components of both the autosomal and X chromosome dataset PCAs (Figure 5 and 6) where majority of African American individuals are intermediately placed between Africans and non-African populations, but closer to Nigerians than other Africans. The phylogenetic tree topology would have shown better and detailed structure if the number of samples per population and SNPs were as high as the ones used in the PCA and admixture plots analyses, however, it was not possible to run such an expanded phylogenomic analysis given the computational and time constraints.

Following the Nigerian ancestry, Gambian and Sierra Leone ancestry is the second greatest proportion among African American populations (Figure 7C). This implies that the second most common ancestry of African Americans is from far West of Africa from the Gambian and Sierra Leone regions. Although (based on sampled populations) the Europeans attained high proportion of slaves in Nigeria, considerable proportion was also abducted from Gambia and Sierra Leone (Bryc et al., 2010; Gates, 2014).

The observed small proportion of East African (Kenyan) admixture in both the African American populations in the admixture pie plots (Figure 7C) and branching of the African Caribbean individual with the East African individuals (Figure 9) contrasts with the well-known West Africa origin of the African American populations. It is possible that this East African admixture in African American populations would be greater than observed if I had more East African populations included in my analyses. There is some degree of Kenyan ancestry in the Nigerian populations due to incomplete lineage sorting or less likely due to introgression (Figure 7C), therefore it is possible that the East African ancestry observed in the African American populations was carried along to the Americas by Nigerians during the slave abduction. However, the observed proportion of East African admixture in African American populations in the admixture plots could be too high to result from the little East African ancestral polymorphisms or admixture observed in the Nigerian populations. Alternatively, East African ancestry in African Americans could also have come from the East African population itself. Though modernization has bridged the migration gap between the continents using ships and aeroplanes, the Kenyan ancestry in African Americans could not have occurred during modern times, as the ancestry is apparent in African Americans from the USA and the Caribbean, as well as at very low frequency in Puerto Rico, Colombia, and Peru. Therefore, a subset of the Kenyan or East African Bantu population may have found its way to the American continent, possibly even during the time of slavery and contributed to the East African ancestry observed in the African Americans. With this regard, the West African subsets could not be the only constituent African ancestry in the American continent.

Both the Y chromosome (Figure 11A and 11B) and mitochondrial DNA (Figure 12A and 12B) phylogenetic trees and networks did not differentiate between East and West Africa, and the African American individuals are clustered within the mixed diversity of West and East Africans. From these datasets it is not clear to where the African American ancestry could be from in Africa. Based on this undefined West and East African structure, African Americans could either be from West , East or both African regions. The Y chromosome and mitochondrial DNA datasets are likely affected by incomplete lineage sorting, as there are simply not enough SNP data to fully resolve geographic groups within Africa. Therefore, if these haploid genetic markers cannot resolve the African structure, it is almost impossible or insufficient to use them separately for inferring population differentiation within Africa and outside Africa, emphasising the importance of whole genome sequences with millions of SNPs that are able to resolve this population differentiation patterns.

Despite the lack of African structure in Y chromosome and mitochondrial DNA datasets the observed West and East African structure from the dataset supports my second hypothesis that the African American genomes will reflect their origins in Africa is accepted. The majority of African Americans evidently originate in West Africa, but there could have been a proportion originating in East Africa.

4.2.4. Native American individuals within the African clades and vice versa

Although both the Y chromosome and mtDNA did not distinguish structure within Africans, Africa was differentiated from out of Africa (Figure 11A and 11B; Figure 12A and 12B). So, it is possible to determine whether some haplotypes are shared through recent gene flow, and not because of incomplete lineage sorting. The observed four Native American individuals (CLM1, CLM10, PUR1, and PUR5) clustered within the African clades in the Y chromosome dataset (Figure 11A and 11B) could only have inherited their African Y chromosome DNA from African paternal ancestry, whereas two African Americans that are sister taxa to Native Americans (ACB3-CLM5 and ASW5-PUR4) in the Native American-European mixture of the same Y chromosome could have inherited the Native American haplotype from Native American paternal ancestry. The six Native Americans (PUR3, PUR4, PUR7, PUR9, MXL8, and PEL5) clustering within the African clade in the mitochondrial chromosome dataset (Figure 12A and 12B) could only have inherited their African haplotype from African maternal ancestry, and the African American (ASW9) clustered within the Native American A group can only have inherited the Native American haplotype from Native American maternal ancestry. Therefore, suggesting historical bidirectional gene flow between the African Americans and Native Americans. Although the ABBA/BABA tests (Table 3) did not identify any significant autosomal gene flow between Puerto Rico (PUR) and either ACB or ASW African American populations, many PUR individuals have clear maternal and paternal African ancestry that could reflect their gene flow with Africans in the Y chromosome and mitochondrial DNA datasets. As explained in genomic diversity Section 4.1, indeed significant D-statistics between P2 and P3 does not mean that there was no gene flow between P1 and P3, but that there was more gene flow between P2 and P3 than could have masked geneflow between P1 and P3 (Baute et al., 2016).

4.2.5. Genetic structure within Iberian Peninsula

The sky-blue colour in the autosomal chromosomes' admixture and pie plots (Figure 7B and 7C) could represent genomic structure within the Iberian Peninsula, and this could be due to numerous possibilities. Firstly, at about 711 years AD the ancestral Moroccans (Moors) under the leadership of Tariq ibn-Ziyad invaded and took over the Iberian Peninsula, ruling Spain for a period of about 800

years (Scobie, 1992). Their wealthy North African civilization advanced and industrialized Spain in many areas of science such as Astronomy, Mathematics, Chemistry, Geography, Philosophy, and Chemistry. The observed sky-blue ancestral component in Spain could be reflecting this North African admixture, which occurred long before the Spanish conquest of the Americas. If the sky-blue ancestry in the Iberian population was due to European substructure, it would possibly have been reflected in the British and Americans with European ancestry. However, it cannot be confirmed that this is a signature of the Moors occupation of the Iberian Peninsula since Moroccan and other Bedouin (North African) populations were not available for sampling for this study. There is also a possibility that the Moor ancestry did not constitute the Iberian Peninsula for a long time but wiped out of Spain when the Moors were eventually expelled by the native Spaniards.

The dominance of the sky-blue colour in the Colombian population in greater proportion than in the Iberian population could represent another possibility. It could be that the sky-blue admixture represents structure within Native American populations, since it is also observed but in low levels in the Mexican, Peruvian, Puerto Rican, and African American populations. This sky-blue admixture could have found its way back to Spain if some admixed Spanish and Native Americans migrated back to Spain from America after the abolishment of slavery. But this would have required massive migration back to Europe to constitute a large of the Iberian local population. A final and most likely possibility is that this admixture could be a unique Iberian genetic signature that is differentiated from the British population as they are geographically isolated and may thus have evolved independently. The unique Basque people of Spain might also be the source of the sky-blue ancestry, although again it would be difficult to determine without known Basque genomes for comparison. If it is a Spanish admixture, it could have found its way to the American continent during the colonization of America by Spanish Europeans. The Spanish slave masters could have had gene flow with the Native Americans and African Americans, explaining the spread of this admixture, but this would be further discussed in the gene flow section below.

4.2.6. Genetic structure within America

The ancestry of the Latin Americans is a product of the post-Columbian admixture among the Native Americans, Africans, and Europeans (Soares-Souza et al., 2018); hence their genetic makeup is a mosaic of segments representing the three ancestries. This mixed ancestry is observed in the K = 7 admixture plot (Figure 7B and 7C) of the autosomal chromosomes and K = 5 Admixture plot (Figure 8B) of X chromosome. The royal-blue and light green (Figures 7B and 7C, K = 7) and red (Figure 8B, K

= 5) ancestries that are dominant in the Latin American populations could only have been inherited through their Native American ancestry, because they are not commonly shared with the African nor European populations. The sky-blue and cyan ancestries (Figures 7B and 7C) could mean that during the early years when the European population colonized America before the time of African slavery, they had gene flow with the Native American populations, and their offspring evolved into the Latin American populations we see today, but such gene flow unravelment is better explained in the gene flow section of this discussion. The sky-blue Spanish admixture, which I proposed from the previous section (Section 4.3.5) could be a unique separation of Iberian from British, is in highest proportion in the Colombian population compared to Mexican, Puerto Rican and Peruvian populations (Figure 7C). This poses the idea that the Spanish shared their defined genetic signature more intensively with the Colombian population than other Native American and African American populations. Along the south-east American coast route, some Spanish populations could have inhabited the island of Puerto Rico, constituting the observed but small Spanish admixture in the Puerto Rico population.

Another alternate idea, although less likely, is that the sky-blue admixture could be a Native American structure, hence explaining its more dominant proportion in the Colombian population than in the Spanish population. The Spanish population could have received this admixture from gene flow with Colombian populations after colonization, but if some of the Spanish-Native American hybrids travelled back to Spain after the genetic contact to constitute the Native American admixture in Spain. But if this was the reality, the Spanish-African American hybrids admixture could also have reflected in the Spanish population if the hybridized Spaniards found their way back to Europe.

Despite the Puerto Rican population being Native American, it is differentiated from the mainland Native American populations (Peruvian, Colombian, and Mexican) in both the autosomal chromosomes PCA (Figure 5, 2nd component) and admixture (Figure 7B and 7C, K = 7) plots and slightly in the X chromosome Admixture plot (Figure 8B, K = 5). A biological explanation that could be assigned to this structure differentiation would be based on the theory of island biogeography (Wilson and MacArthur, 1967 and 2016). Island populations experience different environmental conditions as opposed to their mainland counterparts. The patterns of gene flow into or out of islands are hugely dependent on the connectivity of the island with the mainland. Unlike the mainland Native American populations which are connected, the Puerto Rican population is isolated in the Caribbean Sea. The large distance between Puerto Rico and the American mainland (North, Central, and South American) possibly negatively influenced connectivity and influenced the Puerto Rican population to evolve

independently with very little gene flow from the mainland Native American populations. Therefore, Puerto Rico might have become a structured Native American population long before the Europeans colonized it. This independent evolution of the Puerto Rican population from the mainland Native American populations could be the basis of their differentiation.

Unlike other American (Native Americans and African Americans) that are highly admixed in the Admixture pie plots (Figure 7C), the Utah Residence population with European ancestry (CEU) is completely composed of European ancestry. Although this population is geographically orientated in the same mainland and not hugely isolated from African Americans in United States (ASW) and Mexican (MXL) populations that should affect migration, there is no African nor Native American admixture represented in this European American population (Figure 7C). However, with reference to Table 3 (Set 1 and 2 and 4), there is indeed significant genetic introgression of Europeans into the native and African American populations. This, together with the complete lack of non-European ancestry in the European American population (CEU, Figure 7C) suggests that the gene flow pattern was one-directional, from the European population into the Native American and African American populations. European-Native American/African American hybrid individuals would thus have been more likely incorporated into the Native Americans and African American populations than into the European American population.

4.3. Flawed X chromosome dataset

Due to variable recombination patterns of differently inherited genetic datasets, the X chromosome has 25% lower effective population size than the autosomal chromosomes, but 50% higher than the Y and the mitochondrial chromosomes (Kaessmann et al., 1999). The effect of genetic drift should be greater in the population with lower effective population size than in the population with higher effective population size, hence genetic drift should be harsh and create more structure in the dataset with lower effective population size. Therefore, the X chromosome dataset should have theoretically and practically given more if not the same structure with the autosomal chromosomes' dataset. The X chromosome dataset was processed bioinformatically the same way as the autosomal chromosomes' dataset, but it gave a relatively unstructured PCA (Figure 6), low levels of structure in the Admixture plot (Figure 8B), and a phylogenetic tree with a strange topology (Figure 10). The 1st and the 2nd PCA (Figure 6A) that should account for most variation of the dataset showed three unlikely groupings of the alignment, each with all 14 populations represented. In addition to this, the X chromosome phylogenetic tree (Figure 10) literally shows two alignments; one with basal African

populations with all 14 populations aligned with each other and the other group of genomes with derived African populations aligned to themselves and after the American/European clades. It seems there are two groups of X chromosome alignments (one within the other) and each one shows African different from European/Native American. Given the known human ancestry, which is faithfully reflected in the phylogenetic tree of the autosomal dataset (Figure 9), I take the above as evidence that there was a problem with the alignment of the X chromosome dataset, which I was unable to resolve, and which is why I cannot rely on the interpretation of its results.

4.4. Gene flow in the Americas

4.4.1. Gene flow between the Europeans (slave masters) and native and African Americans (African slaves)

Although the gene flow patterns inferred between the European slave masters (CEU), Native Americans and African Americans from the autosomes may reflect a great influence the American slavery had in shaping the genetic structure of the involved populations, the ABBABABA gene flow test have been proven to show misleading results in some cases where it cannot differentiate between gene flow and incomplete lineage sorting (Moodley et al., 2020). Therefore, it was necessary to incorporate the Y and mitochondrial datasets in support of the ABBABABA results to improve the power of gene flow inference.

The Y chromosome phylogenetic tree (Figure 11A) did not differentiate between the Native American and European structure but depending on which genomes have a sister relationship you can tell what the paternal ancestry was within the Native American-European mixture. There are two African American genomes that are sister (most closely related) with Europeans (ACB1-GBR8, ASW3-CEU3), and seven Native American genomes sister with Europeans (MXL1-GBR6, PEL7-IBS10, PUR6-GBR10, PUR7-IBS9, PUR9-CEU6, MXL3-IBS3, PEL10-GBR2). These African American and Native American individuals most likely inherited the genetic affinity that made them sister taxa with Europeans from European paternal ancestry, thus suggesting a paternal unidirectional gene flow from Europeans to African Americans and Native Americans. The sharing of Y chromosome genetic information between the British, European American, African American, and Native American genomes was anticipated as their gene flow was inferred in the autosomal ABBABABA test (Table 3, Set 1, Set 2, and Set 4), but the observed sharing of haplotypes between Spanish and Native American individuals was not inferred from the ABBABABA test, this could be from D-statistics reasons explained in the previous sections (4.1 and 4.2.4).

Contrasting to the Y chromosome dataset, the mitochondrial dataset (Figure 12A and 12B) did not only differentiate between African and out of Africa populations but differentiated also the Native American and European structure from each other, therefore gene flow direction can be concluded more simply from this genetic marker. There are two Native American (MXL1, PUR5) and one African American (ASW10) genomes in both the European groups (European A and European B), but no European genomes are clustered within the Native American clades. The Native American and African American individuals in the European clades could only have inherited the European haplotypes from European maternal ancestry because mtDNA is uniparentally and maternally transferred from parents to either male or female offsprings. Although higher sample and SNPs size could improve the power of inference, these mitochondrial results shows very clearly that some European women had relationships with both Native American and African American slave men, whereas there is no evidence from this mtDNA dataset that the opposite was true.

Combining the results from autosomes (ABBABABA tests) and Y chromosome tree and network with the European maternal genetic affinity in the African American and Native American genomes, it is clear that both male and female Europeans were having gene flow with the Native Americans and African Americans. But the European-African American/Native American hybrids may not have become part of the European but African American and Native American populations because the European populations are not admixed with African American nor Native American genetic affinity (Figure 7C); thus, supporting the unidirectional gene flow from European to African American and Native American populations.

4.4.2. Gene flow between African Americans and Native Americans

Furthermore, during slavery the African slaves could also have been in close contact not only with the European slave masters, but with the Native Americans that were in common landscape during the time of slavery. This African Americans are from Oklahoma, which is also inhabited by tribes from the south-east that were forced to relocate there (e.g. Cherokee); they are known to have copied their European neighbours, which included the ownership and trade of African slaves. However, the African American-Native American ABBABABA gene flow test was only significant between African Americans, Colombian, and Mexican populations, and not with Puerto Ricans and Peruvians (Table 3, Set 3). But as explained in the previous sections (4.1 and 4.2.4), the lack of significant D-statistics between African Americans and Puerto Ricans/ Peruvians does not necessary mean there was no gene flow between them. Therefore, it can be concluded that not only were the African Americans and Native Americans

having gene flow with the European slave masters, but also between themselves. This could explain a small proportion of Native American royal-blue, and Puerto Rican light green admixture in both the African American populations. Despite all the hardships and abuse the African Americans and Native Americans suffered from the European slave masters, they possibly did not stop some of them from having sexual relationships with each other.

During the Spanish colonization of Puerto Rico, majority of indigenous Puerto Ricans (Tainos) were wiped out of existence by the Spanish forces and new diseases brough by the Spanish colonization (Yaz, 2001). Both the Peruvian and Puerto Rico Spanish colonizers imported limited African slaves from Africa and other regions in Americas for labour in their colonies (Stark, 2009). Despite the ABBABABA test not detecting significant gene flow between African Americans and Peruvian/Puerto Rican populations, the African slaves shared the landscape with the colonized Peruvians and Puerto Ricans in the Spanish colonies. Looking at the African American individual (ASW9) that is sister taxa with the Peruvian individual (PEL8) in the Native American clade (Figure 9), the ASW5 that is sister taxa to PUR4 (Figure 11A), and the African American (ASW9) that is clustered within many Native Americans (Figure 12A), significant gene flow was expected between both the Peruvian and Puerto Rican populations with the African American populations because their clustering can only be based on gene flow not recent common ancestry. Although ABBABABA tests (Table 1, Set 3) found significant gene flow between African Americans and some Native American populations, it was much lower than the gene flow estimated between Europeans to either, thus promoting the idea that sexual relationships between Native Americans and African slaves was not as common as sexual relations they had with Europeans.

Finally, my 3rd hypothesis which states that gene flow could have occurred between African slaves and both their European slave masters and Native American counterparts is accepted. In summary, the gene flow pattern during the time of slavery was triangular in fashion (Figure 14) among the European slave masters, African American slaves, and Native Americans, but unidirectional from European slave masters to the African slaves and Native Americans, and bidirectional between the African slaves and Native American populations.

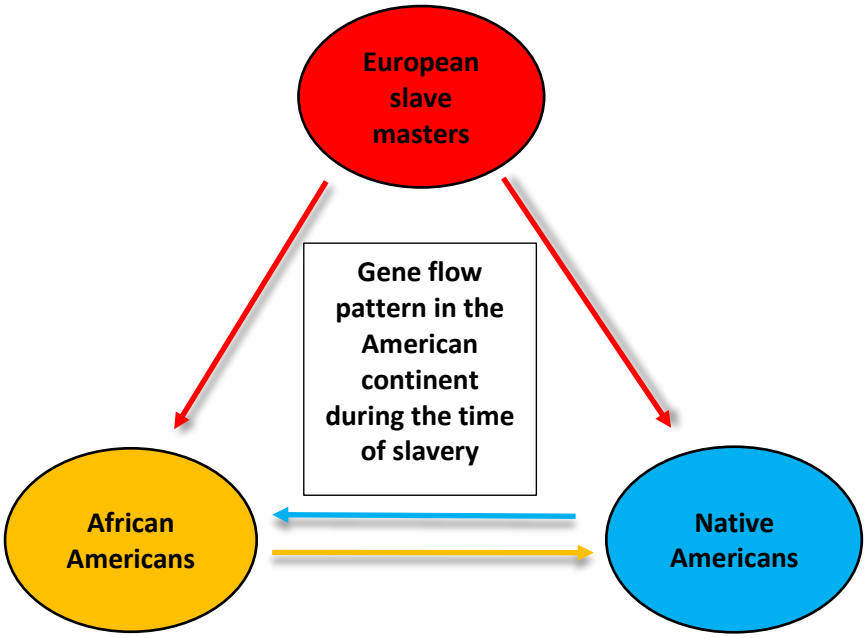

Figure 14. Inferred gene flow pattern among the European slave masters, African slaves, and Native American populations in the American continent during the time of slavery. The coloured arrowed lines indicate gene flow direction between the populations.

5. Conclusion

Although the colonization of the Americas by Europeans, their treatment of the Native Americans, and the whole dark episode of African slavery were a combined human tragedy, significant genetic interactions among the three population groups have resulted in an ethnically and genetically diverse human population of the American continent. Although it was expected to find that European males produced offspring that were absorbed into the African American and Native American populations, it was surprising to find the same true also for European women. This study is evidence that the Y and mitochondrial chromosomes genetic markers that were used in the past genetic studies could not infer a wider human evolution story, although they are no less informative than the much larger autosomal chromosomes because they provide sex-specific information that can be linked to cultural practices, and that their integration with whole genomes provide a much wider resolution. The emergence of Whole Genome Sequences (WGS) to infer human populations evolution and structure has proven that there is still a lot to be learnt from human DNA. Unlike with the past use of small microsatellite genetic markers, deep investigation of the forces of evolution operating among individuals, populations and species is now achievable. But scientists and software developers still owe more effort in production of software and biological methods that will liberate the full advantage

of high coverage WGS and millions of samples. Any human population inference is only as good as the quantity and quality of the genetic data.

6. References

1000 Genomes Project Consortium, 2015. A global reference for human genetic variation. *Nature*, *526*(7571), pp.68-74.

Adcock, G.J., Dennis, E.S., Easteal, S., Huttley, G.A., Jermiin, L.S., Peacock, W.J. and Thorne, A., 2001. Mitochondrial DNA sequences in ancient Australians: implications for modern human origins. *Proceedings of the National Academy of Sciences*, *98*(2), pp.537-542.

Aguirre, E. and Carbonell, E., 2001. Early human expansions into Eurasia: the Atapuerca evidence. *Quaternary International*, *75*(1), pp.11-18.

Alexander, D.H., Novembre, J. and Lange, K., 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), pp.1655-1664.

Baker, J.L., Rotimi, C.N. and Shriner, D., 2017. Human ancestry correlates with language and reveals that race is not an objective genomic classifier. *Scientific Reports*, *7*(1), pp.1-10.

Bandelt, H.J., Forster, P. and Röhl, A., 1999. Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, *16*(1), pp.37-48.

Baute, G.J., Owens, G.L., Bock, D.G. and Rieseberg, L.H., 2016. Genome-wide genotyping-by-sequencing data provide a high-resolution view of wild Helianthus diversity, genetic structure, and interspecies gene flow. *American Journal of Botany*, *103*(12), pp.2170-2177.

Beltrame, M.H., Rubel, M.A. and Tishkoff, S.A., 2016. Inferences of African evolutionary history from genomic data. *Current Opinion in Genetics and Development*, *41*, pp.159-166.

Bentley, D.R., Balasubramanian, S., Swerdlow, H.P., Smith, G.P., Milton, J., Brown, C.G., Hall, K.P., Evers, D.J., Barnes, C.L., Bignell, H.R. and Boutell, J.M., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, *456*(7218), pp.53-59.

Berlin, I., 1980. Time, space, and the evolution of Afro-American society on British mainland North America. *The American Historical Review*, *85*(1), pp.44-78.

Bhatia, G., Patterson, N., Pasaniuc, B., Zaitlen, N., Genovese, G., Pollack, S., Mallick, S., Myers, S., Tandon, A., Spencer, C. and Palmer, C.D., 2011. Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection. *The American Journal of Human Genetics*, *89*(3), pp.368-381.

Blundell, S., 2016. *The Origins of Civilization in Greek and Roman Thought (Routledge Revivals)*. Routledge.

Bonatto, S.L. and Salzano, F.M., 1997. Diversity and age of the four major mtDNA haplogroups, and their implications for the peopling of the New World. *The American Journal of Human Genetics*, *61*(6), pp.1413-1423.

Boyeldieu, P., Dimmendaal, GJ., Fleisch, A., Frajzyngier, Z., Güldemann, T., Nougayrol, P., Porkhomovsky, V., Vossen, R., 2008. Problems of linguistic-historical reconstruction in Africa (SUGIA Sprache und Geschichte in Afrika). 1st edition edn: R Köppe.

Bryc, K., Auton, A., Nelson, M.R., Oksenberg, J.R., Hauser, S.L., Williams, S., Froment, A., Bodo J.M., Wambebe, C., Tishkoff, S.A. and Bustamante, C.D., 2010. Genome-wide Patterns of Population Structure and Admixture in West Africans and African Americans. The Proceedings of the National Academy of Sciences. 107(2). pp. 786-791.

Bulley, C.A., Osei-Bonsu, N. and Rasaq, H.A., 2017. Attributes of leadership effectiveness in West Africa. *AIB Insights*, *17*(1), p.11.

Buvit, I., Izuho, M., Terry, K., Konstantinov, M.V. and Konstantinov, A.V., 2016. Radiocarbon dates, microblades and late Pleistocene human migrations in the Transbaikal, Russia and the Paleo-Sakhalin-Hokkaido-Kuril Peninsula. *Quaternary International*, *425*, pp.100-119.

Byrska-Bishop, M., Evani, U.S., Zhao, X., Basile, A.O., Abel, H.J., Regier, A.A., Corvelo, A., Clarke, W.E., Musunuri, R., Nagulapalli, K. and Fairley, S., 2021. High coverage whole genome sequencing of the expanded 1000 Genomes Project Cohort Including 602 Trios. Available at SSRN: https://ssrn.com/abstract=3967671 or http://dx.doi.org/10.2139/ssrn.3967671.

Campbell, M.C. and Tishkoff, S.A., 2008. African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. Annual Review of Genomics and Human Genetics, 9, pp.403-433.

Campbell, M.C., Hirbo, J.B., Townsend, J.P. and Tishkoff, S.A., 2014. The peopling of the African continent and the diaspora into the new world. *Current Opinion in Genetics and Development*, *29*, pp.120-132.

Cann, R.L., Stoneking, M. and Wilson, A.C., 1987. Mitochondrial DNA and human evolution. *Nature*, *325*(6099), pp.31-36.

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

Cavalli-Sforza, L.L., Piazza, A., Menozzi, P. and Mountain, J., 1988. Reconstruction of human evolution: bringing together genetic, archaeological, and linguistic data. *Proceedings of the National Academy of Sciences*, *85*(16), pp.6002-6006.

Clarke, R.J., 2000. Out of Africa and back again. *International Journal of Anthropology*, *15*(3-4), pp.185-189.

Clayborne, C. ed. 2011. *The Struggle for Freedom* . Prentice Hall, 38.

Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T. and McVean, G., 2011. The variant call format and VCFtools. *Bioinformatics*, *27*(15), pp.2156-2158.

DeCorse, C.R., 1989. *An archaeological study of Elmina, Ghana: trade and culture change on the Gold Coast between the fifteenth and nineteenth centuries* (Doctoral dissertation, University of California, Los Angeles).

DeCorse, C.R., 1991. West African archaeology and the Atlantic slave trade. *Slavery and Abolition*, *12*(2), pp.92-96.

Derricourt, R., 2005. Getting "Out of Africa": sea crossings, land crossings and culture in the hominin migrations. *Journal of World Prehistory*, *19*(2), pp.119-132.

Dingemanse, M., 2004. Map of the Distribution of African Language Families and some Major African Languages. *Creative Commons Attribution 2.5 License.*

Dobzhansky, T., 2013. Nothing in biology makes sense except in the light of evolution. *The american biology teacher*, *75*(2), pp.87-91.

Durand, E.Y., Patterson, N., Reich, D. and Slatkin, M., 2011. Testing for ancient admixture between closely related populations. *Molecular Biology and Evolution*, *28*(8), pp.2239-2252.

Ehret, C. and Posnansky, M. eds., 1982. *The archaeological and linguistic reconstruction of African history*. Univ of California Press.

Ellegren, H., 2009. The different levels of genetic diversity in sex chromosomes and autosomes. *Trends in Genetics*, *25*(6), pp.278-284.

Eltis, D and Richardson, D. 1997. West Africa and the Transatlantic Slave Trade: New Evidence of Long-run Trends. Slavery and Abolition, 18 (1), pp.16-35.

Eltis, D. 1999. *The Trans-Atlantic Slave Trade*. Cambridge University Press. New York.

Eltis, D. and Engerman, S.L., 1993. Fluctuations in Sex and Age Ratios in the Transatlantic SlaveTrade, 1663-1864. Blackwell Publishers. The Economic History Review. New Series, 46 (2), pp. 308-323.

Fage, J.D., 1989. African Societies and the Atlantic Slave Trade. Oxford University Press. No. 125, pp. 97-115.

Fan, S., Kelly, D.E., Beltrame, M.H., Hansen, M.E., Mallick, S., Ranciaro, A., Hirbo, J., Thompson, S., Beggs, W., Nyambo, T. and Omar, S.A., 2019. African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology*, *20*(1), pp.1-14.

Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *evolution*, *39*(4), pp.783-791.

Fisher, L.D., 2017. "Why shall wee have peace to bee made slaves": Indian Surrenderers During and After King Philip's War. *Ethnohistory*, *64*(1), pp.91-114.

Fix, A.G., 1995. Malayan paleosociology: Implications for patterns of genetic variation among the Orang Asli. *American Anthropologist*, *97*(2), pp.313-323.

Gannon, M.J. and Pillai, R., 2015. *Understanding global cultures: Metaphorical journeys through 34 nations, clusters of nations, continents, and diversity*. Sage Publications.

Gates, H.L., 2014. *How Many Slaves Landed in the US*? *The Root*. Retrieved 8 July 2018.

Gemery, H.A. and Hogendorn, J.S., 1974. *The Atlantic Slave Trade: A Tentative Economic Model*. Cambridge University Press. The Journal of African History, Vol. 15, No. 2, pp. 223-246.

Gibbs, R.A., Belmont, J.W., Hardenbol, P., Willis, T.D., Yu, F.L., Yang, H.M., Ch'ang, L.Y., Huang, W., Liu, B., Shen, Y. and Tam, P.K.H., 2003. The international HapMap project.

Green, R.E., Krause, J., Briggs, A.W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H.Y. and Hansen, N.F., 2010. A draft sequence of the Neandertal genome. *Science*, *328*(5979), pp.710-722.

Greenberg, J.H., 1963. The languages of Africa. *International journal of American linguistics*.

Gurdasani, D., Carstensen, T., Tekola-Ayele, F., Pagani, L., Tachmazidou, I., Hatzikotoulas, K., Karthikeyan, S., Iles, L., Pollard, M.O., Choudhury, A. and Ritchie, G.R., 2015. The African genome variation project shapes medical genetics in Africa. *Nature*, *517*(7534), pp.327-332.

Haring, C.H., 1985. *The Spanish Empire in America*. Harvest Books.

Heine, B. and Nurse, D. eds., 2000. *African languages: An introduction*. Cambridge University Press.

Henn, B.M., Cavalli-Sforza, L.L. and Feldman, M.W., 2012. The great human expansion. *Proceedings of the National Academy of Sciences*, *109*(44), pp.17758-17764.

Henn, B.M., Gignoux, C.R., Jobin, M., Granka, J.M., Macpherson, J.M., Kidd, J.M., Rodríguez-Botigué, L., Ramachandran, S., Hon, L., Brisbin, A. and Lin, A.A., 2011. Hunter-gatherer genomic diversity suggests a southern African origin for modern humans. *Proceedings of the National Academy of Sciences*, *108*(13), pp.5154-5162.

Hoang, D.T., Chernomor, O., Von Haeseler, A., Minh, B.Q. and Vinh, L.S., 2018. UFBoot2: improving the ultrafast bootstrap approximation. *Molecular biology and evolution*, *35*(2), pp.518-522.

Huson, D.H. and Bryant, D., 2006. Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, *23*(2), pp.254-267.

Jin, W., Xu, S., Wang, H., Yu, Y., Shen, Y., Wu, B and Jin, L. 2012. *Genome-wide detection of natural selection in African Americans pre- and post-admixture*. Cold Spring Harbor Laboratory Press.

Jobling, M.A., Hurles, M. and Tyler-Smith, C., 2019. *Human evolutionary genetics: origins, peoples and disease*. Garland Science.

Jolliffe, I. and Lovric, M., 2011. International encyclopedia of statistical science. In *Principal Component Analysis* (pp. 1094-1096). Berlin, Heidelberg: Springer Berlin Heidelberg.

Jorde, L.B., Watkins, W.S., Bamshad, M.J., Dixon, M.E., Ricker, C.E., Seielstad, M.T. and Batzer, M.A., 2000. The distribution of human genetic diversity: a comparison of mitochondrial, autosomal, and Y chromosome data. *The American Journal of Human Genetics*, *66*(3), pp.979-988.

Kaessmann, H., Heißig, F., Haeseler, A.V. and Pääbo, S., 1999. DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nature Genetics*, 22(1), pp.78-81.

Kalyaanamoorthy, S., Minh, B.Q., Wong, T.K., Von Haeseler, A. and Jermiin, L.S., 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature methods*, *14*(6), pp.587-589.

Krings, M., Capelli, C., Tschentscher, F., Geisert, H., Meyer, S., Von Haeseler, A., Grossschmidt, K., Possnert, G., Paunovic, M. and Pääbo, S., 2000. A view of Neandertal genetic diversity. *Nature Genetics*, *26*(2), pp.144-146.

Lavretsky, P., Engilis Jr, A., Eadie, J.M. and Peters, J.L., 2015. Genetic admixture supports an ancient hybrid origin of the endangered Hawaiian duck. *Journal of Evolutionary Biology*, *28*(5), pp.1005-1015.

Leigh, J.W. and Bryant, D., 2015. popart: full-feature software for haplotype network construction. *Methods in Ecology and Evolution*, *6*(9), pp.1110-1116.

Li, A., 2020. Resources Feature: Slave Voyages Website Releases New and Updated Lesson Plans. *Emory Center for Digital Scholarship*.

Li, H. and Durbin, R., 2010. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, *26*(5), pp.589-595.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The sequence alignment/map format and SAMtools. *Bioinformatics*, *25*(16), pp.2078-2079.

Linck, E. and Battey, C.J., 2019. Minor allele frequency thresholds strongly affect population structure inference with genomic data sets. *Molecular Ecology Resources*, *19*(3), pp.639-647.

Liu, H., Prugnolle, F., Manica, A. and Balloux, F., 2006. A geographically explicit genetic model of worldwide human-settlement history. *The American Journal of Human Genetics*, *79*(2), pp.230-237.

Llamas, B., Fehren-Schmitz, L., Valverde, G., Soubrier, J., Mallick, S., Rohland, N., Nordenfelt, S., Valdiosera, C., Richards, S.M., Rohrlach, A. and Romero, M.I.B., 2016. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances*, *2*(4), p.e1501385.

Lovejoy, P.E., 2011. *Transformations in slavery: a history of slavery in Africa* (Vol. 117). Cambridge University Press.

Lovell, A., Moreau, C., Yotova, V., Xiao, F., Bourgeois, S., Gehl, D., Bertranpetit, J., Schurr, E. and Labuda, D., 2005. Ethiopia: between sub-Saharan Africa and western Eurasia. *Annals of Human Genetics*, *69*(3), pp.275-287.

Lowy-Gallego, E., Fairley, S., Zheng-Bradley, X., Ruffier, M., Clarke, L., Flicek, P. and 1000 Genomes Project Consortium, 2019. Variant calling on the GRCh38 assembly with the data from phase three of the 1000 Genomes Project. *Wellcome Open Research*, *4*.

MacArthur, R.H. and Wilson, E.O., 2016. *The theory of island biogeography*. Princeton university press.

Malinsky, M., Matschiner, M. and Svardal, H., 2021. Dsuite-Fast D-statistics and related admixture evidence from VCF files. *Molecular Ecology Resources*, *21*(2), pp.584-595.

Mallick, S., Li, H., Lipson, M., Mathieson, I., Gymrek, M., Racimo, F., Zhao, M., Chennagiri, N., Nordenfelt, S., Tandon, A. and Skoglund, P., 2016. The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature*, *538*(7624), pp.201-206.

McCartney, M.W., 2003. A study of the Africans and African Americans on Jamestown Island and at Green Spring, 1619-1803.

McMillan, B. ed., 2002. *Captive passage: the transatlantic slave trade and the making of the Americas*. Smithsonian Institution Press.

Mellars, P., 2006. Going east: new genetic and archaeological perspectives on the modern human colonization of Eurasia. *Science*, *313*(5788), pp.796-800.

Mendes, M., Alvim, I., Borda, V. and Tarazona-Santos, E., 2020. The history behind the mosaic of the Americas. *Current Opinion in Genetics and Development*, *62*, pp.72-77.

Moodley, Y., Westbury, M.V., Russo, I.R.M., Gopalakrishnan, S., Rakotoarivelo, A., Olsen, R.A., Prost, S., Tunstall, T., Ryder, O.A., Dalén, L. and Bruford, M.W., 2020. Interspecific Gene Flow and the Evolution of Specialization in Black and White Rhinoceros. *Molecular Biology and Evolution*, *37*(11), pp.3105-3117.

Moreno-Mayar, J.V., Potter, B.A., Vinner, L., Steinrücken, M., Rasmussen, S., Terhorst, J., Kamm, J.A., Albrechtsen, A., Malaspinas, A.S., Sikora, M. and Reuther, J.D., 2018. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. *Nature*, 553(7687), pp.203-207.

Morgan, K., 2007. *Slavery and the British empire: from Africa to America*. Oxford University Press on Demand.

Nguyen, L.T., Schmidt, H.A., Von Haeseler, A. and Minh, B.Q., 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), pp.268-274.

Nunn, N., 2008. The long-term effects of Africa's slave trades. *The Quarterly Journal of Economics*, *123*(1), pp.139-176.

Oldfield, J.R., 2012. Repairing historical wrongs: Public history and transatlantic slavery. *Social and Legal Studies*, *21*(2), pp.243-255.

Ortiz, E.M., 2019. vcf2phylip v2. 0: convert a VCF matrix into several matrix formats for phylogenetic analysis. *URL https://doi org/105281/zenodo*, *2540861*.

Ostler, J., 2019. *Surviving Genocide: Native Nations and the United States from the American Revolution to Bleeding Kansas*. Yale University Press.

Ovchinnikov, I.V., Götherström, A., Romanova, G.P., Kharitonov, V.M., Liden, K. and Goodwin, W., 2000. Molecular analysis of Neanderthal DNA from the northern Caucasus. *Nature*, *404*(6777), pp.490-493.

Pérez-Reche, F.J., Rotariu, O., Lopes, B.S., Forbes, K.J. and Strachan, N.J., 2020. Mining whole genome sequence data to efficiently attribute individuals to source populations. *Scientific Reports*, *10*(1), pp.1-16.

Potter, B.A., Baichtal, J.F., Beaudoin, A.B., Fehren-Schmitz, L., Haynes, C.V., Holliday, V.T., Holmes, C.E., Ives, J.W., Kelly, R.L., Llamas, B. and Malhi, R.S., 2018. Current evidence allows multiple models for the peopling of the Americas. *Science Advances*, *4*(8), p.eaat5473.

Price, T.D., 2015. *Ancient Scandinavia: an archaeological history from the first humans to the Vikings*. Oxford University Press, USA.

Prugnolle, F., Manica, A. and Balloux, F., 2005. Geography predicts neutral genetic diversity of human populations. *Current Biology*, *15*(5), pp.R159-R160.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A., Bender, D., Maller, J., Sklar, P., De Bakker, P.I., Daly, M.J. and Sham, P.C., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, *81*(3), pp.559-575.

R Core Team., 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Raghavan, M., DeGiorgio, M., Albrechtsen, A., Moltke, I., Skoglund, P., Korneliussen, T.S., Grønnow, B., Appelt, M., Gulløv, H.C., Friesen, T.M. and Fitzhugh, W., 2014. The genetic prehistory of the New World Arctic. *Science*, *345*(6200).

Rambaut, A. 2016. *FigTree v1. 4.0, a graphical viewer of phylogenetic trees*. Edinburgh: University of Edinburgh.

Reed, F.A. and Tishkoff, S.A., 2006. African human diversity, origins and migrations. *Current Opinion in Genetics and Development*, *16*(6), pp.597-605.

Relethford, J.H., 2001. Ancient DNA and the origin of modern humans. *Proceedings of the National Academy of Sciences*, *98*(2), pp.390-391.

Reséndez, A., 2016. *The other slavery: The uncovered story of Indian enslavement in America*. Houghton Mifflin Harcourt.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A. and Feldman, M.W., 2002. Genetic structure of human populations. *Science*, *298*(5602), pp.2381-2385.

Rotimi, C.N., Bentley, A.R., Doumatey, A.P., Chen, G., Shriner, D. and Adeyemo, A., 2017. The genomic landscape of African populations in health and disease. *Human Molecular Genetics*, *26*(R2), pp.R225-R236.

Rotimi, C.N., Tekola-Ayele, F., Baker, J.L. and Shriner, D., 2016. The African diaspora: history, adaptation and health. *Current Opinion in Genetics and Development*, 41, pp.77-84.

Salas, A., Richards, M., Lareu, M.V., Scozzari, R., Coppa, A., Torroni, A., Macaulay, V. and Carracedo, Á., 2004. The African diaspora: mitochondrial DNA and the Atlantic slave trade. *The American Journal of Human Genetics*, *74*(3), pp.454-465.

Salas, A., Richards, M., Lareu, M.V., Sobrino, B., Silva, S., Matamoros, M., Macaulay, V. and Carracedo, Á., 2005. Shipwrecks and founder effects: divergent demographic histories reflected in Caribbean mtDNA. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *128*(4), pp.855-860.

Sands, B.E., 1998. *Eastern and southern African Khoisan: evaluating claims of distant linguistic relationships* (Vol. 14). R Koppe.

Scheinfeldt, L.B. and Tishkoff, S.A., 2013. Recent human adaptation: genomic approaches, interpretation and insights. *Nature Reviews Genetics*, 14(10), pp.692-702.

Scheinfeldt, L.B., Soi, S. and Tishkoff, S.A., 2010. Working toward a synthesis of archaeological, linguistic, and genetic data for inferring African population history. *Proceedings of the National Academy of Sciences*, 107(Supplement 2), pp.8931-8938.

Scobie, E., 1992. The Moors and Portugal's Global Expansion. *Golden age of the Moor*, *11*, pp.333-340.

Semino, O., Santachiara-Benerecetti, A.S., Falaschi, F., Cavalli-Sforza, L.L. and Underhill, P.A., 2002. Ethiopians and Khoisan share the deepest clades of the human Y-chromosome phylogeny. *The American Journal of Human Genetics*, 70(1), pp.265-268.

Shriner, D., Tekola-Ayele, F., Adeyemo, A. and Rotimi, C.N., 2014. Genome-wide genotype and sequence-based reconstruction of the 140,000 year history of modern human ancestry. *Scientific Reports*, *4*(1), pp.1-9.

Simmonds, D., 1973. A note on the excavations in Cape Coast Castle. *Transactions of the Historical Society of Ghana*, *14*(2), pp.267-269.

Skotte, L., Korneliussen, T.S. and Albrechtsen, A., 2013. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, *195*(3), pp.693-702.

Smith, F.H., Falsetti, A.B. and Donnelly, S.M., 1989. Modern human origins. *American Journal of Physical Anthropology*, *32*(S10), pp.35-68.

Soares-Souza, G., Borda, V., Kehdy, F. and Tarazona-Santos, E., 2018. Admixture, genetics and complex diseases in Latin Americans and US Hispanics. *Current Genetic Medicine Reports*, *6*(4), pp.208-223.

Stark, D.M., 2009. A New Look at the African Slave Trade in Puerto Rico Through the Use of Parish Registers: 1660–1815. *Slavery and Abolition*, *30*(4), pp.491-520.

Stoneking, M. and Soodyall, H., 1996. Human evolution and the mitochondrial genome. *Current Opinion in Genetics and Development*, *6*(6), pp.731-736.

Sylvester, M., 1998. The African American: A journey from slavery to freedom. *An online exhibit at Long Island University. Retrieved May*, *18*, p.2004.

Szathmary, E.J., 1993. mtDNA and the peopling of the Americas. *American Journal of Human Genetics*, *53*(4), p.793.

Tamm, E., Kivisild, T., Reidla, M., Metspalu, M., Smith, D.G., Mulligan, C.J., Bravi, C.M., Rickards, O., Martinez-Labarga, C., Khusnutdinova, E.K. and Fedorova, S.A., 2007. Beringian standstill and spread of Native American founders. *PloS One*, *2*(9), p.e829.

Templeton, A., 2002. Out of Africa again and again. *Nature*, *416*(6876), pp.45-51.

Templeton, A.R., Hedges, S.B., Kumar, S., Tamura, K. and Stoneking, M., 1992. Human origins and analysis of mitochondrial DNA sequences. *Science*, *255*(5045), pp.737-739.

Thomas, H. 1999. *The Slave Trade. the story of the Atlantic Slave Trade*. New York. Simon and Schuster. pp. 1440–1870.

Tishkoff, S.A., Reed, F.A., Friedlaender, F.R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J.B., Awomoyi, A.A., Bodo, J.M., Doumbo, O. and Ibrahim, M., 2009. The genetic structure and history of Africans and African Americans. *Science*, *324*(5930), pp.1035-1044.

Vishnoi, A., Roy, R., Prasad, H.K. and Bhattacharya, A., 2010. Anchor-based whole genome phylogeny (ABWGP): a tool for inferring evolutionary relationship among closely related microorganims. *PLoS One*, *5*(11), p.e14159.

Vitti, J.J., Cho, M.K., Tishkoff, S.A. and Sabeti, P.C., 2012. Human evolutionary genomics: ethical and interpretive issues. *Trends in Genetics*, *28*(3), pp.137-145.

Wallace, D.C., Brown, M.D. and Lott, M.T., 1999. Mitochondrial DNA variation in human evolution and disease. *Gene*, *238*(1), pp.211-230.

Wang, J., 2018. Effects of sampling close relatives on some elementary population genetics analyses. *Molecular Ecology Resources*, *18*(1), pp.41-54.

Weidenreich, F., 1946. *Apes, giants and man* (pp. 83-4). Chicago: University of Chicago Press.

Wheeler, D.A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y.J., Makhijani, V., Roth, G.T. and Gomes, X., 2008. The complete genome of an individual by massively parallel DNA sequencing. *Nature*, *452*(7189), pp.872-876.

Wilson, E.O. and MacArthur, R.H., 1967. The theory of island biogeography. *Princeton University Press,* 1.

Wolf, Y.I., Rogozin, I.B., Grishin, N.V. and Koonin, E.V., 2002. Genome trees and the tree of life. *TRENDS in Genetics*, *18*(9), pp.472-479.

Wolpoff, M.H., Hawks, J. and Caspari, R., 2000. Multiregional, not multiple origins. *American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists*, *112*(1), pp.129-136.

Wright, S., 1943. Isolation by distance. *Genetics*, *28*(2), p.114.

Yaz., 2001. "PUERTO RICO" GRILLA'S HOMEPAGE". *Angelfire.com*.

Zakharia, F., Basu, A., Absher, D., Assimes, T.L., Go, A.S., Hlatky, M.A., Iribarren, C., Knowles, J.W., Li, J., Narasimhan, B. and Sidney, S., 2009. Characterizing the admixed African ancestry of African Americans. *Genome Biology*, *10*(12), pp.1-11.

7. Supplementary

Appendix A. Showing the script steps used for plotting the PCA in R studio

```
# read in the eigenvectors, produced in PLINK
1. options(scipen=100, digits=3)
2. eigenvec <- read.table('plink.eigenvec', header = FALSE, skip=0, sep = ' ')
3. rownames(eigenvec) <- eigenvec[,2]
4. eigenvec <- eigenvec[,3:ncol(eigenvec)]
5. colnames(eigenvec) <- paste('Principal Component ', c(1:20), sep = '')
# read in the PED data
6. PED <- read.table('Samples.ped', header = TRUE, skip = 0, sep = '\t')
7. PED <- PED[which(PED$Individual.ID %in% rownames(eigenvec)), ]
8. PED <- PED[match(rownames(eigenvec), PED$Individual.ID),]
9. all(PED$Individual.ID == rownames(eigenvec)) == TRUE
[1] TRUE
# set colours
10. require('RColorBrewer')
11. PED$Population <- factor(PED$Population, levels=c("ESN","YRI","LWK","MSL","GWD","ACB","ASW","MXL","PEL","PUR","CLM","CEU","IBS","GBR"))
12. col <- colorRampPalette(c("black","grey","yellow","brown","orange","navyblue","lightskyblue","darkgreen","lightpink","green","deeppink3","burlywood","blueviolet","red"))(length(unique(PED$Population)))[factor(PED$Population)]
# generate PCA bi-plots
13. project.pca <- eigenvec
14. summary(project.pca)
15. par(mar = c(5,5,5,5), cex = 2.75, cex.main = 2.75, cex.axis = 2.75, cex.lab = 2.0, mfrow = c(1,2))
#plot PCA
16. plot(project.pca[,1], project.pca[,2],type="n",
  main="A",
  adj=0.5,
  xlab="1st component",
  ylab="2nd component",
  font=2,
  font.lab=2)
17. points(project.pca[,1], project.pca[,2], col = col, pch = 20, cex = 1)
#Set legends for the plots
18. legend('bottomright', bty = 'n', cex = 0.65, title = '', c('ESN','YRI','LWK','MSL','GWD','ACB','ASW','MXL','PEL','PUR','CLM','CEU','IBS','GBR'), fill = c('black','grey','yellow','brown','orange','navyblue','lightskyblue','darkgreen','lightpink','green','deeppink3','burlywood','blueviolet','red'))
```

Appendix B. Showing the CHPC server job submission script for the phylogenetic trees of autosomal and X chromosome datasets.

```
#!/bin/bash
### Select  nodes 1 with 56 CPUs
#PBS -l select=1:ncpus=56:mpiprocs=56:nodetype=haswell_fat
```

```
### Job Name
#PBS -N iqtree_chrX
### Project code
#PBS -P CBBI0911
#PBS -l walltime=48:00:00
#PBS -q bigmem
#PBS -W group_list=bigmemq
#PBS -o /home/mndou/lustre/iqtree/chrX/log.out
#PBS -e /home/mndou/lustre/iqtree/chrX/log.err
### Send email on abort, begin and end
#PBS -m abe
### Specify mail recipient
#PBS -M ndoumannda.mn@gmail.com
module add chpc/BIOMODULES
module add iqtree/1.6.6
NP=`cat ${PBS_NODEFILE} | wc -l`
### Run the executable
EXE="iqtree"
ARGS="-s Renamed_Autosomal_tree_samples_OG3.min4.phy -bb 1000 -nt 56"
cd /home/mndou/lustre/iqtree/chrX
${EXE} ${ARGS}
```

Appendix C. Showing the genotype stats of all the 4 datasets before filtering

	Dataset type			
	Autosomal chromosomes	**X chromosome**	**Y chromosome**	**Mitochondrial chromosome**
Number of samples:	728	728	364	728
Number of records:	123574606	4960150	199843	3892
Number of no-ALTs:	0	0	0	0
Number of SNPs:	111860496	4468198	176147	3858
Number of MNPs:	0	0	0	88
Number of indels:	13177046	546319	25276	34
Number of others:	2079960	85853	1987	0
Number of multiallelic sites:	9659825	364364	14100	275
Number of multiallelic SNP sites:	468477	152142	5514	184

Appendix D. Showing the heterozygosity and nucleotide genomic diversity for the 14 African, American, and European populations.

Population	Average Heterozygosity	Heterozygosity %	Average pi	Nucleotide diversity %
ESN	2610238.02	7.853854351	0.000929293	7.859465093
YRI	2619685.13	7.882279447	0.000930386	7.868711907
LWK	2596079.85	7.811254316	0.000922085	7.798509464
MSL	2614175.31	7.865701121	0.000930391	7.868755937
GWD	2595215.98	7.808655062	0.000926559	7.836348158
ACB	2632857.48	7.921913261	0.000937958	7.932752032
ASW	2627049.12	7.904436673	0.000937543	7.929243242
MXL	2122995.96	6.387808677	0.000761206	6.437874551
PEL	1993984.81	5.999631505	0.000710348	6.007745287
PUR	2270453.83	6.831489517	0.000805972	6.816486553
CLM	2216131.27	6.668040264	0.000789188	6.674534544
CEU	2107883.1	6.342336102	0.000745006	6.30086297
IBS	2124016.69	6.390879918	0.000753548	6.373113979
GBR	2104354.75	6.331719784	0.000744383	6.295596285

Appendix E. Showing the significant and non-significant ABBABABA tests among all the 14 populations.

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
ASW	CEU	ACB	0.140641	22.0984	0	242852	219673	165502
ASW	CLM	ACB	0.115833	22.8566	0	238114	216103	171236
ACB	ASW	ESN	0.016351	9.40739	0	208474	200047	193610
ASW	GBR	ACB	0.140319	22.25	0	242863	219669	165608
ACB	ASW	GWD	0.018116	10.5515	0	208016	201191	194031
ASW	IBS	ACB	0.136259	23.2165	0	241691	219008	166482
ACB	ASW	LWK	0.019609	11.3355	0	209185	201423	193676
ACB	ASW	MSL	0.016683	9.80317	0	212681	198608	192090
ASW	MXL	ACB	0.115542	20.6804	0	238757	215936	171205
ASW	PEL	ACB	0.109217	17.8494	0	238196	215028	172683
ASW	PUR	ACB	0.111826	23.0064	0	236640	215557	172196
ACB	ASW	YRI	0.016329	9.47456	0	208061	200212	193778
CLM	CEU	ACB	0.029416	14.4162	0	308712	162811	153506
ACB	CEU	ESN	0.124067	17.6102	0	239552	218820	170516
GBR	CEU	ACB	0.000364	0.308422	0.757761	321738	150919	150809
ACB	CEU	GWD	0.133026	18.5736	0	237770	221664	169614
IBS	CEU	ACB	0.005402	4.25514	2.09E-05	318732	153109	151464
ACB	CEU	LWK	0.141622	19.9842	0	237702	223465	168022

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
ACB	CEU	MSL	0.125522	17.515	0	243759	217512	168997
MXL	CEU	ACB	0.029764	10.875	0	308603	163306	153866
PEL	CEU	ACB	0.036794	9.03171	2.17E-19	305749	166628	154802
PUR	CEU	ACB	0.033876	14.9523	0	305643	164970	154160
ACB	CEU	YRI	0.124702	17.785	0	238976	219110	170522
ACB	CLM	ESN	0.104563	17.6092	0	234707	215842	174976
CLM	GBR	ACB	0.029073	13.6116	0	308781	162738	153542
ACB	CLM	GWD	0.112712	18.7979	0	233078	218490	174226
CLM	IBS	ACB	0.024087	13.0233	0	306501	162830	155171
ACB	CLM	LWK	0.12056	20.1202	0	233122	220110	172747
ACB	CLM	MSL	0.106134	17.6256	0	238853	214573	173397
MXL	CLM	ACB	0.000425	0.246091	0.805612	303488	159454	159319
PEL	CLM	ACB	0.007956	2.4845	0.012974	304199	159747	157226
PUR	CLM	ACB	0.004595	4.10377	4.06E-05	296416	164603	163097
ACB	CLM	YRI	0.105155	17.7727	0	234149	216129	175000
ACB	GBR	ESN	0.123674	17.5504	0	239568	218808	170643
GWD	ACB	ESN	0.012006	9.22225	0	200444	199565	194830
ACB	IBS	ESN	0.120946	18.3507	0	238499	218194	171109
ESN	ACB	LWK	0.014197	11.5378	0	200993	200348	194739
ESN	ACB	MSL	0.013427	10.9011	0	203778	198051	192803
ACB	MXL	ESN	0.104082	16.3904	0	234924	216065	175329
ACB	PEL	ESN	0.098813	14.8258	0	233982	215539	176773
ACB	PUR	ESN	0.102211	17.6459	0	233502	215187	175277
YRI	ACB	ESN	0.009491	7.08869	1.35E-12	199238	198657	194921
ACB	GBR	GWD	0.132555	18.6065	0	237798	221631	169752
IBS	GBR	ACB	0.005043	4.30219	1.69E-05	318832	153019	151483
ACB	GBR	LWK	0.141235	19.8768	0	237706	223440	168136
ACB	GBR	MSL	0.125241	17.5726	0	243747	217515	169095
MXL	GBR	ACB	0.029425	10.612	0	308737	163216	153886
PEL	GBR	ACB	0.036457	8.65162	5.10E-18	305884	166554	154837
PUR	GBR	ACB	0.033536	16.0553	0	305745	164896	154195
ACB	GBR	YRI	0.124259	17.6714	0	238999	219082	170654
ACB	IBS	GWD	0.13003	19.441	0	236699	221062	170188
GWD	ACB	LWK	0.01241	10.1839	0	201812	200289	195378
GWD	ACB	MSL	0.007182	5.85592	4.74E-09	203687	197082	194271
ACB	MXL	GWD	0.112094	17.5872	0	233312	218680	174596
ACB	PEL	GWD	0.106366	15.8015	0	232475	218077	176145
ACB	PUR	GWD	0.110298	18.9093	0	231871	217798	174526
GWD	ACB	YRI	0.011409	8.71257	2.93E-18	199920	199622	195119
ACB	IBS	LWK	0.138246	20.7025	0	236689	222766	168654
ACB	IBS	MSL	0.122555	18.2883	0	242671	216918	169554
MXL	IBS	ACB	0.024408	9.70673	0	306105	163583	155788
PEL	IBS	ACB	0.031416	7.55433	4.21E-14	302981	167133	156952
PUR	IBS	ACB	0.028597	16.1683	0	303660	164835	155669

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
ACB	IBS	YRI	0.121542	18.4704	0	237936	218481	171127
LWK	ACB	MSL	0.016256	12.917	0	204644	199562	193178
ACB	MXL	LWK	0.120105	18.6782	0	233322	220334	173083
ACB	PEL	LWK	0.114462	16.9042	0	232451	219734	174598
ACB	PUR	LWK	0.117512	20.064	0	232044	219295	173175
YRI	ACB	LWK	0.012865	9.59549	0	201182	199957	194877
ACB	MXL	MSL	0.105734	16.4434	0	239052	214813	173731
ACB	PEL	MSL	0.100496	14.9276	0	238106	214312	175170
ACB	PUR	MSL	0.10377	17.7518	0	237633	213901	173682
YRI	ACB	MSL	0.011431	8.89065	6.51E-19	203837	197530	193065
PEL	MXL	ACB	0.007868	3.3229	0.000891	312841	152838	150452
PUR	MXL	ACB	0.0042	1.99708	0.045816	297708	163842	162472
ACB	MXL	YRI	0.104656	16.4928	0	234368	216348	175354
PEL	PUR	ACB	0.003104	0.888684	0.374173	296819	164138	163122
ACB	PEL	YRI	0.099376	14.7249	0	233444	215836	176816
ACB	PUR	YRI	0.102817	17.8056	0	232942	215477	175299
CLM	CEU	ASW	0.032431	15.7893	0	289406	166684	156212
ASW	CEU	ESN	0.11115	17.6284	0	253179	209269	167402
GBR	CEU	ASW	0.000307	0.260328	0.794611	301599	153959	153864
ASW	CEU	GWD	0.118616	18.4242	0	250955	211670	166780
IBS	CEU	ASW	0.006887	5.52759	3.25E-08	298917	156473	154332
ASW	CEU	LWK	0.125997	19.8	0	250537	213121	165425
ASW	CEU	MSL	0.112313	17.4718	0	257389	207963	165966
MXL	CEU	ASW	0.030272	10.8168	0	288855	166737	156939
PEL	CEU	ASW	0.036088	8.35149	6.74E-17	285868	169926	158089
PUR	CEU	ASW	0.039606	17.7405	0	286881	169387	156480
ASW	CEU	YRI	0.11183	17.7975	0	252598	209553	167398
ASW	CLM	ESN	0.090837	17.9414	0	247597	206720	172292
CLM	GBR	ASW	0.032142	14.8437	0	289470	166621	156243
ASW	CLM	GWD	0.097444	18.9142	0	245538	208939	171835
CLM	IBS	ASW	0.025661	13.2814	0	287490	166503	158172
ASW	CLM	LWK	0.104025	20.3186	0	245246	210222	170607
ASW	CLM	MSL	0.092116	17.9272	0	251747	205455	170796
CLM	MXL	ASW	0.00207	1.14485	0.252272	284411	163063	162389
PEL	CLM	ASW	0.00422	1.24736	0.212266	284924	162484	161119
PUR	CLM	ASW	0.007272	6.10005	1.06E-09	278382	168581	166147
ASW	CLM	YRI	0.091471	18.1192	0	247034	207002	172306
ASW	GBR	ESN	0.110752	17.5554	0	253204	209249	167521
GWD	ASW	ESN	0.028291	13.5338	0	204636	203034	191862
ASW	IBS	ESN	0.107812	18.4884	0	251826	208837	168190
LWK	ASW	ESN	0.032073	15.2493	0	204848	204182	191492
ESN	ASW	MSL	0.030014	14.0786	0	207537	201891	190126
ASW	MXL	ESN	0.09054	16.2091	0	248249	206571	172270
ASW	PEL	ESN	0.085199	14.0845	0	247500	205889	173560

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
ASW	PUR	ESN	0.088207	18.0276	0	245875	206477	173004
YRI	ASW	ESN	0.025772	12.2433	0	203017	202439	192267
ASW	GBR	GWD	0.118136	18.4365	0	250994	211633	166914
IBS	GBR	ASW	0.006584	5.6326	1.78E-08	299010	156391	154345
ASW	GBR	LWK	0.125604	19.684	0	250551	213091	165534
ASW	GBR	MSL	0.112028	17.5239	0	257389	207962	166061
MXL	GBR	ASW	0.029987	10.5574	0	288983	166655	156952
PEL	GBR	ASW	0.035805	8.01575	1.09E-15	285993	169857	158114
PUR	GBR	ASW	0.039321	18.6985	0	286974	169319	156507
ASW	GBR	YRI	0.11138	17.6446	0	252628	209517	167523
ASW	IBS	GWD	0.115396	19.4102	0	249589	211269	167554
GWD	ASW	LWK	0.031942	15.5586	0	205396	204459	191802
GWD	ASW	MSL	0.023807	11.55	0	207833	200586	191257
ASW	MXL	GWD	0.097021	17.1968	0	246205	208752	171828
ASW	PEL	GWD	0.091213	14.8045	0	245559	207993	173221
ASW	PUR	GWD	0.09473	19.0881	0	243821	208665	172552
GWD	ASW	YRI	0.027672	13.4076	0	204112	203088	192151
ASW	IBS	LWK	0.122372	20.6209	0	249231	212625	166260
ASW	IBS	MSL	0.109131	18.3608	0	256001	207566	166719
MXL	IBS	ASW	0.023492	8.95459	3.25E-19	286660	166822	159164
PEL	IBS	ASW	0.029315	6.59638	4.21E-11	283408	170244	160547
PUR	IBS	ASW	0.03289	17.8639	0	285187	169045	158279
ASW	IBS	YRI	0.108452	18.5926	0	251255	209116	168196
LWK	ASW	MSL	0.032855	15.569	0	209115	202803	189901
ASW	MXL	LWK	0.103784	18.3089	0	245879	210071	170566
ASW	PEL	LWK	0.09807	15.9437	0	245201	209315	171927
ASW	PUR	LWK	0.100644	20.2657	0	243655	209822	171450
LWK	ASW	YRI	0.032169	15.427	0	204456	204368	191630
ASW	MXL	MSL	0.091907	16.2223	0	252383	205323	170759
ASW	PEL	MSL	0.086601	14.1841	0	251628	204666	172043
ASW	PUR	MSL	0.089468	18.0944	0	250008	205193	171492
YRI	ASW	MSL	0.028035	13.5945	0	207595	201372	190389
PEL	MXL	ASW	0.006587	2.63178	0.008494	292976	155794	153755
PUR	MXL	ASW	0.009326	4.37268	1.23E-05	279218	168172	165064
ASW	MXL	YRI	0.091157	16.3144	0	247685	206845	172284
PUR	PEL	ASW	0.003198	0.870326	0.384122	278160	167631	166563
ASW	PEL	YRI	0.085808	13.9816	0	246951	206176	173589
ASW	PUR	YRI	0.088855	18.2054	0	245309	206760	173015
CLM	CEU	ESN	0.023965	12.2425	0	325548	158916	151478
CEU	GBR	CLM	9.81E-05	0.078555	0.937386	195560	170675	170641
CLM	CEU	GWD	0.024996	12.3766	0	321649	159643	151856
IBS	CEU	CLM	0.011237	8.62733	6.29E-18	194099	174376	170501
CLM	CEU	LWK	0.025959	12.9693	0	319812	159675	151594
CLM	CEU	MSL	0.023819	11.7864	0	329861	157714	150375

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
CLM	CEU	MXL	0.029465	12.7119	0	185224	184620	174052
CLM	CEU	PEL	0.011131	4.87645	1.08E-06	187287	180507	176533
CLM	CEU	PUR	0.050827	23.989	0	193004	187776	169611
CLM	CEU	YRI	0.024017	11.8974	0	324779	159012	151553
GBR	CEU	ESN	0.000469	0.397397	0.691074	339779	148228	148089
GWD	CEU	ESN	0.133212	17.7378	0	230221	225596	172557
IBS	CEU	ESN	0.004076	3.31705	0.00091	336493	150138	148919
LWK	CEU	ESN	0.137604	18.0877	0	232014	225519	170962
ESN	CEU	MSL	0.135137	17.2262	0	231317	225802	172039
MXL	CEU	ESN	0.024309	8.89897	5.42E-19	325450	159422	151855
PEL	CEU	ESN	0.03024	7.46879	8.09E-14	322337	162485	152947
PUR	CEU	ESN	0.026809	12.4661	0	322158	160753	152359
YRI	CEU	ESN	0.130089	16.8377	0	226899	226034	173995
GBR	CEU	GWD	0.000573	0.479894	0.631303	335664	148738	148568
CEU	GBR	IBS	0.000232	0.175971	0.860317	178221	173105	173025
GBR	CEU	LWK	0.000468	0.390387	0.69625	333633	148577	148438
GBR	CEU	MSL	0.000326	0.281417	0.77839	344166	147100	147004
CEU	GBR	MXL	0.000336	0.264566	0.791344	196214	170806	170692
CEU	GBR	PEL	0.000291	0.23382	0.815125	201885	170286	170186
CEU	GBR	PUR	0.000196	0.149853	0.88088	199988	169908	169842
GBR	CEU	YRI	0.000539	0.450861	0.652089	338982	148296	148136
IBS	CEU	GWD	0.003917	3.24828	0.001161	332352	150622	149447
LWK	CEU	GWD	0.14678	18.9821	0	230334	228465	169981
GWD	CEU	MSL	0.12996	17.0246	0	233271	223131	171805
MXL	CEU	GWD	0.025502	9.28731	0	321568	160165	152199
PEL	CEU	GWD	0.03196	7.92429	2.29E-15	318583	163356	153238
PUR	CEU	GWD	0.027933	12.5884	0	318284	161505	152728
GWD	CEU	YRI	0.133252	17.814	0	229521	225761	172669
IBS	CEU	LWK	0.00444	3.60336	0.000314	330424	150564	149233
IBS	CEU	MSL	0.003878	3.19097	0.001418	340858	148988	147837
IBS	CEU	MXL	0.012809	10.485	0	195074	174747	170327
IBS	CEU	PEL	0.014298	12.4801	0	200977	174474	169555
IBS	CEU	PUR	0.010174	7.76127	8.41E-15	198343	173393	169900
IBS	CEU	YRI	0.004125	3.40296	0.000667	335701	150211	148976
LWK	CEU	MSL	0.139471	17.9551	0	236273	224263	169363
MXL	CEU	LWK	0.026245	9.96218	0	319694	160161	151969
PEL	CEU	LWK	0.032583	8.19347	2.54E-16	316681	163323	153016
PUR	CEU	LWK	0.029689	13.9727	0	316587	161678	152354
LWK	CEU	YRI	0.138346	18.2588	0	231467	225838	170945
MXL	CEU	MSL	0.024059	8.98622	2.17E-19	329732	158189	150756
PEL	CEU	MSL	0.029941	7.62741	2.40E-14	326595	161227	151853
PUR	CEU	MSL	0.026695	12.1343	0	326454	159535	151239
YRI	CEU	MSL	0.133368	17.0361	0	231493	225113	172133
PEL	MXL	CEU	0.017566	6.32271	2.57E-10	184144	178879	172703

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
MXL	CEU	PUR	0.045378	15.6783	0	191767	187143	170896
MXL	CEU	YRI	0.024384	8.73849	2.39E-18	324682	159519	151925
PUR	CEU	PEL	0.047755	20.7551	0	190799	189247	171996
PEL	CEU	YRI	0.030313	7.34212	2.10E-13	321583	162596	153029
PUR	CEU	YRI	0.026843	12.4494	0	321387	160848	152439
CLM	GBR	ESN	0.02352	11.4184	0	325635	158831	151531
GWD	CLM	ESN	0.114458	17.7728	0	226279	222001	176401
CLM	IBS	ESN	0.019933	11.5408	0	323095	159120	152900
LWK	CLM	ESN	0.118697	18.2541	0	227896	222043	174924
ESN	CLM	MSL	0.116703	17.3533	0	227528	222113	175688
MXL	CLM	ESN	0.00041	0.236929	0.812712	319308	156410	156282
PEL	CLM	ESN	0.006754	2.12805	0.033333	319825	156509	154409
PUR	CLM	ESN	0.002974	2.7994	0.00512	311845	161167	160211
YRI	CLM	ESN	0.111489	16.8339	0	223169	222324	177723
CLM	GBR	GWD	0.024451	11.4196	0	321755	159544	151929
IBS	GBR	CLM	0.011337	9.71119	0	194120	174350	170441
CLM	GBR	LWK	0.025516	11.931	0	319892	159582	151641
CLM	GBR	MSL	0.023511	11.2012	0	329922	157646	150403
CLM	GBR	MXL	0.029787	13.4081	0	185186	184664	173981
CLM	GBR	PEL	0.01141	5.24636	1.55E-07	187262	180548	176475
CLM	GBR	PUR	0.051016	25.1017	0	192992	187797	169566
CLM	GBR	YRI	0.023505	10.9997	0	324870	158910	151611
CLM	IBS	GWD	0.02111	11.7779	0	319184	159877	153266
LWK	CLM	GWD	0.127101	19.3254	0	226364	224788	174091
GWD	CLM	MSL	0.111301	17.0355	0	229292	219599	175612
MXL	CLM	GWD	0.000573	0.335543	0.737216	315633	157012	156832
PEL	CLM	GWD	0.007473	2.3784	0.017388	316259	157220	154888
PUR	CLM	GWD	0.003073	2.77615	0.005501	308185	161785	160793
GWD	CLM	YRI	0.114454	17.8183	0	225598	222165	176532
CLM	IBS	LWK	0.021569	11.8059	0	317424	159831	153082
CLM	IBS	MSL	0.019979	11.1069	0	327362	157940	151752
CLM	IBS	MXL	0.017038	8.13687	4.06E-16	184654	183506	177357
IBS	CLM	PEL	0.00263	1.34208	0.179571	186980	180101	179156
CLM	IBS	PUR	0.040808	21.6133	0	191950	187105	172432
CLM	IBS	YRI	0.019938	11.1143	0	322335	159209	152985
LWK	CLM	MSL	0.120683	18.2172	0	232098	220830	173269
MXL	CLM	LWK	0.000355	0.20487	0.837673	313913	156867	156756
PEL	CLM	LWK	0.007141	2.23718	0.025275	314508	157044	154817
PUR	CLM	LWK	0.003855	3.47379	0.000513	306649	161824	160581
LWK	CLM	YRI	0.119399	18.4305	0	227364	222356	174921
MXL	CLM	MSL	0.000304	0.177459	0.859148	323545	155231	155137
PEL	CLM	MSL	0.006596	2.13393	0.032848	324059	155328	153292
PUR	CLM	MSL	0.003002	2.84702	0.004413	316062	159969	159011
YRI	CLM	MSL	0.114897	17.1763	0	227700	221441	175799

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
MXL	PEL	CLM	0.001217	0.45963	0.645782	187448	172032	171614
CLM	MXL	PUR	0.005318	2.36033	0.018259	191581	181278	179361
MXL	CLM	YRI	0.000432	0.247346	0.80464	318561	156507	156372
PEL	CLM	PUR	0.00178	0.47592	0.634131	192220	179582	178944
PEL	CLM	YRI	0.006778	2.08578	0.036999	319086	156615	154506
PUR	CLM	YRI	0.002956	2.72065	0.006515	311097	161264	160313
GWD	GBR	ESN	0.132808	17.7555	0	230203	225609	172709
IBS	GBR	ESN	0.003612	3.08523	0.002034	336616	150042	148961
LWK	GBR	ESN	0.137214	18.1099	0	232001	225506	171088
ESN	GBR	MSL	0.134853	17.3095	0	231290	225818	172151
MXL	GBR	ESN	0.023868	8.67059	4.34E-18	325604	159322	151894
PEL	GBR	ESN	0.029802	7.14429	9.05E-13	322488	162397	152998
PUR	GBR	ESN	0.026368	13.1655	0	322279	160670	152414
YRI	GBR	ESN	0.129693	16.8321	0	226884	226039	174139
GWD	IBS	ESN	0.130285	18.6249	0	229449	224780	172961
ESN	GWD	LWK	0.00177	1.81283	0.069859	198491	197671	196972
ESN	GWD	MSL	0.006242	6.28192	3.34E-10	200265	196463	194026
GWD	MXL	ESN	0.113962	16.7186	0	226466	222240	176768
GWD	PEL	ESN	0.108814	15.1331	0	225679	221634	178134
GWD	PUR	ESN	0.112288	17.9484	0	225360	221118	176473
ESN	YRI	GWD	0.002068	3.93873	8.19E-05	197011	196824	196011
LWK	IBS	ESN	0.134629	18.8966	0	231146	224763	171425
ESN	IBS	MSL	0.132368	17.9177	0	230488	225041	172428
MXL	IBS	ESN	0.020254	8.1953	2.50E-16	322705	159878	153530
PEL	IBS	ESN	0.026161	6.4051	1.50E-10	319319	163166	154847
PUR	IBS	ESN	0.022819	13.9925	0	319939	160809	153633
YRI	IBS	ESN	0.127153	17.5226	0	226100	225250	174430
LWK	ESN	MSL	0.0029	3.31725	0.000909	201186	196465	195329
LWK	MXL	ESN	0.118213	16.9996	0	228119	222249	175258
LWK	PEL	ESN	0.113055	15.6017	0	227338	221612	176593
LWK	PUR	ESN	0.116454	18.4656	0	226850	221284	175121
ESN	YRI	LWK	0.001349	2.49492	0.012599	198550	196561	196032
ESN	MXL	MSL	0.116299	16.2757	0	227733	222352	176022
ESN	PEL	MSL	0.111264	15.0026	0	227023	221673	177283
ESN	PUR	MSL	0.114537	17.5228	0	226633	221216	175749
ESN	YRI	MSL	0.002016	4.13058	3.62E-05	201102	194646	193863
PEL	MXL	ESN	0.006626	2.844	0.004455	328620	149759	147787
PUR	MXL	ESN	0.002585	1.21528	0.224258	313167	160442	159615
YRI	MXL	ESN	0.111011	15.8163	0	223381	222543	178071
PEL	PUR	ESN	0.003565	1.02191	0.306823	312056	161061	159917
YRI	PEL	ESN	0.10593	14.5255	0	222696	221860	179359
YRI	PUR	ESN	0.10932	16.9424	0	222294	221444	177799
IBS	GBR	GWD	0.003351	2.92099	0.003489	332486	150506	149500
LWK	GBR	GWD	0.146312	19.0904	0	230336	228435	170121

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
GWD	GBR	MSL	0.129666	17.1203	0	233227	223161	171931
MXL	GBR	GWD	0.024962	8.95269	3.25E-19	321740	160052	152256
PEL	GBR	GWD	0.031425	7.49138	6.82E-14	318752	163255	153307
PUR	GBR	GWD	0.027394	13.1856	0	318420	161404	152797
GWD	GBR	YRI	0.132799	17.7681	0	229507	225757	172826
IBS	GBR	LWK	0.003977	3.38584	0.00071	330534	150455	149263
IBS	GBR	MSL	0.003555	3.089	0.002008	340950	148904	147849
IBS	GBR	MXL	0.013144	11.4607	0	195062	174770	170235
IBS	GBR	PEL	0.014588	13.5644	0	200986	174502	169484
IBS	GBR	PUR	0.01037	8.78477	1.52E-18	198349	173384	169825
IBS	GBR	YRI	0.003592	3.04877	0.002298	335827	150097	149022
LWK	GBR	MSL	0.139189	18.0712	0	236236	224269	169465
MXL	GBR	LWK	0.025806	9.55872	0	319840	160053	152000
PEL	GBR	LWK	0.032149	7.76004	8.49E-15	316823	163227	153059
PUR	GBR	LWK	0.029251	14.5004	0	316698	161583	152399
LWK	GBR	YRI	0.137907	18.2326	0	231459	225809	171076
MXL	GBR	MSL	0.023754	8.80771	1.30E-18	329859	158105	150769
PEL	GBR	MSL	0.029638	7.29405	3.01E-13	326721	161159	151881
PUR	GBR	MSL	0.02639	12.9658	0	326545	159464	151264
YRI	GBR	MSL	0.133079	17.1189	0	231453	225137	172253
PEL	MXL	GBR	0.017609	6.30261	2.93E-10	184049	178899	172707
MXL	GBR	PUR	0.045574	16.4582	0	191806	187134	170821
MXL	GBR	YRI	0.023876	8.39618	4.61E-17	324839	159401	151967
PUR	GBR	PEL	0.048032	24.2318	0	190807	189288	171938
PEL	GBR	YRI	0.029811	6.96934	3.18E-12	321738	162492	153084
PUR	GBR	YRI	0.026336	12.9242	0	321511	160745	152496
LWK	IBS	GWD	0.143934	19.8819	0	229448	227734	170425
GWD	IBS	MSL	0.127174	17.868	0	232472	222356	172181
MXL	IBS	GWD	0.02159	8.57317	1.01E-17	318816	160657	153867
PEL	IBS	GWD	0.028016	6.80388	1.02E-11	315557	164073	155130
PUR	IBS	GWD	0.024089	14.0346	0	316051	161589	153988
GWD	IBS	YRI	0.130286	18.6338	0	228760	224941	173084
LWK	GWD	MSL	0.009106	7.65266	1.97E-14	200992	198010	194437
LWK	MXL	GWD	0.126484	18.07	0	226603	224960	174442
LWK	PEL	GWD	0.120877	16.4829	0	225925	224244	175879
LWK	PUR	GWD	0.124803	19.6185	0	225317	223993	174287
YRI	GWD	LWK	0.000429	0.478155	0.63254	198803	197171	197001
GWD	MXL	MSL	0.110893	16.0167	0	229462	219854	175961
GWD	PEL	MSL	0.105775	14.4829	0	228680	219284	177332
GWD	PUR	MSL	0.109104	17.2331	0	228371	218712	175682
YRI	GWD	MSL	0.004241	4.6502	3.32E-06	200453	195838	194184
PEL	MXL	GWD	0.007206	3.06359	0.002187	325046	150412	148260
PUR	MXL	GWD	0.002526	1.16855	0.242585	309523	161025	160213
GWD	MXL	YRI	0.113944	16.701	0	225785	222397	176899

P1	P2	P3	Dstatistic	Z-score	p-value	BBAA	ABBA	BABA
PEL	PUR	GWD	0.004162	1.19268	0.232994	308529	161775	160434
GWD	PEL	YRI	0.108782	14.9241	0	225021	221812	178288
GWD	PUR	YRI	0.112297	17.9797	0	224679	221287	176605
LWK	IBS	MSL	0.136648	18.7618	0	235371	223542	169793
MXL	IBS	LWK	0.02183	8.86576	7.59E-19	317018	160574	153713
PEL	IBS	LWK	0.028144	6.84715	7.53E-12	313728	163958	154982
PUR	IBS	LWK	0.025342	15.0338	0	314429	161681	153689
LWK	IBS	YRI	0.135334	19.0093	0	230608	225075	171416
MXL	IBS	MSL	0.020195	8.15055	3.62E-16	326944	158670	152388
PEL	IBS	MSL	0.026051	6.47971	9.19E-11	323535	161936	153713
PUR	IBS	MSL	0.022895	13.5211	0	324190	159613	152468
YRI	IBS	MSL	0.130585	17.7541	0	230661	224364	172535
PEL	MXL	IBS	0.01907	6.86597	6.60E-12	188033	178347	171672
MXL	IBS	PUR	0.035354	13.8908	0	190454	186758	174003
MXL	IBS	YRI	0.020281	7.88653	3.11E-15	321948	159971	153611
PUR	IBS	PEL	0.033968	17.8904	0	190671	187693	175360
PEL	IBS	YRI	0.026188	6.24665	4.19E-10	318574	163271	154938
PUR	IBS	YRI	0.022807	13.7405	0	319176	160895	153720
LWK	MXL	MSL	0.12028	16.9924	0	232301	221050	173584
LWK	PEL	MSL	0.115152	15.6622	0	231515	220439	174913
LWK	PUR	MSL	0.118431	18.4875	0	231038	220055	173452
LWK	YRI	MSL	0.004902	6.3129	2.74E-10	201218	196751	194831
PEL	MXL	LWK	0.007089	2.92083	0.003491	323253	150263	148147
PUR	MXL	LWK	0.003524	1.62601	0.103947	307950	161096	159965
LWK	MXL	YRI	0.118898	17.1225	0	227587	222555	175256
PEL	PUR	LWK	0.003056	0.865862	0.386566	306927	161498	160513
LWK	PEL	YRI	0.113727	15.5477	0	226825	221935	176610
LWK	PUR	YRI	0.11717	18.6454	0	226317	221601	175118
PEL	MXL	MSL	0.006572	2.89456	0.003797	332888	148647	146706
PUR	MXL	MSL	0.002717	1.28783	0.197804	317366	159261	158398
YRI	MXL	MSL	0.114496	16.1435	0	227897	221678	176130
PEL	PUR	MSL	0.003384	0.993964	0.32024	316246	159834	158756
YRI	PEL	MSL	0.109445	14.8648	0	227204	221018	177412
YRI	PUR	MSL	0.112724	17.347	0	226810	220544	175860
PEL	MXL	PUR	0.007459	2.74243	0.006099	198763	172627	170071
PEL	MXL	YRI	0.006629	2.8302	0.004652	327870	149847	147873
PUR	MXL	YRI	0.002546	1.15134	0.249594	312417	160531	159715
PEL	PUR	YRI	0.003606	0.994099	0.320175	311316	161170	160012