Faculty of Science, Engineering, and Agriculture

Department: Biological sciences

University of Venda

Thohoyandou Limpopo

South Africa

Genomic diversity and structure among South and East Asian human populations.

Masters dissertation

By

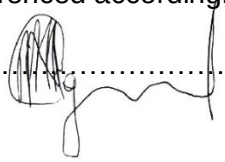
Marcia Dunisani Ngoveni (14013870)

Supervisor: Professor Yoshan Moodley

Co-supervisor: Dr. Andrinajoro Rakotoarivelo

Declaration

I, Marcia Dunisani Ngoveni of student number 14013870, declare that this research dissertation is my original work and has not been submitted for any other degree at any university or institution. The dissertation does not contain any other persons' writing specifically unless acknowledged and referenced accordingly.

Student……  ……Date…….03/03/2022……

Contents

Abstract i

1. Introduction 1

1.1. Genetic diversity of modern humans 1

1.2. The structure of South Asia 3

1.3. The structure of East Asia 4

1.4. Island populations 6

1.5. Purpose of study 7

2. Methods 9

2.1. Samples 9

2.2. Bioinformatics 10

2.2.1 Mapping of hominin and outgroup genomes 10

2.2.2. Filtering 11

2.3. Genetic diversity 13

2.4. Population structure 13

2.4.1 Principal component Analysis 13

2.4.2. Admixture 14

2.4.3. Phylogeny 14

2.4.4 Phylogenetic networks 16

2.5 Testing for introgression 16

3. Results 18

3.1. Genetic Diversity 18

3.1.1. Population heterozygosity and nucleotide diversity 18

3.2. Population structure of x chromosome and autosomal datasets 18

3.2.1. Principal component analysis 19

3.2.2. Admixture plots 21

3.2.3 Phylogeny 24

3.3 Gene flow 31

4. Discussion 33

4.1 Genetic diversity 33

4.2 Genetic structure of Asia 35

4.2.1 Genomic structure in South Asian Populations 35

4.2.2 Genomic structure in East Asia 36

4.2.3 Colonization of Islands 37
4.2.4. The structure differences from each genetic marker 38
4.3 Introgression 39
4.3.1. Introgression within South Asia 39
4.3.3. Gene flow between South and East Asia 40
5. Conclusion 41
6. Acknowledgements 41
7.References 43
Appendix A: Command lines 53
Appendix B: Job submission script 55
Appendix C: Admixture 56
Appendix D: introgression between South and East Asian populations. 58

Abstract

Genetic variability is a complex concept that is affected by many factors. Studying the genetic variability within the human population provides a better understanding of how populations evolve. The genetic variability within modern humans resulted from different conditions experienced in different routes used by out of Africa migrations to other parts of the world. These events included mating with other human species (Neanderthals and Denisovans) and the development of cultural practices as they arrived in other parts of the world. Asia is more diverse in terms of physical geography, cultures, and languages, which makes it an interesting region to study human population genetic diversity and population structure. From animal studies, colonization of islands is often associated with loss of genetic diversity resulting in isolation. Although humans can easily move around, various aspects could result in isolation between populations. To test the concept of low genetic variability in island populations, I used 1000 genomes project variant calls datasets (autosomal, X-chromosome, Y-chromosome, and mitochondrial data), which include 983 individuals from ten Asian populations. I estimated the genetic diversity from autosomal dataset of South and East Asian populations to compare the genetic affinities between mainland and island populations. Reduced levels of genetic diversity were observed in Japan suggesting a possible genetic bottleneck during the initial colonization of the island. Despite the diverse cultural practices and South Asia being colonized early the populations are not more internally structured nor more genetically diverse compared to East Asia. Most East Asian populations had higher heterozygosity compared to South Asian populations. It is clear from the results that physical geography plays a crucial role in shaping the genetic diversity between populations. Significant levels of gene flow were observed between populations that are at close geographical distance compared to populations at distant proximity.

1. Introduction

1.1. Genetic diversity of modern humans

Human evolution is a complex process that involves migrations, founder effects and bottlenecks. Previous studies showed that modern human migrations began from Africa around a hundred thousand years ago (Cann *et al*, 1987; Vigilant *et al*, 1991). As humans moved out of Africa they moved in a group and continued to colonize the new unoccupied environment and expanded into populations (Majumder, 2010). Evidence from a microsatellite loci study suggests that the process of inhabiting new environments had occurred in a stepwise manner until the whole world was colonized (Ramachandran *et al*, 2005). It is commonly known that colonization and growth of an area from a single source might lead to a genetic bottleneck if the colonizing group is small and becomes isolated. This is because the power of genetic drift becomes significant, resulting in a loss of genetic variation within the population (Ramachandran *et al*, 2005). This correlation between genetic differentiation and geography was investigated by measuring pairwise FST from the Human Genome Diversity Project-Centre d'Etude du Polymorphisme Humain (HGDP-CEPH) dataset (Ramachandran *et al*, 2005).

The migrating human populations are believed to have struggled to adapt to the environmental changes, which resulted in a genetic bottleneck that occurred around 84,000-60,000 years ago (Marth *et al*, 2004). Even though we know that humans can almost always move around or across geographical barriers and that genetic diversity can be maintained by only an average of one migrant per generation (Cavalli-Sforza and Feldman, 2003), there is still a possibility that a partially isolated population could suffer from reduced levels of genetic diversity. Apart from the demographic effect of genetic drift, colonization of new environments also comes with different selection pressures which affect genetic variability (Lacy, 1987).

The evidence from mitochondrial DNA and Y-chromosome studies (Pierson *et al*, 2006; Kayser *et al*, 2006) suggests that only a few individuals from the migrating groups survived the change in environmental conditions. This reduction in population numbers may have contributed greatly to the recent observed low levels of heterogeneity (confirmed by Y-chromosome data and mtDNA)

in the out of Africa populations as compared to Africa populations (Garrigan *et al*, 2007; Li *et al*, 2008).

The colonization of islands by small groups of people occurred a bit later compared to the initial colonization of the respective continents (Keegan *et al,* 1987; Gamble *et al*, 1993; Pattom, 2013). Kaifu *et al*, (2015) proposed that the last glacial maximum, which occurred 21,000 years ago, aided in the migration of larger groups of people onto the islands. However, the colonization of some islands still required the use of boats even when the sea levels were low (Kaifu *et al*, 2015).

Physical geography played a crucial role in influencing the pattern of spread across the globe. Some routes were not possible to take during early human migrations because humans had not developed means to deal with them. Two main routes were suggested about the spread of humans into Asia. In line with Stringer (2000), the first route was through the Bab al Mandab strait between Africa and Arabia into southern Arabia. The other route involved the crossing of Nile-Sinai corridor in Egypt into Mediterranean Levant (Derricourt, 2005). These different groups of modern humans experienced different selection and population pressures as they continued moving and occupying areas. They also developed different traits and practices which helped them adapt to the new environment. Some alleles were surely lost along the way as a result of both natural selections and drift acting upon the populations (Cavalli-Sforza and Feldman, 2003). This is because the developments of new cultures might have led to isolation between the populations, resulting in adaptive evolution and genetic differentiation (Saeb and Al-Naqeb, 2016). Klein (1999) further suggests that the development of new and more cultures may have also resulted in rapid expansions of populations.

According to Mallick *et al*, (2015), modern humans admixed with Neanderthals and Denisovans around 50,000 years ago before replacing other Homo lineages (Klein 2009; Stringer 2012) as they spread out of Africa and over the globe (Sankararaman *et al,* 2012; Venoland Akey, 2014; Kim *et al,* 2015). Reich *et al*, (2011) suggest that Denisova interbreeding occurred in mainland Asia and spread through Southeast Asia. A recent archeological study by Li *et al*, (2017) proved that Neanderthal characters are displayed on Chinese specimens. According to Ingman *et al,* (2000) early human migrations still resulted in a loss of genetic diversity across moving populations, notwithstanding all of the ancient gene flow episodes.

1.2. The structure of South Asia

South Asia is inhabited by more than 20% of the overall world population, of which the bulk is living in India (Majumder, 2010). The region also comprises a variety of environmental aspects which includes, the eastern and western Ghats Mountain ranges, the Deccan plateau, and large river valleys. The modern-day demographic make-up of the Indian subcontinent is notably heterogeneous coming from the result of complicated human migration and interaction. The region has served as one of the main and earliest pathways for early human migration out of Africa (Stoneking and Delfin, 2010). For the colonization of this region, two routes were proposed. The southern route into southern Iran, which crossed the Arabian Sea around 75,000 years ago, is the first route (Lahr and Forey, 1994; Mellers 2006). The second route proposed by Boivin et al, (2013) proposed that humans migrated through Central Asia's northern riverine corridors. These initial migrations involved two ancestral groups: the Dravidian language speakers who are now the inhabitors of South Asia, known as the ancestral South Indians (ASI) and the Indo-Aryan language speakers in North India, known as the ancestral North Indians (ANI) (Wong *et al*, 2014). According to Stokowski *et al*, (2007) the two groups can be differentiated by skin complexion, with South Indians being darker in complexion.

India is a very diverse area in terms of linguistic and cultural diversity (Xing *et al*, 2010). According to Singh, (1992) India alone consists of 450 tribal communities from which 750 languages are spoken. Based on the previous study using DNA markers (mitochondrial, autosomal, and, Y-chromosomal) India comprises four distinct language families namely: Dravidian, Tibeto-Burman, Austro-Asiatic, and Indo-European, that are largely nonoverlapping geographically (Basu *et al*, 2003). Dravidian-speaking groups inhabit southern India, Tibeto-Burman speakers are found in northeastern India, Austro-Asian speakers in dispersed geographic areas of central and eastern India, and lastly Indo-European speakers are found in northern India (Manjumder and Basu, 2015). However, genetic evidence from whole genomes studies, Y-chromosome analysis and mitochondrial studies (Metspalu *et al*, 2011; Brahmachari *et al*, 2008; Sahoo and Kashyap, 2006) suggest that most of the linguistic groups in this region are from a combination of the two groups (ancestral South Indians (ASI) and ancestral North Indians (ANI).

Genome-wide study on single nucleotide polymorphisms suggested that India experienced a demographic shift several thousand years ago, where closely related groups became isolated because of cultural practices (Moorjani *et al*, 2013). This study showed that there was admixture among all groups before the establishment of the caste system. The caste system is a social rank used to divide human populations based on culture and religions (Moorjani *et al*, 2013). According to Ali *et al*, (2014) the establishment of caste systems in this region strongly affected the migration patterns resulting in isolation between different ethnic groups. Most populations in India are said to be highly endogamous (Majumder, 2008) and this endogamy might have played a huge role in the genetic diversity of the region. Marrying within ethnic groups can gradually reduce the genetic diversity and can also cause genetic disorders. Genetic studies on Y-chromosomes (Sengupta *et al*, 2006; Cai *et al*, 2011) observed higher variation in Y-chromosomes in India resulting from the cultural practices, suggesting that there are more male migrants in India than females (Bamshad *et al*, 2008). South Asians are expected to be highly structured because of these cultural practices.

1.3. The structure of East Asia

East Asia contributes approximately 38% of the Asian population (Wang *et al*, 2018), and 22 % of the world population, making it a suitable region to investigate genetic relationships between human populations. Based on ancient evidence and fossil records (Shang *et al*, 2007) East Asians were already present in eastern China 40 000 years ago. The region also includes a variety of environments, including the Pacific Ocean in the east, the Himalayan Plateau in the southwest, the Ural Mountains in the west, and the Bering Strait in the north. These physical aspects may have played a role in guiding the migration pathways. Two main routes have been proposed for this dispersal, from the northern or southern part of the Himalayas (Chu *et al,* 1998; Su *et al,* 1999; Yao *et al,* 2002; Pope and Terrel, 2008; Joeng *et al,* 2016; Bae *et al*, 2017).

The direct migration from India to China was unlikely because the Himalayas were too high (8.850 meters) and too long for modern humans to cross without tools. Unlike today, when it is possible to fly through mountains, no means of transportation were used during early human migrations. To travel across the globe, populations had to consider closer and less risky environments.

The southern route was supported by studies on mitochondrial DNA and Y-chromosome data (Hugo Pan Asian SNP consortium,2009; Jeong *et al*, 2016). Based on mitochondrial DNA, whole-genome sequencing, ancient data, and NRY studies, this south to north migration resulted in two ancestral populations in northern East Asia and the pre-Neolithic hunter-gathers from Southeast Asia (Wang *et al*, 2018; Pan and Xu, 2020). This south to north direction was supported by the findings from HUGO Pan-Asian consortium, (2009). The study used single nucleotide polymorphism dataset from 71 Asian populations to reconstruct maximum likelihood tree that showed a south to north migrations.

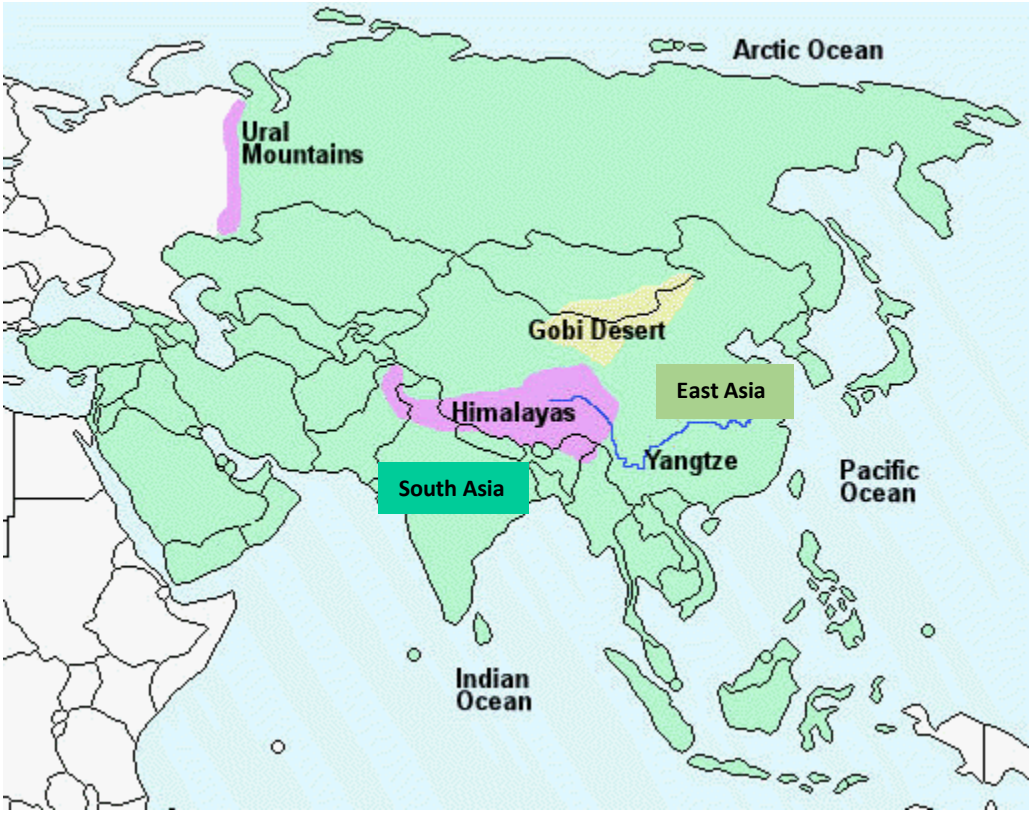


Figure 1. Map of Asia showing the geographical complexity of East Asia (https://www.naturalhistoryonthenet.com/Continents/images/asia.gif).

The eastern part of Asia is also known to be highly diverse in terms of ethnicity (Liu *et al*, 2020). According to Wang *et al*, (2018) East Asians have gone through uneven population expansion around 20,000 to 10, 000 years ago. Pan and Xu, (2020) suggest that this event might have led to greater population diversity. According to geographical distribution and ethnicity, East Asia populations can be divided into three groups: Chinese, Japanese and Korean. Chinese people make up the largest ethnic group in East Asia, with 56 officially recognized ethnic groupings and, Han Chinese constituting the majority 90% (Pan and Xu, 2020). Studies from mitochondrial DNA and Y-chromosome (Yao *et al*, 2002; Su *et al*, 1999) variations suggest that the initial colonization of China resulted from Southeast Asian migrations which were later followed by expansions.. According to Wen *et al,* (2004) Han experienced population expansion around 2,000 years ago to southern China. This expansion was also observed from phylogenetic trees reconstructed from various studies (Cavalli-Sforza *et al*, 1994; Ji and Su, 2000) where the populations formed two district clusters (South and North). Furthermore, Chen *et al*, (2009) gave evidence about the expansion of Han Chinese by running genomic structure analysis on genome-wide SNP variation from over 6000 Han Chinese samples. The Southern Han Chinese resulted from the admixture between Han and native South China populations (Austro-Asiatic phyla, Hmong-Mien, and Daic).

The second most diverse population in East Asia is Vietnam comprising of 54 ethnic groups (Liu *et al*, 2020). The importance of Vietnam is associated with being colonized by early modern human migrations from India (Underhill, 2004). Recent studies from complete mitochondrial genomes, Y-Chromosome, and genome-wide SNPs (Duong *et al*, 2018; Liu *et al*, 2020; Machodlt *et al*, 2020) suggest a diverse genetic profile in Vietnam.

1.4. Island populations

Due to random chance and heavy natural selection, migration to islands can have a negative impact on population size (Wilson and MacArthur,1967), resulting in the loss of genetic diversity within Island populations. This concept is known as island biogeography which was tested in animals and plants (Wilson and MacArthur, 1993). Although, no physical geography can be considered as isolation in humans the possibility of islands suffering from low genetic variability can still be plausible. Other factors, such as poverty and culture (Basu *et al*, 2003; Xing *et al*, 2009; Basu *et al*, 2016), may also prevent migrations. However, water might have served as physical barrier during early human migrations. This is because the archeological evidence

suggesting that the development of marine capability only occurred around 30,000-12,000 years ago (Fujita *et al*, 2016). Another possibility could be that modern humans experienced bottlenecks as they colonized the islands. So, in this study, I will make use of the SNP data from two main Asian islands (Sri Lanka and Japan) to test the idea of island biogeography in humans.

Sri Lanka is a small island that comprises a variety of environmental aspects, which include arid zones, dry, lowland tropical rainforests (Ratnayaka, 2016). Furthermore, the island is separated from mainland India by 50 kilometers strait. This suggests that there was an easy human migration from India during the sea-level drop (Ratnayaka, 2016), which may have led to the people of this island. Mellers *et al*, (2013) proposed that when the island was still united to mainland South Asia around 48, 000 years ago, it was colonized by migrants arriving from Africa via a southern coastal route. The initial migrants were the Vedda which was later followed by two major Indian groups the Tamils and Sinhalese (Ranaweera *et al*, 2013). The rise in the sea levels during the interglacial phase provided a physical barrier that stopped migrations between Sri Lanka and India (Ratnayake, 2016).

Japan is located on Asia's east coast, and mountains cover 80% of its landmass. The island is 200 kilometers from the mainland, implying that Japanese colonization might have necessitated the use of watercraft (Kaifu et al, 2015). Unlike Sri Lanka, Japan was first colonized by modern people traveling along the coast from Southeast Asia around 35,000 to 40,000 years ago (Nakazawa, 2017) which is about 13,000 to 8,000 years later after the colonization of Sri Lanka. Present-day genetic studies (Jinam *et al*, 2012; Nakagome *et al*, 2015; Jinam *et al*, 2021) suggest that Japan was at least colonized by two waves of migrations (Chooke *et al*, 2021). The Jomon were the first to arrive in Japan from Southeast Asia around 40, 000 years ago, and were followed by the Yayoi, who came from mainland Northeast Asia around 2,300 years ago (Nakazawa, 2017).First migration was isolated from the mainland for 38,000 years and this long-term isolation which might have resulted in genetic bottleneck that gradually reduced the genetic diversity of this area.

1.5. Purpose of study

Various studies on mitochondrial DNA, single nucleotide polymorphism, Y-chromosome and X-chromosome data provided enough evidence for the loss of genetic diversity associated with early modern humans in colonizing unoccupied regions out of Africa (Vigilant *et al*, 1991; Nei, *et al*, 1993, Jorde *et al*, 1997). Asia is a suitable area to study and compare the genetic diversity and structure in relation to the initial migration date and cultural developments. This is because the history, origins, and culture of populations in this region is well documented by using archeological evidence, mitochondrial DNA, Y-chromosome, single nucleotide polymorphism, and ancient genome data (Manjumder, 2010; Chaubey *et al*, 2008; Wang *et al*, 2021; Roychaubery *et al*, 2021). Previous studies on genetic diversity and population relationships in Asia were either focused on regional (South or East) or used single genetic marker (HuGO Pan-Asian consortium, 2009; Manjumder, 2010; Wang *et al*, 2014; Liu *et al*, 2020). This study will use different makers to compare the genetic diversity of South and East Asian human populations in relation to the migration dates and cultural development. Expecting to see high genetic diversity in areas that were first to be occupied, given that modern humans' genetic diversity was lost along the way. The use of different genetic markers could be essential when studying population relationships. This is because their effective population sizes are different (X-chromosome is ¾ of every four autosomes, and Y-chromosome and mitochondrial DNA are both a quarter of autosomes, which results in different rate of genetic drift on each. Different markers can portray different structure since the effect of genetic drift is also different.

The development of next generation sequencing tools now allows the use of thousands of individuals which provide more information for better understanding of population relationships. Although large genomic data were generated, most of the programs that were used require a large RAM which limits the use of many individuals, or the amount of data used per study. However, this can be solved by using fewer individuals and a large number of SNPs.

In this study, I will use 1000 genomes variant call dataset which contains 983 individuals from ten Asian (South and East) human populations that is freely available online. The data will be divided into four datasets (Y-chromosome, mitochondrial, X-chromosome, and autosomal) to reduce the size for better computation. These datasets will help to investigate and test:

1. Which of the regions (South or East) have higher genetic diversity.

2. There is a more genetic affinity between island and closer mainland populations.

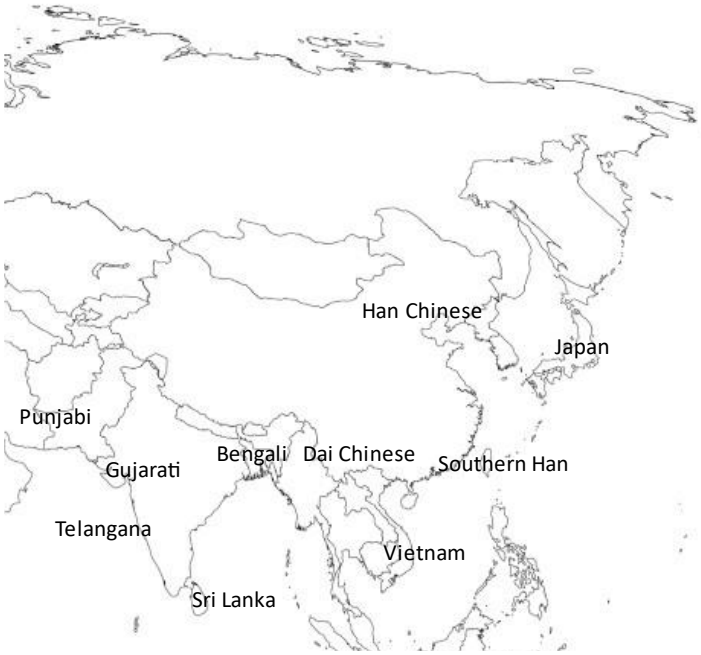
2. Methods

2.1. Samples

This study used data from 30x coverage datasets (autosomal, X-chromosome, and Y-chromosome) and low coverage mitochondrial data from the 1000 genomes project browser, which included 983 individuals from ten Asian human populations (Table 1): Punjabi (PJL), Gujarati Indians in Houston (GIH), Indian Telugu (ITU), Sri Lankan Tamil (STU), Bengali Indians (BEB), Chinese Dai (CDX), Kin, Vietnam (KHV), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB) and Japanese in Tokyo (JPT). These populations were chosen as a better representation for testing how geography shapes the mode of gene flow and structure between and within regions. Only datasets with the unrelated samples were downloaded from the 1000 genomes variant calls datasets, which contains 3115 samples from 26 populations (https://www.internationalgenome.org/data-portal/data-collection/phase-3).

We added two ancient genomes to this study because according to previous studies (Reich *et al*, 2011: Mallik *et al*, 2015; Li *et al*, 2017), ancient hominins also contributed to the gene pool of the present Asian populations. The additional short-read sequences for Neanderthal, Denisovan, and a Chimpanzee outgroup were downloaded from ENA (European Nucleotide Archive) in fastq format. The chimpanzee was chosen as an outgroup because it was a better representation for a divergent genome for both modern humans and hominin genomes. For data analysis, a human reference genome GRCh38_full_analysis_set_plus_decoy_hla. fa was also downloaded from the 1000 genome browser http://www.1000genomes.org. A reference genome assembly is a nucleic acid sequence that has been constructed to represent all of species genetic material. According to Almal *et al*, (2019) it is important to have a reference genome because a reference genome provides a framework for which any individual can be compared to reveal the individual's genetic makeup.

Table 1. Number of genomes used from each population and region (South or East Asia) they belong to.



Population	Population code	No. of samples	Region	Main/Island
Punjabi	PJL	100	SouthAsia	Mainland
Gujarati	GIH	103	SouthAsia	Mainland
Telangana	ITU	104	SouthAsia	Mainland
Sri Lankan	STU	99	SouthAsia	Island
Bengali	BEB	90	SouthAsia	Mainland
Dai Chinese	CDX	93	East Asia	Mainland
Vietnam	KHV	100	East Asia	Mainland
Southern Han	CHS	104	East Asia	Mainland
Han Chinese	CHB	86	East Asia	Mainland
Japan	JPT	104	East Asia	Island
Denisovan	N/A	1	Not specified	N/A
Neanderthal	N/A	1	Not specified	N/A
Chimpanzee	N/A	1	Not	N/A

2.2. Bioinformatics

The SNP datasets were downloaded in variant calls format (VCF) (Denecek *et al*, 2011) which is a text format for storing marker and genotype data. The vcf files were limited to chromosomes and each chromosome contained all 1000 genome phase, 3 individuals. After downloading all the chromosomal vcf files, all the autosomal vcf files were concatenated into one vcf file using bcftools (Denecek *et al*, 2015) while keeping the X-chromosome, Y-chromosome, and mitochondrial data separate. To pull out the ten populations from the four vcf files, a text sample list was created using a text editor to specify only the individuals that belong to the target Asian populations. The sample list was used as an input file in bcftools together with the -S option to subset the datasets (command 19) in appendix A.

2.2.1 Mapping of hominin and outgroup genomes

We first indexed the reference genome with BWA (Burrows-Wheeler aligner) version 0.7.15-r1140 (Li and Durbin, 2008) before mapping the short reads against the human reference genome SAMtools version 1.9 outputting BAM (Binary Alignment Map) files. Samtools is a software and a

library package used to manipulate SAM/BAM files generated from different platforms (Li *et al*, 2009). The bam files were first sorted normally in samtools sort, then sorted by fixate, by read name, and lastly sorted by coordinates. PCR duplicates were marked using the samtools markdup command. The marked dup bam file was then converted into the final bam file. Specific filters can be found in the appendix A (command 1-6).

Bcftools mpileup in bcftools version 1.9 was performed on individual bam files to call for SNPs. All variants with a quality value of less than or equal to 1 were discarded from the vcf files. The multiallelic sites were merged into single rows then later discarded leaving on biallelic sites. The detailed conversion steps are given by the commands in appendix A command (7-12).

2.2.2. Filtering

The merging and quality filtering was done with bcftools. For phylogeny reconstructions 1000 genomes datasets (mitochondrial DNA, Y-chromosome, X-chromosomal, and autosomal) were downsampled to ten individuals per population because of the challenges explained fully in the phylogeny section 2.4.3. The downsampled datasets were then merged to the additional genomes in bcftools.

Both datasets used for phylogenetic reconstruction and other analysis were filtered the same way. The datasets were filtered out for minor allele frequency by setting threshold of 0,02 which exclude all sites in which no alternative alleles are called for any sample and all sites with minor allele frequency (MAF) of less than 2% in appendix A (command 14). However, for PCA analysis I used MAF =0.05. The choice of cutoff thresholds was influenced by the previous study on SNP data that showed that including singletons when inferring population structure can affect the true structure of the data (Linck and Battey, 2018).

Although the datasets are of different coverages, high coverage (autosomal, X-chromosome and Y-chromosome data) and the low coverage mitochondrial dataset, the same threshold value for missing data was used. The data was filtered for missing data by setting 0.2 thresholds which allowed only 20% missing data in appendix A (command line16). This filtering was performed to ensure that the data used portray a true structure of the genomes. Missing data are associated with plenty of uncertainties that prevent parallel comparison of genomes. Biallelic SNPs are the most commonly used makers in genetic analysis software. Command (line 15) in appendix A was

used to ensure that only biallelic were retained, the multiallelic, and indels were discarded. The indels we discarded because they carry no information about evolutionary history.

The data was also filtered for linkage disequilibrium (LD) because high number of linkage disequilibrium within datasets can affect the evolutionary analysis. The SNP in high linkage disequilibrium also contain no information about the main SNP. This process was carried out using command lines (22 and 23) in appendix A respectively in plink version 1.9 (Purcell *et al*, 2007). Command line 22 in appendix A compute a list of makers that are in linkage equilibrium by taking 50 SNPs at a time with a shift of 10 window size based on the threshold value of 0.5. The amount of the data filtered out depend on the threshold value, where a smaller value compared to the LD value allows larger amount of data pruning while larger threshold result in small amount of pruning. The last command (line 23) in appendix A recode the original data based on the file produced by (command 22) in appendix A which contains the information on the SNPs that remained after filtering. Table 2 below shows the SNPs statistics for various datasets used in different analyses.

Table 2: The statistics table showing the number of SNPs and samples before filtering and after filtering for various analyses.

		Types of datasets			
		X-chromosome	autosomal	mitochondrial	Y-chromosome
		Unfiltered datasets			
	Number of samples	983	983	983	453
	Number of SNPs	4960150	8900504	3892	41088
Analysis	Data after filtering				
Genetic Diversity	Number of samples		983		
	Number of SNPs		7662032		
PCA	Number of samples	983	983		
	Number of SNPs	2405064	7662032		
Admixture and pie plots	Number of samples	983	983		
	Number of SNPs	2405064	7662032		
	Number of samples	103	103	103	103

Phylogenetic trees	Number of SNPs	591786	7639271	119	2000
Genetic Networks	Number of samples			100	100
	Number of SNPs			119	2000
ABBA/BABA	Number of samples		983		
	Number of SNPs		7662032		

2.3. Genetic diversity

To determine the genetic diversity within Asian populations we used two methods (heterozygosity and nucleotide diversity) to overcome some uncertainties associated with methods of inferring genetic diversities i.e heterozygosity (Nielsen *et a*l, 2012). This study states that the multiple SNP calling is always associated with high-rate errors which may cause bias. However, to avoid this bias we applied allele filtering by filtering for missing data. Heterozygosity is known as the difference between paternal and maternal alleles. Vcftools version 4.1 was used to estimate both heterozygosity and nucleotide diversity (Pi). The input vcf files used contained all the individuals that belong to a single population. To estimate heterozygosity, we used –het option in vcftools, the --het parameter calculates the heterozygosity per individual basis, and the inbreeding coefficient F. The output file contained the observed homozygosity, expected homozygosity, number of sites, and F estimates of each individual. To obtain the population heterozygosity estimates, I first averaged the estimated individual's homozygosity and subtracted the total number of sites. This was done by taking the sum of the observed homozygosity and dividing it by the total number of sites. The nucleotide divergence was calculated using the parameter --window-pi 100,000. The program estimates the nucleotide diversity per site from multiple individuals. These values were given for each individual, so they were averaged at the end. Formula: O_HET = N_sites / O_HOM.

2.4. Population structure

2.4.1 Principal component Analysis

To test the effect of geography on the genomic relationships I conducted principal component analysis (PCA) (Alexander *et al*, 2009) on autosomal and X-chromosome datasets containing 983 individuals from ten populations using plink V1.9 (Purcell *et al*, 2007). PCA is a model-free(non-parametric) way of reconstructing population structure. Plink is a C++ whole genome sequence

program that is able to manipulate thousands of genome markers from thousands of individuals. The program can easily execute analysis from binary formats which is .bim/.bed formats. We first indexed the reference genome and the vcf files using tabix(Li, 2011) to ensure proper alignment and grouping of data. The indexed data were then converted into binary variant call format (BCF). We used the bcf files to create simple plink file formats. We conducted a PCA in plink from X-chromosome data with 2,405,064 SNPs and autosomal 7,662,032 SNPs. I then plotted the results using R. This process was done according to the tutorial published at https://www.biostars.org/p/335605/.

2.4.2. Admixture

To further determine the genomic relationships between South and East Asian genomes I, inferred population structure using admixture v1.3 (Alexander *et al*, 2015). This program uses a maximum likelihood estimation to infer individual ancestries and identify the number of clusters (K). We excluded the Chimpanzee, Neanderthal, and Denisovan genomes from the analysis because one of the admixture limitations is that it cannot detect clusters that are under-sampled (Alexander *et al*, 2015). In such a way that it also fails to pick up very divergent individuals. The program takes a plink format input. So, I first converted the vcf files into plink format. I then generated the binary biallelic genotype table (.bed files) using plink. The analysis was ran through a loop using (command 17) in the appendix A from K=2 to K=11.The analysis was performed with default cross-validation fivefold and a bootstrapping of 5 for each value of K from K=2 and K=11. We chose the cluster with the lowest cross-validation value for the best value of K. The results were visualized in the Linux terminal using the script plot admixture.r downloaded from https://github.com/speciationgenomics/scripts/raw/master/plotADMIXTURE.r. The output file containing the allele frequency for each population (.p file) , and the population information file specifying which population the individuals belong to as an input file were used as input files to plot the results (command 17) in appendix A.

2.4.3. Phylogeny

Lastly, to further understand the relationship between South and East Asian populations we reconstructed phylogenetic trees for X-chromosome and autosomal data using iqtree v1.6.12 (Nguyen *et al*, 2015). In this software, the maximum likelihood method searches for topologies with the highest probabilities by using model parameters (Nixon, 2001). However, for iqtree to compute a tree it first needs to find the best suitable substitution model for the data. The software

takes fasta, phylip, nexus and, clustal files as input. We converted the datasets into phylip format using a script vfc2phylip.py (Ortiz, 2019), which is freely available online.

I first tried running iqtree locally, but this was not possible due to the enormous memory requirements for such a large autosomal data set (983 genomes, 7,862,032 filtered SNPs). The same dataset was uploaded from my local Linux into the chpc server via ssh using command line 20. I then tried running iqtree with the same input file on the Lengau cluster at the CHPC (Center for high-performance computing) server, but this also failed. As I was not sure whether I had too much data (SNPs) or too many individuals (genomes), I pruned the autosomal SNP data set to reduce the number of SNPs using the vcfrandomsample option from different r values 0,1; 0.2; 0.12; 0.05;0.04 and 0.03 using the command line 19. This resulted in six different autosomal datasets with 285,370; 570,448; 341,181; 143,276; 114,093, and 85,843 SNPs respectively. Unfortunately none of these reduced data sets ran the iqtree analyses either, despite using the big,mem option which allows the use of large number of nodes on the chpc sever (the job submission script can be found in (Appendix B).

The same problem was encountered the with X-chromosome dataset. I then randomly downsampled the original(unfiltered) datasets to ten individuals per population plus the three additional hominin and outgroup genomes which sum up to a total of 103 genomes for the phylogenetic dataset. The downsampled data was subjected to the same filtering process. From the downsampled datasets, we computed phylogenetic trees from the X-chromosome dataset with 591,786 SNPs and the autosomal dataset with 7,639, 271 SNPs with a bootstrap value of 1000. Bootstrap estimation helps in assessing the uncertainty of the tree topology by estimating support values for each node (Felsenstein, 1985). The best fit model was the probability matrix from block reversed Blosum matrix (PMB+F+R5) for the autosomal data and (PMB+F+10) for the X-chromosome. Phylogenetic trees were also reconstructed for mitochondrial and Y-chromosome data. VCF files from mitochondrial (203 SNPs) and Y-Chromosome (2,000 SNPs) were converted to phylip format using the same script that was used to convert the X-chromosome and autosomal datasets. The best fit model for the mitochondrial data set was a combination of unequal base frequent and empirical base frequencies (TIM3+F) and the transversion and rate model (TVM+R3) for the Y-chromosome data.

2.4.4 Phylogenetic networks

Networks are better used to visualize the genealogical relationships of non-recombining markers at the intraspecific level (Cassens *et al*, 2005). Therefore, networks were inferred from mitochondrial and Y-chromosome data. I used networks for these markers because reconstructing phylogenies from intraspecific data is associated with small genetic distances and large sample sizes between populations (Bandelt *et al*, 1999). The networks were generated in PopART (Population Analysis with Reticulate Trees,Leigh and Bryant, 2015) V1.7 using median joining (Bendelt *et al*, 1999) as it is time efficient because it only counts the differences in the character states between sequences (Bandelt *et al*, 1999). The program takes alignments in phylip formats. The converted phylip files for mitochondrial and Y-chromosome data were used by PopART to draw the networks. The network samples were then colored according to populations by importing a trait file specifying the populations which population each sample belong to.

2.5 Testing for introgression

We used Dsuite (Malinsky *et al*, 2020) method to determine gene flow between Island populations and mainland populations, expecting to find a higher genetic affinity between mainland than Island populations. The method uses allele frequency estimate per site to allow multiple individuals to be used. Dtrios was used to estimate the D and F statistics from the autosomal VCF file with 983 individuals because the program does not require any prior knowledge about the populations under study. The D statistic is a parsimonious way of testing for gene flow between closely related species, the method assumes that the number of ABBA and BABA estimates should not be different if there is Incomplete lineage sorting (Durant *et al*, 2011).The Dtrio program calculates the statistics across biallelic SNPs of four Taxa: P1, P2, P3, wherein according to Durand *et al*, (2011), P1 and P2 represent the closely related species/population while P3 represents a population which could have gene flow with P1/P2. The fourth taxon O is the outgroup that represent a distant species. Chimpanzee was used as an outgroup in this study. Equal frequencies between BABA and ABBA suggest that there is no gene flow and according to Durant *et al*, (2011) resultant significant test suggest possible gene flow from either P2/P1 and P3.

To test the gene flow between mainland and island populations we used a population map file containing a list of all samples specifying the populations to which they belong. We conducted the test for Sri Lanka and Japan respectively. To test geneflow between Sir Lanka and closer mainland populations we chose a combination with the mainland (Telangana, Punjabi, Bengali, and Gujarati) at P1 and P3 keeping Sri Lanka at P3. For Japan (Vietnam, Han Chinese, Southern Han, and Dai Chinese) were used as combinations for P1 and P2.

The last set of tests was designed to test the regional geneflow signal depicted from admixture analysis in (Figure 6) results section. The two populations which show a combination of South and East ancestries (Bengali and Dai Chinese) we kept at P2. To test which East Asians (Japan, Vietnam, Han Chines, Southern Han, and Dai Chinese) were responsible for the geneflow with Bengali, we placed the South Asians at P1 and any East Asian at P3. For Dai Chinese, we placed East Asians at P1 and South Asians at P3. This is because in the program Dsuite, the statistics are always positive for each trio suggesting gene flow between P2 and P3.

3. Results

3.1. Genetic Diversity

3.1.1. Population heterozygosity and nucleotide diversity

It was found that some South Asian populations (Gujarati and Bengali) have lower heterozygosity than East Asians, with Gujarati having the lowest among South Asians Figure 2A. Among East Asians, Vietnam had the highest heterozygosity, followed by Han Chinese, Dai Chinese and Southern Han Chinese, with Japan, being the lowest Figure 2A. South Asian populations all had more nucleotide diversity than East Asian populations. When compared to other South Asians, Gujarati has the lowest and Bengali has the highest nucleotide diversity. In East Asia, the genetic diversity of the South China (Vietnam and Dai Chinese and Southern Han Chinese) populations is relatively higher than the North China (Han Chinese and Japan. Vietnam has the most nucleotide diversity in East Asia, while Japan has the lowest.

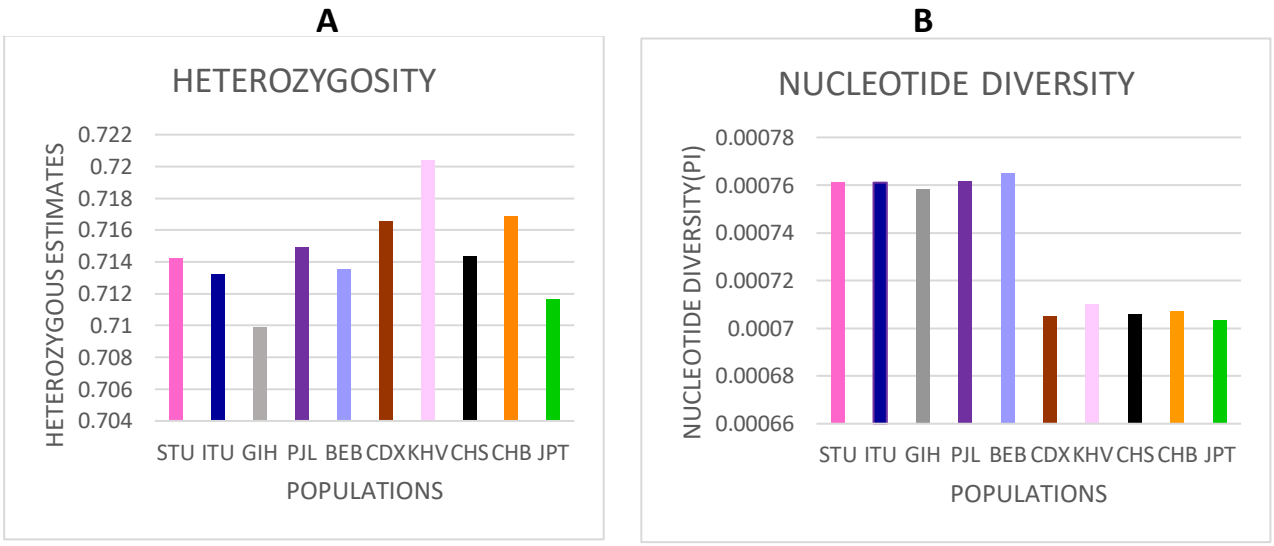


Figure 2: Genetic diversity: graphs showing population-wide heterozygosity estimates(A) and nucleotide diversity(B) of South and East Asian populations. Sri Lankan Tamil (STU), Indian Telugu (ITU), Gujarati Indians in Houston (GIH), Punjabi Pakistan (PJL), Bengali Indian (BEB), Chinese Dai (CDX), Kin Vietnam (KHV), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB) and, Japanese in Tokyo (JPT),

3.2. Population structure of x chromosome and autosomal datasets

3.2.1. Principal component analysis

The principal component analysis in Figure 3 was constructed from the X-chromosome dataset with 2,405,064 SNPs. The PCA plots were orientated by plotting PC1 against all other PCs with PC1 on the x-axis. The populations in East Asia (and South Asia) overlap significantly. Bengali cluster towards East Asians in all PCS suggesting that among other South Asian populations Bengali shows a closer relation with East Asians. PC3 suggests a structure within Sri Lanka.

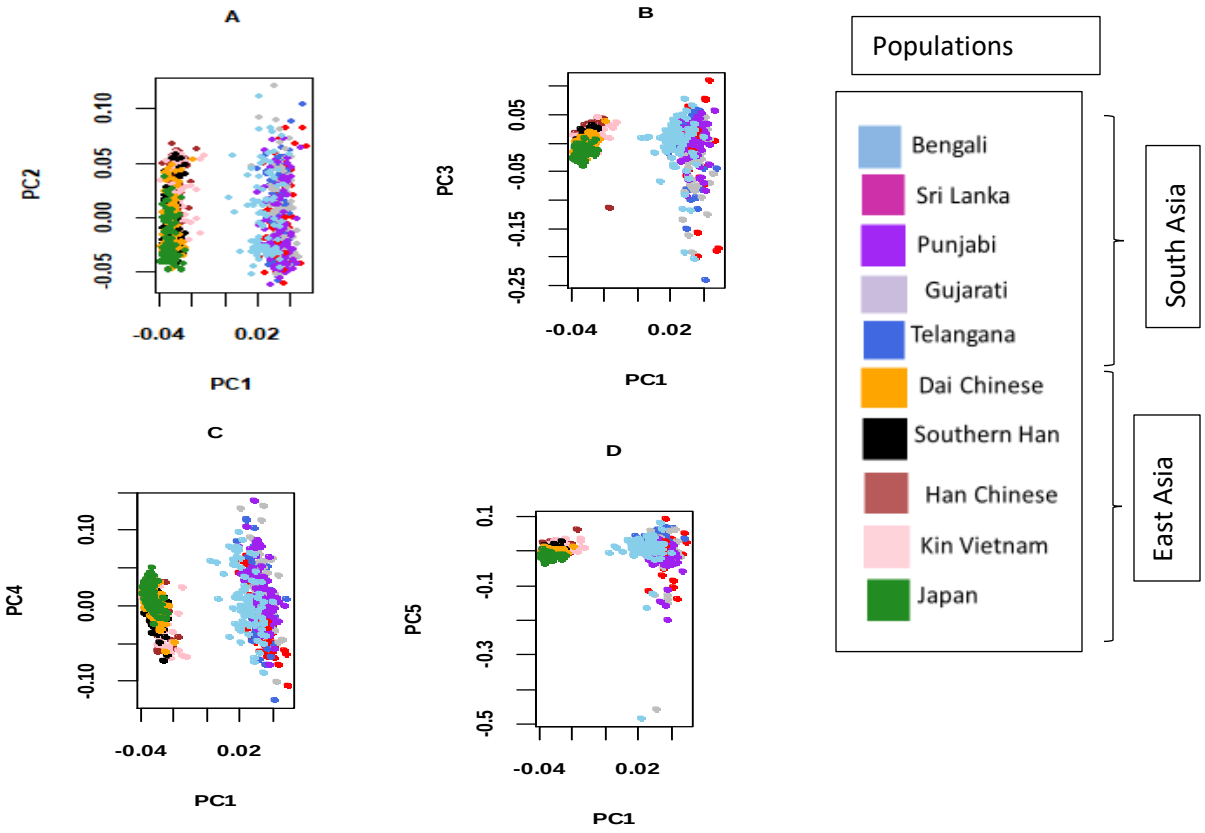


Figure 3: Principal Component Analyses clustering computed from X- chromosomes SNPs from South and East Asian populations. South Asians, Sri Lankan Tamil (STU), Indian Telugu (ITU), Gujarati Indians in Houston (GIH), Punjabi Pakistan (PJL), and Bengali Indian (BEB). East Asians Chinese Dai (CDX), Kin Vietnam (KHV), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB) and, Japanese in Tokyo (JPT).

In Figure 4, the plots (A, B, C, and D) were reconstructed from the autosomal dataset. There is an overlap between South Asians and East Asians in all the PCAS. PC1 and PC2 reveals a distinct structure within East Asian populations. Han and Southern Han cluster together, while Dai Chinese and Kin in Vietnam cluster closer together suggesting a closer relationship between them. Furthermore, there is a geographical distribution within East Asian populations, South China population (Kin in Vietnam, Dai Chinese). In South Asia PC1 and PC3 depict the population structure within South Asian populations. Bengali cluster towards East Asian populations in all the PCAs. The plot in PC4 and PC5 outlines a structure within Telangana (ITU). PC1 and PC4 suggest a differentiation and structure within Gujarati. Lastly PC1 and PC5 depicts a North to South separation of South Asian populations.

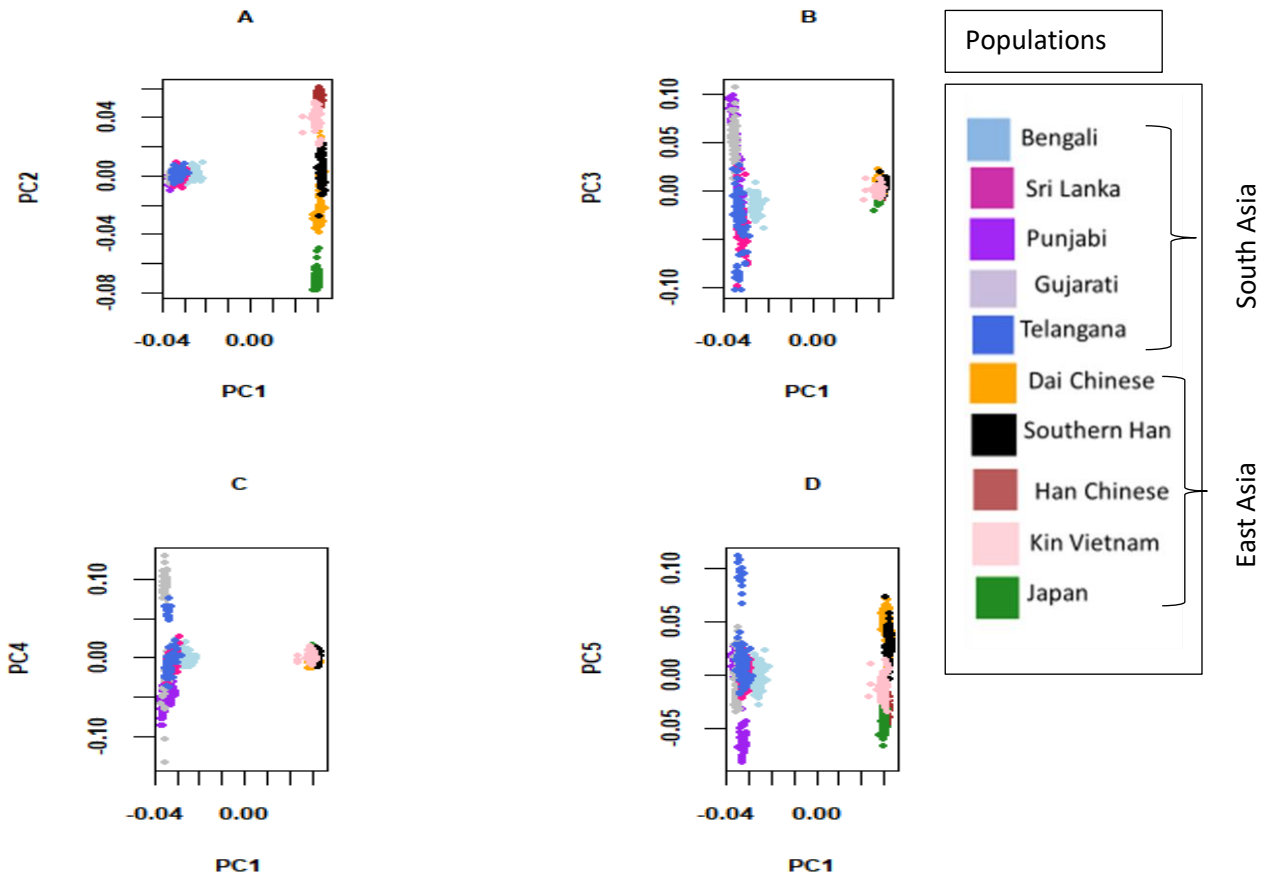


Figure 4: Principal Component Analyses computed from Autosomal SNPs of South and East Asian populations PCAs constructed from and Autosomal 7,662,032 SNPs. Abbreviations: Sri Lankan Tamil (STU), Indian Telugu (ITU), Gujarati Indians in Houston (GIH), Punjabi Pakistan

(PJL) and Bengali Indian (BEB). Chinese Dai (CDX), Kin Vietnam (KHV), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB) and, Japanese in Tokyo (JPT).

3.2.2. Admixture plots

The Admixture analyses were carried out from Autosomal and X-chromosome data from South and East Asian populations consisting of 983 samples from 1000 genomes project. The analysis was carried out using varying ancestral populations from K=2 to K=11(not shown), from the default cross-validation and bootstrapping of five. The best K value from the log-likelihood was five for both autosomal and X-chromosome data. The best K values were obtained by plotting the cross-validations of each value of K in Figure 13 and Figure 14 respectively. From the admixture analysis in Figure 5 at K=2, there is a clear separation of the two regions South (light blue) and East (red) Asia. K=3 suggest different ancestral composition that is strongest in Japan. In comparison to the rest of South Asians, Bengali has a significant level of admixture with East Asians at K=2. The level of admixture (outlined by the blue colour) decreases from Vietnam (KHV) to the Japanese (JPT) in East Asia at K=2.

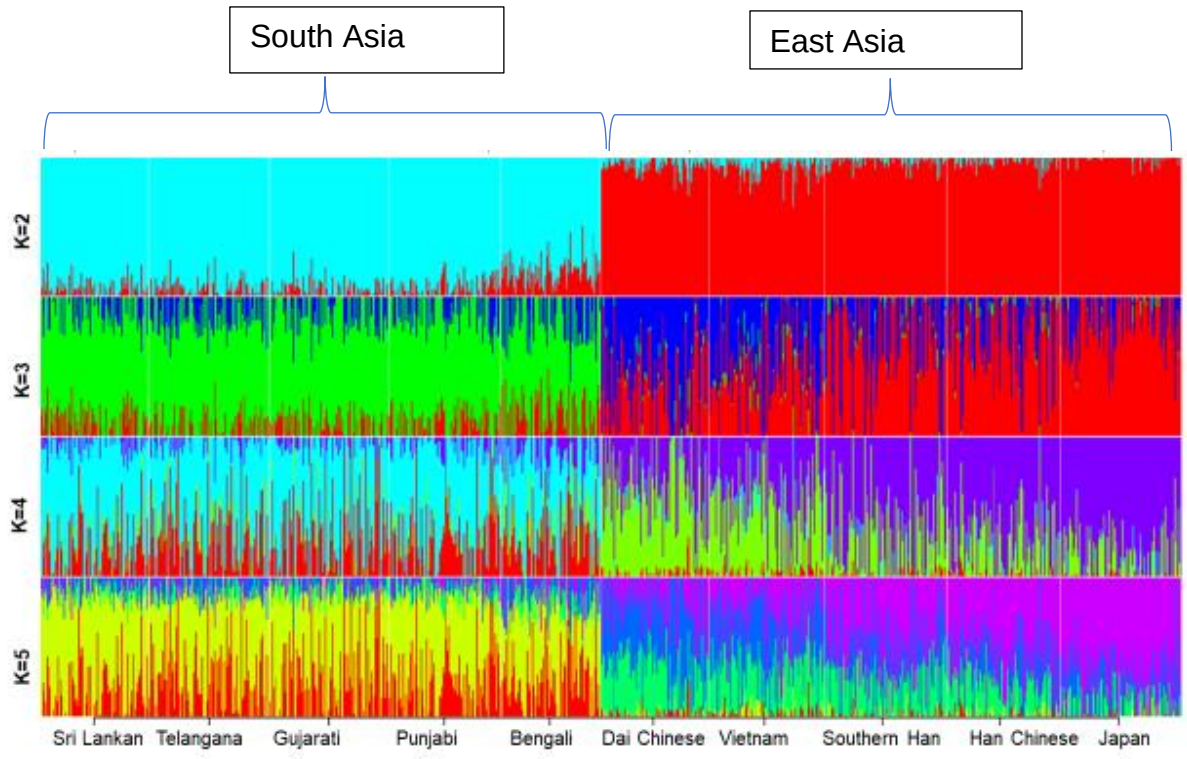


Figure 5: Admixture analysis showing admixture proportions of X-chromosome data from South and East Asian populations.

To investigate the genetic structure within South and East Asia I, carried out admixture analysis from autosomal dataset containing 7,662,032 SNPs. Figure 6 at K=2 shows a clear differentiation between South (light blue) and East Asian (red) populations. At K=3 there is a structure within East Asian populations, the populations are divided into three ancestral groups: the first consist of Dai Chinese and Vietnamese, the second group: Han Chinese and Southern Han, and the last group Japan. K=3 and 4 indicates that Japan shares the most ancestral diversity with Han Chinese. At K=4 there is a structuring within South Asians grouping the populations into three groups: the first group consist of Sri Lanka and Telangana (ITU), the second group include Gujarati and Punjabi and the last group includes Bengali Japan shows a clear distinct ancestral composition than the rest of East Asian populations at K=5. The structure of Bengali at each value of K suggest a possible gene flow with east Asians. K=2 also suggests that there is gene flow between Vietnam (KHV), Han Chinese and, South Asians presented by the light blue colour.

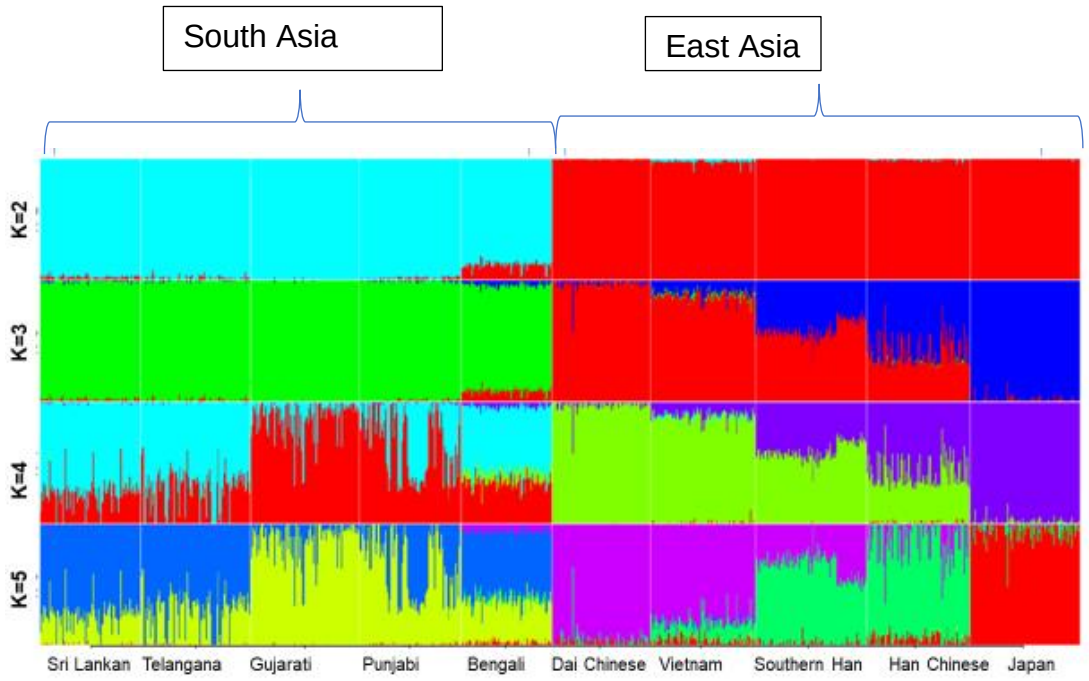


Figure 6: Showing Admixture analyses of autosomal SNPs data from South and East Asian populations. The analysis was carried out from dataset containing 7,662,032 SNPs. The populations are divided by the vertical lines.

The pie charts in Figure 7 were computed based on the ancestral compositions suggested by the admixture plot in Figure 6 at K=5. The color proportions for the pie charts differentiate South from East Asian populations. However, the charts also suggest that Bengali shares the ancestral diversity with East Asians. South Asian populations are characterized by two ancestral proportions given by blue and yellow colors. According to the ancestral compositions, East Asians can be grouped into three groups: the first consist of Dai Chinese and Vietnamese, the second group: Han Chinese and Southern Han, and the last group Japan. The pie charts suggest sharing of ancestral compositions between Bengali and Dai Chinese.

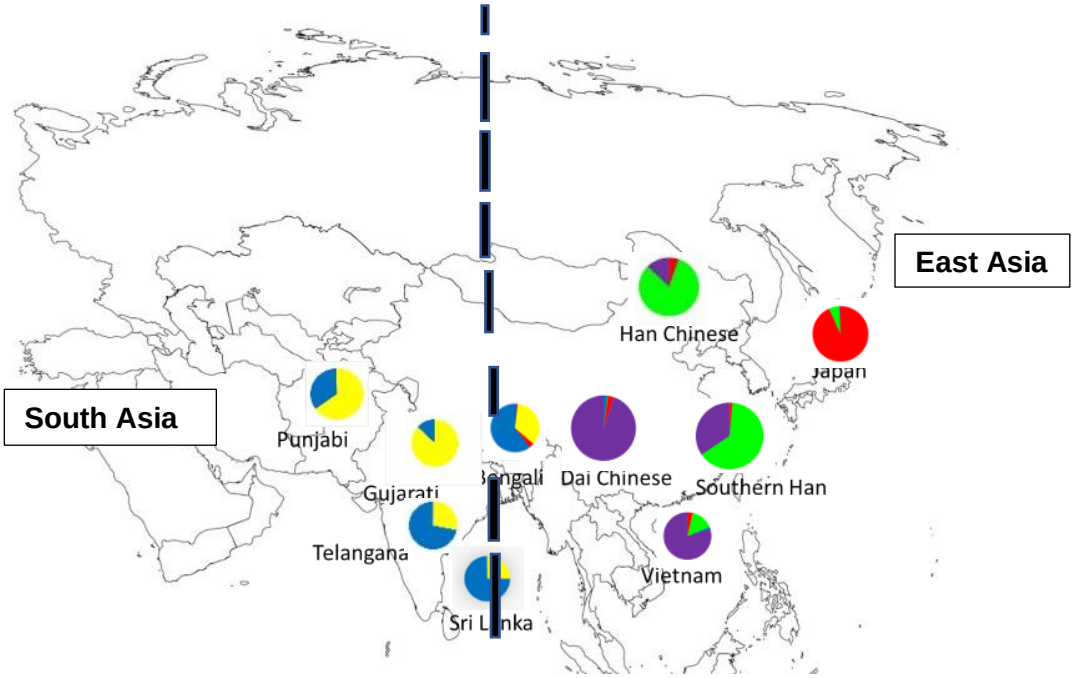


Figure 7: Showing geographical locations of the Asian populations with admixture proportions computed from autosomal data, K=5. The pie charts were colored according to the colors predicted by the admixture program for K=5. The dotted line divides South Asia from East Asia.

3.2.3 Phylogeny

3.2.3.1 Phylogenetic trees for X-chromosome and autosomes

The phylogenetic tree in Figure 8 was reconstructed from X-chromosome data including ten samples from each population, using iqtree. A Chimpanzee genome was used as an outgroup for better rooting of the trees. The tree suggests a closer relationship between Japan and Han Chinese. The intermediates genomes between the South and East Asian clade suggest geneflow between the two regions.

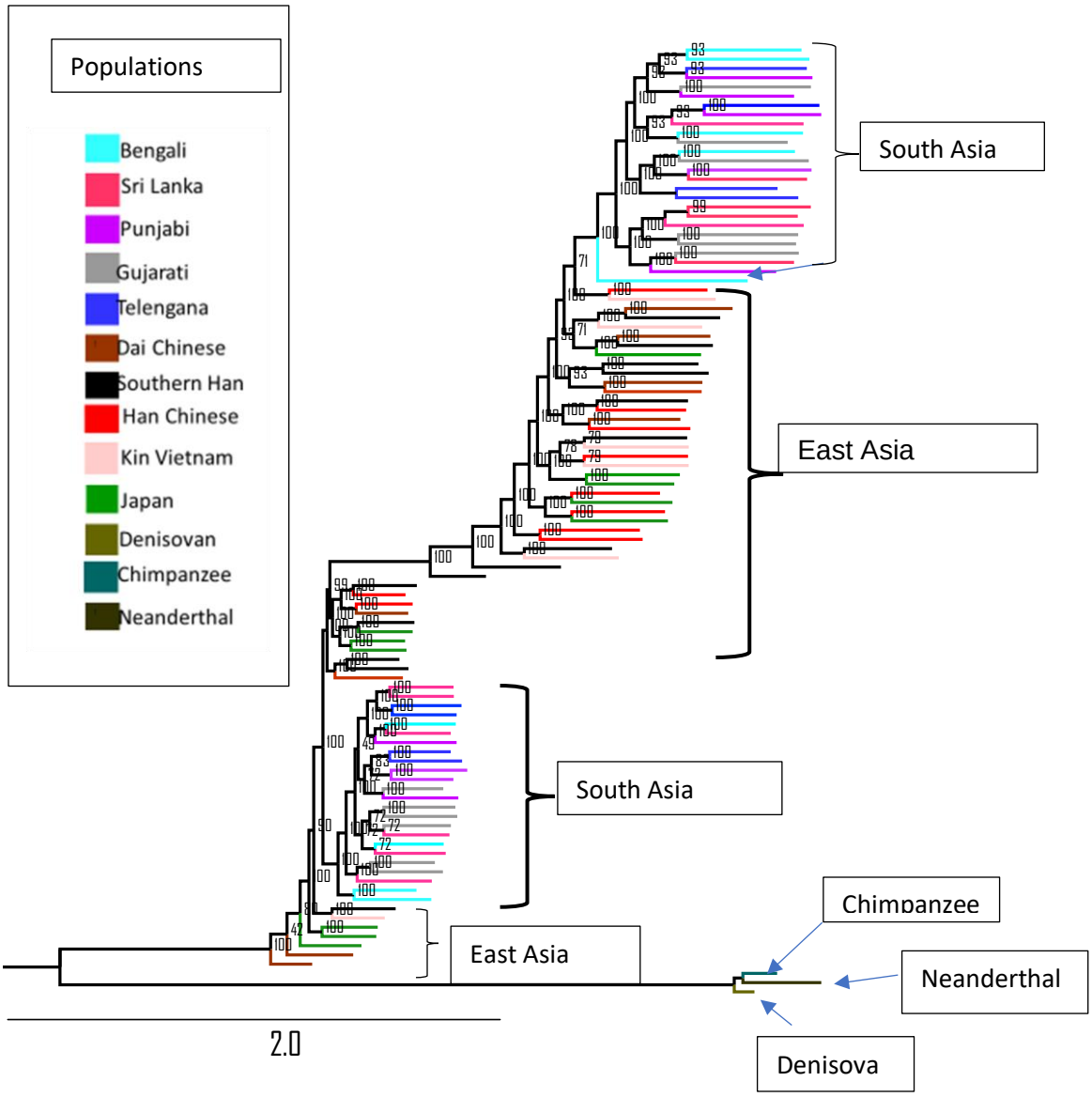


Figure 8: phylogenetic tree reconstructed from X-chromosome data based on the PMB+F+10 model in iqtree. The plot shows a phylogenetic relationship between South and East Asian genomes. The samples were colored according to the populations. A chimpanzee genome was used as an outgroup to root the tree. The arrow was used to indicate the intermediates genomes between South and East Asia.

In Figure 9 there is a paraphyletic relationship between South and East Asian genomes. Most Sri Lankan genomes do not form part of the South Asian clade. Some of the Bengali genomes also fall outside the main South Asian clade suggesting a possible gene flow with East Asians. East Asians form a clear polyphyletic clade. However, two Japan genomes fall outside this clade. Japan genomes are paraphyletic with just two genomes clustering outside this inclusive clade. The substructure within the Asian clade suggests a closer relationship between Dai Chinese and Vietnam. Southern Han is closely related to Han Chinese than other East Asians. There is more structure within the autosomal tree than X-Chromosome tree.

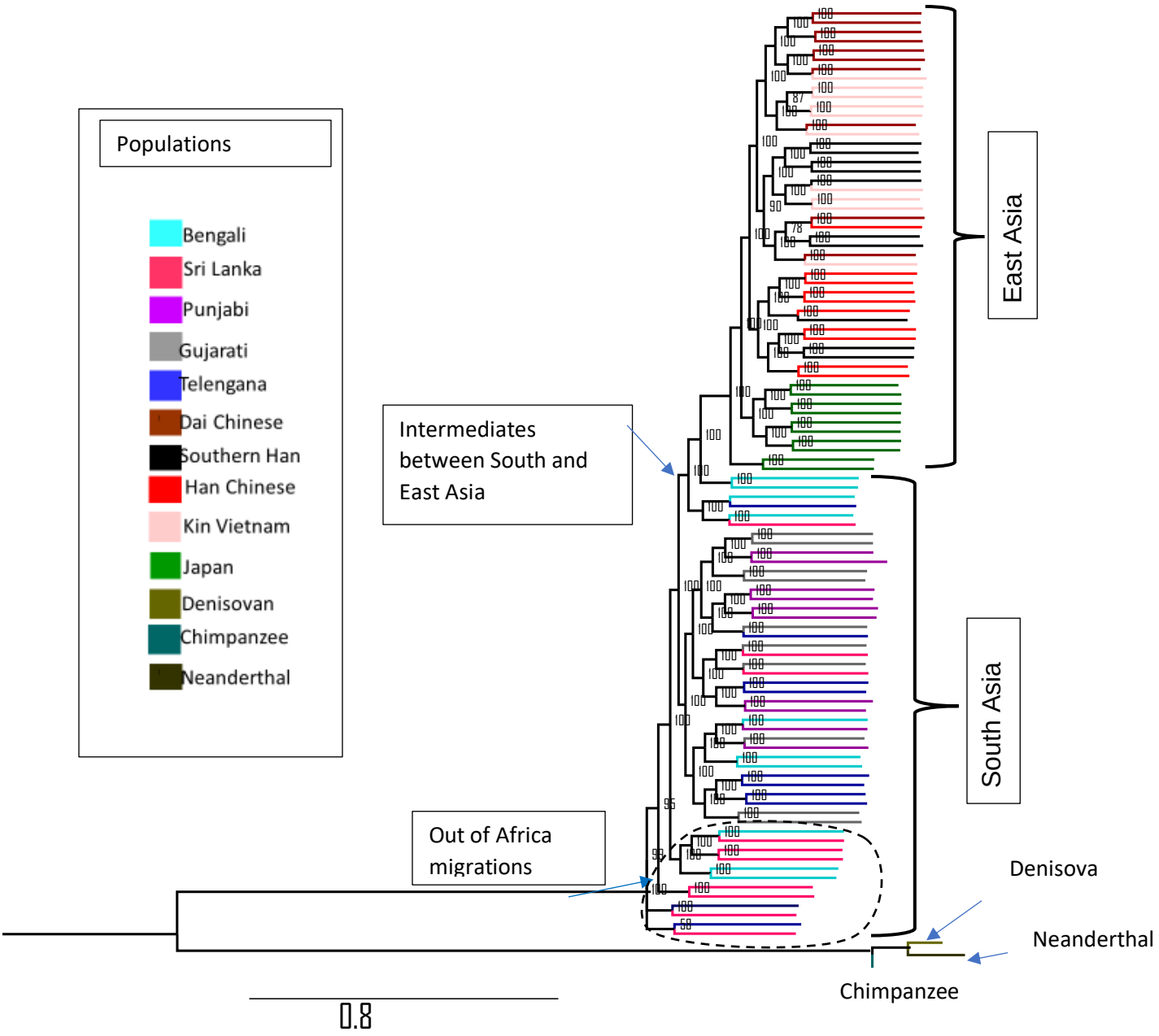


Figure 9: showing a phylogenetic tree reconstructed from autosomal dataset using maximum likelihood based on PMB+F+R5 model from South and East Asians genomes. Ten individuals per population were chosen randomly to reconstruct the tree due to factors explained in the method section. Three outgroup genomes were included namely: Neanderthal, Denisova and Chimpanzee. The tree was colored according to the populations the individuals belong to. Some genomes are not clustered within their clade of origin (labelled as intermediates) on the plot.

3.2.3.2. Mitochondrial and Y-chromosome structure

The phylogenies in Figure 10 were reconstructed from mitochondrial dataset with 119 SNPs. The phylogenetic network in Plot A was used to determine the major clade South and East Asian clade supported by the 100% bootstrap values. There are a lot of shared haplotypes between South and East Asia in general and also a lot of shared haplotypes between populations within each region. An intermediate positioned genome from Telangana (blue arrow, Figure 10A and B) suggests it is a hybrid genome possibly showing ancestral geneflow between South and East Asia. In East Asia, Vietnam and Dai Chinese haplotypes cluster together in Figure 10, they also can be found elsewhere in the network, even on the South Asia side.

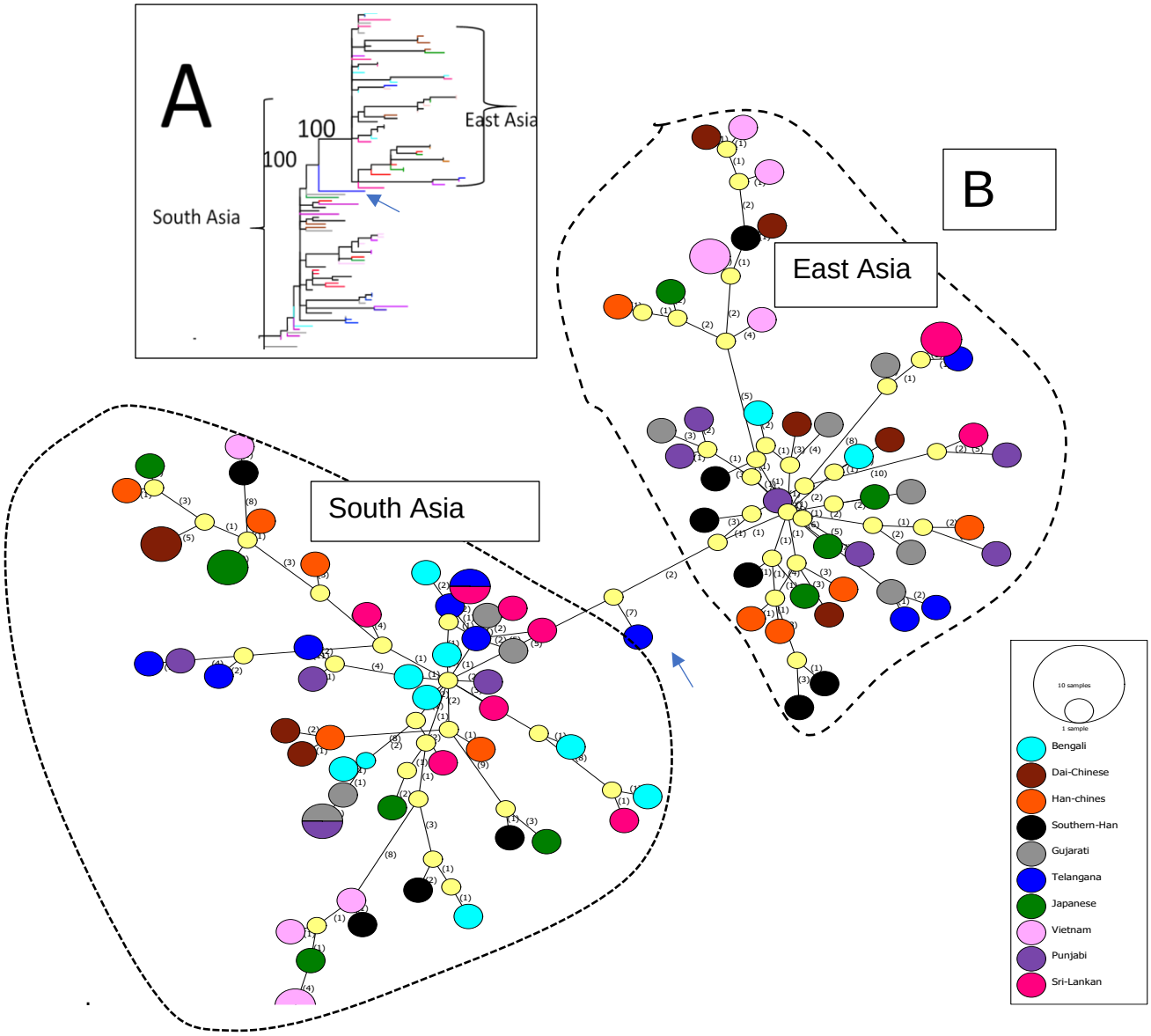


Figure 10: A shows a phylogenetic tree reconstructed from mitochondrial dataset of ten Asian population using the maximum likelihood method based on (TIM3+F) model in Iqtree. The yellow color was used to represent the original color vertex. B: shows a network reconstructed from mitochondrial data of South and East Asian populations based on median joining method. The samples are colored according to the populations they belong to. The numbers represent the mutations between each sample. The dashed lines were used to outline the two major clade South and East Asian based on the bootstrap support values in plot A.

In Figure 11, the key applies to both plot A and plot B. The tree in plot A was used to determine the major clade South Asia and East Asia based on the high bootstrap support values. The numbers (1-9) on the branches represent the bootstrap support values. 1 suggest a 100 % support bootstrap value. Although, two genomes are not clustered within their cluster of origin, there is a clear distinction between the two regions. In East Asia, just one Telangana genome clustered together with East Asians. A sub-clade of East Asian genomes (five Japanese, one Han Chinese, and one Southern Han) within the main South Asian cluster form a mini cluster with one Bengali genome. The sub-cluster within the East Asian clade suggests a closer relationship between Vietnam, Dai Chinese, and Southern Han Chinese.

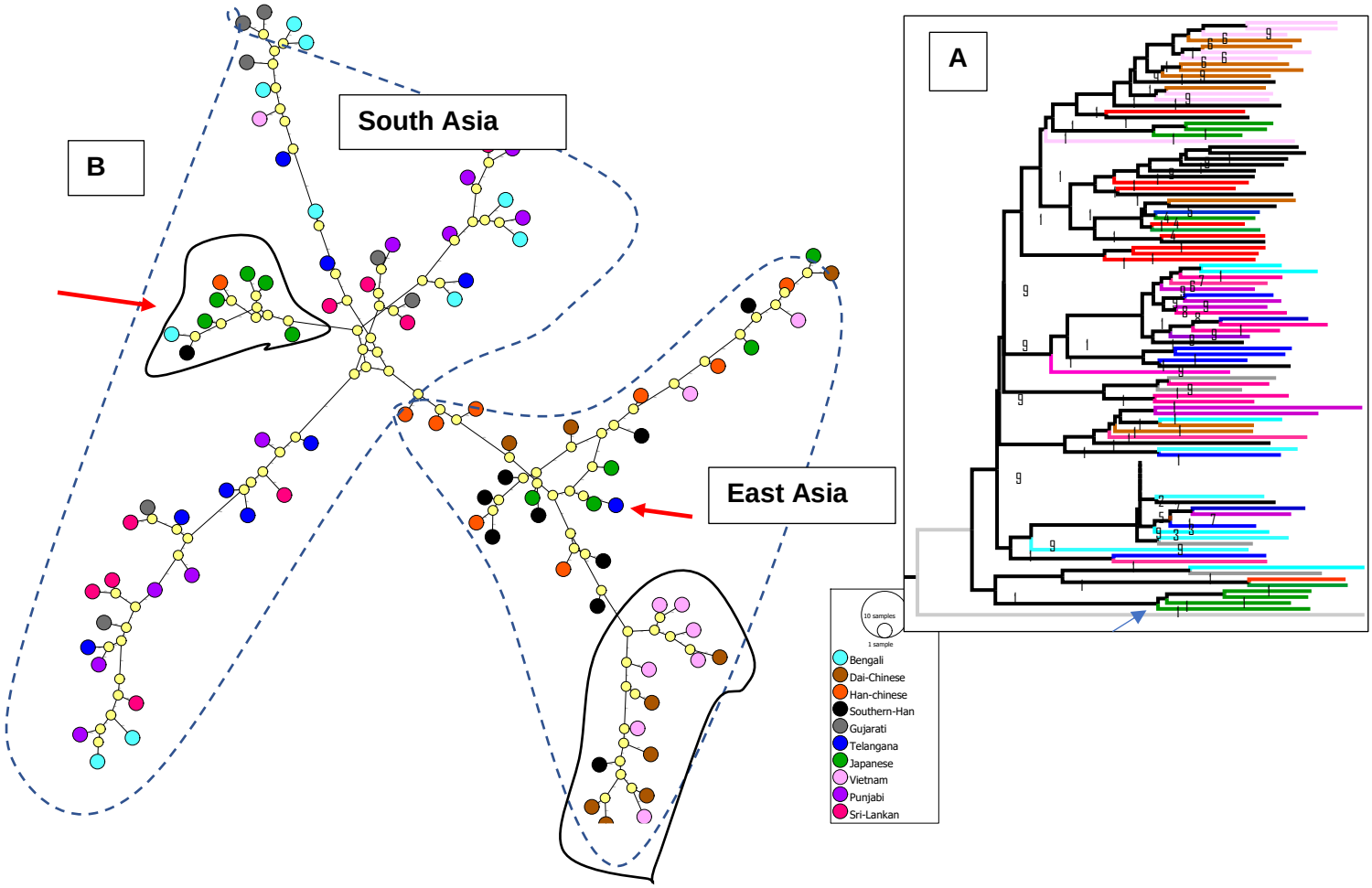


Figure 11: A: Shows a phylogenetic tree reconstructed from the Y-chromosome dataset based on the transversion and rate model (TVM+3) in Iqtree. B: shows a network reconstructed from the same Y-chromosome dataset. I used the yellow color to represent the original color vertex. The dashed lines were used to outline the major clades South and East Asian, and the solid lines were used to outline the subclades based on the support bootstrap values. Each genome in both plot

A and plot B was colored according to the population color specified in the legend. The arrows were used to point out the Telangana (ITU), Japanese, Han Chinese and Southern Han Chinese haplotypes that are not clustering within their clade of origin.

3.3 Gene flow

A significant Z-score is any value between +3 and -3. There was significant introgression between island and Mainland populations of Asia in four cases (Table 3). The results suggest that there is a significant gene flow between Gujrati and Sri Lanka. The results also suggest significant gene flow between Southern Han and Japan.

Table 3: Introgression between island and closer mainland populations in the South and East Asia independently. Populations as abbreviated: Sri Lankan Tamil (STU), Indian Telugu (ITU), Gujarati Indians in Houston (GIH), and Bengali Indian (BEB). EA: East Asians, Chinese Dai (CDX), Kin in Vietnam (KHV), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB), and Japanese in Tokyo (JPT). The populations that are having gene flow are highlighted by the yellow colour.

P1	P2	P3	Dstatistic	Z-score	p-value	f4-ratio	BBAA	ABBA	BABA
ITU	GIH	STU	0.00362601	3.60676	0.000155	16.6793	176381	175883	174612
PJL	GIH	STU	0.00231443	3.01979	0.0012648	0	177757	175846	175034
CDX	CHS	JPT	0.00821543	7.23276	2.37E-13	0.497837	161344	160227	157616
KHV	CHS	JPT	0.0085855	7.1267	5.14E-13	0.509517	161217	160614	157880

Results in (Table 4) confirms the prediction that was observed from admixture results which suggested a possible gene flow between Bengali and East Asians. There was significant gene flow between Bengali and East Asian (Dai Chinese, Vietnam, Southern Han Chinese, and Japan) populations. The results also suggest that the gene flow is from East Asia into Bengali because all the test between Dai Chinese and South Asians were not significant (not shown). The results are inconclusive as to which of the East Asians is the gene flow coming from. The zero P-values suggest that the tests are significant.

Table 4: Introgression between Bengali and East Asian populations. Abbreviations: Sri Lankan Tamil (STU), Indian Telugu (ITU), Gujarati Indians in Houston (GIH), Punjabi Pakistan (PJL), and Bengali Indian (BEB). EA: East Asians Chinese Dai (CDX), Kin in Vietnam (KHV), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB), and Japanese in Tokyo (JPT). The colour yellow was used to indicate the populations that are having gene flow.

P1	P2	P3	Dstatistic	Z-score	p-value	f4-ratio	BBAA	ABBA	BABA
GIH	BEB	CDX	0.00972165	9.35388	0	0.0963346	185172	176453	173056
PJL	BEB	CDX	0.0103144	10.0769	0	0.101859	184939	177021	173406
STU	BEB	CDX	0.0109152	12.0965	0	0.106921	183502	176703	172887
ITU	BEB	CDX	0.0113677	11.1454	0	0.110821	183796	176707	172735
GIH	BEB	CHB	0.00929505	9.10188	5.42E-20	0.0941522	184950	176412	173162
ITU	BEB	CHB	0.0111801	10.7089	0	0.111115	183620	176711	172804
PJL	BEB	CHB	0.00970047	9.47845	0	0.0980852	184679	176941	173541
STU	BEB	CHB	0.010865	11.2596	0	0.108364	183357	176739	172939
GIH	BEB	CHS	0.0095246	9.4331	0	0.0944357	185028	176436	173107
ITU	BEB	CHS	0.0112917	10.9968	0	0.110017	183676	176715	172769
PJL	BEB	CHS	0.0100612	9.80409	0	0.0994627	184782	176991	173465
STU	BEB	CHS	0.0109458	11.523	0	0.107054	183409	176737	172910
GIH	BEB	JPT	0.00889432	8.87775	3.25E-19	0.0880258	185119	176198	173092
GIH	BEB	KHV	0.0092798	8.96816	1.63E-19	0.0987571	184840	176431	173187
ITU	BEB	JPT	0.0108059	10.6035	0	0.104944	183795	176504	172730
ITU	BEB	KHV	0.0109919	10.737	0	0.114863	183483	176705	172863

The results in (Table 5) outlined significant gene flow between South and East Asian populations. In East Asia significant levels of gene flow were observed between Dai Chinese and Southern Han Chinese. In south Asia, significant levels of gene flow were observed between Bengali and the north Indian populations (Punjabi and Gujarati).

Table 5. Introgression between South Asian populations and East Asian populations respectively. Abbreviations: Sri Lankan Tamil (STU), Indian Telugu (ITU), Gujarati Indians in Houston (GIH), Punjabi in Pakistan (PJL), and Bengali Indian (BEB). EA: East Asians Chinese Dai (CDX), Kin in

Vietnam (KHV), Southern Han Chinese (CHS), Han Chinese in Beijing (CHB), and Japanese in Tokyo (JPT). The colour yellow was used to indicate the populations that are having gene flow.

P1	P2	P3	Dstatistic	Z-score	p-value	f4-ratio	BBAA	ABBA	BABA
JPT	CHB	CDX	0.007257	10.1533	0	0.449938	161127	160535	158222
CHB	CHS	CDX	0.004464	4.80804	7.62E-07	0.500079	160153	159202	157787
PJL	GIH	ITU	0.002501	3.31165	0.000464	0	177366	175953	175075
JPT	CHB	KHV	0.0067223	9.68116	0	0.646304	161399	160514	158371
CHB	CHS	KHV	0.003763	4.35766	6.57E-06	1.01442	160625	159159	157966
STU	PJL	BEB	0.003181	4.243	1.10E-05	3.21066	177100	176579	175459
ITU	PJL	BEB	0.00236	3.08489	0.001018	15.6523	177246	176293	175463
STU	GIH	BEB	0.005469	5.62079	9.50E-09	5.50452	177019	176486	174566
ITU	GIH	BEB	0.004649	4.56613	2.48E-06	30.7357	177194	176163	174533

4. Discussion

4.1 Genetic diversity

Understanding the genetic diversity among human populations is one of the important aspects of biology. Genetic variability can be affected by various factors including geography, culture, and other factors. Asia is suitable for such kind of study because of its geography and cultural diversity. Low genetic diversity means that there is a limited variety of alleles for genes within the species (Human). The first aim was to test whether regions that were colonized first (South Asia) have higher genetic diversity than those colonized later (East Asia). The nucleotide diversity results support this, a South Asians had high nucleotide diversity than East Asia populations Figure 2B. However, the heterozygosity results suggest high genetic diversity in some of the East Asian compared to South Asians.

In south Asia, Bengali had the highest (0.00076) nucleotide diversity and Gujarati (0.000758) the lowest. Sri Lanka (island) does not attain either low levels of heterozygosity or nucleotide diversity. In almost all results This observed structure could also be attributed to the fact that the Gujarati population sampled for this study was drawn from the United Kingdom (UK). Both Gujarati and Sri Lankan populations were sampled in the UK, which explains the latter relationship. (PCA, admixture, phylogenetic trees, networks), there is clear evidence that Sri Lanka does not form a

separate cluster from Telangana (the closest mainland population). This might be because there was a continuous introgression between Sri Lanka and the closest mainland populations due to the small distance between India and Sri Lanka. This must be the reason for high genetic diversity in Sri Lanka. However, this gene flow is not detected in the ABBA-BABA analysis. Result in Table 3 but all other evidence points to there being substantial gene flow between Sri Lanka and Telangana, supporting the hypothesis stating that there should be gene flow between Sri Lanka with the closest main land population. The difference between the ABBA-BABA results and the rest of the results could be due to a technical error, as there is no combination of Telangana and Sri Lanka in any of the ABBA-BABA combinations. In South Asia, Punjabi had the highest heterozygosity followed by Sri-Lankan, Bengali, Telangana, and Gujarati populations. Low levels of heterozygosity observed in Gujarati might have resulted from the observed structuring of the Gujarati at PC4 in Figure 4.

Vietnam had the highest heterozygosity in East Asia, followed by Han Dai Chinese, Southern Han, and Japan. Asia Vietnam (0.00071) had the highest and Japan (0.000701) had the lowest nucleotide diversity. According to Pan and Xu, (2020), the regional migrations and admixture introduced genetic variations among East Asians. High levels of genetic diversity were expected in the Vietnamese population because the region was colonized by early Indian migrations (Underhill, 2004). This is confirmed by significant levels of gene flow between Bengali and Vietnam (KHV) in Table 4. According to Liu *et al,* (2020), Vietnam harbors multiple sources of genetic diversity. This diverse profile of Vietnam was also confirmed by (Macholdt *et al*, 2020).

The different trends observed between heterozygosity estimates and nucleotide diversity also might have been influenced by the different approaches we used. We used a sliding window approach for nucleotide diversity and the averaging approach for the heterozygosity estimates. The sliding widow approach was used to estimate nucleotide diversity to reduce the bias associated with large datasets because, according to Ramakrishna, (2013) calculating nucleotide diversity on a very large number of markers per SNP estimate can be noisy. Furthermore, SNPs which are closer together are not always independent due to recombination (Bush and Moore, 2012). Unlike the sliding window approach used to estimate nucleotide diversity, the method we used to estimate heterozygosity compares multiple loci at the same time which might results in over estimations. Both nucleotide diversity and heterozygosity estimation are sensitive to

population sizes and according to (Nielson *et al,* 2012; Hague and Routman, 2016) the smaller size produce larger estimates. However, these results suggest that the differences observed is not because of population size because based on the number of samples used in this study some populations with a larger number attained high genetic diversity.

4.2 Genetic structure of Asia

4.2.1 Genomic structure in South Asian Populations

Human migrations are known to have involved complex processes that reduced genetic diversity gradually (Vigilant *et al*, 1991; Nei *et al*, 1993, Jorde *et al*, 1997). Although South Asia experienced a lot of migrations it was suggested that the development of caste systems brought isolation between populations. I predicted that South Asia is highly structured because of the current caste system. However, from the results the caste system isolations are reflected within populations while the relationship between populations reflects the ancestral variation. The principal component analysis in Figure 4 PC4 and PC5 suggest a structure within Telangana. This observed structure within Telangana might have been shaped by the migrations, because the northern part of Telangana was colonized by the Indo-Aryans populations from Europe migrations while the southern part is dominated by out of Africa populations. It is also most likely that this observed structure within Telangana indicate endogamy. However, it is difficult to make any conclusions since no information is known regarding caste from the sampled genomes. The results in Figure 4 at PC4 and admixture results depicted a structure within Gujarati.. Gujarati retained its original genetic structure because according to Cordaux *et al*, (2004), the colonization of Gujarati also involved recent migration by Indo-European speakers. This observed structure could also be attributed to the fact that the Gujarati population sampled for this study was drawn from the United Kingdom (UK). Both Gujarati and Sri Lankan populations were sampled in the UK, which explains the observed relationship from (PCA, admixture, phylogenetic trees and networks) results.

Despite the current caste systems in South Asia, the gene flow results in Table 5 suggests that there is still more gene flow within the populations in South Asia. This prediction is also supported by the phylogeny results in Figure 9. Although Punjabi is one of the populations that consist of

high number of caste systems, the principal component analysis in Figure 4 at PC1 and PC4 suggest a close relation between Gujarati and Punjabi. This close relationship was also supported by the phylogenetic results in Figure 9 and the admixture results at K=4 in Figure 6. The similarity between Punjabi and Gujarati might have resulted from migrations because both populations have experienced the migrations from European populations.

4.2.2 Genomic structure in East Asia

Despite the cultural practices in South Asia, the results suggest that East Asia is highly structured compared to South Asia. The distinct structuring within the East Asian populations depicted by the principal component analysis results in Figure 4 at PC2, the phylogenetic results in Figure 9 and the admixture results in Figure 6 at K=5. Principal component analysis at PC2 in Figure 4 suggest the geographical distributions of the populations. The geographical distribution of populations is also supported by FST results from (Wang *et al*, 2018). The observed structure also suggest that East Asian populations are isolated.

The genetic difference within East Asians populations was also observed from the heterozygosity results in Figure 2A. These results outlined a relatively higher genetic diversity in Southern China (Vietnam, Dai Chinese and Southern Han Chinese) populations than Northern China (Han Chinese and Japan) populations. A high level of heterogeneity in Sothern Chinese populations than Northern population was also found in (Cavalli-Sforza, 1998).

Dai Chinese is more related to Vietnam than any East Asian populations in Figure 9. This relationship was supported by the principal component analysis results in Figure 4 PC1 and PC2, the phylogeny results in Figure 9 and the sub-clade within the East Asian clade in (Figure 11). These results are consistent with the F3 results from Wang *et al*, (2018) that suggested a possible common origin for Vietnam and Dai Chinese populations. The Han expansion might have added to this similarity between Vietnam and Dai Chinese because both populations were colonized by Han migrations. The colonization of Dai Chinese and Vietnam by Han Chinese migrations was supported by significant gene flow between Han Chinese and Vietnam or Dai Chinese in Table 5.

The PCAs in Figure 4 at PC1 and PC2 also suggest a closer relationship between Han and Japanese. This relationship is also supported by the admixture plots in Figure 6. A similar relationship was observed from genomic structure analysis by previous studies (Nei,1996; HUGO Pan Asia, 2009; Chen *et al*, 2009; Kim *et al*, 2020). The relationship between Han and Japan could be because according to the previous study (Nakazawa, 2017), Japan originated from East Asian populations. The significant gene flow between Han Chinese and Japan outlined by the gene flow results in Table 3 might have caused similarity between them. Genome wide study on single nucleotide polymorphism traced the common ancestry of Japan and Han back to 30,000 to 36,000 years ago (Wang *et al*, 2018). These findings suggest that Japan did not originate from Han Chinese which is consistent with previous studies (Nakazawa, 2017; Chooke *et al*, 2020) that suggest possible Southeast origin.

4.2.3 Colonization of Islands

The genetic structures of the islands populations are consistent with the origins of the initial migrations, and they display ancestral relationships (Mellers *et al*, 2013; Ranaweera *et al*, 2013; Ratnayake, 2016). Previous studies on mitochondrial DNA and Y-chromosome data have shown that both India and Sri Lanka were colonized by modern humans from Africa (Lahr and Forey, 1994; Mellers 2006; Mellers *et al*, 2013). This ancestral relationship is supported by the results from the principal component analysis plots in Figure 4 and admixture plots in Figure 6. However, not much can be said on the ancestral relationship because an African genome was not included in these analyses. These results suggest that there was continuous gene flow between Sri Lanka and mainland Indian populations after the initial colonization by showing less difference between Sri Lanka and other South Asians. The clustering of Sri Lanka with other South Asian populations in both the phylogenetic trees and the genetic networks also supports this observed relationship because, none of the results had shown a differentiation of Sri Lanka from other South Asians. The geographical distance between Sri Lanka and India might have influenced this structure by allowing easy migrations between these regions. A study based on mitochondrial DNA, Y-chromosome, and X-chromosome data observed high levels of reciprocal gene flow between Sri Lanka and the Indian populations (Garrigan *et al*, 2007).

The separation of Japan in Figure 4 at PC1 and PC2 from the rest of the East Asians confirms the hypothesis which expects the island population to be different from mainlanders. This difference is also supported by the formation of a distinct subclade in Figure 9 and admixture plots in Figure 6 and Figure 7 where Japan has distinct ancestral composition from the rest of East Asians. Although geography is not considered a barrier in modern human populations, the genetic diversity results showed low levels of diversity in Japan. These low levels of genetic diversity support the idea that, unlike Sri Lanka which has shorter distance from the mainland, migrations into Japan required watercraft (Kaifu *et al*, 2015). The results also suggest that Japan might have suffered a genetic bottleneck at initial colonization period which might have resulted in reduced levels of genetic diversity observed in Figure 2. The initial colonization of Japan might have involved longer term isolation as compared to Sri Lanka.

4.2.4. The structure differences from each genetic marker

The results show different patterns of structure for all four different datasets. X-chromosome results portrayed lower structure than autosomal datasets results. Based on the effective size X-chromosome is three quarters of the autosomal size, which results in low-rate polymorphism in X-chromosome than autosomes. The reduced effective population's size of X-chromosome data suggest that the effect of drift is higher in X-chromosome than autosomes which suggests there should be more structure in X-chromosome than in autosomes (Schaffner, 2004). This contradicts the observed structure in all results because the X-chromosome data was only able to differentiate South from East Asia, and no further. This might be caused by the differences in the number of SNPs that were used for each dataset. The number of SNPs used for the X-chromosome data were smaller compared to the autosomes. Meaning that only fewer characters were compared from the X-chromosome.

Both mitochondrial DNA and Y-chromosome data are uniparental (maternal and paternal lineages) with no recombination. However, unlike Y-chromosome, the mitochondrial DNA is found in both males and females but transferred through maternal lineage. The results from genetic network suggest that there is a sex-biased gene flow between the regions. This sex-biased geneflow is reflected by the more within variance on the Y-chromosome network than in mitochondrial DNA. From the results, this difference in variance or lack of clear structuring might

have resulted from over filtering of the mitochondrial dataset, only 119 SNPs were used for the phylogenetic reconstructions. However, the results in Figure 11 further suggest that the rate of migrating males in South Asia is relatively low than in East Asia. Only one paternal genome from Telangana was found in the East Asian clade which support the idea of endogamy in this region. This suggest that there was gene flow between South Asian ancestral paternal lineage from Telangana with ancestral East Asian populations. This is also supported by the phylogenetic structure in Figure 9. Some East Asians (Southern Han, Han Chinese, and Japan) have paternal ancestry from South Asia in Figure 11. However, the formation of a subclade within the South Asian clade suggests the possibility of an ancient polymorphism that is still maintained within the East Asians by incomplete lineage sorting.

4.3 Introgression

4.3.1. Introgression within South Asia

The phylogenetic networks and trees suggest a high level of intraregional gene flow. Although the caste system was developed recently, it is safe to say that the observed structure in South Asia resulted from gene flow and not incomplete lineage sorting. This is because if the lack of monophyly observed was due to incomplete lineage sorting there was supposed to be more structure among South Asian genomes. The geographical distribution of populations in Figure 4 at PC1 and PC5 also rules out the concept of incomplete lineage sorting South Asia. The results in Figure 4 at PC5 separate southern Indian population (Telangana) from northern Indian population (Punjabi). The gene flow results in Table 3 predict a significant level of gene flow between Sri Lanka and Gujarati other than between Telangana and Sri Lanka which rules out the hypothesis which suggest that there should be more gene flow between islands and closer mainland populations.

4.3.2. Introgression within East Asian populations

The dispersals of modern humans across the globe were shaped by the physical geographical aspect of the environment and this is clearly outlined by the genetic relationships between

populations. Unlike in South Asia where lack of monophyly is due to geneflow, the lack of monophyly in Figure 9 within East Asian clade suggest incomplete lineage sorting between the genomes. The admixture results in Figure 6 at K=2 suggest similar ancestral composition between Han and Southern Han Chinese as expected since Southern Han resulted from Han expansion (Wen *et al*, 2004). The closer relationship between Han and Southern Han is also supported by the PCA plots in Figure4 at PC1 and PC5).

However, the lack of monophyletic clades that include both populations could also suggest that genetic drift did not have enough time to make the populations reciprocally monophyletic after population expansion. This is because according to Hudson and Coyne, (2002), genetic drift requires 9-12 generations to make populations monophyletic. The clustering of some Southern Han within the subclade which is dominated by Vietnam and Dai Chinese suggest possible gene flow events with the two populations. These gene flow events between Southern Han and Vietnam or Dai Chinese were supported by significant gene flow results in Table 5. The test between Han Chinese, Japan and Vietnam or Dai Chinese showed significant gene flow between Han and Vietnam or Dai Chinese. The gene flow between these populations might have brought similarity between them resulting in the observed southern Chinese incline. Where the southern Chinese populations are being separated from the northern Chinese populations in Figure 4 at PC1 and PC2.

4.3.3. Gene flow between South and East Asia

It is clear from the results that the effect of geography on the genetic diversity and admixture in modern humans cannot be ruled out completely. Admixture results strongly portray a relationship between populations which are closely located geographically. The results in Figure 6 at K=5, and Figure 7 depicts an admixture event from East Asian populations into Bengali which also suggests that it was influenced by the distance between Indian populations and China since this introgression was only significantly observed in the Bengali population. This gene flow event between Bengali and East Asian is also depicted by the clustering of some Bengali genomes outside the South Asian clade in Figure 9. The fact that Bengalis have East Asian ancestors could also explain their clustering towards East Asia (Khan, 2018). The gene flow may be from Dai Chinese into Bengali because the estimated distance, which is 1,085 KM, this distance is smallest compared to the distance between Bengali and the rest of East Asian populations. However, we

cannot conclude that the gene flow is only from Dai Chinese because the tests between Bengali and East Asian populations from gene flow results in Table 4 were all significant. It is also clear that the observed gene flow in Figure 7 was only from one direction (East Asian populations to South Asian populations), because all the gene flow tests between Dai Chinese and South Asian populations were non-significant.

However, the separation between South and East Asia observed from the principal component analysis in Figure 3 and Figure 4 and admixture results in Figure 6 at K= 2 suggests a reduced level of admixture between the two regions. The reduced level of gene flow between South Asian populations and East Asian populations was also supported by the formation of distinct clades on a phylogenic tree in Figure 9and the phylogenetic networks in Figure 10and Figure 11. If the gene flow between the two regions (South and East) was high, we would expect to see less structure because gene flow destroys the structure.

5. Conclusion

The results outline a structure within Asian populations. Despite the diverse cultural practices and South Asia being colonized early the populations are not more internally structured nor more genetically diverse compared to East Asia. Most East Asian populations had higher heterozygosity than south Asian populations. However, nucleotide diversity demonstrated opposition, making it impossible to determine which region is more diverse than the other. The gene flow results suggested that effect of geography in human populations cannot be ruled out because, the significant levels of gene flow were observed between populations that are at close geographical distance compared to populations in distant proximity. Low levels of genetic diversity observed in Japan supports the idea that suggest that the initial colonizers of islands might have experienced a genetic bottleneck.

6. Acknowledgements

I would like to acknowledge National Research foundation for funding this project. I am really grateful to my supervisor professor Yoshan Moodley and co-supervisor Dr Andrinajoro

Rokotoarivelo for guiding and helping me throughout this project. I would also like to thank Thabelo Rambuda and Mannda Ndou for assisting me and providing clarity on some concept of Bioinformatics. I also thank my family for supporting me in pursuing this project.

7.References

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.

Ali, M., Liu, X., Pillai, E.N., Chen, P., Khor, C.C., Ong, R.T.H. and Teo, Y.Y., 2014. Characterizing the genetic differences between two distinct migrant groups from Indo-European and Dravidian speaking populations in India. BMC genetics, 15(1), pp.1-11.

Almal, S., Jeon, S., Agarwal, M., Patel, S., Patel, S., Bhak, Y., Jun, J., Bhak, J. and Padh, H., 2019. Sequencing and analysis of the whole genome of Indian Gujarati male. Genomics, 111(2), pp.196-204.

Bae, C.J., Douka, K. and Petraglia, M.D., 2017. On the origin of modern humans: Asian perspectives. Science, 358(6368), pp.9067.

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.

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.

Basu, A., Mukherjee, N., Roy, S., Sengupta, S., Banerjee, S., Chakraborty, M., Dey, B., Roy, M., Roy, B., Bhattacharyya, N.P. and Roychoudhury, S., 2003. Ethnic India: a genomic view, with special reference to peopling and structure. Genome research, 13(10), pp.2277-2290.

Basu, A., Sarkar-Roy, N. and Majumder, P.P., 2016. Genomic reconstruction of the history of extant populations of India reveals five distinct ancestral components and a complex structure. Proceedings of the National Academy of Sciences, 113(6), pp.1594-1599.

Boivin, N., Fuller, D.Q., Dennell, R., Allaby, R. and Petraglia, M.D., 2013. Human dispersal across diverse environments of Asia during the Upper Pleistocene. Quaternary International, 300, pp.32-47.

Bush, W.S. and Moore, J.H., 2012. Chapter 11: Genome-wide association studies. PLoS computational biology, 8(12), p.e1002822.

Brahmachari, S. and Choudhury, K., 2008. 6th March, 1971. Indian Literature, 52(2 (244), pp.79-79.

Cai, X., Qin, Z., Wen, B., Xu, S., Wang, Y., Lu, Y., Wei, L., Wang, C., Li, S., Huang, X. and Jin, L., 2011. Human migration through bottlenecks from Southeast Asia into East Asia during Last Glacial Maximum revealed by Y chromosomes. PLoS One, 6(8), p.e24282.

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

Cassens, I., Mardulyn, P. and Milinkovitch, M.C., 2005. Evaluating intraspecific "network" construction methods using simulated sequence data: do existing algorithms outperform the global maximum parsimony approach. Systematic biology, 54(3), pp.363-372.

Cavalli-Sforza, L.L. and Feldman, M.W., 2003. The application of molecular genetic approaches to the study of human evolution. Nature genetics, 33(3), pp.266-275.

Cavalli-Sforza, L.L., Cavalli-Sforza, L., Menozzi, P. and Piazza, A., 1994. The history and geography of human genes. Princeton university press.

Chaubey, G., Karmin, M., Metspalu, E., Metspalu, M., Selvi-Rani, D., Singh, V.K., Parik, J., Solnik, A., Naidu, B.P., Kumar, A. and Adarsh, N., 2008. Phylogeography of mtDNA haplogroup R7 in the Indian peninsula. BMC Evolutionary Biology, 8(1), pp.1-12.

Chaubey, G., Metspalu, M., Choi, Y., Mägi, R., Romero, I.G., Soares, P., Van Oven, M., Behar, D.M., Rootsi, S., Hudjashov, G. and Mallick, C.B., 2011. Population genetic structure in Indian Austroasiatic speakers: the role of landscape barriers and sex-specific admixture. Molecular biology and evolution, 28(2), pp.1013-1024.

Chen, J., Zheng, H., Bei, J.X., Sun, L., Jia, W.H., Li, T., Zhang, F., Seielstad, M., Zeng, Y.X., Zhang, X. and Liu, J., 2009. Genetic structure of the Han Chinese population revealed by genome-wide SNP variation. The American Journal of Human Genetics, 85(6), pp.775-785.

Chu, J.Y., Huang, W., Kuang, S.Q., Wang, J.M., Xu, J.J., Chu, Z.T., Yang, Z.Q., Lin, K.Q., Li, P., Wu, M. and Geng, Z.C., 1998. Genetic relationship of populations in China. Proceedings of the National Academy of Sciences, 95(20), pp.11763-11768

Cordaux, R., Aunger, R., Bentley, G., Nasidze, I., Sirajuddin, S.M. and Stoneking, M., 2004. Independent origins of Indian caste and tribal paternal lineages. Current Biology, 14(3), pp.231-235.

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.

Danecek, P., McCarthy, S. and Li, H., 2015. bcftools—utilities for variant calling and manipulating vcfs and bcfs.

Debnath, M., Palanichamy, M.G., Mitra, B., Jin, J.Q., Chaudhuri, T.K. and Zhang, Y.P., 2011. Y-chromosome haplogroup diversity in the sub-Himalayan Terai and Duars populations of East India. Journal of human genetics, 56(11), pp.765-771.

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.

Duong, N.T., Macholdt, E., Ton, N.D., Arias, L., Schröder, R., Van Phong, N., Thi Bich Thuy, V., Ha, N.H., Thi Thu Hue, H., Thi Xuan, N. and Thi Phuong Oanh, K., 2018. Complete human mtDNA genome sequences from Vietnam and the phylogeography of Mainland Southeast Asia. Scientific reports, 8(1), pp.1-13.

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.

Felsenstein, J., 1985. Phylogenies and the comparative method. The American Naturalist, 125(1), pp.1-15.

Gamble, C., 1993. Timewalkers: the prehistory of global colonization. Penguin Books.

Garrigan, D., Kingan, S.B., Pilkington, M.M., Wilder, J.A., Cox, M.P., Soodyall, H., Strassmann, B., Destro-Bisol, G., De Knijff, P., Novelletto, A. and Friedlaender, J., 2007. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data. Genetics, 177(4), pp.2195-2207.

Hague, M.T. and Routman, E.J., 2016. Does population size affect genetic diversity? A test with sympatric lizard species. Heredity, 116(1), pp.92-98.

Harry, B., 1992. Cultural diversity. Families, and the Special Education System.

HUGO Pan-Asian SNP Consortium, Abdulla, M.A., Ahmed, I., Assawamakin, A., Bhak, J., Brahmachari, S.K., Calacal, G.C., Chaurasia, A., Chen, C.H., Chen, J. and Chen, Y.T., 2009. Mapping human genetic diversity in Asia. Science, 326(5959), pp.1541-1545.

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

Jeong, C., Ozga, A.T., Witonsky, D.B., Malmström, H., Edlund, H., Hofman, C.A., Hagan, R.W., Jakobsson, M., Lewis, C.M., Aldenderfer, M.S. and Di Rienzo, A., 2016. Long-term genetic stability and a high-altitude East Asian origin for the peoples of the high valleys of the Himalayan arc. Proceedings of the National Academy of Sciences, 113(27), pp.7485-7490.

Jin, L. and Su, B., 2000. Natives or immigrants: modern human origin in East Asia. Nature Reviews Genetics, 1(2), pp.126-133.

Jinam, T.A., Hong, L.C., Phipps, M.E., Stoneking, M., Ameen, M., Edo, J., HUGO Pan-Asian SNP Consortium and Saitou, N., 2012. Evolutionary history of continental Southeast Asians:"Early train" hypothesis based on genetic analysis of mitochondrial and autosomal DNA data. Molecular biology and evolution, 29(11), pp.3513-3527.

Jinam, T.A., Kawai, Y. and Saitou, N., 2021. Modern human DNA analyses with special reference to the inner dual-structure model of Yaponesian. Anthropological Science, 129(1), pp.3-11.

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.

Kaifu, Y., Izuho, M. and Goebel, T., 2015. Modern human dispersal and behavior in Paleolithic Asia. Emergence and diversity of modern human behavior in Paleolithic Asia, pp.535-566.

Kayser, M., Brauer, S., Cordaux, R., Casto, A., Lao, O., Zhivotovsky, L.A., Moyse-Faurie, C., Rutledge, R.B., Schiefenhoevel, W., Gil, D. and Lin, A.A., 2006. Melanesian and Asian origins of Polynesians: mtDNA and Y chromosome gradients across the Pacific. Molecular biology and evolution, 23(11), pp.2234-2244.

Keegan, W.F. and Diamond, J.M., 1987. Colonization of islands by humans: A biogeographical perspective. In Advances in archaeological method and theory (pp. 49-92). Academic Press.

Klein, R.G., 2009. The human career: Human biological and cultural origins. University of Chicago Press.

Lacy, R.C., 1987. Loss of genetic diversity from managed populations: interacting effects of drift, mutation, immigration, selection, and population subdivision. Conservation biology, 1(2), pp.143-158.

Lahr, M.M. and Foley, R., 1994. Multiple dispersals and modern human origins. Evolutionary Anthropology: Issues, News, and Reviews, 3(2), pp.48-60.

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, H., Ruan, J. and Durbin, R., 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research, 18(11), pp.1851-1858.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L. and Myers, R.M., 2008. Worldwide human relationships inferred from genome-wide patterns of variation. science, 319(5866), pp.1100-1104.

Li, Z.Y., Wu, X.J., Zhou, L.P., Liu, W., Gao, X., Nian, X.M. and Trinkaus, E., 2017. Late Pleistocene archaic human crania from Xuchang, China. Science, 355(6328), pp.969-972

Linck, E. and Battey, C.J., 2018. Minor allele frequency thresholds dramatically affect population structure inference with genomic datasets. Biorxiv, p.188623.

Liu, D., Duong, N.T., Ton, N.D., Van Phong, N., Pakendorf, B., Van Hai, N. and Stoneking, M., 2020. Extensive ethnolinguistic diversity in Vietnam reflects multiple sources of genetic diversity. Molecular biology and evolution, 37(9), pp.2503-2519.

Macholdt, E., Arias, L., Duong, N.T., Ton, N.D., Van Phong, N., Schröder, R., Pakendorf, B., Van Hai, N. and Stoneking, M., 2020. The paternal and maternal genetic history of Vietnamese populations. European Journal of Human Genetics, 28(5), pp.636-645.

Maji, S., Krithika, S. and Vasulu, T.S., 2009. Phylogeographic distribution of mitochondrial DNA macrohaplogroup M in India. Journal of genetics, 88(1), pp.127-139.

Majumder, P.P. and Basu, A., 2015. A genomic view of the peopling and population structure of India. Cold Spring Harbor perspectives in biology, 7(4), p.a008540.

Majumder, P.P., 2010. The human genetic history of South Asia. Current Biology, 20(4), pp.R184-R187.

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), p.201

Marth, G.T., Czabarka, E., Murvai, J. and Sherry, S.T., 2004. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. Genetics, 166(1), pp.351-372.

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

Mellars, P., Gori, K.C., Carr, M., Soares, P.A. and Richards, M.B., 2013. Genetic and archaeological perspectives on the initial modern human colonization of southern Asia. Proceedings of the National Academy of Sciences, 110(26), pp.10699-10704.

Metspalu, M., Romero, I.G., Yunusbayev, B., Chaubey, G., Mallick, C.B., Hudjashov, G., Nelis, M., Mägi, R., Metspalu, E., Remm, M. and Pitchappan, R., 2011. Shared and unique components of human population structure and genome-wide signals of positive selection in South Asia. The American Journal of Human Genetics, 89(6), pp.731-744.

Moorjani, P., Thangaraj, K., Patterson, N., Lipson, M., Loh, P.R., Govindaraj, P., Berger, B., Reich, D. and Singh, L., 2013. Genetic evidence for recent population mixture in India. The American Journal of Human Genetics, 93(3), pp.422-438.

Nakagome, S., Sato, T., Ishida, H., Hanihara, T., Yamaguchi, T., Kimura, R., Mano, S., Oota, H. and Asian DNA Repository Consortium, 2015. Model-based verification of hypotheses on the origin of modern Japanese revisited by Bayesian inference based on genome-wide SNP data. Molecular Biology and Evolution, 32(6), pp.1533-1543.

Nakazawa, Y., 2017. On the Pleistocene population history in the Japanese Archipelago. Current anthropology, 58(S17), pp.S539-S552.

Nei, M. and Takahata, N., 1993. Effective population size, genetic diversity, and coalescence time in subdivided populations. Journal of Molecular Evolution, 37(3), pp.240-244.

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.

Nielsen, R., Korneliussen, T., Albrechtsen, A., Li, Y. and Wang, J., 2012. SNP calling, genotype calling, and sample allele frequency estimation from new-generation sequencing data. PloS one, 7(7), p.e37558.

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.

Pan, Z. and Xu, S., 2020. Population genomics of East Asian ethnic groups. Hereditas, 157(1), pp.1-10.

Patton, M., 2013. Islands in time: island sociogeography and Mediterranean prehistory. Routledge.

Pierson, M.J., Martinez-Arias, R., Holland, B.R., Gemmell, N.J., Hurles, M.E. and Penny, D., 2006. Deciphering past human population movements in Oceania: provably optimal trees of 127 mtDNA genomes. Molecular biology and evolution, 23(10), pp.1966-1975.

Pope, K.O. and Terrell, J.E., 2008. Environmental setting of human migrations in the circum-Pacific region. Journal of Biogeography, 35(1), pp.1-21.

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.

Ramachandran, S., Deshpande, O., Roseman, C.C., Rosenberg, N.A., Feldman, M.W. and Cavalli-Sforza, L.L., 2005. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences, 102(44), pp.15942-15

Ranaweera, L., Kaewsutthi, S., Win Tun, A., Boonyarit, H., Poolsuwan, S. and Lertrit, P., 2014. Mitochondrial DNA history of Sri Lankan ethnic people: their relations within the island and with the Indian subcontinental populations. Journal of Human Genetics, 59(1), pp.28-36.

Reich, D., Patterson, N., Kircher, M., Delfin, F., Nandineni, M.R., Pugach, I., Ko, A.M.S., Ko, Y.C., Jinam, T.A., Phipps, M.E. and Saitou, N., 2011. Denisova admixture and the first modern human

dispersals into Southeast Asia and Oceania. The American Journal of Human Genetics, 89(4), pp.516-528.

Saeb, A. and Al-Naqeb, D., 2016. The impact of evolutionary driving forces on human complex diseases: a population genetics approach. Scientifica, 2016.

Sahoo, S. and Kashyap, V.K., 2006. Phylogeography of mitochondrial DNA and Y-Chromosome haplogroups reveal asymmetric gene flow in populations of Eastern India. American Journal of Physical Anthropology: The Official Publication of the American Association of Physical Anthropologists, 131(1), pp.84-97.

Sankararaman, S., Patterson, N., Li, H., Pääbo, S. and Reich, D., 2012. The date of interbreeding between Neandertals and modern humans.

Sengupta, S., Zhivotovsky, L.A., King, R., Mehdi, S.Q., Edmonds, C.A., Chow, C.E.T., Lin, A.A., Mitra, M., Sil, S.K., Ramesh, A. and Rani, M.U., 2006. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of Central Asian pastoralists. The American Journal of Human Genetics, 78(2), pp.202-221.

Shang, H., Tong, H., Zhang, S., Chen, F. and Trinkaus, E., 2007. An early modern human from tianyuan cave, zhoukoudian, china. Proceedings of the National Academy of Sciences, 104(16), pp.6573-6578.

Singh, K.S., 1992. People of India (Vol. 43). Anthropological Survey of India.

Stokowski, R.P., Pant, P.K., Dadd, T., Fereday, A., Hinds, D.A., Jarman, C., Filsell, W., Ginger, R.S., Green, M.R., van der Ouderaa, F.J. and Cox, D.R., 2007. A genomewide association study of skin pigmentation in a South Asian population. The American Journal of Human Genetics, 81(6), pp.1119-1132.

Stoneking M, Delfin F. The human genetic history of East Asia: weaving a complex tapestry. Curr Biol. 2010;20:R188–93.

Stringer, C., 2000. Coasting out of Africa. Nature, 405(6782), pp.25-27.

Stringer, C., 2012. What makes a modern human. Nature, 485(7396), pp.33-35.

Su B, Xiao J, Underhill PA, Deka R, Zhang W, Akey J, Huang W, Shen D, Lu D, Luo J, Chu J, Tan J, Shen P, Davis R, Cavalli-Sforza LL, Chakraborty R, Xiong M, Du R, Oefner P, Chen Z, Jin

L .,1999. Y-Chromosome evidence for a northward migration of modern humans into Eastern Asia during the last ice age. Am J Hum Genet 65:1718–1724

Vernot, B. and Akey, J.M., 2014. Resurrecting surviving Neandertal lineages from modern human genomes. Science, 343(6174), pp.1017-1021.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. and Wilson, A.C., 1991. African populations and the evolution of human mitochondrial DNA. Science, 253(5027), pp.1503-1507.

Wang, T., Wang, W., Xie, G., Li, Z., Fan, X., Yang, Q., Wu, X., Cao, P., Liu, Y., Yang, R. and Liu, F., 2021. Human population history at the crossroads of East and Southeast Asia since 11,000 years ago. Cell, 184(14), pp.3829-3841

Wang, Y., Lu, D., Chung, Y.J. and Xu, S., 2018. Genetic structure, divergence and admixture of Han Chinese, Japanese and Korean populations. Hereditas, 155(1), pp.1-12.

Watkins, W.S., Thara, R., Mowry, B.J., Zhang, Y., Witherspoon, D.J., Tolpinrud, W., Bamshad, M.J., Tirupati, S., Padmavati, R., Smith, H. and Nancarrow, D., 2008. Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms. BMC genetics, 9(1), pp.1-17.

Wen, B., Xie, X., Gao, S., Li, H., Shi, H., Song, X., Qian, T., Xiao, C., Jin, J., Su, B. and Lu, D., 2004. Analyses of genetic structure of Tibeto-Burman populations reveals sex-biased admixture in southern Tibeto-Burmans. The American Journal of Human Genetics, 74(5), pp.856-865.

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

Wong, L.P., Lai, J.K.H., Saw, W.Y., Ong, R.T.H., Cheng, A.Y., Pillai, N.E., Liu, X., Xu, W., Chen, P., Foo, J.N. and Tan, L.W.L., 2014. Insights into the genetic structure and diversity of 38 South Asian Indians from deep whole-genome sequencing. PLoS genetics, 10(5), p.e1004377.

Xing, J., Watkins, W.S., Hu, Y., Huff, C.D., Sabo, A., Muzny, D.M., Bamshad, M.J., Gibbs, R.A., Jorde, L.B. and Yu, F., 2010. Genetic diversity in India and the inference of Eurasian population expansion. Genome biology, 11(11), pp.1-13.

Xing, J., Watkins, W.S., Witherspoon, D.J., Zhang, Y., Guthery, S.L., Thara, R., Mowry, B.J., Bulayeva, K., Weiss, R.B. and Jorde, L.B., 2009. Fine-scaled human genetic structure revealed by SNP microarrays. Genome research, 19(5), pp.815-825.

Yao YG, Kong QP, Bandelt HJ, Kivisild T, Zhang YP., 2002. Phylogeographic differentiation of mitochondrial DNA in Han Chinese. Am J Hum Genet 70:635–651

Zhang, F., Su, B., Zhang, Y.P. and Jin, L., 2007. Genetic studies of human diversity in East Asia. Philosophical Transactions of the Royal Society B: Biological Sciences, 362(1482), pp.987-996.

Appendix A: Command lines

1. bcftools view -S –sample-file file.vcf -o v > newfile.vcf2.bwa mem -t 6 -B 4 -O 6 -E 1 -M GRCh38_full_analysis_set_plus_decoy_hla.fa 1.fastq.gz 2.fastq.gz | k8 bwa-postalt.js GRCh38_full_analysis_set_plus_decoy_hla.fa.alt | samtools view -1 - > denisova_y.bam

3. samtools sort -n -@ 15 -O BAM --reference GRCh38_full_analysis_set_plus_decoy_hla.fa file.bam -o sorted.bam.

4.samtools fixmate -m -@ 15 --reference GRCh38_full_analysis_set_plus_decoy_hla.fa -O BAM sorted_by_name.bam

5.samtools sort -@ 15 -O BAM --reference GRCh38_full_analysis_set_plus_decoy_hla.fa file.bam -o sorted.bam

6.samtools markdup -@ 15 -O BAM --reference GRCh38_full_analysis_set_plus_decoy_hla.fa sorted_HGDP00473.bam markdup_HGDP00473.bam

7.samtools view -@ 15 -b -T GRCh38_full_analysis_set_plus_decoy_hla.fa markdup_HGDP00473.bam -o final_HGDP00473.bam.

8.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 <sample1.bam> <sample2.bam> <sample3.bam> | bcftools call --threads 30 -mv -O z --ploidy GRCh38 -o file.vcf.gz.

9.bcftools filter --threads 15 -sQUALFILTER -e'QUAL<1' file.vcf.gz -o filtered.vcf.gz -O z.

10.bcftools norm --threads 30 -f GRCh38_full_analysis_set_plus_decoy_hla.fa -o 2_samples_norm.vcf.gz -m -both filtered_samples.vcf.gz -O z.

11./usr/lib/vcflib/bin/vcfallelicprimitives norm.vcf.gz --keep-info --keep-geno | vt sort - | vt uniq - | bgzip -c > HGDP01201.aprimitives.vcf.gz.

12.bcftools norm --threads 10 -f GRCh38_full_analysis_set_plus_decoy_hla.fa -o HGDP01201.aprimitives.merged.vcf.gz -m +both HGDP01201.aprimitives.vcf.gz -O z.

13.bcftools view --threads 10 -o HGDP01201.aprimitives.merged.biallelic.vcf.gz -O z -m2 -M2 .aprimitives.merged.vcf.g

14.bcftools view --threads 2 -O z -i 'MAF>0.02' Unrelated_Y_chr_subsampled_males.vcf.gz -o Unrelated_Y_chr_subsampled_males1.vcf.gz

15.bcftools view --types snps Unrelated_Y_chr_subsampled_males1.vcf.gz -O z -o Unrelated_Y_chr_subsampled_males2.vcf.gz --threads 2

16.bcftools view --threads 2 -e 'AC==0 || F_MISSING > 0.2' -m2 -M2 -O z -o Unrelated_Y_chr_subsampled_males3.vcf.gz Unrelated_Y_chr_subsampled_males2.vcf.gz

17. for i in {2..11} do admixture --cv $FILE.bed $i > log${i}.out done

18. Rscript plotADMIXTURE.r -p $FILE -i $FILE.list -k 5 -l Bengali,Japan,Punjabi,Sri Lanka, Southern Han,Han Chinese,Telangana,Vietnam

19. bcftools view file.vcf.gz | vcfrandomsample -r 0.012 > new.vcf.gz

20. bcftools view -Oz -S samples.txt file.vcf.gz > new.vcf.gz

21. scp filetocopy.tar.gz yourusername@scp.chpc.ac.za:/mnt/lustre/users/mngoveni/run1/

22. plink –vcf file.vcf --allow-extra-chr –indep-pairwise 50 10 0.5m –out newfile.vcf

23.plink –allow-extra-chr –extract newfile.prune.in –make-bed –out final_file –recode vcf-lid –vcf file.vcf

Appendix B: Job submission script

```
#!/bin/bash

### Select 10 nodes with 24 CPUs

#PBS -l select=1:ncpus=56:mpiprocs=56:nodetype=haswell_fat

### Job Name

#PBS -N iqtree

### Project code

#PBS -P CBBI0911

#PBS -l walltime=48:00:00

#PBS -q bigmem

#PBS -W group_list=bigmemq

#PBS -o /mnt/lustre3p/users/mngoveni/marcia/iqtree_log.out

#PBS -e /mnt/lustre3p/users/mngoveni/marcia/iqtree_log.err

### Send email on abort, begin and end

#PBS -m abe

### Specify mail recipient

#PBS -M duni1673@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 topo01.phy -alrt 1000 -bb 1000 -nt AUTO"

cd /mnt/lustre/users/mngoveni/fasta/run3_iqtree

${EXE} ${ARGS}
```

Appendix C: Admixture

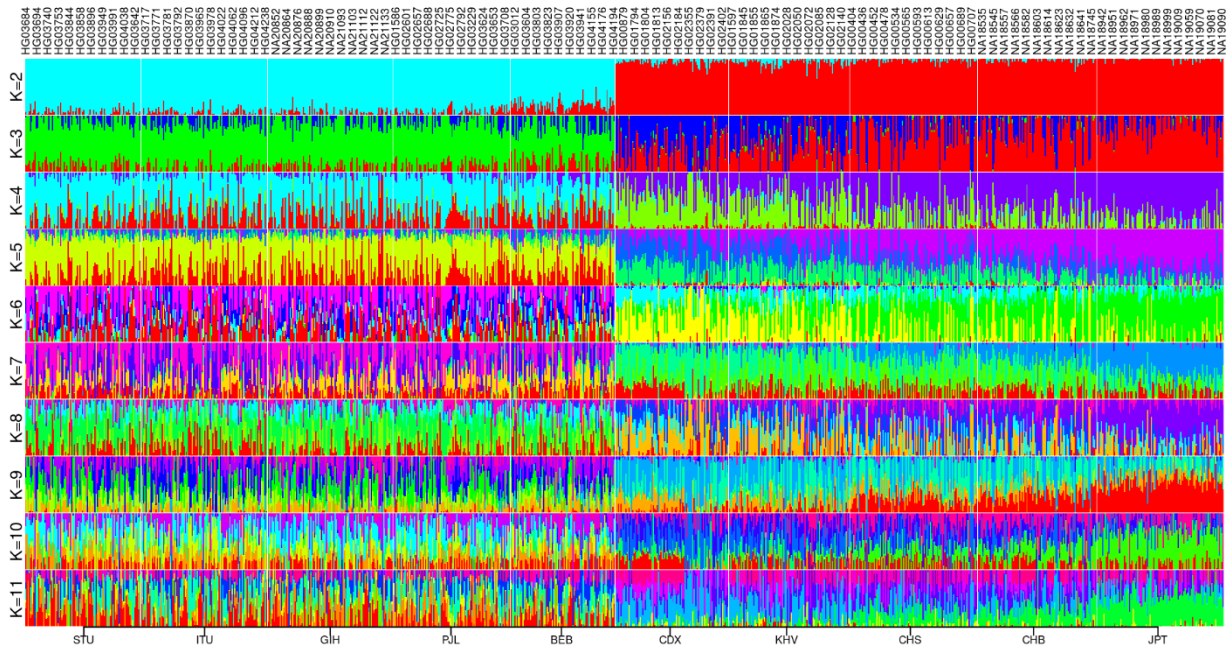


Figure12: admixture plot showing admixture proportion of X-chromosome data from K=2 to K=11

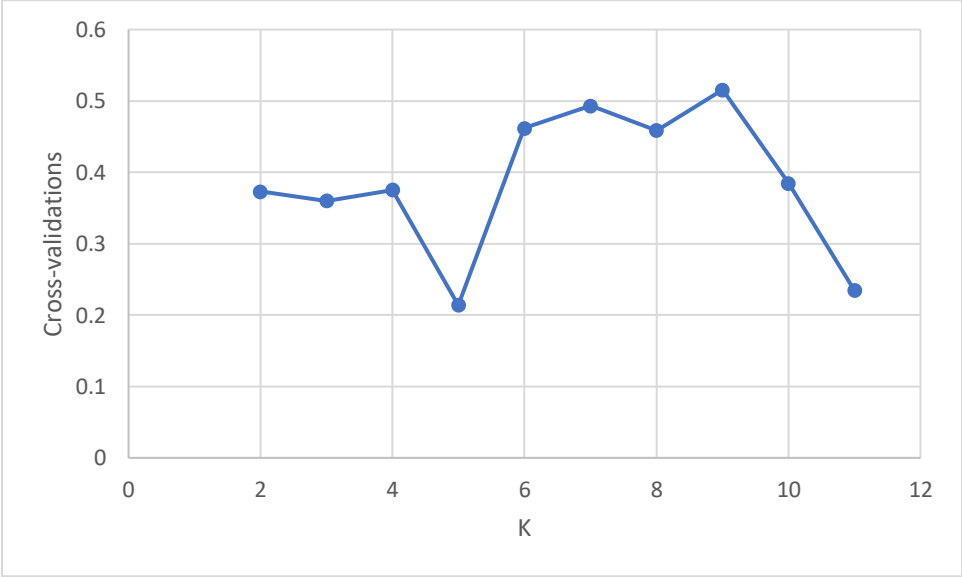


Figure 13: Cross-validation plot for X-chromosome dataset

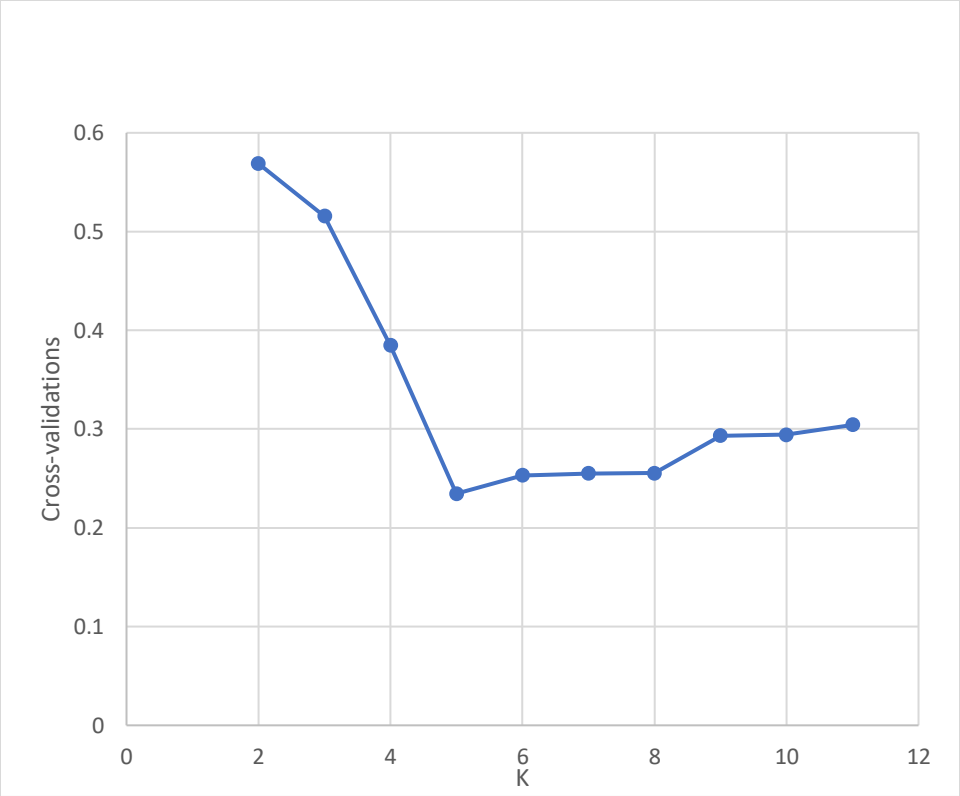


Figure 14: cross-validation plot for the autosomal dataset

Appendix D: introgression between South and East Asian populations.

P1	P2	P3	Dstatistic	Z-score	p-value	f4-ratio	BBAA	ABBA	BABA
CDX	CHB	BEB	0.00055746	0.558228	0.288344	0.0252462	190715	161855	161674
CDX	CHS	BEB	0.00039542	0.376578	0.353244	0.0178217	191382	161053	160926
GIH	BEB	CDX	0.00972165	9.35388	0	0.0963346	185172	176453	173056
ITU	BEB	CDX	0.0113677	11.1454	0	0.110821	183796	176707	172735
JPT	CDX	BEB	0.00062171	0.557158	0.28871	0.0275665	189674	162946	162744
CDX	KHV	BEB	0.00096273	1.04705	0.147537	0.0434185	191262	161247	160937
PJL	BEB	CDX	0.0103144	10.0769	0	0.101859	184939	177021	173406
STU	BEB	CDX	0.0109152	12.0965	0	0.106921	183502	176703	172887
CHS	CHB	BEB	0.00016518	0.206732	0.41811	0.0075606	191805	160579	160526
GIH	BEB	CHB	0.00929505	9.10188	5.42E-20	0.0941522	184950	176412	173162
ITU	BEB	CHB	0.0111801	10.7089	0	0.111115	183620	176711	172804
JPT	CHB	BEB	0.0011862	1.66102	0.0483543	0.0521202	191017	161565	161182
CHB	KHV	BEB	0.00040073	0.388069	0.348983	0.018641	190484	162046	161916
PJL	BEB	CHB	0.00970047	9.47845	0	0.0980852	184679	176941	173541
STU	BEB	CHB	0.010865	11.2596	0	0.108364	183357	176739	172939
GIH	BEB	CHS	0.0095246	9.4331	0	0.0944357	185028	176436	173107
ITU	BEB	CHS	0.0112917	10.9968	0	0.110017	183676	176715	172769
JPT	CHS	BEB	0.00102102	0.996938	0.159397	0.0448993	190883	161671	161341
CHS	KHV	BEB	0.00056674	0.496977	0.309603	0.0260659	191050	161418	161236
PJL	BEB	CHS	0.0100612	9.80409	0	0.0994627	184782	176991	173465
STU	BEB	CHS	0.0109458	11.523	0	0.107054	183409	176737	172910
GIH	BEB	JPT	0.00889432	8.87775	3.25E-19	0.0880258	185119	176198	173092
GIH	BEB	KHV	0.0092798	8.96816	1.63E-19	0.0987571	184840	176431	173187
PJL	GIH	BEB	0.0022788	2.95596	0.0015585	0	178384	175941	175141
ITU	BEB	JPT	0.0108059	10.6035	0	0.104944	183795	176504	172730
ITU	BEB	KHV	0.0109919	10.737	0	0.114863	183483	176705	172863
STU	ITU	BEB	0.00082575	1.14473	0.126161	0.830279	176382	175490	175201
JPT	KHV	BEB	0.00157281	1.42139	0.0776011	0.069794	189527	163233	162720
PJL	BEB	JPT	0.00936988	9.10579	5.42E-20	0.0925141	184862	176741	173459
STU	BEB	JPT	0.0103328	10.7303	0	0.100862	183507	176505	172895
PJL	BEB	KHV	0.0100101	9.67993	0	0.105956	184633	177025	173516

STU	BEB	KHV	0.0105638	11.0678	0	0.110945	183199	176711	173016
CDX	CHB	GIH	0.0010111	1.04131	0.148865	0.0196359	194895	162785	162456
CDX	CHB	ITU	0.00075728	0.789052	0.215041	0.0193974	194727	161959	161714
KHV	CDX	CHB	0.00092047	0.967871	0.166554	0.109158	160545	159261	158968
CDX	CHB	PJL	0.00121408	1.23823	0.107815	0.026494	195125	162864	162469
CDX	CHB	STU	0.00060858	0.633533	0.263193	0.0165905	194567	161907	161710
CDX	CHS	GIH	0.00060511	0.575493	0.282479	0.011695	195595	161937	161741
CDX	CHS	ITU	0.00047629	0.464663	0.321087	0.0121411	195408	161133	160980
CHS	CDX	KHV	0.00120558	1.11063	0.133363	0	159299	159168	158784
CDX	CHS	PJL	0.00066648	0.633482	0.263209	0.0144744	195849	161995	161779
CDX	CHS	STU	0.00036016	0.344942	0.365069	0.0097708	195240	161084	160968
ITU	GIH	CDX	0.00164808	1.62429	0.0521571	0.0160291	187747	174600	174025
CDX	JPT	GIH	0.00027069	0.249449	0.401507	0.0052918	193782	163745	163657
CDX	KHV	GIH	0.00143132	1.59594	0.0552516	0.0276814	195437	162178	161714
PJL	GIH	CDX	0.00062068	0.721952	0.235162	0.0061055	189197	174637	174420
STU	GIH	CDX	0.00119793	1.20609	0.113892	0.0117138	187345	174696	174278
JPT	CDX	ITU	1.19E-05	0.010929	0.49564	0.0003078	193612	162912	162908
CDX	KHV	ITU	0.0013654	1.57551	0.0575698	0.0348301	195247	161390	160950
ITU	PJL	CDX	0.00102369	1.31884	0.093611	0.0099849	187478	174992	174634
ITU	STU	CDX	0.00044907	0.596167	0.275532	0.0043665	186145	174349	174193
KHV	CDX	JPT	0.0003873	0.414434	0.339278	0.0229586	162725	159128	159005
CDX	JPT	PJL	0.00039891	0.343964	0.365437	0.0087627	194022	163810	163680
CDX	JPT	STU	9.45E-06	0.0087	0.496529	0.0002594	193426	162885	162882
CDX	KHV	PJL	0.00128299	1.38709	0.0827075	0.0278823	195727	162204	161788
CDX	KHV	STU	0.00133929	1.55745	0.0596814	0.036357	195055	161346	160914
STU	PJL	CDX	0.00057555	0.81355	0.207951	0.005643	187108	175054	174853
CHS	CHB	GIH	0.000412	0.520543	0.301342	0.0080351	195936	161460	161327
CHS	CHB	ITU	0.00028544	0.361562	0.35884	0.0073464	195800	160665	160574
CHS	CHB	JPT	0.00092668	1.13006	0.129225	0.111498	160348	158574	158280
CHS	CHB	PJL	0.00055498	0.715596	0.23712	0.012191	196156	161530	161351
CHS	CHB	STU	0.00025204	0.322802	0.373423	0.0068876	195669	160643	160562
ITU	GIH	CHB	0.00188798	1.89662	0.0289393	0.0187191	187487	174669	174011
JPT	CHB	GIH	0.00074047	1.05861	0.144888	0.0144184	195035	162333	162093

CHB	KHV	GIH	0.00041374	0.400876	0.344256	0.0082054	194585	162902	162768
PJL	GIH	CHB	0.00043119	0.512781	0.304052	0.0043426	188851	174620	174469
STU	GIH	CHB	0.00157559	1.59841	0.0549758	0.0156858	187117	174796	174246
JPT	CHB	ITU	0.00077121	1.05132	0.146556	0.0196993	194920	161560	161311
CHB	KHV	ITU	0.00060166	0.583965	0.279622	0.0157379	194439	162159	161964
ITU	PJL	CHB	0.00145219	1.87925	0.0301049	0.0144389	187176	175086	174578
ITU	STU	CHB	0.00031081	0.417542	0.338141	0.0030811	185963	174363	174255
JPT	CHB	KHV	0.0067223	9.68116	0	0.646304	161399	160514	158371
JPT	CHB	PJL	0.00081458	1.08221	0.139579	0.0178762	195242	162390	162125
JPT	CHB	STU	0.00060051	0.827766	0.203901	0.0163331	194768	161517	161323
CHB	KHV	PJL	6.35E-05	0.061202	0.475599	0.0014252	194841	162894	162873
CHB	KHV	STU	0.00072407	0.692306	0.244373	0.0200951	194284	162152	161917
STU	PJL	CHB	0.00114133	1.58699	0.0562574	0.0113925	186839	175180	174781
ITU	GIH	CHS	0.00176971	1.77018	0.0383486	0.0172014	187583	174632	174015
JPT	CHS	GIH	0.00033028	0.322502	0.373536	0.0064364	194940	162399	162292
CHS	KHV	GIH	0.00082543	0.722302	0.235054	0.0161761	195186	162310	162043
PJL	GIH	CHS	0.00056359	0.689067	0.245391	0.0055494	188994	174630	174434
STU	GIH	CHS	0.00142664	1.44844	0.0737476	0.0139277	187209	174756	174258
JPT	CHS	ITU	0.00048682	0.480072	0.315588	0.0124452	194804	161647	161489
CHS	KHV	ITU	0.00088817	0.787612	0.215462	0.0229705	195019	161545	161259
ITU	PJL	CHS	0.00120204	1.51122	0.0653656	0.0117169	187298	175029	174608
ITU	STU	CHS	0.00034164	0.460187	0.322691	0.0033201	186034	174354	174234
JPT	CHS	PJL	0.00026218	0.252588	0.400293	0.0057584	195172	162435	162349
JPT	CHS	STU	0.00034946	0.348261	0.363822	0.0095125	194646	161607	161494
CHS	KHV	PJL	0.00061608	0.537859	0.295337	0.0135992	195468	162328	162128
CHS	KHV	STU	0.00097788	0.863775	0.193856	0.0268506	194858	161532	161217
STU	PJL	CHS	0.0008607	1.17992	0.119016	0.0084248	186957	175119	174818
ITU	GIH	JPT	0.00191482	1.95823	0.0251015	0.0185507	187580	174522	173855
ITU	GIH	KHV	0.00171487	1.70168	0.0444078	0.0178772	187349	174666	174068
JPT	KHV	GIH	0.00114439	1.04368	0.148318	0.0225057	193556	164017	163642
PJL	GIH	JPT	0.00050071	0.640373	0.260965	0.0049238	188951	174479	174305
STU	GIH	JPT	0.00144401	1.46207	0.0718607	0.014069	187184	174623	174120
PJL	GIH	KHV	0.00075774	0.932963	0.175419	0.0079889	188806	174709	174445

STU	GIH	KHV	0.00128916	1.29173	0.0982259	0.0135148	186957	174772	174322
JPT	KHV	ITU	0.00136159	1.21821	0.111572	0.0351273	193408	163271	162827
ITU	PJL	JPT	0.00140952	1.88291	0.0298562	0.0136941	187283	174928	174436
ITU	STU	JPT	0.00046936	0.650121	0.257807	0.0045464	186037	174244	174080
ITU	PJL	KHV	0.00095346	1.21927	0.111371	0.0099684	187106	175036	174703
ITU	STU	KHV	0.00042444	0.566224	0.285621	0.004424	185778	174413	174265
ITU	PJL	STU	0.0013051	1.7844	0.0371793	6.02145	176438	176006	175547
JPT	KHV	PJL	0.00086969	0.757754	0.224299	0.019276	193821	164018	163733
JPT	KHV	STU	0.00131429	1.18431	0.118145	0.0361019	193226	163237	162808
STU	PJL	JPT	0.00094076	1.29935	0.0969123	0.00919	186920	174997	174668
CHB	CHS	KHV	0.00376314	4.35766	6.57E-06	1.01442	160625	159159	157966
CHS	CHB	PJL	0.00055498	0.715596	0.23712	0.012191	196156	161530	161351
CHS	CHB	STU	0.00025204	0.322802	0.373423	0.0068876	195669	160643	160562
ITU	GIH	CHB	0.00188798	1.89662	0.0289393	0.0187191	187487	174669	174011
JPT	CHB	GIH	0.00074047	1.05861	0.144888	0.0144184	195035	162333	162093
CHB	KHV	GIH	0.00041374	0.400876	0.344256	0.0082054	194585	162902	162768
PJL	GIH	CHB	0.00043119	0.512781	0.304052	0.0043426	188851	174620	174469
STU	GIH	CHB	0.00157559	1.59841	0.0549758	0.0156858	187117	174796	174246
JPT	CHB	ITU	0.00077121	1.05132	0.146556	0.0196993	194920	161560	161311
CHB	KHV	ITU	0.00060166	0.583965	0.279622	0.0157379	194439	162159	161964
ITU	PJL	CHB	0.00145219	1.87925	0.0301049	0.0144389	187176	175086	174578
ITU	STU	CHB	0.00031081	0.417542	0.338141	0.0030811	185963	174363	174255
JPT	CHB	KHV	0.0067223	9.68116	0	0.646304	161399	160514	158371
JPT	CHB	PJL	0.00081458	1.08221	0.139579	0.0178762	195242	162390	162125
JPT	CHB	STU	0.00060051	0.827766	0.203901	0.0163331	194768	161517	161323
CHB	KHV	PJL	6.35E-05	0.061202	0.475599	0.0014252	194841	162894	162873
CHB	KHV	STU	0.00072407	0.692306	0.244373	0.0200951	194284	162152	161917
STU	PJL	CHB	0.00114133	1.58699	0.0562574	0.0113925	186839	175180	174781
ITU	GIH	CHS	0.00176971	1.77018	0.0383486	0.0172014	187583	174632	174015
JPT	CHS	GIH	0.00033028	0.322502	0.373536	0.0064364	194940	162399	162292
CHS	KHV	GIH	0.00082543	0.722302	0.235054	0.0161761	195186	162310	162043
PJL	GIH	CHS	0.00056359	0.689067	0.245391	0.0055494	188994	174630	174434
STU	GIH	CHS	0.00142664	1.44844	0.0737476	0.0139277	187209	174756	174258

JPT	CHS	ITU	0.00048682	0.480072	0.315588	0.0124452	194804	161647	161489
CHS	KHV	ITU	0.00088817	0.787612	0.215462	0.0229705	195019	161545	161259
ITU	PJL	CHS	0.00120204	1.51122	0.0653656	0.0117169	187298	175029	174608
ITU	STU	CHS	0.00034164	0.460187	0.322691	0.0033201	186034	174354	174234
KHV	CHS	JPT	0.0085855	7.1267	5.14E-13	0.509517	161217	160614	157880
JPT	CHS	PJL	0.00026218	0.252588	0.400293	0.0057584	195172	162435	162349
JPT	CHS	STU	0.00034946	0.348261	0.363822	0.0095125	194646	161607	161494
CHS	KHV	PJL	0.00061608	0.537859	0.295337	0.0135992	195468	162328	162128
CHS	KHV	STU	0.00097788	0.863775	0.193856	0.0268506	194858	161532	161217
STU	PJL	CHS	0.0008607	1.17992	0.119016	0.0084248	186957	175119	174818
ITU	GIH	JPT	0.00191482	1.95823	0.0251015	0.0185507	187580	174522	173855
ITU	GIH	KHV	0.00171487	1.70168	0.0444078	0.0178772	187349	174666	174068
PJL	GIH	ITU	0.00250109	3.31165	0.0004637	0	177366	175953	175075
ITU	GIH	STU	0.00362601	3.60676	0.000155	16.6793	176381	175883	174612
JPT	KHV	GIH	0.00114439	1.04368	0.148318	0.0225057	193556	164017	163642
PJL	GIH	JPT	0.00050071	0.640373	0.260965	0.0049238	188951	174479	174305
STU	GIH	JPT	0.00144401	1.46207	0.0718607	0.014069	187184	174623	174120
PJL	GIH	KHV	0.00075774	0.932963	0.175419	0.0079889	188806	174709	174445
STU	GIH	KHV	0.00128916	1.29173	0.0982259	0.0135148	186957	174772	174322
PJL	GIH	STU	0.00231443	3.01979	0.0012648	0	177757	175846	175034
JPT	KHV	ITU	0.00136159	1.21821	0.111572	0.0351273	193408	163271	162827
ITU	PJL	JPT	0.00140952	1.88291	0.0298562	0.0136941	187283	174928	174436
ITU	STU	JPT	0.00046936	0.650121	0.257807	0.0045464	186037	174244	174080
ITU	PJL	KHV	0.00095346	1.21927	0.111371	0.0099684	187106	175036	174703
ITU	STU	KHV	0.00042444	0.566224	0.285621	0.004424	185778	174413	174265
ITU	PJL	STU	0.0013051	1.7844	0.0371793	6.02145	176438	176006	175547
JPT	KHV	PJL	0.00086969	0.757754	0.224299	0.019276	193821	164018	163733
JPT	KHV	STU	0.00131429	1.18431	0.118145	0.0361019	193226	163237	162808
STU	PJL	JPT	0.00094076	1.29935	0.0969123	0.00919	186920	174997	174668