UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di BIOLOGIA

Corso di laurea magistrale in molecular biology

Tesi di laurea

Collecting Denisova Ancestry Tracts From Human Genomes Across The Globe

Relatore: Prof. Luca Pagani
Dipartment of Biology

Correlatore: Leonardo Vallini
Dipartment of Biology

Laureanda:
Sara Hosseini

Anno Accademico 2023/2024

Collecting Denisova Ancestry Tracts From Human Genomes Across The Globe

Acknowledgements

Firstly, I would like to Acknowledge Archaic Humans who taught me the law of Survival which is one of our basic desires. In order to Survive, humans first need to learn to be adaptable. The next stage would be evolution in all aspects, both internally and physically. Secondly, I want to Acknowledge the first homo-chimp ancestor who left its community. Thanks to them, modern humans created useful majors and fields that start with the past but will end to us. So, at the end we should learn more about ourselves.

I would like to honor my supervisor, Luca Pagani, an amazing teacher who is able to make difficult things easy, and Leonardo Vallini for helping me with all the practical processes. Without his help, I wouldn't have been able to figure out the entire process on my own.

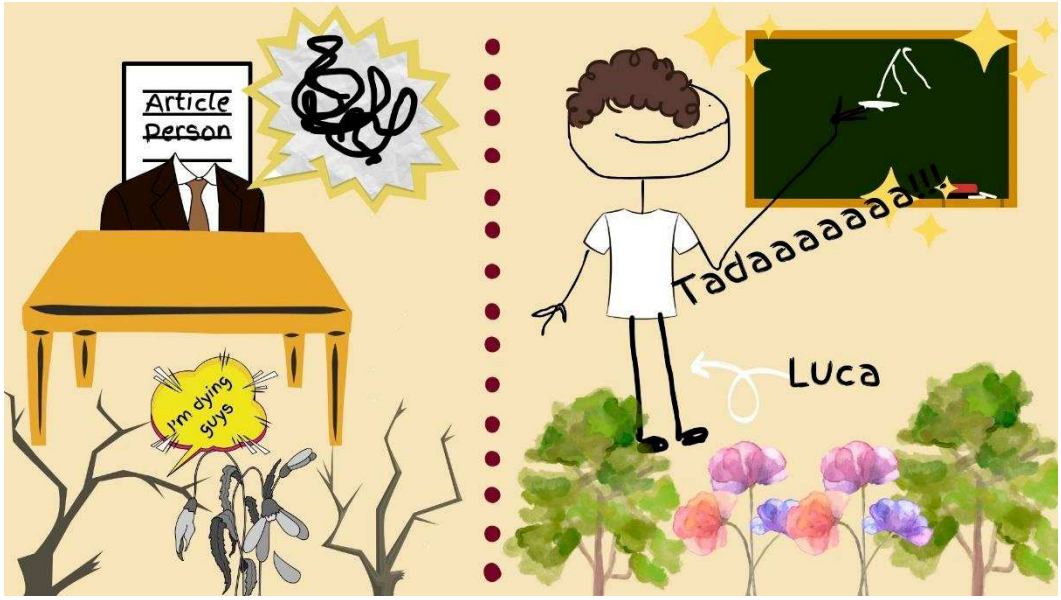

Table Of Figures

Abstract

Admixture, as well as gene flow with now-extinct hominins like Neanderthals and Denisovans, has significantly shaped the patterns of genetic variation in modern humans. In this research, we used IBDmix approach, a unique probabilistic method for collecting putative introgressed hominin sequences. Unlike other methods, IBDmix does not rely on a contemporary reference population. We used IBDmix to find and examine putative Denisovan ancestry that are present in contemporary humans in a dataset of 744 individuals from East Asia, South Asia, South East Asia, America and Oceanian populations. Our results show a greater Denisovan ancestry signal among Oceanian. These findings support the previous studies.

Keywords: IBDmix, Denisovan, Asia, Oceania

1 Introduction

1.1 Human History Through Genetic: From Genetic Alphabet To Ancestral Origins

In 1953 Francis Crick, Rosalind Franklin, James Watson, and Maurice Wilkins demonstrated that the DNA is written out in two chains of about three billion chemical building blocks. DNA (deoxyribonucleic acid) is a material that we inherit from our parents, and it exists in almost all other organisms. In a person's body almost all the cells have the same DNA. Most DNA resides in the cell nucleus (nuclear DNA), but a small proportion of DNA can also be found in the mitochondria (mitochondrial DNA or mtDNA). The information in DNA is stored as a code made up of four chemical bases named adenine (A), guanine (G), cytosine (C), and thymine (T) that can be thought of as the letters of an alphabet (Figure 1.1).

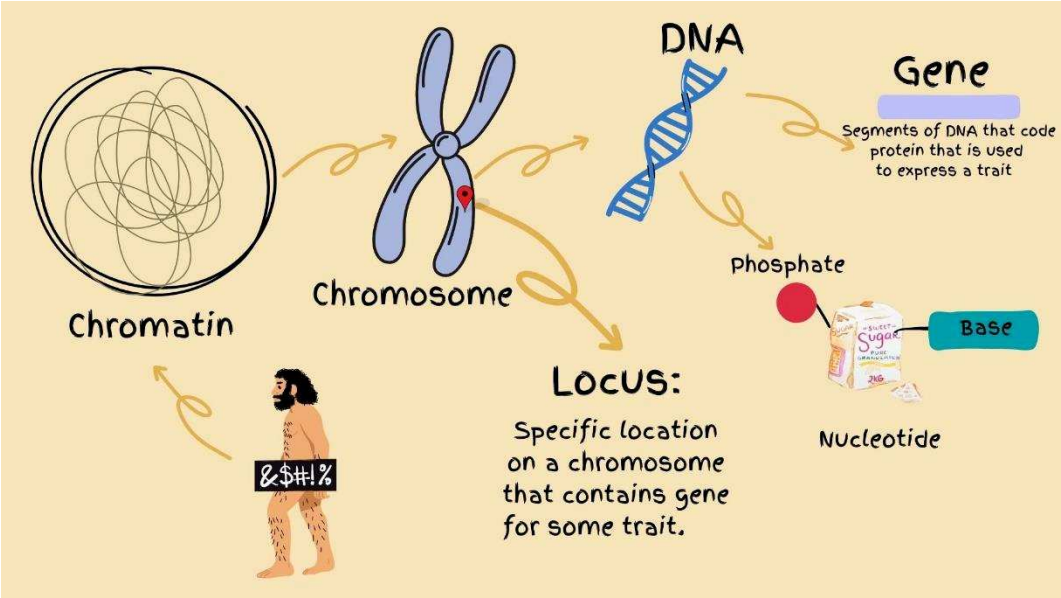

Figure 1. 1 Chromosome Structure

Genes are small segments of these chains that create the genome; they serve as templates for protein synthesis, the main task carried out by cells. Variation in the DNA sequence can arise from mutations –random errors that occur during genome replication, external factors such as certain types of chemicals or excessive radiation or internal factors such as an event that causes a problem with DNA replication in interphase. Normal rate of mutation caused by spontaneous errors in DNA replication in humans from parent to child is about $1.25\ \ 10^{-8}$ [1] mutations per base pair per generation. These mutations exert a substantial influence on our capacity to comprehend the past (figure 1.2). By analyzing these genetic variants, geneticists have the ability to gain insight into historical events. In addition, the analysis of

ancient DNA provides a window into the past, allowing us to investigate the genetic components of extinct species. Ancient DNA is any DNA that we can retrieve from the environment or from ancient remains of animals, which can be used to infer their demographic history. It's worth noting that the field of ancient DNA research includes the extraction of DNA sequences from a variety of unusual sources such as archaeological discoveries, museum specimens and fossil remains. Through the study of aDNA we have answered significant questions including the admixture between archaic and modern humans, gene flow and introgression.

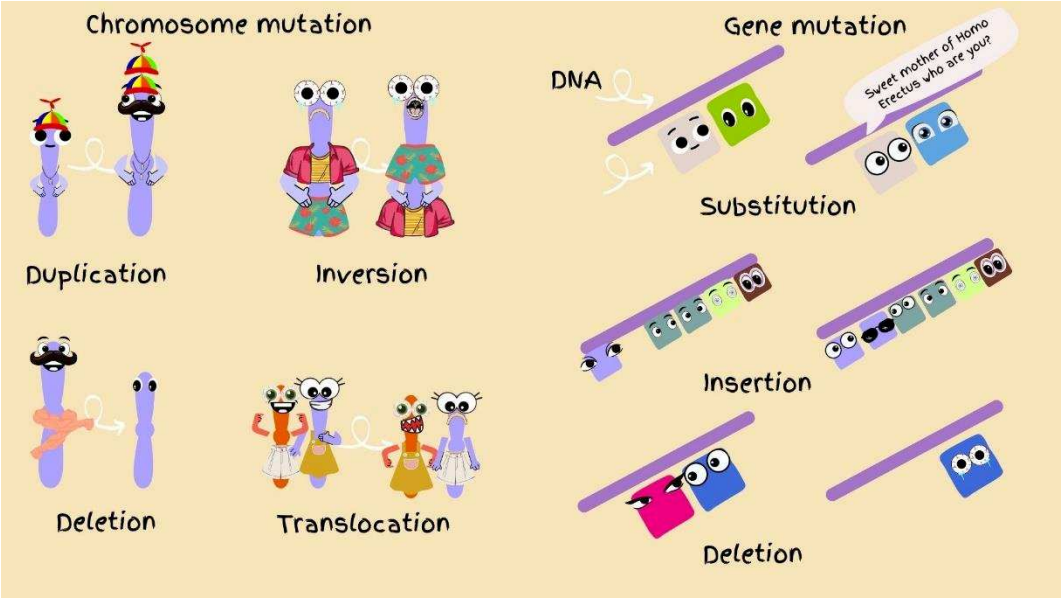

Figure 1. 2 Two Different Types of Mutations

Mitochondrial DNA played a crucial role in the first striking application of genetics to the study of the past. A few hundred mitochondrial DNA bases from people all around the world were studied in 1987 by Allan Wilson and his team. They created a family tree that identified maternal ties by comparing the distinctive mutations in these mitochondrian genomes. The most interesting finding was that sub-Saharan Africans belonged to the deepest branching mitochondrial haplotypes [2]. This shows that Africa is where modern humans' ancestors first appeared. Present day's non-Africans, on the other hand, all descend from a younger branch of the tree. Wilson and his colleagues calculated that "Mitochondrial Eve"[2], the most recent common ancestor of all mitochondrial lineages, lived roughly 200,000 years ago by looking at the rate at which mutations accumulate. But the most accurate estimate at the moment dates this coalescence to around 160,000 years ago [3]. (Figure 1.3)

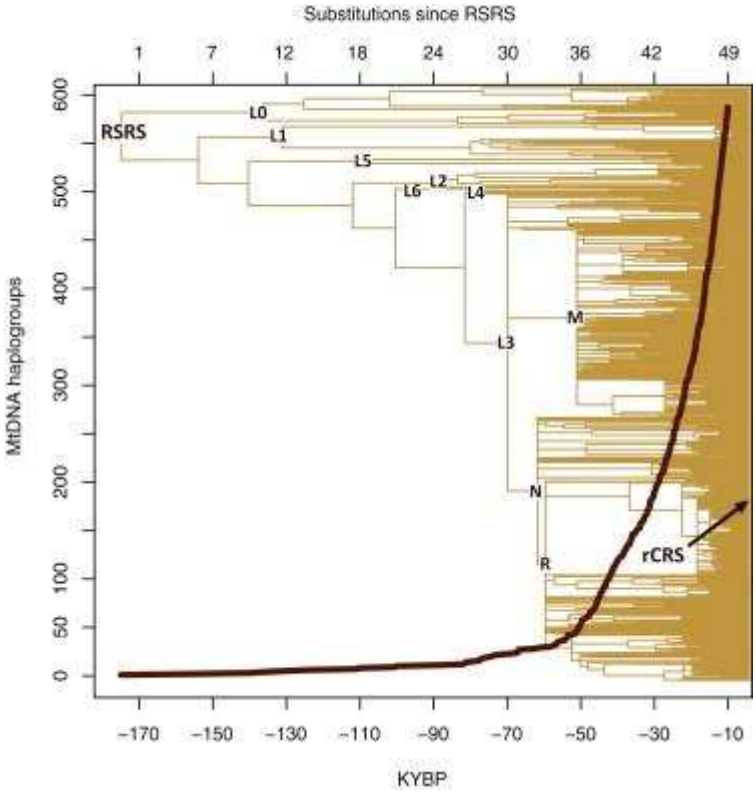

Figure 1. 3 A phylogenetic tree of human mitochondrial lineages adapted from Behar et al. 2012

In 2001, the human genome was sequenced for the first time, meaning most of its chemical letters were read. Around 70 percent of this sequence came from a single African American individual, with some contributions from others. By 2006, companies started selling machines that drastically reduced the cost of sequencing, making it affordable to map the genomes of many more people. This breakthrough allowed for the comparison of sequences from the entire genome, not just isolated regions like mitochondrial DNA. Consequently, it became possible to reconstruct the numerous ancestral lines of descent for each individual. The genome does not represent a continuous sequence from a single ancestor; instead, it is a mosaic consisting of 23 pairs of chromosomes, with one of each pair inherited from each parent. Since each person carries two pairs of each chromosome, one from each parent, the total count is 46. The driving force behind the mosaic structure of our genome is recombination which is a process that occurs during the formation of a person's sperm or eggs, swapping DNA segments of parental DNA to create novel recombined chromosomes that are subsequently inherited by the offspring. (Figure 1.4)

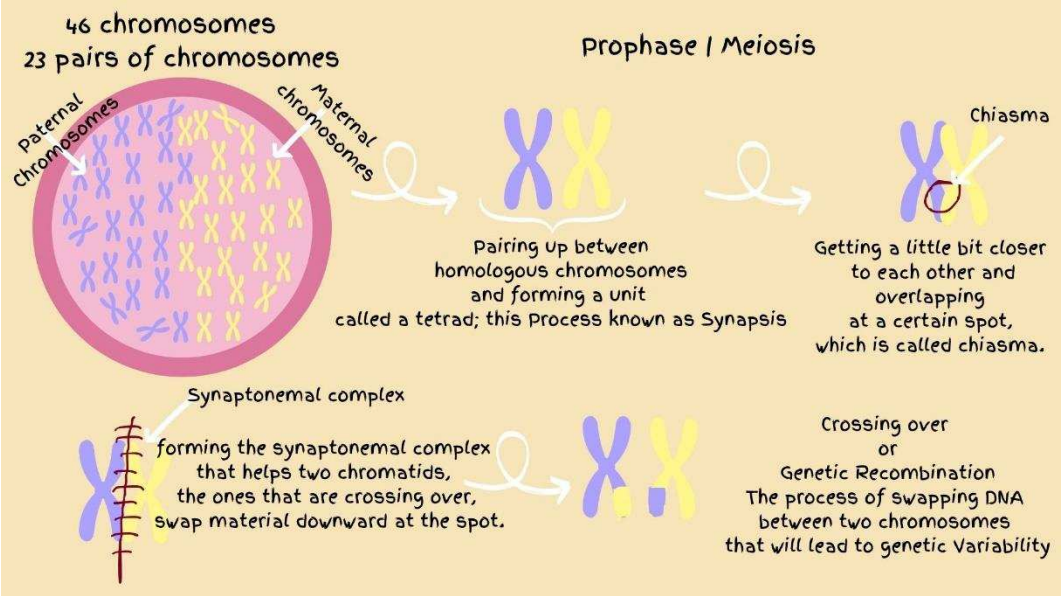

Figure 1. 4 The Process of Recombination

Over the past few years, thanks to aDNA many unexpected connections between human populations have been uncovered and now we know that the story of our ancestors' evolution is more complicated than what is common in popular culture or what we've learned in school or watched in movies. This story is not about some beast-like beings that lacked understanding. It is about beings who took care of their elders, buried their dead and were able to create art.

1.2 Ancient Human: Neanderthal and Denisova

The first modern humans, *Homo sapiens*, evolved between 200 and 300 kya in Africa [8,9,10] and about 70000 years ago they began leaving Africa [8,9,10]. Meanwhile, other species of the Homo genus such as Neandethals (named after "Neander Thal" or "Neander Tal", the German name for Neander Valley where the first fossil has been found) were in West Eurasia and Denisovans were mostly in East Eurasia and Asia between 440 and 40 thousand years ago. Neanderthal and Denisovans existed in Eurasia for hundreds of thousands of years which led to interaction between them and modern humans, however after 40 kya they were replaced by modern humans.

Scientific evidence supports the notion that modern humans and Neanderthals interacted [4]. A significant piece of this evidence comes from western Europe, where Neanderthals disappeared roughly 39 kya [4]. Modern humans arrived in western Europe at least several thousand years earlier, as demonstrated at Fumane in northern Italy. Around 44 kya, Neanderthal-style stone tools gave way to tools characteristic of modern humans. In western Europe, tools made in the Châtelperronian style were discovered alongside Neanderthal remains dating from

44 to 39 kya [4]. This suggests the possibility that Neanderthals either mimicked modern human tool-making techniques or that the two groups engaged in tool or material trade. However, it's important to note that not all archaeologists concur with this interpretation, and there is an ongoing debate regarding whether Châtelperronian artifacts were crafted by Neanderthals or modern humans. (Figure 1.5)

Figure 1. 5 Chatelperronian Tools and Their Location

Meetings between Neanderthals and modern humans occurred both in Europe and the Near East. Notably, the Near East had already been inhabited by modern humans, as evidenced by remains found at Skhul Cave on the Carmel Ridge in Israel and Qafzeh Cave in the Lower Galilee, dating back to about 130,000 to 100,000 years ago [5]. Neanderthals also occupied the region, with one Neanderthal skeleton discovered at Kebara Cave on the Carmel Ridge, dating back to a period between 60,000 and 48,000 years ago [6]. Today's non-African populations have 1.5–2.1 percent of Neanderthal ancestry in their genome [7]. Interestingly, despite Europe being the Neanderthals' native continent, East Asians have a higher proportion of Neanderthal heritage compared to the one of present day Europeans [7]. Europeans are the result of multiple admixtures, and their demographic history might have influenced levels of Neanderthal ancestry. The earliest Europeans, who encountered European Neanderthals, were more closely related to East Asians than modern European populations are [7], and they were later replaced by other migrants after all Neanderthals went extinct [7]. Europeans further received gene flow from other Eurasian populations and maintained ongoing gene flow with African populations. (Figure 1.6)

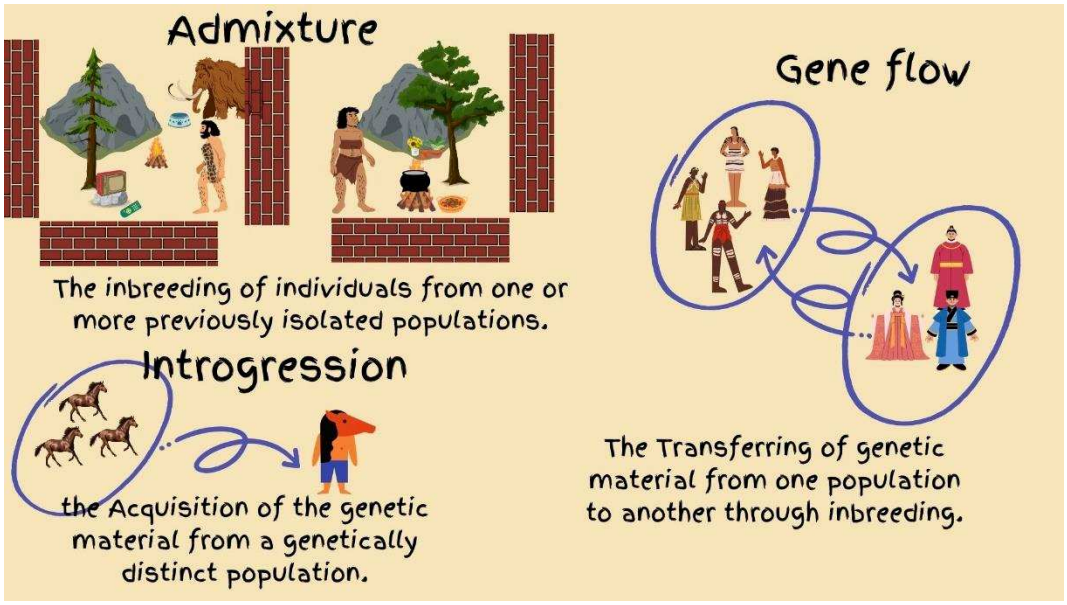

Figure 1. 6 Important Definitions

In 2010, another significant discovery emerged as the Denisovans, a distinct human group, were unveiled through the examination of DNA extracted from a finger bone (Figure 1.7) found in Denisova Cave [8]. This discovery stands as one of the most remarkable contributions of ancient DNA research. By sequencing mitochondrial DNA (mtDNA) from the distal phalanx of a pinky finger (known as Denisova 3) located in layer 11.2 of Denisova Cave, which dates back to a period between 63 and 55 thousand years ago. The mitochondrial genome of Denisova 3 differed from that of contemporary humans at twice as many genetic positions compared to the disparity between Neanderthal and present-day human mtDNA sequences [9].

Figure 1. 7 Denisova 3 Pinky Finger Bone

In 2010 Krause and colleagues [8] reported that Neanderthals exhibited distinctions from contemporary humans at an average of 202 nucleotide positions, while the Denisova showed variances at an average of 385 positions. In comparison, chimpanzees differed at 1,462 positions. The Denisova hominin's mitochondrial DNA (mtDNA) displayed nearly twice the number of distinctions from present-day humans than Neanderthal mtDNAs. Notably, A phylogenetic analysis reveals that the Denisova hominin mtDNA lineage diverges significantly earlier than both modern humans and Neanderthals. Assuming an average divergence of human and chimpanzee mtDNAs around 6 million years ago [8], the estimated date for the most recent common mtDNA ancestor shared by the Denisova hominin, Neanderthals, and modern humans is approximately one million years ago [8]. The nuclear genome sequence demonstrates a comparable average divergence between Denisovans and contemporary humans similar to what is observed between humans and Neanderthals. Denisova3 possesses ancestral alleles at 11.7% of the positions that differ between modern humans and chimpanzees, compared to 12.2% in the Neanderthal genome [10].
While we are uncertain about the exact geographical extent of Denisovan area, the genetic dissimilarity observed between Neanderthal and Denisovan genomes implies that these populations remained separate for a minimum of 300,000 years [9,10]. Nevertheless, genetic evidence suggests that there were sporadic interactions and interbreeding between Denisovans and Neandertals, especially within the vicinity of Denisova Cave [10]. It appears that these encounters between Neandertals and Denisovans were more frequent among the early Denisovan population residing in Denisova Cave than among their later counterparts [10].

1.3 Ancient Encounters in Asia and Beyond

Preliminary examinations of the Denisovan genome uncovered the presence of Denisovan DNA in contemporary Oceanian populations. This finding implies that the ancestors of these present-day individuals encountered Denisovans and interbred with them. By delving into the details of this interaction, scientists have gained valuable insights into the population history of modern humans in this region and have expanded our knowledge of the geographic range and diversity of Denisovans.

The initial comparison of Denisovan and Neanderthal genomes with those of modern humans yielded noteworthy insights. It was observed that Melanesians from Papua New Guinea and Bougainville Island in Oceania had a greater genetic affinity with Denisovans compared to other non-African populations [11]. Given that Denisova 3 predated the arrival of modern humans in southern Siberia, it was more reasonable to assume that Denisovan ancestry in Melanesians resulted from interactions between Denisovans and the ancestors of present-day Melanesians, rather than the other way around. This Denisovan genetic contribution to the

ancestors of modern Melanesians occurred subsequent to the Neanderthal gene flow into the ancestors of all non-Africans. (Figure 1.8)

Figure 1. 8 Who has Denisovan Ancestry?

Subsequent analysis of a broader selection of contemporary human genomes from Asia and Oceania confirmed the widespread presence of Denisovan ancestry. Indigenous Oceanian populations to the east of Wallace's Line, such as Melanesians, Australians, and Indigenous Philippine groups, were confirmed to carry Denisovan ancestry [12]. The methods used to detect this ancestry demonstrated a high level of confidence. Asians and Native Americans were also identified as having Denisovan ancestry, albeit at lower levels, approximately 0.2%, which is significantly less than the levels observed in Melanesians [13,16].

The extent of Denisovan ancestry varies among present-day populations, with Philippine Negritos, like the Ayta, displaying the highest levels [10]. Other Oceanian groups exhibit intermediate levels of Denisovan ancestry, which are generally higher than those in mainland Asians and Native Americans. Additionally, traces of Denisovan ancestry are present in some European populations, potentially inherited from Neandertals with Denisovan ancestry or due to admixture with Asian populations [10].

1.3.1 Denisovan Imprints Across Oceania and SouthEast Asia

Oceania and Southeast Asia, regions where Denisovan genetic heritage is prevalent, were initially settled by early modern humans approximately 50-45,000 years ago [10]. Descendant populations of this initial human dispersal include Negrito groups in Southeast Asia and Indigenous communities in areas such as Australia, New Guinea, the Solomon Islands, New Caledonia, Fiji, and Vanuatu. However, these groups exhibit varying degrees of Denisovan genetic influence.

In 2021 Larena and colleagues [16] focused on the genetic affiliations and population structures in Philippine Negritos. Their study revealed that Neanderthal ancestry is uniformly detectable in all Philippine ethnic groups, however Denisovan ancestry varies significantly. Their research confirmed that Ayta Magbukon Negritos have the highest levels of Denisovan ancestry in the world, with levels that are 34%–40% higher than those of Australians or Papuans. High-coverage whole-genome sequences of Ayta Magbukon Negritos were compared with Australasians, and the results confirmed this significant difference in Denisovan ancestry. It was also noted that the high levels of Denisovan ancestry in Ayta Magbukon are not due to a recent Denisovan admixture event, as the tract lengths are similar to those in Papuans. This suggests that the Denisovan admixture in Ayta Magbukon is of similar age to that in Papuans. This observation aligns with recent research indicating multiple instances of Denisovan introgression into human populations and the widespread presence of Denisovans throughout Island Southeast Asia (ISEA), where Ayta Negritos likely experienced a second Denisovan introgression event.

1.3.2 Tracing Denisovan Ancestry from Asia to America

In 2018 Browning et al [17] plotted a two-way densities of match rate (a measure of how closely certain genetic segments in East Asians align with the corresponding segments in the genome of Altai neanderthal and Altai Denisovan) to the Altai Neanderthal and Altai denisovan genomes to gain insight into archaic segments in modern humans. They focused on segments with at least ten positions that could be compared to both Altai Neanderthal and Altai Denisovan genomes. In East Asian populations (Dai, Beijing and Southern Han) they observed a cluster of genetic segments inherited from Denisovans. These segments had a wide and bimodal distribution of similarity to the Altai Denisovans. They performed a statistical test to understand if there are two distinct components of Denisovan ancestry and showed that there are two distinct components in these populations. About one-third of the Denisovans segments in Japanese and Chinese populations were more closely related to the Altai Denisovan genome. The specific genetic variations in the high-affinity component showed a similarity of around 80% to the Altai Denisovan genome, like how Neanderthal-introgressed genetic variations match the Altai Neanderthal genome. The other component had genetic variations with similarity of about 50% to the Altai Denisovan genome. This suggested a more complex Denisovan ancestry in these populations.

They also found out some populations such as Finns and Native Americans seem to carry a small proportion of segments inherited from Denisovan.

1.4 Aim of the Thesis

The aim of this study is to utilize IBDmix, a unique probabilistic method for locating introgressed hominin sequences. Unlike other methods, IBDmix does not require an unadmixed reference population. We employ IBDmix to identify and analyze Denisovan segments within a dataset of 744 individuals from EastAsian, SouthAsian, SouthEastAsian, Oceanian and American populations, with the goal of gaining a deeper understanding of Denisovan admixture and gene flow among them.

2 Materials & Methods

2.1 Ancestry Quantification: Population Genetics Technique

Studies of ancient DNA are changing our understanding of human evolutionary history, specifically how admixture has formed past of human genomic variation. There are methods that identify introgressed segments in modern human genome. In this section we will look at how Sprime and IBDmix work by shedding light on their methodologies.

2.1.1 Sprime Methodology

Sprime method is used to analyze the genetic information from different groups of people to identify segments of DNA that might have come from archaic human populations like Neanderthals and Denisovans.
This method looks at the genome of specific individuals and compares it to the one of an outgroup chosen because it is believed to have had no direct intermixing with the archaic population we're interested in. (Figure 2.1)

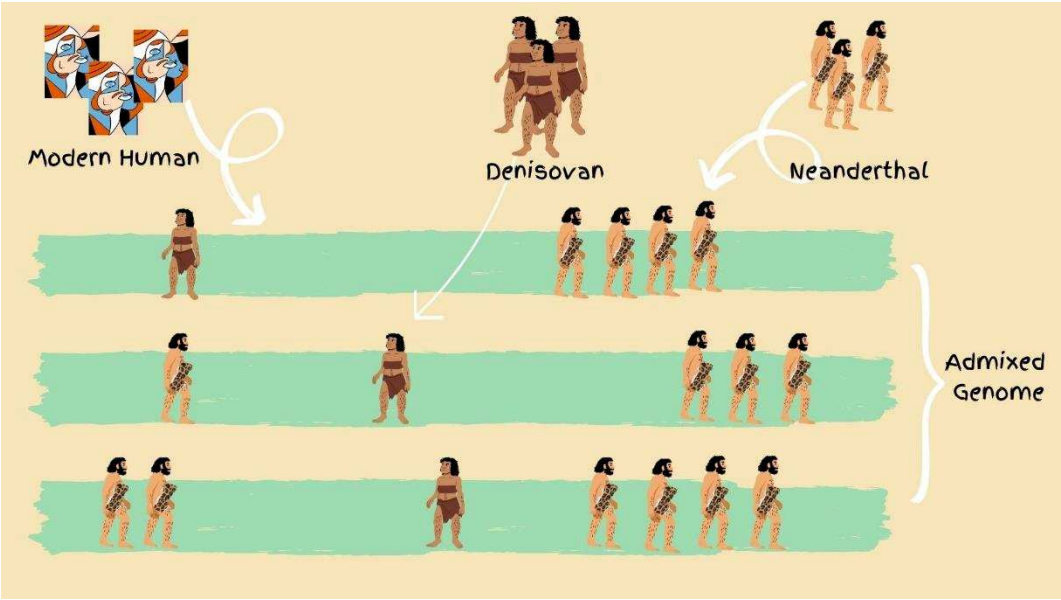

Figure 2. 1 Schematic Representation of an Admixed Genome

Sprime's result is a list of DNA segments that came from the archaic population also we can recognize the genetic variations within those segments. By using this informations we can understand which parts of a person's genome might come from archaic populations. It is worth noting that Sprime categorizes genetic variations into three groups: 1) common in outgroup, 2) not seen in the group, 3) uncommon but present in the outgroup. The focus is on the uncommon ones.

Sprime Algorithm

Identifying Alleles and Frequencies:

- Scientists calculate a score called "T(v1, v2)" for a pair of genetic alleles called v1 and v2. This score tells us about the relationship or recombination between these alleles in a group of individuals we're studying and it will give a specific score for a pair of alleles.
- If these two alleles come from the same ancient source and haven't changed due to recombination, we expect Xi1 to be the same as Xi2 for everyone (Xi1 = Xi2). But sometimes, there might be changes in between because of recombination. We count these changes as "D." D tells us how many of these changes we see. $D=\Sigma|Xi1 - Xi2|$
- Sprime's Score divides into two parts: 1) sum of the positive part that depends on the local mutation and recombination rate. 2) The negative part that penalizes the score when D is larger than the introgression frequency. The value n is defined as the number of haplotype carrying the alleles. When there are many introgression haplotypes we expect more recombination consequently a larger D value. So, they used the normalized value D/n in their pairwise score.
- Alleles will be chosen based on allele frequency in outgroup. The focus is on rare alleles in the outgroup ($f \leq 0.01$). These rare alleles get a lower score because more common alleles are often not of ancient ancestry. (Figure 2.2)

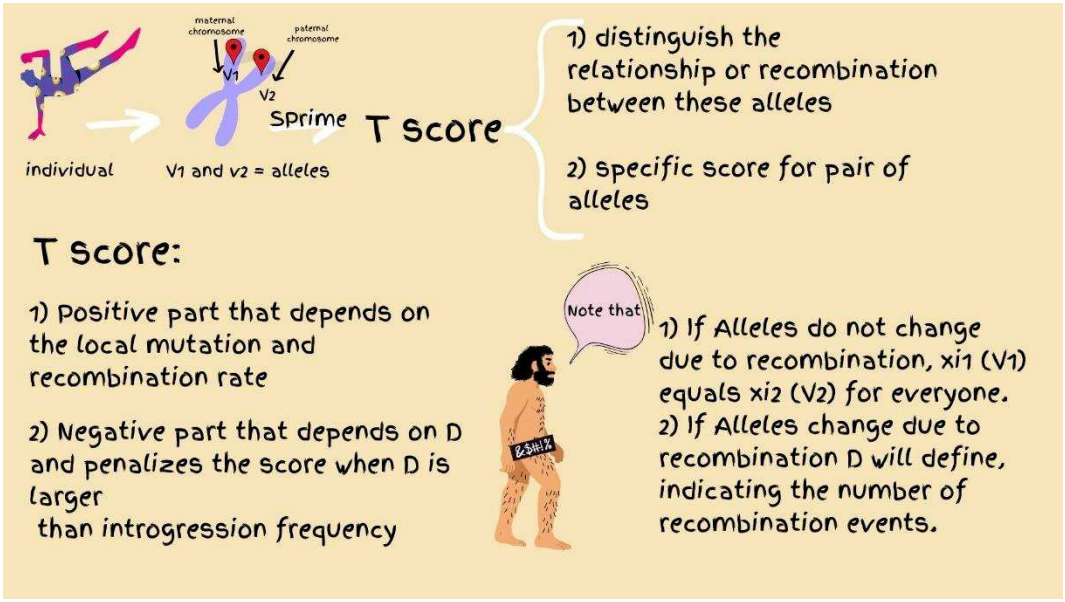

Figure 2. 2 Sprime TScore

Incorporating Mutation and Recombination Rates:

- The consideration of differences in mutation and recombination rates in different regions by Sprime. In regions where mutations happen frequently but recombination is rare, many pairs of high-LD alleles can be found, but this might not be due to ancient introgression. In Contrast, regions with low mutation rates

and high recombination rates will have fewer high-LD pairs, whether they are archaic or not. The ratio of the mutation rate to the recombination rate determines the rates at which we can expect to find high-LD pairs of alleles, both in introgressed and non-introgressed segments. 'm' represents the local rate of mutation per centimorgan per meiosis. If high-LD pairs of alleles are in a region with a low mutation rate per centimorgan, they will score high. However, when 'm' is too low to prevent excessive variability, they do not score alleles. Therefore, a specific function is created for the local rate of mutation per centimorgan. The function is roughly proportional to 1/ (100 m) for large 'm,' and it has a maximum value of around 1.6. (Figure 2.3)

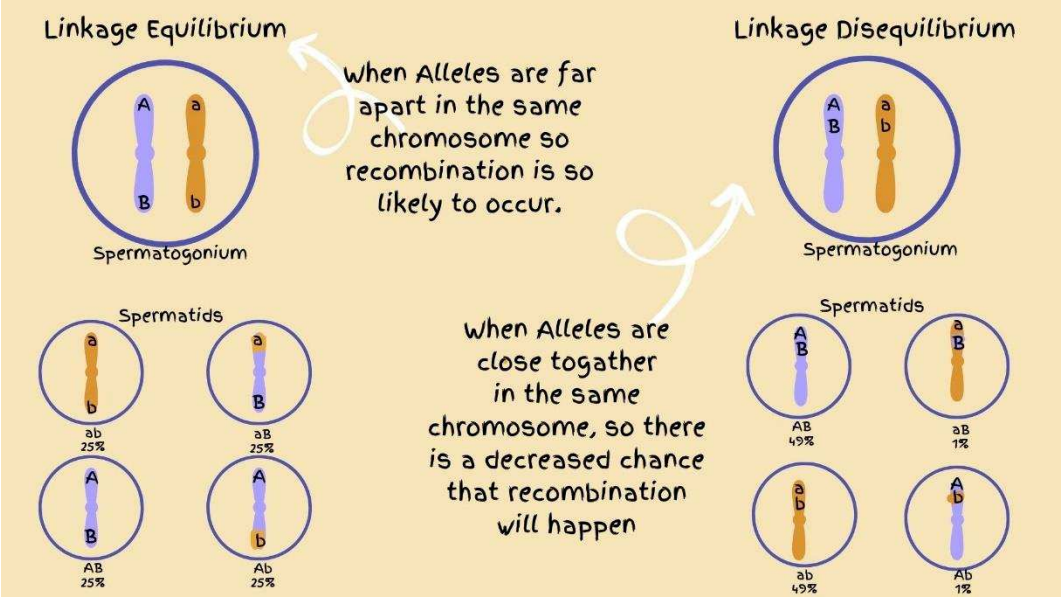

Figure 2. 3 Schematic Representation of the Differences Between LD and LE

m = (local rate of mutations per bp per meiosis) * (local bp per cM rate)

- local rate of mutations per bp per meiosis = local variant density / global variant density * Mg. For the local rate of mutation, it is possible to local rate of assayed variation or use the number of differences between human reference and a primate reference sequence. Mg is a genome-wide mutation rate ($1.2 * 10^{-8}$). The local and global densities are attainable from vcf input files because the number of variant positions in the region is divided by the number of base pairs at that location. For the local rate of recombination, they used HapMap genetic map [18], that is a map based on LD patterns.
- The estimated mutation rate and recombination rate may not be very accurate and can be different between different populations when they deal with small genetic distances. So, they combine estimates from a small region of interest with estimates from larger surrounding regions, going up to 10,000 bp in both directions. They also consider regions with at least 6 genetic variations and stop when they find a region with at least 10 variants. Their approach leads to overestimating rather than underestimating the mutation rate ('m').

2.1.2 IBDmix Methodology

IBDmix is a method use to analyze genetic data from whole-genome sequences. It will take the genetic data of one archaic reference individual and a group of modern humans as an input. Unlike most other methods, IBDmix doesn't require an unadmixed modern human reference population to account for shared genetic material between archaic and modern humans. (Figure 2.4)

The basis of IBDmix is the idea of "identity by descent" (IBD), which suggests that two people share a genetic sequence that was passed down from a common ancestor. One pair of the ancient and contemporary human genomes are compared at a time as the genomic material is processed site by site. IBDmix computes the likelihood of IBD between the archaic and contemporary samples based on allele frequencies and expresses this as a LOD score for each site that meets specific requirements (variant filtering).

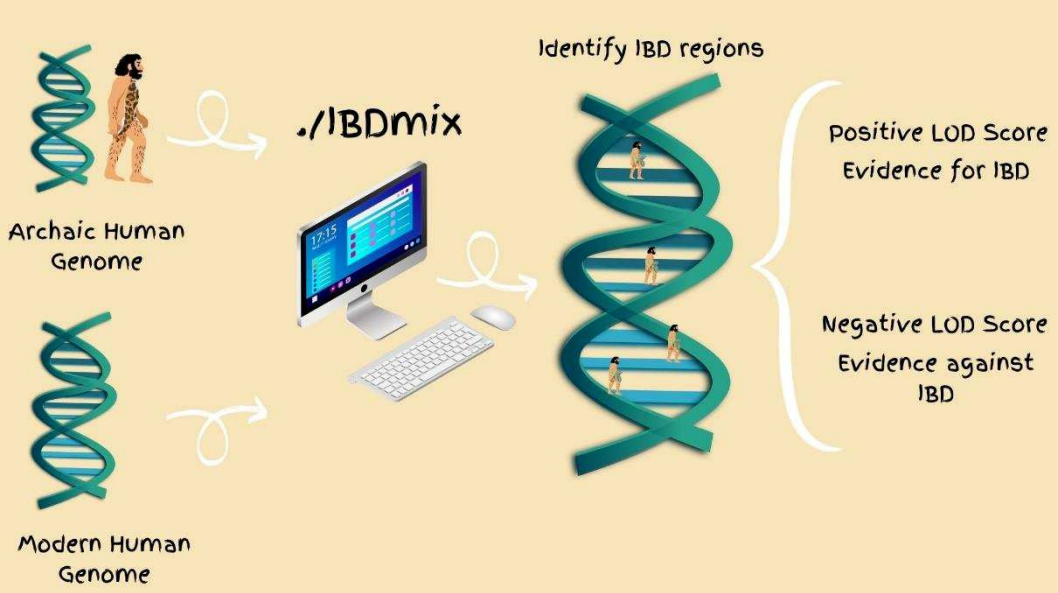

Figure 2. 4 How IBDmix Works

IBDmix uses a scanning algorithm based on dynamic programming to detect the putatively introgressed segments in the modern human genomes. This algorithm will maximize the sum of LOD scores for regions that have a threshold above the predefined threshold. So, the dynamic part comes in when it decides where to stop look for introgressed regions. It will add positions until the total score starts to become negative. In this way, it can find the longest stretch of positions with high score that are more likely be the introgressed regions.

IBDmix Algorithm

IBDmix LOD Score Calculation

- IBDmix calculates LOD scores for each allele. These scores tell us the likelihood of a shared genetic sequence (IBD) versus a non-shared sequence (non-IBD). Positive LOD scores suggest evidence for shared ancestry (IBD), while negative scores suggest otherwise.
- To compute these scores, the software uses allele frequencies and compares IBD and non-IBD models. So, in simple terms it looks at the genomes of both the ancient human and the modern human. It compares their genomes and uses the probability that they share IBDs from the same ancestor.
- IBDmix only applies these scores to variants that pass its filtering criteria, so it doesn't consider the ones that are excluded.
- If modern human genotype data is missing and the archaic genome has alternative alleles, the missing data is imputed.
- IBDmix reports segments where the sum of LOD scores reaches a maximum for each pair of alleles; it suggests that this region likely comes from a common ancestor using a dynamic programming-based scanning algorithm.

Allele Error Rates in IBDmix Calculation

- In IBDmix, independent allele errors are assumed.
- For archaic genomes, they set the allele error rate (ɲ) to 0.01.
- For modern human genomes, the error rate (Ɛ) depends on the minor allele frequency (MAF).

Estimating Likelihoods for IBD and Non-IBD Genotypes with Allele Errors

- The likelihood of observed genotype is then calculated, considering both IBD and non-IBD scenarios.
- IBDmix assumes errors may occur when observing alleles for both human and archaic genomes ($ɲ \geq 0$), and these errors are independent.
- Under the IBD model, denoted as P0(.|IBD) and P0(.| I), IBDmix considers cases where one archaic individual and one modern human share an allele from a common ancestor.
- In the non-IBD model (P0(.|nonIBD) and P0(.|nI)), there is no shared ancestry, and individuals are ordered (archaic first, modern human second), but genotype order doesn't matter.
- P0(.|IBD) represents the probabilities of observing genotypes with errors under the IBD model, while P0(.|I) represents the probabilities for true genotypes without errors.
- IBDmix works with variants that have two alleles: a reference allele A and an alternative allele B. PA and PB are the observed frequencies of these alleles in the modern population.

Variant Filtering

- Before using IBDmix, the genotype data of the ancient human should be filtered. It is necessary to remove multi-allelic SNPs and indels from the archaic genome.

Also, eliminate all variants with one or fewer minor allele counts in the target sample.

- If a gene variation is missing in the archaic genome, IBDmix doesn't look at it. If it's present in the ancient genome but missing in modern humans, it only considers it if the ancient sample has at least one alternative allele. In this case, the modern human genotypes are 'imputed' as homozygous for the reference allele.

Sample size in IBDmix

- IBDmix needs a minimum of 10 individuals to ensure that the estimates are accurate. More individuals are even better.

2.1.3 Compare Sprime and IBDmix

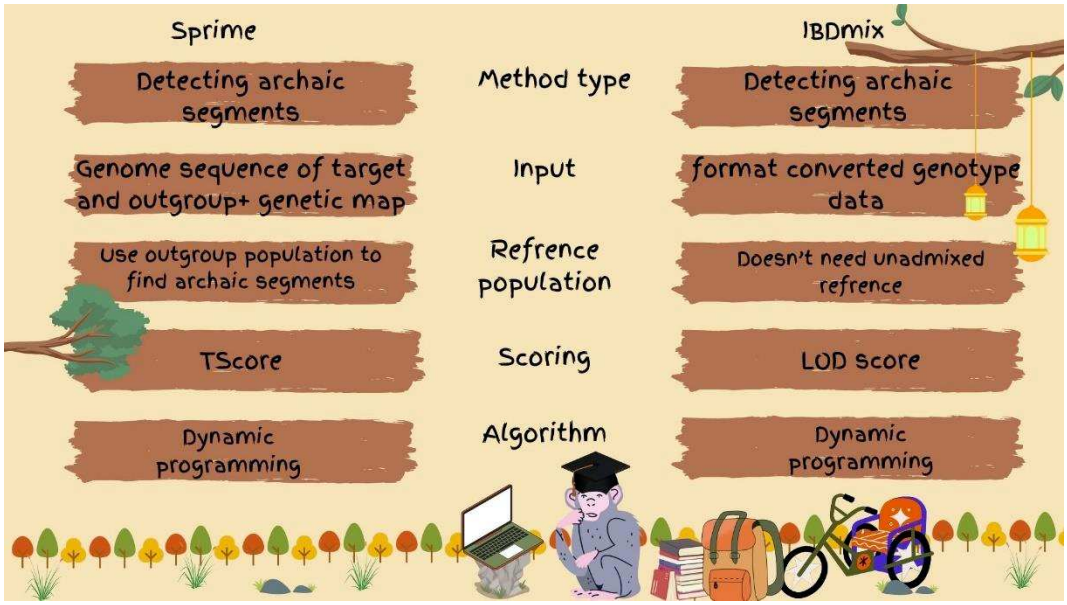

Figure 2. 5 Comparing Sprime and IBDmix

Sprime

- **Method Type:** Sprime is a method for detecting segments of archaic introgression in modern human genomes.
- **Data Input:** It requires a whole genome sequence of target and outgroup individuals with a genetic map.
- **Outgroup Population:** It uses an outgroup population (Yoruban individuals) to detect archaic introgression.
- **Scoring:** Sprime uses an T Score to identify introgressed segments and archaic-specific alleles.
- **Algorithm:** It uses a dynamic programming algorithm.
- **Output:** Sprime provides a list of detected introgressed segments and the associated archaic-specific alleles.

IBDmix

- **Method Type:** IBDmix is for detecting putatively introgressed archaic segments in modern human genomes.
- **Data Input:** It requires format-converted genotype data from whole genome sequencing for an archaic reference individual and a group of modern humans.
- **Reference Population:** IBDmix does not require a modern human unadmixed reference.
- **Scoring:** IBDmix calculates LOD scores to estimate the likelihood of shared ancestry (IBD) versus non-shared ancestry (non-IBD).
- **Algorithm:** It uses a dynamic programming-based scanning algorithm.
- **Output:** IBDmix provides a list of putatively introgressed segments with maximized LOD scores. (Figure 2.5)

2.2 Overview of our Study

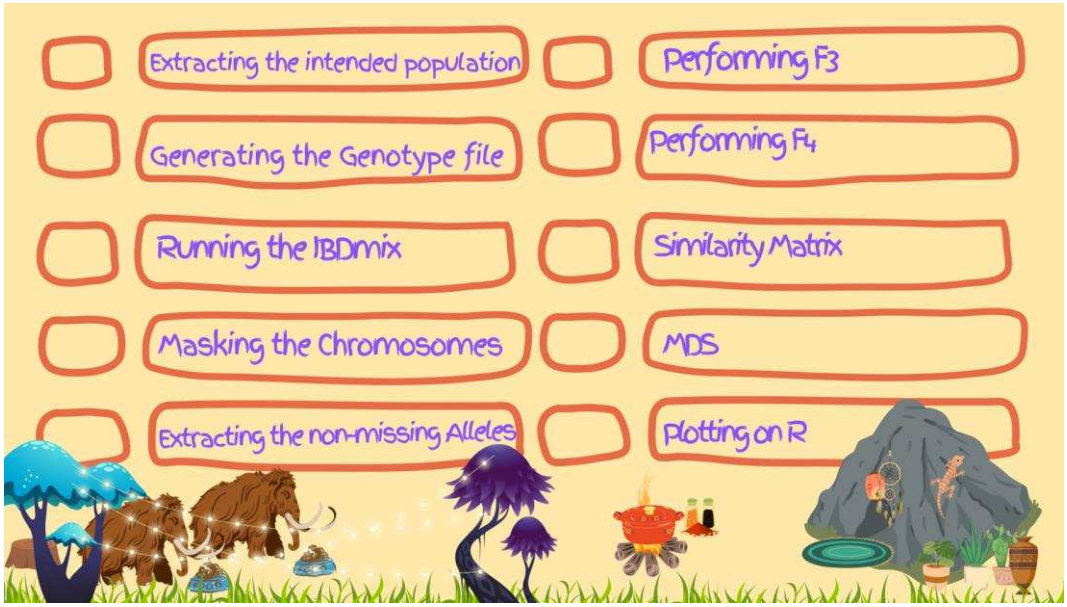

Figure 2. 6 Our Study in a Nutshell

744 individuals from East Asia, Oceania, South Asia, SouthEastAsia, and America were included in this study to increase the geographical coverage of Denisovan. (Figure 2.6)

2.2.1 Extracting Intended Population

In this step, we used bcftools to extract the intended population from modern human chromosomes. Then, we change the format of the first column from alphabet to an integer, as it is an acceptable format for IBDmix.

```
for n in {1…22}; do
    echo $n;
    bcftools view -O z -o chr$n.output.vcf.gz -S
Sample.txt chr$n.vcf.gz;
done
```

2.2.2 Generating Genotype File

In this step we execute the `generate_gt` program, which generates a genotype file with archaic and modern human genomes. The format of this file is required by IBDmix and uses two input VCF files (`-a` for the archaic VCF file and `-m` for the modern human VCF file) to process genetic variant data.

Information on genetic variants for both ancient and contemporary human populations can be found in the genotype file used by IBDmix. The output genotype file consists of 5 columns, including chromosome, chromosome position, reference allele, alternative allele and PinkyDenisova, respectively. PinkyDenisova is a column with binary numbers. 1 represents the reference allele, and zero represents the alternative allele. The last numerical column represents the genotypes of SNPs for each individual at that specific position. (Figure 2.7)

1. Archaic VCF File (-a): the archaic VCF file that contains multiple archaic samples such as Neanderthals or Denisovans.
2. Modern Human VCF File (-m): the modern VCF file
3. Output (-o): The merged genotype file output.

```
./generate_gt -a archaic.vcf -m modernhuman.vcf -o
genotype.txt
```

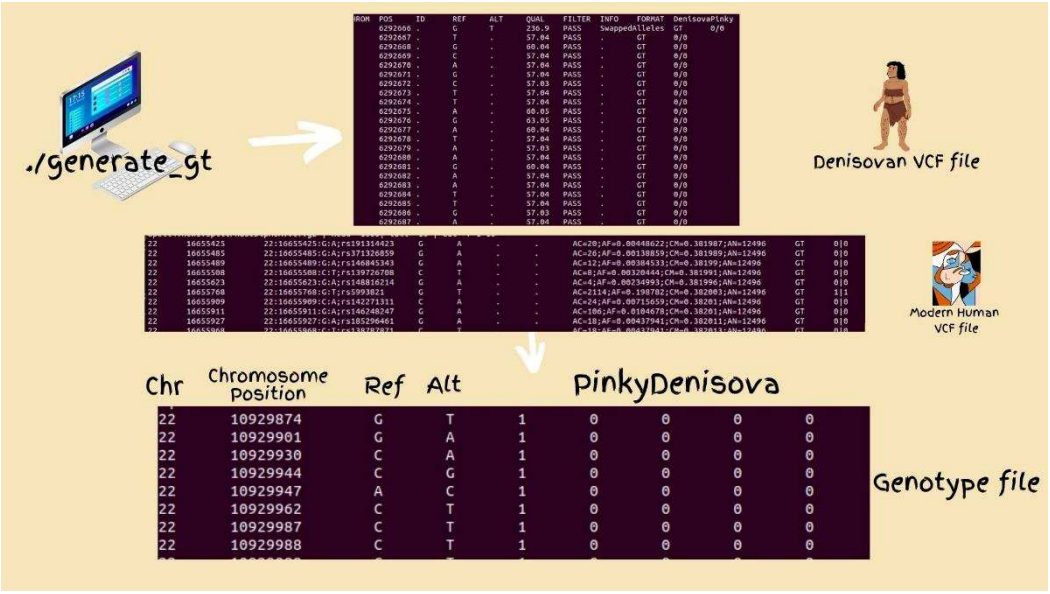

Figure 2. 7 Schematic Representation of the input VCF files and Genotype file

2.2.3 Running IBDmix

In this step IBDmix will detect putative introgressed regions. Running IBDmix requires to specify a number of parameters, the ones that we used are:

- **./ibdmix:** This is the command to execute the "IBDmix" program. The `./` indicates that the program is located in the current directory.
- **-g :** This specifies the input genotype file to be used by the "IBDmix" program.
- **-o:** This specifies the output file where the results of the analysis will be saved.
- **-n DenisovaPinky:** This provides a label or identifier for the analysis.
- **-d 3.0:** This sets the threshold for detecting segments of shared IBD. In this case, segments with a length of at least 3.0 cM (centimorgans) will be considered.
- **-m 1:** This parameter specifies the minimum number of markers required to identify a shared IBD segment. A value of 1 means that even single-marker segments will be considered.
- **-a 0.01:** This sets the significance level for the IBD detection. In this case, segments will be considered significant if they have a probability of less than or equal to 0.01 of being due to chance.
- **-e 0.0025:** This parameter controls the estimation of population allele frequencies. The value of 0.0025 indicates that the program will estimate allele frequencies every 0.0025 cM.

```
./ibdmix -g genotype.txt -o ibdmix.txt -n DenisovaPinky
-d 3 -m 1 -a 0.01 -e 0.0025 -c 2
```

2.2.4 Masking the Chromosomes

In this step, we utilized a Perl script developed by Prof. Luca Pagani. This script is specifically designed to process modern human VCF.gz files and the output generated by IBDmix as its input. The primary objective of this script is to perform masking based on the provided variant information. It effectively masks the non-introgressed segments within the dataset, resulting in the creation of a new VCF file that contains only the putatively introgressed genomic regions. It is noteworthy that we maintained the data integrity of one African individual from the Yoruba population, identified as NA18486, by not applying the masking procedure to their genetic data. (Figure 2.8)

```
perl MaskVcfFromIBDmix_Keepoutgroup.pl chr.vcf.gz
ibdmix.txt masked.output.vcf NA18486
```

Figure 2. 8 Masking Process Using the Perl Script

2.2.5 Extracting The Non-Missing Alleles

In this step, our objective is to extract non- missing SNP list for each individual that exhibits complete data availability based on information from the lmiss files. The `lmiss` file is a text file with a header and n lines for each variant, containing 5 columns, including CHR (chromosome number), SNPs (identifier for the variant), NMISS (indicating the count of missing genotypes at a particular locus), NGENO (representing the number of genotypes, which remains constant at zero due to the haploid nature of individuals, but following chromosome phasing, they are duplicated to diploid status to conform with IBDmix requirements), and F-MISS (reflecting the absence of missing data for the SNP), respectively. (figure 2.9) We used Plink2 [20] with the following command to create lmiss txt files:

```
plink --bfile file  --keep one_individual_to_keep.txt --missing --out output
```

Figure 2. 9 How lmiss Text File looks like?

We examine the lmiss file and isolate SNPs that are characterized by the "010" pattern within the final three columns.
We undertook this step to enhance the efficiency of our calculations. Calculating the f3 statistic on a single comprehensive Tped file is a time-consuming process. Therefore, in each iteration, we create a Tped file consisting of three intended individuals (indi, indj, NA18486) and only at the positions that are non-missing in all three; we subsequently run the f3 statistic on this smaller Tped file, which significantly increases the speed of the computation.

```
#prepare list of individuals to process
cut -f 1 -d " " merge_all.fom > indlist.txt

for ind in $(cat indlist.txt); do

#prepare file with single individual to keep for this iteration
echo ${ind} ${ind} > ind2keep.txt

#compute missingness for this single individual
plink –bfile merge_all –keep ind2keep.txt –missing –out $ind
```

```
# extract line where the genotype is non-missing and
save SNP name to file
grep $0 $ind.lmiss | awk '{print $2}' >
$ind.nonmissingssnps

#remove unnecessary files
rm $ind.lmiss $ind.imiss $ind.nosex $ind.txt
done
```

2.2.6 Running F3

To perform this step, we utilize popstats.py [19], Python version 2, Plink 2 [20].The "three-population test" often known as the f3-outgroup statistic, is a statistical method frequently used to evaluate genetic closeness between three populations or individuals. A higher value of F3-outgroup indicates that the intended individuals/populations are closer to one another. (Figure 2.10)

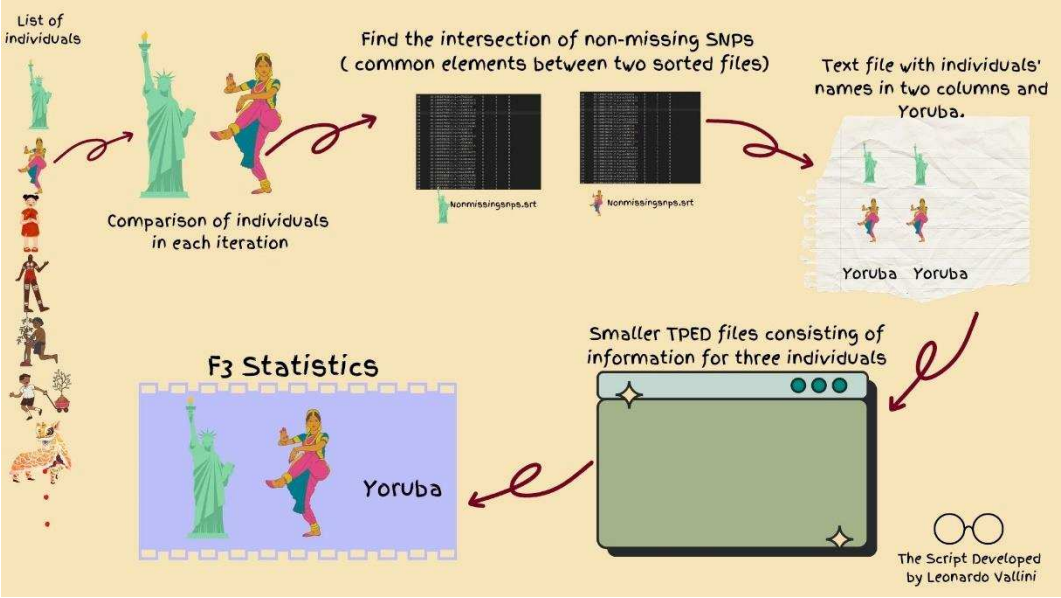

Figure 2. 10 Schematic Representation of the Script We Used to Perform the F3 Statistic

```
#!/bin/bash
# Read the file once and store its contents in an array
mapfile -t indiv < onlyAind.txt

# Loop over indiv array
for ((i=0; i<${#indiv[@]}; i++)); do
  for ((j=i+1; j<${#indiv[@]}; j++)); do
```

```
    # Compare individual i to each individual in
complist
    echo "Comparison of individual ${indiv[$i]} to
individual ${indiv[$j]}"

    intersection=$(comm -12 --nocheck-order
"${indiv[$i]}.nonmissingsnps.srt"
"${indiv[$j]}.nonmissingsnps.srt")

    printf "%s\t%s\n%s\t%s\n%s\t%s\n" "${indiv[$i]}"
"${indiv[$i]}" "${indiv[$j]}" "${indiv[$j]}" "NA18486"
"NA18486_A" "NA18486" "NA18486_B" >
"${indiv[$i]}.${indiv[$j]}.txt"

    plink2 --bfile "merge_all" --keep
"${indiv[$i]}.${indiv[$j]}.txt" --extract <(echo
"$intersection") --out "${indiv[$i]}.${indiv[$j]}" --
export tped --silent

    python /home/tools/scripts/popstats.py --file "
${indiv[$i]}.${indiv[$j]}" --pops
"${indiv[$i]},${indiv[$j]},NA18486" --f3 >> "f3.txt"

    # Clean up
    rm "${indiv[$i]}.${indiv[$j]}.txt"
    rm "${indiv[$i]}.${indiv[$j]}".*
  done
done
```

2.2.7 Running F4

The D-statistics also known as the ABBA-BABA test. The idea behind it is to consider ancestral (A) and derived (B) alleles on the genome of four individuals/ populations. Two allelic patterns of ABBA and BABA should happen equally when there is no introgression. On the other hand, when there is a gene flow between two individuals/ populations we will see an excess of either ABBA or BABA that will result in a D- statistic significantly different from zero. A positive D-statistic indicates introgression between individual/Population A and individual/Population C (i.e an excess of BABA), while a negative D-statistic shows an introgression between individual/Population B and individual/Population C (i.e an excess of ABBA). To evaluate the significance of the D-statistic, one can compute a Z-score. A Z-score bigger than 3 or smaller than -3 can be interpreted as a noteworthy outcome. (Figure 2.11)

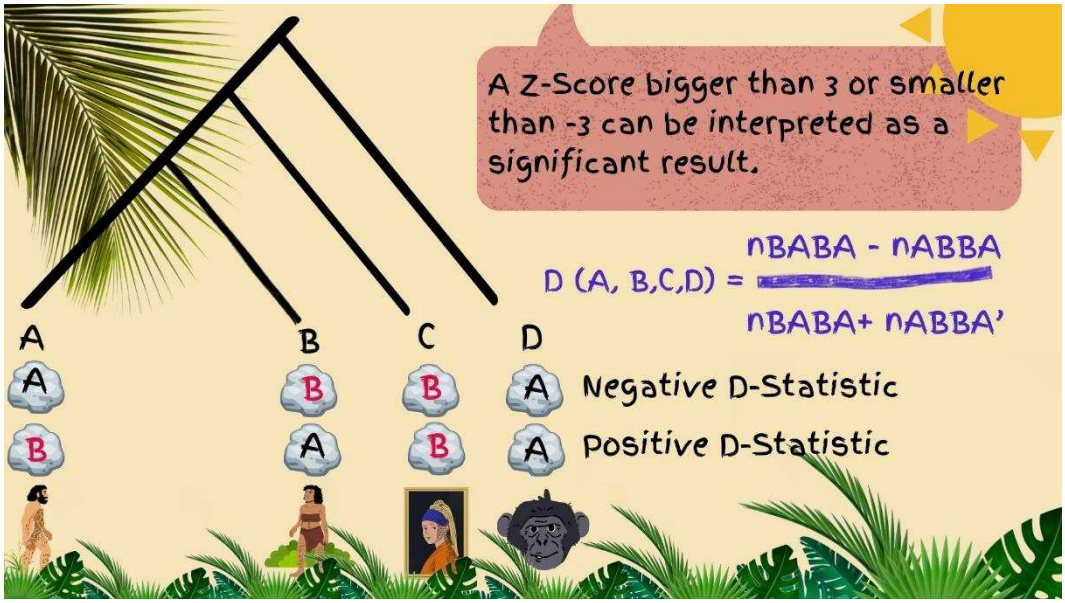

Figure 2. 11 Schematic representation of D-Statistics

The F4-Statistics calculates the shared genetic drift between test Population/Individual and Population/Individual A and Population/Individual B. Tree topologies are balanced at zero when there are no recent interactions between these four groups. Any deviation from zero indicates a deviation from the proposed tree. A negative value of Z-Score in F4 statistics indicates that population/ individual B is closer to C (the test population) while a positive value of Z-Score indicates closer genetic similarities between A and C. Similar to F3 one can use popstats.py to run it. (Figure 2.12)

```
python /home/tools/scripts/popstats.py --file merge_all
--pops
AltaiNeanderthal,Denisovan,testPopulation(X),Ancestral
--f4 >> f4.txt
```

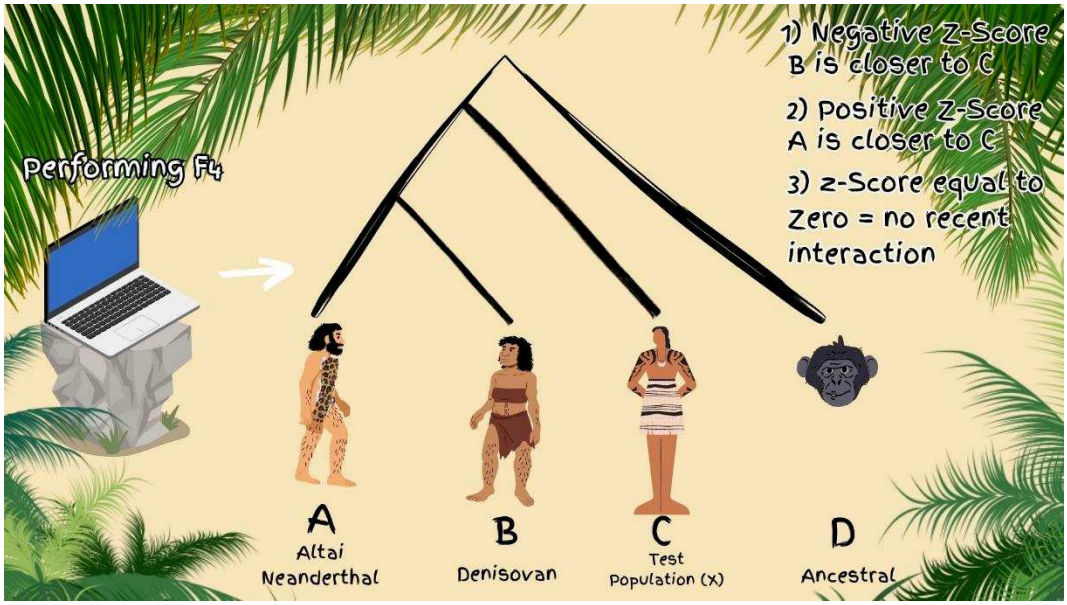

Figure 2. 12 Schematic Representation of the F4 Test

2.2.8 Similarity Matrix, Multidimensional Scaling, Plotting on R

Similarity Matrix

A similarity matrix is useful when we want to gain understanding on how similar two different datasets to one another are. We used the R code below to create a similarity matrix of pairwise similarities between individuals based on the F3 values in the input dataset:

```
library(spaa)

#f3.txt as an input
f3 <- read.delim("C:/Users/f3.txt", header = FALSE)
df <- subset(f3, select = -c(V3, V4, V6, V7, V8, V9,
V10, V11, V12, V13))
names(df) <- c("ind1", "ind2", "f3")

#normalized the f3
max_value <- max(df$f3, na.rm = TRUE)
min_value <- min(df$f3, na.rm = TRUE)
df$f3_normalized <- (df$f3 - min_value) / (max_value -
min_value)
df$f3 = NULL

#creating a similarity matrix
mat= as.matrix(list2dist(df))
```

```
#converting similarity matrix to distance matrix
D <- 1 - mat
```

MultiDimensional Scaling

Multidimensional scaling (MDS) is statistical method use to find the structure of similarity and dissimilarity data. It represents complicated, high-dimensional data in a lower-dimensional space with accuracy. It is preserving the pairwise distances or dissimilarities.

```
#MDS
mds1 <- cmdscale(D, k = 2)
mds = as.data.frame(mds1)
names(mds) = c("Dim1", "Dim2")
mds$ind = rownames(mds)
rownames(mds) = NULL
```

R

R is a free software that is used for statistical analysis, data analysis and visualization. We can visualize the data through using packages such as ggplot2. It enables us to plot high-quality graphs, plots, and charts. (Figure 2.13)

```
library(ggplot2)

ggplot() + geom_point(a , mapping = aes(Dim1, Dim2),
color = "blue", alpha = 1, show.legend = TRUE)+
geom_text(a, mapping = aes(x=Dim1, y=Dim2 - 0.003,
label = pops), size =2)+
  labs(title = "MDS Configuration with Population
Names")
```

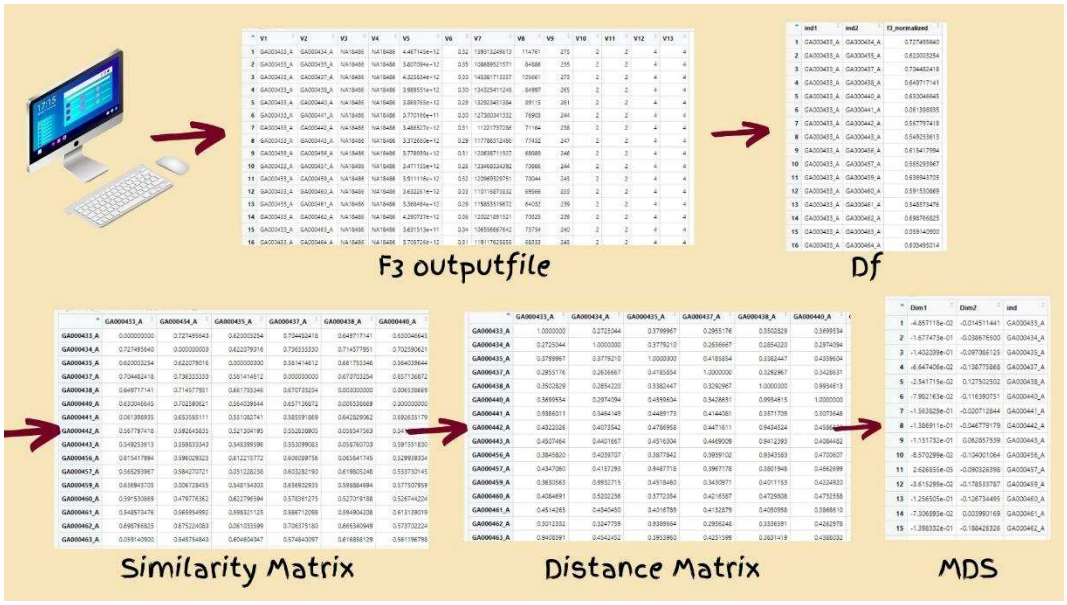

Figure 2. 13 Output Files in R

3 Results

3.1 F3

We ran f3 outgroup statistic in the form of f3-out(ind.i, ind.j, yoruba) for all possible pairwise comparison of individuals.

3.2 F4

We ran F4(Denisova, YRI, Test population, Ancestral) to gain understanding of sharing drifts between Test population and African and Denisova and test whether we correctly identified Denisovan introgressed segments in *Homo sapiens* genomes. All populations showed significantly positive f4 values, showing that the segments identified were indeed largely Denisovan-like. (Figure 3.1)

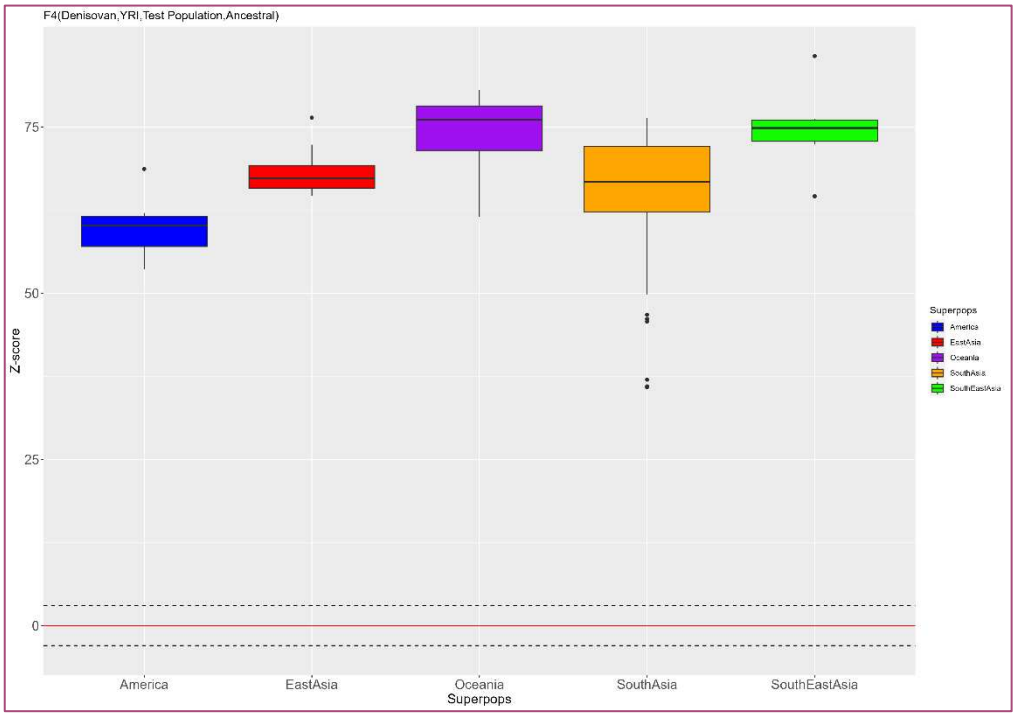

Figure 3. 1 Boxplot for F4(Denisova, Yoruba, Population X, Ancestral) Z-Score vs Superpops

The test we just performed would yield a positive value even if the segments identified by IBDmix were of Neanderthal origin, because of the shared drift between the two archaic species with respect to modern humans. To distinguish between the two we performed the test F4(Altai Neanderthal, Denisovan, Test Population, Ancestral). Our analysis indicated a consistently significant negative Z-Score across all populations, indicating that the majority of the identified segments were of Denisovan rather than Neanderthal origin. (Figure 3.2)

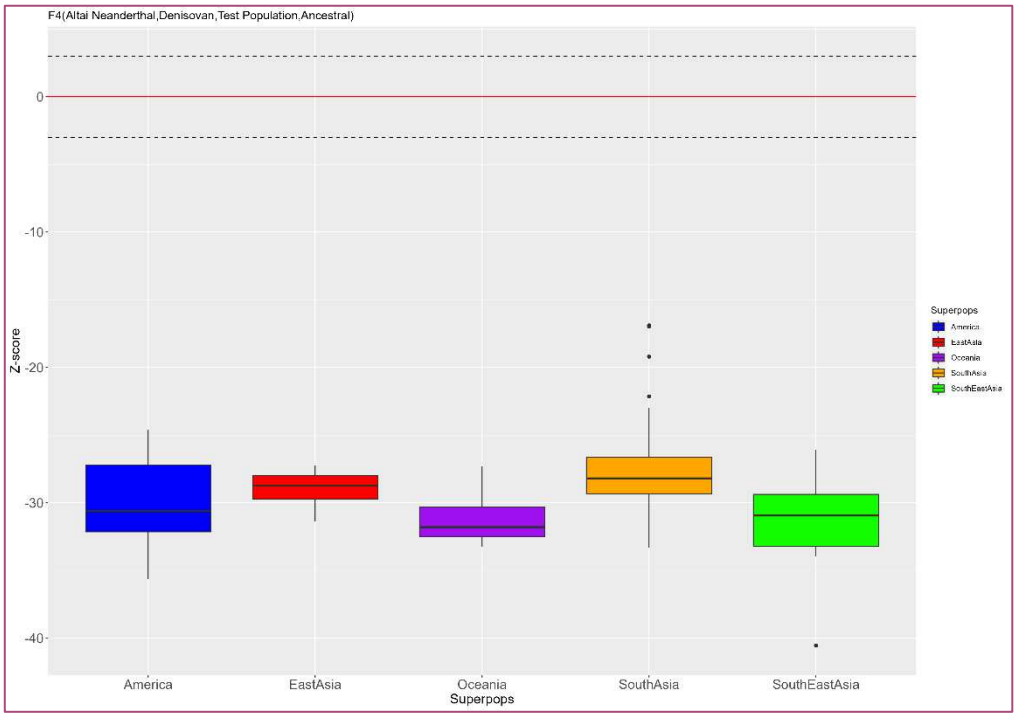

Figure 3. 2 Boxplot for F4(Altai Neanderthal, Denisova, Population X, Ancestral) Z-Score vs Superpops

We performed f4(Denisovan, Test Population, Altai Neanderthal, Ancestral) to gain deeper insight into accuracy of putative introgressed segments, assessing whether IBDmix incorrectly distinguishes Altai Neanderthal introgression as Denisovan. We observed ABBA allelic pattern as all the populations showed the negative Z-Score, showing all the population have an excess of allele sharing with Altai Neandetrhal. Among the populations, East Asians showed the highest level of negative Z-score and America showed the smallest. The Puerto Rican group exhibited a non-significant positive Z-score of 2.815, while the Colombian and Makrani groups showed non-significant negative Z-scores of -2.940 and -0.074, respectively. (Figure 3.3)

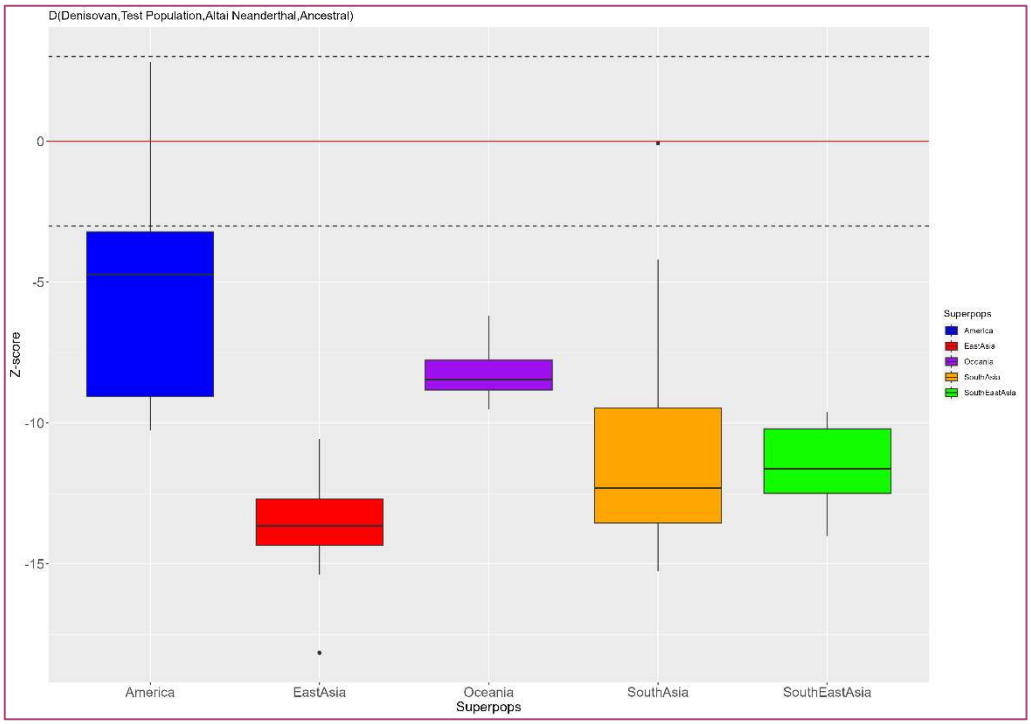

Figure 3. 3 Boxplot for F4(Denisova, Population X, Altai Neanderthal, Ancestral) Z-Score vs Superpops

Subsequently, we performed f4-Statistic(Han, Populationx, Yoruba (YRI), Ancestral) to check if IBDmix is incorrectly identifying some African ancestry among these populations as Denisovan. We have seen a non-significant positive value of Z-score for Oceania and SEA and non-significant negative value of Z-Score for America, East Asia and South Asia. however, Brahui in South Asia (Z-Score = -3.982) and Colombian in America (Z-Score = -4.267) revealed a significant value of Z-Score, indicating Brahui and Colombian introgressed denisovan-like segments have more excess of allele sharing with Africans compared to other populations, indicating these segments are not really denisovans,

but enriched for African ones. (Figure 3.4)

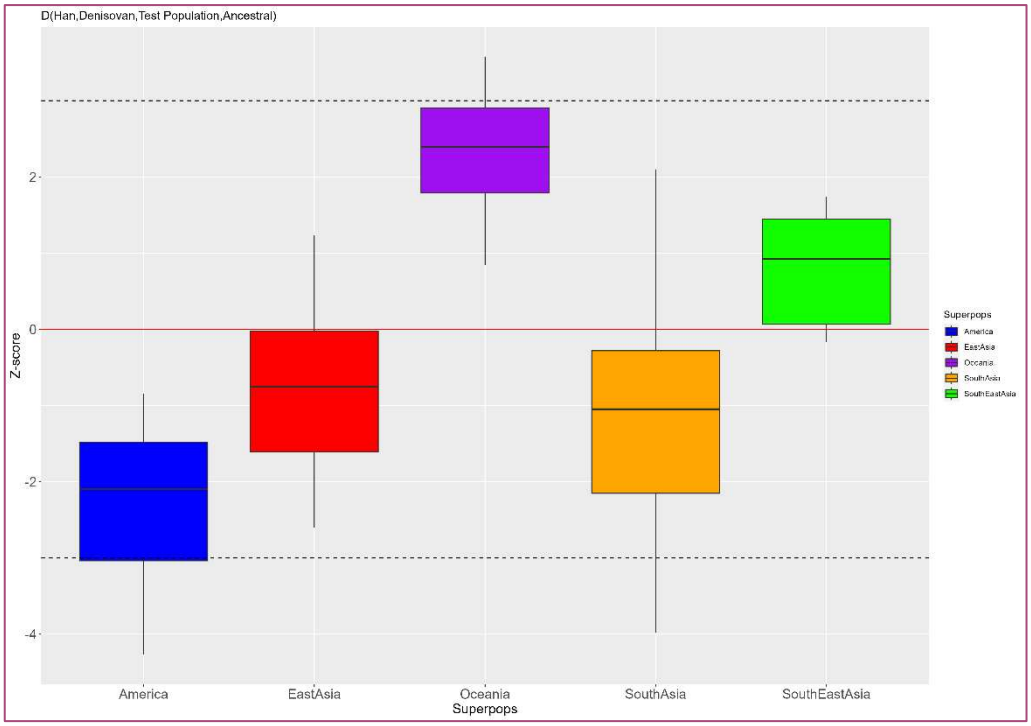

Figure 3. 4 Boxplot For D(Han, Population X, Yoruba, Ancestral) Z-Score vs Superpops

In our analysis of the scatter plot for Z-score D(Han, test population, YRI, Ancestral) against Z-score D(Denisova, Test Population, Altai Neanderthal, Ancestral) for Native Americans and Makrani, we observed a trend: the more putatively introgressed segments with Neanderthal nature, the more African ancestry they exhibited. (Figure 3.5)

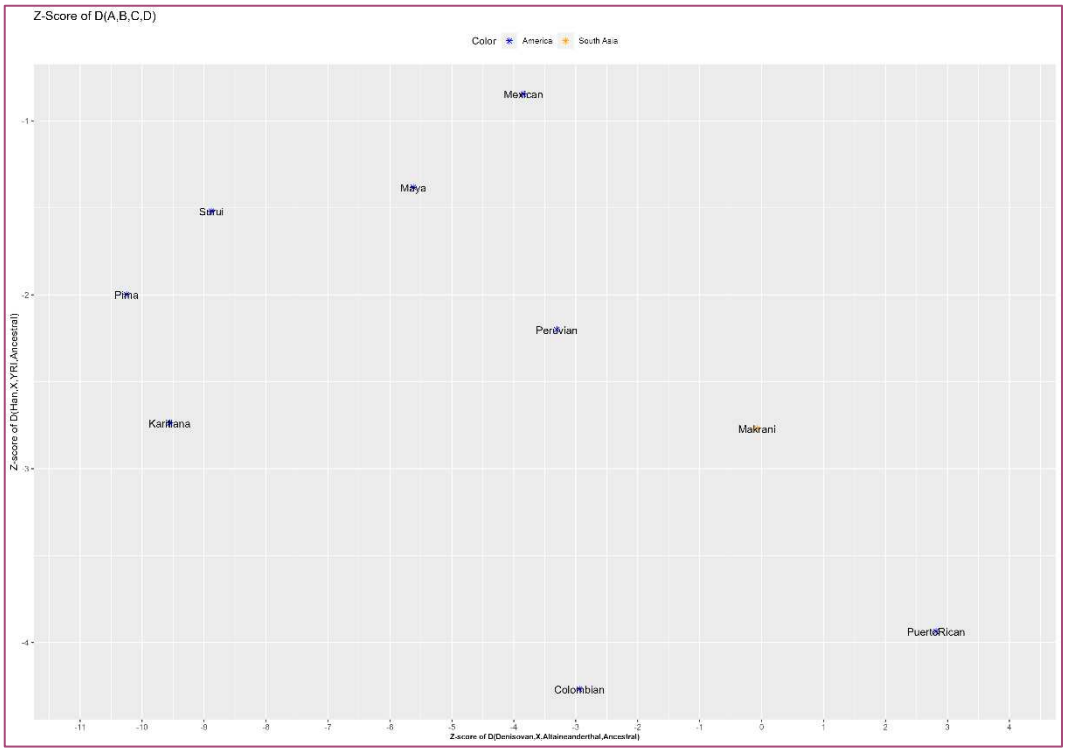

Figure 3. 5 Z-score of D(Han, Test Population, YRI, Ancestral) vs Z-score D(Denisova, Test Population, Altai Neanderthal,Ancestral)

3.3 MDS

To visualize the genetic affinities between the putatively introgressed denisovan segments in the study populations we computed a MDS on the similarity matrix we obtained.We plotted Americans, South Asians, East Asians, Southeast Asians and Oceanians in blue, orange, red, green and purple, respectively. When plotting the first dimension against the second dimension of MDS, our analysis reveals a closer genetic affinity among some Ati, Austronesian and Flores Cibal individuals in Southeast Asia and East Asians. We observed a closer genetic similarity among some Papuan Highlands, Papuan Sepik, Bougainville and Papuan individuals in Oceania and South Asians. Additionally, there is a genetic affinity between American and South Asian populations, as they cluster closer to one another. However, certain individuals from American populations, such as Maya, Peruvian and Pima exhibit genetic proximity to East Asians, as far as their Denisova genetic segments are concerned. (Figure 3.6)

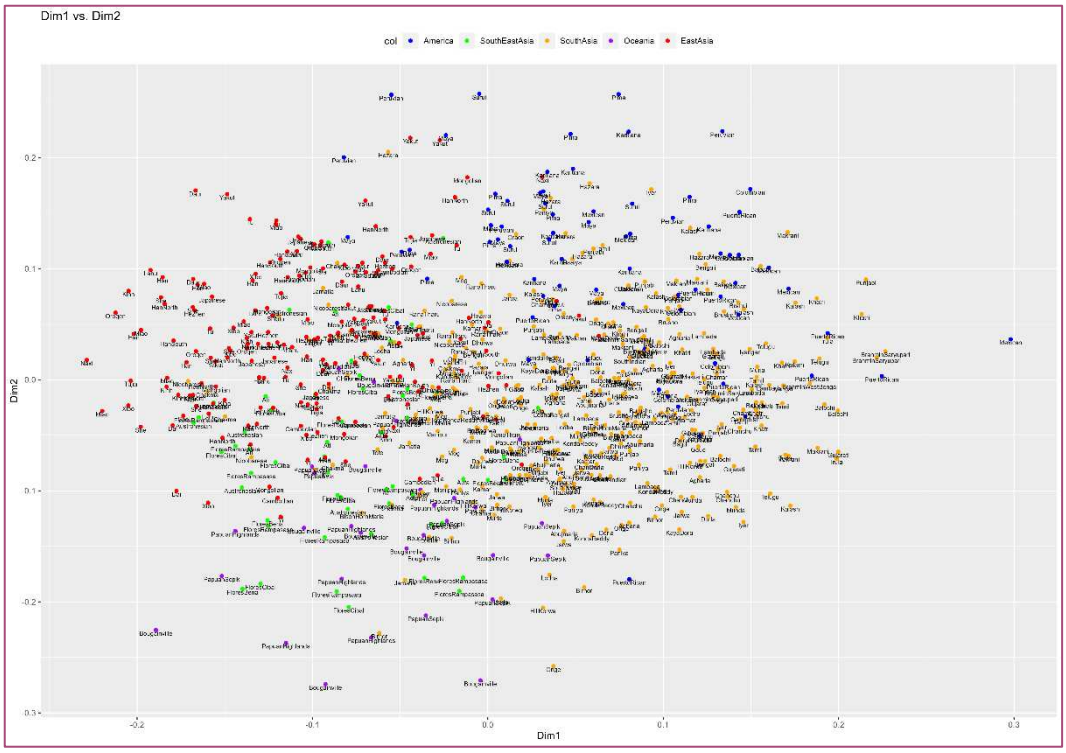

Figure 3. 6 MDS configuration of Dimension 1 against Dimension 2 for 744 individuals

When it comes to the MDS plot of dimension 1 against dimension 3 we observed that dim1 ranges from East Asians to South Asians with oceanian in the middle and Americans close to South Asians. Additionally, Flores Bena, Austronesian, Ati and Flores Rampasasa individuals in Southeast Asia were close to East Asians. However, Dai, Yi, Yakut and Mongolian individuals in East Asia, Austronesian and Aeta individuals in Southeast Asia, as well as Bougainville and Papuan Sepik individuals in Oceania were distant from South Asians. (Figure 3.7)

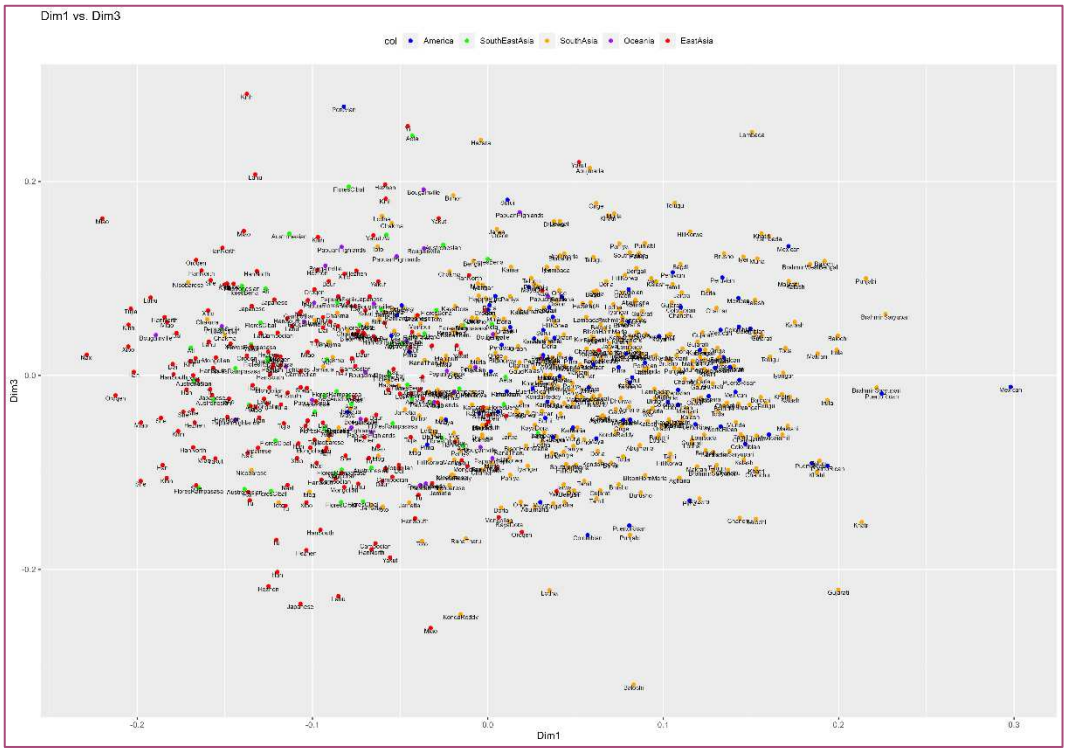

Figure 3. 7 MDS configuration of Dim1 against Dim3 for 744 individuals

For Dimension 1 against Dimension 4 we have observed that Americans are closer to South Asian. On the other hand, SEA and Oceanian are closer to East Asians. Additionally, we observed that Paniya, Chanchu, Tamil, and Konda Reddy individuals in South Asia, Japanese, Oroqen, Mongolian, Lahu, Kinh, Tujia and Naxi individuals in East Asia, as well as a Papuan Highlands individual and a Mexican, are distant from other individuals in their respective regions of South Asia, East Asia, Oceania and America. (Figure 3.8)

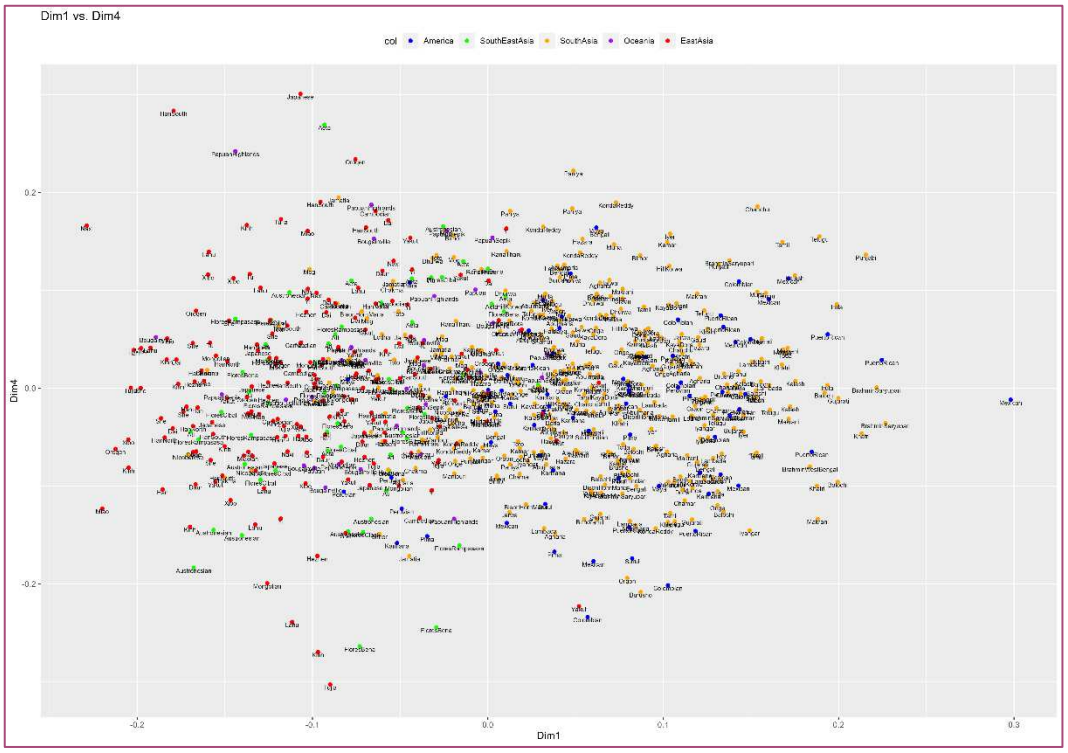

Figure 3. 8 MDS configuration of Dim1 against Dim4 for 744 individuals

Our observation for East Asians, South Asians, Southeast Asians and Americans in Dimension 2 against Dimension 3 showed a closer cluster among these individuals. However, individuals from Bougainville and Papuan Highlands in Oceania, Surui, Pima, Peruvian and Karitiana in America, as well as Hazara, Birhor, Hilkonwa, Onge and Dorla in South Asia, were distant compared to others.(Figure3.9)

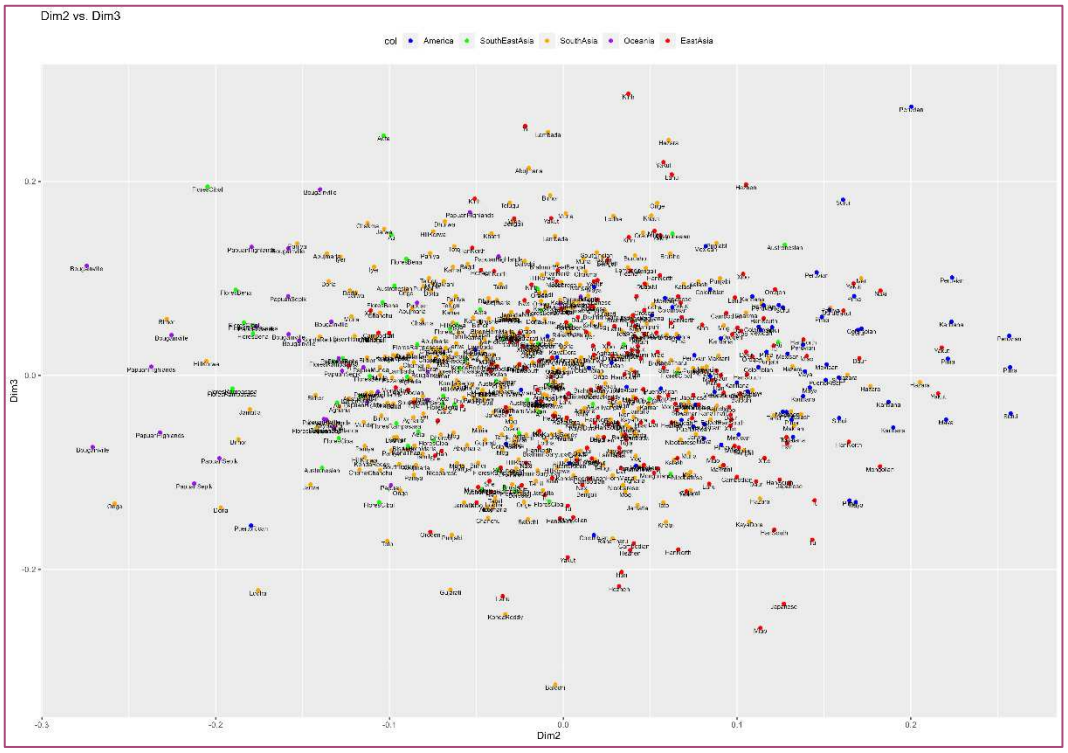

Figure 3. 9 MDS configuration of Dim2 against Dim3 for 744 individuals

When it comes to the MDS configuration of Dimension 2 against Dimension 4 we have observed that Oceanians and Southeast Asians close to South Asians and Americans demonstrate genetic affinity to East Asians. However, individuals from Bougainville, Papuan Sepik, Surui and Pima were distant compared to others. (Figure 3.10)

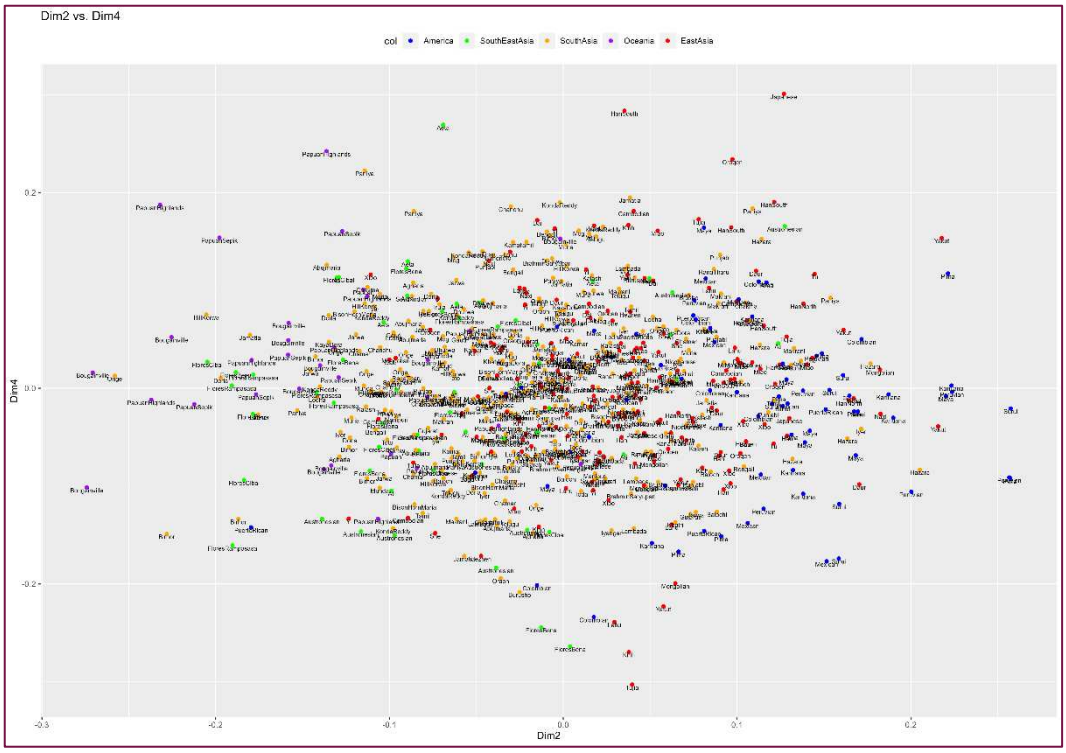

Figure 3. 10 MDS configuration of Dim2 Vs Dim4 for 744 individuals

We plotted the SNPs (putative introgressed archaic segments considered by IBDmix) of each individual against Dim1 of the MDS. Oceanians and 8 Aeta individuals in SEA, one Puerto Rican individual showed more putative introgressed Denisovan SNPs compared to other populations with a negative value for Dim1. Additionally, two Balochi, a Makarani, three Puerto Ricans and five Colombians demonstrate a higher level of introgressed segments compared to others. East Asians showed a negative value for Dim1 and lower level of introgressed segments compared to SEA and Oceanians, respectively. (Figure 3.11)

Figure 3. 11 MDS configuration of SNPs against Dimension 1 for 744 individuals

Subsequently, we plotted Dim1 against Dim2 of 744 individuals based on their Centroids which helps us to observe the general trend or central locations of the clusters and simplify the complex patterns. We highlighted the number of each individual for each population in parentheses (Figure 3.12). We have seen the Americans top, South Asians right, Oceania bottom, East Asia left and SEA between East Asia and Oceania. Our observations show that the Surui, Pima, Peruvian, Karitiana and Maya in America and Hazara in South Asia exhibit distinct genetic relationships compared to other populations. In contrast, Colombian and Puerto Rican populations exhibit genetic affinities closer to South Asians, such as the Balochi and Makrani. Nicobarese, Jmatia, Toto, Mog revealed genetic proximity to East Asians. Introgressed segments of Ati and Austroneasian in SEA were closer to Chakma in South Asia and they were all closer to the East Asian Cluster. Papuan Highlands, Papuan Sepik, Bougainville, Flores Bena, Flores Cibal and Flores Rampasasa demonstrate a closer similarity to each other and form a

cluster more distinct from other populations.

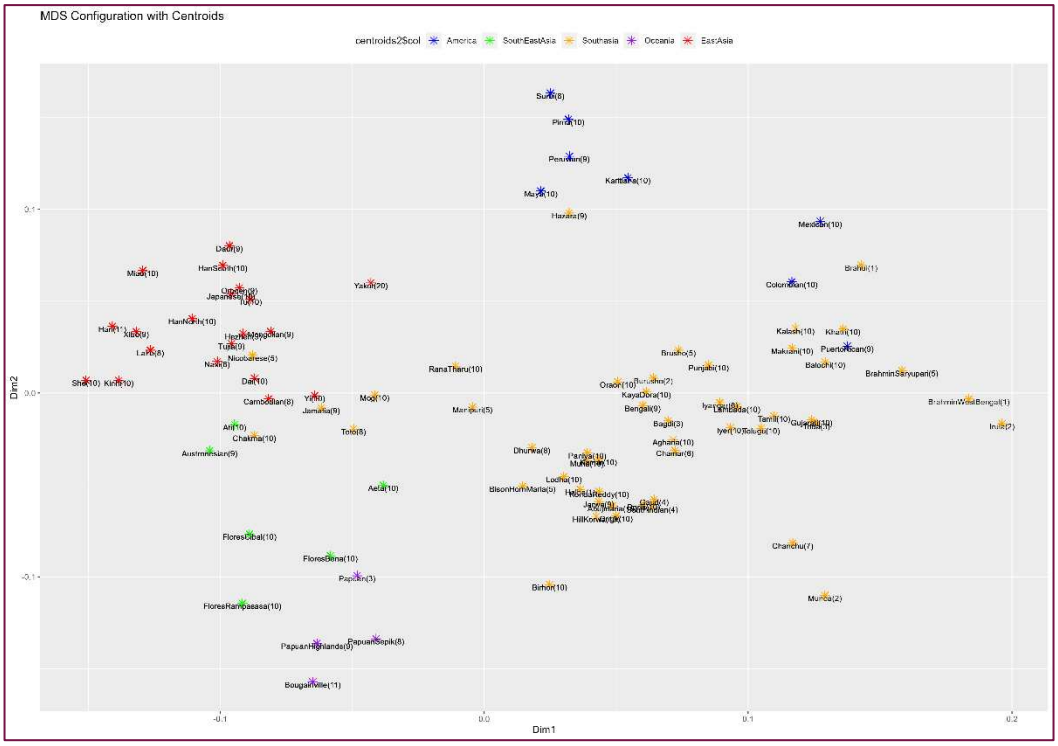

Figure 3. 12 MDS Configuration with Centroids for 88 populations in America, SouthEast Asia, South Asia, Oceania and East Asia.

Future perspectives

IBDmix comes with various limitations. Notably, it needs an archaic reference genome, making it unsuitable for identifying introgressed sequences from hominin lineages that are unknown or unsequenced. Additionally, IBDmix mandates the separate analysis of populations and the use of a sufficiently large sample size to robustly estimate population allele frequencies, assign LOD scores, and determine IBD. When we performed IBDmix for geographically diverse populations, IBDmix revealed unexpected findings. We observed a stronger than anticipated signal of Denisovan ancestry among South Asians and some Native Americans. This was particularly noteworthy in populations where based on previous studies we were confident about the higher levels of African and Neanderthal ancestry. Additionally, to generate a genotype file, IBDmix offers an option to choose one archaic human and one modern human. Later, it detects putative introgressed archaic segments from the genotype file. It would be better if we could specify more than one archaic human. In the later steps, considering specific thresholds for each archaic would allow different archaic segments, such as Denisovan and Neanderthal, to be distinguishable without repeating the entire process separately for each archaic.

Discussion

In this study, we applied IBDmix, a reference-free method for detecting putative introgressed archaic segments, in 744 individuals across East Asia, South Asia, America, Oceania, and SEA, aiming to identify putative Denisovan ancestry. Our analysis of the f3-outgroup statistic revealed that all individuals exhibit a greater degree of genetic similarity with each other than with Yoruba. We demonstrated that the studied populations possess both putative introgressed Denisovan and Neanderthal ancestry, as we observed an ABBA allelic pattern for D(Denisova, Test Population, Altai Neanderthal, Ancestral). However, Makrani, Puerto Rican and Colombian were the only populations that did not show a significant Z-score. Additionally, our analysis did not show significant shared ancestry with Africans, except for Colombians, Papuan Sepik, and Brahui.

To check the putative Neanderthal and African signals among Native Americans and Makrani, we plotted the Z-score of D(Denisova, test population, Altai Neanderthal, Ancestral) against D(Han, Test Population, YRI, Ancestral). We observed a trend in the Makrani population and Americans, where more putatively introgressed segments with Neanderthal nature correspond to a higher level of African ancestry. Therefore, we cannot confirm that Makrani, Puerto Rican and

Colombian populations lack putative introgressed Neanderthal segments, as they did not show a significant Z-Score for D(Denisovan, Test population, Altai Neanderthal, Ancestral). This means that these putative introgressed Denisovan segments are not real Denisovan, but they have more African nature.

Among the populations, Oceanians show the highest level of shared drift with Denisovan and a lower level of allele sharing with Altai Neanderthal. In contrast, East and South Asians exhibit the smallest shared drift with Denisovan and the largest level of allele sharing with Altai Neanderthal; these findings confirm those of previous studies [10].

5 Reference

1. Campbell, C. D., Chong, J. X., Malig, M., Ko, A., Dumont, B. L., Han, L., Vives, L., O'Roak, B. J., Sudmant, P. H., Shendure, J., Abney, M., Ober, C., & Eichler, E. E. (2012). Estimating the human mutation rate using autozygosity in a founder population. *Nature Genetics 2012 44:11*, *44*(11), 1277–1281. https://doi.org/10.1038/ng.2418
2. Cann, R. L., Stoneking, M., & Wilson, A. C. (1987). Mitochondrial DNA and human evolution. *Nature 1987 325:6099*, *325*(6099), 31–36. https://doi.org/10.1038/325031a0
3. Fu, Q., Mittnik, A., Johnson, P. L. F., Bos, K., Lari, M., Bollongino, R., Sun, C., Giemsch, L., Schmitz, R., Burger, J., Ronchitelli, A. M., Martini, F., Cremonesi, R. G., Svoboda, J., Bauer, P., Caramelli, D., Castellano, S., Reich, D., Pääbo, S., & Krause, J. (2013). A Revised Timescale for Human Evolution Based on Ancient Mitochondrial Genomes. *Current Biology*, *23*(7), 553–559. https://doi.org/10.1016/J.CUB.2013.02.044
4. Higham, T., Jacobi, R., Julien, M., David, F., Basell, L., Wood, R., Davies, W., & Ramsey, C. B. (2010). Chronology of the Grotte du Renne (France) and implications for the context of ornaments and human remains within the Châtelperronian. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(47), 20234–20239. https://doi.org/10.1073/PNAS.1007963107/-/DCSUPPLEMENTAL/PNAS.201007963SI.PDF
5. Grün, R., Stringer, C., McDermott, F., Nathan, R., Porat, N., Robertson, S., Taylor, L., Mortimer, G., Eggins, S., & McCulloch, M. (2005). U-series and ESR analyses of bones and teeth relating to the human burials from Skhul. *Journal of Human Evolution*, *49*(3), 316–334. https://doi.org/10.1016/J.JHEVOL.2005.04.006
6. Valladas, H., Joron, J. L., Valladas, G., Arensburg, B., Bar-Yosef, O., Belfer-Cohen, A., Goldberg, P., Laville, H., Meignen, L., Rak, Y., Tchernov, E., Tillier, A. M., & Vandermeersch, B. (1987). Thermoluminescence dates for the Neanderthal burial site at Kebara in Israel. *Nature 1987 330:6144*, *330*(6144), 159–160. https://doi.org/10.1038/330159a0
7. Meyer, M., Kircher, M., Gansauge, M. T., Li, H., Racimo, F., Mallick, S., Schraiber, J. G., Jay, F., Prüfer, K., de Filippo, C., Sudmant, P. H., Alkan, C., Fu, Q., Do, R., Rohland, N., Tandon, A., Siebauer, M., Green, R. E., Bryc, K., … Pääbo, S. (2012). A high-coverage genome sequence from an archaic Denisovan individual. *Science (New York, N.Y.)*, *338*(6104), 222–226. https://doi.org/10.1126/SCIENCE.1224344
8. Krause, J., Fu, Q., Good, J. M., Viola, B., Shunkov, M. v., Derevianko, A. P., & Pääbo, S. (2010). The complete mitochondrial DNA genome of an

unknown hominin from southern Siberia. *Nature*, *464*(7290), 894–897. https://doi.org/10.1038/NATURE08976

9. Green, R. E., Malaspinas, A. S., Krause, J., Briggs, A. W., Johnson, P. L. F., Uhler, C., Meyer, M., Good, J. M., Maricic, T., Stenzel, U., Prüfer, K., Siebauer, M., Burbano, H. A., Ronan, M., Rothberg, J. M., Egholm, M., Rudan, P., Brajković, D., Kućan, Ž., … Pääbo, S. (2008). A complete Neandertal mitochondrial genome sequence determined by high-throughput sequencing. *Cell*, *134*(3), 416–426. https://doi.org/10.1016/J.CELL.2008.06.021
10. Peyrégne, S., Slon, V., & Kelso, J. (2023). More than a decade of genetic research on the Denisovans. Nature Reviews Genetics 2023, 1–21. https://doi.org/10.1038/s41576-023-00643-4
11. Reich, D., Green, R. E., Kircher, M., Krause, J., Patterson, N., Durand, E. Y., Viola, B., Briggs, A. W., Stenzel, U., Johnson, P. L. F., Maricic, T., Good, J. M., Marques-Bonet, T., Alkan, C., Fu, Q., Mallick, S., Li, H., Meyer, M., Eichler, E. E., … Pääbo, S. (2010). Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*, *468*(7327), 1053–1060. https://doi.org/10.1038/NATURE09710
12. 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., Saitou, N., Wollstein, A., Kayser, M., Pääbo, S., & Stoneking, M. (2011). Denisova admixture and the first modern human dispersals into Southeast Asia and Oceania. *American Journal of Human Genetics*, *89*(4), 516–528. https://doi.org/10.1016/j.ajhg.2011.09.005
13. Peter, B. M. (2020). 100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains. *BioRxiv*. https://doi.org/10.1101/2020.03.13.990523
14. Qin, P., & Stoneking, M. (n.d.). *Denisovan Ancestry in East Eurasian and Native American Populations*. https://doi.org/10.1093/molbev/msv141
15. Prüfer, K., de Filippo, C., Grote, S., Mafessoni, F., Korlević, P., Hajdinjak, M., Vernot, B., Skov, L., Hsieh, P., Peyrégne, S., Reher, D., Hopfe, C., Nagel, S., Maricic, T., Fu, Q., Theunert, C., Rogers, R., Skoglund, P., Chintalapati, M., … Pääbo, S. (2017). A high-coverage Neandertal genome from Vindija Cave in Croatia. *Science (New York, N.Y.)*, *358*(6363), 655–658. https://doi.org/10.1126/SCIENCE.AAO1887
16. Larena, M., McKenna, J., Sanchez-Quinto, F., Bernhardsson, C., Ebeo, C., Reyes, R., Casel, O., Huang, J. Y., Hagada, K. P., Guilay, D., Reyes, J., Allian, F. P., Mori, V., Azarcon, L. S., Manera, A., Terando, C., Jamero, L., Sireg, G., Manginsay-Tremedal, R., … Jakobsson, M. (2021). Philippine Ayta possess the highest level of Denisovan ancestry in the world. *Current Biology*, *31*(19), 4219-4230.e10. https://doi.org/10.1016/J.CUB.2021.07.022/ATTACHMENT/54FDACEB-B806-4FFD-A107-F628136DC7EC/MMC3.XLSX

17. Browning, S. R., Browning, B. L., Zhou, Y., Tucci, S., & Akey, J. M. (2018). Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*, *173*(1), 53-61.e9. https://doi.org/10.1016/j.cell.2018.02.031
18. International HapMap Consortium, 2007 ,http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/
19. https://github.com/pontussk/popstats#popstats
20. https://www.cog-genomics.org/plink/2.0/