AARHUS
UNIVERSITY
NATURAL SCIENCES

Master's Thesis

Specifying the meetings between anatomically modern humans, Neanderthals, and Denisovans

Bioinformatics Research Centre
15 July 2023

Cæcilia Lind Skov-Jensen
201806070

Supervisors:
Mikkel Heide Schierup
Moisès Coll Macià

Abstract

About 60,000 years ago, our ancestors, the anatomically modern humans, migrated out of Africa and into Eurasia. Here they met other hominin groups, such as Neanderthals and Denisovans, that they interbred with. These meetings left a genetic mark in the form of introgressed archaic DNA into the genomes of people today. All present-day non-African populations have around 2% introgressed Neanderthal DNA scattered in smaller fragments in their genomes. The proportion of Denisova ancestry is a bit more complex, as the distribution of Denisova genetic signals is highly variable. The highest proportion of Denisovan ancestry is found in Oceanian populations (3-6%), while mainland Eurasian and Native American populations have under 1% of Denisovan content. Although much is already known about anatomically modern humans and their meetings with Neanderthals and Denisovans, many things remain unclear. Usually, archaic content is compared to sequenced archaic genomes. However, in this study, archaic content will be compared among present-day human populations by creating artificial archaic genomes for each population to compare the introgressing population's genomes more specifically. This will be done by comparing the nucleotide divergence patterns in the archaic content between populations. The Neanderthal divergence measurements supported a single introgression event from one Neanderthal population that happened before the split of non-African populations. On the contrary, the Denisova divergence measurements suggested multiple introgressions from at least two divergent Denisova population sources.

Future work could involve other strategies for sampling archaic content, such as downsampling to have the same number of individuals in all samples, sampling the longest non-overlapping fragments to get more information, or sampling all non-overlapping intervals between fragments to get as much information as possible.

Acknowledgments

First, I would like to thank Prof. Mikkel Heide Schierup and postdoc Moisès Coll Macià for their contribution to this project. I could not have finished the project without their guidance and support throughout the whole process.

Doing a Master's thesis during a severe depression is one of the hardest things I have ever had to do, and I will forever be thankful for their understanding and occasional motivational speeches whenever I needed them.

I would also like to say a special thank you to my partner, who has been my rock throughout my whole degree. You and our future together have been my biggest motivation when things were difficult. Now we can finally begin our real adult lives, and I cannot wait to see what is in store for us.

Table of Content

Introduction

Since the first discovery of Neanderthals in 1864[1], it has been of great interest to uncover how they lived, why they disappeared some 30,000 years ago, and how the anatomically modern humans, our ancestors, interacted with them[2]. In 2010, the first Neanderthal genome was sequenced based on 21 Neandertal bones from the Vindija Cave in Croatia[3]. A few months later, a finger bone found in the Denisova Cave located in the Altai Mountains in southern Siberia was sequenced[4]. This sequencing showed that the bone came from another group of archaic humans related to Neanderthals that got the name Denisovans. Thus, at least two distinct human groups, Neandertals and the related Denisovans, inhabited Eurasia when anatomically modern humans emerged from Africa. Multiple other remains have been found and later sequenced. Thus, Neanderthals are believed to have lived in Europe and Western Asia until around 30,000 to 40,000 years ago[5], while Denisovans are believed to have lived in Asia, particularly in Siberia, until about 30,000 to 50,000 years ago[6].

Approximately 60,000 years ago, anatomically modern humans migrated from Africa into Eurasia. Here they met and interbred with Neanderthals, which left a genetic mark of around 2% introgressed Neanderthal DNA scattered in small fragments in the genomes of all non-African populations today[3]. This genetic mark is consistent with the fact that anatomically modern humans interbred with Neanderthals before the split of the non-African populations[7]. However, studies have shown that present-day West Eurasian populations have a lower relative proportion of Neanderthal ancestry than East Eurasian populations. In contrast, populations in South and Central Asia have intermediary levels[8]. This discrepancy has been suggested to be a result of additional admixture events private to East Eurasians or due to differences in selection among these groups, with West Eurasians purging Neanderthal variants more efficiently. Still, most likely, the discrepancy results from a dilution of the Neanderthal ancestry in West Eurasians due to a subsequent admixture event with a group that carried much less or no Neanderthal ancestry[8].

However, the patterns of Denisovan ancestry in anatomically modern humans suggest a more complex history of interactions than those of Neanderthals. The genetic signal distribution of Denisovan DNA is highly variable across modern populations. The highest proportion of Denisovan ancestry is found in Oceanian populations (3-6%) depending on the amount of Papuan-related ancestry, thus maximizing in New Guinea Papuan Highlanders. On the contrary, mainland Eurasian and Native American populations have <1% of Denisovan content[7]. Comparing the Denisova content found in modern humans to the sequenced Denisova genome suggested that at least two "Denisovan-like" populations introgressed into modern humans, one closely related to the Altai Denisova and the other much more divergent[9]. While South Asian and Australo-Papuan populations exclusively carry signals from the "distantly related" Denisovan group, East Asian populations seem to consist of both "distantly related" and "closely related" Denisovan ancestry components[7,9]. The two Denisovan groups contributed with different proportions to the Asian and Oceanian populations. One-third of the Denisovan components in East Asia are derived from the Denisovan group genetically similar to the Altai Denisovan. In contrast, two-thirds are derived from the distantly related Denisovan group. On the contrary, almost all Denisovan components in Oceanian and South Asian populations come from the distantly related Denisovan group.

The extent of the admixture events with the Denisovan groups leading to the signals of Denisovan ancestry is still very speculative, and there are different hypotheses.

One possible explanation is an admixture event common to all Asian and Australo-Papuan populations, which resulted in a shared genomic signal of 2.6% to 3.4%, and then an additional admixture event private to the Australo-Papuan populations, generating the additional 1.6% genomic contribution[7].

Although nowadays we know more about the meetings between modern and archaic humans, many insights about these encounters still need to be clarified, such as the exact migration patterns, as well as the number and timing of the admixture events. Usually, archaic content is compared to sequenced archaic genomes. However, in this study, archaic content will be compared among present-day human populations to compare the genomes of the introgressing population more specifically. This will be done by using the divergence patterns in the archaic content and comparing it to the divergence in the whole genome. In this way, it will be possible to see whether the archaic components that in previous studies are thought to be the same are actually stemming from the same population or even the same event. The best approximation for this would be to compare all individuals pairwise. However, this would be computationally costly, and as each individual only has little archaic content, very little of the archaic component would be compared. Therefore, artificial genomes are made population-wise, making it possible to compare populations. If the theory about one major admixture event with Neanderthals common to all non-African populations is true, it would be expected that the divergence pattern in the Neanderthal content would mimic that for the whole genome. In the same way as the divergence pattern in Denisovan content, based on the migration and admixture theory, it would be expected to observe a high divergence between Australo-Papuan populations and the rest of the world.

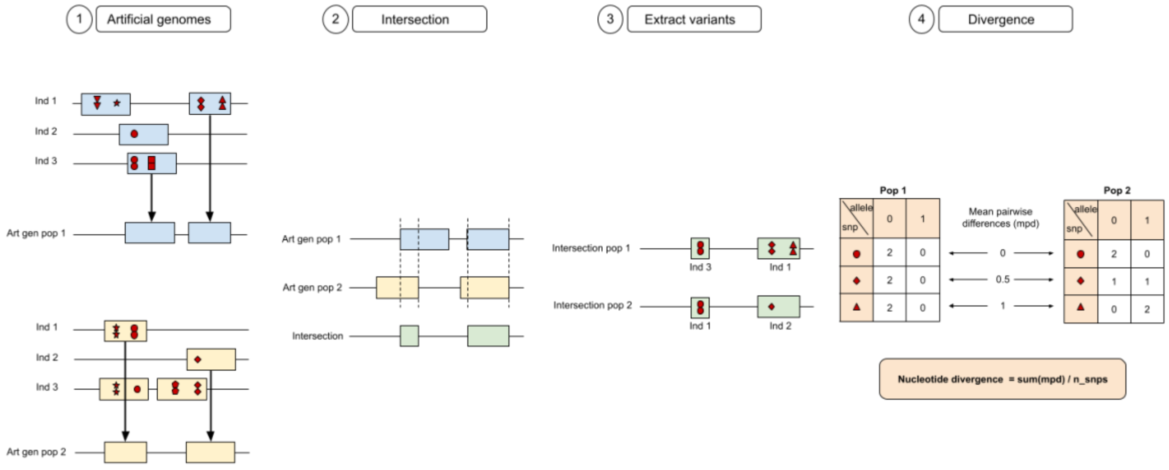
Bergström et al. (2020)[10] tried this pairwise population comparison on nucleotide divergence. However, they did not go further with the analysis than presenting the patterns.

Methods

Dataset and archaic fragments

This project builds upon ongoing research conducted by postdoc Moisès Coll Macià and Prof. Mikkel Heide Schierup, both supervising this master thesis. In their project, they have called archaic fragments using a Hidden Markov Model independent of archaic reference sequences[11] on 825 individuals across 47 non-African populations of the HGDP dataset[10].

At the beginning of my project, I received bed-like files containing information on the regions where archaic fragments are found per individual. Each region was classified as Neanderthal (Altai, Vindija, or Chagyrskaya), Denisova, non-DAVC, or ambiguous, depending on the number of shared variants with the archaic genome sequences. Besides that, I was given access to the original Variant Calling Files (VCF) and the formatted Zarr files to access information on the polymorphisms, such as genotype information.



Pop 1

snp \ allele	0	1
●	2	0
◆	2	0
▲	2	0

Pop 2

snp \ allele	0	1
●	2	0
◆	1	1
▲	0	2

Figure 1: Workflow. **(1)** Artificial genomes. Individuals from each population were sampled, and their archaic fragments were used to create so-called artificial genomes. Non-overlapping fragments were randomly chosen for the artificial genome. Archaic fragments are shown as squares in blue (population 1) or yellow (population 2). The different red shapes represent the variation that is present in the fragments. **(2)** Intersection. The overlapping intervals between the two populations are found between the two artificial genomes. The intersection is represented in green. **(3)** Extract variants. The fragments and individuals that contributed to the intersection are found for each population, and the variants in this specific fragment are notated. These are then used to find the genotypes at each position in the fragment. Variants with two red shapes represent homozygosity for the derived allele (allele 0), one shape represents heterozygosity, and no shapes represent homozygosity for the ancestral allele (allele 1). **(4)** Divergence. The alleles are then counted for each population, and the mean pairwise differences (mpd) are calculated. The nucleotide divergence between the two populations is then calculated by summing the mpd and dividing by the number of compared variants.

Creating artificial archaic genomes

The archaic fragments from individuals of each population were then used to assemble so-called artificial genomes for each population, as shown in the first step of Figure 1. Depending on the number of shared variants the fragments had with the sequenced archaic genomes, two artificial genomes were created per population; One that explicitly contained Neanderthal-like fragments and one explicitly with Denisova-like fragments. All fragments were randomly sampled without replacement for each archaic ancestry to construct the artificial genome. Bedtools v2.30.0[12] was then used to see if fragments were overlapping. When a sampled fragment overlapped with a previously sampled fragment in terms of genomic coordinates, the newly sampled fragment would be discarded.

This process was repeated for all the fragments for each population, ending up with one Neanderthal artificial genome and one Denisova artificial genome per population. In this way, the artificial genomes represented one archaic individual from each population.

Finding the intersection

The intersection between the artificial genomes was found using bedtools v2.30.0[12] to compare the archaic content between populations. The archaic content is compared by looking at the fragments of the artificial genomes. By then pairwise comparing artificial genomes from different populations and screening for overlaps, it is possible to find the shared intervals (i.e., fragments) between the two populations. These shared intervals correspond to the intersection, as shown in the second step of Figure 1. The intersections were made pairwise for all populations for Denisova and Neanderthal artificial genomes separately.

Knowing the shared intervals between populations, the amount of intersection among artificial genomes was computed as the intersection/artificial genome ratio to analyze the amount of "archaic overlap" among populations. Two highly divergent populations are expected to share fewer archaic segments and therefore have a smaller archaic overlap. In contrast, two populations with lower divergence levels are expected to show a more significant amount of overlap.

Extracting variants

As shown in the third step of Figure 1, the variants within the intersected segments were then pulled for the individuals from whom the archaic fragments in the artificial genomes of each population came. This made it possible to obtain the genotype information for each variant in the intersected regions.

Computing the nucleotide divergence

When the genotypes for each variant in the intersected regions were found, the mean number of pairwise differences (mpd) between the two populations was calculated for each variant using scikit-allel v1.3.5[13]. The number of pairwise differences was calculated by subtracting the number of pairwise comparisons where there is no difference from each variant's total number of pairwise comparisons. The pairwise differences are then divided by the total number of pairwise comparisons for each variant to get the mpd:

$$mpd = \frac{n_{pairs} - n_{same}}{n_{pairs}}$$

The fourth step in Figure 1 shows the three potential cases when only two individuals are compared (here, two artificial genomes).

Once the mpd was calculated for all chromosomes between all pairs of populations, the nucleotide divergence, D_{xy}, could then be calculated between each pair of populations x and y as follows[14]:

$$D_{xy} = \frac{\pi_{xy}}{n_{snps}}$$

Where π_{xy} is the sum of mpd for each population, pair across all chromosomes, and n_{snps} is the total number of variants that are compared between each pair of populations x and y

For each pairwise comparison, three different nucleotide divergences were calculated: values calculated from only the Neanderthal segments in the genomes, from only the Denisova segments in the genomes, and from the entire genome. The total number of comparisons between populations when calculating the nucleotide divergence for either Neanderthal or Denisova segments corresponded to the total length of the intersection between the two populations. When calculating the divergence for the whole genome, the callable region of the whole genome of 2,252,286,208 base pairs was used as the total number of comparisons.

Pipeline setup

The results presented in the following section were obtained using the Python package gwf v1.8.5[15] to schedule jobs in a slurm backend. In total, 130,293,092 jobs were run to complete the following results.

The code is available on GitHub.

Results

Artificial genomes

Artificial genomes were obtained for the 47 non-African populations for the Neanderthal and Denisova component, respectively. Figure 2 shows summary statistics for the produced artificial genomes.

From the mean fragment length distribution, it is observed that the Oceanian populations, on average, have longer fragments in their artificial genomes for both the Neanderthal (Figure 2A) and the Denisova artificial genomes (Figure 2B). Furthermore, the difference in the mean fragment length across populations is higher for the Denisova artificial genomes than the Neanderthal artificial genomes. The plots showed more variation in the number of fragments in the artificial genomes and the total sequence. For Neanderthals, there seemed to be a pattern between the number of fragments (and thereby the total sequence) and the number of individuals, as the populations with fewer individuals had artificial genomes with fewer fragments and shorter total sequences (Figures 2C and 2E). The difference in the total sequence in the Neanderthal artificial genome could be explained by the number of individuals from which sampling was possible per population, as there seems to be a correlation between the number of individuals and the total sequence (Figure 3A).

For Denisovans, the pattern showed that the most Denisova fragments in the artificial genomes and the longest artificial genome were found in the Oceanian populations (Figures 2D and 2F), as expected since they are reported to have the most significant proportion of Denisova ancestry[7]. Furthermore, when comparing this region with the other regions included, the number of fragments and the total length of the artificial genome seemed to be very low in the other regions. From Figure 3B, it is clear that there is no correlation between the total sequence and the number of individuals, as the linear model between the two variables is consistent with no correlation.

Other unexplained variations in any of the statistics plots could be due to some populations having fewer archaic fragments in general, or they contain fragments in high frequency, resulting in many overlapping fragments across individuals.

In general, both the number of fragments and the total length of the artificial genome are higher in the Neanderthal artificial genomes compared to those in the Denisova artificial genomes. Besides Oceania, which seems to have approximately the same number of fragments and the same sequence length for both the Neanderthal and Denisova artificial genomes, the other regions generally have under 500 fragments in the Denisova artificial genomes compared to Neanderthal artificial genomes, where most contain above 2000 fragments. As for the total sequence, a similar pattern appears. Besides Oceania, the other regions have a total sequence of 50 million base pairs or below for the Denisova artificial genomes. In contrast, the total sequences of the Neanderthal artificial genomes are generally 200 million base pairs or above.

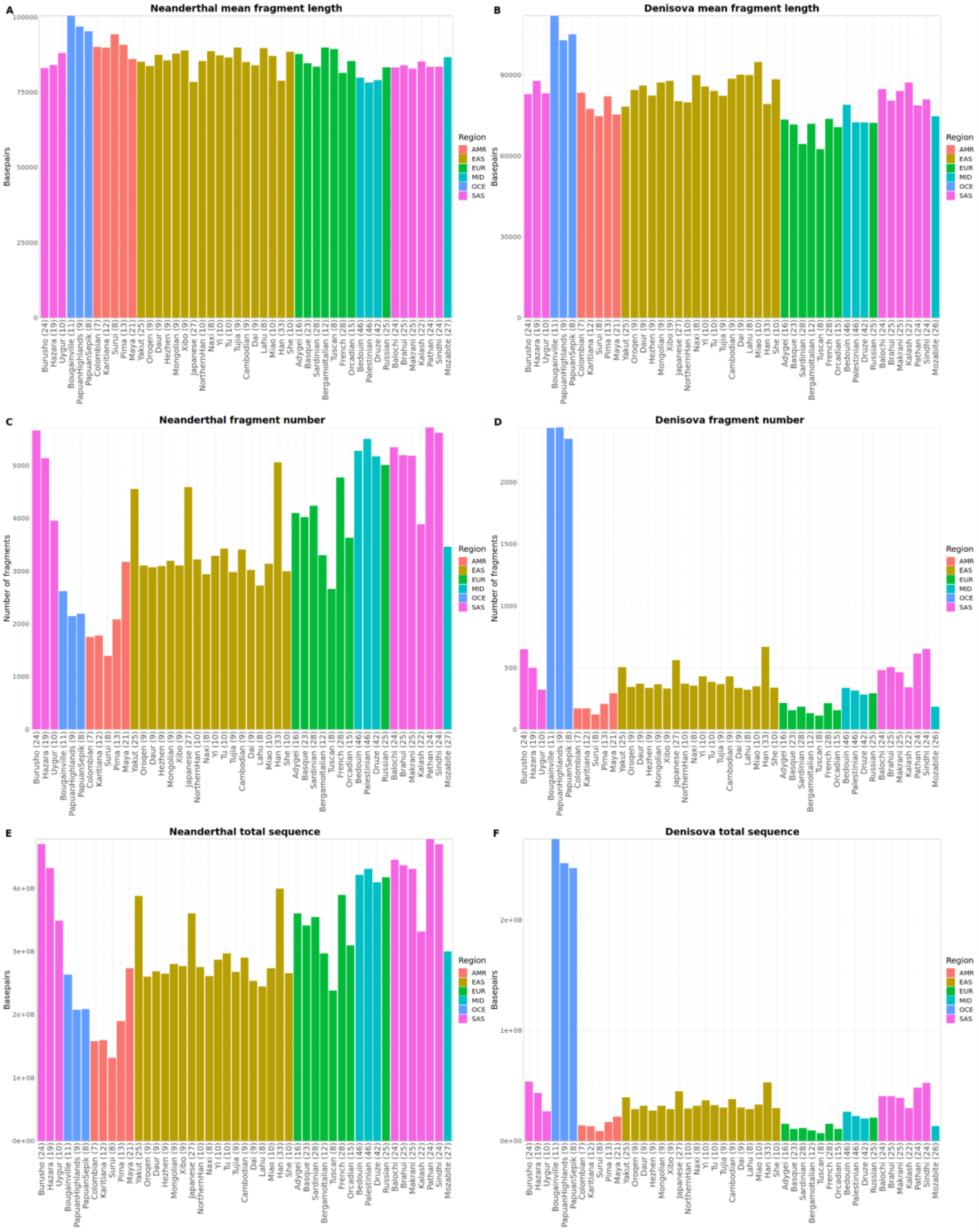


Figure 2: Artificial genome summary statistics for each of the 47 non-African populations. The populations are ordered according to the neighbor-joining tree built in Bergström et al. (2020)[10] using San as an outgroup. The populations are shown on the x-axis along with the number of individuals in parenthesis and are colored according to region. **(A)** Mean fragment length in the Neanderthal artificial genomes in base pairs. **(B)** Mean fragment length in the Denisovan artificial genomes in base pairs. **(C)** Number of fragments in the Neanderthal artificial genomes. **(D)** Number of fragments in the Denisovan artificial genomes in base pairs. **(E)** Total sequence for the Neanderthal artificial genomes in base pairs. **(F)** Total sequence for the Denisovan artificial genomes in base pairs.

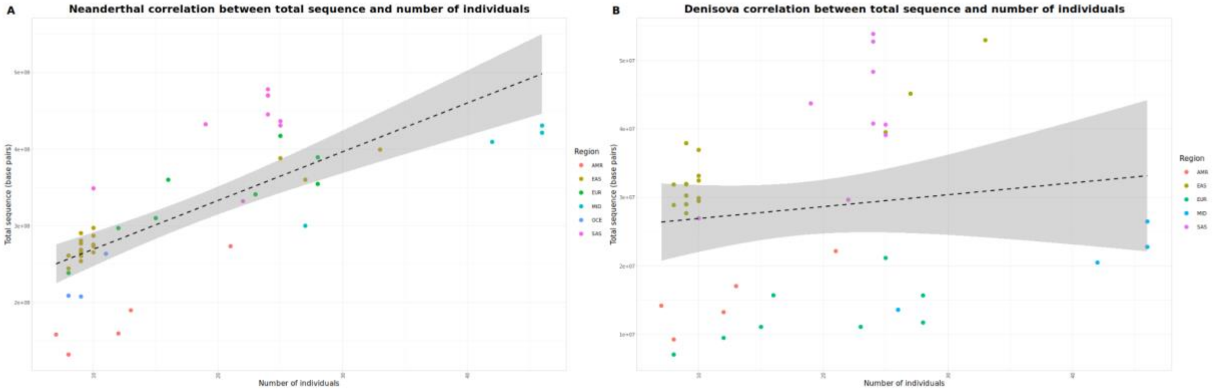


Figure 3: (A) Correlation between the total sequence of Neanderthal artificial genomes in base pairs and the number of individuals. **(B)** Correlation between the total sequence of Denisova artificial genomes in base pairs and the number of individuals. The Oceanian region has been removed, as it skewed the correlation. Data points are colored according to region.

Divergence patterns

The nucleotide divergence is visualized below in Figure 4 for the whole genome, Neanderthals and Denisovans.

Overall, the heatmap for the whole genome (Figure 4A) shows major clusters that separate South Asian, Oceanian, American, East Asian, European, and Middle Eastern populations[16]. The divergence is smallest within the same regions, while the most significant divergence is observed when comparing Oceania with the rest of the world.

When looking at the Neanderthal divergence (Figure 4B), it overall shows the same pattern as for the whole genome; however, there is a lot more noise. This is reflected by the less clear cluster division compared to the whole genome. In contrast to the whole genome divergence pattern, the Neanderthal heatmap indicates that an American population (Surui) is highly divergent from the other populations.

For the Denisova divergence (Figure 4C), there is even more noise compared to the whole genome divergence making it more challenging to separate the clusters. Despite this, it is still possible to see a pattern. The most divergence is between Oceania and Europe/Middle East, Oceania and America, and East Asia and Europe/Middle East.

Looking at the scales, the divergence is much higher for Denisova segments than for Neanderthal segments. Besides that, the pattern is similar; however, the divergence is higher between Europe/Middle East and East Asia for the Denisova segments compared to the Neanderthal segments.

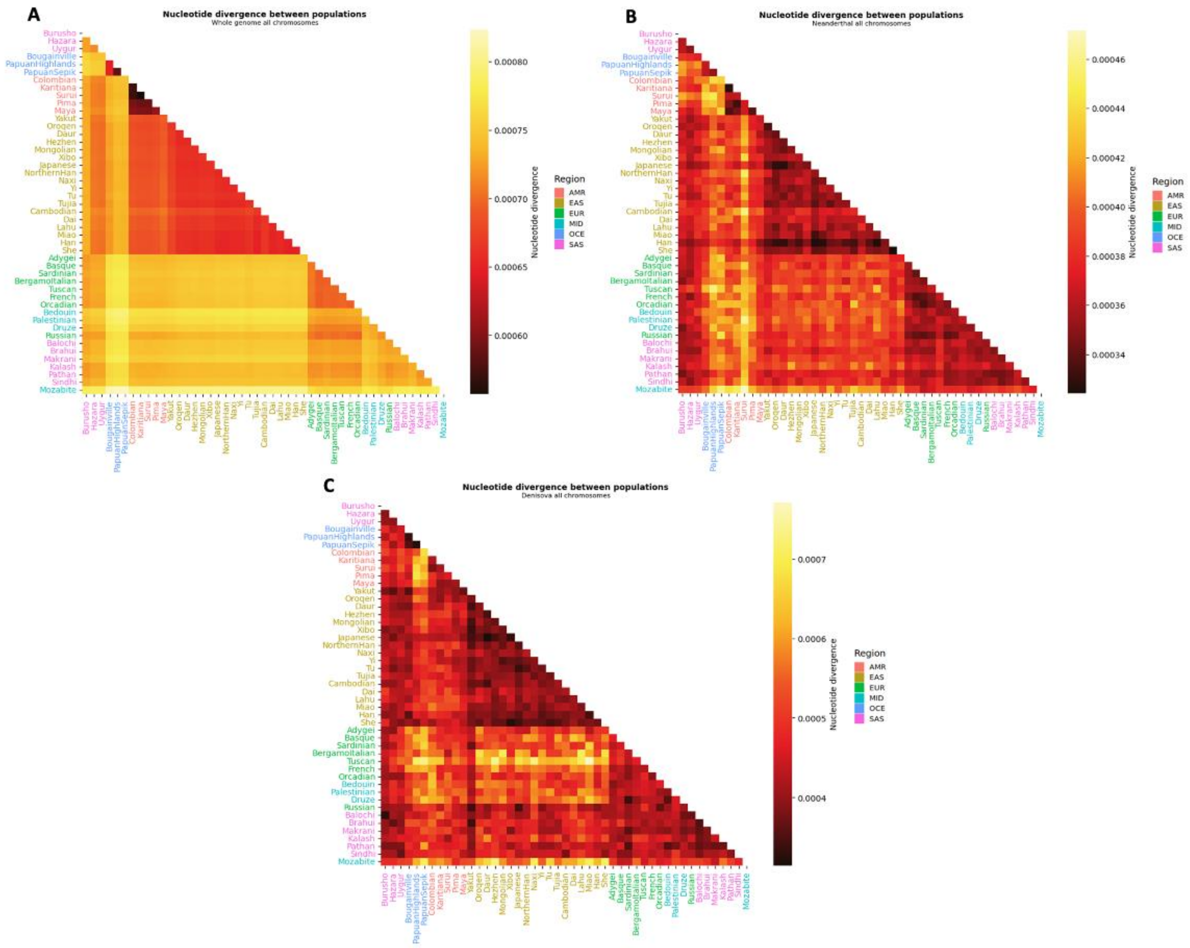


Figure 4: Heatmap showing the nucleotide divergence between populations for the **(A)** whole genome, **(B)** Neanderthal segments, and **(C)** Denisova segments, respectively. The populations are colored according to their region.

Neanderthal and whole genome divergence

Looking further into the correlation between the Neanderthal divergence and the whole genome divergence for the different region comparisons (Figure 5), it is clear from the plots that there is a positive correlation between the Neanderthal divergence and the divergence for the whole genome for all comparisons, which is visualized by the dashed lines. This means that as the whole genome divergence increases, the Neanderthal divergence also increases.

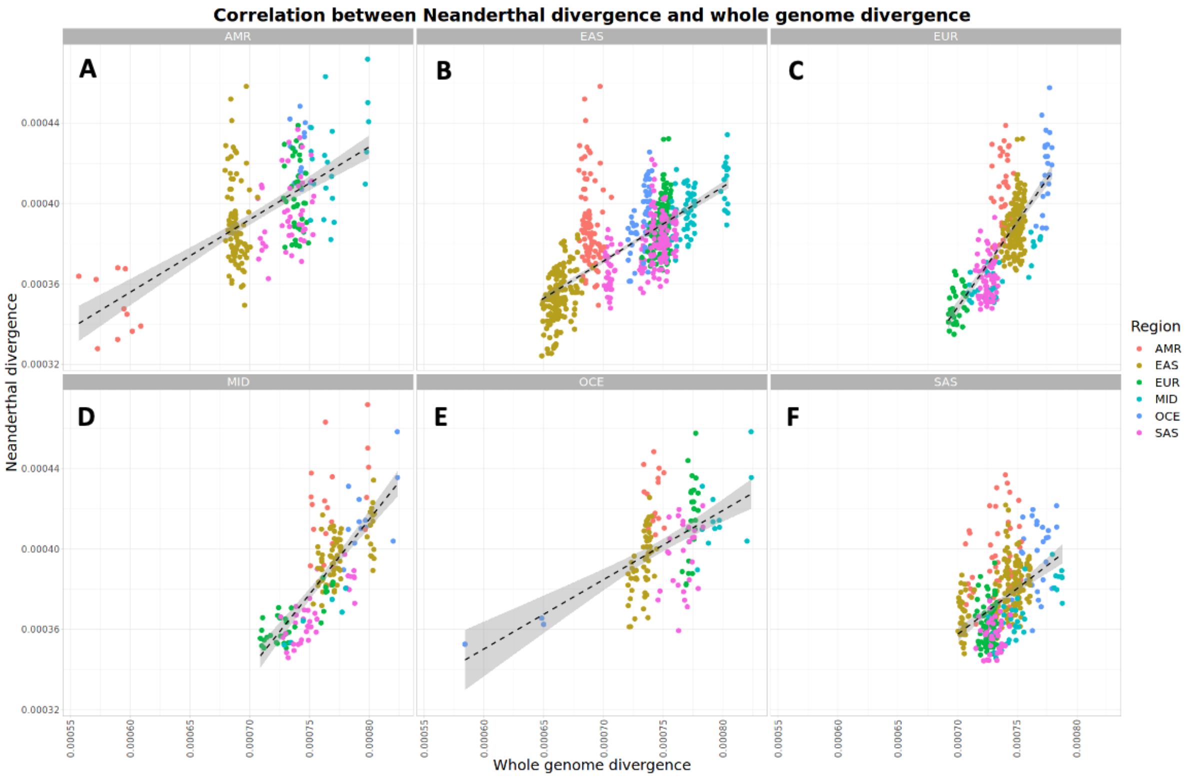


Figure 5: Correlation between Neanderthal divergence and whole genome divergence. The plot is divided into six regions **(A-F)**, which are then compared to all regions. Data points are colored according to region. A linear model is fitted to all regions (dashed line). Populations with an intersection/artificial genome ratio of 1.00 were removed from the plot.

Looking at the plot for America (Figure 5A), it is clear that Americans are very alike regarding their genomes, which can be seen by the small whole genome and Neanderthal divergence when comparing American populations with other American populations. When looking at the whole genome, the American region is most divergent with populations in the Middle Eastern region. The Neanderthal divergence is widely spread (i.e., on the y-axis) among East Asians, South Asians, Europeans, and Middle Easterners. In contrast, the Neanderthal divergence between America and Oceania does not spread as much. The highest Neanderthal divergence is found when comparing American populations with Middle East, East Asian, and Oceanian populations.

Looking at the plot for East Asia (Figure 5B), the lowest overall divergence is found when comparing East Asian populations with East Asian populations. When looking at the whole genome, East Asian populations are most divergent with Middle Eastern populations, followed by Oceania, Europe, and South Asia. The Neanderthal divergence, when compared to American populations, is very spread out, whereas when compared to the other regions, the Neanderthal divergence is denser. Besides that, it seems there are two lines of South Asian and Middle East divergence based on the whole genome divergence.

When looking at the whole genome divergence between Europe and other regions (Figure 5C), the highest divergence is found between Europe and Oceania and some Middle Eastern populations. When comparing Neanderthal content, the highest divergence is found when Europe is compared to Oceania, followed by East Asia and America. The Middle Eastern populations seem to have more similar Neanderthal divergences but are instead spread by the whole genome divergence.

For the Middle Eastern populations (Figure 5D), it can be observed that they share the most sequence with Europeans and South Asians. The highest Neanderthal divergence was found when compared to America, Oceania, and East Asia. Middle Eastern, South Asian, and European populations have similar Neanderthal divergence and are only spread by the whole genome divergence.

Looking at the plot for Oceania (Figure 5E), it is evident that the lowest overall divergence is observed between Oceanian populations. The rest of the regions are in the same cluster with America, Europe, and the Middle East with the highest Neanderthal divergences. When looking at the whole genome divergence, Oceania is most different from the Middle East, followed by Europe and South Asia.

At last, when comparing the South Asian region with the other regions (Figure 5F), the same spread of regions is not observed as in the other plots, and the regions are all in one cluster. The highest whole genome divergence is observed when compared to Oceania and the Middle East. The lowest Neanderthal divergence is observed when comparing South Asia with Europe and South Asia, respectively. In contrast, the highest Neanderthal divergence is found when comparing with America and Oceania, as well as some East Asian populations. It seems as if there are two different clusters of East Asian populations spread by the whole genome divergence.

Denisova and whole genome divergence

Looking further into the correlation between the Denisova divergence and the whole genome divergence for the different region comparisons (Figure 6), there can also be observed positive correlation, which essentially means that as the divergence for the whole genome increases between the regions, the divergence for the Denisovan segments of the genome also increases. When looking at the whole genome divergence, we observe the same pattern as the plots above.

The plot in Figure 6A shows that American populations have the highest Denisova divergence with Oceania, Europe, and the Middle East. In contrast, the lowest Denisova divergence is observed when compared to some East Asian populations and within the American region.

The plot for East Asia (Figure 6B) shows the highest Denisova divergence when comparing East Asia to Europe and the Middle East. The Denisova divergence is about the same for South Asia, Oceania, and America, while the lowest Denisova divergence is observed within the East Asian region.

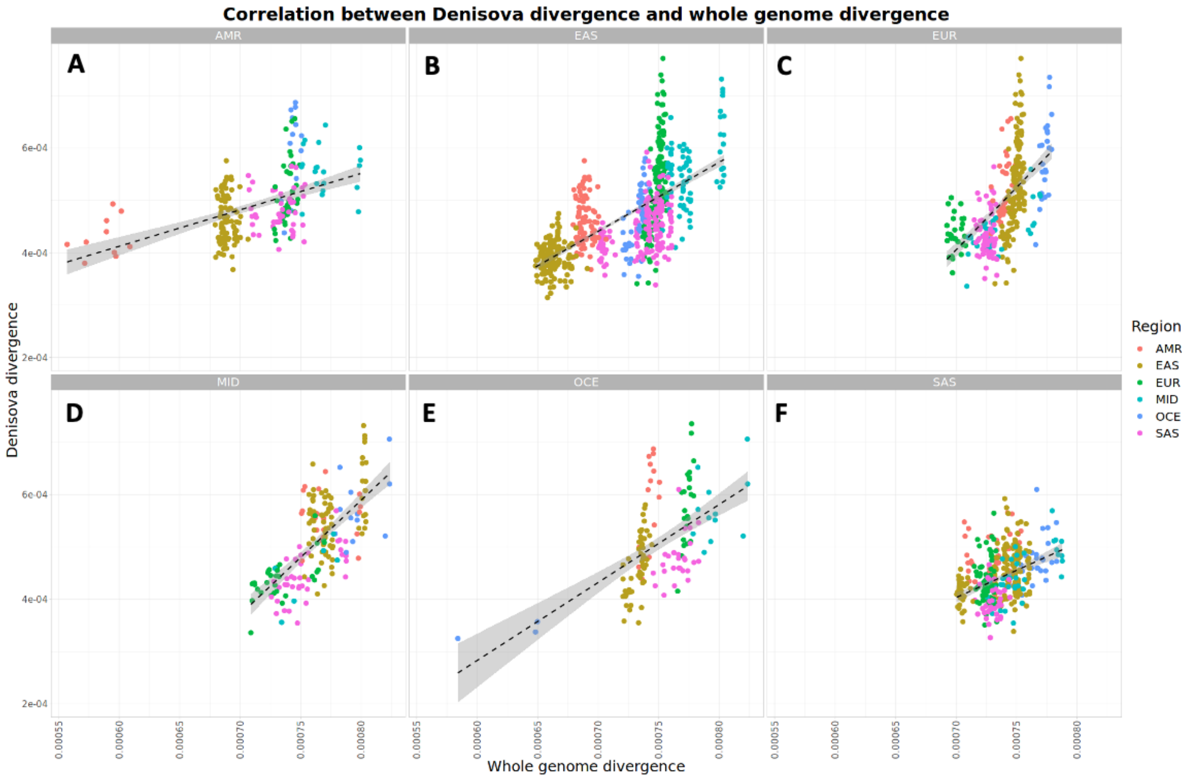


Figure 6: Correlation between Denisova divergence and whole genome divergence. The plot is divided into six regions **(A-F)**, which are then compared to all regions. Datapoints are colored according to region. A linear model is fitted to all regions (dashed line). Populations with an intersection/artificial genome of 1.00 and a Denisova divergence over 6e-05 were removed from the plot.

From the European plot (Figure 6C), it can be observed that the Europeans have the highest Denisova divergence with the entire Oceania region, the America region, and about half of the East Asian region. The lowest Denisova divergence is found within the European region and when comparing European populations with South Asia, the Middle East, and East Asia.

The Middle Eastern populations have the highest Denisova divergence when compared to East Asian populations, Oceanian populations, and American populations. In contrast, the least Denisova divergence is found when compared to European and South Asian populations (Figure 6D).

Looking at the Oceanian region (Figure 6E), the Denisova divergence within the region is the lowest relative to when the Oceanian populations are compared to other regions. Hereafter, the lowest Denisova divergence is observed when compared to East Asia and South Asia. The highest divergence is found when Oceania is compared to Europe, America, and the Middle East.

For South Asia (Figure 6F), distinguishing between the regions is more difficult as they all appear in one big cluster. However, the Denisova divergence overall is lower than for the other plots.

Neanderthal divergence and amount of overlap

Looking at the correlation in Figure 7 between the Neanderthal divergence and the amount of overlap between Neanderthal segments (depicted as the intersection/artificial genome ratio) between the regions, it is observed that there is a negative correlation for all regions. This means that as the Neanderthal divergence increases, the amount of overlap decreases, which can be seen by the dashed lines.

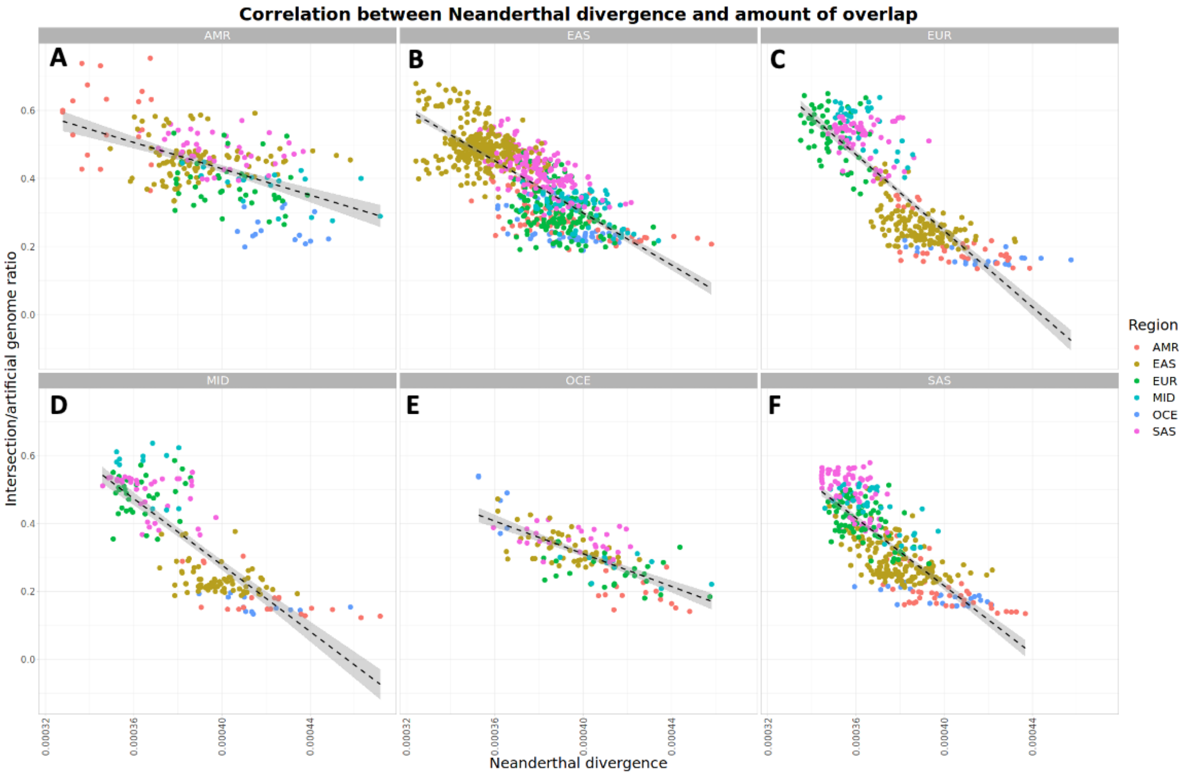


Figure 7: Correlation between Neanderthal divergence and the amount of overlap between the intersection and the artificial genome for Neanderthal segments. The plot is divided into six regions, which are then compared to all regions. Data points are colored according to region. A linear model is fitted to all regions (dashed line). Populations with an intersection/artificial genome of 1.00 were removed from the plot.

In Figure 7A, the data points are scattered between each other, and there is not a very clear partitioning of regions. However, it can be observed that the Americans have the most significant amount of overlap of Neanderthal segments with other Americans (>60%). This is followed by East Asia, South Asia, Europe, and the Middle East (30-60%), and the least amount of overlap is found between America and Oceania (<30%). When looking at the amount of overlap between East Asia and other regions (Figure 7B), it is observed that East Asia has the most significant overlap of Neanderthal segments within the East Asian region of around 40-70%, followed by South Asia. The least amount of overlap is found when comparing East Asia with Europe, America, and Oceania, where the overlap is between 20-40%.

For the Europe region (Figure 7C), the most significant overlap is observed within the Europe region and when compared to the Middle East and South Asia (40-60%). The smallest overlap, below 20%, is with America and Oceania.

The Middle East region (Figure 7D) has the most significant overlap of Neanderthal segments with South Asia and Europe, where the overlap is above 40%. In contrast, the least amount of overlap is found when compared to America, Oceania, and East Asia, with an overlap of around 20%.

For Oceania (Figure 7E), the most overlap, besides within Oceania, is found with East and South Asia, which is only about 30-40%, while the least overlap is with America (20%).

When looking at the plot for South Asia (Figure 7F), the most significant overlap, besides South Asia itself, is observed for the Middle East and Europe, and the lowest is for America and Oceania, which is below 20%.

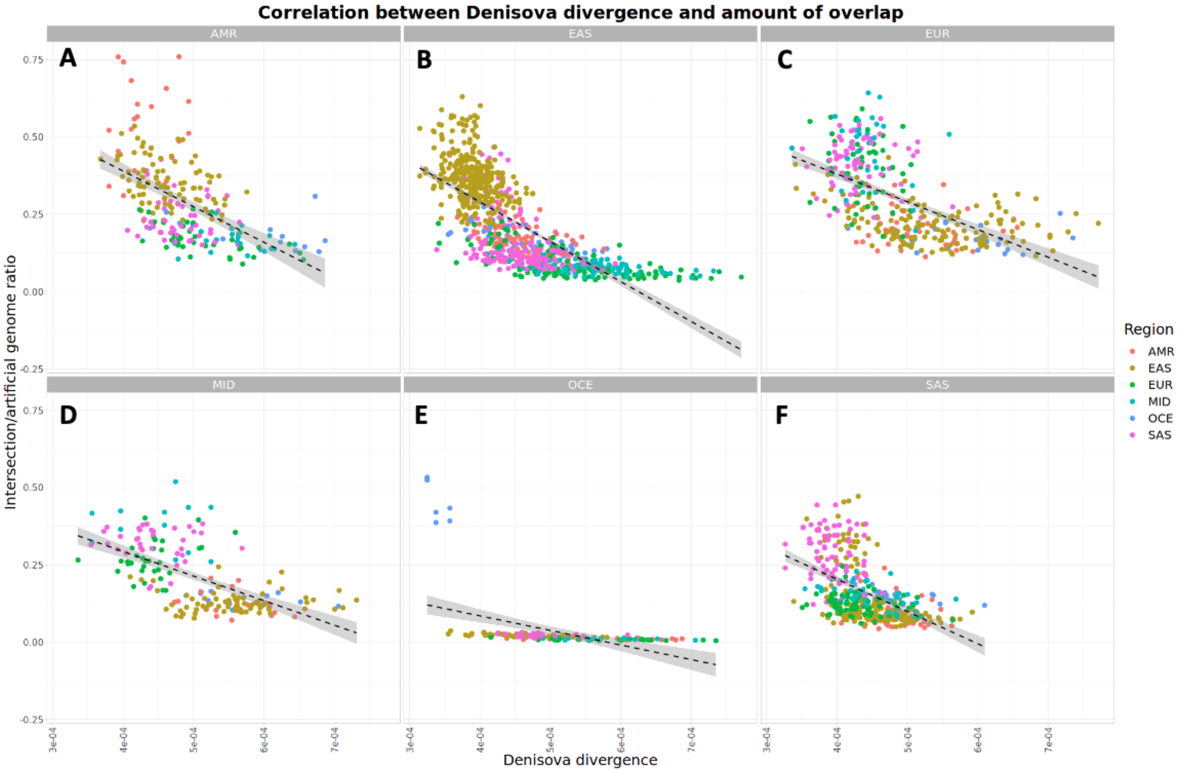


Figure 8: Correlation between Neanderthal divergence and the amount of overlap between the intersection and the artificial genome. The plot is divided into six regions, which are then compared to all regions. Data points are colored according to region. A linear model is fitted to all regions (dashed line). Populations with an intersection/artificial genome ratio of 1.00 and a Denisova divergence over 6e-05 were removed from the plot.

Denisova divergence and amount of overlap

The correlation in Figure 8 between the Denisova divergence and the amount of overlap between Denisova segments (depicted as the intersection / artificial genome ratio) between the regions shows a negative correlation for all regions. This means that as the Denisova divergence increases, the amount of overlap decreases, which can be seen by the dashed lines.

American populations have the highest amount of overlap within the American region, up to over 75% overlap (Figure 8A), followed by East Asians with 25-50% overlap and the least overlap with Europe, the Middle East, and Oceania.
East Asian populations (Figure 8B) have a 25-50% overlap within the region, while the overlaps with the other regions are below 25%.
For European populations (Figure 8C), the highest overlap is observed with Europeans, Middle Easterners, and South Asians (25-50%), while the overlaps with East Asia, Oceania, and America are around 12.5-25%.
Middle East populations (Figure 8D) overlap most with other Middle East populations, European populations, and South Asian populations at around 25-50%. In contrast, the lowest overlap is observed with America, East Asia, and Oceania at about 12.5%.
Oceanian populations (Figure 8E) only have overlapping Denisova segments with other Oceanian populations (40-50%), as the overlaps with all other regions are close to zero.
At last, when looking at the South Asia region (Figure 8F), it is clear that the most significant overlap is with East Asia and South Asia (25-50%), while the overlaps with the other regions are all around 6-12.5%.

Discussion

Artificial genomes summary statistics

The artificial genomes summary statistics showed a correlation between the total sequence and the number of individuals in the populations for the Neanderthal artificial genomes. This correlation indicates that the number of fragments and the total sequence in the resulting artificial genomes are dependent on sample size, which essentially means that the variation observed in those summary statistics might be due to the difference in sample size between the populations rather than actual variation in the amount of Neanderthal content between populations. On the contrary, we can see that the same test for the Denisovan artificial genomes shows no correlation. This is expected since the amount of Denisova component in the analyzed populations is highly variable. Thus, the amount of Denisova content in the artificial genomes will depend on the Denisova content in each population and not sample size, as seen with Neanderthals, which varies to a much lesser extent. Based on already established knowledge about Denisova content and the variation among populations, populations in Oceania, on average, have more Denisova fragments than the rest of the world, which is also reflected in the artificial genomes summary statistics as the Denisova artificial genomes for the Oceanian populations are longer than the other populations.

Strategy of artificial genomes

To summarize the archaic content in each population, I create the artificial genomes in this study. With this strategy, the archaic fragments were sampled randomly among all individuals from a given population, and overlapping fragments were discarded. This strategy was chosen to maximize the archaic content to be compared, in opposition to other methods, such as sampling just a single individual per population. This latter strategy could also bias the results towards the specific archaic content in the selected individuals. However, summarizing the archaic content per population could have been done in other ways. An example could be to order the archaic fragments by size to sample the longest non-overlapping fragment first to the artificial genome, which could increase archaic content but potentially sample fragments from individuals with generally longer fragments.

Another strategy could be, instead of removing entire archaic fragments if they overlap with another sampled fragment, then remove the overlapping intervals of the fragments, and add the non-overlapping part of the fragments to the artificial genome. The benefit of this strategy is getting the maximal information on the archaic content within a population; however, this strategy is a bit more complicated practically.

Future work could be to test different ways of sampling to compare populations in terms of archaic content. One potential variation to the so-called artificial genome strategy would be to subsample individuals belonging to the same population so that all populations would consist of the same number of individuals so that the dependency seen between total archaic content and the sample size is avoided. However, in this way, the samples are limited by the minimum sample size (7 individuals). In this way, much information is thrown away as some populations have more than 20 individuals, and most have more than 10-12 individuals.

Divergence patterns

Overall, heatmaps for the whole genome, Neanderthal, and Denisova divergence showed similar patterns. However, the Denisova divergence differed more compared to the other two in terms of the high divergence between Oceania and Europe/Middle East, Oceania and America, as well as East Asia and Europe/Middle East. Essentially, as the admixture events with Denisovan population sources are suggested to have occurred in South/East Asia, it was also expected that a higher divergence would occur between Europe and South/East Asia. The discrepancy observed in Denisova divergence between Europe and East Asia compared to Europe and South Asia could be caused by several things. First, the more significant divergence between East Asia and Europe compared to South Asia and Europe could be because East Asia has at least two different sources of introgression.

Second, South Asia is suggested to have multiple contributions from different ancestral components, some of which are related to West Eurasian populations[7], which essentially might have resulted in a lower Denisovan divergence between Europe and South Asia.

The highest Denisovan divergence would be expected to be observed between populations with different sources of introgression. However, as there is limited data on Denisova content to compare between populations, it affects the divergence pattern making it more unclear. This is also observed in the Neanderthal heatmap; however, the pattern is not as unclear as for Denisova, which could be explained by the fact that more data on Neanderthal content is available. Besides this, the Neanderthal heatmap is similar to the whole genome. Overall, the divergence range is much smaller for Neanderthal content than for Denisova content. The small range of Neanderthal divergence indicates one contributing Neanderthal population that introgressed before the split of non-African populations. The variation in Neanderthal content after this event might be caused by differences in selection among these groups or additional subsequent admixture events private to some populations. The higher range of Denisova divergence might stem from introgression from at least two divergent Denisova populations, resulting in higher divergence when comparing Denisova content among modern human populations.

The results are based on polymorphism data where the African polymorphisms are included. If the African variants were removed, it might reduce the noise and make the results more easily interpretable. Because African populations exhibit a high level of genetic diversity, including their polymorphisms provides a more comprehensive representation of the overall genetic variation within the human populations. This means that when including the African polymorphisms in the analyses of divergence between non-African populations, the African polymorphisms might dominate the divergence measurements, and the divergence measures are likely to increase.

Correlation between divergences and overlap

From the correlation plots, it is clear that there is a positive correlation between Neanderthal divergence and whole genome divergence. Similarly, a positive correlation exists between Denisova divergence and whole genome divergence. On average, The Denisova divergence is higher than the Neanderthal divergence, which could be explained by introgression from multiple divergent Denisovan populations into extant individuals. This makes the interpretation more ambiguous because different populations might have been involved in various admixture events with Denisovans. In that way, the divergence observed between populations could be because some populations have had more interbreeding events with Denisovans than others. Furthermore, the divergence could be caused by interbreeding events with different Denisovan-like groups, reflecting the difference in Denisova content between populations.

When looking at Oceania, the Denisova divergence within Oceania is much smaller than when comparing Oceania to the other regions. This confirms the hypothesis that Oceanian populations only have traces of introgressed DNA from a distantly related Denisovan group, and most likely, the admixture event only involved the populations from Oceania[7].

Overall, there is a negative correlation between Neanderthal/Denisova divergence and the amount of overlap of Neanderthal/Denisova segments between the populations. This correlation seems to reflect the migrations and splits of the anatomically modern humans and the interbreeding with Neanderthals and Denisovans, as explained above.

As the Denisova segments, on average, have a smaller overlap between populations compared to the Neanderthal segments, it could reflect that multiple independent introgression events occurred for Denisovans. In contrast, the introgression event for Neanderthals was common for all non-African populations because the amount of overlap is more significant between populations.
Specifically for the Americas, it seems that populations within this region have a significant overlap of both Neanderthal and Denisovan segments between them. Multiple factors could explain this. Firstly, the migration to America occurred relatively recently (around 15,000 to 20,000 years ago)[16], meaning there has been less time for a genetic divergence to accumulate among the American populations than in other regions. Secondly, the American populations may have experienced genetic bottleneck effects, as the initial migration only involved a small population, and thereby only a limited number of individuals contributed to the gene pool. This would reduce the genetic diversity between populations, increasing homogeneity within American populations. Lastly, the Americas is geographically located separately from other continents, which means that the amount of gene flow and genetic exchange with other populations is limited. This also leads to lower levels of divergence between American populations and populations in other regions with a more comprehensive gene flow.

When looking at the Denisovan content for Oceania, there is little to no overlap with other regions. An admixture event unique to the Oceanian populations can explain this lack of overlap. Furthermore, like America, Oceania is isolated geographically from the other continents, resulting in limited gene flow and genetic exchange with the rest of the world. As a result, the genetic diversity within populations in Oceania has evolved independently over time, contributing to the increased divergence from other populations.

Concluding remarks

The Neanderthal divergence measurements supported a single introgression event from a Neanderthal population that happened before the split of non-African populations. On the contrary, the Denisova divergence measurements suggested multiple introgressions from multiple Denisova population sources.

Future work could involve other strategies for sampling archaic content, such as downsampling to have the same number of individuals in all samples, sampling the longest non-overlapping fragments to get more information, or sampling all non-overlapping intervals between fragments to get as much information as possible.

References

1. King W. The reputed fossil man of the Neanderthal. *Q J Sci*. 1864;1:88-97.
2. Pääbo S. *Neanderthal Man: In Search of Lost Genomes*. First paperback edition. Basic Books; 2015.
3. Green RE, Krause J, Briggs AW, et al. A Draft Sequence of the Neandertal Genome. *Science*. 2010;328(5979):710-722. doi:10.1126/science.1188021
4. Reich D, Green RE, Kircher M, et al. Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature*. 2010;468(7327):1053-1060. doi:10.1038/nature09710
5. Reilly PF, Tjahjadi A, Miller SL, Akey JM, Tucci S. The contribution of Neanderthal introgression to modern human traits. *Curr Biol*. 2022;32(18):R970-R983. doi:10.1016/j.cub.2022.08.027
6. Krause J, Fu Q, Good JM, et al. The complete mitochondrial DNA genome of an unknown hominin from southern Siberia. *Nature*. 2010;464(7290):894-897. doi:10.1038/nature08976
7. Teixeira JC, Cooper A. Using hominin introgression to trace modern human dispersals. *Proc Natl Acad Sci*. 2019;116(31):15327-15332. doi:10.1073/pnas.1904824116
8. Bergström A, Stringer C, Hajdinjak M, Scerri EML, Skoglund P. Origins of modern human ancestry. *Nature*. 2021;590(7845):229-237. doi:10.1038/s41586-021-03244-5
9. Browning SR, Browning BL, Zhou Y, Tucci S, Akey JM. Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture. *Cell*. 2018;173(1):53-61.e9. doi:10.1016/j.cell.2018.02.031
10. Bergström A, McCarthy SA, Hui R, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*. 2020;367(6484):eaay5012. doi:10.1126/science.aay5012
11. Skov L, Hui R, Shchur V, et al. Detecting archaic introgression using an unadmixed outgroup. *PLOS Genet*. 2018;14(9):e1007641. doi:10.1371/journal.pgen.1007641
12. bedtools: A powerful toolset for genome arithmetic — bedtools 2.30.0 documentation. Accessed April 20, 2023. https://bedtools.readthedocs.io/en/latest/
13. scikit-allel: Explore and analyse genetic variation — scikit-allel 1.3.3 documentation. Accessed April 19, 2023. https://scikit-

allel.readthedocs.io/en/stable/

14. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc Natl Acad Sci*. 1979;76(10):5269-5273. doi:10.1073/pnas.76.10.5269

15. gwf 1.8.5 documentation. Accessed April 19, 2023. https://gwf.app/

16. Nielsen R, Akey JM, Jakobsson M, Pritchard JK, Tishkoff S, Willerslev E. Tracing the peopling of the world through genomics. *Nature*. 2017;541(7637):302-310. doi:10.1038/nature21347