UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS SUPERIORES
UNIDAD JURIQUILLA

DETECTING NEANDERTHAL AND DENISOVAN INTROGRESSED GENOMIC SEGMENTS IN PRESENT-DAY MEXICO

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

Licenciada en Ciencias Genómicas

PRESENTA:

Valeria Alejandra Añorve Garibay

TUTORA:

Dra. Emilia Huerta-Sánchez

CO-TUTOR:

Dr. Vicente Diego Ortega Del Vecchyo

ENES
JURIQUILLA

Juriquilla, Querétaro, México, 2023

Declaración de autenticidad

Por la presente declaro que el contenido de esta tesis es original y no ha sido presentado total o parcialmente para su consideración en ningún otro medio o Universidad. Esta tesis ha sido desarrollada íntegramente por la autora que la suscribe y no es producto de ninguna colaboración, a menos que el texto lo especifique.

Valeria Alejandra Añorve Garibay. Juriquilla, Querétaro, México, 2023.

To my parents… May the world ever be kind and full of gratitude!

Abstract

To my advisor, Emilia Huerta-Sánchez. Thank you for opening your lab doors to me. I have learned so much since I joined the lab. I greatly appreciate your mentorship and support throughout.

To my co-advisor, Diego Ortega Del-Vecchyo. Thank you, Diego, for your mentorship, for introducing me to human populations genetics and for your support through all my career.

To my committee members, Mashaal Sohail, Maria Ávila-Arcos, Federico Sánchez-Quinto and Andrés Moreno-Estrada. Thank you for your feedback and for your availability during this process. I specially thank Mashaal for her continuous help navigating this project and for being a great mentor.

To the members of the Huerta-Sánchez Lab and to the members of the Computational Population Genetics Group. Thank you all for the input and help over the year.

To the Center for Computational Molecular Biology (CCMB) and to the program Apoyo a Proyectos de Investigación e Innovación Tecnológica (DGAPA-PAPIIT, IA206222) for funding this thesis. I also thank Jair S. Garcia-Sotelo, Luis Alberto Aguilar Bautista, Alejandra Castillo, and Carina Uribe for their technical assistance.

To my parents, Olga and Jorge, for their unconditional love and support through this process. Thank you for encouraging me to work towards my dreams.

To my partner, Hugo, and my closest friends, Natalia, Victor and Luis. Thank you for your love and for being with me at every step of this journey. You are wonderful people and I am grateful for having you in my life.

Finally, to the Genomics Sciences program. Thank you to all my professors for introducing me to science. I specially thank Margareta and Memo for being such a great guide and for all the love and support.

Thank you!

Abstract

Genomic analyses have revealed that anatomically modern humans (AMH) interbred with Neanderthal and Denisovan archaic populations about 50,000 years ago. Characterizing the impact of archaic introgression on the genomes of present-day modern humans is important to understand the functional consequences of archaic-introgressed genetic variants. The detection of archaic introgression has been carried worldwide using genomic data from the 1,000 Genomes Populations which include Mexican individuals living in Los Angeles (MXL). However, we lack a comprehensive understanding of the distribution of archaic introgression in present-day Mexico. Here I applied SPrime, a reference-free method for detecting archaic introgression, to 5,833 genome-wide genotypes from the Mexican Biobank Project (MXB). I show that imputed genomes are a valuable resource for inferring archaic-introgressed segments. I inferred a set of 146,875 SNPs that are likely to be from Neanderthal or Denisovan origin. I find that individuals in the southern part of Mexico have on average more archaic alleles compared to individuals in the northern part of Mexico, which is consistent with the observed positive correlation between Indigenous American genetic ancestries and the number of sites with archaic alleles, and the observed negative correlation between European and African genetic ancestries and the number of sites with archaic alleles. I also find an enrichment of Denisovan genetic ancestry in MXB individuals compared to MXL. I provide a valuable resource for characterizing the genomic landscape of archaic introgression in present-day Mexico.

Table of contents

List of Abbreviations

MXB: The Mexican Biobank Project.

MXB-genotyped: The Mexican Biobank Project genotyped data presented and analyzed in Sohail et al., 2022.

MXB-imputed: The Mexican Biobank Project imputed genomes for 5,833 unrelated individuals.

Indigenous MXB: 50 MXB individuals that have whole-genome sequenced data and imputed data.

AFR: Yorubas from 1000 Genomes Project representing African genetic ancestries.

EUR: Iberians from 1000 Genomes Project representing European genetic ancestries.

AMR: Peruvians from 1000 Genomes Project with Indigenous American genetic ancestry greater than 99% representing Indigenous American genetic ancestries.

List of Figures

represents the average value among chromosomes. B) Distribution of the length of introgressed segments.

Figure 7. Putatively archaic-introgressed segments inferred in MXB using YRI as an outgroup. Results in B, C and D show results obtained after applying filters to isolate regions with Neanderthal and Denisovan ancestry. A) The distribution of the length of introgressed segments. B) The number of SNPs per segment that are comparable to the archaic genome. C) Proportion of SNPs per segment matching the Altai Neanderthal allele. D) Proportion of SNPs per segment matching the Altai Denisovan allele.

Figure 8. Contour density plot of match proportion of introgressed segments to the Altai Neanderthal and to the Altai Denisovan genomes. For a given segment, a match rate of 0 denotes that for that segment, none of the alleles match the corresponding allele in the archaic-human reference. A) Density distribution of match rate to archaic individuals (Neanderthal and Denisovan). B) Density distribution of match rate to archaic individuals after isolating regions with Neanderthal and Denisovan ancestry.

Figure 9. The number of sites with archaic alleles (either Neanderthal, Denisovan) found in 1,411 segments detected by SPrime in 5,833 MXB individuals. Individuals are grouped according to their sampling location. Color coding is by archeological region. Boxplots represent the distribution of the number of sites with archaic alleles for each state.

Figure 10. The correlation between percentages of African (AFR), European (EUR) and American (AMR) genetic ancestries and the number of sites with archaic alleles. The X-Axis is the genetic ancestry proportion for a given ancestry type (AFR as pink, EUR as yellow and AMR as purple). The Y-Axis is the number of sites with archaic alleles. The first row represents results at individual level and the second row represents results by grouping individuals by sampling location to account for states in Mexico.

Figure 11. Heatmap representation of the p-values of multiple pairwise-comparison of the average number of sites with archaic alleles by state in Mexico. States are depicted from north to south. Significance levels are expressed by asterisks (* for p value < 0.01 and ** for p value < 0.001).

Figure 12. Distribution of the number of sites with archaic alleles for all MXB individuals and all individuals from MXL, PEL, CLM and PUR in 1000 Genomes Project. Color-coding

is for MXB (blue) and Admixed American populations (black). A) Archaic SNPs (either Neanderthal or Denisovan) B) Neanderthal-specific. C) Denisovan-specific.

Supplementary Figure 1. Relationship between imputation accuracy, R^2, and the averaged metric for all SNPs in all autosomes. A) Genotype accuracy. B) Heterozygotic precision. C) Homozygotic precision.

Supplementary Figure 2. Principal component analysis (PCA) plots of a panel of 868,091 LD-pruned SNPs. Individuals are color-coded by population (archeological region). Results shown are from MXB-imputed data. A) PC1 vs PC3. B) PC2 vs PC3.

Supplementary Figure 3. The number of sites with Altai Neanderthal-specific alleles found in 1,411 segments detected by SPrime in 5,833 MXB individuals. Individuals are grouped according to their sampling location. Color coding is by archeological region. Boxplots represent the distribution of the number of sites with archaic alleles for each state.

Supplementary Figure 4. The number of sites with Altai Denisovan-specific alleles found in 1,411 segments detected by SPrime in 5,833 MXB individuals. Individuals are grouped according to their sampling location. Color coding is by archeological region. Boxplots represent the distribution of the number of sites with archaic alleles for each state.

List of Tables

1. Introduction

The analysis of whole-genome sequences of Neanderthal genomes (Green et al., 2008, Green et al., 2010, Prüfer et al., 2014, Prüfer et al., 2017) and Denisovan genomes (Reich et al., 2010, Meyer et al., 2012, Prüfer et al., 2014, Sawyer et al., 2015) has elucidated a complex admixture landscape and multiple events of introgression between archaic hominins and anatomically modern humans (AMH) (Prüfer et al., 2014, Villanea and Schraiber, 2018). Genomic analyses have revealed that modern humans interbred with archaic hominins such as Neanderthals and Denisovans since 50,000 years ago (Plagnol et al., 2006, Wall et al., 2009, Sankararaman et al., 2014, Villanea and Schraiber, 2018). This introgression event introduced archaic variants into the ancestral out-of Africa (OoA) human gene pool. Subsequently, all present-day individuals in Eurasia inherited ~2% of their genome from Neanderthals (Green et al., 2010), and individuals from Oceania inherited ~5% of their genome from Denisovans (Reich et al., 2010). Additionally, it has been suggested that ghost populations (unidentified hominin species) interbred with African populations (Wall et al., 2009, Hammer et al., 2011) but this result is still contentious (Ragsdale et al., 2023).

Genetic variants inherited from archaic hominins (archaic-introgressed variants) have functional consequences that have helped humans to adapt to the hypoxic environment of the high-altitude Tibetan plateau (Huerta-Sánchez et al., 2014) or to have increased risk of type 2 diabetes (The SIGMA Type 2 Diabetes Consortium et al., 2014). Therefore, it is necessary to identify specific haplotypes that were inherited from archaic hominins (Huerta-Sánchez et al., 2014, Sankararaman et al., 2014, Vernot and Akey, 2014, Browning et al., 2018) to understand the functional, phenotypic and evolutionary consequences of archaic introgression.

We currently have archaic-introgression maps from worldwide populations based on data from the 1,000 Genomes Project (Browning et al., 2018) but there is a lack of a fine-scale introgression map from Mexican individuals living in Mexico. This is important to understand the impact of archaic introgression on the adaptation to different environments (Gower, Graham, et al., 2021) or on the evolution of complex traits of medical interest (McArthur et al., 2021, Koller et al., 2022, Wei et al., 2023).

Several methods have been developed to identify archaic alleles in present-day genomes (Plagnol, Vincent, and Jeffrey D. Wall 2006, Racimo et al., 2015, Browning et al., 2018, Skov et al., 2018). A noteworthy method is SPrime (Browning et al., 2018), a reference-free method for inferring archaic-introgressed variants. SPrime is optimized for detecting Neanderthal and Denisovan haplotypes and it is based on the observation that linkage disequilibrium (LD) is higher in introgressed genomic regions than in non introgressed regions. SPrime has better accuracy than previous methods such as S* (Plagnol and Wall, 2006), and brings the possibility of comparing the detected archaic-introgressed segments to reference archaic genomes to identify the archaic introgression source (Browning et al., 2018).

SPrime has been applied to several populations from Eurasia and Oceania (Browning et al., 2018, Zhang et al., 2021). For Neanderthals, studies have focused on modern Eurasians, as hundreds of genomes are available from the 1000 Genomes Project (Browning et al., 2018). However, we lack a comprehensive understanding of how introgression has shaped the complex admixture landscape of Admixed populations from the Americas. Recent research has shown that there is a large potential for discovering novel adaptive archaic introgression in Admixed genomes (Villanea et al., 2022), as European colonization may have impacted the distribution of archaic-introgressed variants (Villanea et al., 2022, Witt et al., 2023).

In this work, I address the lack of studies by analyzing archaic-introgressed variation in present-day Mexico. I applied SPrime to 5,833 imputed genomes of individuals from the Mexican Biobank Project to identify putative archaic-introgressed alleles. I identified a set of archaic-introgressed variants that are putatively from Neanderthal or Denisovan origin. I leverage the resulting maps of Neanderthal or Denisovan ancestry to characterize the distribution of Neanderthal or Denisovan introgression in present-day Mexico. I find that individuals sampled in the southern part of Mexico have, on average, more sites with archaic-specific alleles than those sampled in the northern part of Mexico, which is consistent with a greater proportion of Indigenous American genetic ancestries in Southern Mexico (Moreno-Estrada et al., 2014, Sohail et al., 2022). In concordance with previous studies (Witt et al., 2023), I find that the number of archaic-introgressed variants is proportional to the proportion of Indigenous American genetic ancestry.

Although MXL individuals from 1000 Genomes Project are commonly used to represent present-day archaic variation in Mexico, I find that individuals in MXB have more Denisovan ancestry than MXL individuals.

Overall, I inferred a map of putatively archaic-introgressed variants identified in individuals residing in present-day Mexico. My results provide resolution to the distribution of Neanderthal and Denisovan archaic-introgressed variants in Mexico, and shed light on the history and evolution of introgression events in Mexico. This will be useful for understanding the phenotypic consequences of archaic introgression in Mexico on future studies.

2. Results

2.1. Imputation performance of the Mexican Biobank Project Genomic Data

I assessed the performance of the imputation algorithm the MXB Team employed for imputing the MXB-genotyped data. For doing so, I leveraged genomic information of a panel of 50 Indigenous MXB individuals broadly distributed across Mexico (Figure 1, Materials and Methods). The 50 Indigenous MXB individuals have whole-genome sequencing data and imputed data. The MXB Team generated the imputed dataset by using the Trans-Omics for Precision Medicine (TOPMed) Imputation Reference Panel (Taliun et al. 2021) through the Michigan Imputation Server. Therefore, I created a dataset containing information for 5,933,348 bi-allelic SNPs that are present in both the imputed and whole-genome sequence data for the 50 Indigenous MXB individuals (Materials and Methods).

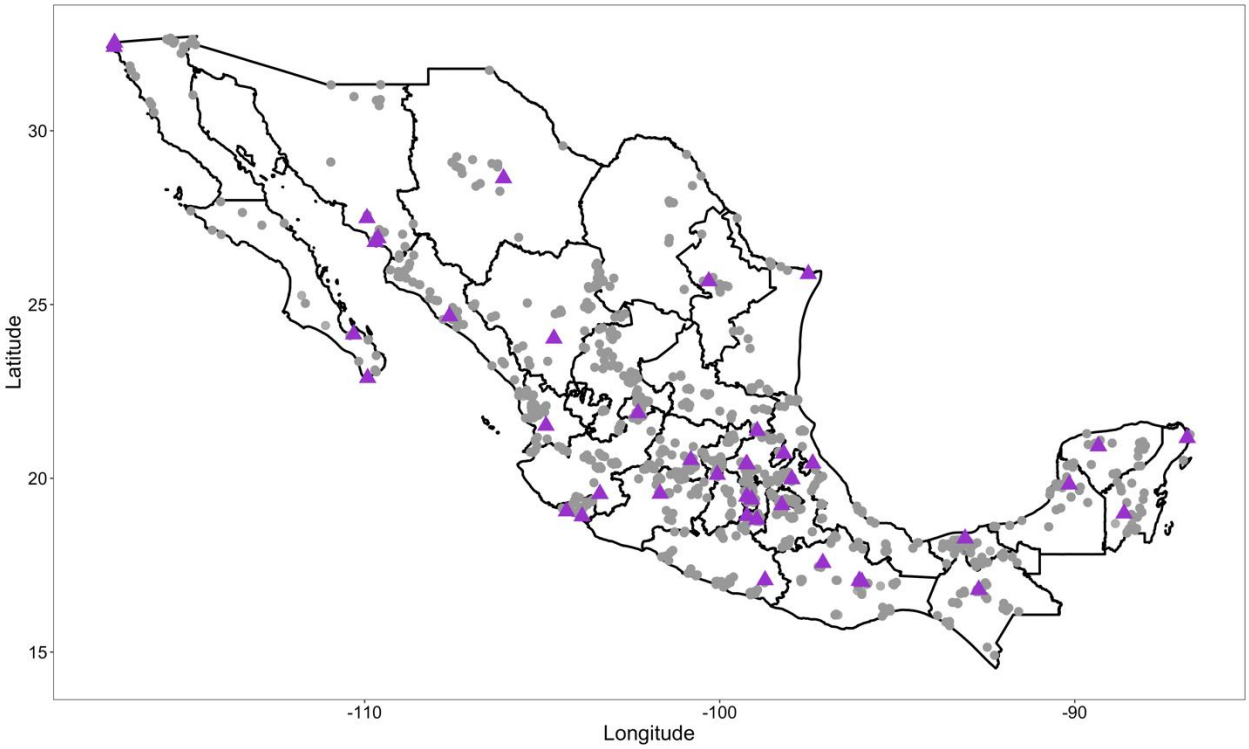


Figure 1. Sampling location of the 6,057 individuals in The Mexican Biobank Project. Gray dots represent samples that were genotyped. Purple triangles represent samples that were both genotyped and whole-genome sequenced (Indigenous MXB). All samples were imputed, however, for all the subsequent analysis I used a subset of 5,833 unrelated individuals (MXB-imputed).

I evaluated imputation performance by computing four different metrics: genotype correlation, genotype accuracy, heterozygotic precision and homozygotic precision. I defined genotype correlation as the Pearson correlation coefficient between the real allele dosages and the imputed allele dosages. Genotype accuracy was defined as the ratio of genotypes that were correctly called in the imputed data to the number of called genotypes. Finally, I defined heterozygotic precision as the ratio of heterozygotes that were correctly called in the imputed data to the number of called heterozygous in the real data and homozygotic precision as the ratio of homozygous that were correctly called in the imputed data to the number of called homozygous in the real data.

The imputation algorithm that the MXB Team employed, *minimac2* (Fuchsberger et al., 2015), returns an imputation accuracy R^2 value for each imputed marker. However, this metric may not be representative of how well the imputed markers are mimicking the whole-genome markers for the same set of individuals. I defined four R^2 cutoff values ($R^2 > 0.05$, 0.1, 0.2 and 0.3) to assess the relationship between R^2 and the four computed metrics. Then, I computed the average value for each imputation metric using all the SNPs that had an R^2 value above the R^2 threshold.

Figure 2 panel A shows the relationship between the average genotype correlation value and the R^2 cutoff values. As expected, I observed an increase in the average correlation as the R^2 value increases. I observed the same pattern for the other three computed statistics (Supplementary Figure 1). However, this increment is very minimal for all metrics, being less than 0.01% for all the metrics considered. Since MXB-imputed markers with $R^2 > 0.05$ have similar average genotype correlation to markers with the other three tested R^2 cutoff values, I decided to use 0.05 as the R^2 cutoff value.

To assess how allele frequency impacts imputation performance, I computed the minor-allele frequency for all markers with $R^2 > 0.05$. Figure 2 Panel B shows the relationship between the minor-allele frequency of each imputed marker and the average genotype correlation. I observed that the genotype correlation increases as the minor-allele frequency increases, which is expected as imputation accuracy tends to be worse for rare variants (MAF < 0.5%) (Jiménez-Kaufmann et al. 2022).

Overall, I have demonstrated that the MXB-imputed markers with $R^2 > 0.05$ were imputed with high accuracy. Based on my results, I conclude that an R^2 cutoff value of 0.05 is reasonable for subsequent analysis that will use imputed data to call introgressed archaic segments.

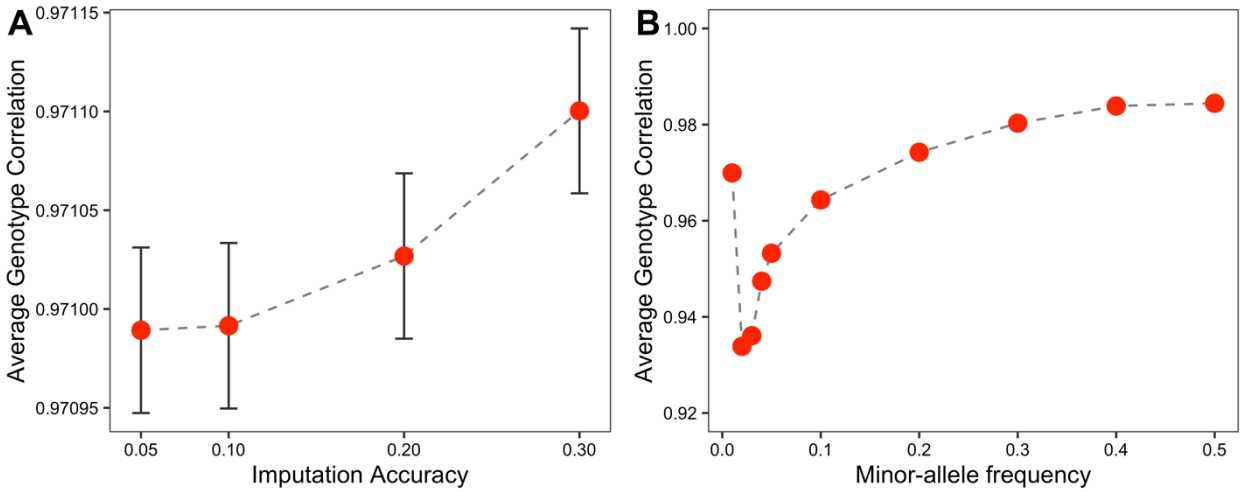


Figure 2. Imputation performance of a panel of 5,933,276 bi-allelic SNPs in 50 Indigenous MXB individuals broadly distributed across Mexico. A) Relationship between imputation accuracy, R^2, and the averaged genotype correlation for all SNPs in all autosomes. Black lines are standard-error bars. B) Relationship between minor-allele frequency and the average genotype correlation across the non-reference interval.

2.2. Population structure of the Mexican Biobank Project imputed markers

I conducted population structure analyses to further evaluate imputation performance of the MXB-imputed markers. This is an important step as imputation can introduce bias if the reference panel does not adequately represent the genetic diversity of the target population (Jiménez-Kaufmann et al. 2022).

I begin by conducting a Principal Component Analysis (PCA) using data from MXB-imputed and three reference populations: Yorubas as Africa (AFR), Iberians as Europe (EUR) and Peruvians (PEL) as Indigenous America (AMR) from the 1000 Genomes Project, to capture predominant axes of population structure and to compare my results with the ones obtained in Sohail et al., 2022 (Materials and Methods).

Figure 3 panel A shows the genetic variation in MXB-imputed in relation to 1000 Genome reference populations. In agreement with Sohail et al., 2022, I found that the values for the first two PCs for most MXB-imputed individuals are similar to those of

present-day individuals from America and Europe, which is consistent with the complex biological admixture process caused by the Spanish colonization. By analyzing MXB-imputed alone (Figure 3: B), I observed that there is subtle genetic substructure within Mexico, which reflects the effects of migration among the different archeological regions. However, PC3 slightly reflects population substructure between the Mayan region and the rest of Mexico (Supplementary Figure 2).

I analyzed the impact of imputation on the population structure of Indigenous MXB individuals by performing a PCA analysis on a data panel comprising whole-genome information for the 50 Indigenous MXB individuals (n SNPs = 351,877) (Materials and methods). Notably, population substructure for Indigenous MXB is well reflected by MXB-imputed data, as it is able to maintain population structure among Indigenous MXB individuals (Figure 3: C).

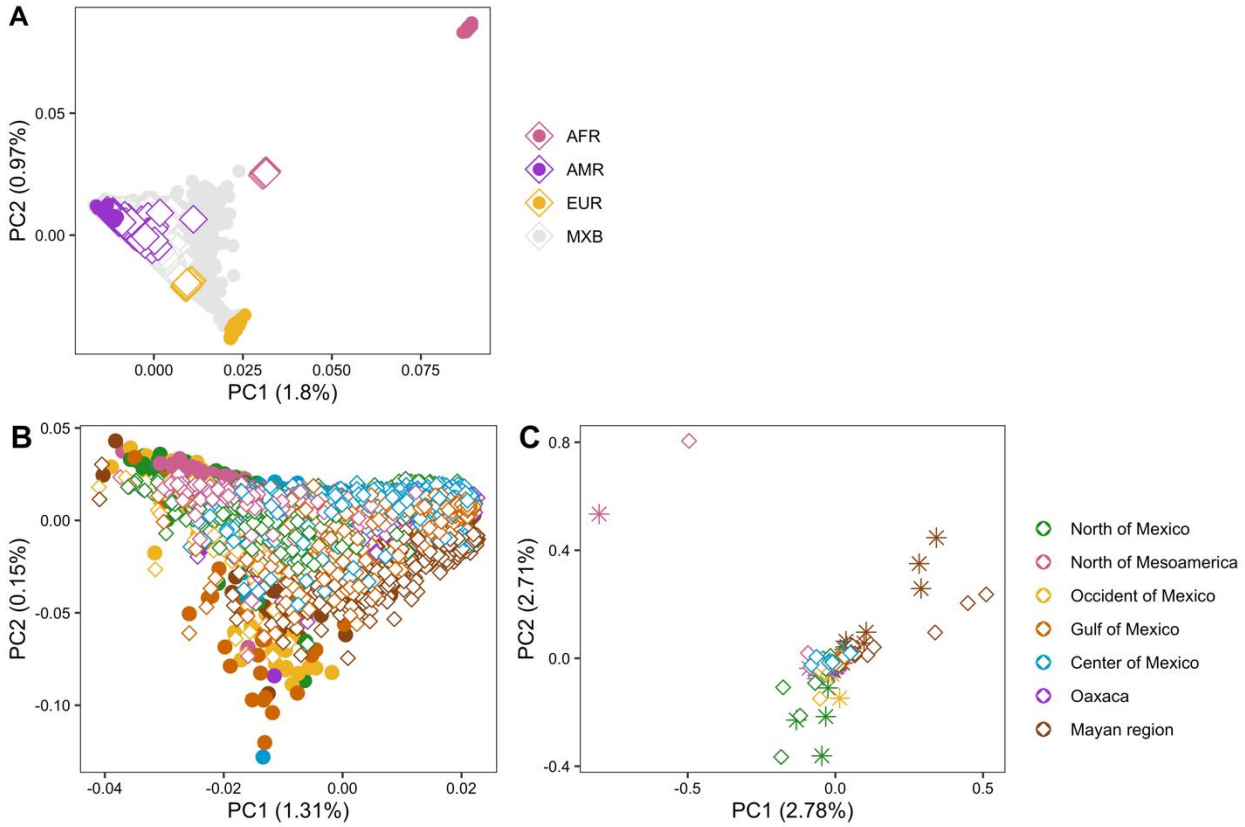


Figure 3. Principal component analysis (PCA) plots of a panel of 868,091 LD-pruned SNPs. Individuals are color-coded by population (archeological region). Dots represent the results obtained from the MXB-imputed data, rhombus represent the results obtained in Sohail et al., 2022 and asterisks represent the results obtained for the whole-genome sequence data from the 50 Indigenous MXB individuals. A) PCA of MXB

with global reference data from the 1000 Genomes Project. Note that Sohail et al., 2022 used a broader set of population references, and therefore PCs scores (rhombus) are slightly different to the ones presented in this thesis (dots). B) PCA of MXB only. C) PCA of Indigenous MXB.

I computed ancestry estimates for MXB-imputed using the same set of LD-pruned SNPs (n = 868,091). This analysis was intended to evaluate the ability of imputation to preserve genetic ancestry proxies by comparing the results with the results obtained in Sohail et al., 2022.

The correlation between global ancestry proportions inferred from ADMIXTURE (Alexander et al., 2009) in MXB-genotyped data and MXB-imputed data is greater than 0.99 for AFR, EUR and AMR ancestry estimates (Figure 4: A, B, C).

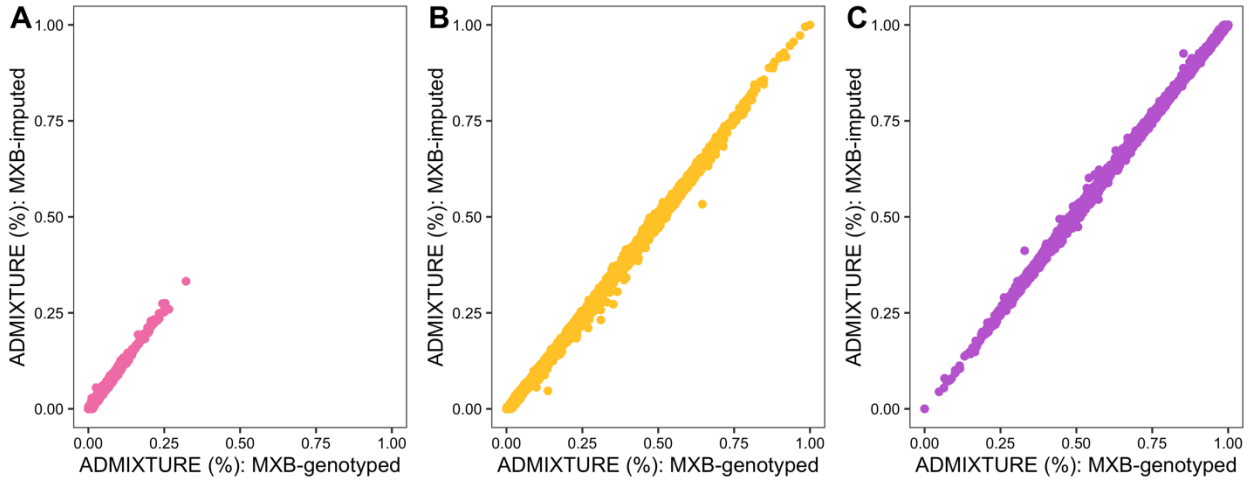


Figure 4. Concordance of ADMIXTURE global genetic ancestry estimates for MXB-genotyped data and MXB-imputed data. A) African genetic ancestries, B) European genetic ancestries, C) Indigenous American genetic ancestries.

Since ADMIXTURE results obtained from the MXB-imputed data are highly consistent with the results in Sohail et al., 2022, I used Gnomix (Hilmarsson et al., 2021) to compute global and local genetic ancestry estimates within MXB-imputed individuals. I summed the contribution of segments in each chromosome with a particular ancestry inferred by Gnomix to define a genome-wide ancestry proportion for each individual (Materials and Methods).

I observed that individuals in MXB-imputed are inferred to be admixed with varying degrees of AFR, EUR and AMR ancestry among states in Mexico. The central and

southern states have higher levels of Indigenous American ancestry. Oaxaca is the state with the highest proportion of Indigenous American ancestry. I also observed that the northern states have on average a greater proportion of European ancestry, and African ancestry tends to be low across all states (Figure 5).

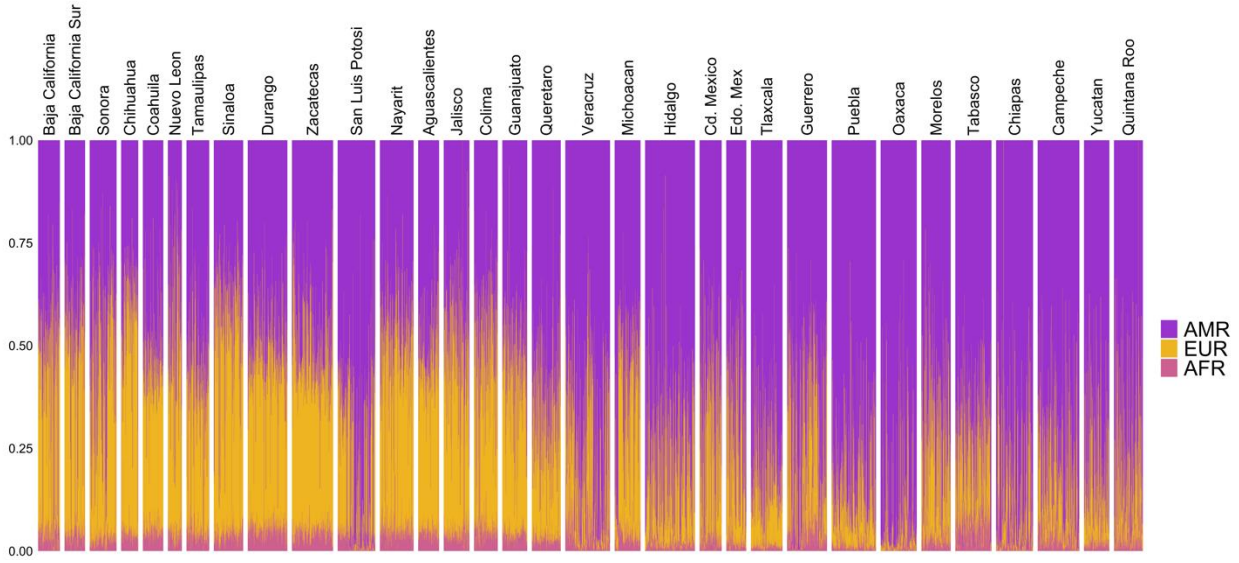


Figure 5. Global genetic ancestry estimates for MXB individuals using Gnomix. Each individual is represented by a vertical bar, and individuals are grouped by state. States are depicted from north to south. Indigenous American, European and African genetic ancestries are colored with pink, yellow and purple, respectively.

2.3. Detecting putative archaic-introgressed segments in present-day Mexico

I applied SPrime (Browning et al. 2018) to characterize the genomic landscape of archaic introgression in MXB-imputed. SPrime is a reference-free method that detects archaic-introgressed segments (i.e., sets of alleles) in the genome. I first apply SPrime to the 50 Indigenous MXB individuals that have whole-genome data and imputed data (Materials and methods). The purpose of this analysis is to evaluate the ability of the imputed data to recover the archaic-introgressed regions in Indigenous MXB individuals.

By applying SPrime to the autosomes of 50 Indigenous MXB sequenced genomes, I inferred 1,171 segments comprising a set of putatively archaic-introgressed SNPs in present-day Mexico. Additionally, I applied SPrime to the 50 Indigenous MXB-imputed genomes and inferred 1,075 regions containing putatively archaic-introgressed SNPs. I

used 108 YRI individuals from 1000 Genomes Project as an outgroup for both analyses (Materials and Methods).

I computed the Positive Predictive Value as the ratio of the detected variants that are present in both the real and imputed calls to the total number of detected variants in the imputed calls to compare the inferred archaic-introgressed variants in the 50 Indigenous MXB sequenced and imputed genomes. I observed that the Indigenous MXB-imputed genomes are able to recover 96% on average of the putatively archaic-introgressed SNPs detected in Indigenous MXB genomes (Figure 6: A). I also observed that the length of the introgressed-segments and the number of introgressed segments are highly concordant, and on average the segments identified in the sequenced genomes are larger (Figure 6: B).

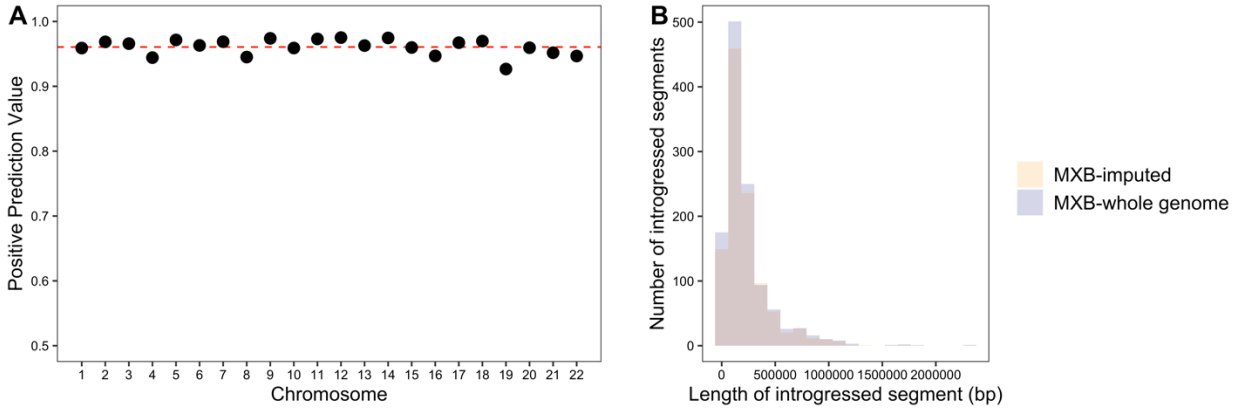


Figure 6. Putatively archaic-introgressed segments inferred in Indigenous MXB using YRI as an outgroup. A) Positive Prediction Value for each chromosome. Red dashed line represents the average value among chromosomes. B) Distribution of the length of introgressed segments.

These results demonstrate that the MXB-imputed data can recover the vast majority of archaic-introgressed variants in Indigenous MXB without adding more than 4% of spurious calls of archaic introgressed variants. This analysis provides evidence that the utilization of the MXB-imputed data is reliable for identifying archaic-introgressed segments in MXB.

Therefore, I applied SPrime to the autosomes of 5,833 MXB-imputed genomes using 108 YRI from 1000 Genomes Project as an outgroup. I inferred 3,117 segments comprising a total of 318,038 putatively archaic-introgressed SNPs. These regions are widely distributed across the autosomes. On average, these regions have a length of

222,183 bp that comprise about 102 SNPs (Figure 7: A). However, the largest introgressed segment is in chromosome 10 and has a size of 4,500,802 bp in 1,459 SNPs, while the smallest introgressed segment is in chromosome 19 and has a size of 1,994 bp in 12 SNPs.

Following Browning et al., I computed the match rate as the proportion of putative archaic alleles that match the archaic reference sequence. I used the genomes of the Altai Neanderthal and Altai Denisovan from Prüfer et al. (Prüfer et al., 2017) to eliminate regions that do not match the archaic references. The match rate reported is the proportion of matches for alleles that have enough coverage and mappability to be compared (Browning et al., 2018) (Materials and Methods).

The overall match rate to the sequenced Altai Neanderthal genome and to the Altai Denisovan genome is 0.53 and 0.22, respectively. The match rate distribution is visualized as a contour plot in Figure 8: A. The majority of the regions detected show a higher affinity to Neanderthal and low affinity to Denisovan. This is consistent with the observation of a higher rate of Neanderthal introgression compared to Denisovan introgression in Mexican in Los Angeles (MXL) from 1000 Genomes populations (Browning et al., 2018).

As recommended by Browning et al., I isolated regions with either Neanderthal or Denisovan ancestry by considering segments identified in MXB individuals that have at least 30 putatively introgressed variants that can be compared to the Altai Neanderthal genome or to the Altai Denisovan genome and have a match rate of at least 30% to the Altai Neanderthal allele or to the Altai Denisovan allele (Browning et al. 2017, McArthur et al. 2022).

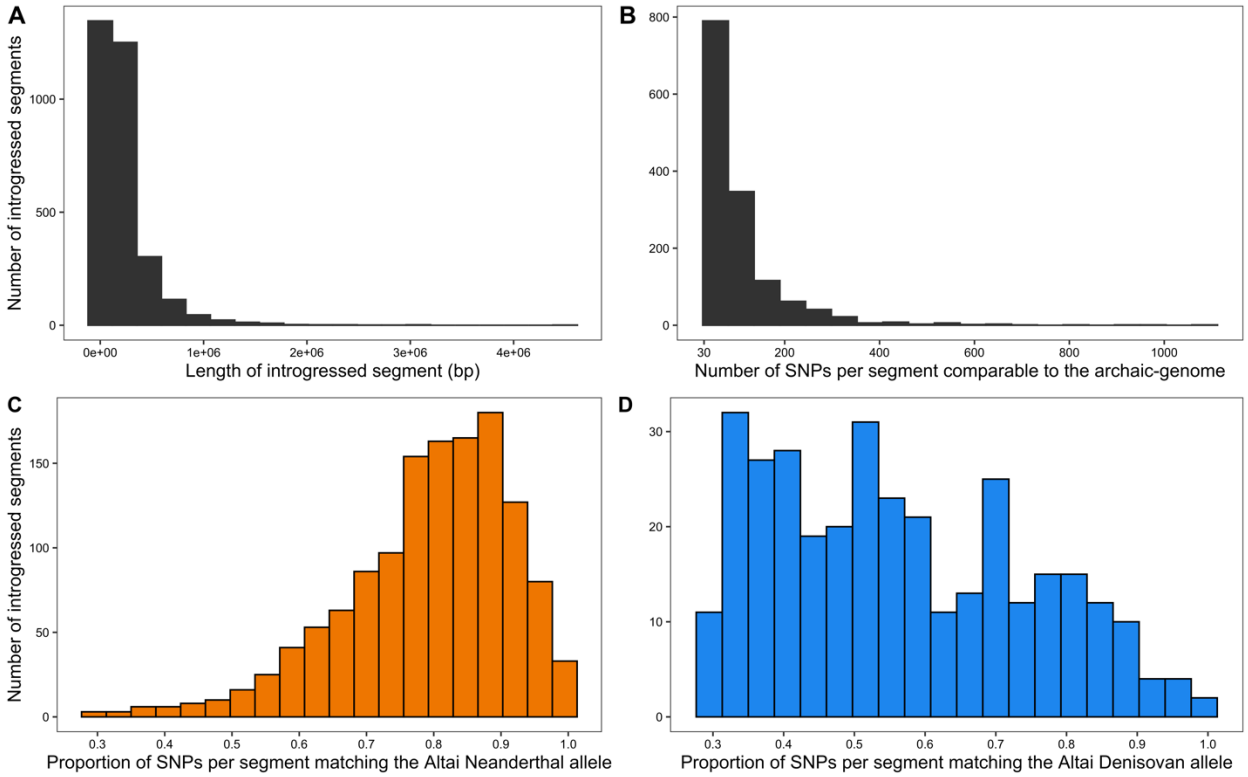


Figure 7. Putatively archaic-introgressed segments inferred in MXB using YRI as an outgroup. Results in B, C and D show results obtained after applying filters to isolate regions with Neanderthal and Denisovan ancestry. A) The distribution of the length of introgressed segments. B) The number of SNPs per segment that are comparable to the archaic genome. C) Proportion of SNPs per segment matching the Altai Neanderthal allele. D) Proportion of SNPs per segment matching the Altai Denisovan allele.

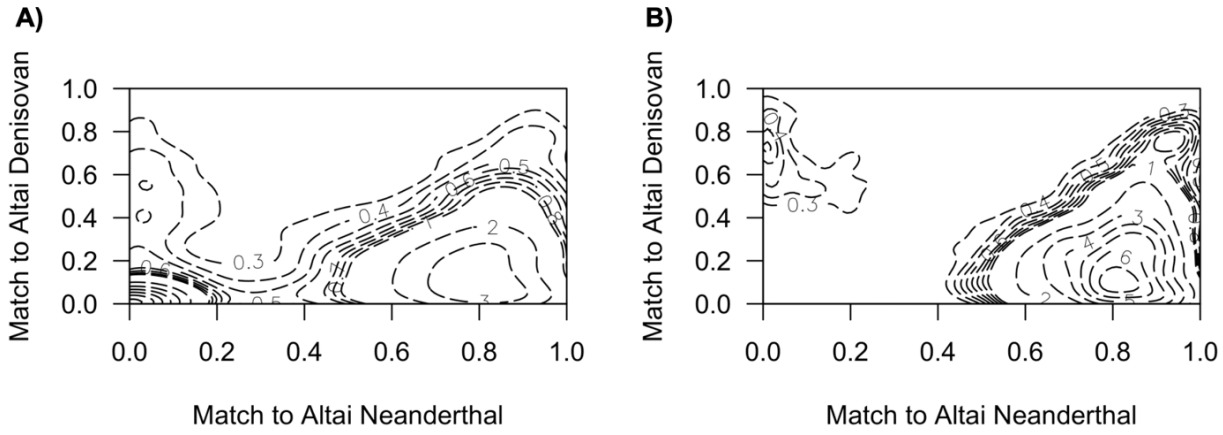


Figure 8. Contour density plot of match proportion of introgressed segments to the Altai Neanderthal and to the Altai Denisovan genomes. For a given segment, a match rate of 0 denotes that for that segment, none of the alleles match the corresponding allele in the archaic-human reference. A) Density distribution

of match rate to archaic individuals (Neanderthal and Denisovan). B) Density distribution of match rate to archaic individuals after isolating regions with Neanderthal and Denisovan ancestry.

After applying these two filters to the segments identified in MXB-imputed, I recovered 1,411 high-quality segments comprising a total of 146,875 putatively archaic-introgressed SNPs (Figure 7: B). The overall match rate of the high-quality segments to the sequenced Altai Neanderthal genome and to the Altai Denisovan genome is 0.79 and 0.56, respectively (Figure 7: C, D). The match rate distribution is visualized as a contour plot in Figure 8: B.

2.4. The distribution of archaic-introgressed variants across states in present-day Mexico

I used the 1,441 segments with isolated Altai Neanderthal and Altai Denisovan ancestry identified by SPrime in MXB-imputed to investigate the distribution patterns of archaic-introgressed variants in Mexico. To do this, I calculated the number of sites with Neanderthal or Denisovan alleles (archaic SNPs), as well as the number of sites with Neanderthal-specific or Denisovan-specific alleles for each individual in MXB-imputed (Witt et al., 2023) (Materials and Methods). To examine how the number of sites with archaic alleles varies among states in Mexico, I grouped the individuals depending on their sampling location to be representative of the 32 states in Mexico. I found that individuals in Oaxaca have the greatest number of sites with archaic alleles (12,083 on average), followed by Puebla (12,006 on average). Individuals in Chihuahua have the fewest number of sites with archaic alleles (10,333 on average) (Figure 9).

The same is true for Neanderthal and Denisovan-specific alleles. Oaxaca is the state with the greatest number of sites with archaic alleles (7,422 and 658 on average) and Chihuahua is the state with fewest number of sites with archaic alleles (6,487 and 104 on average) (Supplementary Figure 3, 4). All states have fewer Denisovan-specific variants than Neanderthal-specific variants, which is consistent with the distribution of the proportion of SNPs per segment that match the Altai Neanderthal allele or the Altai Denisovan allele (Figure 7: C, D).

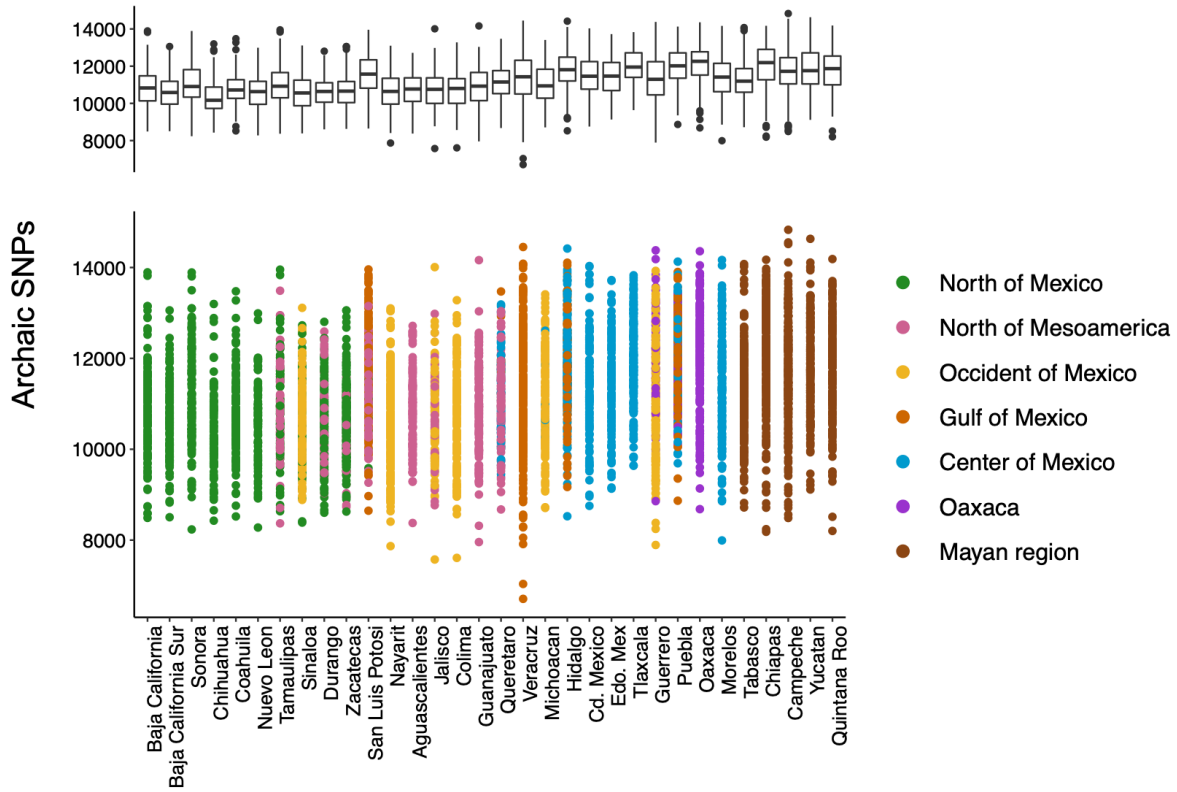


Figure 9. The number of sites with archaic alleles (either Neanderthal, Denisovan) found in 1,411 segments detected by SPrime in 5,833 MXB individuals. Individuals are grouped according to their sampling location. Color coding is by archeological region. Boxplots represent the distribution of the number of sites with archaic alleles for each state.

Witt et al. show that there is a correlation between the number of archaic-alleles and the proportions of African, European and American ancestry in admixed populations (Witt et al., 2023). Based on that observation, I decided to compute the correlation between the amount of AFR, EUR and AMR ancestry and 1) the total number of archaic sites for each individual, and 2) the total number of archaic sites for each state in Mexico (Figure 10). As expected, at a global (all individuals) level, there is a positive correlation between Indigenous American ancestry and the total number of sites with archaic alleles (r^2 = 0.70), and a negative correlation between African and European ancestry and the number of sites with archaic alleles (r^2 = -0.53 and -0.66, respectively). This is consistent with previous work that has shown that Yorubas from 1000 Genomes Project are expected to have no traces of Neanderthal or Denisovan ancestry. The negative correlation for European ancestry could be explained by the fact that admixture from Europe due to

Spanish colonization has diluted the archaic-introgressed signal in MXB (Witt el al., 2023). The same is true at a per state-level, r^2 = -0.76, -0.96 and 0.98 for AFR, EUR and AMR.

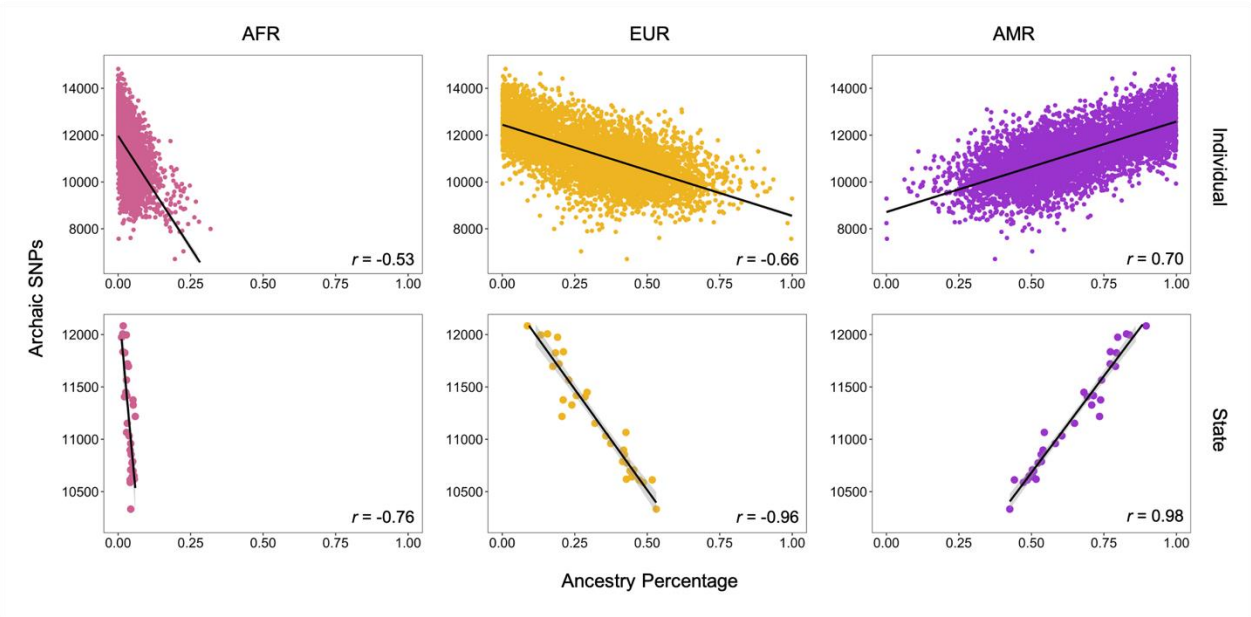


Figure 10. The correlation between percentages of African (AFR), European (EUR) and American (AMR) genetic ancestries and the number of sites with archaic alleles. The X-Axis is the genetic ancestry proportion for a given ancestry type (AFR as pink, EUR as yellow and AMR as purple). The Y-Axis is the number of sites with archaic alleles. The first row represents results at individual level and the second row represents results by grouping individuals by sampling location to account for states in Mexico.

I computed a One-Way Analysis of Variance (ANOVA) to test if there is a significant difference between the average number of archaic alleles according to sampling location. The p-value was < $2e^{-16}$, which leads to conclude that there are significant differences between states. I performed a multiple pairwise-comparison to detect if the mean difference between specific pairs of states is statistically significant. I used the *pairwise.t.test()* function in R with Bonferroni correction for multiple testing (Figure 11).

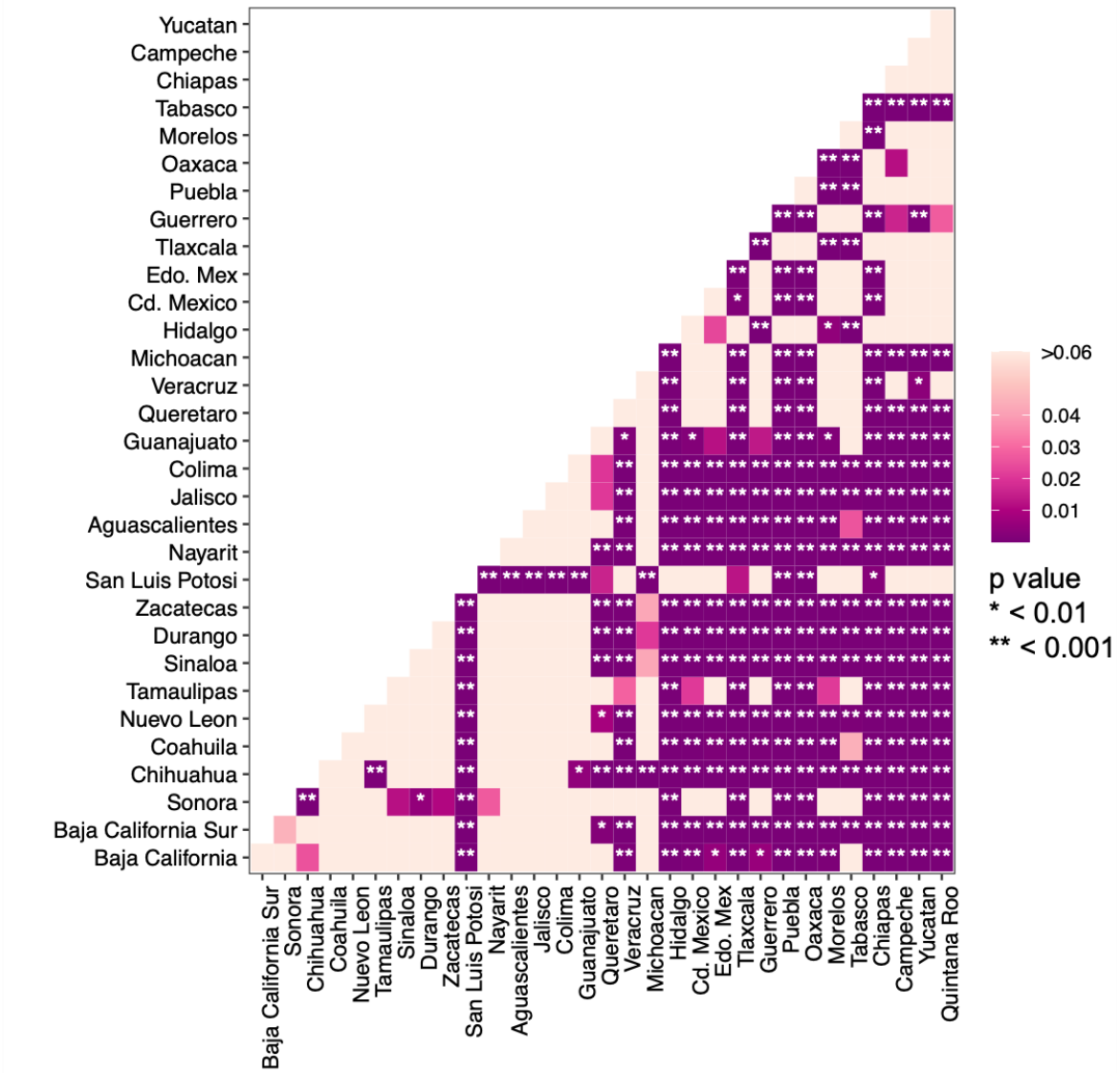


Figure 11. Heatmap representation of the p-values of multiple pairwise-comparison of the average number of sites with archaic alleles by state in Mexico. States are depicted from north to south. Significance levels are expressed by asterisks (* for p value < 0.01 and ** for p value < 0.001).

I found that there is a statistically significant difference between the number sites with archaic alleles in Oaxaca and the rest of states in Mexico apart from Hidalgo, Tlaxcala and Puebla. Interestingly, the majority of the pairs that have a statistically significant difference involve a comparison between one state located in the northern part of Mexico and another state located in the southern part of Mexico.

To further understand the distribution of archaic counts in MXB-imputed in contrast to other admixed populations, I computed the number of sites with archaic alleles for all

individuals from MXL, PEL, Colombians in Medellin (CLM) and Puerto Ricans in Puerto Rico (PUR) from the 1000 Genomes Project populations. I used the putative archaic-introgressed segments identified by Browning et al (Browning et al. 2018) and restricted them to have at least 30 putatively introgressed variants that can be compared to the Altai Neanderthal or Denisovan genome and to have a match rate of at least 30% to the Altai Neanderthal or Denisovan allele (Browning et al. 2017, McArthur et al. 2022).

I observed that the number of sites with archaic alleles (either Neanderthal or Denisovan) in MXB-imputed is significantly different (*p-value* < 0.05) from the four tested 1000 Genomes American populations but PEL (*p-value* 0.094). This can be explained by the fact that the genome-wide proportions of AFR, EUR and AMR ancestry for MXB-imputed and PEL are similar. On average, MXL individuals have 0.038, 0.47 and 0.46 of AFR, AMR and EUR ancestry while PEL individuals have 0.007, 0.77 and 0.19 of AFR, AMR and EUR ancestry (Witt et al., 2023 using data from Martin et al., 2017). However, MXB-imputed individuals have 0.03, 0.63 and 0.31 of AFR, AMR and EUR ancestry. I observed that the number of sites with Neanderthal-specific alleles in MXB-imputed is similar to MXL and PEL and significantly different from CLM and PUR. This could be explained by the fact that MXL and PEL have a greater proportion of Indigenous American Ancestry compared to CLM and PUR (Witt et al., 2022). Interestingly, the number of sites with Denisovan-specific alleles in MXB-imputed is highly variable (min: 57, median: 558, max: 1088) (Figure 12). This can in part be explained by the fact that MXB individuals have variable proportions of Indigenous American and European ancestry (Figure 5).

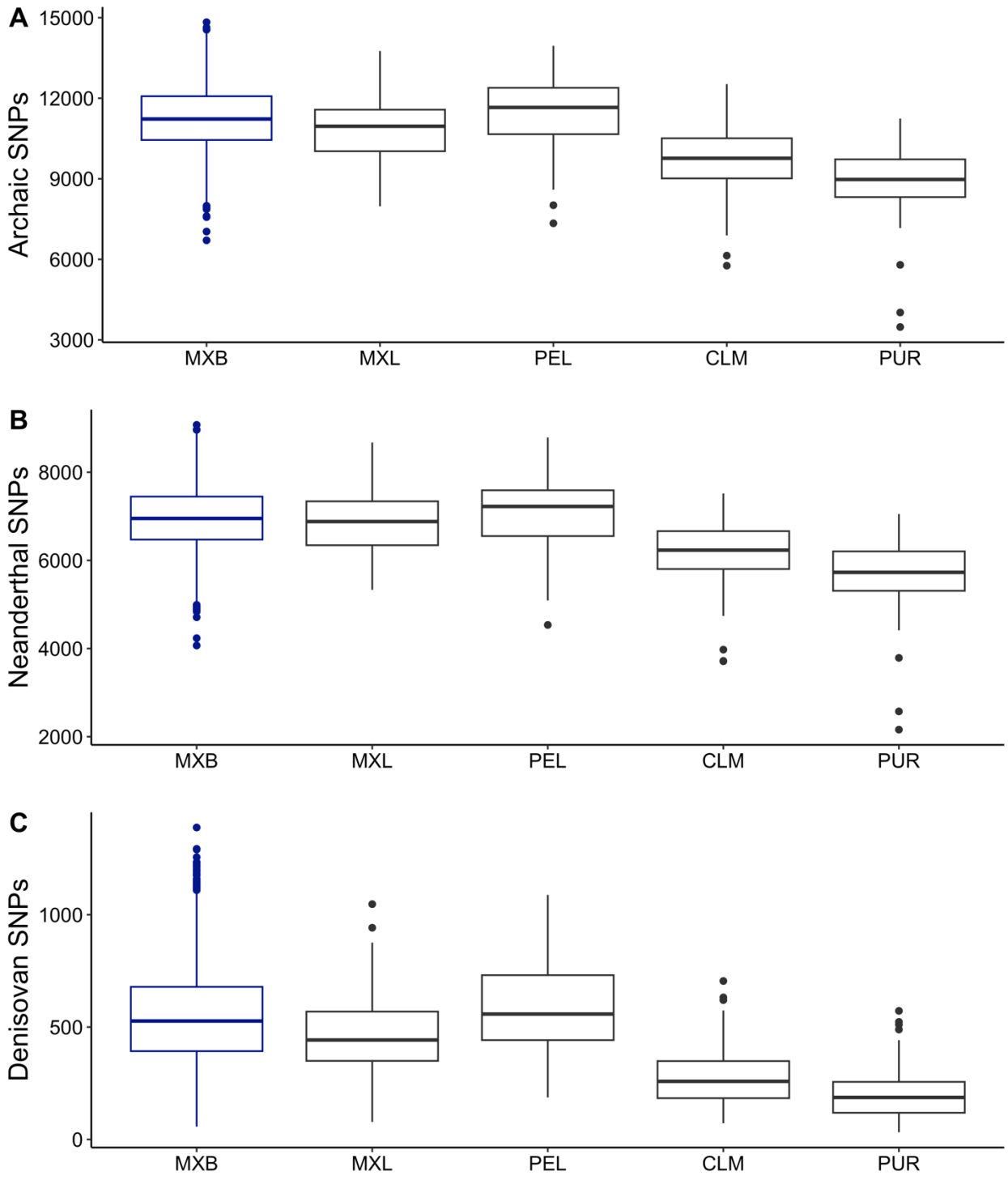


Figure 12. Distribution of the number of sites with archaic alleles for all MXB individuals and all individuals from MXL, PEL, CLM and PUR in 1000 Genomes Project. Color-coding is for MXB (blue) and Admixed American populations (black). A) Archaic SNPs (either Neanderthal or Denisovan) B) Neanderthal-specific. C) Denisovan-specific.

3. Discussion

Previous studies have revealed that present-day non-African human populations have inherited about 1 to 4% of their genetic ancestry from introgression with archaic individuals (Green et al., 2010, Reich et al., 2010). Here, I used the imputed genomes of 5,833 present-day individuals broadly distributed across Mexico to address a series of questions regarding the accuracy of imputed data and the distribution of archaic introgression across Mexico. My work presents a rigorous pipeline to evaluate the performance of imputation on genotyping array data from Mexico and provides evidence on the reliability of the MXB-imputed genomes to conduct genetic analyses. Additionally, I provide a set of putative archaic-introgressed alleles in each introgressed segment detected in MXB, as well as a set of archaic-variants that are likely to be from Neanderthal or Denisovan origin.

Neanderthal-match rates are consistent among MXB-imputed and Mexican in Los Angeles (MXL) from 1000 Genomes Populations (Browning et al. 2018). Interestingly, MXB-imputed have a higher proportion of Denisovan archaic-introgressed segments compared to MXL. This could be explained by the fact that MXB-imputed genomes have more Indigenous American ancestry than MXL, and therefore the archaic-introgression signal in MXB-imputed has not been diluted by admixture as in MXL (Villanea et al., 2022, Witt et al., 2023). Further research needs to be done to test this hypothesis.

Previous work from Witt et al. showed that there is a positive correlation between Indigenous American ancestry and the total number of sites with archaic alleles (archaic allele counts) (Witt et al., 2023). Since individuals sampled from the Mexican Biobank Project had a sample bias towards the representation of Indigenous American ancestry (Sohail et al., 2022), they have on average 20% more of Indigenous American ancestry than MXL. This suggests that the complex admixture processes that MXL individuals faced have diluted the contribution from archaic introgressed alleles (Villanea et al., 2022).

Furthermore, Witt et al. demonstrated that European ancestry is negatively correlated with archaic allele counts in populations with high proportions of American ancestry (Witt et al., 2023). Here, I observed the same pattern as individuals who have more European ancestry had fewer sites with archaic alleles.

I provide insights into the distribution of archaic introgressed variants across states in Mexico. I demonstrated that there is a statistically significant difference in the number of sites with archaic alleles in individuals from the northern part of Mexico and the southern part of Mexico. This is explained by the fact that individuals from the northern part of Mexico have an increased proportion of European ancestry, which is consistent with previous work (Martínez-Cortés et al., 2012, Moreno-Estrada et al., 2014, Sohail et al., 2022). This suggests that the signal of archaic introgression may vary according to the geographical location in Mexico.

I acknowledge that I have made several assumptions and choices in my work. First, I rely on the imputed data of MXB individuals for doing the archaic-introgression analysis. Even though several methods have been developed for inferring archaic introgression in present-day individuals, they all rely on whole-genome sequencing data (Browning et al., 2018, Skov et al., 2018). However, underrepresented populations usually lack sequenced data (Bentley et al., 2017), such that there is a necessity of imputing genotyped data for increasing power to perform population genetic analyses (Jiménez-Kaufmann et al., 2022). I conducted several analyses to demonstrate the reliability of the MXB-imputed data. I recognize that the panel I used is restricted to data of Indigenous MXB individuals that may not be able to fully capture the genetic variation of Mexico. Second, I assumed MXB-imputed genomes are grouped according to their sampling location to be representative of the 32 states in Mexico. I acknowledge that the genetic makeup of a specific sampling location may not accurately represent the genetic composition of the individuals present at the sampling location.

Overall, I inferred a map of putatively Neanderthal and Denisovan introgressed segments identified in present-day Mexico. This thesis provides the first rigorous basis for detecting archaic introgression in present-day Mexico and it is a valuable resource for future studies to investigate the history of archaic-introgression events in Mexico. These archaic-introgressed haplotype maps can be used for elucidating the complex landscape of archaic introgression events in Mexico, as well as to investigate the phenotypic effects of archaic-introgressed variants in Mexican individuals.

4. Materials and methods

4.1. The Mexican Biobank Project

To investigate the genomic landscape of archaic introgression in present-day Mexico, I employed the Mexican Biobank Project (MXB) Dataset. This dataset was obtained through a collaboration agreement with CINVESTAV (MXB project lead: Dr. Moreno-Estrada) for all the analyses conducted here. This dataset gathers genotyping information of 6,057 individuals from 32 states and 898 sampling localities in Mexico (Sohail et al., 2022) (Figure 1). These individuals were recruited as part of the National Health Survey in 2000 (ENSA2000), which sampled around 40,000 participants across Mexico. In order to select individuals for genomic characterization, the MXB Team selected those individuals that can speak an Indigenous language in each state. They also maximize the representation of rural localities to account for representation of Indigenous American ancestry. The MXB is 70% female and comprises individuals born between 1910 and 1980, all sampled in 2000 (Sohail et al. 2022). All individuals in MXB were genotyped at ~1.8 million SNPs using Illumina's Multi-Ethnic Global Array (MEGA) (Sohail et al. 2022).

In addition to the genotyping data, I used whole-genome sequencing data generated by the MXB Team from 50 individuals with the highest proportion of Native American ancestry of the 6,057 individuals genotyped (Jiménez-Kaufmann et al. 2022).

The genotyping data can be used to perform whole-genome imputation. To do this, the MXB Team used the Trans-Omics for Precision Medicine (TOPMed) Imputation Reference panel on GRCh38 (Taliun et al. 2021) through the Michigan Imputation Server. As a result, genomic imputation was performed on the genotyped information of all 6,057 individuals, which resulted in the availability of additional genetic information (imputed genomes) that can enhance genetic analysis. It must be emphasized that the 50 Indigenous MXB individuals that had whole-genome sequencing data also contain whole-genome imputed information based on data from the genotyping array.

The MXB-imputed data was generated using a reference file on the latest genome assembly, GRCh38, whereas both the genotyped and whole-genome data were mapped to the previous assembly, GRCh37.

I used a bioinformatic technique to convert genomic information between reference builds to convert the imputed data from GRCh38 to GRCh37. This conversion involved the use of CrossMap (Zhao et al., 2014) to convert genome coordinates between assemblies.

Reference fasta files for GRCh37 were downloaded from The Broad Data Resources and the chain file for GRCh38 was downloaded using Ensembl FTP.

Table 1 shows the number of variants in the initial files, the number of variants that are called in GRCh38 but are not present in GRCh37 (unmapped variants) and the percentage of variants that are called in GRCh38 and are present in GRCh37 (mapped variants).

Chromosome	**Total Variants**	**Failed to Map**	**% of Mapped Variants**
1	23,7347,13	730,781	96.92
2	25,543,692	141,999	99.44
3	20,864,167	151,554	99.27
4	20,250,718	93,761	99.54
5	18,909,563	83,963	99.56
6	17,626,672	304,181	98.27
7	16,990,285	632,800	96.28
8	16,305,107	125,947	99.23
9	13,263,350	320,395	97.58
10	14,212,906	577,340	95.94
11	14,563,388	423,812	97.09
12	13,986,366	137,257	99.02
13	10,435,997	30,227	99.71
14	9,280,421	142,905	98.46
15	8,667,566	77,497	99.11
16	9,840,732	134,102	98.64

17	8,494,641	599,642	92.94
18	8,253,586	3,906	99.95
19	6,497,881	164,595	97.47
20	6,631,675	148,960	97.75
21	3,829,414	43,448	98.87
22	4,140,642	144,354	96.51

Table 1. LiftOver Performance of converting the MXB-imputed data in GRCh38 to the GRCh37 genome assembly.

As CrossMap (Zhao et al., 2014) disrupts the order of mapped variants, I used the bcftools (Danecek et al., 2021) command *sort* to rearrange the MXB-imputed files in a consistent manner.

4.2. Imputation performance analysis

I conducted a genomic imputation performance analysis to validate the reliability of the MXB-imputed genomes. Doing so involves comparing the imputed genotypes to actual genotypes from whole-genome sequences for a subset of the data. I used the intersection between the SNP positions of the Indigenous MXB whole-genome data and the MXB-imputed data for the same 50 individuals.

4.2.1. Data pre-processing

The MXB-imputed markers are classified into three groups: 1) *imputed*, indicating that a marker was imputed but not genotyped, 2) *typed*, indicating that a marker was both genotyped and imputed and 3) *typed_only*, to indicate that a marker was genotyped but not imputed.

I restricted the imputation performance analysis to markers that are *imputed*. I extracted the 50 Indigenous MXB individuals from the MXB-imputed dataset. Additionally, I filtered out multi-allelic sites, indels, structural variants and duplicated sites in all autosomes. I also excluded monomorphic sites. Thus, I generated an Indigenous MXB-imputed dataset comprising 6,359,016 bi-allelic SNPs. Similarly, for the Indigenous MXB whole-genome sequencing data, I removed multi-allelic sites, indels, structural variants in

all autosomes. I filtered out variants with missing call rates exceeding 0.05. The resulting files comprise 9,088,970 bi-allelic SNPs.

I computed the intersection between the SNPs position in the whole-genome sequencing dataset and the Indigenous MXB-imputed dataset. I named this intersection as the *shared sites* between the whole-genome and the imputed data for Indigenous MXB. I subsetted both the filtered Indigenous MXB whole-genome sequencing dataset and the Indigenous MXB-imputed dataset to contain only the shared sites between them. The data-processing was done using *bcftools* (Danecek et al., 2021) version 1.9, *vcftools* (Danecek et al., 2011) version 1.16 and *plink2* (Chang et al., 2015).

4.2.2. Analysis

I employed the *read_vcf()* function from the scikit-allel (Miles et al., 2021) python package to compare the Indigenous MXB-imputed genotypes to the actual Indigenous MXB genotypes. I extracted genotype calls from the Variant Call Format (VCF) files for both the imputed and whole-genome dataset.

Subsequently, I computed four statistics: genotype correlation, genotype accuracy, heterozygotic precision, and homozygotic alternative precision, which were defined as: the correlation coefficient between both genotype vectors, the ratio of correctly called genotypes to the number of called genotypes, the ratio of correctly called heterozygotes to the number of called heterozygotes, and the ratio of correctly called homozygotes alternate to the number of homozygotes alternate, respectively. I computed each statistic for each SNP in the dataset.

Additionally, the MXB-imputed markers have a parameter called R^2 to reflect the imputation performance (Fuchsberger et al., 2015). I defined an imputation quality threshold of $R^2 = \{0.05, 0.1, 0.2, 0.3\}$. I removed variants in the panel that had a R^2 value smaller than the four quality thresholds and estimated the four-imputation metrics using the retained variants. I computed the mean value for each imputation metric based on the retained variants for each of the four imputation quality thresholds.

4.3. Population structure analysis for MXB-imputed data

I performed population structure analyses to ensure that the MXB-imputed data accurately maintains the genetic diversity of MXB. The purpose of these analyses is to compare the obtained results with the ones obtained for the MXB-genotyped data in Sohail et al., 2022.

For these and all the following analyses, I removed 224 individuals from the cohort that were categorized as related in Sohail et al. 2022. Therefore, the MXB-imputed panel comprises 5,833 unrelated samples.

4.3.1. Target

I created a qc-ed version of the MXB-imputed data by keeping all sites with imputation quality greater than 0.05 ($R^2 > 0.05$). I removed duplicated sites and restricted our analysis to bi-allelic sites in all autosomes. I also applied a Hardy-Weinberg equilibrium (HWE) filter of $1e^{-8}$ to account for batch effects, and a minor-allele frequency (MAF) filter of 0.005 to remove rare variants. The qc-ed version of the MXB-imputed data has 9,219,234 SNPs and 5,833 samples.

4.3.2. References

I used reference samples for Africa (AFR), Europe (EUR) and America (AMR) from the 1000 Genomes Project (1KGP) dataset. I selected 60 Yorubas (YRI) from Ibadan Nigeria as AFR, 60 Iberians (IBS) from Spain as EUR, and 27 Peruvians from Lima with more than 96% ancestry from the Americas as AMR. I removed indels and structural variants, multi-allelic sites and duplicated sites. I applied a HWE filter of $1e^{-3}$ to remove batch effects and a MAF filter of 0.005 to remove rare variants. The qc-ed reference dataset has 15,821,810 variants and 147 samples.

4.3.3. Merged File

For the analysis of population structure, I merged the qc-ed MXB-imputed dataset and the qc-ed reference dataset using the bcftools (Danecek et al., 2021) *merge* function. This merged file is an intersection of the qc-ed MXB-imputed SNP positions and the SNP positions from the qc-ed reference dataset. The merged file comprises 8,129,767 SNPs and 5,980 samples.

I used the *plink2* (Chang et al., 2015) command --indep-pairwise to produce a pruned subset of variants that are in approximate linkage disequilibrium (LD) with each

other. I created a LD-pruned version of the merged file by removing pairs of SNPs that have an LD value greater than 0.1 in windows of 50 SNPs. The LD-pruned merged dataset comprises 868,091 SNPs and 5980 samples.

4.3.4. Principal component analysis (PCA)

I performed two analyses of principal components (PCA). One was performed on the LD-pruned merged dataset including all the samples (MXB-imputed and references), and the second was performed on the same dataset but including only MXB-imputed genomes.

I used smartpca from eigensoft (Patterson et al., 2006, Price et al., 2006) software to obtain both PCAs.

4.3.5. ADMIXTURE

I employed the ADMIXTURE (Alexander et al., 2009) software version 1.3 to estimate individual ancestry proportions for K = {3,4,5,6} number of clusters. I used the LD-pruned merged dataset to run ADMIXTURE.

4.3.6. Local ancestry inference

To estimate local ancestry along the MXB-imputed genomes, I used Gnomix (Hilmarsson et al., 2021). I trained Gnomix using the qc-ed reference dataset containing only the shared SNPs positions between the qc-ed MXB-imputed dataset and the reference dataset. I applied the default settings since they are optimal for whole-genome data and allowed Gnomix to re-phase the genomes using the predicted-local ancestry. I employed the trained model to infer local ancestry tracts for the MXB-imputed genomes.

I calculated genome-wide global ancestry proportions using a script created by myself. This script is based on an Alicia Martin script (Martin et al. 2017) to compute global ancestry proportions for a specific chromosome.

4.3.7. Principal component analysis (PCA) for Indigenous MXB

To further validate the reliability of the MXB-imputed data, I repeated the principal component analysis but considering only the 50 Indigenous MXB individuals that have imputed and whole-genome data.

I extracted the 50 Indigenous MXB individuals from the LD-pruned merged dataset to apply smartpca from the eigensoft (Patterson et al., 2006, Price et al., 2006) software.

For the Indigenous MXB whole-genome data, I limited the dataset to have bi-allelic SNPs only in all autosomes. Then, I restricted the dataset to comprise only the SNPs positions that are present in the LD-pruned merged dataset. 351,877 variants are present in the whole genome-dataset and the LD-pruned merged dataset. I applied smartpca on the restricted dataset with 351,877 variants.

4.4. Detecting archaic-introgressed segments in Mexico using SPrime

I applied SPrime (Browning et al. 2018) to infer archaic-introgressed segments on MXB. The SPrime software detects variations in a target present-day population that are introgressed from an archaic source in the past. I used SPrime (Browning et al., 2018) since it is optimized for detecting introgression from Neanderthals and Denisovans in modern populations and is more accurate than previous methods (Browning et al., 2018). SPrime is reference-free and it is able to detect archaic introgression by comparing the target population to an outgroup. The outgroup must be a population that is not expected to contain introgressed variants. I used the 108 YRI from Ibadan Nigeria in the 1000 Genomes Project dataset because they are thought to have no direct admixture from Neanderthals (Green et al. 2010).

The SPrime software requires a Variant Call Format (VCF) file with genotypes for all autosomes and samples. Both target and outgroup samples must be in the input VCF file.

4.4.1. SPrime on Indigenous MXB

I tested SPrime performance on the Indigenous MXB individuals that have both imputed and whole-genome data.

For the Indigenous-MXB whole genome data, I created a target dataset that contains bi-allelic sites in all autosomes and no monomorphic sites. I removed SNPs with a missing genotype rate below 0.05.

For the Indigenous MXB-imputed data, I created a target dataset by removing all markers with imputation quality less or equal to 0.05. I restricted the dataset to bi-allelic

sites in all autosomes and removed all monomorphic sites. I applied a Hardy-Weinberg equilibrium (HWE) filter of $1e^{-8}$ to account for batch effects.

For the outgroup, I extracted the 108 Yoruba (YRI) samples from the 1000 Genomes Project (1KGP). I removed multi-allelic sites and indels or structural variants.

Then, I merged the target and the outgroup datasets using the bcftools merge function. I named the merged files as 1) Indigenous MXB-Real Genomes + YRI and 2) Indigenous MXB-Imputed Genomes + YRI. Both merged files are the intersection of the target SNP positions and the SNP positions from the outgroup dataset. The Indigenous MXB-Real Genomes + YRI dataset has 6,972,338 variants and the Indigenous MXB-Imputed Genomes + YRI has 6,517,652 variants.

4.4.2. SPrime on MXB-imputed

For the MXB-imputed genomes, I created a dataset containing only bi-allelic sites in all autosomes and no monomorphic sites.

For the outgroup, I extracted the 108 Yoruba (YRI) samples from the 1000 Genomes Project and removed multi-allelic sites and indels or structural variants.

I computed the intersection between the MXB-imputed SNP positions and the outgroup SNP positions to obtain the shared sites between datasets. I subsetted both datasets to contain only the shared sites between them and merged them using the bcftools (Danecek et al., 2021) *merge* function. I named the merged file as MXB-Imputed Genomes + YRI. This file comprises 31,355,288 SNPs.

4.4.3. SPrime description

SPrime (Browning et al., 2018) requires a PLINK format genetic map with cM units to estimate genetic positions between map positions. I downloaded the HapMap genetic maps in GRCh37 and concatenated them to have a whole-genome PLINK format genetic map.

I runned SPrime in the three different merged datasets: 1) the Indigenous MXB-Real Genomes + YRI dataset, 2) the Indigenous MXB-Imputed Genomes + YRI dataset, and 3) the MXB-Imputed Genomes + YRI dataset.

4.4.4. Calculate match rates to a sequenced archaic-genome

Since SPrime (Browning et al., 2018) is a reference-free method, it is possible to use a relevant archaic genome that has been sequenced to map the detected archaic-variants to the archaic genome of interest to confirm the source of introgression. I used the genome of the Altai Neanderthal and the genome of the Altai Denisovan from Prüfer et al. 2017 to represent two possible sources of archaic introgression.

Each variant detected by SPrime can be mapped to the archaic genome, resulting in a match, mismatch or not comparable to the archaic genome. Each state represents that the detected variant is present in the archaic genome, is not present in the archaic genome and is not comparable to the archaic genome because genotype quality is low for that locus or it has poor mappability. To do this, I used the *map_arch* script created by Y Zhou's script (Zhou et al., 2021). This script reads in SPrime output and returns the match status for each detected variant.

After obtaining the match status for each archaic-introgressed variant detected by SPrime, I calculated the match rate for each reported introgression segment (i.e., sets of alleles). The match rate for each segment detected by SPrime is the number of matching positions divided by the sum of matching and mis-matching positions. Match rate is undefined if there are no variants that can be comparable to the archaic genome in the segment (i.e., all variants in the segment are non-comparable to the archaic genome). This is helpful to detect variants that are likely to be false-positives. (Browning et al., 2018)

I modified Y Zhou's script to compute the match rate to obtain the number of total variants detected by the segment, and the number of variants matching the Neanderthal or Denisovan genomes.

Contour plots were created using Y Zhou's script and the *MASS* package in R.

4.5. Genomic regions with archaic ancestry

To define genomic regions with Neanderthal or Denisovan ancestry, I used the segments identified by the SPrime run on the MXB-Imputed Genomes + YRI dataset. To isolate regions with Neanderthal or Denisovan ancestry, I considered segments that have 1) at least 30 putatively archaic-introgressed variants that could be comparable to the Altai

Neanderthal or Altai Denisovan genome and 2) had a match rate of at least 30% to the Altai Neanderthal or Altai Denisovan allele. (Browning et al. 2018) I used this stringent set of archaic-matching segments for the following analysis.

4.5.1. Distribution of archaic-introgression in Mexico

To better understand the distribution of archaic-introgression across Mexico, I count the number of archaic-variants that each individual in MXB has. For doing so, I used the *archaic_snps_perind_sprime.py* script created by Kelsey-Witt (Witt et al., 2023) to count the number of sites with archaic-SNPs, and the total number of archaic SNPs for each individual, using the SPrime calls to define archaic sites.

Witt's script uses the SPrime archaic calls to define archaic SNPs and counts the total number of archaic SNPs for each individual. This script can examine different sets of archaic SNPs. I analyzed archaic-SNPs detected by SPrime that were Altai Neanderthal unique, Altai Denisovan unique and either Altai Neanderthal or Altai Denisovan (all archaic alleles).

I used MXB panel information to group individuals by state to better understand the distribution of the total number of archaic SNPs for individuals across Mexico.

I used the genome-wide ancestry proportions to compute the correlation between percentages of African (AFR), European (EUR), and American (AMR) ancestry and count of sites with archaic alleles.

All scripts needed to reproduce this work are deposited on GitHub:
https://github.com/vagaribay/archaic-segments-MXB.

5. References

1. Alexander, David H., et al. “Fast Model-Based Estimation of Ancestry in Unrelated Individuals.” *Genome Research*, vol. 19, no. 9, Sept. 2009, pp. 1655–64. *DOI.org (Crossref)*, https://doi.org/10.1101/gr.094052.109.

2. Bentley, Amy R., et al. “Diversity and Inclusion in Genomic Research: Why the Uneven Progress?” *Journal of Community Genetics*, vol. 8, no. 4, Oct. 2017, pp. 255–66. *DOI.org (Crossref)*, https://doi.org/10.1007/s12687-017-0316-6.

3. Browning, Sharon R., et al. “Analysis of Human Sequence Data Reveals Two Pulses of Archaic Denisovan Admixture.” *Cell*, vol. 173, no. 1, Mar. 2018, pp. 53-61.e9. *DOI.org (Crossref)*, https://doi.org/10.1016/j.cell.2018.02.031.

4. Chang, Christopher C., et al. “Second-Generation PLINK: Rising to the Challenge of Larger and Richer Datasets.” *GigaScience*, vol. 4, no. 1, Dec. 2015, p. 7. *DOI.org (Crossref)*, https://doi.org/10.1186/s13742-015-0047-8.

5. Danecek, Petr, et al. “The Variant Call Format and VCFtools.” *Bioinformatics*, vol. 27, no. 15, Aug. 2011, pp. 2156–58. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btr330.

6. Danecek, Petr, et al. “Twelve Years of SAMtools and BCFtools.” *GigaScience*, vol. 10, no. 2, Jan. 2021, p. giab008. *DOI.org (Crossref)*, https://doi.org/10.1093/gigascience/giab008.

7. Fuchsberger, Christian, et al. “Minimac2: Faster Genotype Imputation.” *Bioinformatics*, vol. 31, no. 5, Mar. 2015, pp. 782–84. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btu704.

8. Green, Richard E., et al. “A Complete Neandertal Mitochondrial Genome Sequence Determined by High-Throughput Sequencing.” *Cell*, vol. 134, no. 3, Aug. 2008, pp. 416–26. *DOI.org (Crossref)*, https://doi.org/10.1016/j.cell.2008.06.021.

9. Gower, Graham, et al. “Detecting Adaptive Introgression in Human Evolution Using Convolutional Neural Networks.” *ELife*, edited by George H Perry and Diego Ortega Del Vecchyo, vol. 10, May 2021, p. e64669. *eLife*, https://doi.org/10.7554/eLife.64669.

10. Green, Richard E., et al. “A Draft Sequence of the Neandertal Genome.” *Science*, vol. 328, no. 5979, May 2010, pp. 710–22. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1188021.

11. Hammer, Michael F., et al. “Genetic Evidence for Archaic Admixture in Africa.” *Proceedings of the National Academy of Sciences*, vol. 108, no. 37, Sept. 2011, pp. 15123–28. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.1109300108.

12. Hilmarsson, Helgi, et al. *High Resolution Ancestry Deconvolution for Next Generation Genomic Data*. bioRxiv, 21 Sept. 2021. *bioRxiv*, https://doi.org/10.1101/2021.09.19.460980.

13. Huerta-Sánchez, Emilia, et al. “Altitude Adaptation in Tibetans Caused by Introgression of Denisovan-like DNA.” *Nature*, vol. 512, no. 7513, Aug. 2014, pp. 194–97. *DOI.org (Crossref)*, https://doi.org/10.1038/nature13408.

14. Jiménez-Kaufmann, Andrés, et al. “Imputation Performance in Latin American Populations: Improving Rare Variants Representation With the Inclusion of Native American Genomes.” *Frontiers in Genetics*, vol. 12, 2022. *Frontiers*, https://www.frontiersin.org/articles/10.3389/fgene.2021.719791.

15. Koller, Dora, et al. “Denisovan and Neanderthal Archaic Introgression Differentially Impacted the Genetics of Complex Traits in Modern Populations.” *BMC Biology*, vol. 20, no. 1, Nov. 2022, p. 249. *BioMed Central*, https://doi.org/10.1186/s12915-022-01449-2.

16. Martin, Alicia R., et al. “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations.” *American Journal of Human Genetics*, vol. 100, no. 4, Apr. 2017, pp. 635–49. *PubMed Central*, https://doi.org/10.1016/j.ajhg.2017.03.004.

17. Martínez-Cortés, Gabriela, et al. “Admixture and Population Structure in Mexican-Mestizos Based on Paternal Lineages.” *Journal of Human Genetics*, vol. 57, no. 9, Sept. 2012, pp. 568–74. *www.nature.com*, https://doi.org/10.1038/jhg.2012.67.

18. McArthur, Evonne, et al. “Quantifying the Contribution of Neanderthal Introgression to the Heritability of Complex Traits.” *Nature Communications*, vol. 12, no. 1, July 2021, p. 4481. *www.nature.com*, https://doi.org/10.1038/s41467-021-24582-y.

19. Meyer, Matthias, et al. “A High-Coverage Genome Sequence from an Archaic Denisovan Individual.” *Science*, vol. 338, no. 6104, Oct. 2012, pp. 222–26. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1224344.

20. Miles, Alistair, et al. *Cggh/Scikit-Allel: V1.3.3*. Zenodo, 14 May 2021. *Zenodo*, https://zenodo.org/record/4759368.

21. Moreno-Estrada, Andrés, et al. “The Genetics of Mexico Recapitulates Native American Substructure and Affects Biomedical Traits.” *Science*, vol. 344, no. 6189, June 2014, pp. 1280–85. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1251688.

22. Patterson, Nick, et al. “Population Structure and Eigenanalysis.” *PLoS Genetics*, vol. 2, no. 12, 2006, p. e190. *DOI.org (Crossref)*, https://doi.org/10.1371/journal.pgen.0020190.

23. Plagnol Vincent, and Jeffrey D. Wall. “Possible Ancestral Structure in Human Populations.” *PLoS Genetics*, edited by Anna Di Rienzo, vol. 2, no. 7, July 2006, p. e105. *DOI.org (Crossref)*, https://doi.org/10.1371/journal.pgen.0020105.

24. Price, Alkes L., et al. “Principal Components Analysis Corrects for Stratification in Genome-Wide Association Studies.” *Nature Genetics*, vol. 38, no. 8, Aug. 2006, pp. 904–09. *DOI.org (Crossref)*, https://doi.org/10.1038/ng1847.

25. Prüfer, Kay, et al. “A High-Coverage Neandertal Genome from Vindija Cave in Croatia.” *Science (New York, N.Y.)*, vol. 358, no. 6363, Nov. 2017, pp. 655–58. *PubMed Central*, https://doi.org/10.1126/science.aao1887.

26. Prüfer, Kay, et al. “The Complete Genome Sequence of a Neanderthal from the Altai Mountains.” *Nature*, vol. 505, no. 7481, Jan. 2014, pp. 43–49. *www.nature.com*, https://doi.org/10.1038/nature12886.

27. Racimo, Fernando, et al. “Evidence for Archaic Adaptive Introgression in Humans.” *Nature Reviews Genetics*, vol. 16, no. 6, June 2015, pp. 359–71. *DOI.org (Crossref)*, https://doi.org/10.1038/nrg3936.

28. Ragsdale, Aaron P., et al. “A Weakly Structured Stem for Human Origins in Africa.” *Nature*, vol. 617, no. 7962, May 2023, pp. 755–63. *www.nature.com*, https://doi.org/10.1038/s41586-023-06055-y.

29. Sankararaman, Sriram, et al. “The Combined Landscape of Denisovan and Neanderthal Ancestry in Present-Day Humans.” *Current Biology*, vol. 26, no. 9, May 2016, pp. 1241–47. *DOI.org (Crossref)*, https://doi.org/10.1016/j.cub.2016.03.037.

30. Sankararaman, Sriram, et al. “The Genomic Landscape of Neanderthal Ancestry in Present-Day Humans.” *Nature*, vol. 507, no. 7492, Mar. 2014, pp. 354–57. *DOI.org (Crossref)*, https://doi.org/10.1038/nature12961.

31. Sawyer, Susanna, et al. “Nuclear and Mitochondrial DNA Sequences from Two Denisovan Individuals.” *Proceedings of the National Academy of Sciences*, vol. 112, no. 51, Dec. 2015, pp. 15696–700. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.1519905112.

32. Skov, Laurits, et al. “Detecting Archaic Introgression Using an Unadmixed Outgroup.” *PLOS Genetics*, edited by Fernando Racimo, vol. 14, no. 9, Sept. 2018, p. e1007641. *DOI.org (Crossref)*, https://doi.org/10.1371/journal.pgen.1007641.

33. Sohail, Mashaal, et al. *Nationwide Genomic Biobank in Mexico Unravels Demographic History and Complex Trait Architecture from 6,057 Individuals*. bioRxiv, 14 July 2022. *bioRxiv*, https://doi.org/10.1101/2022.07.11.499652.

34. Taliun, Daniel, et al. “Sequencing of 53,831 Diverse Genomes from the NHLBI TOPMed Program.” *Nature*, vol. 590, no. 7845, Feb. 2021, pp. 290–99. *www.nature.com*, https://doi.org/10.1038/s41586-021-03205-y.

35. The 1000 Genomes Project Consortium, et al. “A Global Reference for Human Genetic Variation.” *Nature*, vol. 526, no. 7571, Oct. 2015, pp. 68–74. *DOI.org (Crossref)*, https://doi.org/10.1038/nature15393.

36. The SIGMA Type 2 Diabetes Consortium. “Sequence Variants in SLC16A11 Are a Common Risk Factor for Type 2 Diabetes in Mexico.” *Nature*, vol. 506, no. 7486, Feb. 2014, pp. 97–101. *DOI.org (Crossref)*, https://doi.org/10.1038/nature12828.

37. Vernot, Benjamin, and Joshua M. Akey. “Resurrecting Surviving Neandertal Lineages from Modern Human Genomes.” *Science*, vol. 343, no. 6174, Feb. 2014, pp. 1017–21. *DOI.org (Crossref)*, https://doi.org/10.1126/science.1245938.

38. Villanea, Fernando A., and Joshua G. Schraiber. “Multiple Episodes of Interbreeding between Neanderthal and Modern Humans.” *Nature Ecology & Evolution*, vol. 3, no. 1, Nov. 2018, pp. 39–44. *DOI.org (Crossref)*, https://doi.org/10.1038/s41559-018-0735-8.

39. Villanea, Fernando A., and Kelsey E. Witt. “Underrepresented Populations at the Archaic Introgression Frontier.” *Frontiers in Genetics*, vol. 13, Feb. 2022, p. 821170. *DOI.org (Crossref)*, https://doi.org/10.3389/fgene.2022.821170.

40. Wall, Jeffrey D., et al. “Detecting Ancient Admixture and Estimating Demographic Parameters in Multiple Human Populations.” *Molecular Biology and Evolution*, vol. 26, no. 8, Aug. 2009, pp. 1823–27. *DOI.org (Crossref)*, https://doi.org/10.1093/molbev/msp096.

41. Wei, Xinzhu, et al. “The Lingering Effects of Neanderthal Introgression on Human Complex Traits.” *ELife*, edited by Graham Coop et al., vol. 12, Mar. 2023, p. e80757. *eLife*, https://doi.org/10.7554/eLife.80757.

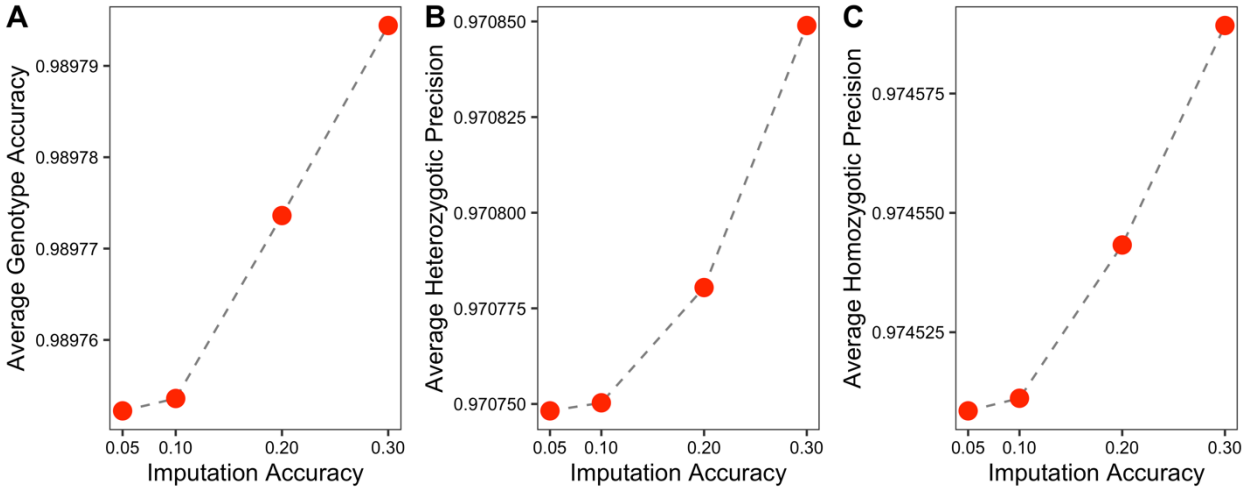
42. Witt, Kelsey E., et al. “The Impact of Modern Admixture on Archaic Human Ancestry in Human Populations.” *Genome Biology and Evolution*, edited by David Enard, vol. 15, no. 5, May 2023, p. evad066. *DOI.org (Crossref)*, https://doi.org/10.1093/gbe/evad066.

43. Zhang, Xinjun, et al. “The History and Evolution of the Denisovan- *EPAS1* Haplotype in Tibetans.” *Proceedings of the National Academy of Sciences*, vol. 118, no. 22, June 2021, p. e2020803118. *DOI.org (Crossref)*, https://doi.org/10.1073/pnas.2020803118.

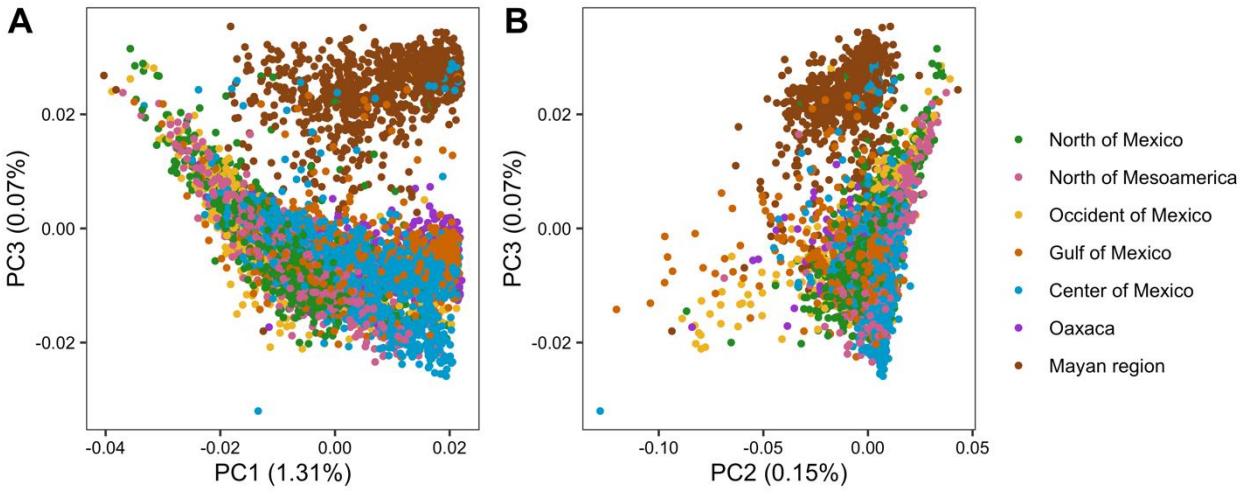
44. Zhao, Hao, et al. “CrossMap: A Versatile Tool for Coordinate Conversion between Genome Assemblies.” *Bioinformatics*, vol. 30, no. 7, Apr. 2014, pp. 1006–07. *DOI.org (Crossref)*, https://doi.org/10.1093/bioinformatics/btt730.

45. Zhou, Ying, and Sharon R. Browning. “Protocol for Detecting Introgressed Archaic Variants with SPrime.” *STAR Protocols*, vol. 2, no. 2, June 2021, p. 100550. *DOI.org (Crossref)*, https://doi.org/10.1016/j.xpro.2021.100550.

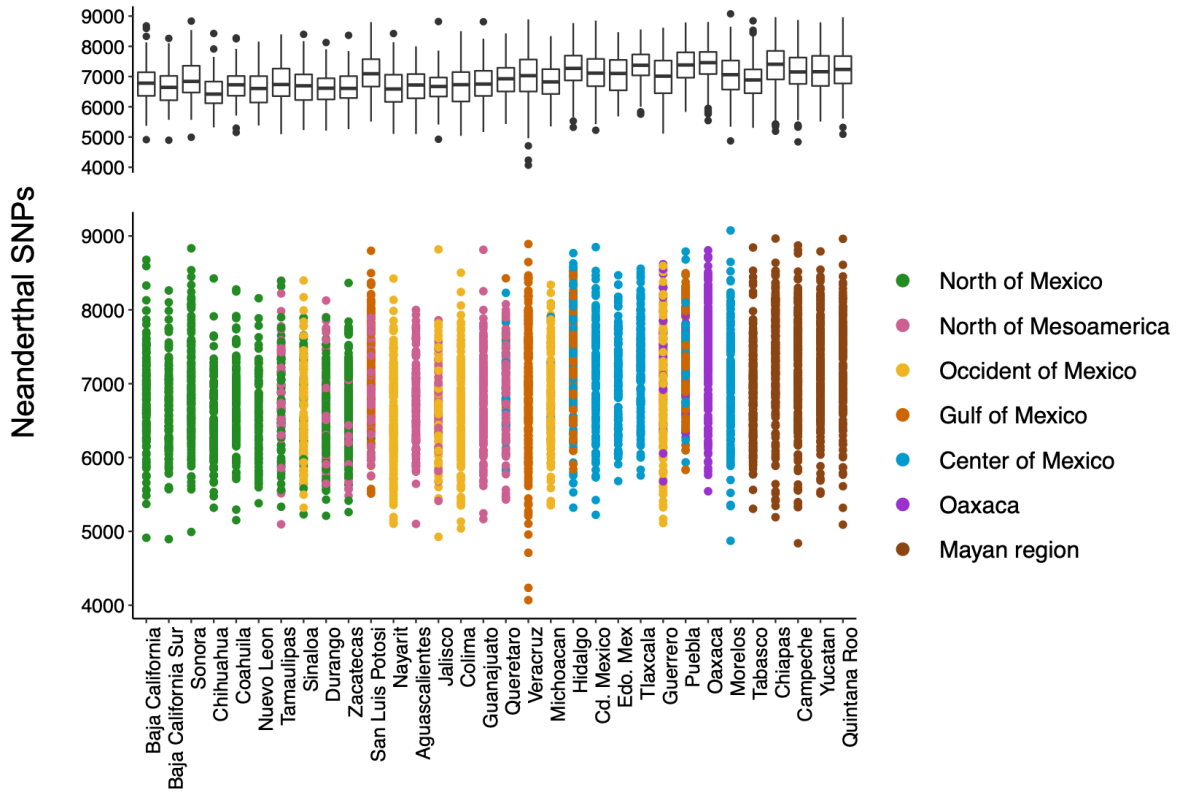
6. Supplementary Figures



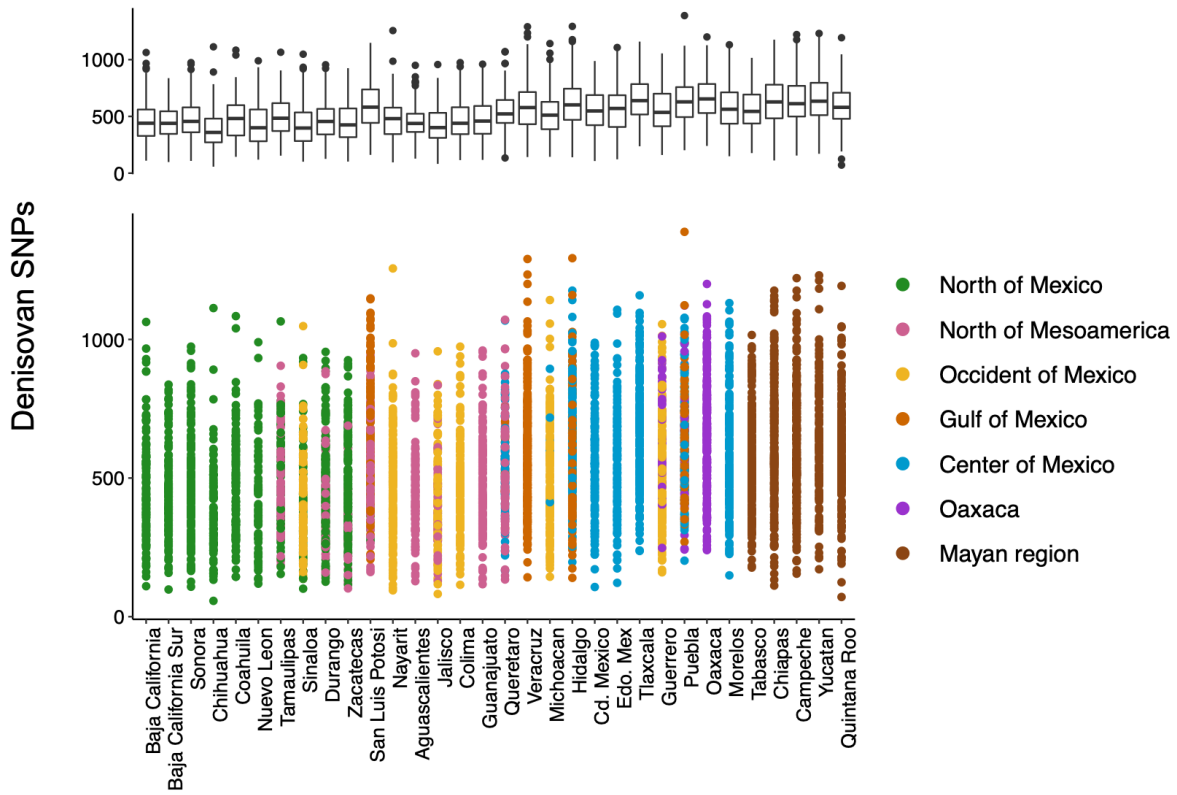
Supplementary Figure 1. Relationship between imputation accuracy, R^2, and the averaged metric for all SNPs in all autosomes. A) Genotype accuracy. B) Heterozygotic precision. C) Homozygotic precision.



Supplementary Figure 2. Principal component analysis (PCA) plots of a panel of 868,091 LD-pruned SNPs. Individuals are color-coded by population (archeological region). Results shown are from MXB-imputed data. A) PC1 vs PC3. B) PC2 vs PC3.



Supplementary Figure 3. The number of sites with Altai Neanderthal-specific alleles found in 1,411 segments detected by SPrime in 5,833 MXB individuals. Individuals are grouped according to their sampling location. Color coding is by archeological region. Boxplots represent the distribution of the number of sites with archaic alleles for each state.



Supplementary Figure 4. The number of sites with Altai Denisovan-specific alleles found in 1,411 segments detected by SPrime in 5,833 MXB individuals. Individuals are grouped according to their sampling location. Color coding is by archeological region. Boxplots represent the distribution of the number of sites with archaic alleles for each state.