# The *MUC19* gene in Denisovans, Neanderthals, and Modern Humans: An Evolutionary History of Recurrent Introgression and Natural Selection

**Authors:** Fernando A. Villanea[1,†], David Peede[2,3,4,†], Eli J. Kaufman[5], Valeria Añorve-Garibay[3,6], Elizabeth T. Chevy[3], Viridiana Villa-Islas[6], Kelsey E. Witt[7], Roberta Zeloni[8], Davide Marnetto[8], Priya Moorjani[9,10], Flora Jay[11], Paul N. Valdmanis[5], María C. Ávila-Arcos[6], Emilia Huerta-Sánchez[2,3,12,*]

**Affiliations:**

1) Department of Anthropology, University of Colorado Boulder.

2) Department of Ecology, Evolution, and Organismal Biology, Brown University.

3) Center for Computational Molecular Biology, Brown University.

4) Institute at Brown for Environment and Society, Brown University.

5) Division of Medical Genetics, Department of Medicine, University of Washington School of Medicine.

6) International Laboratory for Human Genome Research, Universidad Nacional Autónoma de México.

7) Center for Human Genetics and Department of Genetics and Biochemistry, Clemson University.

8) Department of Neurosciences “Rita Levi Montalcini”, University of Turin.

9) Department of Molecular and Cell Biology, University of California, Berkeley,

10) Center for Computational Biology, University of California, Berkeley.

11) Université Paris-Saclay, CNRS, INRIA, Laboratoire Interdisciplinaire des Sciences du Numérique, 91400, Orsay, France.

12) Smurfit Institute of Genetics, Trinity College Dublin, Dublin 2, Ireland.

† These authors contributed equally to this work

**Abstract:**

We study the gene MUC19, for which modern humans carry a Denisovan-like haplotype. MUC19 is a mucin, a glycoprotein that forms gels with various biological functions. We find the diagnostic variants for the Denisovan-like MUC19 haplotype at high frequencies in admixed Latin American individuals among global populations, and at highest frequency in 23 ancient Indigenous American individuals, all predating population admixture with Europeans and Africans. We find that the Denisovan-like MUC19 haplotype carries a higher copy number of a 30 base-pair variable number tandem repeat, and that copy numbers of this repeat are exceedingly high in American populations and are under positive selection. This study provides the first example of positive selection acting on archaic alleles at coding sites and VNTRs. Finally, we find that some Neanderthals carry the Denisovan-like MUC19 haplotype, and that it was likely introgressed into human populations through Neanderthal introgression rather than Denisovan introgression.

**One-Sentence Summary:** Modern humans and Neanderthals carry a Denisovan variant of the *MUC19* gene, which is under positive selection in populations of Indigenous American ancestry.

**Main Text:**

It is widely accepted that most modern humans of non-African ancestry carry both Neanderthal and Denisovan genomic variants [*1–3*]. While most of these variants are putatively neutral, some archaic variants found in modern humans have been targets of positive natural selection [*4–9*]. Interbreeding with Neanderthals and Denisovans may have thereby facilitated adaptation to the myriad of novel environments that modern humans encountered as they populated the globe [10]. Indeed, several studies have identified signatures of adaptive introgression in Eurasian and Oceanian populations [*11–20*]. Indigenous American populations, however, present the greatest potential for studying the underlying evolutionary processes of local adaptation, as they are descendants of individuals who populated the American continent [*21*]. In the 25,000 years since, these populations would have encountered manifold novel environments, far different from the Beringian steppe, to which their ancestral population was adapted [*22*].

In a previous study, we computed the Population Branch Statistic (*PBS*, [*23*]) using SNPs within archaic introgressed tracts in admixed populations from the Americas to identify targets of adaptive introgression. We found the region surrounding *MUC19*—a gene involved in immunity—harbors several Denisovan variants with outlying *PBS* scores in Mexicans (MXL) from the 1000 Genomes Project (TGP), suggesting that the archaic alleles are at high frequency in Mexicans [*24*]. Earlier studies had reported that this region has one of the largest densities of Denisovan alleles in Mexicans [*25*], and *MUC19* was also reported to be under positive selection in North American Indigenous populations using *PBS* and integrated Haplotype Scores (*iHS*—a haplotype-based method for detecting positive selection [*26*]).

In this study, we confirm and further characterize signatures of both introgression and positive selection at *MUC19* in MXL. Notably, we find an archaic haplotype segregating at high frequency in most populations on the American continent, which is also present in two of the late high-coverage Neanderthal genomes—Chagyrskaya and Vindija. MXL individuals harbor Denisovan-specific coding mutations in *MUC19* at high frequencies, and exhibit elevated copy number of a tandem repeat region within *MUC19* compared to other worldwide populations. Our results point to a complex pattern of multiple introgression events, from Denisovans to Neanderthals, and Neanderthals to modern humans, which may have played a unique role in the evolutionary history of Indigenous American populations.

## Results

### Signatures of adaptive introgression at *MUC19* in admixed populations from the Americas

We compiled introgressed tracts that overlap the NCBI RefSeq coordinates for *MUC19* (hg19, Chr12:40787196-40964559) by at least one base pair. Figure 1A shows the density of introgressed tracts for all non-African populations in the region, using introgression maps

inferred with `hmmix` [*27*]. All non-African populations harbor introgressed tracts overlapping this region, but at much lower frequencies than the admixed populations from the Americas (AMR tract frequency: ~18.3%, Non-AMR tract frequency: ~8.7%; Proportions *Z*-test, *P-value*: 5.011e-14; Fisher's Exact Test, *P-value*: 2.144e-12; Table S1); MXL exhibits the highest frequency of the introgressed tracts (~30%; Table S2). Given this result, we then took a 742kb window containing the longest introgressed tract found in Mexicans (hg19, Chr12:40272001-41014000; Figure S1), a population with a large component of Indigenous American genetic ancestry (~48%; [*28*]). This region contains 135 Denisovan-specific SNPs: rare or absent in African populations (<1%), present in MXL (>1%), and shared uniquely with the Altai Denisovan. Remarkably, all 135 of these SNPs are sequestered within a core 72kb region (hg19, Chr12:40759001-40831000; shaded gray region in Figure 1A) that has the highest introgressed tract density amongst individuals in the TGP (see Methods), making both the 742kb and 72kb region outliers for Denisovan-specific SNP density in MXL (742kb region *P-value:* <3.164e-4; 72kb region *P-value:* <3.389e-5; Figure S2; Table S3-S4). In contrast, there are 80 Neanderthal-specific SNPs in MXL found within the larger 742kb region (*P-value:* 0.159; Figure S3; Table S5), with only four located in the 72kb region (*P-value:* 0.263; Figure S3; Table S6).

To test if natural selection is acting on this region, we computed three statistics; one that has been developed to detect adaptive introgression ($U_{A,B,C}(w, x, y)$, *A:* African super population, *B:* non-African populations, *C:* Altai Denisovan; *(w, x, y)* are allele frequency thresholds in *A*, *B* and *C*), and two for positive selection (*PBS*, and *iHS*). For each gene, we computed $U_{AFR,B,Denisovan}(w=1\%, x=30\%, y=100\%)$, which measures the number of Denisovan alleles found in the homozygous state (100%) that are almost absent in Africans (<1%) and reach a frequency of at least 30% in a given non-African population. Figure 1B shows that *MUC19* in MXL is an extreme outlier, as no other gene in any non-African population exhibits such a large value of $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$. When we compute the same statistic in windows instead of per gene, the *MUC19* region is an outlier only in MXL and is zero for all other non-African populations (*P-value* 72kb region: <3.3284e-5; *P-value* 742kb region: <3.139e-4: Figure S4; Table S7-S8). Furthermore, we compared the windowed $U_{AFR,B,Denisovan}(w=1\%, x=30\%, y=100\%)$ results with their corresponding $Q95_{AFR,B,Denisovan}(w=1\%, y=100\%)$ value, which quantifies the 95th percentile of the Denisovan allele frequencies found in a given non-African population *B* for the Denisovan alleles found in the homozygous state (100%), that are almost absent in Africans (<1%), we find that for both the 72kb and 742kb *MUC19* regions that $Q95_{AFR,MXL,Denisovan}(w=1\%, y=100\%) = ~30\%$, which suggests that both the 72kb and 742kb *MUC19* regions exhibit signals consistent with adaptive introgression that are not observed in any other TGP population (Figure S5-S6; Table S9). We next computed $PBS_{MXL:CHB:CEU,}$ where the Han Chinese (CHB) and Central European (CEU) populations were used as control populations, for both the region corresponding to the longest introgressed tract in MXL—742kb—and the 72kb region in *MUC19,* and find that both regions exhibit statistically significant $PBS_{MXL:CHB:CEU}$ values compared to other 742kb ($PBS_{MXL:CHB:CEU}$: 0.066; *P-value:* 0.004) and 72kb ($PBS_{MXL:CHB:CEU}$: 0.127; *P-value:* 0.002) windows of the genome respectively (see Methods; Figure S7; Table S10-S11). We then computed $PBS_{MXL:CHB:CEU}$ for each SNP in the 742kb region. Figure 1C shows that in MXL there are many SNPs with statistically significant PBS values in that region (417 out of 6144 SNPs), all which present values above the 99.95th percentile of genome-wide $PBS_{MXL:CHB:CEU}$ values (Benjamini-Hochberg corrected *P-values:* <0.01; see Supplementary Section S1). We note that some SNPs have a larger $PBS_{MXL:CHB:CEU}$ value near the *SLC2A13* gene than within the 72kb *MUC19* region, but this is

due to changes in the archaic allele frequency in CHB and CEU, as the introgressed tracts in these populations are more sparse than the introgressed tracts in MXL (see tracts in Figure 1C). When we partition the MXL population into two demes, consisting of individuals with more than 50% and those with less than 50% Indigenous American ancestry genome-wide [28], and recompute *PBS*, we find that *PBS* values for archaic variants are elevated among individuals with a higher proportion of Indigenous American ancestry, suggesting that this region was likely targeted by selection before admixture with European and African populations (Figure S8; Table S10-S11).

To exclude the possibility that demographic events such as a founder effect explain the observed signatures of positive selection, we simulated the best fitting demographic parameters inferred for the MXL population [29] to obtain the expected null distribution of *PBS* values under this demographic model. We first showed that *PBS* has power to detect adaptive introgression under this demographic model (see Supplementary Section S1). We found that demographic forces alone result in lower *PBS* values compared to what is observed at this gene region (see Supplementary Section S1), even when we consider a very conservative null model of heterosis, which assumes deleterious mutations are recessive. Furthermore, to also consider haplotype-based measures of positive selection, we computed the integrated haplotype score (*iHS*) for every TGP population using `selscan` [30] to provide a haplotype-based evidence of natural selection (see Methods). Among all TGP populations, MXL is the only population with an elevated proportion of SNPs with normalized |*iHS* scores| > 2 in either the 742kb (599 out of 2248 SNPs) or 72kb region (229 out of 425 SNPs; see Supplemental Section S2; Table S12-S13). Notably, in MXL we find that 130 out of the 135 Denisovan-specific SNPs in the 72kb region have normalized |*iHS* scores| > 2, reflective of positive selection (Figure S9; Table S12-S13, see Supplementary Section S2), which corroborates all of our previous allele frequency-based tests of natural selection. In summary, we find that *MUC19* in MXL exhibits signals consistent with positive selection, through the use of both allele frequency and haplotype tests of selection as previously suggested [8, 26].

## Admixed individuals exhibit an elevated number of variable number tandem repeats at *MUC19*

*MUC19* is structurally similar to other mucins, containing a variable number tandem repeat (VNTR), in the case of MUC19, a 30 base pair repeat motif (hg19, Chr12:40876395-40885001; Figure S10) 45.4kb away from the core 72kb haplotype, but within the larger 742kb introgressed region. To test if individuals who harbor an introgressed tract overlapping the repeat region differ in the number of repeats copies compared to individuals who do not harbor introgressed tracts, we calculated the number of repeat copies of the 30bp motif in the TGP individuals (see Methods; Figure S11; Table S14-S15). For each individual, we first report the average number of repeat copies between their two chromosomes. Among all individuals from the TGP, we identified outlier individuals with elevated repeat copies above the 95th percentile (>487 repeat copies; dashed line in Figure 2). We found that MXL individuals have on average ~493 copies and individuals from the admixed American super population have on average ~417 copies (Figure 2A; Table S16-S17). In contrast, non-admixed American populations have an average of ~341 to ~365 repeats (Figure 2A; Table S16). Remarkably, out of all the outlier individuals from the TGP (>487 repeat copies), a significant proportion of them (~77%) are from admixed

American populations (Proportions *Z*-test, *P-value*: 3.971e-17; Table S18-S21; Figure S12). Outlier individuals from the Americas also carry a significantly higher copy number of tandem repeats compared to the other outlier individuals from non-admixed American populations (Mann-Whitney U, *P-value*: 5.789e-7; Figure S12; Table S18-S21). In particular, in MXL, we find that exactly 50% of individuals exhibit an elevated copy number of tandem repeats (Table S16).

Surprisingly, within individuals exhibiting an outlier number of repeat copies (>487), a significant proportion (~86%) have an introgressed tract overlapping the repeat region and these individuals harbor an elevated number of repeat copies compared to outlying individuals who do not harbor an introgressed tract overlapping the VNTR region (Proportions *Z*-test, *P-value:* 2.127e-29; Mann-Whitney U, *P-value:* 1.398e-06; Figure S13; Table S18-S21). Most strikingly, all of the outlying MXL individuals carry at least one introgressed tract that overlaps with the VNTR region (Figure 2). Notably, MXL has more individuals exhibiting an elevated copy number (>487 repeat copies) than any other TGP population, and there is a positive correlation between the number of repeat copies and the number of introgressed tracts that overlap with the VNTR present in a MXL individual (Spearman's $\rho$: 0.885; *P-value*: 2.839e-22; Figure 2B; Figure S14; Table S22). Furthermore, we find that among MXL individuals, the number of repeat copies and the Indigenous American ancestry proportion at the repeat region is significantly positively correlated (Spearman's $\rho$: 0.483; *P-value*: 2.940e-4; Figure 2C; Figure S15, Table S23-S24), while the African (Spearman's $\rho$: -0.289; *P-value*: 2.072e-2; Figure S15, Table S23-S24) and European (Spearman's $\rho$: -0.353; *P-value*: 4.191e-3; Figure S15, Table S23-S24) ancestry proportions have a significant negative correlation. Taken together, in MXL, we find that an individual's VNTR copy number is highly predicted by the the number of introgressed tracts that overlap the VNTR. To a lesser extent, the VNTR copy number is also predicted by the Indigenous American ancestry proportion in the repeat region, indicating that individuals with elevated VNTR copy number have higher proportions of Indigenous American ancestry and harbor the introgressed haplotype. These results show that individuals who carry an elevated number of the *MUC19* VNTR are likely to also carry the archaic haplotype, especially in admixed American populations where the archaic haplotype of *MUC19* is found at highest frequencies (Mann-Whitney U, *P-value:* 1.597e-87; Figure S13; Figure 2; Table S18-S21).

Importantly, long-read sequence data from the Human Pangenome Reference Consortium (HPRC) and Human Genome Structural Variant Consortium (HGSVC) corroborated our findings (Figure S10; Figure S16), revealing an extra 424 copies of the 30bp *MUC19* tandem repeat exclusively in American samples, arranged in four additional segments of 106 repeats (at 3,171 bp each). This structural variant is exceptionally large; it effectively doubles the size of the ~12kb coding exon that harbors the tandem repeat (Figures S10). This suggests that functional differences between the *Human-like* and archaic haplotypes may lie in the elevated number of the 30 base-pair motifs carried in the archaic haplotype, and that the positive selection detected in American populations may be acting on haplotypes carrying elevated copy numbers of the 30 base-pair motif.

Introgression introduced missense variants at *MUC19*

Inspecting the 135 Denisovan-specific SNPs and 4 Neanderthal-specific SNPs in the core 72kb region reveals that some modern humans carry two Denisovan-specific synonymous sites and nine Denisovan-specific non-synonymous sites, where the archaic allele codes for a different amino acid than the non-introgressed variant (Table S25). We quantified the allele frequencies for these nine Denisovan-specific missense variants in present-day populations and in 23 ancient Indigenous American genomes that predate European colonization and the African slave trade (Figure 3A; Table S26-S33). In the admixed American superpopulation, we find that the Denisovan-specific missense mutations are segregating at the highest frequencies (AMR, Denisovan-specific missense mutation frequency range: ~0.154 - ~0.157) compared to all other TGP superpopulations (non-AMR, Denisovan-specific missense mutation frequency range: ~0 - ~0.108; Table S27-S28). When we stratify by population instead of by superpopulation, we find the Denisovan-specific missense mutations are segregating at frequencies between ~0.069 and ~0.305 amongst admixed American populations, at varying frequencies between ~0.005 and ~0.157 throughout European, East Asian, and South Asian populations, and at the highest frequency in MXL where all nine Denisovan-specific missense mutations are segregating at a frequency of ~0.305 (Figure 3A; Table S29). We find the mean Denisovan-specific missense mutation frequency to be positively correlated with the introgressed tract frequency per population (Pearson's $\rho$: 0.976; *P-value*: 5.306e-16; Figure S17).

We then evaluate the frequency of the nine Denisovan-specific missense mutations in 23 ancient pre-European colonization American individuals, and find that each of the nine Denisovan-specific missense mutations are segregating in the ancient individuals at higher frequencies than in any admixed American population in the TGP, but at statistically similar frequencies with respect to MXL (see Methods; Figure 3A; Table S29-S32). These ancient individuals were sampled from a wide geographic and temporal range (Figure S18; Table S26) and thus do not comprise a meaningful population, yet we detect the presence of the Denisovan-specific missense mutations in sampled individuals from Alaska, Montana, California, Ontario, Central Mexico, Peru, and Patagonia (Table S30). When we quantify the frequency of these mutations in 22 unadmixed Indigenous Americans from the Simons Genome Diversity Project (SGDP), we find that all nine Denisovan-specific missense variants are segregating at a frequency of ~0.364, which is statistically similar to the ancient American frequencies (see Methods; Table S31-S32), and higher than any admixed American population in the TGP, albeit at statistically similar frequencies with respect to MXL (Table S31-S32). Given that all nine of the missense mutations are found within a ~17.5kb region, we quantified the frequency of the Denisovan-specific missense mutation at position Chr12:40808726 in both the ancient individuals and admixed Americans in the TGP, as this position has genotype information in 20 out of the 23 ancient American individuals (Table S30). We then assessed the relationship between Indigenous American ancestry proportion at the 72kb region, and this Denisovan-specific missense mutation frequency. We find a positive and significant relationship (Pearson's $\rho$: 0.489; *P-value*: 1.982e-23; Figure S19) between an individual's Indigenous American Ancestry proportion and their respective Denisovan-specific missense mutation frequency, which suggests that recent admixture in the Americas may have diluted the introgressed ancestry at the 72kb region. We also quantify the frequency of these variants in 44 African individuals from the SGDP, and find all nine Denisovan-specific missense variants at a frequency of ~0.011, in a single chromosome from a Khomani San individual (Table S33).

To estimate the potential effect of these missense mutations on the MUC19 protein, we relied on Grantham scores. Grantham scores quantify the distance between amino acids based on their physicochemical properties, and are used to predict an amino acid substitution's impact on the structure or function of the translated protein [*36*]. One of the Denisovan-specific missense mutations found at position 40821871 (rs17467284 in Figure 3B) results in an amino acid change with a Grantham score of 102. This substitution is classified as moderately radical [*37*], and suggests that the amino acid introduced through introgression is likely to impact the translated protein's structure or function. Interestingly, this Denisovan-specific missense mutation falls within an exon that is highly conserved across vertebrates (PhyloP score: 5.15, *P-value:* 7.08e-6; Figure 3B) [*38*], indicating that this amino acid residue is likely functionally important, and that the amino acid change introduced by the Denisovan-specific missense mutation may have a significant structural or functional impact. Furthermore, this missense mutation falls between two Von Willebrand factor D domains, which play an important role in the formation of mucin polymers and gel-like matrices [*39*]. Our results suggest that this Denisovan-specific missense mutation is a strong candidate for impacting its translated protein, and may affect the polymerization properties of *MUC19* and the viscosity of the mucin matrix.

Identification of the most likely donor of the introgressed haplotype at *MUC19*

To identify the most likely archaic donor, we investigate the patterns of haplotype divergence at *MUC19* by comparing the modern human haplotypes in the TGP in the 72kb region (see Methods; shaded region in Figure 1A)—to the high-coverage archaic humans. We calculated the sequence divergence—the number of pairwise differences normalized by the effective sequence length—between all haplotypes in the TGP and the genotypes for the Altai Denisovan and the three high-coverage Neanderthal individuals (Figure S20; Tables S34-S35). Haplotypes from the Americas exhibit a bimodal distribution of sequence divergence for affinities to the Altai Denisovan, which we do not observe for the African haplotypes (Figure 4A), as expected for an introgressed region. Notably, there is a clear pattern of sequence divergence for the introgressed haplotypes found in the American super-population of the TGP (AMR) with respect to the four high-coverage archaic genomes at the 72kb region (Figure 4B).

Interestingly, despite our $U_{AFR,MXL,Denisovan}(1\%, 30\%, 100\%)$ and archaic SNP density results demonstrating that the introgressed haplotype at the 72kb region shares the most alleles with the Altai Denisovan (Figure 4B), we find that this region is not statistically significantly closer to the Altai Denisovan individual than expected from the genomic background of sequence divergence (sequence divergence from the Denisovan: 0.00097, *P-value:* 0.237, Figure S21, Table S36). However, this is not unusual, given that the Altai Denisovan is not genetically closely related to Denisovan introgressed segments in modern humans (see Supplemental Section S5, and [33]), which might suggest that the Denisovan donor population of the 72kb region in *MUC19* is not closely related to the Altai Denisovan individual. Furthermore, the 72kb region is also not statistically significantly closer to Neanderthals than expected from the genomic background of sequence divergence (sequence divergence from the Altai Neanderthal: 0.003648, *P-value:* 0.995; Chagyrskaya Neanderthal: 0.001818, *P-value:* 0.811; Vindija Neanderthal: 0.001816, *P-value:* 0.806; Figure S21, Table S36).

As an additional approach, we used the *D+* statistic to assess which archaic human exhibits the most allele sharing with the introgressed haplotype at the 72kb region in *MUC19*, as it relies only on the sites that are most informative for introgression [31, 32]. We performed *D+ (P1, P2; P3, Outgroup)* tests with the following configurations: the Yoruban population (YRI) as *P1*, the focal MXL individual (NA19664) with two copies of the introgressed haplotype with an affinity to the Altai Denisovan as *P2*, and one of the four high-coverage archaic genomes as *P3*; we use the EPO ancestral allele call from the six primate alignment as the Outgroup. Notably, we exclusively observe a positive and significant *D+* value (*D+*: 0.743, *P-value*: 1.386e-5; Figure S22; Table S37) only when the Altai Denisovan is used as *P3* (the putative donor population). Conversely, when any of the three Neanderthals are used as *P3*, we observe non-significant *D+* values (*P3*: Altai Neanderthal, *D+*: -0.622, *P-value*: 0.999; *P3*: Chagyrskaya Neanderthal, *D+*: 0.175, *P-value*: 0.183; *P3*: Vindija Neanderthal, *D+*: 0.182, *P-value*: 0.174; Figure S22; Table S37). These *D+* suggest that the introgressed haplotype at the 72kb *MUC19* region shares more alleles with the Altai Denisovan, which is not observed with any of the three Neanderthals and provides evidence that the introgressed haplotype found in modern humans is *Denisovan-like*.

When we consider the longest introgressed tract in MXL—the 742kb region—we find that it is closest to the Chagyrskaya and Vindija Neanderthals, and significantly closer than expected from the genomic background (sequence divergence from the Chagyrskaya Neanderthal: 0.000661, *P-value*: 0.006; sequence divergence from the Vindija Neanderthal: 0.000656, *P-value*: 0.007; Figure S23-S24; Table S38-40). We also tested whether this region is statistically significantly closer to the Altai Denisovan than expected from the genomic background and found that the longest introgressed tract in MXL is also significantly closer than expected to the Altai Denisovan, albeit not as close when compared to the Chagyrskaya and Vindija Neanderthals (sequence divergence from the Altai Denisovan: 0.000806, *P-value*: 0.019; Figure Figure S23-S24; Table S38-401). We then performed *D+* analyses for the 742kb region with identical configurations as for the 72kb region and observe positive and significant *D+* values when *P3* is Chagyrskaya (*D+*: 0.381, *P-value*: 7.375e-6; Figure S25; Table S41), and Vindija Neanderthals (*D+*: 0.383, *P-value*: 7.505e-6; Figure S25; Table S41), but notably *D+* is not significant when the Altai Neanderthal is *P3* (*D+*: 0.091, *P-value*: 1.442e-1; Figure S25; Table S41). *D+* is, however, significant when the Altai Denisovan is *P3* (*D+*: 0.377, *P-value*: 9.889e-8; Figure S25; Table S41). These *D+* results are consistent with our sequence divergence results, which indicate that the introgressed haplotype at the 742kb *MUC19* region has a high affinity for the Altai Denisovan and the two late Neanderthals, but strikingly not the Altai Neanderthal (Figures S20-S25; Tables S34-S41).

Given the exceedingly high density of Denisovan-specific alleles (Figure S2; Table S4), the sequence divergence, and *D+* results for the 72kb and 742kb region, the most parsimonious explanation is that a Denisovan population could have introduced this haplotype into non-Africans. However, our 742kb results also suggests a late Neanderthal population could have introduced the introgressed haplotype. This is further supported by the sequence divergence results at the 72kb region where late Neanderthals exhibit intermediate distance to the introgressed haplotype (Figure 4B), suggesting they must harbor some of the Denisovan alleles.

In the next section, with evidence from additional analysis, we propose a more complicated evolutionary history than a simple case of unidirectional introgression at *MUC19*.

Neanderthals introduce *Denisovan-like* introgression into non-African modern humans

Based on sequence divergence, the Chagyrskaya and Vindija Neanderthals carry a 742kb haplotype that is most similar to the Altai Neanderthal, with the exception of the 72kb region. To understand why the Chagyrskaya and Vindija Neanderthals exhibit intermediate levels of sequence divergence with the introgressed haplotype present in MXL at the 72kb region in *MUC19* relative to the Altai Denisovan and Altai Neanderthal (see the $\alpha$ ellipse in Figure 4B), we computed the number of heterozygous sites for each archaic human. Because the Chagyrskaya and Vindija Neanderthals present intermediate sequence divergences, we expected these two individuals to have more heterozygosity than the Altai Neanderthal. Strikingly, at the 72kb region in *MUC19*, we observe that the Chagyrskaya and Vindija Neanderthals carry an elevated number of heterozygous sites (Chagyrskaya heterozygous sites: 168, *P-value:* 2.307e-4; Vindija heterozygous sites: 171, *P-value:* 2.282e-4; Figure 5A; Figure S26; Table S42) that is higher than those of the Altai Neanderthal (heterozygous sites: 1, *P-value*: 0.679; Figure5A; Figure S26; Table S42) and the Altai Denisovan (heterozygous sites: 6, *P-value*: 0.455; Figure5A; Figure S26; Table S42). The Chagyrskaya and Vindija Neanderthals carry a higher number of heterozygous sites than all African individuals (~75, *P-value*: 0.424; Figure 5A; Figure S27; Table S43), and have a more similar pattern to non-African individuals carrying exactly one *Denisovan-like* haplotype (~287, *P-value*: 3.157e-4; yellow X's in Figure 5A; Figure S27; Table S43). This observation runs opposite to the genome-wide expectation for Neanderthals, as archaic humans have much lower heterozygosity than modern humans (genome-wide heterozygosity is ~0.00014 - ~0.00017 for Chagyrskaya and Vindija Neanderthals, ~0.00019 for the Denisovan, and ~0.001 for Africans modern humans; Figure S28; Table S44).

Within modern humans, we find that individuals carrying exactly one *Denisovan-like* haplotype at the 72kb region harbor significantly more heterozygous sites at *MUC19* compared to the rest of their genome (average number of heterozygous sites: ~287, *P-value*: 3.157e-4; Figure S27; Table S43) which surpasses the number of heterozygous sites at *MUC19* of any African individual (Figure 5A). Individuals carrying two *Denisovan-like* haplotypes harbor significantly fewer heterozygous sites than expected at *MUC19* relative to the rest of their genome (average number of heterozygous sites: ~4, *P-value*: 6.945e-4; Figure S27; Table S43), while African individuals harbor the expected number of heterozygous sites (average number of heterozygous sites: ~75, *P-value*: 0.424; Figure S27; Table S43). Given that the Chagyrskaya and Vindija Neanderthals and non-African individuals who harbor one copy of the *Denisovan-like* haplotype exhibit an excess of heterozygous sites at the 72kb region, we hypothesized that the Chagyrskaya and Vindija Neanderthals also harbor one *Denisovan-like* haplotype. This arrangement would explain the elevated number of heterozygous sites and the intermediary sequence divergences with respect to the introgressed haplotype.

To test this hypothesis we first performed additional tests for gene flow between the archaic individuals using the *D+* statistic within the 72kb *MUC19* region that provided evidence that the Chagyrskaya and Vindija Neanderthals harbor one copy of the *Denisovan-like* haplotype. For these comparisons the Altai Neanderthal is *P1*, either the Chagyrskaya or Vindija Neanderthals are *P2*, and the Altai Denisovan is *P3*, we observe significant and positive *D+* values supporting gene flow between the Denisovan and the Chagyrskaya (*D+:* 0.783; *P-value:* 0.029) and Vindija (*D+:* 0.819; *P-value:* 0.018) Neanderthals (Figure S29; Table S45). To further investigate whether the Chagyrskaya and Vindija Neanderthals harbor one *Denisovan-like* haplotype in the 72kb region, we used `BEAGLE` to phase the 72kb region. As no phasing has been done for archaic humans, we first tested the reliability of using the TGP as a reference panel. To do so we constructed a synthetic Neanderthal for the 72kb region with one copy of the *Denisovan-like* haplotype by sampling one allele from the Altai Neanderthal and the other other allele from the Altai Denisovan at sites that are heterozygous in either the Chagyrskaya or Vindija Neanderthals and are fixed differences between the Altai Neanderthal and the Altai Denisovan. We found that we could phase the synthetic individual perfectly at this region (see Supplementary Section S3). Encouraged by these results, we phased the Chagyrskaya and Vindija Neanderthals at the 72kb region, and confirmed they carry one haplotype that is similar to the Altai Neanderthal, and one haplotype that is similar to the *Denisovan-like* haplotype in MXL. Relative to the Altai Neanderthal, the Chagyrskaya *Neanderthal-like* haplotype exhibits 3.5 differences, and the Vindija exhibits 4 differences (Figure 5B; Table S46). Relative to the Altai Denisovan, the Chagyrskaya *Denisovan-like* haplotype exhibits 43 differences, and the Vindija haplotype exhibits 41 differences (Figure 5B; Table S46). As expected, the phased *Denisovan-like* haplotype in these two Neanderthals is closest to the *Denisovan-like* haplotype in MXL; the Chagyrskaya exhibits 5 differences, and the Vindija Neanderthal exhibits 4 differences (Figure 5B; Table S47). Furthermore, we show that, in the 72kb region, the introgressed haplotype in MXL is statistically significantly closer to the phased *Denisovan-like* haplotype present in Chagyrskaya and Vindija Neanderthals (sequence divergence from Chagyrskaya Neanderthal haplotype: 0.000104, *P-value:* 0.003; sequence divergence from Vindija Neanderthal haplotype: 0.000083, *P-value:* 0.002; Figure S30; Table S47). Due to the potential introduction of biases when phasing ancient DNA data, to investigate if the Chagyrskaya and Vindija Neanderthals carry a *Denisovan-like* haplotype we developed an approach called Pseudo-Ancestry Painting (*PAP*, see Methods) to assign the two alleles at a heterozygous site to two source individuals. We found that using a MXL (NA19664) and a YRI (NA19190) individuals as sources maximizes the number of heterozygous sites in the Chagyrskaya (*PAP* Score: 0.94, *P-value:* 3.683e-4) and Vindija (*PAP* Score: 0.929, *P-value:* 8.679e-05) Neanderthals, whose alleles are present in the two sources (Figure S31; Table S48).

In sum, our analyses suggest that some non-Africans carry a mosaic region of archaic ancestry: a small *Denisovan-like* haplotype (72kb) embedded in a larger Neanderthal haplotype (742kb), that was inherited through Neanderthals, who themselves acquired Denisovan ancestry from an earlier introgression event (Figure S32). This is consistent with the literature, where Denisovan introgression into Neanderthals is rather common [*34*, *35*]. Thus we refer to the mosaic haplotype found in modern humans as the archaic haplotype.

## Discussion

The study of adaptive archaic introgression is in its infancy, but has already illuminated candidate genomic regions that affect the health and overall fitness of global populations. In this study, we pinpointed several aspects of the gene *MUC19* that highlight its importance as a candidate to study adaptive introgression: one of the haplotypes that span this gene in modern humans is of archaic origin; modern humans inherited this haplotype from Neanderthals, who in turn inherited it from Denisovans; the haplotype introduced nine missense mutations that are at high frequency in both Indigenous and Admixed American populations; individuals with the archaic haplotype carry a massive coding VNTR expansion relative to the non-archaic haplotype, and their functional differences may help explain how mainland Indigenous Americans adapted to their environments, which remains under-explored. To our knowledge, this study provides the first example of natural selection acting on archaic alleles at coding sites, and the first example of natural selection acting on VNTRs.

A larger implication of our findings is that archaic ancestry could have been a useful source of standing genetic variation as the early Indigenous American populations adapted to new environments, with genes like *MUC19* and other mucins possibly mediating important fitness effects [*40*]. The staggering variation in the *MUC19* coding VNTR in global populations dovetails with this idea, and adds to a growing body of evidence for the important role of structural variants in human genomics and evolution [*41-42*]. Yet, in American populations, particular haplotypes carrying the most extreme copy numbers were selected and are now very frequent; effectively doubling the functional domain of this mucin, indicating an adaptive role driven by environmental pressures particular to the Americas.

Another interesting aspect of *MUC19* is the evolutionary history of the introgressed region. Our observation of a 72kb Denisovan haplotype found in Neanderthals and non-African modern humans that is nested within a larger Neanderthal haplotype, suggests that the smaller Denisovan haplotype was first introgressed into Neanderthals, who later admixed with modern humans to introduce the full 742 kb haplotype. While the Altai Neanderthal does not harbor the Denisovan haplotype at the 72kb region, the other two chronologically younger Neanderthals (Chagyrskaya and Vindija) do. We phased these younger Neanderthals (see Supplementary Section X) and showed that they harbor exactly one Denisovan-like haplotype, which explains why they exhibit an excess of heterozygosity. The *Denisovan-like* haplotype in the younger Neanderthals is also statistically significantly closer to the archaic haplotype present in MXL (Figure SX; Table SX), providing additional evidence that modern humans obtained this haplotype through an interbreeding event with Neanderthals. Despite the introgressed archaic haplotype having an excessive amount of shared alleles with the Altai Denisovan at the 72kb region, the Altai Denisovan harbors several private mutations—18 and 6 mutations in the homozygous and heterozygous state respectively—that are absent across all 287 *Denisovan-like* haplotypes in the TGP, suggesting that the introgressing Denisovan population may not be closely related to Altai Denisovan (see Supplemental Section S3). Indeed, the introgressed haplotype in the 72kb region is present at low frequencies in other non-African populations including Papuans—where the genome-wide Denisovan ancestry of Papuans has been estimated to originate from a population of Denisovans that was not closely related to the Altai Denisovan [*33*].

Finally, we find a single San individual who carries the nine Denisovan missense variants in heterozygous form, uniquely among all African individuals considered here. The sequence divergence between this San haplotype and the archaic MXL haplotype at the 72kb region is high (0.001342), further supporting the origin of the archaic haplotype in non-Africans as introgressed. Khoe-San populations are estimated to have diverged from other African groups 120 thousand years ago [43]. This may indicate that the ancestral modern human population also carried two *MUC19* haplotypes in polymorphism, one harboring the nonsynonymous variants, only to be lost in most human populations and reintroduced through archaic introgression in non-African populations. Finding a divergent haplotype in the San is consistent with a previous study [44], as ~1% of their ancestry can be attributed to lineages diverged from the main human lineage beyond 1 million years ago. We note that this San individual does not harbor an extended number of repeat copies of the VNTR (301), which further supports the importance of the VNTR expansion in the Americas. Regardless of its origin, it is striking that two highly divergent haplotypes were maintained in polymorphism in the San and in two Neanderthal populations, and this may be the result of balancing selection [*45*]. Balancing selection in the form of heterozygous advantage may also explain why the archaic haplotype is found at high frequencies in American populations yet did not reach fixation in any sampled population. More generally, the evolutionary history of this region suggests a complex history that involves recurrent introgression and natural selection, and it parallels complex introgression patterns from other regions of the genome (Neanderthal mtDNA, Y chromosome, and *KNL1* spindle gene [*46–48*]).

Perhaps the largest knowledge gap concerning why the archaic haplotype of *MUC19* would be under positive selection is its underlying function. Mucins are secreted glycoproteins responsible for the gel-like properties and the viscosity of the mucus [*49*]. Mucins are characterized by proline, threonine and serine (PTS) tandem repeats, which in *MUC19* are structured into 30bp tandem repeats. The massive difference in copy numbers of the 30bp PTS tandem repeat domains carried by individuals harboring the *Human-like* and archaic haplotypes strongly suggests *MUC19* variants differ in function as a consequence of different molecular binding affinities between variants. This is the case in other mucins, such as *MUC7*, where variants carrying different numbers of PTS repeats exhibit different microbe-binding properties [*40*]. If the two variants of *MUC19* also have differential binding properties, this would lend support to why positive selection would increase the frequency of the archaic haplotype in American populations. Yet, there is limited medical literature associating variation in *MUC19* with human fitness. Further experimental validation of how VNTRs and the Denisovan-specific missense mutations affect MUC19 function is necessary to understand the effect the archaic haplotype may exert on the translated *MUC19* protein, and how it modifies its function during the formation of mucin polymers.

Methods developed in evolutionary biology can be useful for identifying candidate variants underlying biological functions. Future functional and evolutionary studies of the *MUC19* region will not only provide insight into specific mechanisms of how variation at this gene confers a selective advantage, but also specific evolutionary events that occurred in the history of humans. Beyond improving our understanding of how archaic variants facilitated adaptation in novel environments, our findings also highlight the importance of studying archaic introgression in understudied populations, such as admixed populations from the Americas [*50*]. Genetic variation in American populations is less well-characterized than other global populations; it is

difficult to deconvolve Indigenous ancestries from European, African, and—to a lesser extent—South Asian ancestries, following 500 years of European colonization [*29*]. This knowledge gap is exacerbated by the high cost of performing genomic studies, building infrastructure, and generating scientific capacity in Latin America—but it is a worthwhile investment—as our study shows that leveraging these populations can lead to the identification of exciting candidate loci that can expand our understanding of adaptation from archaic standing variation.

## Methods

### Data Processing

*Modern Human Data*

Sequence data for the *MUC19* locus were obtained from a publicly available global reference panel, the 1,000 Genomes Project Phase III (TGP), which contains a diverse set of individuals from multiple populations [1000 Genomes Project Consortium, 2015]. The autosomal variant sites from the integrated callset VCF files for the TGP were downloaded from http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502 and the local ancestry calls for admixed American individuals [Martin et al., 2017] were downloaded from https://personal.broadinstitute.org/armartin/tgp_admixture. Data for *MUC19* in the Papuan and present day Indigenous American individuals was obtained from the Simons Genome Diversity Project (SGDP) [Mallick et al., 2016; Wong et al., 2020]. The autosomal variant sites VCF files for the SGDP were downloaded from https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/phased_data2021. Both the TGP and SGDP datasets were filtered to remove multi-allelic and structural variant sites. For the *iHS* analyses conducted using the TGP dataset, we removed additional sites that did not have an ancestral allele call and were no longer bi-allelic after considering the ancestral allele call, where we annotated the dataset using the ancestral allele calls in fasta format for the hg19 assembly using the Enredo, Pecan, Ortheus (EPO) pipeline, which was download from http://ftp.ensembl.org/pub/release-74/fasta/ancestral_alleles [Paten et al. 2008a, Paten et al. 2008b, Herrero et al. 2016]. Given that the modern human genotypes were imputed and only include information for variable sites, any site that was not removed during the filtering process was assumed to be homozygous reference as was done in Huerta-Sánchez et al 2014. As the ACB and ASW admixed populations have a high proportion of African ancestry, individuals from these populations were removed and not considered in any analysis. Note that the analyses in the section “Copy number polymorphism of a 30bp tandem repeat motif between the *Human-like* and a *archaic* haplotypes" were conducted on the TGP data aligned to the hg38 reference assembly, while all other analyses were completed using the hg19 reference which was soft masked for repetitive regions. Data aligned to hg38 was downloaded from the Human Pangenome Reference Consortium: https://projects.ensembl.org/hprc/ and the
Human Genome Structural Variation Consortium: https://www.internationalgenome.org/data-portal/data-collection/hgsvc2.

*Archaic Human Data*

The autosomal all-sites VCF files and the BED files consisting of the suggested general filtering best practices for the four high-coverage archaic genomes were downloaded from https://www.eva.mpg.de/genetics/genome-projects. The autosomal VCF files for the archaic

genomes were initially filtered with their respective BED files to exclude regions of the genome that are prone to alignment errors as was done in the original [Prüfer et al., 2017; Mafessoni et al., 2020]. The initially filtered archaic VCF files were then merged using *BCFtools v1.16*, and after the initial merging the resulting VCF files were filtered to only include sites that were mono-allelic or bi-allelic where at least one archaic had a MQ >= 25 and GQ >= 40—archaics that did not meet this threshold were coded as missing data. Additionally for our *D*+ analysis, we generated another merged archaic dataset by adding the additional requirement that at a given site we could determine the ancestral and derived state as defined by the hg19 ancestral sequence.

*Combined Data*

The autosomal VCF files for each modern human dataset (i.e., TGP and SGDP) and each filtered archaic genomes were initially merged using *BCFtools v1.16*, after the initial merging the resulting VCF files were filtered to only include sites with mono-allelic or bi-allelic SNPs and where the respective archaic had a MQ >= 25 and GQ >= 40. For *D*+ analyses we generated another combined dataset per archaic, with the additional requirement that the given site must have an ancestral allele call present as defined by the hg19 ancestral sequence. Additionally to assign sites into the different SNP set partitions (see the Archaic SNP Density section) and phasing analyses we merged the TGP dataset with all four filtered archaic genomes archaics with the same filtering scheme for the merged datasets with the single archaic. We retained sites that had at least one archaic pass the filtering criteria and for any archaic that did not meet the filtering criteria was coded as missing data. Since the modern human genotypes were imputed and only include information for variable sites—unlike the archaic data which contains information for all sites—any site that was originally absent in the TGP or SGDP datasets but present in archaic data we assumed to be homozygous for the reference allele in the modern humans [Huerta-Sánchez et al 2014]. After each of the combined datasets were curated, we annotated coding sites for NCBI RefSeq genes [Pruitt et al., 2007] using the NCBI RefSeq Select transcripts (downloaded from https://hgdownload.cse.ucsc.edu/goldenpath/hg19/database/ncbiRefSeqSelect.txt.gz) and *SnpEff v5.1* [Cingolani et al., 2012]. Lastly, for better computational efficiency when performing analyses all of the final VCF files were converted to Zarr arrays using *scikit-allele v1.3.5* (10.5281/zenodo.597309).

*Pre-Contact Indigenous American Genomes*

Genomic data for *MUC19* in ancient individuals was generated by combining high coverage (>1X) pre-European contact genomes from the literature, including nine individuals from California, one from Ontario [Scheib et al., 2018], four from Peru [Lindo et al., 2018] four from Patagonia [de la Fuente et al., 2018], one from Alaska [Moreno-Mayar et al., 2018], one from Montana [Rasmussen et al., 2014], and three from Central Mexico [Villa-Islas et al., 2023]. Sequence reads were downloaded in fastq format and converted to bam format using *bwa v7.17* [Li and Durbin, 2009]. Reads were then sorted, duplicates were removed, and all non-autosomal chromosomes were removed using *SAMtools v1.9* [Li et al., 2009]. Using *ANGSD v0.92* we further filtered out reads that had a quality score less than 30 and then determined the read depth of the alleles present at the Denisovan-specific coding sites. Then, for each ancient individual we determined the genotype at each of the Denisovan-specific coding sites which we had sequencing information for by first considering any allele that had a read depth of two or greater,

and then ensured that site was mono-allelic or bi-allelic for only the hg19 reference and/or Denisovan-specific alternative allele.

## Identification of the *MUC19* Introgressed Region

To identify the genomic coordinates of the introgressed region in *MUC19*, we inferred introgressed tracts using *hmmix* [Skov et al. 2018] for chromosome 12. Specifically, we inferred introgressed tracts using the *-haploid* option which generates two sets of inferred tracts per individual—i.e., one for each haplotype—and only retained inferred archaic tracts that had a posterior probability greater than or equal to 0.8, and that overlapped with at least one base pair of the *MUC19* NCBI RefSeq coordinates for the hg19 assembly (Chr12:40787196-40964559). As we were only interested in the tracts that overlap *MUC19*, if an individual's haplotype had two inferred tracts overlapping *MUC19* we stitched them together following the approach implemented in [Coll Macià et al., 2021] by taking the union of the two respective tracts—i.e., the minimum of the two start positions and the maximum of the two end positions. We then performed a Proportions *Z*-test to determine if the proportion of introgressed tracts in the admixed American populations is greater than in non-admixed American populations using the statsmodels.stats.proportion.proportions_ztest function implemented in statsmodels v0.13.2 [Seabold and Perktold 2010], and a Fisher's Exact Test to assess if introgressed tracts are overrepersented in admixed American populations than non-admixed American populations using the scipy.stats.fisher_exact as implemented in *scipy v1.7.2* [Virtanen et al., 2020]—for both statistical tests African populations were not included among non-admixed American populations and *P-values* less than 0.05 were considered statistically significant. Lastly, we identified two distinct regions, the 742kb region containing the longest archaic tract in any MXL individual (Chr12:40272001-41014000, in NA19725), and a focal 72kb region (Chr12:40759001-40831000) which has the highest density of inferred introgressed tracts amongst non-African populations in the TGP that is larger than 40kb, which is the length needed to confidently differentiate between introgression and incomplete lineage sorting [Huerta-Sánchez et al., 2014]. It should be noted that in Figure 1A we omitted the 2.613 Mb tract in the PUR individual HG01108 for visual clarity, but the version of this plot including this tract along with population specific plots can be viewed in Figures S33-S37.

## Archaic SNP Density

Throughout this study we compute different subsets of archaic specific alleles found in non-African populations. For a SNP to first be considered an archaic allele, we require that an allele must be rare in the African superpopulation (i.e., at a frequency less than 0.01) and at a frequency greater than 0.01 in the respective non-African population as was done in [Witt et al., 2023]. For the SNP to be considered Denisovan-specific we further required the archaic allele to be fixed in the sequenced Denisovan and not fixed in any of the three high-coverage Neanderthals. Similarly, for a SNP to be considered Neanderthal-specific we further required the archaic allele to be fixed in at least one Neanderthal and not fixed in the sequenced Denisovan. For an SNP to be considered as a shared archaic allele we further required the archaic allele to be fixed in the sequenced Denisovan and in at least one Neanderthal. For the SNPs that were not identified as archaic SNPs—i.e., the union of the Denisovan-specific, Neanderthal-specific, and shared archaic SNPs—we classified alleles segregating in the TGP and absent in all archaic

individuals as Human-specific. Lastly, non-Archaic SNPs that are segregating in at least one archaic and are also segregating in the TGP are considered to be shared hominin alleles. To determine if non-African populations at the 742kb longest introgressed tract region and at the core 72kb *MUC19* region harbor more archaic SNPs than expected we computed the number of Denisovan-specific and Neanderthal-specific SNPs for each region. To assess if our 742kb and 72kb *MUC19* regions have a higher archaic SNP density than expected, we compared the observed archaic SNP density to a distribution of archaic SNP densities from the genomic background of 742kb and 72kb non-overlapping windows with comparable effective sequence length density—i.e., within one standard deviation from the mean for each windowed effective sequence length distribution. To calculate *P-values*, we determined the proportion of windows from the genomic background with an archaic-speacific SNP density greater than or equal to what we observed at the 742kb and 72kb *MUC19* regions, respectively, where a *P-value* less than 0.05 is considered significant.

## Positive Selection

### *Population Branch Statistic*

We utilized the Population Branch Statistic (*PBS*) to assess if the archaic haplotype has been subjected to positive selection in Admixed American populations. *PBS* uses the logarithmic transformation of pairwise estimates of $F_{ST}$ to measure the branch length in the target population since its divergence from the two control populations [Yi et al., 2010]. We chose MXL as target populations and CEU and CHB as our control populations. To account for differences in sample sizes between populations we used Hudson's estimator of $F_{ST}$ as it has been shown to not only be a conservative estimator, but is also robust to differences in sample sizes [Bhatia et al., 2013]. Not that since $F_{ST}$ and *PBS* both represent branch lengths on unrooted trees, in all *PBS* computations we set negative $F_{ST}$ and *PBS* values to zero to be conservative. To assess if there is evidence of positive selection at the 742kb longest introgressed tract region and the 72kb densest introgressed tract region we computed $PBS_{MXL:CHB:CEU}$ at each region and then to compute *P-values* we compared each observed $PBS_{MXL:CHB:CEU}$ value to a distribution of $PBS_{MXL:CHB:CEU}$ values from the genomic background of 742kb and 72kb non-overlapping windows with comparable effective sequence length—i.e., within one standard deviation from the mean for each windowed effective sequence length distribution—and determined the proportion of windows from the genomic background with a per-region $PBS_{MXL:CHB:CEU}$ value greater than or equal to what we observed at the 742kb and 72kb region, respectively, where a *P-value* less than 0.05 is considered significant. Additionally, we computed $PBS_{MXL:CHB:CEU}$ for every SNP in the genome and identified outlier SNPs in the 742kb region by setting the significance threshold at the genome-wide 99.95th $PBS_{MXL:CHB:CEU}$ percentile—for information on how we assessed significance for SNPs within the 742kb region see supplementary section S1.

### *Integrated Haplotype Score*

To corroborate our signals of selection we additionally performed a haplotype-based test for selection by computing integrated haplotype scores (*iHS*) [Voight et al., 2006] for all populations in the TGP. *iHS* measures the decay in linkage disequilibrium from a core SNP due to new mutations and recombination events, as recent positive selection is expected to result in haplotypes that are long, frequent, and have a high haplotype homozygosity in a population. Normalized |*iHS*| scores > 2 reflect that the haplotype is longer than expected under neutrality—and is commonly considered the threshold for evidence of positive selection at a locus [Voight et

al., 2006; Szpiech et al., 2024]—with extreme positive and negative values indicating that the derived and ancestral haplotypes are unusually long, respectively. To compute *iHS* we used the TGP dataset with ancestral allele information. Using selscan v2.0.0 [Szpiech et al., 2024] and the recombination maps from [International HapMap Consortium, 2007] we first computed the unstandardized *iHS* value for every SNP with a minor allele frequency > 0.05 per TGP population and then normalized *iHS* values by derived frequency bins as described in [Voight et al., 2006]. As there is no formal way to assess significance for normalized *iHS*, to determine if the focal 742kb longest introgressed tract region and the 72kb densest introgressed tract region showed signals consistent with positive selection we assessed if these regions harbored clusters of extreme *iHS* scores (i.e., normalized |*iHS*| scores > 2) [Voight et al., 2006; Szpiech et al., 2024]. Specifically, for every TGP population we binned the genome into 742kb and 72kb non-overlapping windows and computed the proportion of SNPs with normalized |*iHS*| scores > 2 for windows with more than 10 SNPs to build a genome-wide distribution of the proportion of SNPs with extreme *iHS* scores. We then assessed if the focal 742kb and 72kb regions fall within the top 1% of windows with the highest fractions of extreme *iHS* scores. Additionally, we repeated these analyses for all archaic SNPs, which are described in supplementary section S2.

*U-Statistics*

To complement our tests for positive selection we also performed explicit tests for adaptive introgression using the $U_{A,B,C}(w, x, y)$ statistic [Racimo et al., 2016]. If a genomic region is adaptively introgressed we would expect there to be many sites within that region where the archaic allele is at high frequency in the non-African population that has experienced adaptive introgression and absent or rare amongst African populations. The $U_{A,B,C}(w, x, y)$ statistic quantifies the number of sites in a given region where the archaic individual (*C*) has an allele frequency of *y%*, that the allele is at a frequency less than *w%* in a control population (*A*), and that the allele is segregating at a frequency greater than *x%* in a target population (*B*). For this study we used $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ which quantifies the number of sites where the Denisovan allele is found in the homozygous state that are segregating at a frequency less than 0.01 in the African super population and is segregating at a frequency greater than 0.3 in a given non-African population (*B*). To assess if the entire *MUC19* gene exhibits signatures of adaptive introgression we computed $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ for every NCBI RefSeq gene that has at least one segregating site amongst the Denisovan and the TGP, for all non-African populations (*B*). Given the variance in the effective sequence lengths and number of segregating sites amongst NCBI RefSeq genes, we decided that it was not feasible to assess statistical significance, but it should be noted that *MUC19* in MXL is the maximum $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ value for all NCBI RefSeq genes amongst all non-African populations (*B*), as no other gene in any non-African population exhibits such a large value of $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ (Figure 1B; Dataset SX). In order to assess statistical significance, we computed $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ for all non-African populations (*B*) for both the 742kb longest introgressed tract in MXL region and the focal 72kb region and to test if the observed $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ is larger than expected, we compared each of our observed values to a distribution of $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ from the genomic background of 742kb and 72kb non-overlapping windows with comparable effective sequence length—i.e., within one standard deviation from the mean for each windowed effective sequence length distribution. To calculate *P-values* for each non-African population (*B*), we determined the proportion of windows from the genomic background with an $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$

greater than or equal to what we observed at the 742kb and 72kb *MUC19* regions, respectively, where a *P-value* less than 0.05 is considered significant.

*Q-Statistics*

To provide an orthogonal line of evidence for adaptive introgression specific to MXL, we computed the $Q95_{A,B,C}(w, y)$ statistic [Racimo et al., 2016]. Under a scenario of adaptive introgression, one would expect that the introgressed alleles in the recipient population are at high frequencies. The $Q95_{A,B,C}(w, y)$ summarizes the site frequency spectrum in the target population (*B*), conditioned on the allele being present at a frequency of at least *y*% in the archaic population (*C*) and less than *w*% in a control population (*A*), which is accomplished by quantifying the 95th percentile of this conditional site frequency spectrum in the target population (*B*). In this study, we computed the $Q95_{AFR,B,Denisovan}(1\%, 100\%)$ statistic, which measures the 95th percentile of allele frequencies in a given non-African population (*B*) for alleles found in a homozygous state in Denisovans and that are segregating at a frequency of less than 1% in the African super population. To provide evidence that there are signals consistent with adaptive introgression exclusive to MXL we computed $Q95_{AFR,B,Denisovan}(1\%, 100\%)$ for all non-African populations (*B*) for both the 742kb longest introgressed tract in MXL region and the focal 72kb region, as well as in non-overlapping windows with comparable effective sequence length—i.e., within one standard deviation from the mean for each windowed effective sequence length distribution. Beyond summarizing how high Denisovan allele frequencies are, we also assessed the number of Denisovan alleles at high frequencies. For this, we analyzed the joint distribution of $Q95_{AFR,B,Denisovan}(1\%, 100\%)$ and $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ statistics for each non-African population (B). We compared the observed values in the 742kb and 72kb *MUC19* regions to the joint distribution of these statistics from the genomic background of non-overlapping windows. Although there is no formal statistical test for significance using this joint distribution, it should be noted that for both the 742kb longest introgressed tract in MXL and the focal 72kb region, we observed $Q95_{AFR,MXL,Denisovan}(1\%, 100\%) = 0.305$ and $U_{AFR,MXL,Denisovan}(1\%, 30\%, 100\%) = 136$. These values are higher than any other non-African population at both focal *MUC19* regions, which is consistent with evidence for adaptive introgression specific to MXL in these regions.

## Sequence Divergence

To assess the extent of divergence between *MUC19* haplotypes harbored by the various individuals in this study, we calculated sequence divergence which corresponds to number of pairwise differences between chromosomes normalized by the effective sequence length—i.e., the total number of sites that passed quality control. We use the term haplotype to refer to a single chromosome from a phased modern human individual and genotypes to refer to the two chromosomes of an archaic individual that are unphased. For information on assessing sequence divergence using the phased archaic data at the focal 72kb region see supplementary section S5.

*Identifying the Donor of the Longest Introgressed Tract found in MXL*

To determine the most likely archaic source of the 742kb longest introgressed tract (Chr12:40272001-41014000) found in an MXL individual (i.e., NA19725), inferred using *hmmix* [Skov et al. 2018] we computed the sequence divergence between each of the NA19725's chromosomes and the genotypes of the four archaic individuals. To identify the archaic donor for

the 742kb longest introgressed tract, we compared the observed sequence divergence to a distribution of sequence divergence from the genomic background of 742kb non-overlapping windows with comparable effective sequence length density—i.e., within one standard deviation from the mean for 742kb windows effective sequence length distribution. To calculate *P-values*, we determined the proportion of windows from the genomic background with a sequence divergence less than or equal to what we observed at the 742kb region. After correcting for two multiple comparisons—i.e., one per haplotype—using the Bonferroni correction, a *P-value* less than 0.025 is considered significant. To assess differences in the frequency of the 742kb introgressed haplotype among TGP populations we then computed the observed sequence divergence and *P-values* for all haplotypes in the TGP. Given that the 742kb introgressed tract found in NA19725 was closest to the two late Neanderthals (i.e., Chagyrskaya and Vindija Neanderthals) we classified a TGP haplotype as introgressed at the 742kb region if it was significantly closer than expected to either of the two late Neanderthals. We then performed a Proportions *Z*-test to determine if the proportion of introgressed haplotypes at the 742kb region in the admixed American populations is greater than in non-admixed American populations using the statsmodels.stats.proportion.proportions_ztest function implemented in statsmodels v0.13.2 [Seabold and Perktold 2010], and a Fisher's Exact Test to assess if introgressed haplotypes at the 742kb region are overrepresented in admixed American populations than non-admixed American populations using the scipy.stats.fisher_exact as implemented in *scipy v1.7.2* [Virtanen et al., 2020]—for both statistical tests African populations were not included among non-admixed American populations and *P-values* less than 0.05 were considered statistically significant.

*Modern Human Haplotype-Archaic Human Sequence Divergence at the Focal 72kb Region*

To determine the haplotype identity—i.e., the most likely donor—for the 72kb *MUC19* region in TGP individuals, we calculated the sequence divergence for all pairwise possibilities between each TGP haplotype and the genotypes of the four archaic individuals. We then characterized a TGP haplotype as being *Denisovan-like* at the 72kb region if the sequence divergence to the sequenced Denisovan was less than 0.00144524 (or 70 pairwise differences between a single TGP chromosome and the two Denisovan chromosomes), corresponding to the lone black bar of the bimodal distribution in Figure 2A and the $\alpha$ ellipse in Figure 2B**.** Seven TGP haplotypes exhibited intermediary sequence divergence levels with respect to the sequenced Denisovan between 0.002023 and 0.002044 and were determined to be recombinant haplotypes (see Figure S38 and the $\gamma$ ellipse in Figure 2B) and all TGP haplotypes with a sequence divergence larger than 0.002044 include all African chromosomes and were considered to be *Human-like* haplotypes (see the $\beta$ ellipse in Figure 2B). We then performed a Proportions *Z*-test to determine if the proportion of *Denisovan-like* haplotypes in the admixed American populations is greater than in non-admixed American populations using the statsmodels.stats.proportion.proportions_ztest function implemented in statsmodels v0.13.2 [Seabold and Perktold 2010], and a Fisher's Exact Test to assess if *Denisovan-like* haplotypes are overrepresented in admixed American populations than non-admixed American populations using the scipy.stats.fisher_exact as implemented in *scipy v1.7.2* [Virtanen et al., 2020]—for both statistical tests African populations were not included among non-admixed American populations and *P-values* less than 0.05 were considered statistically significant.

## Site Patterns Tests of Introgression

To further corroborate claims of introgression based on our sequence divergence results we used the *D+* statistic to formally test hypotheses about local introgression [Lopez-Fang et al., 2022; Peede et al., 2022]. The *D+* statistic utilizes observed site patterns from three populations and an outgroup—Newick format: *(((P1, P2), P3), O)*; site pattern format: *(P1's allelic state, P2's allelic state, P3's allelic state, O's allelic state)*—as a proxy for gene tree frequencies, where *P1* represents a population assumed to have not received gene flow from *P3*, *P2* represent a potential recipient population of introgression from the *P3* donor population, and an outgroup is used to polarize the ancestral states. *D+* specifically utilizes four site patterns: *ABBA, BABA, BAAA,* and *ABAA* (where an *A* denotes the ancestral allele and *B* denotes the derived allele) to test for asymmetries in site pattern frequencies. Under a scenario of no gene flow the *D+* statistic is expected to be zero, a significant and positive *D+* value indicates that *P2* and *P3* share more derived and ancestral alleles than expected, which may be explained by introgression. For all *D+* tests we used the ancestral allele calls from the six primate alignment inferred from EPO pipeline to polarize ancestral states [Paten et al., 2008a; Paten et al., 2008b; Herrero et al., 2016]. Since the *D+* statistic is normally distributed, to assess if observed *D+* values significantly differed from zero at the 742kb longest introgressed tract region and at the focal 72kb *MUC19* region, for each test we constructed *Z*-distributions of *D+* values from 742kb and 72kb non-overlapping windows with comparable effective sequence length and computed *P-values* using the *scipy.stats.norm.sf* function implemented in *scipy v1.7.2* [Virtanen et al., 2020] and a *P-value* less than 0.05 is considered significant.

### *Patterns of Allele Sharing Between the Introgressed Haplotype in MXL and the Archaics*

To further investigate the signals of introgression at *MUC19* we used *D+* as a complementary approach to our sequence divergence results. While using all sites to compute sequence divergence uses more information, this extra information may add noise due to new mutations arising in both the recipient and donor population since the introgression event, thus we calculated *D+* for the 742kb longest introgressed tract region in MXL and the focal 72kb *MUC19* region to further identify the most likely donor of the introgressed haplotye in MXL. To test hypotheses about introgression at each region we calculated *D+* for all combinations of *P1 = {YRI}, P2 = {MXL individual (NA19664) who harbors two copies of the introgressed haplotype}, and P3 = {Altai Denisovan, Altai Neanderthal, Chagyrskaya Neanderthal, Vindija Neanderthal}* at the focal regions, and compared the observed *D+* values to the genomic background as previously described above.

### *Gene Flow between Denisovans and Late Neanderthals*

To assess the possibility of gene flow between Denisovans and the late Neanderthals at the 72kb *MUC19* region we performed *D+* tests of introgression among the archaic individuals. To test this hypothesis we computed *D+* for all combinations of *P1 = {Altai Neanderthal}, P2 = {Chagyrskaya Neanderthal, Vindija Neanderthal}, and P3 = {Altai Denisovan}* at the focal 72kb region, and compared the observed *D+* values to the genomic background as previously described above.

## Heterozygosity in the Archaics and TGP

To understand if the 72kb *MUC19* region is an outlier for heterozygosity in the four archaic genomes we counted the number of heterozygous sites for the 72kb *MUC19* region and compared the observed number of heterozygous sites to a distribution of heterozygous site counts from the genomic background of 72kb windows with comparable effective sequence length density—i.e., within one standard deviation from the mean of the windowed effective sequence length distribution. To calculate *P-values*, we determined the proportion of windows from the genomic background where the number of heterozygous sites is greater than or equal to what we observed at the 72kb *MUC19* region, where a *P-value* less than 0.05 is considered significant. Additionally, we were interested in understanding if African individuals and individuals carrying the *Denisovan-like* haplotype at the 72kb region are also outliers for heterozygosity in our focal 72kb *MUC19* region. To do so we computed the average number of heterozygous sites amongst all African individuals (n = 504), heterozygous individuals (n = 255) who carry exactly one copy of the *Denisovan-like* haplotype at the 72kb region, and homozygous individuals (n = 16) who carry two copies of the *Denisovan-like* haplotype at the 72kb region, and compared our observed values to distributions of average heterozygous site counts from the genomic background of 72kb windows with comparable effective sequence length density—i.e., within one standard deviation from the mean of the windowed effective sequence length distribution. To calculate *P-values* for the African and heterozygous individuals, we determine the proportion of windows from the genomic background with an average number of heterozygous sites greater than or equal to what is observed at the 72kb *MUC19* region, and to calculate *P-values* for homozygous individuals we determine the proportion of windows from the genomic background with an average number of heterozygous sites less than or equal to what is observed at the 72kb *MUC19* region, where a *P-value* less than 0.05 is considered significant.

## Phasing the Archaic Genomes at *MUC19*

### *Benchmarking Phasing with a Synthetic Neanderthal*

To evaluate the feasibility of phasing a late Neanderthal at the core 72kb region, we first generated a Synthetic Neanderthal from sampling one allele from the Altai Denisovan and the other allele from the Altai Neanderthal using the TGP and all archaics combined dataset. This Synthetic Neanderthal harbors 155 heterozygous sites, representing the intersection of sites where the Altai Denisovan and Altai Neanderthal are fixed for different allelic states and those that are heterozygous in at least one of the late Neanderthals—i.e., the Chagyrskaya and Vindija Neanderthals. We then used BEAGLE v5.4 [Browning et al., 2021] to phase this Synthetic Neanderthal at the 72kb region, using TGP individuals as the reference panel and including the Synthetic Neanderthal, Altai Denisovan, and Altai Neanderthal in the phasing panel. Our analysis demonstrated that the Synthetic Neanderthal could be perfectly phased at the core 72kb region, which provided the motivation to phase each of the late Neanderthals at this region (see Supplementary Section S3 for a detailed discussion).

### *Phasing the Late Neanderthals at the 72kb MUC19 region*

Building on our benchmarking results, which demonstrated that we could perfectly phase a Synthetic Neanderthal at the core 72kb region (see Supplementary Section S3), we implemented a two-step approach to phase the late Neanderthals—i.e., the Chagyrskaya and Vindija Neanderthals—at the 72kb region using the combined dataset of TGP and all archaic genomes. In the first step, we statistically phased the heterozygous sites for each late Neanderthal that overlapped with segregating sites in the TGP or with sites where the Altai Denisovan and Altai

Neanderthal were fixed for different allelic states. To do so, we utilized BEAGLE v5.4, with TGP individuals serving as the reference panel, and included the late Neanderthal, Altai Neanderthal, and Denisovan in the phasing panel [Browning et al., 2021]. In the second step, we attempted to resolve the remaining heterozygous sites in each late Neanderthal that could not be statistically phased—i.e., the sites that are invariant in the TGP and did not overlap with a fixed difference between the Altai Denisovan and Altai Neanderthal. For these sites, we used a read-based phasing approach. Using IGV v2.8.10, we inspected reads from the BAM files and inferred haplotypes based on reads overlapping adjacent heterozygous sites that had been statistically phased. These inferred read-based haplotypes were then validated by checking their consistency with the phase of adjacent heterozygous sites determined by BEAGLE v5.4 (see Supplementary Section S4 for a detailed discussion).

Psuedo-Ancestry Painting (*PAP*) in the late Neanderthals

As genome-wide phasing is not currently possible for archaic genomes, we calculated Pseudo-Ancestry Painting (*PAP*) scores in order to assign alleles from a target heterozygous individual to haplotypes with fixed differences from two source individuals. Specifically, let $T = \{het_1, \ldots, het_n\}$ denote the set of all $n$ heterozygous sites (i.e., $het_i$) in a target individual for a given region, and let $S^1 = \{aac_1, \ldots, aac_n\}$ and $S^2 = \{aac_1, \ldots, aac_n\}$ denote the alternative allele count where $aac_i \in \{-1, 0, 1, 2\}$ (note that $-1$ represents a missing genotype due to not passing quality control in that given individual), at all of the heterozygous sites in $T$ for two source individuals, respectively. The *PAP* score corresponds to the number of heterozygous sites in $T$ that can be explained by the two source individuals, normalized over the total number of heterozygous sites, and is defined as:

$$PAP\ Score = \frac{1}{n}\sum_{i=1}^{n} \quad 1_{\{S_i^1 \neq S_i^2\}},$$

where $1_{\{S_i^1 \neq S_i^2\}}$ is an indicator variable that is defined as:

$$1_{\{S_i^1 \neq S_i^2\}} = 1, \text{ if } S_i^1 = 0 \text{ and } S_i^2 = 2,$$
$$1_{\{S_i^1 \neq S_i^2\}} = 1, \text{ if } S_i^1 = 2 \text{ and } S_i^2 = 0,$$
$$1_{\{S_i^1 \neq S_i^2\}} = 0, \text{ otherwise.}$$

We aimed to explain the excess of heterozygosity in the Chagyrskaya and Vindija Neanderthals targets, by calculating *PAP* scores using a pairing of the Altai Neanderthal and Denisovan as sources, as well as a pairing of an MXL individual (i.e., NA19664) who is homozygous for the *Denisovan-like* haplotype at the 72kb region and an YRI individual (i.e., NA19190) who is homozygous for the *Human-like* haplotype. To ensure that the *PAP* scores are properly behaved we computed additional *PAP* configurations where the Altai Neanderthal and Denisovan are the target individuals and the focal MXL (i.e., NA19664) and YRI (i.e., NA19190) individuals are sources, as a negative control experiment. For each configuration, to assess if *PAP* scores are significantly elevated at the focal 72kb *MUC19* region we compared the observed *PAP* scores, per configuration, to a distribution of *PAP* scores from the genomic background of 72kb non-overlapping windows with comparable effective sequence lengths—i.e., within one standard deviation from the mean of the windowed effective sequence length distribution. To calculate *P*-

*values* per configuration, we determined the proportion of windows from the genomic background with a *PAP* score greater than or equal to what we observed at the 72kb *MUC19* region, where a *P-value* less than 0.05 is considered significant.

## Copy Number Polymorphism of a 30bp Tandem Repeat Motif Between the *Human-like* and *Archaic* haplotypes

### *Repeat Counts from Short-Read Data*

We followed previously established methods to estimate repeat length from short-read data [Course et al., 2021]. We used the view command implemented in SAMtools v1.9 to extract and count reads from CRAM files for each sample that maps to the repeat region (hg38, Chr12:40482139-40491565). This process was repeated for two non-repetitive control regions of the human genome (hg38, Chr7:5500000-5600000 and Chr12:6490000-6590000) to calculate the average read density for each sample. The fraction of enrichment or depletion of reads in the repeat region, relative to the control regions, was used to estimate the repeat length and average number of repeat copies compared to the reference human genome, which contains 287.5 copies of the 30bp repeat. After estimating the number of repeat copies for each TGP individual, we determined the number of inferred introgressed tracts overlapping the repeat region (hg19, Chr12:40876395-40885001) in the same manner as described in the *"Identification of the MUC19 Introgressed Region"* Methods section. We defined outlier individuals with an elevated number of repeat copies as those having more than 487 repeat copies, corresponding to the 95th percentile of the TGP repeat copy number distribution at *MUC19*.

To investigate the relationship between copy number variation and introgression at *MUC19*, we first partitioned all TGP individuals into two subsets, each consisting of two groups for hypothesis testing: 1) admixed American vs non-admixed American individuals; and 2) individuals with at least one introgressed tract overlapping the repeat region vs individuals with no introgressed tracts overlapping the repeat region. For each subset, we calculated the proportion of outlier individuals within each group and performed a Proportions *Z*-test to determine if there was an enrichment of individuals with an elevated number of repeat copies between the groups, using the statsmodels.stats.proportion.proportions_ztest function implemented in statsmodels v0.13.2 [Seabold and Perktold 2010]. Additionally, for each subset we conducted a Mann-Whitney *U*-test to assess whether there was a significant difference in the distributions of repeat copies between the two groups, using the scipy.stats.mannwhitneyu function implemented in scipy v1.7.2 [Virtanen et al., 2020]. For both sets of statistical tests, *P-values* less than 0.05 were considered statistically significant.

Following the analysis using all TGP individuals, we then sought to further investigate the relationship between copy number variation and introgression at *MUC19*, by only considering the outlying individuals. To do so, we partitioned the 118 individuals with an elevated number of repeat copies into two different subsets, both consisting of two groups for hypothesis testing: 1) admixed American vs non-admixed American outlying individuals, and 2) outlying individuals with at least one introgressed tract overlapping the repeat region vs outlying individuals with no introgressed tracts overlapping the repeat region. For each subset, we computed the proportion of individuals within each group with respect to all outlier individuals and again performed a Proportions *Z*-test, using the statsmodels.stats.proportion.proportions_ztest function implemented

in statsmodels v0.13.2 [Seabold and Perktold 2010]. Similarly, we conducted a Mann-Whitney *U*-test to assess differences in the distributions of repeat copies between the two outlier groups, using the scipy.stats.mannwhitneyu function implemented in scipy v1.7.2 [Virtanen et al., 2020]. For both sets of statistical tests, *P-values* less than 0.05 were considered statistically significant.

To directly test the relationship between copy number variation and introgression at *MUC19*, we performed a series of correlation tests to investigate the association between an individual's number of introgressed tracts overlapping the repeat region (i.e., 0, 1, or 2 tracts) and the number of repeat copies. Specifically, we calculated Spearman's correlation coefficient for all TGP individuals, as well as for each super population and population, using the scipy.stats.spearmanr function implemented in scipy v1.7.2 [Virtanen et al., 2020]. Correlation coefficients were not computed for super populations or populations with no introgressed tracts overlapping the repeat region, and *P-values* less than 0.05 were considered statistically significant. Lastly, to directly assess the relationship between copy number variation and admixture in the Americas, we first computed the ancestry proportions for the repeat region for each admixed individual, which was done by intersecting each admixed American individual's local ancestry call BED files with the repeat region (hg19, Chr12:40876395-40885001) using BEDTools v2.31.0 [Quinlan and Hall 2010]. For each admixed American population, we then computed Spearman's correlation coefficient between an individual's ancestry proportion (i.e., 0%, 50%, or 100%) per ancestry component (i.e., Indigenous American, European, and African ancestry) and the number of repeat copies, again using the scipy.stats.spearmanr function implemented in scipy v1.7.2 [Virtanen et al., 2020]. After applying a Bonferroni correction to account for three multiple comparisons—i.e., one per ancestry component—*P-values* less than 0.0167 were considered statistically significant.

*Repeat Counts from Long-Read data*

Phased long-read genomes were obtained from the HPRC and HGSVC. A region corresponding to hg38 Chr12:40482543-40491234 was extracted from each of the FASTA files, and FASTA files with variation at those coordinates were extracted from Chr12:40479026-40491234. These regions were then trimmed to match the start and end of hg38 Chr12:40482593-40491199, to match the coordinates of the largest Simple Tandem Repeat from Tandem Repeat Finder [Benson, 1999]. The repeat length was divided by 30 and then averaged between the two haplotypes for each individual to calculate the estimated repeat copies. To ensure that the repeat copies inferred from long-read data was comparable to our inferences from short-read data we performed a least-squares linear regression and corresponding Pearson’s correlation coefficient between the estimated repeat copies for individuals with both types of sequencing data available using the scipy.stats.linregress function implemented in scipy v1.7.2 [Virtanen et al., 2020], where a *P-value* less than 0.05 was considered statistically significant.

## Denisovan-specific Coding Mutations

The focal 72kb *MUC19* region contains two Denisovan-specific synonymous mutations and nine Denisovan-specific missense mutations relative to the hg19 reference genome. To estimate the potential impact of these coding mutations, we annotated each mutation with its respective Grantham score. Grantham scores quantify the physicochemical distance between amino acids and are informative about how a mutation alters the protein’s functional properties. Following

the classification by Li et al. [1985], we categorized the effects of each mutation as follows: Grantham score < 50 = conservative, 51–100 = moderately conservative, 101–150 = moderately radical, and > 150 = radical. To assess the conservation of each substituted base, we annotated each coding mutation with the PhyloP score of the reference base (obtained from the UCSC genome browser 100 vertebrates Basewise Conservation track [Pollard et al., 2010]).

Since missense mutations change the amino acid identity, which may potentially affect the protein's function, we first aimed to determine whether admixed American populations are more likely to harbor Denisovan-specific missense mutations than non-admixed American populations. For each of the nine Denisovan-specific missense mutations, we computed the frequency of the mutation within both groups and performed a Proportions *Z*-test using the statsmodels.stats.proportion.proportions_ztest function implemented in statsmodels v0.13.2 [Seabold and Perktold 2010] to assess if there is an enrichment of these mutations among admixed populations. Additionally, we used Fisher's Exact Test using the scipy.stats.fisher_exact function as implemented in scipy v1.7.2 [Virtanen et al., 2020] to determine whether these missense mutations are overrepresented in admixed American populations compared to non-admixed Americans. African individuals were excluded from the non-admixed American population group in both analyses and *P-values* less than 0.05 were considered statistically significant. Next, we evaluated if there is an enrichment of Denisovan-specific missense mutations between three focal groups: MXL individuals from the TGP, Indigenous American individuals from the SGDP, and ancient Indigenous American individuals. For each of the nine Denisovan-specific missense mutations, we conducted pairwise comparisons between these three groups using the Proportions Z-test and Fisher's Exact Test, following the same approach as outlined above. For comparisons including ancient Indigenous American individuals, we only considered individuals who passed quality control at the respective site, and again, *P-values* less than 0.05 were considered statistically significant. To assess the relationship between the frequency of introgressed tracts overlapping *MUC19* and the frequency of Denisovan-specific missense mutations among TGP populations, we performed a least-squares linear regression and corresponding Pearson's correlation coefficient between the introgressed tract frequency and mean Denisovan-specific missense mutation frequency among TGP populations using the scipy.stats.linregress function implemented in scipy v1.7.2 [Virtanen et al., 2020], where a *P-value* less than 0.05 was considered statistically significant. Lastly, to test if recent admixture in the Americas has diluted the introgressed ancestry at *MUC19* we quantified the relationship between an individual's Indigenous American ancestry proportion at the focal 72kb region and the frequency of a Denisovan-specific missense mutation at position Chr12:40808726. Given that all nine of the Denisovan-speciifc missense mutations are found within a ~17.5kb region, we used the missense mutation at Chr12:40808726 as this position has genotype information in 20 out of the 23 ancient American individuals (Table S30). For each admixed American individual in the TGP we computed their respective Indigenous American ancestry proportion for the focal 72kb region in a similar manner as described in the previous methods section, and since all of the ancient American individuals pre-date the colonization events in the Americas we assumed their Indigenous American ancestry proportion is 100%. We then assessed the relationship between an individual's Indigenous American ancestry proportion and the frequency of the Denisovan-specific missense mutation at position Chr12:40808726 by performing a least-squares linear regression and corresponding Pearson's correlation coefficient using the scipy.stats.linregress function implemented in scipy v1.7.2 [Virtanen et al., 2020], where a *P-value* less than 0.05 was considered statistically significant.

## References and Notes

[1] KD Ahlquist, Mayra M Banuelos, Alyssa Funk, Jiaying Lai, Stephen Rong, Fernando A Villanea, and Kelsey E Witt. Our tangled family tree: new genomic methods offer insight into the legacy of archaic admixture. Genome biology and evolution, 13(7):evab115, 2021.

[2] Sharon R Browning, Brian L Browning, Ying Zhou, Serena Tucci, and Joshua M Akey. Analysis of human sequence data reveals two pulses of archaic denisovan admixture. Cell, 173(1):53–61, 2018.

[3] Fernando A Villanea and Joshua G Schraiber. Multiple episodes of interbreeding between neanderthal and modern humans. Nature ecology & evolution, 3(1):39, 2019.

[4] Melinda A Yang, Anna-Sapfo Malaspinas, Eric Y Durand, and Montgomery Slatkin. Ancient structure in Africa unlikely to explain neanderthal and non African genetic similarity. Molecular biology and evolution, 29(10):2987– 2995, 2012.

[5] Sriram Sankararaman, Swapan Mallick, Nick Patterson, and David Reich. The combined landscape of Denisovan and Neanderthal ancestry in present day humans. Current Biology, 26(9):1241–1247, 2016.

[6] Martin Petr, Svante Pääbo, Janet Kelso, and Benjamin Vernot. Limits of long-term selection against Neandertal introgression. Proceedings of the National Academy of Sciences, 116(5):1639–1644, 2019.

[7] Xinjun Zhang, Bernard Kim, Kirk E Lohmueller, and Emilia Huerta Sánchez. The impact of recessive deleterious variation on signals of adaptive introgression in human populations. Genetics, 215(3):799–812, 2020.

[8] Fernando Racimo, Sriram Sankararaman, Rasmus Nielsen, and Emilia Huerta-Sánchez. Evidence for archaic adaptive introgression in humans. Nature Reviews Genetics, 16(6):359, 2015.

[9] Xinjun Zhang, Bernard Kim, Armaan Singh, Sriram Sankararaman, Arun Durvasula, and Kirk E Lohmueller. Maladapt reveals novel targets of adaptive introgression from neanderthals and denisovans in worldwide human populations. Molecular Biology and Evolution, 40(1):msad001, 2023.

[10] Shaohua Fan, Matthew EB Hansen, Yancy Lo, and Sarah A Tishkoff. Going global by adapting local: A review of recent human adaptation. Science, 354(6308):54–59, 2016.

[11] Fernando L Mendez, Joseph C Watkins, and Michael F Hammer. A haplotype at STAT2 introgressed from Neanderthals and serves as a candidate of positive selection in papua new guinea. The American Journal of Human Genetics, 91(2):265–274, 2012.

[12] Sriram Sankararaman, Swapan Mallick, Michael Dannemann, Kay Prüfer, Janet Kelso, Svante Pääbo, Nick Patterson, and David Reich. The genomic landscape of Neanderthal ancestry in present-day humans. Nature, 507 (7492):354–357, 2014.

[13] Benjamin Vernot and Joshua M Akey. Resurrecting surviving Neandertal lineages from modern human genomes. Science, 343(6174):1017–1021, 2014.

[14] Rachel M Gittelman, Joshua G Schraiber, Benjamin Vernot, Carmen Mikacenic, Mark M Wurfel, and Joshua M Akey. Archaic hominin admixture facilitated adaptation to out-of-africa environments. Current Biology, 26(24):3375–3382, 2016.

[15] Aaron J Sams, Anne Dumaine, Yohann Nédélec, Vania Yotova, Carolina Alfieri, Jerome E Tanner, Philipp W Messer, and Luis B Barreiro. Adaptively introgressed Neandertal haplotype at the OAS locus functionally impacts innate immune responses in humans. Genome biology, 17(1):1–15, 2016.

[16] Michael Dannemann and Janet Kelso. The contribution of Neanderthals to phenotypic variation in modern humans. The American journal of human genetics, 101(4):578–589, 2017.

[17] Davide Marnetto and Emilia Huerta-Sánchez. Haplostrips: revealing population structure through haplotype visualization. Methods in Ecology and Evolution, 8(10):1389–1392, 2017.

[18] Emilia Huerta-Sánchez, Xin Jin, Zhuoma Bianba, Benjamin M Peter, Nico las Vinckenbosch, Yu Liang, Xin Yi, Mingze He, Mehmet Somel, Peixi ang Ni, et al. Altitude adaptation in tibetans caused by introgression of denisovan-like DNA. Nature, 512(7513):194, 2014.

[19] Fernando Racimo, David Gokhman, Matteo Fumagalli, Amy Ko, Tor ben Hansen, Ida Moltke, Anders Albrechtsen, Liran Carmel, Emilia Huerta-Sánchez, and Rasmus Nielsen. Archaic adaptive introgression in tbx15/wars2. Molecular biology and evolution, 34(3):509–524, 2017.

[20] Xinjun Zhang, Kelsey E Witt, Mayra M Bañuelos, Amy Ko, Kai Yuan, Shuhua Xu, Rasmus Nielsen, and Emilia Huerta-Sánchez. The history and evolution of the denisovan-epas1 haplotype in tibetans. Proceedings of the National Academy of Sciences, 118(22):e2020803118, 2021.

[21] Erika Tamm, Toomas Kivisild, Maere Reidla, Mait Metspalu, David Glenn Smith, Connie J Mulligan, Claudio M Bravi, Olga Rickards, Cristina Martinez-Labarga, Elsa K Khusnutdinova, et al. Beringian standstill and spread of Native American founders. PloS one, 2(9):e829, 2007.

[22] Hylke E Beck, Niklaus E Zimmermann, Tim R McVicar, Noemi Vergopolan, Alexis Berg, and Eric F Wood. Present and future köppen-geiger climate classification maps at 1-km resolution. Scientific data, 5(1):1–12, 2018.

[23] Xin Yi, Yu Liang, Emilia Huerta-Sánchez, Xin Jin, Zha Xi Ping Cuo, John E Pool, Xun Xu, Hui Jiang, Nicolas Vinckenbosch, Thorfinn Sand Korneliussen, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. science, 329(5987):75–78, 2010.

[24] Kelsey E Witt, Alyssa Funk, Valeria Añorve-Garibay, Lesly Lopez Fang, and Emilia Huerta-Sánchez. The impact of modern admixture on archaic human ancestry in human populations. Genome Biology and Evolution, 15 (5):evad066, 2023.

[25] Fernando Racimo, Davide Marnetto, and Emilia Huerta-Sánchez. Signatures of archaic adaptive introgression in present-day human populations. Molecular biology and evolution, 34(2):296–317, 2016.

[26] Austin W Reynolds, Jaime Mata-Míguez, Aida Miró-Herrans, Marcus Briggs-Cloud, Ana Sylestine, Francisco Barajas-Olmos, Humberto Garcia Ortiz, Margarita Rzhetskaya, Lorena Orozco, Jennifer A Raff, et al. Comparing signals of natural selection between three indigenous North American populations. Proceedings of the National Academy of Sciences, page 201819467, 2019.

[27] Laurits Skov, Ruoyun Hui, Vladimir Shchur, Asger Hobolth, Aylwyn Scally, Mikkel Heide Schierup, and Richard Durbin. Detecting archaic introgression using an unadmixed outgroup. PLoS Genetics, 14(9):e1007641, 2018.

[28] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E

Kenny. Human demographic history impacts genetic risk prediction across diverse populations. The American Journal of Human Genetics, 100(4):635–649, 2017.

[29] Medina-Muñoz SG, Ortega-Del Vecchyo D, Cruz-Hervert LP, Ferreyra-Reyes L, García-García L, Moreno-Estrada A, Ragsdale AP. Demographic modeling of admixed Latin American populations from whole genomes. The American Journal of Human Genetics;110(10):1804-16, 2023.

[30] Szpiech ZA. selscan 2.0: scanning for sweeps in unphased data. Bioinformatics. 40(1):btae006, 2024.

[31] Lesly Lopez Fang, Diego Ortega-Del Vecchyo, Emily Jane McTavish, and Emilia Huerta-Sánchez. Leveraging shared ancestral variation to detect local introgression. bioRxiv, 2022. doi: 10.1101/2022.03.21.485082.

[32] David Peede, Diego Ortega-Del Vecchyo, and Emilia Huerta-Sánchez. The utility of ancestral and derived allele sharing for genome-wide inferences of introgression. bioRxiv, 2022. doi: 10.1101/2022.12.02.518851.

[33] Linda Ongaro, Emilia Huerta-Sánchez. A history of multiple Denisovan introgression events in modern humans. Nature Genetics, 5:1-1, 2024.

[34] Viviane Slon, Fabrizio Mafessoni, Benjamin Vernot, Cesare De Filippo, Steffi Grote, Bence Viola, Mateja Hajdinjak, Stéphane Peyrégne, Sarah Nagel, Samantha Brown, et al. The genome of the offspring of a Neanderthal mother and a Denisovan father. Nature, 561(7721):113–116, 2018.

[35] Benjamin M Peter. 100,000 years of gene flow between Neandertals and Denisovans in the Altai mountains. bioRxiv, 2020. doi: 10.1101/2020.03.13.990523.

[36] Richard Grantham. Amino acid difference formula to help explain protein evolution. science, 185(4154):862–864, 1974.

[37] Wen-Hsiung Li, Chung-I Wu, and Chi-Cheng Luo. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Molecular biology and evolution, 2(2):150–174, 1985.

[38] Katherine S Pollard, Melissa J Hubisz, Kate R Rosenbloom, and Adam Siepel. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research, 20(1):110–121, 2010.

[39] Gabriel Javitt, Lev Khmelnitsky, Lis Albert, Lavi Shlomo Bigman, Nadav Elad, David Morgenstern, Tal Ilani, Yaakov Levy, Ron Diskin, and Deborah Fass. Assembly mechanism of mucin and von willebrand factor polymers. Cell, 183(3):717–729, 2020.

[40] Duo Xu, Pavlos Pavlidis, Recep Ozgur Taskent, Nikolaos Alachiotis, Colin Flanagan, Michael DeGiorgio, Ran Blekhman, Stefan Ruhl, and Omer Gokcumen. Archaic hominin introgression in Africa contributes to functional salivary MUC7 genetic variation. Molecular Biology and Evolution, 34(10):2704–2715, 2017.

[41] Almarri MA, Bergström A, Prado-Martinez J, Yang F, Fu B, Dunham AS, Chen Y, Hurles ME, Tyler-Smith C, Xue Y. Population structure, stratification, and introgression of human structural variation. Cell, 182(1):189-99, 2020.

[42] Ebert P, Audano PA, Zhu Q, Rodriguez-Martin B, Porubsky D, Bonder MJ, Sulovari A, Ebler J, Zhou W, Serra Mari R, Yilmaz F. Haplotype-resolved diverse human genomes and integrated analysis of structural variation. Science, 372(6537):eabf7117, 2021.

[43] Ragsdale, A. P., Weaver, T. D., Atkinson, E. G., Hoal, E. G., Möller, M., Henn, B. M., & Gravel, S. A weakly structured stem for human origins in Africa. Nature, 617(7962), 755-763, 2023

[44] Wang, K., Mathieson, I., O'Connell, J., & Schiffels, S. Tracking human population structure through time from whole genome sequences. PLoS genetics, 16(3), e1008552, 2020.

[45] Lucas Henriques Viscardi, Vanessa Rodrigues Paixao-Cortes, David Comas, Francisco Mauro Salzano, Diego Rovaris, Claiton Dotto Bau, Carlos Eduardo G Amorim, and Maria Catira Bortolini. Searching for ancient balanced polymorphisms shared between Neanderthals and modern humans. Genetics and Molecular Biology, 41:67–81, 2018.

[46] Posth, Cosimo, Christoph Wißing, Keiko Kitagawa, Luca Pagani, Laura van Holstein, Fernando Racimo, Kurt Wehrberger et al. "Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals." Nature communications 8(1): 16046, 2017.

[47] Petr, Martin, Mateja Hajdinjak, Qiaomei Fu, Elena Essel, Hélène Rougier, Isabelle Crevecoeur, Patrick Semal et al. "The evolutionary history of Neanderthal and Denisovan Y chromosomes." Science 369(65110): 1653-1656, 2020.

[48] Stéphane Peyrégne, Janet Kelso, Benjamin M Peter, and Svante Pääbo. The evolutionary history of human spindle genes includes back-and-forth gene flow with Neandertals. Elife, 11:e75464, 2022.

[49] Petar Pajic, Shichen Shen, Jun Qu, Alison J May, Sarah Knox, Stefan Ruhl, and Omer Gokcumen. A mechanism of gene evolution generating mucin function. Science advances, 8(34):eabm8757, 2022.

[50] Fernando A Villanea and Kelsey E Witt. Underrepresented populations at the archaic introgression frontier. Frontiers in Genetics, 13:821170, 2022.

**Supplement References**

[51] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature, 526(7571):68, 2015.

[52] Swapan Mallick, Heng Li, Mark Lipson, Iain Mathieson, Melissa Gymrek, Fernando Racimo, Mengyao Zhao, Niru Chennagiri, Susanne Nordenfelt,Arti Tandon, et al. The simons genome diversity project: 300 genomes from 142 diverse populations. Nature, 538(7624):201–206, 2016.

[53] Karen HY Wong, Walfred Ma, Chun-Yu Wei, Erh-Chan Yeh, Wan-Jia Lin, Elin HF Wang, Jen-Ping Su, Feng-Jen Hsieh, Hsiao-Jung Kao, Hsiao Huei Chen, et al. Towards a reference genome that captures global genetic diversity. Nature communications, 11(1):5482, 2020.

[54] Benedict Paten, Javier Herrero, Kathryn Beal, Stephen Fitzgerald, and Ewan Birney. Enredo and pecan: genome-wide mammalian consistency based multiple alignment with paralogs. Genome research, 18(11):1814– 1828, 2008.

[55] Benedict Paten, Javier Herrero, Stephen Fitzgerald, Kathryn Beal, Paul Flicek, Ian Holmes, and Ewan Birney. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome research, 18(11):1829–1843, 2008.

[56] Javier Herrero, Matthieu Muffato, Kathryn Beal, Stephen Fitzgerald, Leo Gordon, Miguel Pignatelli, Albert J Vilella, Stephen MJ Searle, Ridwan Amode, Simon Brent, et al. Ensembl comparative genomics resources. Database, 2016:bav096, 2016.

[57] Prüfer K, De Filippo C, Grote S, Mafessoni F, Korlević P, Hajdinjak M, Vernot B, Skov L, Hsieh P, Peyrégne S, Reher D. A high-coverage Neandertal genome from Vindija Cave in Croatia. Science, 358(6363):655-8, 2017.

[58] Mafessoni F, Grote S, De Filippo C, Slon V, Kolobova KA, Viola B, Markin SV, Chintalapati M, Peyrégne S, Skov L, Skoglund P. A high-coverage Neandertal genome from Chagyrskaya Cave. Proceedings of the National Academy of Sciences, 117(26):15132-6, 2020.

[59] Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research, 35(suppl_1):D61-5, 2007.

[60] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. fly, 6(2):80–92, 2012.

[61] CL Scheib, Hongjie Li, Tariq Desai, Vivian Link, Christopher Kendall, Genevieve Dewar, Peter William Griffith, Alexander Mörseburg, John R Johnson, Amiee Potter, et al. Ancient human parallel lineages within North America contributed to a coastal expansion. Science, 360(6392):1024–1027, 2018.

[62] John Lindo, Randall Haas, Courtney Hofman, Mario Apata, Mauricio Moraga, Ricardo A Verdugo, James T Watson, Carlos Viviano Llave, David Witonsky, Cynthia Beall, et al. The genetic prehistory of the andean highlands 7000 years bp though european contact. Science advances, 4(11): eaau4921, 2018.

[63] Constanza De la Fuente, María C Ávila-Arcos, Jacqueline Galimany, Meredith L Carpenter, Julian R Homburger, Alejandro Blanco, Paloma Contreras, Diana Cruz D´avalos, Omar Reyes, Manuel San Roman, et al. Genomic insights into the origin and diversification of late maritime hunter gatherers from the chilean patagonia. Proceedings of the National Academy of Sciences, 115(17):E4006–E4012, 2018.

[64] J Víctor Moreno-Mayar, Ben A Potter, Lasse Vinner, Matthias Steinrücken, Simon Rasmussen, Jonathan Terhorst, John A Kamm, Anders Albrechtsen, Anna-Sapfo Malaspinas, Martin Sikora, et al. Terminal pleistocene alaskan genome reveals first founding population of native americans. Nature, 553 (7687):203, 2018.

[65] Morten Rasmussen, Sarah L Anzick, Michael R Waters, Pontus Skoglund, Michael DeGiorgio, Thomas W Stafford Jr, Simon Rasmussen, Ida Moltke, Anders Albrechtsen, Shane M Doyle, et al. The genome of a late pleistocene human from a clovis burial site in western montana. Nature, 506(7487): 225, 2014.

[66] Viridiana Villa-Islas, Alan Izarraras-Gomez, Maximilian Larena, Elizabeth Mej´ıa Perez Campos, Marcela Sandoval-Velasco, Juan Esteban Rodríguez Rodríguez, Miriam Bravo-Lopez, Barbara Moguel, Rosa Fregel, Ernesto Garfias-Morales, et al. Demographic history and genetic structure in pre hispanic central mexico. Science, 380(6645):eadd6142, 2023.

[67] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows–wheeler transform. bioinformatics, 25(14):1754–1760, 2009.

[68] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. Bioinformatics, 25(16):2078– 2079, 2009.

[69] Coll Macià M, Skov L, Peter BM, Schierup MH. Different historical generation intervals in human populations inferred from Neanderthal fragment lengths and mutation signatures. Nature Communications,12(1):5317, 2021.

[70] Seabold, S., & Perktold, J. Statsmodels: econometric and statistical modeling with python. SciPy, 7(1), 2010.

[71] Witt KE, Villanea F, Loughran E, Zhang X, Huerta-Sanchez E. Apportioning archaic variants among modern populations. Philosophical Transactions of the Royal Society B, 377(1852):20200411, 2023.

[72] Gaurav Bhatia, Nick Patterson, Sriram Sankararaman, and Alkes L Price. Estimating and interpreting fst: the impact of rare variants. Genome research, 23(9):1514–1521, 2013.

[73] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS biology, 4(3):e72, 2006.

[74] Szpiech ZA. selscan 2.0: scanning for sweeps in unphased data. Bioinformatics, 40(1):btae006, 2024.

[75] International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature, 449(7164):851, 2007.

[76] Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. Nature methods, 17(3):261–272, 2020.

[77] Quinlan, A. R., & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics, 26(6), 841-842, 2010.

[78] Benson, G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic acids research, 27(2), 573-580, 1999.

[79] Pollard, Katherine S., Melissa J. Hubisz, Kate R. Rosenbloom, and Adam Siepel. "Detection of nonneutral substitution rates on mammalian phylogenies." Genome research 20(1): 110-121, 2010.

[80] Meredith M Course, Arvis Sulovari, Kathryn Gudsnuk, Evan E Eichler, and Paul N Valdmanis. Characterizing nucleotide variation and expansion dynamics in human-specific variable number tandem repeats. Genome Research, 31(8):1313–1324, 2021.

[81] Benjamin C Haller and Philipp W Messer. Slim 3: Forward genetic simulations beyond the wright–fisher model. Molecular biology and evolution, 36(3):632–637, 2019.

[82] Jennifer Harrow, Adam Frankish, Jose M Gonzalez, Electra Tapanari, Mark Diekhans, Felix Kokocinski, Bronwen L Aken, Daniel Barrell, Amonida Zadissa, Stephen Searle, et al. Gencode: the reference human genome an notation for the encode project. Genome research, 22(9):1760–1774, 2012.

[83] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Car los D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. PLoS genetics, 5 (10):e1000695, 2009.

[84] Anna-Sapfo Malaspinas, Michael C Westaway, Craig Muller, Vitor C Sousa, Oscar Lao, Isabel Alves, Anders Bergström, Georgios Athanasiadis, Jade Y Cheng, Jacob E Crawford, et al. A genomic history of aboriginal australia. Nature, 538(7624):207–214, 2016.

[85] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Ga bor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, 1000 Genomes Project, Carlos D Bustamante, et al. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences, 108(29):11983–11988, 2011.

[86] David Reich, Richard E Green, Martin Kircher, Johannes Krause, Nick Patterson, Eric Y Durand, Bence Viola, Adrian W Briggs, Udo Stenzel, Philip LF Johnson, et al. Genetic history of an archaic hominin group from denisova cave in siberia. Nature, 468(7327):1053–1060, 2010.

[87] Bernard Y Kim, Christian D Huber, and Kirk E Lohmueller. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. Genetics, 206(1):345–361, 2017.

[88] Anjali G Hinch, Arti Tandon, Nick Patterson, Yunli Song, Nadin Rohland, Cameron D Palmer, Gary K Chen, Kai Wang, Sarah G Buxbaum, Ermeg L Akylbekova, et al. The landscape of recombination in african americans. Nature, 476(7359):170–175, 2011.

[89] Santiago G Medina-Munoz, Diego Ortega-Del Vecchyo, Luis Pablo Cruz Hervert, Leticia Ferreyra-Reyes, Lourdes Garcia-Garcia, Andres Moreno Estrada, and Aaron Ragsdale. Demographic modeling of admixed latin american populations from whole genomes. bioRxiv, pages 2023–03, 2023.

[90] Alistair Miles and NJ Harding. scikit-allel: A python package for exploring and analysing genetic variation data, 2016.

[91] Guy S Jacobs, Georgi Hudjashov, Lauri Saag, Pradiptajati Kusuma, Chelzie C Darusallam, Daniel J Lawson, Mayukh Mondal, Luca Pagani, Fran¸cois-Xavier Ricaut, Mark Stoneking, et al. Multiple deeply divergent denisovan ancestries in papuans. Cell, 2019.

[92] Rose MC, Voynow JA. Respiratory tract mucin genes and mucin glycoproteins in health and disease. Physiological reviews, 86(1):245-78, 2006.

[93] Yin Chen, Yu Hua Zhao, Tejas Baba Kalaslavadi, Edward Hamati, Keith Nehrke, Anh Dao Le, David K Ann, and Reen Wu. Genome-wide search and identification of a novel gel-forming mucin muc19/muc19 in glandular tissues. American journal of respiratory cell and molecular biology, 30(2): 155–165, 2004.

[94] Joseph Edward Kerschner. Mucin gene expression in human middle ear epithelium. The Laryngoscope, 117(9):1666–1676, 2007.

[95] DF Yu, Y Chen, JM Han, H Zhang, XP Chen, WJ Zou, LY Liang, CC Xu, and ZG Liu. Muc19 expression in human ocular surface and lacrimal gland and its alteration in Sjögren syndrome patients. Experimental eye research, 86(2):403–411, 2008.

[96] Takagawa T, Kitani A, Fuss I, Levine B, Brant SR, Peter I, Tajima M, Nakamura S, Strober W. An increase in LRRK2 suppresses autophagy and enhances Dectin-1–induced immunity in a mouse model of colitis. Science translational medicine, 10(444):eaan8162, 2018.

[97] Hamid, I., Korunes, K. L., Beleza, S., & Goldberg, A. Rapid adaptation to malaria facilitated by admixture in the human population of Cabo Verde. Elife, 10, e63177, 2021.

[98] Langergraber KE, Prüfer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, Robbins MM. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. Proceedings of the National Academy of Sciences, 109(39):15716-21, 2012.

[99] Prüfer K, Racimo F, Patterson N, Jay F, Sankararaman S, Sawyer S, Heinze A, Renaud G, Sudmant PH, De Filippo C, Li H. The complete genome sequence of a Neanderthal from the Altai Mountains. Nature;505(7481):43-9, 2014

**Acknowledgments:** We would like to thank Alyssa Funk for contributing to the development of the PBS analysis, Ratchanon Pornmongkolsuk for early visualizations of global frequencies of MUC19, and Diego Ortega del Vecchyo and Paolo Provero for their insightful comments and discussion. We would also like to thank the Crawford and Ramachandran laboratories, especially Ria Vinod, Julian Stamp, Chibuikem Nwizu, Cole Williams, and Leah Darwin for their invaluable feedback and support throughout the duration of this project. Part of this research was conducted using computational resources and services at the Center for Computation and Visualization, Brown University.

**Funding:**

The Leakey Foundation (to FAV).

National Institutes of Health (1R35GM128946-01 to EHS).

Alfred P. Sloan Foundation (to EHS).

Brown University Predoctoral Training Program in Biological Data Science (NIH T32 GM128596 to DP and ETC).

National Institutes of Health (R35GM142978 to PM).

Burroughs Wellcome Fund (Career Award at the Scientific Interface to PM).

National Institutes of Health (R01NS122766 to PNV).

Human Frontier Science Program (to EHS, FJ, and MAA).

**Author contributions:**

Conceptualization: FAV, DP, EHS

Formal analysis: FAV, DP, EJK, VAG, KEW, VVI, RZ, DM, PM, FJ, PNV, MAA, EHS

Supervision: DM, PM, FJ, PNV, MAA, EHS

Writing – original draft: FAV, DP, EHS

Writing – review & editing: EJK, VAG, KEW, VVI, RZ, DM, PM, FJ, PNV, MAA

**Competing interests:** Authors declare that they have no competing interests.

**Data and materials availability:** The 1,000 Genomes Project Phase III, Simons Genome Diversity Project, high-coverage archaic genomes, Human Pangenome Reference Consortium, and Human Genome Structural Variant Consortium datasets are all publicly available. Ancient American genomes are available after signing data agreements from the original publications. All software used in this study is publicly available, and all statistical tests are described in the methods. All the information needed to reproduce the results in this study is described in the methods and supplemental methods; additionally, the original code can be found at: https://github.com/David-Peede/MUC19/tree/main.

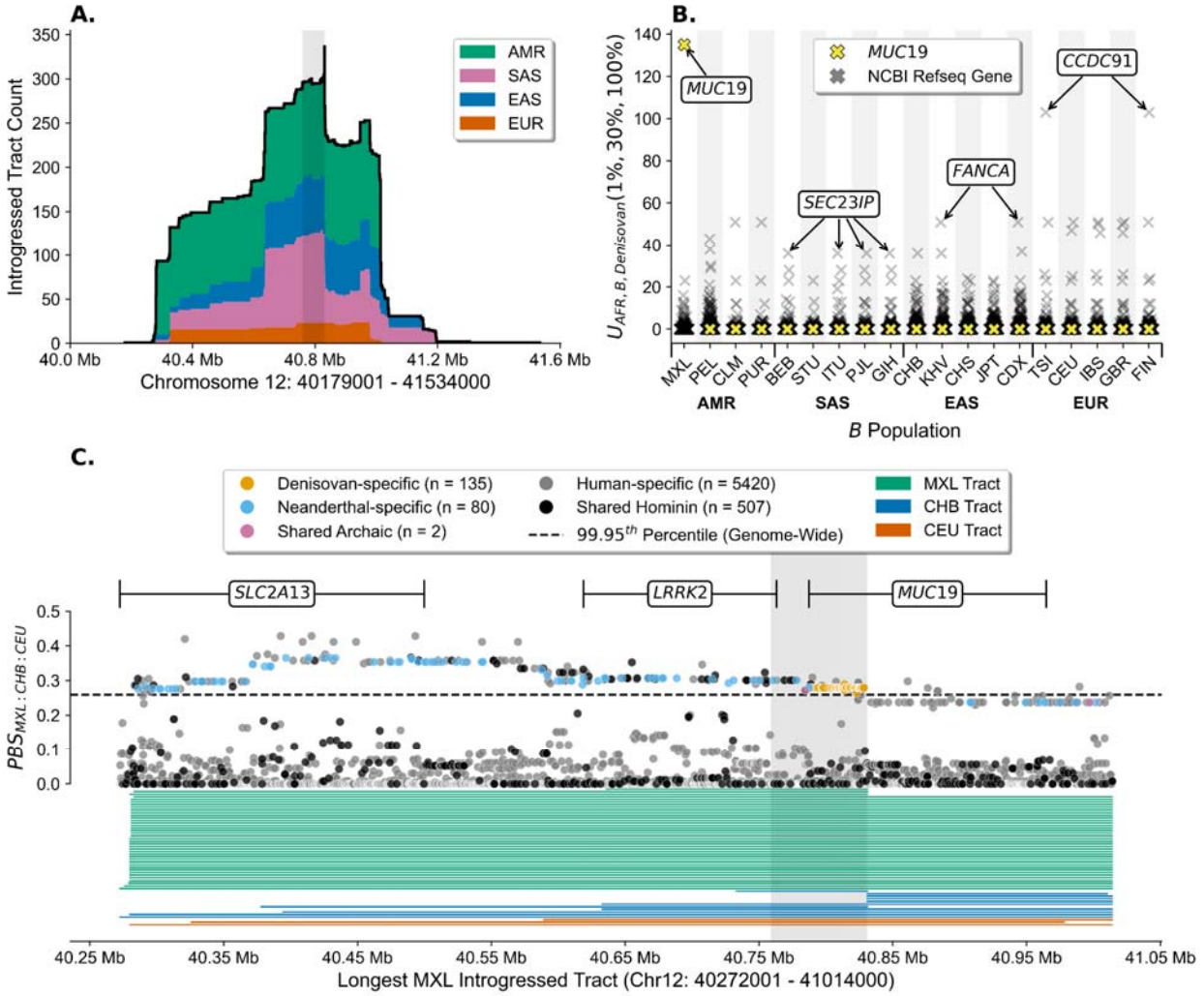



**Figure 1. Signals of adaptive introgression at *MUC19*.**

(**A**) Density of introgressed tracts inferred using `hmmix` that overlap *MUC19* for the TGP (black outline) and stratified by super population—Admixed Americans (AMR) in bluish green, South Asians (SAS) in reddish purple, East Asians (EAS) in blue, and Europeans (EUR) in vermillion. The gray shaded region corresponds to the focal 72kb region, which is the densest contiguous region of introgressed tracts longer than 40kb. (**B**) $U_{AFR,B,Denisovan}(1\%, 30\%, 100\%)$ values for each non-African population, stratified by super population, per NCBI Refseq gene (gray X's), where *MUC19* is denoted as a yellow X. (**C**) Population Branch Statistic (*PBS*) for the Mexican population (MXL) in the TGP using the Han Chinese (CHB) and Central European (CEU) populations in the TGP as control populations ($PBS_{MXL:CHB:CEU}$) for all SNPs in the 742kb region that corresponds to the longest introgressed tract found in MXL. The orange, sky blue, and reddish purple points represent SNPs that are rare or absent in Africa (<1%), present in MXL (>1%), and are, respectively, either shared uniquely with the Denisovan, uniquely with the Neanderthals, or shared with both the Denisovan and Neanderthals (see Methods). The black points represent SNPs present across both modern human populations and the archaics, while the gray points represent SNPs private to modern humans. The black dashed line represents the 99.95th percentile of $PBS_{MXL:CHB:CEU}$ scores for all SNPs genome-wide, and the gray shaded region corresponds to the focal 72kb region—the same gray shaded region in panel A. The

*MUC19* and *LRRK2* genes are fully encompassed within the 742kb region, while ~65% of *SLC2A13* overlaps the 742kb region. Below the $PBS_{MXL:CHB:CEU}$ points are the introgressed tracts for MXL (bluish green), CHB (blue), and CEU (vermillion) sorted from shortest to longest within each population. Note that the $PBS_{MXL:CHB:CEU}$ values for SNPs above the 99.95th percentile before *LRRK2* are larger than the values for the Denisovan-specific SNPs in the focal 72kb region due to larger allele frequency differences between MXL and the two control populations. It should also be noted that all 135 Denisovan-specific SNPs are sequestered within the focal 72kb region, while only 4 out of the 80 Neanderthal-specific SNPs are found within the focal 72kb region.

Python code to replicate this figure is available at: https://github.com/David-Peede/MUC19/blob/main/figure_nbs/figure_1_v_revisions.ipynb

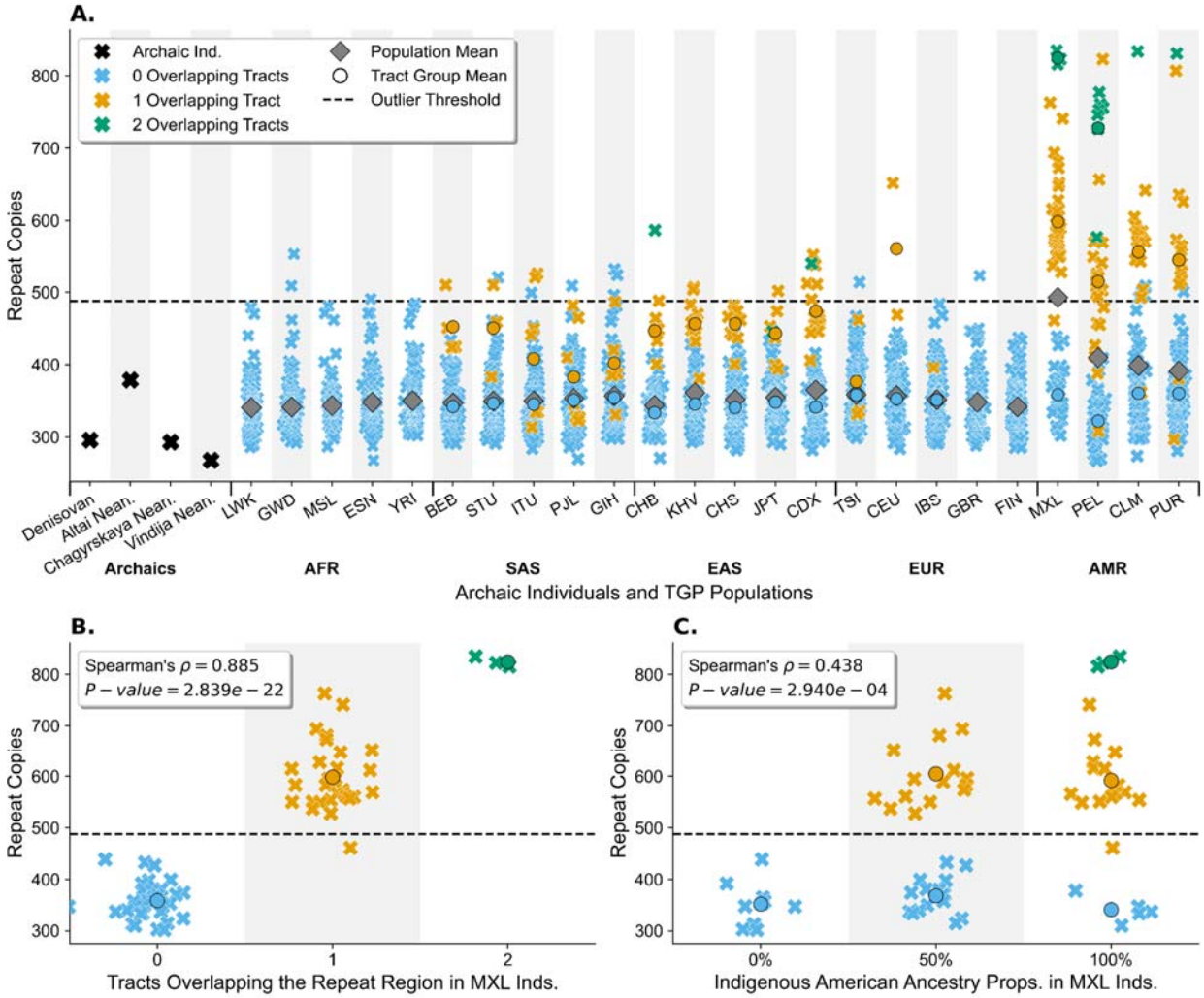



**Figure 2. Copy number variation of a 30 base pair variable number tandem repeat motif in the TGP individuals at *MUC19*.**

(**A**) Average number of repeat copies between an individual's two chromosomes for archaic individuals (black X's), individuals who do not harbor an introgressed tract (sky blue X's), individuals with one introgressed tract (yellow X's), and individuals with two introgressed tracts (bluish green X's) determined by the number of introgressed tracts inferred using `hmmix` overlapping the *MUC19* VNTR, for each population in the TGP. The mean number of repeat copies stratified by population is denoted by a grey diamond and the average number of repeat copies amongst individuals who carry exactly zero, one, and two introgressed tracts are denoted by sky blue, yellow, and bluish green circles respectively and are stratified by population. The black dashed line denotes the outlier threshold, which corresponds to the 95th percentile of the TGP repeat copies distribution. Repeat copies appeared similar to the reference human genome (287.5 copies) in the Altai Denisovan (296 copies) and Altai (379 copies), Vindija (268 copies), and Chagyrskaya (293 copies) Neanderthal archaic genomes. (**B**) The relationship between the average number of repeat copies between a MXL individual's two chromosomes and the number of introgressed tracts overlapping the *MUC19* VNTR region. Note that there is a significant positive correlation between the number of repeat copies and the number of introgressed tracts

present in an MXL individual (Spearman's $\rho$: 0.885; *P-value*: 2.839e-22). (**C**) The relationship between the average number of repeat copies between a MXL individual's two chromosomes and the proportion of Indigenous American ancestry at the *MUC19* VNTR region. Note that there is a significant positive correlation between the number of repeat copies and the proportion of Indigenous American ancestry in an MXL individual (Spearman's $\rho$: 0.438; *P-value*: 2.940e-4).

Python code to replicate this figure is available at: https://github.com/David-Peede/MUC19/blob/main/figure_nbs/figure_2_v_revisions.ipynb

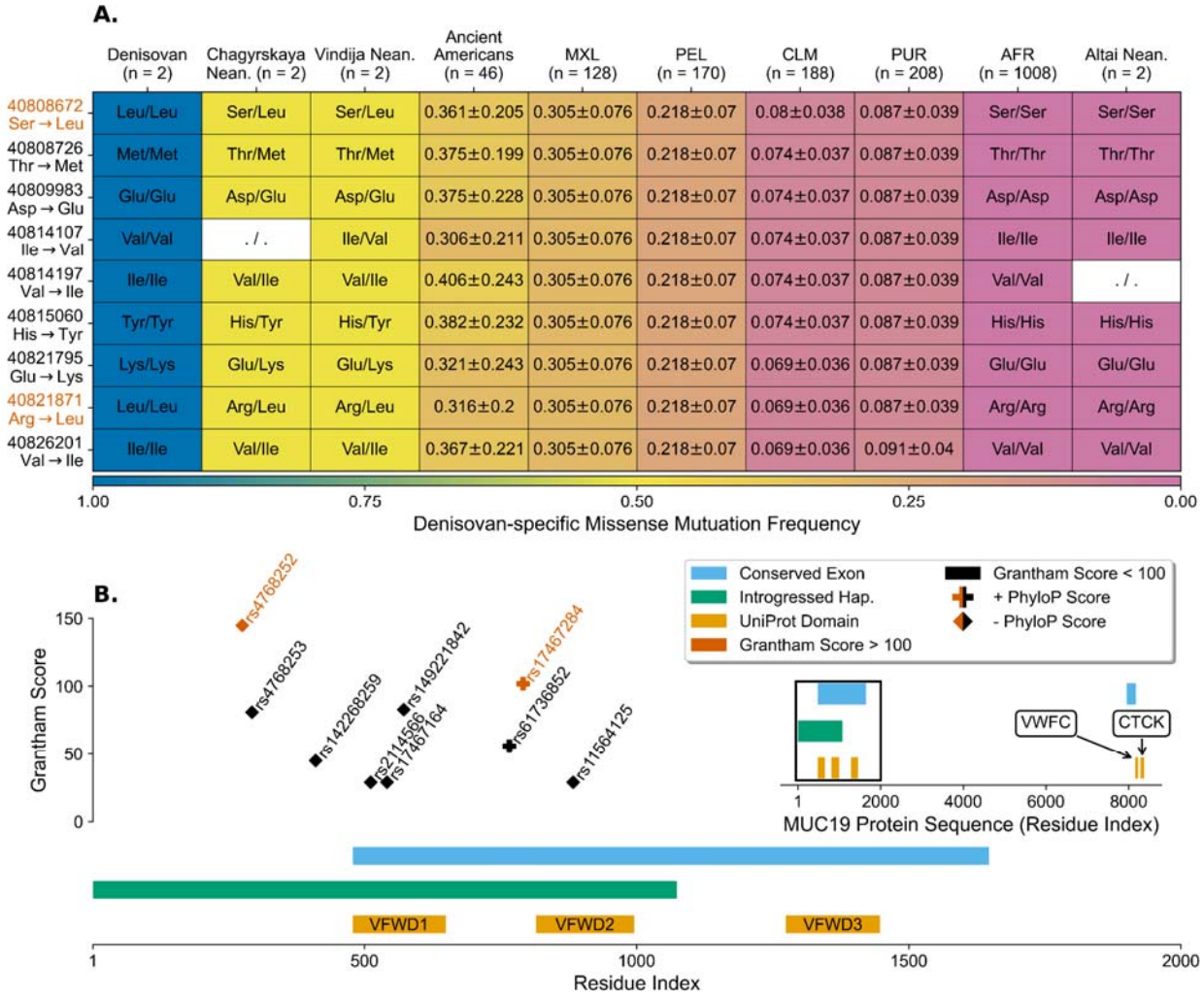



**Figure 3. Frequency and protein sequence context of the nine Denisovan-specific missense mutations at the 72kb region in *MUC19*.**

(**A**) Heatmap depicting the frequency of Denisovan-specific missense mutations (columns) amongst the four archaic individuals, 23 ancient pre-European colonization American individuals, the entire African superpopulation in the TGP (AFR), and admixed American populations in the TGP—Mexico (MXL), Peru (PEL), Colombia (CLM), Puerto Rico (PUR)—where the "n" represents the number of chromosomes in each population. The left hand side of each row denotes one of the nine Denisovan-specific missense mutations where the position and amino acid substitution (hg19 reference amino acid → Denisovan-specific amino acid) are denoted in black if its respective Grantham score is less than 100 or vermillion if its respective Grantham score is greater than 100. For the ancient and admixed American populations, the text in each cell represents the Denisovan-specific missense mutation frequency plus or minus the 95% confidence interval—for information on the number of ancient American chromosomes that passed quality control at every position see Table S30. For the archaic individuals, each cell is denoted with the individual's amino acid genotype—note that ". / ." represent sites that did not pass quality control in that given archaic individual—and each AFR cell is denoted by the homozygous hg19 reference amino acid genotype as no Denisovan-

specific missense mutation is present in any AFR individuals. (**B**) Denisovan-specific missense mutations in the context of the MUC19 protein sequence. The full MUC19 protein sequence is displayed in the smaller embedded subplot where the boxed region corresponds to the first 2000 residues and is depicted as the main plot, which encompasses the 72kb introgressed haplotype (in bluish green). Conserved exons are colored as sky blue and the UniProt domains are colored orange, where the text corresponds to specific UniProt domain identity—Von Willebrand factor (VFW) D domains, VWFC domain, and C-terminal cystine knot-like (CTCK) domain. Each of the nine Denisovan-specific missense mutations are denoted by their rsID, plotted with respect to residue index on the x-axis and their corresponding Grantham score on the y-axis. The color of each Denisovan-specific missense mutation denotes whether the mutation has a Grantham score less than 100 (black) or a Grantham score greater than 100 (vermillion, and the marker denotes whether their respective exon has a negative PhyloP score (diamonds) or a positive PhyloP score (crosses).

Python code to replicate this figure is available at: https://github.com/David-Peede/MUC19/blob/main/figure_nbs/figure_3_v_revisions.ipynb

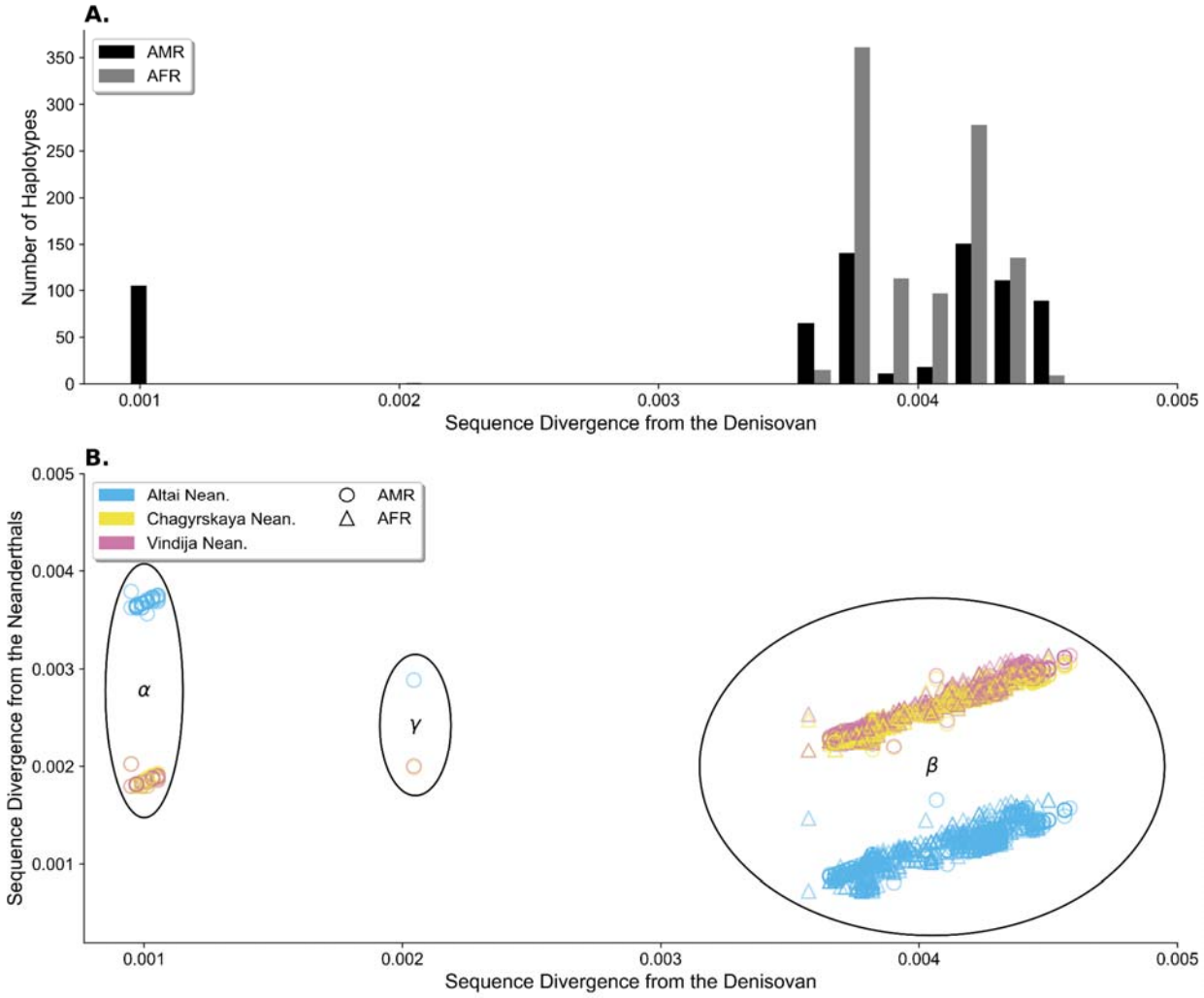



**Figure 4. Haplotype divergence at the 72kb region in *MUC19*.**

(**A**) Distribution of haplotype divergence—number of pairwise differences between a modern human haplotype and an archaic genotype normalized by the effective sequence length—with respect to the Altai Denisovan for all individuals in the Admixed American (AMR, black bars) and African (AFR, gray bars) super populations. (**B**) Joint distribution of haplotype divergence from the Altai Denisovan (x-axis) and the Neanderthals (y-axis)—Altai Neanderthal in sky blue, Chagyrskaya Neanderthal in yellow, and Vindija Neanderthal in reddish purple—for all individuals in the AMR (circles) and AFR (triangles) super populations. The three black ellipses ( , , and ) represent the three distinct haplotype groups segregating in the TGP. The ellipse represents the introgressed haplotypes which exhibit a low sequence divergence from the Altai Denisovan, a high sequence divergence from the Altai Neanderthal, and an intermediate sequence divergence—higher compared to the Altai Denisovan but lower compared to the Altai Neanderthal—with respect to the Chagyrskaya and Vindija Neanderthals. The ellipse represents the non-introgresssed haplotypes which exhibit a high sequence divergence from the Altai Denisovan, a low sequence divergence from the Altai Neanderthal, and an intermediate sequence divergence—lower compared to the Altai Denisovan but higher compared to the Altai Neanderthal—with respect to the Chagyrskaya and Vindija Neanderthals. Note that the AMR haplotype within the ellipse is positioned at intermediate sequence divergence values with

respect to the $\alpha$ and $\beta$ ellipses, which represents one of seven recombinant haplotypes segregating in the TGP (see Figure S35).

Python code to replicate this figure is available at: https://github.com/David-Peede/MUC19/blob/main/figure_nbs/figure_4_v_revisions.ipynb

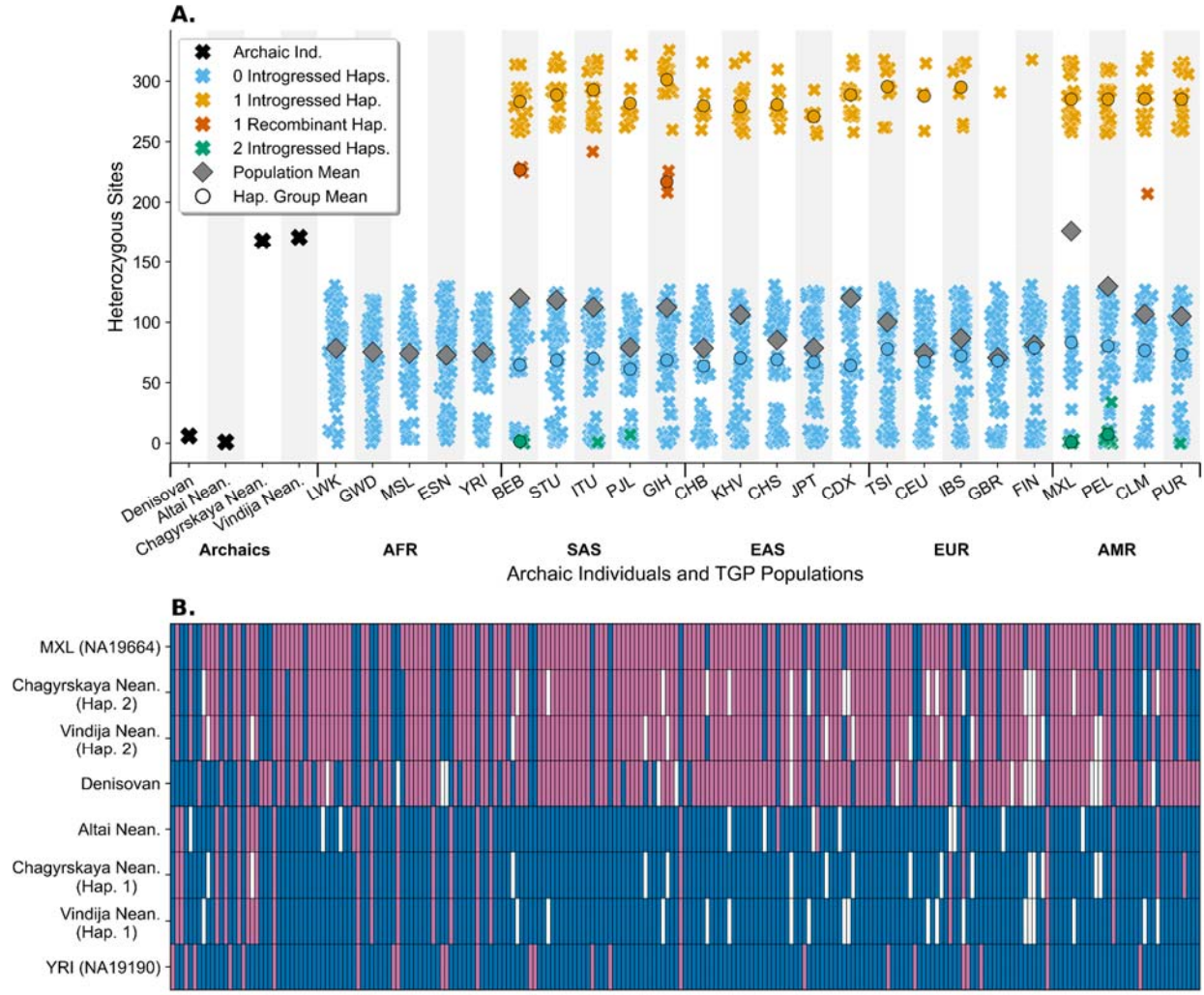
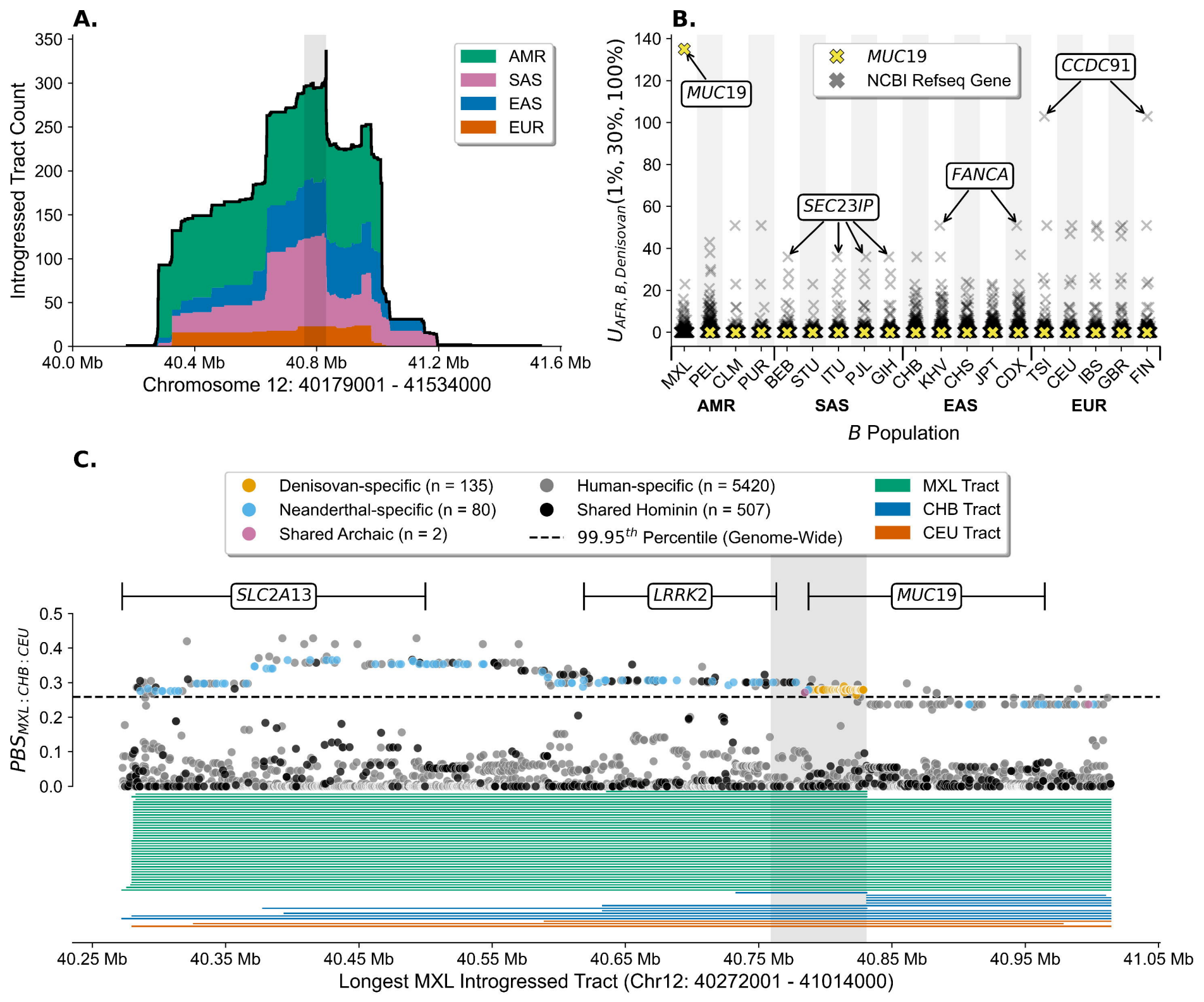


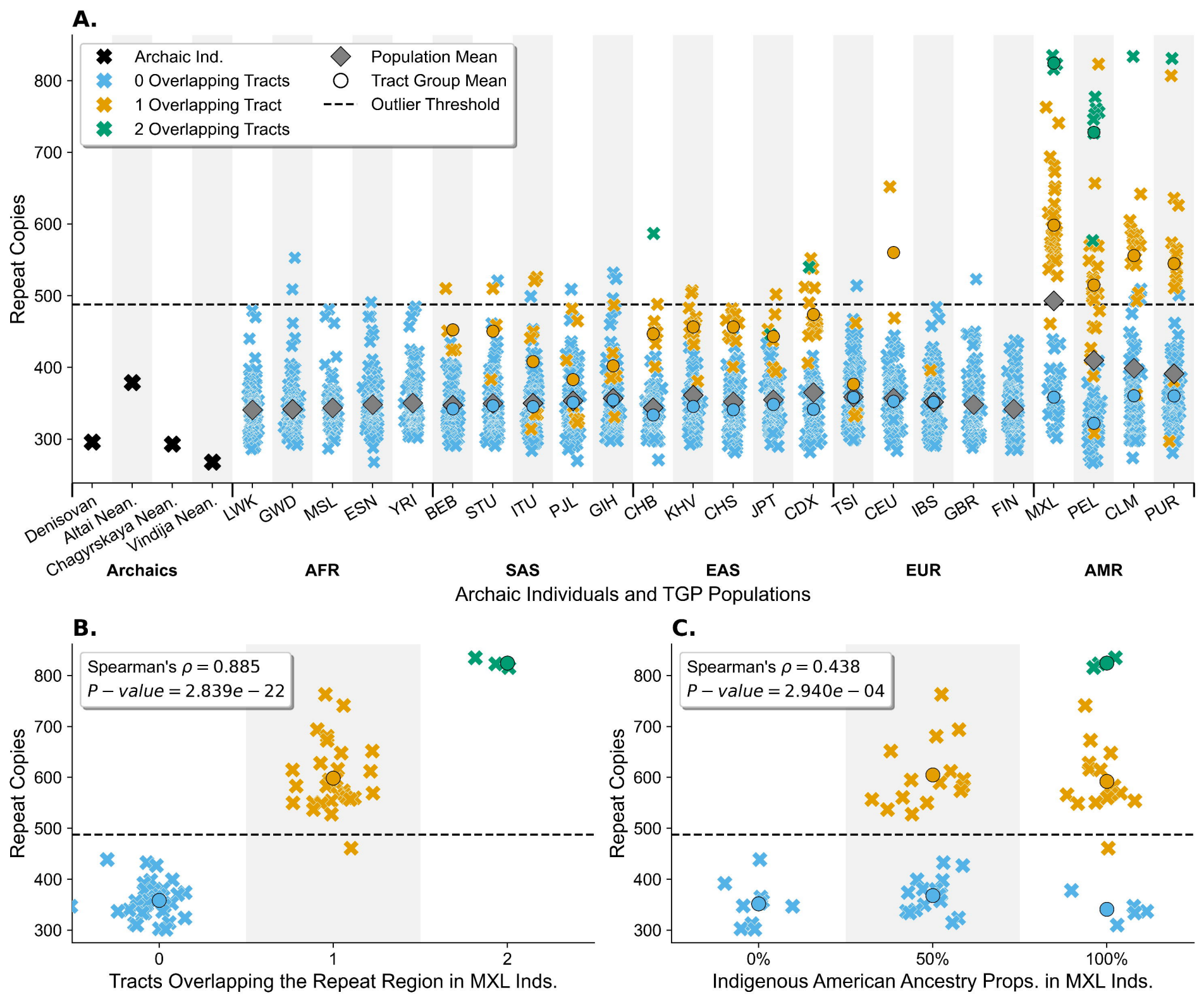

**Figure 5. The high levels of heterozygosity in the Chagyrskaya and Vindija Neanderthals are explained by *Denisovan-like* ancestry at the 72kb region in *MUC19*.**

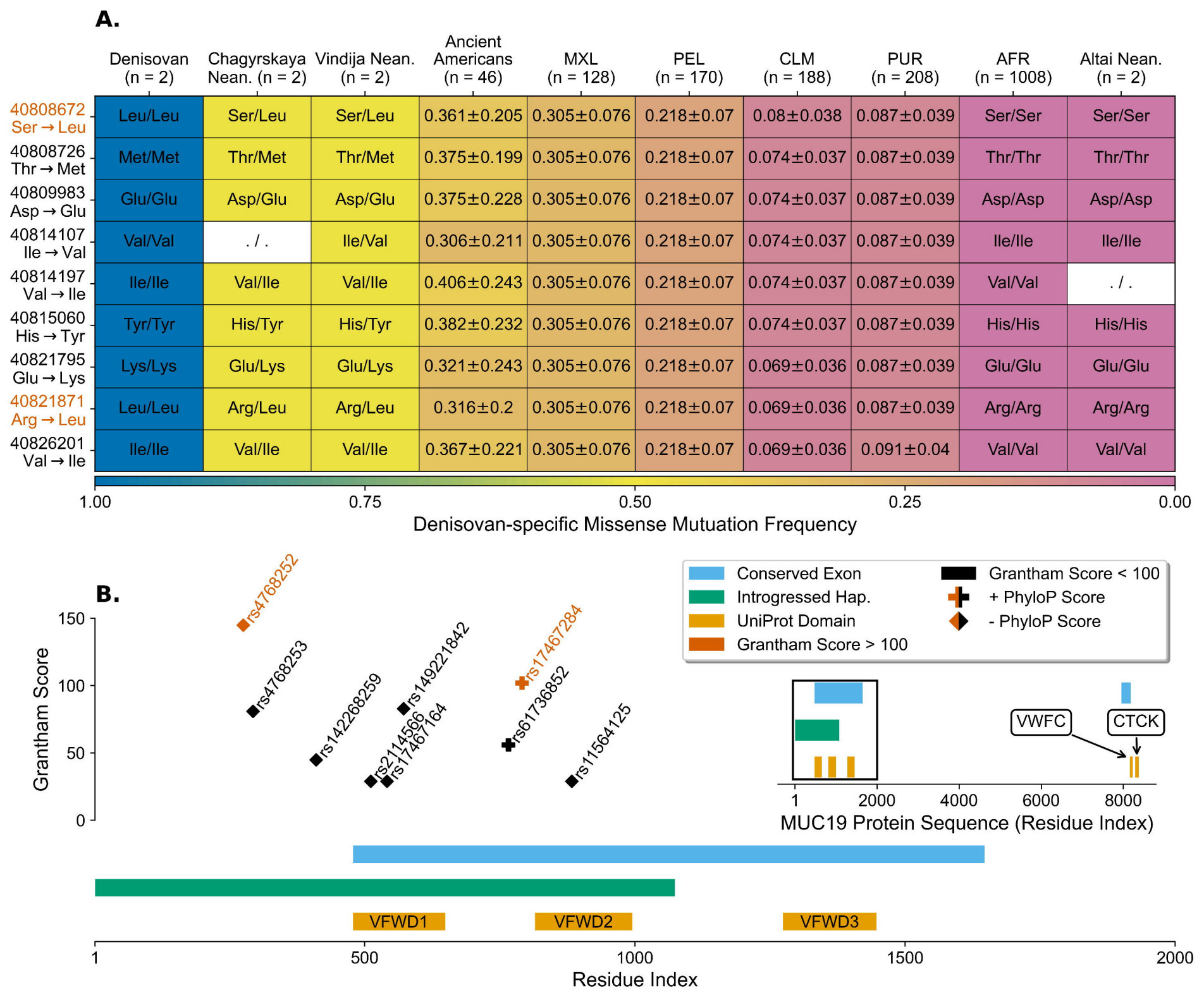
(**A**) Number of heterozygous sites at the 72kb region in *MUC19* per archaic individual (black X's), TGP individuals without the introgressed haplotype (sky blue X's), TGP individuals with exactly one copy of the introgressed haplotype (yellow X's), TGP individuals with a recombinant introgressed haplotype (vermillion X's), and TGP individuals with two copies of the introgressed haplotype (bluish green X's). The average number of heterozygous sites stratified by population are denoted by the grey diamonds and the average number of heterozygous sites amongst individuals who carry exactly zero, one, and two introgressed haplotypes are denoted by sky blue, yellow, and bluish green circles respectively and are stratified by population. (**B**) Haplotype matrix of the 233 segregating sites (columns) amongst the focal MXL individual (NA19664) with two copies of the introgressed haplotype; the focal YRI individual (NA19190) without the introgressed haplotype; the Altai Denisovan; the Altai Neanderthal; and the two phased haplotypes for the Chagyrskaya and Vindija Neanderthals, respectively. Cells shaded blue denote the hg19 reference allele, cells shaded reddish purple denote the alternative allele, and cells shaded white represent sites that did not pass quality control in the given archaic individual. Note that the focal MXL and YRI individuals are homozygous for every position in the 72kb region in *MUC19* and that the heterozygous sites for

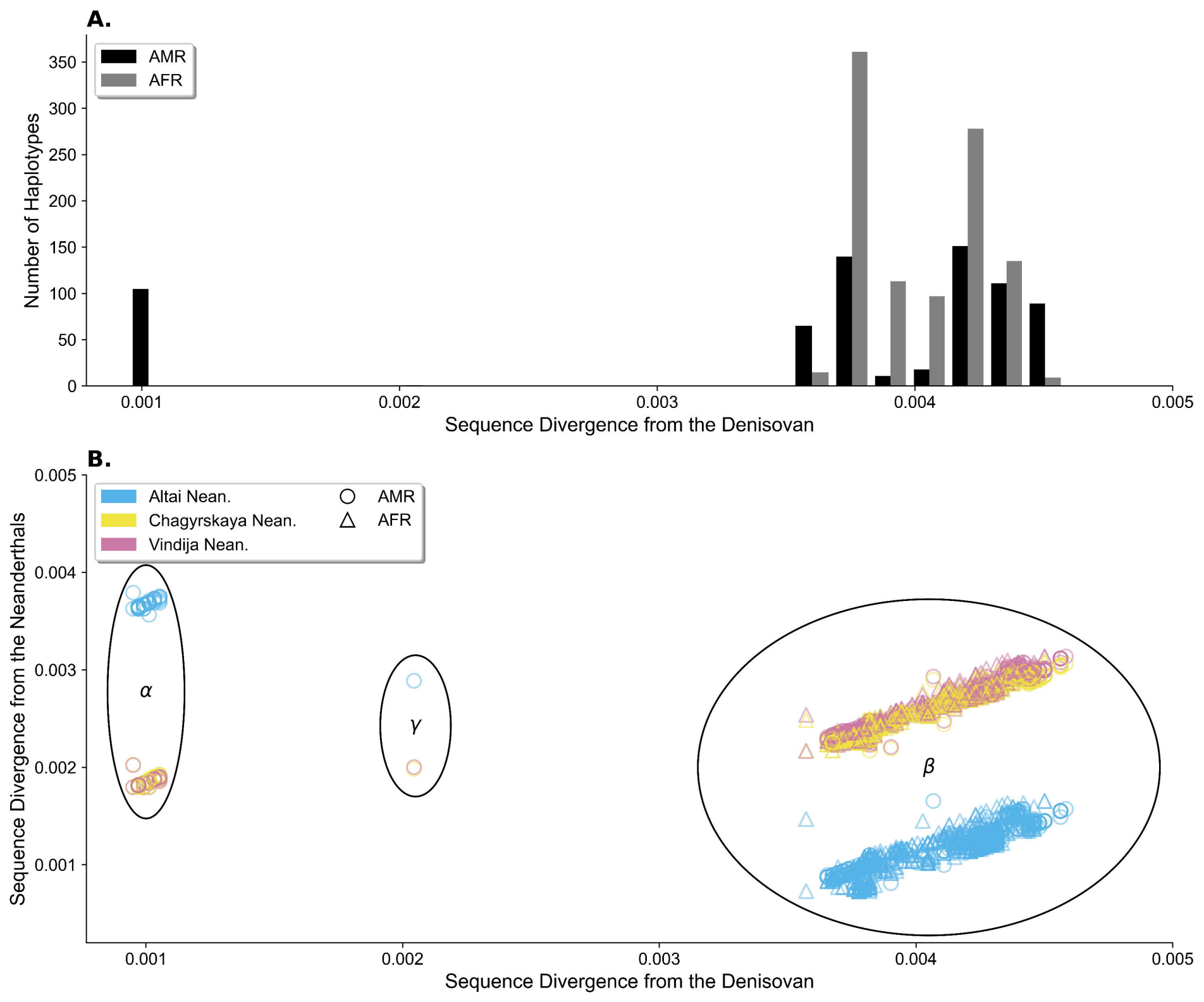
the Altai Denisovan and Altai Neanderthal—six and one heterozygous sites respectively—are omitted.

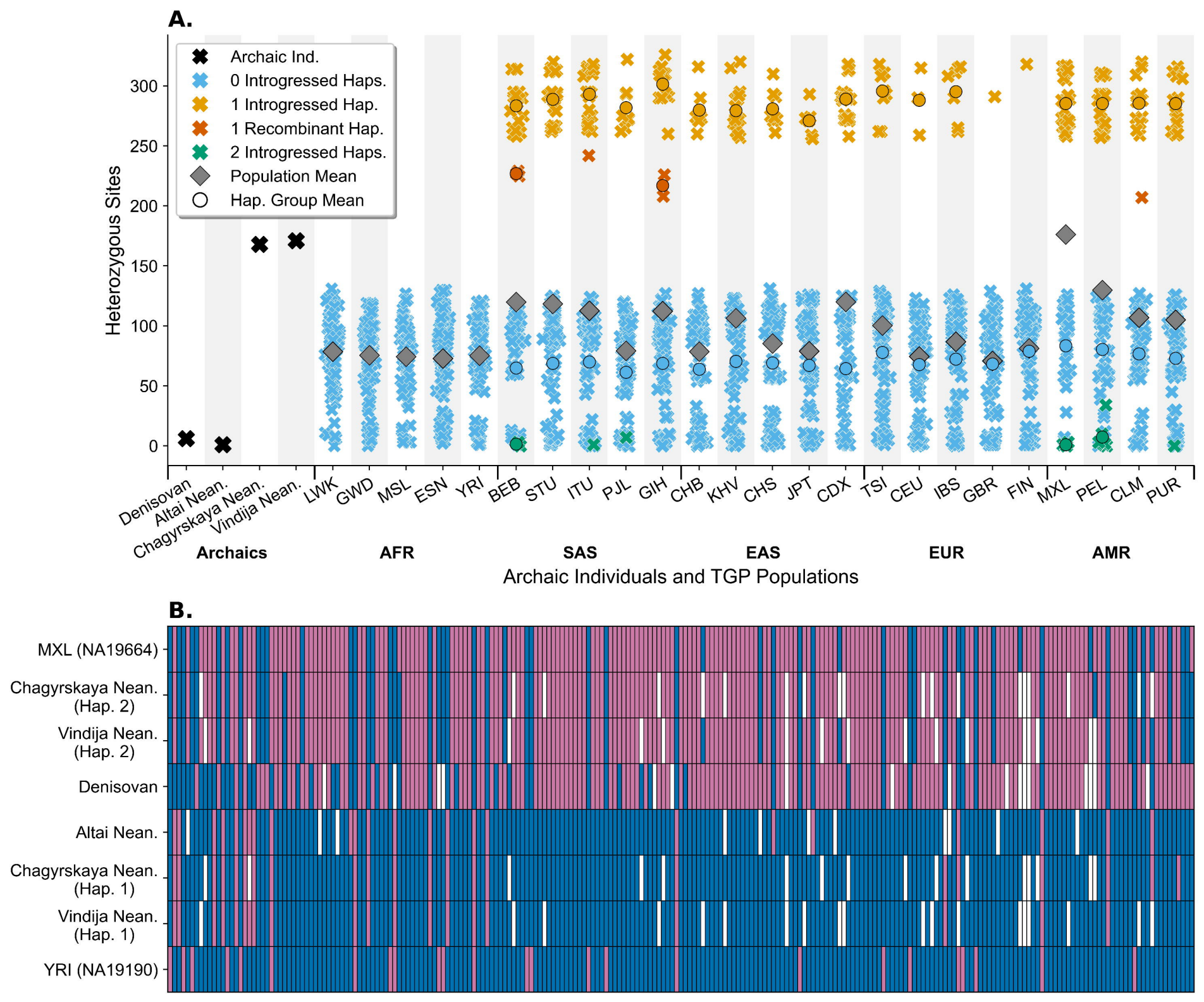
Python code to replicate this figure is available at: https://github.com/David-Peede/MUC19/blob/main/figure_nbs/figure_5_v_revisions.ipynb


A.
Introgressed Tract Count
AMR
SAS
EAS
EUR
40.0 Mb
40.4 Mb
40.8 Mb
41.2 Mb
41.6 Mb
Chromosome 12: 40179001 - 41534000
B.
$U_{AFR, B, Denisovan}(1\%, 30\%, 100\%)$
MUC19
NCBI Refseq Gene
CCDC91
FANCA
SEC23IP
MXL PEL CLM PUR BEB STU ITU PJL GIH CHB KHV CHS JPT CDX TSI CEU IBS GBR FIN
AMR
SAS
EAS
EUR
B Population
C.
Denisovan-specific (n = 135)
Neanderthal-specific (n = 80)
Shared Archaic (n = 2)
Human-specific (n = 5420)
Shared Hominin (n = 507)
99.95th Percentile (Genome-Wide)
MXL Tract
CHB Tract
CEU Tract
SLC2A13
LRRK2
MUC19
$PBS_{MXL:CHB:CEU}$
40.25 Mb
40.35 Mb
40.45 Mb
40.55 Mb
40.65 Mb
40.75 Mb
40.85 Mb
40.95 Mb
41.05 Mb
Longest MXL Introgressed Tract (Chr12: 40272001 - 41014000)


A.
Archaic Ind.
0 Overlapping Tracts
1 Overlapping Tract
2 Overlapping Tracts
Population Mean
Tract Group Mean
Outlier Threshold
Repeat Copies
800
700
600
500
400
300
Denisovan
Altai Nean.
Chagyrskaya Nean.
Vindija Nean.
LWK
GWD
MSL
ESN
YRI
BEB
STU
ITU
PJL
GIH
CHB
KHV
CHS
JPT
CDX
TSI
CEU
IBS
GBR
FIN
MXL
PEL
CLM
PUR
Archaics
AFR
SAS
EAS
EUR
AMR
Archaic Individuals and TGP Populations
B.
Spearman's ρ = 0.885
P − value = 2.839e − 22
Repeat Copies
0
1
2
Tracts Overlapping the Repeat Region in MXL Inds.
C.
Spearman's ρ = 0.438
P − value = 2.940e − 04
Repeat Copies
0%
50%
100%
Indigenous American Ancestry Props. in MXL Inds.


A.
Denisovan (n = 2)
Chagyrskaya Nean. (n = 2)
Vindija Nean. (n = 2)
Ancient Americans (n = 46)
MXL (n = 128)
PEL (n = 170)
CLM (n = 188)
PUR (n = 208)
AFR (n = 1008)
Altai Nean. (n = 2)
40808672 Ser → Leu
Leu/Leu Ser/Leu Ser/Leu 0.361±0.205 0.305±0.076 0.218±0.07 0.08±0.038 0.087±0.039 Ser/Ser Ser/Ser
40808726 Thr → Met
Met/Met Thr/Met Thr/Met 0.375±0.199 0.305±0.076 0.218±0.07 0.074±0.037 0.087±0.039 Thr/Thr Thr/Thr
40809983 Asp → Glu
Glu/Glu Asp/Glu Asp/Glu 0.375±0.228 0.305±0.076 0.218±0.07 0.074±0.037 0.087±0.039 Asp/Asp Asp/Asp
40814107 Ile → Val
Val/Val ./. Ile/Val 0.306±0.211 0.305±0.076 0.218±0.07 0.074±0.037 0.087±0.039 Ile/Ile Ile/Ile
40814197 Val → Ile
Ile/Ile Val/Ile Val/Ile 0.406±0.243 0.305±0.076 0.218±0.07 0.074±0.037 0.087±0.039 Val/Val ./.
40815060 His → Tyr
Tyr/Tyr His/Tyr His/Tyr 0.382±0.232 0.305±0.076 0.218±0.07 0.074±0.037 0.087±0.039 His/His His/His
40821795 Glu → Lys
Lys/Lys Glu/Lys Glu/Lys 0.321±0.243 0.305±0.076 0.218±0.07 0.069±0.036 0.087±0.039 Glu/Glu Glu/Glu
40821871 Arg → Leu
Leu/Leu Arg/Leu Arg/Leu 0.316±0.2 0.305±0.076 0.218±0.07 0.069±0.036 0.087±0.039 Arg/Arg Arg/Arg
40826201 Val → Ile
Ile/Ile Val/Ile Val/Ile 0.367±0.221 0.305±0.076 0.218±0.07 0.069±0.036 0.091±0.04 Val/Val Val/Val
1.00 0.75 0.50 0.25 0.00
Denisovan-specific Missense Mutuation Frequency
B.
Grantham Score
150 100 50 0
rs4768252
rs4768253
rs142268259
rs2114566
rs17467164
rs149221842
rs17467284
rs61736852
rs11564125
Conserved Exon
Introgressed Hap.
UniProt Domain
Grantham Score > 100
Grantham Score < 100
+ PhyloP Score
- PhyloP Score
VWFC
CTCK
1 2000 4000 6000 8000
MUC19 Protein Sequence (Residue Index)
VFWD1
VFWD2
VFWD3
1 500 1000 1500 2000
Residue Index


A.
AMR
AFR
Number of Haplotypes
0
50
100
150
200
250
300
350
0.001
0.002
0.003
0.004
0.005
Sequence Divergence from the Denisovan
B.
Altai Nean.
Chagyrskaya Nean.
Vindija Nean.
AMR
AFR
α
γ
β
Sequence Divergence from the Neanderthals
0.001
0.002
0.003
0.004
0.005
Sequence Divergence from the Denisovan


A.
Archaic Ind.
0 Introgressed Haps.
1 Introgressed Hap.
1 Recombinant Hap.
2 Introgressed Haps.
Population Mean
Hap. Group Mean
Heterozygous Sites
0
50
100
150
200
250
300
Denisovan
Altai Nean.
Chagyrskaya Nean.
Vindija Nean.
LWK
GWD
MSL
ESN
YRI
BEB
STU
ITU
PJL
GIH
CHB
KHV
CHS
JPT
CDX
TSI
CEU
IBS
GBR
FIN
MXL
PEL
CLM
PUR
Archaics
AFR
SAS
EAS
EUR
AMR
Archaic Individuals and TGP Populations
B.
MXL (NA19664)
Chagyrskaya Nean. (Hap. 2)
Vindija Nean. (Hap. 2)
Denisovan
Altai Nean.
Chagyrskaya Nean. (Hap. 1)
Vindija Nean. (Hap. 1)
YRI (NA19190)