Imperial College London

Discerning Ancestry-Related Assortative Mating from Migration by their Genomic Imprints upon Admixed Populations

Ben Nouhan

August 26, 2021

A thesis submitted for the partial fulfillment of the requirements for the degree of Master of Research in Computational Methods in Ecology and Evolution
at Imperial College London

Many human populations throughout history have been socially stratified. Individuals of human populations show a tendency to mate with similar individuals, a phenomenon called assortative mating, which contributes to this stratification. In admixed populations, these similarities can also be genetic, leading to a non-random admixture process wherein disproportionate mating takes place within socioculturally distinct subpopulations. These social strata display different proportions of genetic ancestry inherited from the pre-admixture source populations. Accordingly, when the ancestral origins of different genomic regions are mapped out, this non-random admixture can be inferred: by the distribution of the size of fragments consecutively assigned to the same ancestry. However, differentiating assortative mating from migration with this method is difficult since past migration patterns manifest similarly in the genomes of an admixed population. Here I show how significant modern-day assortative mating can be detected in, and compared between, admixed North and South American populations using population genomic techniques. This empirical evidence of assortative mating may bring the validity of certain genome-wide association studies and research of the genomic impacts of migration into question. Furthermore, I help to lay the groundwork for a technique that can integrate migration data with genomic data to accurately quantify and timestamp past assortative mating. If successful in doing so, we can hopefully reveal the historical context for modern-day ancestry-related social stratification present in populations throughout the world, and perhaps track said stratification, and the efficacy of efforts to combat it, in real time.

Contents

1 Introduction

Positive assortative mating, a phenomenon wherein individuals are more likely to mate with those phenotypically similar to themselves, is widely accepted to occur in human populations (Norris et al., 2019). This has the potential to alter population structure by introducing social stratification and, in turn, create social constructs upon which further assortative mating can be based, such as wealth, class or social policies (Risch et al., 2009).

From a genetics standpoint, this multigenerational non-random admixture between genetically distinct groups leaves a genomic imprint in the individuals comprising the population, one that is in stark contrast to populations more closely following Hardy-Weinberg equilibrium (HWE) (Zaitlen et al., 2017). However, the genomic imprint on the population structure left by either sociocultural barriers or geographical barriers limiting admixture is difficult to discern. Afterall, large scale migration of a population will genetically manifest itself similarly to the change of societal rules or norms that condition social interaction, such as the revocation of racial segregation policies.

Single nucleotide polymorphisms (SNPs) can be used as indicators of ancestry (Risch et al., 2009). Population genomic techniques allow us to generate a large array of SNPs which can be analysed using local ancestry inference to map ancestries to positions and regions along the genome. Further analysis can then indicate past assortative mating in a population (Schubert et al., 2020).

One such analysis is that of continuous ancestry tract lengths: the lengths of genomic regions consecutively assigned to the same ancestry. Looking at the distribution of these lengths, the ancestry to which they belong and the overall ancestry proportion of individuals within a population can indicate how long ago the admixture occurred and to what extent. Recombination of the DNA of admixing individuals leads to a decrease in continuous ancestry tract lengths, as those within the parents' genomes interrupt one another upon recombination. Hence, admixture more generations ago will manifest as distributions of shorter continuous ancestry tracts and vice versa (Gravel, 2012).

Genotype frequency is another indicator of population admixture; one would expect a more admixed population to have higher heterozygous genotype frequencies at a given position. While this alone does not inherently indicate assortative mating, the extent to which the observed genotype frequency deviates from what would be expected under HWE can also be considered. The assortative mating index (AMI) quantifies the relative local ancestry homozygosity-to-heterozygosity ratio at a given position based on this concept; this can be used as a proxy for the extent of assortative mating at said position (Norris et al., 2019).

HWE is commonly used in population genomics as a quality check for genetic markers - SNPs chosen for being particularly informative for certain pathological research - in genome-wide association study (GWAS). Alleles with frequencies deviating too far from it are removed and deemed sequencing misreads (Linares-Pineda et al., 2012). This does not take into account stratification, present in most if not all societies, within the studied populations. Showing here that the deviation of allelic frequencies from HWE is not an artefact but rather an intrinsic quality of some populations may serve as a warning against this practice.

Populations of the Americas such as Colombia, Barbados, Mexico or the US provide appropriate and well-researched case studies integrating migration, admixture and assortative mating. Many such populations have different but connected histories: a Native American population is

colonised by Europeans; the Native population shrinks due to war, hard labour and disease, while the European population grows via migration. These phenomena continue such that African slaves are transported to the region as a source of additional labour, after which the population continues evolving with the lingering impacts of colonialism (Bryc et al., 2010; e Silva et al., 2020; Mas-Sandoval et al., 2019).

These North and South American populations are far from the only examples of where migration and assortative mating coincide and as such can be studied; indeed most human populations are the result of admixture between multiple populations. However, the three source populations giving rise to the admixed population - African, European and Native - being genetically distinct facilitates the identification of local ancestry fragments and enables the study of the complex admixture process.

By analysing the length of the local ancestry fragments it is possible to evaluate both the admixture dates and the strength of the ancestry-related assortative mating. Said assortative mating can be understood as the degree of impermeability of the socioeconomic and cultural barriers between subgroups of the admixed population with differentiable genetic ancestries. Further understanding and ideally quantifying ancestry-related assortative mating, and using it as a proxy for ancestry-related social stratification, will not only help us better understand how such stratification historically and presently influence mating behaviours in the Americas, but could also be used to track or predict it in present and future admixed populations.

To accurately estimate the extent of assortative mating in a population using genomic techniques, the genomic impact of migration on said population must be accounted for, despite them being difficult to differentiate. Previous research has either studied genomic impact of migration while assuming otherwise random admixture (Borda et al., 2020; Gravel, 2012; Norris et al., 2020), or studied assortative mating while assuming a single pulse of migration from each immigrating ancestry (Norris et al., 2019; Risch et al., 2009; Zaitlen et al., 2017). However, for reasons outlined, studies on the effects of migration on population genomics must consider assortative mating, and when studying assortative mating one must consider migration as a continuous process rather than a single event. Equally, comparing measured assortative mating levels of different populations and cross-referencing this with their histories and current socioeconomic climates could yield interesting insights as to causes and long-term effects of ancestry-related social stratification.

Hence, the aims of this project are twofold. Firstly, to use genomic data from admixed populations of the Americas to explore different analytical methods designed to reveal non-random admixture in a population. This will enable me to compare these methods by their potential to distinguish between migration and assortative mating as sources for this non-random admixture. Secondly, to use the results of these analyses to compare the admixed populations by the level of assortative mating revealed. My hypotheses are that each population will exhibit significant positive assortative mating, and that the level of said assortative mating in each population will be significantly different to that of the others.

Only by reconciling migration and assortative mating can we confidently infer assortative mating from genomic data, which can then be used to draw conclusions about past and detect trends of future ancestry-related social stratification.

2 Methods

2.1 Studied Populations

For the initial analyses, all African, European and American populations from the 1000 Genomes Project (1KGP) and the Human Genome Diversity Project (HGDP) were used (**Table 1**), with the exception of the Russian and Finnish populations. These were excluded owing to minimal colonial-era migration to the Americas from these populations and the relative genetic similarities between these populations, Siberians and, by extension, Native Americans.

Table 1: Details of the populations used throughout this study. Populations abbreviated as initialisms are from the 1000 Human Genome Project dataset, while full-word abbreviated populations are from the Human Genome Diversity Project dataset. The number of samples used from each population is denoted by n.
*The Tuscan and Yoruba populations comprise samples from both datasets.

Superpopulation	Population	Abbreviation	n
Admixed	African Ancestry in Southwest USA	ASW	61
	African Caribbean in Barbados	ACB	96
	Colombian in Medellin, Colombia	CLM	94
	Mexican Ancestry in Los Angeles, California	MXL	64
	Peruvian in Lima, Peru	PEL	85
	Puerto Rican in Puerto Rico	PUR	104
African	Bantu in Kenya	BantuKenya	11
	Bantu in South Africa	BantuSouthAfrica	8
	Biaka in Central African Republic	Biaka	22
	Esan in Nigeria	ESN	99
	Gambian in Western Division, The Gambia	GWD	113
	Luhya in Webuye, Kenya	LWK	99
	Mandenka in Senegal	Mandenka	22
	Mbuti in Democratic Republic of Congo	Mbuti	13
	Mende in Sierra Leone	MSL	85
	San in Namibia	San	6
	Yoruba in Nigeria	YRI/Yoruba*	129
European	Basque in France	Basque	23
	Bergamo Italian in Bergamo, Italy	BergamoItalian	12
	British in England and Scotland	GBR	91
	Northern and Western European Ancestry in Utah	CEU	99
	French in France	French	28
	Orcadian in Orkney	Orcadian	15
	Sardinian in Italy	Sardinian	28
	Iberian in Spain	IBS	107
	Toscani in Italy	TSI/Tuscan*	115
Native American	Colombian in Colombia	Colombian	7
	Karitiana in Brazil	Karitiana	12
	Maya in Mexico	Maya	21
	Pima in Mexico	Pima	13
	Surui in Brazil	Surui	8

2.2 Data Preparation with BCFtools

Using BCFtools v1.9, the 30x coverage 1KGP and high-coverage HGDP datasets were merged, and all populations except those listed in **(Table 1)** were removed. All C→G, G→C, A→T and T→A SNPs were filtered out as they are harder to assign and are hence prone to error (Danecek et al., 2021). SNPs were further filtered with a minor allele frequency threshold of 5%, as to reduce the dataset and remove rare and thus uninformative SNPs. Following this, all 22 filtered VCF files,

one per autosome, were indexed for phasing.

2.3 Haplotype Estimation with SHAPEIT4

Phasing was carried out using SHAPEIT v4.2.0, which efficiently assigns haplotype estimates for each genotype by cross-referencing the genomic region in question with the corresponding region of a pre-phased reference panel and of the other genomes being phased (Delaneau et al., 2019). The programme was run using default parameters, the B38 genetic map recommended by the developers, and an appropriate high-coverage phased reference genome from the 1KGP website was used to improve haplotype estimation accuracy (see: **Data and Code Availability**). The individually phased chromosomes were then merged into a single VCF file with BCFtools.

2.4 Ancestry Estimation with PLINK & ADMIXTURE

Linkage disequilibrium pruning was performed with PLINK v2.0 on the genomes in VCF format, which creates a subset of largely independent SNPs - thereby significantly reducing the computational power needed for subsequent analyses with minimal information loss - before converting the pruned dataset to PLINK format (Purcell et al., 2007). These SNPs form the basis of this study.

The programme ADMIXTURE v1.3.0 used cluster analysis and principal component analysis to estimate the proportions of African, European and Native American ancestry for each remaining sample, with the number of ancestries parameter set at three. (Alexander et al., 2009).

2.5 Local Ancestry Inference with RFMIX v2

The ADMIXTURE outputs were subsequently used to filter out all significantly admixed samples, with a minimum threshold of 99% African, European or Native American ancestry. This subsetting was executed using BCFTools, yielding a subset VCF of >99% non-admixed samples. THis was used as a reference panel for local ancestry assignment with the programme RFMIX. A query subset was created correspondingly, containing all samples in the "Admixed" superpopulation in **(Table 1)**.

RFMIX v2.03-r0, based on concepts developed in RFMIX v1, assigns ancestries to segments of an individual's genome, which not only yields ancestry proportions as with ADMIXTURE, but also effectively maps out each genome in terms of each genomic region's estimated ancestry or origin. It does this by progressively modelling ancestry along each chromosome using discriminant random forests, conditional random field modelling and observed haplotype sequences of ancestry inferred from an input reference panel (Maples et al., 2013).

The RFMIX run was performed using the aforementioned query and reference VCF files, and a sample map linking the sample codes to their respective populations. Parameters used were three iterations of the algorithm and 20 generations. Before 20 generations ago, assuming an average generation length of 25 years, no known European-Native American admixture had taken place.

2.6 Assortative Mating Index Calculation

One measure of assortative mating is the assortative mating index (AMI), a log odds ratio test for the relative local ancestry homozygosity and heterozygosity:

$$AMI = ln\left(\frac{hom^{obs}/hom^{exp}}{het^{obs}/het^{exp}}\right) \tag{1}$$

Three ancestries are being investigated, hence expected homozygous and heterozygous allelic frequencies can be thought of in terms of the biallelic (**Equation 2**) or triallelic (**Equation 3**) Hardy-Weinberg models (Norris et al., 2019):

$$(x + \bar{x})^2 = x^2 + 2\bar{x}x + \bar{x}^2 \tag{2}$$

$$(a + e + n)^2 = a^2 + e^2 + n^2 + 2ae + 2an + 2en \tag{3}$$

The left side of each of these models are haplotype frequencies, while the right sides are genotype frequencies, each side of the equation summing to one. In the triallelic model, a, e and n are the initials of the ancestry they represent, while x and $\bar{x}$ in the biallelic model correspond to a given ancestry - African, European or Native - and all other ancestries respectively. Hence, while AMI is calculated only once using the triallelic model, the AMI using the biallelic model must be calculated three times: once with respect to each ancestry. For example, with respect to African ancestry, the homozygous genotype would be both African alleles or both non-African alleles, and the heterozygous genotype would be one African allele and one allele of one of the other ancestries.

The outputs of RFMIX v2 were analysed by a series of R Studio scripts I created for this project (see: **Data and Code Availability**). Firstly, the forward-backward (*.fb.tsv*) output files are read by the script "rfmix.fb.tsv_genotype_assign_HPC.R". These files contain the estimated haplotype probabilities at each genomic position for each sample. The script then assigns the genotype for each genomic position in each sample, with a probability threshold of 0.9, and returns the frequencies of each of the six triallelic genotypes at each position across samples as a table. This genotype frequency table is then read by the script "rfmix.fb.tsv_genotype_analysis.R" before calculating the triallelic AMI and the three biallelic AMIs, at each position with respect to each ancestry.

2.7 Continuous Ancestry Tract Length Analysis

Ancestry assignments of lower certainty in the forward-backward file, using the 0.9 probability threshold, have the potential to fracture continuous ancestry tracts thereby completely alter the distribution of their lengths. Hence the *.msp.tsv* RFMIX output files were used instead, equivalent to the forward-backward files but with automatic haplotype assignment to haplotype with highest estimated likelihood.

To generate the fragment length distributions, the script "rfmix.msp.tsv_window_size.R" reads the *.msp.tsv* files, sums the length of consecutive genomic windows assigned to the same ancestry, and appends the lengths to the vector containing the lengths of other fragments corresponding to the fragment's ancestry and population. The script "rfmix.msp.tsv_inbred_window_size.R" works similarly, but generates fragment length distributions of consecutive homozygous genotype assignments, rather than haplotype assignments.

2.8 Timeline of Admixture Estimation with TRACTS

TRACTS is a software for modelling migration histories using ancestry tracts data, incorporating time-dependent gene-flow theory and correcting for chromosomal end effects and haplotype assignment errors. In doing so, it predicts how many generations prior to the query genomes the migration events bringing the different populations together occured (Gravel, 2012).

The software uses the *.bed* file format as input, a file output of the original RFMIX but not of RFMIX v2, hence I created a script to convert *.msp.tsv* to *.bed*, "msp2bed_conversion.R". This merges together each chromosome from the 22 .msp.tsv files, and merges each consecutive intrachromosomal fragment - pre-defined by RFMIX - of the same ancestry into single fragments, whereby adjacent fragments can be of vastly different lengths and always different assigned ancestries. It then recalculates each cell based on this merging of fragments assigned to the same ancestry, reshuffles and reformats the columns, and saves one *.bed* file per query sample, each *.bed* file containing fragments constituting the entire genome of one individual, as required to run TRACTS.

Because in each of the admixed query populations there was initial admixture between Native Americans and Europeans populations, followed by African and further European ancestry being added to the gene pool, none of the models provided by TRACTS were entirely appropriate. I therefore adjusted the provided four population model, which assumes admixture of two initial populations and subsequently two further populations with three migration events. The adjusted version instead assumes initial admixture between two populations and subsequent admixture with one of those two populations (European) and a third population (African). Said adjusted model is encoded in the Python 2 script "models_4pop.py", which is run by "taino_ppxx_xxpp.py" for each admixed query populations with 25 bootstraps.

3 Results

3.1 Ancestry Proportion

Pruning led to the dataset being reduced to 4,111,226 SNPs per sample. ADMIXTURE was used on these SNPs to estimate the ancestry proportion of three ancestries - African, European and Native - for all 1690 individuals represented in **Table 1**.

The averaged output for each of the 31 populations is visualised in **Fig. 1**, which displays Peruvians and LA Mexicans as predominantly of Native ancestry and minimally African; Colombians and Puerto Ricans as predominantly European but more Native than African; and Barbadians and African Americans from US Southwest (ASWs) as predominantly African with minimal Native ancestry.

The distribution of ancestry proportions on the individual level within these six admixed populations is shown in **Fig. 2**, which largely corresponds with **Fig. 1**. This suggests approximately 25% of LA Mexicans and Colombians have no African ancestry, fewer than 5% and 50% of Barbadians and ASWs respectively have Native ancestry, and that only around 20% of Peruvians have African ancestry while around 25% of them are of exclusively Native ancestry.

These individual-level ancestry proportion distributions are further presented in **Fig. S1**, with the distribution for each population of a given ancestry displayed side-by-side in box plots. All

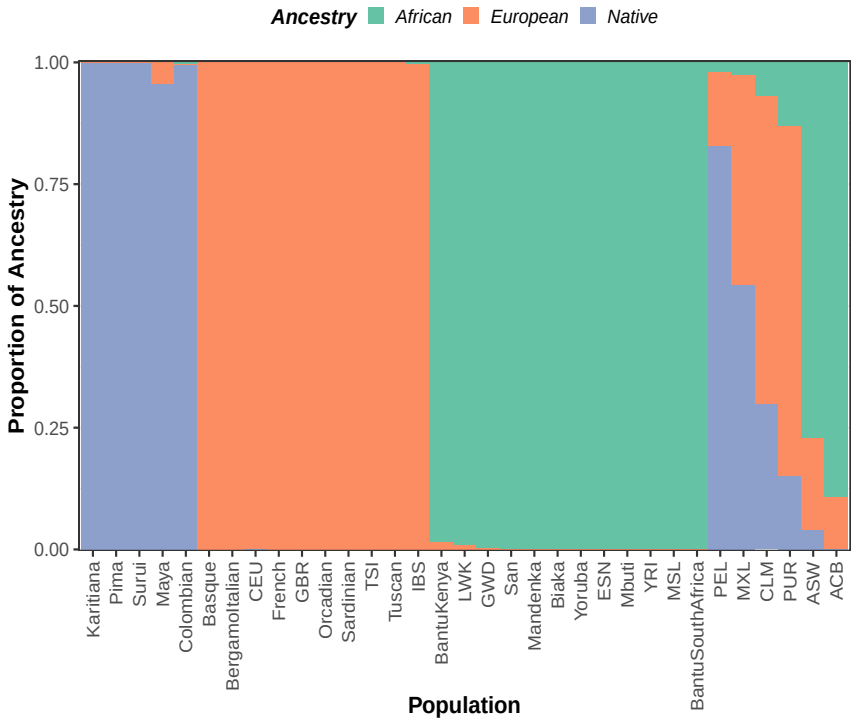

Figure 1: Stacked barplots showing the proportions of the three ancestries of each reference or query population used throughout the study, generated by ADMIXTURE. Genomic data from individuals of selected populations from the 1000 Genomes Project and the Human Genome Diversity Project were processed and subjected to ADMIXTURE, the output of which was averaged for all individuals of a given population. Populations 1-5 are Native, 6-15 are European, 16-27 are African, and 28-33 are admixed populations from the Americas.

distributions were different at the 5% significance level - except for Barbadian and Peruvian European ancestry proportion distributions. However, their Native and African ancestry distributions contrast starkly, supporting the assumption that all six studied admixed populations have highly different ethnological structures.

Following the admixture run, the 25 reference populations were filtered to remove all samples with less than 99% of the corresponding ancestry. This left a reference panel of 72, 507 and 550 people of 99% or more Native, European and African ancestry respectively for use in the local ancestry inference by RFMIX of the 504 query samples from the admixed populations.

3.2 Assortative Mating Index

One of the outputs of RFMIX is equivalent to that of ADMIXTURE, and a comparison of their relative performance on the 1690 studied individuals is shown in **Fig. S2**. Briefly, RFMIX tends to give lower African ancestry proportion estimates than ADMIXTURE in samples which both deem to have higher African Ancestry, and higher European ancestry proportion estimates in samples which both deem to have lower European Ancestries.

The main RFMIX output was used to calculate assortative mating index values for each SNP in each population. The triallelic AMI values for each position and population are plotted in **Fig. 3**. In a population without assortative mating, we would expect the mean AMI value to be zero. With a sample size of 4,111,226 SNPs, and the standard deviations being of similar sizes to the corresponding means, the standard errors of the means are negligible and hence the sample means can be considered accurate estimates of the true means. Based on this, we can see all means are

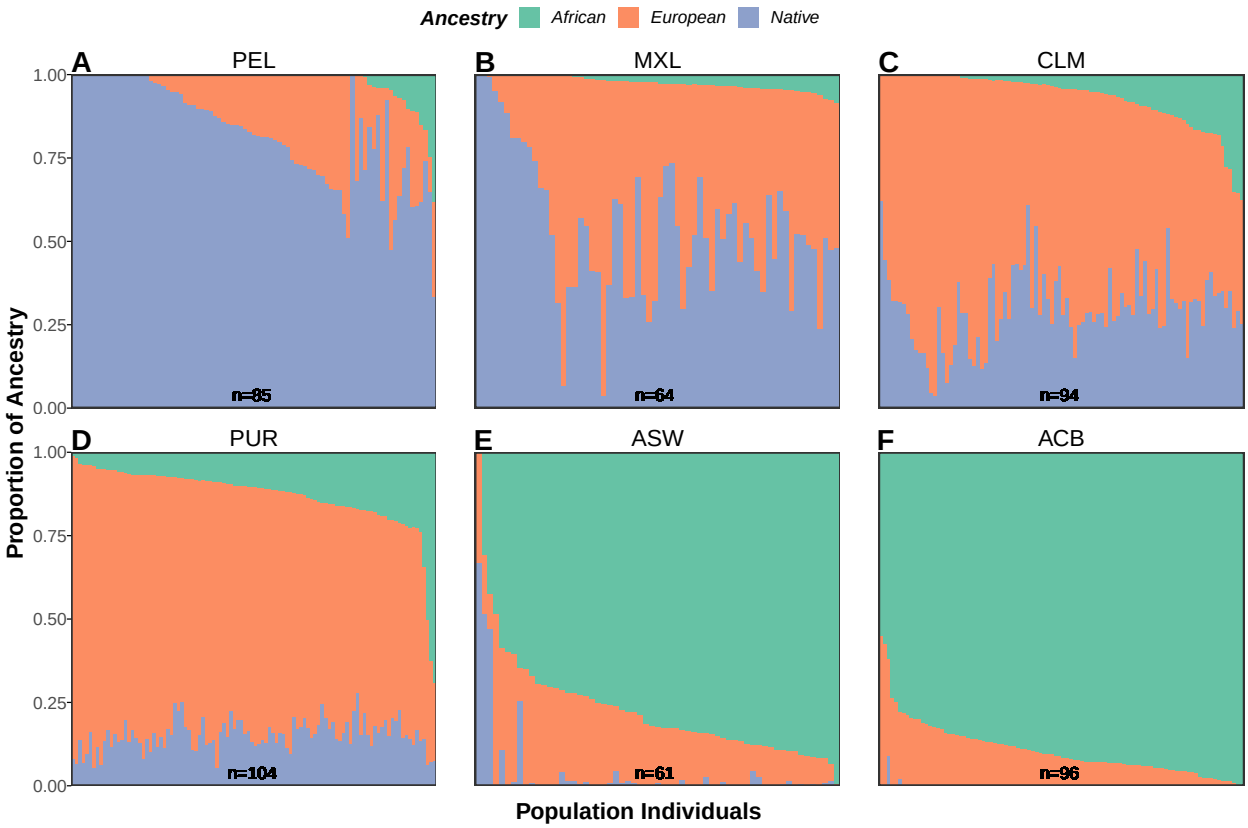

Figure 2: Stacked barplots showing the proportions of the three ancestries of each individual comprising the six query admixed populations, generated by ADMIXTURE. Genomic data from individual samples of the six admixed populations from the 1000 Genomes Project were processed and subjected to ADMIXTURE. The number of samples comprising each population is denoted by n, and individuals are ordered within the plot of each respective admixed population by increasing African and then European ancestry.

significantly higher than zero, indicating positive assortative mating in all admixed populations.

Wilcoxon tests were performed to also ascertain whether the AMI distribution of each population are significantly different from the other populations: this was confirmed to be the case (**Fig. S4A**).

The same analyses were carried out for the biallelic ancestry-specific AMI values. With the same large sample size, the distribution of each population is significantly higher than zero for all three ancestries, confirming that assortative mating has occurred in each population with respect to all three ancestries.

Wilcoxon tests were then performed to compare the AMI distributions of each ancestry by population and of each population by ancestry (**Fig. S4B** and **C** respectively). With the exception of European-specific AMI distributions for Puerto Rico and Peru, all combinations of ancestries or populations were significantly different.

To test whether mean ancestry-specific AMI value is correlated with or driven by mean ADMIXTURE-estimated ancestry proportion, they were plotted for each admixed population (**Fig. S3**). However, no significant correlation was found (p-value = 0.133).

3.3 Continuous Ancestry Tract Lengths

The final use of the RFMIX output was analysing the lengths of continuous ancestry tracts. Displaying the haplotype continuous ancestry tracts in a histogram allows visual comparison between

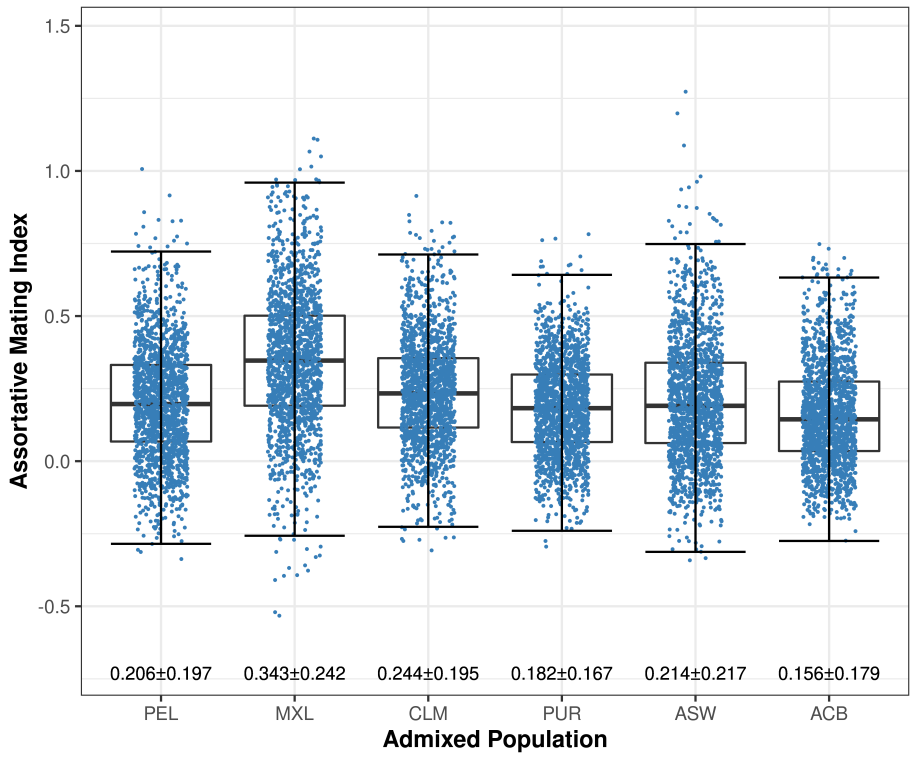

Figure 3: Comparative box plots displaying the distribution of the triallelic assortative mating index calculated for each studied single nucleotide polymorphism for each studied admixed population. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers signify the last data point within the closest quartile value plus 150% of the interquartile range. Mean ± standard deviation is given beneath; the standard error of the mean is negligible owing to the sample size of 4,111,226. Horizontal jitter is used simply to better display the distribution.

the tract length distributions of the different ancestries (**Fig. 5A**), while box plots better display descriptive statistics of the data (**Fig. S5**).

As would be expected, there is a clear correlation between the relative heights and x-axis positions of the distributions in a given population and the corresponding mean ancestry proportion. Skewed distributions, such as the right-skewed African distributions of the ASW and ACB plots, suggest some form of deviation from HWE, but whether these are caused by migration, assortative mating or some other phenomenon is unclear.

A supplementary approach is finding and plotting homozygous continuous ancestry tract lengths, as in **Fig. 5B**. This exaggerates HWE deviations, and provides additional peaks to some of the distributions. These peaks are more informative than just skewness: they show different tract length distributions of the same ancestry that have been merged, essentially representing two populations of the same ancestry merging into one. Hence significant same-ancestry migration is the likely cause of corresponding skewness in the haplotype continuous ancestry tract length visualisations, for example with ASW and ACB.

Less intense right skewness, such as with European ancestry in the ASW population, could indicate either minor but sustained European migration, or European assortative mating, where at least some of the European population disproportionately interbred thereby preserving longer continuous ancestry tracts than would be expected under HWE.

Each homozygous continuous ancestry tract length distribution has a left tail absent in their haplotype counterparts, likely an artefact of heterozygous alleles breaking large homozygous tracts which would leave one of the two haplotype tracts intact.

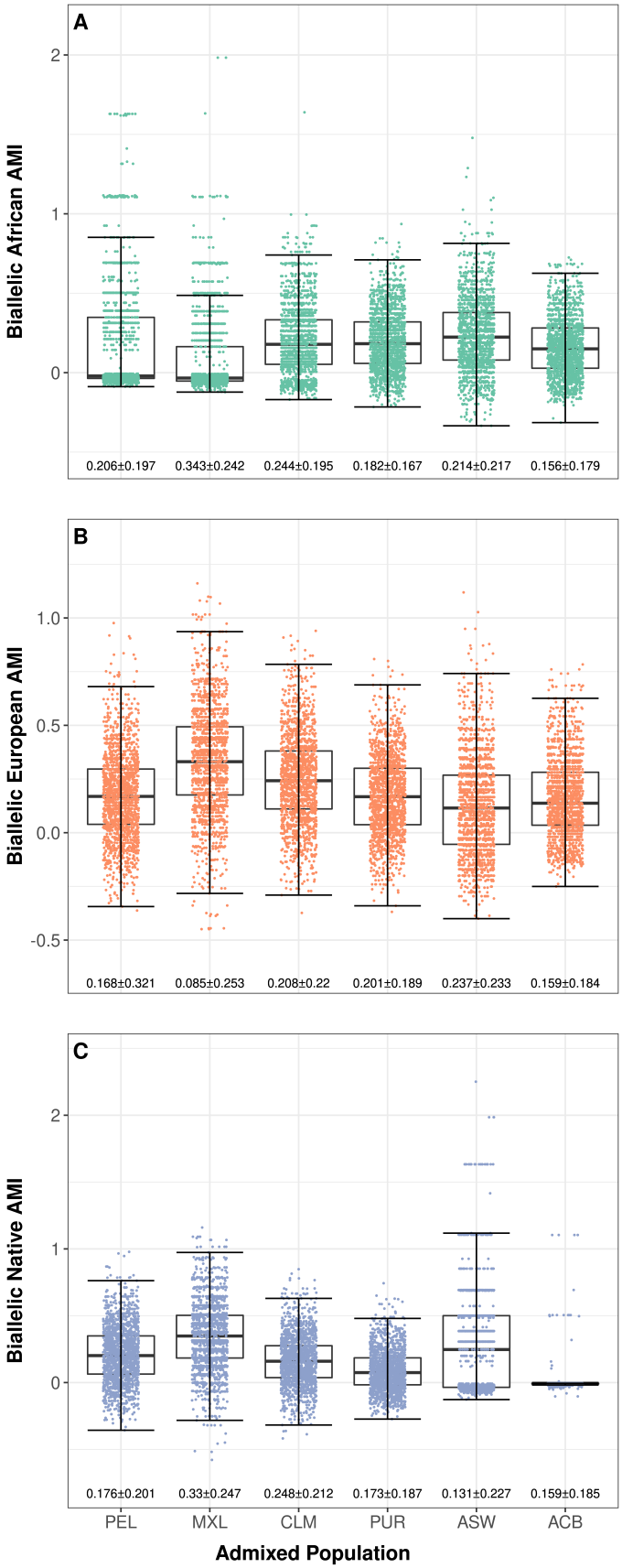

Figure 4: Comparative box plots displaying the distribution of biallelic ancestry-specific assortative mating indices (AMIs) calculated for each studied single nucleotide polymorphism for each studied admixed population. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers signify the last data point within the closest quartile value plus 150% of the interquartile range. Mean ± standard deviation is given beneath; the standard error of the mean is negligible owing to the sample size of 4,111,226. Horizontal jitter is used simply to better display the distribution.

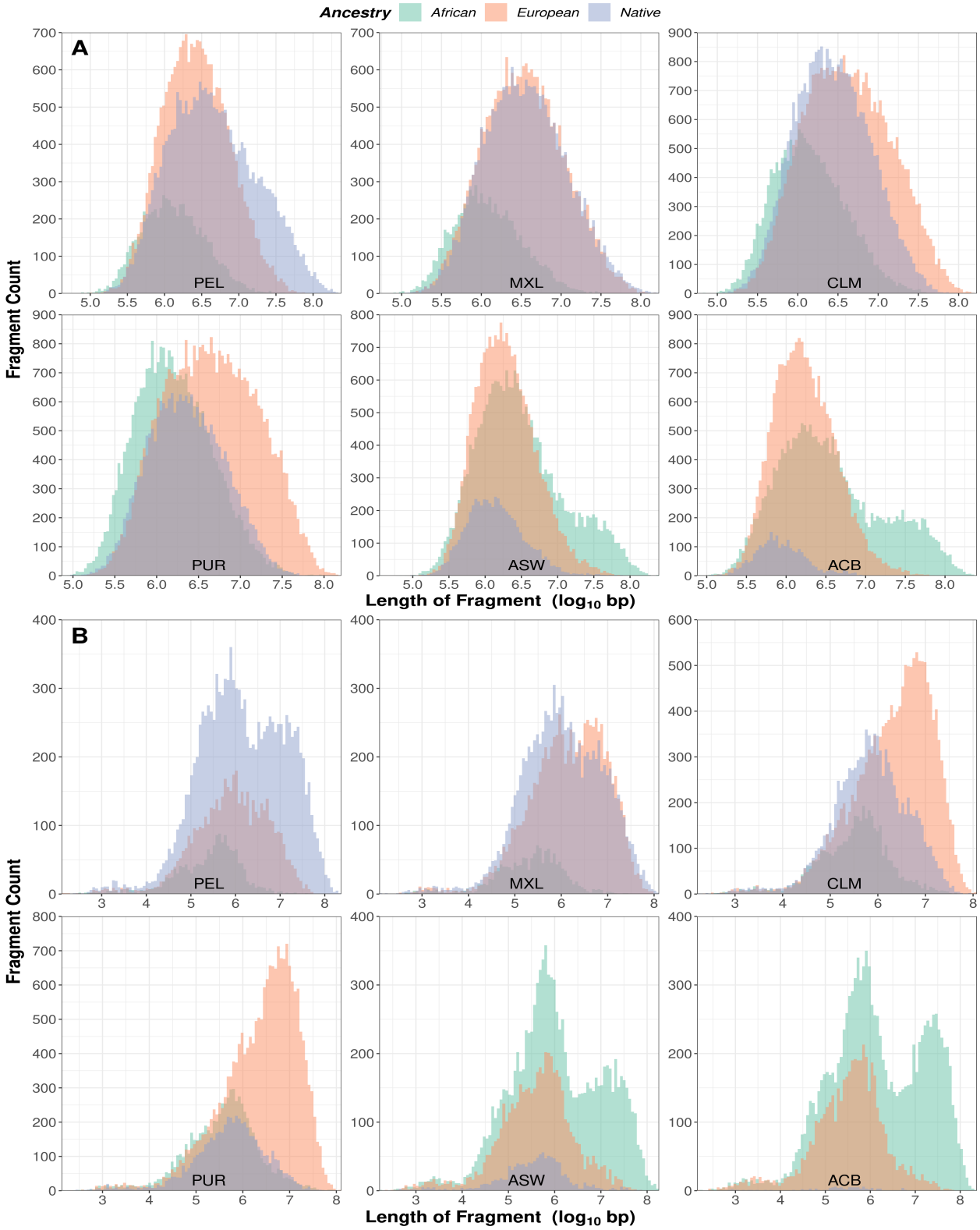

Figure 5: Histograms of continuous ancestry tract lengths of each of the three ancestries for each admixed population. Fragment lengths are measured in base pairs in log10 scale, and are separated into 100 bins in each plot. Fragment length is considered either the number of consecutive haplotype assignments of a given ancestry on a single strand (A), or the number of consecutive homozygous genotype assignments of a given ancestry on both strands taken together (B).

A more sophisticated software for continuous ancestry tract length analysis is TRACTS, which uses tract length distributions to infer how many generations ago migration events took place. Cross-referencing this with relevant slave migration data provides a picture of delays between

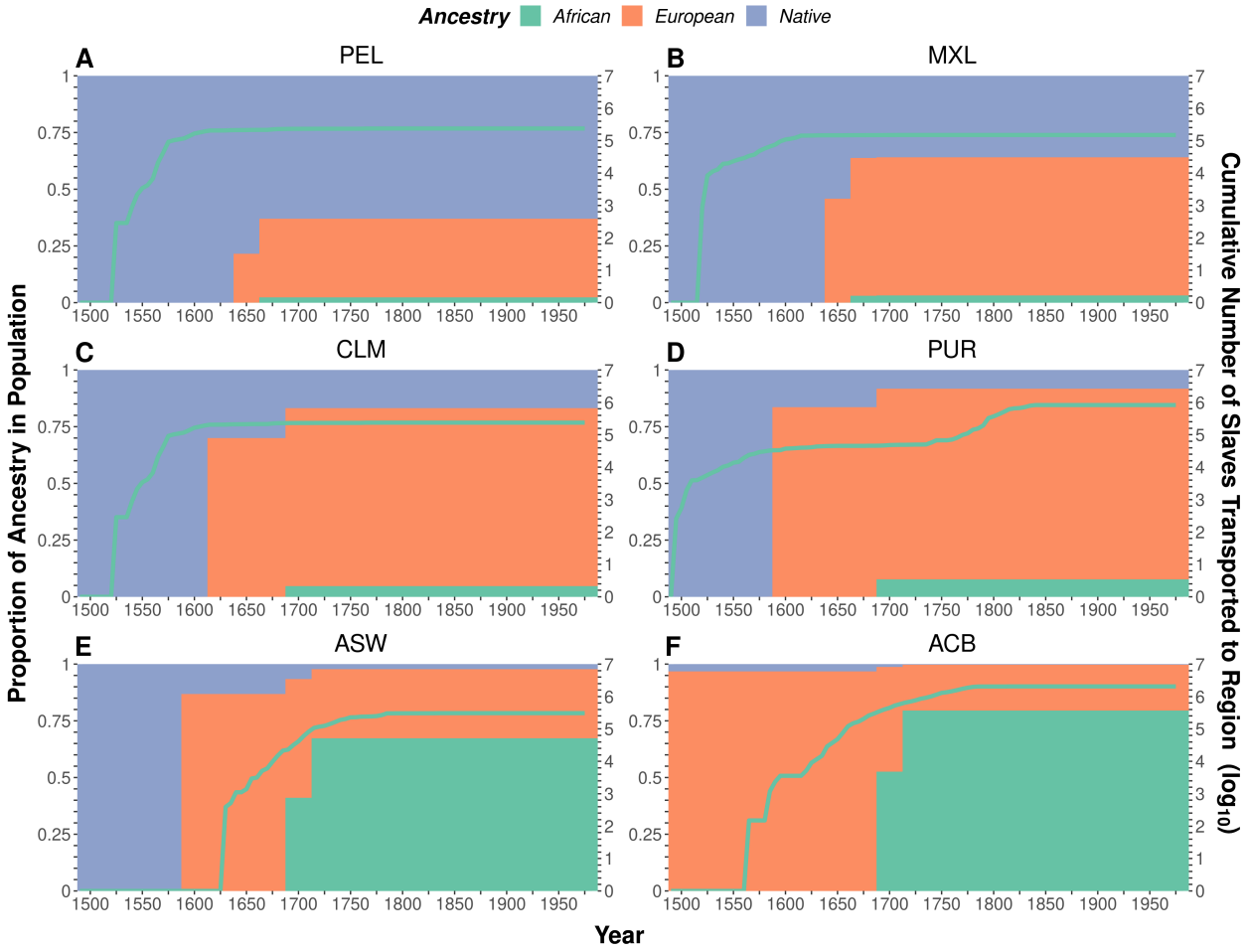

Figure 6: Number of slaves transported to the general regions of the studied admixed populations every five years, and stacked barplots showing how the proportion of the three ancestries changed in said populations generation to generation as estimated by TRACTS, between the years 1500 and 2000 CE. Data on the number of slaves transported to each region, in log10 scale, are cumulative estimates of slaves disembarked there every 5 years based on records of trans-atlantic slave voyages from https://www.slavevoyages.org. Regions used are ports in North-Eastern South America for PEL & CLM, ports in what is now Mexico for MXL, ports on Spanish Caribbean islands for PUR, ports north of the Rio Grande in North America for ASW, and ports on British Caribbean islands for ACB. Genomic data of the individuals from each admixed population was analysed with TRACTS with 25 bootstraps, with generations being estimated as 25-year periods.

migration and significant admixture: assortative mating (**Fig. 6**). As it was Europeans that transported slaves across the Atlantic, we know Europeans arrived in the region the same generation that Africans began to arrive, or earlier.

Therefore, for example in Peru, we can see that Europeans and Africans began arriving around 1525, the majority of Africans had arrived by 1575. Significant admixture between Europeans and Natives occurred around 1650, and significant admixture between Africans and the rest of the population occurred around 1675. This suggests extreme assortative mating for 4-6 generations in Europeans and a similar length, albeit lagging by a generation, in Africans.

While the Mexican and Colombian plots can be interpreted similarly, the other three seem to suggest that the most significant African admixture occurred prior to 80-95% of the slaves being transported to the region, and spuriously that Europeans arrived at Barbados long before evidence suggests.

4 Discussion

Assortative mating is regularly disregarded in population genetics, however most past and present societies are stratified. As such, in studies investigating admixture in a population, one cannot assume negligible assortative mating. Demonstrating the presence of assortative mating in different populations with vastly different ancestral compositions should highlight this point.

The complexity of the demographic history of the colonisation of the Americas between various different spatiotemporal contexts has led to vast differences in ancestry proportions of the populations there. They vary widely between majority Native, European or African (**Fig. 1**) depending on the size of the pre-Colombian Native population, the history of slavery in the society, how many Europeans settled there and a plurality of other factors. Significant differences in ancestry proportions are also observed between individuals (**Fig. 2**), which points to sustained patterns of assortative mating.

However, sampled populations like ASW (Americans of Sub-Saharan African Ancestry from Oklahoma in Southwest USA) and MXL (Mexican Ancestry in Los Angeles California) could potentially present significant sample biases. The ASW samples will likely have more African than the average US Southwest population, and MXL samples more European and possibly African ancestry than the average Mexican.

At the same time, errors in the local ancestry inference with RFMIX could lead to biases both in the proportions of the three genetic ancestries and in the length of the local ancestry fragments associated with them. In this sense, an imbalanced reference panel (72 Native samples to 507 European and 550 African) could underrepresent Native American ancestry. Similarly, based on an RFMIX correlation with ADMIXTURE **Fig. S2**, ADMIXTURE seems to estimate 100% African or 0% European more readily than RFMIX, suggesting it may be less sensitive at those two extremes. Hence RFMIX ancestry proportion estimates might have been the better choice to proceed with, although an instrumental systematic error like this is unlikely to significantly impact subsequent analyses.

Assortative mating is present in all admixed populations across the Americas studied herein. That was the initial hypothesis, and it was supported in full by the results of the AMI analysis, both with global and ancestry-specific AMI (**Fig. 3-4**). Despite differing demographic histories, the levels of assortative mating in the studied populations were both significant, and significantly different to one another.

AMI is calculated based on deviations from Hardy-Weinberg equilibrium. HWE is expected in a population following a single generation of truly random admixture (Smithjohn et al., 2015). This means two generations after large-scale migration, a population without ancestry-related social stratification, and insignificant levels of assortative mating, should exhibit AMI values of zero. Given all query populations exhibited significant levels, if we assume none of the samples are from first-generation immigrants then we can conclude there has been ancestry-related assortative mating at least during the most recent generations. As such, filtering for HWE might be an inappropriate quality control in population genetics studies, a practice still widely used (Linares-Pineda et al., 2012; Smithjohn et al., 2015).

However, the AMI analysis method only allows us to reach that conclusion for the present-day populations: it tells us little about their past. The analysis of the length of continuous local

ancestry tracts is better-suited to evaluating the whole admixture process, from the arrival of first Europeans until the present, but the results are less conclusive. The right skewness of the distribution in the histogram visualisation shows the admixture event cannot be explained by a single pulse of gene flow followed by random mating (**5A**). However, discerning between multiple migration pulses from assortative mating remains difficult. Using the length of genomic fragments that have the same local ancestry in both homologous chromosomes in each sample as a summary statistic shows more promise for the disentangling of the two scenarios (**5B**). Here, a second peak of longer fragments cannot be introduced by assortative mating alone: it requires at least a second pulse of migration. Hence, we can see ASW and ACB populations experienced more than one migration pulse from African populations, while PEL received at least two gene flow pulses from Native American populations.

The method integrating TRACTS with slave voyage data also informs us about the past by quantifying time since admixture, assuming random mating. Deviations in the dates of admixture inferred by TRACTS from the historical records used for the slave voyage data might indicate assortative mating. The **Fig. 6** plots for the Peruvian, Mexican and Colombian populations are intuitive and historically plausible, although more detailed research into whether documentation of the period corroborates the projections should be conducted. In the other query populations the majority of the slaves were purportedly transported after the generation of most significant admixture, approximately by an order of magnitude in all three cases. It is perhaps no coincidence that these are the three query populations for which the slave voyages data regions used were most geographically unspecific. Hence, more thoroughly researching the history of trans-Atlantic slavery of these three regions, and thus more accurately determining the ports at which slaves destined for Barbados, Puerto Rico and the US Southwest initially disembarked from their voyage, may bring those plots more into line with the others.

Specifically for the Barbados plot, even with the concept of an initial Native population hard-coded into the model, the algorithm could only explain the genomic pattern by predicting that Europeans arrived 50-100 generations ago. This is 500 or more years before it is known to have occurred. This is likely related to ADMIXTURE estimating that only 2 out of the 96 samples contain any Native DNA, both with a proportion of less than 0.1: a stark reminder that assortative mating and migration were not the only population-shaping phenomena at play in the colonial-era Americas.

These are not the only issues with this TRACTS analysis. The analysis uses ancestry proportion, not absolute quantity of genetic material. This means Native populations shrinking due to disease and other consequences of colonialism would have the same effect of increasing European ancestry proportion in the population as European migration. When interpreting TRACTS plots it must also be remembered that TRACTS is constrained to only one pulse per ancestry, at the generation it deems to have had the biggest effect on the proportion of that ancestry. Finally, an inherent flaw in analysing social stratification using generation as the unit of time - albeit unavoidable in genetic research - is that generation length is likely to differ significantly by subgroup, and indeed over time. In a truly stratified society, one would expect different stratas to have different generation lengths.

In this paper, a method for detecting assortative mating in a population, a method that highlights past migration events, and a method that suggests periods of assortative mating following

pulses of migration were established. But ultimately, migration and the decline of assortative mating in the past - both the removal of barriers to admixture - manifest themselves near-identically. Therefore, the two are seemingly inextricable when projecting into the past, absent accurate migration data to explain the contribution of migration to admixture. Without more comprehensive inter- and intracontinental migration data, limited to the records remaining from that era, that problem is presently unsolved.

However, it may be possible to use artificial neural networks to circumvent this need for migration data. Firstly, a model to predict continuous ancestry tract length distribution based on input parameters such as level of assortative mating must be created. This model can be used to simulate tract length distributions with every combination of input parameters. An artificial neural network can then be trained to learn the patterns between these distributions and the corresponding parameters. In theory, it may subsequently be able to accurately predict the parameters, including level of assortative mating in the population, when applied to the tract length distributions generated in this study with empirical data. Artificial neural networks have been successfully trained in this way before (Sheehan & Song, 2016).

The AMI analysis, having been established as a legitimate technique for distinguishing between migration and assortative mating, could be used to monitor ancestry-related social stratification. As genome sequencing gets cheaper, larger and more selected sample sizes will enable more reliable results. Samples could be taken from those in small age windows in increments of say 10 years to get a picture of such stratification in a population for the past few generations. Following this, samples could be taken from young people every 10 years to keep track of it in the long-term, perhaps to inform governmental policy.

While artificial neural networks have potential in bypassing the need for migration data, integrating such data into the simulation model would make the method even more powerful. This may not be possible with the current records of migration in the colonial-era Americas, but could be used in modern populations. Much higher-quality migration data is available, although globalisation is leading to increasing ancestral diversity in populations, and accounting for more ancestries adds complexity. Like the AMI analysis, this could have promise in the monitoring of ancestry-related social stratification in modern populations.

To further increase the accuracy of these methods in quantifying assortative mating, another factor must be considered. Admixture of two ancestries may seem antithetical to ancestry-related social stratification, but not all admixture in a population is mutually voluntary. Many instances of admixture between slave master and slave, or colonist and Native, were the result of rape and thus actually symptomatic of social stratification. To prevent such events from counteracting assortative mating as a proxy for ancestry-related social stratification, this would ideally be quantified and integrated into the model. If it were assumed that negligible instances of this occurred between European Females and Native or African Males relative to the inverse, similar analyses testing the recombining section of the X chromosome rather than autosomes could be conducted. Including the sex chromosomes would add another layer of complexity to the study of the admixture process by unveiling sex bias patterns. Sex bias reflects differences in ancestry proportions between mates. Analysing assortative mating together with sex bias would not only allow us to analyse social stratification patterns, but also reveal the direction of the social hierarchies linked to gender and race (Micheletti et al., 2020).

Assortative mating has long been neglected as a factor influencing admixture, whether in research into the effects of migration on a population's genome or in genetic marker selection for genome-wide association studies. By improving upon and developing the methods utilised and suggested in this paper, powerful tools for the estimation of past and present assortative mating may be possible. Not only would this allow us to correct for assortative mating in the aforementioned studies, but it would enable us to better understand the history of human societies and may even enable us to monitor, highlight and thus perhaps discourage present-day ancestry-related social stratification.

5 Data and Code Availability

5.1 Data

1KGP Samples:
https://www.internationalgenome.org/data-portal/data-collection/30x-grch38

HGDP Samples:
https://www.internationalgenome.org/data-portal/data-collection/hgdp

Phasing Reference Panel:
http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/20201028_3202_phased/

Phasing Genetic Map:
https://github.com/odelaneau/shapeit4/blob/master/maps/genetic_maps.b38.tar.gz

Slave Voyage Data:
https://www.slavevoyages.org/voyage/database#tables (see tracts_mig_plots.R for details)

5.2 Code

Code Repository:
https://github.com/Bennouhan/cmeecoursework/tree/master/project/code

A detailed visualisation of the project's workflow can be found in **Fig. S6**, indicating which script(s) were used during each step in the analyses. See the README.md for further details.

References

Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, *19*(9), 1655–1664. https://doi.org/10.1101/gr.094052.109

Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Giordano, B. S. S., Leal, T. P., Furlan, V., Scliar, M. O., Zamudio, R., Zolini, C., Araújo, G. S., Luizon, M. R., Padilla, C., Cáceres, O., Levano, K., Sánchez, C., Trujillo, O., Flores-Villanueva, P. O., Dean, M., ... Tarazona-Santos, E. (2020). The genetic structure and adaptation of Andean highlanders and Amazonians are influenced by the interplay between geography and culture. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(51), 32557–32565. https://doi.org/10.1073/pnas.2013773117

Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M., Bustamante, C. D., & Ostrer, H. (2010). Genome-wide patterns of population structure and admixture among Hispanic/Latino populations. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(SUPPL. 2), 8954–8961. https://doi.org/10.1073/pnas.0914618107

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2)arXiv 2012.10295, 1–4. https://doi.org/10.1093/gigascience/giab008

Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accurate, scalable and integrative haplotype estimation. *Nature Communications*, *10*(1), 24–29. https://doi.org/10.1038/s41467-019-13225-y

e Silva, M. A. C., Nunes, K., Lemes, R. B., Mas-Sandoval, À., Amorim, C. E. G., Krieger, J. E., Mill, J. G., Salzano, F. M., Bortolini, M. C., da Costa Pereira, A., Comas, D., & Hünemeier, T. (2020). Genomic insight into the origins and dispersal of the Brazilian coastal natives. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(5), 2372–2377. https://doi.org/10.1073/pnas.1909075117

Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, *191*(2)arXiv 1202.4811, 607–619. https://doi.org/10.1534/genetics.112.139808

Linares-Pineda, T. M., Cañadas-Garre, M., Sánchez-Pozo, A., Calleja-Hernández, M., D'Haens, G. R., Panaccione, R., Higgins, P. D., Vermeire, S., Gassull, M., Chowers, Y., Hanauer, S. B., Herfarth, H., Hommes, D. W., Kamm, M., Löfberg, R., Quary, A., Sands, B., Sood, A., Watermayer, G., ... Yang, J. (2012). Quality Control Procedures for Genome Wide Association Studies. *American Journal of Human Genetics*, *573*(6), 5–22. https://doi.org/10.1002/0471142905.hg0119s68.Quality

Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A discriminative modeling approach for rapid and robust local-ancestry inference. *American Journal of Human Genetics*, *93*(2), 278–288. https://doi.org/10.1016/j.ajhg.2013.06.020

Mas-Sandoval, A., Arauna, L. R., Gouveia, M. H., Barreto, M. L., Horta, B. L., Lima-Costa, M. F., Pereira, A. C., Salzano, F. M., Hünemeier, T., Tarazona-Santos, E., Bortolini, M. C., & Comas, D. (2019). Reconstructed Lost Native American Populations from Eastern Brazil

Are Shaped by Differential Jê/Tupi Ancestry. *Genome Biology and Evolution*, *11*(9), 2593–2604. https://doi.org/10.1093/gbe/evz161

Micheletti, S. J., Bryc, K., Ancona Esselmann, S. G., Freyman, W. A., Moreno, M. E., Poznik, G. D., Shastri, A. J., Agee, M., Aslibekyan, S., Auton, A., Bell, R., Clark, S., Das, S., Elson, S., Fletez-Brant, K., Fontanillas, P., Gandhi, P., Heilbron, K., Hicks, B., ... Mountain, J. L. (2020). Genetic Consequences of the Transatlantic Slave Trade in the Americas. *American Journal of Human Genetics*, *107*(2), 265–277. https://doi.org/10.1016/j.ajhg.2020.06.012

Norris, E. T., Rishishwar, L., Chande, A. T., Conley, A. B., Ye, K., Valderrama-Aguirre, A., & Jordan, I. K. (2020). Admixture-enabled selection for rapid adaptive evolution in the Americas. *Genome Biology*, *21*(1), 1–29. https://doi.org/10.1186/s13059-020-1946-2

Norris, E. T., Rishishwar, L., Wang, L., Conley, A. B., Chande, A. T., Dabrowski, A. M., Valderrama-Aguirre, A., & King Jordan, I. (2019). Assortative mating on ancestry-variant traits in admixed Latin American populations. *Frontiers in Genetics*, *10*(APR), 1–14. https://doi.org/10.3389/fgene.2019.00359

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., Maller, J., Sklar, P., De Bakker, P. I., Daly, M. J., & Sham, P. C. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela, R., Rodriguez-Santana, J. R., Rodriguez-Cintron, W., Avila, P. C., Ziv, E., & Gonzalez Burchard, E. (2009). Ancestry-related assortative mating in Latino populations. *Genome Biology*, *10*(11). https://doi.org/10.1186/gb-2009-10-11-r132

Schubert, R., Andaleon, A., & Wheeler, H. E. (2020). Comparing local ancestry inference models in populations of two- And three-way admixture. *PeerJ*, *8*, 1–19. https://doi.org/10.7717/peerj.10090

Sheehan, S., & Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLoS Computational Biology*, *12*(3), 1–28. https://doi.org/10.1371/journal.pcbi.1004845

Smithjohn, M. U., Smith, M. U., & Baldwin, J. T. (2015). Making Sense of Hardy-Weinberg Equilibrium What Is Hardy-Weinberg Equilibrium ? The H-W eq principle is , of course , the cornerstone of introductory population genetics . *The American Biology Teacher*, *77*(8), 577–582. https://doi.org/10.1525/abt.2015.77.8.3.THE

Zaitlen, N., Huntsman, S., Hu, D., Spear, M., Eng, C., Oh, S. S., White, M. J., Mak, A., Davis, A., Meade, K., Brigino-Buenaventura, E., LeNoir, M. A., Bibbins-Domingo, K., Burchard, E. G., & Halperin, E. (2017). The effects of migration and assortative mating on admixture linkage disequilibrium. *Genetics*, *205*(1), 375–383. https://doi.org/10.1534/genetics.116.192138

Supplementary Material

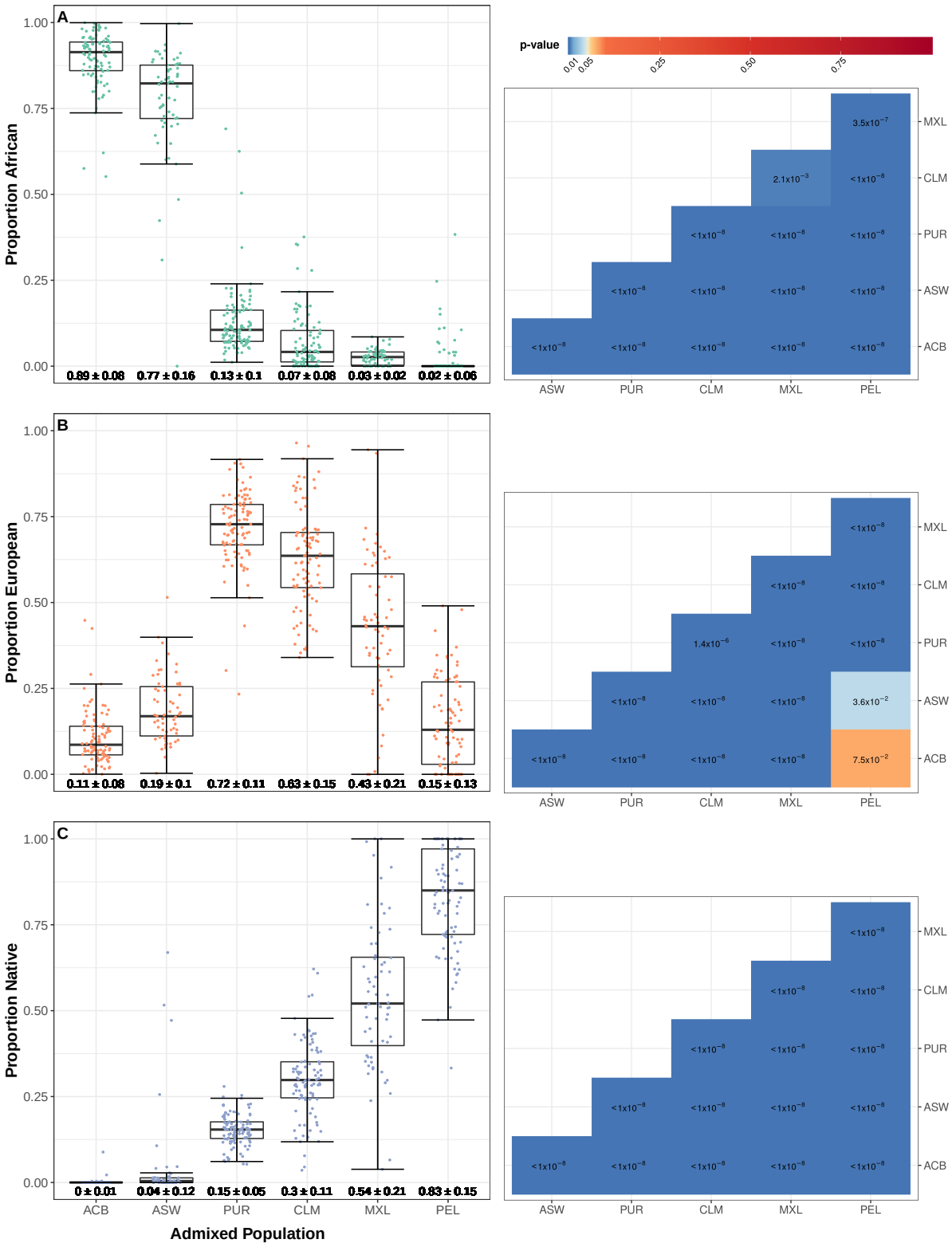

Figure S1: Comparative box plots displaying the distributions of the three ancestry proportions for each individual of each admixed population, with corresponding p-value heatmaps comparing populations statistically. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers signify the last data point within the closest quartile value plus 150% of the interquartile range. Mean ± standard deviation is given beneath. Horizontal jitter is used to better display the distribution. To the right of the boxplots for each ancestry is a corresponding p-value heatmap. These show the results of Wilcoxon tests conducted between every combination of two admixed populations, with shades of blue indicating differences between populations are significant at the 5% level.

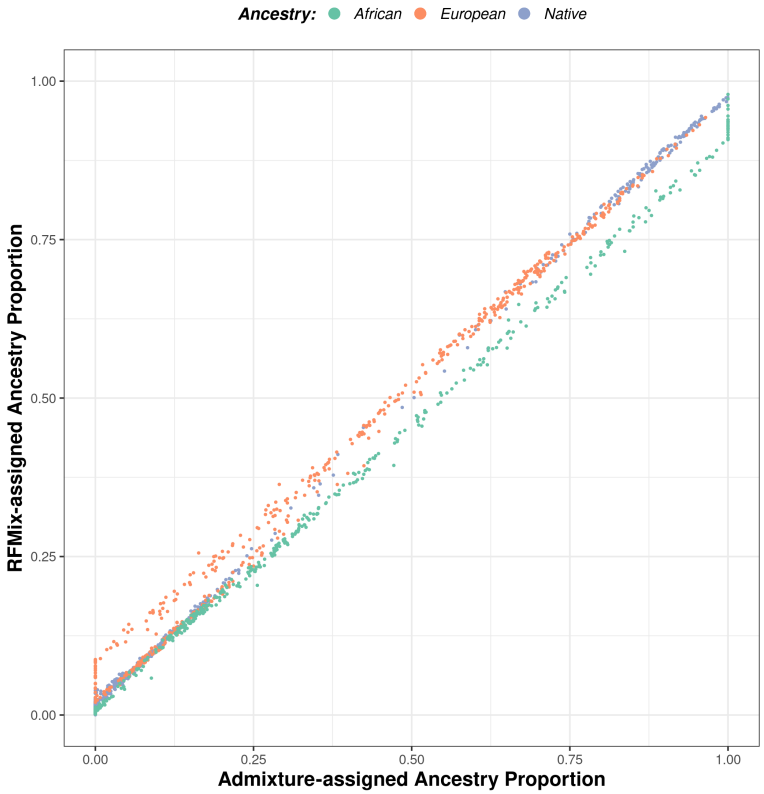

Figure S2: Scatterplot correlating ancestry proportions assigned by RFMIX for all 1690 query and reference individuals against those assigned by ADMIXTURE.

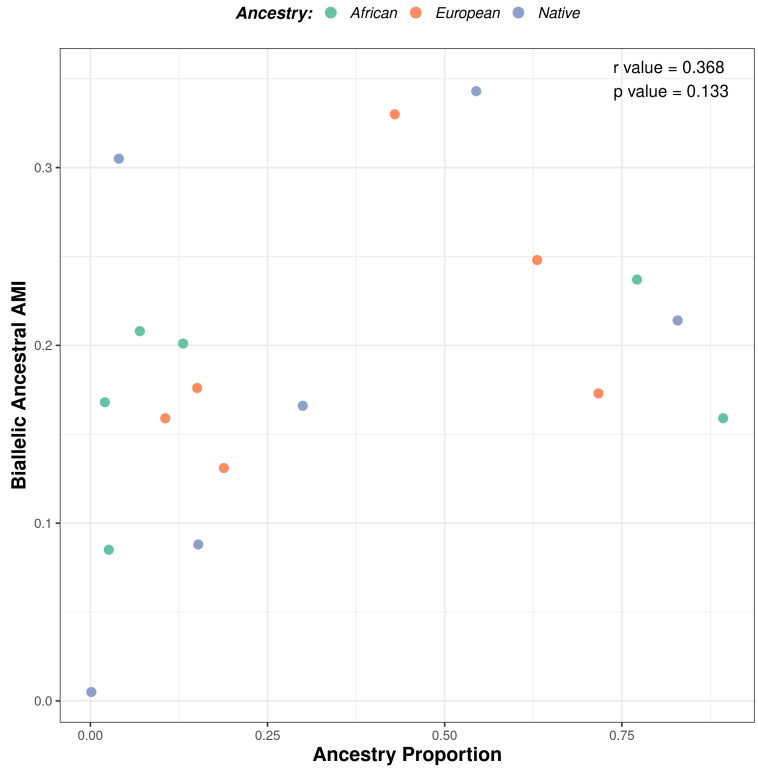

Figure S3: Scatterplot charting all three mean biallelic ancestry-specific AMI against all three ancestry proportions for each of the six admixed populations.

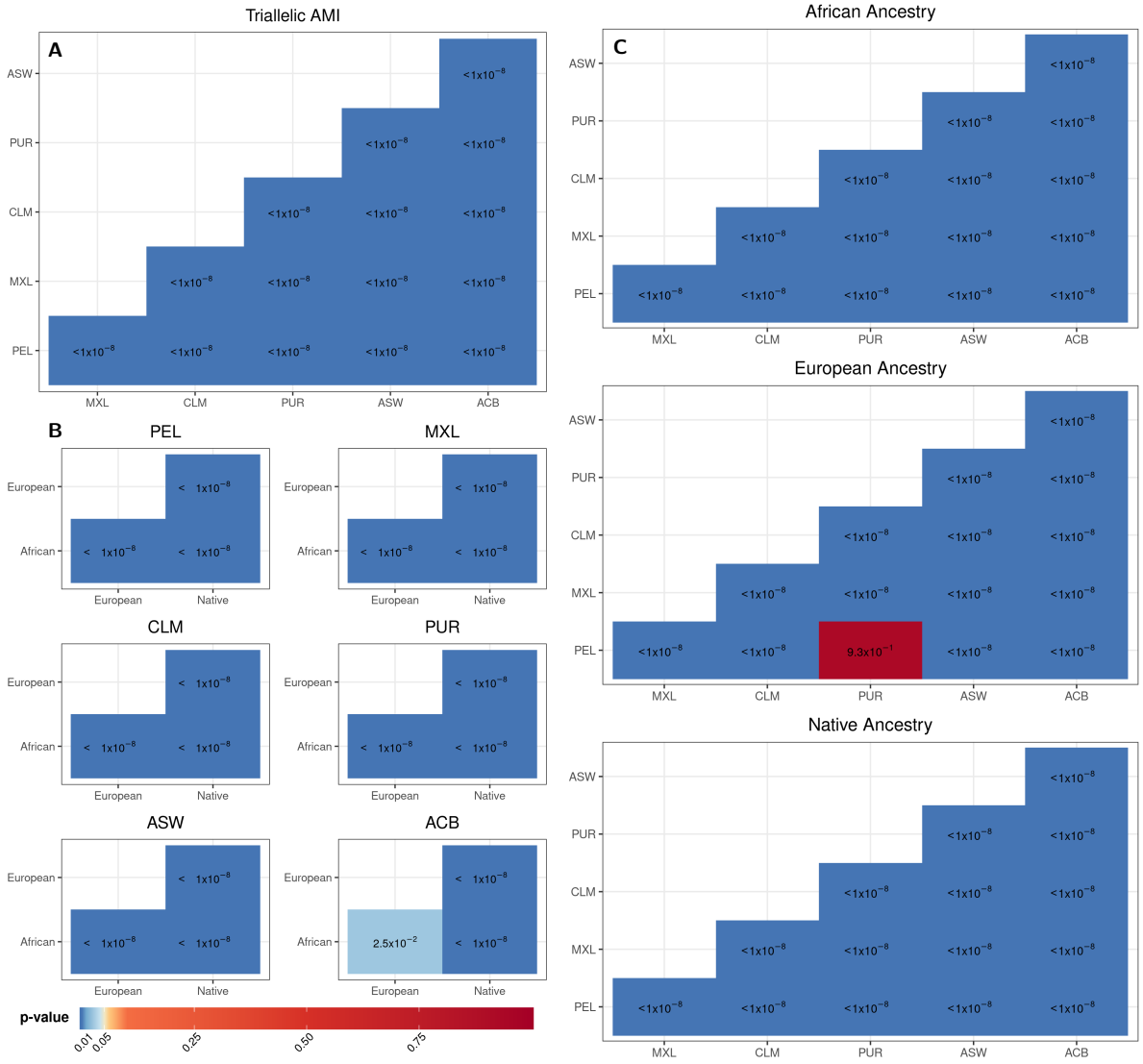

Figure S4: Heatmaps displaying p-value results of Wilcoxon tests used to compare assortative mating index values of different populations and ancestries. Each set of heatmaps correspond to a different set of comparisons between all combinations of assortative mating index (AMI) distributions. **A** compares all combinations of the six admixed populations with regards to their triallelic AMI distributions, shown in **Fig. 3**. **B** compares all combinations of the three ancestries with regards to their biallelic ancestry-specific AMI distributions, for each of the six admixed populations. **C** compares all combinations of the six admixed populations with regards to their biallelic ancestry-specific AMI distributions, for each of the three ancestries, shown in **Fig. 4A-C**. Shades of blue indicate differences between populations or ancestries are significant at the 5% level.

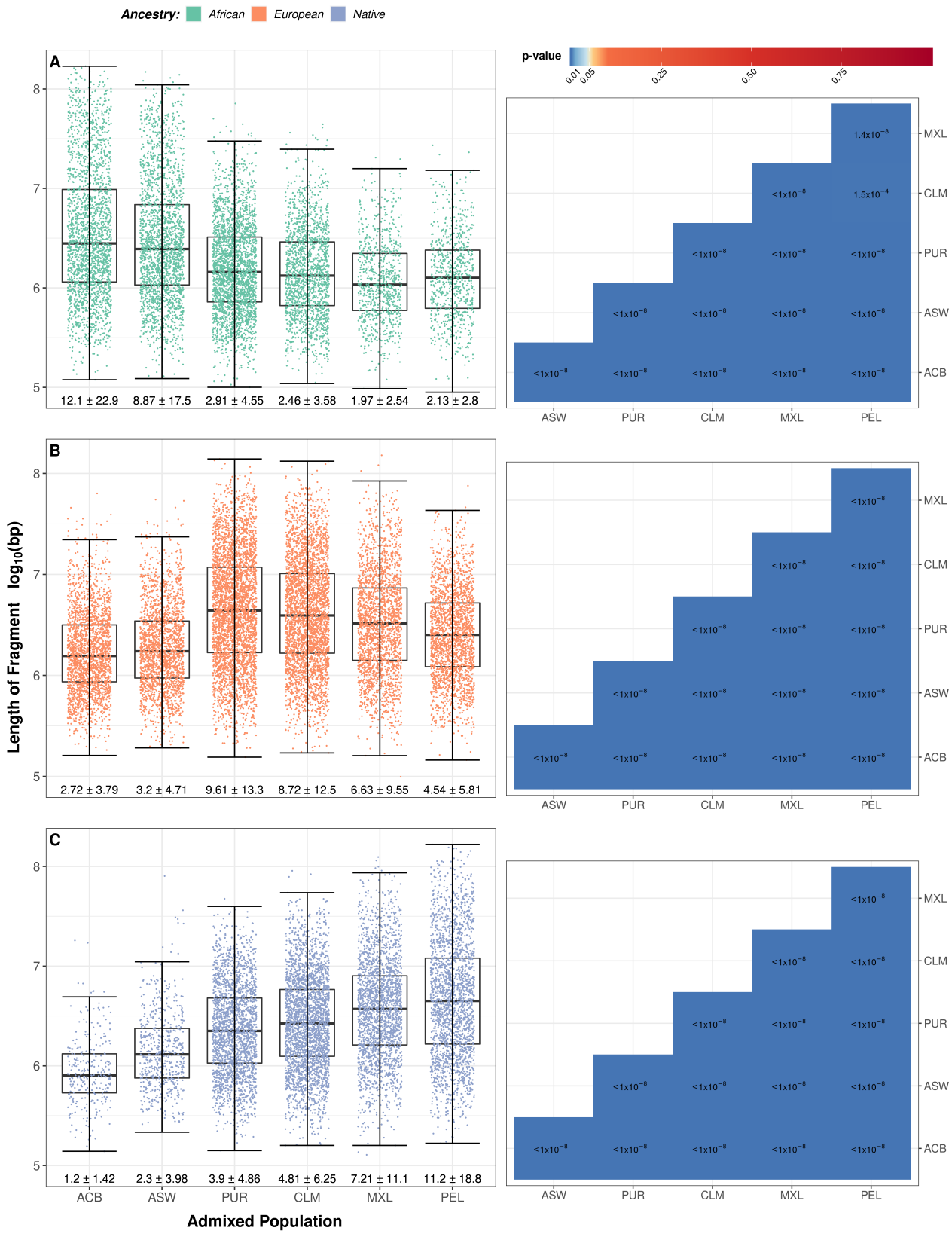

Figure S5: Comparative box plots displaying the distributions of continuous ancestry tract lengths of each ancestry for all individuals of each admixed population, with corresponding p-value heatmaps comparing populations statistically. Fragment length, that is the number of consecutive haplotype assignments of a given ancestry on a single strand, are measured in base pairs in log10 scale. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers signify the last data point within the closest quartile value plus 150% of the interquartile range. Mean ± standard deviation is given beneath in units of Mbp. Horizontal jitter is used to better display the distribution. To the right of the boxplots for Afrinca, European and Native ancestries (**A-C**) is a corresponding p-value heatmap. These show the results of Wilcoxon tests conducted between every combination of two admixed populations, with shades of blue indicating differences between populations are significant at the 5% level.

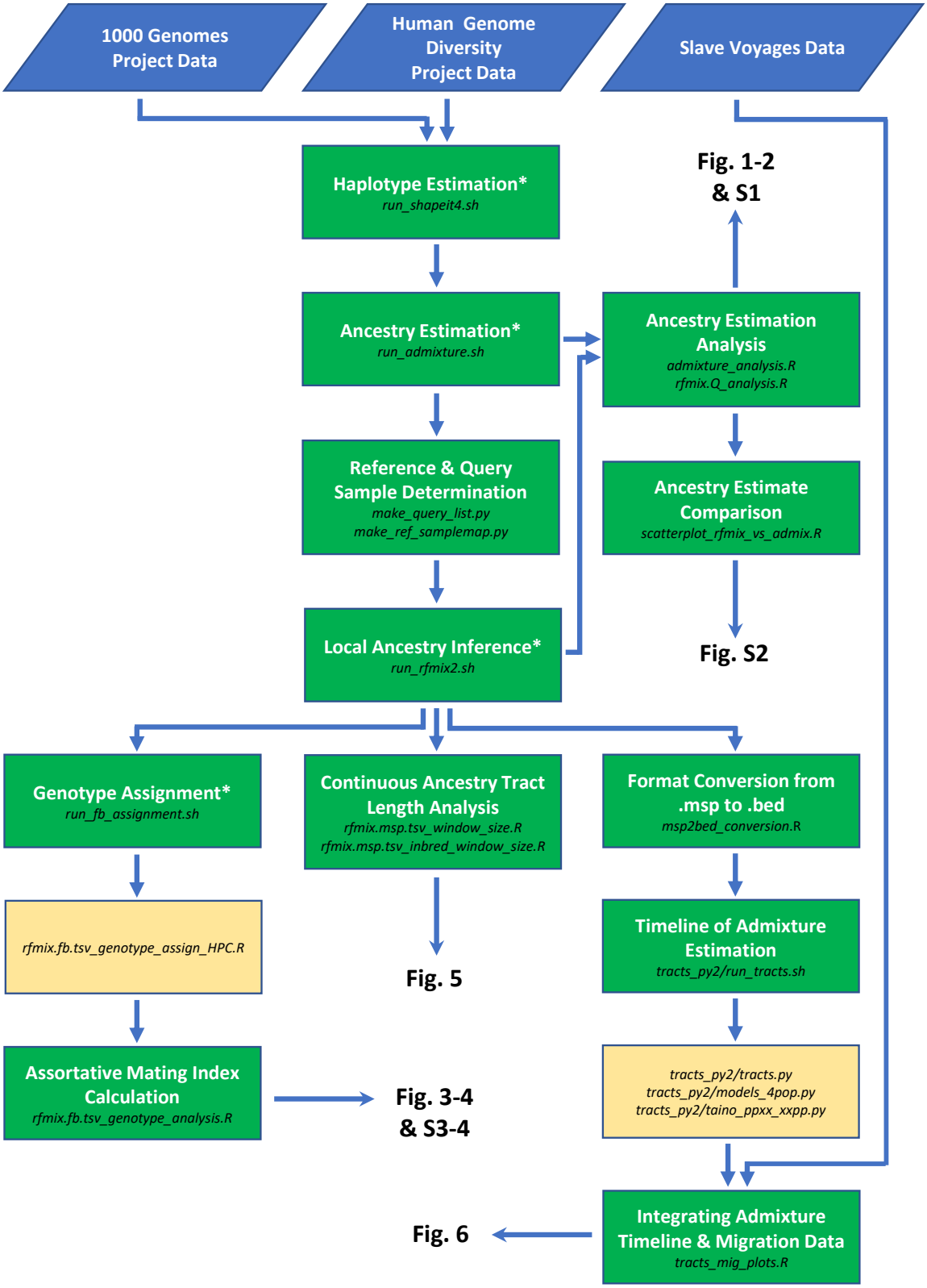

Figure S6: Flowchart representing the analysis workflow of the project, from input data to the output figures. Arrows indicate that the output from one step is the input for the next. Below the label of each step is the script(s) from the provided github repository required to run that step. The scripts named in the unlabelled yellow boxes are run automatically by the script in the previous step. Asterisked step labels indicate this step was performed on a high-performance computer due to the computational power required.