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

Ben Nouhan, bjn20@ic.ac.uk

Imperial College London

August 23, 2021

Word Count: 4586

250 word abstract placeholder: 10000 10001 10002 10003 10004 10005 10006 10007 10008 10009 10010 10011 [13] 10012 10013 10014 10015 10016 10017 10018 10019 10020 10021 10022 10023 [25] 10024 10025 10026 10027 10028 10029 10030 10031 10032 10033 10034 10035 [37] 10036 10037 10038 10039 10040 10041 10042 10043 10044 10045 10046 10047 [49] 10048 10049 10050 10051 10052 10053 10054 10055 10056 10057 10058 10059

[61] 10060 10061 10062 10063 10064 10065 10066 10067 10068 10069 10070 10071 [73] 10072 10073 10074 10075 10076 10077 10078 10079 10080 10081 10082 10083 [85] 10084 10085 10086 10087 10088 10089 10090 10091 10092 10093 10094 10095 [97] 10096 10097 10098 10099 10100 10101 10102 10103 10104 10105 10106 10107

[109] 10108 10109 10110 10111 10112 10113 10114 10115 10116 10117 10118 10119 [121] 10120 10121 10122 10123 10124 10125 10126 10127 10128 10129 10130 10131[133] 10132 10133 10134 10135 10136 10137 10138 10139 10140 10141 10142 10143[145] 10144 10145 10146 10147 10148 10149 10150 10151 10152 10153 10154 10155

[157] 10156 10157 10158 10159 10160 10161 10162 10163 10164 10165 10166 10167 [169] 10168 10169 10170 10171 10172 10173 10174 10175 10176 10177 10178 10179 [181] 10180 10181 10182 10183 10184 10185 10186 10187 10188 10189 10190 10191

[193] 10192 10193 10194 10195 10196 10197 10198 10199 10200 10201 10202 10203 [205] 10204 10205 10206 10207 10208 10209 10210 10211 10212 10213 10214 10215 [217] 10216 10217 10218 10219 10220 10221 10222 10223 10224 10225.

Contents

1 Introduction

Positive assortative mating, a genetic phnomenon 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).

This multigenerational non-random admixture between genetically distinct groups leaves a genomic imprint in the individuals comprising the population, in stark contrast to populations more closely following Hardy-Weinberg equilibrium (HWE) (Zaitlen et al., 2017). However, when studying the admixture or lack thereof between two or more groups, geographical barriers such as oceans must be considered alongside social ones. Afterall, large scale immigration of a new ethnic group will genetically manifest itself similarly to, for example, the revocation of racial segregation policies applying to that same ethnic group: both events facilitate future admixture between those groups.

Ancestry-informative markers (AIMs) are single nucleotide polymorphisms (SNPs) than indicate an individual is of a certain ancestry (Risch et al., 2009). Population genomics techniques allow us to generate a large array of AIMs which can be analysed using local ancestry inference to map ancestries to positions and regions along the genome, after which further analysis can indicate past assortative mating in a population (Schubert et al., 2020).

One such analysis is that of continuous ancestry tract (CAT) lengths: the lengths of genomic regions wherein AIMs are consecutively assigned to the same ancestry. Looking at the distribution of these lengths, the ancestry to which they belong and the overall ancestry proprtion 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 CAT lengths, as CATs within the parents' genomes interupt one another upon recombination, hence admixture more generations ago will manifest as distributions of shorter CATs and vice versa (Gravel, 2012).

AIM 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:heterozygosity ratio at a given position based on this concept, which 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 in genome-wide association study (GWAS) - equivalent to AIMs but for pathological research - whereby alleles with frequencies deviating too far from it are removed and considered sequencing misreads (Linares-Pineda et al., 2012). This does not take into account stratificatiom, present in most if not all societies, within the populations studied herein. Showing allelic deviation from HWE is not an artefact but rather an intrinsic quality may serve as a warning against this practice.

Populations of the Americas such as Colombia, Barbados, Mexico or the US provide apropriate and well-researched case studies integrating migration, admixture and assortative mating. Many of 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 Eu-

ropean population grows via migration. These phenomena continue such that African slaves are transported to the region to supplement or replace the Natives; and colonialism eventually ends in the region after which the population continues evolving with these historical scars (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 meeting points between assortative mating and migration, and with the current geopolitical and economic landscape - globalisation, industrialisation, wars, shifting demographics and global warming - they will not be the last. Further understanding and ideally quantifying ancestry-based assortative mating, and using it as a proxy for ancestry-based 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 of each constiuent 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 rathen 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 casues and long-term effects of ancestry-based 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 unveil non-random admixture in a population. This will enable me to compare these methods by their potential to distigunish 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 are significantly different to that of the others.

Only by reconciling migration and assortative mating can we confidently infer assortative mating from genomic data, and use this to draw conclusions about past and make predictions about future ancestry-based 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, alongside the genetic similarities between

these populations, Siberans and, by extension, Native Americans.

Table 1: Details of the populations used throughout this study. Populations abbreviated as three capitalised letters 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 SHAPEIT4.2.0, 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 the B38 genetic map recommended by the developers and default

parameters, plus an apropriate high-coverage phased reference genome from the 1KGP website (see: **Data and Code Availability**) to inmprove haplotype estimation accuracy. 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 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). This effectively generates a large array of AIMs upon which to base 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 default parameters and three ancestries to be detected (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% pure samples 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 (Maples et al., 2013). It does this by subjecting the chromosomes to a combination of machine learning methods: discriminant random forests and conditional random field modelling.

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 run-throughs of the algorithm and 20 generations, before which, assuming an average generation length of 25 years, no known European-Native American admixtire had taken place.

2.6 Assortative Mating Index Calculation

One measure of assortative mating is the assortative mating index (AMI), which takes a log odds ratio of 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):

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

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

Here a, e and n in the triallelic model are initials of the ancestries they signify, 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) ouput files were read by the script "rfmix.fb.tsv_genotype_assign_HPC.R". These files contain the estimated haplotype probabilities at each genolmic 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 with respect to each ancestry, at each position.

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 CATs thereby completely alter the CAT length distribution. Hence the .msp.tsv RFMIX output files were used instead, equivalent to the forward-backward files but with automatic haplotype assignment to haploytype with highest estimated liklihood.

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 assigmnets.

2.8 Timeline of Admixture Estimation with TRACTS

TRACTS is a software for modelling migration histories using ancestry tracts data, incorporating the theory of time-dependent gene-flow 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

adjhacent fragments can be of vastly different lengths and always different assigned ancestries. It then recalculates each cell based on this mergeing 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 constituing 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 apropriate. I therefore adjusted the provided four population model, which assumes admixture of two initial populations and subsequently two further populations with three migration events, to instead assume 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.

The _mig file outputs of TRACTS contain what proportion each newly introduced ancestry contributes to the query population after each migration event, and how many generations ago that migration event occurred. The script "tracts_mig_plots.R" uses this data to calculate the estimated relative proportion of the three ancestries during each of the past 25 generations for each query population.

3 Results

3.1 Ancestry Proportion

Pruning led to the dataset being reduced to 4,111,226 AIMs per sample. ADMIXTURE was used on these AIMs to estimate 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 minimially 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** while suggesting aproximately 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 visualised in **Fig. S1**, with the distribution for each population of a given ancestry displayed side-by-side in box plots. All 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, lending credence to the assumption all six admixed populations have entirely different

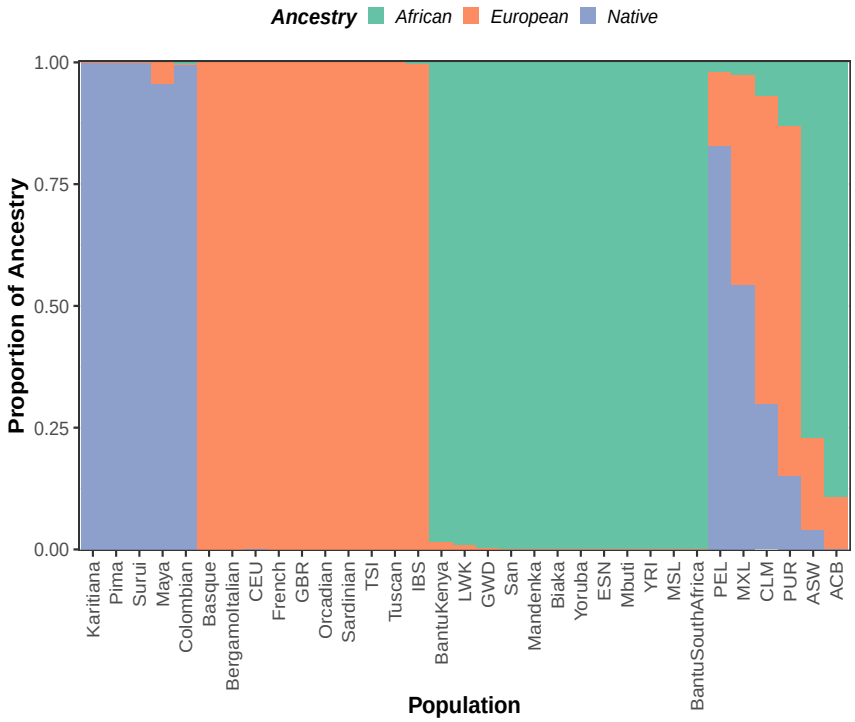

Figure 1: Stacked barplots showing the proportions of the three ancestries of each reference or query population used throughout the study, generated by ADMIXTURE. Genomic data from individuals of selected populations from the 1000 Genomes Project and the Human Genome Diversity Project were processed 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 from the Americas.

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 - somewhat imbalanced - reference panel of 72, 507 and 550 people of 99% or more Native, European and African ancestry 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 those both deem to have higher African Ancestry, and higher European ancestry proportion estimates in those both deem to have lower European Ancestries. ADMIXTURE seems to estimate 100% African or 0% European earlier more readily than RFMIX, suggesting it may be less sensitive at those two extremes. The main RFMIX output was used to calculate assortative mating index values for each AIM 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 AIMs, and the standard deviations being of similar sizes to the corresponding means, the standard errors of the means are negligable and hence the sample means are accurate estimates of the true means. Based on this, we can see all means are 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

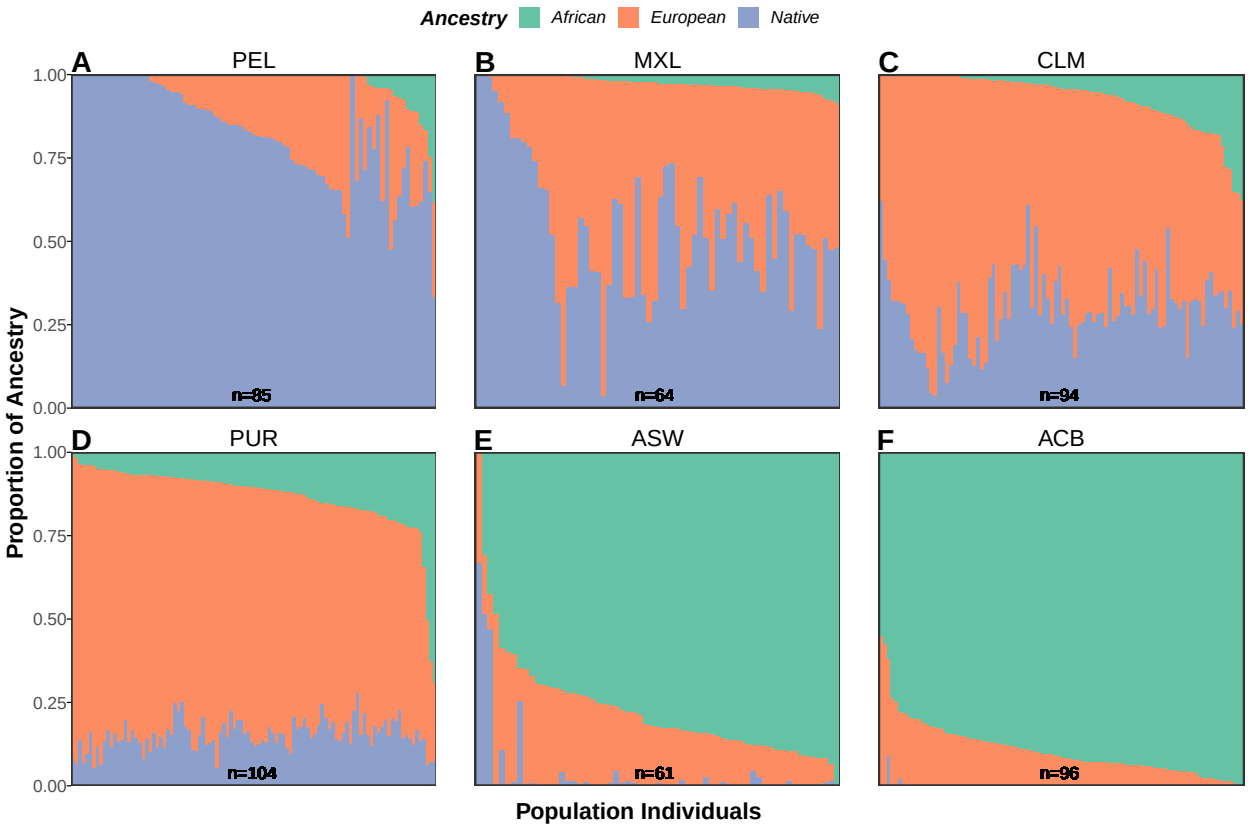

Figure 2: Stacked barplots showing the proportions of the three ancestries of each individual comprising the six query admixed populations, generated by ADMIXTURE. Genomic data from individuals of the six from the 1000 Genomes Project and the Human Genome Diversity Project were processed and subjected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

are significantly different from the other populations, which 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 the 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 distriutions for Puerto Rico and Peru, all combinations of ancestries or populations were significantly different. To test whether mean ADMIXTURE-estimated ancestry proportion is correlated with mean ancestry-specific AMI value they were plotted for each admixed population (**Fig. S3**) but, with a p-value 0f 0.133, no significant correlation was found.

3.3 Continuous Ancestry Tract Lengths

The final use of the RFMIX output was analysisng the lengths of continuous anestry tracts. Displaying the haplotype CATs in a histogram allows visual comparison between the CAT length distriutions of the different ancestries (**Fig. 5A**), while box plots better visualise descriptive statistics of the data (**Fig. S5**). As would be expected, there's a clear correlation between the relative heights and x-axis positions of the distributions in a given population and the corrrespinding mean ancestry proportion. Skewed distributions, such as the right-skewed African distributions of the

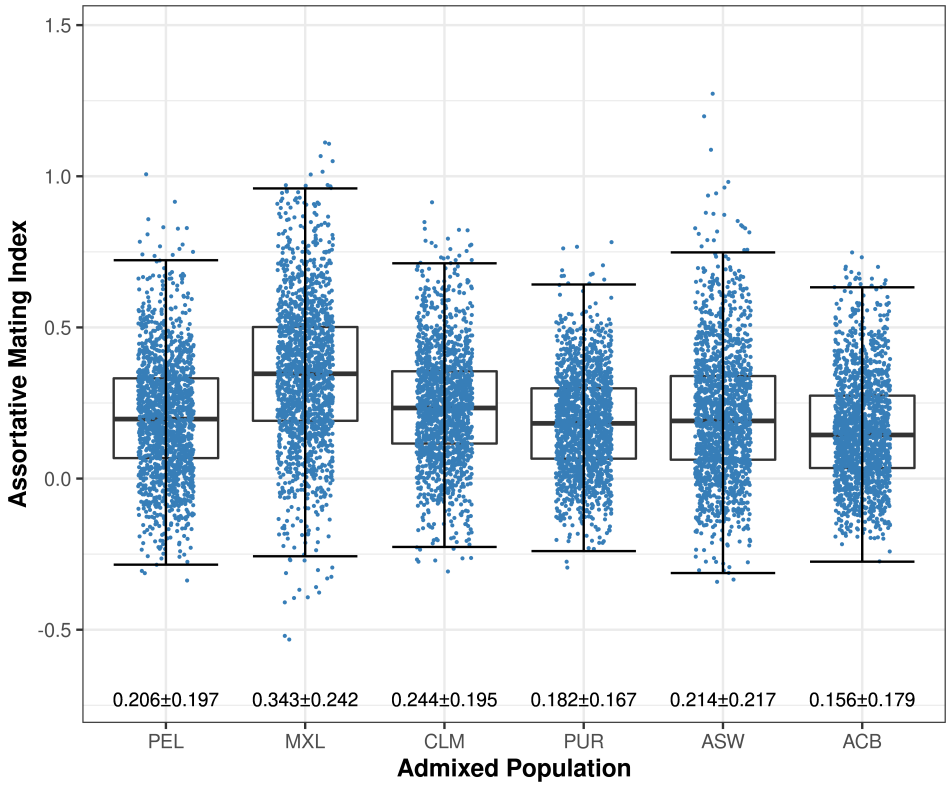

Figure 3: Comparative box plots displaying the distribution of the triallelic assortative mating index calculated for each ancestry-informative marker for each admixed population. The boxes signify upper quartile, median and lower quartile values of the distribtuion, while the whiskers signify the last 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 negligable owing to the sample size of 4,111,226. Horizontal jitter is used to better display the distribution.

ASW and ACB plots, suggest some form of deviation from HWE but whether they are caused by migration, assortative mating or some other phenomenon is unclear. A supplementary approach is finding and plotting homozygous CAT lengths, as in **Fig. 5B**. This exagerates HWE deviations, and provides additional peaks to some of the distributions. These peaks are more informative than just skewness: they show different CAT length distributions of the same ancestry that have been merged, suggesting significant migration is the likely cause of corresponding skewness in the haplotype CAT length visualisation, for example with ASW and ACB. Less intense right skewness, such as with European ancestry in the ASW population, could indicate either minor and sustained European migration, or European assortative mating, where at least some of the European population disproportionately interbred thereby preserving longer CATs than would be expected under HWE. Each homozygous CAT length distribution has a left tail absent in the haplotype CAT lengths, likely an artefact of heterozygous alleles breaking large homozygous CATs which would leave one of the two haplotype CATs intact. A more sophisticated software for CAT length analysis is TRACTS, which uses them to infer how many generations ago migration events took place. Cross-referencing this with relevent slave migration data provides a picture of delays between migration and significant admixture, ie assortative mating (**Fig. 6**). As it was Europeans transporting slaves accross the Atlantic, we know Europeans arrived in the region the same generation 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

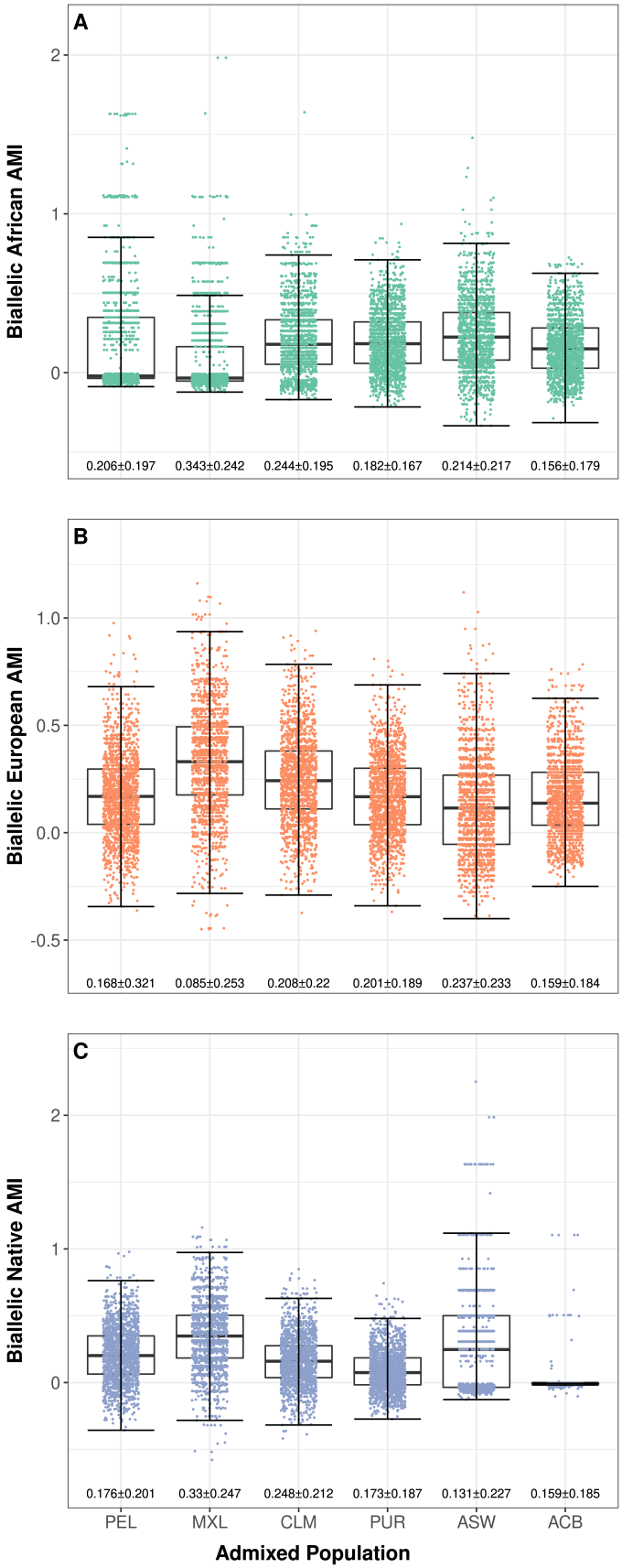

Figure 4: Comparative box plots displaying the distribution of the biallelic ancestry-specific assortative mating indicies calculated for each ancestry-informative marker for each admixed population. The boxes signify upper quartile, median and lower quartile values of the distribtuion, while the whiskers 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 negligable owing to the sample size of 4,111,226. Horizontal jitter is used to better display the distribution.

rest of the populations 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 interpretted 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 that Europeans arrived at Barbados long before evidence suggests.

4 Discussion

!!! !!!!!!!!!!!!!! DISCUSSION NOT COMPLETE YET, THESE ARE JUST NOTES !!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!!! !!

4.1 Data Preparation & Ancestry Proportion

Remember above 99pc PEL individuals are included in ref sample, hence assignment of PEL ancestry will be a bit weird - definitely a bias. I went forward with the population in both populations, but the one with will match native very perfectly, and the one without will A. show as less native than it truly is as a population as v native ones are excluded, and B. might have a native bias in assignment, where there are similarities between two PEL individuals which may not be a general native trait but a specific PEL trait.

Remember, ASW is Americans of Sub-Saharan African Ancestry in Oklahoma, Southwest USA, MXL is Mexican Ancestry in Los Angeles CA United States - so significant sample bias; ASW will have more african than the average US SW resident; MXL will have more European, and possibly african could have come later when in LA vs generations ago in Mexico

More NAT samples would be better; 72 vs 507 Eur and 550 Afr - remove "imbalanced" remark from results and say here instead. More NAtive samples would make the algorithm more likely to assign AIM aleles Native and not the other two - could be responsible for the assigned aprox 5% European ancestry in the mayan population, and lead to more accurate assignments for the query samples. Not available atm tho.

Our assumption when studying social stratification is that population-wide genetic assortative mating in humans is negligable - ie inherently being attracted to a certain hair colour for based purely on instinctual attraction vs social bias, is negligable (source backing this up?). These can't be distinguised as it would require a human population with zero social structure or biases, impossible naturally and unethical artificially.

4.2 Assortative Mating Index

So the data shows significant assortative mating, but is the data/method reliable? Get reference saying HWE is reached in 2 or so generations, which shows the AMI plot does a decent job of deviating. Also try to talk a bit about HWE as quality control being bad.

refer to RFMIX vs ADMIXTURE plot here unless have in results already. saying one overestimates blah blah blah or the other underestimates blah. UPDATE: have mentioned, develop further here (inc if result are reliable)

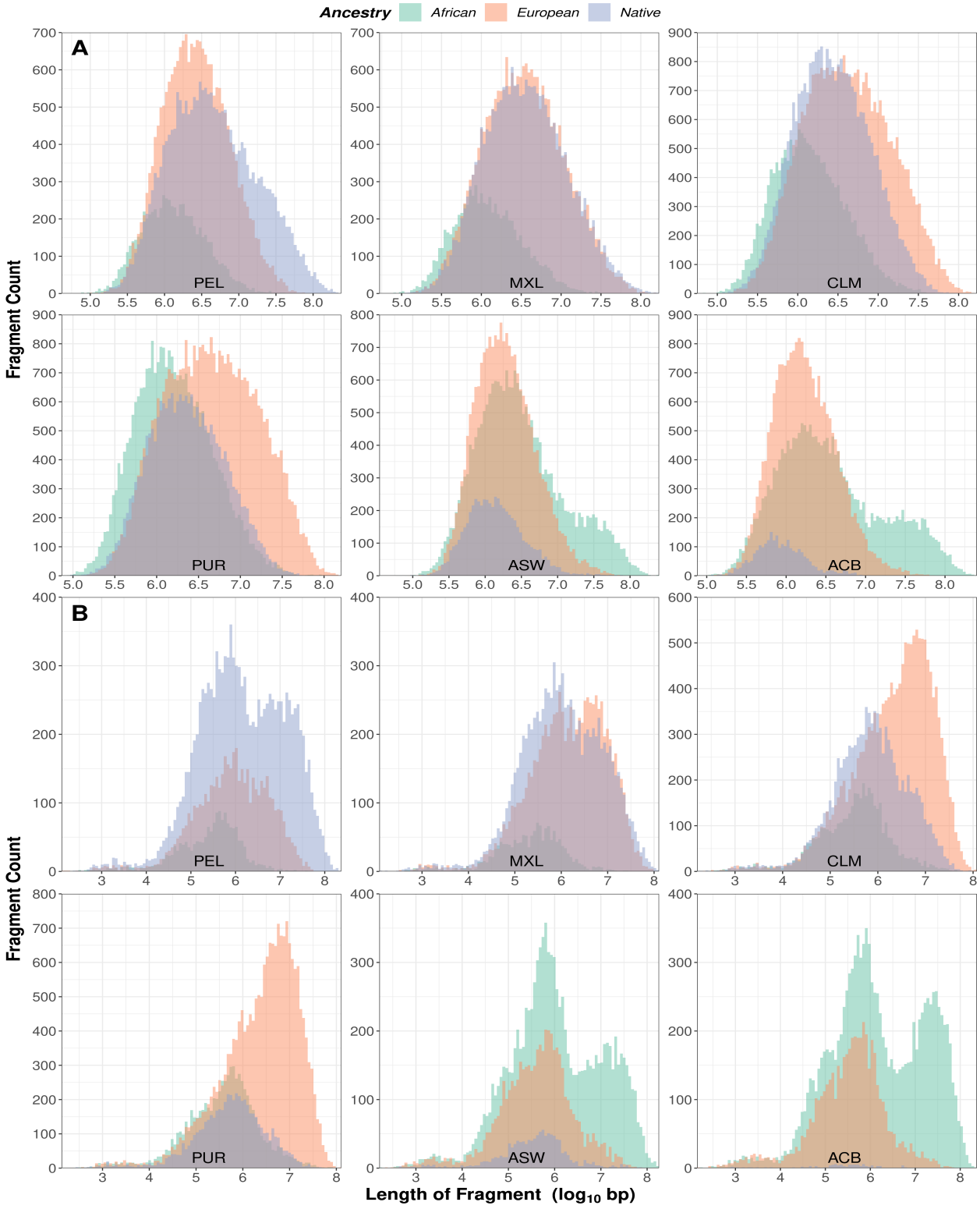

Figure 5: Histograms of continuous ancestry tract lengths of each of the three ancestries for each admixed population. Tract lengths are measured in base pairs in log10 scale, and are seperated into 100 bins in each plot. Tract length is considered either the number of consecutive haplotype asignments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

could do this analysis with sex chromosomes, see if assortative mating varied by gender, as a way to distinguish between voluntary or involuntary admixture, the latter not being incompatible with social stratification.

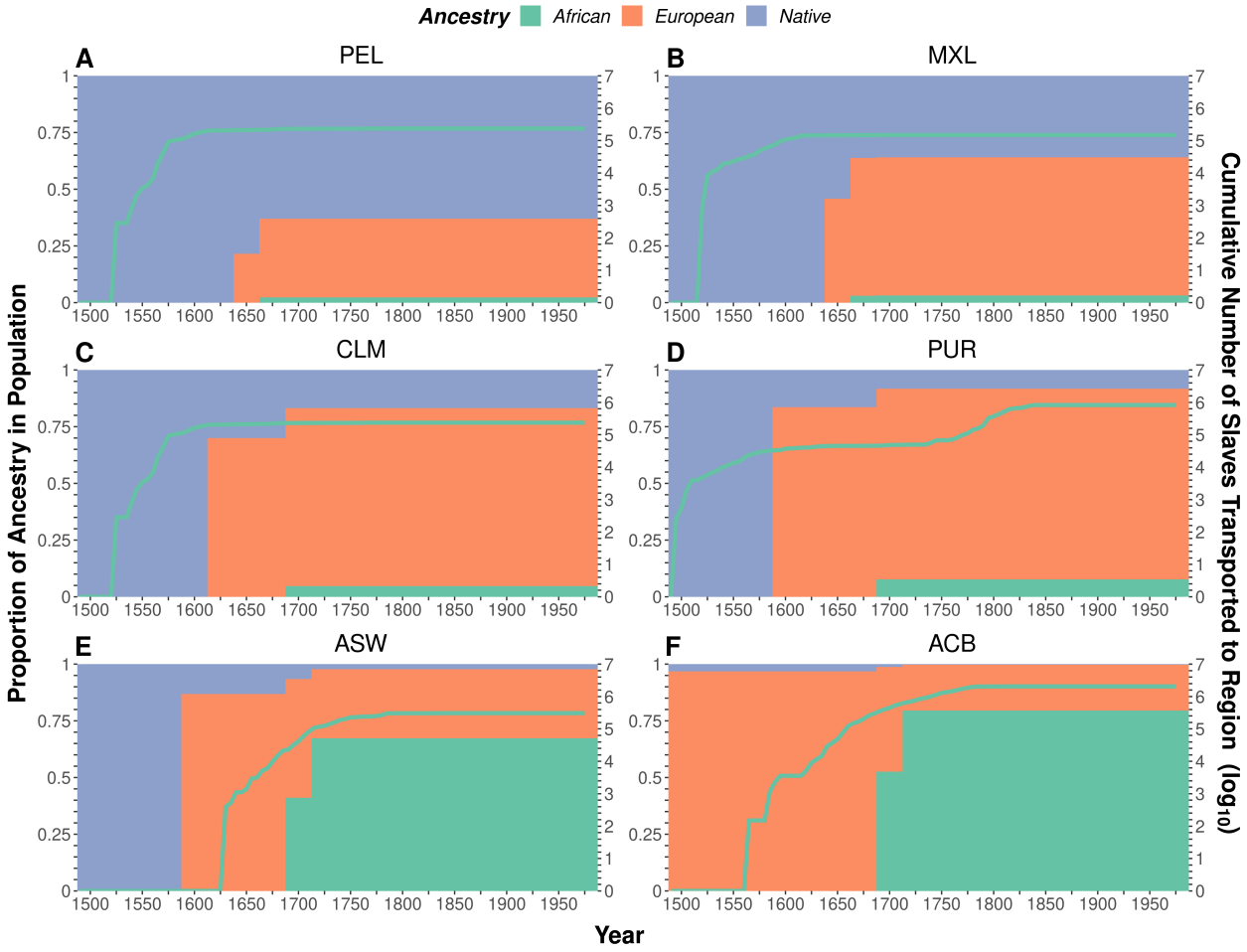

Figure 6: Number of slaves transported to the general regions of the admixed populations every five years, and stacked barplots showing how the proportion of the three ancestries changed in said populations generation to generation as estimated by TRACTS, between the years 1500 and 2000 CE. Data on the 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 Carribean islands for PUR, ports north of the Rio Grande in North America for ASW, and ports on British Carribean islands for ACB. Genomic data of the individuals from each admixed population was analysed with TRACTS to 25 bootstraps, with generations being estimated as 25-year periods.

4.3 Continuous Ancestry Tract Lengths

(basically mentioned in discussion) For frag length histograms, 2nd peaks suggest another migration event, ie new gene flow, as one would expect it to simply be a normal distribution if only one migration event occured. Similarly, right-tailed distributions suggest constant strean of subsequent immigrants after main migration event (and vice versa). test.

Homozygous CAT not only tells us some right-skewness is due to migration (ASW and ACB) but also shows how 2 peaks can mascerade as 1 (PEL and MXL)

Simulations with various migration events and AM parameters to match the distributions?

TRACTS output likely still a function of migration and assortative mating to some extent

Barbados ACB TRACTS - even with the concept of an initital Native population hard-coded into the model, the algorithm could only explain the genomic pattern by predicting that Europeans arrived 50-100 generations ago, 500+ years before it really occured. This will be because ADMIXTURE estimated that only 2 out of the 96 samples contain any Native DNA, both less that 10% - a stark reminder assortative mating and migration weren't the only population-shaping phenomena

at play.

TRACTS does not give lots of pules, only 1 per that it deems to have had the biggest effect per ancestry, model simply doesnt allow more. Ideally this would be expanded

TRACTS results make more sence (ie vast majority of slaves transported to region a few generations before African admixture) in oppulations where we could be more geographically specific - PEL, CLM and MXL. In the others, the majority of the slaves were transported after admixture, roughly by an order of magnitude in all three cases. More thoroughly researching the histories of these regions and more accurately determining the ports at which slaves that ended up in Barbados, Puerto Rico and the US Southwest initially disembarked from their voyage should help. Also generation lengths may well be significantly different in the different populations, and indeed over time. In fact in slave-based societies in the southern US and carribean, slaves breeding was a cheaper method of procuring additional slaves, hence one might expect African subpopulations to have shorter generation lengths.

in discussion, TRACTS does not tell us much about and subsequent assortative mating, and migration data for Europeans would be useful rather than using slave migration as a proxy. Also, it's ancestry proportion, not absolute quantity of genetic material - Native populations shrinking due to disease and colonisation would have the same effect of increasing European ancestry proportion in the population as European migration.

4.4 Concluding Remarks

End of the day, migration and the ending of AM, both the removal of barriers to admixture, manifest themselves identically - hence the two are inextricable absent accurate migration data which can be used to explain the contribution of migration to admixture, leaving information as to the impact of assortative mating.

But can machine learning bypass this need? Talk to alex and/or matteo about this

As for tracking present and predicting future anectry-based social stratification, much higher quality migration data is available so should be less of an issue compared with learning from history of migration and AM in the Americas.

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

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

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

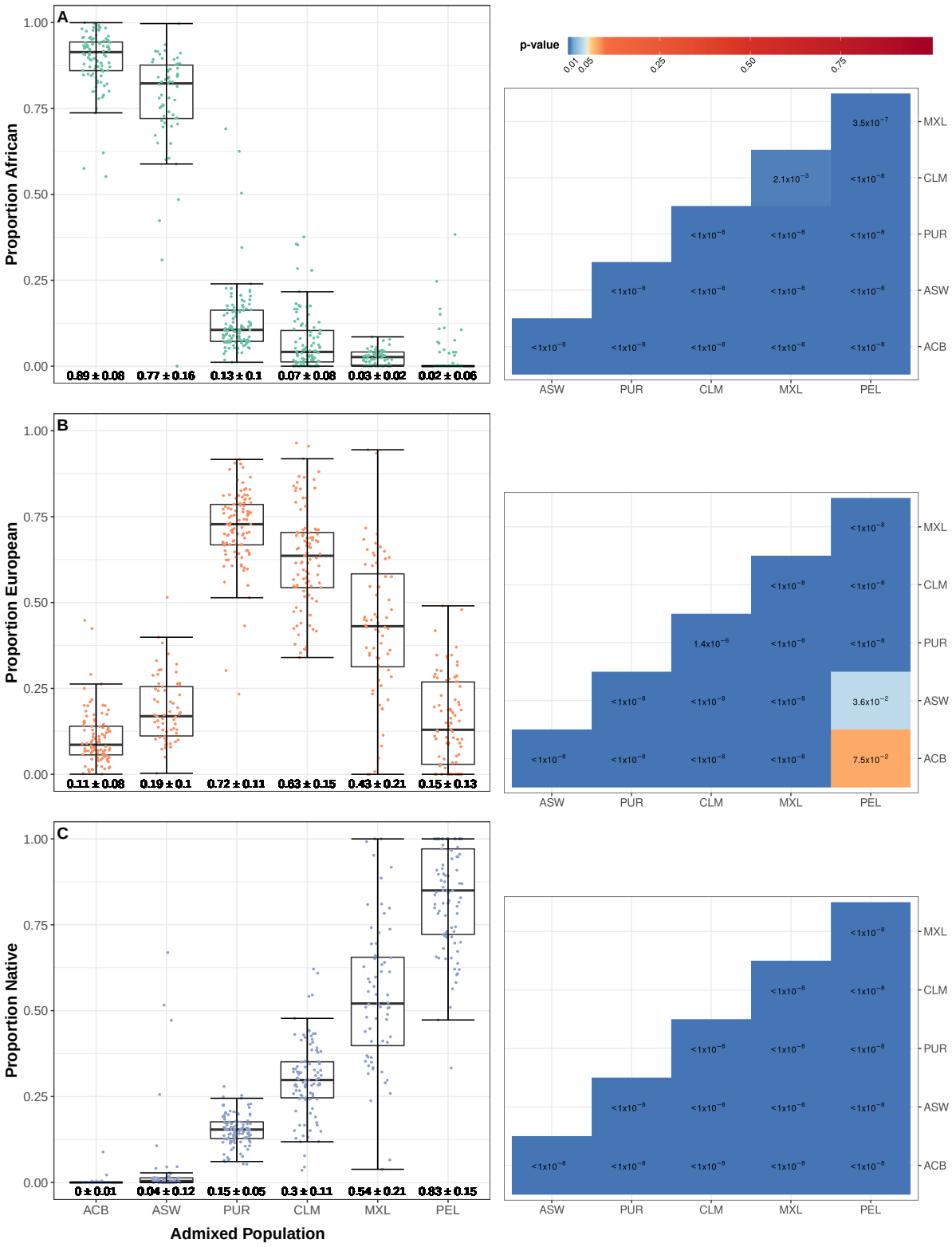

Figure S1: Comparative box plots displaying the distributions of the three ancestry proportions for each individual of each admixed population, with corresponding p-value heatmaps comparing populations statistically. The boxes signify upper quartile, median and lower quartile values of the distribtuion, 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.

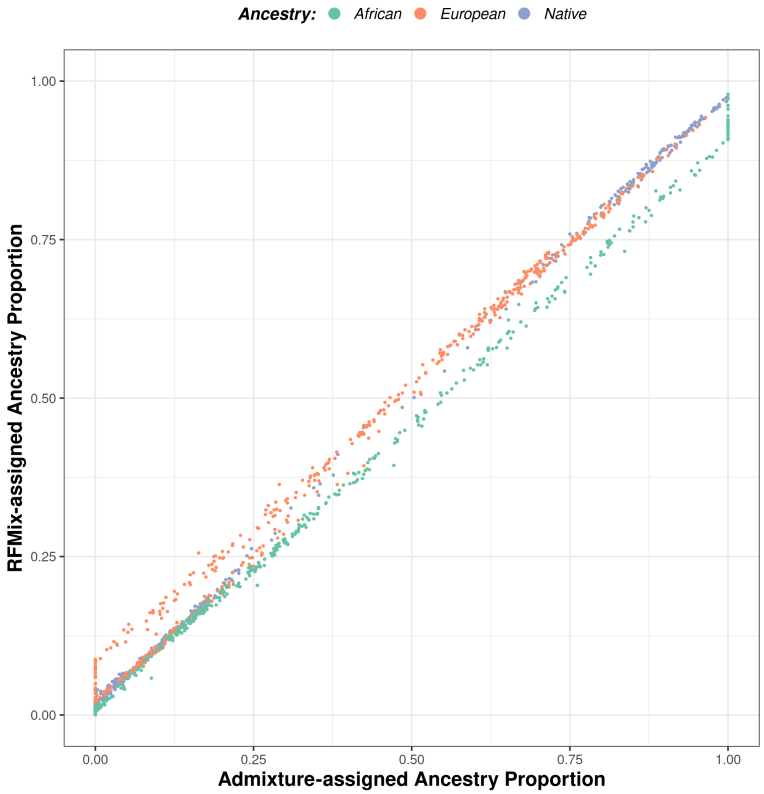

Figure S2: Scatterplot correlating ancestry proportions assigned by RFMIX for all 1690 query and reference individuals against those assigned by ADMIXTURE.

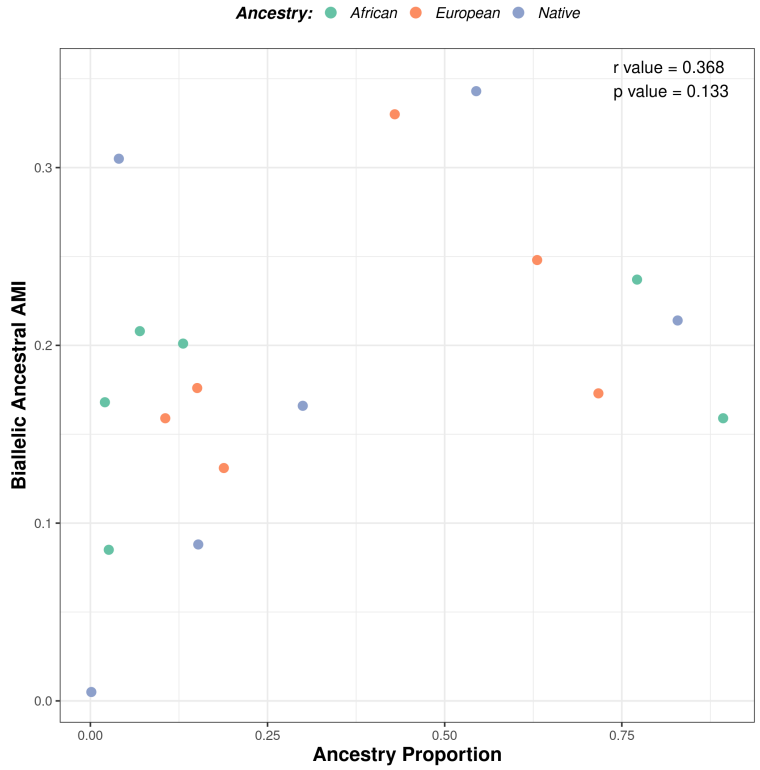

Figure S3: Scatterplot charting all three mean biallelic ancestry-specific AMI against all three ancestry proportion for each of the six admixed populations.

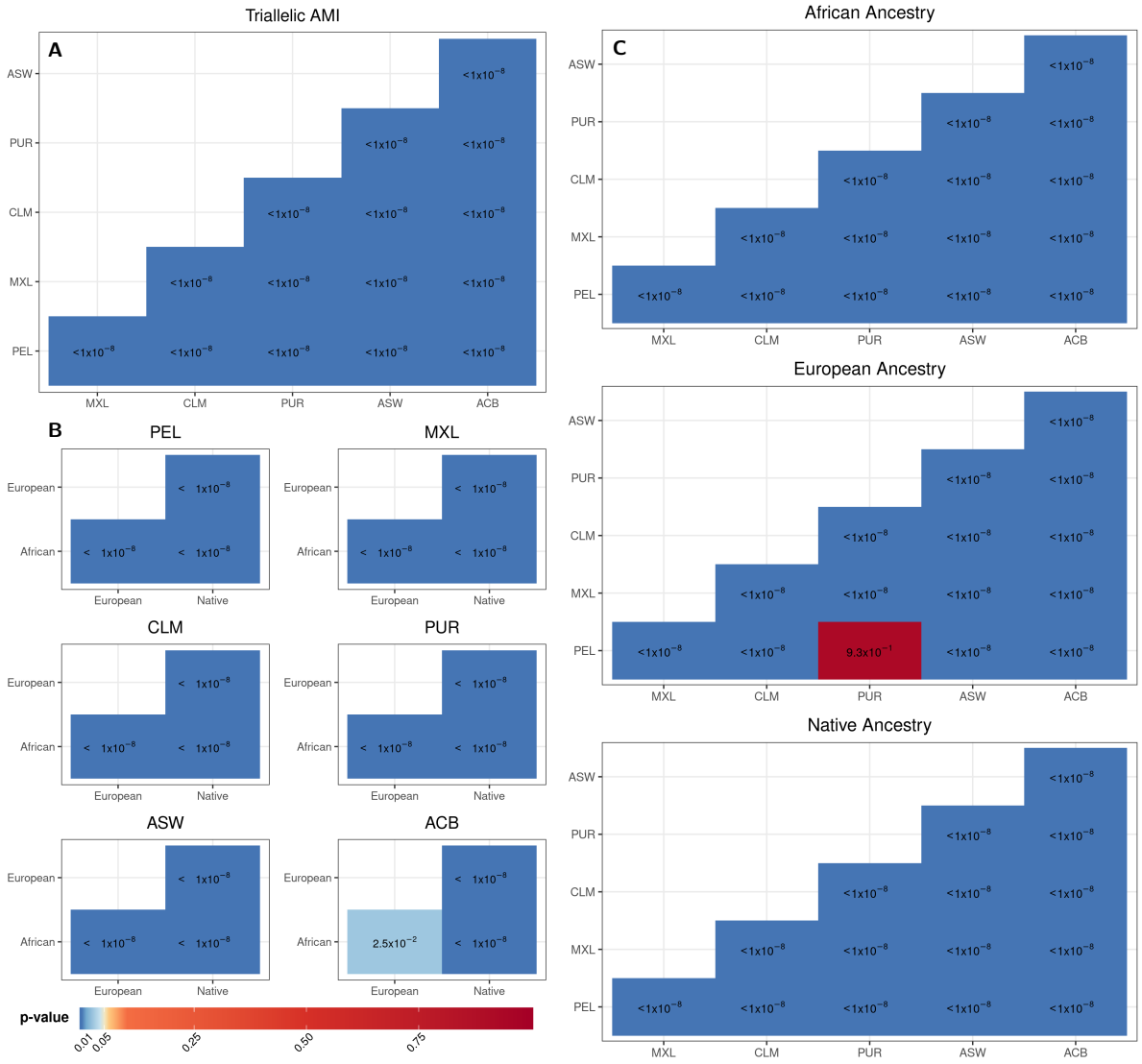

Figure S4: Heatmaps displaying p-value results of Wilcoxon tests used to compare assortative mating index values of different populations and ancestries. Each set of heatmaps correspond to a different set of comparisons between all combinations of assortative mating index (AMI) distributions. **A** compares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

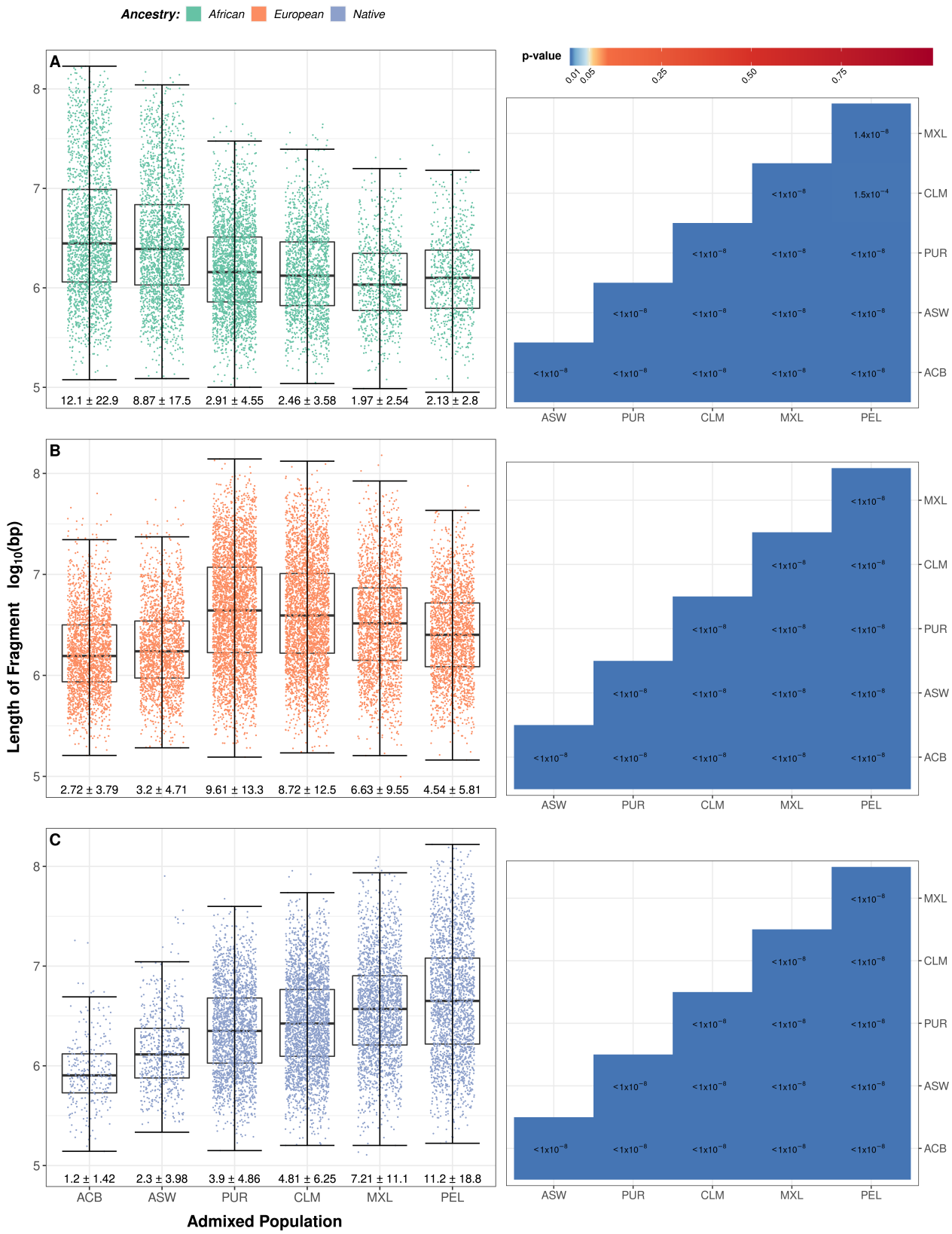

Figure S5: Comparative box plots displaying the distributions of continuous ancestry tract lengths of each ancestry for all individuals of each admixed population, with corresponding p-value heatmaps comparing populations statistically. Tract length, that is the number of consecutive haplotype asignments of a given ancestry on a single strand, are measured in base pairs in log10 scale. The boxes signify upper quartile, median and lower quartile values of the distribtuion, 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.

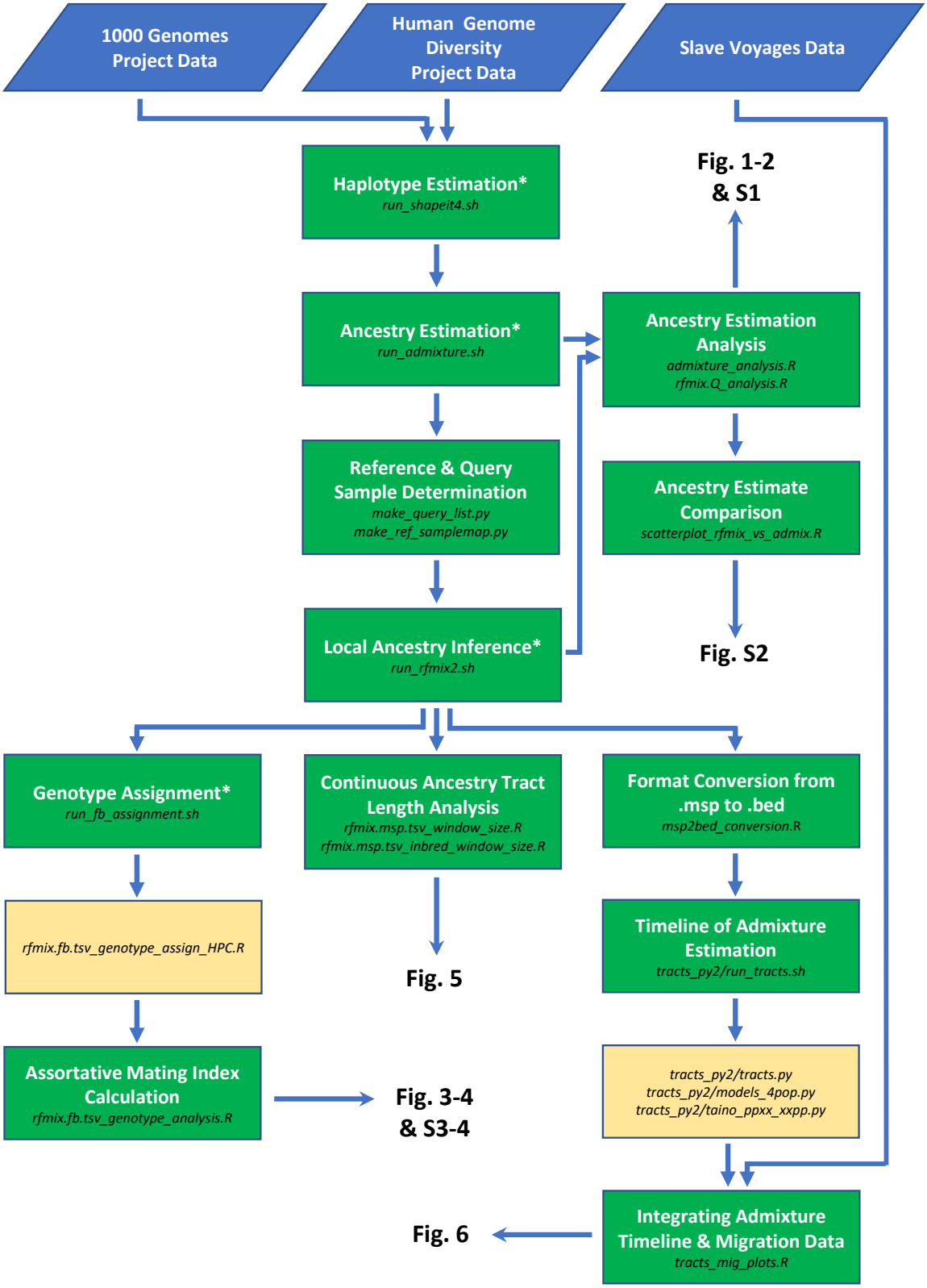

Figure S6: Flowchart representing the analysis workflow of the project, from input data to the output figures. Arrows indicate that the output from one step is the input for the next. Below the label of each step is the script(s) from the provided github repository required to run that step. The scripts 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.