

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

Ben Nouhan, bjn20@ic.ac.uk

Imperial College London

August 24, 2021

Word Count: 5292

Many human populations have become socially stratified based on perceived race throughout history. Assortative mating, the tendency to mate with those similar to you, is a key reason for this. When we map out the ancestral origin of different regions of a genome, the effects of 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 is difficult since past migration events manifest similarly in the genomes of a population. Here I show how modern-day assortative mating can be detected in and compared between admixed North and South American populations. 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 unveil the historical context for modern-day ancestry-related social stratification present in populations throughout the world, and perhaps track it in real time.

Contents

1	Introduction	3
2	Methods	5
2.1	Studied Populations	5
2.2	Data Preparation with BCFtools	5
2.3	Haplotype Estimation with SHAPEIT4	6
2.4	Ancestry Estimation with PLINK & ADMIXTURE	6
2.5	Local Ancestry Inference with RFMIX v2	6
2.6	Assortative Mating Index Calculation	6
2.7	Continuous Ancestry Tract Length Analysis	7
2.8	Timeline of Admixture Estimation with TRACTS	8
3	Results	8
3.1	Ancestry Proportion	8
3.2	Assortative Mating Index	9
3.3	Continuous Ancestry Tract Lengths	11
4	Discussion	13
5	Data and Code Availability	18
5.1	Data	18
5.2	Code	18
	References	19
	Supplementary Material	21

¹⁷ **1 Introduction**

¹⁸ Positive assortative mating, a phenomenon wherein individuals are more likely to mate with those
¹⁹ phenotypically similar to themselves, is widely accepted to occur in human populations (Norris et
²⁰ al., 2019). This has the potential to alter population structure by introducing social stratification
²¹ and, in turn, create social constructs upon which further assortative mating can be based, such as
²² wealth, class or social policies (Risch et al., 2009).

²³ From a genetics standpoint, this multigenerational non-random admixture between genetically
²⁴ distinct groups leaves a genomic imprint in the individuals comprising the population, in stark
²⁵ contrast to populations more closely following Hardy-Weinberg equilibrium (HWE) (Zaitlen et
²⁶ al., 2017). However, the genomic imprint on the population structure left by either sociocultural
²⁷ barriers or geographical barriers limiting admixture is difficult to discern. Afterall, large scale
²⁸ migration of a population will genetically manifest itself similarly to the change of societal rules or
²⁹ norms that condition social interaction, such as the revocation of racial segregation policies.

³⁰ Single nucleotide polymorphisms (SNPs) can be used as indicators of ancestry (Risch et al.,
³¹ 2009). Population genomics techniques allow us to generate a large array of SNPs which can be
³² analysed using local ancestry inference to map ancestries to positions and regions along the genome,
³³ after which further analysis can indicate past assortative mating in a population (Schubert et al.,
³⁴ 2020).

³⁵ One such analysis is that of continuous ancestry tract lengths: the lengths of genomic regions
³⁶ consecutively assigned to the same ancestry. Looking at the distribution of these lengths, the
³⁷ ancestry to which they belong and the overall ancestry proportion of individuals within a population
³⁸ can indicate how long ago the admixture occurred and to what extent. Recombination of the DNA
³⁹ of admixing individuals leads to a decrease in continuous ancestry tract lengths, as those within the
⁴⁰ parents' genomes interrupt one another upon recombination. Hence, admixture more generations
⁴¹ ago will manifest as distributions of shorter continuous ancestry tracts and vice versa (Gravel,
⁴² 2012).

⁴³ Genotype frequency is another indicator of population admixture; one would expect a more
⁴⁴ admixed population to have higher heterozygous genotype frequencies at a given position. While
⁴⁵ this alone does not inherently indicate assortative mating, the extent to which the observed geno-
⁴⁶ type frequency deviates from what would be expected under HWE can also be considered. The
⁴⁷ assortative mating index (AMI) quantifies the relative local ancestry homozygosity:heterozygosity
⁴⁸ ratio at a given position based on this concept, which can be used as a proxy for the extent of
⁴⁹ assortative mating at said position (Norris et al., 2019).

⁵⁰ HWE is commonly used in population genomics as a quality check for genetic markers in genome-
⁵¹ wide association study (GWAS) - SNPs chosen for being particularly informative for pathological
⁵² research - whereby alleles with frequencies deviating too far from it are removed and considered
⁵³ sequencing misreads (Linares-Pineda et al., 2012). This does not take into account stratification,
⁵⁴ present in most if not all societies, within the populations studied herein. Showing allelic deviation
⁵⁵ from HWE is not an artefact but rather an intrinsic quality may serve as a warning against this
⁵⁶ practice.

⁵⁷ Populations of the Americas such as Colombia, Barbados, Mexico or the US provide appropriate
⁵⁸ and well-researched case studies integrating migration, admixture and assortative mating. Many of

59 such populations have different but connected histories: a Native American population is colonised
60 by Europeans; the Native population shrinks due to war, hard labour and disease, while the Eu-
61 ropean population grows via migration. These phenomena continue such that African slaves are
62 transported to the region as a source of additional labour; after which the population continues
63 evolving with these scars of colonialism (Bryc et al., 2010; e Silva et al., 2020; Mas-Sandoval et al.,
64 2019).

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

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

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

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

97 Only by reconciling migration and assortative mating can we confidently infer assortative mating
98 from genomic data, and use this to draw conclusions about past and make predictions about future
99 ancestry-related social stratification.

¹⁰⁰ **2 Methods**

¹⁰¹ **2.1 Studied Populations**

¹⁰² For the initial analyses, all African, European and American populations from the 1000 Genomes
¹⁰³ Project (1KGP) and the Human Genome Diversity Project (HGDP) were used **Table 1**, with the
¹⁰⁴ exception of the Russian and Finnish populations. These were excluded owing to minimal colonial-
¹⁰⁵ era migration to the Americas from these populations, alongside the genetic similarities between
¹⁰⁶ 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
	Basque in France	Basque	23
European	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
	Toscane in Italy	TSI/Tuscan*	115
	Colombian in Colombia	Colombian	7
	Karitiana in Brazil	Karitiana	12
Native American	Maya in Mexico	Maya	21
	Pima in Mexico	Pima	13
	Surui in Brazil	Surui	8

¹⁰⁷ **2.2 Data Preparation with BCFtools**

¹⁰⁸ Using BCFtools v1.9, the 30x coverage 1KGP and high-coverage HGDP datasets were merged, and
¹⁰⁹ all populations except those listed in (**Table 1**) were removed. All C→G, G→C, A→T and T→A
¹¹⁰ SNPs were filtered out as they are harder to assign and are hence prone to error (Danecek et al.,
¹¹¹ 2021). SNPs were further filtered with a minor allele frequency threshold of 5%, as to reduce the
¹¹² dataset and remove rare and thus uninformative SNPs. Following this, all 22 filtered VCF files,

113 one per autosome, were indexed for phasing.

114 2.3 Haplotype Estimation with SHAPEIT4

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

122 2.4 Ancestry Estimation with PLINK & ADMIXTURE

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

127 The programme ADMIXTURE v1.3.0 used cluster analysis and principal component analysis
128 to estimate the proportions of African, European and Native American ancestry for each remaining
129 sample, with default parameters and three ancestries to be detected (Alexander et al., 2009).

130 2.5 Local Ancestry Inference with RFMIX v2

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

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

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

146 2.6 Assortative Mating Index Calculation

147 One measure of assortative mating is the assortative mating index (AMI), which takes a log odds
148 ratio of the relative local ancestry homozygosity and heterozygosity:

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

149

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

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

$$(a + e + n)^2 = a^2 + e^2 + n^2 + 2ae + 2an + 2en \quad (3)$$

153

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

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

171 2.7 Continuous Ancestry Tract Length Analysis

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

177 To generate the fragment length distributions, the script "rfmix.msp.tsv_window_size.R" reads
 178 the .msp.tsv files, sums the length of consecutive genomic windows assigned to the same ancestry,
 179 and appends the lengths to the vector containing the lengths of other fragments corresponding to
 180 the fragment's ancestry and population. The script "rfmix.msp.tsv_inbred_window_size.R" works
 181 similarly, but generates fragment length distributions of consecutive homozygous genotype assign-

182 ments, rather than haplotype assignments.

183 2.8 Timeline of Admixture Estimation with TRACTS

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

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

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

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

210 3 Results

211 3.1 Ancestry Proportion

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

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

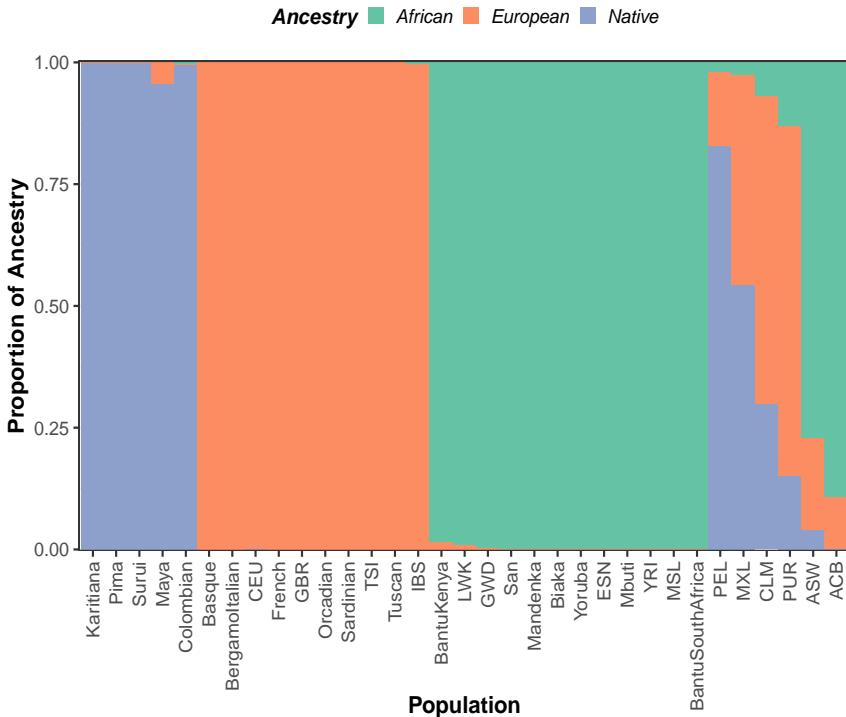


Figure 1: Stacked barplots showing the proportions of the three ancestries of each reference or query population used throughout the study, generated by ADMIXTURE. Genomic data from individuals of selected populations from the 1000 Genomes Project and the Human Genome Diversity Project were processed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

220 The distribution of ancestry proportions on the individual level within these six admixed pop-
 221ulations is shown in **Fig. 2**, which largely corresponds with **Fig. 1** while suggesting approximately
 222 25% of LA Mexicans and Colombians have no African ancestry, fewer than 5% and 50% of Barba-
 223 dians and ASWs respectively have Native ancestry, and that only around 20% of Peruvians have
 224 African ancestry while around 25% of them are of exclusively Native ancestry.

225 These individual-level ancestry proportion distributions are further visualised in **Fig. S1**, with
 226 the distribution for each population of a given ancestry displayed side-by-side in box plots. All dis-
 227 tributions were different at the 5% significance level, except for Barbadian and Peruvian European
 228 ancestry proportion distributions. However, their Native and African ancestry distributions con-
 229 trast starkly, lending credence to the assumption all six admixed populations have entirely different
 230 ethnological structures.

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

235 3.2 Assortative Mating Index

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

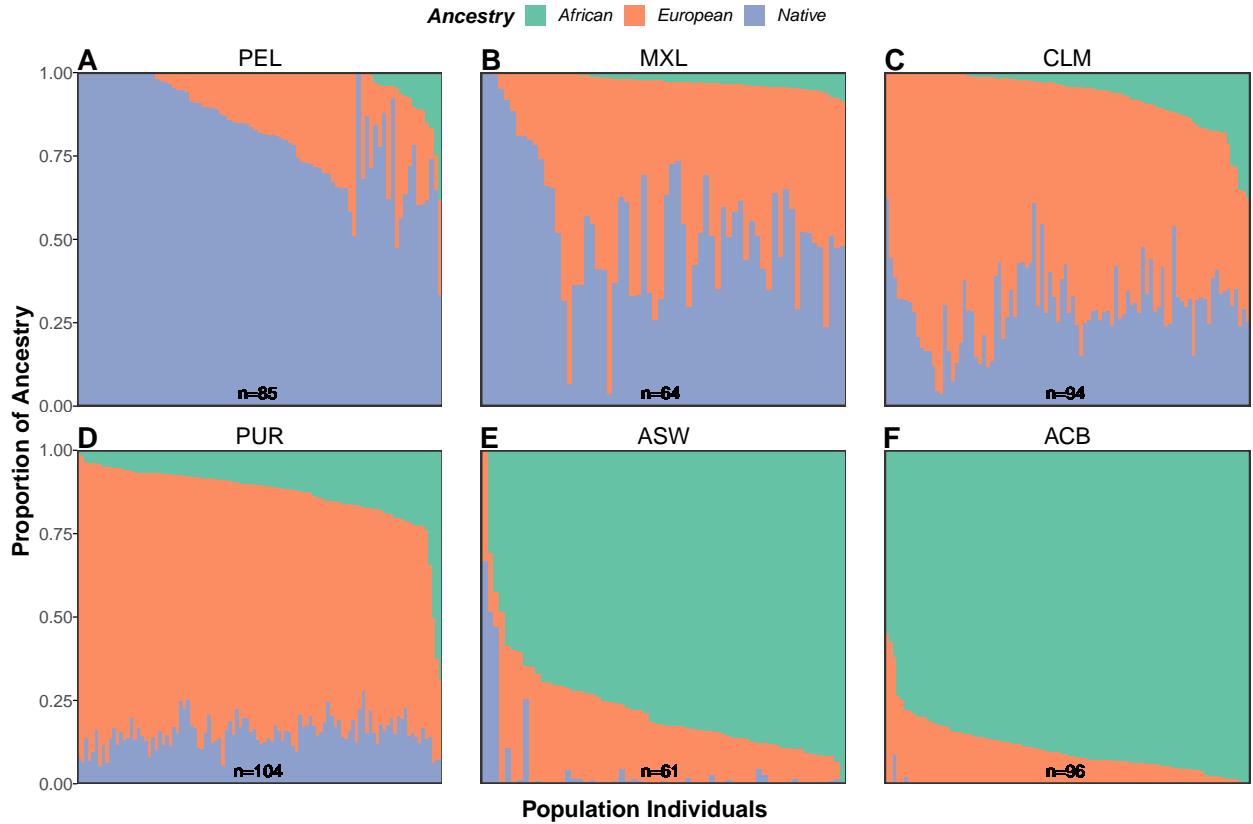


Figure 2: Stacked barplots showing the proportions of the three ancestries of each individual comprising the six query admixed populations, generated by ADMIXTURE. Genomic data from individuals of the six from the 1000 Genomes Project and the Human Genome Diversity Project were processed and subjected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

240 to have lower European Ancestries.

241 The main RFMIX output was used to calculate assortative mating index values for each SNP
 242 in each population. The triallelic AMI values for each position and population are plotted in **Fig.**
 243 **3**. In a population without assortative mating, we would expect the mean AMI value to be zero.
 244 With a sample size of 4,111,226 SNPs, and the standard deviations being of similar sizes to the
 245 corresponding means, the standard errors of the means are negligible and hence the sample means
 246 are accurate estimates of the true means. Based on this, we can see all means are significantly
 247 higher than zero, indicating positive assortative mating in all admixed populations.

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

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

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

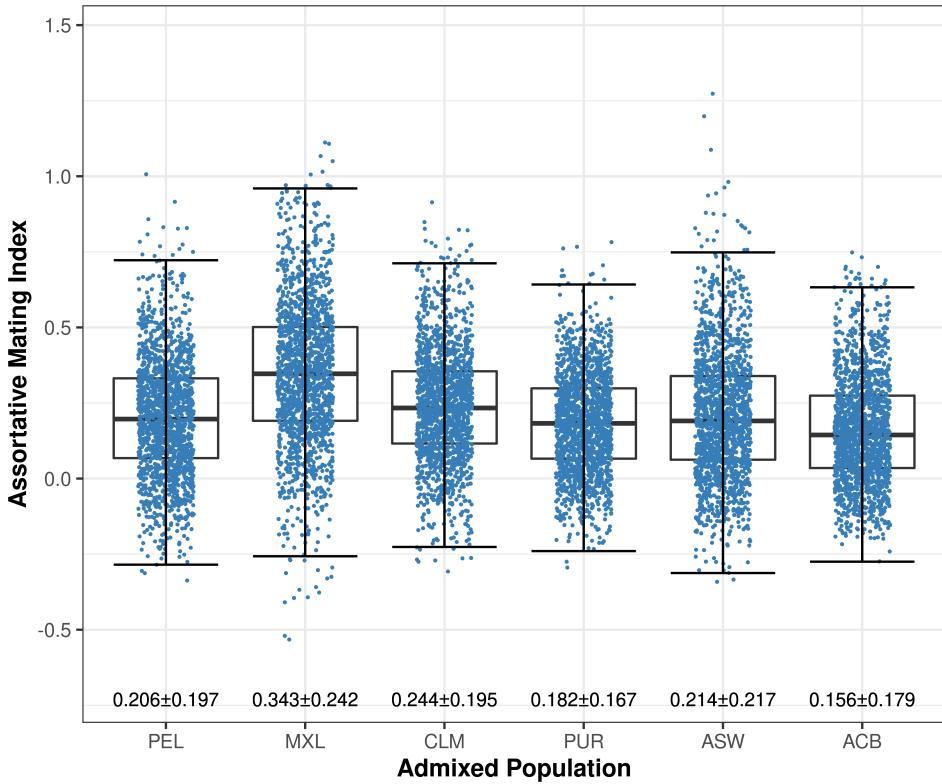


Figure 3: Comparative box plots displaying the distribution of the triallelic assortative mating index calculated for each studied single nucleotide polymorphism for each admixed population. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers signify the last data point within the closest quartile value plus 150% of the interquartile range. Mean \pm 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 to better display the distribution.

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

262 3.3 Continuous Ancestry Tract Lengths

263 The final use of the RFMIX output was analysing the lengths of continuous ancestry tracts. Dis-
 264 playing the haplotype continuous ancestry tracts in a histogram allows visual comparison between
 265 the tract length distributions of the different ancestries (**Fig. 5A**), while box plots better visualise
 266 descriptive statistics of the data (**Fig. S5**).

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

272 A supplementary approach is finding and plotting homozygous continuous ancestry tract
 273 lengths, as in **Fig. 5B**. This exaggerates HWE deviations, and provides additional peaks to some of
 274 the distributions. These peaks are more informative than just skewness: they show different tract
 275 length distributions of the same ancestry that have been merged, suggesting significant migration
 276 is the likely cause of corresponding skewness in the haplotype continuous ancestry tract length
 277 visualisation, for example with ASW and ACB, effectively causing two populations of the same

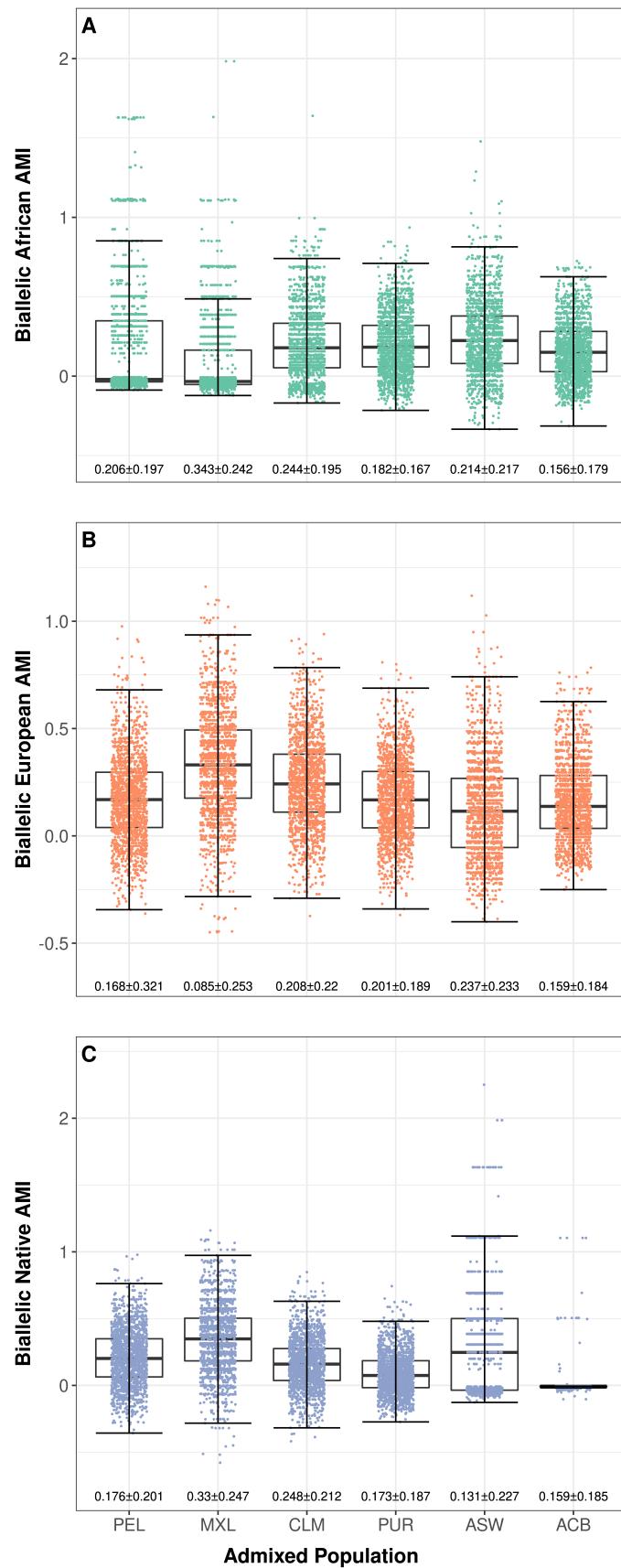


Figure 4: Comparative box plots displaying the distribution of the biallelic ancestry-specific assortative mating indices calculated for each studied single nucleotide polymorphism for each admixed population. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers signify the last data point within the closest quartile value plus 150% of the interquartile range. Mean \pm 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 to better display the distribution.

278 ancestry to merge into one.

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

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

286 A more sophisticated software for continuous ancestry tract length analysis is TRACTS, which
287 uses them to infer how many generations ago migration events took place. Cross-referencing this
288 with relevant slave migration data provides a picture of delays between migration and significant
289 admixture, ie assortative mating (**Fig. 6**). As it was Europeans transporting slaves across the
290 Atlantic, we know Europeans arrived in the region the same generation Africans began to arrive or
291 earlier.

292 Therefore, for example in Peru, we can see that Europeans and Africans began arriving around
293 1525, the majority of Africans had arrived by 1575, significant admixture between Europeans and
294 Natives occurred around 1650, and significant admixture between Africans and the rest of the
295 population occurred around 1675. This suggests extreme assortative mating for 4-6 generations in
296 Europeans and a similar length, albeit lagging by a generation, in Africans.

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

301 4 Discussion

302 Before drawing conclusions from analyses the validity of the underlying data must be questioned.
303 ASW the code for Americans of Sub-Saharan African Ancestry in Oklahoma, Southwest USA, while
304 MXL is Mexican Ancestry in Los Angeles CA United States, which have significant sample biases.
305 The ASW samples will likely have more African than the average US Southwest resident, and MXL
306 samples more European and possibly African ancestry than the average Mexican.

307 The RFMIX reference panel was heavily imbalanced, with 72 Native samples to 507 European
308 and 550 African. This few Native samples will make the algorithm more likely to assign SNP
309 alleles to European or African despite being more indicative of Native ancestry. This could be
310 responsible for the assignment of approximately 5% European ancestry in the Mayan population
311 (**Fig. 1**), and can be resolved by sequencing more Native American genomes. Also, based on **Fig. S2**,
312 ADMIXTURE seems to estimate 100% African or 0% European more readily than RFMIX,
313 suggesting it may be less sensitive at those two extremes. Hence RFMIX ancestry proportion
314 estimates might have been the better choice to go forward with, but an instrumental systematic
315 error like this is unlikely to significantly impact subsequent analysis.

316 One aim of this project was to test the hypothesis that the levels of assortative mating in the

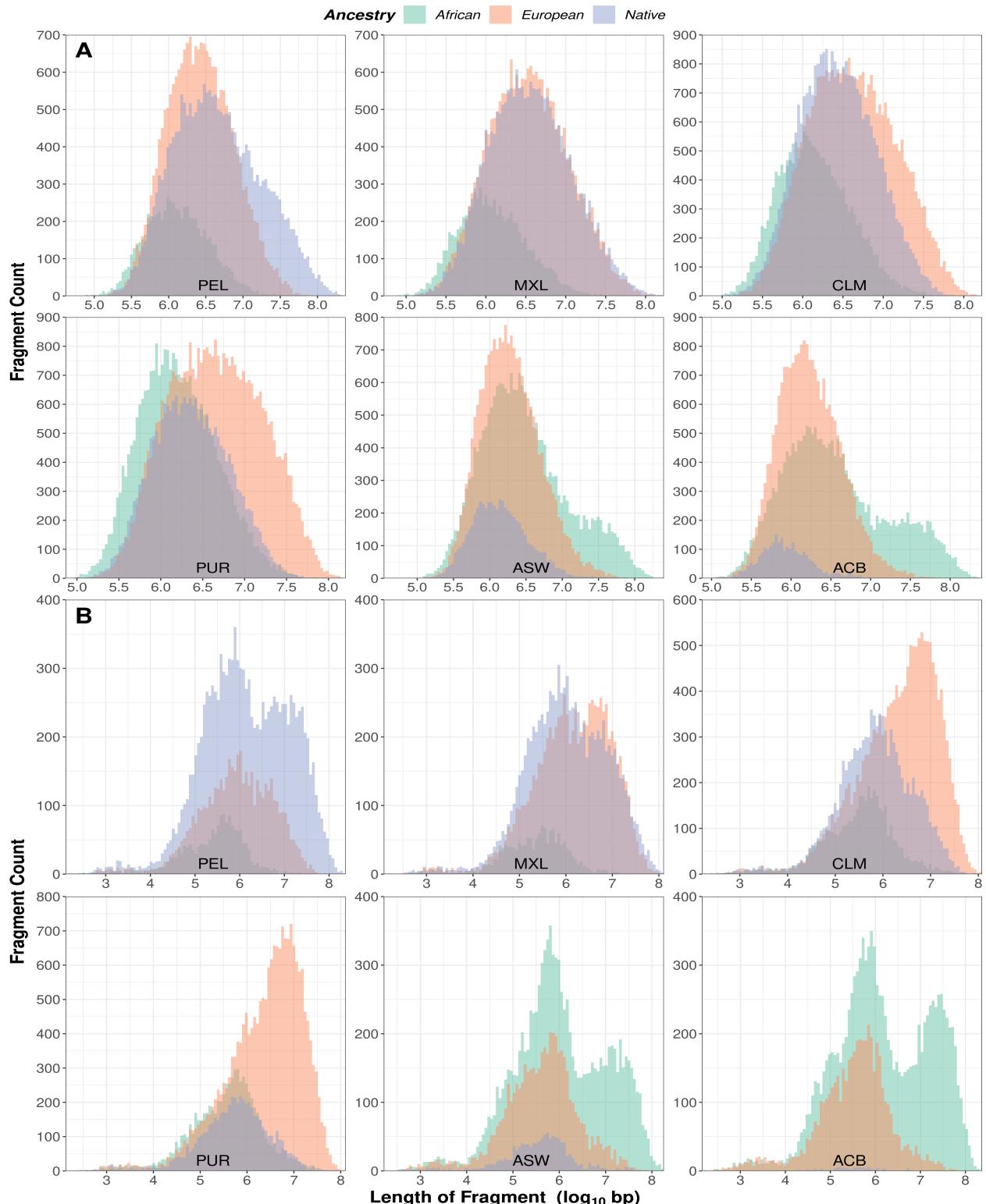


Figure 5: Histograms of continuous ancestry tract lengths of each of the three ancestries for each admixed population. Fragment lengths are measured in base pairs in \log_{10} scale, and are separated into 100 bins in each plot. Fragment length is considered either the number of consecutive haplotype 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 (B).

317 studied populations were significant, and significantly different to each other. This hypothesis was
 318 supported in full by the results of the AMI analysis, both with triallelic and ancestry-specific AMI
 319 (**Fig. 3-4**). The other aim was to assess the methods used throughout, including this AMI analysis,
 320 by their ability to differentiate assortative mating from migration, and thus whether the results are

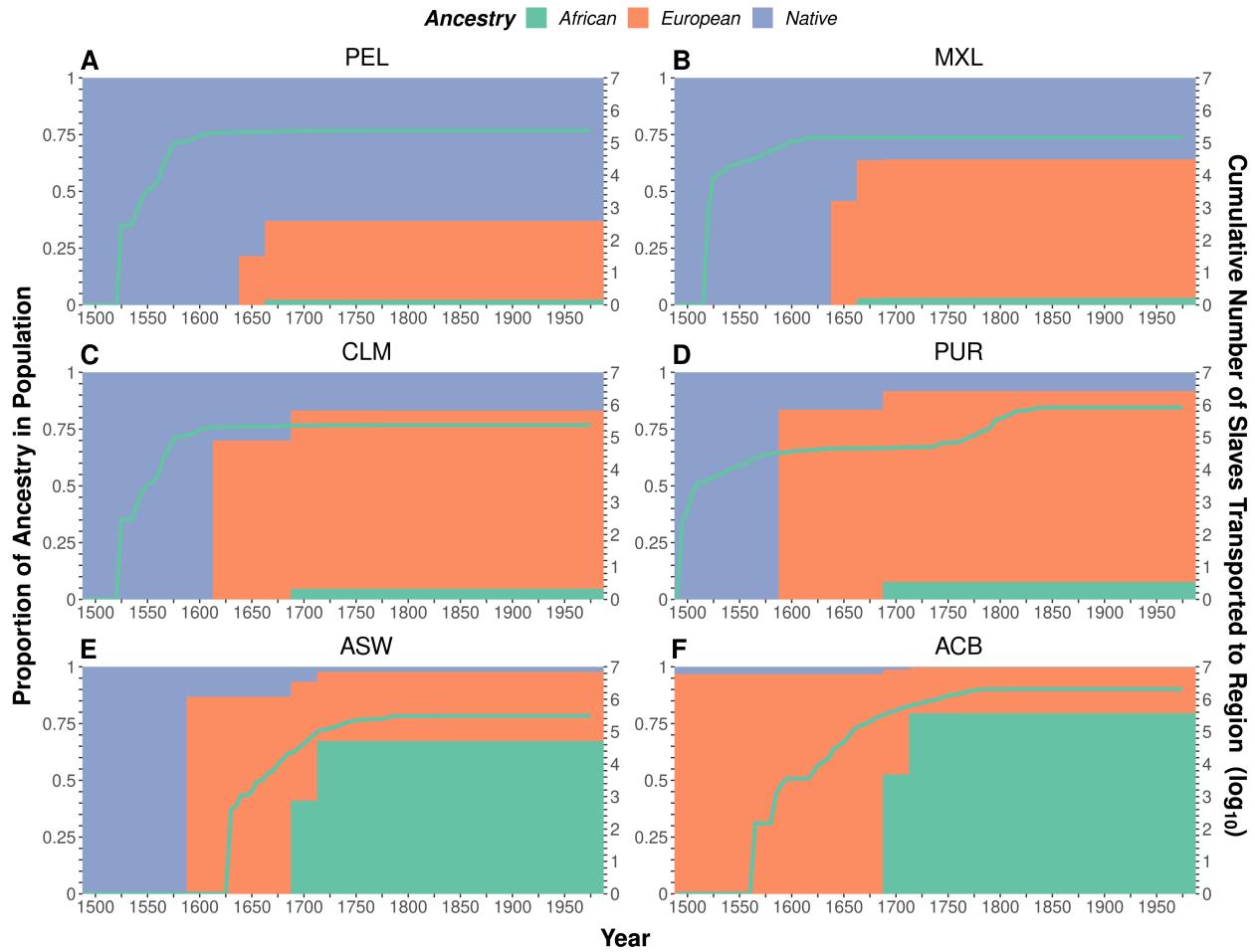


Figure 6: Number of slaves transported to the general regions of the admixed populations every five years, and stacked barplots showing how the proportion of the three ancestries changed in said populations generation to generation as estimated by TRACTS, between the years 1500 and 2000 CE. Data on the 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 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 to 25 bootstraps, with generations being estimated as 25-year periods.

321 valid - a far more nuanced task.

322 In theory, Hardy-Weinberg equilibrium is established in a population following a single generation of fully random admixture (Smithjohn et al., 2015). This means two generations after 323 large-scale migration, a population without ancestry-related social stratification should exhibit 324 insignificant levels of assortative mating. Given all query populations had significant levels, if we 325 assume none of the samples are from first-generation immigrants then we can conclude that not only 326 are the AMI analysis results valid, but that the method successfully distinguishes between assortative 327 mating and migration. It also highlights HWE as inappropriate as a concept for quality-checking 328 genetic markers in genome-wide association studies, for which it is still widely used (Linares-Pineda 329 et al., 2012), owing to the non-random admixture in human populations demonstrated herein con- 330 tradicting the HWE assumptions (Smithjohn et al., 2015).

332 However, the AMI analysis method only allows us to reach that conclusion for the present-day 333 populations, it tells us little about their past. Continuous ancestry tract analysis is better-suited 334 for this, but the results are less conclusive. The histogram visualisation of the homozygous tract 335 analysis (**Fig. 5B**) not only indicates that the right skewness in the **Fig. 5A** African ASW and

336 ACB plots and Native PEL plot are due to migration rather than assortative mating, but also
337 shows in the Native MXL distribution that two peaks can mascerade as one. This method does
338 not quantify the length of time passed since admixture, and hence can only be used to highlight
339 major same-ancestry migration events in the past for further investigation. It fails to differentiate
340 minor and prolonged migration from assortative mating.

341 The method integrating TRACTS with slave voyage data does inform us about the past, by
342 quantifying time since admixture. The **Fig. 6** plots for the Peruvian, Mexican and Colombian
343 populations are intuitive and historically plausible, although more detailed research into whether
344 documentation of the period corroborates the projections should be conducted. Perhaps relatedly,
345 these are the three query populations for which the slave voyage used was most geographically
346 accurate.

347 In the others, the majority of the slaves were allegedly transported after the generation of
348 most significant admixture, roughly by an order of magnitude in all three cases. More thoroughly
349 researching the histories of these three regions and more accurately determining the ports at which
350 slaves that ended up in Barbados, Puerto Rico and the US Southwest initially disembarked from
351 their voyage may bring them more into line with the others. Specifically for the Barbados plot,
352 even with the concept of an initital Native population hard-coded into the model, the algorithm
353 could only explain the genomic pattern by predicting that Europeans arrived 50-100 generations
354 ago, 500+ years before it really occured. This is likely because ADMIXTURE estimated that only
355 2 out of the 96 samples contain any Native DNA, both with a proportion of less than 0.1 - a stark
356 reminder that assortative mating and migration weren't the only population-shaping phenomena
357 at play in the colonial-era Americas.

358 These are not the only issues with this TRACTS analysis. The analysis uses ancestry proportion,
359 not absolute quantity of genetic material. This means Native populations shrinking due to disease
360 and other consequences of colonialism would have the same effect of increasing European ancestry
361 proportion in the population as European migration. Additionally, when interpreting TRACTS
362 plots it must be remembered that TRACTS is constrained to only one pulse per ancestry: at
363 the generation it deems to have had the biggest effect on the proportion of that ancestry. Until
364 a superior software permits predicted ancestry proportion to change each generation, comparing
365 TRACTS output with migration data has limited utility. Likewise, using African migration timing
366 as a proxy for European migration rather than integrating European migration data will limit the
367 potential of the method.

368 Finally, an inherent flaw in analysing social stratification using generation as the unit of time,
369 albeit unavoidable in genetic research, is that generation length is likely to differ significantly by
370 subgroup, and indeed over time. In slave-based societies of the southern US and carribean, slaves
371 breeding was a cheaper method of procuring additional slaves, hence one might expect African
372 subpopulations to have had shorter generation lengths earlier on. In a truly stratified society, one
373 would expect different stratas to have different generation lengths.

374 Ultimately, migration and the halting of assortative mating in the past, both the removal of
375 barriers to admixture, manifest themselves identically. Therefore, the two are seemingly inextric-
376 cable when projecting into the past absent accurate migration data to explain the contribution
377 of migration to admixture, thereby leaving only assortative mating. However, it may be possible
378 to use artificial neural networks to circumvent this need for migration data. Firstly, a model to

379 predict continuous ancestry tract length distribution based on input parameters such as level of
380 assortative mating must be created. This model can be used to simulate tract length distributions
381 with every combination of input parameters. An artificial neural network can then be trained to
382 learn the patterns between these distributions and the corresponding parameters. In theory, it may
383 subsequently be able to accurately predict the parameters, including level of assortative mating in
384 the population, when applied to the tract length distributions generated in this study with empirical
385 data - artificial neural networks have been successfully trained in this way before (Sheehan &
386 Song, 2016).

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

394 While artificial neural networks have potential in bypassing the issue of inadequate migration
395 data, integrating migration into the simulation model would make the method even more powerful.
396 This may not be possible to effectively with the current records of migration in the colonial-
397 era Americas, but could be used in modern populations: much higher-quality migration data is
398 available, although globalisation is leading to increasing ancestral diversity in populations, and
399 accounting for more ancestries adds complexity. Like the AMI analysis, this could have promise in
400 keeping track of ancestry-related social stratification in modern populations.

401 To increase the accuracy of these methods to quantify assortative mating, another factor must
402 be considered. Admixture of two ancestries may seem antithetical to ancestry-related social strat-
403 ification, not all admixture in a population is mutually voluntary. Many instances of admixture
404 between slavemaster and slave, or colonist and Native, were involuntary and thus symptomatic of
405 social stratification. To prevent involuntary admixture from counteracting assortative mating as a
406 proxy for ancestry-related social stratification, this would ideally be quantified and integrated into
407 the model. If it were assumed that negligible involuntary admixture occurred between European
408 Females and Native or African Males, similar analyses but with the recombining section of the X
409 chromosome rather than autosomes could be conducted. Discrepancies between assortative mating
410 in European female to non-European male admixture and in European male to non-European fe-
411 male admixture could shed light on this phenomenon, as could sex-specific contributions from each
412 ancestry in each population (Micheletti et al., 2020).

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

420 **5 Data and Code Availability**

421 **5.1 Data**

422 **1KGP Samples:**

423 <https://www.internationalgenome.org/data-portal/data-collection/30x-grch38>

424 **HGDP Samples:**

425 <https://www.internationalgenome.org/data-portal/data-collection/hgdp>

426 **Phasing Reference Panel:**

427 http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/
428 20201028_3202_phased/

429 **Phasing Genetic Map:**

430 https://github.com/odelaneau/shapeit4/blob/master/maps/genetic_maps.b38.tar.gz

431 **Slave Voyage Data:**

432 <https://www.slavevoyages.org/voyage/database#tables> (see tracts_mig_plots.R for details)

433 **5.2 Code**

434 **Code Repository:**

435 <https://github.com/Bennouhan/cmeecoursework/tree/master/project/code>

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

438 **References**

- 439 Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in
440 unrelated individuals. *Genome Research*, 19(9), 1655–1664. [https://doi.org/10.1101/gr.
441 094052.109](https://doi.org/10.1101/gr.094052.109)
- 442 Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Giordano, B. S. S., Leal, T. P., Furlan, V.,
443 Sciliar, M. O., Zamudio, R., Zolini, C., Araújo, G. S., Luizon, M. R., Padilla, C., Cáceres,
444 O., Levano, K., Sánchez, C., Trujillo, O., Flores-Villanueva, P. O., Dean, M., ... Tarazona-
445 Santos, E. (2020). The genetic structure and adaptation of Andean highlanders and Ama-
446 zonians are influenced by the interplay between geography and culture. *Proceedings of the
447 National Academy of Sciences of the United States of America*, 117(51), 32557–32565. <https://doi.org/10.1073/pnas.2013773117>
- 449 Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M.,
450 Bustamante, C. D., & Ostrer, H. (2010). Genome-wide patterns of population structure
451 and admixture among Hispanic/Latino populations. *Proceedings of the National Academy
452 of Sciences of the United States of America*, 107(SUPPL. 2), 8954–8961. [https://doi.org/
453 10.1073/pnas.0914618107](https://doi.org/10.1073/pnas.0914618107)
- 454 Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A.,
455 Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and
456 BCFtools. *GigaScience*, 10(2)arXiv 2012.10295, 1–4. [https://doi.org/10.1093/gigascience/
457 giab008](https://doi.org/10.1093/gigascience/giab008)
- 458 Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accu-
459 rate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1), 24–29.
460 <https://doi.org/10.1038/s41467-019-13225-y>
- 461 e Silva, M. A. C., Nunes, K., Lemes, R. B., Mas-Sandoval, À., Amorim, C. E. G., Krieger, J. E.,
462 Mill, J. G., Salzano, F. M., Bortolini, M. C., da Costa Pereira, A., Comas, D., & Hünemeier,
463 T. (2020). Genomic insight into the origins and dispersal of the Brazilian coastal natives.
464 *Proceedings of the National Academy of Sciences of the United States of America*, 117(5),
465 2372–2377. <https://doi.org/10.1073/pnas.1909075117>
- 466 Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, 191(2)arXiv 1202.4811,
467 607–619. <https://doi.org/10.1534/genetics.112.139808>
- 468 Linares-Pineda, T. M., Cañadas-Garre, M., Sánchez-Pozo, A., Calleja-Hernández, M., D’Haens,
469 G. R., Panaccione, R., Higgins, P. D., Vermeire, S., Gassull, M., Chowers, Y., Hanauer,
470 S. B., Herfarth, H., Hommes, D. W., Kamm, M., Löfberg, R., Quary, A., Sands, B., Sood,
471 A., Watermayer, G., ... Yang, J. (2012). Quality Control Procedures for Genome Wide
472 Association Studies. *American Journal of Human Genetics*, 573(6), 5–22. [https://doi.org/
473 10.1002/0471142905.hg0119s68.Quality](https://doi.org/10.1002/0471142905.hg0119s68.Quality)
- 474 Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A discriminative mod-
475 eling approach for rapid and robust local-ancestry inference. *American Journal of Human
476 Genetics*, 93(2), 278–288. <https://doi.org/10.1016/j.ajhg.2013.06.020>
- 477 Mas-Sandoval, A., Arauna, L. R., Gouveia, M. H., Barreto, M. L., Horta, B. L., Lima-Costa, M. F.,
478 Pereira, A. C., Salzano, F. M., Hünemeier, T., Tarazona-Santos, E., Bortolini, M. C., &
479 Comas, D. (2019). Reconstructed Lost Native American Populations from Eastern Brazil

- 480 Are Shaped by Differential Jê/Tupi Ancestry. *Genome Biology and Evolution*, 11(9), 2593–
481 2604. <https://doi.org/10.1093/gbe/evz161>
- 482 Micheletti, S. J., Bryc, K., Ancona Esselmann, S. G., Freyman, W. A., Moreno, M. E., Poznik, G. D.,
483 Shastri, A. J., Agee, M., Aslibekyan, S., Auton, A., Bell, R., Clark, S., Das, S., Elson, S.,
484 Fletez-Brant, K., Fontanillas, P., Gandhi, P., Heilbron, K., Hicks, B., ... Mountain, J. L.
485 (2020). Genetic Consequences of the Transatlantic Slave Trade in the Americas. *American
486 Journal of Human Genetics*, 107(2), 265–277. <https://doi.org/10.1016/j.ajhg.2020.06.012>
- 487 Norris, E. T., Rishishwar, L., Chande, A. T., Conley, A. B., Ye, K., Valderrama-Aguirre, A., & Jor-
488 dan, I. K. (2020). Admixture-enabled selection for rapid adaptive evolution in the Americas.
489 *Genome Biology*, 21(1), 1–29. <https://doi.org/10.1186/s13059-020-1946-2>
- 490 Norris, E. T., Rishishwar, L., Wang, L., Conley, A. B., Chande, A. T., Dabrowski, A. M.,
491 Valderrama-Aguirre, A., & King Jordan, I. (2019). Assortative mating on ancestry-variant
492 traits in admixed Latin American populations. *Frontiers in Genetics*, 10(APR), 1–14.
493 <https://doi.org/10.3389/fgene.2019.00359>
- 494 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
495 association and population-based linkage analyses. *American Journal of Human Genetics*,
496 81(3), 559–575. <https://doi.org/10.1086/519795>
- 497 Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela,
498 R., Rodriguez-Santana, J. R., Rodriguez-Cintron, W., Avila, P. C., Ziv, E., & Gonzalez
500 Burchard, E. (2009). Ancestry-related assortative mating in Latino populations. *Genome
501 Biology*, 10(11). <https://doi.org/10.1186/gb-2009-10-11-r132>
- 502 Schubert, R., Andaleon, A., & Wheeler, H. E. (2020). Comparing local ancestry inference models
503 in populations of two- And three-way admixture. *PeerJ*, 8, 1–19. <https://doi.org/10.7717/>
504 peerj.10090
- 505 Sheehan, S., & Song, Y. S. (2016). Deep Learning for Population Genetic Inference. *PLoS Compu-
506 tational Biology*, 12(3), 1–28. <https://doi.org/10.1371/journal.pcbi.1004845>
- 507 Smithjohn, M. U., Smith, M. U., & Baldwin, J. T. (2015). Making Sense of Hardy-Weinberg Equi-
508 librium What Is Hardy-Weinberg Equilibrium ? The H-W eq principle is , of course , the
509 cornerstone of introductory population genetics . *The American Biology Teacher*, 77(8),
510 577–582. <https://doi.org/10.1525/abt.2015.77.8.3.THE>
- 511 Zaitlen, N., Huntsman, S., Hu, D., Spear, M., Eng, C., Oh, S. S., White, M. J., Mak, A., Davis,
512 A., Meade, K., Brigino-Buenaventura, E., LeNoir, M. A., Bibbins-Domingo, K., Burchard,
513 E. G., & Halperin, E. (2017). The effects of migration and assortative mating on admixture
514 linkage disequilibrium. *Genetics*, 205(1), 375–383. <https://doi.org/10.1534/genetics.116.192138>

516 Supplementary Material

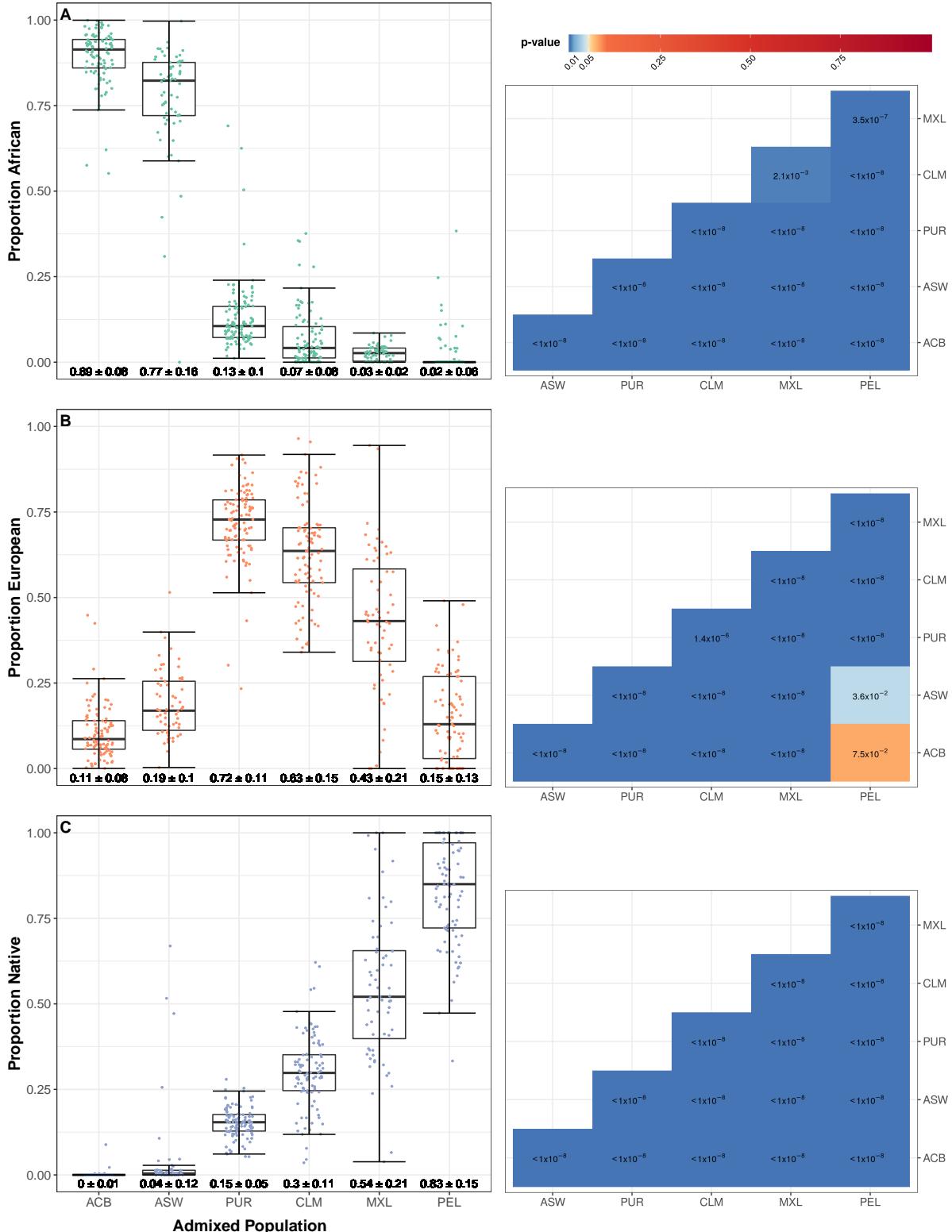


Figure S1: Comparative box plots displaying the distributions of the three ancestry proportions for each individual of each admixed population, with corresponding p-value heatmaps comparing populations statistically. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers signify the last data point within the closest quartile value plus 150% of the interquartile range. Mean \pm 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.

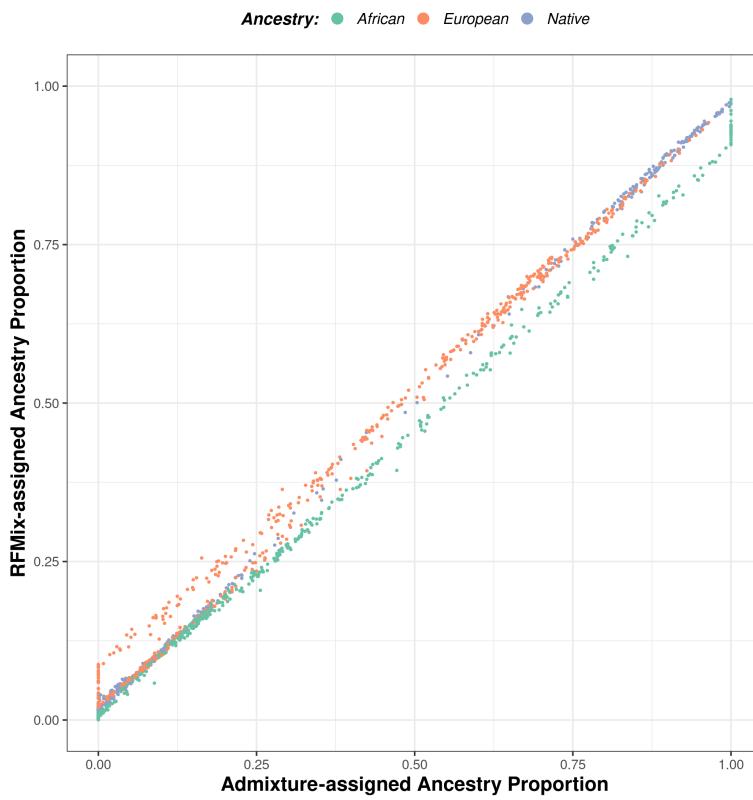


Figure S2: Scatterplot correlating ancestry proportions assigned by RFMIX for all 1690 query and reference individuals against those assigned by ADMIXTURE.

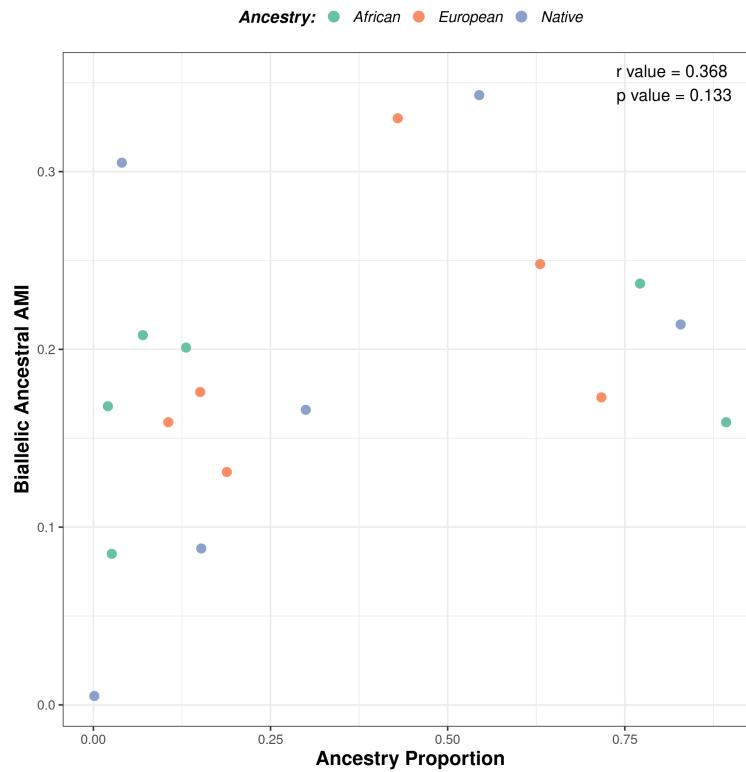


Figure S3: Scatterplot charting all three mean biallelic ancestry-specific AMI against all three ancestry proportion for each of the six admixed populations.

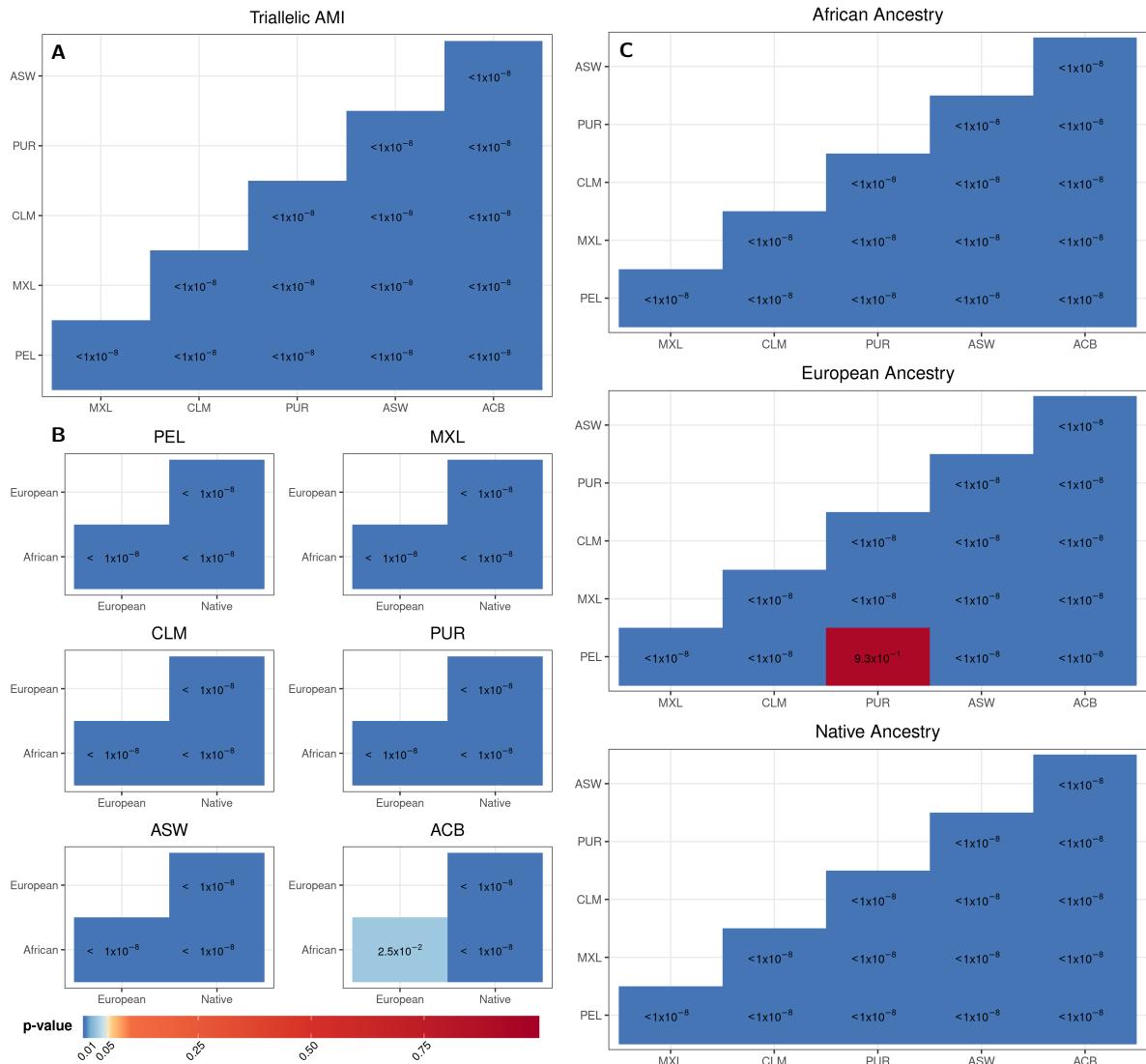


Figure S4: Heatmaps displaying p-value results of Wilcoxon tests used to compare assortative mating index values of different populations and ancestries. Each set of heatmaps correspond to a different set of comparisons between all combinations of assortative mating index (AMI) distributions. **A** compares all combinations of the six admixed populations with regards to their triallelic AMI 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.

Ancestry: African (green) European (orange) Native (blue)

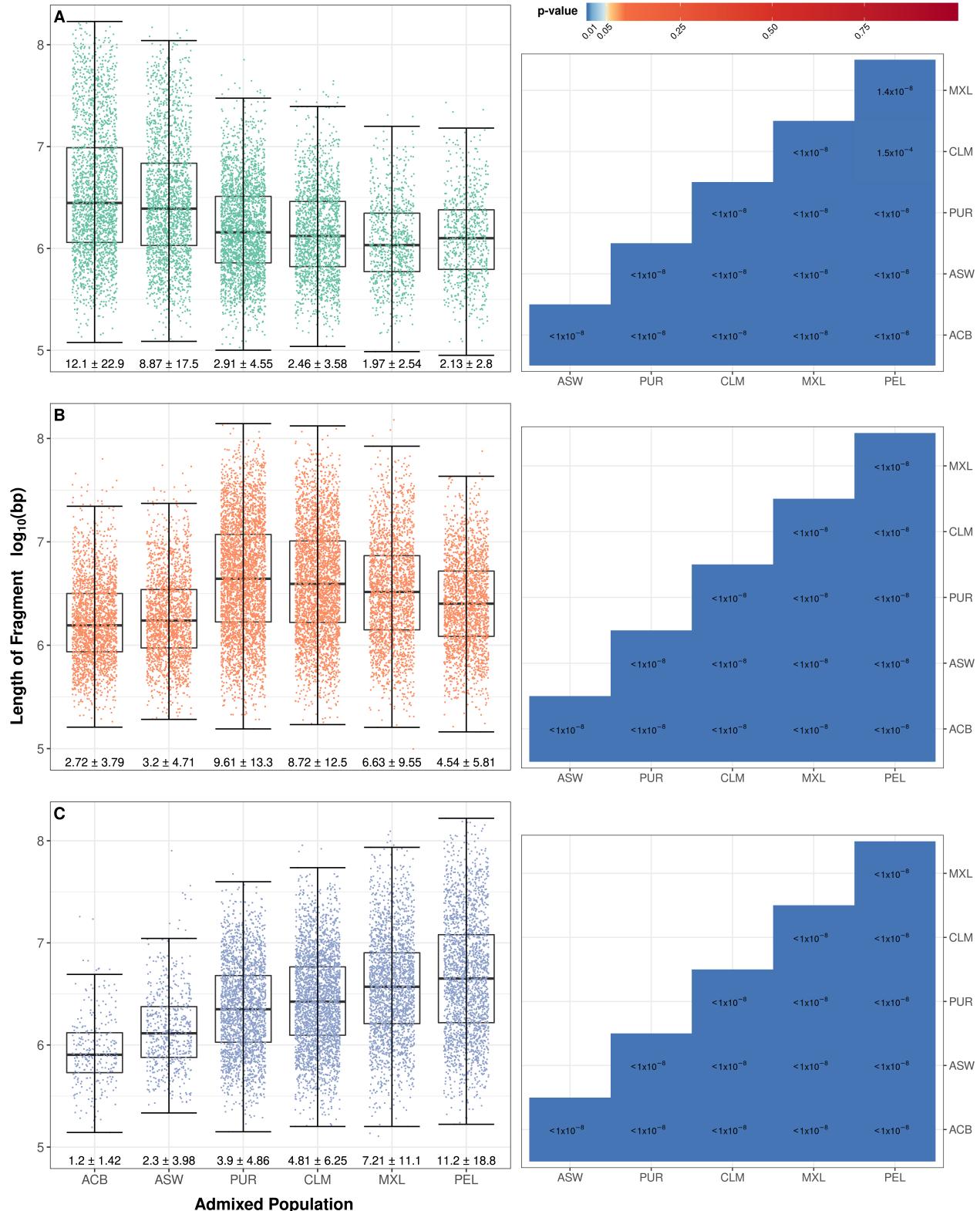


Figure S5: Comparative box plots displaying the distributions of continuous ancestry tract lengths of each ancestry for all individuals of each admixed population, with corresponding p-value heatmaps comparing populations statistically. Fragment length, that is the number of consecutive haplotype assignments of a given ancestry on a single strand, are measured in base pairs in \log_{10} 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 \pm 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 African, 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.

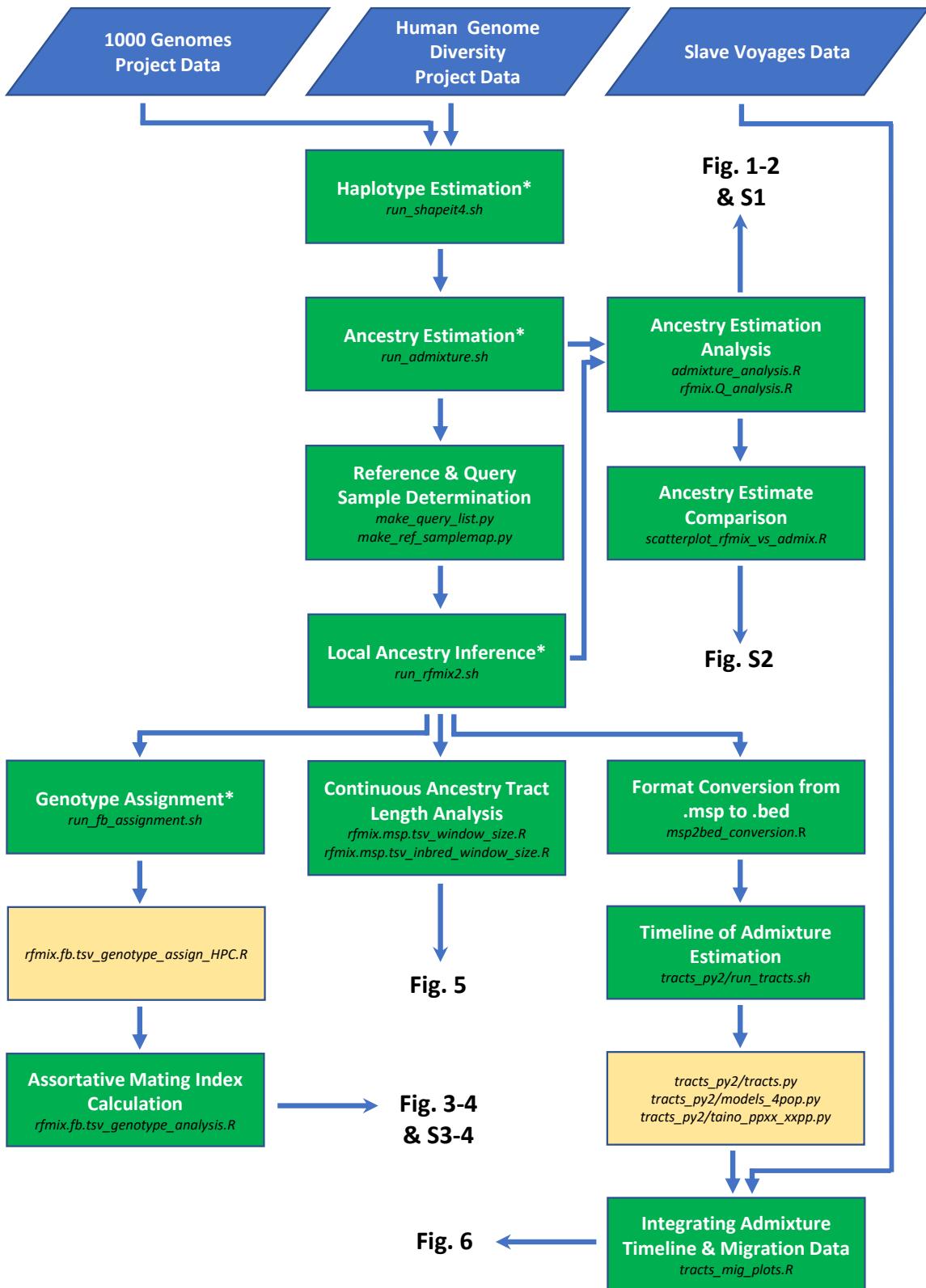


Figure S6: Flowchart representing the analysis workflow of the project, from input data to the output figures. Arrows indicate that the output from one step is the input for the next. Below the label of each step is the script(s) from the provided github repository required to run that step. The scripts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