

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

Ben Nouhan, bjin20@ic.ac.uk

Imperial College London

August 24, 2021

Word Count: 5178

1 250 word abstract placeholder: 10000 10001 10002 10003 10004 10005 10006 10007
2 10008 10009 10010 10011 [13] 10012 10013 10014 10015 10016 10017 10018 10019
3 10020 10021 10022 10023 [25] 10024 10025 10026 10027 10028 10029 10030 10031
4 10032 10033 10034 10035 [37] 10036 10037 10038 10039 10040 10041 10042 10043
5 10044 10045 10046 10047 [49] 10048 10049 10050 10051 10052 10053 10054 10055
6 10056 10057 10058 10059
7 [61] 10060 10061 10062 10063 10064 10065 10066 10067 10068 10069 10070 10071
8 [73] 10072 10073 10074 10075 10076 10077 10078 10079 10080 10081 10082 10083
9 [85] 10084 10085 10086 10087 10088 10089 10090 10091 10092 10093 10094 10095
10 [97] 10096 10097 10098 10099 10100 10101 10102 10103 10104 10105 10106 10107
11 [109] 10108 10109 10110 10111 10112 10113 10114 10115 10116 10117 10118
12 10119 [121] 10120 10121 10122 10123 10124 10125 10126 10127 10128 10129 10130
13 10131[133] 10132 10133 10134 10135 10136 10137 10138 10139 10140 10141 10142
14 10143[145] 10144 10145 10146 10147 10148 10149 10150 10151 10152 10153 10154
15 10155
16 [157] 10156 10157 10158 10159 10160 10161 10162 10163 10164 10165 10166
17 10167 [169] 10168 10169 10170 10171 10172 10173 10174 10175 10176 10177 10178
18 10179 [181] 10180 10181 10182 10183 10184 10185 10186 10187 10188 10189 10190
19 10191
20 [193] 10192 10193 10194 10195 10196 10197 10198 10199 10200 10201 10202
21 10203 [205] 10204 10205 10206 10207 10208 10209 10210 10211 10212 10213 10214
22 10215 [217] 10216 10217 10218 10219 10220 10221 10222 10223 10224 10225.

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 | 17 |
| 5.1 | Data | 17 |
| 5.2 | Code | 18 |
| | References | 19 |
| | Supplementary Material | 21 |

23 **1 Introduction**

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

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

36 Single nucleotide polymorphisms (SNPs) can be used as indicators of ancestry (Risch et al.,
37 2009). Population genomics techniques allow us to generate a large array of SNPs which can be
38 analysed using local ancestry inference to map ancestries to positions and regions along the genome,
39 after which further analysis can indicate past assortative mating in a population (Schubert et al.,
40 2020).

41 One such analysis is that of continuous ancestry tract lengths: the lengths of genomic regions
42 consecutively assigned to the same ancestry. Looking at the distribution of these lengths, the
43 ancestry to which they belong and the overall ancestry proportion of individuals within a population
44 can indicate how long ago the admixture occurred and to what extent. Recombination of the DNA
45 of admixing individuals leads to a decrease in continuous ancestry tract lengths, as those within the
46 parents' genomes interrupt one another upon recombination. Hence, admixture more generations
47 ago will manifest as distributions of shorter continuous ancestry tracts and vice versa (Gravel,
48 2012).

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

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

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

65 such populations have different but connected histories: a Native American population is colonised
66 by Europeans; the Native population shrinks due to war, hard labour and disease, while the Eu-
67 ropean population grows via migration. These phenomena continue such that African slaves are
68 transported to the region as a source of additional labour; after which the population continues
69 evolving with these scars of colonialism (Bryc et al., 2010; e Silva et al., 2020; Mas-Sandoval et al.,
70 2019).

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

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

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

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

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

¹⁰⁶ **2 Methods**

¹⁰⁷ **2.1 Studied Populations**

¹⁰⁸ For the initial analyses, all African, European and American populations from the 1000 Genomes
¹⁰⁹ Project (1KGP) and the Human Genome Diversity Project (HGDP) were used **Table 1**, with the
¹¹⁰ exception of the Russian and Finnish populations. These were excluded owing to minimal colonial-
¹¹¹ era migration to the Americas from these populations, alongside the genetic similarities between
¹¹² 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 |

¹¹³ **2.2 Data Preparation with BCFtools**

¹¹⁴ Using BCFtools v1.9, the 30x coverage 1KGP and high-coverage HGDP datasets were merged, and
¹¹⁵ all populations except those listed in (**Table 1**) were removed. All C→G, G→C, A→T and T→A
¹¹⁶ SNPs were filtered out as they are harder to assign and are hence prone to error (Danecek et al.,
¹¹⁷ 2021). SNPs were further filtered with a minor allele frequency threshold of 5%, as to reduce the
¹¹⁸ dataset and remove rare and thus uninformative SNPs. Following this, all 22 filtered VCF files,

119 one per autosome, were indexed for phasing.

120 2.3 Haplotype Estimation with SHAPEIT4

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

128 2.4 Ancestry Estimation with PLINK & ADMIXTURE

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

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

136 2.5 Local Ancestry Inference with RFMIX v2

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

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

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

152 2.6 Assortative Mating Index Calculation

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

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

155

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

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

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

159

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

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

177 2.7 Continuous Ancestry Tract Length Analysis

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

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

188 ments, rather than haplotype assignments.

189 2.8 Timeline of Admixture Estimation with TRACTS

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

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

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

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

216 3 Results

217 3.1 Ancestry Proportion

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

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

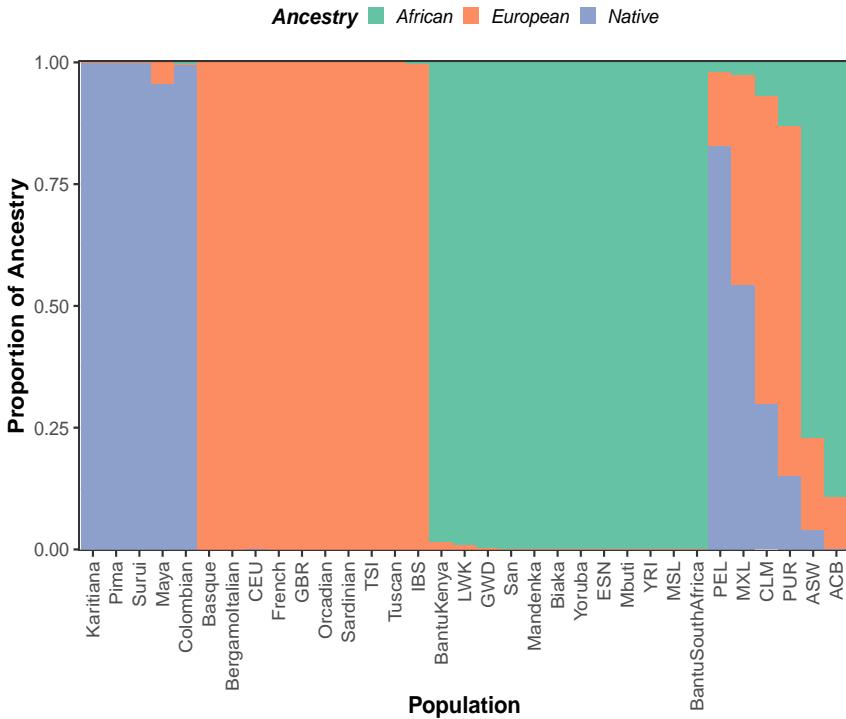


Figure 1: Stacked barplots showing the proportions of the three ancestries of each reference or query population used throughout the study, generated by ADMIXTURE. Genomic data from individuals of selected populations from the 1000 Genomes Project and the Human Genome Diversity Project were processed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

226 The distribution of ancestry proportions on the individual level within these six admixed pop-
 227ulations is shown in **Fig. 2**, which largely corresponds with **Fig. 1** while suggesting approximately
 228 25% of LA Mexicans and Colombians have no African ancestry, fewer than 5% and 50% of Barba-
 229 dians and ASWs respectively have Native ancestry, and that only around 20% of Peruvians have
 230 African ancestry while around 25% of them are of exclusively Native ancestry.

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

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

241 3.2 Assortative Mating Index

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

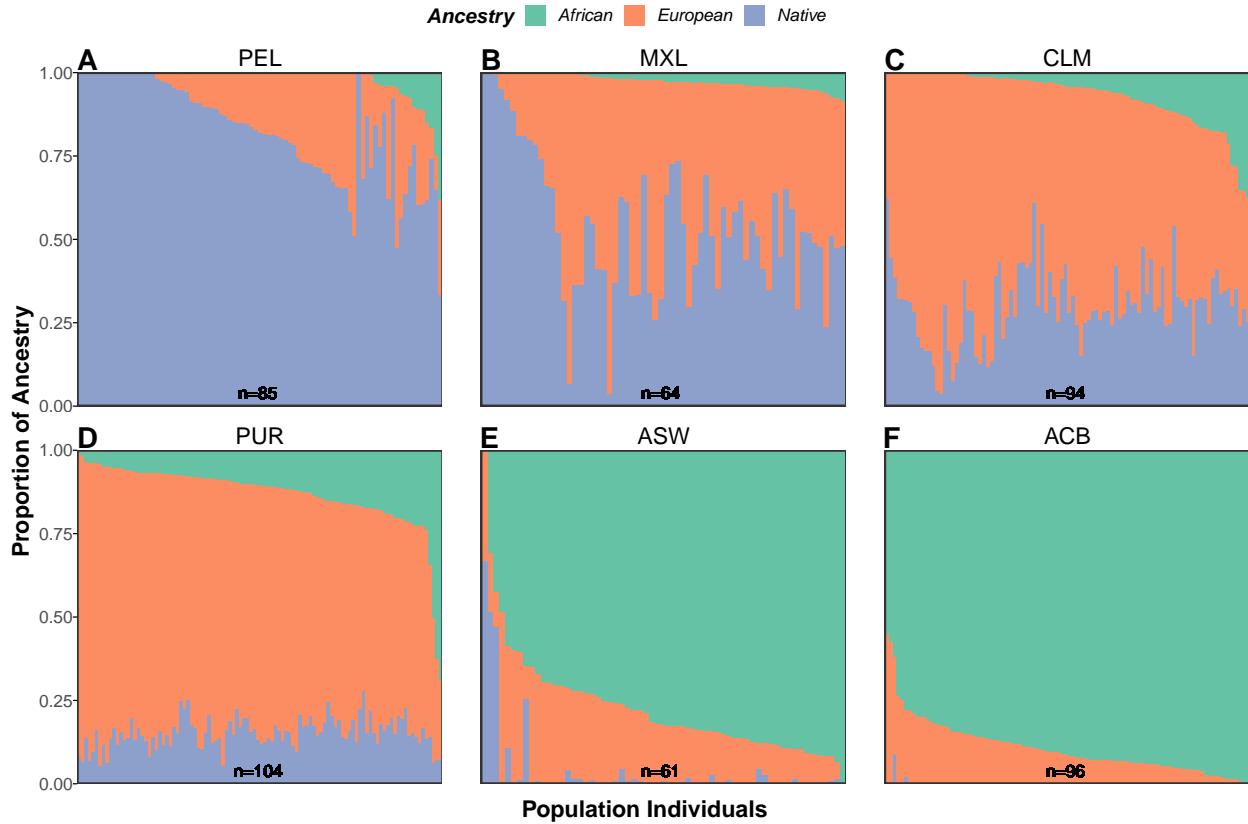


Figure 2: Stacked barplots showing the proportions of the three ancestries of each individual comprising the six query admixed populations, generated by ADMIXTURE. Genomic data from individuals of the six from the 1000 Genomes Project and the Human Genome Diversity Project were processed and subjected 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.

246 to have lower European Ancestries.

247 The main RFMIX output was used to calculate assortative mating index values for each SNP
 248 in each population. The triallelic AMI values for each position and population are plotted in **Fig.**
 249 **3**. In a population without assortative mating, we would expect the mean AMI value to be zero.
 250 With a sample size of 4,111,226 SNPs, and the standard deviations being of similar sizes to the
 251 corresponding means, the standard errors of the means are negligible and hence the sample means
 252 are accurate estimates of the true means. Based on this, we can see all means are significantly
 253 higher than zero, indicating positive assortative mating in all admixed populations.

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

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

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

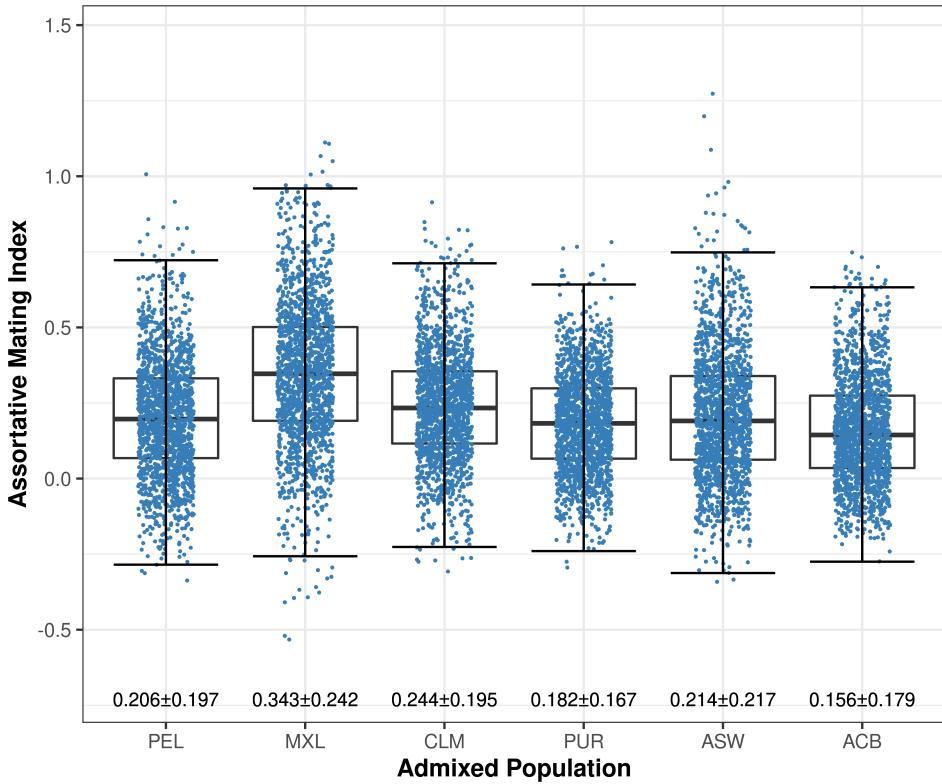


Figure 3: Comparative box plots displaying the distribution of the triallelic assortative mating index calculated for each studied single nucleotide polymorphism for each admixed population. The boxes highlight the upper quartile, median and lower quartile values of the distribution, while the whiskers 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.

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

268 3.3 Continuous Ancestry Tract Lengths

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

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

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

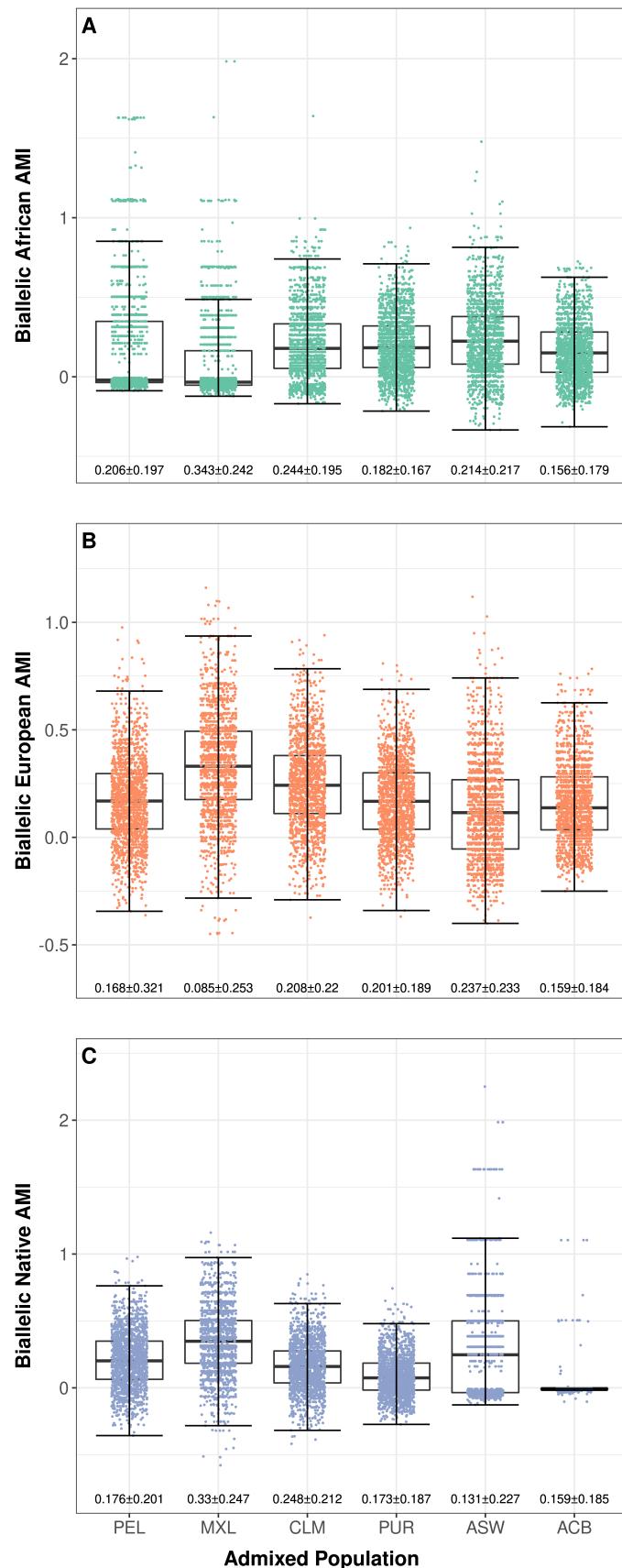


Figure 4: Comparative box plots displaying the distribution of the biallelic ancestry-specific assortative mating indices calculated for each studied single nucleotide polymorphism for each admixed population. The boxes highlight the upper quartile, median and lower quartile values of the distribution, 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.

284 ancestry to merge into one.

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

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

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

298 Therefore, for example in Peru, we can see that Europeans and Africans began arriving around
299 1525, the majority of Africans had arrived by 1575, significant admixture between Europeans and
300 Natives occurred around 1650, and significant admixture between Africans and the rest of the
301 population occurred around 1675. This suggests extreme assortative mating for 4-6 generations in
302 Europeans and a similar length, albeit lagging by a generation, in Africans.

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

307 4 Discussion

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

313 The RFMIX reference panel was heavily imbalanced, with 72 Native samples to 507 European
314 and 550 African. This few Native samples will make the algorithm more likely to assign SNP
315 alleles to European or African despite being more indicative of Native ancestry. This could be
316 responsible for the assignment of approximately 5% European ancestry in the Mayan population
317 (**Fig. 1**), and can be resolved by sequencing more Native American genomes. Also, based on **Fig. S2**,
318 ADMIXTURE seems to estimate 100% African or 0% European more readily than RFMIX,
319 suggesting it may be less sensitive at those two extremes and that the RFMIX ancestry proportion
320 estimate might have been the better choice.

321 One aim of this project was to test the hypothesis that the levels of assortative mating in the
322 studied populations were significant, and significantly different to each other. This hypothesis was

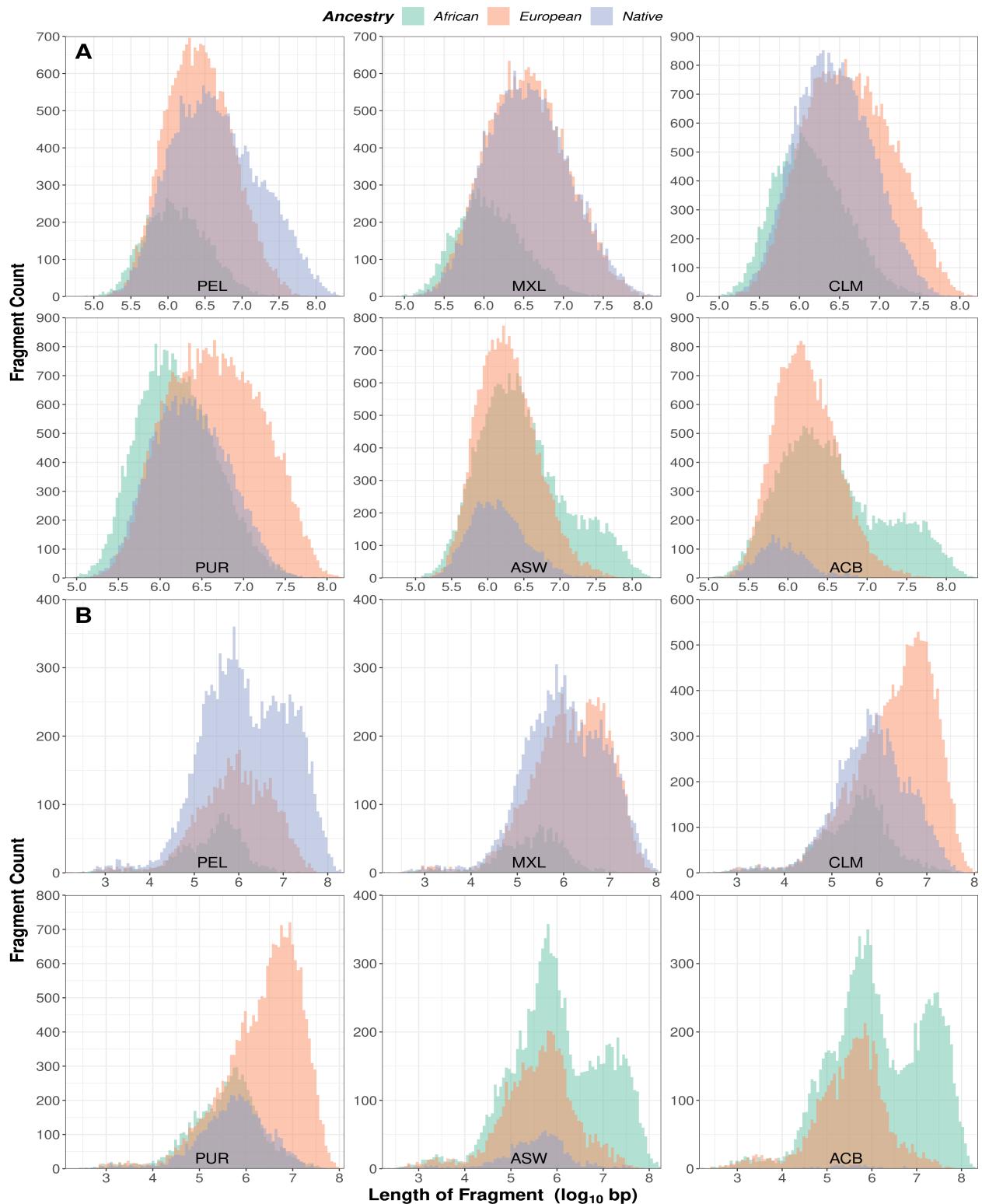


Figure 5: Histograms of continuous ancestry tract lengths of each of the three ancestries for each admixed population. Fragment lengths are measured in base pairs in \log_{10} scale, and are separated into 100 bins in each plot. Fragment length is considered either the number of consecutive haplotype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

323 supported in full by the results of the AMI analysis, both with triallelic and ancestry-specific AMI
 324 (**Fig. 3-4**). The other aim was to assess the methods used throughout, including this AMI analysis,
 325 by their ability to differentiate between assortative mating and migration, and thus whether the
 326 results are valid - a far more nuanced task.

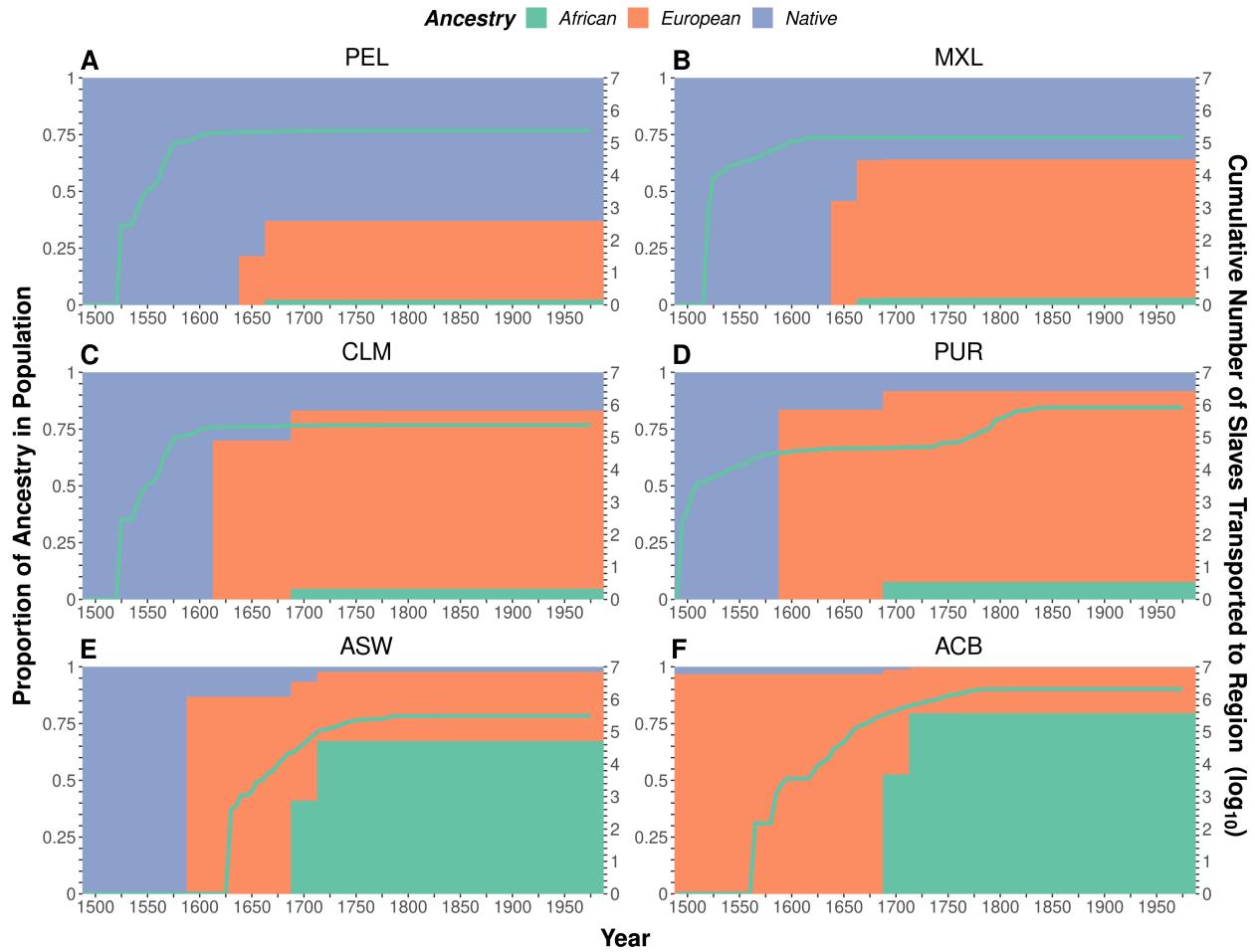


Figure 6: Number of slaves transported to the general regions of the admixed populations every five years, and stacked barplots showing how the proportion of the three ancestries changed in said populations generation to generation as estimated by TRACTS, between the years 1500 and 2000 CE. Data on the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

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

337 However, the AMI analysis method only allows us to reach that conclusion for the present-day
 338 populations, it tells us little about their past. Continuous ancestry tract analysis is better-suited
 339 for this, but the results are less conclusive. The histogram visualisation of the homozygous tract
 340 analysis (**Fig. 5B**) not only indicates that the right skewness in the **Fig. 5A** African ASW and
 341 ACB plots and Native PEL plot are due to migration rather than assortative mating, but also

342 shows in the Native MXL distribution that two peaks can mascerade as one. This method does not
343 quantify the length of time passed since admixture, and hence can only be used to highlight major
344 same-ancestry migration events in the past for further investigation. It fails differentiate between
345 minor and prolonged migration and assortative mating.

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

352 In the others, the majority of the slaves were allegedly transported after the generation of
353 most significant admixture, roughly by an order of magnitude in all three cases. More thoroughly
354 researching the histories of these three regions and more accurately determining the ports at which
355 slaves that ended up in Barbados, Puerto Rico and the US Southwest initially disembarked from
356 their voyage may bring them more into line with the others. Specifically for the Barbados plot,
357 even with the concept of an initital Native population hard-coded into the model, the algorithm
358 could only explain the genomic pattern by predicting that Europeans arrived 50-100 generations
359 ago, 500+ years before it really occured. This is likely because ADMIXTURE estimated that only
360 2 out of the 96 samples contain any Native DNA, both with a proportion of less than 0.1 - a stark
361 reminder that assortative mating and migration weren't the only population-shaping phenomena
362 at play in the colonial-era Americas.

363 These are not the only issues with this TRACTS analysis. The analysis uses ancestry proportion,
364 not absolute quantity of genetic material. This means Native populations shrinking due to disease
365 and other consequences of colonialism would have the same effect of increasing European ancestry
366 proportion in the population as European migration. Additionally, when interpretting TRACTS
367 plots it must be remembered that TRACTS is constrained to only one pulse per ancestry: at
368 the generation it deems to have had the biggest effect on the proportion of that ancestry. Until
369 a superior software permits predicted ancestry proportion to change each generation, comparing
370 TRACTS output with migration data has limited utility. Likewise, using African migration timing
371 as a proxy for European migration rather than integrating European migration data will limit the
372 potential of the method. Finally, an inherent flaw in analysing social stratification using generation
373 as the unit of time, albeit unavoidable in genetic research, is that generation length is likely to
374 differ significantly by subgroup, and indeed over time. In slave-based societies of the southern US
375 and carribean, slaves breeding was a cheaper method of procuring additional slaves, hence one
376 might expect African subpopulations to have had shorter generation lengths earlier on. In a truly
377 stratified society, one would expect different stratas to have different generation lengths.

378 As genome sequencing gets cheaper, can use this to keep track of AM levels in a population, to
379 see if increasing or decreasing, e.g. testing 18-25 year olds every 10 or so years. Harder in socities
380 with high proportions of lots of ancestries, but could still test the most prevalent ancestries.

381 Any historical backing for mexico having highest AM? Is quite a segregated society of wealthy
382 majoirty european mexicans with lifestyles equivalent to their couterparts in developed countries,
383 vs far poorer majority Native Mexicans.

384 While admixture of two ancestries could be seen as antithetical of ancestry-related social strat-

385 ification, not all admixture in a population is mutually voluntary. Many instances of admixture
386 between slavemaster and slave, or colonist and Native, were involuntary and thus symptomatic
387 of social stratification. To prevent involuntary admixture from counteracting assortative mating
388 as a proxy for ancestry-related social stratification, this would ideally be quantified or integrated
389 into the model. If it were assumed that negligible involuntary admixture occurred between Euro-
390 pean Females and Native or African Males, analyses similar to those herein, with the recombining
391 section of the X chromosome rather than autosomes, could be conducted. Discrepancies between
392 assortative mating in European female to non-European male admixture and in European male
393 to non-European female admixture could shed light on this phenomenon, as could sex-specific
394 contributions from each ancestry in each population (Micheletti et al., 2020).

395 End of the day, migration and the ending of AM, both the removal of barriers to admixture,
396 manifest themselves identically - hence the two are inextricable absent accurate migration data
397 which can be used to explain the contribution of migration to admixture, leaving information
398 as to the impact of assortative mating. It may be possible to use artificial neural networks to
399 circumvent this need for migration data. Firstly, a model to predict continuous ancestry tract
400 length distribution based on input parameters such as level of assortative mating must be created.
401 This model can be used to simulate tract length distributions with every combination of input
402 parameters. An artificial neural network can then be trained to learn the patterns between these
403 distributions and the corresponding parameters. In theory, it may subsequently be able to accurately
404 predict the parameters, including level of assortative mating in the population, when applied to
405 the tract length distributions generated in this study with empirical data.

406 While this has potential in bypassing the issue of inadequate migration data, integrating mi-
407 gration into the simulation model would make the method even more powerful. This may not be
408 possible to be done effectively with the current records of migration in the colonial-era Americas,
409 but could be used in modern populations: much higher-quality migration data is available, although
410 globalisation is leading to increasing ancestral diversity in populations, and accounting for more
411 ancestries adds complexity.

412 5 Data and Code Availability

413 5.1 Data

414 1KGP Samples:

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

416 HGDP Samples:

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

418 Phasing Reference Panel:

419 http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000G_2504_high_coverage/working/
420 20201028_3202_phased/

421 Phasing Genetic Map:

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

423 **Slave Voyage Data:**

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

425 **5.2 Code**

426 **Code Repository:**

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

428 A detailed visualisation of the project's workflow can be found in **Fig. S6**, indicating which

429 script(s) were used during each step in the analyses. See the README.md for further details.

430 **References**

- 431 Alexander, D. H., Novembre, J., & Lange, K. (2009). Fast model-based estimation of ancestry in
432 unrelated individuals. *Genome Research*, 19(9), 1655–1664. [https://doi.org/10.1101/gr.
433 094052.109](https://doi.org/10.1101/gr.094052.109)
- 434 Borda, V., Alvim, I., Mendes, M., Silva-Carvalho, C., Giordano, B. S. S., Leal, T. P., Furlan, V.,
435 Scliar, M. O., Zamudio, R., Zolini, C., Araújo, G. S., Luizon, M. R., Padilla, C., Cáceres,
436 O., Levano, K., Sánchez, C., Trujillo, O., Flores-Villanueva, P. O., Dean, M., ... Tarazona-
437 Santos, E. (2020). The genetic structure and adaptation of Andean highlanders and Ama-
438 zonians are influenced by the interplay between geography and culture. *Proceedings of the
439 National Academy of Sciences of the United States of America*, 117(51), 32557–32565. <https://doi.org/10.1073/pnas.2013773117>
- 440 Bryc, K., Velez, C., Karafet, T., Moreno-Estrada, A., Reynolds, A., Auton, A., Hammer, M.,
441 Bustamante, C. D., & Ostrer, H. (2010). Genome-wide patterns of population structure
442 and admixture among Hispanic/Latino populations. *Proceedings of the National Academy
443 of Sciences of the United States of America*, 107(SUPPL. 2), 8954–8961. [https://doi.org/
444 10.1073/pnas.0914618107](https://doi.org/10.1073/pnas.0914618107)
- 445 Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A.,
446 Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and
447 BCFtools. *GigaScience*, 10(2)arXiv 2012.10295, 1–4. [https://doi.org/10.1093/gigascience/
448 giab008](https://doi.org/10.1093/gigascience/giab008)
- 449 Delaneau, O., Zagury, J. F., Robinson, M. R., Marchini, J. L., & Dermitzakis, E. T. (2019). Accu-
450 rate, scalable and integrative haplotype estimation. *Nature Communications*, 10(1), 24–29.
451 <https://doi.org/10.1038/s41467-019-13225-y>
- 452 e Silva, M. A. C., Nunes, K., Lemes, R. B., Mas-Sandoval, À., Amorim, C. E. G., Krieger, J. E.,
453 Mill, J. G., Salzano, F. M., Bortolini, M. C., da Costa Pereira, A., Comas, D., & Hünemeier,
454 T. (2020). Genomic insight into the origins and dispersal of the Brazilian coastal natives.
455 *Proceedings of the National Academy of Sciences of the United States of America*, 117(5),
456 2372–2377. <https://doi.org/10.1073/pnas.1909075117>
- 457 Gravel, S. (2012). Population genetics models of local ancestry. *Genetics*, 191(2)arXiv 1202.4811,
458 607–619. <https://doi.org/10.1534/genetics.112.139808>
- 459 Linares-Pineda, T. M., Cañadas-Garre, M., Sánchez-Pozo, A., Calleja-Hernández, M., D’Haens,
460 G. R., Panaccione, R., Higgins, P. D., Vermeire, S., Gassull, M., Chowers, Y., Hanauer,
461 S. B., Herfarth, H., Hommes, D. W., Kamm, M., Löfberg, R., Quary, A., Sands, B., Sood,
462 A., Watermayer, G., ... Yang, J. (2012). Quality Control Procedures for Genome Wide
463 Association Studies. *American Journal of Human Genetics*, 573(6), 5–22. [https://doi.org/
464 10.1002/0471142905.hg0119s68.Quality](https://doi.org/10.1002/0471142905.hg0119s68.Quality)
- 465 Maples, B. K., Gravel, S., Kenny, E. E., & Bustamante, C. D. (2013). RFMix: A discriminative mod-
466 eling approach for rapid and robust local-ancestry inference. *American Journal of Human
467 Genetics*, 93(2), 278–288. <https://doi.org/10.1016/j.ajhg.2013.06.020>
- 468 Mas-Sandoval, A., Arauna, L. R., Gouveia, M. H., Barreto, M. L., Horta, B. L., Lima-Costa, M. F.,
469 Pereira, A. C., Salzano, F. M., Hünemeier, T., Tarazona-Santos, E., Bortolini, M. C., &
470 Comas, D. (2019). Reconstructed Lost Native American Populations from Eastern Brazil

- 472 Are Shaped by Differential Jê/Tupi Ancestry. *Genome Biology and Evolution*, 11(9), 2593–
473 2604. <https://doi.org/10.1093/gbe/evz161>
- 474 Micheletti, S. J., Bryc, K., Ancona Esselmann, S. G., Freyman, W. A., Moreno, M. E., Poznik, G. D.,
475 Shastri, A. J., Agee, M., Aslibekyan, S., Auton, A., Bell, R., Clark, S., Das, S., Elson, S.,
476 Fletez-Brant, K., Fontanillas, P., Gandhi, P., Heilbron, K., Hicks, B., ... Mountain, J. L.
477 (2020). Genetic Consequences of the Transatlantic Slave Trade in the Americas. *American
478 Journal of Human Genetics*, 107(2), 265–277. <https://doi.org/10.1016/j.ajhg.2020.06.012>
- 479 Norris, E. T., Rishishwar, L., Chande, A. T., Conley, A. B., Ye, K., Valderrama-Aguirre, A., & Jord-
480 dan, I. K. (2020). Admixture-enabled selection for rapid adaptive evolution in the Americas.
481 *Genome Biology*, 21(1), 1–29. <https://doi.org/10.1186/s13059-020-1946-2>
- 482 Norris, E. T., Rishishwar, L., Wang, L., Conley, A. B., Chande, A. T., Dabrowski, A. M.,
483 Valderrama-Aguirre, A., & King Jordan, I. (2019). Assortative mating on ancestry-variant
484 traits in admixed Latin American populations. *Frontiers in Genetics*, 10(APR), 1–14.
485 <https://doi.org/10.3389/fgene.2019.00359>
- 486 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
487 association and population-based linkage analyses. *American Journal of Human Genetics*,
488 81(3), 559–575. <https://doi.org/10.1086/519795>
- 489 Risch, N., Choudhry, S., Via, M., Basu, A., Sebro, R., Eng, C., Beckman, K., Thyne, S., Chapela,
490 R., Rodriguez-Santana, J. R., Rodriguez-Cintron, W., Avila, P. C., Ziv, E., & Gonzalez
491 Burchard, E. (2009). Ancestry-related assortative mating in Latino populations. *Genome
492 Biology*, 10(11). <https://doi.org/10.1186/gb-2009-10-11-r132>
- 493 Schubert, R., Andaleon, A., & Wheeler, H. E. (2020). Comparing local ancestry inference models
494 in populations of two- And three-way admixture. *PeerJ*, 8, 1–19. <https://doi.org/10.7717/peerj.10090>
- 495 Smithjohn, M. U., Smith, M. U., & Baldwin, J. T. (2015). Making Sense of Hardy-Weinberg Equi-
496 librium What Is Hardy-Weinberg Equilibrium ? The H-W eq principle is , of course , the
497 cornerstone of introductory population genetics . *The American Biology Teacher*, 77(8),
498 577–582. <https://doi.org/10.1525/abt.2015.77.8.3.THE>
- 499 Zaïtlen, N., Huntsman, S., Hu, D., Spear, M., Eng, C., Oh, S. S., White, M. J., Mak, A., Davis,
500 A., Meade, K., Brigino-Buenaventura, E., LeNoir, M. A., Bibbins-Domingo, K., Burchard,
501 E. G., & Halperin, E. (2017). The effects of migration and assortative mating on admixture
502 linkage disequilibrium. *Genetics*, 205(1), 375–383. <https://doi.org/10.1534/genetics.116.192138>

506 Supplementary Material

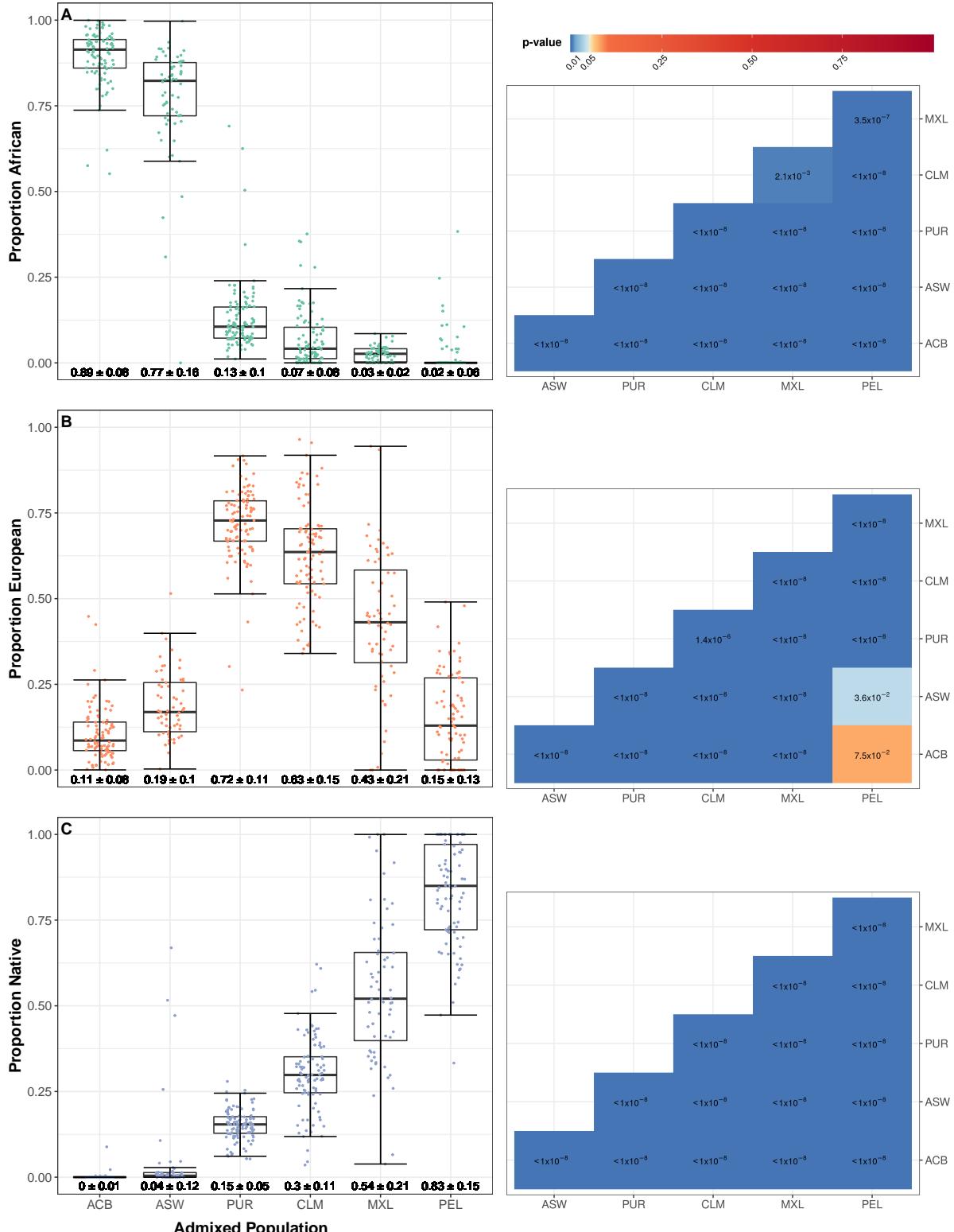


Figure S1: Comparative box plots displaying the distributions of the three ancestry proportions for each individual of each admixed population, with corresponding p-value heatmaps comparing populations statistically. The boxes highlight the upper quartile, median and lower quartile values of the distribution,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

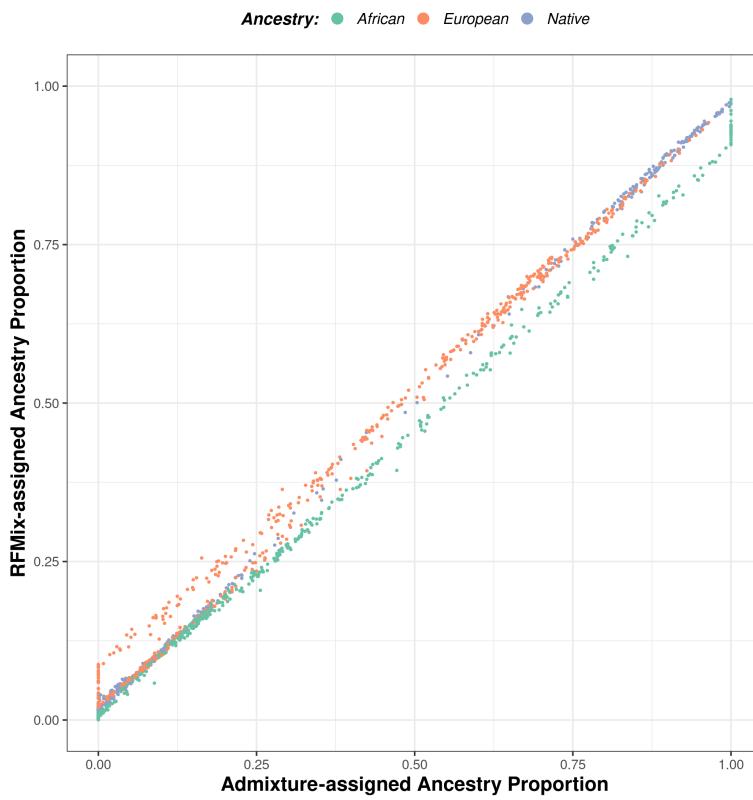


Figure S2: Scatterplot correlating ancestry proportions assigned by RFMIX for all 1690 query and reference individuals against those assigned by ADMIXTURE.

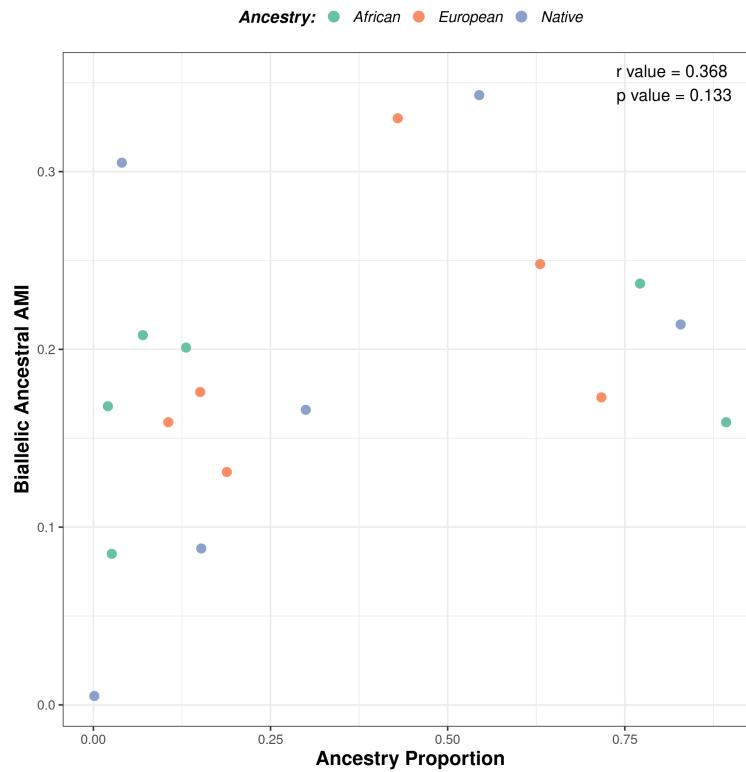


Figure S3: Scatterplot charting all three mean biallelic ancestry-specific AMI against all three ancestry proportion for each of the six admixed populations.

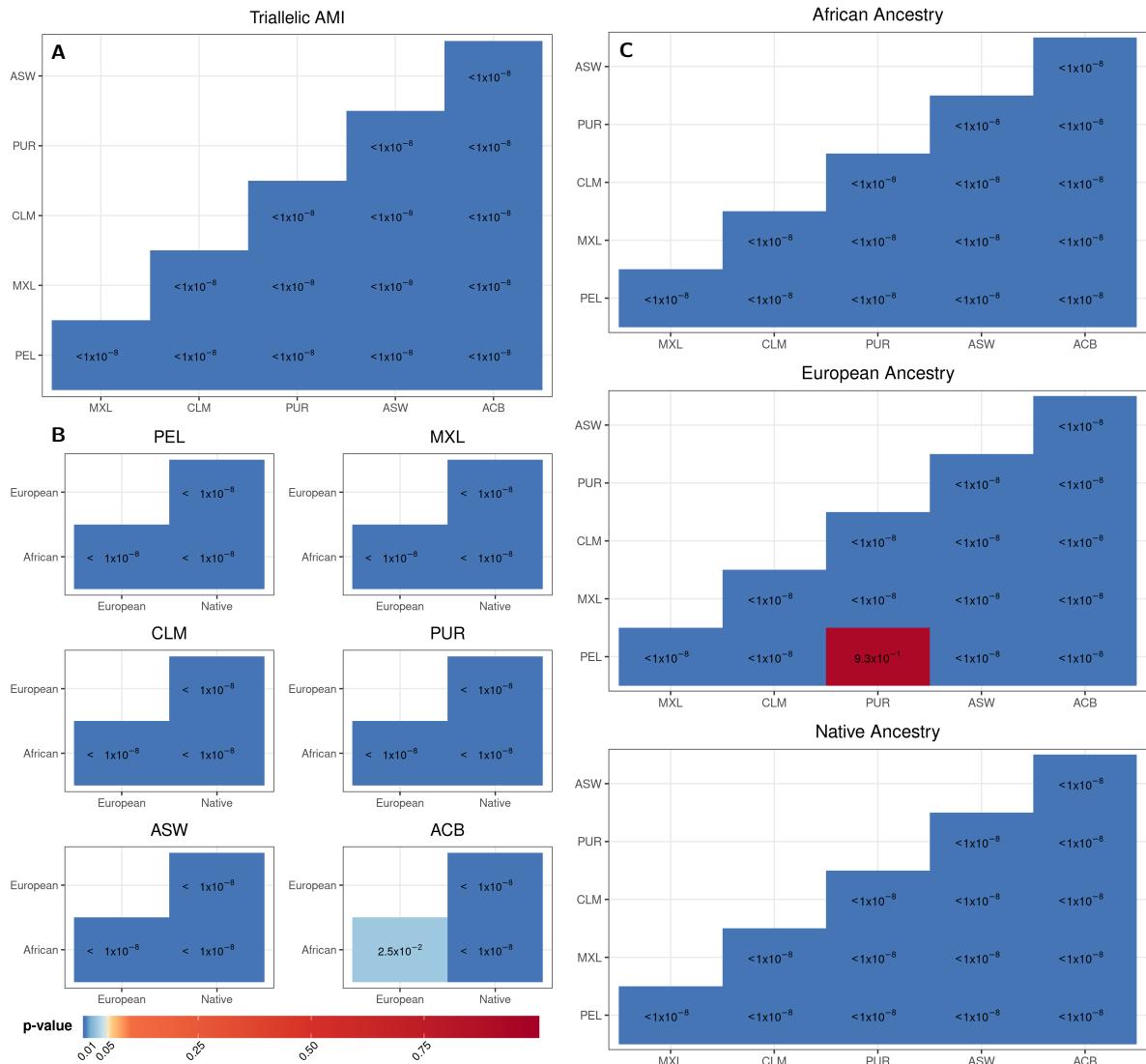


Figure S4: Heatmaps displaying p-value results of Wilcoxon tests used to compare assortative mating index values of different populations and ancestries. Each set of heatmaps correspond to a different set of comparisons between all combinations of assortative mating index (AMI) distributions. **A** compares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

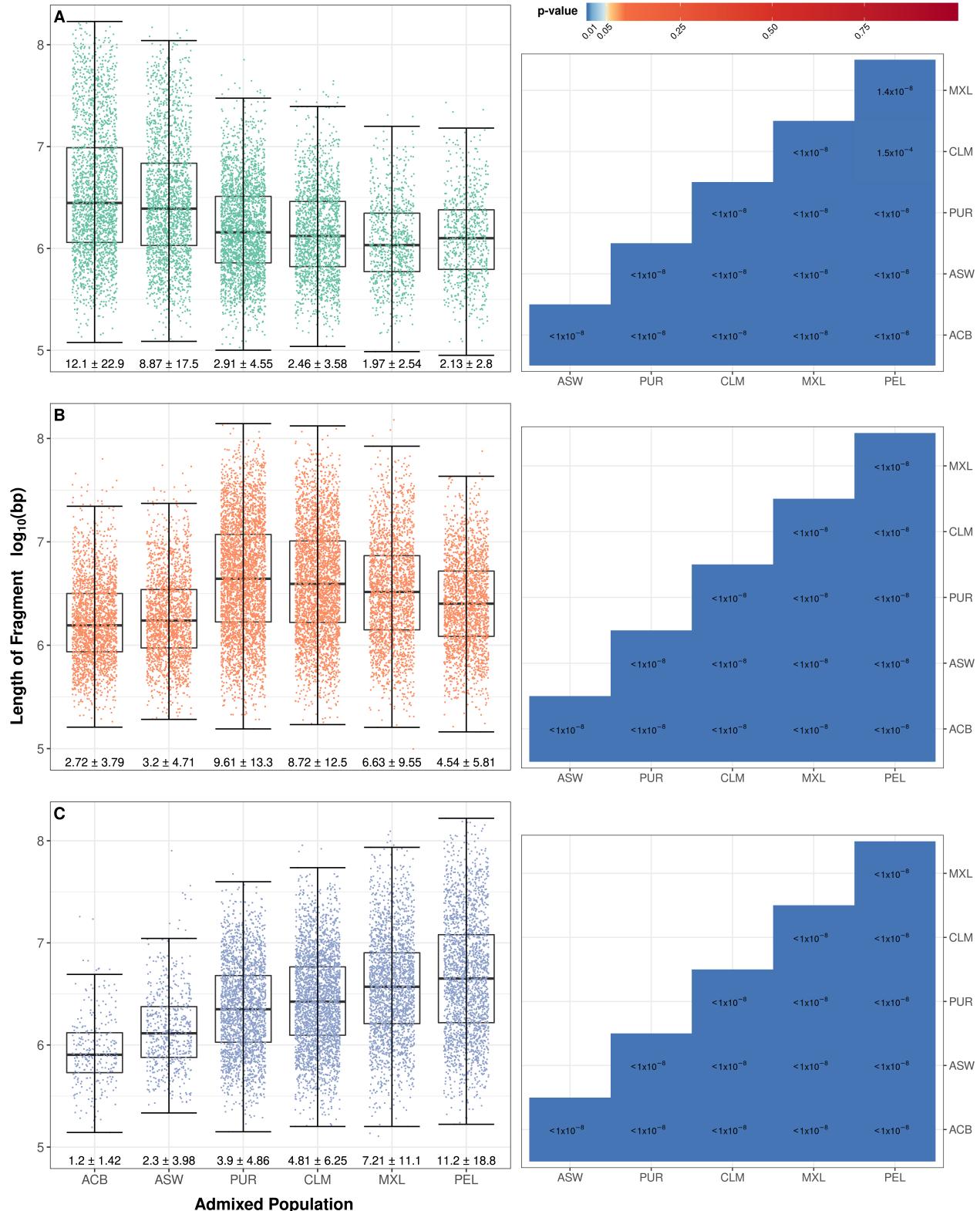


Figure S5: Comparative box plots displaying the distributions of continuous ancestry tract lengths of each ancestry for all individuals of each admixed population, with corresponding p-value heatmaps comparing populations statistically. Fragment length, that is the number of consecutive haplotype assignments 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.

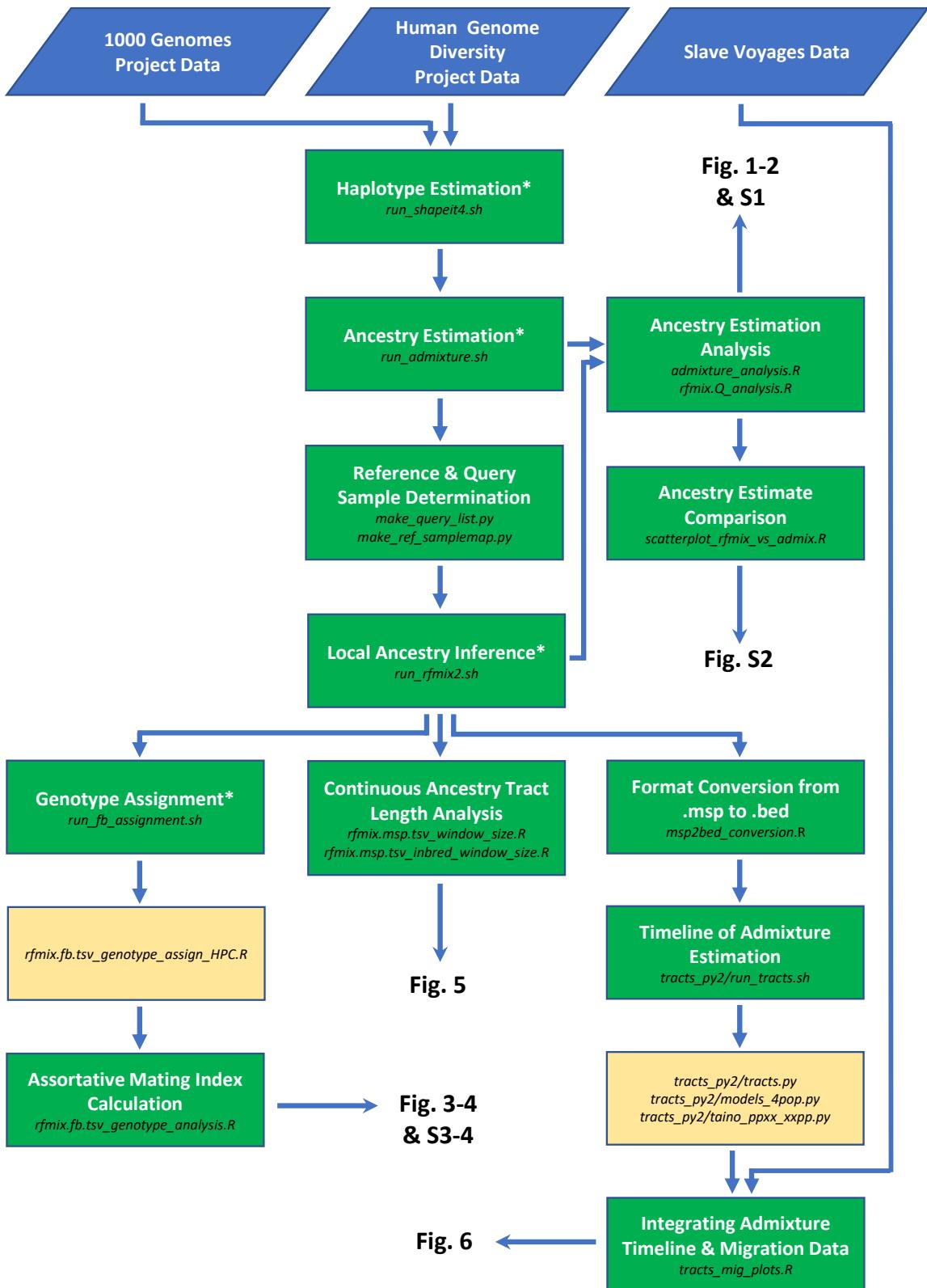


Figure S6: Flowchart representing the analysis workflow of the project, from input data to the output figures. Arrows indicate that the output from one step is the input for the next. Below the label of each step is the script(s) from the provided github repository required to run that step. The scripts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