Chapter 1

# Investigating the sub-continental ancestry of ethnic minorities within the U.K. Biobank from sparse genotype data

## 1.1 Introduction

From a genetic standpoint, the British population is one of the most studied in the world, with many studies sequencing or genotyping individuals from across the U.K (e.g. [1–4]). These projects have been primarily aimed at researching the genetic basis of disease, but have also been used to investigate population history, substructure and the relationship of different sub-populations in the U.K. to other European countries [2, 5, 6].

The U.K. is also an ethnically diverse country, with 13.8% of individuals belonging to ethnic minority groups (source: ONS survey). Groups of people from across the world have migrated to the U.K. at different periods in the previous three centuries, driven by the legacy of colonialism [7], the transatlantic Slave Trade and a variety of other reasons. Despite this, the roughly 9 million ethnic minorities within the U.K. remain relatively understudied in the context of genetics. For example, every one of the 27 papers in the GWAS catalogue with "U.K. Biobank" in the title, and two others presently in the catalog curation queue, limited their analyses to subgroups described in various terms as "White British", "British", "European", "White European", "Caucasian" or "White" [8]. The primary reason for this is reasonable concerns over the confounding effect of population substructure within a cohort [9]; retaining a more genetically homogeneous cohort is one strategy to mitigate this.

However, removing ethnic minorities from GWAS analyses is problematic, as evidence is mounting that the results from GWAS, including Polygenic Risk Scores (PRS), may not be transferrable to other populations if they have been conducted in cohorts of exclusively European individuals [10–12]. The reasons for this are not yet fully understood, but it thought that differences in LD structure may be at least partially responsible [13]. Ethnic minorities may therefore miss out on the advances in healthcare driven by large-scale genomic projects.

Understanding, and correcting for, population structure is an important step towards including a diversity of ancestries in GWAS. Several recent studies have shown the power of methods which explicitly model linkage between neighbouring markers when controlling for population structure, relative to traditional approaches such as PCA. Zaidi and Mathieson (2020) [14] showed that whilst it is not possible to correct for recent population stratification using principal components of common variants, correcting using a matrix of pairwise IBD sharing is effective. Similarly, it has been shown (S.Hu, personal communication of unpublished data) that incorporating principle components did not eliminate significant associations between genetic variants and birth location in UK Biobank partcipants. However these significant hits dissappeared when using a ChromoPainter coancestry matrix, here generated by painting target samples against a set of reference individuals and using the resulting painting profile as covariates in the association test. Byrne et al also eliminated significant associations with birth place in a cohort of Dutch individuals, by painting samples using PBWT matching [15].

Other recent studies have leveraged advances in algorithm development, such as the positional Burrows-Wheeler transform, to perform haplotype-based analyses on Biobank-scale datasets. Saada et al (2020) detected around 214 billion IBD segments across 487,409 individuals in the U.K. Biobank, obtaining enough information to estimate birth location to within 45 km, demonstrating the power of haplotype-based approaches on large datasets. However, their method only estimated pairwise IBD between individuals rather than comparing each individual to *all* other individuals in the dataset. The latter approach is more powerful at detecting recent shared because it finds who an individual shares ancestry with overall [16]. Additionally, Saada et al only considered self-identified White British individuals. Zhou et al (2020) recovered a similar number of IBD segments within the U.K. Biobank (231.5bn), also using a PBWT-based method [17].

Recent studies have outlined the power of haplotype-based approaches in inferring the population histories of different African ethnic groups [18–20]. Therefore, it seems natural to extend the approaches of Saada et al and Byrne et al to exploring the ancestry and structure

of individuals of recent African ancestry in the U.K. Biobank as a first step to including a wider diversity of ethnicities in association studies.

Additionally, but no less importantly, there is intrinsic value in exploring the ancestry of individuals (ethnic minorities in the U.K.) who have typically been excluded from analyses. Excluding individuals based upon their ethnicity presents ethical issues; individuals

To achieve both of these aims, I will leverage the a recently compiled dataset, hereafter referred to as 'Human Origins'. At the time of writing, it is the most detailed dataset of genotype data from African individuals in terms of the number of ethnolinguist groups represented. Whilst the dataset contains individuals from across Africa, it contains particularly large numbers of individuals from South Africa (n=104), Cameroon (n=567) and Ghana (n=211), which are countries known to have contributed immigrants to the U.K. Of the 5998 samples in the Human Origins dataset, 1,518 are previously unpublished, including all samples and 188 populations from Sudan, Nigeria, Ghana and The Congo. Therefore, this dataset is ideal for use as a reference panel to investigate the ancestry of ethnic minorities within the U.K. Biobank. In particular, given our newly acquired data comes from parts of west Africa that may well represent sources of African ancestry among UK minority groups, I chose to investigate individuals with recent African ancestry. However, these results should in theory be equally applicable to other non-European populations, such as those from east and south Asia.

However, one potential issue is that only 70,776 SNPs overlap between the U.K. Biobank and Human Origins genotyping arrays. This is much lower than the number used in a typical ChromoPainter analysis, which is usually betweeen 500,000 and 700,000. Using a low number of SNPs in the analysis may reduce the power to infer accurate ancestry proportions, in particular for haplotype-based methods since haplotype information depends on SNP density, as shown in Chapter 2. Therefore, one option is to impute the non-overlapping SNPs using a reference panel. However, the effect of imputation on ChromoPainter-style analyses has yet to be fully investigated. It is possible that imputing a large number of positions may introduce biases, particularly towards populations which are present in the reference panel. Studies have shown repeatedly that genotypes in non-European individuals are imputed less accurately compared to European individuals when using a primarily European reference panel [21, 22]. Accordingly, we can ask whether it is preferable to retain a smaller number of non-imputed SNPs or a larger number SNPs, some of which have been imputed. My work in Chapter 2 showed that imputation introduced bias towards European populations prevalent in the reference panel; in this chapter, I will extend that analysis to determine the effect of imputation on population assignment in African ethnic groups.

This chapter will focus on two questions. Firstly, I will evaluate the effect of using imputed genotypes on the validity of ChromoPainter analysis in African individuals, similar to analyses I performed in Chapter 2 but tailored to my U.K. Biobank analysis. Secondly, I will compare genetic variation patterns of U.K. Biobank participants with recent African ancestry to the Human Origins dataset populations, in order to shed light on their ancestral origins.

## 1.2 Methods

### 1.2.1 U.K. Biobank data access and initial processing

The U.K. Biobank dataset contains extensive phenotype data for 488,378 individuals and 6994 phenotypic measurements at the time of writing (`https://www.U.K.biobank.ac.U.K./`). Access was obtained to study the U.K. Biobank dataset via UCL Genetics Institute (ref number 51119, principal investigator = D.Curtis).

I obtained the U.K. Biobank genotype data, consisting of 488,377 individuals genotyped at 784,256 genome-wide SNPs on the U.K. Biobank Axiom Array. I will hereafter refer to these data as the 'non-imputed' data, as all SNPs were directly genotyped with imputation. I used plink2 [23] to convert the binary plink files to `.bcf` format.

I also obtained U.K. Biobank data, which had already been imputed to approximately 96m SNPs using the combined references of the Haplotype Reference Consortium (HRC) and UK10K haplotype resource. I will hereafter refer to these data as the 'imputed' data. Full details of imputation can be found in the paper of McCarthy et al (2016) [24]. The imputed data was downloaded and converted from `.bgen` to `.bcf` format using qctool2 (`https://www.well.ox.ac.U.K./~gav/qctool_v2/`).

I therefore had two separate datasets; 'imputed' and 'non-imputed', containing the same individuals and differing only in whether or not imputation had been used to increase the total number of SNPs.

### 1.2.2 ADMIXTURE analysis

I am primarily interested in using ChromoPainter [25] to explore the ancestry of ethnic minorities in the U.K. Biobank. However performing ChromoPainter analysis on the entire U.K. Biobank dataset (n=488,377 individuals) is computationally infeasible. Thus, I chose to analyse only those individuals with more than 50% non-European ancestry. ADMIXTURE is a fast and accurate way to estimate continental-scale ancestry proportions [26] and is

therefore ideal for this task.

I LD-pruned the non-imputed U.K. Biobank dataset using using `plink -indep-pairwise 50 10 0.02` [23]. This left a total of 70,776 bi-allelic SNPs. I then subsetted the 1000 Genomes dataset down to the 70,776 SNPs retained in the U.K. Biobank dataset and merged the two datasets using `bcftools -merge`. Thus, I had a dataset containing all U.K. Biobank and 1000 Genomes individuals, genotyped at 70,776 SNPs.

I ran ADMIXTURE in supervised mode using the argument `-supervised` and fixed the four reference populations as GBR British, Nigeria Yoruba, Han Chinese and Gujarati Indian from the 1000 Genomes dataset. These populations were chosen as they represent a broad division of worldwide populations into African, European, East Asian and South Asian; for the purposes of this particular analysis, it was not necessary to include finer-scale populations. The rest of the arguments were left to default. I used the resulting `.Q` files to determine the ancestry proportions of each reference population in each U.K. Biobank individual.

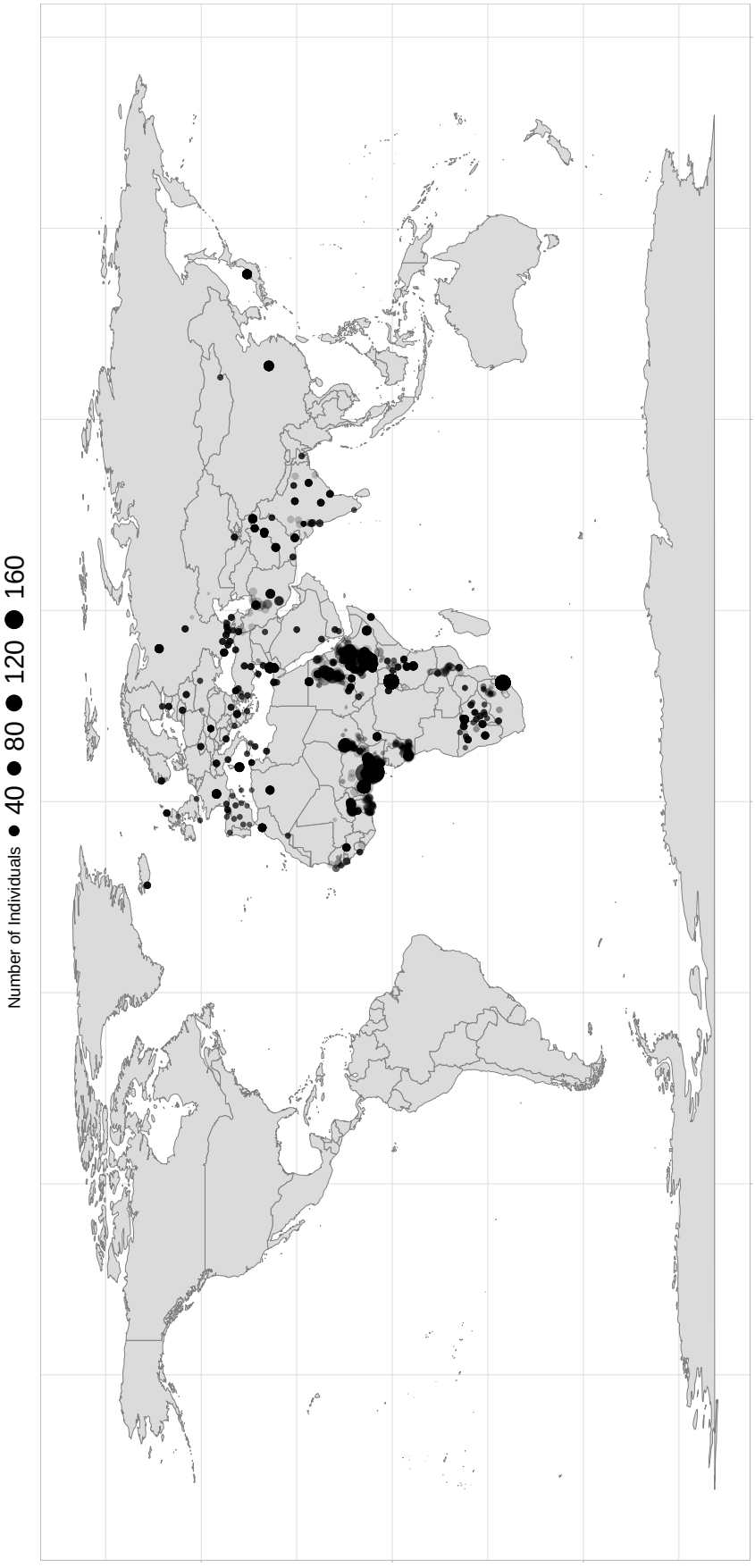
Individuals with at least 50% ancestry from Nigeria Yoruba were carried into later analysis; I refer to these as 'selected' Biobank individuals.

### 1.2.3 Data preparation - Human Origins

To determine the ancestry of U.K. Biobank individuals, I compared their SNP patterns to populations/ethnic groups from different parts of the world to infer which populations they share recent ancestry with. As I am particularly interested in studying individuals with recent African ancestry, I used the so-called "Human Origins" reference dataset (appendix A.20) for this purpose, as it contains individuals from 349 different ethnic groups from across Africa and 535 world-wide groups in total (Fig. 1.1). Full details of processing can be found in Appendix A.20 (**??**).

### 1.2.4 Data merge - non-imputed data and Human Origins

I used `bcftools -merge` to merge 5,998 reference "Human Origins dataset" individuals with 8,476 UK Biobank participants that had ≥50% African ancestry, using the gt-conform utilty from Beagle (`https://faculty.washington.edu/browning/conform-gt.html`) to remove any inconsistent positions. This dataset contained 65,749 non-imputed SNPs that overlap between the Human Origins and UK Biobank arrays. I phased these data with shapeit4 [21] using `-pbwt-depth 8`, the b37 genetic map and otherwise default parameters.


Number of Individuals
40
80
120
160

### 1.2.5 Data preparation - imputed data

I similarly merged the imputed UK Biobank data with the Human Origins reference dataset at 525,566 SNPs that were genotyped in Human Origins, and phased these data with shapeit4, using the same settings as for the non-imputed data.

### 1.2.6 Chromopainter

For both of the imputed and non-imputed datasets, I used CHROMOPAINTER to infer the proportion of genome-wide DNA that each UK Biobank and Human Origins reference individual matches to individuals from each Human Origins reference population. Using this CHROMOPAINTER output, I then used SOURCEFINDv2 [27] to infer the proportion of ancestry that each UK Biobank individual shares most recently with each of the 553 Human Origin reference populations.

An alternative option to using Origins would be to use PBWT (positional Burrows-Wheeler transform) paint `https://github.com/richarddurbin/pbwt/blob/master/pbwtPaint.c`), a fast approximation to ChromoPainter which provides approximately the same output and is scalable to large sample sizes [15]. However, it is not possible to provide a reference panel and each haplotype must be compared to all others in turn. This would be much less efficient and would not allow me to take full advantage of the Human Origins dataset.

### 1.2.7 SOURCEFIND

I estimated ancestry proportions for each of the selected U.K. Biobank individuals using SOURCEFINDv2 [27]. I used the combined painting from the section above. I analysed each U.K. Biobank individual with more than 50% African ancestry separately, using all Human Origins populations as surrogates. I left all parameters as default.

### 1.2.8 Imputation bias test

The imputed U.K. Biobank dataset was imputed using a reference panel containing the Haplotype Reference Consortium. Whilst this reference panel contains many European populations, it contains relatively few from Africa. Imputing variants in non-European individuals using a reference panel that is primarily composed of European individuals may lead to biased or inaccurate imputation [28]. Given I am particularly interested in analysing individuals with recent African ancestry in the U.K. Biobank, it is important to determine whether this is the case.

An obvious way to test this would be to compare a painting on the **U.K. Biobank** individuals using datasets comprised of a majority imputed and non-imputed SNPs. However, this is not possible; the samples in the U.K. Biobank dataset do not have any associated population or ethnic group labels beyond broad self-identified categories. Accordingly, it would not be possible to mask their ethnic group and attempt to guess it using only the genetic data, an approach which I use for the Human Origins data in this chapter.

Therefore, I used the Human Origins dataset, where I could control whether or not SNPs are imputed and mask population labels. I submitted the full Human Origins reference dataset (5998 individuals and 560,420 SNPs) to the Sanger Imputation Server (`https://imputation.sanger.ac.U.K./`), which uses the full Haplotype Reference Consortium (HRC) as a reference panel for imputation. This reference panel was chosen because it was the same one used for imputing the U.K. Biobank individuals.

I subsetted the imputed Human Origins dataset down to SNPs present in the U.K. Biobank array, leaving 727,325 positions present in the imputed Human Origins dataset and then randomly removed SNPs until 500,000 remained. Although the number of SNPs still differ, my previous research in Chapter 2 shows that increasing the number of SNPs beyond 400,000 does not affect the ability to correctly assign individuals to populations. I phased the imputed and non-imputed datasets separately using shapeitv4 at default settings.

To answer these questions, and therefore determine whether using the imputed or the 70,000 SNP Human Origins dataset is better in this scenario, I performed a painting using (i) the full 560,442 genotyped SNPs, (ii) 64,762 genotyped SNPs overlapping UK Biobank, and (iii) 500,000 SNPs that include the 70,000 genotyped SNPs [SO I HAVE MIS-UNDERSTOOD WHAT YOU DID HERE. I THOUGHT YOU HAD REDUCED HO TO 70K SNPS PRIOR TO SUBMITTING TO HRC. BUT YOU'RE SAYING YOU INSTEAD SUBMITTED THE WHOLE DATASET (ANALOGOUS TO WHAT YOU DID WITH UKBIOBANK) BUT THEN REPLACED SOME OF THE GENOTYPED SNPS WITH IMPUTED ONES? SO THIS WOULD MEAN THAT THIS ISN'T AS "HARSH" AS I THOUGHT IN MY EARLIER COMMENTS. IT STILL HAS AN ISSUE WHEREBY THE REFERENCE DATA ALSO CONTAINS IMPUTED SNPS, WHILE THAT'S NOT TRUE WITH THE U.K. BIOBANK ANALYSIS? CAN YOU MAKE THIS CLEAR, AND BRIEFLY DISCUSS POTENTIAL RAMIFICATIONS?]and 430,000 SNPs imputed using the HRC reference. I performed painting (ii) in both linked and unlinked mode to determine whether there is haplotype information using 70,000 SNPs.

For each of the three datasets described above, I selected all ethnic groups from Nigeria,

Cameroon and Ghana which had five or more individuals (n=51 populations, n=1203 individuals) and split each population randomly in half, into 'donors' and 'recipients'. I painted all recipient populations (n=51) using all donor populations (n=51) using a leave-one-out approach (motivation for this approach given in appendix B.2). I tested the information content of each painting by counting how often individuals copy more from individuals in their own populations than individuals from other populations. I also counted the number of times a population had the lowest TVD (motivation and description of TVD given in appendix X) with its own population (Table 1.1).

## 1.3 Results

### 1.3.1 4% of U.K. Biobank individuals have at least 50% non-European ancestry

Performing ChromoPainter analysis on the 488,378 individuals in the U.K. Biobank would be computationally unfeasible; therefore I first performed supervised ADMIXTURE on all U.K. Biobank individuals. In order to identify individuals with at least 50% African ancestry, I set $K = 4$ supervision clusters that were defined using European (CEU), Gujarati, Han Chinese and Yoruban reference individuals from the 1000 genomes dataset. I then carried forward individuals with more than 50% ancestry from Yoruba to later ChromoPainter analyses.

In total, there were 8476, 2653, 9171 individuals with at least 50% inferred ancestry related to Yoruba, Han Chinese and Gujarati reference populations respectively, corresponding to 4.16% of the total U.K. Biobank individuals. Although I use these population labels for convenience, I note that an individual with e.g. 50% 'Han Chinese' ancestry does not necessarily derive 50% of their ancestry from the Han Chinese population, but that 50% of their ancestry most closely matches Han China relative to the other reference populations. Thus, a Japanese individual may be modeled as 100% Han Chinese whilst not being Han Chinese in an ethnic sense. Similarly, for brevity, I will refer to individuals who have more than 50% of their ancestry from Yoruba as being 'African' Biobank individuals, whilst acknowledging that 'African' as a broad label encompasses a large diversity of ancestries and ethnicities.

I validated the ADMIXTURE results to ensure that there was not any mixing of sample labels and that enough ADMIXTURE EM iterations had been performed. To do this, I selected all individuals who self-identified as being either "Caribbean", "African" or "Black or Black British" (n=7527) and plotted the distribution of ADMIXTURE ancestry proportions, under the assumption that these individuals should contain more African than other kinds of

ancestry. On average this was the case, with the mean proportion of African ancestry among these individuals being 0.88 (Fig. 1.2), compared to 11 % British, 0.22% Han Chinese and 0.19% Gujarati.

However, there was substantial variation in the ancestry proportions for those who self-identified as being either "Caribbean", "African" or "Black or Black British". Proportions of Yoruban and British ancestry ranged from 0 to 1, Han Chinese from 0 to 0.53 and Gujarati from 0 to 0.759, reflecting the diverse array of genetic ancestries that can fall under a given ethnic label. This follows from previous research which has shown self-reported ethnicity can be an unreliable proxy for genetic ancestry [29,30]. This suggests that relying on self-reported ethnicity may yield variable results when e.g. used as a covariate in a GWAS. For example, there were 48 people who self identified as being either "Caribbean", "African" or "Black or Black British", but had less than 1% African ancestry.

### 1.3.2 To impute or not?

In order to use ChromoPainter and the Human Origins dataset as a reference to infer fine-scale ancestry in U.K. Biobank individuals, the datasets must be merged. The overlap of SNPs genotyped in each dataset is only 70,776 SNPS, or an average of ≈1 SNP per 40Kb. Given linkage disequilibrium (e.g. as measured by Pearson's correlation) between pairs of SNPs decays to background levels by 100Kb within most populations [31], analysing 70,000 SNPs may substantially decrease any potential power gains from modeling haplotypes to detect fine-scale differences between populations. [HAVE YOU DESCRIBED THIS PREVIOUSLY? I.E. DO YOU NEED TO DESCRIBE HOW YOU IMPUTED UK BIOBANK USING HRC, AND THEN REDUCED TO THE SNPS OVERLAPPING UK BIOBANK? ALSO NOTE HOW MANY GENOTYPED SNPS YOU HAD IN UK BIOBANK FOR USE IN THE IMPUTATION.]In contrast, the imputed U.K. Biobank dataset has 535,544 SNPs in total, all of which are genotyped in the Human Origins reference dataset and 87.7% of which are imputed in UK Biobank individuals. While this may boost power over using only 70,000 SNPs, including a high percentage of imputed SNPs may bias ancestry inference. Therefore, I needed to determine a) whether there is a loss of power when 70,000 SNPs relative to the a full 500,000 SNP dataset and b) whether there is bias when using a dataset which contains a majority of imputed SNPs.

To answer these questions, [THINK YOU NEED TO ADD MORE DETAILS HERE, LIKE "I returned to the imputed and unimputed Human Origins datasets I describe in Section 2.xx. Recall here I reduced the Human Origins dataset to 70K SNPs and then imputed to XX SNPs using HRC..." THEN YOU LIKELY ALSO WANT TO REMIND

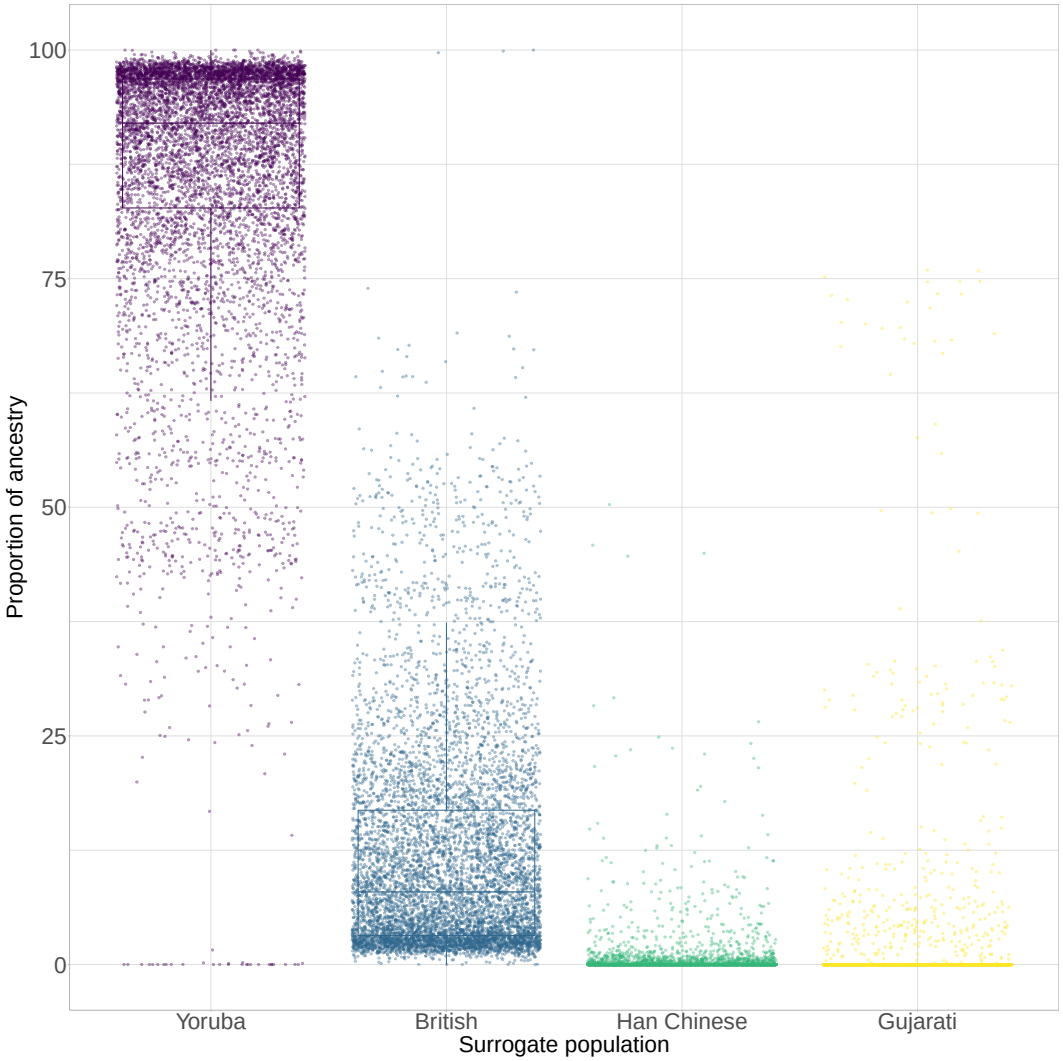



**Figure 1.2:** Ancestry proportions inferred from supervised Admixture run (k=4) for all individuals who self identified as being either "Caribbean", "African" or "Black or Black British". Points within each column are given random jitter to improve visual clarity.

THAT THIS IS A WORSE THAN THE UK BIOBANK SCENARIO, SINCE THEY WERE IMPUTED USING FAR MORE THAN 70K SNPS]and therefore determine whether using the imputed or the 70,000 SNP Human Origins dataset is better in this scenario, I performed a painting using (i) the full 560,442 genotyped SNPs, (ii) only the 64,762 genotyped SNPs overlapping UK Biobank, and (iii) 500,000 SNPs that include the [IS IT A FULL 70K, RATHER THAN THE 64.47K? FINE EITHER WAY, JUST WANT TO MAKE SURE THIS IS THE PRECISE NUMBER]70,000 genotyped SNPs and 430,000 SNPs imputed using the HRC reference. I performed painting (ii) in both linked and unlinked mode to determine whether there is any haplotype information using 70,000 SNPs.

For each of the three datasets described above, I selected all ethnic groups from Nigeria,

| painting | TVD | copying |
|---|---|---|
| 70K (linked) | 44% | 24% |
| 70K (unlinked) | 20% | 17% |
| imputed (linked) | 14% | 14% |
| full (linked) | 38% | 23% |

**Table 1.1:** Percentage of populations which had lowest TVD (TVD) or copied the most (copying) from their own population under different paintings. 70K linked used 70,000 SNPs in linked mode, 70K used 70,000 SNPs in unlinked mode, imputed used 430,000 imputed and 70,000 non-imputed SNPs in linked mode and full used 500,000 non-imputed SNPs in linked mode.

Cameroon and Ghana which had five or more individuals (n=51 populations, n=1203 individuals) and split each population randomly in half, into 'donors' and 'recipients'. I painted all recipient populations (n=51) using all donor populations (n=51) using a leave-one-out approach (description and motivation of this approach given in appendix B.2). I only considered populations of five individuals or more because any fewer individuals would likely result in very weak power to assign individuals to that population. I tested the information content of each painting by counting how often individuals copy more from individuals in their own populations than individuals from other populations. I also counted the number of times a population had the lowest TVD (motivation and description of TVD given in appendix X) with its own population (Table 1.1).

Populations in the 70,000 non-imputed painting matched more to and had a lower $TVD$ with their own population than the 500,000 non-imputed painting. Whilst it seems counter-intuitive[AGREE THIS IS very COUNTER-INTUITIVE! FOR TABLE 3.1, WHAT DO YOU MEAN BY "population which had lowest TVD"? SO THIS ISN'T IDNIVIDUAL ASSIGNMENT, BUT INSTEAD YOU AVERAGE COPY VECTORS ACROSS INDIVIDUALS? YOU NEED TO MAKE CLEAR WHAT YOU DID HERE. MAYBE YOU CAN ADD ANOTHER COLUMN THAT IS THE "SUMMED TVD" ACROSS ALL 51 POPULATIONS, SHOWING THAT (HOPEFULLY) THESE VALUES ARE QUITE CLOSE BETWEEN 70K AND FULL LINKED] that there is more power using a smaller number of SNPS, this is broadly consistent with my previous findings in Chapter 2, which showed that metrics of painting information plateau [REFERENCE A FIGURE OR TABLE HERE, I.E. WHICH SCENARIO DO YOU MEAN?]after approximately 50,000 SNPs (i.e. there is no clear benefit to using more than 50,000 SNPs in terms of assinging individuals to a population). This is reassuring and suggests there is no loss of power when using the 70,000 SNP set. These data also shows that there is a fairly dramatic loss of power when using imputed data relative to non-imputed data, as over 3x the number of populations had a lower TVD with their own population when using imputed compared to non-imputed data.

Given the above results suggested that imputing data results in a loss of information, I was interested in whether this constituted a 'bias' towards certain populations. Imputation methods rely on identifying reference haplotypes which are closest to the target haplotypes. However, if the ethnic groups that the target individuals derive ancestry from are not present in the imputation reference panel, missing variants are imputed from populations in the reference panel which are most closely related to the target samples. In this case, two target populations may be imputed to appear more genetically similar to that reference population, reducing the differentiation between them. [AGAIN CAN YOU REFER BACK TO CHAP 2, WHERE YOU HAVE SEEN THIS?]In theory, this artificial similarity would be propagated through to the ChromoPainter analysis. In particular, we would expect populations present in the reference panel to donate more to all other individuals than they would if no imputation had taken place.

For example, in the case of the Haplotype Reference Consortium, the closest reference population to two African target samples from e.g. Cameroon may be the Yoruba from Nigeria, which is the [ISN'T THERE ALSO GAMBIANS? PERHAPS OTHERS OUTSIDE OF 1KGP? ALSO NIGERIAN ESAN, I THINK? AND SIERRA LEONE?]only west African group in the reference. These samples would appear more similar to the Yoruba ethnic group than if they had not been imputed. In a ChromoPainter analysis, the Yoruba donor population would donate more than than when using non-imputed SNPs.

Comparing the imputed and non-imputed coancestry matrices revealed biases consistent with the above expectation. If the coancestry matrix columns are combined into populations, then the sum of each column gives the total length of genome that population contributes to all recipient individuals in the dataset. Therefore, comparing the column sums between the imputed and non-imputed matrices informs us about which populations contribute more when using imputed compared to non-imputed SNPs. Fig 1.3 shows the amount of differential haplotype donation on a per-population basis, with populations highlighted based on their presence or absence in the 1000 genomes dataset. It is clear that populations present in the 1000 genomes are primarily clustered towards the right hand side, rather than randomly distributed across figure. This strongly suggests that imputation causes a bias towards those populations present in a reference panel.

To formally test whether the ordering of populations was likely significantly different to the ordering expected under the null model of no impact of being present in the 1000 genomes dataset, I performed a non-parametric permutation test. If we order the populations based on their differential haplotype donation and assign a rank value to each population, we can calculate the sum, $S$ of the ranks values of all populations present in the 1000 genomes[IS

THIS THE WILCOXON RANK SUM TEST?]. If the 1000 genomes populations are clustered at the higher end of the ordering, we would expect the value of $S$ to be smaller than if the populations are randomly distributed across the ordering. I performed 100,000 replications of randomly ordering the population labels and calculating the value of $S$. Of the 100,000, 26 had $S$ greater than the true empirical value calculated from the data, showing the ordering of the populations is unlikely to be due to chance ($p = 0.00026$).

Put together, these results suggest that using imputed data would introduce a level of bias and loss of information. Therefore, in all later analysis, I chose to use the approximately 70,000 non-imputed SNPs which overlap between the Human Origins and U.K. Biobank datasets.

### 1.3.3 African ancestry in the U.K. Biobank samples is concentrated in Ghana and Nigeria

Using ≈70,000 directly genotyped SNPS, I painted all U.K. Biobank individuals with at least 50% African ancestry (n=8475) using all Human Origins individuals as donors (n=5577).

Principal component analysis on the resulting chunkcounts coancestry matrix reveals the general structure of the selected individuals, alongside the reference populations (Fig. 1.4). Three clines are present; one of similarity to Southern African populations typified by the Zulu ethnic group from South Africa, one of similarity to West African populations such as Yoruba and Cameroon_Dii, and the last to East African populations such as those from Ethiopia. The majority of U.K. Biobank individuals are positioned near West African populations; in particular between Yoruba and Cameroon_Arabe. The presence of a broad cluster of West African individuals is consistent with prior expectations that West African ancestry should be prevalent in a sample of British individuals, due to the history of migration from this region [32]. A second cluster of UK Biobank individuals is located along the Southern African cline, close to the Bantu_SA label.

Aggregating the columns of the co-ancestry matrix by reference population and taking the sum of each column gives the total length of genome for which a U.K. Biobank individual shares recent ancestry with individuals from that donor population. This can be visualised on a map, where each point represents a reference population and the colour corresponds to the total amount that reference population contributes towards the ancestry of all retained U.K. Biobank individuals (Fig. 1.5). Higher values correspond to more ancestry from that population in the U.K. Biobank sample. However, it should be noted that raw ChromoPainter output can be influenced strongly by sample size and so the values shown in Fig. 1.5 should

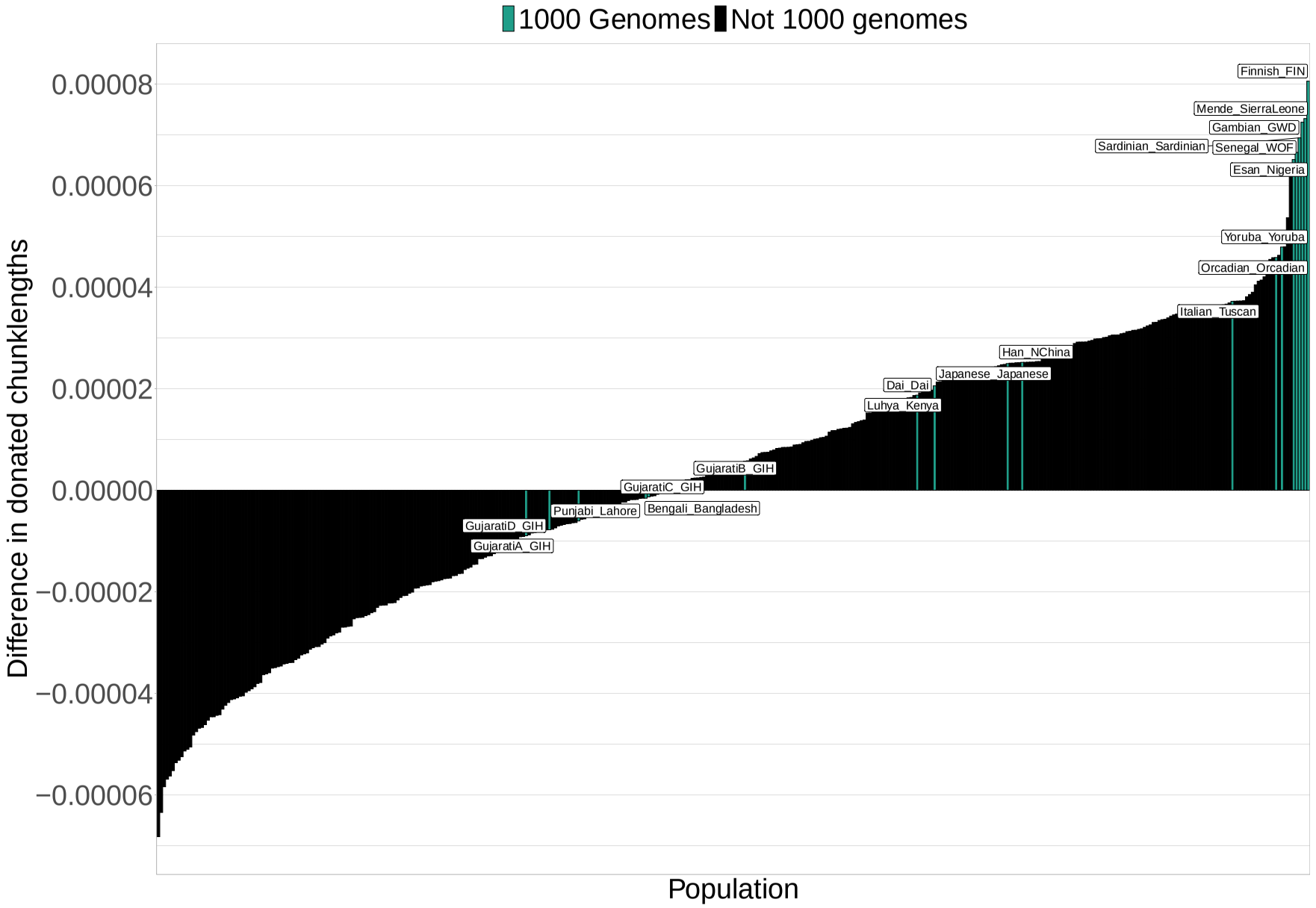



**Figure 1.3:** Differences in the amount donated by populations when using imputed and non-imputed data. Each vertical bar corresponds to a single population (N=395), with the height of the bar corresponding to the difference in haplotype donation, with positive values indicating increased donation and negative values indicating reduced donation. Bars coloured in green are also present in the 1000 genomes imputation panel and black bars are not present in the 1000 genomes.


**Figure 1.4:** Principle component analysis of chunklengths matrix for U.K. Biobank individuals with $\geq$50% inferred recent African ancestry and human origins array. Individuals are coloured dependent on whether they are U.K. Biobank (green) or Human Origins (purple) samples. Labels indicate mean principle component coordinates for individuals in that population. A random sample of populations were chosen to have labels to prevent the figure from being too cluttered.

not be taken literally as an exact reflection of the ancestry distribution.

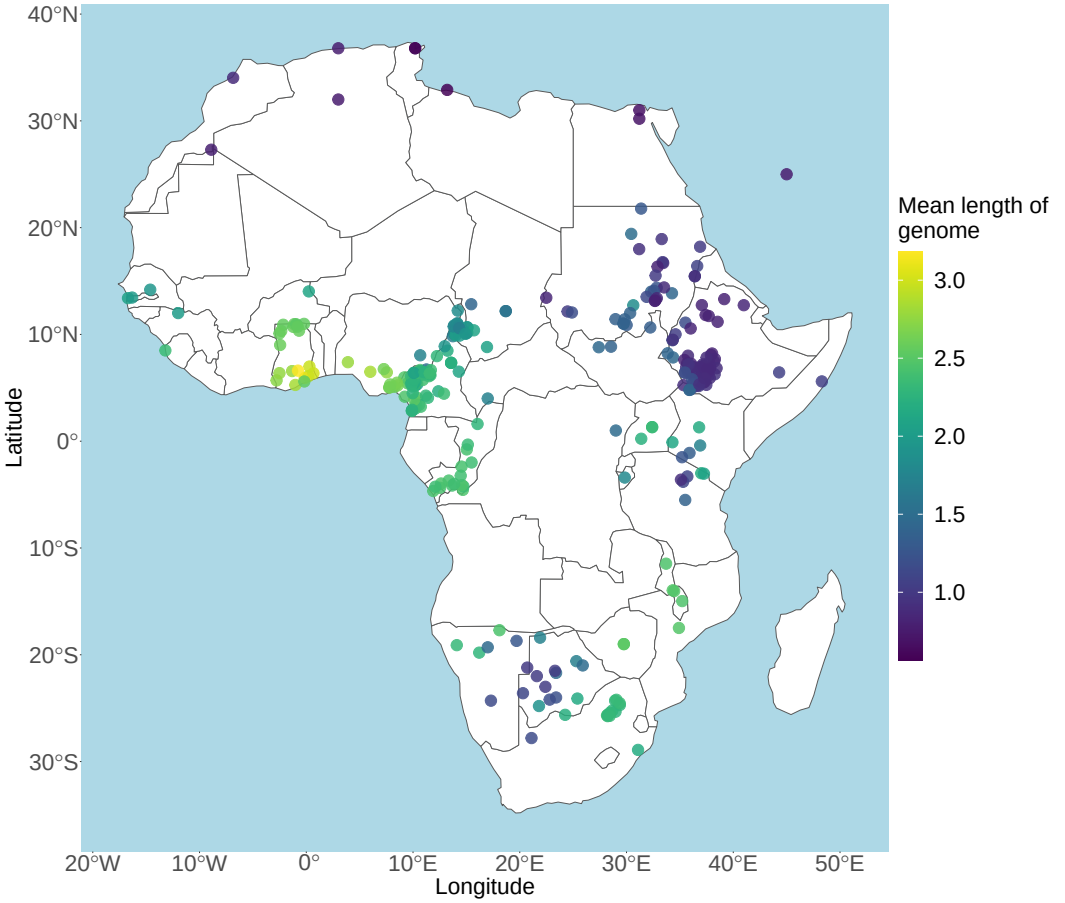



**Figure 1.5:** Map of haplotype donation to U.K. Biobank individuals. Each point represents a different African population. Colour corresponds to the mean length (cM) that populatation donated to all African U.K. Biobank individuals.

The map supports the findings from the PCA in Fig. 1.4; the populations with the largest contribution are those from West Africa (Fig. 1.5). In particular, populations from Ghana and Nigeria contribute the most to the ancestry of Biobank individuals. On the other hand, populations in East and North Africa contribute relatively little, with Southern / South-East Africa being approximately intermediate. This is consistent with two different historical events.

Firstly, it is known from historical and genetic studies that a majority of the individuals who were forcibly transported from Africa to the Americas during the transatlantic slave trade were from the west coast of Africa [33]. Given the U.K. Biobank sample contains many

individuals who were either born in, or trace their ancestry from the Caribbean, a region that had a large influx of slaves [34], we would expect there to be a large contribution of ancestry from West Africa. Secondly and more recently, there has been a relatively large amount of historical immigration from countries in West Africa, such as Ghana and Nigeria, to the U.K [32]. Although there are a number of immigrants from other parts of Africa, reflected in the nonzero contributions from other ethnic groups, these contributions are small compared to those from West Africa.

I performed the same visualisation using the painting using imputed SNPs and the ancestry distribution was qualitatively the same.

I used SOURCEFIND to infer the proportion of ancestry that each UK Biobank individual shares most recently with each of the 535 surrogate groups, as this accounts for uneven donor population sizes. A map of proportions is given in Fig. 1.6, with each point corresponding to the mean percentage of ancestry of that particular group across all African U.K. Biobank individuals. Similar to the copyvector map, the ancestry is focused around Nigeria and Ghana, with Yoruba (39.8%) and Ghana Fante (7.31%) having the highest mean proportions. The distribution of colour on this figure is focused around a smaller number of populations compared to Fig. 1.5. This is because SOURCEFIND attempts to narrow down the set of populations which most likely contribute towards the ancestry of a given individual and so appear 'cleaner' than raw ChromoPainter results.

Fig. 1.7 displays the 30 ethnic groups with the highest mean proportions of ancestry within the U.K. Biobank individuals, and the distribution of values within each group. Yoruba was a clear standout for the most represented population; 3604/8309 individuals had at least 50% Yoruba ancestry. This is compared to the next most common ancestry, Ghana_Fante, which had an average of 7.3% per person and 373/8309 individuals with at least 50% ancestry. It is not clear what the reason for the large amount of Yoruban ancestry relative to all other populations is. One possible answer may come from considering the birth country of the U.K. participants. Of all the individuals for which we have country of birth data for (n=6190), more of them were born in the Caribbean (n=2263) relative to any other country. This should not be surprising given the history of migration from the Caribbean to the U.K. Of the individuals born in the Caribbean, over half were assigned to the Yoruban ethnicity, a much higher proportion than any other country of birth. Therefore, one could tentatively explain the abundance of Yoruba ancestry as resulting from the transatlantic Slave Trade, where individuals from the Yoruba ethnic group were taken to the Caribbean at a higher frequency than other nearby ethnic groups in the Human Origins reference. This may be in part because Yoruba is the second largest ethnic group in Nigeria and individuals

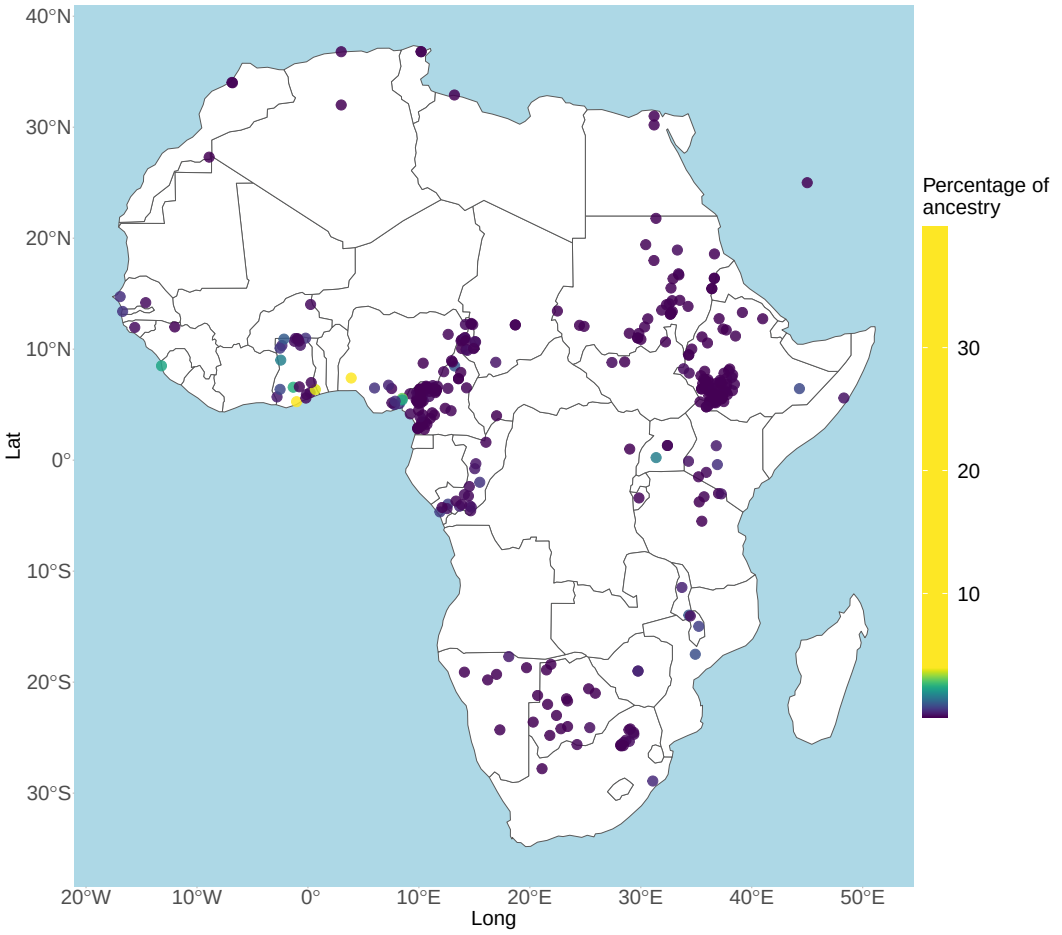



**Figure 1.6:** Map displaying the mean proportion of SOURCEFIND estimated ancestry of each African reference population within U.K. Biobank individuals. Each point is an African reference population with the colour corresponding to the mean ancestry proportion for that population across selected U.K. Biobank individuals. The colour-bar has been rescaled as two populations, Yoruba and Ghana_Fante have substantially higher proportions than all other populations.

belonging to it live primarily in coastal areas where the Slave Trade operated. The relatively large number of individuals from the Caribbean in the U.K. could thus have brought Yoruban ancestry to the U.K.

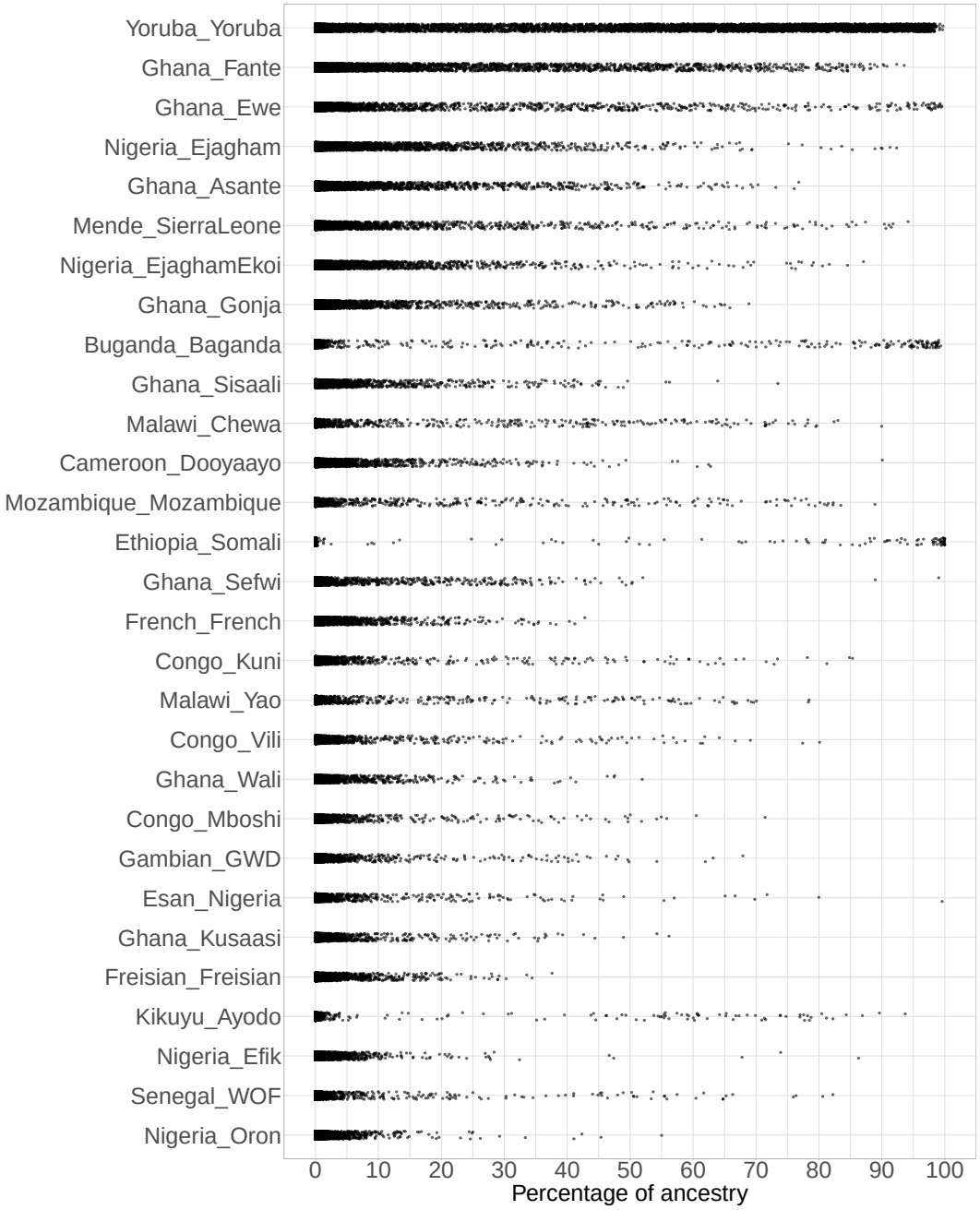



**Figure 1.7:** The 30 Human Origins populations which have the highest contribution to all U.K. Biobank individuals with at least 50% African Ancestry, based on SOURCEFIND analysis. Each row of points contains 8476 individuals and their position corresponds to the percentage of ancestry from that population.

There are other instances of an over and under-representation of one ethnic group from a particular country (Fig. 1.8). For example, Nigeria is dominated by a single ethnic group,

despite having data for 31 different ethnic groups. On the other hand, the individuals from Sudan are more evenly distributed across ethnicities. This may be caused because there are more reference ethnic groups in Sudan to assign individuals to. Further, it is known (personall communication N.Bird, 2021) that inability to distinguish individuals in closely related Sudanese populations.

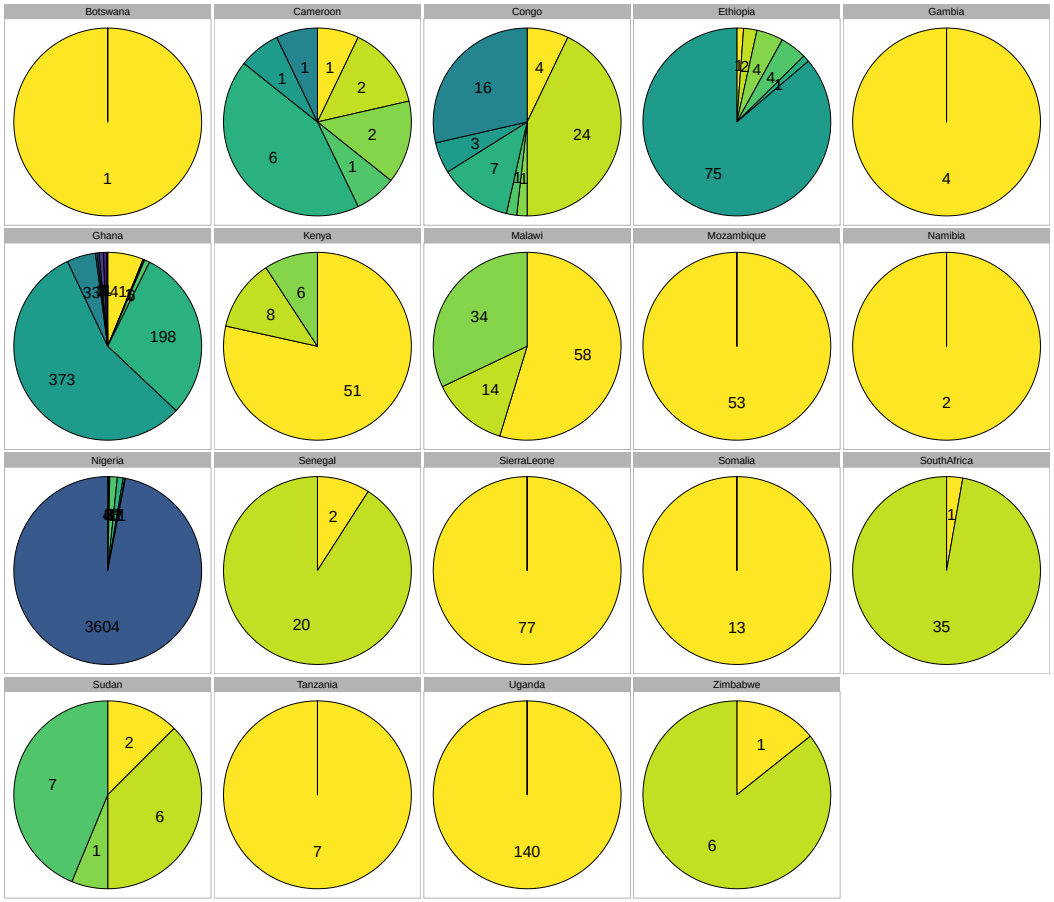



**Figure 1.8:** Variation of individuals assigned to different ethnic groups by country of assigned group. Each panel represents all individuals assigned to an ethnic group from that country, with proportions of each pie corresponding to proportion of individuals of that ethnic group in that country. Numbers within each slice correspond to total number of individuals within a given ethnic group.

Some other patterns can be noted. Whilst many individuals have intermediate levels of ancestry from West African populations (e.g. Ghana_Fante or Yoruba_Yoruba), much fewer individuals have intermediate levels of Ethiopia_Somali ancestry (Fig. 1.7). This may be because Somalis are more recent immigrants to the UK [cite] and therefore tend to be less admixed with Europeans relative to other immigrant populations which have been in

the U.K. longer and hence can be modeled as a mixture of almost entirely Ethiopia_Somali ancestry.

To test whether this was the case, I selected individuals assigned to either Ethiopia_Somali, Yoruba or Ghana_Fante and estimated their proportions of total African, European and Asian ancestry using SOURCEFIND. Individuals from Yoruba and Ghana_Fante had, on average, 6.2% and 5.2% European ancestry respectively, whereas individuals from Ethiopia_Somali had 0.21% on average, suggesting they are indeed less mixed than other populations, which is consistent with them being more recent migrants.

### 1.3.4 Verifying painting accuracy

Not all individuals within the U.K. Biobank were born in the U.K.; visualising the ancestry distribution of these individuals allows ensures us that the painting is accurate and may reveal insights into population history. For instance, the ancestry distribution of individuals born in the Caribbean may provide evidence for where in Africa slaves forcibly transported to the Caribbean during the transatlantic slave trade originated from. This is important, as disembarkation records from the Slave Trade are often sparse, meaning many people with African ancestry who currently live in the Americas may not have knowledge of where their ancestors originated from.

I subsetted the coancestry matrix to contain only U.K. Biobank individuals who provided data on birth location (n=6153/8472). We would expect that individuals who were born in a particular country would copy the most from reference populations from that country. For example, we would expect individuals who were born in South Africa to copy the most from sampled Bantu and Zulu ethnic groups from South Africa. This may not always be the case, as some ethnic groups have crossed borders in their history, or we may not have sampled representative groups from some countries, but it may broadly be expected to be true. We also have birth place data for individuals who were not born in Africa (e.g. the Caribbean and Brazil).

Fig. 1.9 shows the map of haplotype donation from reference groups to U.K. Biobank individuals born in South Africa. It is clear that reference populations from South Africa, in particular the Zulu ethnic group, contribute the most to these individuals. The pattern is qualitatively the same for all countries which had a reasonable number of donor populations, suggesting that the painting had good resolution down to at least the level of individual countries [OUGHT TO REFERENCE FIG 3.9 HERE?].

There are several interesting results. For example, there are 2,263 individuals who were

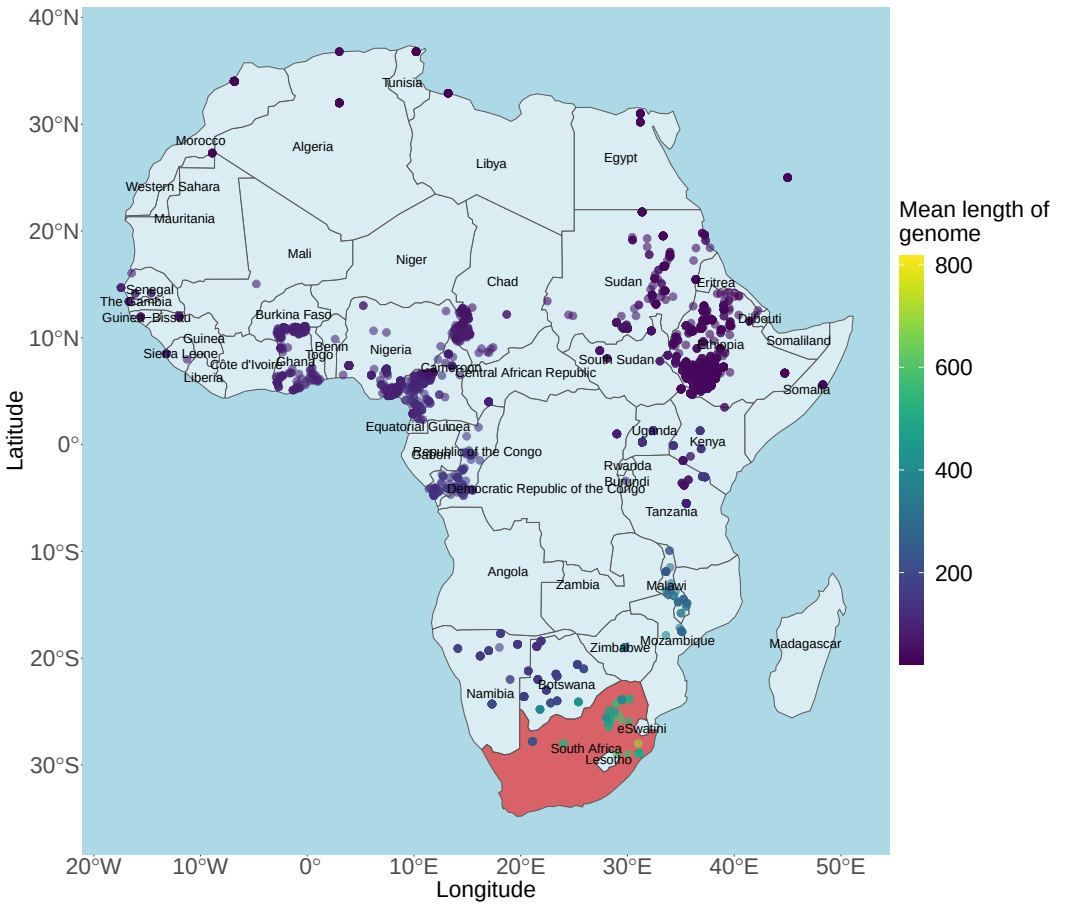



**Figure 1.9:** Map of haplotype donation to U.K. Biobank individuals born in South Africa. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in South Africa.

born in the Caribbean. Visualising the haplotype donation map for these individuals shows that they are primarily of West African ancestry (supplementary figure D3) , consistent with historical evidence [33]. Individuals born in Brazil have ancestry from further South, again consistent with historical evidence (supplementary figure D2). Of the nine individuals born in Brazil, six of them had a majority SOURCEFIND component from an ethnic group in The Republic of the Congo. However, it should be noted that there is a relatively small sample size from individuals born in Brazil (n=9), and that these individuals may not be representative of the Brazilian population as a whole.

As a formal test of the painting accuracy, I estimated SOURCEFIND ancestry proportions in each retained U.K. Biobank individual. An individual was 'assigned' to a particular ethnic

group if they had 75% or more of their total ancestry from that group. If the country the assigned reference population is from matches the birth location of the individual, then I considered that a 'success' and a 'fail' otherwise. Individuals who were born in the U.K. or who had no birth country were excluded from this analysis. 75% was chosen as an arbitrary threshold.

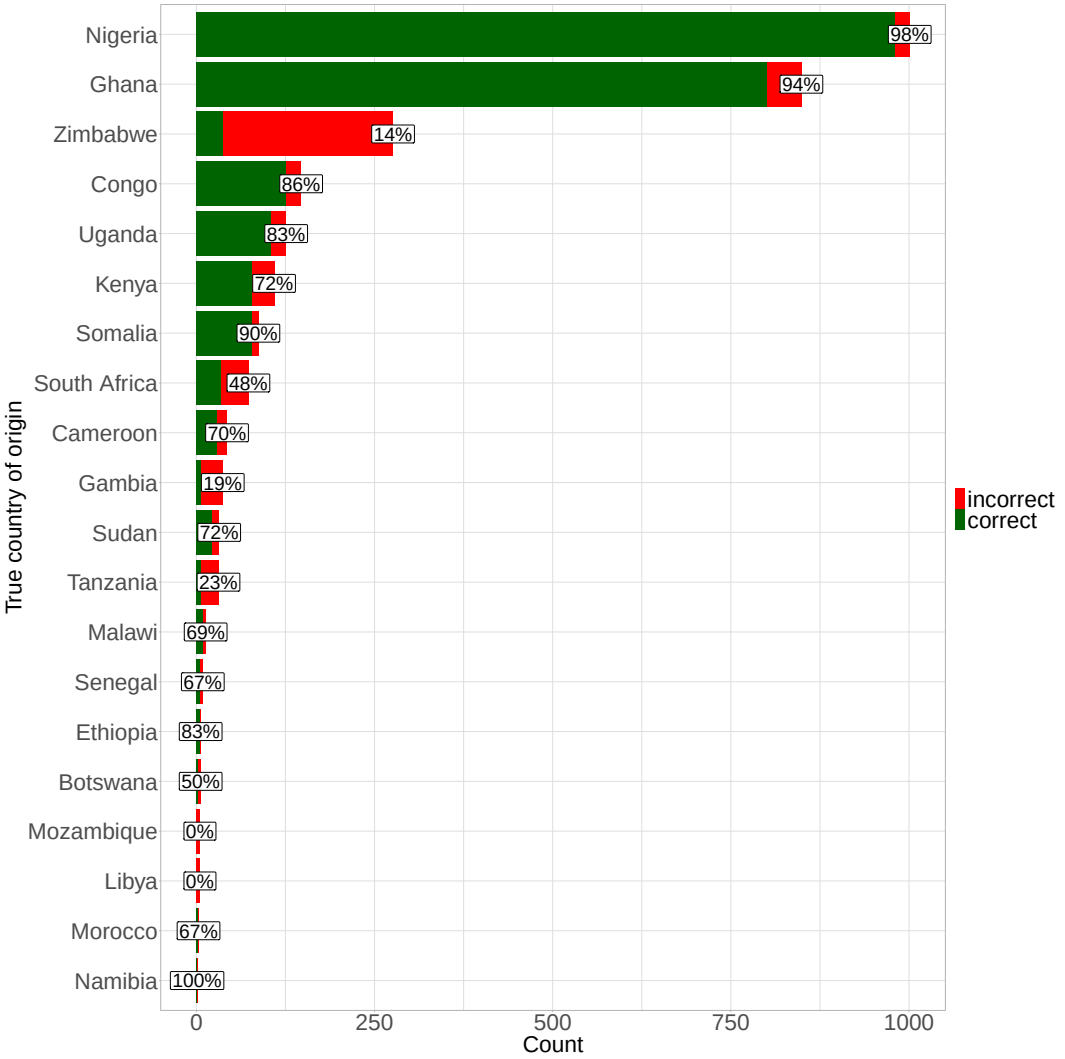



**Figure 1.10:** Correspondence of true birth country with estimated birth country. Each bar corresponds to a true birth country, with the length of the bar corresponding to the total number of people in our dataset born in that country. The green section corresponds to the total number of individuals where the birth country was correctly guessed and the red section to those who were incorrectly guessed. Percentage labels give percentage correct for that country.

The overall accuracy at predicting birth location across all individuals was 81.63%, suggesting there was substantial information within the coancestry matrix. For certain countries where there was large number of surrogate populations, such as Ghana and Nigeria, the prediction accuracy was high. For other countries, the prediction accuracy was much

lower. For example, Tanzania, which is only represented by a single reference population, had a prediction accuracy of 23%. Zimbabwe had by far the lowest prediction accuracy (14%) out of countries with more than 100 U.K. Biobank individuals. Of the 266 individuals born in Zimbabwe, 194 were assigned to an ethnic group from outside Zimbabwe; 74 to Malawi_Chewa, 71 to Mozambique_Mozambique and 49 to Malawi_Yao. Individuals from the ethnic groups from Malawi are found across Malawi, Zimbabwe and other countries, showing the possible weakness of this approach which aims to categorise individuals into a single country, as ethnic groups often transcend countries. Indeed we only have data from one (partially) Zimbabwean group, the Zulu, who may not well-reflect the ancestors of U.K. Biobank participants born in Zimbabwe.

I performed the same analysis but using the data which had been imputed. This stands as a practical test of whether it is preferable to impute or retain a smaller number of non-imputed SNPs when estimating country-level haplotype variation. This yielded an accuracy of 81.89%, a value almost identical to that obtained with the dataset containing approximately 70,000 non-imputed SNPs, despite my earlier results indicating that sub-country SOURCEFIND results are less accurate if using imputed data due to reference bias [CAN YOU REFERENCE THE CHAPTER SECTION HERE?]. This may be because this broad-scale assignment of individuals to countries is not as affected by imputation as a more subtle dissection of sub-country ancestry.[WE CAN TEST WHETHER THIS MIGHT BE THE CASE – I.E. DO SUB-COUNTRY ASSIGNMENT RESULTS DIFFER BETWEEN THE 70K AND IMPUTED? PERHAPS YOU CAN CREATE A SCATTERPLOT WHERE EACH DOT IS A REFERENCE POP, AND THE AXES GIVE THE AVERAGE INFERRED ANCESTRY FROM (OR AMOUNT DONATED BY) THAT GROUP FOR 70K AND IMPUTED.]

### 1.3.5 Patterns of African ancestry across the U.K.

The U.K. Biobank dataset contains data on the testing centre that each individual registered at. I used this information to determine whether there was structure in how individuals with recent African ancestry related to different African ethnic groups are distributed across the U.K. There were no apparent outliers in terms of any centers with substantially larger proportion of individuals who had at least 50% African ancestry Supplementary Fig. D.4) than others. However, as expected, centers in large cities such as Barts, Croydon and Hounslow (London), Birmingham and Manchester had the highest proportion of individuals with at least 50% African ancestry.

I then plotted the distribution of people with recent ancestry related to african ethnic groups at different centers on a map of the U.K (Fig. 1.11). No clear pattern was apparent,

other than Yoruban ancestry dominating most centres, with some smaller testing centers only containing individuals inferred as having Yoruba-related ancestry.

I estimated the information entropy, $E$, of each assessment centre based on the SOURCEFIND proportions, similar to previous work performed by van Dorp et al (2018), who used the principle of entropy to determine the extent to which individuals from different ethnic groups were scattered across different clusters [35].

To evaluate the extent to which individuals assigned to each ethnic group registered at different testing centers, I calculated entropy given by Schutze et al (2008) as $\sum_{i=1}^{L}[p_{i,j} \cdot log(p_{i,j})]$ [36], where $p_{i,j} = \frac{m_{i,j}}{m_j}$, $m_j$ is the number of individuals from testing center $j$ assigned to ethnic group $i$ and $m_{ij}$ is the number of ethnic groups to which individuals from center $j$ are assigned. Testing centers in large cities such as London and Birmingham had the highest information entropy, consistent with prior expectations that large cities would contain a higher diversity of ancestries (Fig. 1.11[LEGEND SEEMS TO HAVE DISAPPEARED HERE??]).

### 1.3.6 Patterns of African ancestry across the U.K.

I also had access to the birth-date of each U.K. Biobank participant. Therefore, it is possible to calculate the increase of the ancestry of a particular ethnic group over time based on birth-year. 1.12. I took all U.K. Biobank individuals with more than 50% African ancestry and split them into 50 bins according to their birth date. Using a rolling window in the `rollyapply` function from the `zoo` R library, I calculated the mean proportion of all ancestries across ancestry for each bin. Fig 1.12 shows the increase of Buganda ancestry over time.

We can observe roughly a doubling of the mean proportion of Buganda_Baganda ancestry between 1950 and 1964. In 1972, then president Idi Amin expelled roughly 60,000 Ugandans to the U.K. Therefore, this increase may tentatively correspond to an increase in the number of individuals between the ages of 7-22 arriving in the U.K. during these dates.

## 1.4 Discussion

In this chapter, I first showed that, in individuals with recent African ancestry, there is enough linkage information across 70,000 genome-wide SNPs to recover a substantial amount of useful haplotype information. Further, I found that using imputed genotypes may significantly reduce the power of a painting and introduce a degree of bias towards populations present in a reference panel used for imputation.


count
10000
20000
30000
40000
(entropy)
2000000
1500000
1000000
500000
Croydon
Birmingham

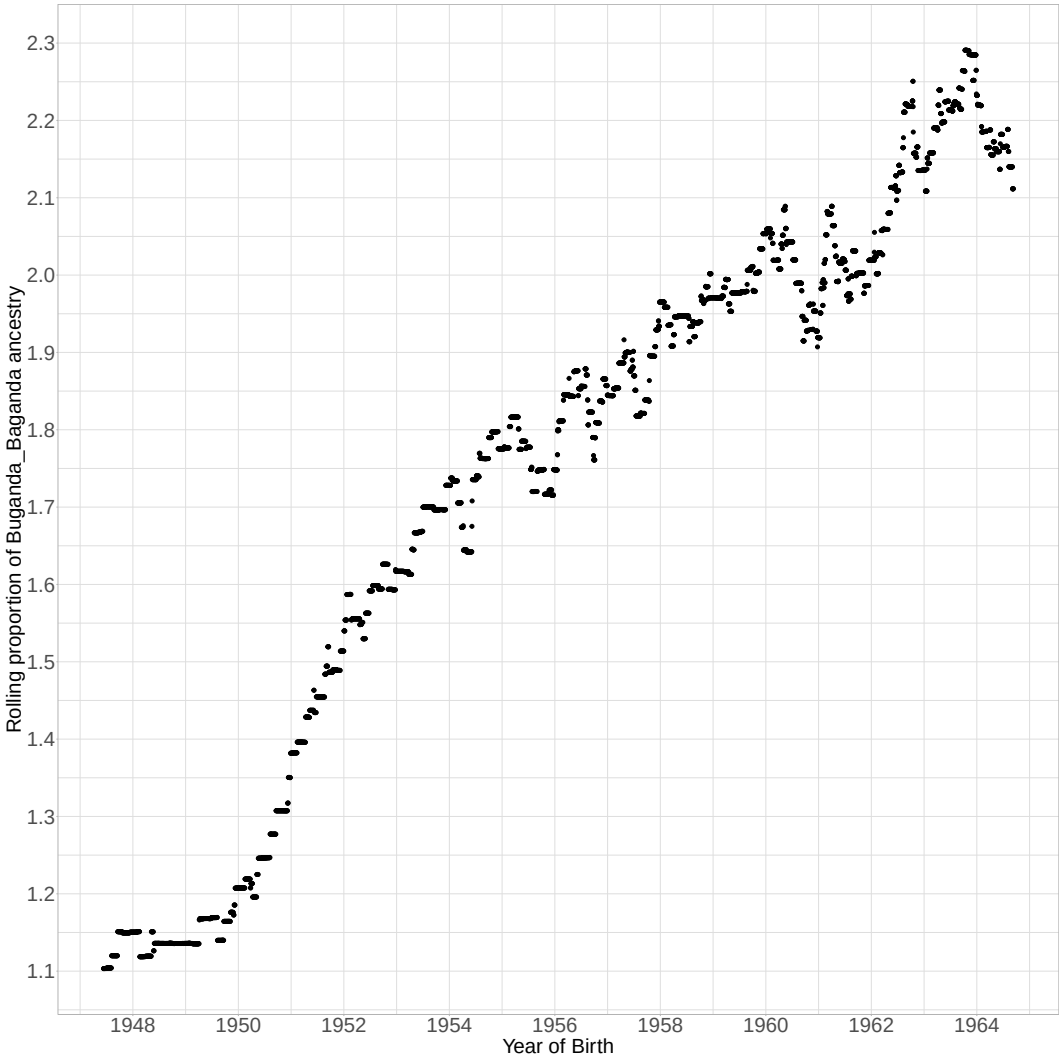



**Figure 1.12:** Increase in the mean proportion of Buganda ancestry between 1948 and 1965. An overlapping sliding window was applied to SOURCEFIND ancestry proportions and mean proportion of Buganda ancestry for each window plotted against the mean birth-date of individuals in that bin.

Future work on using Biobanks to explore population structure and history could focus on two points. Firstly, development of efficient methods to paint a single sample using a reference panel containing many thousands of samples, which also scales to Biobank-scale sample sizes (100,000+)[CAN PBWT NOT DO THIS? IN GENERAL WORTH DISCUSSING WHY PBWT MAY NOT BE GOOD ENOUGH (I.E. LIMITATIONS)]. Secondly, larger reference panels of worldwide populations and more ethnic groups will allow for a more detailed characterisation of genetic variation.

# Bibliography


[1] Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[2] Stephen Leslie, Bruce Winney, Garrett Hellenthal, Dan Davison, Abdelhamid Boumertit, Tammy Day, Katarzyna Hutnik, Ellen C. Royrvik, Barry Cunliffe, Daniel J. Lawson, Daniel Falush, Colin Freeman, Matti Pirinen, Simon Myers, Mark Robinson, Peter Donnelly, and Walter Bodmer. The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314, 2015.

[3] Clare Turnbull. Introducing whole-genome sequencing into routine cancer care: the genomics england 100 000 genomes project. *Annals of Oncology*, 29(4):784–787, 2018.

[4] UK10K consortium et al. The uk10k project identifies rare variants in health and disease. *Nature*, 526(7571):82, 2015.

[5] Stephan Schiffels, Wolfgang Haak, Pirita Paajanen, Bastien Llamas, Elizabeth Popescu, Louise Loe, Rachel Clarke, Alice Lyons, Richard Mortimer, Duncan Sayer, et al. Iron age and anglo-saxon genomes from east england reveal british migration history. *Nature communications*, 7(1):1–9, 2016.

[6] Xiaoming Liu. Human prehistoric demography revealed by the polymorphic pattern of cpg transitions. *Molecular biology and evolution*, 37(9):2691–2698, 2020.

[7] Susheila Nasta. *'Voyaging in': colonialism and migration*. Cambridge University Press, 2005.

[8] Teri A Manolio. Using the data we have: improving diversity in genomic research. *The American Journal of Human Genetics*, 105(2):233–236, 2019.

[9] Jacklyn N Hellwege, Jacob M Keaton, Ayush Giri, Xiaoyi Gao, Digna R Velez Edwards, and Todd L Edwards. Population stratification in genetic association studies. *Current protocols in human genetics*, 95(1):1–22, 2017.

[10] Karoline Kuchenbaecker, Nikita Telkar, Theresa Reiker, Robin G Walters, Kuang Lin, Anders Eriksson, Deepti Gurdasani, Arthur Gilly, Lorraine Southam, Emmanouil Tsafantakis, et al. The transferability of lipid loci across african, asian and european cohorts. *Nature communications*, 10(1):1–10, 2019.

[11] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, Genevieve L Wojcik, Benjamin M Neale, Simon Gravel, Mark J Daly, Carlos D Bustamante, and Eimear E Kenny. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

[12] Carlos D Bustamante, M Francisco, and Esteban G Burchard. Genomics for the world. *Nature*, 475(7355):163–165, 2011.

[13] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, Alexander Gusev, Sara Lindström, Stephan Ripke, Giulio Genovese, Po-Ru Loh, Gaurav Bhatia, Ron Do, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592, 2015.

[14] Arslan A Zaidi and Iain Mathieson. Demographic history mediates the effect of stratification on polygenic scores. *Elife*, 9:e61548, 2020.

[15] Ross P Byrne, Wouter van Rheenen, Leonard H van den Berg, Jan H Veldink, and Russell L McLaughlin. Dutch population structure across space, time and gwas design. *Nature communications*, 11(1):1–11, 2020.

[16] Daniel John Lawson and Daniel Falush. Population identification using genetic data. *Annual Review of Genomics and Human Genetics*, 13(1):337–361, 2012. PMID: 22703172.

[17] Ying Zhou, Sharon R. Browning, and Brian L. Browning. A fast and simple method for detecting identity-by-descent segments in large-scale data. *The American Journal of Human Genetics*, 106(4):426–437, 2020.

[18] Saioa López, Ayele Tarekegn, Gavin Band, Lucy van Dorp, Nancy Bird, Sam Morris, Tamiru Oljira, Ephrem Mekonnen, Endashaw Bekele, Roger Blench, et al. Evidence of the interplay of genetics and culture in ethiopia. *Nature communications*, 12(1):1–15, 2021.

[19] Garrett Hellenthal, Nancy Bird, and Sam Morris. Structure and ancestry patterns of Ethiopians in genome-wide autosomal DNA. *Human Molecular Genetics*, 30(R1):R42–R48, 02 2021.

[20] Deepti Gurdasani, Tommy Carstensen, Segun Fatumo, Guanjie Chen, Chris S Franklin, Javier Prado-Martinez, Heleen Bouman, Federico Abascal, Marc Haber, Ioanna Tachmazidou, et al. Uganda genome resource enables insights into population history and genomic discovery in africa. *Cell*, 179(4):984–1002, 2019.

[21] Olivier Delaneau, Jean-François Zagury, Matthew Robinson, Jonathan Marchini, and Emmanouil Dermitzakis. Integrative haplotype estimation with sub-linear complexity. *bioRxiv*, page 493403, 2018.

[22] Daniel Taliun, Daniel N Harris, Michael D Kessler, Jedidiah Carlson, Zachary A Szpiech, Raul Torres, Sarah A Gagliano Taliun, André Corvelo, Stephanie M Gogarten, Hyun Min Kang, et al. Sequencing of 53,831 diverse genomes from the nhlbi topmed program. *Nature*, 590(7845):290–299, 2021.

[23] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[24] Shane McCarthy, Sayantan Das, Warren Kretzschmar, Olivier Delaneau, Andrew R Wood, Alexander Teumer, Hyun Min Kang, Christian Fuchsberger, Petr Danecek, Kevin Sharp, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279, 2016.

[25] Daniel John Lawson, Garrett Hellenthal, Simon Myers, and Daniel Falush. Inference of population structure using dense haplotype data. *PLoS Genetics*, 8(1):11–17, 2012.

[26] David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

[27] Juan C. Chacon-Duque, Kaustubh Adhikari, Macarena Fuentes-Guajardo, Javier Mendoza-Revilla, Victor Acuna-Alonzo, Rodrigo Barquera Lozano, Mirsha Quinto-Sanchez, Jorge Gomez-Valdes, Paola Everardo Martinez, Hugo Villamil-Ramirez, Tabita Hunemeier, Virginia Ramallo, Caio C. Silva de Cerqueira, Malena Hurtado, Valeria Villegas, Vanessa Granja, Mercedes Villena, Rene Vasquez, Elena Llop, Jose R. Sandoval, Alberto A. Salazar-Granara, Maria-Laura Parolin, Karla Sandoval, Rosenda I. Penaloza-Espinosa, Hector Rangel-Villalobos, Cheryl Winkler, William Klitz, Claudio

Bravi, Julio Molina, Daniel Corach, Ramiro Barrantes, Veronica Gomes, Carlos Resende, Leonor Gusmao, Antonio Amorim, Yali Xue, Jean-Michel Dugoujon, Pedro Moral, Rolando Gonzalez-Jose, Lavinia Schuler-Faccini, Francisco M. Salzano, Maria-Catira Bortolini, Samuel Canizales-Quinteros, Giovanni Poletti, Carla Gallo, Gabriel Bedoya, Francisco Rothhammer, David Balding, Garrett Hellenthal, and Andres Ruiz-Linares. Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. *Nature Communications*, page 252155, 2018.

[28] Lucy Huang, Mattias Jakobsson, Trevor J Pemberton, Muntaser Ibrahim, Thomas Nyambo, Sabah Omar, Jonathan K Pritchard, Sarah A Tishkoff, and Noah A Rosenberg. Haplotype variation and genotype imputation in african populations. *Genetic epidemiology*, 35(8):766–780, 2011.

[29] Roman Shraga, Sarah Yarnall, Sonya Elango, Arun Manoharan, Sally Ann Rodriguez, Sara L Bristow, Neha Kumar, Mohammad Niknazar, David Hoffman, Shahin Ghadir, et al. Evaluating genetic ancestry and self-reported ethnicity in the context of carrier screening. *BMC genetics*, 18(1):1–9, 2017.

[30] Y. V. Louwers, O. Lao, B. C. J. M. Fauser, M. Kayser, and J. S. E. Laven. The Impact of Self-Reported Ethnicity Versus Genetic Ancestry on Phenotypic Characteristics of Polycystic Ovary Syndrome (PCOS). *The Journal of Clinical Endocrinology & Metabolism*, 99(10):E2107–E2116, 10 2014.

[31] Elena Bosch, Hafid Laayouni, Carlos Morcillo-Suarez, Ferran Casals, Andrés Moreno-Estrada, Anna Ferrer-Admetlla, Michelle Gardner, Araceli Rosa, Arcadi Navarro, David Comas, et al. Decay of linkage disequilibrium within genes across hgdp-ceph human samples: most population isolates do not show increased ld. *BMC genomics*, 10(1):1–9, 2009.

[32] Michael Banton. Recent migration from west africa and the west indies to the united kingdom. *Population Studies*, 7(1):2–13, 1953.

[33] Steven J Micheletti, Kasia Bryc, Samantha G Ancona Esselmann, William A Freyman, Meghan E Moreno, G David Poznik, Anjali J Shastri, M Agee, S Aslibekyan, A Auton, et al. Genetic consequences of the transatlantic slave trade in the americas. *The American Journal of Human Genetics*, 107(2):265–277, 2020.

[34] James A Rawley and Stephen D Behrendt. *The transatlantic slave trade: a history*. U of Nebraska Press, 2005.

[35] Lucy Van Dorp, Sara Lowes, Jonathan L Weigel, Naser Ansari-Pour, Saioa López, Javier Mendoza-Revilla, James A Robinson, Joseph Henrich, Mark G Thomas, Nathan Nunn,

et al. Genetic legacy of state centralization in the kuba kingdom of the democratic republic of the congo. *Proceedings of the National Academy of Sciences*, 116(2):593–598, 2019.

[36] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press Cambridge, 2008.