Harnessing haplotype sharing information from low coverage sequencing and sparsely genotyped data

Sam Morris

A dissertation submitted in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

of

University College London.

UCL Genetics Insitute

University College London

March 12, 2022

I, Sam Morris, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Accounting for linkage information has been shown to enhance power to detect fine-scale population structure, particularly when considering recent shared ancestry. In particular, ChromoPainter has been shown to be a successful method at identifying shared haplotypes between samples. It has also been used widely on ancient DNA samples. However, coverage is an issue, and it is possible that analysing low-coverage samples may provide biased results, for example if using phasing and imputation algorithms to pre-process data. Whilst a small number of studies have tested the utility of using ChromoPainter on ancient DNA, none have tested a range of samples across different coverages, at all steps of the analysis pipeline. In this work, I assess the impact of coverage on each step of the ChromoPainter analysis pipeline. I show that bias can exist when exploring (e.g.) population structure using low-coverage samples, and I investigate a series of modifications and strategies to reduce the extent of this bias. I also address a related challenge of analysing haplotype information in sparsely genotyped data in present-day individuals, e.g. when analysing only variants that overlap multiple genotyping arrays. Using these findings, I infer fine-scale African ancestry in U.K. Biobank participants using a new reference panel of data from 349 African ethnolinguistic groups, demonstrating how imputation of sparsely genotyped samples can substantially harm the estimation of sub-continental ancestry. Furthermore, I analyse a novel ancient DNA dataset from Bavaria in order to determine the extent of continuity between the Late Neolithic and Iron Ages, as well as the age of east-west

structure in Europe. I also analyse novel ancient DNA samples from Slavic-speaking regions, exploring the genetic relationship between samples from the Migration Era to the Early Middle Ages, and the signatures of these ancient populations in present-day Slavic speaking populations. Finally, I summarise my findings and recommend approaches for future work on haplotype-based studies using low-coverage or sparsely genotyped data.

Acknowledgements

I would like to thank the following people.

First, my supervisor Garrett for his dedication and skill, not only in helping me through my PhD, but with all of the other small things along the way.

My Mum, Dad and sister for always supporting me and their voluntary proof-reading efforts.

All gang in office 212 who I had a lot of fun with; Mislav, Lucy, Nancy, Magnifica, Arturo, Dave, Mike, Chris, Camus and all the others who came by, even for a bit. It was sad to be cut short, but I hope to see you all in the future. All of the LIDo guys as well, too many to name, who made being at UCL so enjoyable.

All the good folk at UCL Computer Science cluster, particularly Ed and David, for putting up with my poor cluster etiquette over the years.

Nadine for being the best administrator ever.

Thank you to Pascal and Jay for looking after me when times were tough.

0.1 Impact statement

I intend that the work presented in this thesis will provide a foundation for other researchers who apply haplotype-based methods for the analysis of low coverage ancient DNA and sparsely genotyped. Specifically, the benchmarks I

provide in Chapter 2 can be followed by scientists in order to perform reliable ancient DNA analyses. This is important, as many studies are now using the aforementioned techniques. I also hope that others will take over up work into adapting ChromoPainter for ancient DNA and make further improvements to the algorithm.Similarly, other researchers can use my results to make decisions on whether to retain a smaller number of SNPs, or impute missing ones, when merging datasets across multiple genotyping arrays. Given previous research has outlined the utility of accounting for haplotypes when accounting for population stratification in GWAS, my findings may be useful looking forward when such approaches become more common.

My empirical work on ancient DNA in chapters 4 and 5 should a grounding for future work, much like the work I referenced in those sections aided me in understanding the historical and genetic context of the current research. For example, future studies may use these results to inform how they sample new ancient DNA samples.

Outside of academia, I believe there is a fundamental benefit to learning about our history as a species, something which the study of ancient DNA has provided tools for in the past decade. Ancient DNA analysis remains a field with popular reach, so I hope my work will go a small way towards providing the public with interesting and scientifically valid findings.

I believe that exploring the ancestry of ethnic minorities within the U.K. Biobank can be of value to those individuals communities, particularly when they have been excluded from many similar kinds of analyses. Lastly, my work should also play a part in the inclusion of a more diverse array of ethnicities in association studies.

Contents

Chapter 1

Introduction

1.1 Chromopainter and ancient DNA

In this introduction I will discuss the following points: i) What are 'haplotype-based' methods and what advantages and disadvantages do they offer over 'unlinked' methods, ii) a summary of different methods used to analyse ancient DNA and iii) the need to merge datasets genotyped on different arrays.

1.1.1 Gains to be made with haplotype information

1.1.1.1 History

Haplotype-based methods are statistical approaches in genetic analysis which explicitly model linkage disequilibrium (LD), or the correlation in frequency, between neighbouring genetic markers along a haplotype [1]. This is in contrast to 'unlinked' methods, which assume a model of linkage equilibrium between SNPs. A 'haplotype' is a contiguous sequence of alleles which are located on the same chromosome. In this thesis, I will concentrate on haplotype-based methods in the context of identifying shared haplotypes between individuals in

[1]Note that other methods, for example `octopus` [1] are referred to as 'haplotype-based' genotype callers, but they represent a distinct group of methods to e.g. ChromoPainter.

order to understand the genetic structure and history of a population(s).

Linkage disequilibrium (LD) is the key concept underpinning haplotype-based approaches. It has been studied since the earliest days of genetics [2, 3] and has since been a fundamental aspect of virtually all areas of genetics [4]. The primary advantage of accounting for LD in a model is that information about the frequency of an allele in a population also provides information about the frequency of neighbouring alleles within the same population.

Some of the earliest uses of LD information for the study of genetic structure came from microsatellite markers, whose linked tandem repeats can be thought of as analogous to linked alleles on a haplotype. Microsatellites were, and still are, commonly applied to study the population structure of wild animal systems; for instance, Amos et al (1993) used microsatellites markers to examine the population structure of whales [5]. Later, microsatellites at the CD4 locus were leveraged to show the preferred model of Human population history was a recent African origin [6]. This was deduced as Sub-Saharan Africans had substantially more variability in haplotype frequency and a higher diversity of STRP alleles associated with the Alu deletion than non-Africans, strongly suggesting Africa was the common origin of these haplotypes. This study outlined the insights into population history that can be obtained from the analysis of a very small number of linked markers.

The next major advance was the development of methods to use LD information between SNP markers rather than within microsatellites, as SNPs are substantially more numerous across the human genome. Studies in the early 2000s utilised the then-new Hap-Map results [7] to show LD varies across the human genome [8] and between worldwide populations [9, 10], and that such variation can be used to make inferences about human populations history [11]. Using 3,024 autosomal SNPs, Conrad et al (2006) calculated the proportion of unique haplotypes that were shared between two geographic regions, and by showing that the number of distinct haplotypes per region declines from

Africa, provided additional evidence to support the previously proposed recent African origin of humanity [12]. It was also shown that isolated Native American populations had approximately 3 times fewer haplotypes per genomic region, indicating that recent endogamy plays a large role in shaping patterns of haplotype variation.

The 2000s also saw a rapid increase in the number of SNP markers and individuals which had been sequenced. Accounting for LD and recombination within a model is necessarily computationally complex and the number of combinations of alleles and their possible evolutionary histories balloons as the number of loci increases. Therefore, the new era of sequencing demanded new and more efficient methods to cope with such data. The development of the Li and Stephens copying model (LSM) [13] was instrumental in the development of such methods [14] and provided an elegant solution to the increased complexity when modelling recombination between linked loci. As such, it has since played a part in virtually all areas of genomic methodology; for example, the LSM was, and still is, the foundation for methods of the haplotype phasing methods needed for haplotype-based methods [15, 16]. LSM provides a way to generate a 'target' haploid conditional upon a set of other observed haploids, specifically by modelling it as a 'mosaic' of the other sampled haploids using a Hidden Markov Model. The conditional probability that the target haploid 'copies' from a particular reference haplotype is obtained by observing whether the alleles at the same position match between the target and reference haplotypes. The mosaic nature of the target haplotype reflects how historical recombination alters the genealogy relating sampled haplotypes along a genetic sequence, which in this model causes so-called 'switches' in which reference haplotype it copies from. In general, if a target haploid matches a DNA segment to a particular reference haploid for a genomic region, the target is inferred to share a most recent ancestor with that reference haploid, relative to all other reference haploids, for that genomic region.

The first paper to use the LSM model explicitly to study human population history was that of Hellenthal et al 2008 [17]. The original LSM was developed to infer recombination rates. It did so by randomly ordering a set of phased haploids, presumed to be sampled from a genetically homogeneous population, and then taking each haploid in turn and forming it as a mosaic of the haploids earlier in the random ordering. They then multiplied the resulting probabilities of generating each haploid, using this so-called "product of approximate-conditional" (PAC) likelihoods as a basis to infer the recombination rate. Hellenthal et al 2008 instead used the mosaic approach to calculate the probability of forming a set of haploids from one population as a mosaic of those from another population(s), using these probabilities to infer the relative order in which populations were formed. While their approach had some flaws, such as not explicitly accounting for admixture, it provided some insights into the power of LSM-based approaches to infer features of human history, using only a modest number of SNPs (n=2,560). For example, similar to the results of Conrad et al (2006), Hellenthal et al's analysis of the structure of global haplotype sharing provided strong evidence of a recent African origin of modern humans. In the same year, Jakobsson et al (2008) analysed a much larger number of SNPs (n=525,910) and 29 worldwide populations [18] to show that haplotype clusters show an elevated ability to determine local structure compared to unlinked SNPs alone; 51% of haplotype clusters were found in at most two regions, in contrast with 4% of SNP alleles.

Building on the copying model proposed by Hellenthal et al (2008), Lawson et al (2015) [19] created ChromoPainter, again based the LSM. ChromoPainter is a more general model than that of Hellenthal 2008; whereas the Hellenthal 2008 model was explicitly formulated to determine the ordering of human colonisation, ChromoPainter efficiently forms a set of target haplotypes as a mosaic of a set of reference haplotypes. In particular, it generates a 'coancestry matrix', which gives information on the level of recent shared ancestry between each donor and recipient individual. ChromoPainter also allowed for the user

to input recombination rate maps containing estimated recombination rates between neighbouring SNPs. Analysis of simulated data showed it to have an enhanced ability to separate closely related populations when plotted on a PCA compared to unlinked methods. It was developed in tandem with its own clustering method fineSTRUCTURE, and has since been extended into methods to detect and date admixture [20], and infer ancestry proportions [20, 21].

The 'next-generation' of chromosome painting methods had to confront the same issue that Li and Stephens did, which was how to adapt methodology to larger and larger sample sizes. ChromoPainter was designed with datasets of <10,000 people in mind, whereas biobank-scale datasets typically contain 500,000+ individuals. As such, ChromoPainter does not scale well to large datasets, especially when there are a large number of donor haplotypes.

One approach is to use the Burrows-Wheeler transform (PBWT) [22, 23] to efficiently find matching haplotypes in large datasets. The insight to apply the PBWT to genetic data has been one of the most crucial insights into computation biology, as it allows for substantial increases in efficiency across a wide range of applications such as sequence alignment [24], phasing [25] and data compression [26]. PBWT has been applied to Chromosome Painting on Biobank-scale datasets in several recent papers [27, 28]. Similarly, methods to detect IBD in Biobank-scale cohorts have leveraged the PBWT [29, 30]. However, PBWT-based approaches are still relatively immature; for example, they do not allow for the use of a reference panel and all haplotypes must be compared to all other haplotypes in an 'all-v-all' manner (further explanation given in Appendix section B.1). Despite their current limitations, it seems that the future of Chromosome Painting will at least in part be based on the PBWT or similar approaches that increase computational efficiency, even if at slight losses in accuracy. Byrne et al used ChromoPainter and PBWT-paint to a subset of Dutch individuals and found eigenvectors of the coancestry matrix to be almost identical ($r^2 = 0.99$) and the correlation between raw coancestry

matrices to be lower at ($r^2 = 0.82$).

1.1.1.2 Advantages of accounting for haplotypes

ChromoPainter can be run in either 'linked' or 'unlinked' mode. In the linked mode, described in detail in sections 2.2.1, LD between neighbouring SNPs is accounted for. Unlinked mode assumes a model of linkage equilibrium between markers and has been shown to be statistically identical to the likelihood model underlying the commonly used ADMIXTURE algorithm [19].

A typical case study, and one which I will return to in later chapters, was a study investigating population structure among individuals from the British Isles [31]. This study, hereafter referred to as POBI, genotyped 2039 people from England, Wales and Scotland [31]. One finding was that it was possible to detect structure between individuals from Devon and Cornwall (two neighbouring counties) using ChromoPainter. On the other hand, this structure was not discernible when using unlinked methods (PCA). This outlines the benefits of incorporating linkage information when attempting to identify fine-scale structure between closely related groups of individuals.

Gattepaille and Jakobson (2012) [32] provided the mathematical foundations for the advantage of using linked markers over unlinked ones. They describe a metric, GIA (gain of informativeness for assignment), a term borrowed from information theory, to describe the additional amount of information gained when using haplotype data instead of unlinked alleles. They showed that whilst combining two markers in linkage equilibrium is not necessarily advantageous for ancestry inference, GIA is often positive for markers in LD with one another, demonstrating the advantage of haplotypes. Under a variety of simulated scenarios, incorrect assignment of individuals into populations was reduced between 26% and 97% when using haplotype data. For example, they showed that using empirical data of individuals from France and Germany, accounting for haplotypes could reduce the rate of mis-assignment by 73%.

Another advantage of using haplotype information is that it may mitigate ascertainment bias. Ascertainment bias occurs when a subset of SNPs are chosen for analysis, most often when selecting markers for a genotpying array. SNPs are typically chosen because they show variation within a population of interest. However, if this variation is identified in one population, e.g. British, then there is no guarantee that the variation will also be seen in another population, e.g. Han Chinese. In this case, including these SNPs can often provide misleading estimates of genetic diversity and commonly estimated parameters such as f_{st} [33]. Conrad et al (2006) showed that, owing to the lack of African individuals used in the SNP discovery process, populations from the Middle East, Europe and South Asia showed the highest levels of SNP-based heterozygosity. These findings were in stark disagreement with the currently accepted model of human history and studies which demonstrated Africans have the highest levels of genetic diversity [12, 17, 34–36]. However, when haplotype heterozygosity rather than SNP heterozygosity was used as a metric for diversity, African populations consistently had the highest values. Therefore, although the ascertainment for a particular SNP may depend strongly upon the ascertainment scheme, the same underlying haplotypes are likely to be observed, regardless of which SNPs are used to tag them.

Haplotype-based methods also rely less on the inclusion of rare alleles. Rare alleles are highly informative about recent, fine-scale population structure. Methods which leverage this information have been used to model the population history of large datasets [37–39]. However, rare alleles are harder to genotype, as they are more difficult to distinguish from sequencing errors and they are often not included on standard genotyping arrays. Because of this, allele-frequency filters are often applied in population genetic studies to reduce the risk of incorporating incorrectly genotyped SNPs. Further, more SNPs need to be sequenced in order to find rare variants in a wide range of populations. Using haplotype information may negate the needs for using rare variants; if individuals share long haplotypes in common, then it is likely that they also

share rare variants that occur on those haplotypes.

However haplotype-based methods are not without their drawbacks. They are typically slower by an of magnitude, as they are more computationally complex than unlinked methods. Secondly, the nature of haplotype-based methods means they require the data to be phased. Phasing is a statistical procedure [2] that requires substantial computation resources. The inconvenience of introducing an additional time and resource intensive step to the analysis means that many studies opt not to use such methods.

Finally, 'switch-errors' may often occur during phasing, when the incorrect ordering of alleles on a haplotype is inferred. Whilst Lawson and Falush (2012) showed that sporadic, randomly distributed switch-errors are unlikely to significantly affect the overall ChromoPainter analysis, systemic errors, where haplotypes from particular individuals are made to look more like each other than they do those of other members of the sample, may be more problematic and provide misleading results [40].

1.2 Methods used to analyse ancient DNA

In this section, I will outline some of the most widely used methods to analyse ancient DNA.

1.2.1 Unlinked methods

The first studies into ancient DNA mostly used statistical methods which compare allele-sharing or allele-frequencies between populations or individuals. These methods, in particular F-statistics and their extensions [41–44] and Principle Component Analysis [45], can address a wide-range of questions pertaining to population structure, admixture and shared drift.

[2]Phasing can also be performed using other methods, such as sequencing family trios. However, this is rarely used in population genetic studies (although see [33] for an example of it being used) and so I will not discuss it here

A key reason why methods based on allele-sharing and allele-frequency differences were, and still are, widely used in ancient DNA is that they can easily be modified to use data in pseudo-haploid format. Pseudo-haploid genotypes are generated by sampling a read at random to represent a single allele at a given SNP. This is often necessary, because ancient samples routinely do not have enough reads covering a SNP to confidently call heterozygous genotypes. Pseudo-haploid calls are therefore used widely, including currently (e.g. [46]), in most studies of ancient humans.

Whilst pseudo-haploid genotype calls circumvent the problem of calling heterozygous genotypes at low coverage positions, they necessarily hold less information relative to true diploid genotypes and are thus less powerful at e.g. identifying population structure or genetic similarity. Further, the use of pseudo-haploid calls may result in an elevated level of reference bias [47–49]. Reference bias occurs because the reference fasta file which is used to align reads only contains a single allele at each position. Therefore, reads which contain a non-reference allele (i.e. an allele not represented in the reference fasta) contain more mismatches with the reference than reads which contain the reference allele, and accordingly are given a lower mapping quality score. Then, when selecting a read at random, reads with the reference allele are more likely to be selected as the pseudo-haploid call, generating a bias towards the reference allele. Attempts are being made to represent non-linear reference genomes as graphs in order to mitigate the effect of reference bias [49, 50].

For many of the early ancient DNA studies, such as that of Green et al 2010 [41] and Lazaridis et al 2014 [51], powerful methods for detecting population substructure and admixture were not required, as the questions asked primarily considered broad questions about human history, such as the nature of human-archaic interactions and whether there was significant genetic differences between the first farmers and the preceding hunter-gatherers. These populations, particularly humans and Neanderthals, are highly diverged and

hence do not require powerful methods to be distinguished. For example, in the case of Lazaridis et al (2014), simply plotting Loschbour and Stuttgart on a PCA of modern individual showed they had substantially different ancestries.

Perhaps the most widely used method amenable to pseudo-haploid data is the family of F-statistics [3], which were first outlined in a 2009 study into the population history of India [53]. These methods use the principle of shared drift in order to estimate genetic similarity (f_2), branch-length and admixture (f_3) and tests of tree-like phylogeny (f_4). Since 2009, F-statistics have been extended into multiple, more advanced, frameworks which are able to answer more complex questions about population history through the generation of population admixture graphs. In particular, qpAdm has been shown to be a flexible and coverage-robust method of estimating individual and population level admixture fractions [44]. An attractive feature of F-statistics is that they explicitly test models of population history and can provide readily interpretable results with associated jackknifed confidence intervals. A related method is the so-called ABBA-BABA test, developed by Green et al (2010) [41] in order to determine whether, and to what extent, admixture between humans and the newly sequenced Neanderthal genome had occurred. This simple test counts the number of times across the genome a 4 population phylogenetic tree shows a particular configuration at a given locus in order to determine whether an admixture event has taken place.

In contrast to the F-statistics, which explicitly tests models of population relationships, Principle Component Analysis (PCA) is a 'model-free' method typically used to obtain a visual summary of the genetic ancestry of the sample being analysed. PCA is commonly used as it is typically fast and easily interpretable. Several methods have been developed which adapt the standard PCA approach (e.g. eigenstrat [45]) to low coverage ancient DNA [54–56]. I note that PCA may also be performed on matrices obtained from linked

[3] Although related, they should not to be confused with Sewall Wright's F-statistics [52].

analysis, such as a matrix of pairwise IBD sharing or ChromoPainter coancestry matrix.

Throughout my thesis, I will make extensive usage of both PCA and F-statistics on both present-day and ancient human populations.

1.2.2 ChromoPainter ancient DNA

1.2.2.1 History

In recent years, many of the 'low hanging fruit' of broad-scale questions regarding the ancient history of humans in Eurasia have mostly been answered and studies into more fine-scale populations structures have become more prevalent. Accordingly, methods which can detect more subtle population structure have been required. However, the incorporation of ChromoPainter analysis into studies of ancient DNA was slow, in part because of the difficult of phasing low-coverage genomes and concerns over introducing bias towards present-day populations during imputation.

ChromoPainter can be used to answer a variety of questions relating to the genetic variation and population history of groups of samples. It can provide an overview of genetic ancestry through Principle Component Analysis of the coancestry matrix. For instance, differential haplotype donation to different worldwide populations, as shown in Fig 1.1, can reveal geographic correlates of genetic variation.

The first use of ChromoPainter on ancient DNA was in the seminal paper of Lazaridis et al (2014) [51]. Through the generation of two high-coverage ancient genomes, they were the first to propose that most present-day Europeans can be modelled as a mixture of three ancestral populations. For the ChromoPainter analysis, they did not impute missing genotypes in the ancient samples, as the possible bias effects had yet to be studied; only positions with non-missing

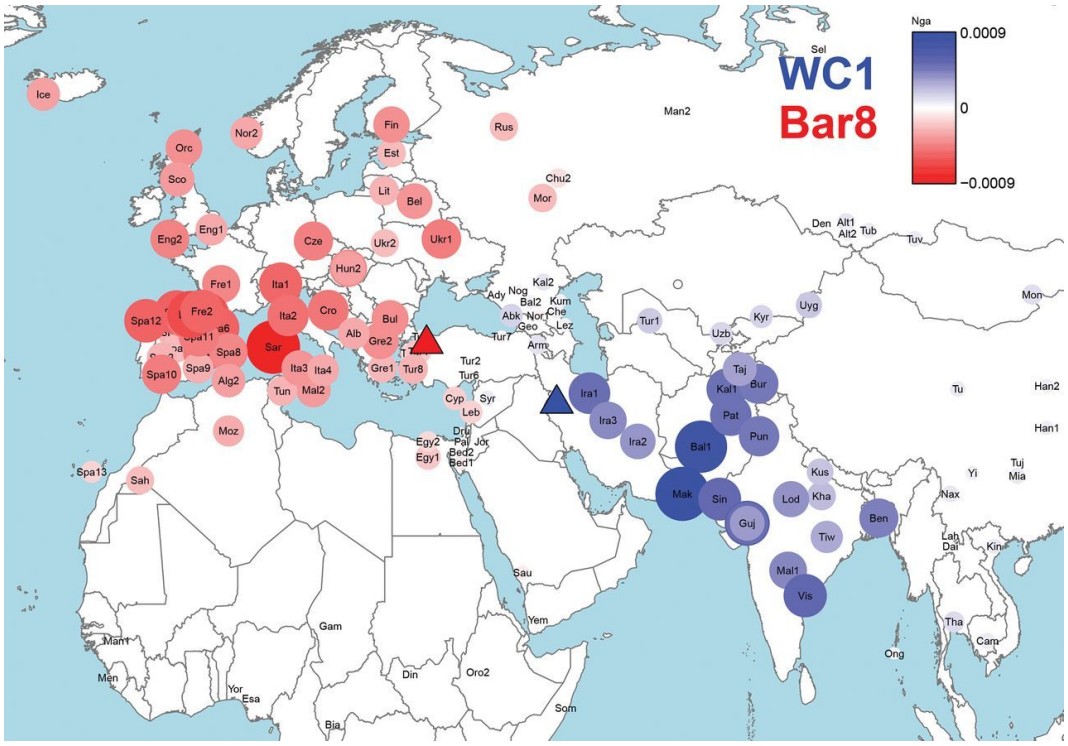

Figure 1.1: Map of differential haplotype sharing with present-day populations between WC1 (Iranian Farmer) and Bar8 (Anatolian Farmer) from Broushaki et al (2016) [57]. Bar8 copies relatively more from red populations and WC1 from blue populations.

genotypes were retained. As the samples were of high coverage, this was not an issue, as 495,357 SNPs were kept. The the ability of fineSTRUCTURE to meaningfully cluster ancient individuals was confirmed by recapitulating previous results that identified different present-day European populations as being more closely related to Early Farmers and hunter-gatherers than others.

In-between 2014 and the present-day, there have been approximately 31 studies which have used ChromoPainter on ancient samples (based on Web of Science search results). As of writing (September 2021), the study of Margaryan et al (2020) is the biggest so far to use ChromoPainter, with over 400 samples used [58]. This study concluded that detecting structure within the dataset using 'traditional' methods was not possible and so opted to use haplotype-based analyses on all samples above 0.5x mean depth. Another recent large study into the genomic history of the Roman Empire and surrounding regions

leveraged ChromoPainter [59].

More recently, ChromoPainter has been used to study aspects of archaic hominin ancestry in present-day humans [60, 61]. Whilst ChromoPainter is not specifically designed to accurately estimate local ancestry, it is possible to infer identify potentially introgressed Denisovan regions of DNA by determining whether a haplotype which is more similar to the Denisovan genome than to a panel of sub-Saharan Africans. ChromoPainter has also been extended to studying the ancient DNA of non-human organisms such as bacteria [62].

1.2.2.2 Benchmarking ChromoPainter and imputation

Many studies which have used ChromoPainter on ancient samples have performed tests and benchmarks to various degrees of detail.

The first study to investigate the reliability of ChromoPainter on ancient DNA was Martiniano et al (2017) [48]. Testing whether including imputed genotypes introduced bias towards particular present-day populations was key, as if it were the case, it may invalidate any results obtained. The authors estimated potential bias by plotting normal quantile-quantile plots of the copyvectors obtained from imputed (after downsampling to 2x coverage) and non-imputed markers. Whilst the differences in amount of copying differed by up to 14%, most percentage differences were substantially lower and there was no evidence of structured bias towards or against particular geographic regions, with the authors concluding "There is no strong evidence for systematic changes being caused by genotype imputation".

The same study also investigated the impact of filtering genotypes based on genotype probabilities by creating two datasets, one containing filtered genotypes and without, and performing fineSTRUCTURE clustering on both. fineSTRUCTURE inferred 7 more clusters when using filtered genotypes; whilst this could be an indication of improved clustering resolution, it is hard to draw

solid conclusions from these data. The overall number of fineSTRUCTURE clusters can not be seen as a direct measurement of performance; for example, the additional clusters inferred may simply be a result of the stochastic nature of MCMC sampling, and given only a single replicate of each test was performed, it is not possible to rule this out. Performing the same analysis on simulated data, where the population labels of individuals are known in advance, would be a more controlled test.

Since the study of Martiniano et al, many papers which incorporated ChromoPainter analysis into studies of ancient DNA have included their own set of benchmarks. Antonio et al (2019) [59] tested imputation accuracy on an ancient sample (NE1) downsampled to different levels of coverage. However, this analysis was only performed on a single sample and the effect of imputation on the ChromoPainter process was not evaluated. Margaryan et al (2020) performed a downsampling test on two high coverage genomes down to 1x mean coverage and concluded that, whilst there was some suggestion that the 1x downsample tended to a more mixed ancestry profile, there was no evidence that incorrect ancestries have been inferred or that major changes in ancestries have occurred.

Imputation is a necessary pre-processing step for ChromoPainter analysis on low-medium coverage ancient DNA samples for two primary reasons. Firstly, ChromoPainter does not allow for missing genotypes and so imputation is required to estimate missing genotypes. Secondly, whilst they are covered by reads, non-missing positions may still be low in coverage and thus require to be re-estimated, particularly when the true genotype is heterozygous. Therefore, it is important to determine to what extent it is possible to accurately impute genotypes at different levels of mean coverage.

The accuracy of imputation on ancient samples has been tested in various studies [48, 63, 64]. There is difficulty in comparing the estimated accuracies between studies, however, due to differences in factors such as samples analyses,

software used to call genotypes and impute samples, the regions analysed and filters applied.

The most systematic and thorough evaluation of imputation in ancient genomes was performed by Hui et al (2020) [63]. This study noted that it is possible to impute using a one or two step approach and, through the use of downsampled genomes, showed that the two-step approach provides more accurate imputed genotypes. This study also showed that whilst most genotype likelihood callers (e.g. GATK, atlas) performed similarly well, atlas was preferred because of it's ability to model post-mortem damage (PMD) in ancient samples. Accordingly, I will use atlas to call genotype likelihoods in the rest of my thesis.

It should be noted that the study only considered a single ancient genome (NE1) and it is therefore unclear how generalisable these results are to ancient samples of different ancestries. However, this study provided important benchmarks for many critical steps in the analysis of low coverage samples which had previously been missing from the literature, such as selection of a reference panel, the feasibility of local imputation and the effects of applying of pre and post imputation filters. One takeaway message was that it is possible to recover nine out of ten common (MAF ≥ 0.3) genotypes in a sample of 0.05x coverage.

In Chapter 2 of my thesis, I will explore the effect of coverage on imputation and ChromoPainter performed on ancient DNA samples.

1.3 Issues and solution to low coverage data

Low sequencing coverage is an issue which has plagued the field of ancient DNA since its inception. Compared to DNA obtained from present-day samples, ancient DNA samples typically have a much lower proportion of endogenous DNA, as DNA degrades over time from environmental factors. Therefore, when the DNA fragments are sequenced, relatively few of them will align to the

human reference.

The primary issue with low-coverage data is the increased uncertainty when calling diploid genotypes, particularly when the true genotype is heterozygous. Several methodological adaptations have been applied to existing methods in order to adapt them to low coverage ancient DNA. These approaches primarily attempt to circumvent making diploid genotype calls; for example, the previously mentioned strategy of pseudo-haploid genotype calling.

Alternatively, methods may avoid making diploid calls by working on genotype likelihoods. Genotype likelihoods represent a posterior estimate of the confidence of the three different genotypes at a bi-allelic locus, and thus allow the method to appropriately propogate that certainty throughout the analysis. A wide array of complex statistical approaches have been developed in order to accurately estimate the posterior genotype likelihoods. These approaches integrate factors such as sequencing-machine reported base-quality scores and estimates of read-mapping / sequencing errors [65]. Common methods to estimate likelihoods include the GATK model [66], SAMtools [67], SOAPsnp [68] and SYK model [69]. Genotype likelihoods can either be estimated prior to the analysis from aligned reads (BAM files), using software such as ANGSD [70], ATLAS [71] or GATK [66]. Other softwares will take BAM files directly as input and estimate genotype likelihoods during the analysis process (e.g. STITCH [72]).

Once genotype likelihoods have been estimated, population level parameters such as inbreeding coefficients and f_{st} can be estimated directly [70] with greater accuracy than direct genotype calls. Similarly, modifications of the ADMIXTURE [73] algorithm and PCA have been developed in order to analyse low coverage samples more effectively [74, 75]. Recent advances have allowed the identification of 1st and 2nd-degree relatives from as low as 0.02x coverage samples [76, 77].

Several methods account for low-coverage data by jointly estimating ancient DNA specific confounding factors, such as contamination and post-mortem damage, alongside the demographic parameter of interest [78]. For instance, Schraiber (2018) [79] developed a novel maximum-likelihood approach which leverages information from different low-coverage samples from within the same population to infer population-level parameters, such as genetic continuity between ancient and modern populations.

Viera et al (2016) developed a method (ngsF-HMM) to infer matching identical-by-descent (IBD) segments from low-coverage data [80]. To account for the uncertainty, all three genotype likelihoods are integrated over in order to estimate whether or not a genomic region is IBD given the likelihoods. This method showed that there is a substantial gain in power when likelihoods are used compared to genotype calls.

As mentioned in the previous paragraph, there are several other characteristics of ancient DNA which should be accounted for when performing genetic analysis.

Present-day humans contaminating ancient genetic samples is of primary concern as it non-trivial to distinguish between sequencing reads originating from the ancient sample and e.g. present-day individuals performing laboratory analysis [81]. A failure to account for such contamination may lead to underestimating the level of divergence between present-day and ancient samples, as well as the introduction of spurious signals of admixture [41, 82, 83]. In addition to the many precautions taken in the laboratory to reduce the risk of human contamination, such as performing analysis in positive-pressure rooms and intensive irradiation of equipment, several bioinformatics approaches have also been developed to estimate the level of contamination in an ancient sample. For example, a recent method leveraged the fact that contaminating sequences are found on different haplotypes to the genuine ancient sequence and so can be detected through a reduction in local levels of linkage disequilibrium relative to

those found in a reference panel [84]. As contaminant sequences are more likely to carry a derived allele [81], searching the genome for significant deviations from the expected equilibrium percentage of dervied allele (0% at homozygous ancestral and 50% at heterozygous sites) allows for the estimation of local contamination rates [41, 85].

Another aspect of ancient DNA that must be considered is that of post-morterm degradation (PMD). For example, DNA fragmentation (hydrolytic depurination resulting in single-strand breaks) means nearly all ancient DNA fragments are between 40-500bp in length [86, 87]. The presence of substantially shorter DNA fragments increases the risk of mis-aligning reads to the incorrect part of the genome [88].

Further, intermolecular cross-links can form between DNA strands [86] and miscoding lesions, caused by hydrolytic deamination of nucleotides, may result in modifications that cause nucleotides to be misread by DNA polymerases [89]. One consequence of this is that it leads to an excess of spurious C->T substitutions after sequencing [86]. Failing to account for such substitutions (usually termed cytosine deamination) may lead to downstream errors in bioinformatic analyses. Therefore, methods have been developed in order to account for cytosine deamination; for example, the atlas suite of tools which are specifically designed to call variants in low-coverage ancient DNA samples [71]. atlas takes advantage of the fact that cytosine deamination is more likely to occur at the beginning of a sequencing read to model the extent of PMD using an exponential decay function (decaying exponentially with respect to the position on the sequencing read). This provides a likelihood that a given C->T substitution is a true mutation or the result of PMD. Integrating this model into the variant-calling process resulted in a subsantially higher proportion of correctly called genotypes relative to an ancient DNA-naive method (GATK) [71].

In this thesis, I will attempt to mitigate any effects of low-coverage data

on ChromoPainter analysis by implementing an approach similar to that of Viera et al (2016), which modifies the ChromoPainter algorithm to account for genotype likelihoods.

1.4 Combining data from multiple chips

An issue similar to that of low-coverage ancient DNA data stems from the development of a large number of different genotyping arrays. Different cohorts are genotyped on different arrays and sets of SNPs, as different SNPs have different characteristics, such as different frequencies in different populations and associations with different phenotypes. Whilst this has meant a wider variety of questions and populations can be studied, it also makes combining data from across different arrays potentially troublesome, as they often have a small overlap in the SNPs upon which they have been genotyped.

For example, in my thesis, I have worked with at least three genotyping arrays, referred to here as 'Human Origins', 'Hell Bus' and the UK Biobank. Often I have wanted to compare populations on different arrays, such as the African populations on the Human Origins array and UK Biobank individuals on the UK Biobank array. After merging the datasets, the overlap was small, only 70,000 SNPs. This is around an order of magnitude fewer SNPs than is analyse a typical ChromoPainter analysis. Having fewer SNPs may reduce power, as there are fewer pieces of information, and less linkage between each neighbouring SNP.

One solution to the issue of a small number of SNPs would be to impute the remaining SNPs using a reference panel and imputation algorithm such as Beagle [90]. However, it is possible that imputation may cause a bias in the data. If missing genotypes are imputed incorrectly more often from one population than another, this will result in an increased, but spurious genetic similarity between the target and reference population. This may be a particular

issue when analysing populations which are not well represented in imputation reference panels, such as non-Europeans. The nature and magnitude of this bias, however, is yet to be fully understood, particularly in the context of ChromoPainter.

Therefore, this thesis will explore whether is it more desirable to impute the missing positions or to use use a smaller number of overlapping SNPs. Accordingly, in chapter 3 of this thesis, I will explore this question with a case study investigating African ancestry in the UK Biobank dataset.

1.5 Summary of thesis aims

In this thesis I will explore the applicability of ChromoPainter to low-coverage ancient DNA samples and sparsely genotyped data resulting from merged genotype arrays. To do this, I will perform a series of tests on both real and simulated data from present-day and ancient samples and apply my findings to two novel (unpublished) datasets of ancient samples from Bavaria and Czechia.

Specifically, in Chapter 2, I will perform downsampling simulations on five high-coverage ancient genomes to assess the impact of coverage on imputation, phasing and ChromoPainter analysis, and determine the feasibility of extracting haplotype information from sparsely genotyped data in practice. In Chapter 3, I will infer African ancestry across samples in the U.K. Biobank dataset, using sparsely genotyped data resulting from the merge of two different genotyping arrays. I will investigate the potential of using imputation to boost power to infer fine-scale ancestry signatures in U.K. Biobank participants, in terms of how closely related they are to individuals in reference data containing a large number of African ethnolinguistic groups. In Chapter 4, I will analyse unpublished ancient genome data from Bavaria, obtained by collaborators at Mainz University, exploring how genetic patterns varied from the Neolithic to the Medieval Era in a small geographic region. In Chapter 5, I will analyse

unpublished ancient Slavic samples from Czechia, obtained by collaborators at Max Planck Institute for Evolutionary Anthropology, to assess the genetic relationships between Migration Era, Middle age and present-day Slavic-speaking peoples. Lastly, my concluding chapter will summarise my work and key findings, including my recommendations for future haplotype-based studies using low-coverage data and/or combining data from multiple SNP arrays.

Chapter 2

ChromoPainter and ancient DNA

2.1 Introduction

This chapter is related to the use of ChromoPainter on low coverage ancient DNA samples.

First, I will describe the existing methodology, ChromoPainterV2, and then a new version I have developed, ChromoPainterUncertainty, which is designed to mitigate bias related to sequencing coverage.

Next I will perform benchmarking tests on all the steps necessary to analyse low-coverage ancient DNA with ChromoPainter. This includes genotype calling and genotype likelihood estimation with atlas [71], phasing and genotype imputation with GLIMPSE [91], ChromoPainter [19] analysis (copy-vector estimation and PCA) and SOURCEFIND ancestry component estimation [21]. I will also describe some of the existing issues pertaining to low coverage ancient DNA and several considered mitigation strategies. Finally, I will simulate, using present-day samples, ancient samples with variable degrees of missing SNPs in order to determine whether ancient samples of a particular coverage have

enough typed SNPs to retain haplotype information.

2.2 Methods

2.2.1 Description of the ChromoPainter algorithm

As discussed in the introduction, ChromoPainter is a method designed to estimate haplotype sharing between individuals [19]. In diploid organisms such as humans and dogs, ignoring copy-number-variation, each genetic region of an individual is represented by two haplotypes. As input, ChromoPainter requires each individual's data to be 'phased' into these two haplotypes. Phasing refers to the process of determining which alleles along a chromosome were inherited together from the same parent.

In ChromoPainter, sampled individuals are split into 'donor' and 'recipient' haplotypes. It employs the widely-used Li and Stephens copying model [13] to model each recipient haplotype as a mosaic of haplotypes observed in the donor panel. Typically (and throughout this thesis) an individual does not act as a donor to themself, e.g. one of the individual's two haplotypes can not act as a donor for the other haplotype. Unlike the original Li and Stephens model, which uses the product of approximate conditionals (PAC) likelihoods, ChromoPainter reconstructs each recipient haplotype as a mosaic of *all* other donor haplotypes. Here, the term 'copying' can be though of as a genealogical process where haplotypes are reconstructed using the genealogically closest haplotype. The copying model is implemented in the form of a Hidden Markov Model (HMM), with the observed states being the genotype data, and the hidden states being the 'nearest-neighbour' haplotype the recipient haplotype copies from. The emission probabilities are given as the probability of a recipient haplotype copying from a particular donor haplotype, given their respective genotypes.

Consider a donor haplotype d and recipient haplotype r, carrying alleles

x and y, respectively, at position (e.g. a SNP) p. There are two possibilities - either the alleles match between the donor and recipient at p, or they do not. The probability of r copying from d is:

$$\Pr(r = x \mid d = y) = [(1 - \theta) * z_{dr}] + [\theta * z_{!dr}], \tag{2.1}$$

where $z_{dr} = 1$ if $x = y$ and $z_{!dr} = 0$ if $x \neq y$, and θ is the probability of a mutation occurring. The mutation probability θ can be estimated using Watterson's estimator [92], or estimated using an iterative EM algorithm.

The transition probabilities of the HMM, which are the probabilities of a change in the donor being copied when moving from one SNP to another, is guided by a recombination rate map, with higher recombination rates leading to a higher probability of transitioning. Switches between donors are interpreted as changes in ancestral relationships due to historical recombination and modelled as a Poisson process.

In ChromoPainterV2, the input genetic data comes in the form of phased genotype calls (i.e. `1|0`). ChromoPainterV2 produces several different output files. The two which most used in this work are those appended with .chunklengths and .chunkcounts. These matrices are also referred to as 'coancestry matrices'. In the chunklengths matrix, cl, the entry $cl_{d,r}$ gives the total expected proportion of haplotype segments (defined as a contiguous set of SNPs copied from a single donor) that recipient r copies from donor d. Thus, higher values of $cl_{d,r}$ indicate that recipient r and donor d share more recent ancestry. The .chunkcounts matrix instead gives the total number of haplotype segments that recipient r copies from donor d.

In this work, 'copyvector' is used to refer to the vector of chunklengths that a single recipient individuals copies from all donors, or a single row of the coancestry matrix. Throughout, I often define donors as populations, so that

each element of the copy vector is the total amount of DNA that the recipient matches to all individuals from a given donor population.

2.2.1.1 Description of ChromoPainterV2Uncertainty

ChromoPainterUncertainty works in a very similar way to ChromoPainterV2, bar two differences. Firstly, the input data is in the form of an allele probability $0 \leq x \leq 1$, which is given as the probability of observing the alternate allele at that SNP. This value is calculated from the posterior likelihood that an allele has been imputed correctly. This is different to ChromoPainterV2, which uses 'hard' allele calls that only take a value of 0 or 1.

Here, I will show how it is possible to incorporate the uncertainty in impute genotype calls into the ChromoPainter input. Consider the following example: we have a phased genotype in the form `0|1`, corresponding to the reference allele on the first haplotype and the alternative allele at the second haplotype. I define G as the sum of the genotypes at a SNP; in this case $G = 0 + 1 = 1$. As GLIMPSE, the imputation and phasing algorithm I will use for this work, provides hard genotype calls, G can be calculated directly.

We also have a posterior genotype likelihood, in the form $GL(p_0, p_1, p_2)$, where p_i is the posterior probabilities that the true genotype is i. Genotype probability dosage, D, is the expected total number of copies of the alternate allele given GL. D can be calculated as $p_1 + [2 * p_2]$. We can calculate U, the uncertainty as $U = |G - D|$. Then, we can assign a probability to each allele; if the allele is `1` then the allele likelihood is simply $1 - U$ and if the allele is `0` then the allele likelihood is $0 + U$. Therefore, when there is no uncertainty in the genotype call, the allele probability will be either 0 or 1. When there is uncertainty, the allele probability will take a value $0 \leq x \leq 1$, with more uncertain genotypes tending towards allele probabilities of 0.5.

The second difference is the incorporation of the allele probability into the

emission probability of the HMM. As before, consider a donor d and recipient r at SNP p. Now we let r_x be the probability that the recipient haploid r carries the alternative allele, with d_x the probability the donor haploid carries the alternative allele.

$$\begin{aligned} p(r_x|d_x) = (1-\theta) * [r_x * d_x + (1-r_x) * (1-d_x)] \\ +\theta * [r_x * (1-d_x) + (1-r_x) * d_x] \end{aligned} \tag{2.2}$$

Note that above (3) reduces (1) if $d_x = \{0, 1\}$ and if $r_x = \{0, 1\}$, i.e there is no uncertainty in the calls.

2.2.2 Generation of downsampled genomes

I created a set of 'downsampled' ancient genomes in order to explicitly quantify the effect of coverage on each stage of the ChromoPainter analysis. I took five high coverage genomes and for each, removed a random subset of reads from the `.bam` file in order to reduce the coverage to a target level. I then performed each stage of a typical ChromoPainter analysis, e.g. mimicking the analyses of new ancient DNA samples I describe in chapters 4 and 5, on the full coverage and downsampled genomes.

Five high coverage ancient genomes were downloaded in the form of aligned `.bam` files from the European Nucleotide Archive:

1. Yamnaya – Yamnaya Bronze Age steppe-pastoralist [93]
2. UstIshim – Siberian Upper Palaeolithic hunter-gatherer [94]
3. sf12 – Scandinavian Hunter-Gatherer [95]
4. LBK – early European farmer from the *Linearbandkeramik* culture from Stuttgart, Germany [51]

5. Loschbour – 8,000 year-old hunter-gatherer from Luxembourg) [51]

These samples were chosen due to their high original coverage (> 18x), and because they represent some of the ancestries present in Western Eurasia over the past 40,000 years.

Each original full-coverage `.bam` file was processed using the atlas (version 1.0, commit f612f28) pipeline [71] (`https://bitbucket.org/wegmannlab/atlas/wiki/Home`). First, the validity of each file was assessed (i.e. ensuring that each `.bam` file was not malformed in any way) using ValidateSamFile command from PicardTools [96]. atlas is a suite of software designed for processing low-coverage ancient DNA and was chosen following the recommendation of Hui et al (2020) [63], as it explicitly accounts for post-mortem damage (PMD) patterns in ancient DNA. The most common form of PMD is C-deamination, which leads to a C->T transition on the affected strand and a G–>A transition on the complimentary strand.

I then downsampled each full-coverage genome using the `atlas downsample` task, resulting in a `.bam` file with coverages 0.1x, 0.5x, 0.8x, 1x, 2x, 3.5x, 5x, 10x and 20x per individual.

For each full coverage and downsampled `.bam` file, I estimated post-mortem damage (PMD) patterns using the `atlas estimatePMD` task. Recalibration parameters were then estimated using the atlas `atlas recal` task. Finally, both the recalibration and PMD parameters were given to the `atlas callNEW` task which produces genotype calls and genotype likelihood estimates for each downsampled and full coverage `.bam`. For this stage, I made calls at the 77,818,345 genome-wide positions present in the phase 3 thousand genomes project [97]. This was done to reduce the risk of calling false-positive (i.e. falsely polymorphic) genotypes in the aDNA samples.

2.2.3 Generation of ancient literature samples

I also generated a set of ancient samples from the literature to use as donors in the ChromoPainter analysis.

This dataset consists of 918 other ancient samples from the literature given in Appendix section A.1. These samples were of variable coverage, ranging from 0.002-72x coverage, and chosen because of their previously reported relevance to understanding past ancestry patterns in European populations like those analysed in chapters 4 and 5. These 918 consist of all samples given in Table A.1 were processed in an identical way to the downsampled target individuals described in the previous section, other than they were not downsampled.

2.2.4 Imputation and phasing - GLIMPSE

Genotype imputation and phasing are two important steps for processing low-coverage ancient DNA. Low coverage (<1x) samples typically lack enough read information to make accurate genotype calls at most positions in the genome, and often do not contain any reads at many positions [98]. Therefore, it can be helpful to use external information from a high-coverage reference panel in order to improve the accuracy of genotype calls and phasing, reducing the impact of errors on downstream analyses [91].

Three different characteristics are desirable for an imputation algorithm in this context. Firstly, it should take genotype likelihoods as input. This is because genotype likelihoods allow for flexible representation of the possible genotypes at a particular position, particularly when there may not be enough coverage to make a hard genotype call. Secondly, it should emit posterior genotype-probabilities which, when accurately calibrated, give the probability that a particular genotype call is correct. This is necessary for estimating the uncertainty values, described in section 2.2.1.1, needed for ChromoPainterUncertainty analysis. Thirdly, the algorithm must be able to complete in a reasonable

running time when using a large number of samples and high number of SNPs. Using a large number of densely positioned SNPs (e.g. such as the approximately 77 million identified in the 1000 Genomes Project) increases the useful linkage-disequilibrium information between each SNP, and it is well-established that increasing the number of individuals used in imputation/phasing reference panels improves accuracy [25, 91, 99, 100].

Two programs, Beagle 4.0 [101] and GLIMPSE [91] fulfil the first and second criteria above, but only GLIMPSE runs quickly enough to analyse SNPs with sequence-level density. GLIMPSE offers up to 1000x reduction in running time compared to Beagle 4.0 [91], so I chose to use this algorithm for the imputation and phasing steps.

Phasing and imputation ideally requires a reference panel of high-coverage present-day individuals. I used the 1000 Genomes Project dataset re-sequenced to 30x average coverage, which contains 3202 individuals from 26 worldwide populations [102]. A description of the processing of this reference dataset can be found in Appendix A.2.

I merged together i) the full coverage individuals, ii) downsampled individuals and iii) 918 ancient samples from the literature into a single bcf file using bcftools (version 1.11-60-g09dca3e) [103] to act as the samples for GLIMPSE to phase. Here, 'target' refers to the individuals being imputed/phased and 'reference' refers to the reference panel.

Following the GLIMPSE tutorial (`https://odelaneau.github.io/GLIMPSE/tutorial_b38.html`), I first used `GLIMPSE_chunk` to split up each chromosome into chunks, keeping both `-window-size` and `-buffer-size` to 2,000,000 base pairs, which is their default settings. I used the b37 genetic map supplied by GLIMPSE for the `-map` argument. Across all chromosomes, this produced 936 chunks that are on average 2.99Mb long.

GLIMPSE then imputed each chunk separately, using `GLIMPSE_phase` us-

ing the same 1000 genomes dataset as a reference and default settings. This stage both imputes missing genotypes and generates a set of haplotype pairs which can be sampled from in a later step to produced phased haplotypes. `GLIMPSE_ligate` then merges the imputed chunks back to form single chromosomes using the default settings. I then used `GLIMPSE_sample` to produce a .vcf with phased haplotypes sampled for each individual, again using default settings. Consequently, the output of GLIMPSE is i) unphased genotype calls with posterior genotype likelihoods and ii) phased haplotypes.

It is important to note that GLIMPSE leverages information from individuals that have been imputed, 'absorbing' them into the reference panel. For example, if there were 100 target samples and 1000 reference samples, each target is phased in turn and then absorbed into the reference panel, so that there would be 1001 reference samples when the second target individual is imputed. This makes it necessary to avoid including the same sample, downsampled to different coverages, in the same set of targets for one imputation run, in order to avoid the confounding effect of allowing an individual to act as the reference to itself. For example, including Loschbour at 0.1x and 10x coverage could mean it imputed itself, a situation which would never occur in reality.

2.2.5 Estimating imputation sensitivity and specificity

I used rtg-tools-3.11 [104] and the `vcfeval` task to estimate the sensitivity and specificity of imputation in the downsampled individuals. Here, 'baseline' (i.e. the truthset) is defined as the genotype calls in the full coverage individual and the 'calls' as the genotype calls in the downsampled individual. Sensitivity and precision are defined as:

$$sensitivity = \frac{V_{call} - FP}{V_{call}} \tag{2.3}$$

$$precision = \frac{V_{baseline} - FN}{V_{baseline}} \tag{2.4}$$

A 'variant' is considered to be a SNP with a genotype that is either `0/1` or `1/1`, with $V_{baseline}$ and V_{call} the number of variants called in the full coverage and downsampled genomes, respectively. False negatives (FN) are where a variant is called in the full coverage genome but not in the downsampled genome. False positives (FP) are cases where a variant is called in the downsampled genome but not in the full-coverage genome.

V, or true-positive, is the number of events where a variant position (i.e. a SNP with a genotype that is either `0/1` or `1/1`) is detected in either the full coverage ($V_{baseline}$) or downsampled ($V_{baseline}$) sample. FN is the number of times that a variant position is called in the full coverage sample and not the downsampled sample. Conversely, FP is the number of times a variant position is called in the downsampled sample and where the same SNP in the full coverage sample is invariant (i.e. `0/0`).

2.2.6 ChromoPainter analysis

It is important to understand the effect of sequencing coverage on the accuracy of ChromoPainter copyvector estimation. A 'copyvector', c_r, is a vector of length D, where each entry gives the total length of genome that recipient individual r most closely matches to each of the D donor individual/populations. I sometimes refer to 'normalised' copyvectors; this simply refers to where each entry of c_r is divided by the sum of all entries, scaling the copyvector to sum to 1.

I painted each downsampled and full coverage ancient individual using a set of 124 ancient individuals, hereafter referred to as the 'standard set', selected because they had a sequencing depth greater than 2x. I compared the copyvectors for the same individual at each level of downsampling, to the

same individual at full coverage. For example, I compared the copyvector of Yamnaya at 0.1x to the copyvector of the same Yamnaya sample at full coverage. A high correspondence, measured by r-squared for example, between the copyvectors of the full coverage and downsampled individual suggests less effect of coverage.

To prepare the data for ChromoPainter, I merged the .vcf containing the posterior genotype likelihoods of i) downsampled, ii) full coverage and iii) 124 ancient samples from the literature together, and did the same for the .vcfs containing the phased haplotypes. I combined the posterior genotype likelihoods with the phased alleles to generate allele likelihoods (described in section 2.2.1.1) in ChromoPainter-uncertainty format, in addition to per-position recombination rate files. This was performed for each chromosome in turn using my own script (`https://github.com/sahwa/vcf_to_ChromoPainter`).

I next used ChromoPainterUncertainty to perform the painting. I assigned the standard set individuals as donors and all downsampled, full coverage and standard set as recipients. The 'standard set' samples from the literature were included in order so that they can be used a surrogates in later SOURCEFIND analysis.

I also performed an identical analysis, but using ChromoPainterV2 and hard genotype calls.

This painting produced a chunklengths matrix for each chromosome which were merged using chromocombine-0.0.4 (`https://people.maths.bris.ac.uk/~madjl/finestructure-old/chromocombine.html`). The resulting chunklengths matrix thus gives the total length of genome in centimorgans that a recipient most closely matches to each donor individual.

2.2.7 ChromoPainter Principle Component Analysis

Principle Component Analysis (PCA) can be used to reduce the underlying structure in the chunklengths coancestry matrix to two dimensions, thus allowing it to be more easily visualised. As individuals cannot paint themselves, the diagonals of each coancestry matrix contain zeros. Therefore, I performed PCA using the fineSTRUCTURE library `https://people.maths.bris.ac.uk/~madjl/finestructure/finestructureR.html`.

2.2.8 SOURCEFIND

The chunklengths coancestry matrix produced by ChromoPainter contains information about the estimated length of genome a recipient most closely matches a given donor individual or population. However, incomplete lineage sorting, where alleles segregate in a way that is discordant to the true phylogeny reflecting the orders in which populations split from one another, means that there are regions in the genome where a recipient individual most closely matches a reference individual that is not from their own population. For example, an individual from France copies non-zero amounts from African donors, despite not having any African ancestry through recent admixture. Furthermore, unequal donor population sizes may bias the aggregated amount copied to a given population.

Therefore, to account for these issues when estimating ancestry proportions, it is necessary to run an additional step, SOURCEFIND [21]. Simulations have shown that SOURCEFIND ancestry proportions correspond well to simulated truth-set values [21]. The ancestry proportions produced by SOURCEFIND should be interpreted as the proportion of ancestry that each individual/population shares most recently with each surrogate. This need not necessarily imply an admixture event; for instance, you might expect *France* to have ancestry recently related to both *Germany* and *Spain* due to isolation-by-distance rather than admixture.

SOURCEFIND models each target copyvector as a linear mixture of copyvectors from a set of surrogate groups, inferring the proportion of ancestry for which the target individual is most recently related to each surrogate group. The parameter space of surrogate ancestry proportions is explored using a Markov chain Monte Carlo algorithm, where the ancestry proportions are updated using a Metropolis-Hastings step. The output of SOURCEFIND for each target individual is therefore an $n * p$ matrix, where n is the number of MCMC samples and p is the total number of surrogate groups.

To test for the effect of coverage on the proportions estimated by SOURCEFIND, I performed two separate analyses, both using the downsampled and full coverage individuals as targets. The first uses three surrogate populations (Yamnaya, Western Hunter-Gatherer and Anatolia Neolithic Farmer), and the second uses an expanded list of 37 surrogate populations. I chose the first set of three surrogates, as these are typically used in ancient DNA analysis to obtain a 'broad' overview of the ancestry of a European individual, as it has been shown that central Europeans within the last 10,000 years can be well modelled as a mixture of those three groups [51, 105]. Note, this does not mean that there was not admixture from other sources, but that a majority of ancestry of ancient central Europeans can be derived from these sources. This stands to act as a relatively straightforward test case, since the three populations are highly genetically differentiated from one another.

For all runs of SOURCEFIND, I used 1,000,000 iterations, of which 50,000 were designated as burn-ins, and then samples were taken every 50 iterations. 2,000,000 iterations were chosen because my previous tests show that is the minimum necessary to provide reasonably confidence of convergence within reasonably running time (Appendix section E.0.1). The rest of the parameters were left as default. Ancestry proportions and credible intervals group were estimated using the CODA R library [106].

2.3 Pre-post GLIMPSE and linked/unlinked PCA test

I wanted to determine at what stage of the analysis pipeline low coverage samples (0.1x) become significantly diverged from the other downsamples when plotted on a PCA. For instance, it may be that the bias is introduced in the imputation stage. To test this, I performed a set four PCAs on all downsampled and equivalent full coverage samples and a set of present-day individuals shown in Table 2.1.

For both the ChromoPainter PCAs, in order to account for the zeros on the diagonals of each coancestry matrix, I used the fineSTRUCTURE R library `https://people.maths.bris.ac.uk/~madjl/finestructure/finestructureR.html`.

The four PCAs were as follows:

1. **Pre-GLIMPSE** Using the genotypes generated by atlas, but before imputation with GLIMPSE, I projected all downsampled ancients of all coverages onto the present-day populations using the eigenstrat library. [107].

2. **Post-GLIMPSE** Using the GLIMPSE generated imputed genotypes generated by atlas, I projected all downsampled ancients of all coverages onto the present-day populations using the eigenstrat library.

3. **ChromoPainter - unlinked** I performed an 'all-v-all' unlinked ChromoPainter painting, using all populations in Table 2.1.

4. **ChromoPainter - linked** I performed an 'all-v-all' unlinked ChromoPainter painting, using all populations in Table 2.1.

Bias present in PCA (2) but not (1) indicates it has been introduced in

Population	Number of samples
HB:croatian	19
HB:cypriot	12
HB:french	28
HB:german	30
HB:germanyaustria	4
HB:greek	20
HB:hungarian	19
HB:irish	7
HB:lithuanian	10
HB:mordovian	15
HB:northitalian	12
HB:norwegian	18
HB:polish	17
HB:romanian	16
HB:scottish	6
HB:siciliane	10
HB:southitalian	18
HB:spanish	34
HB:tsi	98
HB:tuscan	8
HB:welsh	4
HB:westsicilian	10

Table 2.1: Population labels and sample sizes of populations included in the pre-post GLIMPSE and linked/unlinked PCA test. All samples are from the Hellenthal and Busby dataset, described in A.4.

the imputation stage. Similarly, bias present in (4) but not (3) suggests that including linkage information introduces bias in low coverage samples.

2.4 Reducing SNP count

One way to mitigate mitigate coverage-related bias would be to exclude imputed SNPs which have a low probability of being imputed correctly or restricting analysis to non-imputed SNPs above a certain coverage.

However, reducing the total number and or density of SNPs used in a painting may reduce the accuracy of the estimated copyvectors. All other

things being equal, there is less linkage information between two SNPs which are separated by a larger genetic distance. Therefore, it is necessary to precisely determine what effect reducing the number of SNPs has. In particular, we would like to know the minimum number and density of SNPs required to retain the advantages of haplotype-based methods over unlinked methods.

Using data from the People of the British Isles (POBI) project, previous work showed it is possible to distinguish between British individuals from neighbouring counties Devon and Cornwall using the fineSTRUCTURE algorithm, but not using unlinked methods (ADMIXTURE [108]) [31]. Therefore, determining whether it is possible to distinguish between individuals from Devon and Cornwall acts as a good test case for reducing SNPs. In particular I tested how many SNPs can we remove before we lose the ability to distinguish between these two populations.

The original POBI dataset contains 2039 individuals from 33 populations from across England, Northern Ireland, Wales and Scotland, genotyped at 452 592 SNPs. Details of the data preparation for this dataset can be found in Appendix section A.4.

Using the `shuf` unix command, I randomly reduced the total number of SNPs down to only the following percentages: 0.2%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%. SNPs were removed from the .vcf files using `bcftools -view`.

For each target level of reduced SNPs, I painted all individuals from Devon and Cornwall using a 'leave-one-out' approach. I then combined the resulting chunklengths matrices across all chromosomes and combined copyvectors columns by donor group, so that each individual was represented by a K-vector of values, with element k denoting the proportion of DNA that person matched to any haploid in donor group k.

2.5 Direct imputation test

To explicitly test the effect of imputation on the copyvectors estimated by ChromoPainter, I created a dataset which simulated a typical imputation scenario; imputing SNPs after merging two datasets with a low SNP overlap. In particular I did this in a way to mimic a real analysis on ancient samples of approximately 0.15 coverage (determined from empirical data), which have approximately 70,000 SNPs out of 500,000 covered by at least a single read.

I took the Human Origins dataset (described in Appendix section A.3), containing 560,240 bi-allelic SNPs and submitted the reduced dataset to the Sanger Imputation Service (`https://www.sanger.ac.uk/tool/sanger-imputation-service/`). The Sanger Imputation Service uses Eagle2 [109] and the Haplotype Reference Consortium as a reference to impute missing variants. Once the data had been imputed, I subsetted the data back to the original set of 560,240 SNPs. I therefore had a dataset which contained 70,000 non-imputed SNPs and 490,240 imputed SNPs. This is hereafter referred to as the 'imputed dataset'. 70,000 non-imputed SNPs was chosen because that is the number of SNPs which overlap between two datasets in Chapter 3 and thus represents a realistic case-study.

For both the imputed dataset and original Human Origins dataset, I performed an all-v-all painting and combined data across chromosomes. An 'all-v-all' painting is where each individual is painted in turn by all other individuals, resulting in an n-by-n coancestry matrix, where n is the number of individuals analysed.

2.6 Results

2.6.1 Imputation accuracy

To estimate how accurately GLIMPSE imputes genotypes in ancient samples of differing coverages, I estimated the sensitivity (Fig. 2.1) and precision (Fig. 2.2) of genotype imputation using rtg-tools [104]. This approach compares genotype calls at each position in each downsampled individual after imputation to the same individual at full coverage without imputation.

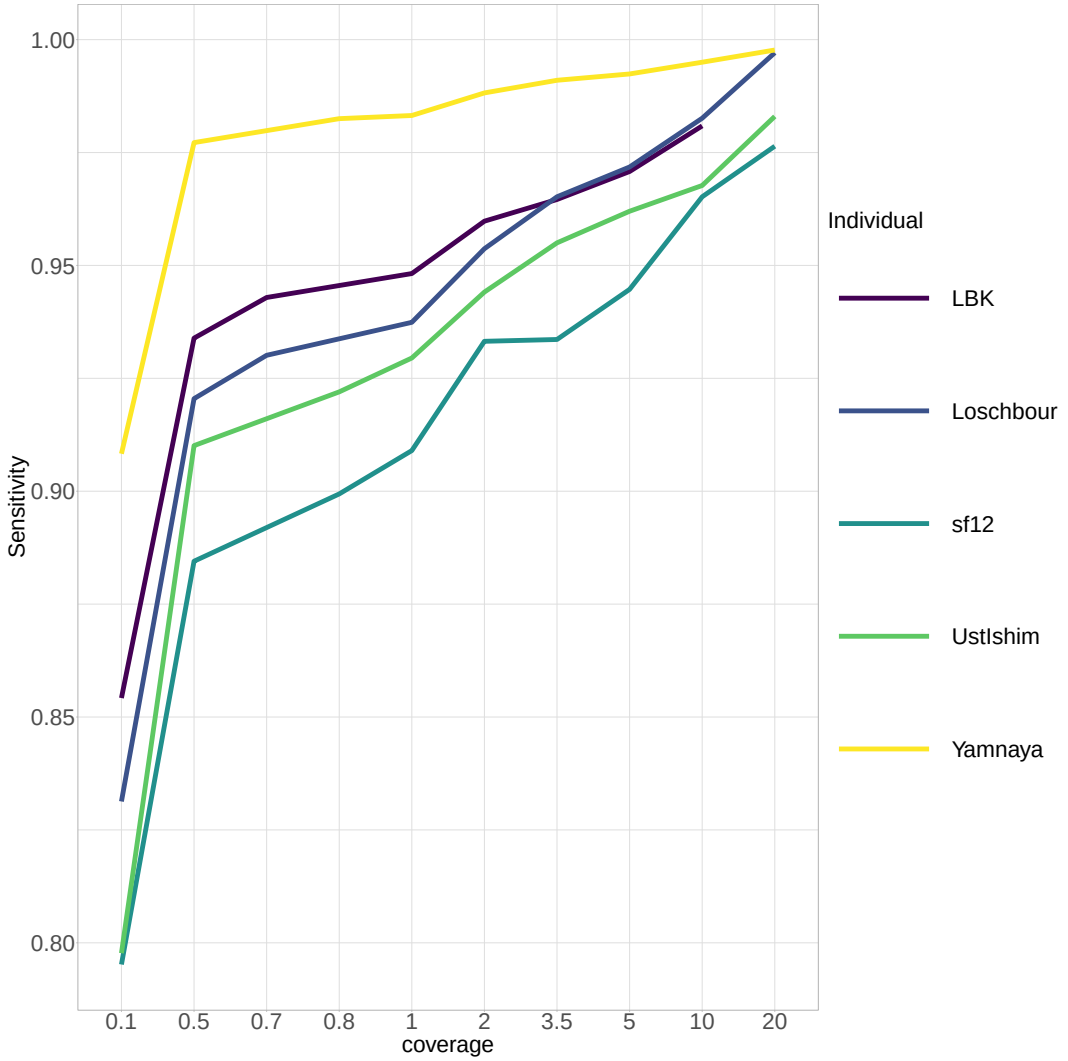

Figure 2.1: Sensitivity of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

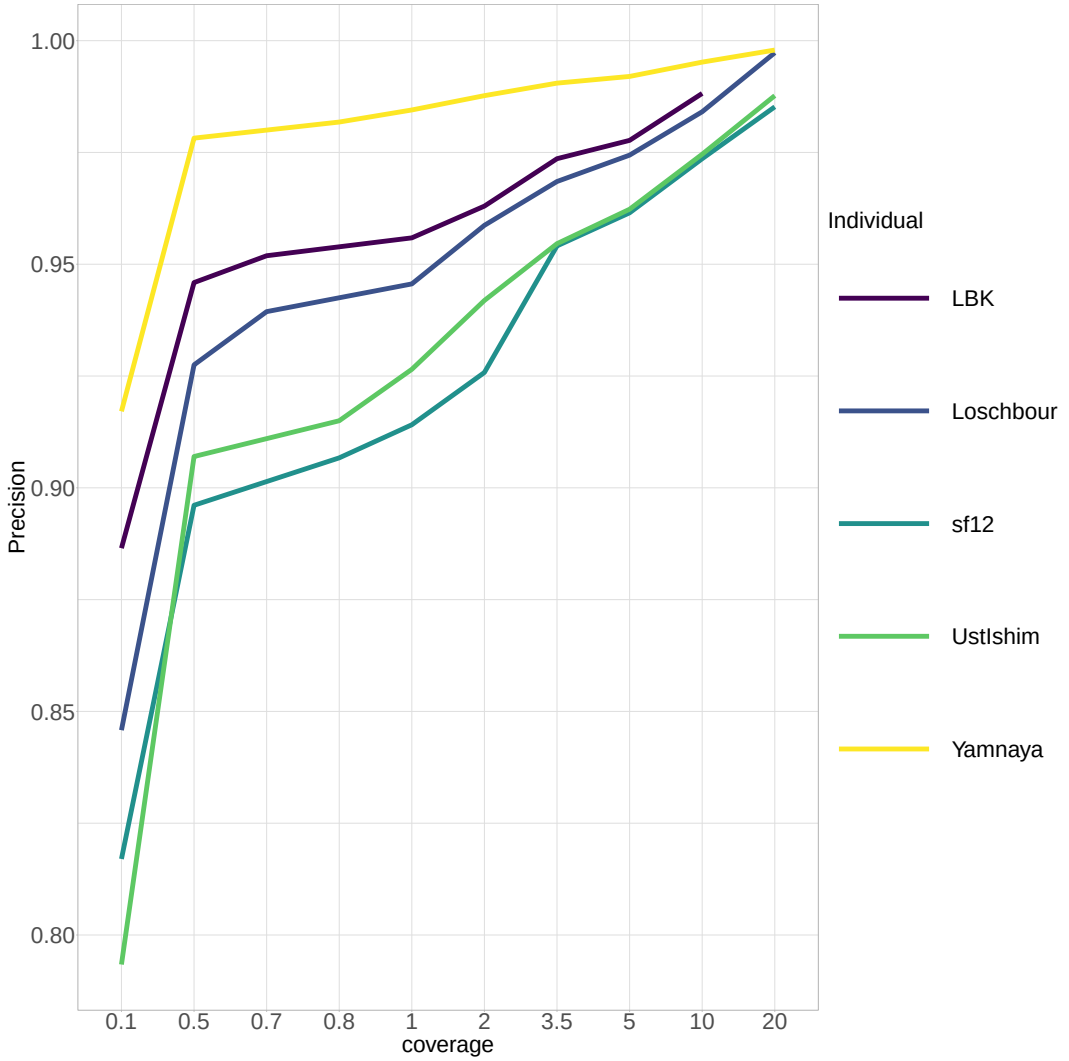

Figure 2.2: Precision of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

As expected, both the overall sensitivity and precision of imputation fell with coverage, with a particularly sharp drop-off in both metrics between 0.5x and 0.1x coverage. Whilst I did not investigate this, other studies have shown the probability of any one SNP in an sample being correctly imputed depends strongly on the frequency in the reference panel [63, 91]. In particular, alleles which are rare in the reference panel are less likely to be imputed correctly.

Different downsampled individuals varied in the precision and sensitivity of genotype imputation. At all coverages, Yamnaya had the both the highest

sensitivity and precision. This may be because the imputation reference panel contains a high proportion of present-day Europeans, who have a relatively higher porportion of recent Yamnaya-like ancestry relative to e.g. Hunter Gatherer-like ancestry [110]. Many studies in present-day individuals have shown that imputation accuracy increases when more haplotypes which are close to the target individual are found in the reference panel [25, 99]. On the other hand, the sample Ust'Ishim is known to have contributed very little genetic ancestry to present-day populations [111] and may therefore have fewer closely matching haplotypes in the reference panel, and a correspondingly lower imputation accuracy.

Imputation accuracy may also be related to demographic history. Populations which are known to have smaller effective population size, such as Western-Hunter Gathers, also contain longer tracts between individuals which are identical by descent (IBD) [112] and fewer heterozygous positions. As imputation relies on matching IBD tracts between individuals, imputation accuracy increases where individuals share more IBD [113]. Additionally, switch-errors during the pre-phasing step of imputation may harm imputation accuracy, so a reduced density of heterozygous positions may result in increased accuracy.

2.6.2 Phasing accuracy

I also used rtg-tools to calculate the number of phased heterozygous genotypes where the downsampled individual has the same phasing as the full coverage individual (Fig 2.3). I note that this should not be considered to be the same as estimating the switch error rate, since we do not know that the phasing in the full-coverage individual is the true phase. However, this can be used as a rough proxy for switch errors, since it is known that phasing in lower coverage individuals is likely to be less accurate than those in the high coverage individuals [91].

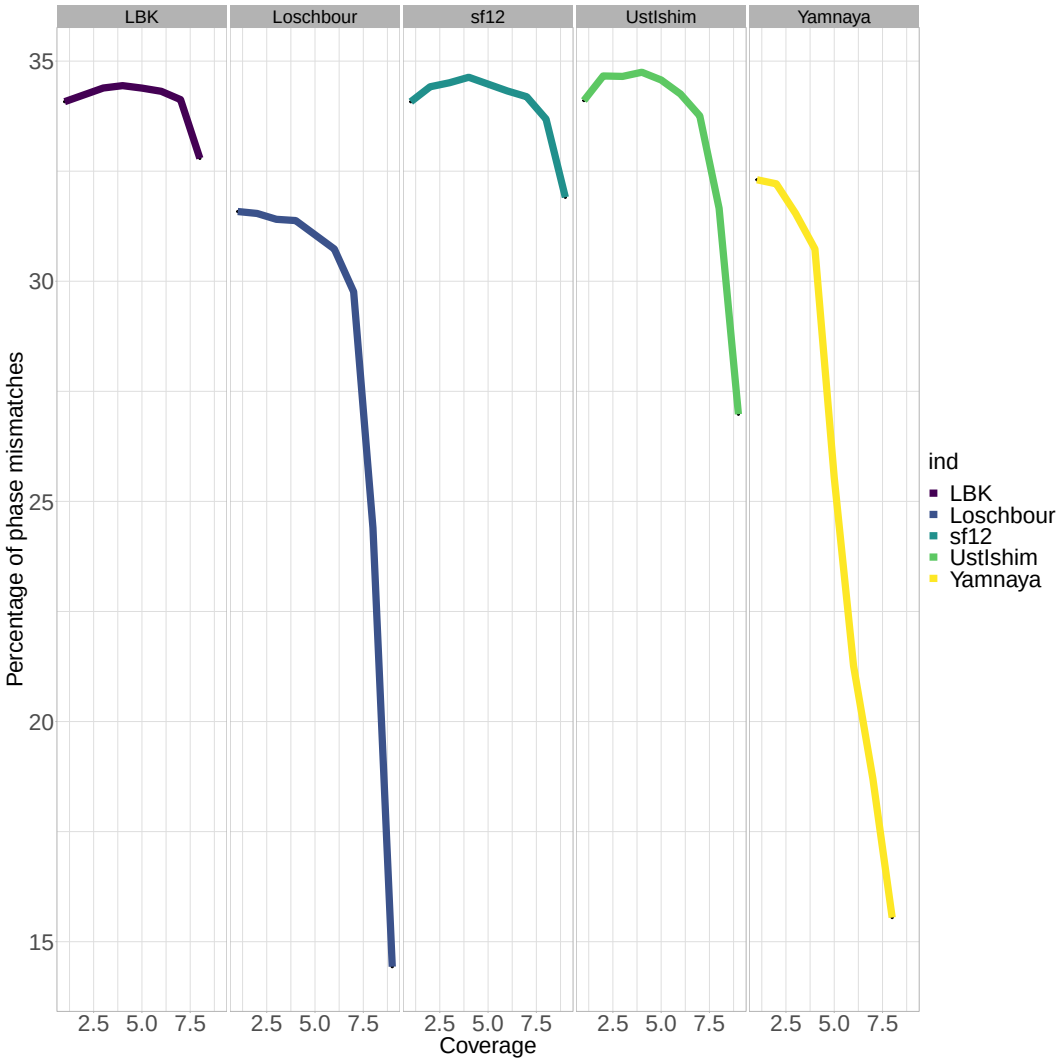

Figure 2.3: Percentage of phased genotypes which agree with the same full-coverage sample for each individual and each level of downsampling. Genotypes with phase deemed unresolvable by rtg-tools were excluded from the calculations. Note that these numbers are given as incorrect / (incorrect + correct - unresolved) and so values are in part driven by the relative heterozygosity of each sample.

2.6.3 Validating posterior probability calibration

GLIMPSE estimates genotype probabilities at each SNP within each individual, giving the posterior probability that a given genotype within a single individual is correctly called. I assessed how well-calibrated these probabilities are in the Yamnaya 0.1x downsampled individual, using the maximum genotype likelihood at each of the approximately 77 million positions which were processed by GLIMPSE. A high $max(GL)$ for a particular genotype (i.e. 0.99) corresponds to a high confidence in the genotype. Alternatively a flat $max(GL)$ (i.e. 0.33) corresponds to no information about the genotype.

I split the genome into 10,000 equally-sized bins according to $max(GL)$. For each bin, I calculated both the proportion of SNPs which were correctly imputed (i.e. that matched the same high coverage individual) and the mean $max(GL)$ (Fig. 2.4). If the genotype probabilities are well calibrated, we would expect to see a clear positive linear relationship between $max(GL)$ probability and the probability that genotype matches the full-coverage sample.

The probabilities are well calibrated (r-squared = 0.981) and could therefore be useful for downstream analysis. It should be noted that they are slightly conservative, in that a majority of the points in Fig. 2.4 are above the line of equality. For example, the mean proportion of correct genotypes within all bins where $0.73 < max(GL) < 0.76$ was 82%. I performed the same analysis using different samples at different levels of coverage and the results were qualitatively similar (Supplementary Figure. D.1).

2.6.4 ChromoPainter analysis

To assess the impact of coverage on ChromoPainter analysis, I merged the dataset of downsampled individuals with the 'standard set' of ancient reference individuals (124 ancient samples > 2X coverage) and performed an 'all-v-all' painting of the merged dataset, which separately paints each individual as a

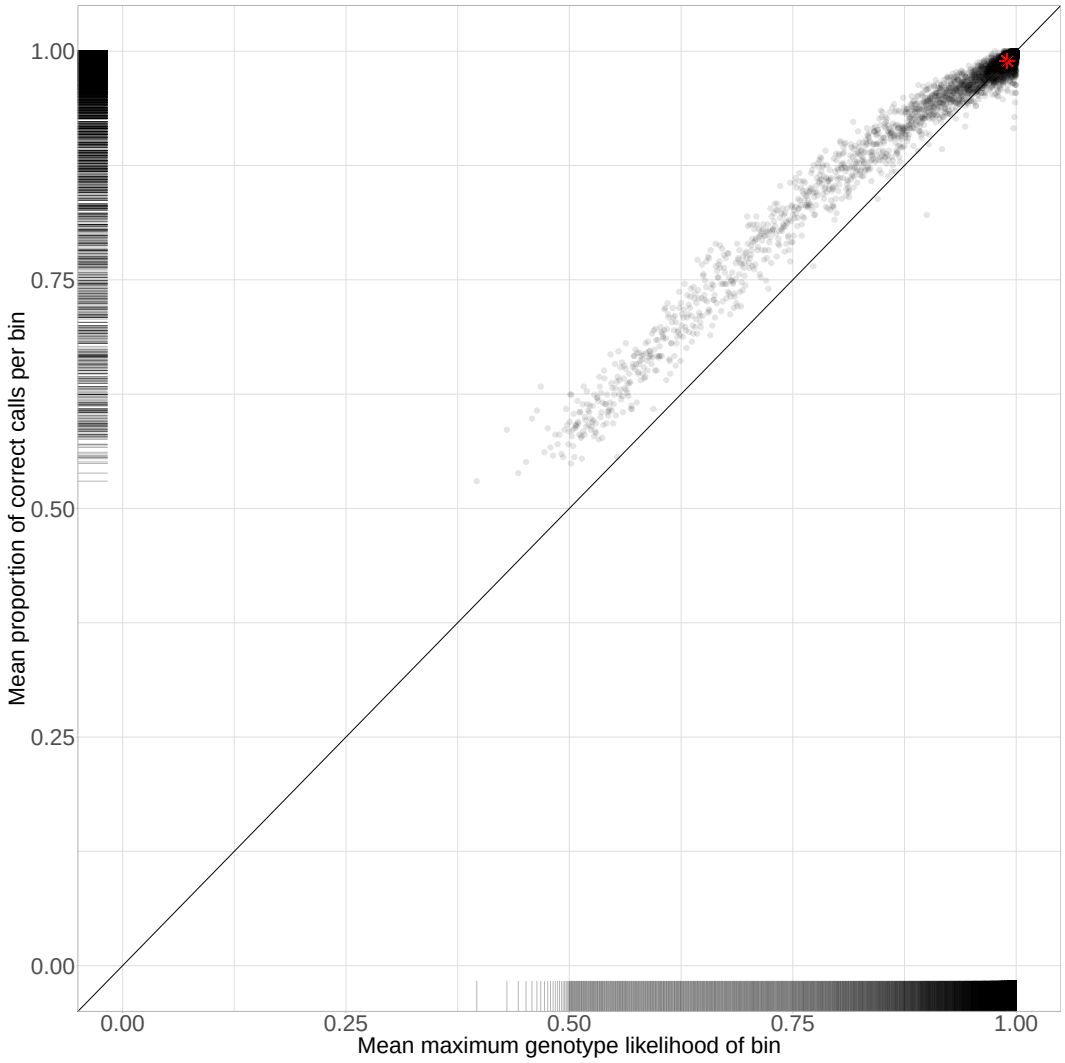

Figure 2.4: Relationship between genotype likelihood and probability of genotype call being correct for Yamnaya downsampled to 0.1x coverage. Genome binned by maximum posterior genotype likelihood and mean maximum posterior genotype likelihood (x-axis) and proportion of correct calls calculated per bin (y-axis). Rugs on each margin show the distribution of x and y values. Black line is $y = x$.

recipient using all other individuals in the dataset as donors. The 'all-v-all' painting was necessary to paint the 124 'standard set' of individuals against one another so that they can act as surrogates in later SOURCEFIND analysis.

I was interested to see whether a downsampled individual and full coverage had similar copyvectors, or in other words, whether they matched similar amounts to the same donor individuals. To do this, I estimated TVD between the copyvectors of the full coverage and downsampled individuals. TVD is a distance metric which gives a measure of dissimilarity between two copyvectors.

Fig. 2.5 displays the relationship between copyvectors for each downsampled individual and the corresponding full coverage individual for both 0.1x and 0.5x coverage. Each individuals' copyvectors were estimated using the same set of ancient samples as donors.

As expected, the TVD between the full-coverage and downsampled copyvectors decreased with coverage. The 0.1x genome had a substantially increased TVD, similar to the much reduced imputation accuracy. For each of the genomes downsampled to 0.1x, a particular difference to the 0.5x downsampled genomes is that the lowest contributing donors contribute more to the 0.1x downsampled genome than to the full coverage genome and that the highest contributing donors contribute less to the 0.1x genome than they do the full coverage genome. Put in other words, the copyvectors at 0.1x are tending towards becoming more 'flat', or copying the same amount from each donor individual.

This can also be seen as 'regressing to the prior'. In this case, the prior is copying an equal amount to each donor individual. This can be visualised explicitly by calculating TVD between each downsampled genome and a flat prior, a vector of length D, where D is the total number of donor individuals and each element of D is equal to 1 / D (Fig. 2.6). This clearly shows the reduced TVD to the flat copyvector for the 0.1x individual relative to other

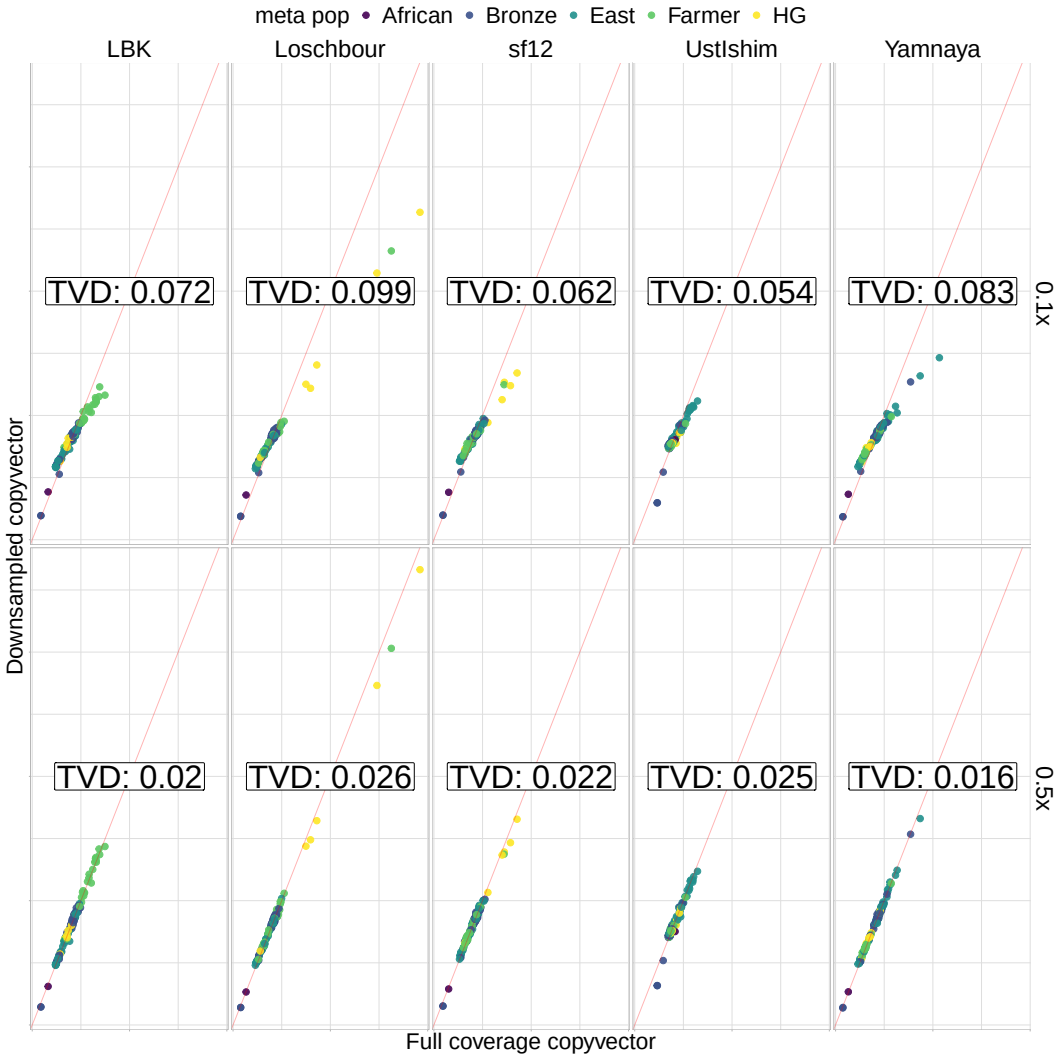

Figure 2.5: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to each of 125 ancient individuals (dots). Results are shown for 0.1x (top row) and 0.5x (bottom row) downsampled genomes. Points coloured by manual assignment to broad-scale populations. Red line is line of equality ($y = x$). x and y units are normalised copying values and thus removed for clarity.

coverages. In later sections, I will discuss whether this is 'noise' or 'bias' induced by imputation, i.e. whether copying is regressing to the prior in a similar manner for all samples.

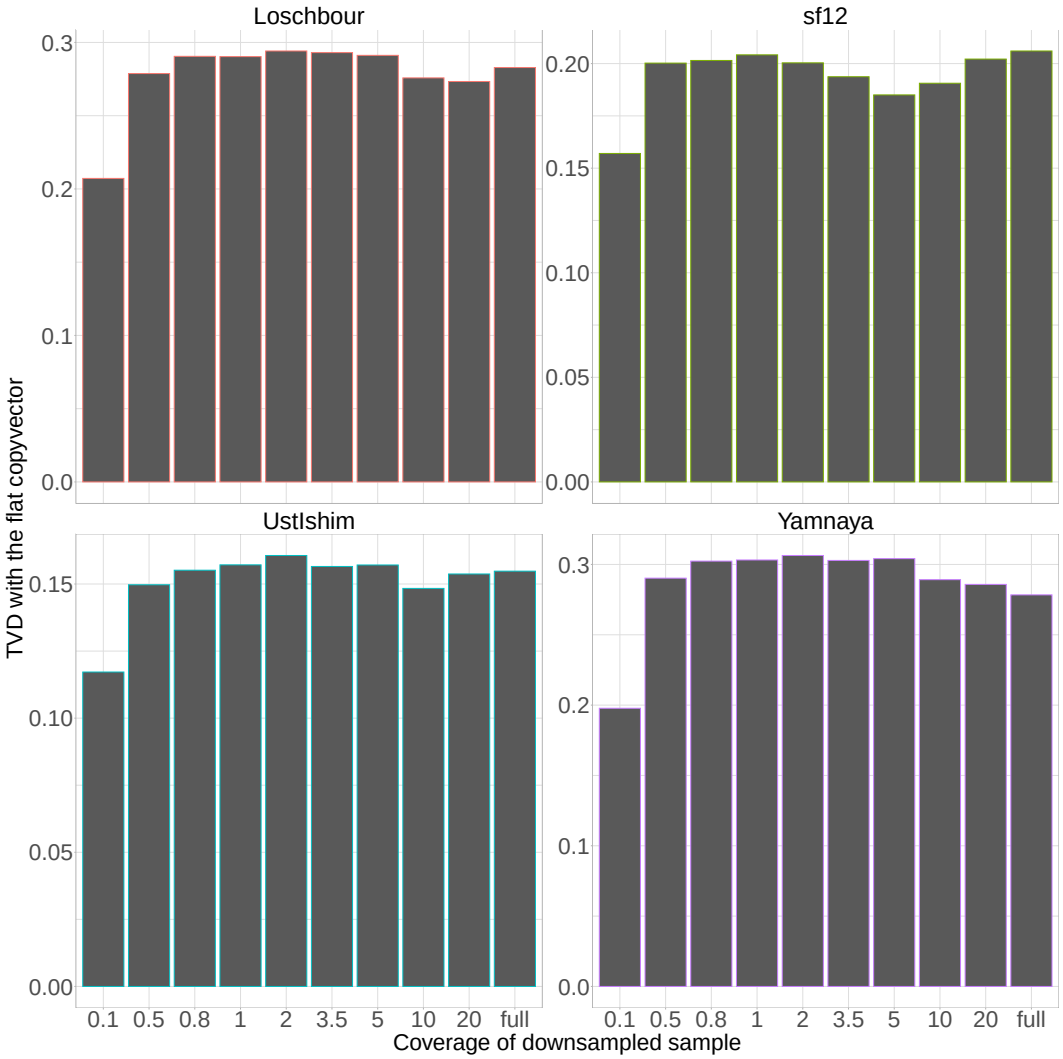

Figure 2.6: TVD (metric of copyvector dissimilarity between two individuals) between each downsampled ancient individual and a flat copyvector. Flat copyvector equivalent to a vector of length N where each element $= 1/N$.

I also considered the effect of coverage on the copyvectors estimated when using present-day individuals from the 1000 genomes project as donors (Fig. 2.7). Painting ancient samples using present-day donors is often useful, particularly with more recent ancient samples, as there may not be enough relevant ancient samples to paint the ancients with. I merged the downsampled and full coverage

ancient individuals with the thousand genomes dataset (described in detail in Appendix section A.2). As was the case with the all-v-all ancients painting, the TVD between copyvectors was highest for the 0.1x individuals. However, the copyvectors show a strong correlation / low TVD for 0.5x individuals.

It should be noted that utility of painting different ancient individuals with a modern reference panel depends on the ancestry and age of the ancient sample. The spread of points along the $y = x$ line in Fig. 2.7 shows how much a particular ancient recipient preferentially copies more from particular modern population over others. LBK, for example, has points which are spread evenly across $y = x$, showing that they copy much more from some populations than others, suggesting modern populations are good for distinguishing this particular ancient sample. On the other hand, the points for Ust'Ishim are shrunk towards lower values of $y = x$, showing that the copyvector is relatively flat and that it does not preferentially copy from some populations to the same degree that LBK does. This is consistent with findings that UstIshim did not contribute ancestry towards present-day populations [94]. Accordingly, relatively less useful information is obtained from painting Ust'Ishim with a modern reference panel than LBK.

Principle component analysis (PCA) is a widely used technique to visualise the relative genetic diversity of different individuals. PCA can be performed on the chunklengths matrix in a similar way to how PCA on the genotype dosage matrix is often employed in ancient DNA studies. Visualising whether downsampled individuals cluster close to the same sample at full-coverage is a useful way of determining whether the copyvectors of the downsampled individual reflect those of the full-coverage individual.

The position of the full coverage individuals are consistent with prior knowledge about their ancestry (Fig. 2.8). For example, Loschbour is positioned alongside other Hunter Gatherers, who are highly differentiated from the later Neolithic farmers and Bronze Age Europeans. sf12 clusters with the other

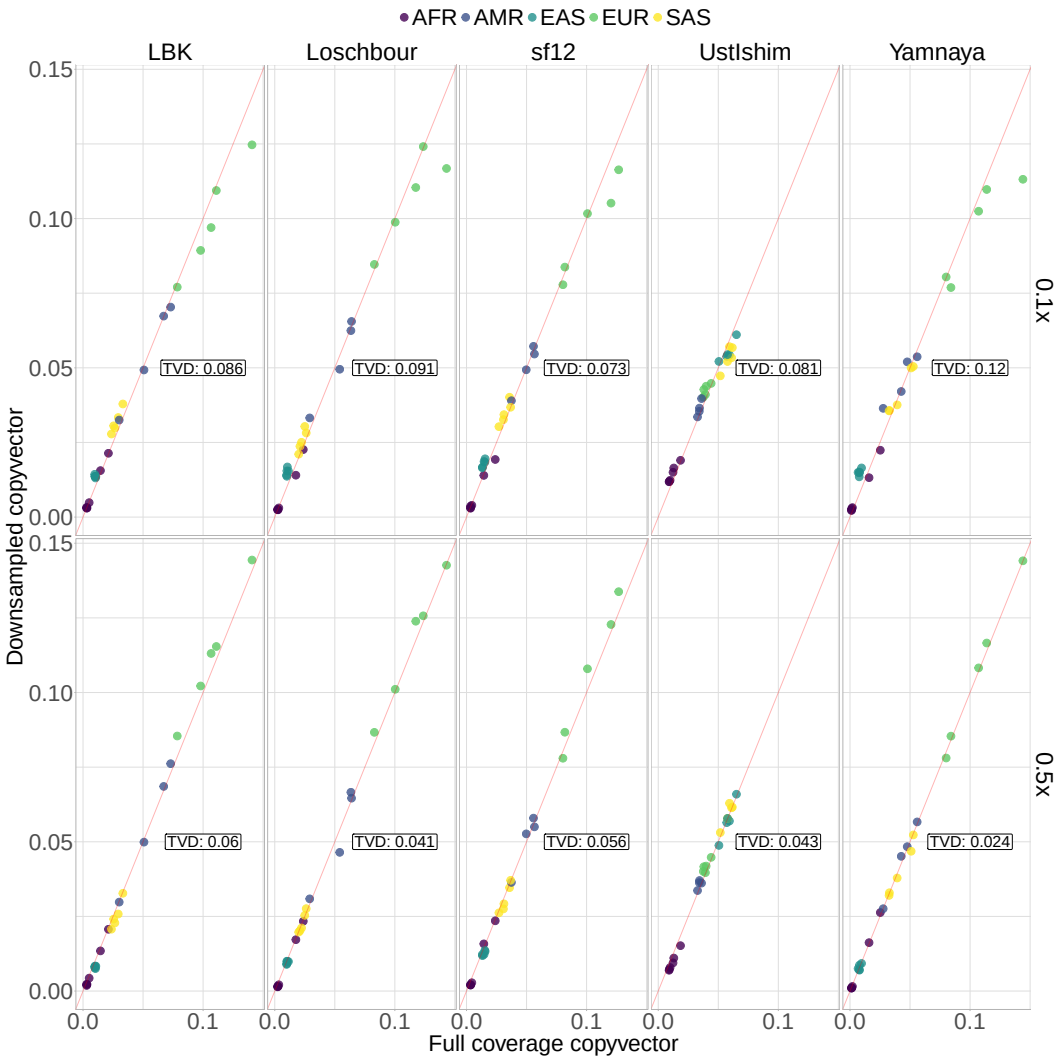

Figure 2.7: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to individuals from each of 26 present-day populations (dots). Red line is $y = x$. x and y units are normalised copying values and thus removed for clarity. Points coloured by meta-population.

Coverage	Loschbour	sf12	UstIshim	Yamnaya
0.1	Iberia_HG	PWC_SwedenNHG.SG	BHeimburg_LN	CordedWare
0.5	Loschbour	sf12	UstIshim	Poltavka
0.8	Loschbour	sf12	UstIshim	Poltavka
1	Loschbour	sf12	UstIshim	Poltavka
2	Loschbour	sf12	UstIshim	YamnayaSamara
3.5	Loschbour	sf12	UstIshim	YamnayaSamara
5	Loschbour	sf12	UstIshim	YamnayaSamara
10	Loschbour	sf12	UstIshim	Yamnaya
20	Loschbour	sf12	UstIshim	Yamnaya

Table 2.2: For each downsampled individual at each level of coverage, each entry gives the closest Cartesian neighbour based upon the PCA in Fig 2.8, not including other downsamples.

Scandinavian Hunter Gatherers in the dataset. Yamnaya is differentiated from the group of Bronze Age individuals and situated close to individuals from the Poltavka and Srubnaya culture. LBK is located with other individuals from the early to middle Neolithic in central Europe. Consistent with sharing little ancestry with any group over another, UstIshim is positioned close to the central Bronze Age mass, where most of the individuals in the PCA are located.

For all levels of downsampling other than the 0.1x, the downsampled and full coverage genomes were positioned very closely to one another on the PCA. When considering all downsampled individuals, a pattern emerges whereby the genome downsampled to 0.1x for each individual is 'pulled' towards the origin of the PCA. This may reflect a 'homogenisation' of low coverage genomes when many genotypes are imputed.

To formally examine the positioning of the samples on the PCA, I calculated the closest Cartesian neighbour to each of the downsampled individuals, not including other downsampled individuals (Table 2.2). Other than at 0.1x coverage, the samples UstIshim, sf12 and Loschbour always were closest to the same sample at full coverage. Up to 5x coverage, Yamnaya was closest to closely related YamnayaSamara and Poltavka samples.

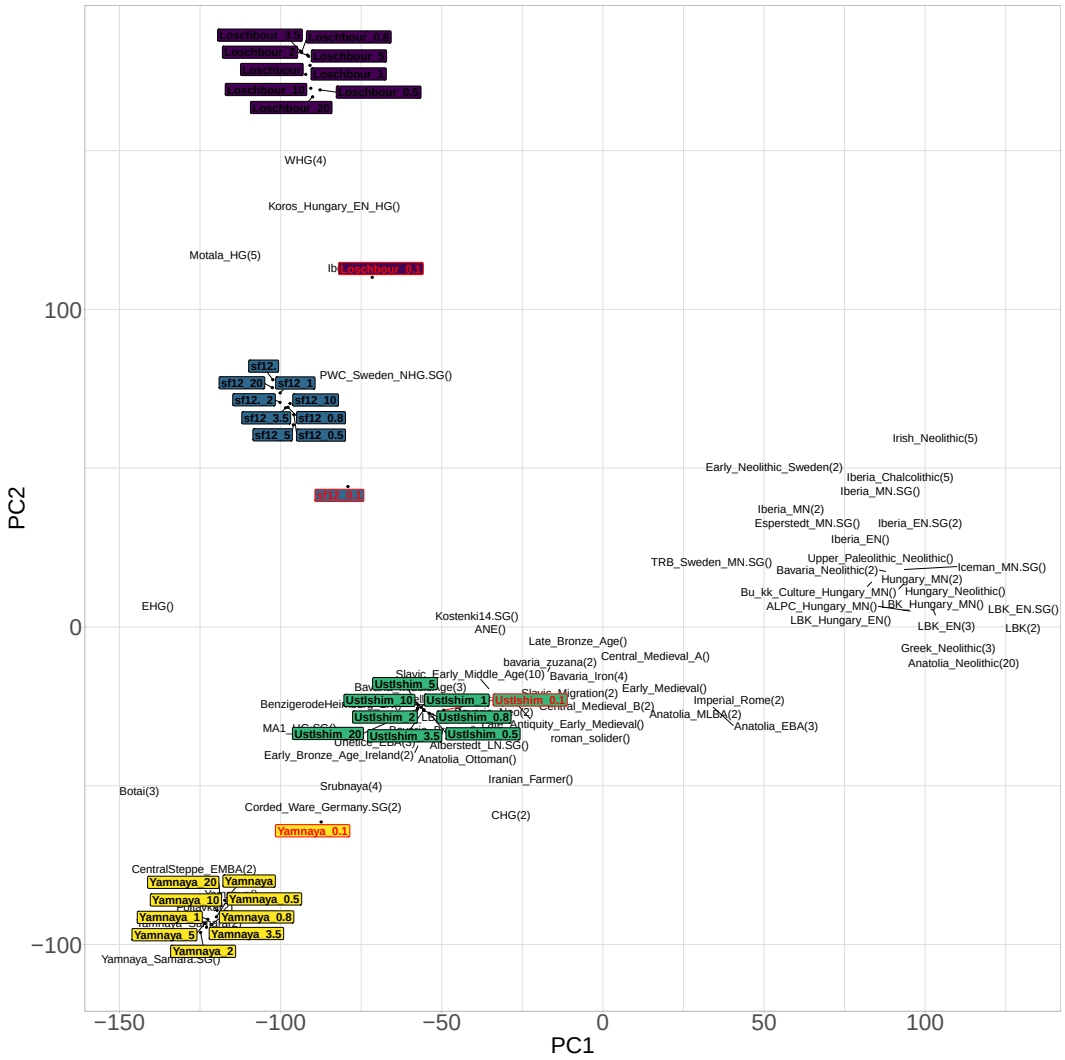

Figure 2.8: Principle component analysis (PCA) of downsampled, full coverage and downloaded ancient individuals generated from the linked chunklengths matrix. Full coverage and downsampled genomes of the same individual are coloured the same. Reference individuals are grouped into populations plotted as the mean principle components for all individuals within the population. Numbers in labels correspond to the number of individuals within the reference population. 0.1x samples have red border for clarity.

Taken together, this data suggests a minimal effect of coverage down to and including 0.5x mean depth. To my knowledge, no other study has evaluated the effect of coverage on ChromoPainter analysis down to a coverage of 0.5x. Margaryan et al (2020) showed a minimal effect of coverage at 1x and that fineSTRUCTURE groupings, containing individuals as low as 0.1x coverage, were not driven by coverage [58].

2.6.5 SOURCEFIND

I next determined the effect of sequencing coverage on the ancestry proportions estimated by SOURCEFIND, which accounts for variable donor group sizes and incomplete lineage sorting to improve interpretability relative to the raw chunklengths matrix.

I began by considering three ancestral sources, or 'surrogates', fixed as Anatolia Neolithic, Western Hunter-Gatherer and Yamnaya steppe pastoralist. I compared inferred proportions for the same individual across different levels of coverage (Fig. 2.9).

Consistent with previous the results, SOURCEFIND estimates are robust down to 0.5-0.8x coverage. At 0.1x coverage, there is an increase in ancestry components that are not present in higher coverage samples, suggesting they are artefacts caused by low coverage. For example, small components of Anatolia Neolithic and Yamnaya ancestry appear in Loschbour at 0.1x coverage, which are not present at any higher coverages. Above 0.5x coverage, the effect of coverage on estimated ancestry proportions appears to be marginal. For example, in sf12, the difference in the minor ancestry component of Anatolia Neolithic is, at most, 2.4%.

However, more than three surrogates are often used, as SOURCEFIND is meant to infer the most important contributors without *a priori* knowledge of the samples' ancestry. Therefore, I re-ran SOURCEFIND using 39 surrogate

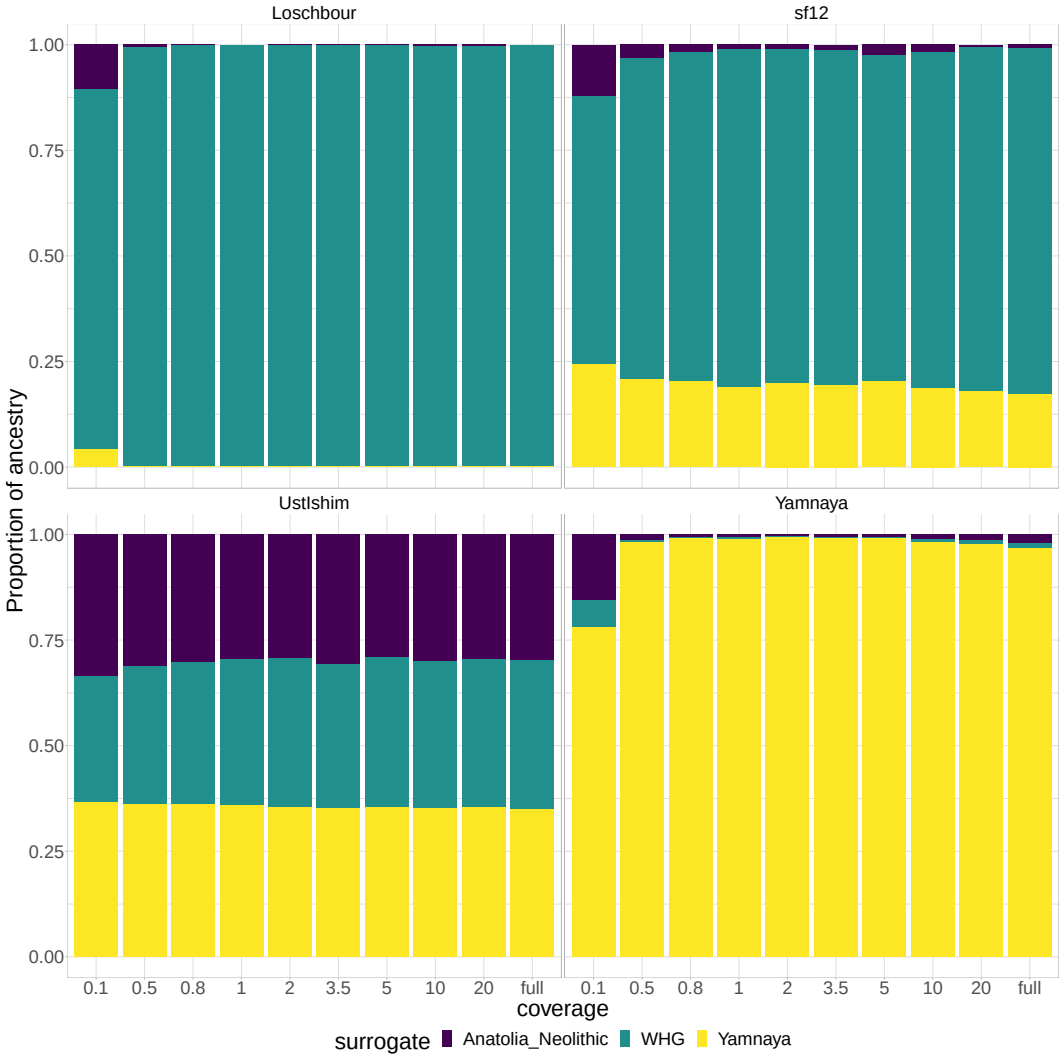

Figure 2.9: Each panel gives SOURCEFIND inferred recent ancestry sharing proportions for a different downsampled genome. Bars represent proportion of ancestry, coloured by different surrogates. Different coverages for the same individual are given within each panel. Three surrogates were used. LBK excluded because of anomalously poor results.

populations. For all downsamples above 0.1x in coverage, the ordering of proportions for each surrogate was the same.

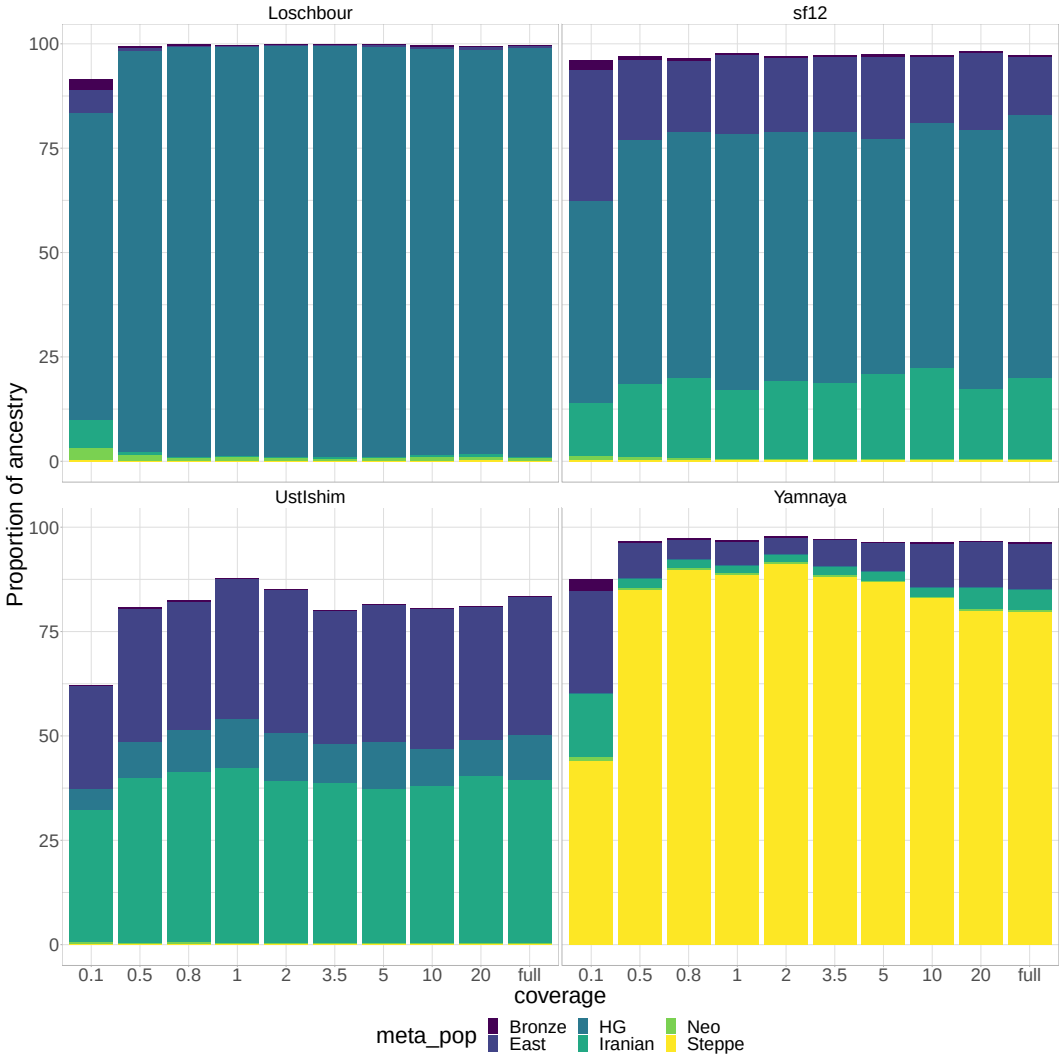

Figure 2.10: Each panel gives information for a different downsampled genome. Bars represent proportion of ancestry inferred by SOURCEFIND, coloured by different surrogates. Different coverages for the same individual are given within each panel. All 39 ancient surrogates were used. Only surrogates with more than 5% are shown. Ancient surrogates grouped into hand-assigned 'meta-populations' for visual clarity. LBK excluded because of anomalously poor results.

Again, Loschbour seems to be the least affected by coverage, with only slight differences between the 0.5x and full coverage samples. It is known that Upper Palaeolithic / Early Neolithic Hunter-Gatherer populations were small and lacked genetic diversity [51, 114, 115]. It is therefore expected that

Hunter-Gatherers would share longer IBD segments than individuals from outbred populations. Accordingly, this may make estimating SOURCEFIND proportions easier.

2.7 Issues and possible solutions for low coverage ancient DNA

The previous section outlined a drawback of performing ChromoPainter analysis on low coverage (<0.5x) ancient DNA samples; low coverage samples appear to be shifted towards the origin of a principle component analysis (PCA) relative to the same sample at higher coverage (Fig. 2.8) and can contain ancestry estimates that are not present in the same full coverage sample (Fig. 2.9). This is evident for the lowest coverage samples at 0.1x and suggests that samples of this coverage cannot be reliably analysed using current methodology.

In order to solve the issue of coverage-related bias, it is first necessary to determine at which stage of the analysis pipeline the bias is introduced. By 'analysis pipeline', I refer to the three stages of (1) variant calling, (2) imputation and phasing, and (3) ChromoPainter described in the methods section.

2.7.1 PCA imputation test

To explicitly test at what stage the bias is introduced, I performed a set of principle component analyses on the downsampled data. First, I performed PCA projections of all downsampled ancient individuals onto a set of present-day European individuals (shown in Table 2.1) using i) pre-GLIMPSE genotypes and ii) post-GLIMPSE (imputed) genotypes (Fig. 2.11). PCA projections are used when the target dataset, in this case downsampled ancients, contain variable levels of missing data.

The results show that there is no apparent coverage-related bias in the pre-GLIMPSE PCA; the 0.1x samples do not substantially differ in their position from the other downsamples of the same individual. However, there is a degree of noise; for example, the LBK downsamples are spread over a small region on the PCA.

On the other hand, downsamples of UstIshim, sf12 and Loschbour are shifted to the centre of the post-GLIMPSE PCA and away from the full coverage individual and other downsamples. This suggests that coverage-related bias is being introduced in the imputation stage. At the same time, GLIMPSE appears to have removed some of the noise in the downsampled individuals of coverage $\geq$0.5x. For instance, the noise observed in the LBK samples in the pre-imputation PCA is substantially reduced and the samples cluster more tightly.

I also performed PCAs based upon an all-v-all ChromoPainter painting using the same set of present-day European samples (Table 2.1) and downsampled ancient individuals as previously, in both linked and unlinked modes. There is an increased amount of noise and evidence of coverage-related bias relative to the post-GLIMPSE genotype PCA. Fig. 2.12) displays the PCA for the same painting, but using the unlinked chunkcounts matrix. Comparing the linked and unlinked PCAs shows the effect of including linkage (i.e. haplotype information) on the amount of bias and noise across each sample. Per-sample, there appears to be reduced noise in the unlinked painting.

These results suggest that imputation introduces a degree of bias into 0.1x samples that is not apparent on non-imputed genotypes. They also suggest that ChromoPainter introduces an additional degree of bias when analysing haplotypes, or that it amplifies bias already present introduced at the imputation stage. Accordingly, removing SNPs which have been poorly imputed may be a way to mitigate such biases.

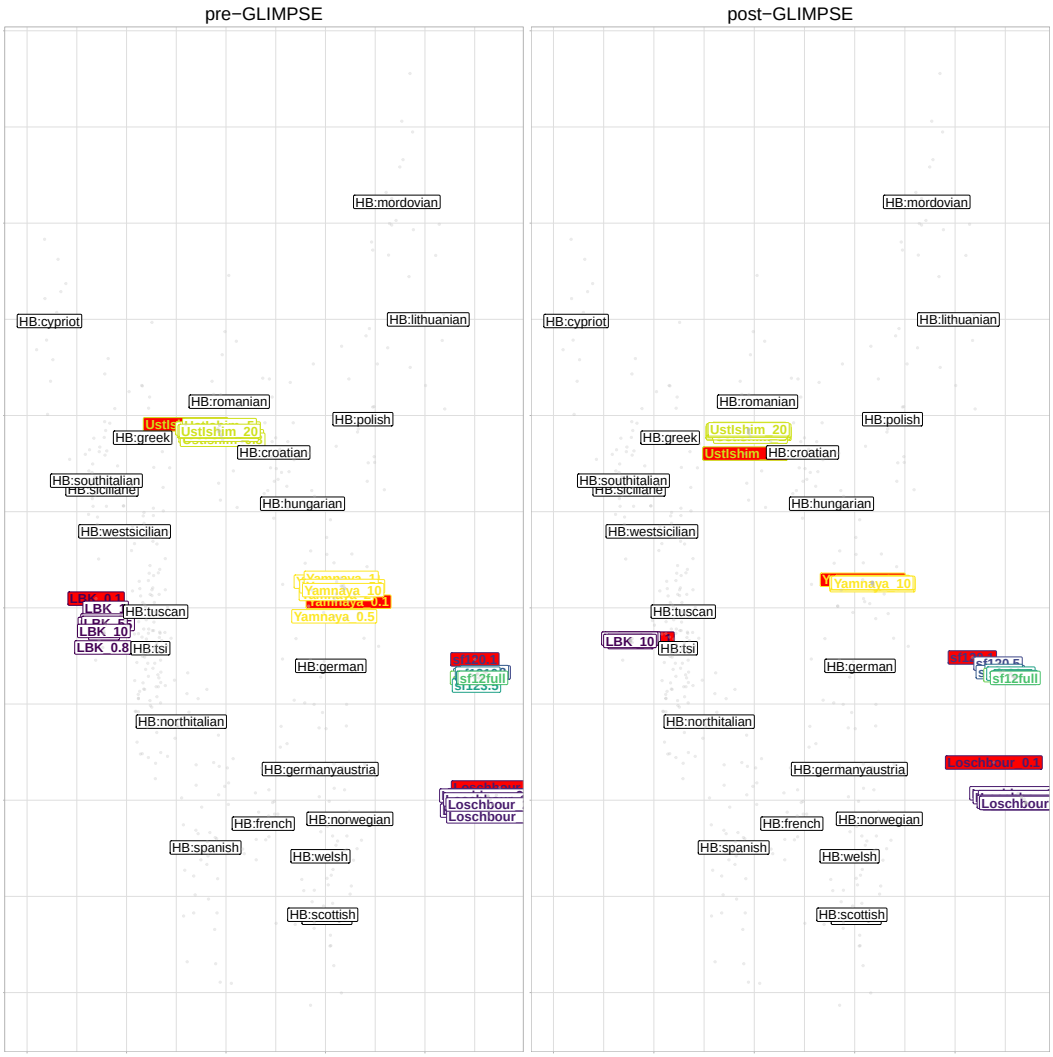

Figure 2.11: Principle Component Analysis. Left - pre-GLIMPSE genotypes. Right - post-GLIMPSE (imputed) genotypes. White labels correspond to the midpoint of all samples from that population, grey points correspond to modern individuals. 0.1x samples highlighted in red for clarity.

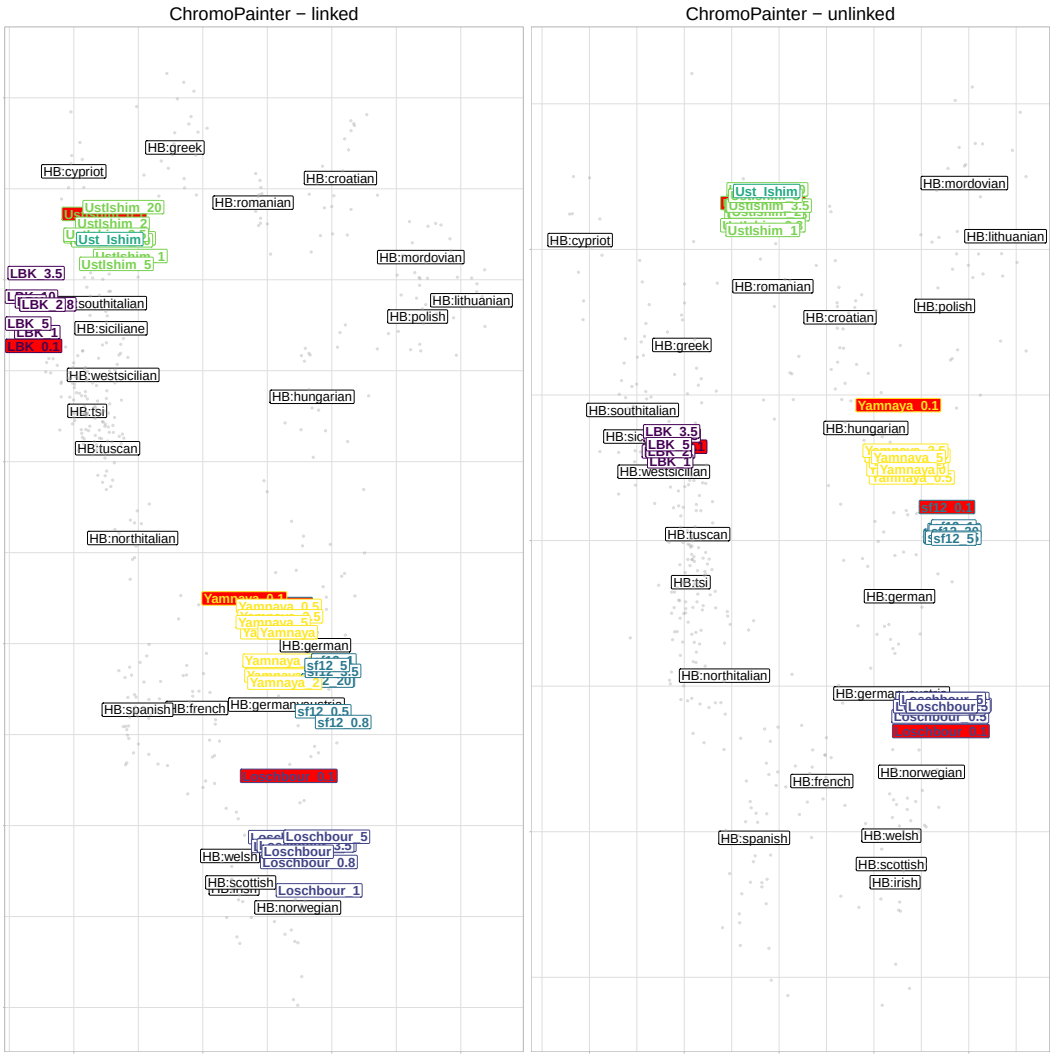

Figure 2.12: Left - ChromoPainter Linked. Right - ChromoPainter Unlinked. White labels correspond to the midpoint of all samples from that population, grey points correspond to modern individuals. 0.1x samples highlighted in red for clarity.

2.7.2 Direct imputation test

The previous section suggested that imputation plays a role in the introduction of coverage-related bias. However, it is not clear whether it is 'bias', i.e. towards the reference population used to assist imputation, or 'noise' due to random incorrect imputation. To directly test whether the effect of imputation is noise or bias, I used the Human Origins dataset (described in Appendix section A.3), containing the genotypes of 5998 present-day individuals from across the world at 560,442 SNPs. I chose to use present-day samples because there is a larger total number of individuals and larger number of individuals per population, giving more power to detect any potential bias. Additionally, the populations in present-day samples are more homogenous and well-defined compared to ancient groups. I set all but 70,000 random SNPs as missing and imputed missing positions using the HRC as a reference, in order to simulate a dataset where the majority of SNPs are imputed. I then performed an all-v-all painting of i) the original Human Origins dataset where none of the 560,442 SNPs had been imputed and ii) the simulated dataset where 430,000 SNPs had been imputed.

Bias occurs when missing genotypes are incorrectly imputed with variants from certain populations more frequently than others. We might expect these populations to be those which are more prevalent in the reference panel. We would correspondingly expect bias to mean that, when painted, some donor populations would donate more than others, relative to if no imputation had taken place. On the other hand, if 'noise' is dominating results, we would expect the incorrectly imputed genotypes to be randomly distributed across populations, and similarly we would not expect to see any populations donating more than others relative to if no imputation had taken place.

Therefore, we can compare the amount different donor groups donate under the dataset where none of the 560,442 SNPs had been imputed versus the dataset where 430,000 (86%) of these SNPs have been imputed by plotting

the mean amount donated by each population using imputed SNPs and non-imputed SNPs (Fig. 2.13). The 20 populations that contribute most are either European or Jewish. Notably, the Haplotype Reference Consortium panel that was used to impute the data consists primarily of individuals of European descent. The two populations which are over-copied the most after imputation are two English populations from Kent and Cornwall. This suggests that there is a most likely a bias towards copying more from European populations when the data has been imputed using the HRC.

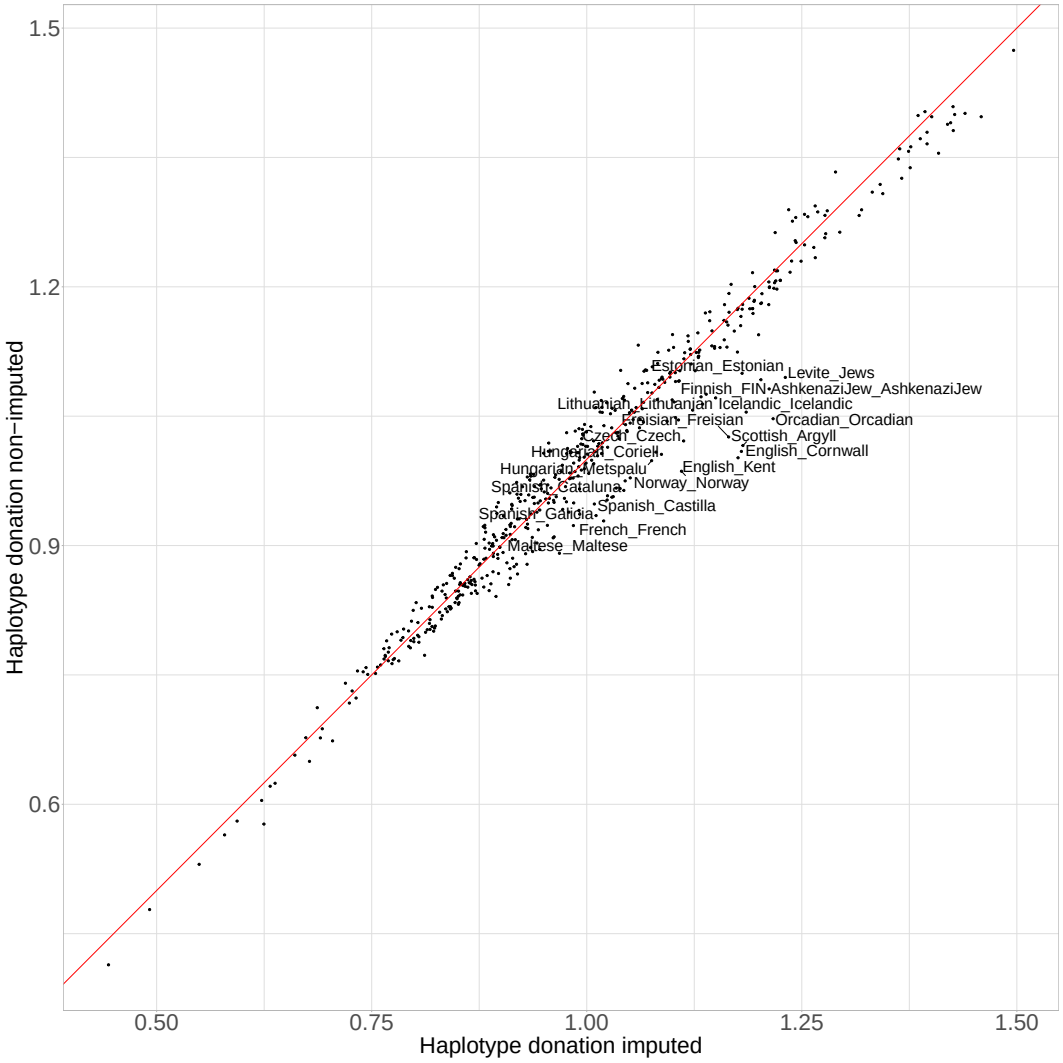

Figure 2.13: Comparison of the mean normalised cM donated by each donor population using the imputed and non-imputed SNP sets. The 20 populations with the largest different between imputed and non-imputed donation are highlighted. Red line is line of equality.

2.8 Solutions

In this section I will explore potential solutions to the issue of coverage-related bias. Based on the findings in previous sections, imputation causes bias towards particular reference populations in modern samples.

2.8.1 Accounting for allele likelihoods

Section 2.2.1.1 describes an improvement to the ChromoPainter algorithm. Instead of assuming that each allele on a haplotype is correct with a probability $1-\theta$, where θ represents an error probability, the posterior genotype probability from GLIMPSE is accounted for in the emission probabilities of the copying model. The motivation behind this update is that the uncertainty associated with genotype calls at low coverage is suitably propagated throughout the painting process, resulting in uncertain alleles contributing less towards the expected copying values than more certain ones. This is similar in spirit to that of Viera et al (2016), who account for genotype likelihoods to infer inbreeding IBD tracts from low coverage sequencing data [116].

To determine whether accounting for allele likelihoods improved the painting accuracy of a low-coverage genome, I painted the individuals downsampled to 0.1x and 0.5x and corresponding full coverage samples using the 'standard set' of ancient reference individuals, using both ChromoPainterV2 and ChromoPainterV2Uncertainty. I then calculated r-squared between the copyvectors of full coverage and downsampled individuals using the two different methods (Fig. 2.14). This shows that at 0.1x, the ChromoPainterV2 method clearly outperforms ChromoPainterV2Uncertainty across all samples, whereas at 0.5x, the new method marginally outperforms the standard method. Therefore, while accounting for allele likelihoods may improve performance in cases of coverage $\geq$0.5x, which has been shown to still capture some haplotype information, it does not help in cases of coverage of 0.1x where bias problems persist.

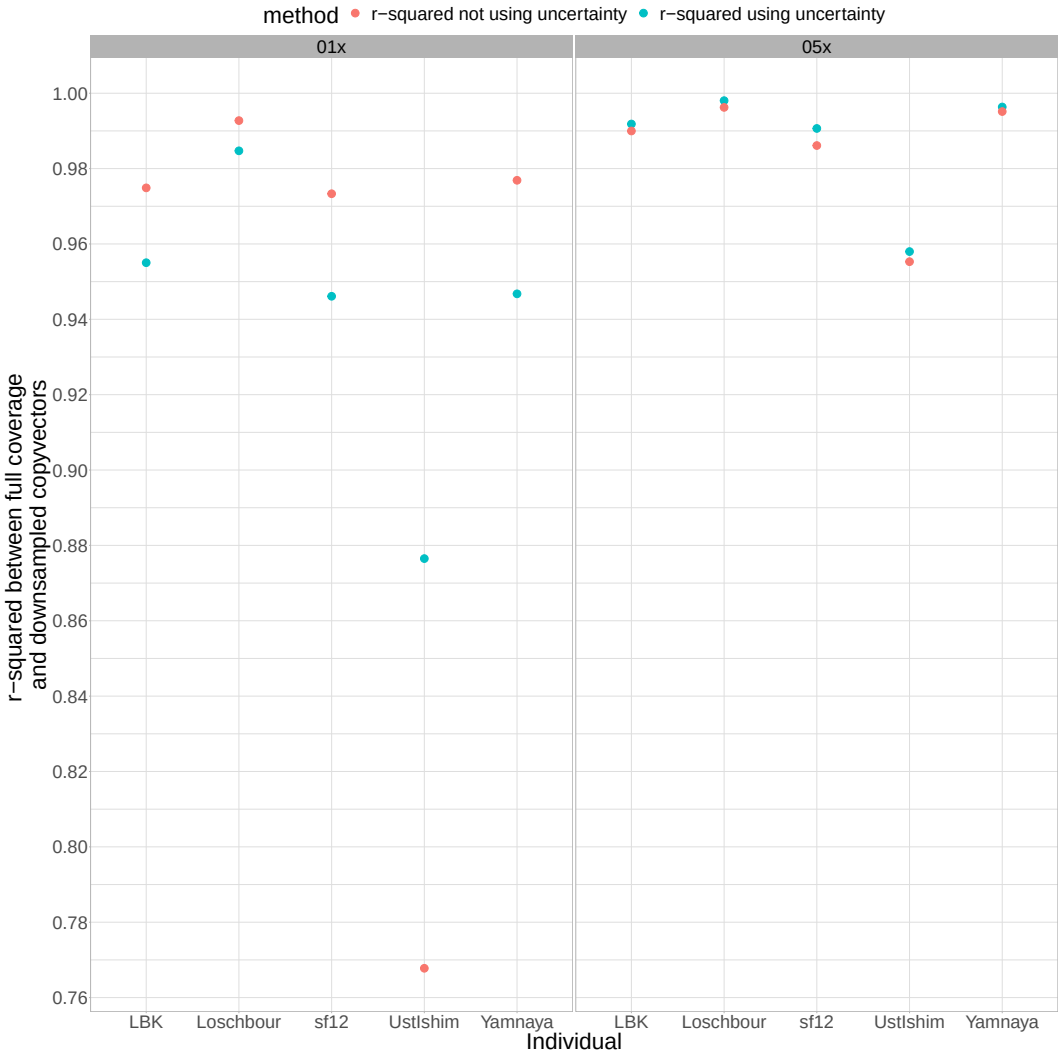

Figure 2.14: Comparison of performance of ChromoPainterV2 and ChromoPainterV2Uncertainty. Panels correspond to samples downsampled to 0.1x (left) and 0.5x (right). Points correspond to the r-squared between the downsampled individual and the same individual at full coverage. Red points are values obtained from ChromoPainterV2 and blue points are those obtained from ChromoPainterV2Uncertainty.

2.8.2 Filtering SNPs

In this section, I will test whether filtering the set of input SNPs on different criteria reduces the effect of coverage related bias.

The frequency of a particular variant in the reference panel (RAF - reference allele frequency) used for imputation is known to affect how accurately that variant can be imputed [25, 63, 90, 91]. Specifically, we expect variants which are less frequent in the reference panel to be imputed at a lower accuracy than those which are more frequent. Therefore, removing variants with a low frequency in the reference panel may mitigate the coverage related bias by removing variants which have been incorrectly imputed. In other words, we want to retain the SNPs where both alleles are relatively common within the population.

For each individual, I took the 428,425 SNPs in the HellBus set and removed SNPs with $0.1 > RAF$ or $RAF > 0.9$, removing an average of 50,187 SNPs per individual. RAF refers to the frequency of the allele in the 1000 genomes reference panel used to phase and impute the HellBus dataset. I then painted individuals downsampled to 0.1x and 0.5x using the standard set of 125 ancient donor individuals.

Comparing the TVD values between the copvectors showed that this did not improve the 0.5x copyvectors (Table 2.3).

I then chose to filter SNPs based on $max(GP)$ at each position. $max(GP)$ correspond to the accuracy with which a SNP has been imputed, with higher values reflecting a higher chance of that genotype being imputed correctly. For each individual downsampled to 0.5x and 0.1x, I only retained positions where the $max(GP) >= 0.990$. For the 0.5x individuals, this resulted in a total of 348,852 SNPs for LBK, 339,949 for Loschbour, 315,075 for sf12, 308,961 for UstIshim and 386,484 for Yamnaya. Because different SNPs were removed from different individuals, each individual was painted separately. The same

sample	u_01x	s_01x	r_01x	gp_01x	u_05x	s_05x	r_05x	gp_05x
LBK	0.926	0.927	0.933	0.746	0.981	0.981	0.982	0.959
Loschbour	0.898	0.898	0.907	0.654	0.980	0.980	0.976	0.925
sf12	0.923	0.923	0.942	0.774	0.981	0.981	0.980	0.950
UstIshim	0.944	0.944	0.945	0.827	0.980	0.980	0.976	0.960
Yamnaya	0.915	0.915	0.920	0.726	0.986	0.986	0.985	0.964

Table 2.3: Table of $1 - TVD$ values between the copyvectors of full coverage and downsampled individuals. 'u' refers to ChromoPainterUncertainty, 's' refers to ChromoPainterV2, 'r' refers to filtering SNPs with reference allele frequency (RAF) $0.1 > RAF$ or $RAF > 0.9$ and 'gp' refers to filtering by $max(GP) >= 0.990$.

standard set of 124 ancient donors was used. Again, this did not improve the accuracy of the copyvectors.

2.8.3 Restricting analysis to non-imputed SNPs

Section 2.7.1 showed that imputation was the likely cause of coverage related bias. Thus, restricting ChromoPainter analysis to non-imputed SNPs above a certain coverage may mitigate such bias.

However, removing SNPs may have negative side-effects; increasing the genetic distance between SNPs reduces linkage information and therefore may reduce the overall power to distinguish between closely related haplotypes. At the most extreme case, retaining only a small number of SNPs may effectively reduce the method to unlinked and lose the advantage given by accounting for haplotypes. This may be important if we decide to restrict analysis to non-imputed SNPs, as low coverage samples may only have a small number of high enough coverage, non-imputed SNPs. Therefore, it is important to determine whether samples of a particular coverage have enough regions containing enough high-coverage SNPs to retain the advantages of haplotype-based methods over unlinked ones.

One case study to test whether a set of SNPs has enough linkage information is to determine whether it is possible to distinguish individuals born in

Devon from those born in Cornwall. This has shown to be possible using the fineSTRUCTURE clustering algorithm using linkage information, but not using unlinked methods (ADMIXTURE [108]) [31]. Therefore, determining whether it is possible to distinguish between individuals from Devon and Cornwall acts as a test case for determining how many high-coverage SNPs would give sufficient SNP density to distinguish between these two populations.

To assess this, I painted individuals from Devon (n=73) and Cornwall (n=89) with all other POBI individuals as donors (n=2,039), using the full set of SNPs (n=452,592). It is necessary to develop a classification score which quantifies to what degree it is possible to distinguish between individuals from Devon and Cornwall. For a classification score, I calculated the proportion of Cornwall individuals whose copy vector had a lower TVD with the mean copyvector of all other Cornwall individuals than with the mean copy vector of all Devon individuals. In other words, this asks whether the individual is genetically closer to the Devon or Cornwall population.

I repeated the analogous procedure to find a classification score for Devon individuals, given in table 2.4. I then painted the same individuals using a reduced set of SNPs, in particular reducing the set of SNPs to 12 different percentages ranging from 0.2% - 90% of the total original number of SNPs (a full list of the reduction levels and details of the painting procedure can be found in the methods section 2.4). Painting using a reduced set of SNPs is intended to simulate an ancient genome where only a subset of the total number of SNPs have been covered by a sufficient number of reads. Defining 'sufficient' isn't precisely defined, but it is possible to calculate the probability of observing both reads given x reads at a given heterozygous positions and assuming equal probability of observing reference and non-reference alleles; for example, 9 reads are needed to obtain at least a 0.995 probability of observing both alleles (Fig. 2.15).

In my painting of 5998 world-wide samples on the Human Origins array

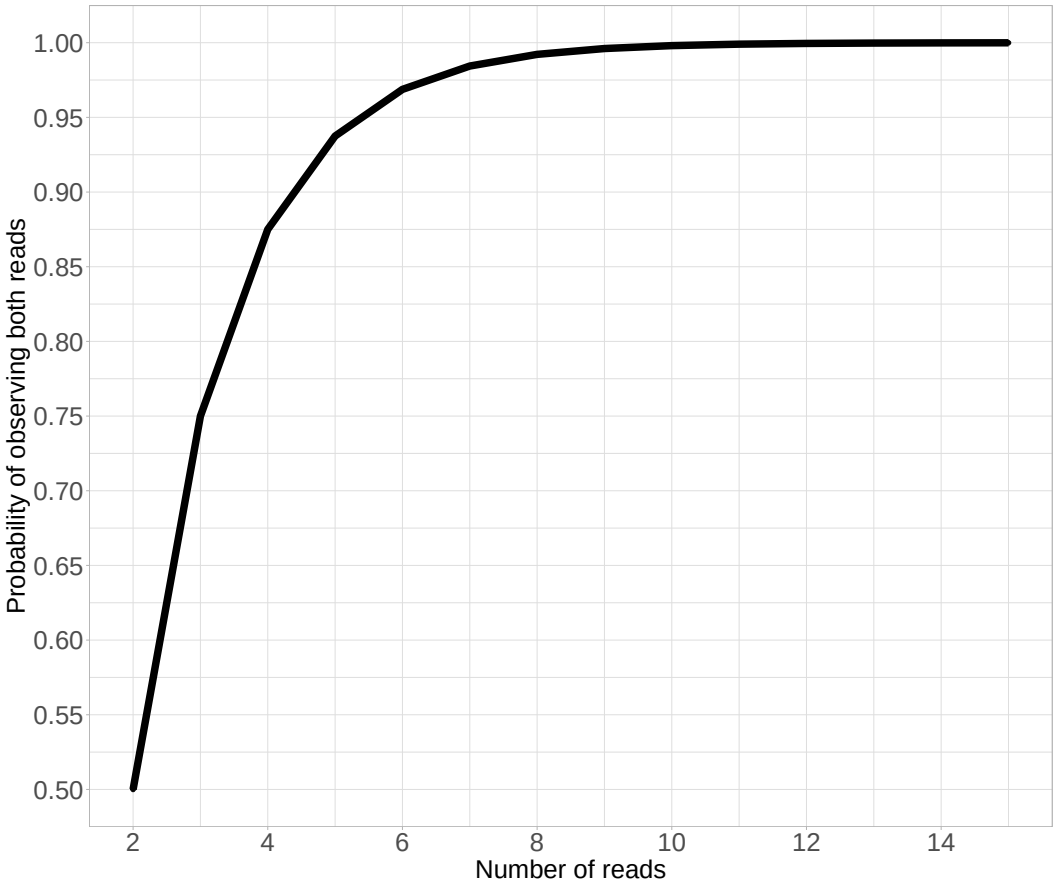

Figure 2.15: Probability of observing both reads at a heterozygous positions, given x reads assuming equal probability of observing reference and non-reference alleles.

(described in Appendix section A.3), the average number of segments that forms a recipient genome is 9764 (range: 1437-18,963). Given a genome-wide size of $\approx$ 3000Mb, this implies that an average 'chunk' size (in Mb) is $3000/9764 = 307.2 \approx$ 500kb, where a 'chunk' is a set of contiguous SNPs matched to a single donor. Therefore, for each of the 12 different levels of SNP reduction used in my Devon/Cornwall analysis, I can calculate the average number of SNPs per 500kb chunks, and determine how many of these 500kb chunks are necessary to accurately distinguish individuals from Devon and Cornwall. To do so, for each reduced SNP percentage, I found the Cornwall/Devon classification score using only data from chromosome 22 (which has only W 500kb chunks), and using only chromosomes 21 and 22 (which has V 500Kb chunks), etc, continuing until the classification scores were equivalent to that when analysing all 22 autosomes

Percentage of SNPs retained	Cornwall	Devon
1 %	0.801	0.945
2 %	0.820	0.986
3 %	0.876	0.973
4 %	0.910	0.973
5 %	0.888	0.973
6 %	0.899	0.973
7 %	0.888	0.973
8 %	0.910	0.973
9 %	0.910	0.973
10 %	0.910	0.973
20 %	0.921	0.973
30 %	0.910	0.973
40 %	0.899	0.973
50 %	0.910	0.973
70 %	0.910	0.973
80 %	0.910	0.973
90 %	0.921	0.973

Table 2.4: Proportion of individuals correctly assigned to their population at different percentages of SNPs retained.

at all 452,592 SNPs. In this way, for each reduced SNP percentage, I found the number of 500Kb chunks necessary to as accurately distinguish between Devon and Cornwall as in the case where we had analysed a full data set of 452,592 SNPs (Table 2.5). I found results to be very similar to if chunk-size were instead defined as 250kb or 1Mb (Table 2.5).

I repeated an identical analysis, including reducing the total number of SNPS, using individuals from the Mandenka and Yoruba ethnic groups rather than Devon and Cornwall.

Guided by these results, for each ancient individual (n=587, median coverage=1.1x), I found the number of non-overlapping windows of sizes 250Kb, 500Kb or 1Mb that had Y SNPs above Z coverage, varying both Y and Z.

Fig 2.16 shows the mean number of 500Kb windows per individual with at least Y SNPs above Z coverage, with individuals grouped into bins based on

Number of SNPs retained	250Kb	500Kb	1Mb	Number of SNPs per 500Kb Window
20,000	9356	4715	2388	3.3
25,000	6954	3509	1781	4.1
30,000	6272	3166	1607	5.0
35,000	4083	2064	1049	5.8
40,000	3099	1565	796	6.6
45,000	3602	1820	925	7.5
50,000	2612	1321	673	8.3
100,000	1304	661	338	16.6
150,000	1005	508	260	25.0
200,000	705	357	183	33.3
250,000	705	357	183	41.6
300,000	506	255	130	50.0
350,000	267	135	69	58.3
400,000	705	357	183	66.6
450,000	136	69	35	75.0

Table 2.5: Number of 250Kb, 500Kb or 1Mb windows required at different levels of SNP reduction to match the TVD assignment power of 500K fully genotyped SNPs for individuals in Devon and Cornwall. Note that the number of necessary 250kb and 500kb windows is roughly four and two times, respectively, the number of 1Mb windows, indicating the definition of window size makes little difference.

their mean coverage. Points are coloured yellow if, within the bin of coverage, samples have at least 2000 windows.

Samples less than 0.5x do not have enough windows if the threshold for a 'good' SNPs is being covered by a single read. As it not possible to call a heterozygous position with only a single read, this suggests that there are not enough non-imputed SNPs with enough coverage to match the power seen in full coverage individuals. For example, samples between 0.3-0.4x have approximately 1000 segments with $\geq$ 10 SNPs above 2x in coverage; Table 2.5 shows that 1565 windows of $\geq$ 8.3 SNPs is enough to match full power. However, as Figure 2.15 shows, 50% of these genotypes may not observe both reads if the position is heterozygous. Indeed, even when there are 3 reads covering a site, there is still a 25% chance of not identifying a heterozygous

Number of SNPs retained	250Kb	500Kb	1Mb	Number of SNPs per 500Kb Window
30,000	6272	3166	1607	5.0
35,000	3099	1565	796	5.8
40,000	3099	1565	796	6.6
45,000	2612	1321	673	7.5
50,000	3099	1565	796	8.3
100,000	1886	956	489	16.6
150,000	1304	661	338	25.0
200,000	506	255	130	33.3
250,000	267	135	69	41.6
300,000	506	255	130	50.0
350,000	506	255	130	58.3
400,000	506	255	130	66.6
450,000	267	135	69	75.0

Table 2.6: Number of 250Kb, 500Kb or 1Mb windows required at different levels of SNP reduction to match the TVD assignment power of 500K fully genotyped SNPs for individuals in from Mandenka and Yoruba ethnic groups.

z

position. Only the samples in the 2-5x coverage bin had enough windows when using a coverage threshold of 4 and 5 reads.

This analysis therefore suggests that there are not enough regions with enough high quality SNPs at mean coverages less than 2x to reliably analyse using ChromoPainter.

2.9 Summary of Results and Discussion

In this section I used a downsampling approach on five high-coverage ancient DNA samples to show that ChromoPainter analysis can be performed on samples down to 0.5x coverage without showing a significant deviation from the same sample at full coverage. In particular, ChromoPainter copyvectors, SOURCEFIND ancestry proportion estimates and Principle Component Analysis position all of 0.5x coverage and higher showed a good correspondence with

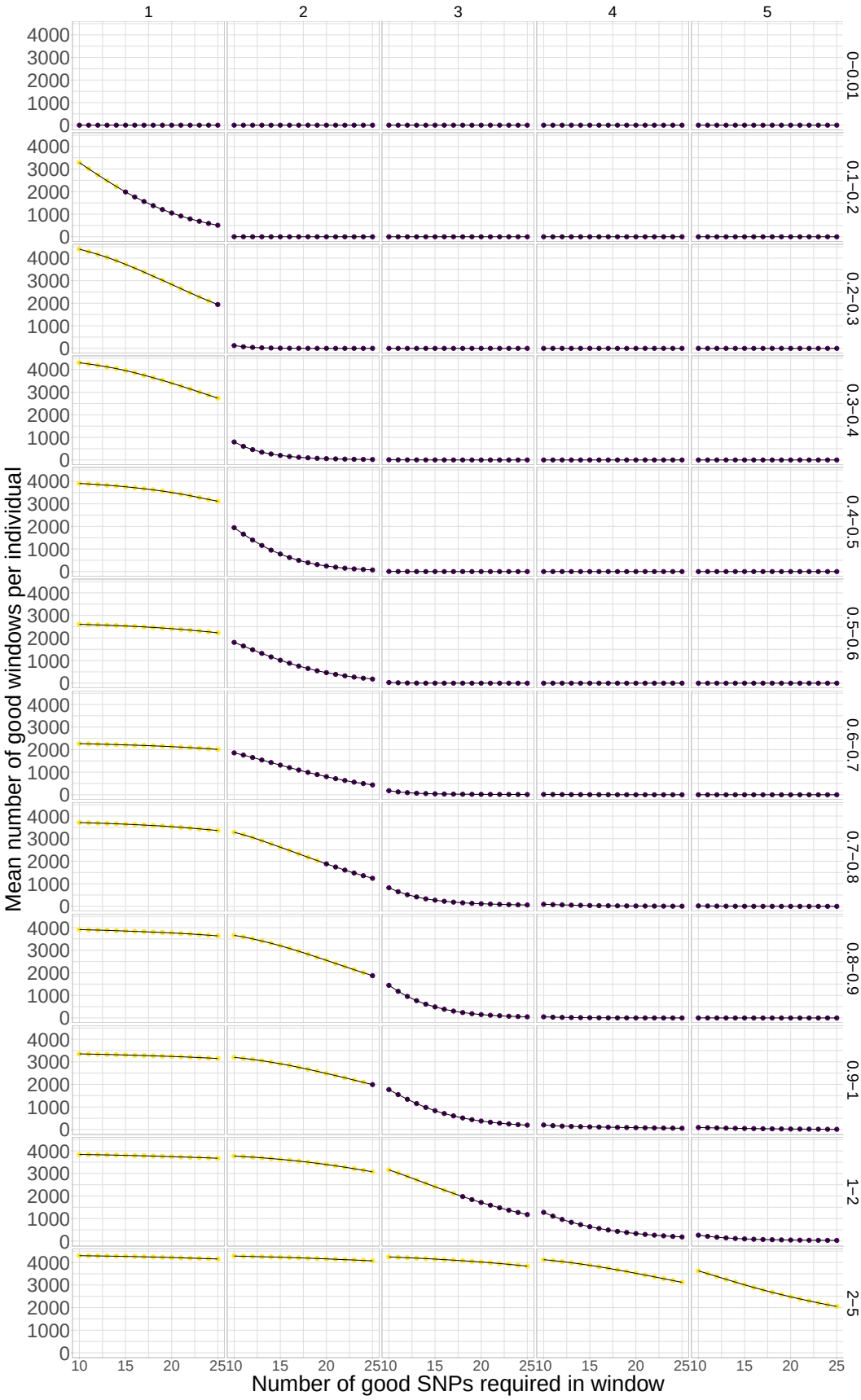

Figure 2.16: Mean number of 500Kb windows (y-axis) within the genome of each ancient individuals within a given range of coverages (rows) with at least Y SNPs (x-axis) above a particular coverage Z (columns)

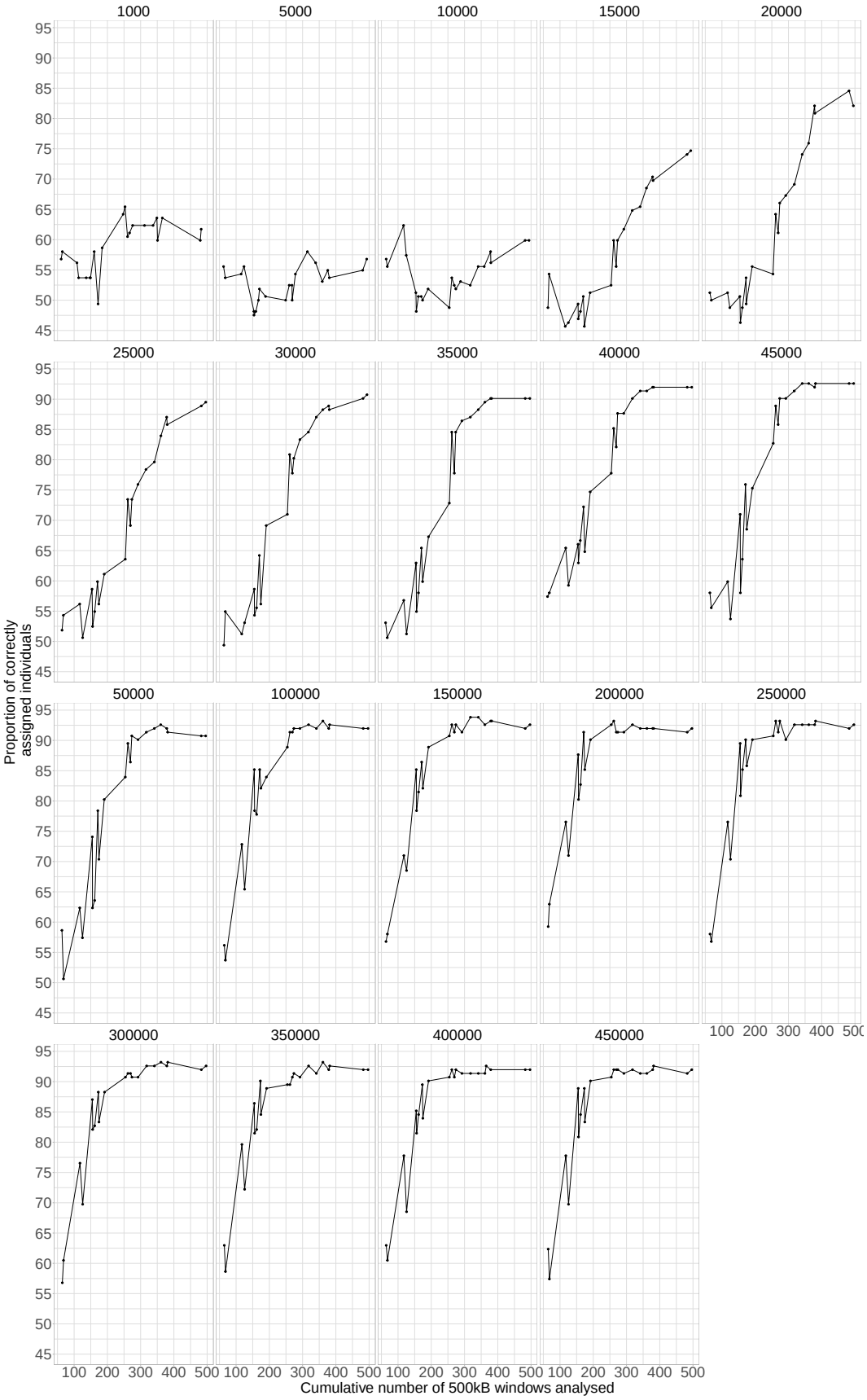

Figure 2.17: The effect of adding 500kB windows on the ability to assign individuals from Devon and Cornwall to their respective populations. Each panel represents a different total number of SNPs used. X-axis gives the cumulative number of 500kB windows used in analysis. Y-axis gives the combined proportion of individuals assigned.

the same metrics at full coverage. The 0.1x downsampled showed deviations from the full coverage samples which meant that they cannot currently be analysed reliably with ChromoPainter and its associated methods. I showed that imputation introduces bias into low-coverage samples that is manifested by those samples being shifted towards the centre of a PCA.

I performed a range of analyses to try and recover useful haplotype information from low coverage samples and improve the performance of the analysis. Counter-intuitively, approaches such as removing SNPs with a low imputation quality and reference allele frequency did not improve the performance of ChromoPainter on low coverage samples. However, this is broadly consistent with a single previous study, which also showed that filtering the dataset for SNPs with a low imputation quality score did not substantially affect fineSTRUCTURE clustering [48]. However, it also runs counter to studies which have shown filtering SNPs based on imputation quality score significantly can significantly reduce the number of incorrectly imputed genotypes [63].

I also developed a modification to the ChromoPainter model which accounted for uncertainty in genotype calls; however it only marginally improved the performance of ChromoPainter on samples of 0.5x or higher. Again, this was surprising, as previously published methodology which accounts for genotype likelihoods when estimating IBD tracts has been shown to be effective [80].

Finally, I used simulated data from present-day individuals to show that samples around 0.5x coverage can in theory be analysed with useful haplotype information, but that imputation is necessary for lower coverage samples.

Many of the analyses performed in this section only used a single target sample, as I did not identify a way to generate multiple downsampled individuals from the same population. For example, the SOURCEFIND analysis I performed used a single target downsample when estimating ancestry proportions. This differs from a typical ancient DNA analysis, such as those of Margaryan et

al [58], where there may be up to 20 low coverage samples per population. This number may increase in the future as the technology to generate ancient DNA improves. Leveraging information across multiple samples from the same population would improve the accuracy of population-wide ancestry or admixture estimates, for example. Thus, he results presented in this section which used a single target individual may underestimate the ability to analyse low-coverage samples. It may be possible to accurately analyse 0.1x samples if there are multiple samples per population.

In this section I used present-day individuals to estimate the number and size of chunks needed to retain haplotype information. This was because present-day individuals are simpler to analyse; the populations are better defined than in ancient samples (i.e. it is possible to only include individuals whose grandparents were born within 100kM of a target location), are of uniform coverage and contain many more individuals per population. Thus, using present-day individuals removes potentially confounding factors that may be present when analysing ancient samples. However, using present-day samples to draw conclusions about ancient samples may lead to underestimating the number of SNPs per window required. As the present-day samples had been genotyped high-quality DNA samples and a genotpying array, each genotype can be called with a high confidence. This is not the case with ancient samples, where each SNP may be covered by a small number (<3) of reads.

For the imputation and phasing reference panel, I used the 1000 genomes dataset which contains around 6000 haplotypes. The Haplotype Reference Consortium contains roughly 10 times as many haplotypes and thus offers substantial gains in the potential accuracy of genotype imputation [91]. I did not use the HRC owing to difficulties in obtaining access to the data; however, I expect that future studies which use this resource will be able to analyse ancient DNA samples of low coverage to a higher degree of accuracy.

Whilst I did not interrogate the range of coverages between 0.1-0.5x, this

could be an avenue for future research.

Chapter 3

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

3.1 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. [31, 117–119]). 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 [31, 37, 120].

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 [121], 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" [122]. The primary reason for this is reasonable concerns over the confounding effect of population substructure within a cohort [123]; 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 transferable to other populations if they have been conducted in cohorts of exclusively European individuals [124–126]. The reasons for this are not yet fully understood, but it thought that differences in LD structure may be at least partially responsible [127]. 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) [128] 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 participants. However the significant hits disappeared when corrected for using a ChromoPainter coancestry matrix, 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-paint, a method closely related to ChromoPainter [27].

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 [40]. 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 [129].

Recent studies have outlined the power of haplotype-based approaches in inferring the population histories of different African ethnic groups [130–132]. 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 other issues; individuals who registered for the U.K. Biobank undertook a series of extensive tests and not including their data in studies seems to be ethically dubious at best [133].

Therefore, to investigate the African ancestry of U.K. Biobank individuals, I will leverage 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 Democratic Republic of 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.

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 between 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. 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 [25, 134]. 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.

3.2 Methods

3.2.1 U.K. Biobank data access and initial processing

The U.K. Biobank dataset contains genotype data for 488,378 individuals 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 this dataset as the 'non-imputed' data, as all SNPs were directly genotyped without imputation. I used plink2 [135] 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 from the original 784,256, using the combined references of the Haplotype Reference Consortium (HRC) and UK10K haplotype

resource. I will hereafter refer to this data as the 'imputed' data. Full details of imputation can be found in the paper of McCarthy et al (2016) [100]. 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.

3.2.2 ADMIXTURE analysis

I am primarily interested in using ChromoPainter [19] 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. The ADMIXTURE algorithm is a fast and accurate way to estimate continental-scale ancestry proportions [108] and is therefore ideal for the task identifying individuals with more than 50% non-European ancestry in a large cohort.

I LD-pruned the non-imputed U.K. Biobank dataset using using `plink -indep-pairwise 50 10 0.02` [135], leaving 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.

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

3.2.3 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 for this purpose, as it contains individuals from 349 different ethnic groups from across Africa and 535 world-wide groups in total (Fig. 3.1). Full details of processing can be found in Appendix section A.3 .

3.2.4 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 utility 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 this dataset with shapeit4 [25] using `-pbwt-depth 8`, the b37 genetic map and all other parameters set as default.

3.2.5 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,

Figure 3.1: Map of Human Origins populations.

and phased this dataset with shapeit4, using the same settings as for the non-imputed data.

3.2.6 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.

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 [27]. 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.

3.2.7 SOURCEFIND

I estimated ancestry proportions for each of the selected U.K. Biobank individuals using SOURCEFINDv2 [21]. 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.

3.2.8 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 [136]. 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. 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 (Appendix section E.0.2). I phased the imputed and non-imputed datasets separately using shapeit4 at default settings.

To 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

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 section). 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 section B.3) with its own population (Table 3.1).

3.3 Results

3.3.1 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% ancestry most closely related to either 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 modelled 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=7,527) 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. 3.2), 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 [137,138]. 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.

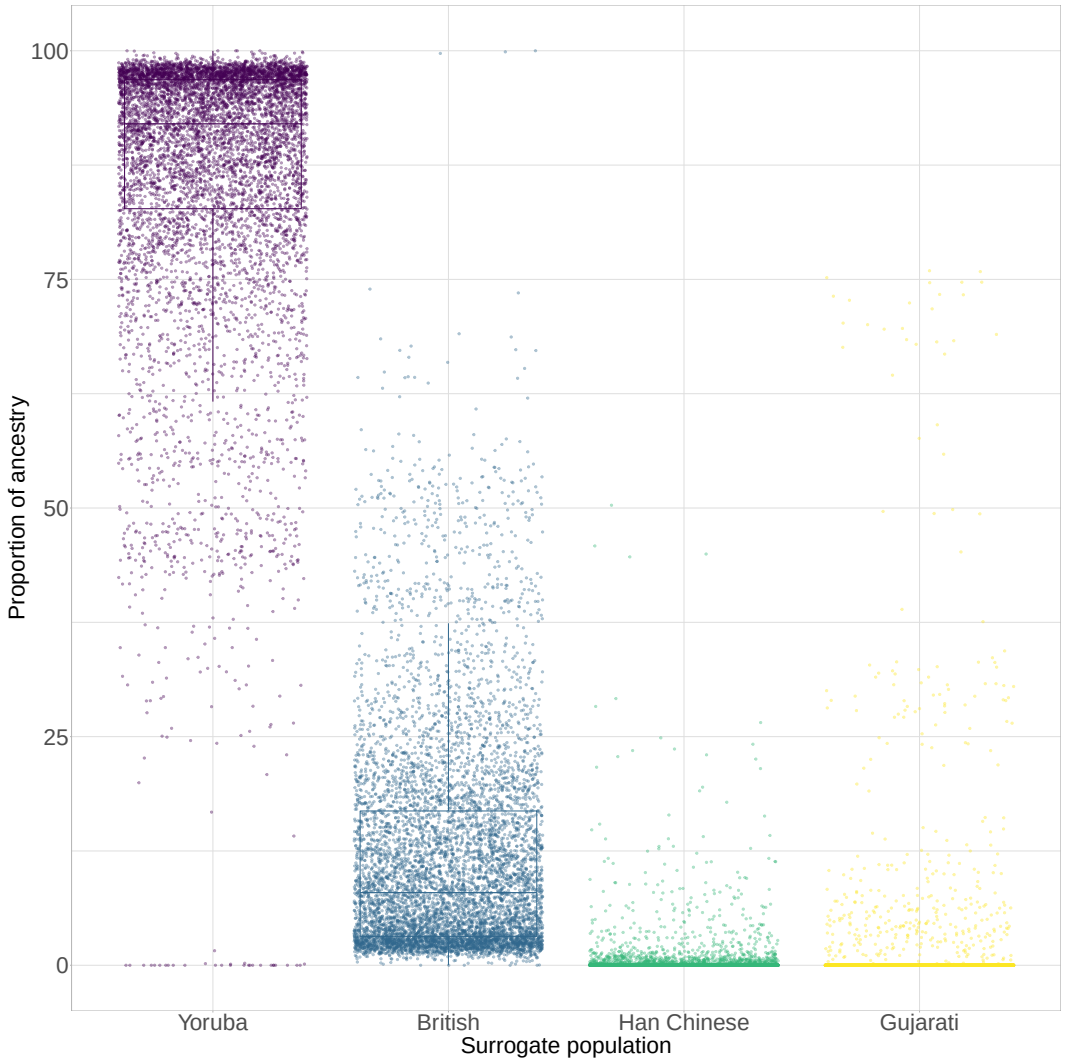

Figure 3.2: Ancestry proportions inferred from supervised ADMIXTURE run (k=4) for all individuals who self identified as being either "Caribbean", "African" or "Black or Black British". Points within each column are given random jitter to improve visual clarity.

3.3.2 To impute or not?

In order to use the Human Origins dataset as a reference in ChromoPainter analysis to 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 [139], analysing 70,000 SNPs may substantially decrease any potential power gains from modeling haplotypes to detect fine-scale differences between populations. 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, I returned to the imputed and unimputed Human Origins datasets I describe in Section 3.2.8. Recall here I reduced the Human Origins dataset to 70K SNPs and then imputed to approximately 500,000 SNPs using HRC 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 64.47K 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, 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 section 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

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

Table 3.1: 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.

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 section B.3) with the mean copyvector of all other individuals in its own population (Table 3.1).

Populations in the 70,000 non-imputed painting matched more to and had a lower TVD with their own mean population copyvector than the 500,000 non-imputed painting. Whilst it seems counter-intuitive 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 (Fig. E.2) (i.e. there is no clear benefit to using more than 50,000 SNPs in terms of assigning individuals to a population). This is reassuring and suggests there is no loss of power when using the 70,000 SNP set. This 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 (Fig 2.13). 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 one of the few 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 3.3 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. 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 ($p = 0.00026$). This permutation test was motivated by the Wilcoxon Rank Sum Test.

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.

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

Using approximately 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=5,577).

Principal component analysis on the resulting chunkcounts coancestry matrix reveals the general structure of the selected individuals, alongside the reference populations (Fig. 3.4). 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

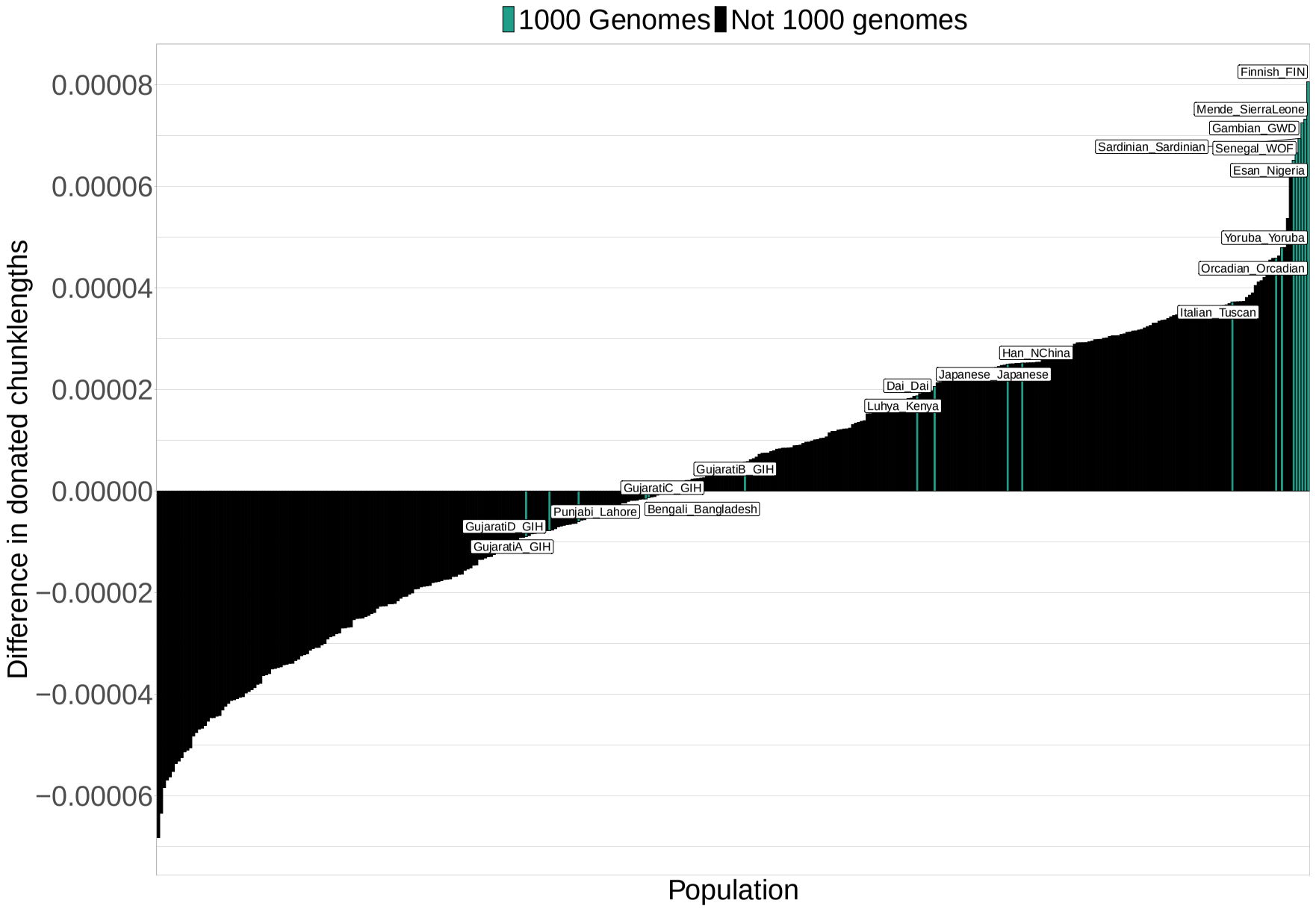

Figure 3.3: Differences in the amount donated by populations when using imputed and non-imputed data. Each vertical bar corresponds to a single population (N=395), with the height of the bar corresponding to the difference in haplotype donation, with positive values indicating increased donation and 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.

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 [140]. 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 coancestry 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. 3.5). 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. 3.5 should not be taken literally as an exact reflection of the ancestry distribution.

The map supports the findings from the PCA in Fig. 3.4; the populations with the largest contribution are those from West Africa (Fig. 3.5). 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 [141]. Given

Figure 3.4: 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.

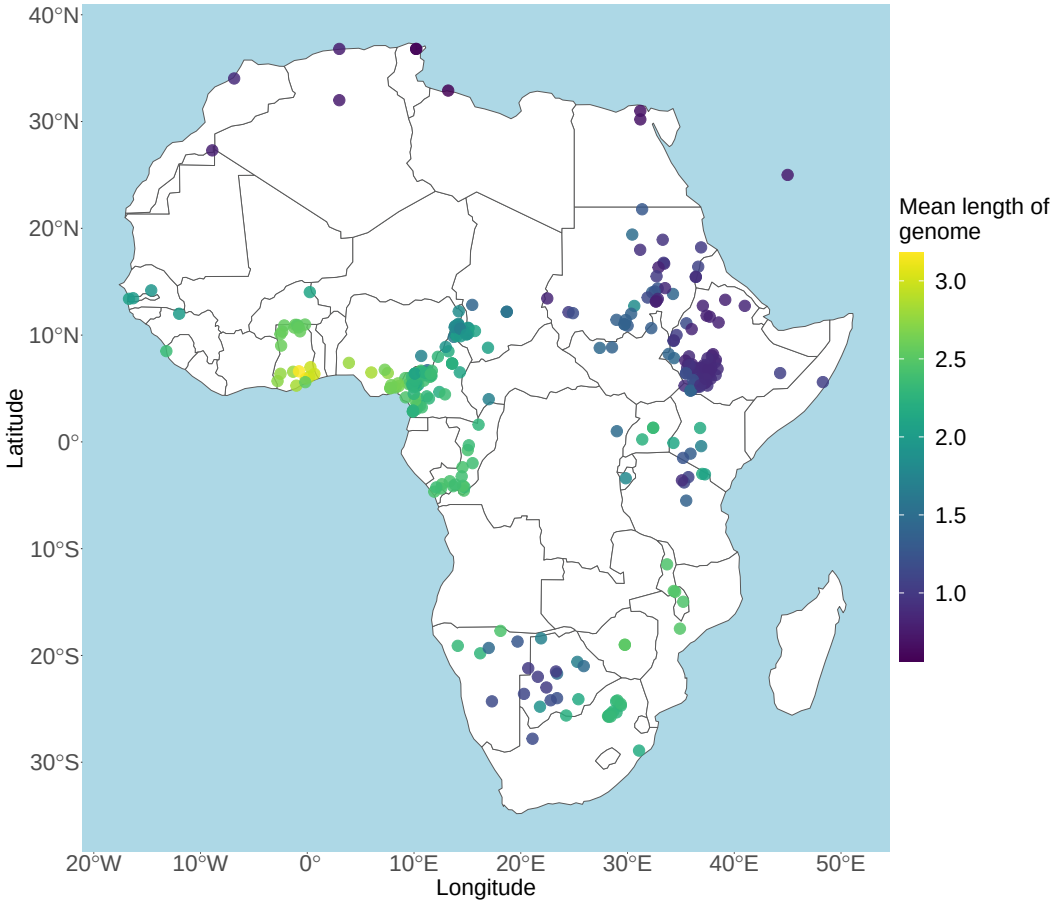

Figure 3.5: Map of haplotype donation to U.K. Biobank individuals. Each point represents a different African population. Colour corresponds to the mean length (cM) that populatation donated to all African U.K. Biobank individuals.

the 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 [142], 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 [140]. Although there are a number of immigrants from other parts of Africa, reflected in the non-zero 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. 3.6, 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. 3.5. 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. 3.7 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

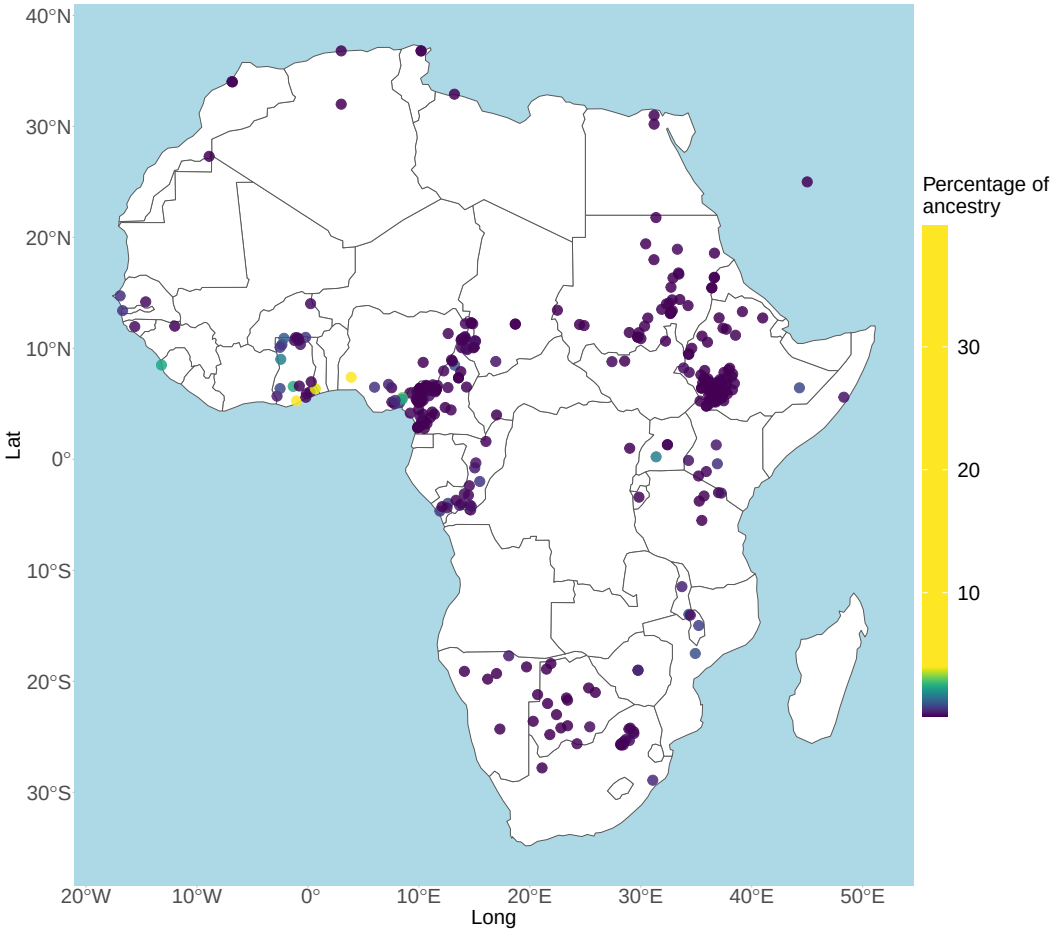

Figure 3.6: Map displaying the mean proportion of SOURCEFIND estimated ancestry of each African reference population within U.K. Biobank individuals. Each point is an African reference population with the colour corresponding to the mean ancestry proportion for that population across selected U.K. Biobank individuals. The colour-bar has been rescaled as two populations, Yoruba and Ghana_Fante have substantially higher proportions than all other populations.

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 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.

There are other instances of an over and under-representation of one ethnic group from a particular country (Fig. 3.8). 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 (personal communication N.Bird, 2021) that using the Human Origins dataset, there is inability to distinguish between individuals in closely related Sudanese populations.

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. 3.7). This may be because Somalis are more recent immigrants to the UK and therefore tend to be less admixed with Europeans relative to other immigrant populations which have been in

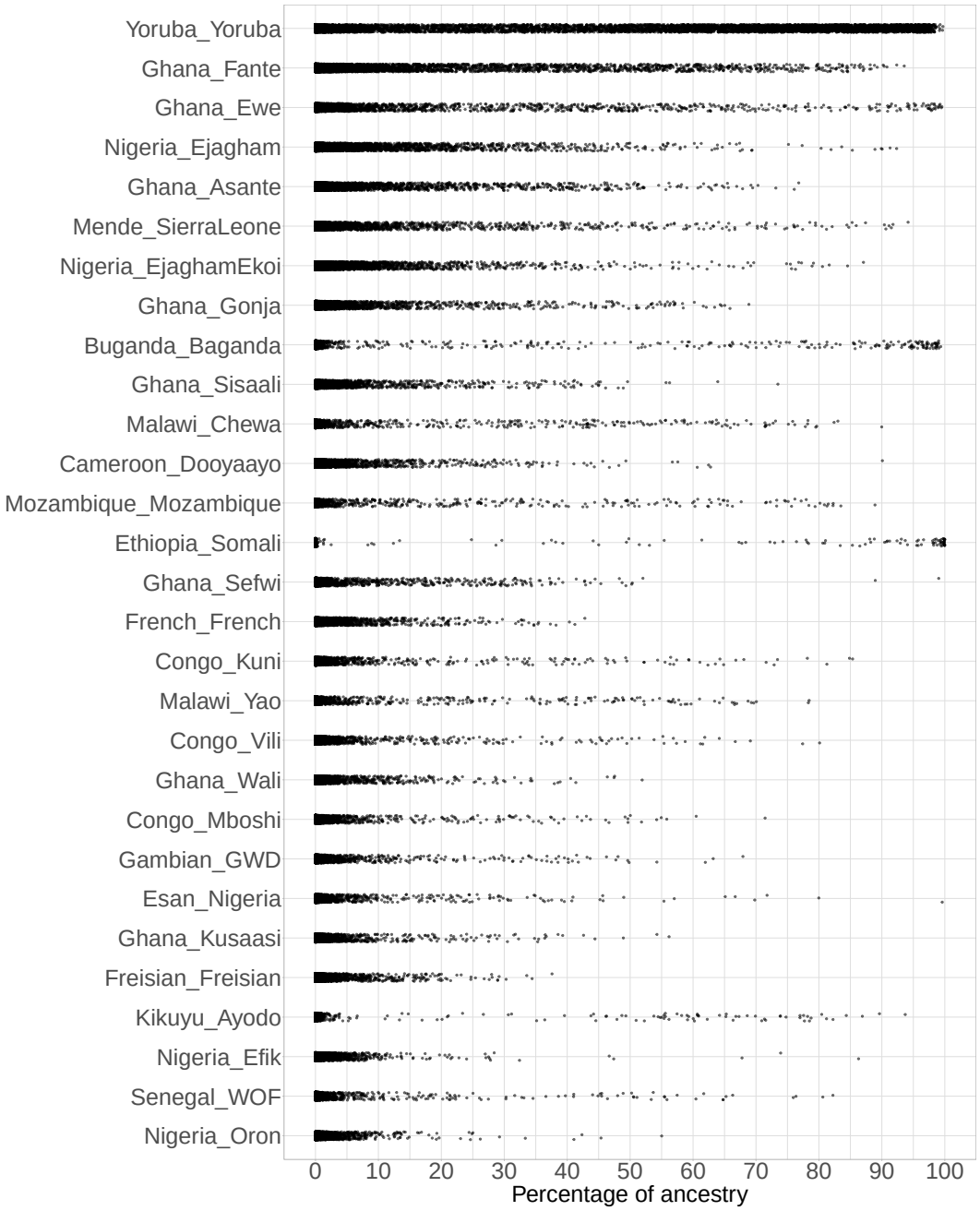

Figure 3.7: The 30 Human Origins populations which have the highest contribution to all U.K. Biobank individuals with at least 50% African Ancestry, based on SOURCEFIND analysis. Each row of points contains 8476 individuals and their position corresponds to the percentage of ancestry from that population.

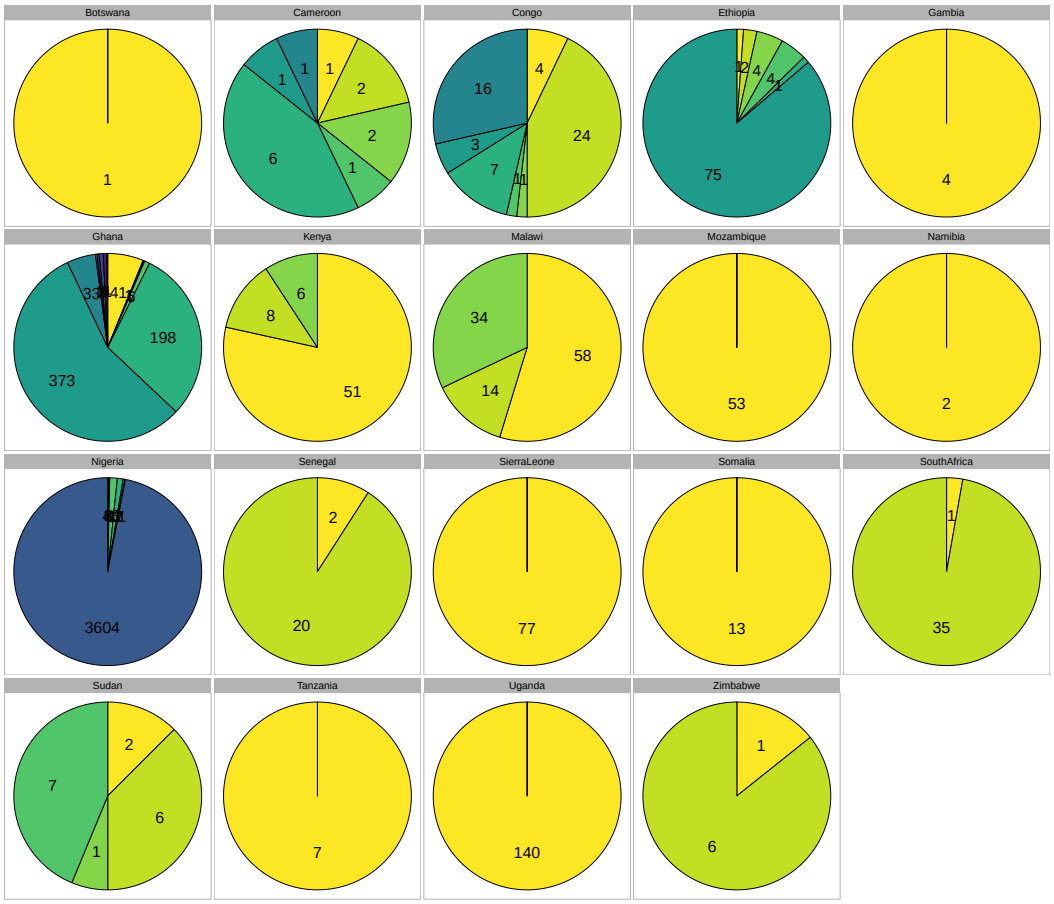

Figure 3.8: Variation of individuals assigned to different ethnic groups by country of assigned group. Each panel represents all individuals assigned to an ethnic group from that country, with proportions of each pie corresponding to proportion of individuals of that ethnic group in that country. Numbers within each slice correspond to total number of individuals within a given ethnic group.

the U.K. longer and hence can be modelled 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.

3.3.4 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).

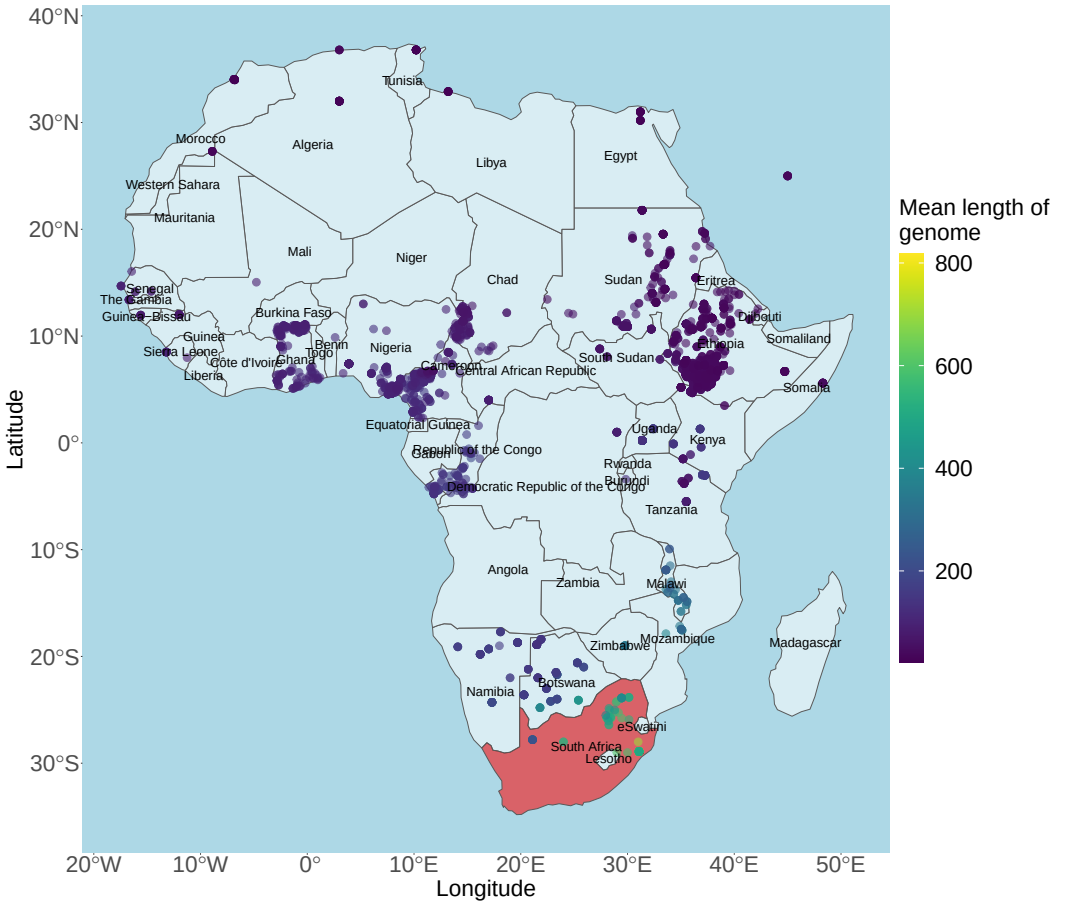

Figure 3.9: Map of haplotype donation to U.K. Biobank individuals born in South Africa. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in South Africa.

Fig. 3.9 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 (Fig 3.10).

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

who were born in the Caribbean. Visualising the haplotype donation map for these individuals shows that they are primarily of West African ancestry (supplementary figure D.4), consistent with historical evidence [141]. Individuals born in Brazil have ancestry from further South, again consistent with historical evidence (supplementary figure D.3). 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.

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

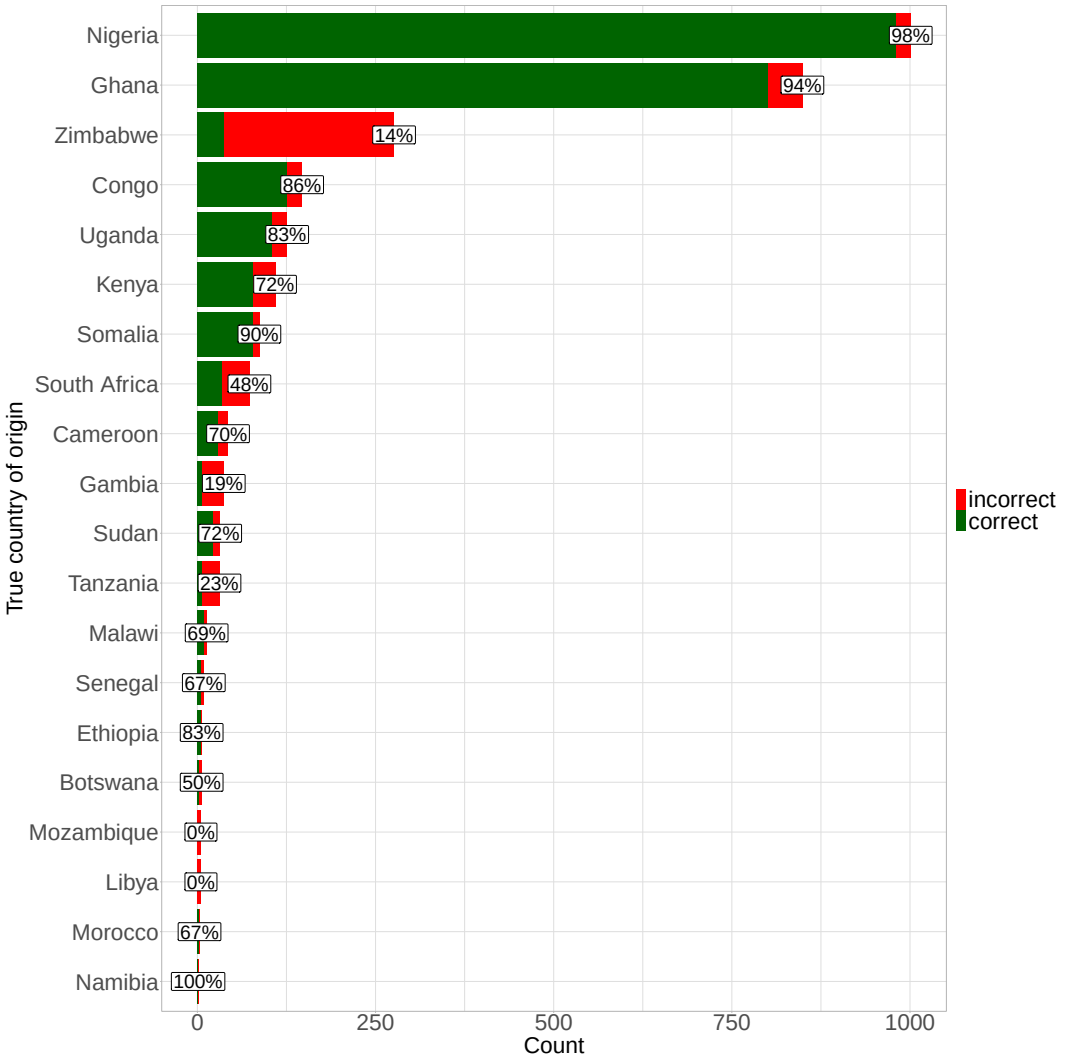

Figure 3.10: Correspondence of true birth country with estimated birth country. Each bar corresponds to a true birth country, with the length of the bar corresponding to the total number of people in our dataset born in that country. The green section corresponds to the total number of individuals where the birth country was correctly guessed and the red section to those who were incorrectly guessed. Percentage labels give percentage correct for that country.

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 population assignment results are less accurate if using imputed data due to reference bias (Table 3.1). 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. To test whether this is the case, I took all ethnic groups from Nigeria, Cameroon and Ghana in the Human Origins dataset which had five or more individuals (n=51 populations, n=1203 individuals), and for each individual, estimated ancestry proportions of each of the 51 populations. I performed this analysis for both datasets containing no imputed SNPs and 70% imputed SNPs. For each dataset, I took the average proportion of ancestry for each ethnic group across all individuals.

Fig 3.11 shows that there are substantial differences between the proportions obtained from imputed and non-imputed datasets, showing sub-country assignment is affected by imputation. In particular, there is less variance across the proportions for the imputed dataset (var=0.67) relative to the non-imputed dataset (var=0.87). This is clear on the figure, as there are many population bunched around the 2% point for the imputed dataset; the same populations are spread across a wider range of values for the non-imputed dataset.

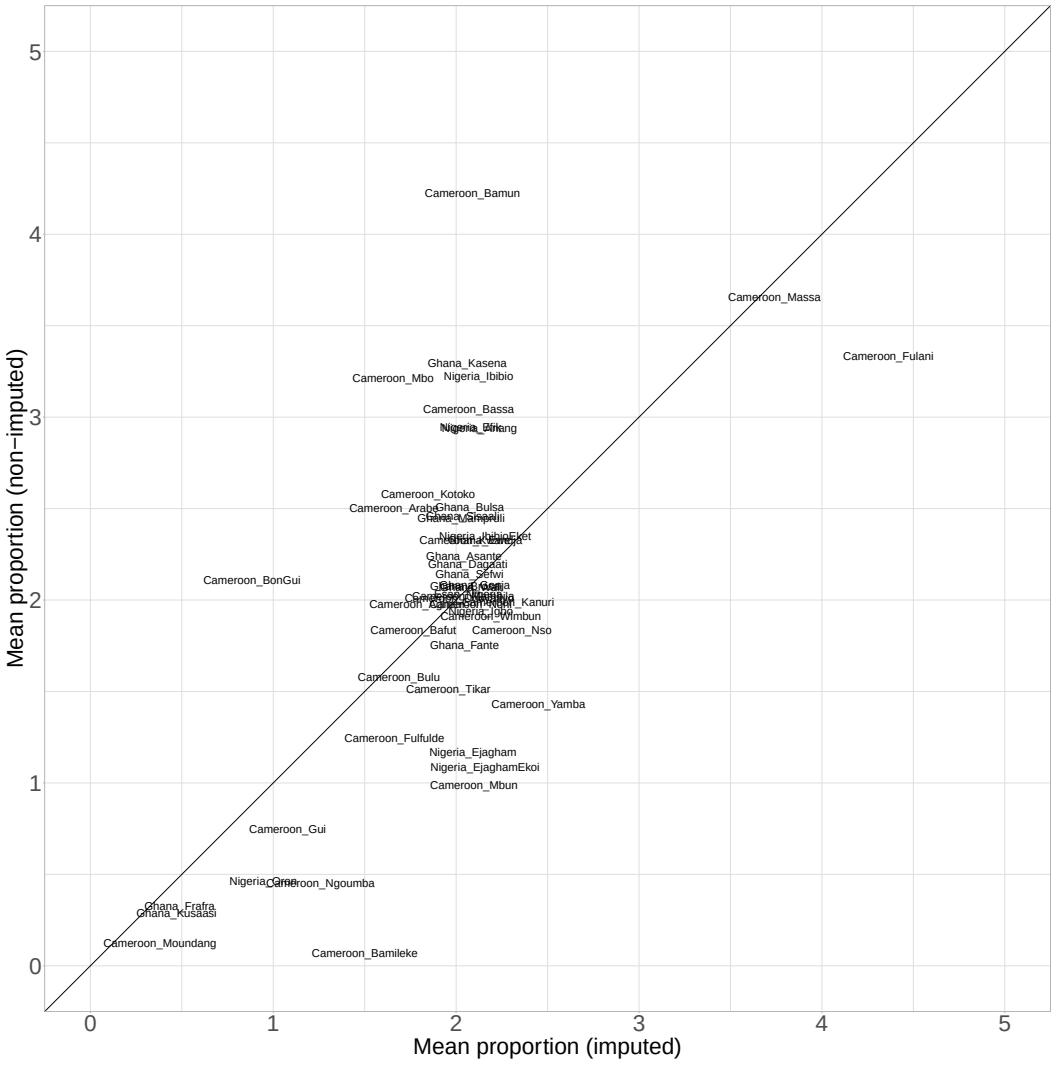

Figure 3.11: Mean ancestry proportions averaged across 1203 individuals from Ghana, Nigeria and Cameroon of 51 populations from the same countries. Proportions obtained from data containing 70% imputed SNPs (x-axis) and no imputed SNPs (y-axis).

3.3.5 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 are distributed across the U.K. There were no apparent outliers in terms of any centres with substantially larger proportion of individuals who had at least 50% African ancestry than others (Supplementary Fig. D.5). However, as expected, centres 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 centres on a map of the U.K (Fig. 3.12). No clear pattern was apparent, other than Yoruban ancestry dominating most centres, with some smaller testing centres 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 [143].

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 $\sum_{i=1}^{L}[p_{i,j} \cdot log(p_{i,j})]$ [144], 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 centres 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. 3.12).


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

3.3.6 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 (Fig. 3.13). 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 3.13 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.

3.4 Summary of Results and Discussion

The aim of this chapter was twofold; firstly, to determine whether using less dense non-imputed or more dense imputed SNPs is preferable when combining genotype data from multiple chips. Secondly, I wanted to explore the diversity of African ancestry in the U.K. Biobank and its relation to population history.

I also 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 and accurately predict the birth country of a sample. 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.

West African ancestry was the most common across samples with recent African ancestry, with ancestry from ethnic groups from Nigeria and Ghana

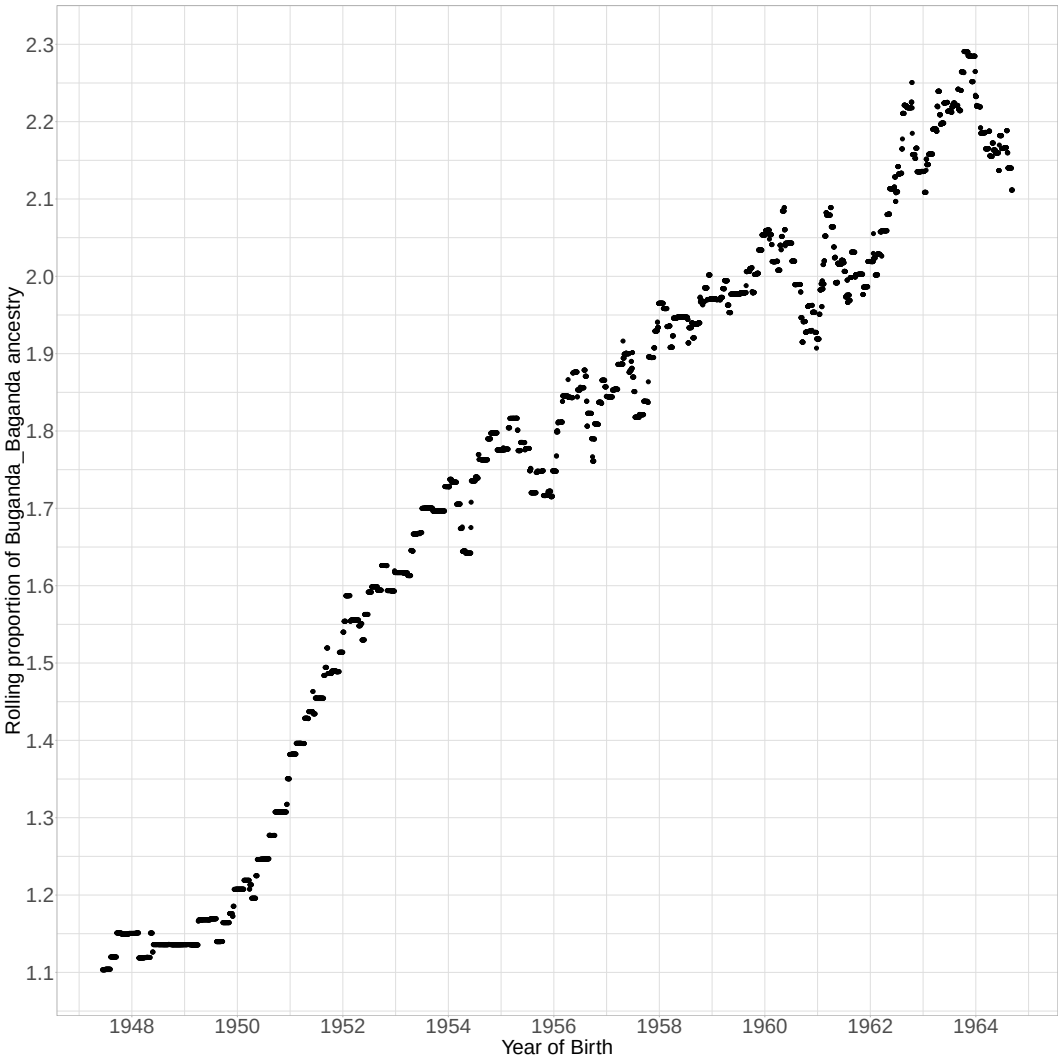

Figure 3.13: Increase in the mean proportion of Buganda ancestry between 1948 and 1965. An overlapping sliding window was applied to SOURCEFIND ancestry proportions and mean proportion of Buganda ancestry for each window plotted against the mean birth-date of individuals in that bin.

being especially prevalent. In particular, individuals had substantially more ancestry from Yoruba than any other ethnicity. I did not find evidence for structure in how African ancestry was distributed across the U.K., based on the testing centre that participants registered at.

Future work on using Biobanks to explore population structure and history could focus on two points. Ideally, I would liked to have painted the entire U.K. Biobank dataset using the Human Origins dataset as a reference panel, rather than restricting analysis to individuals with 50% recent African ancestry. This would have allowed me to analyse a substantially higher amount of African haplotypes across the entire dataset and this give a more complete extent of African ancestry in the U.K. Thus, the development of efficient methods, likely based on the PBWT, which allow for Biobank-scale datasets to be painted by large reference panels would accelerate research into ethnic minority ancestries.

Secondly, larger reference panels of worldwide populations and more ethnic groups will allow for a more detailed characterisation of genetic variation. Similarly, including details on ethnic identity in Biobank projects would improve the resolution at which analysis could be carried out.

Chapter 4

Bavaria ancient DNA

4.1 Introduction

Throughout the Pleistocene and Holocene, Germany has been the setting for many population movements and admixture events of modern humans. The Swabian Alps is home to some of the earliest pieces of symbolic art, dated to at least 32kya [145] and musical instruments dated to 40kya [146], both assigned to the Aurignacian tradition.

Later, the region was also home to one of the first Neolithic traditions in the *Linearbandkeramik* (LBK), a key culture in the Neolithisation of Europe. Early LBK populations across Germany mixed with the preceding Mesolithic hunter gatherer populations [105, 147–150]. At the end of the Neolithic, a new ancestry was detected [105, 151] in concert with the arrival of the Corded Ware Complex [152], most closely related to the Yamnaya Pastoralists from the Pontic-Capsian Steppe. Recent studies using ancient DNA have shown that the arrival of Steppe-related ancestry in Europe occurred no earlier than 2700BC [153] and spread widely shortly after.

During the Bronze Age, cultures closely related to Yamnaya, such as Bell Beakers, Corded Ware and Unetice [105] appeared across Germany at sites

such as Kromsdorf [154] and Tollense [155, 156]. It was later dominated by Iron Age cultures such as Hallstatt and La Tène, which have been shown to be partially continuous with the preceding Bell Beaker culture [157].

In the present-day, Germany represents a boundary point between East and West Europe, with a relatively sharp genetic boundary occurring between Germany and Poland to the east, given their close geographic proximity [158–160]. However, within Germany, SNP-based studies have shown that there is only very weak substructure [161]. Questions remain as to the origin of this East-West structure; is it recent structure, or does it persist to the Middle Ages or earlier?

Cherry-Tree cave, or *Kirschbaumhöhle*, represents a unique opportunity to study a transect of southern German samples from the Neolithic to the present-day. The cave represents a relatively untouched layer of stratigraphy, with a large series of radiocarbon dates revealing that human and animal inhabitation of the cave stretches back until at least the Michelsberg Culture in the Early Neolithic [162].

Here, I analyse novel data from 11 medium-to-high coverage samples from two sites from Southern Germany and one site from from Southern Austria. In particular, the samples from Kirschbaumhöhle span from the Late Neolithic to the Iron Age, providing an opportunity to study a time transect in a narrow geographic region (Table. 4.1).

A collaborator, Prof. Joachim Burger, Johannes Gutenberg University Mainz, posed the following three questions.

1. **Second Neolithic immigration wave**. One of the samples (Erg1) is thought to have belonged to the first wave of farmers carrying farming technology from the near-east to Europe, and another (DIN2) to the second wave. Do we observe genetic differences between the two waves of

samples and do they show evidence of previously reported hunter-gatherer admixture?

2. **Cherry Tree Cave**. Do we see evidence of genetic continuity from from the Late Neolithic through to the Iron Age in Cherry Tree Cave?

3. **Germanic / Slavic divide**. Is there a distinction between the Germanic and Slavic samples from the Middle Age samples? How do these populations compare to the preceding samples from the Bronze and Iron ages?

4.2 Methods

4.2.1 Data generation

Eleven whole-genomes of ancient individuals were generated by collaborators at the Johannes Gutenberg, University of Mainz, Germany. The estimated radiocarbon dates range from 5200B to 1060AD (Fig. 4.2). Six of the samples were found in Cherry-Tree Cave in the Bavarian district of Forchheim, four from futher South in the region of Dingolfing/Essenbach and one sample from Molzbichl in southern Austria (Fig. 4.1). The samples had a median coverage of 4.84x and ranged from 0.7x to 17.52x. Full details of coverage, location and dates are given in Table 4.1.

I was given the data of each newly sequenced sample in `vcf` format.

4.2.2 Genotype imputation and phasing using GLIMPSE

In order to compare the genetic variation in the newly sequenced samples to a reference dataset, I merged them with the 942 ancient samples from the literature detailed in Appendix section A.1, resulting in a total of 955 samples in `.bcf` format with genotype likelihood data at 77,213,942 genome-wide SNPs.

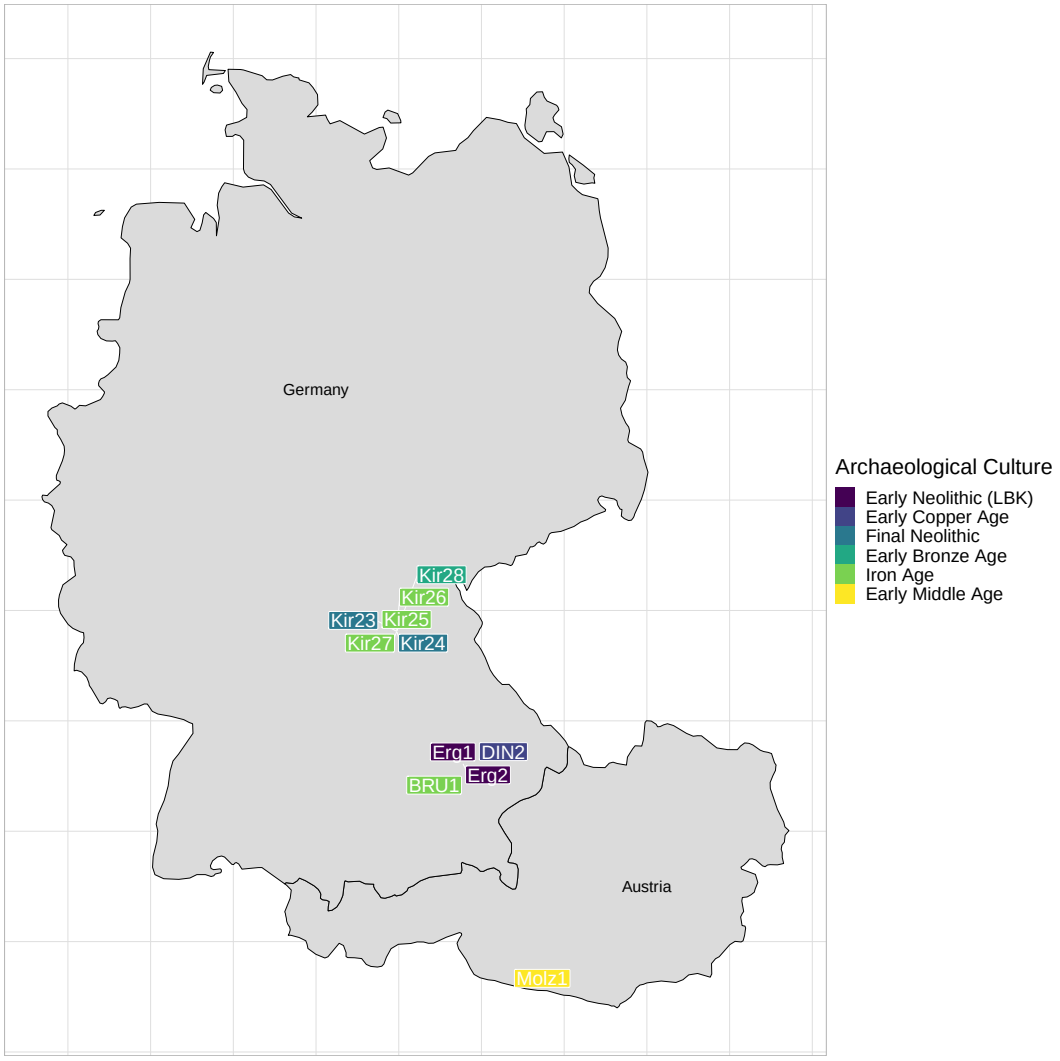

Figure 4.1: Map of newly sequenced ancient individuals, positioned according to where they were excavated. Colour on label corresponds to archaeological culture which they were found.

I followed the recommended GLIMPSE [91] imputation and phasing pipeline (`https://odelaneau.github.io/GLIMPSE/tutorial_b38.html`), using the 30x-coverage 1000 genomes dataset [102] as a reference panel. This resulted in phased haplotypes and posterior genotype likelihoods for each of the 955 individuals.

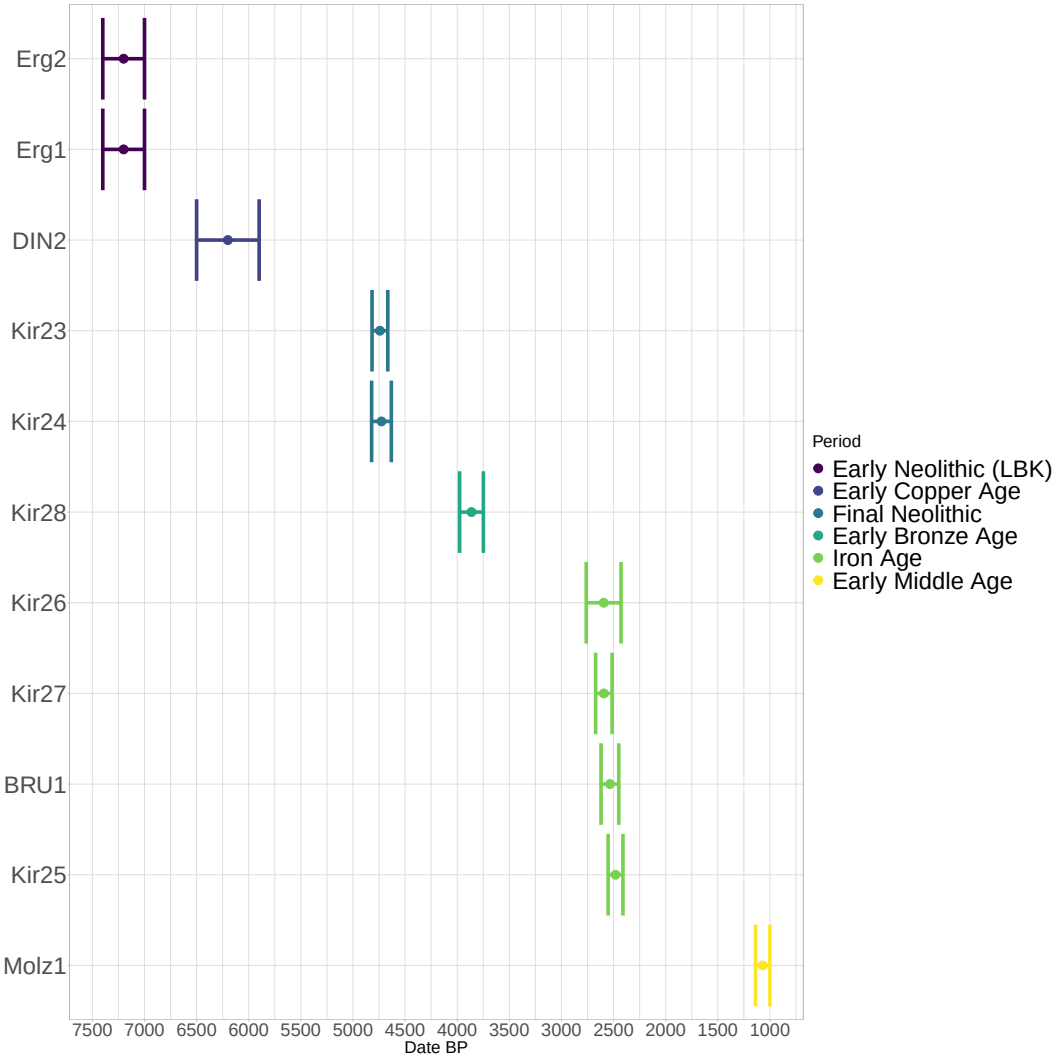

Figure 4.2: Estimated radiocarbon dates for each newly sequenced ancient individual, grouped by archaeological period. Error bars correspond to upper and lower 95% quantiles of the mean date.

4.2.3 Uniparental haplogroups

To determine the mtDNA and y-chromosome haplogroups for each newly sequenced ancient sample, I used Haplogrep (`https://haplogrep.i-med.ac.at/`) [163] on the raw `.fastq` file for each sample.

4.2.4 IBD sharing

I used hap-IBD [30] to estimate IBD segments greater than 2cM in length between all pairs of ancient individuals above 1.5x coverage (n=466),

Sample.ID	Location	Date	UQ	LQ	Period	Sequencing Depth
Erg1	Ergoldsbach	5200	5400	5000	Early Neo (LBK)	4.52
Erg2	Ergoldsbach	5200	5400	5000	Early Neo (LBK)	0.71
DIN2	Dingolfing	4200	4500	3900	Early Copper Age	1.71
Kir24	Cherry Tree Cave	2762	2821	2632	Final Neo	3.98
Kir23	Cherry Tree Cave	2741	2817	2666	Final Neo	17.52
Kir28	Cherry Tree Cave	1863	1977	1749	EBA	17.30
Kir26	Cherry Tree Cave	595	762	428	Iron Age	4.84
Kir27	Cherry Tree Cave	593	672	514	Iron Age	16.60
BRU1	Bruckberg	535	620	450	Iron Age	11.54
Kir25	Cherry Tree Cave	481	552	410	Iron Age	4.55
Molz1	Molzbichl	1069	1138	1000	Early Middle Age	13.22

Table 4.1: Details of newly sequenced ancient DNA samples. UQ and LQ give upper and lower 95% quantile estimates for radiocarbon dates. EBA is Early Bronze Age.

using the phased output from GLIMPSE as input haplotypes, the genetic maps from (`http://bochet.gcc.biostat.washington.edu/beagle/genetic_maps/plink.GRCh37.map.zip`) and leaving all parameters as default.

4.2.5 plink PCA

To obtain a broad overview of the ancestry of the newly sequenced individuals in the context of the 942 literature samples detailed in Appendix section A.1, I performed PCA on the pre-imputation genotypes using plink2 [164].

I retained the 500,000 markers with the lowest amount of missingness across all samples and LD-pruned the resulting SNPs using the settings `-maf 0.01` and `-indep-pairwise 50 5 0.2`. PCA was performed using default settings from plink2.

4.2.6 ChromoPainter and fineSTRUCTURE analysis

To characterise the ancestry of the newly sequenced ancient samples in the context of other ancient individuals, I first selected all newly sequenced samples

and literature samples above 1.5x coverage (n=466) and performed an 'all-v-all' painting where each sample was painted using all other samples. 1.5x was somewhat arbitrarily chosen as my previous work has shown this is a suitable threshold for the inclusion of samples for ChromoPainter analysis (section 2.6.4); whilst I show 0.5x as the cut-off for coverage-related effects, I chose to be conservative and opt for a higher threshold, given all but one of the 11 newly sequenced samples have average coverage > 1.5x. I used this painting, hereafter referred to as 'ancient' painting, to perform fineSTRUCTURE clustering and tree building on the ancient samples.

I performed Principle Component Analysis on the coancestry matrix of the 'ancients' painting using the `prcomp_irlba` function from the irlba R library. To account for the fact that the diagonals of the coancestry matrix are always zeros (as an individual cannot be painted by themselves), I set the diagonal of each row to be the mean of that row, following Lawson et al 2012 [19]. Although there were 466 individuals in the 'ancients' painting, not all of these were included in the chunklengths PCA. This was because many individuals in that set were not relevant to exploring the ancestry of the Bavarian individuals. For instance, when plotted, samples such as those from the Xiong Nu, a 3rd century BC culture from inner Mongolia, dominate the variation in a PCA to the point where identifying structure between the samples of interest becomes challenging. Therefore I removed 327 individuals based on visual inspection of the first two principal components.

To determine the genetic similarity between the newly sequenced ancient samples and present-day populations, I performed an 'all-v-all' painting using a selected group of 26 present-day European populations (Table 4.2) from the HellBus dataset (described in Appendix section A.4) and the 11 newly sequenced ancient individuals, hereafter referred to as 'present-day painting'.

I applied fineSTRUCTURE (v0.0.5) [19] to cluster the chunkcounts ChromoPainter output for the 'ancients' painting. fineSTRUCTURE assigns individ-

Population	Number of samples
HB:belorussian	9
HB:bulgarian	31
HB:croatian	19
HB:cypriot	12
HB:french	28
HB:german	30
HB:germanyaustria	4
HB:greek	20
HB:hungarian	19
HB:irish	7
HB:lithuanian	10
HB:mordovian	15
HB:northitalian	12
HB:norwegian	18
HB:polish	17
HB:romanian	16
HB:russian	25
HB:scottish	6
HB:siciliane	10
HB:southitalian	18
HB:spanish	34
HB:tsi	98
HB:tuscan	8
HB:ukrainian	20
HB:welsh	4
HB:westsicilian	10

Table 4.2: Name of population and number of samples used in the present-day ChromoPainter, MOSAIC and qpAdm analyses.

uals to clusters, estimates the number of clusters and builds a dendrogram of genetic similarity based on a tree-building algorithm. This is particularly useful when combining many samples from different studies, as is the case with the 'ancients' painting; the population label identifiers used by different studies may not be consistent with one another. Therefore, we can use fineSTRUCTURE groupings as population labels rather than group labels. fineSTRUCTURE was first run in MCMC mode using 1,000,000 burn-in MCMC iterations and 2,000,000 main MCMC iterations. It was then run in tree-building mode (`-m`

T) using 100,000 burn-in and 100,000 main iterations.

Tree figures, coancestry matrix figures and principle component plots were generated using the fineSTRUCTURE R library (`https://people.maths.bris.ac.uk/~madjl/finestructure/FinestructureRcode.zip`).

4.2.7 SOURCEFIND

I used SOURCEFIND [21] to infer the proportions of ancestry by which each newly sequenced ancient individual is most related to a set of surrogate populations. While this method does not explicitly attempt to identify admixture, in contrast to e.g. ALDER [165] or GLOBETROTTER [20], it can reflect admixture proportions [21] but more generally reflects recent ancestry sharing patterns.

The first analysis used the ancients painting and only three surrogates: Western Hunter-Gatherers, Neolithic farmers from Anatolia and Yamnaya, to mimic previous research suggesting many ancient Europeans descend from the mixture of three sources well-represented by these groups [51]. The second analysis attempted to characterise more fine-scale ancestry patterns, by modelling each target ancient individual (using the same ancients painting) as a mixture of all sampled ancient populations above 1.5x coverage (n=466) that had an average sample age no more than 100 years younger than that of the target individual. The third analysis used the "modern" painting and formed each ancient individual as a mixture of all present-day populations shown in Table 4.2. For each of these analyses, I found the mean and 95% credible interval of ancestry estimates across 2,000,000 posterior samples combined from three independent SOURCEFIND runs that each sampled every 10,000 MCMC iterations after discarding the first 10,000 MCMC iterations as "burn-in".

4.2.8 MOSAIC admixture analysis

I inferred admixture events, dates and proportions in newly sequenced ancient samples using MOSAIC, a haplotype-based method [166]. While MOSAIC cannot infer multiple pulses of admixture from the same admixing sources as GLOBETROTTER [20] can, in theory it is unlikely we would have adequate power to identify such multiple pulses when analysing only a single ancient sample, as is the case in this study. Furthermore, the 'painting' step and admixture inference step in MOSAIC are combined, providing a simpler pipeline and more flexible assignment of different surrogates relative to GLOBETROTTER (i.e. the set of surrogates can be changed without repainting the samples).

I performed two MOSAIC analyses that correspond to two of the SOURCEFIND analyses described in Section 4.2.7. First, I performed an 'ancient surrogates' analysis where the all ancient samples above 1.5x coverage (n=466) were used as surrogates to admixing sources. I used the fineSTRUCTURE groupings to categorise ancient samples into surrogate populations. Second, I also performed a 'present-day surrogates' analysis where a selected set of present-day populations (Table 4.2) were used as surrogates. While using present-day populations to reflect ancestry patterns in ancient individuals may be counter-intuitive, the larger sample sizes and larger variety of present-day populations can provide more clean results relative to using ancients

I ran MOSAIC using default settings, assuming two or three admixing sources per target individual/population. For populations with more than one sampled individual, MOSAIC provided bootstrap-based 95% confidence quantiles around date estimates. MOSAIC also estimates f_{st} between the set of surrogates and the estimated 'true' mixing source, which is useful when a close proxy for the 'true' mixing source is not available

4.2.9 F-statistics

Many of the relevant samples in the literature were of very low coverage (< 0.1). As my work in section 2.6.4 indicated that samples with less than 0.5x coverage cannot reliably be analysed using ChromoPainter, I also used F-statistics [42] that are mostly robust to coverage related effects [44]. In particular I used Admixtools (`https://uqrmaie1.github.io/admixtools`) to analyse 942 individuals from 143 populations (Appendix section A.1, including many low-coverage samples from relevant LBK cultures presented in Rivollat et al (2020) that would not have been suitable for use with ChromoPainter [167]. This analysis also incorporated 2280 present-day individuals from 144 populations from the HellBus dataset as putative ancestry surrogates for tested ancient individuals. Populations shown in Table 4.2.

For the input to ADMIXTOOLS, I used the genotyped imputed from GLIMPSE, as it has been shown that using imputed markers reduced reference bias relative to using pseudo-haploid markers [48]. I then used the f_4 branch test to test whether two populations form a clade relative to two other populations. For example, the expected value of $f_4(french, german; yoruba, mbuti)$, which tests whether {french,german} form a clade relative to {yoruba,mbuti}, should not give a score significantly different to zero. In contrast, exchanging $french$ with *yoruba* would yield a significantly positive f_4 scores, with strength of evidence to reject the null ($f_4 = 0$) measured using standardised Z-statistics.

I also used the f_3 test, denoted $f_3(A, B; C)$, to (i) estimate the branch length between A and B after their divergence from C, or (ii) test whether C descends from an admixture event between sources represented by A and B. The latter can occur if C has a substantial number of SNPs with allele-frequencies which are intermediate between A and B.

Finally, I used qpAdm to infer ancestry proportions, following the protocol described in Olalde et al (2018) by choosing the following populations/samples

as outgroups: *Mota*, *Kostenki14*, *papuan*, *han*, *hannchina*, *mbutipygmy*, *sannamibia*, *yakut*. These outgroups were suitable for use in investigating ancient Eurasians, since they are asymmetrically related to many ancient populations, but do not show evidence of recent gene flow with them.

4.3 Results

4.3.1 Broad-scale ancestry changes in Bavaria reflect those found elsewhere in Europe

The newly sequenced samples from the Early Neolithic (Erg1 and Erg2, approx 5200BC) and Copper Age (DIN2, approx 4200BC) cluster with other literature samples from European Neolithic on the plink2 PCA (Fig. 4.3). As in previously reported PCA results [147], the earliest Neolithic samples, from Anatolia and Greece, and who are thought to be the source population from which all subsequent Neolithic farmers derive [51, 150, 168–170], are positioned at the end of the cluster farthest away from the hunter-gatherer samples (for example, WHG on Fig. 4.3). This likely reflects the fact that they are unadmixed with respect to the later Neolithic samples. As the Neolithic progressed, farmers from the near-east mixed with local hunter-gatherer groups in central Europe [147] and acquired local hunter-gatherer ancestry. Accordingly, these samples are shifted away from the earlier Neolithic samples towards the hunter-gatherers. With this in mind, the position of the new Early Neolithic sample Erg1, shifted north away from the contemporaneous sample Erg2, is suggestive of hunter-gatherer admixture.

There are four key observations from the Figure 4.3 PCA regarding the new samples:

1. The two Late Neolithic individuals are genetically separate, with Kir24 positioned close to Yamnaya and Kir23 clustering with Neolithic Euro-

peans.

2. The Bronze Age sample Kir28 clusters with other European Bronze Age samples

3. The four Iron Age samples (Kir25, Kir26, Kir27 and BRU1) cluster towards the Neolithic individuals and other European Iron Age samples

4. The three Medieval period samples (Alh1, Alh10, Molz1) cluster with the Bronze Age sample Kir28 instead of the Iron Age samples.

4.3.2 Early Neolithic

The three Early/Middle Neolithic samples, Erg1, Erg2 and DIN2, all display a strong affinity to Anatolian farmers, consistent with the prevailing theory that near-eastern farmers were responsible for the spread of early agricultural technology across Europe, and that all Neolithic farmers share recent common ancestry [51, 168–170]. fineSTRUCTURE grouped Erg1 with two samples from Upper Palaeolithic/Neolithic Italy and DIN2 with Early/Middle Neolithic samples from Germany, Greece, Anatolia and Hungary. Despite their age, the genetic variation of the Early Neolithic samples falls well within the variation of present-day individuals; when painted using present-day samples, the three Early Neolithic individuals cluster with present-day Italians, consistent with findings from previous research [51, 105] (Fig. 4.4). Erg1 was assigned to mtDNA haplogroup K which has been found in Neolithic and pre-pottery sites across Europe [150, 171] and Western Asia [172, 173].

Erg1 is from the *Linearbandkeramik* (LBK) culture and is speculated to have belonged to the first wave of immigrants carrying farming technology from south-eastern Europe or Anatolia into central Europe. DIN2 is from a nearby site, around 500 years more recent, and is thought to potentially belong to a second wave of farmers who migrated along the Danube. It is unclear to

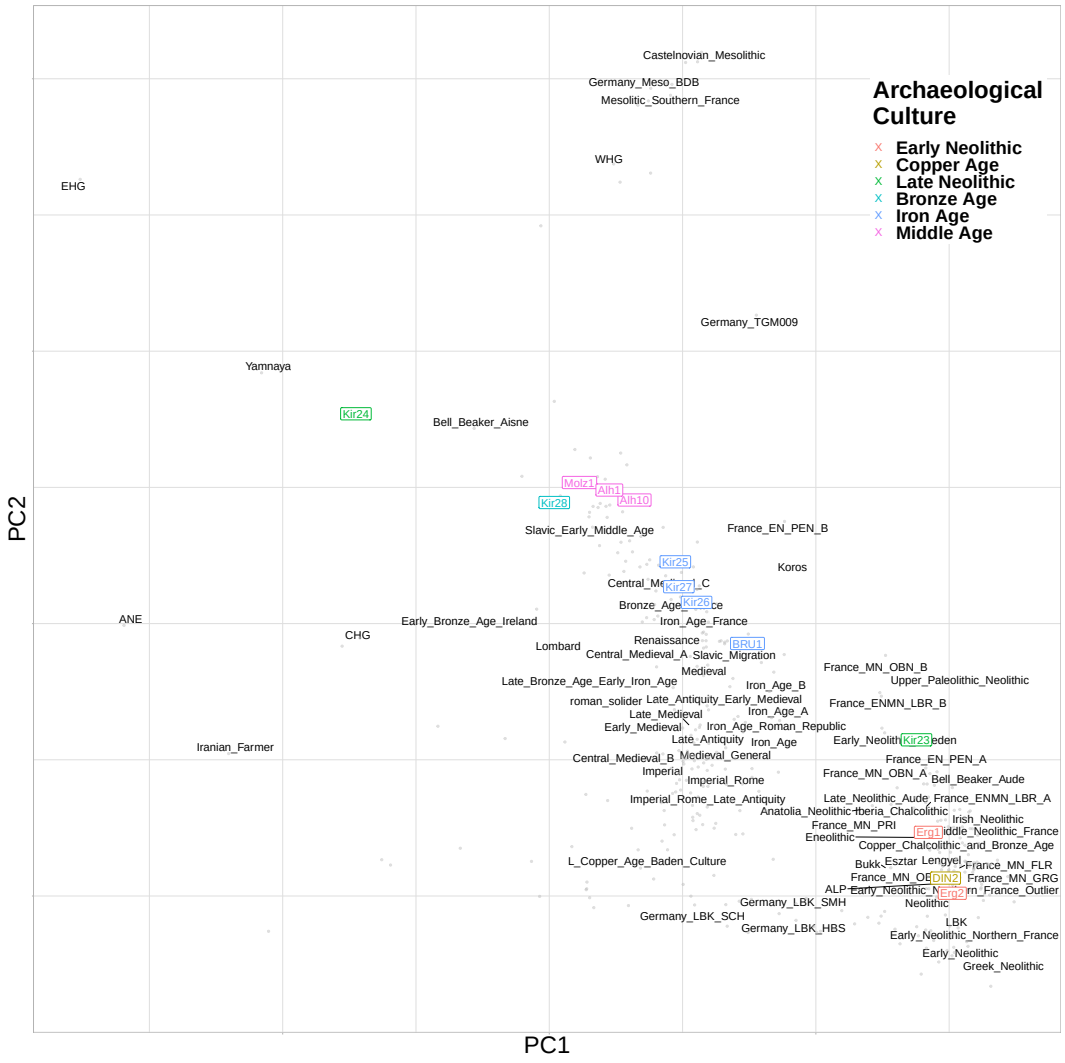

Figure 4.3: Principle component analysis of genotype matrix using plink2. Grey points indicate principle component coordinates for each sample. Black text indicated mean principle component coordinates for all individuals within that group. Coloured labels represent newly sequenced ancient samples.

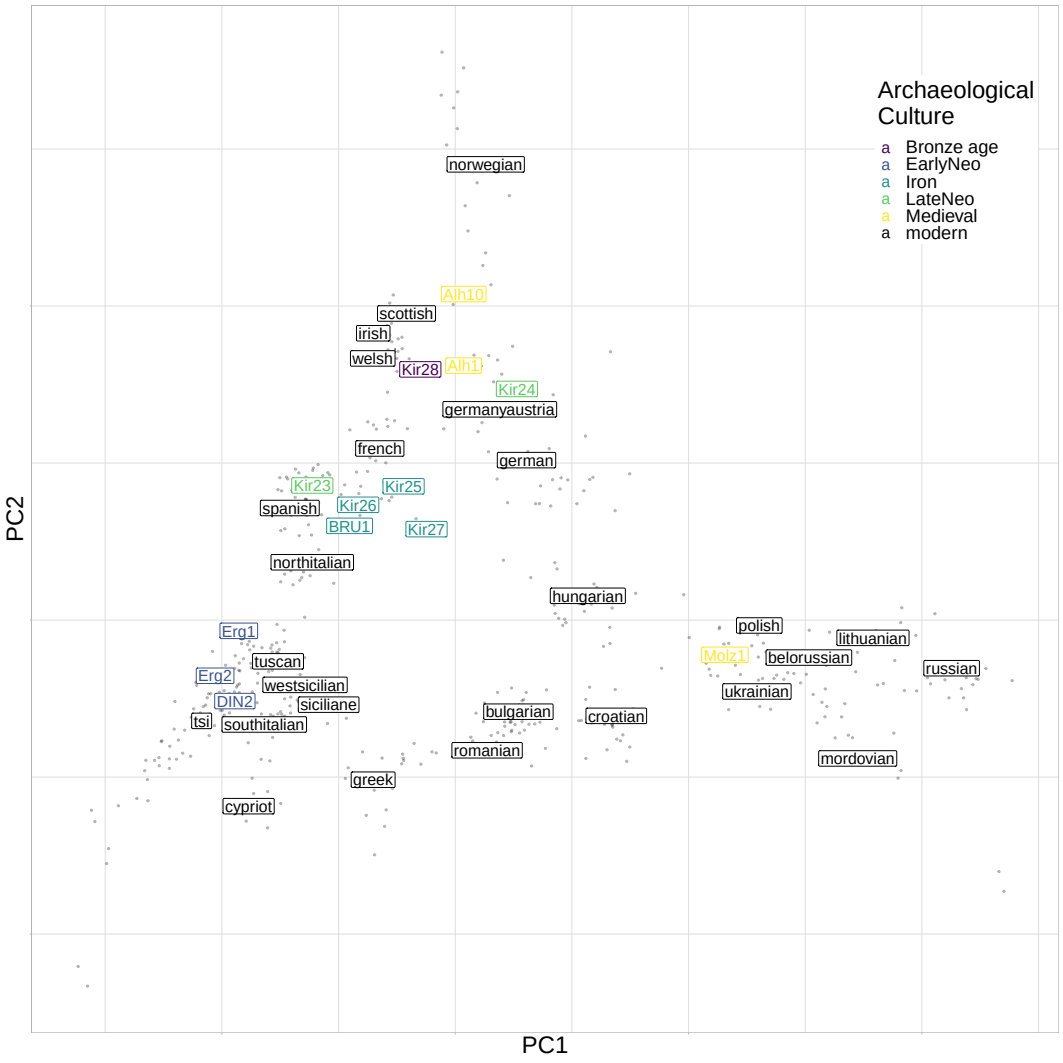

Figure 4.4: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations. The position of each reference label is the mean PC coordinates of all individuals within that population. Transparent coloured points correspond to present-day individuals.

what extent these different waves corresponded to populations with different ancestries.

When painted using 465 ancient samples from the literature and the newly sequenced samples, Erg1 had the lowest TVD (TVD is a distance metric based on ChromoPainter copyvectors; calculation and justification outlined in Appendix section B.3) with DIN2, supporting the hypothesis that they were from similar source population. DIN2 has the lowest TVD with NE5, NE4 and NE7, samples assigned to Middle and Late Neolithic cultures on the Hungarian plane, and was assigned to mitochondrial haplogroup (J1C) alongside NE4 and NE5. Both the autosomal and mtDNA link to Neolithic Hungary supports the hypothesis that DIN2 migrated along the Danubian route.

To explicitly test whether Erg1 and DIN2 group together to the exclusion of other ancient samples and therefore, whether they likely originated from a similar source population, I performed f_4 tests in the form of $f_4(W = Erg1, X = DIN2; Y = test, Z = Mbuti)$, where *test* is 143 ancient populations used in the F-statistics analysis. This tests whether Erg1 and DIN2 form a clade to the exclusion of *test* or not. Of the 143 comparisons, only the population labelled as WHG had a $|Z| > 3$, ($Z = 3.057$), suggesting that Erg1 and DIN2 originate from the same local population. Note that one test with $|Z| > 3$ may be expected when doing 143 tests, even if the null is true.

To determine whether Erg1 showed increased genetic similarity to local farming populations, I also performed combinations of f_3 in the form of $f_3(A = Erg1, B = test, C = Mbuti)$, where *test* iterates across 143 ancient populations. This tests the branch length, or the amount of genetic drift that has occurred on the branch shared by Erg1 and *test* since their divergence from an outgroup (Mbuti). The sample/population with the highest f_3 statistic was NE7, a sample from 4,360 – 4,490 BC and the Lengyel culture (a Neolithic culture centered on the Danube River, known to be an offshoot of the LBK culture Erg1 belonged to). On the other hand, DIN2 shows a clear affinity to samples

from Neolithic France.

My dataset included data from several other LBK populations local to Erg1 and DIN2; samples from Schwetzingen, Stuttgart-Mullhausen and Halberstadt. These samples appear to form a distinct cluster on the plink PCA and are shifted away from the primary cluster of Neolithic individuals and towards samples from the Anatolian Bronze Age and Baden Culture (a central European Chalcolithic culture) (Fig. 4.3). I wanted to know which LBK population Erg1 and DIN2 were closest to. I found strong evidence ($|Z| = 7.97$) that Erg1 shared more alleles with LBK populations from Schwetzingen than with Stuttgart-Mühlhausen, suggesting the early LBK populations showed relatively fine-scale geographic structure. Given the lack of Hunter Gatherer ancestry in the Rivollat LBK samples, this structure seems unlikely to be driven by variable amounts of Hunter-Gatherer admixture (Fig. 4.7).

4.3.3 Variable amounts of local hunter-gather ancestry in Neolithic farmers indicates a structured population

Prior research has shown that admixture occurred between newly arrived farming immigrants from Anatolia and local hunter-gatherers [105, 147, 174–176]. The position of Erg1 on the PCA, shifted slightly north towards the majority of the Bronze Age samples, suggests that it may have a component of Hunter-Gatherer ancestry. Indeed an f_3 admixture test, using $f_3(A = CastelnovianMesolithic, B = LBK; C = Erg1)$ to test for admixture in C from two sources related to surrogates A and B, yielded a significantly negative result ($|Z| = 4.25$), as expected in the case of admixture [42]. Furthermore, qpAdm also concluded that Erg1 can be modelled as a mixture of Anatolia Neolithic (66%, se=8.1) and WHG (33%, se=8.1). In contrast, qpAdm modelled Erg2 as descending solely from sources related to Anatolian Neolithic farmers. MOSAIC also inferred admixture in Erg1, dated to 5.3 generations prior to it

sample date (i.e. approximately 5288 years ago), between WHG and Anatolia Neolithic sources. I caution that the admixture date may be unreliable due to only targeting a single individual, and given MOSAIC bootstraps over individuals, it was not possible to obtain confidence intervals around admixture date.

Estimated Hunter-gatherer related ancestry in Erg1 ranged from 18-38% among MOSAIC, qpAdm, with SOURCEFIND inferring 27.2% (se=1.41) when using six surrogates {Anatolian Neolithic, Loschbour, LaBrana, Bichon, and the two 'Iron Gates' samples}. MOSAIC indicated the cluster of Italian hunter-gatherers as the closest population to the true mixing source (Fig. 4.6). However, SOURECFIND indicated Iron Gates individuals from Serbia as the largest contributors of hunter-gatherer related ancestry.

4.3.4 Spatially and temporally close samples in Late Neolithic display highly distinct ancestries

This dataset included two individuals found in the same stratigraphical layer of Cherry-Tree cave; Kir23 and Kir24 were both dated to the Late Neolithic (approx 4700 BP). Despite their temporal and spatial closeness, they show highly different ancestry profiles (Fig. 4.8).

On both the plink PCA and fineSTRUCTURE clustering, Kir24 clusters with individuals from populations present around the Eurasian Steppe during the Bronze-Age, such as those from the Yamnaya and Afanasievo cultures. These are the populations thought to be in part responsible for the spread of Indo-European languages across Europe [105]. That the Yamnaya and Afanasievo samples were sampled in Russia suggests that Kir24 may have been a recent migrant from the Eurasian Steppe. This is supported by IBD analysis; of all the ancient samples in the dataset Kir24 shares the most IBD (31.12cM) with the Yamnaya type-specimen and the lowest TVD with 2 other members

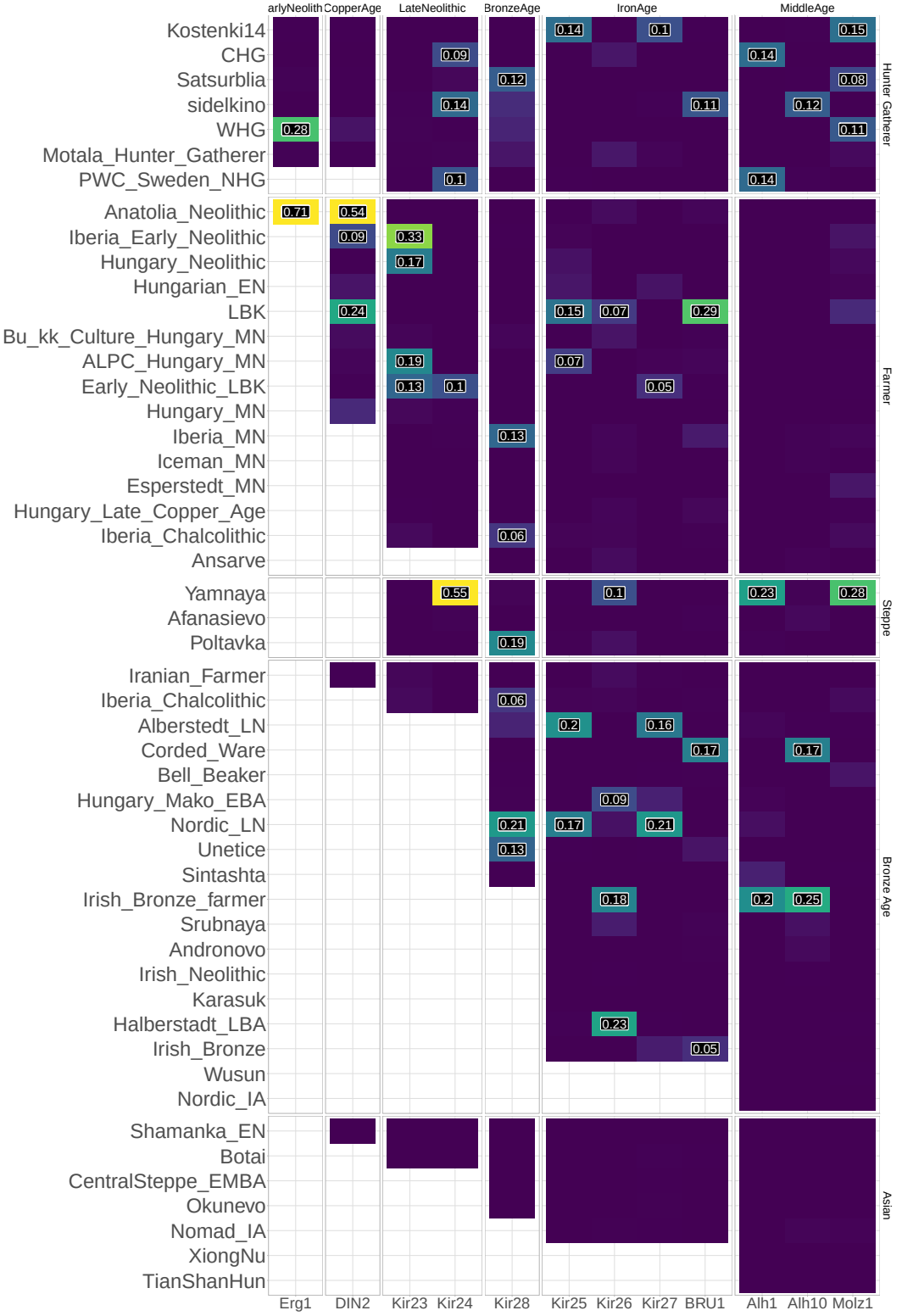

Figure 4.5: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture of only populations which are dated to being older or contemporaneous as the the target. Numbers within each cell correspond to the ancestry proportion estimate.

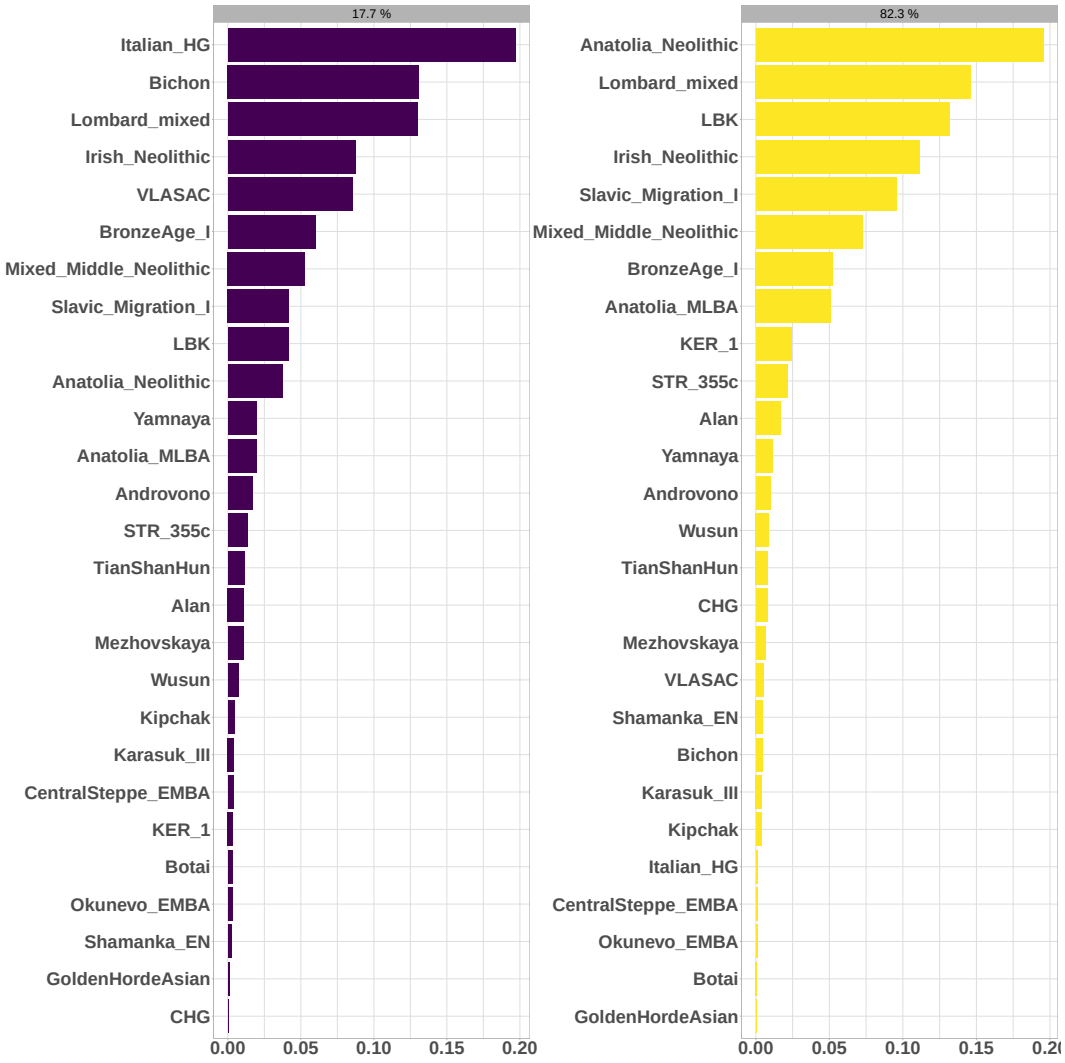

Figure 4.6: Copying matrix plot for sources in 2-way admixture event for Erg1. Each panel represents one of the 2 mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars within each panel represent the amount that mixing source copied from a particular population.

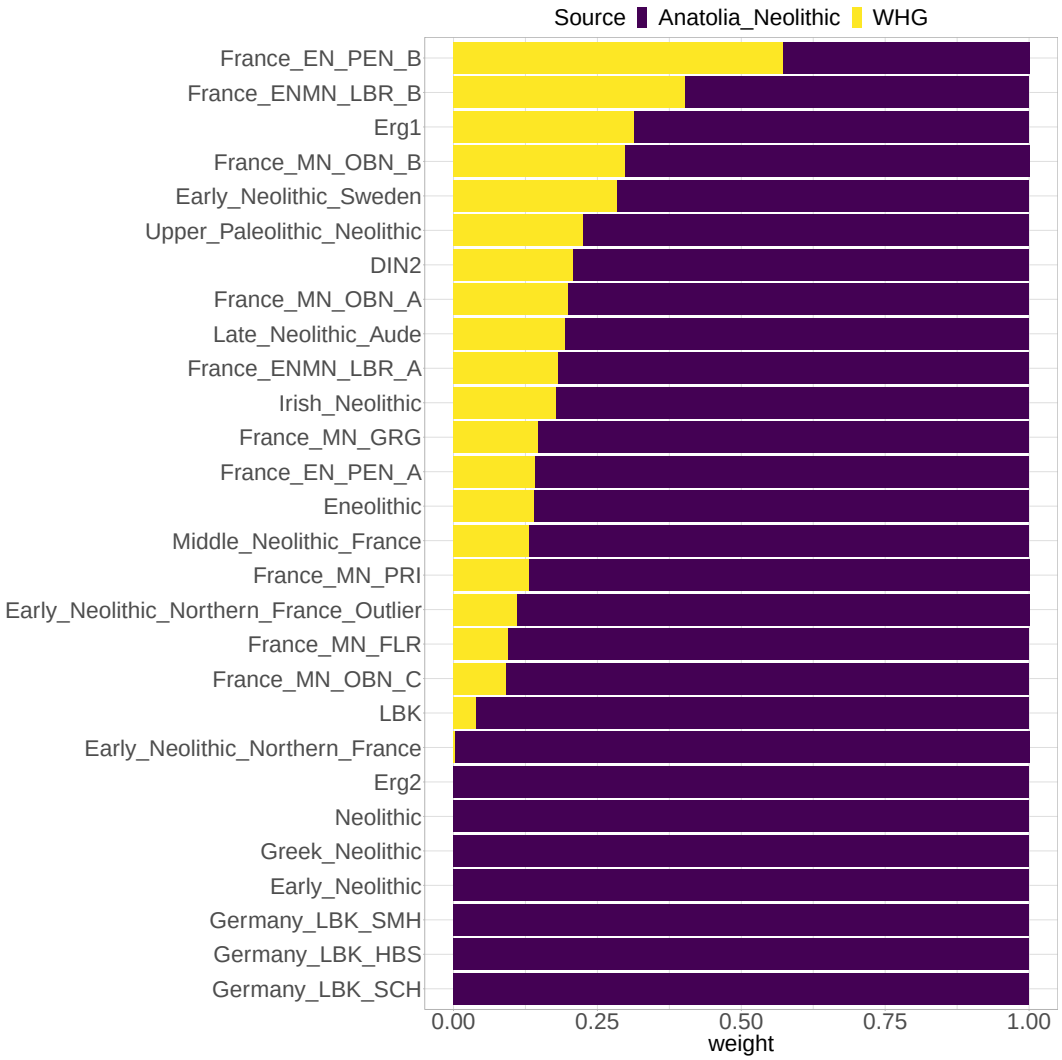

Figure 4.7: qpAdm ancestry proportion estimates for a selection of European Neolithic individuals. All individuals were modeled as a 2-way mixture between Anatolian Neolithic farmers and Western-Hunter Gatherers (WHG). Outgroups used are *Mota*, *Kostenki14*, *papuan*, *han*, *hannchina*, *mbutipygmy*, *sannamibia*, *yakut*

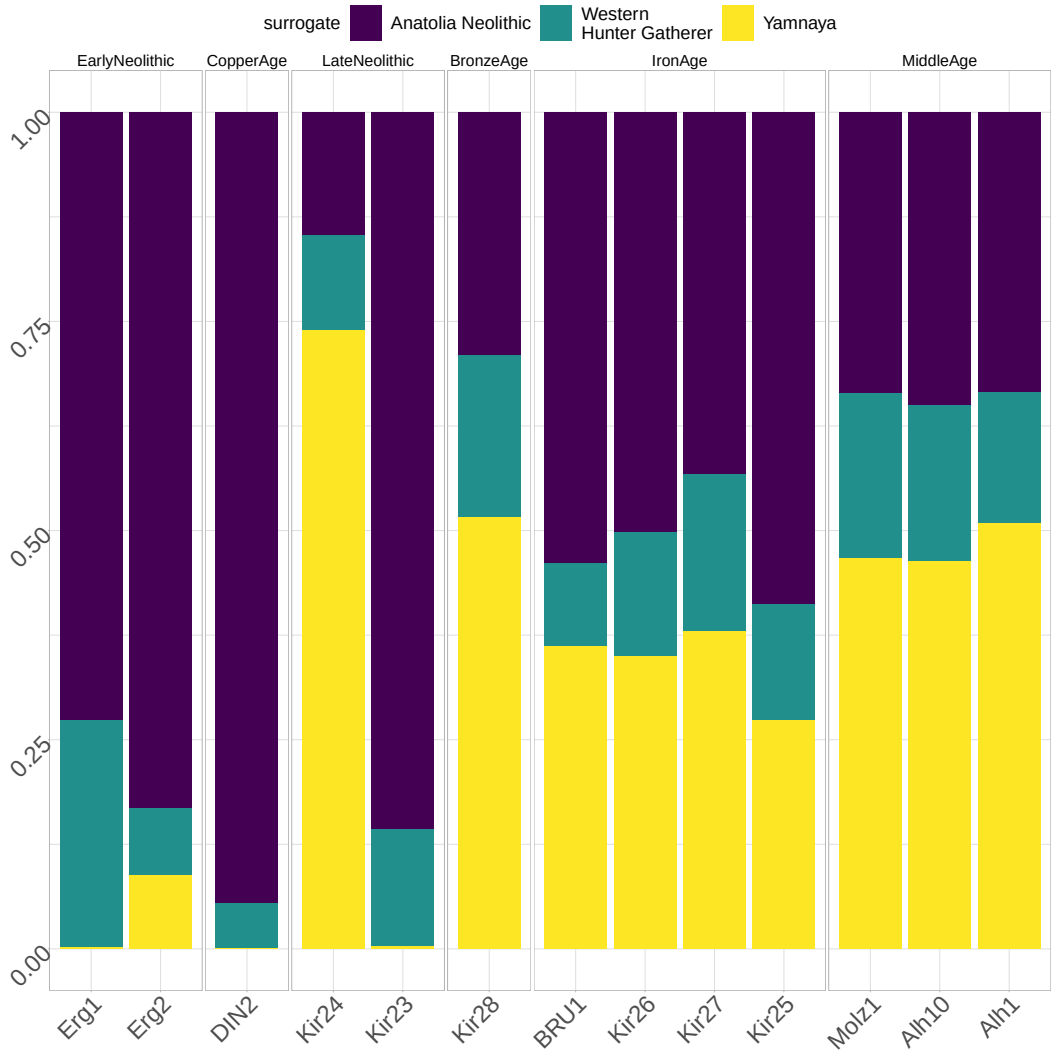

Figure 4.8: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture of only populations which are dated to being older or contemporaneous as the the target. Numbers within each cell correspond to the ancestry proportion estimate.

of the Yamnaya population. This timing (Kir24 is dated to approximately 4700 BP) corresponds to some of the earliest appearance of Yamnaya-like ancestry in central Europe [177]. Using qpAdm, Kir24 could be modelled as a mixture of Yamnaya (93%, se=12) and WHG (6%, se=8) without any Neolithic ancestry.

Kir24 was assigned to mtDNA haplogroup T1a1, which has been found in Yamnaya samples from the Middle Volga region and Bulgaria [178]; the same study found the frequency of T1a1 to be higher in the Yamnaya peoples than in any other ancient or modern population.

On the other hand, Kir23 is found in a fineSTRUCTURE cluster with Ballynahatty, from Neolithic Ireland (3343-3020 BC), and is positioned on both plink and ChromoPainter PCAs with other late Neolithic samples. It is found in adjacent fineSTRUCTURE groups to samples from Neolithic Spain and Ireland. As is the case with other Neolithic samples of this era, Kir23 has a component of Hunter-Gatherer ancestry; it is known that Middle Neolithic individuals are characterised by admixture with the existing Hunter-Gatherer populations. qpAdm modelling showed that Kir23 could be formed from a mixture of Neolithic Anatolia (96%, se=14) and Hunter Gatherer (6.25, se=0.91) without the need for additional Steppe ancestry.

To test whether the source of Neolithic ancestry in Kir23 was most similar to local populations, I performed f_4 tests in the form $f_4(W = Kir23, X = mbutipygy; Y = test, Z = Erg2)$, which tests whether Kir23 forms a clade with Erg2, a local farmer individual, or *test*, where *test* was one of several different farmer populations. Erg2 was chosen as the local group because it did not infer any potentially confounding Hunter Gather ancestry. Kir23 always formed a clade with Erg2, suggesting that the source of ancestry into Kir23 was local and that there was a degree of continuity within the region.

4.3.5 'Southern' ancestry to Cherry-Tree Cave during the Iron Age is Italian in origin

The plink PCA shows that the four Iron Age samples are shifted towards the cluster of Neolithic individuals and away from the samples typical of the European Bronze Age. The same pattern is also seen in the present-day PCA, where the Iron Age samples are shifted substantially towards Spain / Northern Italy relative to the preceding Bronze Age sample which is situated among Northern / Western European populations (Germany, Wales) (Fig. 4.4).

In fineSTRUCTURE, all four Iron Age individuals were grouped alongside several Lombard samples and a Roman soldier from 300AD. qpAdm modelling showed that the Iron Age samples can be well formed from a mixture of the preceding Bavarian Bronze age sample and those from either Renaissance Italy, Imperial Rome, Imperial Rome Late Antiquity or 'Roman Solider' from Veeramah et al (2018), with all other possible sources included with Bronze Age giving a poorly fitting models (Table 4.3). This suggests a model of admixture from populations best represented by those from post Iron-Age Italy. SOURCEFIND using all ancients as surrogates, inferred 26% of the IA samples' ancestry was most closely related to the "Rennaisance" Italy population from 1500CE, with no such inferred ancestry in the temporally flanking Bronze and Middle Age samples.

MOSAIC inferred the Iron Age samples could be formed of a mixture of $\approx$ 18% ancestry from a source closest to an Alamannic-Frankish sample (510 – 530 AD) and $\approx$82% ancestry from a source closest to Anatolian Neolithic / LBK samples, with admixture dated to 9.2 generations ago (bootstrapped 95% CI: 7.86-11.31). The estimated F_{st} between the two mixing sources was 0.016, approximately equivalent between present-day Germans and Palestinians [179].

Based on SOURCEFIND and qpAdm modelling with selected ancient and present-day East Asian samples, unlike Gamba et al (2014) [174], I found no

Target	Left	Weight	SE	Z
Bavaria Iron	Bavaria Bronze	1.458	0.732	1.992
Bavaria Iron	HallstattBylany	-0.458	0.732	-0.625
Bavaria Iron	Bavaria Bronze	0.956	0.426	2.245
Bavaria Iron	Renaissance	0.044	0.426	0.103
Bavaria Iron	Bavaria Bronze	0.986	0.202	4.871
Bavaria Iron	Imperial Rome Late Antiquity	0.014	0.202	0.070
Bavaria Iron	Bavaria Bronze	0.990	0.173	5.738
Bavaria Iron	Imperial Rome	0.010	0.173	0.056
Bavaria Iron	Bavaria Bronze	0.981	0.280	3.505
Bavaria Iron	Roman Solider	0.019	0.280	0.069

Table 4.3: Selected qpAdm results for estimating proportions of ancestry in the four Bavarian Iron Age samples. Each two rows is one test, with left populations as Bavaria Bronze and other. 'Weight' gives proportion of ancestry, 'SE' jackknifed standard error of Weight. Note negative Weight for model involving HallstattBylany, showing that the model does not fit well

evidence of East-Asian or East-Asian-like admixture (Fig. 4.5).

4.3.6 Present-day genomes unpick genetic differences between early Germanic and Slavic populations

Lastly, my dataset included three samples (1 newly sequenced) from the Middle Age period. The two genomes from Altheim, Germany, date to around 500AD and were found in a Roman context. The single individual from Molzbichl, Austria, dates to around 300 years later, and has been assigned to a 'Slavic' cultural context. It is currently unknown whether, in addition to cultural and linguistic differences, genetic differentiation exists between the 'Germanic' peoples represented by the two Altheim samples, and the 'Slavic' peoples represented by the Molzbichl sample.

The three Middle Age samples appear to share common ancestry based on the plink PCA and are located next to other spatially and temporally close samples from the Middle Ages. Similarly, they have almost indistinguishable

SOURCEFIND ancestry proportions (Fig 4.8).

f_4 in the form $f_4(mbutipygymy, Bavaria\_Iron; Bavaria\_Slav, Bavaria\_Germanic)$ returned a non-significant result, consistent with 'Germanic' and 'Slavic' populations splitting post Iron Age. However this non-significant result could be caused by low sample sizes in the Middle Age populations or a lack of power in allele-frequency based methods.

However, the two Germanic samples fall into a fineSTRUCTURE cluster with a set of contemporaneous samples from Northern Europe, including 10-11$^{\text{th}}$ century Vikings from Estonia, Sweden and Iceland, whereas Molz1 clusters with other individuals known to be from Early Slavic populations. Interestingly, the Slavic cluster also containing a sample DA29, also know as 'GoldenHordeEuro'. This sample is from Karasuyr, Kazakhstan, and has was dated to 1200-1400 CE. The Golden Horde was a Mongol khanate established in the 13th Century CE. Given this sample shows clear evidence of European ancestry and clusters alongside individuals from Early Middle Age Europe, it has been proposed that this individual was captured in Europe during the Mongol raids of the 13th Century, when they assaulted the Kievan Rus' federation [180]. That 'GoldenHordeEuro' clusters with Molz1 suggests the location of capture in Europe may have been from Austria where Molz1 was found.

On a haplotype-based PCA with modern samples, Molz1 clusters with present-day Slavic speaking populations such as Poland, Ukraine and Belarus, while the two Germanic samples cluster with present-day individuals from Germanic-speaking countries in Western Europe, such as Scotland, Germany and Wales (Fig. 4.4). Plotting differential haplotype sharing between the Slavic and Germanic sample makes this pattern clear (Fig 4.9). There is a clear division down the centre of Europe, dividing it into East and West that shows the structure in present-day Europeans has existed since at least the Early Middle Ages.

In SOURCEFIND, the two samples from Altheim derived a large proportion of their ancestry to modern day Germans (81.8%, se=12.8), whereas the Molzbichl sample derived a large proportion of its ancestry from modern day Polish (77.85%, se=20.3) and Croatians (11.7%, se=9.1).

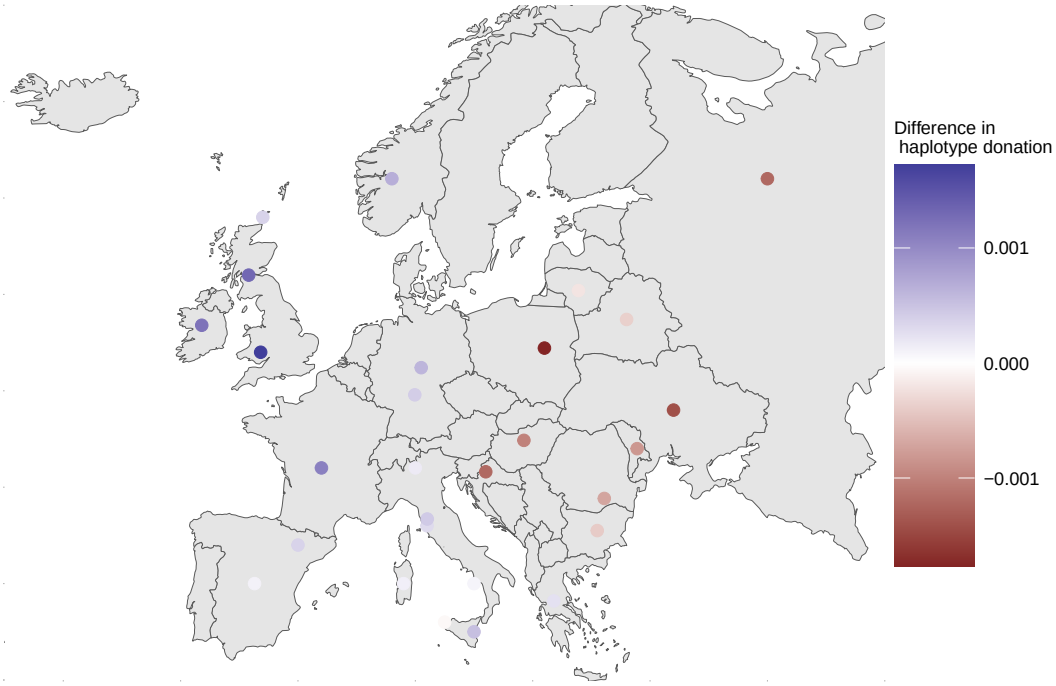

Figure 4.9: Differential haplotype-donation between Germanic and Slavic samples. Each coloured point is one present-day population. Points are coloured based on whether they donate relatively more to Germanic (blue) or Slavic (red) ancient samples.

4.3.7 Summary of Results and Discussion

Drawing back to the questions asked at the beginning.

Whilst the two samples from the Early and Middle Neolithic, Erg1 and DIN2,

showed some signs of being from at least closely related source populations, they also displayed variation suggestive of different population histories. Consistent with the hypothesis that DIN2 may have migrated along the Danubian route, it shares the lowest TVD and is found in a fineSTRUCTURE cluster with other samples from the Hungarian Plane. Most importantly, Erg1 and DIN2 two samples also showed differences in the degree of Hunter-Gatherer ancestry; whilst DIN2 showed no evidence of admixture, Erg1 likely had a recent Hunter-Gatherer ancestor.

I found evidence of population discontinuity in Cherry-Tree Cave from the Late Neolithic through to Iron Age. I identified a incoming signal of 'southern' ancestry during the Iron Age, which was not present in the single sample from the preceding Bronze Age. The most plausible source of this ancestry is from Italy, with the best source in the dataset being the cluster of Renaissance samples from Antonio et al (2019) [59], date to between 282 - 354 AD. This, combined with evidence the Iron Age samples cluster with present-day individuals from north Italy and historical evidence of Lombard migrations to Southern Germany [181], suggests they may be the admixing source. Whilst collaborators proposed that the source may be related to the local Hallstatt culture, qpAdm modelling rejected this scenario (Table 4.3). Wherever the source originated from, this admixture event provides strong evidence against continuity in Cherry-Tree Cave.

Lastly, I used present-day genomes of individuals from across Europe to show that there are clear genetic differences between the Middle Age Germanic and Slavic samples, with the Germanic samples showing a strong affinity to western European countries and the Slavic samples showing a strong affinity to eastern European samples (Fig. 4.9). However, in the context of ancient samples, all three Middle Age samples clustered with local samples from the Bronze Age rather than the Iron Age (Fig. 4.3).

This dataset revealed that temporally and spatially close samples may

have very distinct genetic ancestry profiles, with Early Bronze Age samples Kir24 and Kir23 showing high levels of Steppe-related and Neolithic ancestry respectively. In particular, Kir24 seemed to be very recently related to the Yamnaya type-specimen sample, sharing 31cM of IBD with it. The arrival of Yamnaya-like ancestry from this early (2762BC) represents one of the earliest known appearances in the literature.

Future studies in this region should focus on obtaining a higher density of samples, in particular from the Bronze and Iron Ages; the low number of samples from these time periods mean any results should be interpreted with caution. More samples would show whether the introduction of 'southern' like ancestry in the Iron Age was a widespread phenomena, or restricted to a smaller geographic region in Southern Germany. Similarly, a wider sampling of Iron Age groups from Germany, Italy and Switzerland may allow for a more accurate identification of this source.

Whilst the utility of using present-day genomes was outlined through the comparison of the Slavic and Germanic samples, the analysis would have been significantly improved with higher resolution data from Germany. The data I have, described in Appendix section A.4, only had country-level details. Data which had labels from different sub-regions in Germany, similar to the POBI dataset, would have allowed for a finer-scale investigation into the current east-west genetic divide in present-day Germany.

Chapter 5

The genomics of the Slavic migration period, Early Middle Ages and their links to the present day

5.1 Introduction

The Slavic peoples originated as a diverse network of tribal societies who lived in Central and Eastern Europe from the first Millennia AD [182] and whose origin, although disputed, is thought to be Polesia (a marshy forested area straddling Poland, Belarus, Russia and Ukraine) [183]. Although various Roman and Greek sources refer to Slavs as *Veneti* and *Spori* as early as the 1st and 2nd centuries AD, the term 'Slavs' was first used in writing by Roman bureaucrat Jordanes at the beginning of the 6th century after their attack on the Byzantine empire [184]. This era, known by historians as The Migration Period, was a period of European history, roughly between 375-568 AD after the fall of the Roman Empire [185], characterised by the large-scale movement of various peoples. The Migration Period began with the Huns moving into

Eastern Europe at the end of the 4th Century, occupying an area including present-day Hungary and Romania. During the 5th century, various Germanic groups invaded and established a homeland across parts of the Western Roman Empire. This was followed by the expansion of Slavic populations into regions of low population density in the sixth century.

Across the next two centuries, these peoples had settled across large parts of Europe. In particular, the Early Slavs had expanded southwards into the Balkans and Alps [182, 186–188]. It has been proposed that these migrations were key to forming the foundations of present-day Slavic (speaking) nations [182].

By the beginning of the 12th century, Slavs constituted a large part of a number of many medieval Christian states across Europe. As from this time period, Slavs could be broadly split up in three groups: the eastern Slavs as part of the Kievan Rus', southern Slavs in the Bulgarian Empire, the Principality of Serbia, Kingdom of Croatia and the Banate of Bosnia, and western Slavs in the Principality of Nitra, Great Moravia, the duchy of Bohemia and the Kingdom and Poland. In addition, Slavic settlement also occurred in the Eastern Alps; Slovenia, large parts of present-day Austria and Friul.

Today 315 million people speak Slavic languages and linguistic evidence suggests that they can be broadly split into these three broad groups; western Slavs (Poles, Czechs and Slovaks), eastern Slavs (Ukrainians, Belarusians and Russians) and southern Slavs (Croatians, Bulgarians, Slovenians, Bosnians, Macedonians, Montenegrins and Serbians) [189].

The history of the Slavic peoples can be artificially be split into three periods: Migration Period (∼375AD - ∼568AD), Early Middle Ages/High Middle Ages (∼600AD - ∼1250AD) and present-day. Several previous studies have investigated the genetics of the transitions between these periods. Juras et al (2014) used uni-parental mtDNA markers from ancient DNA samples

Figure 5.1: Slavic tribes from the 7th to 9th centuries AD in Europe. Source: (`https://commons.wikimedia.org/wiki/File:Slavic_tribes_in_the_7th_to_9th_century.jpg`)

from Poland to show continuity between both Roman Iron Age period (200 BC – 500 AD) and Medieval Age (1000–1400AD) with present-day Poles, Czechs and Slovaks [190]. However, whilst informative about sex-biased migrations, uniparental markers carry only a fraction of the information that autosomal markers do, and therefore may provide misleading or incomplete information about the relationship between samples [191, 192], especially when admixture is prevalent (although see [193]). For example, it is know that mtDNA and nuclear DNA may have different evolutionary histories and thus display discordant phylogenetic trees [194].

Kushniarevich et al (2015) [195] combined results from mtDNA, non-recombining Y and autosomal DNA to investigate the population structure of a wide range of present-day Balto-Slavic populations. They proposed that incoming Slavic speakers admixed with peoples in the regions they occupied during the Migration Period.

More recently, Macháček et al (2021) [196] analysed a cattle rib from Lány, Czechia, dated to approximately 600AD, that is inscribed with Germanic runes. The bone was found in a location where Slavs were thought to have arrived at the end of the Migration Period, after the Germanic tribes had disappeared and the use of a Slavic language is historically confirmed as of the 9th century. However, whether there was early genetic contact as well is yet to be determined.

Several studies into present-day Slavic populations have detected signatures of admixture from East-Asia [20, 166, 197–199]. Whether or not these signals can be observed in ancient individuals is yet to be seen and could further refine the admixture date. For example, different admixture dates in different Slavic populations may reveal structure among present-day Slavs.

Finally, several studies have used haplotype-based methods to explore the structure of present-day Slavic populations. Ralph and Coop [200] compared regions of IBD matching across different European populations. They found a relatively high degree of IBD sharing among pairs of individuals from Eastern Europe, suggestive of expansion from a smaller, common source population. This expansion was tentatively estimated to between 0-1000AD. Consistent with estimates of a small population size, Hellenthal et al (2014) [20] inferred an excess of among Eastern European individuals and an admixture event, albeit with a more constrained admixture date of 440 - 1080 CE. However, this could also be interpreted in terms of a small effective population size [201, 202]. Salter-Townshend and Myers (2019) also identified admixture in the Chuvash people between east Europeans and east Asians approximately 1224 CE [166].

In this chapter, I will analyse 17 new medium to high coverage whole ancient genomes from Czech Republic, spanning from the Migration Period to Early Middle Ages (384-950 AD). These are, to my knowledge, the first high-coverage whole ancient-genomes from this period. I will merge the newly sequenced samples with reference data from other ancient individuals and a large reference set of relevant present-day European individuals in order to understand their ancestry in the context of both present-day and ancient samples. In particular, I am interested in considering the following questions:

1. Do the labels "Migration Period" and "Early Middle Ages" make sense from a genetic standpoint (i.e. do samples from either period cluster with another to the exclusion of the other). Is there evidence of continuity between these two eras?

2. Is there evidence of continuity between Early Middle Ages and present-day?

3. How are present-day Slavic speakers structured, and do the different ancient Slavic samples have different affinities to different present-day Slavic language groups?

5.2 Methods

5.2.1 Description of samples

Whole-genome sequence data were generated from 17 ancient individuals (Table 5.1). Five samples from Líbivá date to the Migration Period (348 AD - 504 AD), while the other 12 samples from Pohansko date to the later Early Middle Ages (724 AD - 995 AD).

The Migration Period and Early Middle Age samples were categorised based upon the style of pottery found in the burial grounds (Z. Hofmanová, personal

Code	Site	Date (AD)	Period	Coverage
LIB11	Břeclav – Líbivá	741.5	Early Middle Ages	5.3
LIB12	Břeclav – Líbivá	475.5	Migration period	6.8
LIB2	Břeclav – Líbivá	495.0	Migration period	6.4
LIB3	Břeclav – Líbivá	509.0	Migration period	5.3
LIB4	Břeclav – Líbivá	472.5	Migration period	6.5
LIB5	Břeclav – Líbivá	348.0	Migration period	7.3
LIB7	Břeclav – Líbivá	830.5	Early Middle Ages	5.6
POH11	Pohansko – Lesní školka	783.0	Early Middle Ages	5.0
POH13	Pohansko – Lesní školka	879.5	Early Middle Ages	6.0
POH27	Pohansko – Jizní Předhradí	783.0	Early Middle Ages	5.9
POH28	Pohansko – Jizní Předhradí	822.5	Early Middle Ages	5.6
POH36	Pohansko – Jizní Předhradí	880.5	Early Middle Ages	5.5
POH39	Pohansko – Jizní Předhradí	866.4	Early Middle Ages	5.3
POH3	Pohansko – Lesní hrúd	956.5	Early Middle Ages	5.4
POH40	Pohansko – Lesní školka	950.5	Early Middle Ages	5.5
POH41	Pohansko – Lesní školka	875.5	Early Middle Ages	5.2
POH44	Pohansko – Pohřebiště U Kostela	NA	Early Middle Ages	5.3

Table 5.1: Information on newly sequenced ancient samples. Date (AD) estimated from radiocarbon dating. 'Migration' corresponds to Migration Period and 'EMA' corresponds to Early Middle Ages. Coverage calculated as the mean depth across all 77,213,942 genome-wide SNPs where genotypes were called at.

communication).

5.2.2 Ancient DNA processing

I merged the 17 newly sequenced individuals with the ancient literature samples given in section A.1, resulting in a total of 959 ancient individuals with genotype likelihoods at 77,213,942 genome-wide autosomal SNPs.

I followed the GLIMPSE [91] imputation and phasing pipeline (`https://odelaneau.github.io/GLIMPSE/tutorial_b38.html`) to generate genotype likelihoods and phased genotypes for each individual. For the reference panel, I used the 30x 1000 genomes dataset [102], described in Appendix section A.2.

5.2.3 Present-day DNA processing

I merged the newly sequenced and published ancient samples with the MS-POBI-HellBus dataset, described in detail in Appendix section A.4, chosen because it contains a high number of relevant samples from central and eastern Europe. I removed samples from Australia, New Zealand and USA.

The present-day and ancient samples were phased separately, as GLIMPSE is designed for sequence-level density of data, and the present-day samples were genotyped on a low-density genotyping array. Therefore, I phased the present-day samples using shapeit4 [25] using default parameters and the supplied genetic map. I note that phasing the datasets separately may reduce power to compare ancient and present-day samples.

The present-day and ancient samples described in section 5.2.2 were merged and converted to ChromoPainter format.

5.2.4 plink PCA

I performed a PCA on the pre-imputation genotypes using plink2 [164]. I chose to use plink2 because recent studies have shown it is substantially better at dealing with samples containing variable amounts of missing data than other methods such as smartPCA [56].

I retained only the 500,000 markers with the lowest amount of missingness. I then LD-pruned the resulting SNPs using the settings `-maf 0.01` and `-indep-pairwise 50 5 0.2` and performed PCA using plink2 under default settings.

5.2.5 Allele-frequency based tests

I used Admixtools [42], implemented in Admixr R library [203] to perform different F-statistics.

5.2.6 ChromoPainter and fineSTRUCTURE analysis

The merged data described in sections 5.2.2 and 5.2.3 contained a total of 959 ancient and 14,795 present-day samples genotyped at 477,417 autosomal bi-allelic SNPs.

I first selected all ancient samples above 2x coverage and performed an 'all-v-all' painting where each haplotype was compared to all other haplotypes in turn, hereafter referred to as 'ancient' painting. I chose to remove samples with <2x coverage because all new samples analysed here had at least 5x coverage, and my previous work indicated little difference in ChromoPainter results among samples >2x coverage (Chapter 2 section 2.6.4).

I also performed an 'all-v-all' painting of the 17 newly sequenced samples and the present-day populations given in table 5.2, hereafter referred to as 'present-day painting'.

The fineSTRUCTURE [19] clustering and tree building algorithm was applied to the ChromoPainter output for both the 'present-day' and 'ancient' paintings, in each case using 2,000,000 MCMC iterations after 1,000,000 iterations of "burn-in". I then ran the tree-building mode (-m T) with 100,000 additional hill-climbing steps before tree building,

Tree figures, coancestry matrix figures and principle component plots were generated using the fineSTRUCTURE R library (`https://people.maths.bris.ac.uk/~madjl/finestructure/FinestructureRcode.zip`).

5.2.7 SOURCEFIND ancestry proportion analysis

I used SOURCEFIND [21] to infer the proportions of ancestry by which each target (e.g. ancient) individual is most related to a set of surrogate ancient populations. Each of the 47 clusters of ancient samples inferred by fineSTRUCTURE was analysed in turn, using the other 46 clusters to act as

Population	Number of Individuals
HB:tsi	98
HB:spanish	34
HB:german	30
HB:french	28
HB:greek	20
HB:croatian	19
HB:hungarian	19
HB:norwegian	18
HB:southitalian	18
HB:polish	17
HB:romanian	16
HB:mordovian	15
HB:cypriot	12
HB:northitalian	12
HB:lithuanian	10
HB:siciliane	10
HB:westsicilian	10
HB:tuscan	8
HB:irish	7
HB:scottish	6
HB:germanyaustria	4
HB:welsh	4

Table 5.2: Name of population and number of samples used in the present-day ChromoPainter analysis

surrogates.

Each cluster was run across three independent MCMC runs, using 50,000 burn-in iterations, 500,000 main iterations, and thinning every 5 iterations All three MCMC runs were then combined to form an MCMC list using the coda R libary [106] and `mcmc` function to jointly estimate ancestry proportions and empirical credible intervals for each target population.

5.2.8 MOSAIC admixture analysis

I inferred admixture events, dates and proportions using MOSAIC [166], performing two different analyses that mimicked the two ChromoPainter "ancient"

and "present-day" paintings described above. In particular I tested each of the 5 fineSTRUCTURE clusters containing the 17 newly sequenced individuals using as surrogates: (i) 46 other fineSTRUCTURE clusters containing ancient individuals (i.e. from the "ancient" painting results) or (ii) only the 5 other Slavic ancient populations plus 49 present-day populations in Table 5.3. I assumed each target population could be formed as a mixture of both two and three admixing sources, with all other parameters left as default.

I then performed a 'present-day surrogates' analysis using a select group of present-day populations 5.3 and all ancient Slavic samples. I analysed each population in turn using all other populations as surrogates.

MOSAIC was run using default settings and the following sets of populations as targets and the following sets as surrogates. I formed each target as a mixture of both 2 and 3 mixing sources, with all other parameters left as default. Upper and lower quantiles for admixture dates were estimated using a bootstrap procedure. Other than changing the number of mixing sources, all other parameters were left as default.

5.3 Results

5.3.1 Mixed ancestry of Migration Period Slavs

The Migration Period samples consisted of five individuals with radiocarbon dates corresponding to the Migration Period (348 - 509AD). Both the unlinked (Fig. 5.2) and linked PCAs (Fig. 5.3) show that the Migration Period samples are heterogeneous are not likely to originate from the same source population. One sample, LIB2 (495AD) is located in the centre of a large cluster of contemporaneous individuals from Iron Age central and northern Europe. fineSTRUCTURE grouped LIB2 with Viking era individuals from Sweden, Denmark, Iceland, Estonia and Norway from 300-1100AD. When painted using

Population	Number of Individuals
HB:han	34
HB:bulgarian	31
HB:japanese	28
HB:sardinian	28
HB:russian	25
HB:yakut	25
HB:greek	20
HB:ukrainian	20
HB:croatian	19
HB:hungarian	19
HB:mongolian	19
HB:southitalian	18
HB:chuvash	17
HB:polish	17
HB:romanian	16
HB:buryat	15
HB:mordovian	15
HB:altai	13
HB:tuva	13
HB:evenk	12
HB:northitalian	12
HB:cambodian	10
HB:dai	10
HB:hannchina	10
HB:lithuanian	10
HB:miao	10
HB:nganassan	10
HB:selkup	10
HB:siciliane	10
HB:tu	10
HB:tujia	10
HB:uygur	10
HB:westsicilian	10
HB:yi	10
HB:belorussian	9
HB:daur	9
HB:oroqen	9
HB:xibo	9
HB:hezhen	8
HB:naxi	8
HB:tuscan	8
HB:dolgan	7
HB:chukchi	5
HB:koryake	5
HB:yukagir	4
HB:myanmar	3
HB:burya	2
HB:ket	2
HB:malayan	1

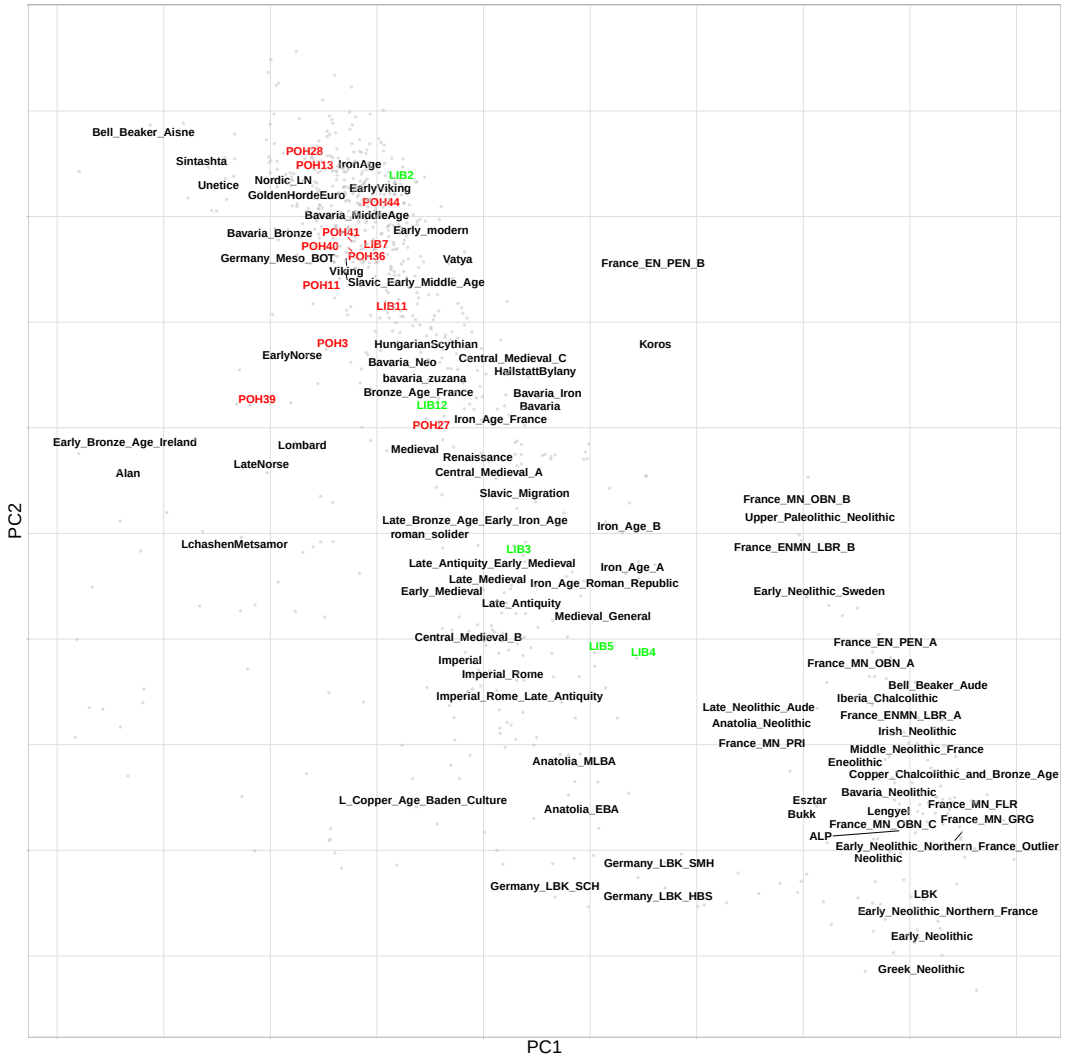

Figure 5.2: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using the plink2. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations. The position of each reference label is the mean PC coordinates of all individuals within that population

a set of present-day reference samples, LIB2 matches the most haplotypes and clusters with Norwegians (Fig. 5.7). Put together, this data suggests LIB2 may be a recent migrant from Viking regions.

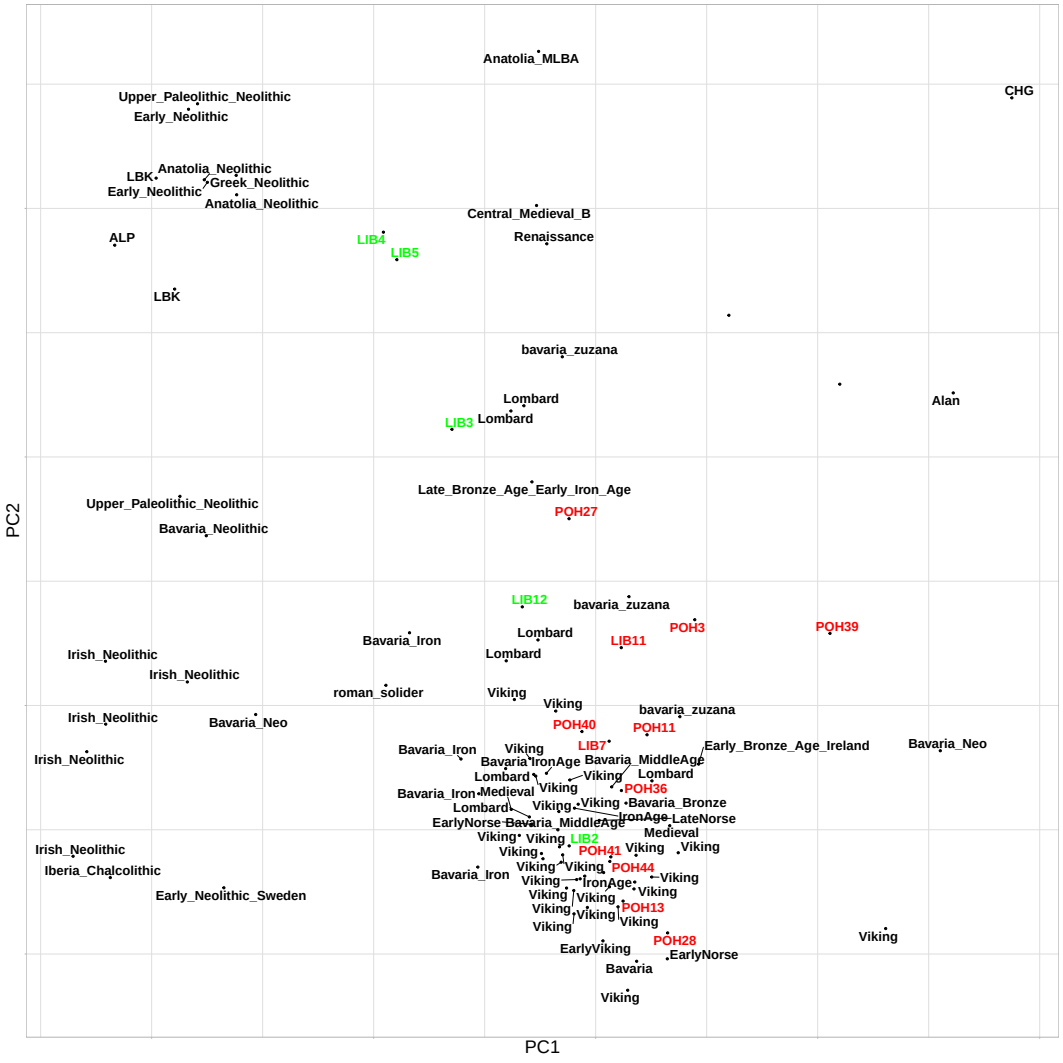

Figure 5.3: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using fineSTRUCTURE. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations.

On the other hand, LIB4 and LIB5 cluster together with Early Iron Age and Renaissance samples from Italy, and generally show an increased affinity Neolithic / Southern European populations relative to the other Migration Period samples (Fig 5.2-5.3).

LIB3 clusters with Lombard samples from Northern Italy (Fig 5.2) in the

'ancient' painting, and with Tuscans in the 'present-day' painting. Historical evidence cites alliances between Slavs and Lombards in the 5th century [204]. Finally, LIB12 displays ancestry which is more typical of the preceding Central European Bronze Age, suggesting it may represent a 'leftover' from a local Bronze Age population which was unaffected by the Antiquity / Iron Age migrations to the region.

5.3.2 Early Middle Age Slavs represent a relatively homogeneous group typical of European Middle Ages

In comparison to the five Migration Period ancient Slavs, the 12 Early Middle Age Slavs (741-956 AD) are more homogeneous. All 12 samples cluster in the same fineSTRUCTURE group (named Slavic Early Middle Age II), alongside Viking/Medieval samples from Ukraine, Poland and Sweden. SOURCEFIND showed that the Slavic Early Middle Age II cluster derives roughly equal parts of ancestry from the clusters Viking 10C Scan I, BronzeAge I and Lombard mixed cluster. Interestingly, these three ancestry sources are similar to those identified by SOURCEFIND analyses in the Migration Period samples (Fig D.7). I tentatively therefore suggest that the Early Middle Age Slavs were formed from the mixture of 'northern' (best represented by Viking) and 'southern' ancestries (best represented by Lombards) onto a substrate of local Bronze Age populations.

MOSAIC admixture analysis on the Early Middle Age samples using ancient surrogates proved inconclusive. However, using present-day individuals as surrogates inferred a three-way admixture event involving sources closest to present-day day north-central Slavs (76.6%), southern-eastern Slavs (21.9%) and East Asians best represented by Mongolians (1.5%) (Fig. 5.4). This admixture event was estimated to have occurred 9.4 (2.5% 5.7gens - 97.5% 17.9gens) generations before the samples (Fig. 5.5), i.e. 476 - 732 AD.

This admixture event is consistent with a signal inferred in both present-day day Eastern European individuals [20, 166]. In previous studies, this admixture event was dated to approximately 1200CE (MOSAIC) and 440-1080 (GLOBETROTTER).

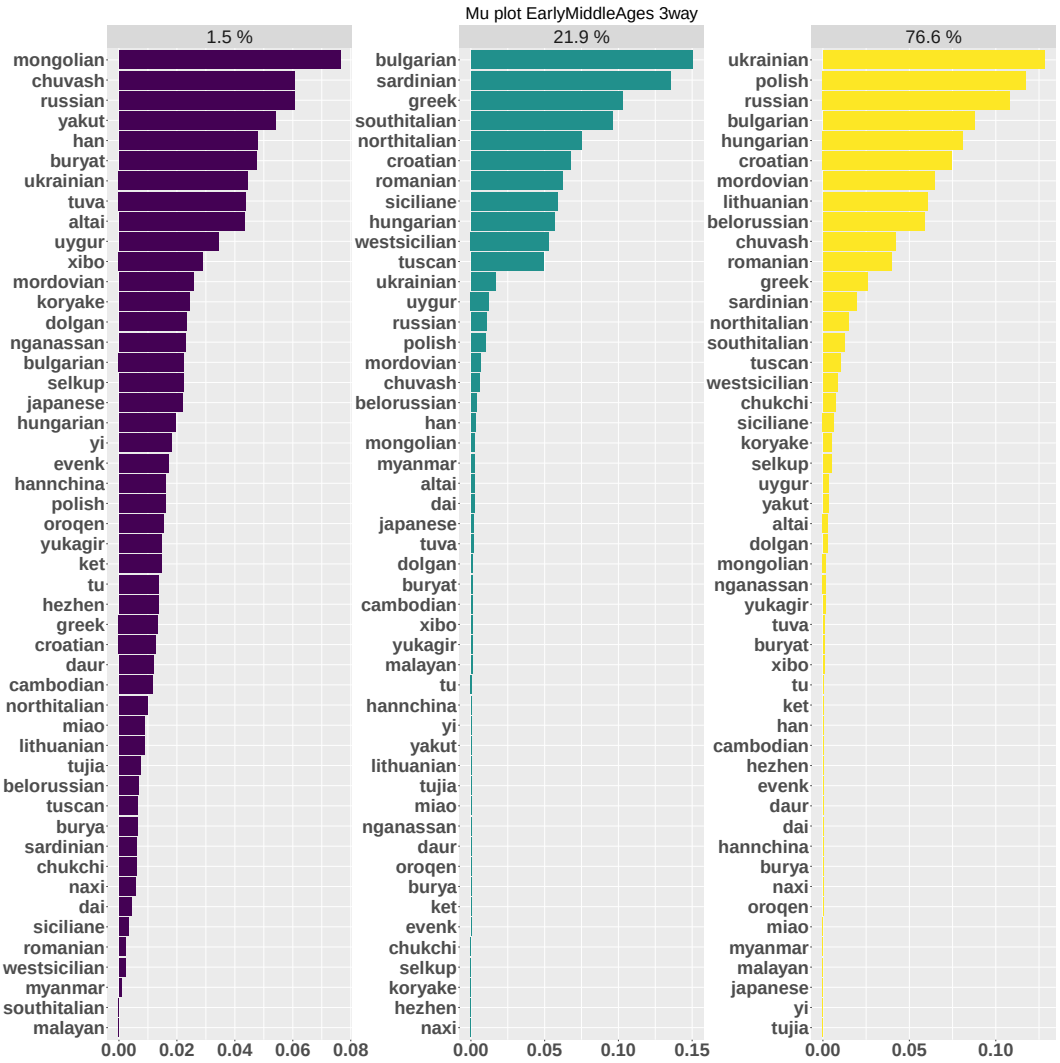

Figure 5.4: Copying matrix plot for sources in 3-way admixture event for Early Middle Age ancient Slavic samples. Each panel represents one of the 3 putative mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars within each panel represent the reflect how to best represent the relative haplotype composition of that source using the surrogate populations.

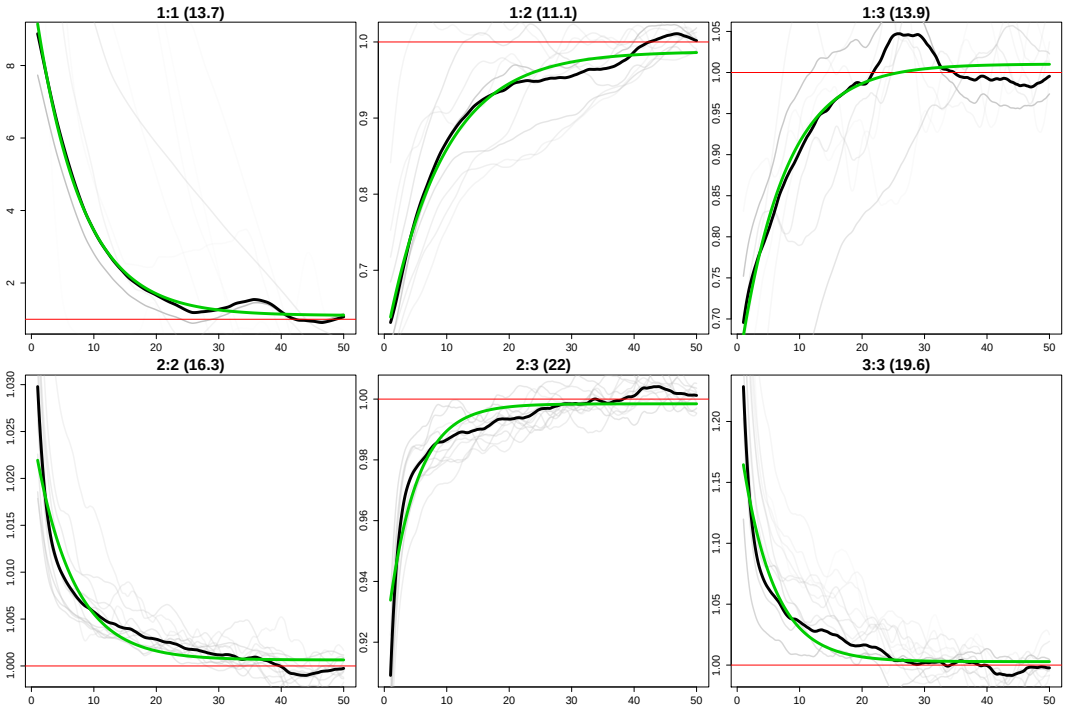

Figure 5.5: Inferred Coancestry Curves obtained from modelling Early Middle Age samples as a 3-way mixture of present-day individuals. Black lines are empirical coancestry curves across all target individuals, light grey are per individual, green is the fitted single-event coancestry curve. x-axis gives genetic distance and y-axis the probability of switching segments from source a to source b. Sources are those given in Fig. 5.4.

5.3.3 Assessing continuity between Early Middle Age and Migration Period samples

To formally establish whether the Early Middle Age and Migration Period samples cluster within their respective populations to the exclusion of the other, following Leslie et al 2015 [31], I performed a TVD permutation test. Full details of TVD justification and calculation are outlined in Appendix section B.3.

Using the ancients chunklengths matrix, I grouped the samples into Migration Period and Early Middle Age and calculated the average copyvectors C_{mp} and C_{ema} across samples within each groups. Here $C_{mp} = \{C_{mp}(1), ..., C_{mp}D\}$, where $C_{mp}(d)$ is the average amount a Migration Period individual copies from (i.e. is painted by) individuals from donor population d. Then, I calculated the

empirical TVD between the two groups as $TVD_{mp,ema} = \sum_d |C_{mp}(d) - C_{ema}(d)|$. For 10,000 iterations, I then randomly permuted the population labels among the samples and then calculated the analogous TVD, $TVD^{rand}_{mp,ema}$, between these two randomised "populations". I then calculated, as a p-value for the null model assuming individuals are exchangeable between the two populations, the number of randomly permuted iterations where $TVD^{rand}_{mp,ema} \geq TVD_{mp,ema}$. This test supported clustering the samples into their respective groups ($p = 0.0013$).

To determine the extent of continuity between the Migration Period and Early Middle Ages, I modelled each Early Middle Ages sample as a mixture of other ancients, including individuals from the preceding Migration Period, using SOURCEFIND. The proportion of ancestry derived from the Migration Period was low (mean 3.4% , range 0.4% - 12.5%), suggesting that there was a relatively large scale population replacement between the two different time periods.

5.3.4 Legacy of Slavic migrations in present-day individuals

Principle component analysis (PCA) of the present-day painting indicates genetic similarity between ancient Slavic samples from the Early Middle Ages and present-day day Slavic speaking populations (Fig. 5.6). The Early Middle Age samples primarily cluster with present-day Polish and Belorussian individuals, but appear to fall on a cline of genetic similarity between Russians and southern Europeans.

As with the ancients PCA, Migration Era Slavs are spread across the present-day PCA. LIB3, LIB4, and LIB5 cluster with present-day Italians, consistent with deriving a substantial ancestry component from southern European sources. LIB4 and LIB5 appear to be positioned closer to southern Italians and Greeks, whereas LIB3 is closer to northern Italian and Tuscan populations. LIB2 shows

a strong affinity to present-day Norwegians, suggesting it may be a recent migrant from Viking regions.

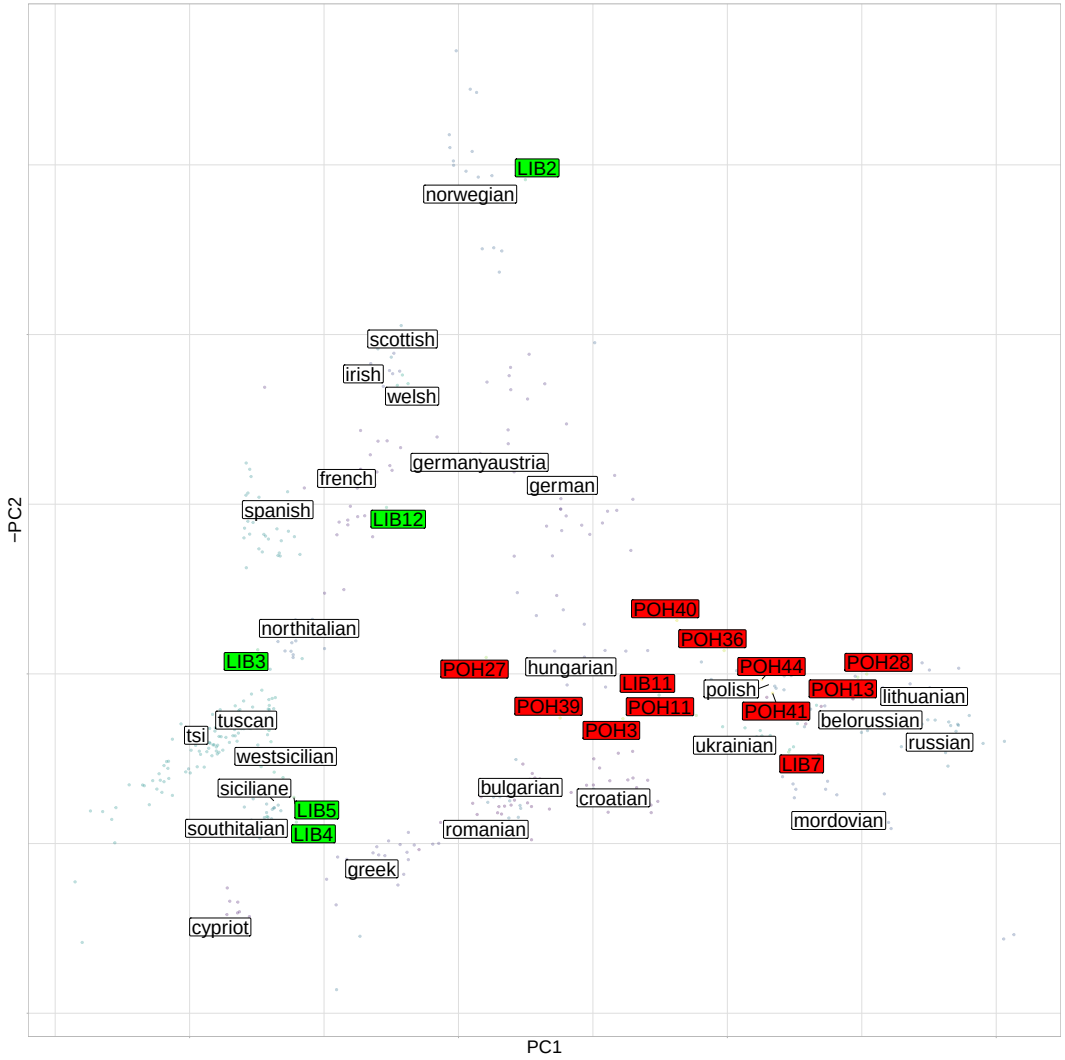

Figure 5.6: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations. The position of each reference label is the mean PC coordinates of all individuals within that population. Transparent coloured points correspond to present-day individuals.

The same pattern can be observed on the raw copyvector output matrix from the present-day painting (Fig. 5.7). In particular, Migration Era samples show little excess affinity to present-day day Slavic populations and more affinity to present-day Greek individuals.

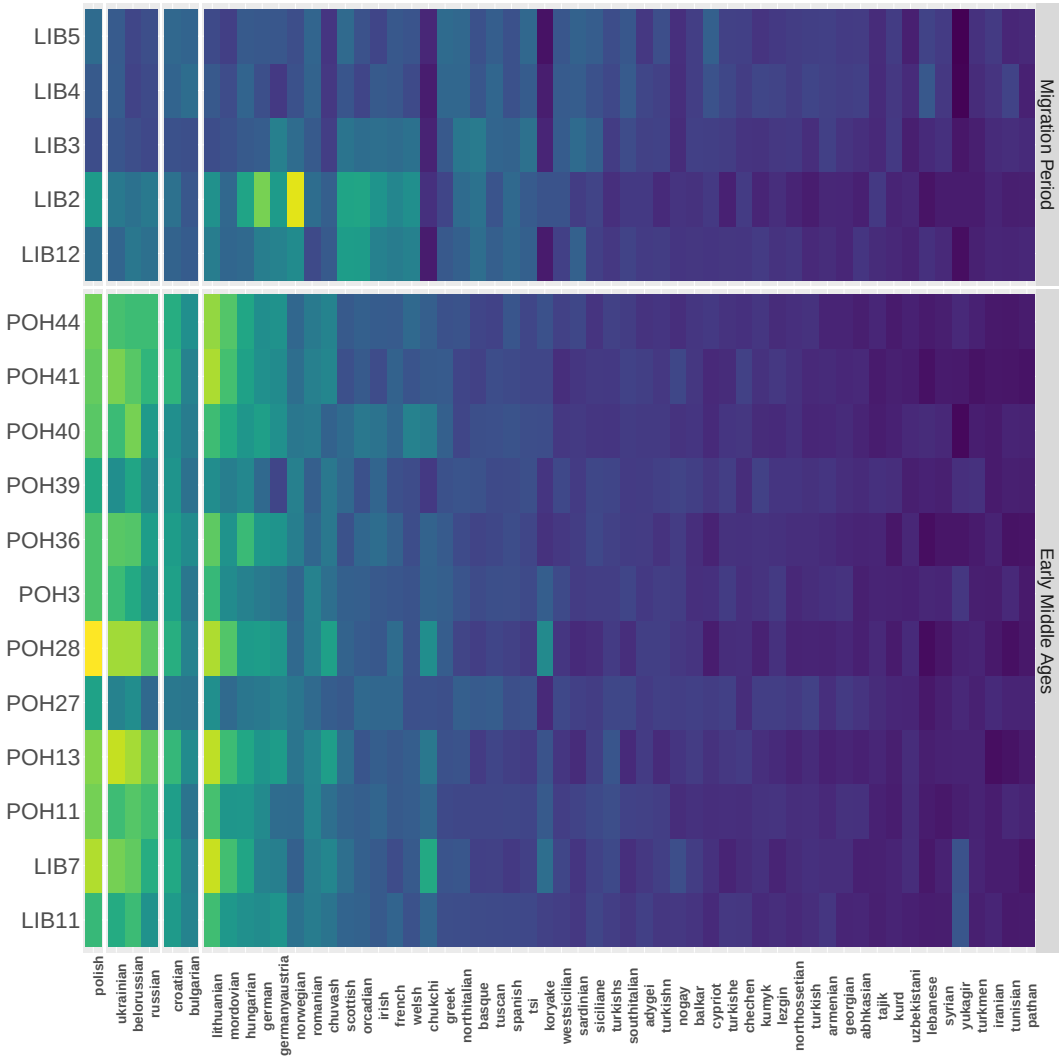

Figure 5.7: Raw chunklengths matrix from the 'present-day' painting. Rows correspond to different ancient recipient individuals, grouped into Migration Period and Early Middle Age period, and columns to different donor populations. Colour of cells corresponds to the total length of genome that a given donor individuals donates to that recipient, with dark/blue indicating less sharing and light/yellow colours indicating more sharing.

In contrast, the Early Middle Age samples showed a strong affinity to present-day day Slavic populations, especially Polish, Lithuanians and Mordovians.

To confirm that the observed results were not a result of phasing or imputing ancient individuals using present-day samples, I calculated f_3 statistics on pre-imputation genotypes. Specifically, I calculated f_3, or the branch length / amount of shared drift, between a set of present-day test populations and the grouped Early Middle Age samples. The results are qualitatively similar to those obtained using ChromoPainter, with Early Middle Age ancient Slavic individuals being closest to samples from Eastern Europe (Fig. 5.8). However, the f_3 results do not appear to show the same degree of geographical structure; for example, Early Middle Age have a more positive f_3 with present-day Irish individuals than with some present-day Slavic-speaking groups such as Croatians, perhaps reflecting relatively higher genetic drift in the Irish population.

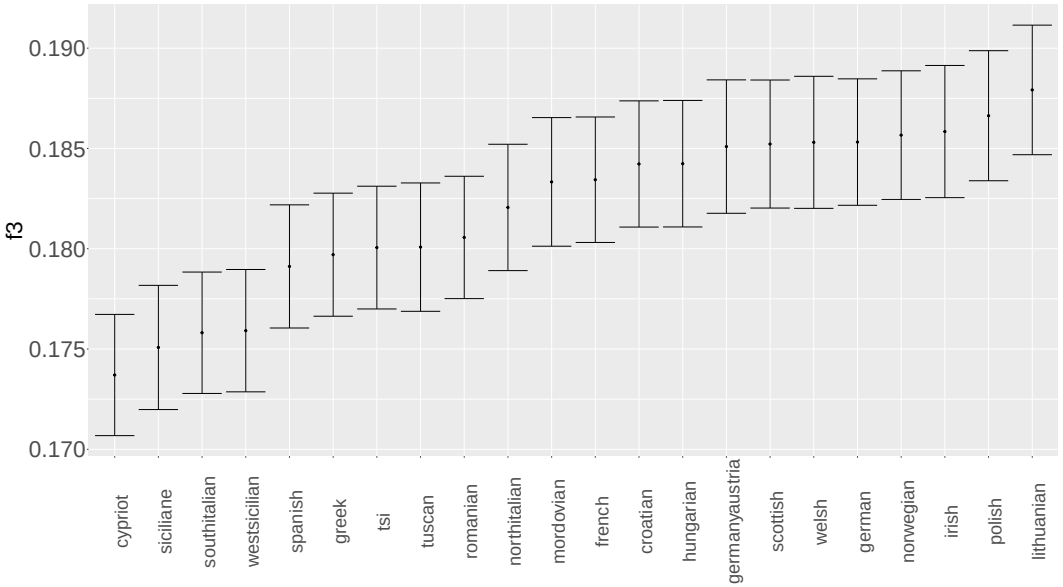

Figure 5.8: f_3 statistics in the form of $f_3(EMA, present{-}day; mbutipygmy)$, where *present-day* is different present-day European population. Error bars rerpesent $\pm * 2$ standard error.

5.3.5 Genetic structure and admixture events of present-day Slavic people

fineSTRUCTURE clustering on the 17 ancient samples with 21 present-day European populations gave results similar to those obtained from visually inspecting the chunklengths matrix in Fig 5.7. Among Migration Period samples, LIB2 and LIB12 cluster with north-west European groups, LIB3 clusters with Tuscany, and LIB4/LIB5 cluster with Spain. The present-day Slavic populations I had data for fall into two fineSTRUCTURE clades consistent with geography: (1) Croatians and Bulgarians ("south-east"), (2) Belarusians, Lithuanians, Polish, Russians and Ukrainians ("east"). Of the Early Middle Age samples, three (POH3, POH39, POH27) cluster into 'south-east' Slavic clade, with the remaining seven clustering into the 'east' clade. These results are consistent with the a hypothesis that the structure in present-day Slavic populations has been present since the Early Middle Ages.

Previous studies have identified admixture events in present-day Slavic populations involving an east-Asian source approximately 440 - 1080 CE [166, 205]. In previous sections, I showed that this signal exists in the Early Middle Age ancient samples and is best characterised by populations from present-day Mongolia (Fig. 5.4). I employed MOSAIC [166] to replicate the results of Hellenthal et al (2014) and Myers and Salter-Townshend (2019) and determine whether a similar admixing source is present in the ancient populations. I analysed all present-day populations (Table 5.3) and ancient Slavic populations in turn. For the ancient Slavic samples, I grouped all Early Middle Age samples together and grouped LIB3, LIB4 AND LIB5 together as the Migration Period samples.

When considering 2-way admixture event, all of the tested populations (both ancient and present-day), bar the Migration Period, showed evidence of an admixture event involving a minor source that has the lowest f_{st} with

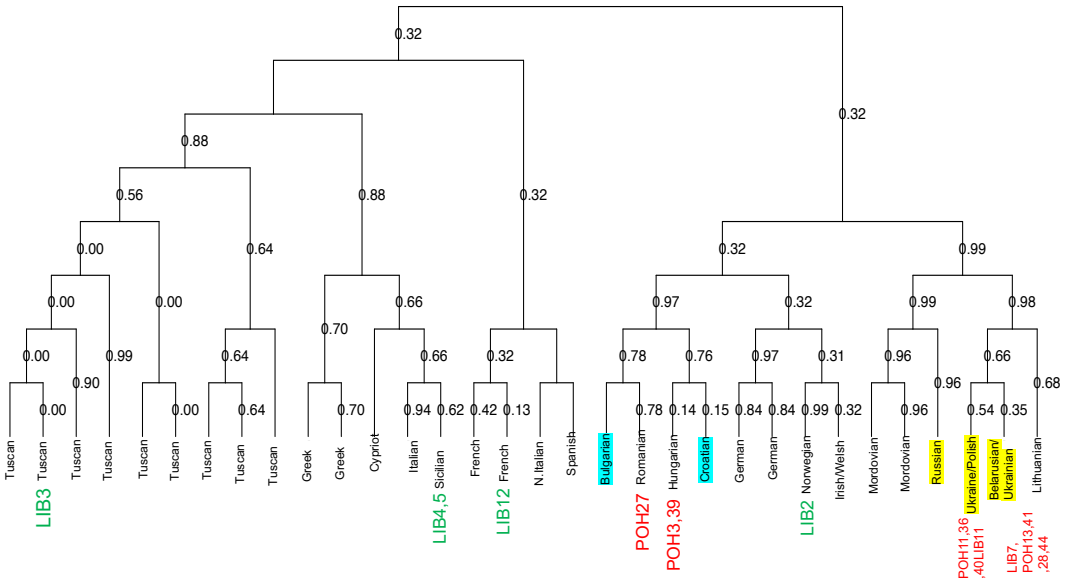

Figure 5.9: Population dendrogram generated by the fineSTRUCTURE tree building algorithm. Labeled tips refer to the primary population(s) represented in that clade. present-day non-Slavic populations shown in black. 'south-east' Slavs highlighted in cyan and 'north-west' Slavs highlighted in yellow. Migration period individuals superimposed in green and Early Middle Age samples superimposed in red. Read fineSTRUCTURE paper for description of edge values. Note: some tips contained more than one population but were not included as labels to save space.

present-day Uygurs. The dates and bootstrapped confidence intervals are given in Fig. 5.10. Other than Norwegians and Croatians, whose dates are later and earlier respectively, the dates for other populations appear to be constrained around 1250 CE. This date is similar, but slightly later than that obtained from Hellenthal et al (2014), who estimate it to be 440 to 1080 CE.

Of the present-day Slavic speaking populations, Belorussian, Polish and Ukrainian, show evidence of a 3-way admixture event, in which the middle component has the lowest f_{st} with Migration Era ancient samples (Fig. 5.11). The major component has a low f_{st} with Early Middle Age Slavs. This suggests that the formation of present-day Slavic populations could have occurred via admixture events involving Migration Era individuals with high levels of Southern European ancestry, Middle Age Era samples which show a strong affinity to present day eastern Europeans, and a small but significant east Asian

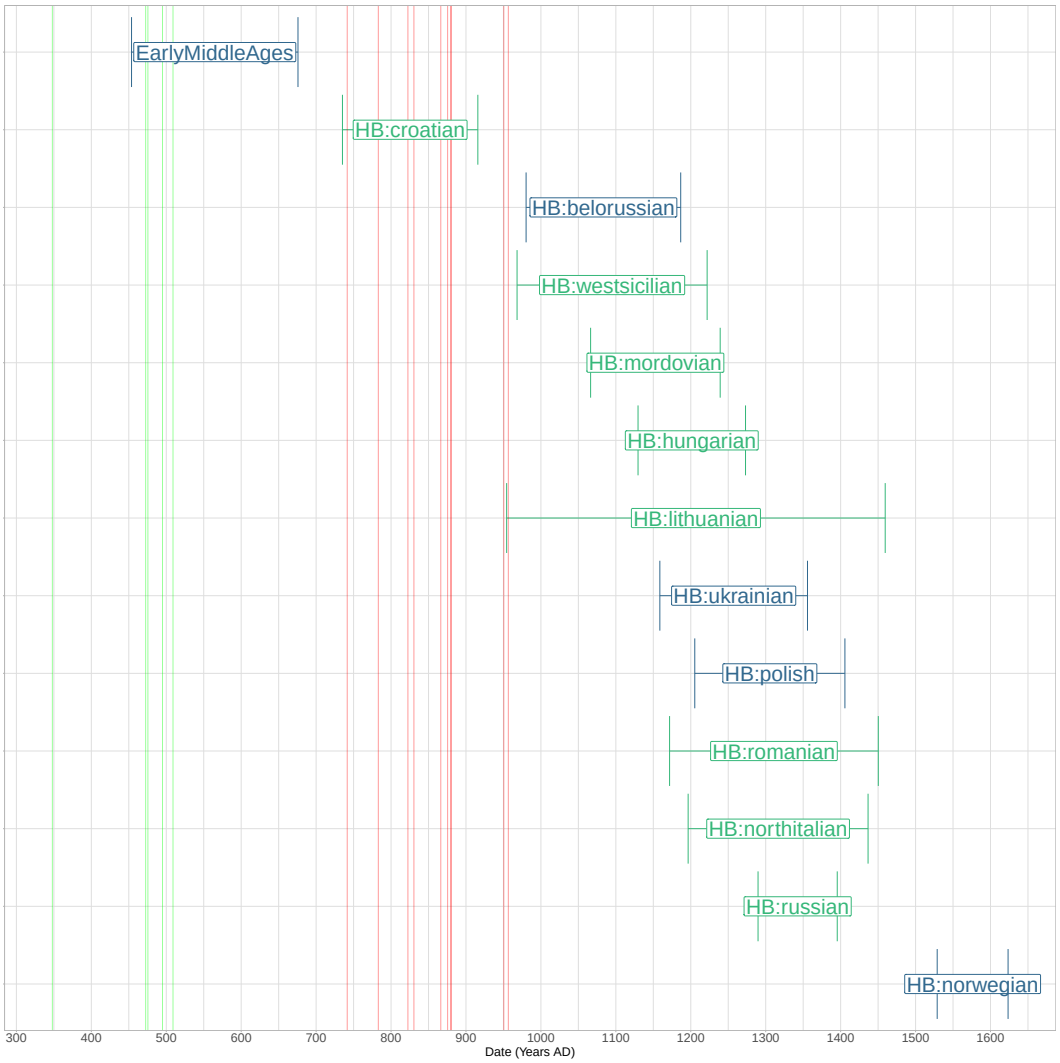

Figure 5.10: MOSAIC inferred 2-way admixture dates with bootstrapped 97.5% and 2.5% CI. Vertical green lines correspond to radiocarbon estimated dates of Migration Period samples and red lines equivalent for Early Middle Age samples. Estimated dates obtained by assuming an average generation time of 26 and date of birth of 1950 for present-day samples.

source best represented by present-day Uygurs. These results are similar to those in the Middle Age samples (Fig. 5.4), though dates are more recent in the present-day samples (Fig 5.10), suggesting recent admixture in present-day populations may be masking the older signals we see in the Early Middle Ages group.

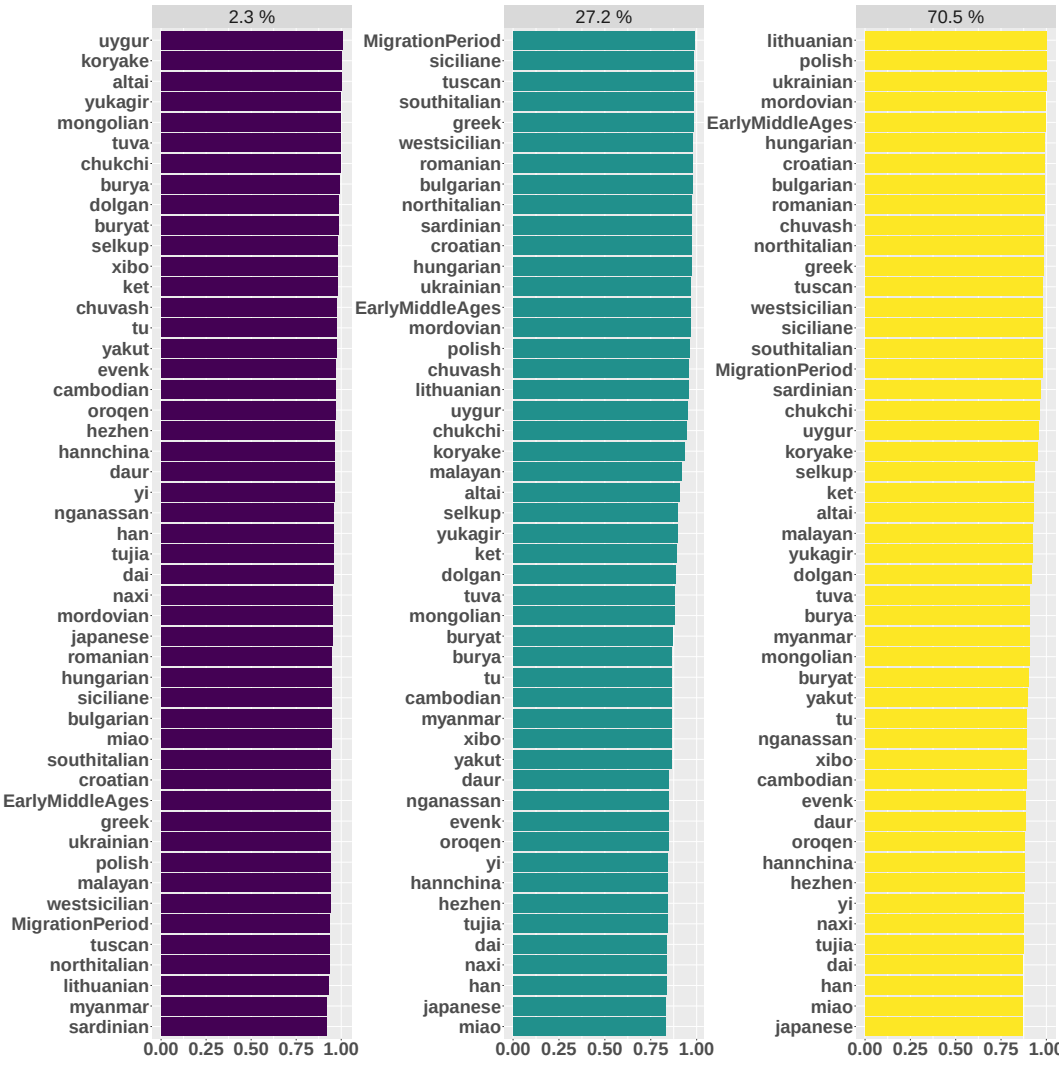

Figure 5.11: $1 - F_{st}$ between 3 inferred mixing sources for present-day Belorussians. Each panel represent a different mixing source. Each bar gives the value $1 - F_{st}$ between that samples population and the mixing source. Higher values of $1 - F_{st}$ suggest that source is well represented by a particular population.

5.4 Summary of Results and Discussion

Referring back to the questions posed in the introduction.

I found that the Migration Period samples, relative to the Early Middle Age samples, show a high degree of diversity in terms of ancestry, with affinities to present-day samples varying from Norway to southern Italy. On the other hand, fineSTRUCTURE analysis on the 'ancients' painting grouped all Early Middle Age samples together, showing that they represent a group of samples which likely share common ancestry. Consistent with this, the Early Middle Age samples showed evidence of east Asian admixture, a signal that was not present in the Migration Period samples. These results suggest a population turnover may have occurred between approximately 500-700 AD, the time period between the Migration Period and Early Middle Age. However, based on MOSAIC results of present day populations, a model of mixture between sources close to Migration Era, Early Middle Age and east-Asians seems plausible (Fig. 5.11).

All of the Early Middle Age samples showed a high genetic similarity to present-day Slavic and non-Slavic speaking populations from eastern Europe, such as Poland and Lithuania (Fig. 5.7). This is in stark contrast to the Migration Period, who all fell on a cline of genetic similarity between present-day Scandinavian and Mediterranean populations (Fig. 5.6). These results provide strong evidence that continuity exists between the Early Middle Ages and the present-day, but not between the Migration Period and Early Middle Ages.

Finally, a joint fineSTRUCTURE analysis which included both ancient and present-day samples showed that present-day Slavic speakers can be split into 'north-west' and 'south-east' groups, and that different Early Middle Ages samples had differing affinities to these groups (Fig 5.9).

I found strong evidence that LIB2 was a recent migrant from Viking regions.

There are many sources which detail the links between the Viking and Slavic peoples towards the end of the first millennium [206, 207]. However, most evidence suggests these links occurred later than the estimated radiocarbon date of LIB2. For example, it is known that the Scandinavian colonists settled in present-day Russia as early as 750 AD, whilst LIB2 was samples at approximately 495 AD. Therefore, we could suggest that this is evidence of an earlier link than previously known. In their large-scale study of ancient DNA of Viking samples from across Europe, Margaryan et al (2020) present Viking samples and ancestry in Estonia, but not until the beginning of the 8th Century, some 200 years after the estimated date of LIB2.

I also found evidence of southern European-like ancestry in three (LIB3, LIB4 and LIB5) Migration Period samples. The appearance of southern European-like ancestry in Central Europe in the first millennium is similar to a signal found in a study exploring the ancestry of individuals with elongated skulls in medieval Bavaria (approximately 500AD) [208]. It was shown that particular individuals harbour substantial Southern-European ancestry from outside of Bavaria, closest to individuals from present-day Greece and Turkey. There are at least two possible explanations for the presence of this ancestry in the Migration Era samples. Firstly, LIB3, LIB4 and LIB5 may be similar migrants to the region. This is consistent with the fact they are all female; Veeramah et al (2018) showed that there was a tendency for females to migrate from southern regions, perhaps related to the formation of strategic alliances. Alternatively, it is possible these individuals are a leftover from a historically described Lombard migration that moved from Northern Germany through Czechia, Slovakia, Hungary and ended up in Lombardia (Z. Hofmanova, personal communication). Accordingly, this could appear as genetic similarity to present-day populations from Northern Italy. This hypothesis is supported by the clustering of LIB3, LIB4 and LIB5 with present-day Italian samples in the 'present-day' fineSTRUCTURE analysis (Fig 5.9).

The results from the analysis of combined ancient and present-day genomes are consistent with those from Kushniarevich et al (2015) [195] who determined that Eastern (Russia, Belarus, Ukraine) and Western (Polish) central European Slavs form a cluster to the exclusion of Southern Slavs (Croatia, Bulgaria), whilst also remaining distinct from geographically proximate Germanic (German/Austrian) and Baltic (Lithuanian) populations. This is also consistent with results from Veeramah et al 2011, who showed that Sorbs, a west-Slavic population found between Poland and Germany, have a much stronger affinity to more distant Slavic populations from Czechia than to more proximate Germans [160].

Chapter 6

General Conclusions

6.1 General summary

In this thesis, I have explored the use of ChromoPainter on ancient DNA samples and present-day samples which contain sparsely genotyped markers. I evaluated the impact of coverage on all steps of the analysis pipeline, from imputation and phasing to ChromoPainter and SOURCEFIND analysis, focussing on the trade-off between potential gains from leveraging haplotype information and potential reference bias. I then applied my findings to two novel and one publicly available dataset(s).

In Chapter 2, I showed that the copyvectors of $\geq$0.5x downsamples show a high correspondence with the same sample at full coverage (Fig. 2.5), when painted using both ancient and present-day donors.

Disappointingly, my several attempts to improve the performance of ChromoPainter on 0.1x and 0.5x samples were not successful, including filtering the SNPs used using different criteria (Section 2.8). This was surprising, as my work and that of others [63] showed that filtering SNPs on e.g. genotype probabilities could substantially reduce the overall fraction of incorrectly imputed genotypes. I also found evidence of imputation bias towards the reference (Fig. 2.13).

Using present-day samples, I also showed that you can gain haplotype information using sparsely genotyped data with (presumably) perfect information at each SNP. Specifically, individuals from Cornwall and Devon can be distinguished genetically with >90% accuracy using only 1565 500-kb regions that contain ≈ 6.6 SNPs on average (i.e. ≈ 40,000 SNPs in total) (Table. 2.5). A similar classification rate was found for distinguishing Mandenka from Senegal and Yoruba from Nigeria, with >90% accuracy when using 1565 500-kb regions of ≈ 8 SNPs (Table 2.6). However, it appears current imputation approaches do not make reliable enough genotype calls on aDNA samples with <0.5x average coverage to provide many 500-kb windows with correctly called (and no incorrectly called) genotypes. Perhaps this is not surprising, as my exploration of 587 available ancient DNA samples revealed that genomes with 0.5x coverage have <1500 500-kb regions with 12 SNPs covered by even two reads (Figure 2.16), making calling heterozygotes challenging (or impossible) throughout the genome.

In Chapter 3, I explored African ancestry in U.K. Biobank samples. Following from my Chapter 2 findings, I showed that it is possible to recover substantial haplotype information with only a fraction of the total number of SNPs usually used. Being able to use fewer SNPs in an analysis will allow different datasets to be merged and jointly analysed, opening up a larger array of questions to be answered, whilst also significantly reducing the computational footprint of an analysis. I found that in terms of fine-scale population assignment, performing imputation on non-European samples using a predominantly European reference panel biases ChromoPainter analyses towards reference populations (Fig 3.3), as does performing analysis in unlinked mode (Table 3.1). Indeed, performing analysis on a majority imputed SNPs is more harmful for accuracy than using 70,000 SNPs in unlinked mode. This suggests that imputing to combine data from different SNP arrays may actually be more harmful than using a relatively small number (<100,000) of overlapping non-imputed SNPs when inferring fine-scale ancestry patterns.

My analyses showed that approximately 4% of U.K. Biobank participants have at least 50% African ancestry. Within this set of individuals, genetic ancestry from West Africa was very prevalent, consistent with historical events (Fig 3.5). In particular, I found that there was over ten times the number of individuals with at least 50% ancestry from Yoruba than there was the next most common ancestry.

In Chapter 4, I analysed novel ancient DNA datasets from Bavaria with the samples spanning almost 7000 years of history. The analysis of ancient Bavarian samples recapitulated previous research which identified admixture events between early farmers and local hunter-gatherers, and the presence of steppe-related ancestry in the Late Neolithic. However, it also provided some less expected results, showing that samples with extremely different ancestries cohabited the same cave and the same time period. I also identified ancestry most closely related to Iron Age Italian sourced which arrived in Bavaria during the Iron Age, but was not present in the preceding Bronze Age, which may be related the migrations of Lombard populations. Future studies could increase the number of ancient sample sequenced from Bronze and Iron Age Bavaria in order to constrain the date the ancestry appears and source of origin. Finally, I showed that early Germanic and Slavic samples from the Middle Ages, which could not be distinguished using other ancient samples, showed strong genetic differences when analysed using present-day data (Fig. 4.9). Whilst I was able to identify structure down to the level of individuals countries, the lack of data from different regions in Germany meant that I was not able to determine whether there was fine-scale differential relatedness to the ancient samples for different German states.

My final Chapter analysed the differences between Migration Era and Early Middle Age samples from Czechia. The data revealed that whilst different Migration Era samples displayed genetic affinities to a wide spectrum of other ancient and present-day populations, the Early Middle Age individuals were

relatively more homogenous and broadly showed strong similarity to present-day Slavic speaking populations (Fig. 5.7). However, fineSTRUCTURE analysis using present-day Slavic and non-Slavic speaking populations clearly showed that present-day Slavic speaking populations can be split into south-east and north-west clusters, with different ancient samples showing different affinities to each cluster. Lastly, I provided evidence that previously reported [20, 166] signals of east-Asian admixture in eastern-European populations was also present in the Early Middle Age ancient Slavic samples (Fig. 5.5. Although the five Migration Era samples represented an array of ancestries present in Czechia during that period, the sample size (n=3 at most) per sub-population was too low to reliably infer admixture events.

6.2 Recommendations

My recommendations for analysing low coverage data are as follows:

1. Include samples with at least 0.5x mean coverage. Samples below this coverage (0.1x) show effects of coverage-related bias in copyvector estimation, SOURCEFIND analysis and positions on a PCA.

2. When merging data from different genotyping arrays, it is preferable only to retain directly genotyped SNPs rather than imputing missing ones using a reference panel. This applies when the total number of directly genotyped SNPs is at least 45,000 (Fig. E.2).

6.3 Limitations of work and future avenues of research

Firstly, I did not consider ancient samples from Africa. This is in part because of a lack of high coverage samples from Africa (Mota being the highest coverage at 10x) and the vast majority of ancient DNA samples from western Eurasia.

I expect results to differ when considering African samples. Africans harbour more diversity and have lower levels of background LD [139] and thus would be expected to match shorter segments to other individuals. Imputation accuracy would likely be lower, in part because of less LD and higher genetic diversity, but also because less of the total proportion of genetic diversity is present in reference panels. Finally, the large population turnovers in Africa (e.g. the Bantu expansions) mean that many pre-Bantu ancient samples may harbour diversity that does not exist in present-day individuals. Therefore, it is possible that coverage greater than 0.5x may be necessary to accurately analyse African samples with ChromoPainter.

I did not evaluate the effect of coverage on either fineSTRUCTURE or GLOBETROTTER analysis. This is because GLOBETROTTER struggles to identify admixture events in single samples and I only had a single downsample for each individual and level of coverage. To accurately estimate admixture events, segments of DNA within an individual copied from different populations need to be identified. Such segments may be particularly hard to identify in low coverage samples, as the segment boundaries may contain low-coverage SNPs.

I didn't use the largest reference panel (HRC) to impute ancient samples, due to technical challenges in obtaining access to the data and so likely underestimate the potential accuracy of imputation on low coverage samples. Thus, future work should examine the scale of improvements in imputation accuracy when using extremely large reference panels. For example, plans to sequence the whole-genomes of 200,000 U.K. Biobank participants would provide an unparalleled resource to impute variants in ancient samples of western European ancestry.

Whilst my attempt at incorporating genotype likelihoods into the ChromoPainter process only provided very modest improvements, the fact that this approach has been successful in other methods [116, 209–211] suggests

that in theory it should also be applicable to Chromosome painting. Future work on ChromoPainter could explore the reason why this did not work and suggest alternate ways in which to account for the uncertainty associated with low coverage data. Studies could also interrogate the performance of ChromoPainter on the range of coverages between 0.1-0.5x. Recent research has argued it is possible to infer ancestral relationships between samples as low as 0.1x in coverage, although only for particular applications such as demographic change [212].

On the other hand, methodological advances in laboratory DNA extraction techniques, DNA enrichment and sequencing technologies and library preparation for ancient samples may mean that all samples can be sequenced to a high enough coverage that coverage-related effects are inconsequential.

Appendix A

Datasets used

This appendix described the different datasets used in analyses performed in this thesis. It includes datasets of both modern and ancient genomes.

A.1 Ancient reference dataset

This section describes the generation of the dataset of reference ancient individuals used in Chapters 2, 4 and 5.

For each of the samples in Table A.1, the following steps were taken to produce ChromoPainter input.

1. Each `.bam` was processed with `PicardTools ValidateBam` [96] task to ensure no files were corrupted or contained incorrect read group information.

2. Each `.bam` file was processed with atlas (version 1.0, commit f612f28) pipeline [71] (`https://bitbucket.org/wegmannlab/atlas/wiki/Home`). For .bam file, I estimated post-mortem damage (PMD) patterns using atlas `estimatePMD` task. Recalibration parameters were then estimated using atlas `recal` task. Finally, both the recalibration and PMD parameters were given to the `callNEW` task which produces genotype calls and

Paper	Number of Samples	Reference
Allentoft 2015	20	[151]
Antonio 2019	134	[59]
Broushaki 2016	1	[213]
Brunel 2020	58	[157]
Cassidy 2015	4	[214]
deBarrosDamgaard 2018a	34	[93]
deBarrosDamgaard 2018b	58	[180]
Gamba 2014	10	[174]
Gunther 2015	2	[149]
Hofmanova 2016	5	[150]
Jones 2015	2	[215]
Lazaridis 2014	1	[51]
Marchi 2020	4	[216]
Margaryan 20	442	[58]
Berger unpublished	14	NA
Olade 2014	1	[217]
Rivollat 20	101	[167]
Sanchez-Quinto 2019	7	[218]
Seguin-Orlando 2014	1	[219]
Veeramah 2018	1	[208]
Hofmanova unpublished	37	NA

Table A.1: Name of paper, number of samples and reference for all literature ancient samples used in analyses

.

genotype likelihood estimates for each downsampled and full coverage .bam. For this stage, I made calls at the 77,818,345 genome-wide positions present in the phase 3 thousand genomes project [97]. This was done to reduce the risk of calling false-positive non-polymorphic sites. This resulted in a `.bcf` file for each ancient sample.

3. All `.bcf` files were split into chromosomes and all samples from the same chromosome were merged. Imputation and phasing was performed with GLIMPSE (version 1.1.1). I followed the steps laid out in the GLIMPSE tutorial (`https://odelaneau.github.io/GLIMPSE/tutorial_b38.html`). First, I used `GLIMPSE_chunk` to split up each reference chromosome into chunks, keeping both `-window-size` and

`-buffer-size` to 2,000,000, their default settings. Across all chromosomes, this produced 936 chunks of an average 2.99Mb long. I used the b37 genetic map supplied by GLIMPSE for the `-map` argument.

Each chunk was then imputed separately using `GLIMPSE_phase` using the same 1000 genomes dataset as a reference. Default settings and the supplied b37 genetic map were used. This stage both imputes missing genotypes and generates a set of haplotype pairs which can be sampled from in a later step to produced phased haplotypes.

`GLIMPSE_ligate` was then used to merge the imputed chunks back to form single chromosomes using the default settings and the supplied b37 genetic map.

Haplotypes were then sampled using `GLIMPSE_sample` to produce a .vcf with phased haplotypes for each individual, again using default settings and the supplied b37 genetic map.

Consequently, the output of GLIMPSE is i) unphased genotype calls with posterior genotype likelihoods and ii) phased haplotypes.

4. Finally, the posterior genotype likelihoods and phased haplotypes were combined to generate ChromoPainterUncertainty output using a custom script (`https://github.com/sahwa/vcf_to_chromopainter`).

A.2 30x 1000 genomes dataset

Samples from [102].

This dataset consists of 3,202 modern individuals from 172 worldwide populations, sequenced to a targeted depth of 30x coverage. The downloaded dataset was aligned to the gr38 reference genome. Samples were downloaded to the UCL Computer Science cluster by myself from the ftp mirror. The following steps were taken to process the data before being used as an imputation reference.

1. Filtered such that SNPs with only 2 alleles were retained
2. Performed a liftover to hg19 using LiftoverVcf from picard tools [96]
3. Filter again for SNPs with only 2 alleles
4. Phase using shapeit4, using the 'sequencing' parameter and setting –pbwt-depth 4.
5. Remove duplicated SNPs using bcftools norm [220]
6. Use Beagle's conform-gt utility to ensure reference alleles were consistent with the previous 1000 genomes build. This was done because all previous datasets I have compiled were also conformed to the previous 1000 genomes build.

Population Name	Number of Individuals	Data Collection
Abkhasian	2	Simons Genome Diversity Project
Adygei	17	Simons Genome Diversity Project
African Ancestry SW	112	1000 Genomes
African Caribbean	123	1000 Genomes
Albanian	1	Simons Genome Diversity Project
Aleut	2	Simons Genome Diversity Project
Altaian	1	Simons Genome Diversity Project
Ami	2	Simons Genome Diversity Project
Armenian	2	Simons Genome Diversity Project
Atayal	1	Simons Genome Diversity Project
Australian	2	Simons Genome Diversity Project
Balochi	24	Human Genome Diversity Project
Bantu Herero	2	Simons Genome Diversity Project
Bantu Kenya	12	Human Genome Diversity Project
Bantu South Africa	4	Human Genome Diversity Project
Bantu Tswana	2	Simons Genome Diversity Project
Basque	24	Human Genome Diversity Project
Bedouin	44	Human Genome Diversity Project
Bedouin B	2	Simons Genome Diversity Project
Bengali	142	1000 Genomes
Bengali,Bengali	2	1000 Genomes
Bergamo	2	Simons Genome Diversity Project
Bergamo Italian	10	Human Genome Diversity Project
Biaka	26	Human Genome Diversity Project

(continued)

Population Name	Number of Individuals	Data Collection
Bougainville	13	Human Genome Diversity Project
Brahmin	2	Simons Genome Diversity Project
Brahui	25	Human Genome Diversity Project
British	105	1000 Genomes
British,English	2	1000 Genomes
Bulgarian	2	Simons Genome Diversity Project
Burmese	2	Simons Genome Diversity Project
Burusho	24	Human Genome Diversity Project
Cambodian	9	Human Genome Diversity Project
Cambodian,Cambodian	1	Simons Genome Diversity Project
CEPH	184	1000 Genomes
Chane	1	Simons Genome Diversity Project
Chechen	1	Simons Genome Diversity Project
Chukchi	1	Simons Genome Diversity Project
Colombian	153	1000 Genomes
Crete	2	Simons Genome Diversity Project
Czech	1	Simons Genome Diversity Project
Dai	9	Human Genome Diversity Project
Dai Chinese	109	1000 Genomes
Daur	10	Human Genome Diversity Project
Dinka	3	Simons Genome Diversity Project
Druze	42	Human Genome Diversity Project
Dusun	2	Simons Genome Diversity Project
Esan	171	1000 Genomes
Esan,Esan	2	1000 Genomes
Eskimo Chaplin	1	Simons Genome Diversity Project
Eskimo Naukan	2	Simons Genome Diversity Project
Eskimo Sireniki	2	Simons Genome Diversity Project
Estonian	2	Simons Genome Diversity Project
Even	3	Simons Genome Diversity Project
Finnish	102	1000 Genomes
Finnish,Finnish	3	1000 Genomes
French	27	Human Genome Diversity Project
Gambian Fula	100	Gambian Genome Variation Project
Gambian Jola	100	Gambian Genome Variation Project
Gambian Mandinka	278	1000 Genomes
Gambian Mandinka,Gambian	2	1000 Genomes
Gambian Wolof	100	Gambian Genome Variation Project
Georgian	2	Simons Genome Diversity Project
Greek	2	Simons Genome Diversity Project
Gujarati	113	1000 Genomes

(continued)

Population Name	Number of Individuals	Data Collection
Han	33	Human Genome Diversity Project
Han Chinese	112	1000 Genomes
Hawaiian	1	Simons Genome Diversity Project
Hazara	20	Human Genome Diversity Project
Hezhen	9	Human Genome Diversity Project
Hungarian	2	Simons Genome Diversity Project
Iberian	160	1000 Genomes
Iberian,Spanish	2	1000 Genomes
Icelandic	2	Simons Genome Diversity Project
Igorot	2	Simons Genome Diversity Project
Iranian	2	Simons Genome Diversity Project
Iraqi Jew	2	Simons Genome Diversity Project
Irula	2	Simons Genome Diversity Project
Itelman	1	Simons Genome Diversity Project
Japanese	133	1000 Genomes
Japanese,Japanese	1	1000 Genomes
Jordanian	3	Simons Genome Diversity Project
Ju|'hoan North	2	Simons Genome Diversity Project
Ju|'hoan North,San	2	Simons Genome Diversity Project
Kalash	23	Human Genome Diversity Project
Kapu	2	Simons Genome Diversity Project
Karitiana	12	Human Genome Diversity Project
Khomani San	2	Simons Genome Diversity Project
Khonda Dora	1	Simons Genome Diversity Project
Kinh Vietnamese	122	1000 Genomes
Kinh,Kinh Vietnamese	2	1000 Genomes
Korean	2	Simons Genome Diversity Project
Kusunda	2	Simons Genome Diversity Project
Kyrgyz	2	Simons Genome Diversity Project
Lahu	8	Human Genome Diversity Project
Lezgin	2	Simons Genome Diversity Project
Luhya	114	1000 Genomes
Luhya,Luhya	2	1000 Genomes
Luo	2	Simons Genome Diversity Project
Madiga	1	Simons Genome Diversity Project
Makrani	25	Human Genome Diversity Project
Mandenka	23	Human Genome Diversity Project
Mansi	2	Simons Genome Diversity Project
Maori	1	Simons Genome Diversity Project
Masai	2	Simons Genome Diversity Project
Maya	19	HGDP Transcriptome

(continued)

Population Name	Number of Individuals	Data Collection
Mayan,Maya	2	Simons Genome Diversity Project
Mbuti	13	Human Genome Diversity Project
Mbuti,Mbuti	2	Simons Genome Diversity Project
Mende	126	1000 Genomes
Mende,Mende	2	1000 Genomes
Mexican Ancestry	107	1000 Genomes
Miao	10	Human Genome Diversity Project
Mixe	3	Simons Genome Diversity Project
Mixtec	2	Simons Genome Diversity Project
Mongola	2	Simons Genome Diversity Project
Mongolian	8	Human Genome Diversity Project
Mozabite	27	Human Genome Diversity Project
Mozabite,Mozabite	1	Simons Genome Diversity Project
Naxi	9	Simons Genome Diversity Project
North Ossetian	2	Simons Genome Diversity Project
Northern Han	10	Human Genome Diversity Project
Norwegian	1	Simons Genome Diversity Project
Orcadian	15	Human Genome Diversity Project
Oroqen	9	Human Genome Diversity Project
Palestinian	46	Human Genome Diversity Project
Papuan	3	Simons Genome Diversity Project
Papuan Sepik	2	Human Genome Diversity Project
Papuan,Papuan Highlands	6	Simons Genome Diversity Project
Papuan,Papuan Sepik	6	Simons Genome Diversity Project
Pathan	23	Human Genome Diversity Project
Pathan,Pathan	1	Simons Genome Diversity Project
Peruvian	130	1000 Genomes
Piapoco	2	Simons Genome Diversity Project
Pima	14	Human Genome Diversity Project
Polish	1	Simons Genome Diversity Project
Puerto Rican	150	1000 Genomes
Punjabi	154	1000 Genomes
Punjabi,Punjabi	4	1000 Genomes
Quechua	3	Simons Genome Diversity Project
Relli	2	Simons Genome Diversity Project
Russian	25	Human Genome Diversity Project
Saami	2	Simons Genome Diversity Project
Saharawi	2	Simons Genome Diversity Project
Samaritan	1	Simons Genome Diversity Project
San	2	Human Genome Diversity Project
Sardinian	27	Human Genome Diversity Project

(continued)

Population Name	Number of Individuals	Data Collection
She	10	Human Genome Diversity Project
Sindhi	24	Human Genome Diversity Project
Somali	1	Simons Genome Diversity Project
Southern Han Chinese	171	1000 Genomes
Surui	8	Human Genome Diversity Project
Tajik	2	Simons Genome Diversity Project
Tamil	128	1000 Genomes
Telugu	118	1000 Genomes
Thai	2	Simons Genome Diversity Project
Tlingit	2	Simons Genome Diversity Project
Toscani	112	1000 Genomes
Tu	10	Human Genome Diversity Project
Tubalar	2	Simons Genome Diversity Project
Tujia	10	Human Genome Diversity Project
Turkish	2	Simons Genome Diversity Project
Tuscan	8	Human Genome Diversity Project
Ulchi	2	Simons Genome Diversity Project
Uygur	10	Human Genome Diversity Project
Xibo	9	Human Genome Diversity Project
Yadava	2	Simons Genome Diversity Project
Yakut	25	Human Genome Diversity Project
Yemenite Jew	2	Simons Genome Diversity Project
Yi	10	Human Genome Diversity Project
Yoruba	207	1000 Genomes
Zapotec	2	Simons Genome Diversity Project

A.3 Human Origins dataset

This dataset consists of 560,420 SNPS and 5998 individuals from 509 worldwide populations. It has a particularly large number of samples from West and East Africa; in particular, Cameroon, Ethiopia, Nigeria and Ghana.

Region	Country	Populations	Ref	sum
Africa	Algeria	Algerian	Lazaridis et al 2014	4
Africa	Algeria	Mozabite	Lazaridis et al 2014	21
Africa	Botswana	Gana	Lazaridis et al 2014	7
Africa	Botswana	Gui	Lazaridis et al 2014	7
Africa	Botswana	Hoan	Lazaridis et al 2014	6

Africa	Botswana	Ju hoan South	Lazaridis et al 2014	5
Africa	Botswana	Kgalagadi	Lazaridis et al 2014	5
Africa	Botswana	Khwe	Lazaridis et al 2014	8
Africa	Botswana	Naro	Lazaridis et al 2014	8
Africa	Botswana	Shua	Lazaridis et al 2014	9
Africa	Botswana	Taa East	Lazaridis et al 2014	6
Africa	Botswana	Taa North	Lazaridis et al 2014	9
Africa	Botswana	Taa West	Lazaridis et al 2014	15
Africa	Botswana	Tshwa	Lazaridis et al 2014	4
Africa	Botswana	Tswana	Lazaridis et al 2014	5
Africa	BotswanaorNamibia	Bantu SA	Lazaridis et al 2014	8
Africa	Cameroon	Cameroon Baka	Fan 2019	2
Africa	Cameroon	Cameroon Bakola	Fan 2019	2
Africa	Cameroon	Cameroon Bedzan	Fan 2019	2
Africa	Cameroon	Cameroon Foulbe	Fan 2019	2
Africa	Cameroon	Cameroon Mada	Fan 2019	2
Africa	Cameroon	Cameroon Ngoumba	Fan 2019	2
Africa	Cameroon	Cameroon Tikar	Fan 2019	2
Africa	Cameroon	Cameroon Aghem	Lipson 2020	28
Africa	Cameroon	Cameroon Bafut	Lipson 2020	11
Africa	Cameroon	Cameroon Bakoko	Lipson 2020	1
Africa	Cameroon	Cameroon Bangwa	Lipson 2020	2
Africa	Cameroon	Cameroon Mbo	Lipson 2020	21
Africa	Cameroon	Cameroon Kotoko	Lopez 2021	7
Africa	CentralAfricanRepublic	BiakaPygmy	Lazaridis et al 2014	20
Africa	CentralAfricanRepublic	Kaba	Fan 2019	2
Africa	Chad	Bulala	Fan 2019	2
Africa	Chad	Laka	Fan 2019	2
Africa	Congo	MbutiPygmy	Lazaridis et al 2014	10
Africa	Egypt	Egyptian Comas	Lazaridis et al 2014	11
Africa	Egypt	Egyptian Metspalu	Lazaridis et al 2014	7
Africa	Ethiopia	Aari	Fan 2019	2
Africa	Ethiopia	Agaw	Fan 2019	2
Africa	Ethiopia	Amhara	Fan 2019	2
Africa	Ethiopia	Ethiopia Afar	Lopez 2021	10
Africa	Ethiopia	Ethiopia Agew	Lopez 2021	30
Africa	Ethiopia	Ethiopia Alaba	Lopez 2021	14
Africa	Ethiopia	Ethiopia Alae	Lopez 2021	46
Africa	Ethiopia	Ethiopia Amhara	Gurdasani et al 2015	24
Africa	Ethiopia	Ethiopia Amhara	Lopez 2021	28
Africa	Ethiopia	Ethiopia Anuak	Lopez 2021	9
Africa	Ethiopia	Ethiopia Arbore	Lopez 2021	14
Africa	Ethiopia	Ethiopia Ari Cultivator	Lopez 2021	14
Africa	Ethiopia	Ethiopia Ari Potter	Lopez 2021	24

Africa	Ethiopia	Ethiopia Ari Smith	Lopez 2021	14
Africa	Ethiopia	Ethiopia Basket	Lopez 2021	14
Africa	Ethiopia	Ethiopia Bena	Lopez 2021	28
Africa	Ethiopia	Ethiopia Bench	Lopez 2021	12
Africa	Ethiopia	Ethiopia Berta	Lopez 2021	13
Africa	Ethiopia	Ethiopia BetaIsrael	Lazaridis et al 2014	7
Africa	Ethiopia	Ethiopia BetaIsrael	Lopez 2021	6
Africa	Ethiopia	Ethiopia Bodi	Lopez 2021	14
Africa	Ethiopia	Ethiopia Burji	Lopez 2021	24
Africa	Ethiopia	Ethiopia Chara	Lopez 2021	17
Africa	Ethiopia	Ethiopia Dasanech	Lopez 2021	15
Africa	Ethiopia	Ethiopia Dawro	Lopez 2021	14
Africa	Ethiopia	Ethiopia DawroManja	Lopez 2021	11
Africa	Ethiopia	Ethiopia Dhime	Lopez 2021	21
Africa	Ethiopia	Ethiopia Dirasha	Lopez 2021	17
Africa	Ethiopia	Ethiopia Dizi	Lopez 2021	14
Africa	Ethiopia	Ethiopia Dorze	Lopez 2021	15
Africa	Ethiopia	Ethiopia Gedeo	Lopez 2021	21
Africa	Ethiopia	Ethiopia GentaGamo	Lopez 2021	15
Africa	Ethiopia	Ethiopia Gidicho	Lopez 2021	11
Africa	Ethiopia	Ethiopia Gofa	Lopez 2021	15
Africa	Ethiopia	Ethiopia Gumuz	Gurdasani et al 2015	20
Africa	Ethiopia	Ethiopia Gumuz	Lopez 2021	2
Africa	Ethiopia	Ethiopia Gurage	Lopez 2021	16
Africa	Ethiopia	Ethiopia Hadiya	Lopez 2021	14
Africa	Ethiopia	Ethiopia Hamer	Lopez 2021	14
Africa	Ethiopia	Ethiopia Honsita	Lopez 2021	17
Africa	Ethiopia	Ethiopia Kafacho	Lopez 2021	16
Africa	Ethiopia	Ethiopia Kambata	Lopez 2021	13
Africa	Ethiopia	Ethiopia Karo	Lopez 2021	14
Africa	Ethiopia	Ethiopia KefaShekaManjo	Lopez 2021	14
Africa	Ethiopia	Ethiopia Komo	Lopez 2021	8
Africa	Ethiopia	Ethiopia Konta	Lopez 2021	16
Africa	Ethiopia	Ethiopia Kore	Lopez 2021	16
Africa	Ethiopia	Ethiopia Kuwegu	Lopez 2021	10
Africa	Ethiopia	Ethiopia Maale	Lopez 2021	11
Africa	Ethiopia	Ethiopia Mao	Lopez 2021	9
Africa	Ethiopia	Ethiopia Masholae	Lopez 2021	19
Africa	Ethiopia	Ethiopia Menit	Lopez 2021	15
Africa	Ethiopia	Ethiopia Mezhenger	Lopez 2021	14
Africa	Ethiopia	Ethiopia Mossiye	Lopez 2021	10
Africa	Ethiopia	Ethiopia Murle	Lopez 2021	13
Africa	Ethiopia	Ethiopia Mursi	Lopez 2021	10
Africa	Ethiopia	Ethiopia Nao	Lopez 2021	17

Africa	Ethiopia	Ethiopia NegedeWoyto	Lopez 2021	9
Africa	Ethiopia	Ethiopia Nuer	Lopez 2021	11
Africa	Ethiopia	Ethiopia Nyangatom	Lopez 2021	12
Africa	Ethiopia	Ethiopia Oromo	Gurdasani et al 2015	24
Africa	Ethiopia	Ethiopia Oromo	Lazaridis et al 2014	4
Africa	Ethiopia	Ethiopia Oromo	Lopez 2021	7
Africa	Ethiopia	Ethiopia OtherGamo	Lopez 2021	16
Africa	Ethiopia	Ethiopia Qimant	Lopez 2021	17
Africa	Ethiopia	Ethiopia Shabo	Lopez 2021	11
Africa	Ethiopia	Ethiopia Shekacho	Lopez 2021	16
Africa	Ethiopia	Ethiopia Sheko	Lopez 2021	15
Africa	Ethiopia	Ethiopia Shinasha	Lopez 2021	18
Africa	Ethiopia	Ethiopia Sidama	Lopez 2021	21
Africa	Ethiopia	Ethiopia Somali	Gurdasani et al 2015	24
Africa	Ethiopia	Ethiopia Somali	Lopez 2021	2
Africa	Ethiopia	Ethiopia Suri	Lopez 2021	14
Africa	Ethiopia	Ethiopia Tigray	Lopez 2021	13
Africa	Ethiopia	Ethiopia Tsemay	Lopez 2021	18
Africa	Ethiopia	Ethiopia Wolayta	Gurdasani et al 2015	21
Africa	Ethiopia	Ethiopia Wolayta	Lopez 2021	4
Africa	Ethiopia	Ethiopia Wolayta Cultivator	Lopez 2021	6
Africa	Ethiopia	Ethiopia Wolayta Potter	Lopez 2021	10
Africa	Ethiopia	Ethiopia Wolayta Smith	Lopez 2021	12
Africa	Ethiopia	Ethiopia Wolayta Tanner	Lopez 2021	8
Africa	Ethiopia	Ethiopia Wolayta Weaver	Lopez 2021	12
Africa	Ethiopia	Ethiopia Yem	Lopez 2021	13
Africa	Ethiopia	Ethiopia Zayse	Lopez 2021	17
Africa	Ethiopia	Ethiopia Zilmamo	Lopez 2021	12
Africa	Ethiopia	Mursi	Fan 2019	2
Africa	Gambia	Gambian GWD	Lazaridis et al 2014	6
Africa	Kenya	BantuKenya	Lazaridis et al 2014	6
Africa	Kenya	Elmolo	Fan 2019	2
Africa	Kenya	Kikuyu	Fan 2019	2
Africa	Kenya	Kikuyu	Lazaridis et al 2014	4
Africa	Kenya	Luhya Kenya LWK	Lazaridis et al 2014	8
Africa	Kenya	Luo	Lazaridis et al 2014	8
Africa	Kenya	Masai Ayodo	Lazaridis et al 2014	2
Africa	Kenya	Masai Kinyawa MKK	Lazaridis et al 2014	9
Africa	Kenya	Ogiek	Fan 2019	2
Africa	Kenya	Rendille	Fan 2019	2
Africa	Kenya	Sengwer	Fan 2019	2

Africa	Khomani	Khomani	Lazaridis et al 2014	9
Africa	Libya	Libyan Jew	Lazaridis et al 2014	9
Africa	Malawi	Malawi Chewa	Skoglund et al 2015	11
Africa	Malawi	Malawi Ngoni	Skoglund et al 2015	4
Africa	Malawi	Malawi Tumbuka	Skoglund et al 2015	10
Africa	Malawi	Malawi Yao	Skoglund et al 2015	9
Africa	Morocco	Moroccan Jew	Lazaridis et al 2014	6
Africa	Morocco	MoroccoBerber	Lopez 2021	19
Africa	Morocco	Saharawi	Lazaridis et al 2014	6
Africa	Namibia	Damara	Lazaridis et al 2014	12
Africa	Namibia	Haiom	Lazaridis et al 2014	7
Africa	Namibia	Himba	Lazaridis et al 2014	4
Africa	Namibia	Ju hoan North	Lazaridis et al 2014	21
Africa	Namibia	Nama	Lazaridis et al 2014	16
Africa	Namibia	Wambo	Lazaridis et al 2014	5
Africa	Namibia	Xuun	Lazaridis et al 2014	13
Africa	Nigeria	Nigeria Esan	Lazaridis et al 2014	8
Africa	Nigeria	Nigeria Yoruba	Lazaridis et al 2014	70
Africa	Saudi-Beduins	SaudiBeduins	Lopez 2021	8
Africa	Senegal	Mandenka	Lazaridis et al 2014	17
Africa	Senegal	Senegal	Lopez 2021	13
Africa	SierraLeone	Mende Sierra Leone MSL	Lazaridis et al 2014	8
Africa	Somalia	Somali	Lazaridis et al 2014	13
Africa	SouthAfrica	Zulu	Gurdasani et al 2015	100
Africa	Sudan	Sudan Dinka	Lazaridis et al 2014	7
Africa	Tanzania	Datog	Lazaridis et al 2014	3
Africa	Tanzania	Hadza	Fan 2019	2
Africa	Tanzania	Hadza	Lazaridis et al 2014	14
Africa	Tanzania	Hadza Henn	Lazaridis et al 2014	3
Africa	Tanzania	Iraqw	Fan 2019	2
Africa	Tanzania	Sandawe	Fan 2019	1
Africa	Tanzania	Sandawe	Lazaridis et al 2014	22
Africa	Tunisia	Tunisian	Lazaridis et al 2014	8
Africa	Tunisia	Tunisian Jew	Lazaridis et al 2014	7
Africa	Uganda	Buganda	Gurdasani et al 2015	96
Africa	Uganda	Uganda Muganda	Lopez 2021	6
Africa	Uganda	Uganda Mussese	Lopez 2021	6
CentralAsiaSiberia	Russia	Russian	Lazaridis et al 2014	22
EastAsia	China	Han	Lazaridis et al 2014	33
EastAsia	China	Han NChina	Lazaridis et al 2014	10
EastAsia	China	Mongola	Lazaridis et al 2014	6

EastAsia	Japan	Japanese	Lazaridis et al 2014	29
SouthAsia	Bangladesh	Bengali Bangladesh BEB	Lazaridis et al 2014	7
SouthAsia	India	Cochin Jew	Lazaridis et al 2014	5
SouthAsia	India	GujaratiA GIH	Lazaridis et al 2014	5
SouthAsia	India	GujaratiB GIH	Lazaridis et al 2014	5
SouthAsia	India	GujaratiC GIH	Lazaridis et al 2014	5
SouthAsia	India	GujaratiD GIH	Lazaridis et al 2014	5
SouthAsia	India	India Hindu	Lopez et al 2017	12
SouthAsia	India	India Zoroastrian	Lopez et al 2017	13
SouthAsia	India	Kharia	Lazaridis et al 2014	8
SouthAsia	India	Lodhi	Lazaridis et al 2014	13
SouthAsia	India	Mala	Lazaridis et al 2014	13
SouthAsia	India	Punjabi Lahore PJL	Lazaridis et al 2014	8
SouthAsia	India	Tiwari	Lazaridis et al 2014	14
SouthAsia	India	Vishwabrahmin	Lazaridis et al 2014	13
SouthAsia	Pakistan	Balochi	Lazaridis et al 2014	5
SouthAsia	Pakistan	Brahui	Lazaridis et al 2014	20
SouthAsia	Pakistan	Burusho	Lazaridis et al 2014	23
SouthAsia	Pakistan	Hazara	Lazaridis et al 2014	13
SouthAsia	Pakistan	Kalash	Lazaridis et al 2014	16
SouthAsia	Pakistan	Makrani	Lazaridis et al 2014	8
SouthAsia	Pakistan	Pathan	Lazaridis et al 2014	19
SouthAsia	Pakistan	Sindhi	Lazaridis et al 2014	18
WestEurasia	Albania	Albanian	Lazaridis et al 2014	6
WestEurasia	Armenia	Armenian	Lazaridis et al 2014	10
WestEurasia	Ashkenazi	Ashkenazi Jew	Lazaridis et al 2014	7
WestEurasia	Belarus	Belarusian	Lazaridis et al 2014	10
WestEurasia	Bulgaria	Bulgarian	Lazaridis et al 2014	9
WestEurasia	Croatia	Croatian	Lazaridis et al 2014	10
WestEurasia	Cyprus	Cypriot	Lazaridis et al 2014	8
WestEurasia	Czechoslovia	Czech	Lazaridis et al 2014	10
WestEurasia	England	English Cornwall GBR	Lazaridis et al 2014	5
WestEurasia	England	English Kent GBR	Lazaridis et al 2014	5
WestEurasia	Estonia	Estonian	Lazaridis et al 2014	10
WestEurasia	Finland	Finnish FIN	Lazaridis et al 2014	7
WestEurasia	France	French	Lazaridis et al 2014	25
WestEurasia	France	French South	Lazaridis et al 2014	7
WestEurasia	Georgia	Abkhasian	Lazaridis et al 2014	9
WestEurasia	Georgia	Georgian Jew	Lazaridis et al 2014	7
WestEurasia	Georgia	Georgian Megrels	Lazaridis et al 2014	10
WestEurasia	Greece	Greek Comas	Lazaridis et al 2014	14
WestEurasia	Greece	Greek Coriell	Lazaridis et al 2014	6
WestEurasia	Hungary	Hungarian Coriell	Lazaridis et al 2014	10
WestEurasia	Hungary	Hungarian Metspalu	Lazaridis et al 2014	10

WestEurasia	Iceland	Icelandic	Lazaridis et al 2014	12
WestEurasia	Iran	Iran Fars	Broushaki et al 2016	17
WestEurasia	Iran	Iran Zoroastrian	Broushaki et al 2016	27
WestEurasia	Iran	Iranian	Lazaridis et al 2014	8
WestEurasia	Iran	Iranian Jew	Lazaridis et al 2014	9
WestEurasia	Iraq	Iraqi Jew	Lazaridis et al 2014	6
WestEurasia	Israel	BedouinA	Lazaridis et al 2014	25
WestEurasia	Israel	BedouinB	Lazaridis et al 2014	19
WestEurasia	Israel	Druze	Lazaridis et al 2014	35
WestEurasia	Israel	Israeli Arabs	Lopez 2021	23
WestEurasia	Israel	IsraeliBedouins	Lopez 2021	6
WestEurasia	Israel	Palestinian	Lazaridis et al 2014	33
WestEurasia	Italy	Italian Bergamo	Lazaridis et al 2014	12
WestEurasia	Italy	Italian EastSicilian	Lazaridis et al 2014	5
WestEurasia	Italy	Italian Tuscan	Lazaridis et al 2014	8
WestEurasia	Italy	Italian WestSicilian	Lazaridis et al 2014	6
WestEurasia	Italy	Sardinian	Lazaridis et al 2014	27
WestEurasia	Jordan	Jordanian	Lazaridis et al 2014	4
WestEurasia	Lebanon	Lebanese	Lazaridis et al 2014	8
WestEurasia	Lithuania	Lithuanian	Lazaridis et al 2014	10
WestEurasia	Malta	Maltese	Lazaridis et al 2014	8
WestEurasia	Norway	Norway	Lazaridis et al 2014	11
WestEurasia	OrkneyIslands	Orcadian	Lazaridis et al 2014	12
WestEurasia	Palestine	PalestinianArabs	Lopez 2021	13
WestEurasia	Russia	Adygei	Lazaridis et al 2014	16
WestEurasia	Russia	Balkar	Lazaridis et al 2014	10
WestEurasia	Russia	Chechen	Lazaridis et al 2014	9
WestEurasia	Russia	Chuvash	Lazaridis et al 2014	10
WestEurasia	Russia	Kumyk	Lazaridis et al 2014	8
WestEurasia	Russia	Lezgin	Lazaridis et al 2014	9
WestEurasia	Russia	Mordovian	Lazaridis et al 2014	10
WestEurasia	Russia	Nogai	Lazaridis et al 2014	9
WestEurasia	Russia	North Ossetian	Lazaridis et al 2014	10
WestEurasia	Saudi Arabia	Saudi	Lazaridis et al 2014	8
WestEurasia	Scotland	Scottish Argyll Bute GBR	Lazaridis et al 2014	4
WestEurasia	Spain	Basque French	Lazaridis et al 2014	20
WestEurasia	Spain	Basque Spanish	Lazaridis et al 2014	9
WestEurasia	Spain	Spanish Andalucia IBS	Lazaridis et al 2014	4
WestEurasia	Spain	Spanish Aragon IBS	Lazaridis et al 2014	6
WestEurasia	Spain	Spanish Baleares IBS	Lazaridis et al 2014	4
WestEurasia	Spain	Spanish Cantabria IBS	Lazaridis et al 2014	5
WestEurasia	Spain	Spanish Castilla la Mancha IBS	Lazaridis et al 2014	5
WestEurasia	Spain	Spanish Castilla y Leon IBS	Lazaridis et al 2014	5
WestEurasia	Spain	Spanish Cataluna IBS	Lazaridis et al 2014	5

WestEurasia	Spain	Spanish Extremadura IBS	Lazaridis et al 2014	5
WestEurasia	Spain	Spanish Galicia IBS	Lazaridis et al 2014	5
WestEurasia	Spain	Spanish Murcia IBS	Lazaridis et al 2014	4
WestEurasia	Spain	Spanish Pais Vasco IBS	Lazaridis et al 2014	5
WestEurasia	Spain	Spanish Valencia IBS	Lazaridis et al 2014	5
WestEurasia	Syria	Syria	Lopez 2021	12
WestEurasia	Syria	Syrian	Lazaridis et al 2014	2
WestEurasia	Turkey	Turkish	Lazaridis et al 2014	4
WestEurasia	Turkey	Turkish Adana	Lazaridis et al 2014	10
WestEurasia	Turkey	Turkish Aydin	Lazaridis et al 2014	7
WestEurasia	Turkey	Turkish Balikesir	Lazaridis et al 2014	6
WestEurasia	Turkey	Turkish Istanbul	Lazaridis et al 2014	10
WestEurasia	Turkey	Turkish Jew	Lazaridis et al 2014	8
WestEurasia	Turkey	Turkish Kayseri	Lazaridis et al 2014	10
WestEurasia	Turkey	Turkish Trabzon	Lazaridis et al 2014	9
WestEurasia	Ukraine	Ukrainian East	Lazaridis et al 2014	6
WestEurasia	Ukraine	Ukrainian West	Lazaridis et al 2014	3
WestEurasia	Uzbekistan	Uzbek	Lazaridis et al 2014	10
WestEurasia	Yemen	Yemen	Lazaridis et al 2014	6
WestEurasia	Yemen	Yemenite Jew	Lazaridis et al 2014	8

Table A.3: Continent, Country, ethnicity, published study and number of individuals in each Human Origins population.

A.3.1 Processing

Only bi-allelic SNPs were retained. To ensure that all datasets, ancient and modern, can be merged together without the confounding effects of strand flips, I then used conform-gt (`https://faculty.washington.edu/browning/conform-gt.html`) to align all alleles to the same strand as the 1000 genomes reference, keeping all parameters as default. Any genotypes which had a genotype likelihood of below 0.990 were set as missing.

Data was phased use `shapeit4` [25], setting `-pbwt 8` and keeping all other parameters as default. The 1000 Genomes was used as as reference (section A.2). Sporadic low quality missing genotypes were imputed.

A.4 MS POBI HellBus dataset

Multiple Sclerosis (MS), People of the British Isles (POBI), Hellenthal and Busby (HB) / MS POBI HellBus contains a total of 14,795 individuals from 211 worldwide populations and genotyped at 477,417 autosomal bi-allelic SNPs.

Samples from Sawcer et al (2011) [221] (10299 individuals from 15 pops), Leslie et al 2015 [31] (2039 individuals from 35 pops) and Busby et al (2457 individuals from 161 pops).

Individuals from MS populations USA, Canada and New Zealand were all removed as the individuals were not native to that country.

The following steps were taken to process the data

1. Filtered such that SNPs with only 2 alleles were retained
2. Phase using shapeit4 [25] setting `-pbwt-depth 8`.
3. Remove duplicated SNPs using bcftools norm [220]
4. Use Beagle's conform-gt utility to ensure reference alleles were consistent with the previous 1000 genomes build. This was done because all previous datasets I have compiled were also conformed to the previous 1000 genomes build.

Table A.4: Populations and corresponding number of individuals for all populations in the 'MS POBI HellBus' dataset

Dataset	Population	Number of Individuals
HB	abhkasian	20
HB	adygei	17
HB	altai	13
HB	armenian	35
HB	balkar	19
HB	balochi	24
HB	bantukenya	11
HB	bantusouthafrica	8

Table A.4: Populations and corresponding number of individuals for all populations in the 'MS POBI HellBus' dataset *(continued)*

Dataset	Population	Number of Individuals
HB	basque	24
HB	bedouin	45
HB	belorussian	9
HB	bengali	1
HB	bhunjia	1
HB	biakapygmy	21
HB	brahmin	11
HB	brahui	25
HB	bulgarian	31
HB	burusho	25
HB	burya	2
HB	buryat	15
HB	cambodian	10
HB	ceu	59
HB	chamar	10
HB	chechen	20
HB	chenchu	4
HB	chukchi	5
HB	chuvash	17
HB	colombian	7
HB	croatian	19
HB	cypriot	12
HB	dai	10
HB	daur	9
HB	dharkar	8
HB	dhurwa	1
HB	dolgan	7
HB	druze	42
HB	dusadh	7
HB	egyptian	12
HB	english	8
HB	ethiopiana	7
HB	ethiopianjew	11
HB	ethiopiano	7
HB	ethiopiant	5
HB	evenk	12
HB	finnish	2
HB	french	28
HB	georgian	20
HB	german	30

Table A.4: Populations and corresponding number of individuals for all populations in the 'MS POBI HellBus' dataset *(continued)*

Dataset	Population	Number of Individuals
HB	germanyaustria	4
HB	gond	4
HB	greek	20
HB	hadza	3
HB	hakkipikki	3
HB	han	34
HB	hannchina	10
HB	hazara	22
HB	hezhen	8
HB	hungarian	19
HB	indian	1
HB	indianjew	8
HB	iranian	20
HB	irish	7
HB	japanese	28
HB	jordanian	20
HB	kalash	23
HB	kanjar	5
HB	karitiana	11
HB	karnataka	8
HB	ket	2
HB	kol	16
HB	koryake	5
HB	kshatriya	7
HB	kumyk	14
HB	kurd	6
HB	kurmi	1
HB	kurumba	4
HB	kyrgyz	16
HB	lahu	8
HB	lambadi	1
HB	lebanese	5
HB	lezgin	18
HB	lithuanian	10
HB	luhya	94
HB	maasai	97
HB	makrani	25
HB	malayan	1
HB	mandenka	22
HB	mawasi	1

Table A.4: Populations and corresponding number of individuals for all populations in the 'MS POBI HellBus' dataset *(continued)*

Dataset	Population	Number of Individuals
HB	maya	21
HB	mbutipygmy	13
HB	meena	1
HB	meghawal	1
HB	melanesian	10
HB	miao	10
HB	mongolian	19
HB	mordovian	15
HB	moroccan	25
HB	mozabite	29
HB	muslim	5
HB	myanmar	3
HB	naga	4
HB	naxi	8
HB	nganassan	10
HB	nihali	2
HB	nogay	16
HB	northitalian	12
HB	northossetian	15
HB	norwegian	18
HB	orcadian	15
HB	oroqen	9
HB	palestinian	46
HB	papuan	17
HB	pathan	22
HB	pima	14
HB	piramalaikallar	8
HB	polish	17
HB	romanian	16
HB	russian	25
HB	sakd	4
HB	sandawe	28
HB	sankhomani	30
HB	sannamibia	5
HB	sardinian	28
HB	saudi	19
HB	scottish	6
HB	selkup	10
HB	she	10
HB	siciliane	10

Table A.4: Populations and corresponding number of individuals for all populations in the 'MS POBI HellBus' dataset *(continued)*

Dataset	Population	Number of Individuals
HB	sindhi	24
HB	southitalian	18
HB	spanish	34
HB	surui	5
HB	syrian	16
HB	tajik	15
HB	tamilnadu	2
HB	tharus	2
HB	tsi	98
HB	tu	10
HB	tujia	10
HB	tunisian	12
HB	turkish	19
HB	turkishe	23
HB	turkishn	20
HB	turkishs	20
HB	turkmen	10
HB	tuscan	8
HB	tuva	13
HB	uae	14
HB	ukrainian	20
HB	upcaste	5
HB	uygur	10
HB	uzbekistani	15
HB	velamas	9
HB	welsh	4
HB	westsicilian	10
HB	xibo	9
HB	yakut	25
HB	yemeni	9
HB	yi	10
HB	yoruba	21
HB	yukagir	4
MS	Belgium	544
MS	Denmark	332
MS	Finland	581
MS	France	479
MS	Germany	1100
MS	Italy	745
MS	NIreland	61

Table A.4: Populations and corresponding number of individuals for all populations in the 'MS POBI HellBus' dataset *(continued)*

Dataset	Population	Number of Individuals
MS	Norway	953
MS	Poland	58
MS	Spain	205
MS	Sweden	1212
MS	UK	1854
POBI	UK	2039

A breakdown of the POBI populations:

Table A.5: Counties and corresponding number of individuals for all counties in the POBI dataset

County	Number of Individuals
Cheshire	33
Cornwall and Isles of Scilly	90
Cumbria	195
Devon	73
Dorset	37
Dumfries and Galloway	42
Durham	54
Dyfed	55
East Riding of Yorkshire Unitary Authority	32
East Sussex	34
Fife	59
Gloucestershire	70
Gwent	31
Gwynedd	76
Hampshire	26
Kent	50
Leicestershire	66
Lincolnshire	104
Merseyside	47
Norfolk	98
North Yorkshire	64
Northamptonshire	37
Northern Ireland	44

Table A.5: Counties and corresponding number of individuals for all counties in the POBI dataset *(continued)*

County	Number of Individuals
Northumberland	50
Nottinghamshire	57
Orkney Islands	96
Oxfordshire	77
Somerset	17
South Yorkshire	77
Staffordshire	28
Suffolk	82
Surrey	24
Tyne and Wear	54
West Sussex	26
Worcestershire	34

Appendix B

Some commonly used terms and their motivation for use

Here are some terms I commonly use.

B.1 'all-v-all'

I use this term when painting each individual in turn is painted using all other individuals as donors. If there are N individuals, the result is an NxN coancestry matrix.

B.2 'Leave-one-out'

Consider a situation where an all-v-all painting is performed on a set of individuals grouped into populations, where 2 of the populations are *Devon* and *Cornwall*. We would like to estimate the proportion of genome each recipient individual matches to both *Devon* and *Cornwall*, so we take the sums across columns, aggregating them by population. However, this means that each individual from, for example, Cornwall, can match to one less individual from Cornwall than other populations, as they cannot paint themselves. To avoid this, we may perform a 'leave-one-out' painting, where each population

is painted separately, and a single individual from each other population is removed from the set of donors.

B.3 Total Variation Distance

Often we would like to estimate how similar the copyvectors, C_x and C_y of two individuals or populations (average) are to one another. Given copvectors are the same length, one way would be to simply estimate Pearsons' correlation. However, this can lead to misleading results because Pearson's r-squared is often over-sensitive to outlying values.

An alternative is to estimate TVD, where $TVD_{x,y}$ between two copyvectors C_x and C_y is given as $TVD = \sum |C_x - C_y|$.

Appendix C

Colophon

This document was produced using the UCL thesis LaTeX template (`https://github.com/UCL/ucl-latex-thesis-templates`). This document was set in the lmodern typeface using LaTeX and BibTeX, composed with a text TexMaker on Linux. `microtype` was also used. All figures were generated using `ggplot2` using `theme_light()`. All tables were generated using the `kbl` function from the `kableExtra` R library The final version of the thesis can be found at `https://github.com/sahwa/thesis`.

Appendix D

Supplementary figures

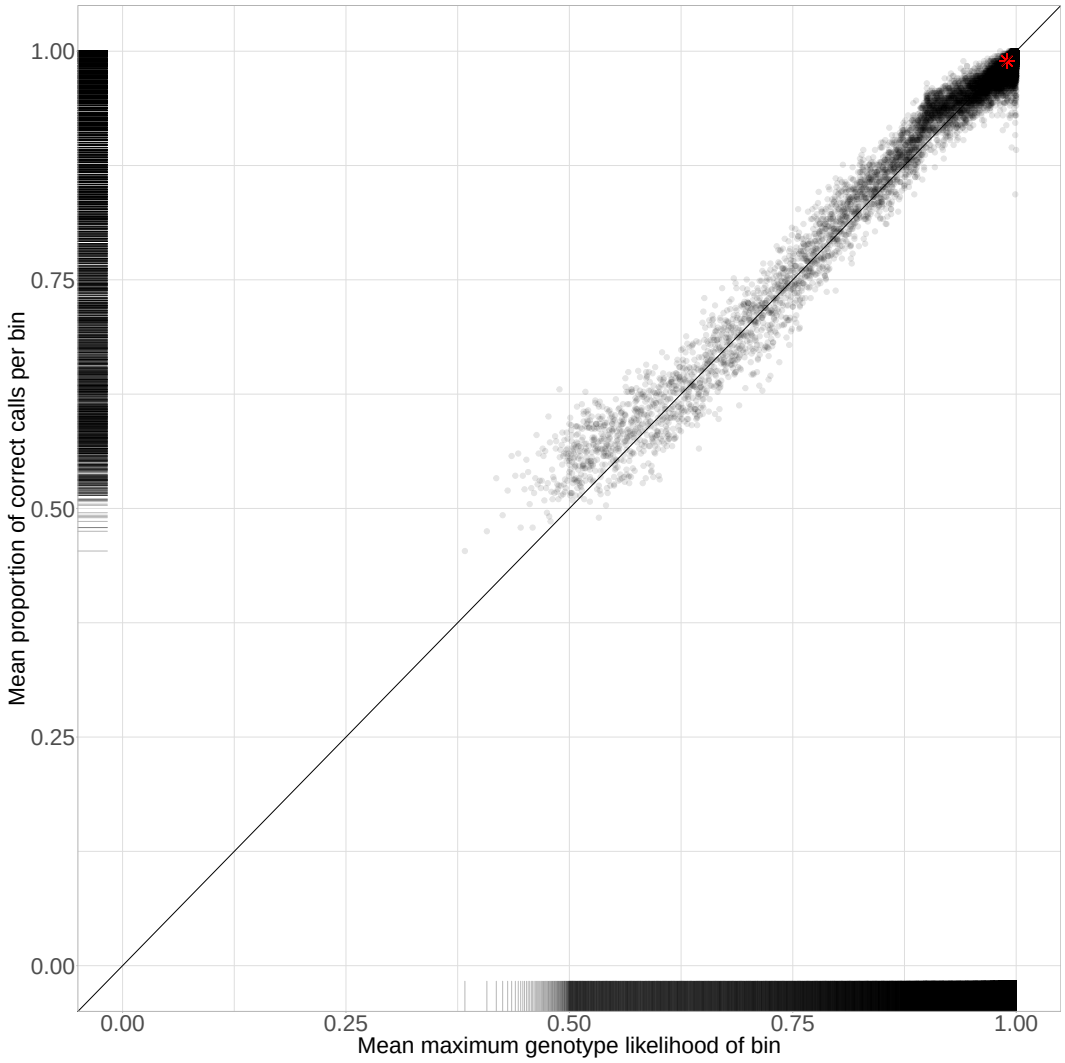

Figure D.1: Relationship between genotype likelihood and probability of genotype call being correct for UstIshim downsampled to 0.1x coverage. Genome binned by maximum posterior genotype likelihood and mean maximum posterior genotype likelihood (x-axis) and proportion of correct calls per bin (y-axis). Rugs on each margin show the distribution of x and y values. Black line is $y = x$.

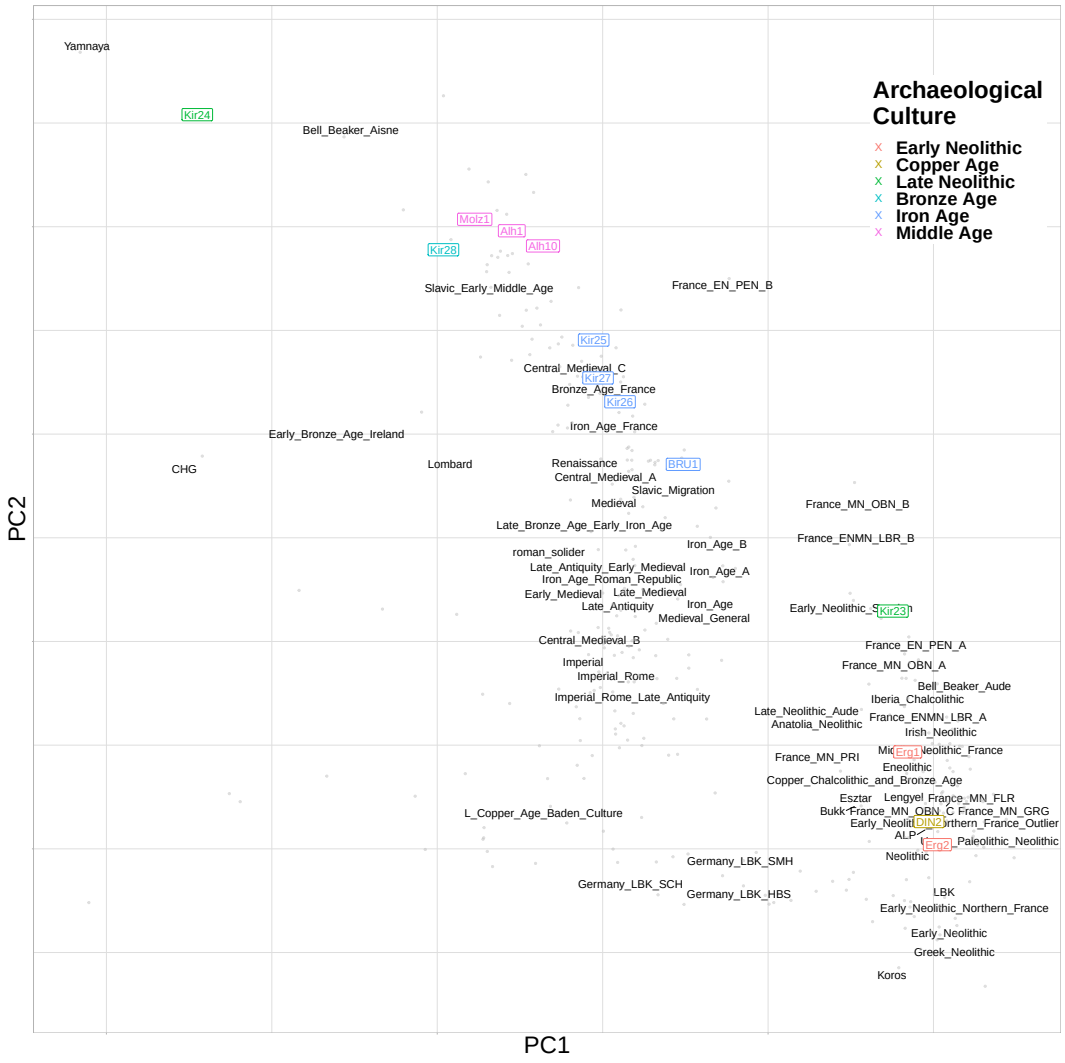

Figure D.2: Principle component analysis of genotype matrix using plink2. Grey points indicate principle component coordiantes for each sample. Black text indicated mean principle component coordinates for all individuals within that group. Coloured labels represent newly sequenced ancient samples.

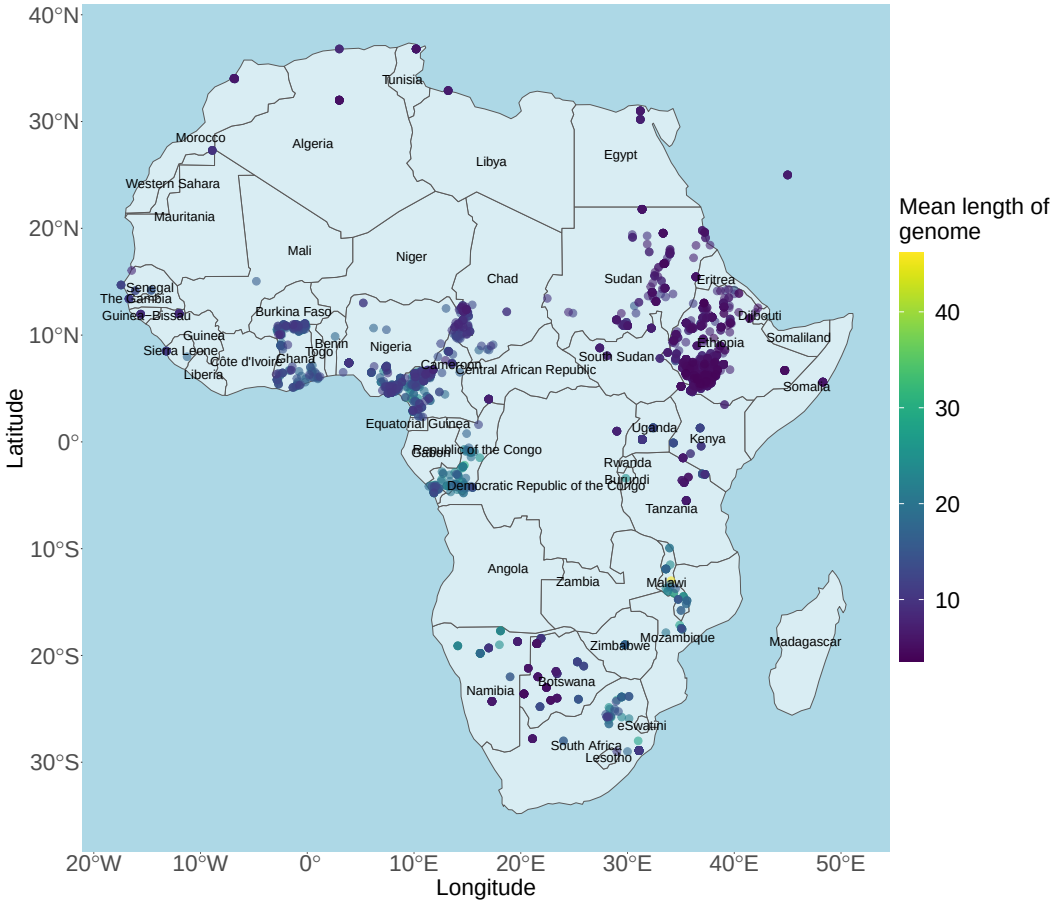

Figure D.3: Map of haplotype donation to U.K. Biobank individuals born in Brazil. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in Brazil.

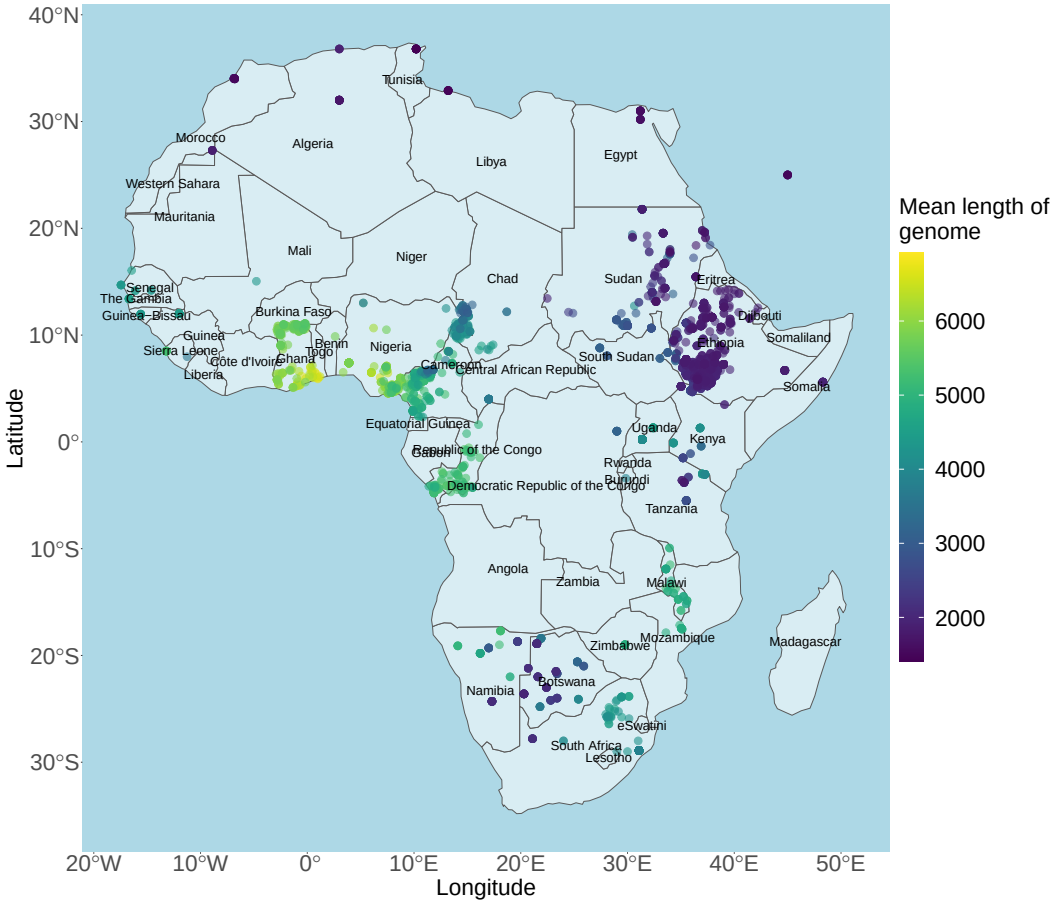

Figure D.4: Map of haplotype donation to U.K. Biobank individuals born in the Caribbean. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in the Caribbean.

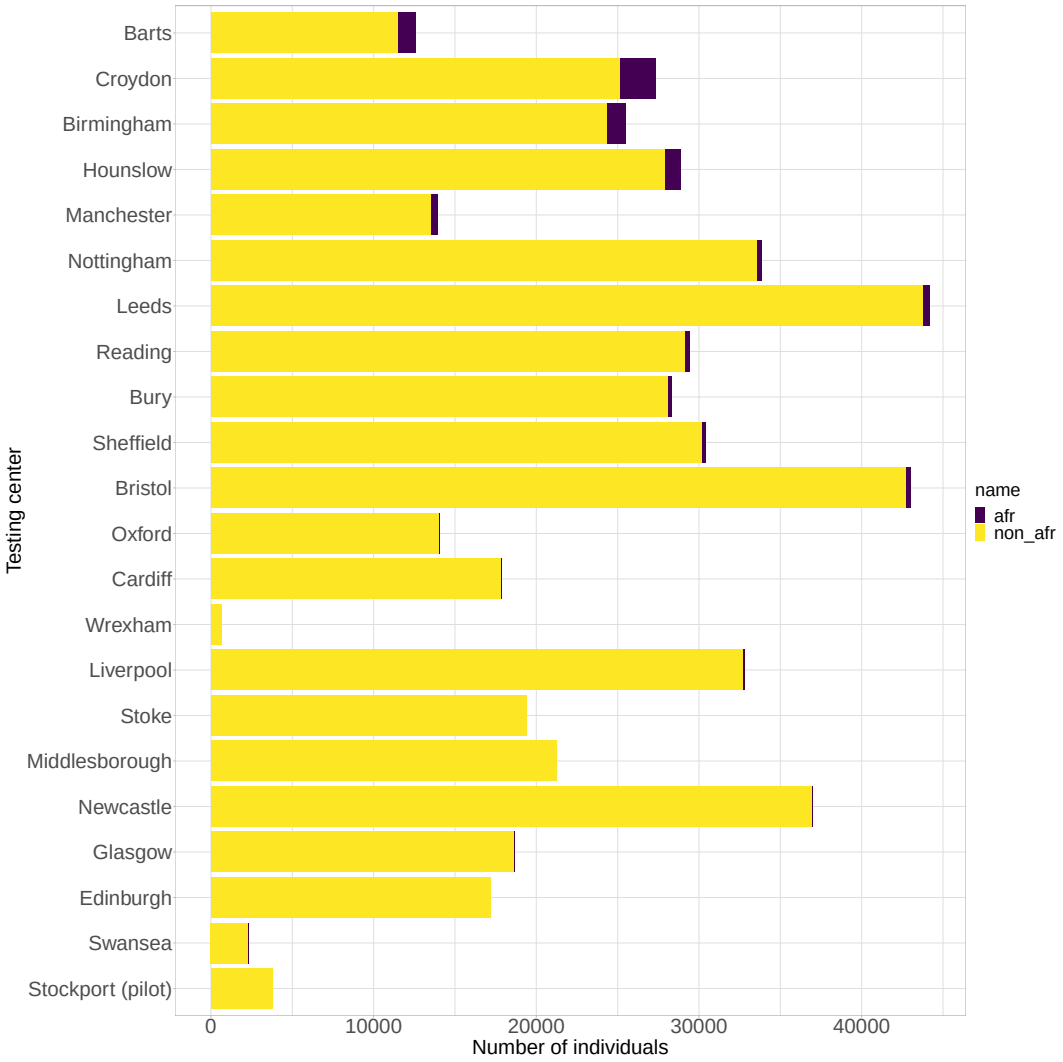

Figure D.5: Number of individuals who have (purple) and have not (yellow) at least 50% African ancestry (purple) by different testing centers. Centers ordered by proportion of individuals who have at least 50% African ancestry.

Figure D.6: Principle Component Analysis on chunklengths matrix of newly sequenced ancient Bavarian samples and selected ancient literature samples.

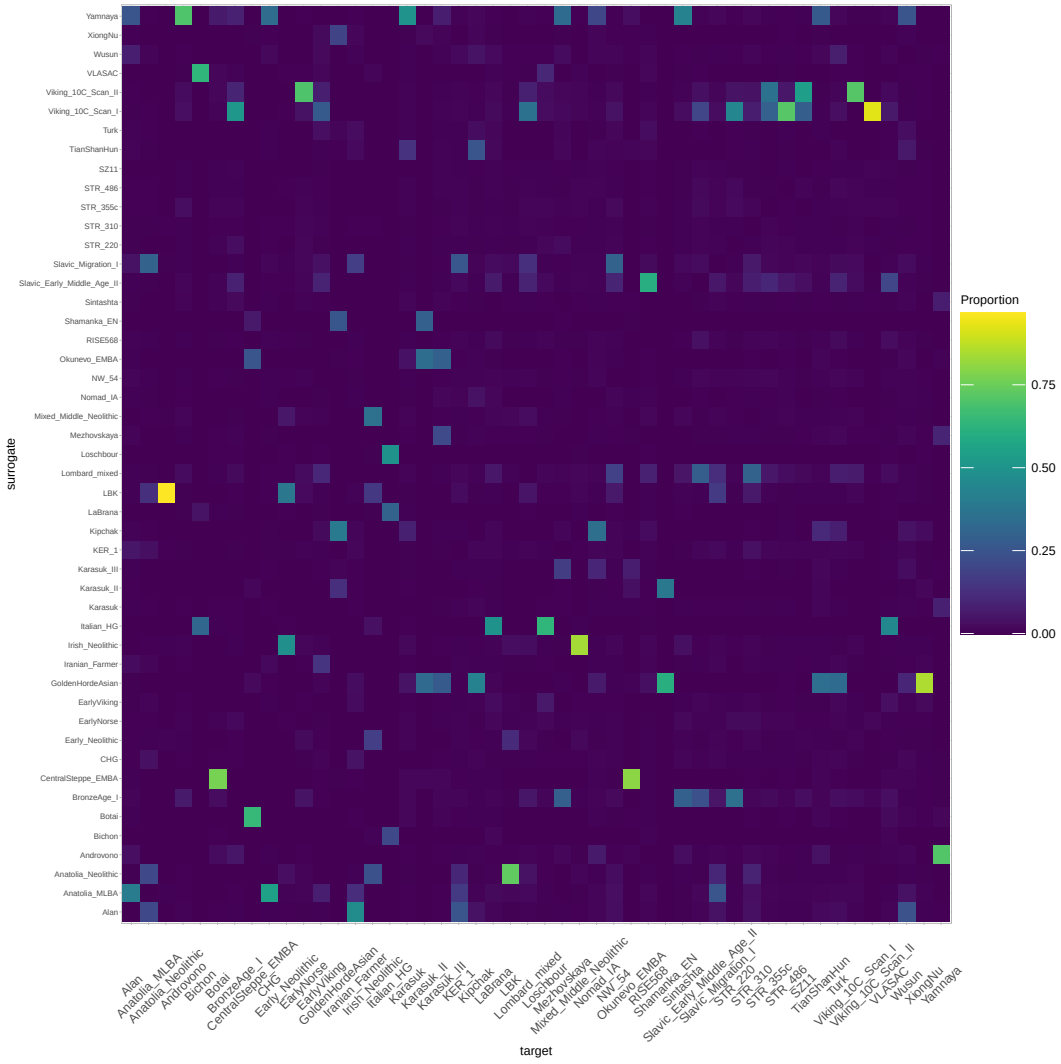

Figure D.7: Heatmap of SOURCEFIND proportions for all clusters in the ancient Slav analysis.

Appendix E

Supplementary results

Auxiliary results.

E.0.1 Determining the number of MCMC iterations required in SOURCEFIND analysis

SOURCEFIND is a haplotype-based method for inferring ancestry. At its heart, SOURCEFIND uses Markov chain Monte Carlo sampling to explore the parameter space of ancestry proportions. As is the case with any method that uses MCMC sampling, it is important to ensure that enough iterations have been performed; if this is not the case, the algorithm may not converge.

To determine what is the minimum number of iterations, I ran SOURCEFIND for 7 different numbers of iterations and 10 runs for each number. Results are presented in Figure E.1. Visually inspecting the results shows that using 50,000 iterations or less leads to variable results. 500,000 iterations appears to be the best balance between running time and accuracy.

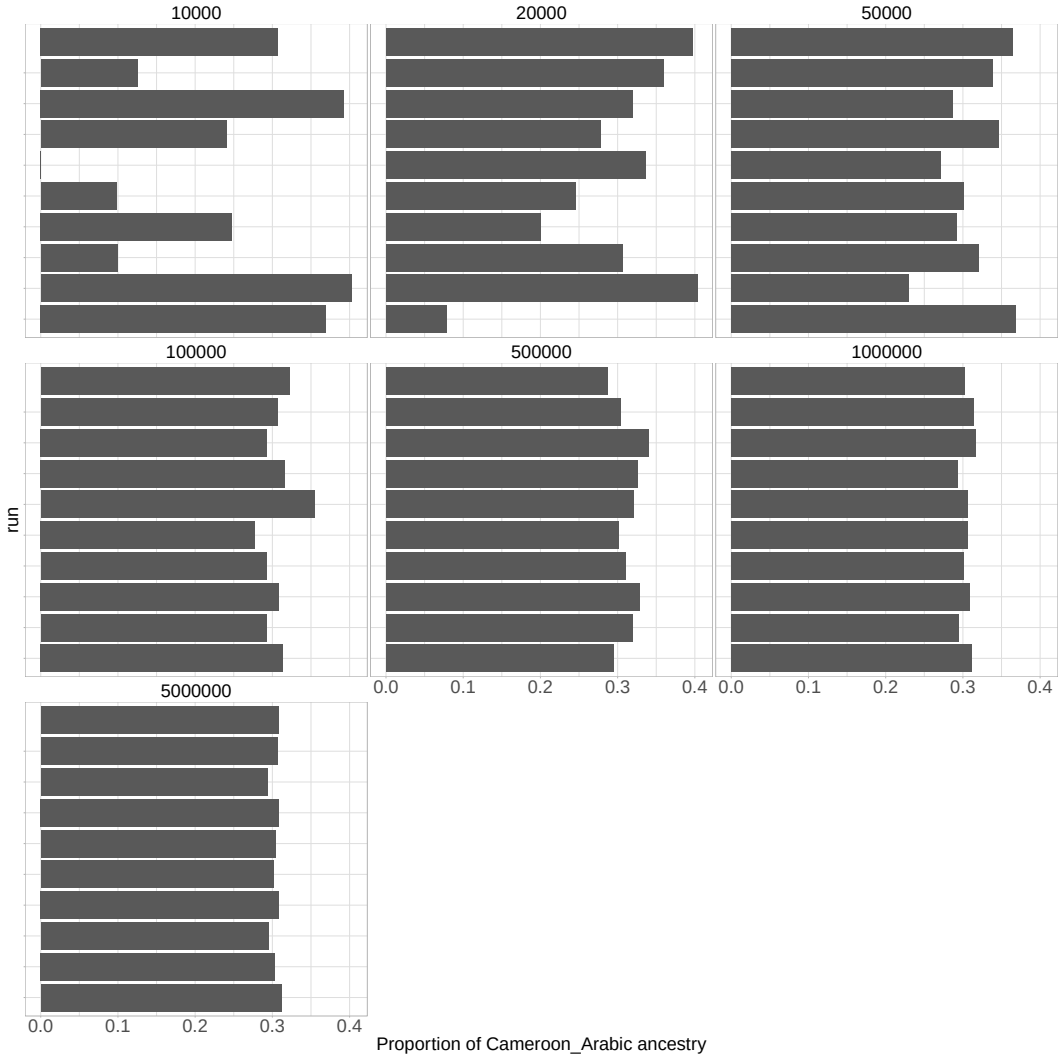

Figure E.1: Proportion of inferred Cameroon Arabic ancestry averaged across individuals from Cameroon Kanuri ethnic group. Each panel contains proportions for a different number of MCMC iterations. Within each panel, each bar is the proportion inferred from each of the 10 independent SOURCEFIND runs.

E.0.2 Determining the number of SNPs required to separate individuals from Devon and Cornwall

This figure shows the how TVD assignment accuracy varies with the total number of SNPs included.

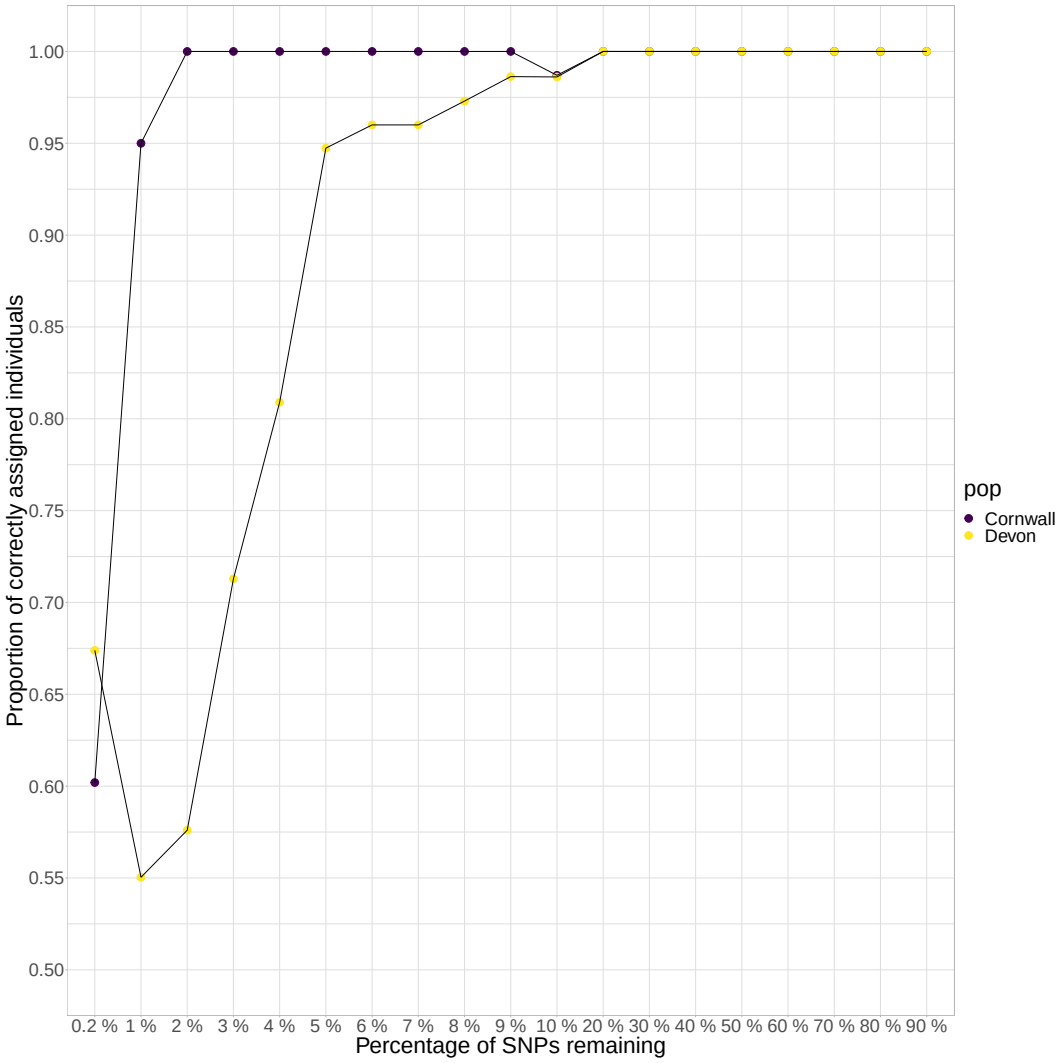

Figure E.2: Proportion of individuals correctly assigned (via TVD) to their correct population (y-axis) using different number of SNPs (x-axis).

Bibliography

[1] Daniel P. Cooke, David C. Wedge, and Gerton Lunter. A unified haplotype-based method for accurate and comprehensive variant calling. *Nature Biotechnology*, 2021.

[2] Thomas Hunt Morgan. Complete linkage in the second chromosome of the male of Drosophila. *Science*, 36(934):719–720, 1912.

[3] William Bateson and Edith Rebecca Saunders. *Experiments [in the Physiology of Heredity]*. Harrison, 1902.

[4] Montgomery Slatkin. Linkage disequilibrium—understanding the evolutionary past and mapping the medical future. *Nature Reviews Genetics*, 9(6):477–485, 2008.

[5] Bill Amos, Christian Schlotterer, and Diethard Tautz. Social structure of pilot whales revealed by analytical DNA profiling. *Science*, 260(5108):670–672, 1993.

[6] Sarah A Tishkoff, Erin Dietzsch, William Speed, et al. Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. *Science*, 271(5254):1380–1387, 1996.

[7] Richard A Gibbs, John W Belmont, Paul Hardenbol, et al. The international HapMap project. 2003.

[8] Dana C Crawford, Tushar Bhangale, Na Li, et al. Evidence for substantial fine-scale variation in recombination rates across the human genome. *Nature genetics*, 36(7):700–706, 2004.

[9] David M Evans and Lon R Cardon. A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *The American Journal of Human Genetics*, 76(4):681–687, 2005.

[10] David E Reich, Michele Cargill, Stacey Bolk, et al. Linkage disequilibrium in the human genome. *Nature*, 411(6834):199–204, 2001.

[11] Donald F Conrad, Mattias Jakobsson, Graham Coop, et al. A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature genetics*, 38(11):1251–1260, 2006.

[12] Rebecca L Cann, Mark Stoneking, and Allan C Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(6099):31–36, 1987.

[13] Na Li and Matthew Stephens. Modeling Linkage Disequilibrium and Identifying Recombination Hotspots Using Single-Nucleotide Polymorphism Data. *Genetics*, 165(4):2213–2233, 2003.

[14] Yun S Song. Na Li and Matthew Stephens on modeling linkage disequilibrium. *Genetics*, 203(3):1005–1006, 2016.

[15] Matthew Stephens and Peter Donnelly. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *The American Journal of Human Genetics*, 73(5):1162–1169, 2003.

[16] Matthew Stephens and Paul Scheet. Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation. *The American Journal of Human Genetics*, 76(3):449–462, 2005.

[17] Garrett Hellenthal, Adam Auton, and Daniel Falush. Inferring human colonization history using a copying model. *PLoS genetics*, 4(5):e1000078, 2008.

[18] Mattias Jakobsson, Sonja W Scholz, Paul Scheet, et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003, 2008.

[19] 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.

[20] Garrett Hellenthal, George B.J. J. Busby, Gavin Band, et al. A Genetic Atlas of Human Admixture History. *Science*, 343(6172):747–751, 2014.

[21] Juan C. Chacon-Duque, Kaustubh Adhikari, Macarena Fuentes-Guajardo, et al. Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. *Nature Communications*, page 252155, 2018.

[22] Michael Burrows and David Wheeler. A block-sorting lossless data compression algorithm. In *Digital SRC Research Report*. Citeseer, 1994.

[23] Richard Durbin. Efficient haplotype matching and storage using the positional Burrows–Wheeler transform (PBWT). *Bioinformatics*, 30(9):1266–1272, 01 2014.

[24] Heng Li and Richard Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 05 2009.

[25] Olivier Delaneau, Jean-François Zagury, Matthew R Robinson, et al. Accurate, scalable and integrative haplotype estimation. *Nature communications*, 10(1):1–10, 2019.

[26] Ardalan Naseri, Degui Zhi, and Shaojie Zhang. Multi-allelic positional Burrows-Wheeler transform. *BMC bioinformatics*, 20(11):1–8, 2019.

[27] Ross P Byrne, Wouter van Rheenen, Leonard H van den Berg, et al. Dutch population structure across space, time and GWAS design. *Nature communications*, 11(1):1–11, 2020.

[28] Juba Nait Saada, Georgios Kalantzis, Derek Shyr, et al. Identity-by-descent detection across 487,409 British samples reveals fine scale population structure and ultra-rare variant associations. *Nature Communications*, 11(1):1–15, 2020.

[29] Ardalan Naseri, Xiaoming Liu, Kecong Tang, et al. RaPID: ultra-fast, powerful, and accurate detection of segments identical by descent (IBD) in biobank-scale cohorts. *Genome biology*, 20(1):1–15, 2019.

[30] 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.

[31] Stephen Leslie, Bruce Winney, Garrett Hellenthal, et al. The fine-scale genetic structure of the British population. *Nature*, 519(7543):309–314, 2015.

[32] Lucie M Gattepaille and Mattias Jakobsson. Combining Markers into Haplotypes Can Improve Population Structure Inference. *Genetics*, 190(1):159–174, 01 2012.

[33] Anders Bergström, Shane A. McCarthy, Ruoyun Hui, et al. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484):eaay5012, 2020.

[34] Noah A Rosenberg, Jonathan K Pritchard, James L Weber, et al. Genetic structure of human populations. *science*, 298(5602):2381–2385, 2002.

[35] Sohini Ramachandran, Omkar Deshpande, Charles C Roseman, et al. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proceedings of the National Academy of Sciences*, 102(44):15942–15947, 2005.

[36] Anne M Bowcock, Andres Ruiz-Linares, James Tomfohrde, et al. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455–457, 1994.

[37] Stephan Schiffels, Wolfgang Haak, Pirita Paajanen, et al. Iron age and Anglo-Saxon genomes from East England reveal British migration history. *Nature communications*, 7(1):1–9, 2016.

[38] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

[39] Timothy D O'Connor, Wenqing Fu, NHLBI GO Exome Sequencing Project, et al. Rare variation facilitates inferences of fine-scale population structure in humans. *Molecular biology and evolution*, 32(3):653–660, 2015.

[40] 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.

[41] Richard E Green, Johannes Krause, Adrian W Briggs, et al. A draft sequence of the Neandertal genome. *science*, 328(5979):710–722, 2010.

[42] Nick Patterson, Priya Moorjani, Yontao Luo, et al. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

[43] Benjamin M Peter. Admixture, population structure, and F-statistics. *Genetics*, 202(4):1485–1501, 2016.

[44] Éadaoin Harney, Nick Patterson, David Reich, and John Wakeley. Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*, 217(4), 01 2021. iyaa045.

[45] Alkes L Price, Nick J Patterson, Robert M Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[46] Kendra A Sirak, Daniel M Fernandes, Mark Lipson, et al. Social stratification without genetic differentiation at the site of Kulubnarti in Christian Period Nubia. *bioRxiv*, 2021.

[47] Torsten Günther and Carl Nettelblad. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, 15(7):1–20, 07 2019.

[48] Rui Martiniano, Lara M. Cassidy, Ros Ó'Maoldúin, et al. The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genetics*, 13(7):1–24, 2017.

[49] Rui Martiniano, Erik Garrison, Eppie R Jones, et al. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome biology*, 21(1):1–18, 2020.

[50] Erik Garrison, Jouni Sirén, Adam M. Novak, et al. Variation graph toolkit improves read mapping by representing genetic variation in the reference, 2018.

[51] Iosif Lazaridis, Nick Patterson, Alissa Mittnik, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

[52] Sewall Wright. The genetical structure of populations. *Annals of eugenics*, 15(1):323–354, 1949.

[53] David Reich, Kumarasamy Thangaraj, Nick Patterson, et al. Reconstructing Indian population history. *Nature*, 461(7263):489–494, 2009.

[54] Olivier François and Flora Jay. Factor analysis of ancient population genomic samples. *Nature communications*, 11(1):1–11, 2020.

[55] Salvador Herrando-Pérez, Raymond Tobler, and Christian D Huber. smartsnp, an r package for fast multivariate analyses of big genomic data. *Methods in Ecology and Evolution*, 2021.

[56] Jonas Meisner, Siyang Liu, Mingxi Huang, and Anders Albrechtsen. Large-scale inference of population structure in presence of missingness using PCA. *Bioinformatics*, 37(13):1868–1875, 01 2021.

[57] Farnaz Broushaki, Mark G Thomas, Vivian Link, et al. eastern Fertile Crescent. *Science*, 353(6298):499–503, 2016.

[58] Ashot Margaryan, Daniel J Lawson, Martin Sikora, et al. Population genomics of the Viking world. *Nature*, 585(7825):390–396, 2020.

[59] Margaret L Antonio, Ziyue Gao, Hannah M Moots, et al. Ancient Rome: a genetic crossroads of Europe and the Mediterranean. *Science*, 366(6466):708–714, 2019.

[60] Guy S. Jacobs, Georgi Hudjashov, Lauri Saag, et al. Multiple Deeply Divergent Denisovan Ancestries in Papuans. *Cell*, 177(4):1010–1021.e32, 2019.

[61] João C Teixeira, Guy S Jacobs, Chris Stringer, et al. Widespread Denisovan ancestry in Island Southeast Asia but no evidence of substantial super-archaic hominin admixture. *Nature Ecology & Evolution*, 5(5):616–624, 2021.

[62] Yoshan Moodley, Andrea Brunelli, Silvia Ghirotto, et al. Helicobacter pylori's historical journey through Siberia and the Americas. *Proceedings of the National Academy of Sciences*, 118(25), 2021.

[63] Ruoyun Hui, Eugenia D'Atanasio, Lara M Cassidy, et al. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Scientific reports*, 10(1):1–8, 2020.

[64] Kristiina Ausmees, Federico Sanchez-Quinto, Mattias Jakobsson, and Carl Nettelblad. An Empirical Evaluation of Genotype Imputation of Ancient DNA. Technical Report 2019-008, Department of Information Technology, Uppsala University, October 2019.

[65] Aaron McKenna, Matthew Hanna, Eric Banks, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 2010.

[66] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 2013.

[67] Heng Li, Bob Handsaker, Alec Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009.

[68] Ruiqiang Li, Yingrui Li, Xiaodong Fang, et al. SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 2009.

[69] Su Y. Kim, Kirk E. Lohmueller, Anders Albrechtsen, et al. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, 2011.

[70] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics*, 15(1):1–13, 2014.

[71] Vivian Link, Athanasios Kousathanas, Krishna Veeramah, et al. ATLAS: Analysis Tools for Low-depth and Ancient Samples. *bioRxiv*, page 105346, 2017.

[72] Robert W. Davies, Jonathan Flint, Simon Myers, and Richard Mott. Rapid genotype imputation from sequence without reference panels. *Nature Genetics*, 48(8):965–969, 2016.

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

[74] Line Skotte, Thorfinn Sand Korneliussen, and Anders Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, 2013.

[75] Miao Zhang, Yiwen Liu, Hua Zhou, et al. A novel nonlinear dimension reduction approach to infer population structure for low-coverage sequencing data. *BMC bioinformatics*, 22(1):1–13, 2021.

[76] Daniel Fernandes, Kendra Sirak, Mario Novak, et al. The Identification of a 1916 Irish Rebel: new approach for estimating relatedness from low coverage homozygous genomes. *Scientific reports*, 7(1):1–10, 2017.

[77] Daniel M Fernandes, Olivia Cheronet, Pere Gelabert, and Ron Pinhasi. TKGWV2: An ancient DNA relatedness pipeline for ultra-low coverage whole genome shotgun data. *Nature Communications*, 2021.

[78] Fernando Racimo, Gabriel Renaud, and Montgomery Slatkin. Joint Estimation of Contamination, Error and Demography for Nuclear DNA from Ancient Humans. *PLoS Genetics*, 2016.

[79] Joshua G. Schraiber. Assessing the relationship of ancient and modern populations. *Genetics*, 2018.

[80] Filipe G. Vieira, Anders Albrechtsen, and Rasmus Nielsen. Estimating IBD tracts from low coverage NGS data. *Bioinformatics*, 2016.

[81] Stéphane Peyrégne and Kay Prüfer. Present-Day DNA Contamination in Ancient DNA Datasets. *BioEssays*, 42(9):2000081, 2020.

[82] Jeffrey D Wall and Sung K Kim. Inconsistencies in Neanderthal genomic DNA sequences. *PLoS Genetics*, 3(10):e175, 2007.

[83] Richard E Green, Adrian W Briggs, Johannes Krause, et al. The Neandertal genome and ancient DNA authenticity. *The EMBO journal*, 28(17):2494–2502, 2009.

[84] Nathan Nakatsuka, Éadaoin Harney, Swapan Mallick, et al. ContamLD: estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium. *Genome biology*, 21(1):1–22, 2020.

[85] Matthias Meyer, Martin Kircher, Marie-theres Gansauge, et al. A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual A High-Coverage Genome Sequence from an Archaic Denisovan Individual. *Science (New York, NY)*, 222(2012):1–14, 2012.

[86] Svante Pääbo. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proceedings of the National Academy of Sciences*, 86(6):1939–1943, 1989.

[87] S Paabo. Miocene DNA sequence-a dream come true? *Curr. Biol.*, 1:45–46, 1991.

[88] Cesare de Filippo, Matthias Meyer, and Kay Prüfer. Quantifying and reducing spurious alignments for the analysis of ultra-short ancient DNA sequences. *BMC biology*, 16(1):1–11, 2018.

[89] Jesse Dabney, Matthias Meyer, and Svante Pääbo. Ancient DNA damage. *Cold Spring Harbor perspectives in biology*, 5(7):a012567, 2013.

[90] Brian L. Browning and Sharon R. Browning. Genotype Imputation with Millions of Reference Samples. *American Journal of Human Genetics*, 98(1):116–126, 2016.

[91] Simone Rubinacci, Diogo M Ribeiro, Robin J Hofmeister, and Olivier Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1):120–126, 2021.

[92] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 1975.

[93] Peter de Barros Damgaard, Rui Martiniano, Jack Kamm, et al. The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science*, 360(6396), 2018.

[94] Qiaomei Fu, Heng Li, Priya Moorjani, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 2014.

[95] Torsten Günther, Helena Malmström, Emma M. Svensson, et al. Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biology*, 2018.

[96] Broad Institute. Picard Tools. `http://broadinstitute.github.io/picard/`, 2018. Accessed: 2018-MM-DD; version X.Y.Z.

[97] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[98] Rasmus Nielsen, Joshua S Paul, Anders Albrechtsen, and Yun S Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443–451, 2011.

[99] Lucy Huang, Yun Li, Andrew B. Singleton, et al. Genotype-Imputation Accuracy across Worldwide Human Populations. *The American Journal of Human Genetics*, 84(2):235–250, 2009.

[100] Shane McCarthy, Sayantan Das, Warren Kretzschmar, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics*, 48(10):1279, 2016.

[101] Sharon R. Browning and Brian L. Browning. Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering. *The American Journal of Human Genetics*, 81(5):1084–1097, 2007.

[102] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, et al. High coverage whole genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *bioRxiv*, 2021.

[103] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[104] John G Cleary, Ross Braithwaite, Kurt Gaastra, et al. Joint variant and de novo mutation identification on pedigrees from high-throughput sequencing data. *Journal of Computational Biology*, 21(6):405–419, 2014.

[105] Wolfgang Haak, Iosif Lazaridis, Nick Patterson, et al. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, 2015.

[106] Martyn Plummer, Nicky Best, Kate Cowles, and Karen Vines. CODA: convergence diagnosis and output analysis for MCMC. *R News*, 6(1):7–11, March 2006.

[107] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, 2006.

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

[109] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nature genetics*, 48(11):1443–1448, 2016.

[110] W Haak, P Forster, B Bramanti, et al. Ancient DNA from the first European farmer in 750-year-old Neolithic sites. *Science*, 310(November):1016–1019, 2005.

[111] Kay Prüfer, Fernando Racimo, Nick Patterson, et al. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, 2014.

[112] Sharon R Browning and Brian L Browning. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. *The American Journal of Human Genetics*, 97(3):404–418, 2015.

[113] Augustine Kong, Gisli Masson, Michael L Frigge, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008.

[114] Laurent Excoffier and Stefan Schneider. Why hunter-gatherer populations do not show signs of Pleistocene demographic expansions. *Proceedings of the National Academy of Sciences*, 96(19):10597–10602, 1999.

[115] Qiaomei Fu, Cosimo Posth, Mateja Hajdinjak, et al. The genetic history of Ice Age Europe. *Nature*, 534(7606):200–205, 2016.

[116] Filipe G Vieira, Anders Albrechtsen, and Rasmus Nielsen. Estimating IBD tracts from low coverage NGS data. *Bioinformatics*, 32(14):2096–2102, 2016.

[117] Clare Bycroft, Colin Freeman, Desislava Petkova, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018.

[118] 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.

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

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

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

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

[123] Jacklyn N Hellwege, Jacob M Keaton, Ayush Giri, et al. Population stratification in genetic association studies. *Current protocols in human genetics*, 95(1):1–22, 2017.

[124] Karoline Kuchenbaecker, Nikita Telkar, Theresa Reiker, et al. The transferability of lipid loci across African, Asian and European cohorts. *Nature communications*, 10(1):1–10, 2019.

[125] Alicia R Martin, Christopher R Gignoux, Raymond K Walters, et al. Human demographic history impacts genetic risk prediction across diverse populations. *The American Journal of Human Genetics*, 100(4):635–649, 2017.

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

[127] Bjarni J Vilhjálmsson, Jian Yang, Hilary K Finucane, et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *The american journal of human genetics*, 97(4):576–592, 2015.

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

[129] 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.

[130] Saioa López, Ayele Tarekegn, Gavin Band, et al. Evidence of the interplay of genetics and culture in Ethiopia. *Nature communications*, 12(1):1–15, 2021.

[131] 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.

[132] Deepti Gurdasani, Tommy Carstensen, Segun Fatumo, et al. Uganda genome resource enables insights into population history and genomic discovery in Africa. *Cell*, 179(4):984–1002, 2019.

[133] Roseann E Peterson, Karoline Kuchenbaecker, Raymond K Walters, et al. Genome-wide association studies in ancestrally diverse populations: opportunities, methods, pitfalls, and recommendations. *Cell*, 179(3):589–603, 2019.

[134] Daniel Taliun, Daniel N Harris, Michael D Kessler, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature*, 590(7845):290–299, 2021.

[135] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, 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.

[136] Lucy Huang, Mattias Jakobsson, Trevor J Pemberton, et al. Haplotype variation and genotype imputation in African populations. *Genetic epidemiology*, 35(8):766–780, 2011.

[137] Roman Shraga, Sarah Yarnall, Sonya Elango, et al. Evaluating genetic ancestry and self-reported ethnicity in the context of carrier screening. *BMC genetics*, 18(1):1–9, 2017.

[138] Y. V. Louwers, O. Lao, B. C. J. M. Fauser, et al. 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.

[139] Elena Bosch, Hafid Laayouni, Carlos Morcillo-Suarez, 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.

[140] Michael Banton. Recent Migration from West Africa and the West Indies to the United Kingdom. *Population Studies*, 7(1):2–13, 1953.

[141] Steven J Micheletti, Kasia Bryc, Samantha G Ancona Esselmann, et al. Genetic consequences of the transatlantic slave trade in the Americas. *The American Journal of Human Genetics*, 107(2):265–277, 2020.

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

[143] Lucy Van Dorp, Sara Lowes, Jonathan L Weigel, 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.

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

[145] Nicholas J Conard. A female figurine from the basal Aurignacian of Hohle Fels Cave in southwestern Germany. *Nature*, 459(7244):248–252, 2009.

[146] Nicholas J Conard, Maria Malina, and Susanne C Münzel. New flutes document the earliest musical tradition in southwestern Germany. *Nature*, 460(7256):737–740, 2009.

[147] Mark Lipson, Anna Szécsényi-Nagy, Swapan Mallick, et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, 551(7680):368–372, 2017.

[148] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.

[149] Torsten Günther, Cristina Valdiosera, Helena Malmström, et al. Ancient genomes link early farmers from Atapuerca in Spain to modern-day Basques. *Proceedings of the National Academy of Sciences*, 112(38):11917–11922, 2015.

[150] Zuzana Hofmanová, Susanne Kreutzer, Garrett Hellenthal, et al. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences*, 2016.

[151] Morten E. Allentoft, Martin Sikora, Karl Göran Sjögren, et al. Population genomics of Bronze Age Eurasia. *Nature*, 2015.

[152] Martin Furholt. The absolute chronological dating of cord ceramics in Central Europe and South Scandinavia. 2003.

[153] Anja Furtwängler, Adam Ben Rohrlach, Thiseas C Lamnidis, et al. Ancient genomes reveal social and genetic structure of Late Neolithic Switzerland. *Nature communications*, 11(1):1–11, 2020.

[154] Esther J Lee, Cheryl Makarewicz, Rebecca Renneberg, et al. Emerging genetic patterns of the European Neolithic: perspectives from a late Neolithic Bell Beaker burial site in Germany. *American journal of physical anthropology*, 148(4):571–579, 2012.

[155] Detlef Jantzen, Ute Brinker, Jörg Orschiedt, et al. A Bronze Age battlefield? Weapons and trauma in the Tollense Valley, north-eastern Germany. *Antiquity*, 85(328):417–433, 2011.

[156] Ute Brinker, Stefan Flohr, Jürgen Piek, and Jörg Orschiedt. Human remains from a Bronze Age site in the Tollense Valley: victims of a battle? In *The Routledge handbook of the bioarchaeology of human conflict*, pages 192–206. Routledge, 2013.

[157] Samantha Brunel, E. Andrew Bennett, Laurent Cardin, et al. Ancient genomes from present-day France unveil 7,000 years of its demographic history. *Proceedings of the National Academy of Sciences*, 117(23):12791–12798, 2020.

[158] John Novembre, Toby Johnson, Katarzyna Bryc, et al. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

[159] Manfred Kayser, Oscar Lao, Katja Anslinger, et al. Significant genetic differentiation between Poland and Germany follows present-day political borders, as revealed by Y-chromosome analysis. *Human genetics*, 117(5):428–443, 2005.

[160] Krishna R Veeramah, Anke Tönjes, Peter Kovacs, et al. Genetic variation in the Sorbs of eastern Germany in the context of broader European genetic diversity. *European Journal of Human Genetics*, 19(9):995–1001, 2011.

[161] Michael Steffens, Claudia Lamina, Thomas Illig, et al. SNP-based analysis of genetic substructure in the German population. *Human heredity*, 62(1):20–29, 2006.

[162] Laura R Botigué, Shiya Song, Amelie Scheu, et al. Ancient European dog genomes reveal continuity since the Early Neolithic. *Nature communications*, 8(1):1–11, 2017.

[163] Hansi Weissensteiner, Dominic Pacher, Anita Kloss-Brandstätter, et al. HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic acids research*, 44(W1):W58–W63, 2016.

[164] Christopher C Chang, Carson C Chow, Laurent CAM Tellier, et al. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015.

[165] Po-Ru Loh, Mark Lipson, Nick Patterson, et al. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, 193(4):1233–1254, 04 2013.

[166] Michael Salter-Townshend and Simon Myers. Fine-Scale Inference of Ancestry Segments Without Prior Knowledge of Admixing Groups. *Genetics*, 212(3):869–889, 05 2019.

[167] Maïté Rivollat, Choongwon Jeong, Stephan Schiffels, et al. Ancient genome-wide DNA from France highlights the complexity of interactions between Mesolithic hunter-gatherers and Neolithic farmers. *Science Advances*, 6(22), 2020.

[168] Wolfgang Haak, Oleg Balanovsky, Juan J. Sanchez, et al. Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. *PLoS Biology*, 8(11), 2010.

[169] Wolfgang Haak, Peter Forster, Barbara Bramanti, et al. Ancient DNA from the first European farmers in 7500-year-old Neolithic sites. *Science*, 310(5750):1016–1018, 2005.

[170] Barbara Bramanti, Mark G Thomas, Wolfgang Haak, et al. Genetic discontinuity between local hunter-gatherers and central Europe's first farmers. *science*, 326(5949):137–140, 2009.

[171] Eva Fernández, Alejandro Pérez-Pérez, Cristina Gamba, et al. Ancient DNA analysis of 8000 BC near eastern farmers supports an early neolithic pioneer maritime colonization of Mainland Europe through Cyprus and the Aegean Islands. *PLoS genetics*, 10(6):e1004401, 2014.

[172] Iosif Lazaridis, Dani Nadel, Gary Rollefson, et al. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, 2016.

[173] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.

[174] Cristina Gamba, Eppie R. Jones, Matthew D. Teasdale, et al. Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5:5257, 2014.

[175] Iain Mathieson, Songül Alpaslan-Roodenberg, Cosimo Posth, et al. The genomic history of southeastern Europe. *Nature*, 555(7695):197–203, 2018.

[176] Gloria González-Fortes, Eppie R Jones, Emma Lightfoot, et al. Paleogenomic evidence for multi-generational mixing between Neolithic farmers and Mesolithic hunter-gatherers in the Lower Danube Basin. *Current Biology*, 27(12):1801–1810, 2017.

[177] Fernando Racimo, Jessie Woodbridge, Ralph M. Fyfe, et al. The spatiotemporal spread of human migrations during the European Holocene. *Proceedings of the National Academy of Sciences*, 117(16):8989–9000, 2020.

[178] Christine Keyser, Caroline Bouakaze, Eric Crubézy, et al. Ancient DNA provides new insights into the history of south Siberian Kurgan people. *Human genetics*, 126(3):395–410, 2009.

[179] authors. Genetic structure of Europeans: a view from the north–east. *PloS one*, 4(5):e5472, 2009.

[180] Peter de Barros Damgaard, Nina Marchi, Simon Rasmussen, et al. 137 ancient human genomes from across the Eurasian steppes. *Nature*, 557(7705):369–374, 2018.

[181] Willi Wegewitz. *The Lombard fire area of Putensen, Harburg district.* Lax, 1972.

[182] Paul M Barford and Paul M Barford. *The early Slavs: culture and society in early medieval Eastern Europe.* Cornell University Press, 2001.

[183] Paul Fouracre, Rosamond McKitterick, David Abulafia, et al. *The New Cambridge Medieval History: Volume 1, C. 500-c. 700.* Number 1. Cambridge University Press, 1995.

[184] Florin Curta, Paul Stephenson, et al. *Southeastern Europe in the middle ages, 500-1250.* Cambridge University Press, 2006.

[185] Guy Halsall. *Barbarian migrations and the Roman West, 376–568.* Cambridge University Press, 2007.

[186] Sebastian Brather. *Archäologie der westlichen Slawen: Siedlung, Wirtschaft und Gesellschaft im früh-und hochmittelalterlichen Ostmitteleuropa*, volume 61. Walter de Gruyter, 2008.

[187] Patrick J Geary. *The myth of nations: the medieval origins of Europe*. Princeton University Press, 2003.

[188] Martin Gojda. *The ancient Slavs: settlement and society*, volume 1989. Edinburgh University Press, 1991.

[189] Roland Sussex and Paul Cubberley. *The slavic languages*. Cambridge University Press, 2006.

[190] Anna Juras, Miroslawa Dabert, Alena Kushniarevich, et al. Ancient DNA Reveals Matrilineal Continuity in Present-Day Poland over the Last Two Millennia. *PLOS ONE*, 9(10):1–9, 10 2014.

[191] Kerry L. Shaw. Conflict between nuclear and mitochondrial DNA phylogenies of a recent species radiation: What mtDNA reveals and conceals about modes of speciation in Hawaiian crickets. *Proceedings of the National Academy of Sciences*, 99(25):16122–16127, 2002.

[192] Milan Malinsky, Hannes Svardal, Alexandra M Tyers, et al. Whole-genome sequences of Malawi cichlids reveal multiple radiations interconnected by gene flow. *Nature ecology & evolution*, 2(12):1940–1955, 2018.

[193] Daniel Rubinoff and Brenden S. Holland. Between Two Extremes: Mitochondrial DNA is neither the Panacea nor the Nemesis of Phylogenetic and Taxonomic Inference. *Systematic Biology*, 54(6):952–961, 12 2005.

[194] Cosimo Posth, Christoph Wißing, Keiko Kitagawa, et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for African gene flow into Neanderthals. *Nature communications*, 8(1):1–9, 2017.

[195] Alena Kushniarevich, Olga Utevska, Marina Chuhryaeva, et al. Genetic Heritage of the Balto-Slavic Speaking Populations: A Synthesis of Autosomal, Mitochondrial and Y-Chromosomal Data. *PLOS ONE*, 10(9):1–19, 09 2015.

[196] Jiří Macháček, Robert Nedoma, Petr Dresler, et al. Runes from Lány (Czech Republic) - The oldest inscription among Slavs. A new standard for multidisciplinary analysis of runic bones. *Journal of Archaeological Science*, 127:105333, 2021.

[197] Vasili Pankratov, Sergei Litvinov, Alexei Kassian, et al. East Eurasian ancestry in the middle of Europe: genetic footprints of Steppe nomads in the genomes of Belarusian Lipka Tatars. *Scientific reports*, 6(1):1–11, 2016.

[198] BA Maliarchuk, MA Perkova, and MV Derenko. Origin of the Mongoloid component in the mitochondrial gene pool of Slavs. *Genetika*, 44(3):401–406, 2008.

[199] Pengfei Qin, Ying Zhou, Haiyi Lou, et al. Quantitating and dating recent gene flow between European and East Asian populations. *Scientific reports*, 5(1):1–8, 2015.

[200] Peter Ralph and Graham Coop. The Geography of Recent Genetic Ancestry across Europe. *PLOS Biology*, 11(5):1–20, 05 2013.

[201] Hussein Al-Asadi, Desislava Petkova, Matthew Stephens, and John Novembre. Estimating recent migration and population-size surfaces. *PLoS genetics*, 15(1):e1007908, 2019.

[202] Harald Ringbauer, Graham Coop, and Nicholas H Barton. Inferring recent demography from isolation by distance of long shared sequence blocks. *Genetics*, 205(3):1335–1351, 2017.

[203] Martin Petr, Benjamin Vernot, and Janet Kelso. admixr—R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics*, 35(17):3194–3195, 01 2019.

[204] F Lotter. Völkerverschiebungen im Ostalpen–Mitteldonau–Raum zwischen Antike und Mittelalter (365–600). *Gra Ergänzungsband*, 39, 2003.

[205] Garrett Hellenthal, Daniel Falush, Simon Myers, et al. The Kalash Genetic Isolate? the Evidence for Recent Admixture. *American Journal of Human Genetics*, 98(2):396–397, 2016.

[206] Wladyslaw Duczko. *Viking Rus: studies on the presence of Scandinavians in Eastern Europe*. Brill, 2004.

[207] Gary Dean Peterson. *Vikings and Goths: A History of Ancient and Medieval Sweden*. McFarland, 2016.

[208] Krishna R. Veeramah, Andreas Rott, Melanie Groß, et al. Population genomic analysis of elongated skulls reveals extensive female-biased immigration in Early Medieval Bavaria. *Proceedings of the National Academy of Sciences*, 2018.

[209] Emma A Fox, Alison E Wright, Matteo Fumagalli, and Filipe G Vieira. ngsLD: evaluating linkage disequilibrium using genotype likelihoods. *Bioinformatics*, 35(19):3855–3856, 03 2019.

[210] Jonas Meisner and Anders Albrechtsen. Inferring Population Structure and Admixture Proportions in Low-Depth NGS Data. *Genetics*, 210(2):719–731, 2018.

[211] Mikhail Lipatov, Komal Sanjeev, Rob Patro, and Krishna R Veeramah. Maximum Likelihood Estimation of Biological Relatedness from Low Coverage Sequencing Data. *bioRxiv*, 2015.

[212] Leo Speidel, Lara Cassidy, Robert W Davies, et al. Inferring Population Histories for Ancient Genomes Using Genome-Wide Genealogies. *Molecular Biology and Evolution*, 38(9):3497–3511, 06 2021.

[213] Farnaz Broushaki, Mark G. Thomas, Vivian Link, et al. Early Neolithic genomes from the eastern Fertile Crescent. *Science*, 2016.

[214] Lara M. Cassidy, Rui Martiniano, Eileen M. Murphy, et al. Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic

genome. *Proceedings of the National Academy of Sciences*, 113(2):368–373, 2016.

[215] Eppie R. Jones, Gloria Gonzalez-Fortes, Sarah Connell, et al. Upper Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, 6:1–8, 2015.

[216] Nina Marchi, Laura Winkelbach, Ilektra Schulz, et al. The mixed genetic origin of the first farmers of Europe. *bioRxiv*, 2020.

[217] Inigo Olalde, Morten E Allentoft, Federico Sánchez-Quinto, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature*, 507(7491):225–228, 2014.

[218] Federico Sánchez-Quinto, Helena Malmström, Magdalena Fraser, et al. Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. *Proceedings of the National Academy of Sciences*, 116(19):9469–9474, 2019.

[219] Andaine Seguin-Orlando, Thorfinn S. Korneliussen, Martin Sikora, et al. Genomic structure in Europeans dating back at least 36 , 200 years. *Science*, 346(6213):1113–1118, 2014.

[220] Heng Li, Bob Handsaker, Alec Wysoker, et al. The sequence alignment/map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[221] Stephen Sawcer, Garrett Hellenthal, Matti Pirinen, et al. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis, 2011.