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

November 24, 2021

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

Acknowledgements

Thanks to all the good folk at UCL Computer Science cluster, particular Ed and David.

Thanks to my Mum and Dad.

Thanks to everyone in office 212, Mislav, Lucy, Nancy, Arturo, Dave, Mike, Chris, Camus.

Thanks to Nadine for being the best and most supportive admin ever.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Chromopainter and ancient DNA

In this introduction I will outline the following: 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 and options for imputation.

1.1.1 Gains to be made with haplotype information

1.1.1.1 History

Haplotype-based methods are those which explicitly model Linkage Disequilibrium (LD) between neighbouring SNPs along a haplotype. This is in contrast to 'unlinked' methods, which assume a model of Linkage Equilibrium between SNPS. A 'haplotype' is a sequence of alleles along chromosome. 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.

Linkage Disequilibrium (LD), the non-independence of alleles carried at

different positions in the genome, has been studied since the earliest days of genetics [2, 3] and has since been a fundamental aspect of virtually all areas of genetics. Fundamentally, accounting for LD provides 'free' information about markers elsehwere in the genome. In this thesis, here I will focus on its application to inference of haplotype sharing between individuals.

Some of the earliest uses of LD information for the study of population history came fropm microsatellite markers, whose linked repeats can be seen 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 to examine the population structure of whales [4]. Later, LD patterns at the CD4 locus were leveraged to show the preferred model of Human population history was a recent African origin [5]. In particular, Sub-Saharan Africans had substantially more variability in frequencies of haplotypes 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.

The next major advance was the development of methods to use LD information between SNP markers rather than within microsatellites; studies in the early 2000s utilised the then-new Hap-Map results [6] to show LD varies markedly across worldwide populations [7]. This study 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 the previously proposed recent African origin of humanity [8]. These findings were obtained despite the number of SNPs being very low by today's standards; only 2,834 autosomal SNPs were retained. This displays the power of using haplotype-based approaches and that the power scales nonlinearly with the total number of SNPs used.

The 2000s saw a rapid increase in the number of SNP markers and individuals which had been sequenced. Accounting for recombination and LD within a

model is necessarily computationally complex, as the number of combinations of alleles and their possible evolutionary histories balloons as the number of loci considered increases (does it scale quadratically?). Therefore, the new era of sequencing demanded new, more efficient methodology to cope with such data. The development of the Li and Stephens copying model (LSM) [9] was instrumental in the development of such methods [10] and provided an elegant solution to the increased complexity modeling recombination between linked loci. As such, it is now a critical model in virtually all areas of genomic methodology, such as gene conversion parameters, admixed populations, human colonization history, local ancestry in admixed populations and imputation. The LSM was, and still is, the foundation for methods of the haplotype phasing methods needed for haplotype-based methods [11, 12].

Perhaps the first paper to formalise a haplotype-based approach for the study of population history was that of Hellenthal et al 2008 [13]. The ancestry model of Hellenthal et al is based upon the LSM, using a Hidden Markov Model to reconstruct each target haploid as a mosaic of *donor* haplotypes. The conditional (emission) probability that a given haploid 'copies' from a particular reference haplotype is given by whether the alleles at the same position match, with switching between copying haplotypes informed by a genetic map.

In the same year, Jakobsson et al (2008) analaysed a much larger number of SNPs (n=525,910) [14]. It was demonstrated that haplotype clusters show an elevated ability to determine local structure than unlinked SNPs alone; 51.87% of haplotype clusters were found in at most two regions, in contrast with 4.66% of SNP alleles. This seems naturally intuitive, as haplotype clusters are formed from different combination of SNP alleles, which are necessarily more unique than single SNPs.

Building on the copying model proposed by Hellenthal et al (2008), Lawson et al (2015) [15] created ChromoPainter, a model which ..... The authors showed that ChromoPainter had an enhanced ability to separate closely related

populations when plotted on a PCA compared to unlinked methods. ChromoPainter was originally developed in tandem with its own clustering method fineSTRUCTURE, and has since been extended into methods to detect and date admixture [16], and infer ancestry proportions [17].

The 'next-generation' of chromosome painting methods had to address the typical issue in population genomics; 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. One approach is to use the Burrows-Wheeler transform [18, 19], which allows efficient matching of haplotypes in large datasets.

Similarly, methods to detect IBD in Biobank-scale cohorts have leveraged PBWT - cite browning and other work.

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 later sections, 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.

A typical case study, and one which I will return to in later chapters, was a study which attempted to identify population structure among individuals from the British Isles. This study, hereafter referred to as POBI, genotyped 2039 people from England, Wales and Scotland. In summary, it was possible to detect structure down to the level of Devon and Cornwall (two neighbouring counties) using ChromoPainter. On the other hand, little structure was apparent when using unlinked methods (PCA). This outlines the benefits of incorporating linkage information when attempting to identify fine-scale structure between closely related populations.

Gattepaille and Jakobson (2012) [20] 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 individuals alleles separately. They showed that whilst combining two markers 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. They showed that using empirical data of individuals from France and Germany, accounting for haplotypes could reduce the rate of mis-assignment by 73%.

One less considered advantage of using haplotype information is that it may mitigate ascertainment bias. Ascertainment bias occurs when a subset of SNPs are chosen for analysis. SNPs are typically chosen because they display variation. If this variation is determined in one population, say British, then there is no guarantee that the variation will also be seen in another population, say Han Chinese. Therefore, including these SNPs can often provide misleading estimates of genetic diversity and commonly estimated parameters such as F_{st} [21]. 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 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 [22–24]. However, when instead of SNP heterozygosity, haplotype heterozygosity is used as a metric for diversity, African populations consistently have the highest values. The reason for this is, 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. Thus, ascertainment

is less likely to ascertain

In a similar manner, another advantage of using haplotype-based methods is that rare alleles are not required. Rare alleles are highly informative about recent, fine-scale population structure, as they are shared by the fewest number of individuals (max n=2) within a dataset. Methods which leverage this information have been used to model the population history of large datasets [25–27]. However, rare alleles are harder to genotype, as they are more difficult to distinguish from sequencing errors. This is particularly the case when using relatively low-coverage genomes. Because of this, allele-frequency filters are often applied in population genetic studies. Further, more SNPs need to be sequenced in order to find rare variants in a wide range of populations. Using haplotype information negates the needs for using rare variants; if individuals share long haplotypes in common, then by default will also share rare variants that occur on those haplotypes.

However, the usage of haplotype-based methods is not without drawbacks. They are typically slower by several orders of magnitude, as the computational complexity is something.

Secondly, the nature of haplotype-based methods means they require the data to be phased. Phasing is a statistical procedure [1] that requires substantial computation resources. Phasing is a procedure which is often error-prone (switch errors). Care must be taken to ensure the appropriate samples are included in the reference panel

1.2 Methods used to analyse ancient DNA

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

[1]Phasing can also be performed using other methods, such as sequencing family trios. However, this is rarely used in population genetic studies and so I will not discuss it here

1.2.1 Unlinked methods

The first ancient DNA papers mostly relied on statistical methods which compare allele-sharing or allele-frequencies between populations or individuals. These methods, in particular f-statistics and their extensions [28–31] and Principle Component Analysis [32], can address a wide-range of questions pertaining, but not limited to, population structure, admixture, genetic similarity and population graphs.

One reason why methods based on allele-sharing and allele-frequency differences are widely used in ancient DNA studies is that they can easily be modified to work with 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 there are often not enough reads covering a SNP to confidently call heterozygous genotypes, and so are particularly suited to ancient DNA studies where there may only be a single read covering a SNP. Pseudo-haploid calls are therefore used widely, including currently (e.g. [33]), in most studies of ancient humans.

Whilst pseudo-haploid genotype calls circumvent the problem of calling heterozgyous genotypes at low coverage positions, there is necessarily a reduction in the information they hold relative to true diploid genotypes and are thus less powerful. Further, the use of pseudo-haploid calls may result in an elevated level of reference bias [34–36].

For many of the early ancient DNA studies, such as that of Green et al 2010 [28] and Lazaridis et al 2014 [37], powerful methods for detecting population substructure and admixture were not required as they primarily considered broad-scale 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. 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 [2], which were first outlined in a 2009 study into the population history of India [39]. 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 treeness (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 [31].

One possible issue of f-statistics is that of drifted populations; f_3 tends to pick out drifted populations.

A similar method is the so-called ABBA-BABA test, developed by Green et al (2010) [28] 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 contrast to the F-statistics, which explicitly tests models of population relationships, Principle Component Analysis (PCA) is typically used to obtain a broad overview of the genetic ancestry of the sample beingn analysed.

[2] Although related, they should not to be confused with Sewall Wright's F-statistics [38].

1.2.2 ChromoPainter ancient DNA

1.2.2.1 History

In recent years, the 'low hanging fruit' of broad-scale questions 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 samples and concerns over introducing bias towards present-day populations.

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 a broad overview of genetic ancestry through Principle Component Analysis of the coancestry matrix. Differential haplotype donation to different worldwide populations, as shown in Fig X, can reveal geographic correlates of genetic variation. The identification of genetic clusters and admixture events is also able.

The first use of chromopainter on ancient DNA was in the seminal paper of Lazaridis et al (2014) [37]. Cautious about imptuting missing alleles in the ancient samples, the effect of which had yet to be studied, the authors opted to retain only positions with non-missing genotypes (as the samples were of high coverage, this was not an issue, as 495,357 SNPs were retained). The authors confirmed the ability of fineSTRUCTURE to meaningfully cluster ancient individuals by recapitulting previous results that identified different present-day European populations as being more closely related to Early Farmers and hunter-gatherers than others.

Inbetween 2014 and the present-day, there have been approximately x studies which have used CP on ancient samples (based on Web of Science

search results). I will briefly describe some of the more notable papers.

Jeong et al (2019) focused on detecting admixture using GLOBETROTTER. This thesis will not consider the use of GLOBETROTTER on ancient DNA, but it is notable because .... [40].

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 [41]. 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 0.1x mean depth.

Another recent large study was that into samples from ancient Rome [42].

The most recent study using CP a

More recently, ChromoPainter has been used to study aspects of archaic hominin ancestry in present-day humans [43, 44]. 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 plants, bacteria [45].

1.2.2.2 Benchmarking ChromoPainter and imputation

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

An early study to explicitly investigate the reliability of ChromoPainter on ancient DNA was Martiniano et al (2017) [35]. This study explored various aspects of ChromoPainter analysis on ancient samples. Testing whether

including imputed genotypes introduced bias towards particular present-day populations was key, as if it were the case, it would potentially invalidate all results obtained from using ChromoPainter on ancient samples. Potential bias was estimated by plotting normal quantile-quantile plots of the copyvectors obtained from imputed 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 impact of filtering genotypes based on genotype probabilities was determined by creating two datasets, one containing hard filtered genotypes and one not, and performing fineSTRUCTURE clustering. They inferred 7 more clusters when using filtered genotytpes. Whilst this could perhaps be an indication of imporved performance, 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 the MCMC sampling scheme, 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 individuls 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 included their own set of benchmarks. Antonio et al (2019) [42] analysed 127 ancient genomes of a mean coverage of 1x and tested imputation accuracy on a single individual (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 occured.

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-estimted, 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 has been tested in various studies. 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. However, all investigations have reported a 'high' concordance between

The most systematic and thorough evaluation of imputation in ancient genomes was performed by Hui et al (2020) [46]. 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.

Furthermroe, Hui et al showed that of the several different methods for estimating genotype likelihoods from read data, atlas provided

It should be noted that the study only considered a single ancient genome (NE1) and it is therefore unclear how generalisable these results are when applied to samples with ancestries more or less prevalent in a reference panel. In particular, the results may not be applicable to ancient samples from Africa,

which would likely harbour more diversity, much of which would be unlikely to be present in any reference panels. However, this study provided important benchmarks for many criticial steps in the analysis of low coverage samples which had previously been missing from the literature, such as selection of a reference panel, the feasability of local imputation and the application of pre and post imputation filters.

1.2.2.3 Mini-conclusion

In total, there have been several efforts to determine whether or not coverage plays a role in CP analysis. The results show that....

However, the studies have been lacking in that....

1.3 Issues and solution to low coverage data

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. This is because 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 coverage of a genome is therefore the mean number of reads mapped to each position in the genome.

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 3 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 [47]. Common methods to estimate likelihoods include the GATK model [48], SAMtools [49], SOAPsnp [50] and SYK model [51]. Genotype likelihoods can either be estimated prior to the analysis from aligned reads (BAM files), using software such as ANGSD [52], ATLAS [53] or GATK [48]. Other softwares will take BAM files directly as input and estimate genotype likelihoods during the analysis process (e.g. STITCH [54]).

Once genotype likelihoods have been estimated, population level parameters such as inbreeding coefficients and F_{st} can be estimated directly [52] with greater accuracy than direct genotype calls. Similarly, modifications of the ADMIXTURE [55] algorithm and PCA have been developed in order to analyse low coverage samples more effectively [56, 57]. Recent advances have allowed the identification of 1st and 2nd-degree relatives from as low as 0.02x coverage samples [58, 59].

Several methods exist which jointly estimating ancient DNA specific confounding factors, such as contamination and post-mortem damnage, alongside the demographic parameter of interest [60]. Schraiber (2018) [61] 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 [62]. This program is mostly designed for demographic inference in the context of conservation

genetics - for example, estimating the relation of inbreeding to fitness decline. To account for the uncertainty, all 3 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. Whilst similar to ChromoPainter in terms of modeling SNPs as linked markers, ngsF-HMM differs in that it estimates pairwise IBD segments rather than comparing each haplotype to all other haplotypes.

1.4 Combining data from multiple chips

A related issue stems from the current practice of developing a large number of genotyping arrays. Different cohorts are genotyped on different arrays and sets of SNPs, as different SNPs have different characteristics. For example, some SNPs are known to be associated with particular phenotypes, some SNPs are known to be more variable (and therefore more informative at identifying structure) in certain populations. Whilst this generation of custom genotyping arrays has meant a wider variety of questions and populations can be studied using genoyping arrays, 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 3 genotyping arrays; '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 a typical ChromoPainter analysis.

Having a smaller number of SNPs may reduce power in two ways. Firstly,

there is simply fewer informative data points to use when comparing the SNP patterns between two populations and therefore fewer possible data points which can be used to identify populations. Secondly, ChromoPainter derives parts of its power from the LD between neighbouring SNPs. LD between two neighbouring SNPs is correlated with their physical distance. Fewer overall SNPs means each neighbouring pair of SNPs are physically further away from one another and thus have less LD information.

One solution to the issue of a small number of SNP would be to impute the remaining SNPs. In this context, imputation refers to estimating missing genotypes using of a model usually based upon the LSM and a large reference panel. Imputation is widely used in e.g. GWAS to generate sequence-level data.

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, one question to ask is the following; is it more desirable to impute the missing positions or to use use a smaller number of overlapping SNPs. This is something which I will investigate in chapter 3 with a case study investigating African ancestry in the UK Biobank dataset.

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 two new versions, ChromoPainterUncertainty and ChromoPainterUncertaintyRemoveRegions, which are designed to to attempt 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 [53], phasing and genotype imputation with GLIMPSE [63], ChromoPainter [15] analysis (copy-vector estimation and PCA) and SOURCEFIND ancestry component estimation [17]. 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 at

[WHAT ABOUT THE DEVON/CORNWALL STUFF – IT'S NOT JUST ABOUT aDNA?]

2.2 Methods

2.2.1 Description of the ChromoPainter algorithm

ChromoPainter is a method designed to infer patterns of haplotype sharing between individuals [15]. 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, which refers to the process of determining which alleles along a chromosome were inherited together from the same parent. Sampled individuals are split into 'donor' and 'recipient' haplotypes, and ChromoPainter employs the widely-used Li and Stephens copying model [9] 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-neighbor' 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 d and recipient r, carrying alleles x and y, respectively, at

position (e.g. 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 mutation probability.The mutation probability θ can be estimated using Watterson's estimator [64], or estimated using an iterative EM algorithm. Begin with an estimate of θ, usually Watterson's estimate, and at each iteration, replace the value of θ with:

$$\theta^* = \frac{\sum_{l=1}^{L}(\sum_{i=1}^{j} \alpha_{il}\beta_{il}I_{[h_{*l} \neq h_{il}}/P(D))}{L} \tag{2.2}$$

The transition probabilities, i.e. 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.

In ChromoPainterV2, the input genetic data comes in the form of genotype calls (i.e. 1/0, A/T/C/G). ChromoPainterV2 produces several different output files. The two which most used in this work are those appended with .chunklengths and .chunkcounts. 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.

In this work, 'copyvector' is used to refer to the vector of chunklengths

that a single recipient individuals copies from all donors. 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.

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 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 genotype probabilities of being genotype i. 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 is `1` then the allele likelihood is simply $1 - U$ and if the allele is `0` then the allele likelihood is $0 + U$.

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.

$$\begin{aligned} p(r_x|d_x) = (1-\theta) * [r_x * d_x + (1-r_x) * (1-d_x)] \\ +\theta * [r_x * (1-d_x) + (1-r_x) * d_x] \end{aligned} \tag{2.3}$$

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 at each stage of the ChromoPainter analysis. I took several 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 [65]

2. UstIshim – Siberian Upper Paleolithic hunter-gatherer [66]

3. sf12 – Scandinavian Hunter-Gatherer [67]

4. LBK – early European farmer from the Linearbandkeramik culture from Stuttgart, Germany [37]

5. Loschbour – 8,000 year-old hunter-gatherer from Luxembourg) [37]

These samples were chosen due to their high original coverage ($> 18x$), and because they are a diverse representation of 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 [53] (`https://bitbucket.org/wegmannlab/atlas/wiki/Home`). First, the validity (i.e. ensuring that each `.bam` file was not malformed in any way) using ValidateSamFile command from PicardTools [68]. atlas is a suite of software designed for processing low-coverage ancient DNA and was chosen following the reccomendation of Hui et al (2020) [46], 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 [69]. 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 samples

I also generated a set of ancient samples to use as donors in the ChromoPainter analysis (Appendix table 1).

This dataset consists of 124 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 from appendices A1, A2, A3, A4, and they 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, often not containing any reads at several sites [70]. 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, and reduce the impact of errors on downstream analyses [63].

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 crucial for estimating uncertainty values, described in section 1.2.11, for including these genotype

probabilities into the painting process. 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 [63, 71–73].

Two programs, Beagle 4.0 [74] and GLIMPSE [63] fulfill 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 [63], 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 [75]. A description of the processing of this reference dataset can be found in appendix A.2.

I next 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) [76] 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.

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.

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 basepairs, 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` using 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.

2.2.5 Estimating imputation sensitivity and specificity

I used rtg-tools-3.11 [77] and the `vcfeval` task to estimate the sensitivity and specificity of variant discovery 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:

$$sensitivity = \frac{V_{call} - FP}{V_{call}} \tag{2.4}$$

$$precision = \frac{V_{baseline} - FN}{V_{baseline}} \tag{2.5}$$

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. 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 1.2.1.1 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 124 ancient samples downloaded from the literature as recipients. The 124 ancient samples from the literature were included in order so that they can be used a surrogates in later SOURCEFIND analysis.

This produces 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 (e.g.) from their own population. For example, an individual from France copies non-zero amounts from African donors, despite not having any recent 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 [17]. Simulations have shown that SOURCEFIND ancestry proportions correspond well to simulated values [17]. The ancestry proportions produced by SOURCEFIND should be interpreted as the proportion of ancestry that each invidual/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 down-sampled 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 (individuals and population labels in Appendix B.x). 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 modeled as a mixture of those three groups [37, 78]. 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 'easy' 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 D.5). The rest of the parameters were left as default. Ancestry proportions, credible intervals and chain mixing/convergence checking for each surrogate group were estimated using the CODA R library [79].

2.3 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 with 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 neighboring counties Devon and Cornwall using the fineSTRUCTURE algorithm, but not using unlinked methods (ADMIXTURE [80]) [81]. 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.

2.4 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 analyses I perform in Chapter 3.

[I SAY THIS, BUT ACTUALLY (HOPEFULLY) THIS ISN'T QUITE TRUE – FOR UK BIOBANK, PRESUMABLY YOU SUBMITTED UK-BIOBANK TO HRC USING ALL GENOTYPED SNPS, THEN REDUCED TO THE SNPS THAT OVERLAP HUMAN ORIGINS? IN THAT CASE, WE COULD PHRASE THIS AS BEING MOTIVATED BY THAT ANALYSIS BUT EXPECTED TO PERFORM (PERHAPS CONSIDERABLY) WORSE – E.G. WE ARE ASSUMING WE ONLY HAD 70K GENOTYPES PRIOR TO IMPUTATION, RATHER THAN A FULL SET OF 1M OR SO SNPS (WHATEVER THE UK BIOBANK ARRAY HAS). IN THIS WAY, IT PROBABLY DOES MIMIC aDNA MORE – CAN WE PHRASE THIS IN A WAY TO SHOW THIS? E.G. DOES AN aDNA SAMPLE OF AVERAGE COVERAGE X HAVE 70K SNPS WITH >2-3x COVERAGE? I.E. WHAT IS "X" HERE?].

DONT EXPLAIN IN TERMS OF CHAPTER 3

I took the Human Origins dataset (appendix A.19), 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 [82] 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.

2.5 Results

2.5.1 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 [77]. This approach compares genotype calls at each position in each downsampled individual after imputation to the same individual at full coverage without imputation.

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 [46, 63]. In particular, alleles which are rare in the reference panel are less likely to be imputed correctly.

Different downsampled individuals differed 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

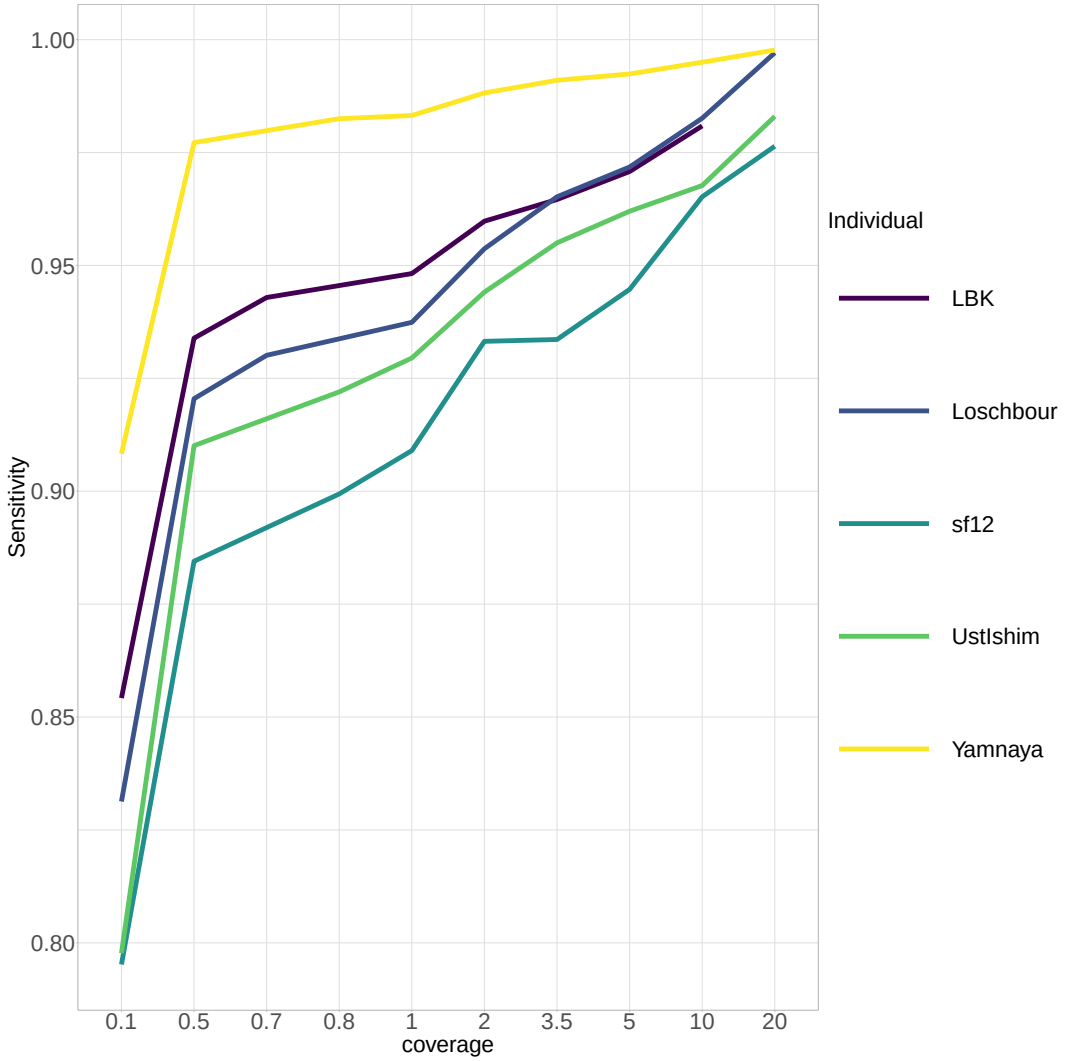


Figure 2.1: Sensitivity of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

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 [83]. 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 [71, 72]. On the other hand, the sample Ust'Ishim is known to have contributed very little genetic ancestry to present-day populations [84] and may therefore have fewer closely matching haplotypes in the reference panel, and a correspondingly lower imputation accuracy.

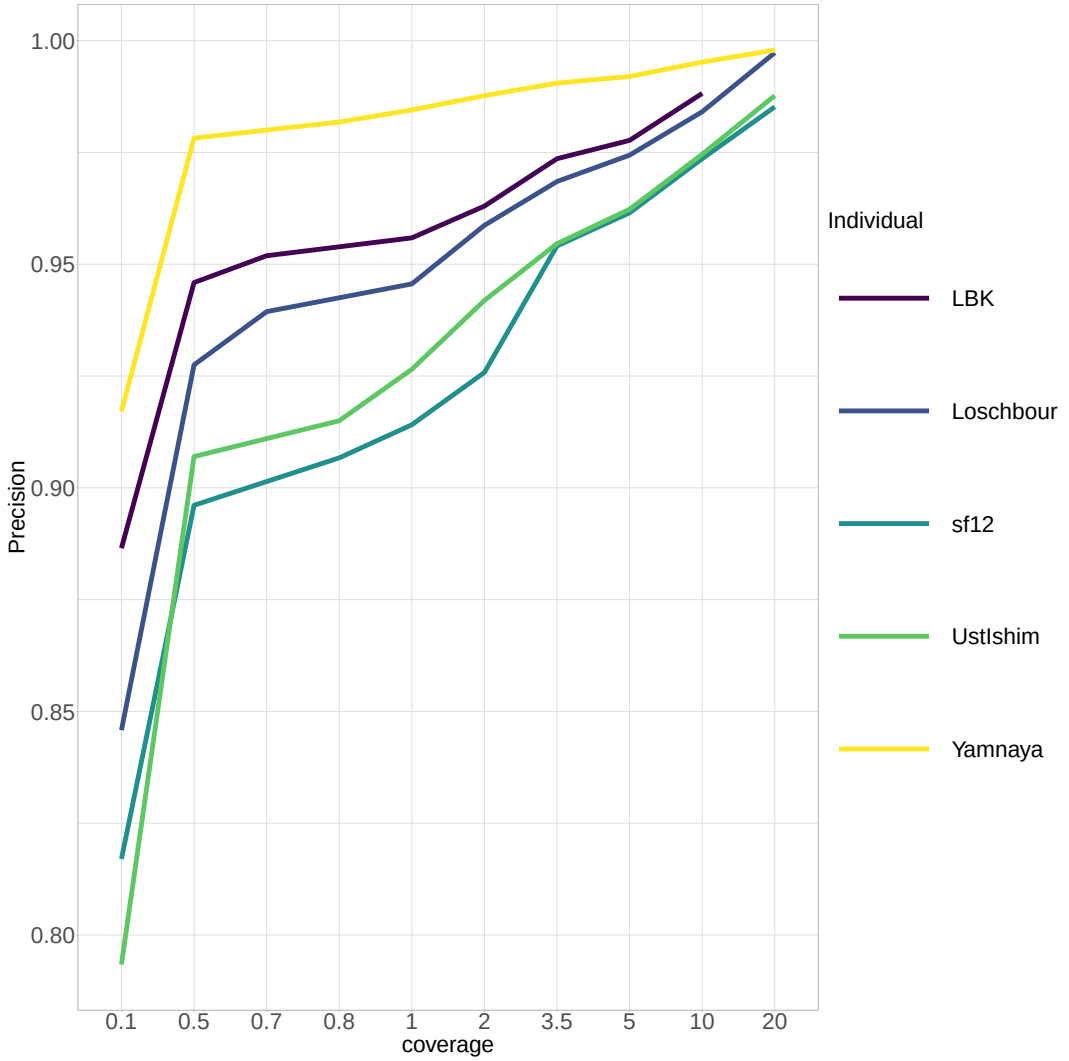


Figure 2.2: Precision of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

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) [85] and fewer heterozygous positions. As imputation relies on matching IBD tracts between individuals, imputation accuracy increases where individuals share more IBD [86]. 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.

2.5.2 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 [63].

2.5.3 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

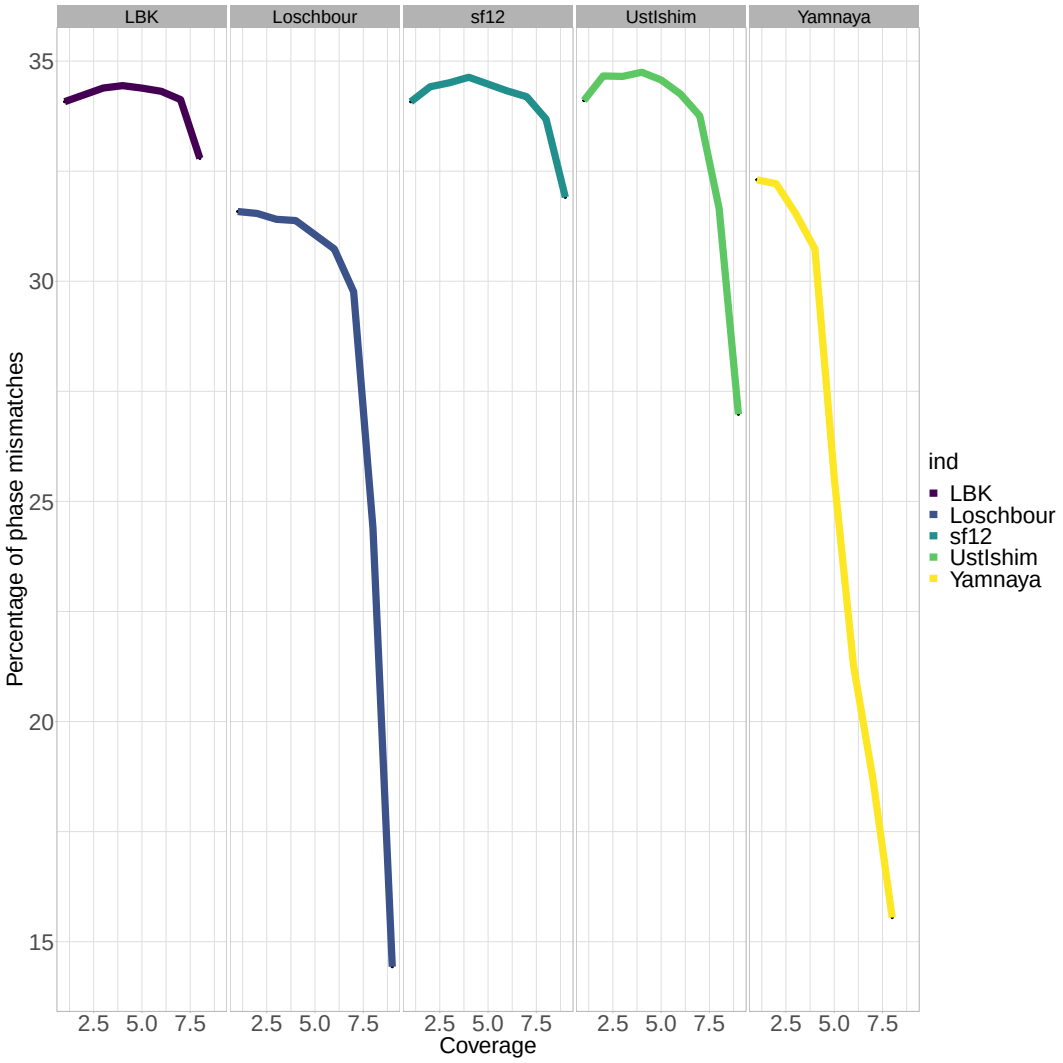


Figure 2.3: Percentage of phased genotypes which agree with the same full-coverage sample? for each individual and each level of downsampling. Genotypes with phase deemed unresolvable by rtg-tools were excluded from the calculations. Note that these numbers are given as incorrect / (incorrect + correct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

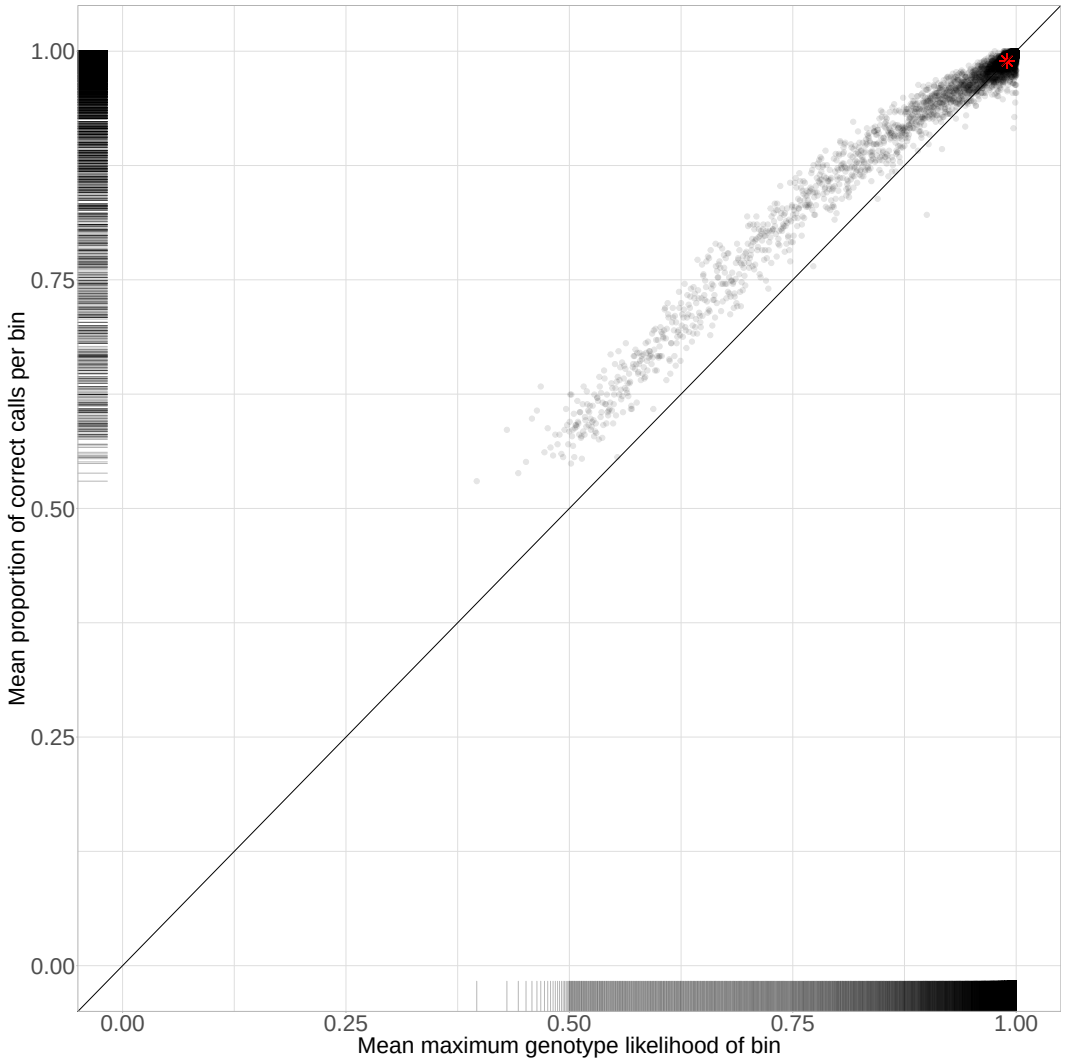


Figure 2.4: Relationship between genotype likelihood and probability of genotype call being correct for Yamnaya downsampled to 0.1x coverage. Genome binned by maximum posterior genotype likelihood and mean maximum posterior genotype likelihood (x-axis) and proportion of correct calls per bin (y-axis).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

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 (result omitted).

2.5.4 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 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 r-squared between the copyvectors of the full coverage and downsampled individuals.

Fig. 2.5 displays the relationship between copyvectors for each downsampled individual 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

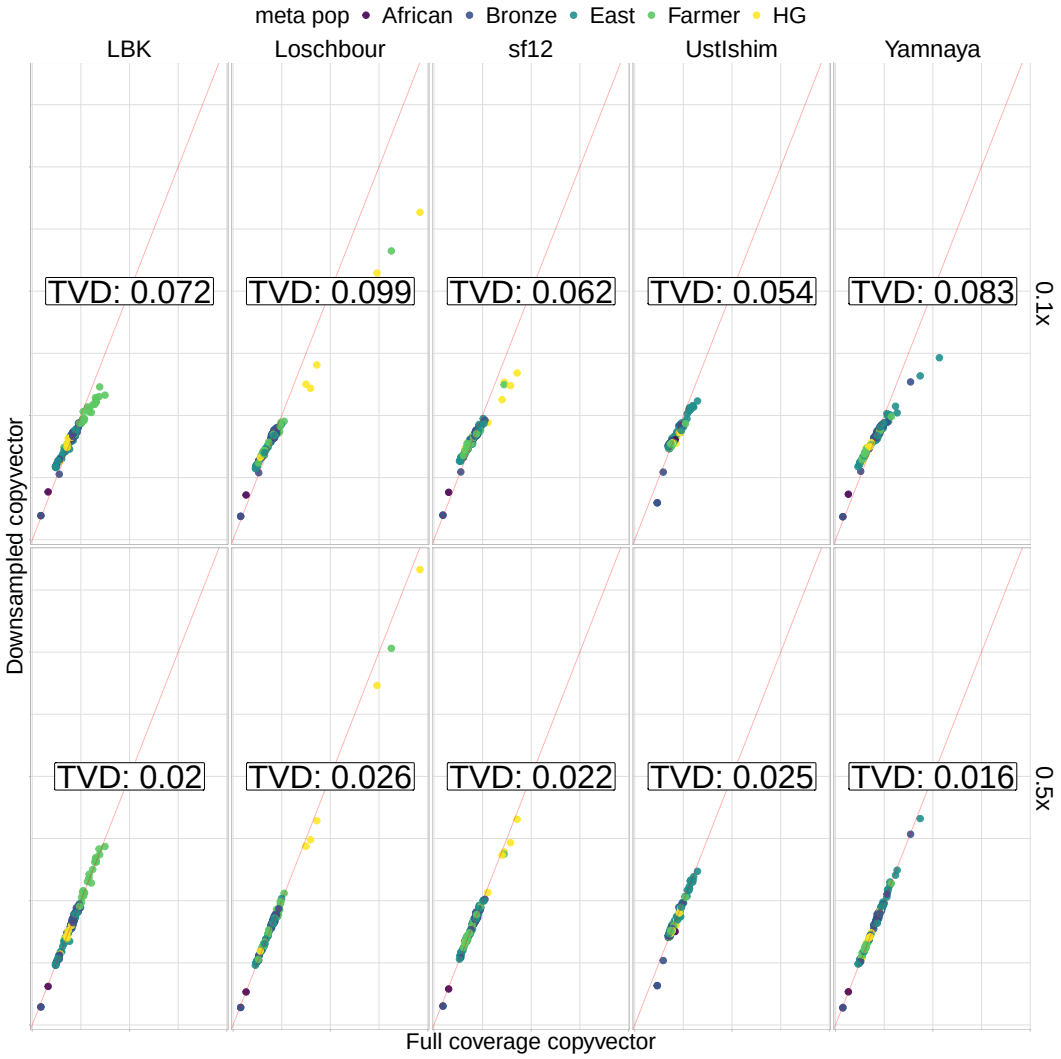


Figure 2.5: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to each of 125 ancient individuals (dots). Results are shown for 0.1x (top row) and 0.5x (bottom row) downsampled genomes. Points coloured by manual assignment 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.

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 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.

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 A.5). 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

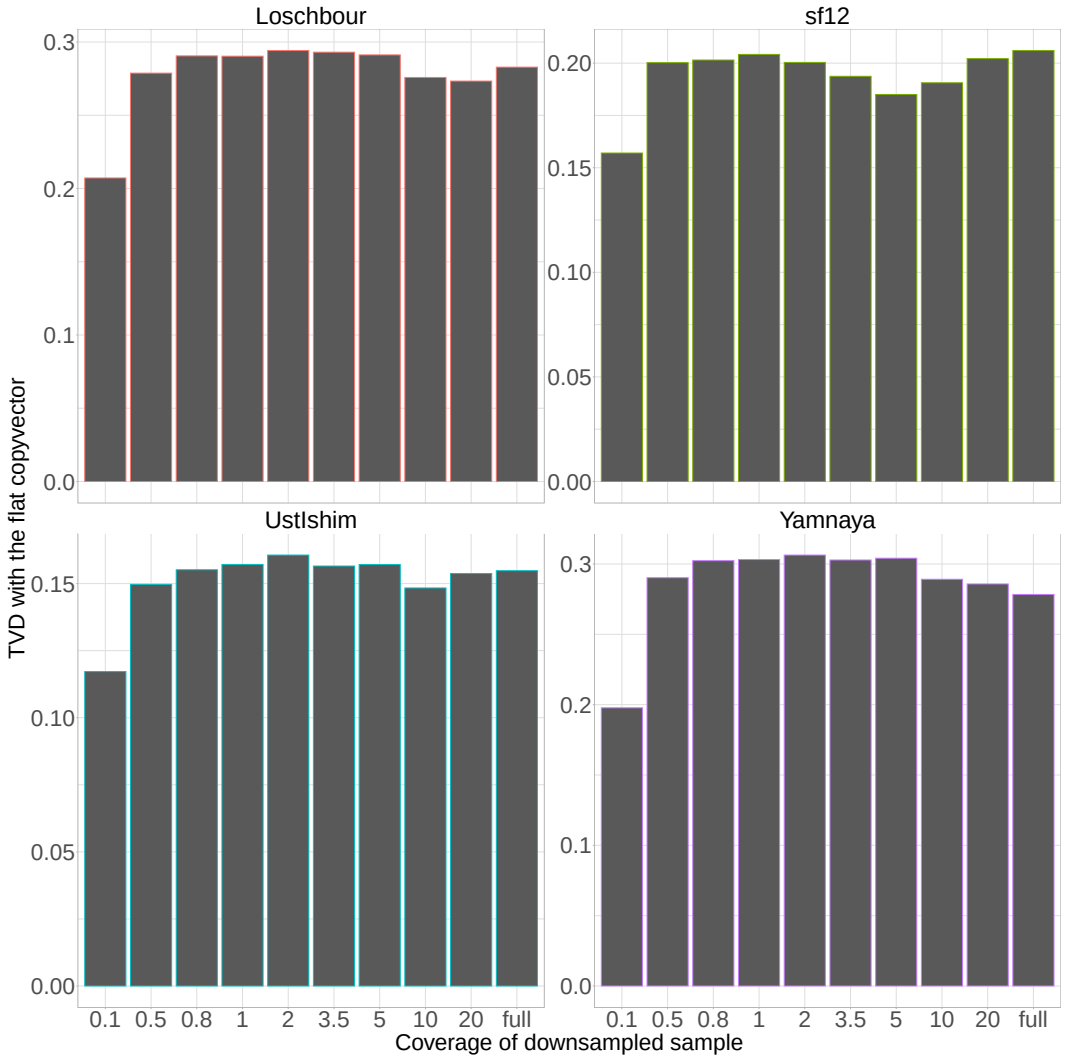


Figure 2.6: TVD (metric of copyvector dissimilarity between two individuals) between each downsampled ancient individual and a flat copyvector. Flat copyvector equivalent to a vector of length N where each element $= 1/N$.

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 [66]. Accordingly, relatively less useful information is obtained from painting Ust'Ishim with a modern reference panel than LBK.

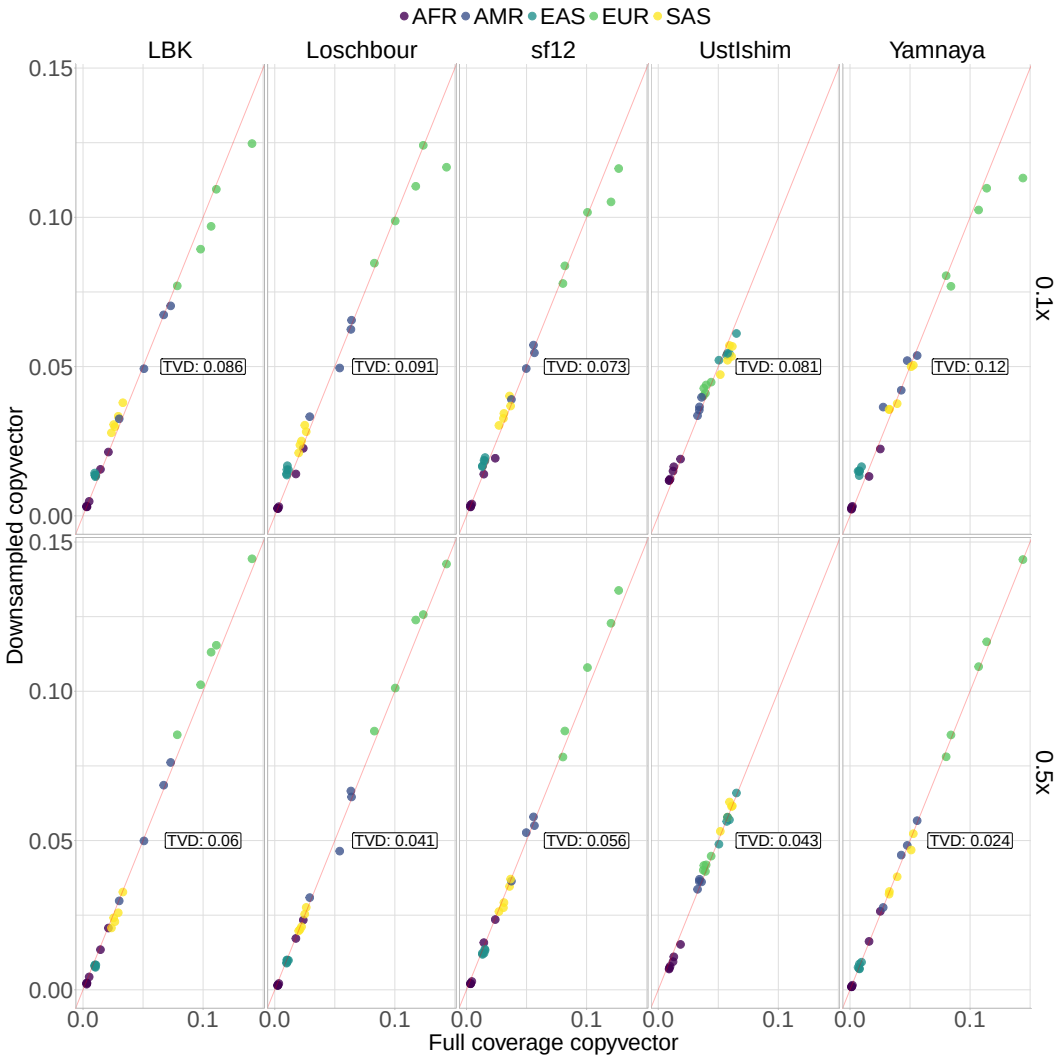


Figure 2.7: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to individuals from each of 26 present-day populations (dots). Red line is $y = x$. x and y units are normalised copying values and thus removed for clarity.

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 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.

Taken together, these data suggest 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 [41].

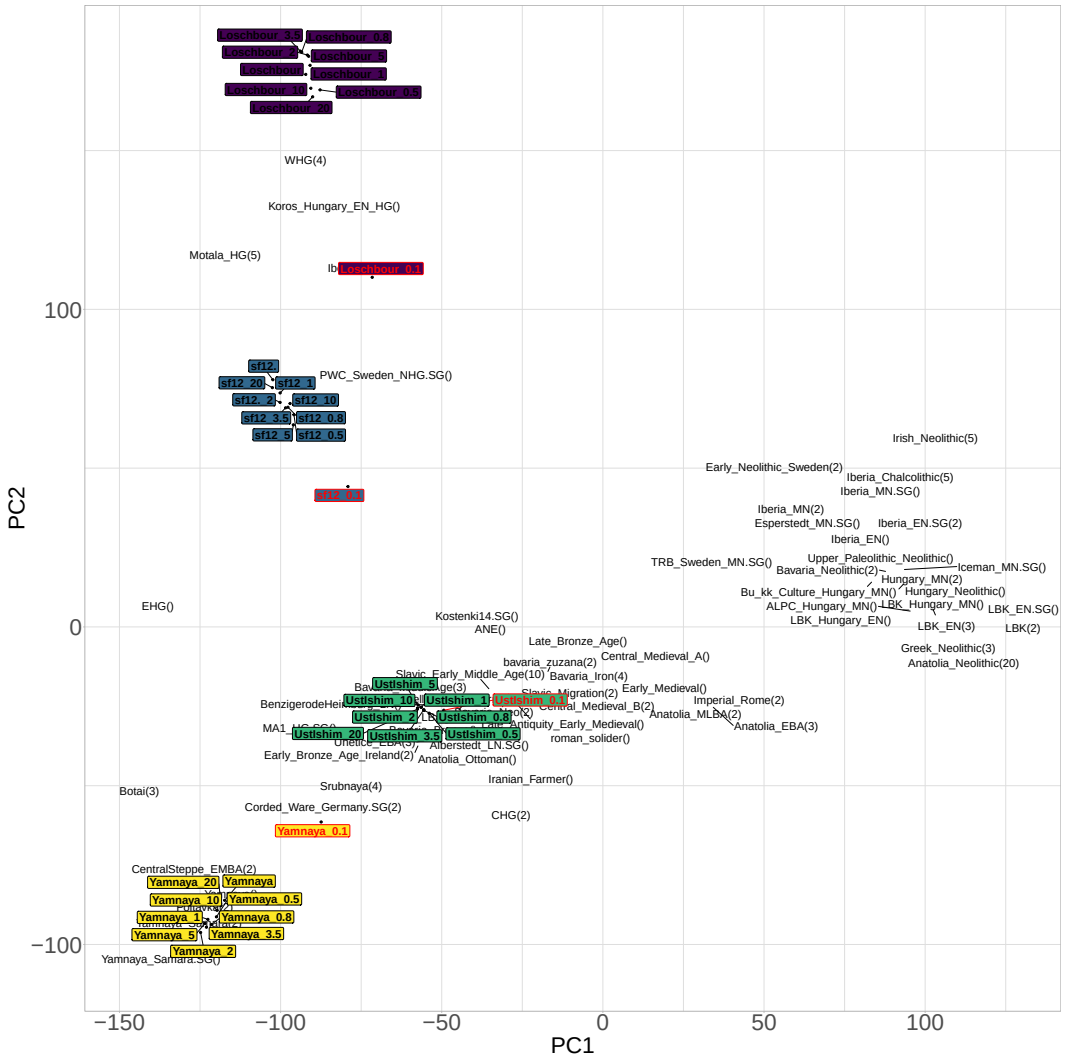


Figure 2.8: Principle component analysis (PCA) of downsampled, full coverage and downloaded ancient individuals generated from the linked chunklengths matrix. Full coverage and downsampled genomes of the same individual are coloured the same. Reference individuals are grouped into populations plotted 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.

2.5.5 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" patterns to improve interpetability 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 appear to be 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 artifacts 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, 2.369%.

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 populations.

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 Paleolithic / Early Neolithic Hunter-Gatherer populations were small and lacked genetic diversity [37, 87, 88]. It is therefore expected that Hunter-Gatherers would share longer IBD segments than individuals from outbred populations. Accordingly, this may make estimating SOURCEFIND

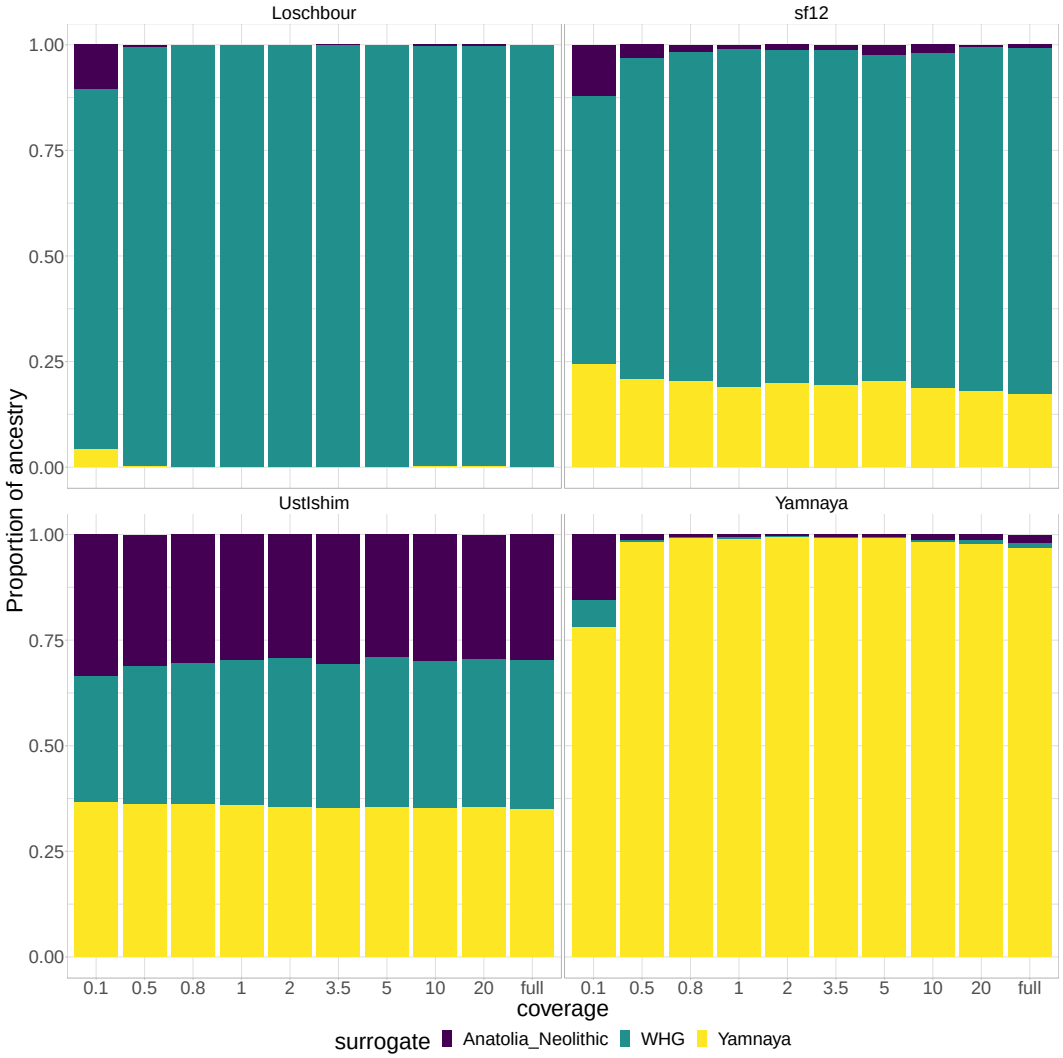


Figure 2.9: Each panel gives inferred recent ancestry sharing proportions for a different downsampled genome. Bars represent proportion of ancestry, coloured by different surrogates. Different coverages for the same individual are given within each panel. Three surrogates were used.

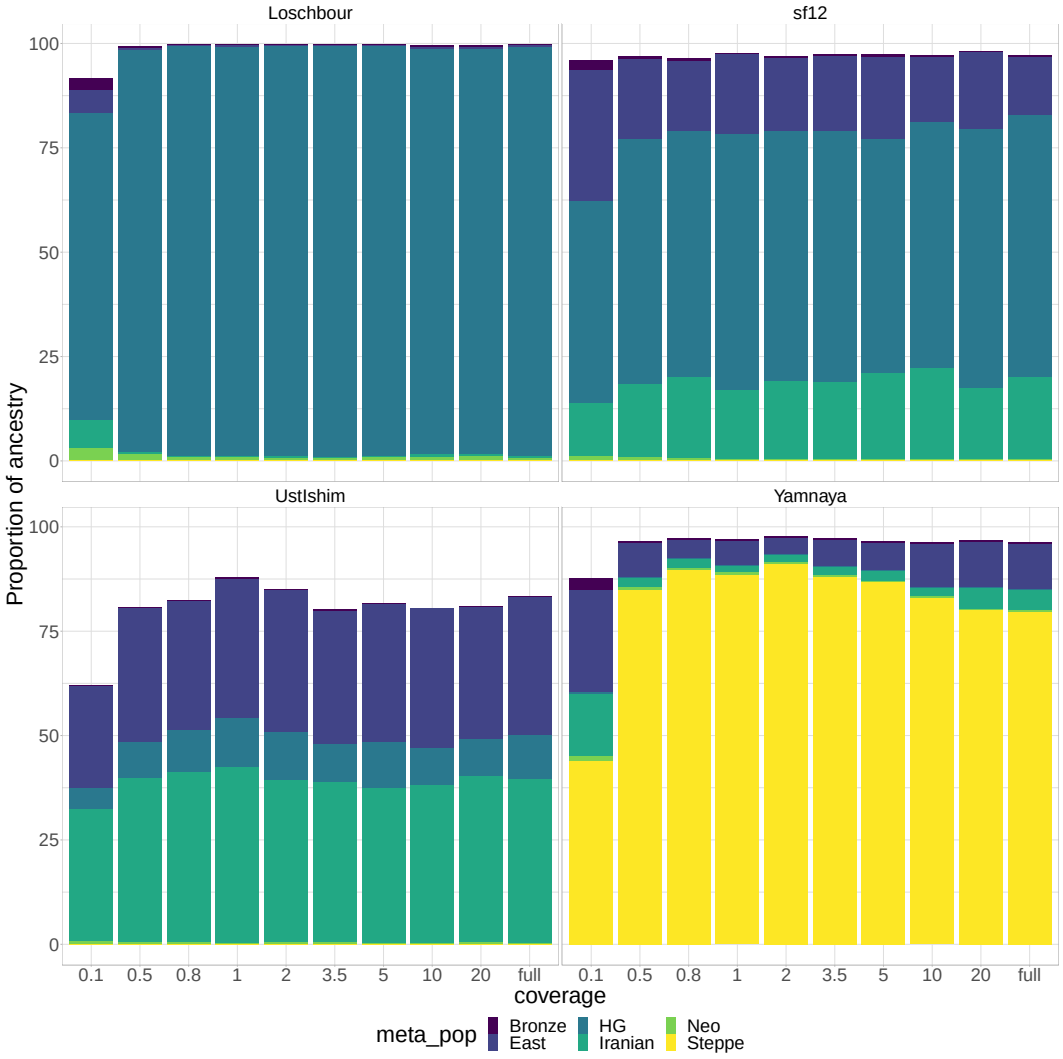


Figure 2.10: Each panel gives information for a different downsampled genome. Bars represent proportion of ancestry, coloured by different surrogates. Different coverages for the same individual are given within each panel. All 39 ancient surrogates were used. Only surrogates with more than 5% are shown.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

proportions easier.

2.6 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). 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 this mis-estimation 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.

2.6.1 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 using i) pre-GLIMPSE genotypes and ii) post-GLIMPSE (imputed) genotypes (Fig. 2.12). 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, the 0.1x samples are clearly shifted to the centre[REALLY HARD TO SEE THIS – CAN YOU HIGLIGHT THEM SOMEHOW; MAYBE PUT A RED BOX AROUND THE 0.1x SAMPLES IN EACH PLOT?] of the post-GLIMPSE PCA, 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 a PCA, using the same set of present-day European samples and downsampled ancient individuals as previously, but on the chunklengths matrix[WHICH PAINTING? IS IT AGAINST THE 26 (1KGP?) EUROPEAN DONOR POPS; IF SO SHOULD RE-ARRANGE THE FIRST PART OF THE SENTENCE TO SAY THIS EXPLICITLY] ChromoPainter output. 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.

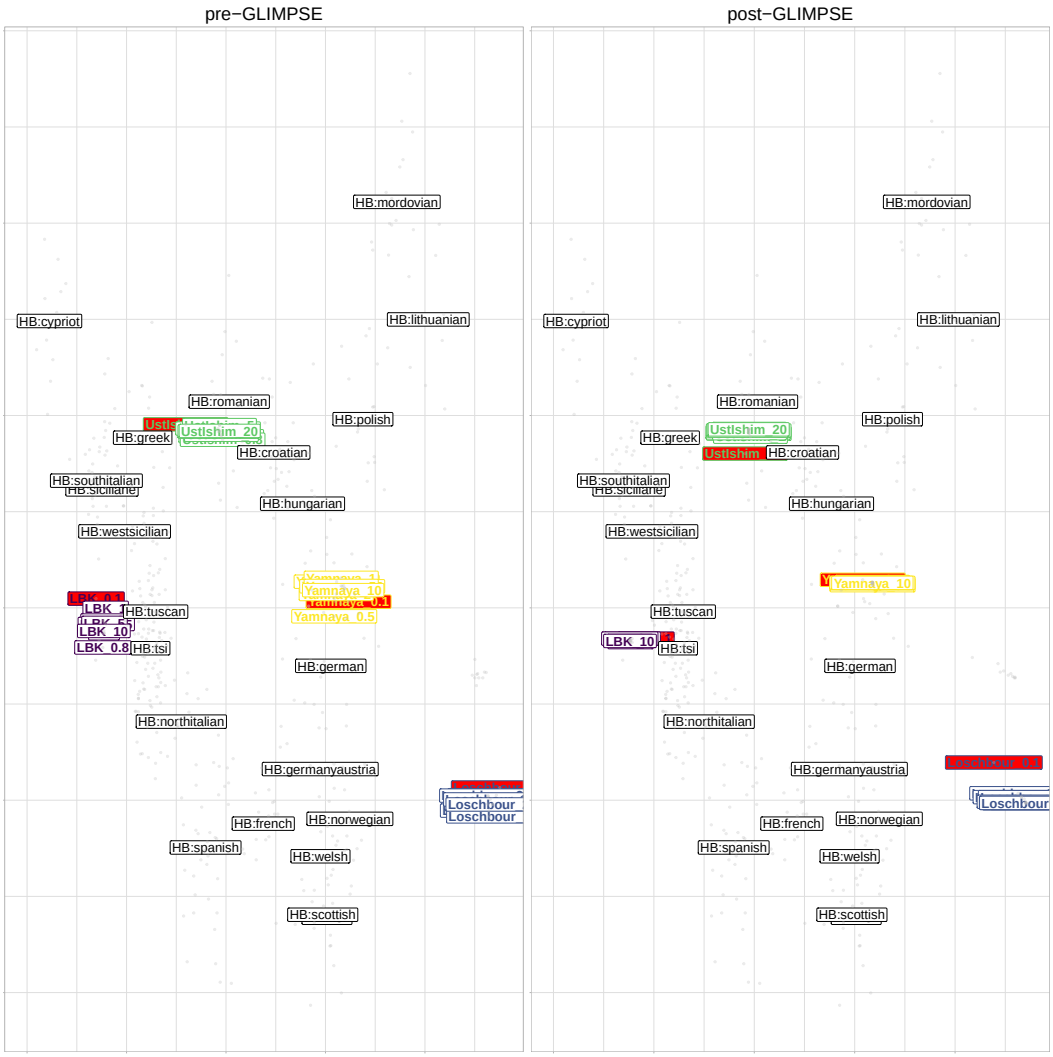


Figure 2.11: Principle Component Analysis. Left - pre-GLIMPSE genotypes. Right - post-GLIMPSE genotypes. White labels correspond to the midpoint of all samples from that population, grey points correspond to modern individuals.

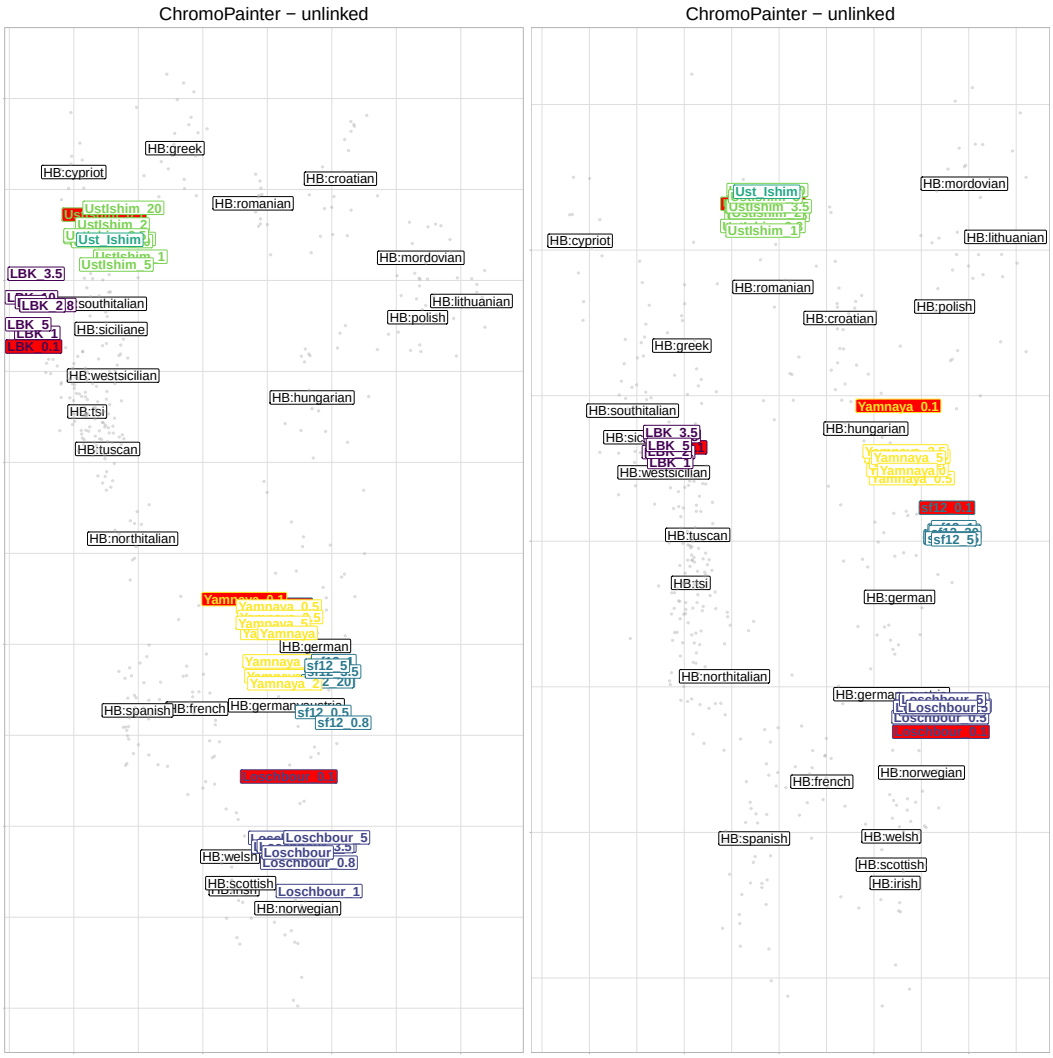


Figure 2.12: Left - ChromoPainter Linked. Right - ChromoPainter Unlinked. White labels correspond to the midpoint of all samples from that population, grey points correspond to modern individuals.

2.6.2 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 A.19), containing the genotypes of 5998 present-day individuals from across the world, genotyped 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 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 imputed and non-imputed SNP set[DO YOU MEAN "under the dataset where all XX SNPs are not imputed versus the dataset where XX (XX%) 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.

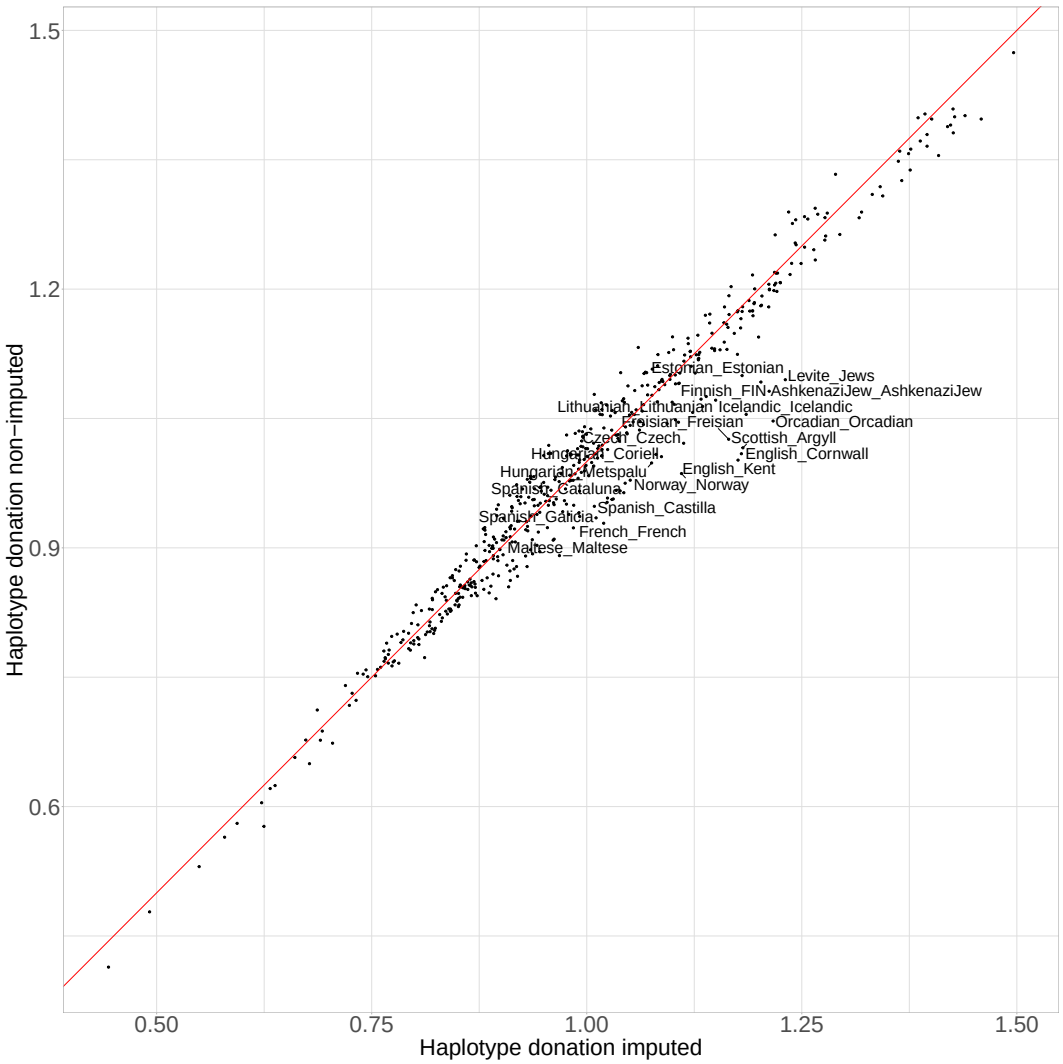


Figure 2.13: Comparison of the mean normalised cM donated by each donor population using the imputed and non-imputed SNP sets. The 20 populations with the largest different between imputed and non-imputed donation are highlighted.

2.7 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.

2.7.1 Accounting for allele likelihoods

Section 2.2.1 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 [89].

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.

2.7.2 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

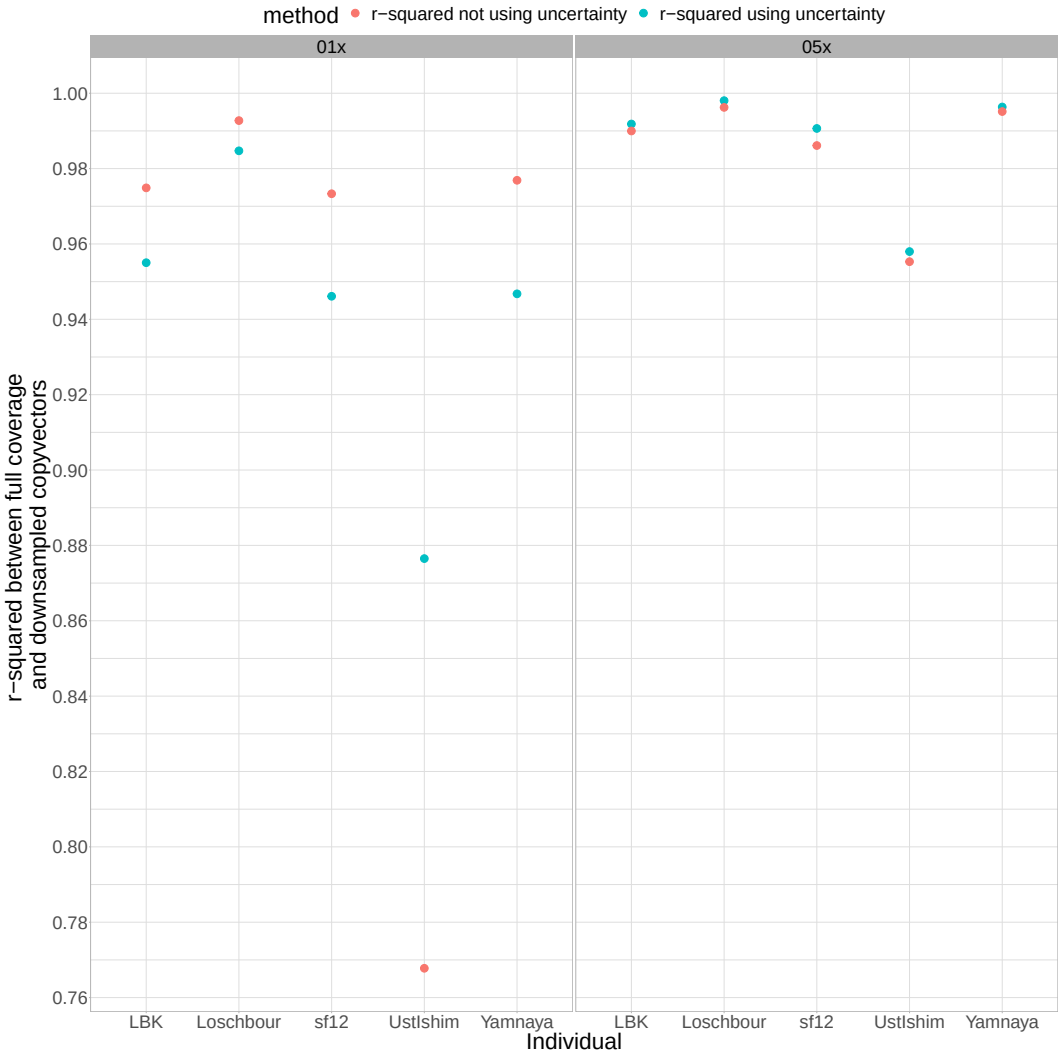


Figure 2.14: Comparison of performance of ChromoPainterV2 and ChromoPainterV2Uncertainty. Panels correspond to samples downsampled to 0.1x (left) and 0.5x (right). Points correspond to the r-squared between the downsampled individual and the same individual at full coverage. Red points are values obtained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

allele frequency) used for imputation is known to affect how accurately that variant can be imputed [46, 63, 71, 90]. 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 I then painted individuals downsampled to 0.1x and 0.5x using the standard set of 125 ancient donor individuals.

Comparing the r-squared values between the copvectors showed that this did not improve the 0.5x copyvectors (Table 1.1).

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, I only retained positions where the $max(GP) >= 0.990$. 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 standard set of 124 ancient donors was used. Again, this did not improve the accuracy of the copyvectors.

sample	u_01x	s_01x	r_01x	gp_01x	u_05x	s_05x	r_05x	gp_05x
LBK	0.989	0.989	0.979	0.819	0.996	0.996	0.997	0.992
Loschbour	0.998	0.998	0.992	0.844	0.999	0.999	0.999	0.994
sf12	0.989	0.989	0.974	0.761	0.995	0.995	0.995	0.982
Yamnaya	0.990	0.990	0.972	0.772	0.999	0.999	0.998	0.995
UstIshim	0.848	0.848	0.930	0.773	0.992	0.992	0.979	0.969

Table 2.1: Table of r-squared 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 $max(GP) >= 0.990$.

2.7.3 Restricting analysis to non-imputed SNPs

Section 1.6.1 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 from the analysis may have side-effects. Increasing the genetic distance between SNPs reduces linkage information and therefore may reduce the overall power of haplotype-based methods 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 haplotype-based methods. 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 previously been achieved using the fineSTRUCTURE clustering algorithm using linkage information, but not using

unlinked methods (ADMIXTURE [80]) [81]. 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=2039), using the full set of SNPs (n=452,592). For a classification score, I found the proportion of Cornwall individuals whose copy vector had a lower TVD with the mean copy vector of all other Cornwall individuals than with the mean copy vector of all Devon individuals. I repeated the analogous procedure to find a classification score for Devon individuals. [WHAT WERE THE CLASSIFICATION PROBABILITIES?] 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.) 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 (e.g. X reads)[WE'VE DISCUSSED THIS AT SOME POINT, BUT SOMEWHERE YOU SHOULD PUT WHAT THE PROBABILITIES OF OBSERVING A HET ARE GIVEN Y READS, I.E. Pr(BOTH ALLELES ARE OBSERVED | HET, Y READS) FOR VARIOUS Y. THIS IS EQUAL TO 1-Pr(one allele is observed | HET, Y READS) = $1 - 2 * 0.5^Y$. THUS YOU WANT 3 READS TO HAVE A 75% CHANGE OF OBSERVING A HET].

In my painting of XX world-wide samples on the Human Origins array (section XX), the average number of segments that forms a recipient genome is Y (range: XX-XX). Given a genome-wide size of ≈3000Mb, this implies that an average "chunk" size (in Mb) is 3000/Y = Z ≈500kb, where a "chunk" is a set of contiguous SNPs matched to a single donor. Therfore, for each of the 12 different

numbers of genome-wide total SNPs used in my Devon/Cornwall analysis, I can calculate the average number of SNPs per 500kb chunk, 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 chromsomes 21 and 22 (which has V 500kb chunks), etc, continuing until the classification scores were equivalent to that when analysing all 22 autosomes 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 **??**). I found results to be very similar to if chunk-size were instead defined as 250kb or 1Mb (Table **??**).

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 2.15 shows the [WHICH WINDOW SIZE DID YOU USE FOR THIS PLOT? 500KB?]mean number of windows per individual with at least Y SNPs above Z coverage, with individuals being grouped into bins based on their mean coverage. Points are coloured yellow if, within the bin of coverage, samples have at least 2000 windows[NOTE WE CAN'T TELL HOW TO READ THIS WITHOUT KNOWING THE CORRESPONDING SNP DENSITIES IN TABLE 2.2? I.E. THAT 40K SNPS CORRESPONDS TO 6-7SNPS/WINDOW. CAN YOU PUT THIS INFO IN TABLE 2.2?]. 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

n_snps	250Kb	500Kb	1Mb
40,000	7691	3879	1967
45,000	6272	3166	1607
50,000	5659	2858	1452
100,000	3602	1820	925
150,000	4083	2064	1049
200,000	4083	2064	1049
250,000	4083	2064	1049
300,000	5659	2858	1452
350,000	4507	2278	1158
400,000	4083	2064	1049
450,000	4083	2064	1049

Table 2.2: 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. ADD IN COLUMNS AFTER N_SNPS TO SAY WHATS THE NUMBER OF SNPS PER 500KB WINDOW

n_snps	250Kb	500Kb	1Mb
30,000	6272	3166	1607
35,000	3099	1565	796
40,000	3099	1565	796
45,000	2612	1321	673
50,000	3099	1565	796
100,000	1886	956	489
150,000	1304	661	338
200,000	506	255	130
250,000	267	135	69
300,000	506	255	130
350,000	506	255	130
400,000	506	255	130
450,000	267	135	69

Table 2.3: 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.

read, this suggests that there are not enough non-imputed SNPs with enough coverage to match the power seen in full coverage individuals.[NOT QUITE

CLEAR IT'S THAT BAD – FOR 0.3-0.4x SAMPLES, THERE ARE 1000 SEGMENTS WITH >=10 SNPS, WHILE TABLE 2.3 SAYS 1565 SNPS WITH >=8.3 SNPS IS ENOUGH (AND THAT'S TO ACHIEVE FULL POWER; ONE QUESTION IS WHETHER YOU CAN REDUCE THIS; I.E. WOULD BE HELPFUL TO HAVE ANOTHER TABLE SHOWING THE CLASSIFICATION RATE YOU GET WITH 500). NOTE ALSO IT WOULD BE HELPFUL TO HAVE VALUES BETWEEN 50K AND 100K, GIVEN THE RANGE COVERED BY THE X-AXIS OF FIG 2.14] Indeed, even when there are 3 reads covering a site, there is still a 25% chance of not identifying a heterozygous 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.

INSTEAD OF COMPARING TO FULL COVERAGE - COMPARED TO UNLINKED

FOR 500KB PLOT CLASSIFICATION RATE (Y-AXIS) V NUMBER OF REGIONS - HAVE A ROW FOR EACH DIFFERENT KIND OF SNP DENSITY - HOW DOES CLASSIFICATION RATE IMPROVE WHEN ADDING REGIONS - HAVE HORIZONTAL LINES FOR LINKED AND UNLINKED MODELS WITH FULL SNPS

2.7.4 Averaging across copyvectors

2.8 Discussion

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

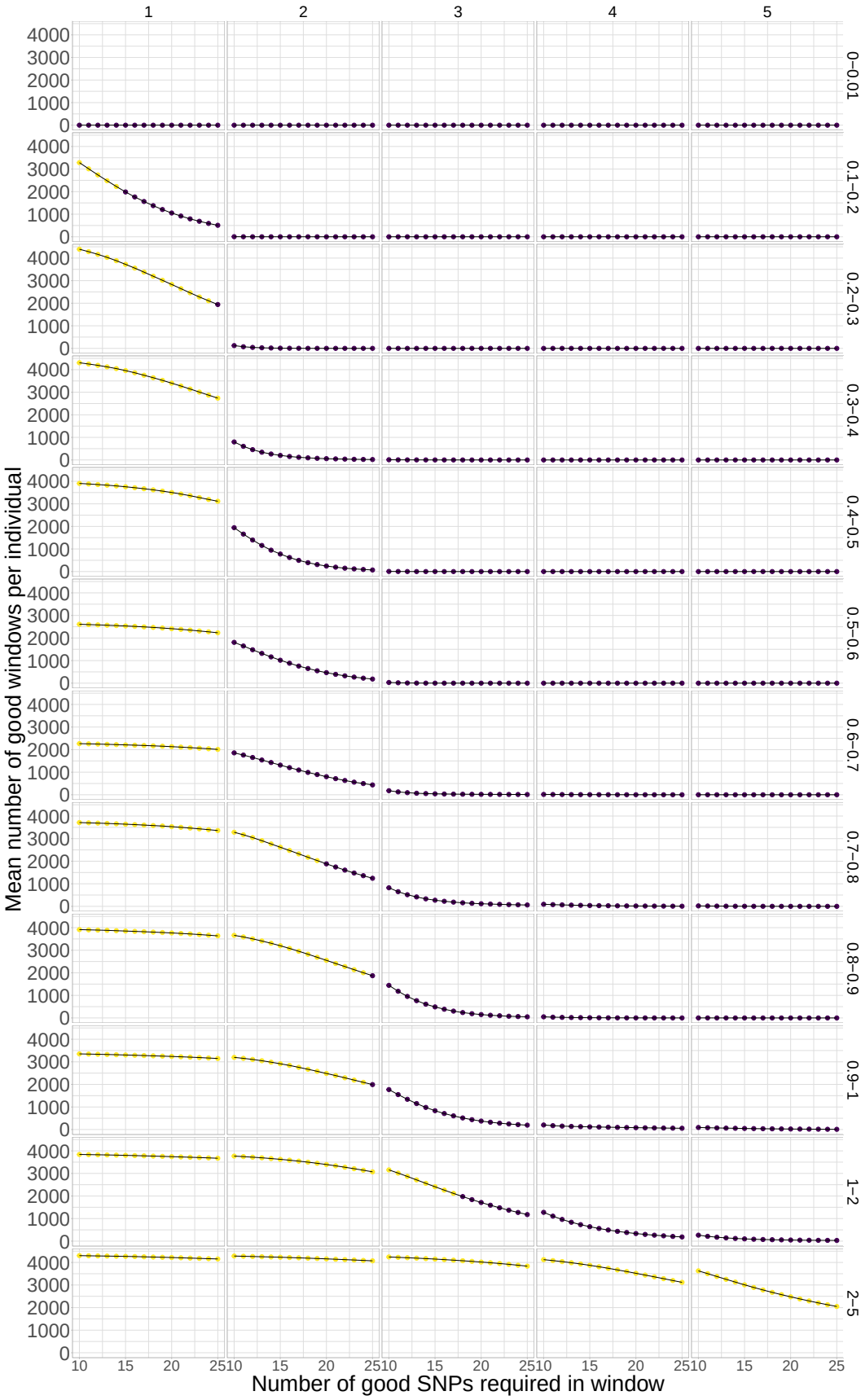


Figure 2.15: Mean number of windows (y-axis) within the genome of each ancient individuals within a given range of coverages (rows) with at least Y SNPs (x-axis) above a particular coverage Z (columns). SAY WHAT WINDOW SIZE 500kb

used a single target downsample when estimating ancestry proportions. This differs from a typical ancient DNA analysis, such as those of Margaryan et al [41], 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.

this is too harsh because 0.5x has zero SNPs above 3x, but we can use it with imputation. maybe discuss and redo figure

2.9 Summary of findings

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. [81, 91–93]). 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 [25, 81, 94].

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 [95], the transatlantic Slave Trade and economic 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 catalog 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" [96]. The primary reason for this is reasonable concerns over the confounding effect of population substructure within a cohort [97]; retaining a more genetically homogeneous cohort is one strategy to mitigate this.

However, removing ethnic minorities from GWAS analyses is problematic, as evidence is mounting that the results from GWAS, including Polygenic Risk Scores (PRS), may not be transferrable to other populations if they have been conducted in cohorts of exclusively European individuals [98–100]. The reasons for this are not yet fully understood, but it thought that differences in LD structure may be the cause [101]. 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 over traditional approaches such as PCA. Zaidi and Mathieson (2020) [102] 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 principle components did not correct for GWAS hits on birth location. However the significant hits dissappeared when using a ChromoPainter coancestry matrix, generated by painting target samples against a reference and use the resulting painting profile as a reference.

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 because [sam: not sure how to word the exact reason]. Additionally, they only considered self identified White British individuals.

Recent studies have outlined the power of haplotype-based approaches in inferring the population histories of different African ethnic groups [103–105]. 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 diverse 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.

To achieve both of these aims, I will leverage the a recently compiled dataset, hereafter referred to as 'Human Origins'. At the time of writing, it is the most detailed dataset of genotype data from African individuals in terms of the number of ethnolinguistic 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. 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 betweeen 500,000 and 700,000. Using a low number of SNPs in the analysis may reduce the power to infer accurate ancestry proportions, in particular for haplotype-based methods since haplotype information depends on SNP density. 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 [71, 106]. 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.

This chapter will focus on two questions. Firstly, evaluate the effect of using imputed genotypes on the validity of ChromoPainter analysis in African individuals. Secondly, relating the genetic variation of individuals in the U.K. Biobank with recent African ancestry to the populations in the Human Origins dataset in an attempt to infer their ancestral origins.

3.2 Results

3.2.1 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 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. Individuals with more than 50% ancestry from Yoruba would then be carried forward to later ChromoPainter analyses.

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

I validated the ADMIXTURE results to ensure that there was not any mixing of sample labels and that enough ADMIXTURE EM iterations had been performed. To do this, I selected all individuals who self-identified as being either "Caribbean", "African" or "Black or Black British" (n=7527)

and plotted the distribution of ADMIXTURE ancestry proportions, under the assumption that these individuals should contain more African than other kinds of ancestry. This was the case, with the mean proportion of African ancestry among these individuals being 0.88 (Fig. 3.1), 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 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.

3.2.2 To impute or not?

In order to use ChromoPainter and the Human Origins dataset as a reference to infer fine-scale ancestry in U.K. Biobank individuals, the datasets must be merged. The overlap of SNPs genotyped in each dataset is only 70,776 SNPS, or an average of ≈1 SNP per 40Kb. Given linkage disequilibrium (e.g. as measured by Pearson's correlation) between pairs of SNPs decays to background levels by 100Kb within most populations [107], 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

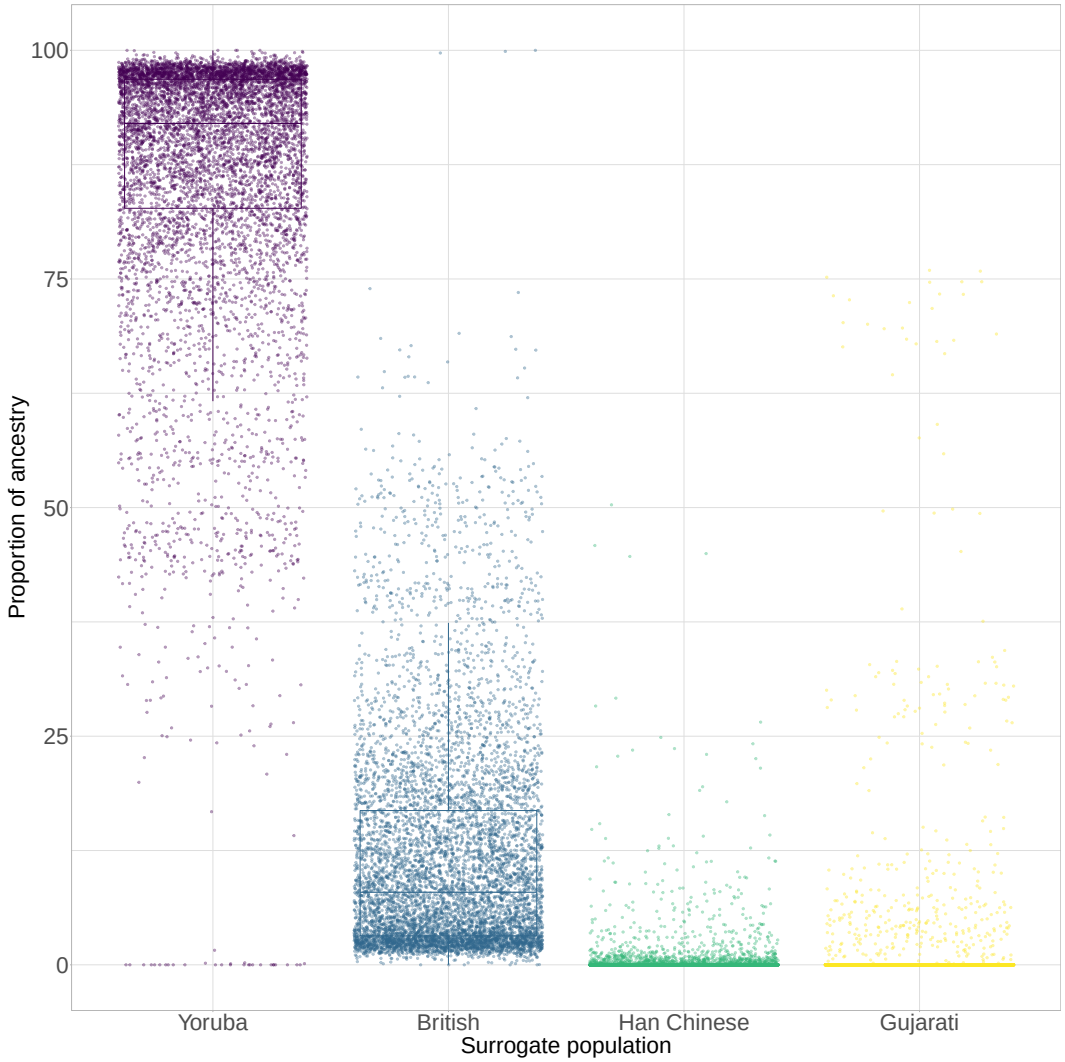


Figure 3.1: Ancestry proportions inferred from supervised Admixture run (k=4) for all individuals who self identified as being either "Caribbean", "African" or "Black or Black British". Points within each column are given random jitter to improve visual clarity.

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, and therefore determine whether using the imputed or the 70,000 SNP Human Origins dataset is better in this scenario, I performed a painting using (i) the full 560,442 genotyped SNPs, (ii) 64,762 genotyped SNPs overlapping UK Biobank, and (iii) 500,000 SNPs that include the 70,000 genotyped SNPs 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 3 datasets described above, I selected all ethnic groups from Nigeria, Cameroon and Ghana which had 5 or more individuals (n=51 populations, n=1203 individuals) and split each population randomly in half, into 'donors' and 'recipients'. I painted all recipient populations (n=51) using all donor populations (n=51) using a leave-one-out approach (motivation for this approach given in appendix B.2). I tested the information content of each painting by counting how often individuals copy more from individuals in their own populations than individuals from other populations. I also counted the number of times a population had the lowest TVD (motivation and description of TVD given in appendix X) with its own population (Table 3.1).

Populations in the 70,000 non-imputed painting matched more to and had a lower TVD with their own population than the 500,000 non-imputed painting. Whilst it seems counter-intuitive 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 after approximately 50,000 SNPs (i.e. there is no clear benefit to using more than 50,000 SNPs in terms of assinging individuals to a population). This is reassuring and suggests there is no loss of power when using the 70,000 SNP set. These data also shows that there is a fairly dramatic loss of power when using imputed data relative to non-imputed data, as over 3x the number of

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.

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 would be imputed to appear more genetically similar to that reference population, reducing the differentiation between them. In theory, this artificial similarity would be propagated through to the ChromoPainter analysis. In particular, we would expect populations present in the reference panel to donate more to all other individuals than they would if no imputation had taken place.

For example, in the case of the Haplotype Reference Consortium, the closest reference population to two African target samples from e.g. Cameroon may be the Yoruba from Nigeria, which is the only west African group in the reference. These samples would appear more similar to the Yoruba ethnic group than if they had not been imputed. In a ChromoPainter analysis, the Yoruba donor

population would donate more than than when using non-imputed SNPs.

Comparing the imputed and non-imputed coancestry matrices revealed biases consistent with the above expectation. If the coancestry matrix columns are combined into populations, then the sum of each column gives the total length of genome that population contributes to all recipient individuals in the dataset. Therefore, comparing the column sums between the imputed and non-imputed matrices informs us about which populations contribute more when using imputed compared to non-imputed SNPs. Fig 3.2 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 ($p = 0.00026$).

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

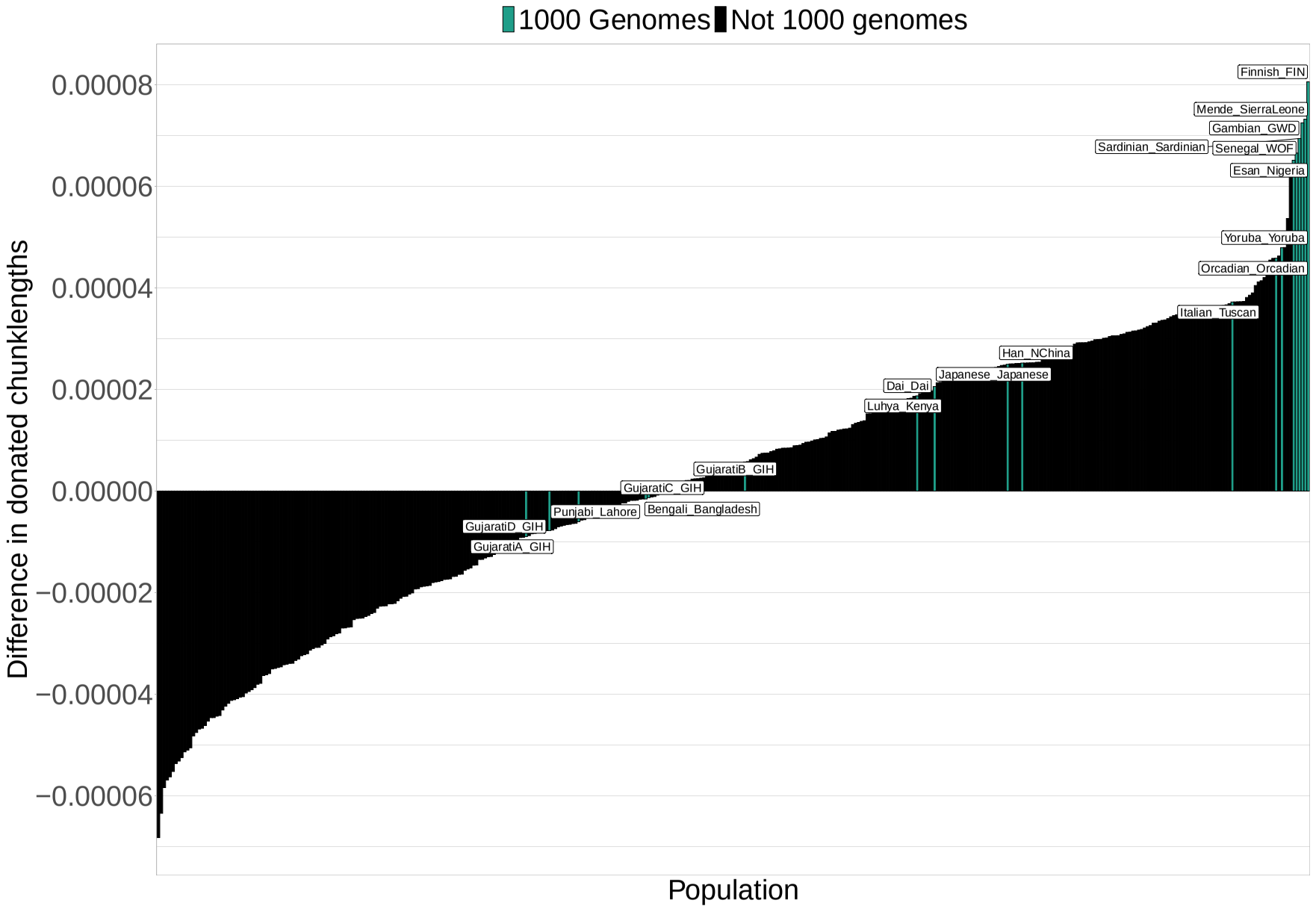


Figure 3.2: Differences in the amount donated by populations when using imputed and non-imputed data. Each vertical bar corresponds to a single population (N=395), with the height of the bar corresponding to the difference in haplotype donation, with positive values indicating increased donation and 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.

Human Origins and U.K. Biobank datasets.

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

Following from the previous section, I decided to use the dataset containing 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=5577).

Principal component analysis on the resulting chunkcounts coancestry matrix reveals the general structure of the selected individuals, alongside the reference populations (Fig. 3.3). Three clines are present; one of similarity to Southern African populations typified by the Zulu ethnic group from South Africa, one of similarity to West African populations such as Yoruba and Cameroon_Dii, and the last to East African populations such as those from Ethiopia, such as Ethiopia_Ari-Potter. 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 [108]. A second cluster of UK Biobank individuals is located along the Southern African cline, close to the Bantu_SA label.

Aggregating the columns of the co-ancestry matrix by reference population and taking the sum of each column gives the total length of genome that donor population has contributed to the selected U.K. Biobank individuals. 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. 3.4). Higher values correspond to more ancestry from that population in the

Figure 3.3: Principle component analysis of chunklengths matrix for all African U.K. Biobank individuals 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.

U.K. Biobank sample.

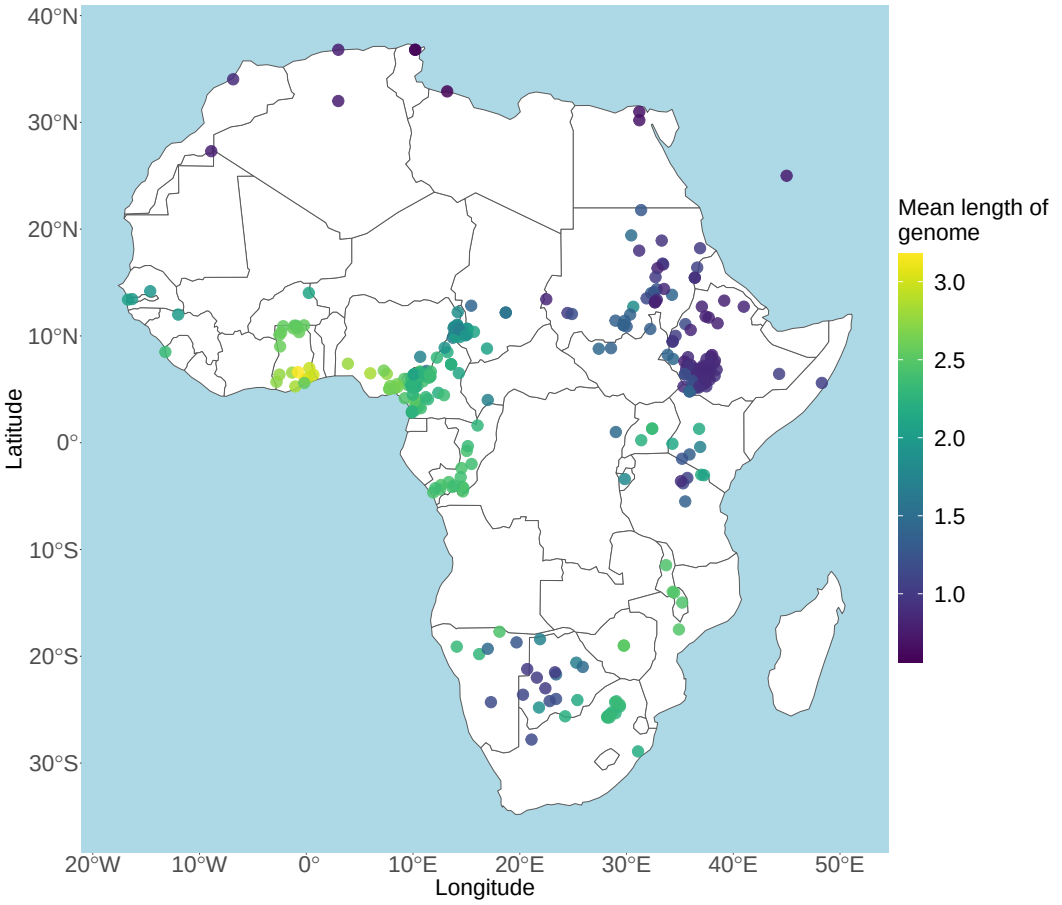


Figure 3.4: Map of haplotype donation to U.K. Biobank individuals. Each point represents a different African population. Colour corresponds to the mean length (cM) that populatation donated to all African U.K. Biobank individuals.

The map supports the findings from the PCA in Fig. 3.3; the populations with the largest contribution are those from West Africa (Fig. 3.4). 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 [109]. Given the U.K. Biobank sample contains many individuals who were either born in, or trace their ancestry from the Caribbean, a region that had a large influx of slaves [110], 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 [108]. Although there are a number of immigrants from other parts of Africa, reflected in the nonzero contributions from other ethnic groups, these contributions are small compared to those from West Africa.

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

I used SOURCEFIND to infer the proportion of ancestry that each UK Biobank individual shares most recently with each of the 535 surrogate groups, as this accounts for uneven donor population sizes. A map of proportions is given in Fig. 3.5, 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. 3.4. 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. 3.6 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; the mean proportion of Yoruba ancestry per individual was 39.80%

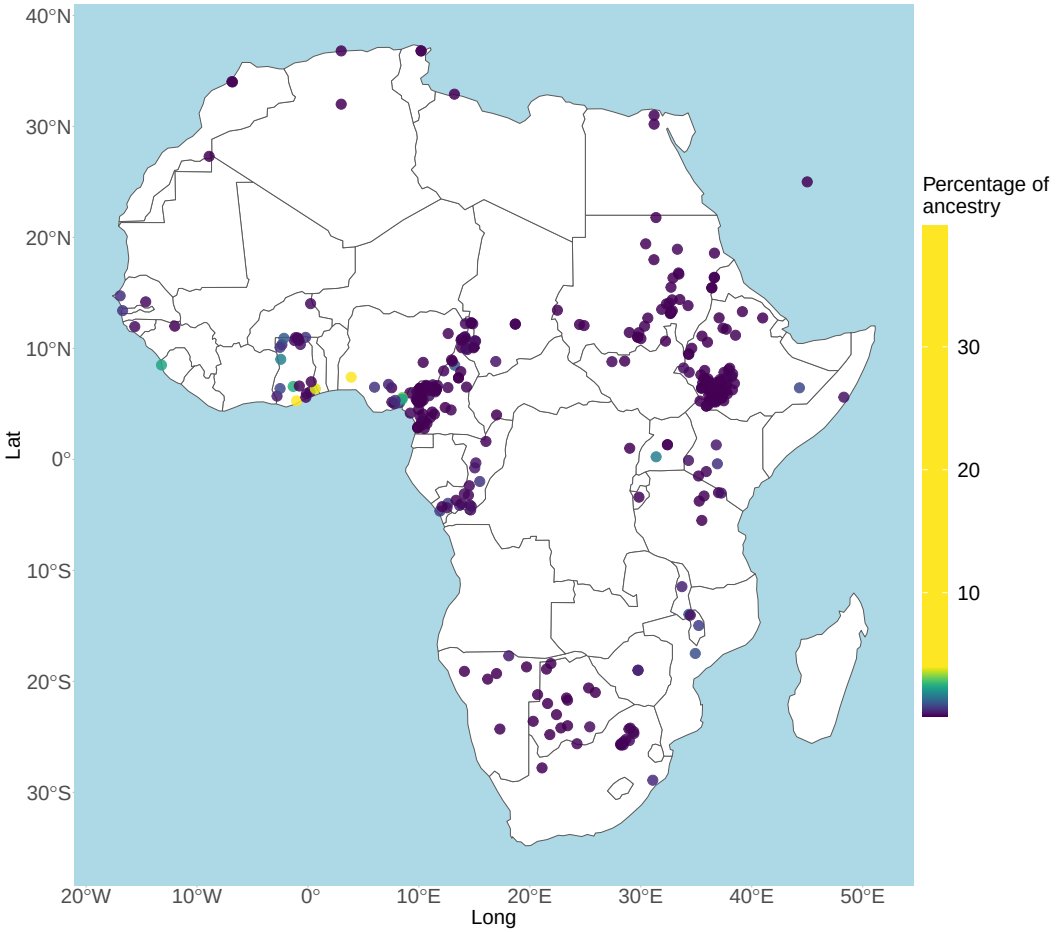


Figure 3.5: Map displaying the mean proportion of SOURCEFIND estimated ancestry of each African reference population within U.K. Biobank individuals. Each point is an African reference population with the colour corresponding to the mean ancestry proportion for that population across selected U.K. Biobank individuals. The colour-bar has been rescaled as two populations, Yoruba and Ghana_Fante have substantially higher proportions than all other populations.

and 3604/8309 individuals had at least 50% Yoruba ancestry. This is compared to the next most common ancestry, Ghana_Fante, which had an average of 7.3% per person and 373/8309 individuals with at least 50% ancestry. It is not clear what the reason for the large amount of Yoruban ancestry relative to all other populations is. One possible answer may come from considering the birth country of the U.K. participants. Of all the individuals for which we have country of birth data for (n=6190), more of them were born in the Caribbean (n=2263) relative to any other country. This should not be surprising given the history of migration from the Caribbean to the U.K. Of the individuals born in the Caribbean, over half were assigned to the Yoruban ethnicity, a much higher proportion than any other country of birth. Therefore, one could tentatively explain the abundance of Yoruba ancestry as resulting from the transatlantic Slave Trade, where individuals from the Yoruba ethnic group were taken to the Caribbean at a higher frequency than other nearby ethnic groups in the Human Origins reference. This may be in part because Yoruba is the second largest ethnic group in Nigeria and individuals 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. would 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. 3.7). For example, Uganda is dominated by a single ethnic group because there is only reference data from a single group in Uganda. 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, or an inability to distinguish 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

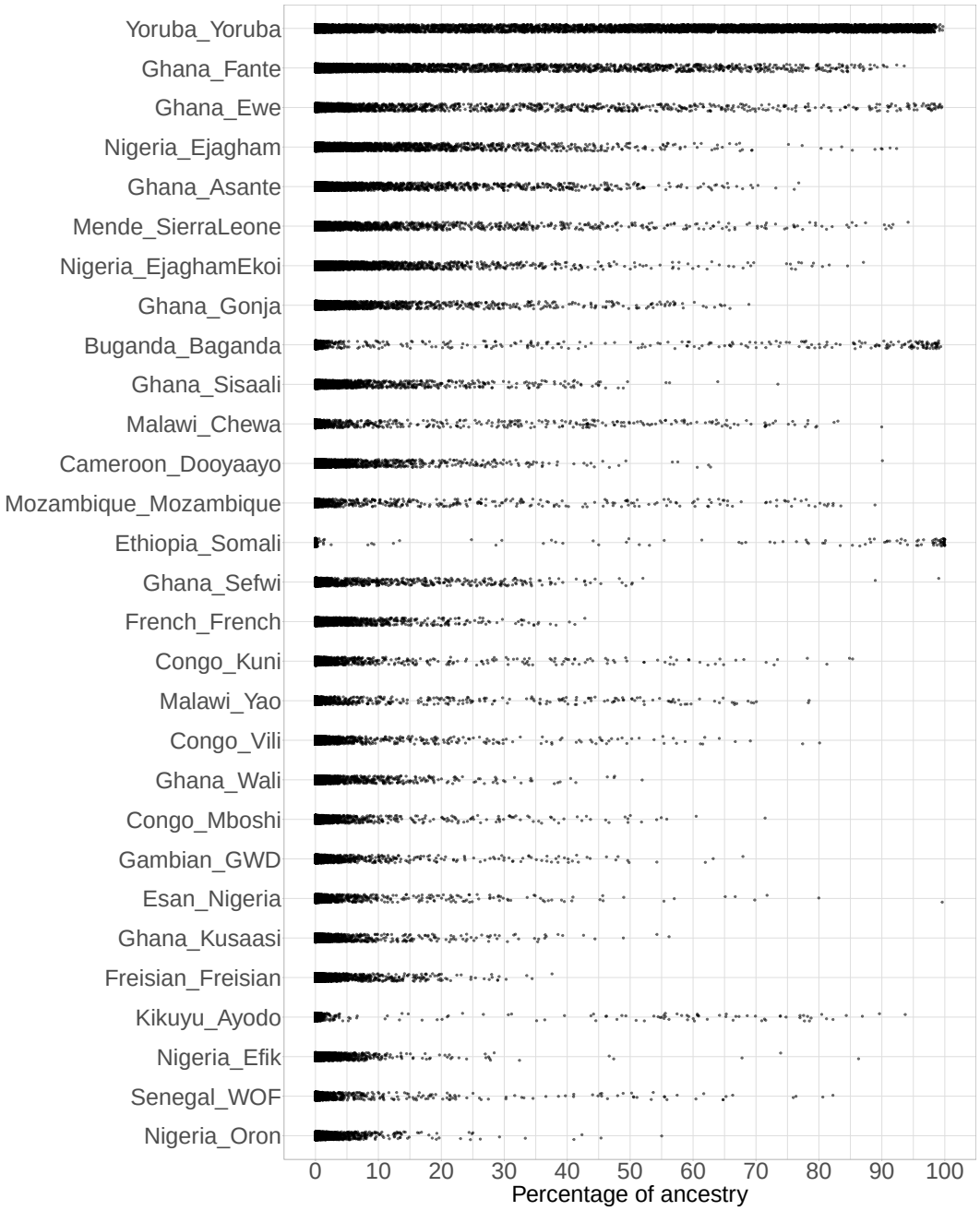


Figure 3.6: The 30 Human Origins populations which have the highest contribution to all U.K. Biobank individuals with at least 50% African Ancestry, based on SOURCEFIND analysis. Each row of points contains 8476 individuals and their position corresponds to the percentage of ancestry from that population.

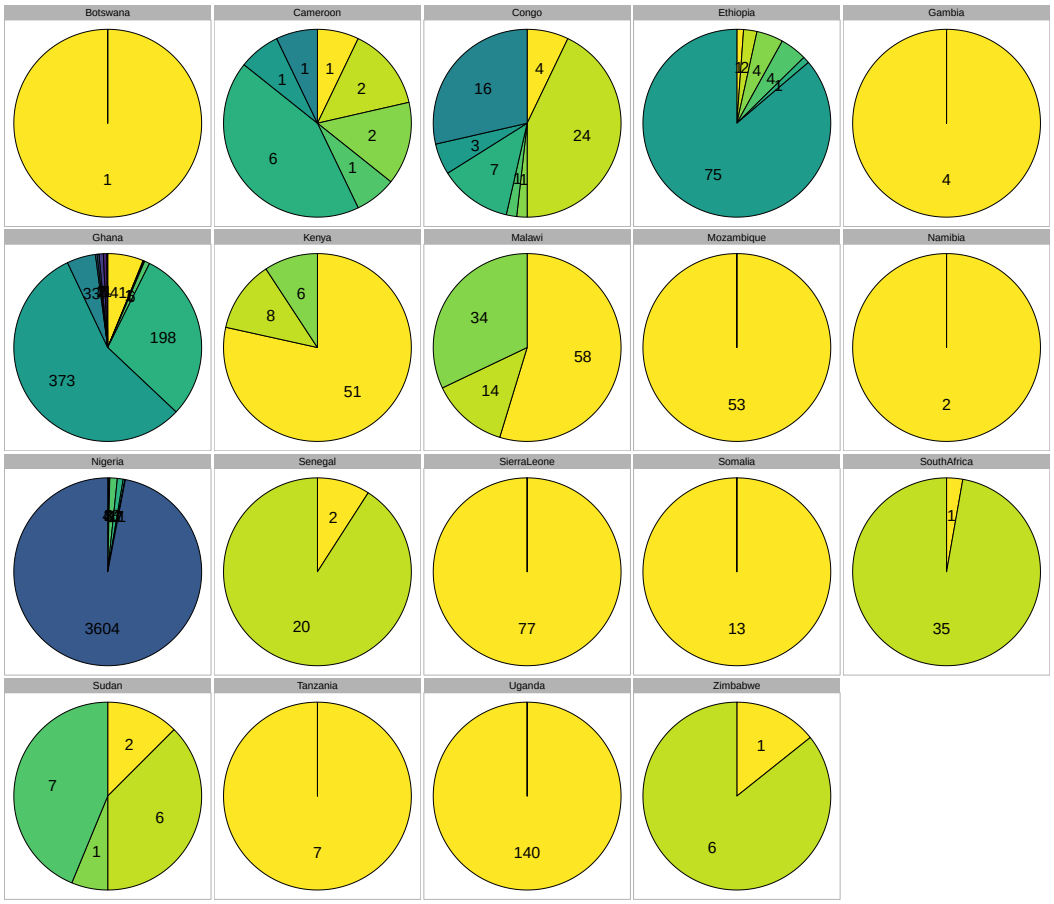


Figure 3.7: Variation of individuals assigned to different ethnic groups by country of assigned group. Each panel represents all individuals assigned to an ethnic group from that country, with proportions corresponding to different ethnic groups.

Ethiopia_Somali ancestry (Fig. 3.6). This may be because Somalis are more recent immigrants to the UK [cite] and therefore tend to be less admixed with Europeans relative to other immigrant populations which have been in the U.K. longer and hence can be modeled as a mixture of almost entirely Ethiopia_Somali ancestry.

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

3.2.4 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 us to ensure 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.

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, but it should broadly be true. We also have birth place data for individuals who were not born in Africa (e.g. the Caribbean and Brazil).

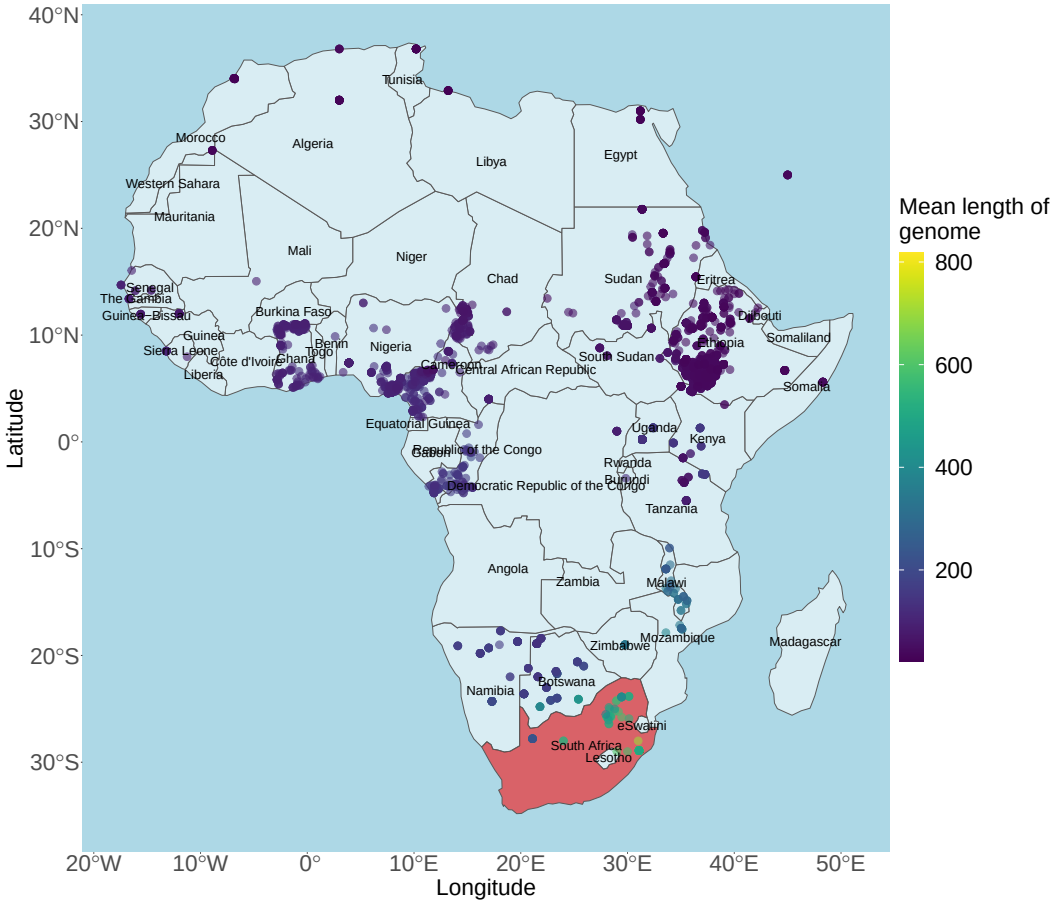


Figure 3.8: Map of haplotype donation to U.K. Biobank individuals born in South Africa. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in South Africa.

Fig. 3.8 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.

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 D3) , consistent with historical evidence [109]. Individuals born in Brazil have ancestry from further South, again consistent with historical evidence (supplementary figure D2). Of the 9 individuals born in Brazil, 6 of them had a majority SOURCEFIND component from an ethnic group in 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 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 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,

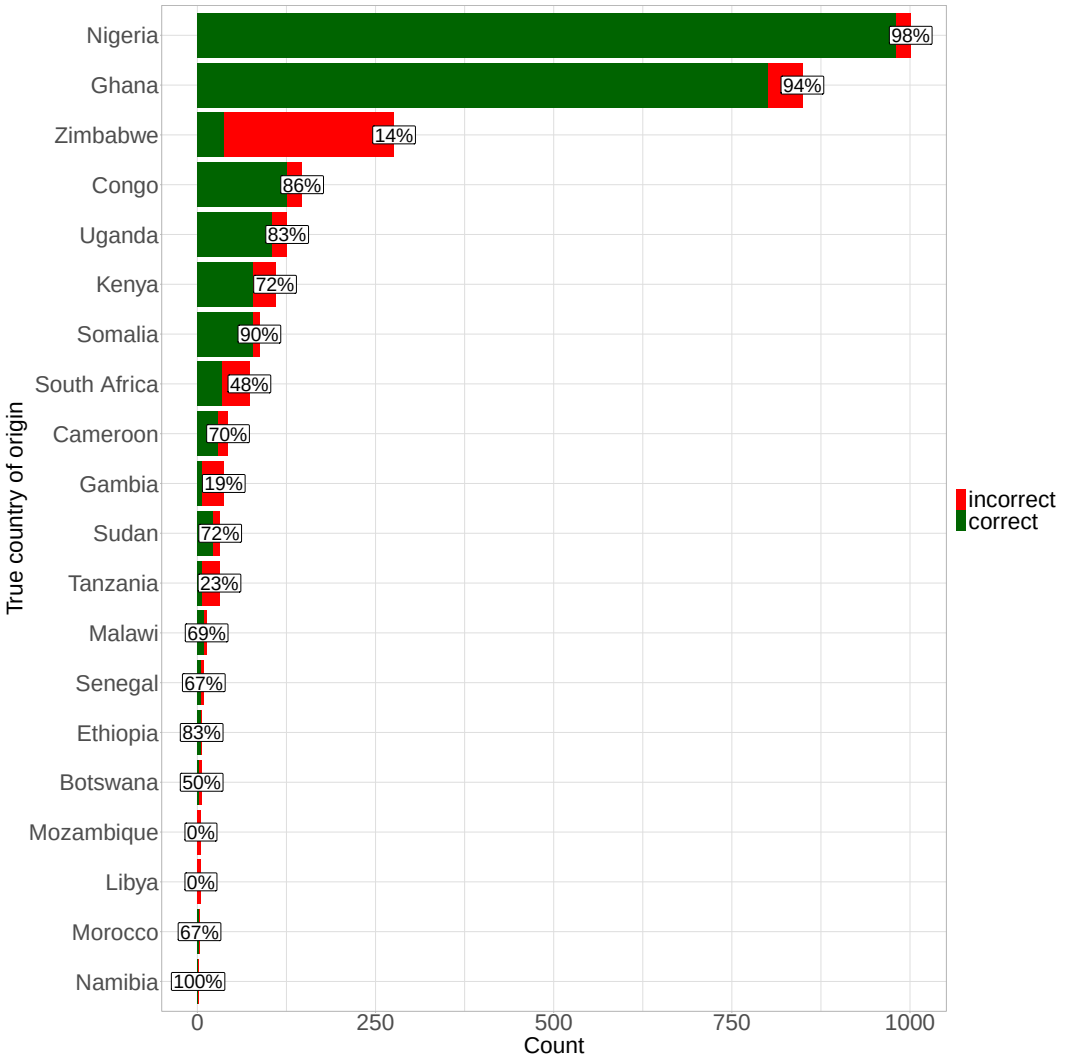


Figure 3.9: Correspondence of true birth country with estimated birth country. Each bar corresponds to a true birth country, with the length of the bar corresponding to the total number of people in our dataset born in that country. The green section corresponds to the total number of individuals where the birth country was correctly guessed and the red section to those who were incorrectly guessed. Percentage labels give percentage correct for that country.

Tanzania, which is only represented by a single reference population, had a prediction accuracy of 23%. Zimbabwe had by far the lowest prediction accuracy (14%) out of countries with more than 100 U.K. Biobank individuals. Of the 266 individuals born in Zimbabwe, 194 were assigned to an ethnic group from outside Zimbabwe; 74 to Malawi_Chewa, 71 to Mozambique_Mozambique and 49 to Malawi_Yao. Individuals from the ethnic groups from Malawi are found across Malawi, Zimbabwe and other countries, showing the possible weakness of this approach which aims to categorise individuals into a single country, as ethnic groups often transcend countries.

I performed the same analysis but using the data which had been imputed. This stands as a practical test of whether it is preferable to impute or retain a smaller number of non-imputed SNPs when estimating country-level haplotype variation. This yielded an accuracy of 81.89%, a value almost identical to that obtained with the dataset containing approximately 70,000 non-imputed SNPs, despite my earlier results indicating that sub-country SOURCEFIND results are less accurate if using imputed data due to reference bias. This suggests that broad-scale ancestry assignment (i.e. assigning individuals to countries) is not affected by imputation, whereas sub-country estimates are.

3.2.5 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 different ethnicities are distributed across the U.K. There was no apparent outliers in terms of centers and the proportion of individuals who had at least 50% African ancestry (Supplementary Fig. D.4). However, as expected, centers in large cities such as Barts, Croydon and Hounslow (London), Birmingham and Manchester had the highest proportion of individuals with at least 50% African ancestry.

I then plotted the distribution of different ethnic groups at different centers

on a map of the U.K (Fig. 3.10). No clear pattern was apparent, other than Yoruban ancestry dominating most centres, with some smaller testing centers only containing individuals assign to the Yoruba ethnic group.

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 [111], who used fineSTRUCTURE clusters

To evaluate the extent to which individuals assigned to each ethnic group registered at different testing centers, I used group label information to calculate the per group entropy statistic. To calculate entropy, for each center j we calculate $p_{i,j}$, the probability that an individual from a testing center j was from a particular ethnic group i, as: $p_{i,j} = \frac{m_{i,j}}{m_j}$, where m_j is the number of individuals from testing center j and m_{ij} is the number of ethnic groups to which individuals from center j are assigned. The entropy of each center is then calculated using the standard formula, $\sum_{i=1}^{L}[p_{i,j} \cdot log(p_{i,j})]$, given by Schütze et al (2008) [112], where L is the total number of ethnic groups. Testing centers in large cities such as London and Birmingham had the highest information entropy, consistent with prior expectations that large cities would contain a higher diversity of ancestries (Fig. 3.10).

3.2.6 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. 3.11. 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 implemented using the `rollyapply` function from the `zoo` R library, I calculated the mean age and mean proportion of all ancestries across ancestry for each bin. Fig 3.11 shows the increase of Buganda ancestry over time.


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

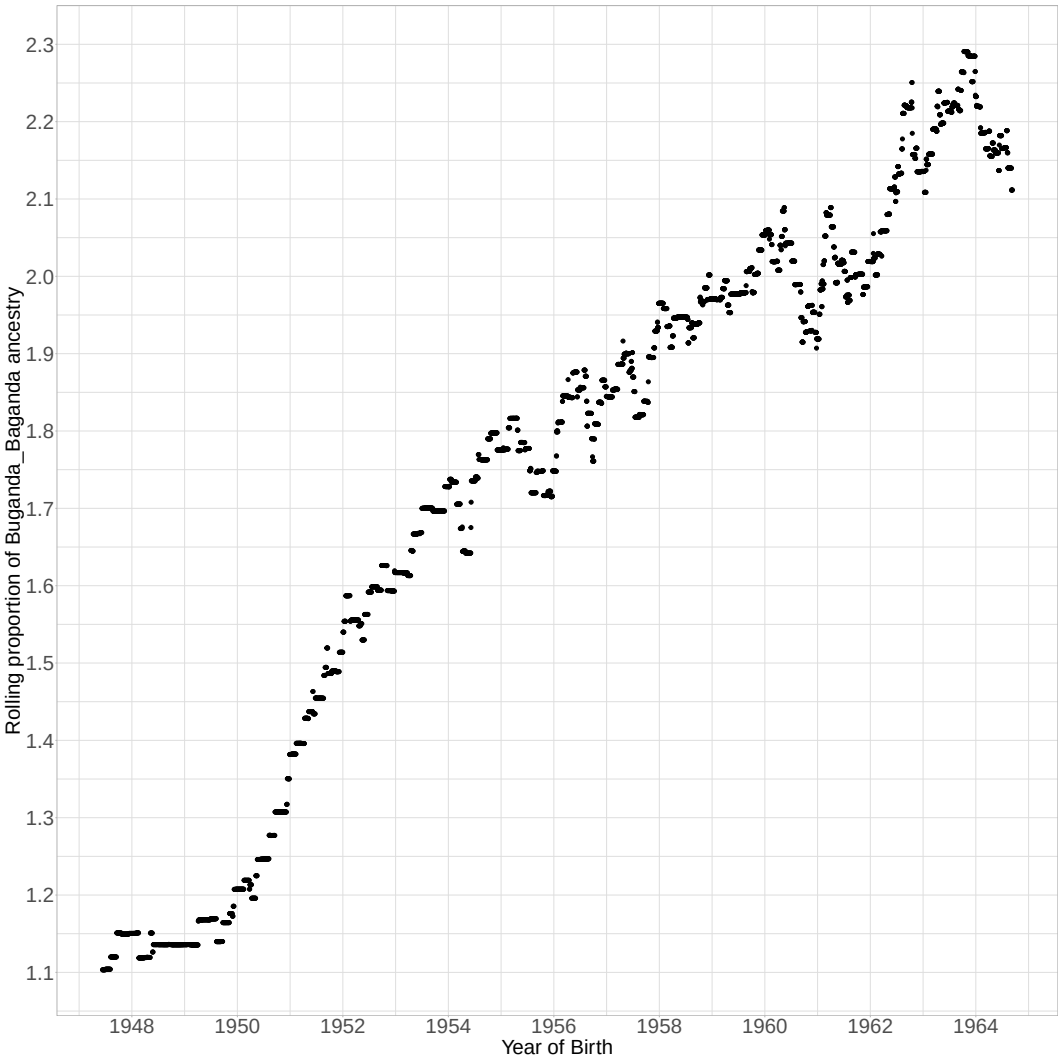


Figure 3.11: Increase in the mean proportion of Buganda ancestry between 1948 and 1965. An overlapping sliding window was applied to SOURCEFIND ancestry proportions and mean proportion of Buganda ancestry for each window plotted against the mean birth-date of individuals in that bin.

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.

3.3 Discussion

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

Future work on using Biobanks to explore population structure and history could focus on two points. Firstly, development of efficient methods to paint a single sample using a reference panel containing many thousands of samples, which also scales to Biobank-scale sample sizes (100,000+). Secondly, larger reference panels of worldwide populations and more ethnic groups will allow for a more detailed characterisation of genetic variation.

3.4 Methods

3.4.1 U.K. Biobank data access and initial processing

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

I obtained the U.K. Biobank genotype data, consisting of 488,377 individuals genotyped at 784,256 genome-wide SNPs on the U.K. Biobank Axiom Array. I will hereafter refer to these data as the 'non-imputed' data, as all SNPs were directly genotyped with imputation. I used plink2 [113] to convert the binary plink files to `.bcf` format.

I also obtained U.K. Biobank data, which had already been imputed to approximately 96m SNPs using the combined references of the Haplotype Reference Consortium (HRC) and UK10K haplotype resource. I will hereafter refer to these data as the 'imputed' data. Full details of imputation can be found in the paper of McCarthy et al (2016) [73]. 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.

3.4.2 ADMIXTURE analysis

I am primarily interested in using ChromoPainter [15] to explore the ancestry of ethnic minorities in the U.K. Biobank. However performing ChromoPainter analysis on the entire U.K. Biobank dataset (n=488,377 individuals) is computationally infeasible. Thus, I chose to analyse only those individuals with more than 50% non-European ancestry. ADMIXTURE is a fast and accurate way to estimate continental-scale ancestry proportions [80] and is therefore ideal for this task.

I LD-pruned the non-imputed U.K. Biobank dataset using using `plink --indep-pairwise 50 10 0.02` [113]. This left a total of 70,776 bi-allelic SNPs. I then subsetted the 1000 Genomes dataset down to the 70,776 SNPs retained in the U.K. Biobank dataset and merged the two datasets using `bcftools`

`-merge`. Thus, I had a dataset containing all U.K. Biobank and 1000 Genomes individuals, genotyped at 70,776 SNPs.

I ran ADMIXTURE in supervised mode using the argument `-supervised` and fixed the 4 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 piece of analysis, it was not necessary to include finer-scale populations. The rest of the arguments were left to default. I used the resulting `.Q` files to determine the ancestry proportions of each reference population in each U.K. Biobank individual.

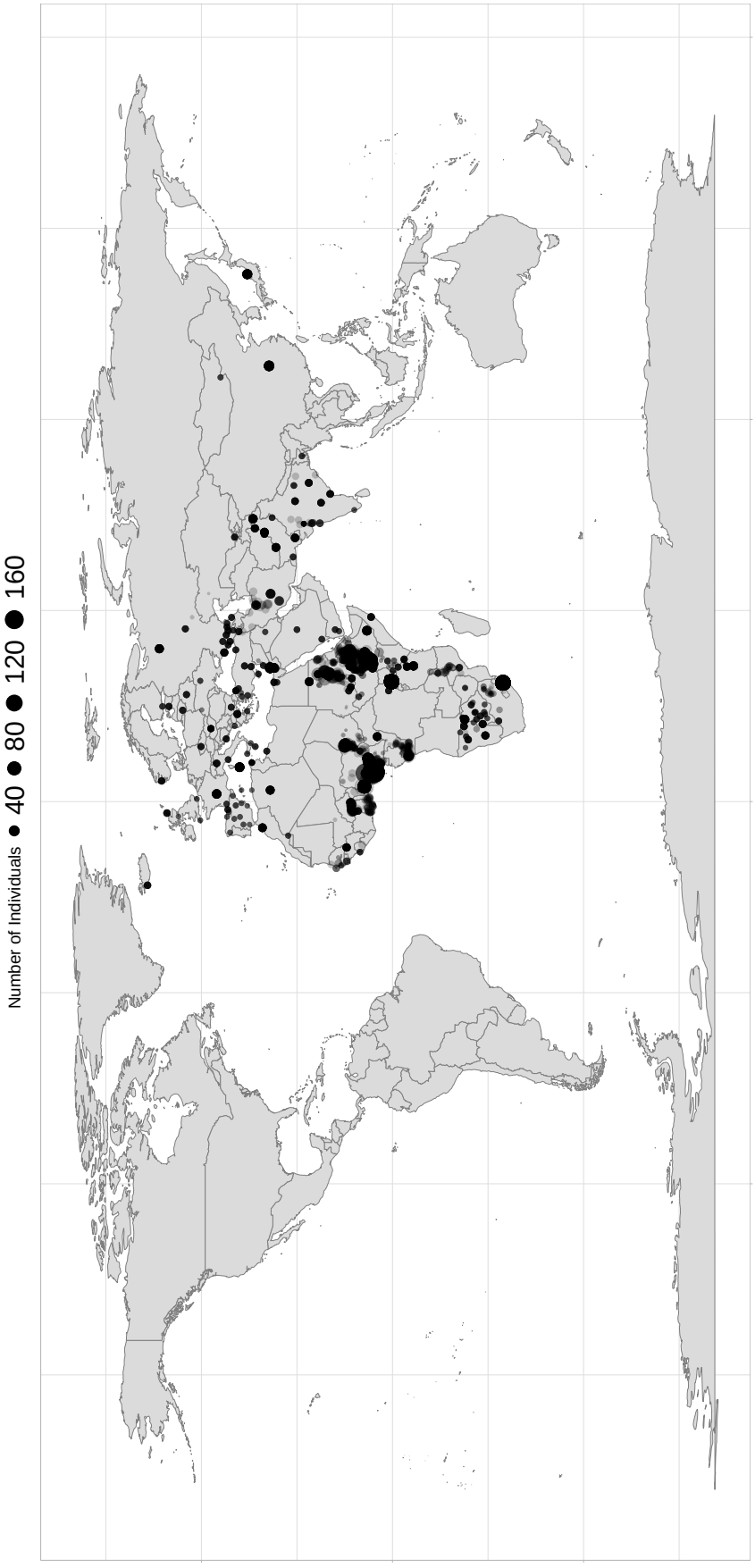
Individuals with at least 50% ancestry from Nigeria Yoruba were carried into later analysis; I refer to these as 'selected' Biobank individuals.

3.4.3 Data preparation - Human Origins

To determine the ancestry of U.K. Biobank individuals, I compared their SNP patterns to populations/ethnic groups from different parts of the world to infer which populations they share recent ancestry with. As I am particularly interested in studying individuals with recent African ancestry, I used the so-called "Human Origins" reference dataset (appendix A.20) for this purpose, as it contains individuals from 349 different ethnic groups from across Africa and 535 world-wide groups in total (Fig. 3.12). Full details of processing can be found in Appendix A.20 (Human Origins dataset).

3.4.4 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 $\geq$50% African ances-


Number of Individuals
40
80
120
160

try, using the gt-conform utilty from Beagle (`https://faculty.washington.edu/browning/conform-gt.html`) to remove any inconsistent positions. This dataset contained 65,749 non-imputed SNPs that overlap between the Human Origins and UK Biobank arrays. I phased these data with shapeit4 [71] using `-pbwt-depth 8`, the b37 genetic map and otherwise default parameters.

3.4.5 Data preparation - imputed data

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

3.4.6 Chromopainter

For both of the imputed and non-imputed datasets, I used CHROMOPAINTER to infer the proportion of genome-wide DNA that each UK Biobank and Human Origins reference individual matches to individuals from each Human Origins reference population. Using this CHROMOPAINTER output, I then used SOURCEFINDv2 [17] to infer the proportion of ancestry that each UK Biobank individual shares most recently with each of the 553 Human Origin reference populations.

An alternative option to using Origins would be to use PBWT (positional Burrows-Wheeler transform) paint `https://github.com/richarddurbin/pbwt/blob/master/pbwtPaint.c`), a fast approximation to ChromoPainter which provides approximately the same output and is scalable to large sample sizes [114]. 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.

3.4.7 SOURCEFIND

I estimated ancestry proportions for each of the selected U.K. Biobank individuals using SOURCEFINDv2 [17]. 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.

3.4.8 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 [115]. 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.

A natural test case would be to compare a painting on the U.K. Biobank individuals using imputed and non-imputed SNPs. However, this is not possible for two reasons. Firstly, the imputed dataset contains approximately 500,000 SNPs, whereas the non-imputed dataset only contains 70,000 SNPs. Having an order of magnitude difference in the number of SNPs means it would not be possible to disentangle the combined effects of SNP count and imputation on the estimates of bias. Secondly, 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 estimate 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 a) the

imputation and b) the total number of SNPs used in the analysis. I submitted the full Human Origins reference dataset (5998 individuals and 560,420 SNPs) to the Sanger Imputation Server (`https://imputation.sanger.ac.U.K./`), which uses the full Haplotype Reference Consortium (HRC) as a reference panel for imputation. This reference panel was chosen because it was the same one used for imputing the U.K. Biobank individuals.

I subsetted the imputed Human Origins dataset down to SNPs present in the U.K. Biobank array, leaving 727,325 positions present in the imputed Human Origins dataset and then randomly removed SNPs until 500,00 remained. Although the number of SNPs still differ, my previous research in Chapter 2 shows that increasing the number of SNPs beyond 400,000 does not affect the ability to correctly assign individuals to populations. I phased the imputed and non-imputed datasets separately using shapeitv4 at default settings.

To answer these questions, and therefore determine whether using the imputed or the 70,000 SNP Human Origins dataset is better in this scenario, I performed a painting using (i) the full 560,442 genotyped SNPs, (ii) 64,762 genotyped SNPs overlapping UK Biobank, and (iii) 500,000 SNPs that include the 70,000 genotyped SNPs 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 3 datasets described above, I selected all ethnic groups from Nigeria, Cameroon and Ghana which had 5 or more individuals (n=51 populations, n=1203 individuals) and split each population randomly in half, into 'donors' and 'recipients'. I painted all recipient populations (n=51) using all donor populations (n=51) using a leave-one-out approach (motivation for this approach given in appendix B.2). I tested the information content of each painting by counting how often individuals copy more from individuals in their own populations than individuals from other populations. I also counted the number of times a population had the lowest TVD (motivation and description

of TVD given in appendix X) with its own population (Table 3.1).

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 one of the earliest symbolic art, dated to at least 32kya [116] and musical instruments dated to 40kya [117], both assigned to the Aurignacian tradition. Later, the region was also home to one of the first Neolithic traditions in the *Linearbandkeramik*, a key culture in the Neolithisation of Europe.

Cherry-Tree cave (Kirschbaumhöhle in German) represents an unique opportunity to study the transect of samples from the Neolithic to the present-day. The cave represents a relatively untouched layer of stratigraphy.

In the present-day, Germany represents a boundary point between East and West Europe; [???]current population structure [PRESUMABLY YOU'RE ADDING DETAILS, + REFS, HERE? THERE IS SOME INFO HERE IN LESLIE ET AL 2015 FIGURES]. 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?

Here, I present novel data from 11 medium-to-high coverage samples from two sites from Southern Germany and one site from one 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.

Previous studies into the genomic history of Bavaria have focused, for example, on the mixed ancestry of migrant females during the Early Middle Ages.

A collaborator, Prof. Joachim Burger, Johannes Gutenberg University Mainz, posed the following 3 questions.

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

2. **Cherry Tree Cave**. How can we make sense of the genetic ancestry changes from the Late Neolithic through to the Iron Age in Cherry Tree Cave? Do we see evidence of genetic continuity between the ages and are they characterised by admixture from outside sources?

3. **Germanic / Slavic divide**. Is there a distinction between the Germanic and Slavic 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. Their estimated

radiocarbon dates range from 1060AD to 5200BC (Fig. 4.2). Six of the samples were found in Cherry-Tree Cave in the Bavarian district of Forchheim (Fig. 4.1), four from futher South in the region of Dingolfing/Essenbach and one sample from southern Austria. 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.

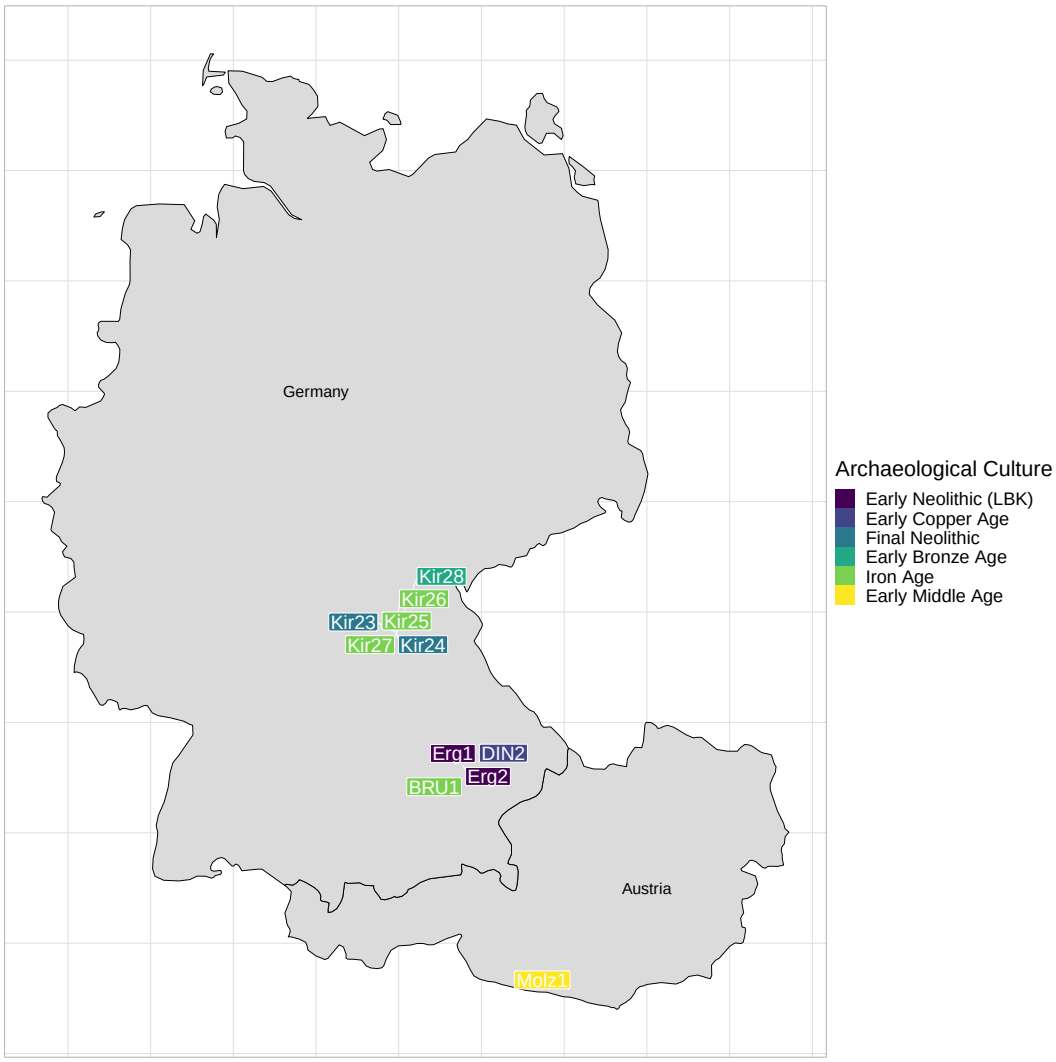


Figure 4.1: Map of newly sequenced ancient individuals, positioned according to where they were excavated. Colour on label corresponds to archaeological culture which they were found.

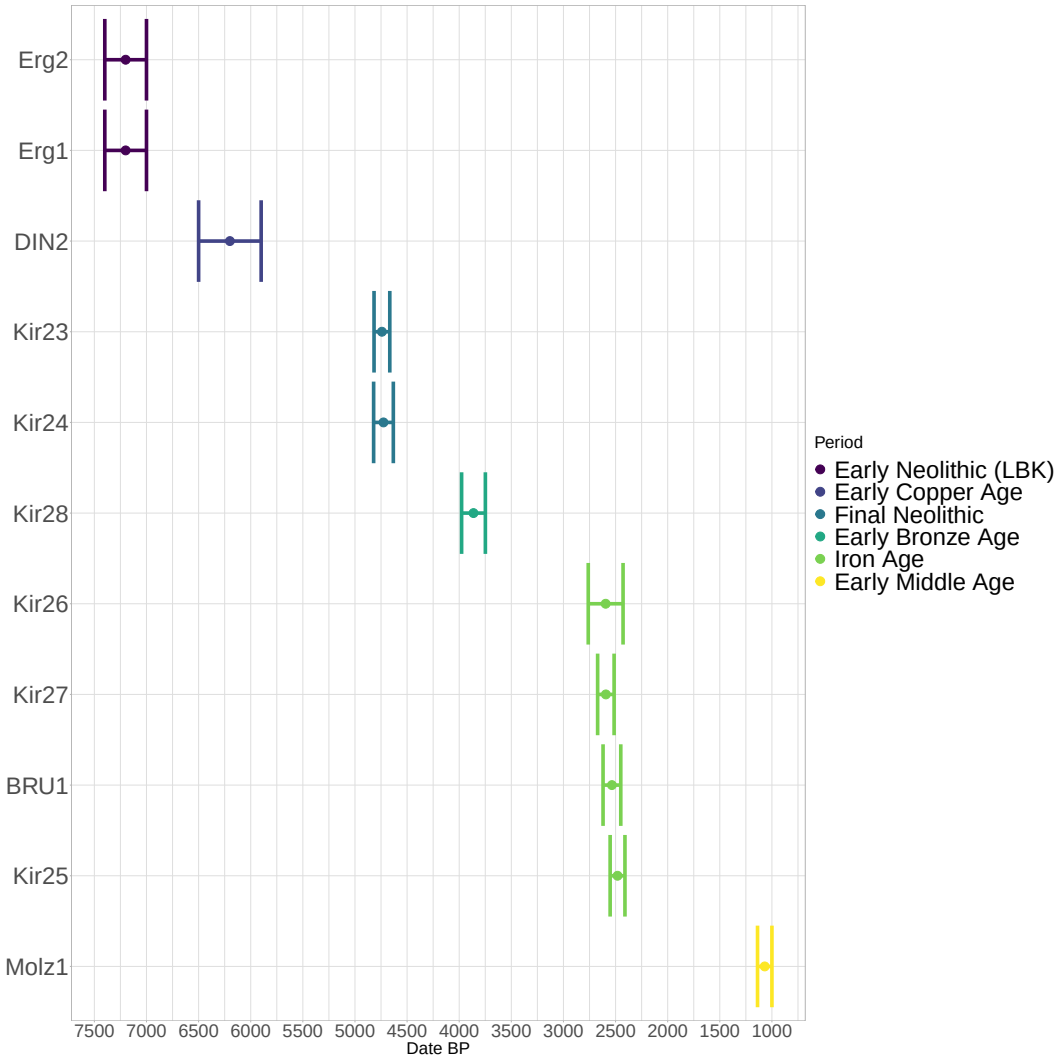


Figure 4.2: Estimated radiocarbon dates for each newly sequenced ancient individual, grouped by archaeological period.

Sample.ID	Location	Date	Period	Coverage
Erg1	Ergoldsbach-Essenbach	5200	Early Neolithic (LBK)	4.52
Erg2	Ergoldsbach-Essenbach	5200	Early Neolithic (LBK)	0.71
DIN2	Dingolfing	4200	Early Copper Age	1.71
Kir24	Cherry Tree Cave	2762	Final Neolithic	3.98
Kir23	Cherry Tree Cave	2741	Final Neolithic	17.52
Kir28	Cherry Tree Cave	1863	Early Bronze Age	17.30
Kir26	Cherry Tree Cave	595	Iron Age	4.84
Kir27	Cherry Tree Cave	593	Iron Age	16.60
BRU1	Bruckberg	535	Iron Age	11.54
Kir25	Cherry Tree Cave	481	Iron Age	4.55
Molz1	Molzbichl	1069	Early Middle Age	13.22

Table 4.1: Table providing details for the newly sequenced Bavarian samples.

4.2.2 Stuff that Jens did (e.g. read aligning)

Collaborators in University of Mainz performed DNA extraction, read alignment and variant calling. [OBVIOUSLY HAVE TO CHANGE TITLE, AND ADD MORE DETAILS.]

4.2.3 Genotype imputation and phasing using GLIMPSE

I merged the 11 newly sequenced individuals with the reference data-sets A.1 to A.17 resulting in a total of 942 individuals in `.bcf` format with genotype likelihood data at 77,213,942 genome-wide SNPs. Data was then split into separate `.bcf` files for each chromosome and indexed using bcftools [118].

As before, I followed the recommended GLIMPSE [63] imputation and phasing pipeline (`https://odelaneau.github.io/GLIMPSE/tutorial_b38.html`), here using the 30x-coverage 1000 genomes dataset [75] as a reference when phasing and imputing data in these 11 newly sequenced individuals and 942 individuals from published datasets A.1-A.17 in Table XX. I used default settings to split the genome into windows for efficient imputation/phasing, in the end generating phased samples with calls at 50,342,061 total bi-allelic autosomal SNPs.

4.2.4 Determination of uniparental haplogroups

I used Haplogrep (`https://haplogrep.i-med.ac.at/`) to identify the mtDNA and y-chromosome haplogroups for each newly sequenced ancient samples [119] from the raw `.fastq` files.

4.2.5 IBD sharing

I used hap-IBD [120] to estimate IBD segments between all pairs of ancient individuals, 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 default parameters. I estimated IBD segments for each chromosome separately and summed their length segments between each pair of individuals across all chromosomes.

4.2.6 plink PCA

To obtain a broad overview of the ancestry of the newly sequenced individuals in the context of 915 other ancient samples, I performed PCA on the pre-imputation genotypes using plink2. Performing a PCA in plink2 allows for both an understanding of genome-wide variation patterns and the identification of any data quality issues that are independent of phasing or ChromoPainter analysis.

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.

4.2.7 Chromopainter analysis

To characterise of the ancestry of the newly sequenced ancient samples in the context of other ancient individuals, I first selected all ancient samples above 1.5x coverage (n=466) and performed an 'all-v-all' painting where each phased haploid was matched to all other phased haplotypes. 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 2.5.4). I used this painting, hereafter referred to as 'ancient' painting, to perform fineSTRUCTURE clustering and tree building on ancient samples.

I performed Principle Component Analysis on the coancestry matrix of the 'ancients' painting using the `prcomp_irlba` function from the irlba R libary. To account for the fact that the diagonals of the coancestry matrix are always

Population	nsamples
HB:tsi	196
HB:spanish	68
HB:bulgarian	62
HB:german	60
HB:french	56
HB:russian	50
HB:greek	40
HB:ukrainian	40
HB:croatian	38
HB:hungarian	38
HB:norwegian	36
HB:southitalian	36
HB:polish	34
HB:romanian	32
HB:mordovian	30
HB:cypriot	24
HB:northitalian	24
HB:lithuanian	20
HB:siciliane	20
HB:westsicilian	20
HB:belorussian	18
HB:tuscan	16
HB:irish	14
HB:scottish	12
HB:germanyaustria	8
HB:welsh	8

Table 4.2: Name of population and number of samples used in the present-day ChromoPainter analysis

zeros (as an individual cannot be painted by themselves), I set the diagonal of each row to be the mean of that row. 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 XX 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 XXX) plus the newly sequenced ancient individuals, hereafter referred to as 'present-day painting'.

I applied fineSTRUCTURE (v0.0.5) [15] to cluster the chunkcounts ChromoPainter output for the 'ancient' painting. This algorithm assigns individuals to genetically homogeneous clusters, estimates the 'true' 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, co-ancestry matrix figures and principle component plots were generated using the fineSTRUCTURE R library (`https://people.maths.bris.ac.uk/~madjl/finestructure/FinestructureRcode.zip`).

4.2.8 SOURCEFIND

I used SOURCEFIND [17] to infer the proportions of ancestry by which each target (e.g. ancient) individual is most related to a set of surrogate populations. While this method does explicitly attempt to identify admixture, in contrast to (e.g.) ALDER [121] or GLOBETROTTER [16], it can reflect admixture proportions [17] but more generally reflects recent ancestry sharing patterns.

The first analysis used the ancients painting and only three surrogates: Western Hunter-Gatherers, Neolithic farmers and Yamnaya. This analyses

reflects previous research suggesting most ancient Europeans, with the exception of some paleolithic Hunter-Gather populations [88], descend from the mixture of three sources well-represented by these groups. The second analysis attempted to characterise more fine-scale ancestry patterns, by modeling each target ancient individual (using the same ancients painting) as a mixture of all sampled ancient populations 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".

4.2.9 MOSAIC admixture analysis

I inferred admixture events, dates and proportions using MOSAIC, a haplotype-based method [122]. While MOSAIC cannot infer multiple pulses of admixture from the same admixing sources as GLOBETROTTER can in theory, it is unlikely we have power to identify such multiple pulses when analysing only a single ancient sample. 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 different kinds of admixture analysis. First, I performed an 'ancient surrogates' analysis where the all ancient samples above 1.5x coverage were used as surrogates to admixing sources. I used the fineSTRUCTURE groupings to categorise ancient samples into surrogate populations.

I also performed a 'present-day surrogates' analysis where a selected set of present-day populations were used to analyse both present-day Slavic popula-

tions and ancient Slavic populations. 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 refined 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 x% confidence intervals 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

4.2.10 F-statistics

Many of the relevant samples in the literature were of either very low coverage (< 0.1) or genotyped on a capture array. As my work in chapter 2 indicated that samples with less than 0.5x coverage cannot reliably be analysed using ChromoPainter, I also used F-statistics [29] that are mostly robust to coverage related effects [31]. In particular I used Admixtools (`https://uqrmaie1.github.io/admixtools`) to analyse 942 individuals from 143 populations (Appendices A1-A18), including many low-coverage samples from LBK cultures from Rivollat et al (2020) that would not have been suitable for use with ChromoPainter [123]. This analysis also incorporated 2280 present-day individuals from 144 populations from the HellBus dataset as putative ancestry surrogates for tested ancient individuals.

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 [35]. 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

As expected, on the first two principal components samples from the Early Neolithic (approx 5200BC) and Copper Age (approx 4200BC) cluster with other samples from the European Neolithic (Fig. 4.3). Previous studies have explained the pattern observed when Neolithic samples are plotted on a PCA [124]; the earliest Neolithic samples, from Anatolia and Greece, who are thought to be the source population from which all subsequent Neolithic farmers derive [37, 125–128], are usually positioned at the end of the cluster which is farthest away from the hunter-gatherer samples (for example, WHG on Fig. 4.3). This likely reflects the fact 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 [124] 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 Erg1, shifted north away from the contemporaneous sample Erg2, is suggestive of hunter-gatherer admixture.

There are four key observations from the ancient 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 Europeans.
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 3 early/middle Neolithic samples 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 [37, 126–128]. fineSTRUCTURE analysis groups Erg1 with 2 samples from Upper Paleolithic/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;

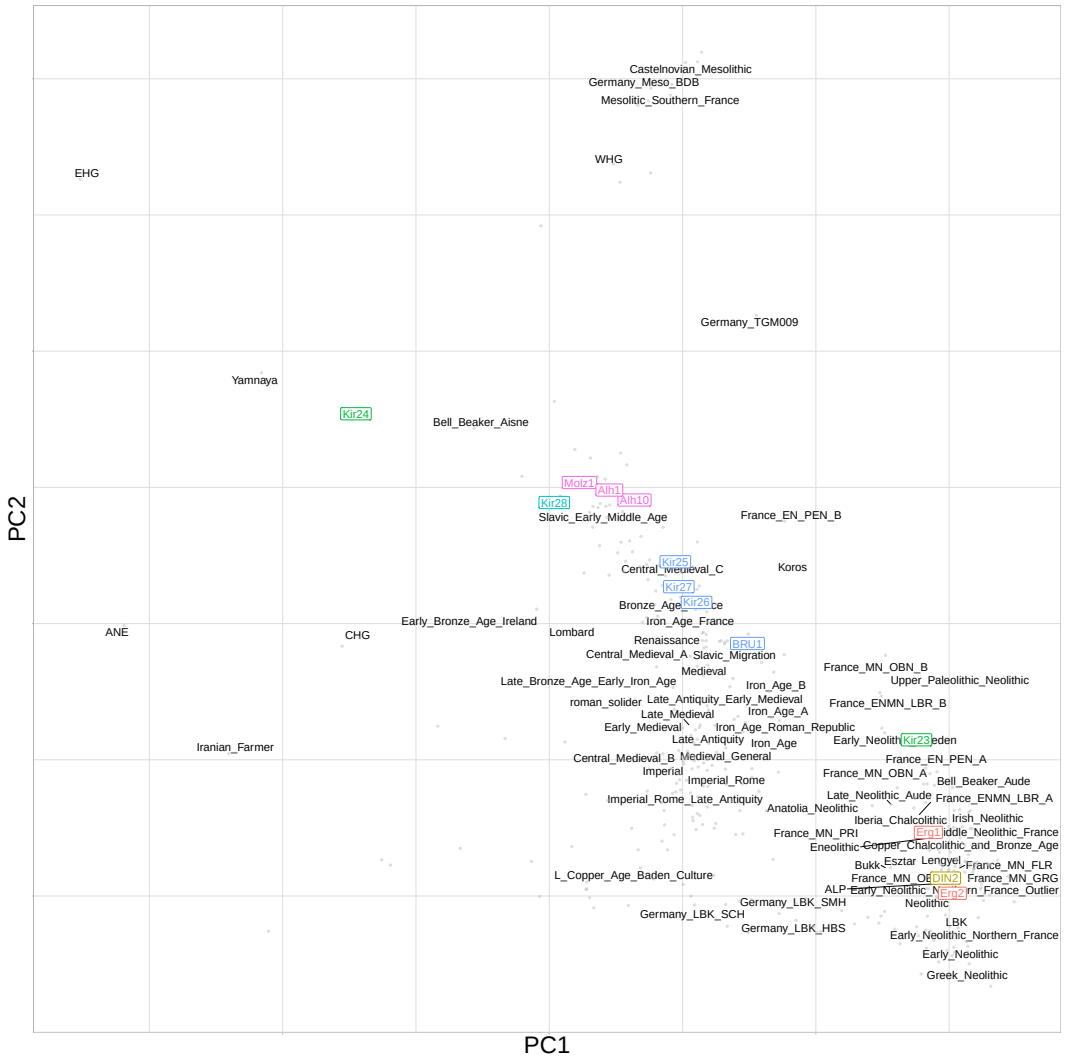


Figure 4.3: Principle component analysis of genotype matrix using plink2. Grey points indicate principle component coordinates for each sample. Black text indicated mean principle component coordinates for all individuals within that group. Coloured labels represent newly sequenced ancient samples.

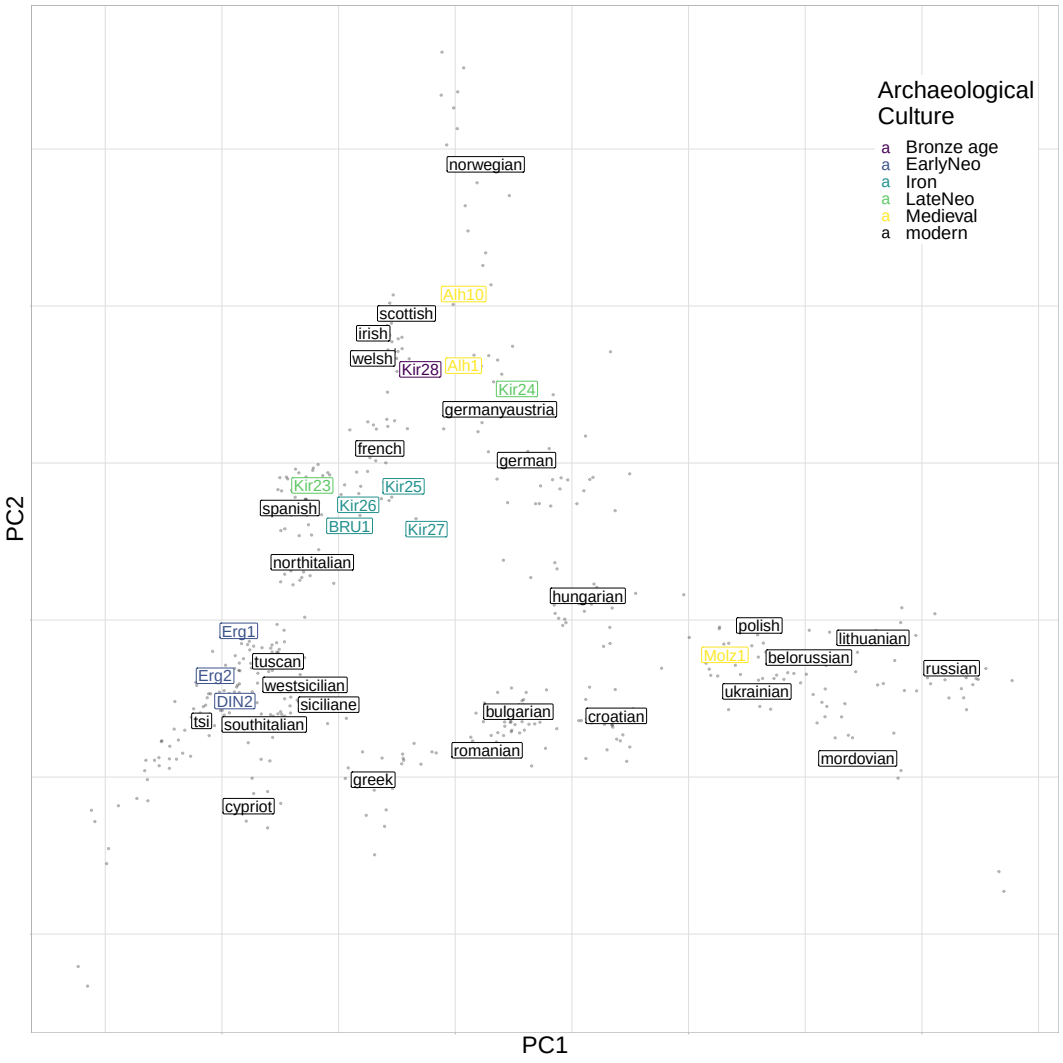


Figure 4.4: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations. 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.

when painted using present-day samples, the 3 Early Neolithic individuals cluster with present-day Italians, consistent with findings from previous research [37, 78] (Fig. 4.4. Erg1 was assigned to mtDNA haplogroup K which has been found in Neolithic and pre-pottery sites across Europe [125, 129] and Western Asia [130, 131].

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 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 with DIN2, supporting the hypothesis that they were from the same source population. Erg1 had the second lowest TVD with Ess7, another LBK sample, from Essenbach, Germany. DIN2 also shares low TVD with Ess7, but has the lowest TVD with NE5, NE4 and NE7, samples assigned to Middle and Late Neolithic cultures on the Hungarian plane. DIN2 was assigned to mitochondrial haplogroup J1C, the same as the samples NE4 and NE5. Both the autosomal and mtDNA link to Neolithic Hungary supports the hypothesis that DIN2 migrated along the Danbian 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 labeled as WHG had a $|Z| > 3$, ($Z = 3.057$), suggesting that Erg1 and DIN2 originate

from the same local population. However, this result was surprising given we would not typically expect an individual from the LBK culture to form a clade with hunter-gatherer populations; this could be indicative of gene flow between a WHG-like source and Erg1. This result was robust to outgroup choice.

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 between Erg1 and *test* since their divergence from an outgroup. 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.

I obtained SNP-capture data from several other local LBK populations; 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). 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 [124, 132]. The position of Erg1 on the PCA suggests that it may have a significant component of Hunter-Gatherer ancestry. I applied the SOURCEFIND algorithm to the 'ancients painting' co-ancestry matrix to infer ancestry proportions for all newly sequenced individuals, fixing 3 surrogate populations at WHG, Yamnaya and Anatolian Neolithic (Fig. 4.8). I inferred 26% WHG ancestry in Erg1, suggesting it may have had a relatively recent ancestor who was a Hunter-Gatherer. I inferred a smaller proportion of WHG ancestry into DIN2 (8%), perhaps suggesting that they were part of a structured local population, where different elements received varying amounts of hunter-gatherer admixture. qpAdm modeling broadly agreed with these estimates and showed that Erg1 can be modeled as a mixture of Anatolia Neolithic (61%, se=0.095) and WHG (0.3855%, se=0.095). Erg2 showed no evidence of hunter-gatherer ancestry and could be modeled directly as Anatolian Neolithic farmer.

To localise the closest source of Hunter-Gatherer admixture into Erg1, I re-performed the 3-population SOURCEFIND analysis, but instead split up the WHG surrogates into Loschbour, LaBrana, Bichon and the 2 individuals from the Iron Gates, leaving 6 surrogate populations in total. I inferred that the two 8800-year-old Iron Gates individuals from Serbia contributed towards 33% of the ancestry of Erg1, showing that it was likely to be closest population to the mixing source in our dataset. To confirm that this was not an artefact of there being 2 Iron Gates individuals (where all of the other WHG populations had a single sample), I removed the lowest coverage Iron Gates individual from the surrogate pool and repeated the analysis. The proportion of ancestry inferred from Iron Gates was similar (31%), suggesting the sampling did not affect

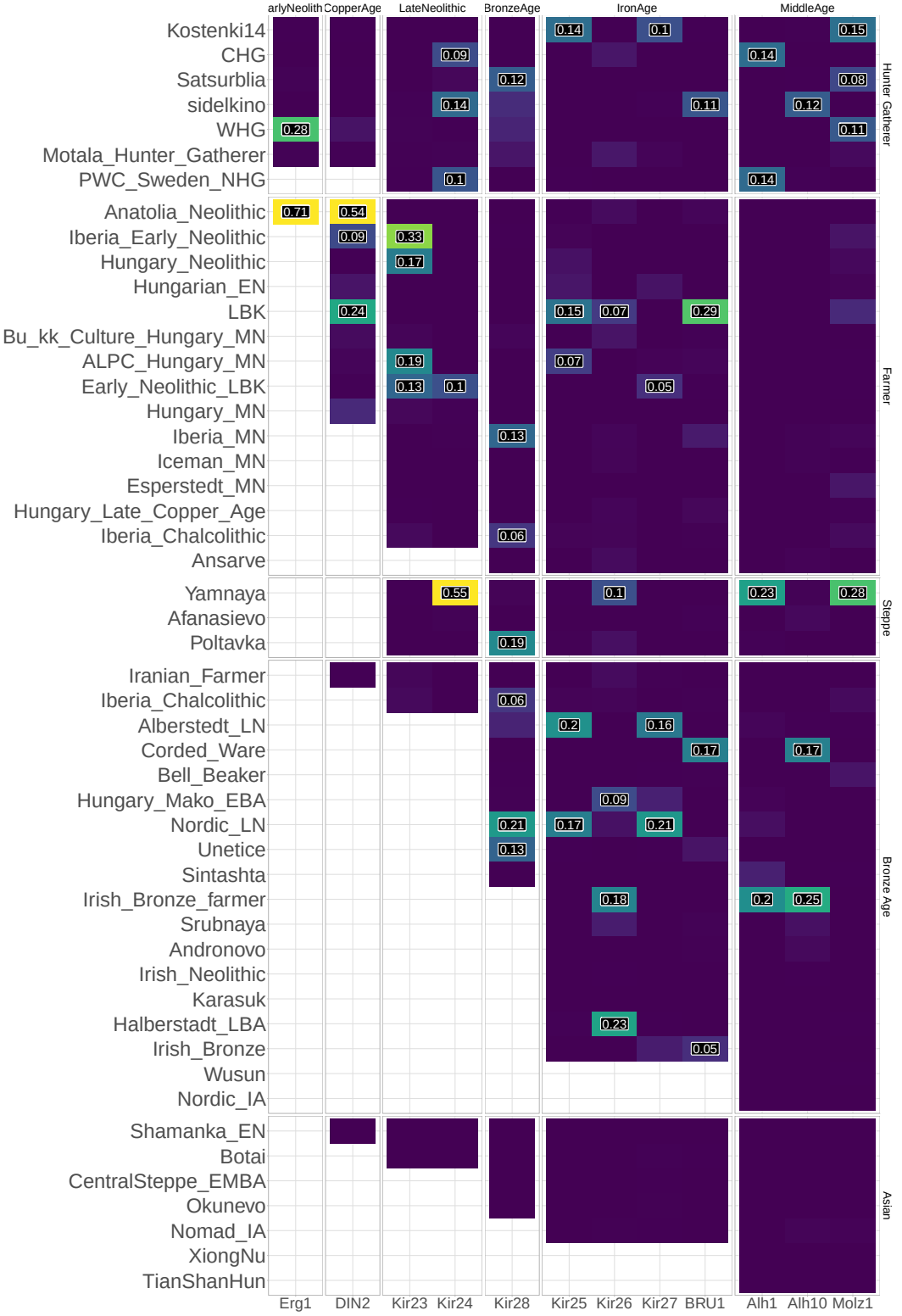


Figure 4.5: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture 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.

the inferred proportion. The same result occured across all 5 indepdendent SOURCEFIND runs.

To determine the date of admixture between an Anatolian Farmer-like and WHG-like source into Erg1, I used MOSAIC [122], which infers admixture events using a similar technique to chromosome painting. MOSAIC is able to model the 'true' admixing sources and determine the genetic differentiation between those and the sampled sources, in addition to the date of admixture. When modeled as a 2-way admixture event, MOSAIC inferred similar WHG and Anatolia Neolithic mixing proportions to SOURCEFIND. It inferred the cluster of Italian hunter-gatherers to be the closest population to the true mixing source (Fig. 4.6). MOSAIC is able to infer the Fst between the 'true' mixing groups and the sampled populations. I inferred very low Fst between the true and source populations, suggesting we had sampled a good proxy for the 'true' mixing sources. I inferred an admixture date of 5.3 generations before the Erg1 was alive. I caution that the admixture date may be unreliable due to only targeting a single individual and given MOSAIC boostraps over individuals (rather than over Chromosomes as in GLOBETROTTER or LD blocks as in qpAdm), it was not possible to obtain confidence intervals around admixture date.

To confirm this admixture event, I performed an f_3 admixture test, which, when significantly negative, provides unambiguous evidence of an admixture event [29]. I performed the test $f_3(A = Castelnovian\_Mesolithic, B = Anatolia\_Neolithic, C = Erg1)$, selecting the A and B populations as those were inferred by MOSAIC to be closest to the admixture sources. This did not yield a significant result ($Z = 1.96$). However, exchanging Anatolia_Neolithic for LBK, a source temporally and geographically more proximate to Erg1 yielded a significant result ($Z = 4.25$).

I also obtained SNP-capture data from several other local LBK populations; samples from Schwetzingen, Stuttgart-Mullhausen and Halberstadt (Rivollat

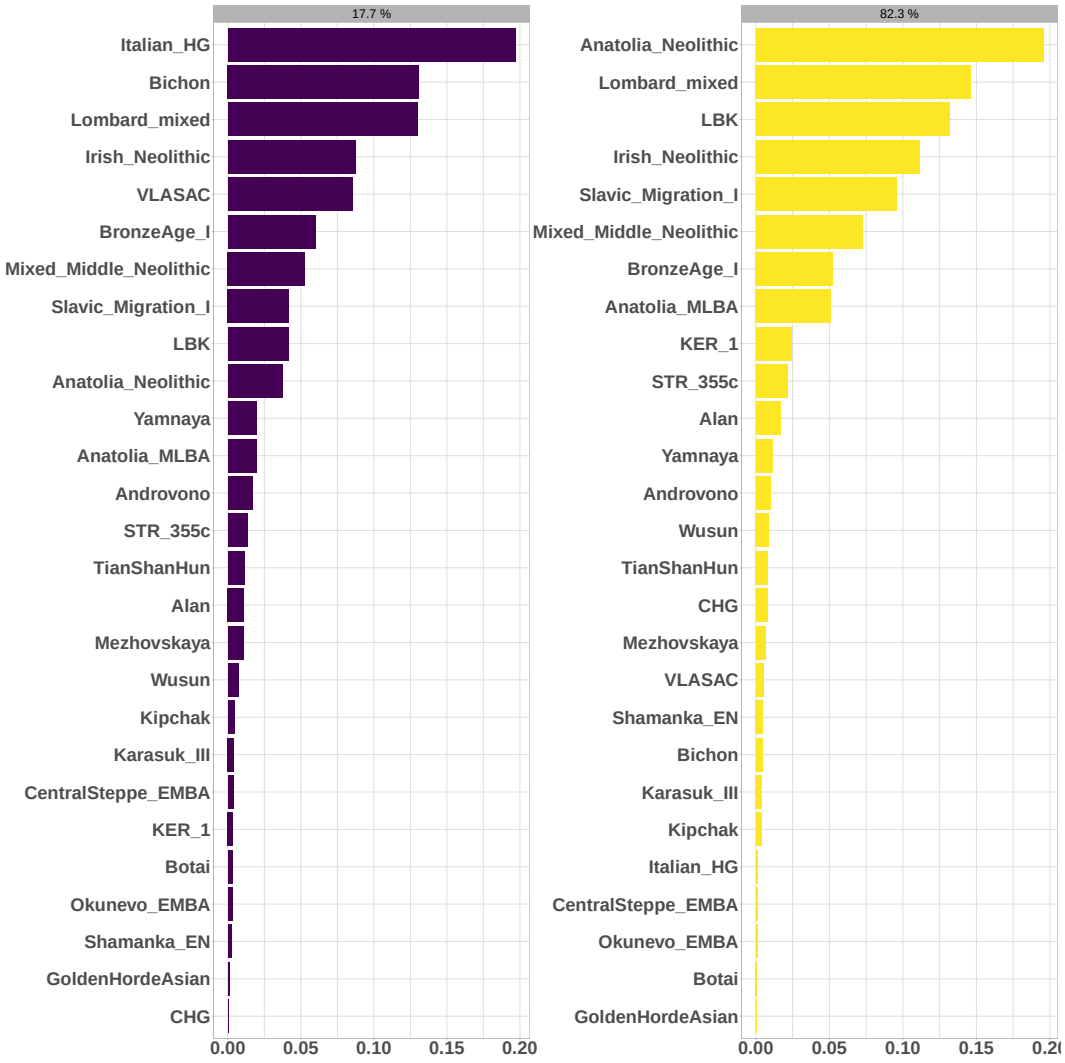


Figure 4.6: Copying matrix plot for sources in 2-way admixture event for Erg1. Each panel represents one of the 2 mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars within each panel represent the amount that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

samples). These samples appear to form a distinct cluster on the unlinked 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 contextualise the amount of Hunter-Gatherer in the newly sequenced samples, compared to different French and German farmer groups. As expected, and shown by previous studies [133], Early Neolithic populations show little sign of Hunter-Gatherer ancestry, which appears more into the Middle Neolithic and further west from Greece and Anatolia. Populations from France, in particular early samples, show the highest amount of HG ancestry. However, contemporaneous populations from Germany display much reduced levels of HG ancestry and can fit a model of purely Anatolia Neolithic ancestry well. Our sample Erg1 appears to be an exception, displaying high levels of HG ancestry comparable to the samples found in France (Fig. 4.7). On the other hand, Erg2, a sample which is contemporaneous and local to Erg1, showed no evidence of Hunter Gatherer admixture

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

This dataset included 2 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 and ChromoPainter 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 known to be responsible for the spread of Indo-European languages across Europe [78]. These results support the findings of Allentoft (2015), who concluded that the Afanasievo

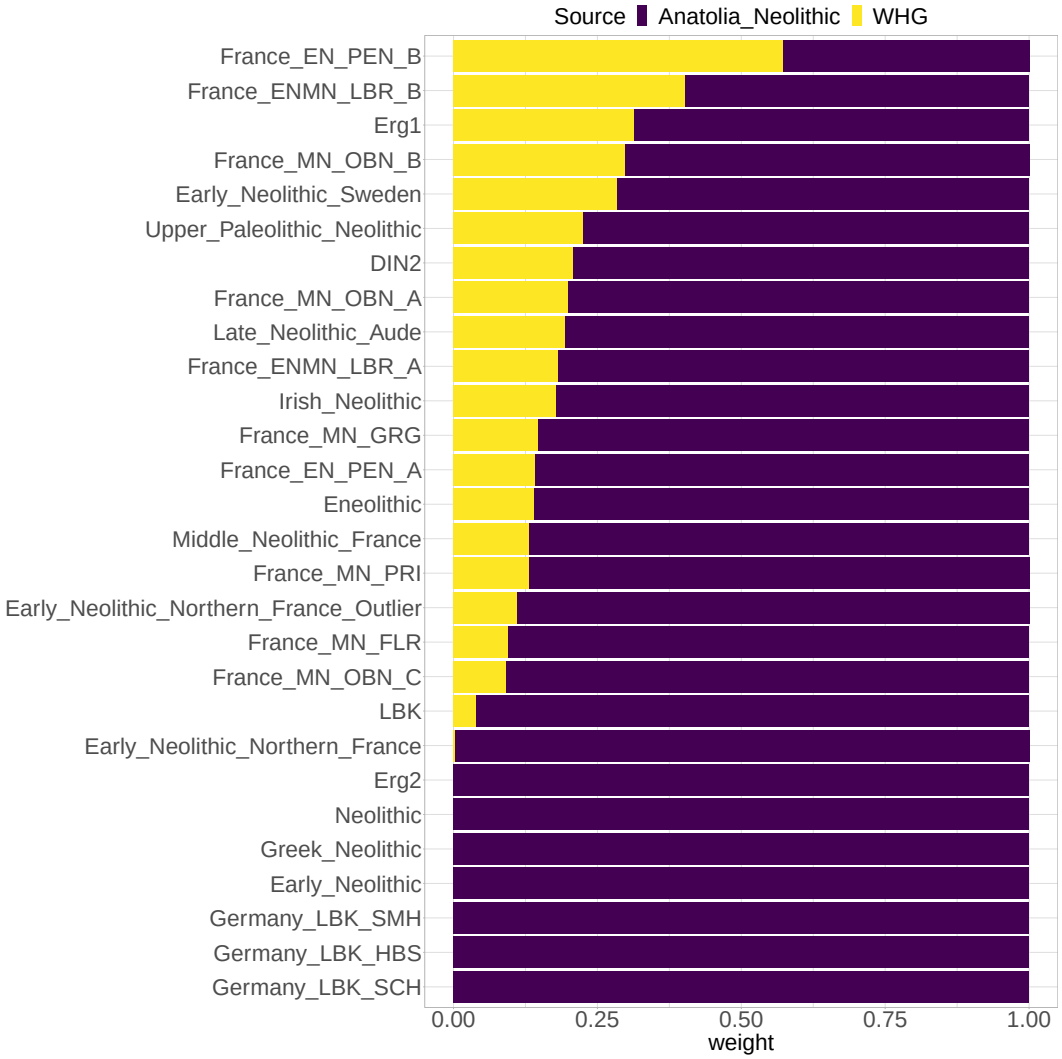


Figure 4.7: qpAdm ancestry proportion estimates for a selection of European Neolithic individuals. All individuals were modeled as a 2-way mixture between Anatolian Neolithic farmers and Western-Hunter Gatherers (WHG). Outgroups given in methods 4.2.9.

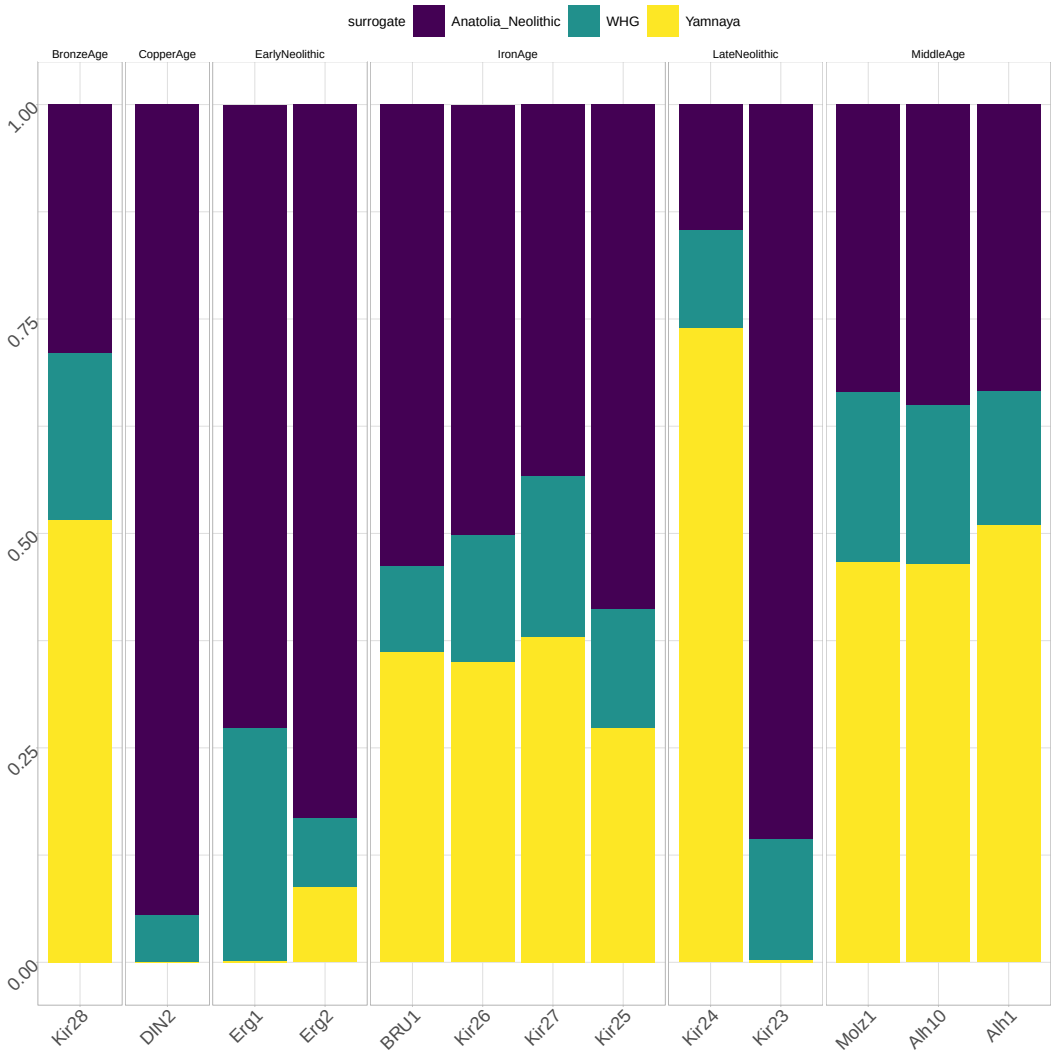


Figure 4.8: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture 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.

Culture were 'genetically indistinguishable' from the Yamnaya Culture. 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 Yamnaya and the lowest TVD with 2 other members 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 [134]. Using qpAdm, Kir24 could be modeled 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 [135]. Additionally, they 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 modeling 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 lacked 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 Introduction of 'southern' ancestry to Cherry-Tree Cave during the Iron Age

Both the plink and ChromoPainter PCAs show that the Iron Age samples appear to be shifted towards the cluster of Neolithic individuals relative to the Bronze Age. The same pattern is also seen in the modern 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. 5.7). Previous studies into the Bronze-Iron Age transition in Western-Europe (France) have shown relative continuity [136]. Other studies in Eastern-Central Europe (Hungary) have shown the Bronze-Iron Age transition was accompanied by an increase in Eastern-European ancestry (albeit from a single sample) [132]. I was interested to see whether the transition in Bavaria had elements of either of these phenomena.

To identify the possible source of 'southern' ancestry in the Iron Age samples, I formed each of the Bronze Age, Iron Age and Middle Age Bavarian populations as a mixture of all other ancient populations using SOURCEFIND. I detected a component represented by 'Renaissance', a population from approximately 1500CE Italy, which contributed towards 26% of the ancestry to Iron Age individuals, but was found in neither the preceding Bronze Age nor following Middle Age. Thus, Renaissance samples appear to be the closest proxy for the 'southern' ancestry source. qpAdm modeling 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). All other

possible sources included with Bronze Age resulted into poorly fitting models. This suggests a model of admixture from populations best represented by those from post Iron-Age Italy.

To determine whether this was an admixture event, I grouped the Iron Age samples together and performed MOSAIC admixture analysis. In the 2-way admixture model, the Iron Age samples could be formed of a mixture of a source closest to an Alamannic-Frankish sample (510 – 530 AD) 17.7% and a source closest to Anatolian Neolithic / LBK samples (82.3%). The estimated F_{st} between the 2 mixing sources was 0.016, approximately equivalent between present-day Germans and Palestinians [137]. Boostrapped dates estimated the date to between 7.86 and 11.31 (95% quantiles) generations ago. This signal is supported by the fineSTRUCTURE groupings; all 4 Iron Age individuals were grouped alongside several Lombard samples and a Roman solider from 300AD.

Based on SOURCEFIND modeling with the extended older surrogates set, unlike Gamba et al (2014) [132], I found no 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

Finally, our dataset included 3 samples 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' context.

The 3 Middle Age samples appear to share common ancestry based on the plink PCA and are located next to other samples from the Middle Ages. Some structure is apparent from the ChromoPainter PCA, with the two Altheim samples clustering more closely together to the exclusion of the Slavic sample; however, this difference appears to be subtle. f_4 in the

form $f_4(mbutipygymy, Bavaria\_Iron; Bavaria\_Slav, Bavaria\_Germanic)$ returned a non-significant result, showing that samples from the Iron Age in Bavaria were symmetrically related to the later Middle Age sample. These results suggest that the differentiation between 'Germanic' and 'Slavic' populations arose 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.

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. On the other hand, 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. That 'GoldenHordeEuro' clusters with Molz1 suggests the location of capture in Europe may have been from Austria where Molz1 was found.

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. All 3 samples are positioned close on the ancients PCA, suggesting they lack differentiation in the context of ancient samples. However, their positions on the modern PCA reveals there was strong differentiation between early Slavic and Germanic peoples (Fig. 5.7). Molz1 clusters with present-day Slavic speaking populations such as Poland, Ukraine and Belarus. On the other hand, the two Germanic samples cluster with present-day individ-

uals from Germanic-speaking countries in Western Europe, such as Scotland, Germany and Wales.

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.

These results were recapitulated using SOURCEFIND, where we modeled each individual as a mixture of different modern-day populations. 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).

4.3.7 Sample heterozygosity and homozygosity

I calculated per-sample heterozygosity and runs-of-homozygosity for all samples above 3x coverage.

4.3.8 Discussion

I found that there was structure even within samples which were extremely spatially and temporally close. For example, Erg1 and Erg2 were found in the same layer and in the same location; yet Erg1 shows evidence of recent Hunter-Gatherer ancestry, whereas Erg2 shows no evidence of admixture. This raises the possibility that admixture between farmers and Hunter-Gatheres occured on an extremely fine geographic and tempral scale. Similarly, the two Late Neolithic samples showed differences in genetic ancestry, with one sample possibly being a recent migrant from the Eurasian Steppe, displaying ancestry typical of the Yamnaya steppe-pastoralists and the other being of primarily farmer ancestry. These results clearly demonstrate that individuals who were

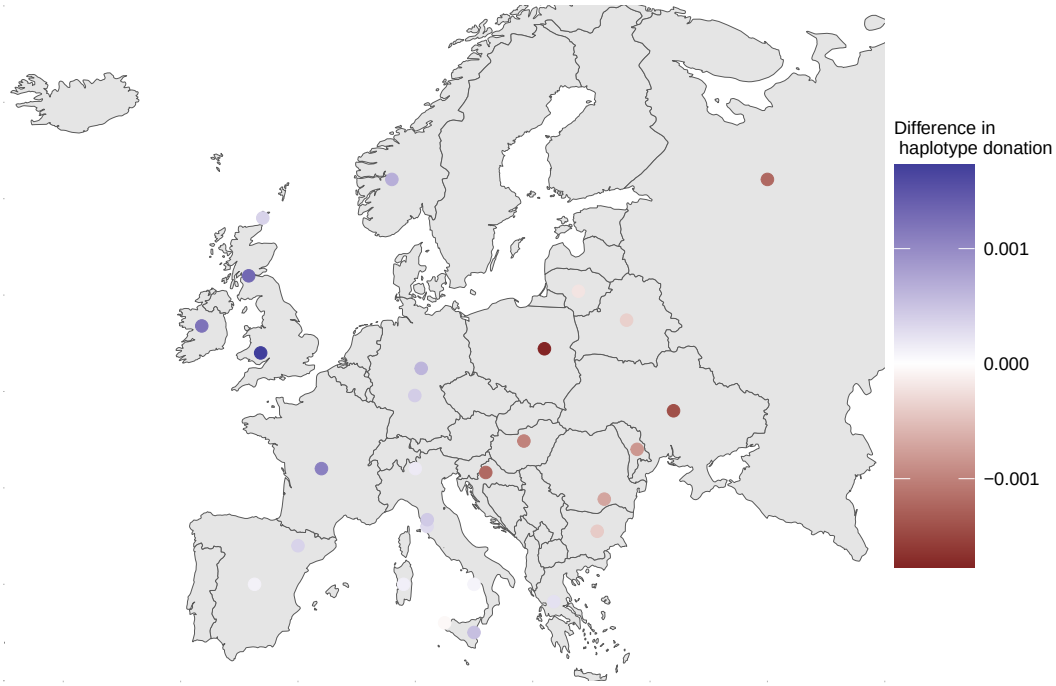


Figure 4.9: Differential haplotype-donation between Germanic and Slavic samples. Each coloured point is one present-day population. Points are coloured based on whether they donate relatively more to Germanic (blue) or Slavic (red) ancient samples.

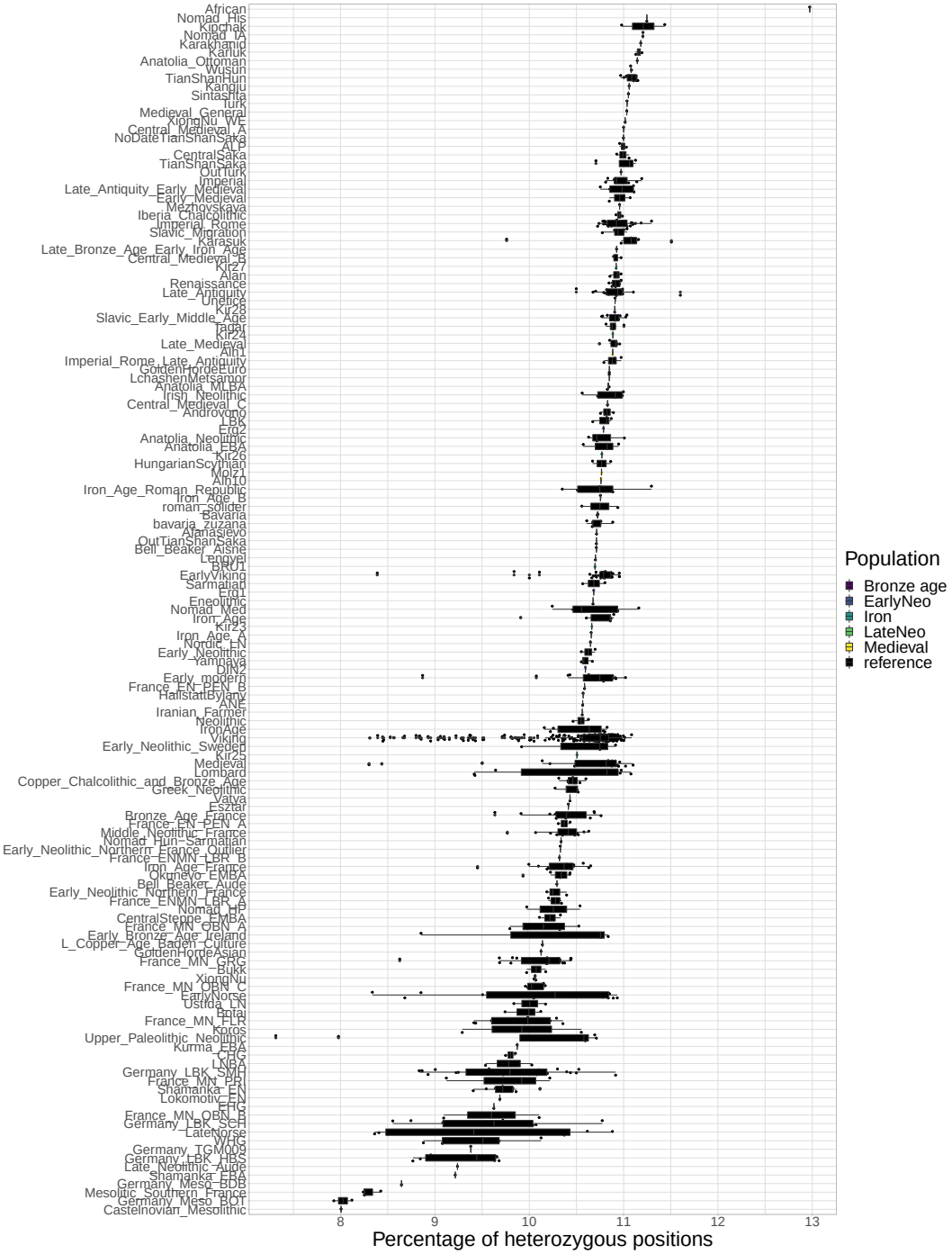

African
Nomad_His
Kipchak
Nomad_IA
Karakhanid
Kanuk
Anatolia_Ottoman
Wusun
TianShanHun
Kangju
Sintashta
Turk
Medieval_General
XiongNu_WE
Central_Medieval_A
NoDateTianShanSaka
ALP
CentralSaka
TianShanSaka
OurTurk
Imperial
Late_Antiquity_Early_Medieval
Early_Medieval
Mezhovskaya
Iberia_Chalcolithic
Imperial_Rome
Slavic_Migration
Karasuk
Late_Bronze_Age_Early_Iron_Age
Central_Medieval_B
Kir27
Alan
Renaissance
Late_Antiquity
Greece
Kir28
Slavic_Early_Middle_Age
Tagar
Kir24
Late_Medieval
Ain1
Imperial_Rome_Late_Antiquity
GoldenHordeEuro
LchashenMetsamor
Anatolia_MLBA
Irish_Neolithic
Central_Medieval_C
Andronovo
LBK
Ero2
Anatolia_Neolithic
Anatolia_EBA
Kir26
HungarianScythian
Mplz1
Ain10
Iron_Age_Roman_Republic
Iron_Age_B
roman_soldier
Bavaria
bavaria_zuzana
Afanasievo
OutTianShanSaka
Bell_Beaker_Aisne
Lengyel
BRO1
EarlyViking
Sarmatian
Ero1
Eneolithic
Nomad_Med
Iron_Age
Kir23
Iron_Age_A
Nordic_LN
Early_Neolithic
Yamnaya
DIN2
Early_modern
France_EN_PEN_B
HallstattBylany
ANE
Iranian_Farmer
Neolithic
IronAge
Viking
Early_Neolithic_Sweden
Kir25
Medieval
Lombard
Copper_Chalcolithic_and_Bronze_Age
Greek_Neolithic
Vatya
Eszar
Bronze_Age_France
France_EN_PEN_A
Middle_Neolithic_France
Nomad_Hun-Sarmatian
Early_Neolithic_Northern_France_Outlier
France_ENMN_LBK_B
Iron_Age_France
Okunevo_EMBA
Bell_Beaker_Aude
Early_Neolithic_Northern_France
France_ENMN_LBK_A
Nomad_HP
CentralSteppe_EMBA
France_MN_CBN_A
Early_Bronze_Age_Ireland
L_Copper_Age_Baden_Culture
GoldenHordeAsian
France_MN_GRG
Bukk
XiongNu
France_MN_CBN_C
EarlyNorse
Ustrda_LN
Botai
France_MN_PLR
Kpros
Upper_Paleolithic_Neolithic
Kurma_EBA
CHG
LNBA
Germany_LBK_SMH
France_MN_PBI
Shamanka_EN
Lokomotiv_EN
EHG
France_MN_OBN_B
Germany_LBK_SCH
LateNorse
WHG
Germany_TGM009
Germany_LBK_HBS
Late_Neolithic_Aude
Shamanka_LBA
Germany_Meso_BDB
Mesolithic_Southern_France
Germany_Meso_BOT
Castelnovian_Mesolithic
8
9
10
11
12
13
Percentage of heterozygous positions
Population
Bronze age
EarlyNeo
Iron
LateNeo
Medieval
reference

Figure 4.10

likely genetically and phenotypically distinct lived amongst one another during the Late Neolithic.

I found that across the different archaeological periods, within Cherry-Tree Cave, there was a degree of continuity, but with evidence of admixture from the outside.

Using 3 ancient genomes, I showed that the distinction between 'Germanic' and 'Slavic' peoples can be outlined in the context of modern samples.

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 [138] and whose origin, although disputed, is thought to be Polesia (a marshy forested area straddling Poland, Belarus, Russiana and Ukraine) [139]. 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 [140]. 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 [141], 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 2 centuries, these peoples had settled across large parts of Europe. In particular, the Early Slavs had expanded southwards into the Balkans and Alps [138, 142–144]. It has been proposed that these migrations were key to forming the foundations of present-day Slavic (speaking) nations [138].

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 3 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.

The differentiation of Slavs into these 3 broad groups can still be seen today in the different language groups. Today 315 million people speak Slavic languages. Linguistic evidence suggests that they can be broadly split into 3 groups; Western Slavs (Poles, Czechs and Slovaks), Eastern Slavs (Ukrainians, Belarusians and Russians) and Southern Slavs (Croatians, Bulgarians, Slovenians, Bosnians, Macedonians, Montenegrins and Serbians) [145].

The history of the Slavic peoples can be artificially be split into 3 periods; Migration Period ($\tilde{3}$75AD - 568AD), Early Middle Ages/High Middle Ages ($\sim\!\!\sim$600AD - $\sim\!\!\sim$1250AD) and present-day. Although previous studies have investigated the genetics of the transitions between these periods, they have

Figure 5.1: 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

been relatively limited in their scope. Juras et al (2014) used uniparental mtDNA markers from ancient DNA samples 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 [146]. 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 present-day and ancient samples [147] (although see [148]). For example, it is know that mtDNA and nuclear DNA may have different evolutionary histories and thus display discordant phylogenetic trees [149].

Kushniarevich et al (2015) [150] 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 in order to understand the historical processes that have formed the present-day genetic structure. They proposed that admixture of incoming Slavic speakers during the Migration Period with the pre-existing substrate of regional genetic components, which differed between South, East and West Slavs. Using this evidence, they propose that the "cultural assimilation of indigenous populations by bearers of Slavic languages as a major mechanism of the spread of Slavic languages to the Balkan Peninsula".

More recently, Macháček et al (2021) [151] analysed ancient rune inscriptions on a cattle rib from Lány, Czechia, dated to approximately 600AD. The bone is inscribed with Germanic runes. Finding Germanic runes in the context of Slavic peoples provides evidence of early interactions between Slavic and Germanic peoples. 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 [16, 122, 152–154]. 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 [155] 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) [16] inferred an excess of IBD-sharing among Eastern European individuals, albeit with a more constrained admixture data of 440 - 1080 CE. However, this could also be interpreted in terms of a small effective population size [156, 157]. Salter-Townshend and Myers (2019) also identified admixture in the Chuvash people betweeen East Europeans and East Asians approximately 1224 CE.

Despite these efforts, no studies have integrated autosomal DNA from ancient and present-day samples whilst applying powerful haplotype-based methods to infer population structure, ancestry proportions and admixture events. Therefore in this chapter, I will analyse 17 new medium to high coverage whole ancient genomes from Czech Republic, spanning the Migration Period and Early Middle Ages. These are, to my knowledge, the first high-coverage whole ancient-genomes from Slavic speakers. 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. Can we gain an understanding of the geographical origins of the Slavic peoples from ancient DNA

2. 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)

3. Was there interactions between Germanic and Slavic peoples during the Early Migration Period.

4. If so, what genetic differences can be observed between these periods? Are they characterised by admixture from outside sources? If so, what are these sources and can the admixture events be dated?

5. What is the relationship between the ancient samples and present-day day Slavic populations. Are they continuous?

6. 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 was generated from 17 ancient individuals. All newly sequenced samples are from Czechia and are split across two different field sites.

The newly sequenced samples are grouped into two temporal categories; 5 samples are from the Migration Period (348 AD - 504 AD) and the Líbivá site, and the other 12 samples are from the later Early Middle Ages (724 AD - 995 AD) and the are from the Pohansko site.

Apart from the age of the samples, the Migration Period and Early Middle

Code	Site	Date (AD)	Period	Coverage
LIB5	Břeclav z Líbivá	348	Migration	7.32
LIB4	Břeclav – Líbivá	472	Migration	6.46
LIB12	Břeclav – Líbivá	475	Migration	6.75
LIB2	Břeclav – Líbivá	495	Migration	6.39
LIB3	Břeclav – Líbivá	509	Migration	5.29
LIB11	Břeclav – Líbivá	741	Migration	5.33
LIB7	Břeclav – Líbivá	830	Migration	5.64
POH11	Pohansko – Lesní školka	783	EMA	4.99
POH27	Pohansko – Jizní Předhradí	783	EMA	5.86
POH28	Pohansko – Jizní Předhradí	822	EMA	5.58
POH41	Pohansko – Lesní školka	875	EMA	5.22
POH13	Pohansko – Lesní školka	879	EMA	5.95
POH36	Pohansko – Jizní Předhradí	880	EMA	5.47
POH40	Pohansko – Lesní školka	950	EMA	5.46
POH3	Pohansko – Lesní hrúd	956	EMA	5.39
POH44	Pohansko – Pohřebiště U Kostela	NA	EMA	5.33
POH39	Pohansko – Jizní Předhradí	866	EMA	5.30

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.

Age samples can be differentiated by the style of pottery found in the burial grounds (Z. Hofmanová, personal communication).

5.2.2 Ancient DNA processing

I merged the 17 newly sequenced individuals with the reference data-sets A.1 to A.17 resulting in a total of 942 ancient individuals in .bcf format, with genotype likelihoods at 77,213,942 genome-wide autosomal SNPs. Data was then split into separate .bcf files for each chromosome and indexed using bcftools.

I followed the GLIMPSE [63] 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 [75], described in appendix A.5.

5.2.3 Present-day DNA processing

I chose the MS-POBI-HellBus dataset, described in detail in appendix A.20, because it contains a high number of relevant samples from central and Eastern Europe. I removed samples from Australia, New Zealand and USA, as these samples were not from native individuals from that country.

The modern and ancient samples were phased separately. This was because GLIMPSE, which is necessary to phase the ancient samples with, is not suitable to phase the modern samples with, for two reasons. Firstly, GLIMPSE is designed to work with sequence-level density of data, and the modern samples have been genotyped on a low-density genotyping array. Secondly GLIMPSE accepts data as genotype likelihoods; these were not available for the modern samples. Therefore, the modern samples were phased using shapeit4 [71].

Appendix A.20 describes the initial filtering that was used to generate this dataset. It was then phased using shapeit4 [71] without the use of a reference panel and setting the number of conditioning haplotypes to 8. It was then converted to ChromoPainter input format using a custom R script and merged with the dataset of ancient samples described in the previous section.

5.2.4 plink PCA

To determine the broad-scale ancestry distribution of the newly sequenced individuals in the context of 915 other ancient samples, I performed PCA on the non-imputed genotypes using plink2. Performing an unlinked PCA also allows us to identify any data quality issues which are independent of phasing / haplotype-based analysis.

I retained only the 500,000 markers with the lowest amount of missingness and then 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 and the first two principle components plotted.

5.2.5 Sample heterozygosity and ROH

I used plink (v1.90p) to calculate the total length (kB) of runs of homozygosity (ROH) within each sample across all ancient and present-day individuals in the combined dataset.

5.2.6 Allele-frequency based tests

I used Admixtools [29], implemented in Admixr R library [158] to employ several different f-statistics.

I converted imputed .vcf to .ped/.map format using plink. It has been shown that using imputed markers reduced reference bias relative to using pseudo-haploid markers [35]. Convertf from the Admixtools library was then used to convert .ped/.map files into Eigenstrat format suitable for use with Admixtools.

5.2.7 ChromoPainter and fineSTRUCTURE analysis

I began with a merged dataset of present-day and ancient individuals, described in sections 5.2.2 and 5.2.3 in ChromoPainter format.

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. 2x was somewhat arbitrarily chosen as a conservative threshold to reduce coverage related bias whilst still retaining a suitable number of individuals. This allows for the characterisation of the ancestry of the newly sequenced ancient samples in the context of other ancient individuals. It is also the painting that can be used to perform fineSTRUCTURE clustering and tree building on ancient samples. Hereafter referred to as 'ancient' painting.

I also performed an 'all-v-all' painting of a selected group of present-day individuals and the newly sequenced ancient individuals. The populations

retained are given in table **??**. Hereafter referred to as 'present-day painting'.

Both the 'present-day' and 'ancient' paintings were merged separately using chromocombine-0.0.4 (`https://people.maths.bris.ac.uk/~madjl/finestructure-old/chromocombine.html`).

The fineSTRUCTURE [15] clustering and tree building algorithm was applied to the chunkcounts ChromoPainter output, for both the 'present-day' and 'ancient' paintings. This algorithm assigns individuals to genetically homogeneous clusters, estimates the 'true' number of clusters and builds a dendrogram of genetic similarity. 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 external group labels.

fineSTRUCTURE (v0.0.5) was applied to the resulting chunkcounts matrices for both the ancients painting and the moderns painting. It 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, co-ancestry matrix figures and principle component plots were generated using the fineSTRUCTURE R library (`https://people.maths.bris.ac.uk/~madjl/finestructure/FinestructureRcode.zip`).

5.2.8 SOURCEFIND ancestry proportion analysis

I used SOURCEFIND [17] to infer the proportions of ancestry by which each target (e.g. ancient) individual is most related to a set of surrogate populations. Each of the 47 clusters of ancient samples inferred by fineSTRUCTURE was analysed in turn, using the other 46 clusters to act as surrogates.

Population	nsamples
HB:tsi	196
HB:spanish	68
HB:bulgarian	62
HB:german	60
HB:french	56
HB:russian	50
HB:greek	40
HB:ukrainian	40
HB:croatian	38
HB:hungarian	38
HB:norwegian	36
HB:southitalian	36
HB:polish	34
HB:romanian	32
HB:mordovian	30
HB:cypriot	24
HB:northitalian	24
HB:lithuanian	20
HB:siciliane	20
HB:westsicilian	20
HB:belorussian	18
HB:tuscan	16
HB:irish	14
HB:scottish	12
HB:germanyaustria	8
HB:welsh	8

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

Each cluster was run across 3 independent MCMC runs, using 50,000 burn-in iterations, 500,000 main iterations, thinned every 5 iterations. All 3 MCMC runs were then combined to form an MCMC list using the coda R libary [79] and `mcmc` function to jointly estimate ancestry proportions and empirical credible intervals for each target population.

5.2.9 MOSAIC admixture analysis

I inferred admixture events, dates and proportions using MOSAIC [122].

I performed 2 different kinds of admixture modeling. First, I performed an

'ancient analysis' using the 47 fineSTRUCTURE clusters of ancient samples to assign groups. I analysed each cluster in turn, using all other 46 clusters as surrogates.

I then performed a 'present-day surrogates' analysis using a select group of present-day populations **??** 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 either 2 or 3 ancestral sources. Upper and lower quantiles for admixture dates were estimated using a bootstrap procedure.

5.3 Results

Principle Component Analysis (PCA) using plink2 showed that the Migration Period samples do not all cluster together and instead fall on a cline of similarity between a cluster of Central European Middle Age/Iron Age samples (top-left) and Neolithic samples (bottom-right) (Fig. 5.2). The Early Middle Age samples are more homogeneous, with all samples occupying the broad region containing European Iron Age samples. However, samples POH39 and POH3 display an elevated affinity to samples from Early Bronze Age Ireland.

5.3.1 Mixed ancestry of migration period Slavs

In order to reveal further structure in the ancient samples, I performed an all-v-all painting of 152 ancient samples with a coverage greater than 2x. I then applied the fineSTRUCTURE clustering algorithm to the samples in order to assign them to genetically homogeneous groups.

The migration period consisted of 5 individuals from Břeclav (Líbivá), Czech Republic, from 5 different burial sites, who had radiocarbon dates corresponding

Population	nsamples
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

Table 5.3: Name of populations and number of samples used in the present-day MOSAIC analysis

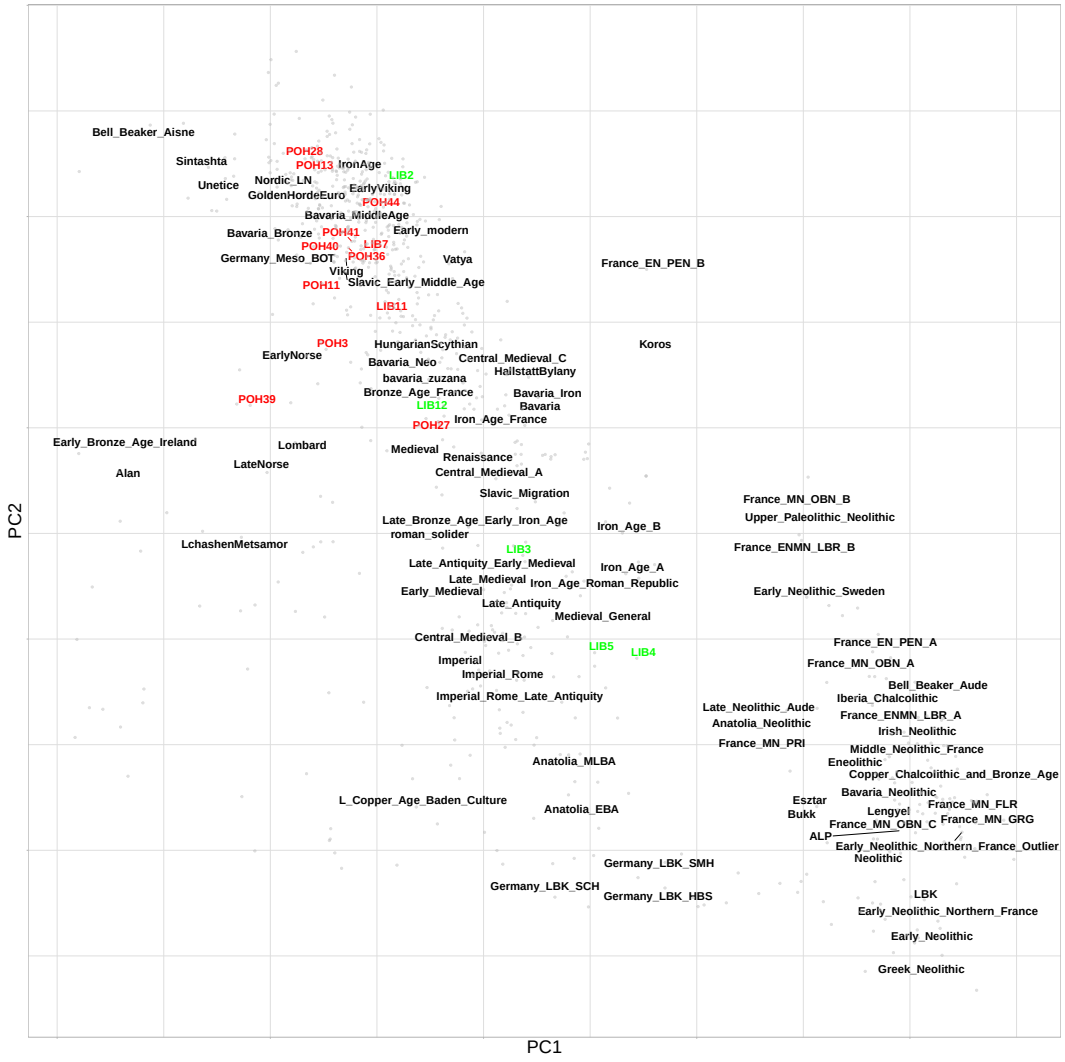


Figure 5.2: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using the plink2. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations. The position of each reference label is the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

to the Migration Period (348 - 509AD). It is apparent from both the unlinked (Fig. 5.2) and linked PCAs (Fig. 5.3) that the Migration Period samples represent a heterogeneous group of individuals who do not originate from the same source population. LIB2 (495AD) is located in the centre of a large cluster of contemporaneous individuals from Iron Age Central and Northern Europe. This individual shares the most haplotypes with Viking individuals from Denmark, Estonia and the UK from roughly the same time period. fineSTRUCTURE analysis grouped LIB2 primarily with Viking era individuals from Sweden, Denmark, Iceland, Estonia and Norway from 300-1100AD. When painted using a set of present-day reference samples, LIB2 matches the most haplotypes and clusters with Norwegians (Fig. 5.8). Put together, these data suggests LIB2 may be 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 [159, 160]. However, most evidence suggests these links occurred later than the date of these samples. For example, it is known that the Scandinavian colonists settled in present-day Russia as early as 750. 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.

On the other hand LIB4 and LIB5, and to a lesser extent LIB3, show an affinity the European Neolithic, indicated by their position on the linked and unlinked PCA. Interestingly, they share the most haplotypes with several Italian Neolithic samples, despite being separated by approximately 6000 years (not a clue why this is). Despite sharing the most haplotypes with these samples, LIB4 and LIB5 are found in fineSTRUCTURE clusters with more recent samples from Italy (Early Iron Age / Renaissance), suggesting the link to Neolithic Italy may have been transmitted by more recent populations (need to expand more

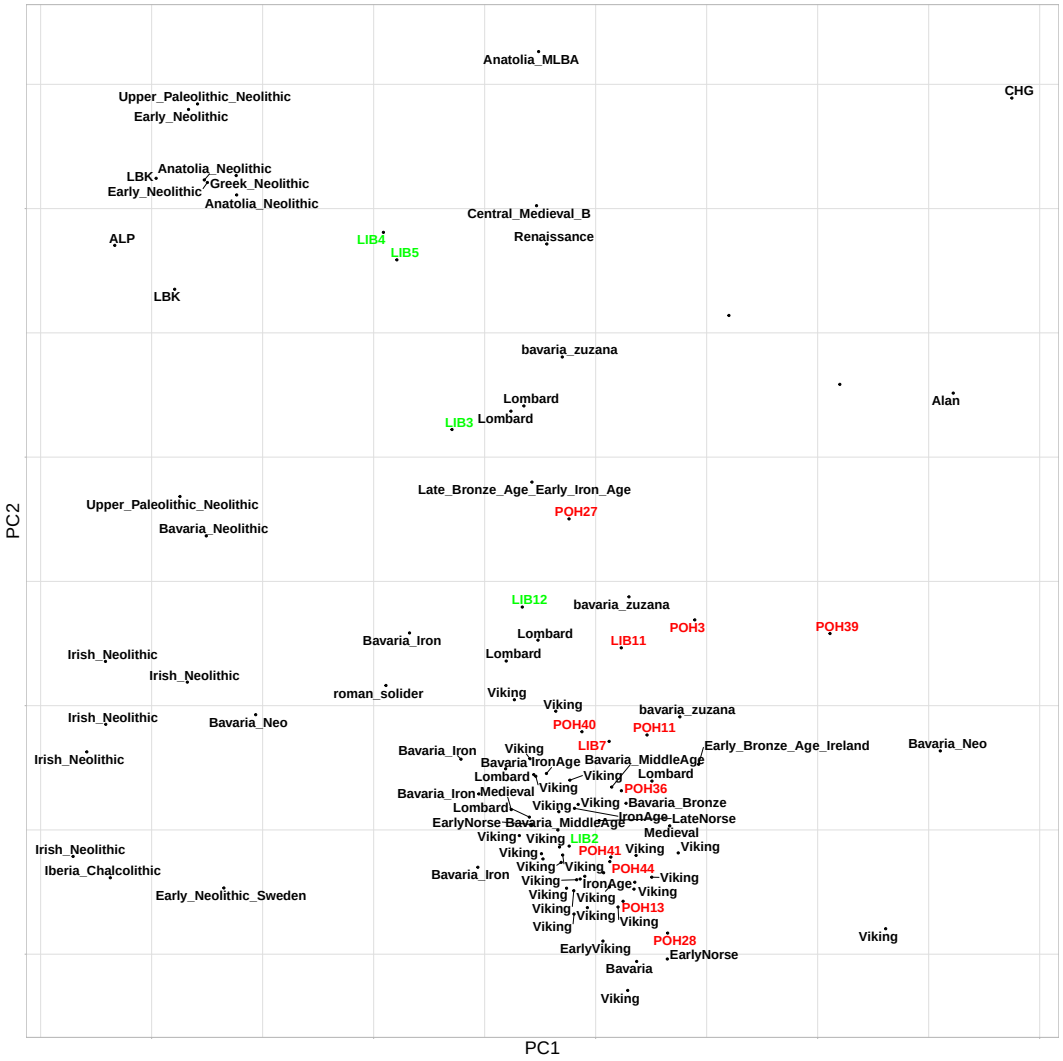


Figure 5.3: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using fineSTRUCTURE. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations.

on this). Both LIB4 and LIB5 share the most haplotypes with one another; this and their consistent positions on PCA and fineSTRUCTURE groupings suggest they are closely related and could be from the same local population.

The appearance of Southern-European ancestry in Central Europe, most closely related to Neolithic farmers, into 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) [161]. 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 that (at least LIB3, need to check others) is 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 (nobody really knows where it started according to Zuzana) through Czechia, Slovakia, Hungary and ended up in Lombardia. 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 5.11).

Ancestry proportion estimation using SOURCEFIND showed that the cluster containing LIB3, LIB4 and LIB5 shares 25% of their ancestry most recently with people from Anatolia, 16% from LBK (Linearbandkeramik) and 12% from a cluster containing Lombard individuals.

I performed MOSAIC admixture modeling using present-day samples as surrogates and the clusters of newly sequenced ancient samples as targets. I did not detect an admixture even when targeting LIB3, LIB4 and LIB5. This

could be due to low power or a low number of samples, or that the samples are unadmixed with respect to the surrogate populations.

On the fineSTRUCTURE PCA, LIB3 clusters with Lombard samples from Northern Italy. Historical evidence cites alliances between Slavs and Lombards in the 5th century [162]. In the 'present-day' painting, LIB3 clusters with and shares the most haplotypes with present-day Tuscans.

Finally, LIB12 displays ancestry which is more typical of the preceding Central European Bronze Age. It copies the most haplotypes from samples from Bronze Age Ireland (Rathlin) and Bavaria and is found in a cluster with several other Bronze Age samples. This suggests 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 5 Migration Period ancient Slavs, the 12 Early Middle Age Slavs (741-956 AD) represent a more homogeneous set of samples. All 12 samples were clustered into the same fineSTRUCTURE group (named Slavic Early Middle Age II), alongside Viking/Medieval samples from Ukraine, Poland and Sweden. SOURCEFIND analysis showed that this cluster derives roughly equal parts of ancestry from the clusters Viking 10C Scan I, BronzeAge I and Lombard mixed cluster. Interestingly, these are 3 ancestry sources which are similar to those identified by SOURCEFIND analysis in the Migration Period samples. We could tentatively therefore suggest that the Early Middle Age Slavs were formed from the mixture of 'Northern' (represented by Viking) and 'Southern' ancestries (represented by Lombards) onto a substrate of local Bronze Age populations. Note that I suggest that these are the most representative populations and not necessarily the 'true' populations that mixed.

MOSAIC admixture modeling using ancient surrogates proved inconclusive. However, using present-day individuals as surrogates provided cleaner results. The best fitting model was a 3-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. 5.6).

This admixture event is consistent with a signal inferred in both present-day day Eastern European individuals [16, 122]. In previous studies, this admixture event was dated to approximately 1200CE (MOSAIC) and 438CE (GLOBETROTTER). Despite the differing dates, the proportion of ancestry is consistent across studies (approximately 2%), suggesting the signal is genuine. To further support the event, the proportion of ancestry from this source is consistent across 2-way and 3-way MOSAIC admixture models.

5.3.3 Do the samples cluster together - TVD permutation test

fineSTRUCTURE analysis suggested that the Migration Era and Early Middle Age samples did not originate from the same source population. To formally establish whether the Early Middle Age and Migration Period samples cluster within their respective populations, following Leslie et al 2015 [81], I performed a TVD permutation test. TVD is a distance metric which can be calculated from the chunklengths matrix and is equivalent to finding the absolute distance between two copyvectors, with larger values meaning two samples have more different ancestry profiles.

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. Then, I calculated the empirical

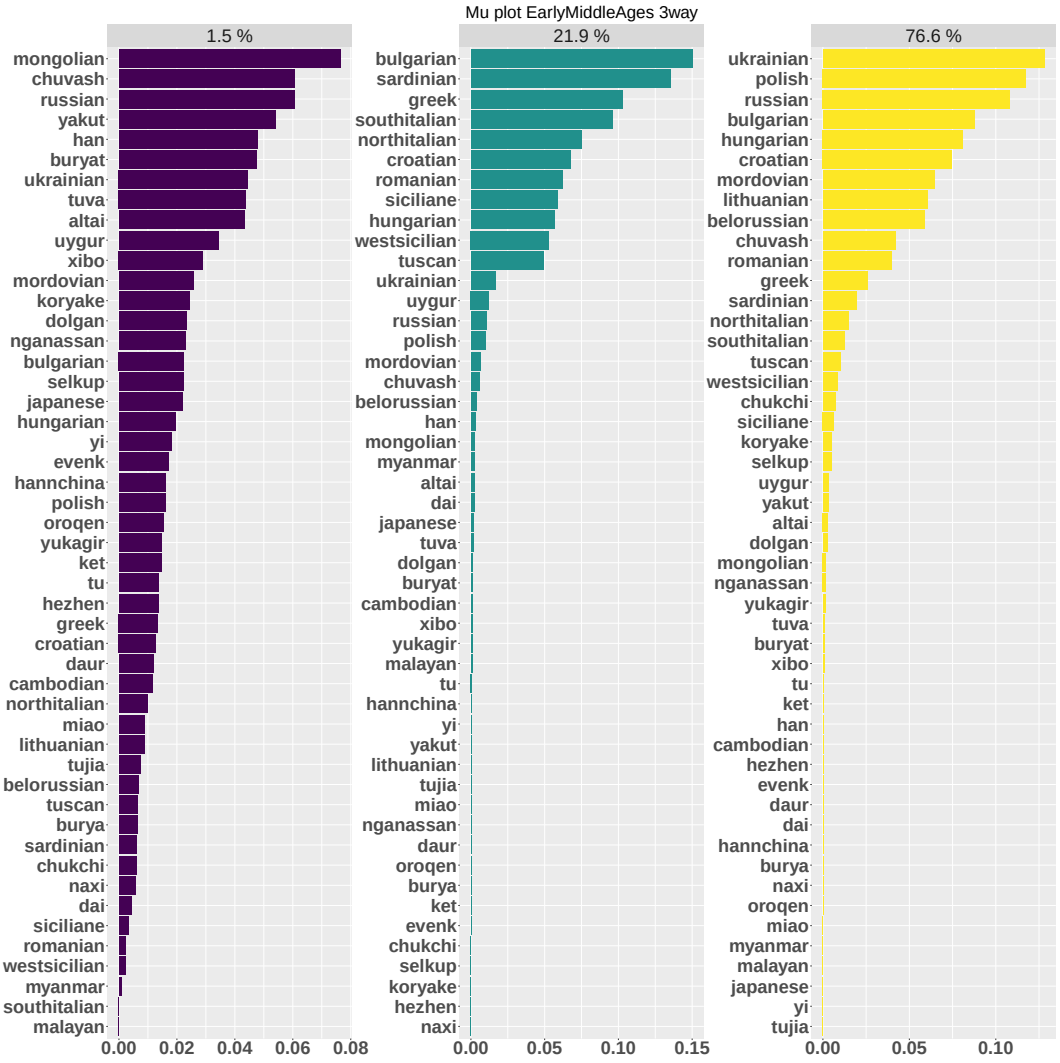


Figure 5.4: Copying matrix plot for sources in 3-way admixture event for Early Middle Age ancient Slavic samples. Each panel represents one of the 3 putative mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars within each panel represent the amount that putative mixing source copied from a particular population.

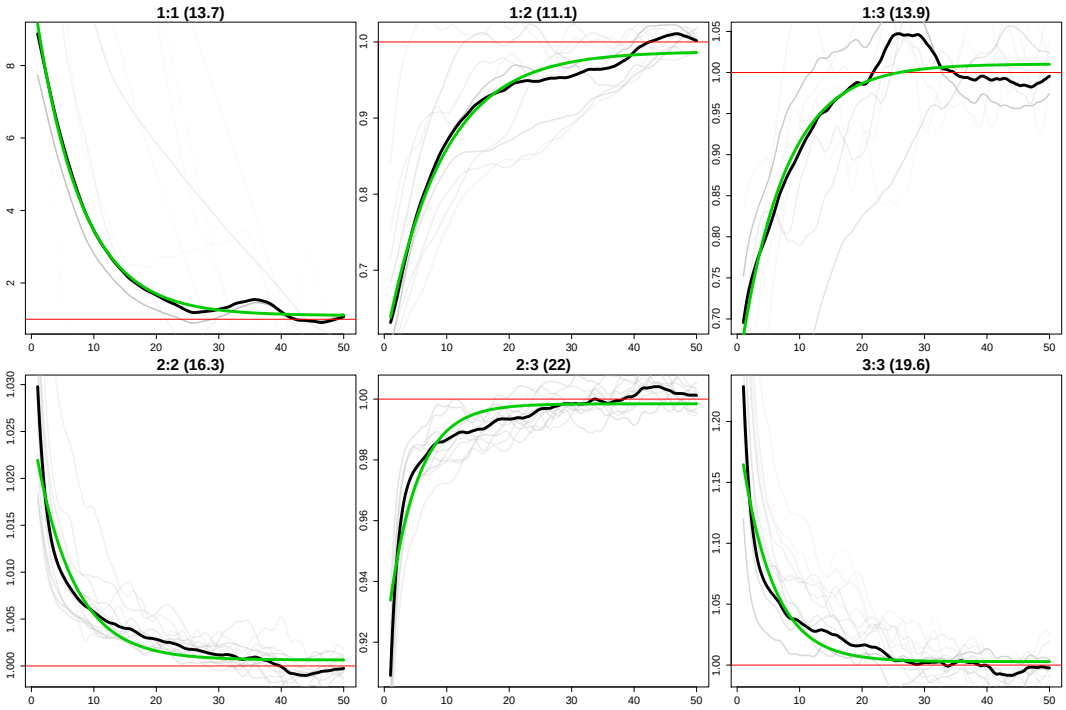


Figure 5.5: Inferred Coancestry Curves obtained from modeling Early Middle Age samples as a 3-way mixture of present-day individuals. Black lines are empirical coancestry curves across all target individuals, light grey are per individual, green is the fitted single-event coancestry curve. Note to self - need to figure out what the numbers mean but doesn't say in the manual anywhere.

TVD between the two groups as $TVD_{mp,ema} = \sum |C_{mp} - C_{ema}|$. For 10,000 iterations, I then randomly permuted the population labels among the samples and then calculated a 'random' TVD, $TVD^{rand}_{mp,ema}$ between the samples with randomly permuted populations. We can then calculate the p-value that we can reject the null model of no significant differences between the groups (not sure if this is the right way of wording it) as the number of randomly permuted iterations where $TVD^{rand}_{mp,ema} > TVD_{mp,ema}$. This test supported clustering the samples into their respective groups ($p = 0.0013$).

5.3.4 Interactions between the two groups

The previous section suggested that individuals from the Migration Period and Early Middle Ages had differing ancestry signals.

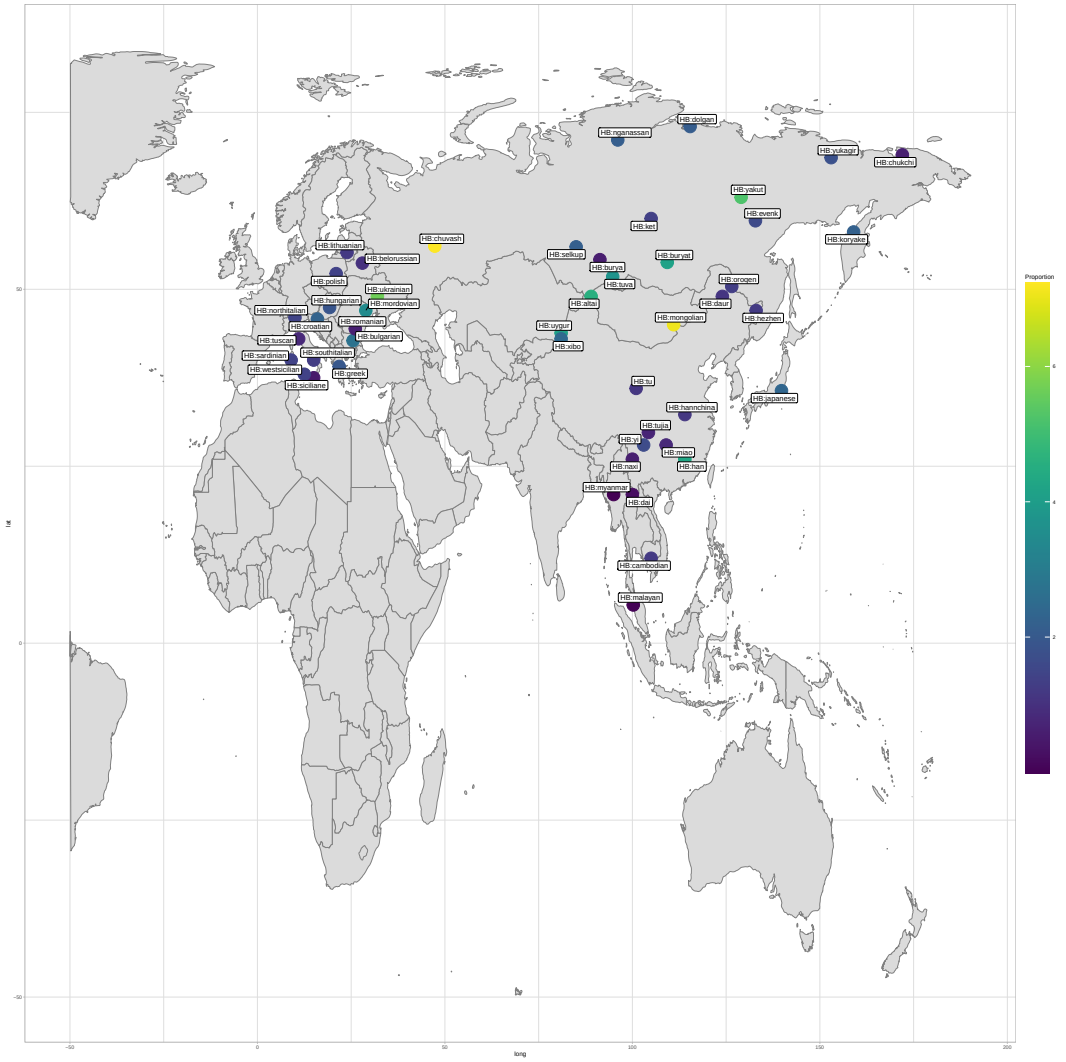


Figure 5.6: Distribution of East-Asian minor ancestry component in Early Middle Age samples.

To determine the extend of mixture and continuity between the Migration Period and Early Middle Ages, I modeled each Early Middle Ages sample as a mixture of other ancients, including individuals from the preceding Migration Period. The proportion of ancestry the individuals derive from the Migration Period clusters could be used as a proxy for the degree of continuity. 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.

Note - I could do something like admixture f3 to see if Early Middle Age

is admixed between Middle Age and any other pop instead as a more explicit test.

5.3.5 Legacy of Slavic migrations in present-day individuals

To understand the genetic legacy the newly sequenced Slavic samples left in different European populations, I painted each sample using the HellBus dataset of present-day individuals. This dataset contains a diverse set of European populations - particularly those from present-day Slavic speaking countries (Polish, Croatian, Bulgarian, Belorussian, Ukrainian, Russian) but also neighbouring non-Slavic speaking countries (Romanian, Lithuanian, Germany and Mordovia).

Principle component analysis (PCA) of the chunklengths matrix, where present-day European samples acted as donors, reveals genetic similarity between ancient Slavic samples from the Early Middle Ages and present-day day Slavic speaking people (Fig. 5.7). The 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. This cline could be mediated by the possible historical admixture event between a source closest to present-day East Asians and a second closest present-day Southern Europeans, with the position of the samples along the cline dependent on the level of admixture from the different sources.

As with previous analyses, Migration Era Slavs are spread across the PCA. 3 samples, 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.

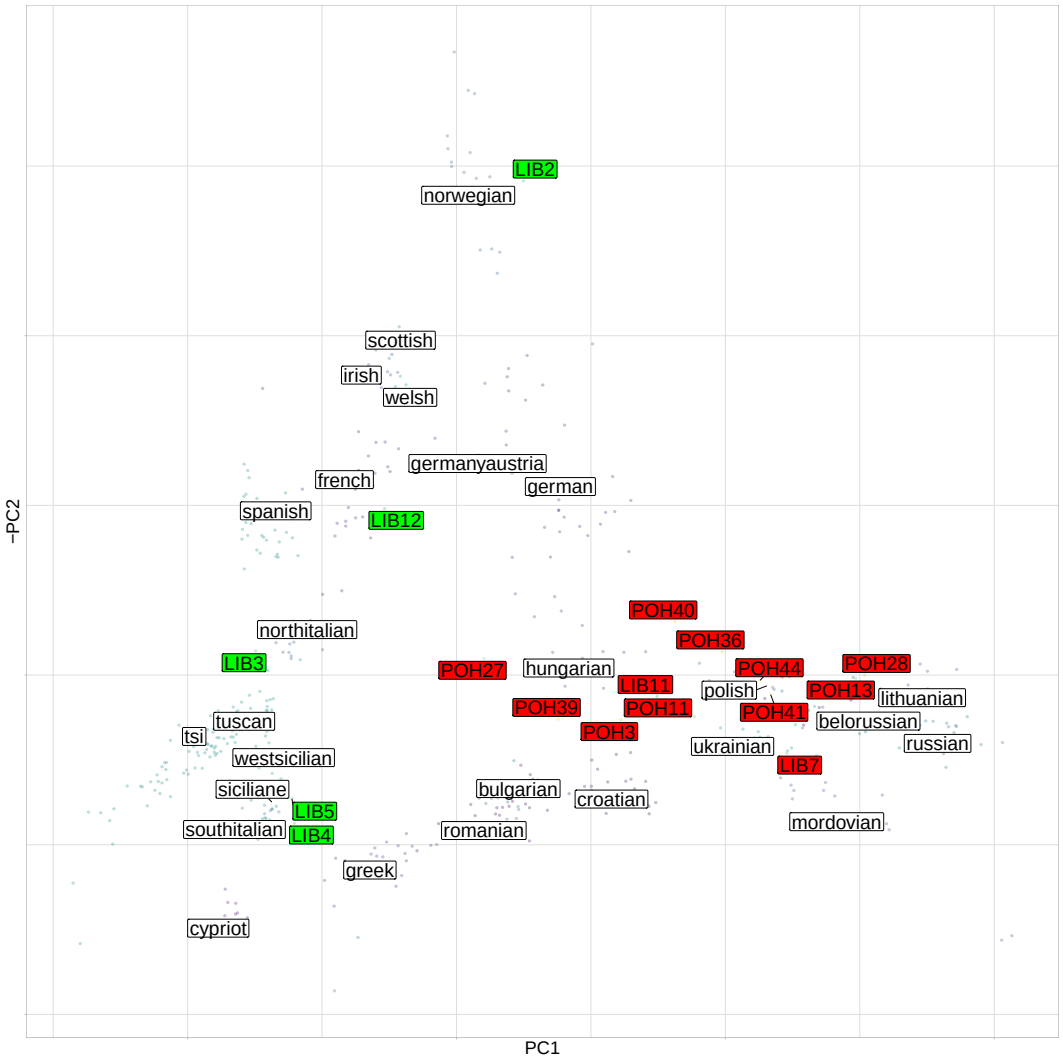


Figure 5.7: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations. 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 (Fig. 5.8). The Migration Era samples appear not to show any excess affinity to present-day day Slavic populations. The two samples who in previous analysis showed a strong genetic relationship to the Neolithic, LIB4 and LIB5, shared the most haplotypes with present-day day Greek individuals. This should not

be surprising given present-day day Greeks have a relatively high proportion of Neolithic ancestry relative to other European populations [163].

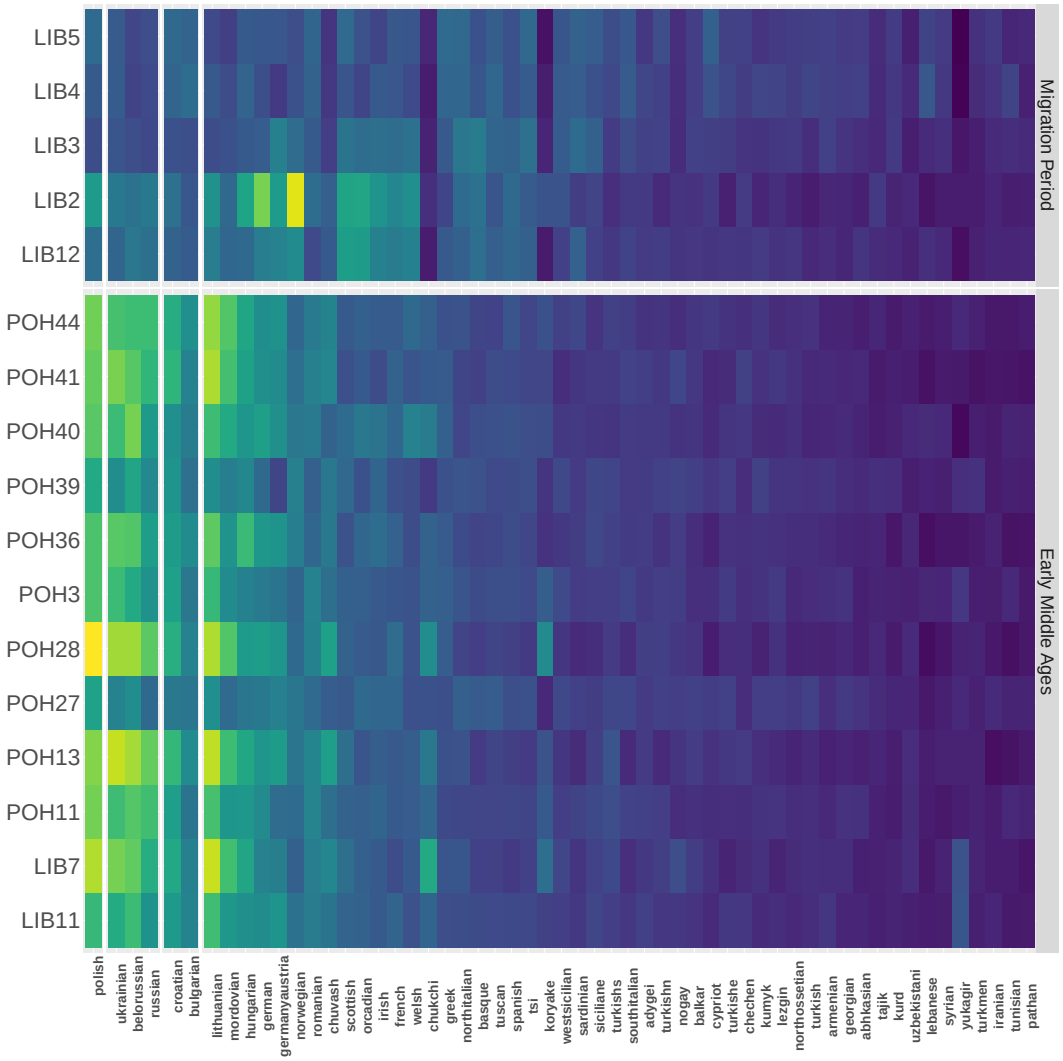


Figure 5.8: Raw chunklengths matrix from the 'present-day' painting. Rows correspond to different ancient recipient individuals, grouped into Migration Period and Early Middle Age period, and columns to different donor populations. Colour of cells corresponds to the total length of genome that a given 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. In particular, we find that samples copy many more haplotypes from present-day day Polish individuals than they do from other populations. This is consistent with previous findings based on uniparental markers. There was also a strong affinity to several non-Slavic

speaking present-day populations - notably Lithuania and Mordovian.

To confirm that the observed results were not a result of phasing or imputing ancient individuals using present-day samples, I utilised f_3 statistics, which were performed on non-imputed 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 haplotype-based methods, with Early Middle Age ancient Slavic individuals being closest to samples from Eastern Europe (Fig. 5.9). 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 present-day Croatians.

It should be noted that f_3 statistics have the potential to be biased towards drifted groups (explain some more about this later).

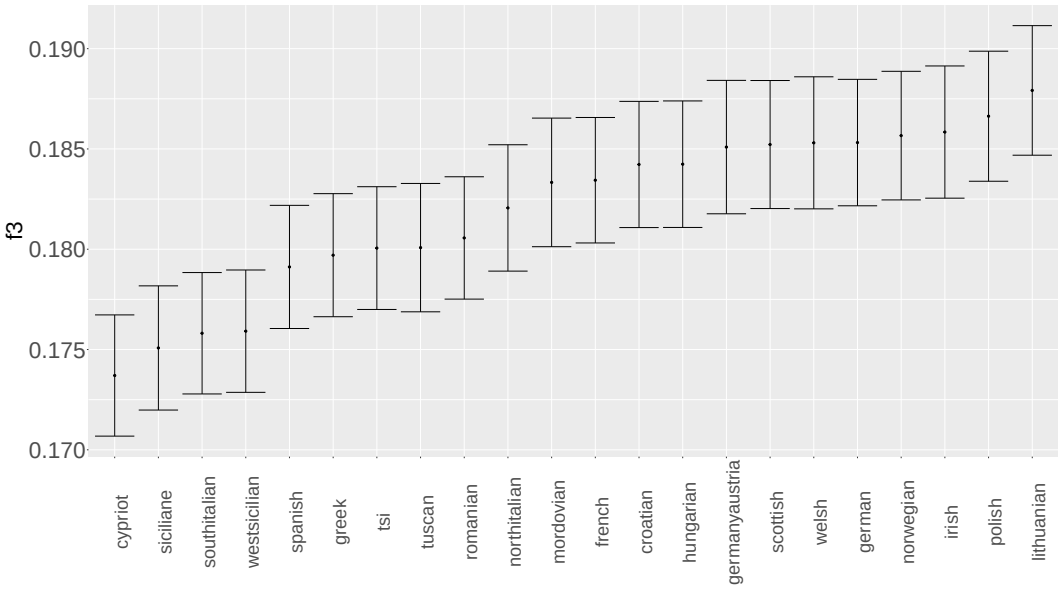


Figure 5.9: f_3 statistics in the form of $f_3(EMA, present{-}day; mbutipygmy)$, where *present-day* is different present-day European population. Error bars rerpesent $\pm * 2$ standard error.

5.3.6 Continuity with present-day day Slavs

The previous section strongly suggests at least some degree of continuity between Early Middle Age samples and present day Slavic populations that is not shared with the samples from the Migration Period, as the Early Middle Age samples share many more haplotypes with present-day Slavs compared to Migration Period.

To explicitly test the hypothesis that the Early Middle Age samples were continuous with the present-day day Slavic populations, I used *qpWave*, which tests the number of streams of ancestry from a set of *right* populations into a set of *left* populations, $qpWave(left = croatian, lithuanian, polish, ukrainian, right = middleage, migration)$. The matrix with rank $r = 0$ can be rejected ($p = 0.112$). Note - not sure how to interpret this.

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

As described in the introduction, several studies have investigated the structure of present-day Slavic populations, but none have integrated autosomal DNA from present-day and ancient samples and analysed them jointly with haplotype-based methods. I performed an all-v-all painting of a selection of present-day European populations and all newly sequenced ancient Slavic samples and applied the fineSTRUCTURE algorithm to the resulting chunkcounts matrix, inferring 32 clusters.

Present-day Slavs do not form a monophyletic group within the fineSTRUCTURE dendrogram to the exclusion of non-Slavic populations (Fig. 5.11), as several non-Slavic speaking populations such as German, Irish and Scottish cluster in the main clade containing Slavic speakers. Within Slavs, structure is apparent; speakers of 'Southern' Slavic languages from Croatia and Bulgaria form a group to the exclusion of 'Eastern' Slavic speaking populations from

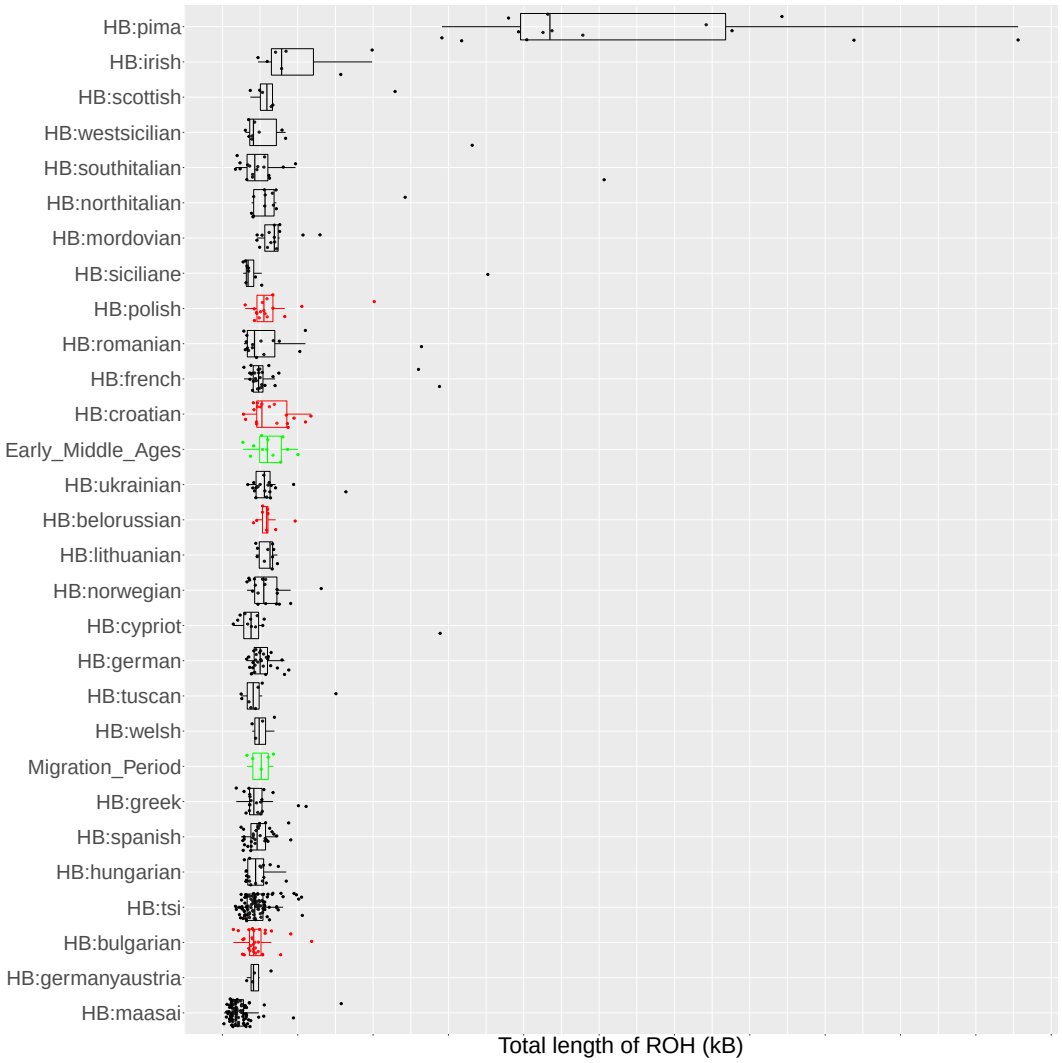


Figure 5.10: Total length of runs-of-homozygosity (ROH) in different present-day and ancient populations. Each point is the total length of ROH (kB) within an individual in that population. Points given jitter to aid visualisation. HB:pima and HB:masasai included to display extremes of ROH in different present-day human populations.

Belarus, Russia and Ukraine. Individuals from Poland cluster with 'Eastern' Slavic speakers, suggesting the principle axis of variation splits populations into 'North-West' and 'South-East' groups.

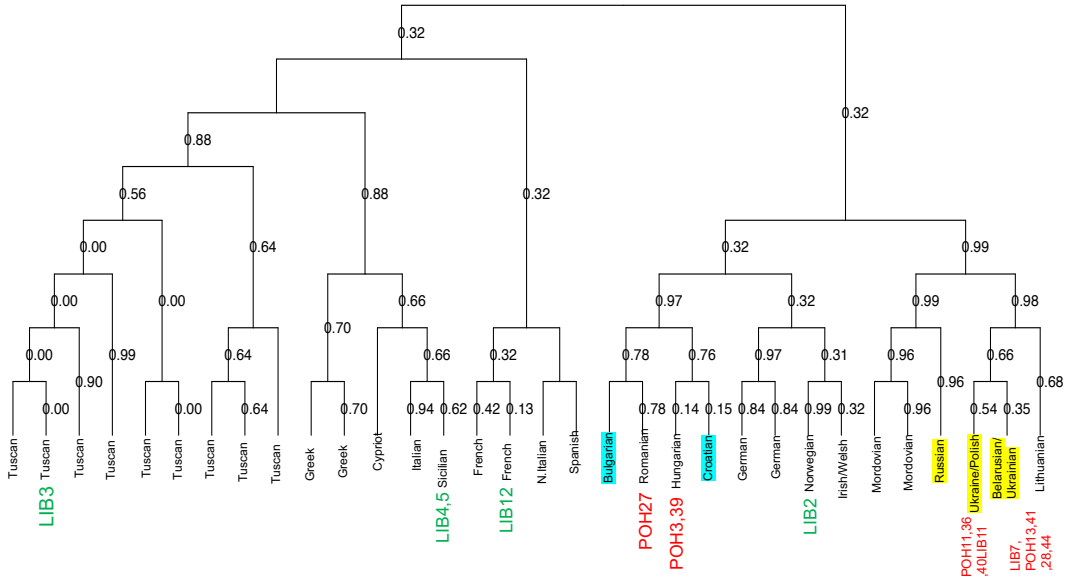


Figure 5.11: Population dendrogram generated by the fineSTRUCTURE tree building algorithm. Labeled tips refer to the primary population(s) represented in that clade. present-day non-Slavic populations shown in black. 'South-East' Slavs highlighted in cyan and 'North-West' Slavs highlighted in yellow. 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.

Of the Early Middle Age samples, 3 samples (POH3, POH39, POH27) were present in the 'South-East' Slavic cluster, falling into a group composed of Bulgarian and Romanian samples. The remaining 7 samples are found in the 'North-West' cluster containing samples from Lithuania, Poland, Ukraine and Belarus. Painting the samples using present-day individuals has thus uncovered structure that was not able to be detected by looking only at ancient samples. It also suggests the structure of Slavic populations into was present at least as early as the date of these samples.

Previous studies have identified admixture events in present-day Slavic

populations involving an East Asian source approximately 440 to 1080 CE [122, 164]. 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 [122] to replicate these results and determine whether a similar admixing source is present in the ancient populations.

When considering 2-way admixture event, all of the tested populations, bar the Migration Period Slavs, showed evidence of an admixture event involving a minor source which has the lowest Fst with present-day Uygurs. The dates and bootstrapped confidence intervals are given in Fig. 5.12. Other than Norwegians and Croatians, whose estimated dates are later and earlier respectively, the admixture dates for other populations appear to be constrained to approximately 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.

Interestingly, most present-day Slavic speaking populations, such as present-day Polish, show evidence of a 3-way admixture event, where the middle component has the lowest F_{st} with Migration Era ancient samples (Fig. 5.13). 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 an admixture event(s) 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 source best represented by present-day Uygurs.

5.4 Discussion

The combined results from the Migration Period suggest the individuals living in Czechia during this time period were of mixed ancestry and did not originate from the same source population. The diverse set of ancestries, spanning from

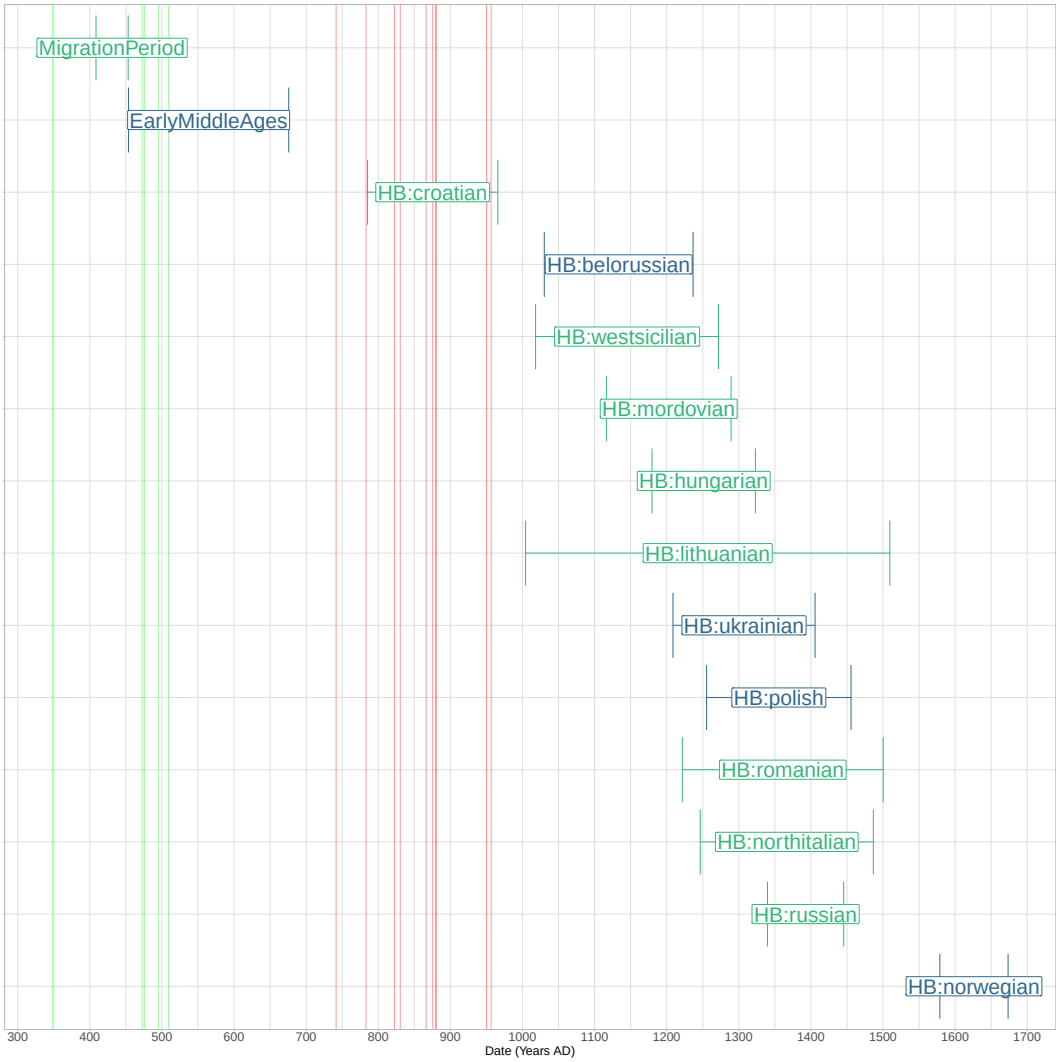


Figure 5.12: MOSAIC inferred 2-way admixture dates with bootstrapped 97.5% and 2.5% CI. Vertical green lines correspond to radiocarbon estimated dates of Migration Period samples and red lines equivalent for Early Middle Age samples. Estimated dates obtained by assuming an average generation time of 26 and date of birth of 1950 for present-day samples.

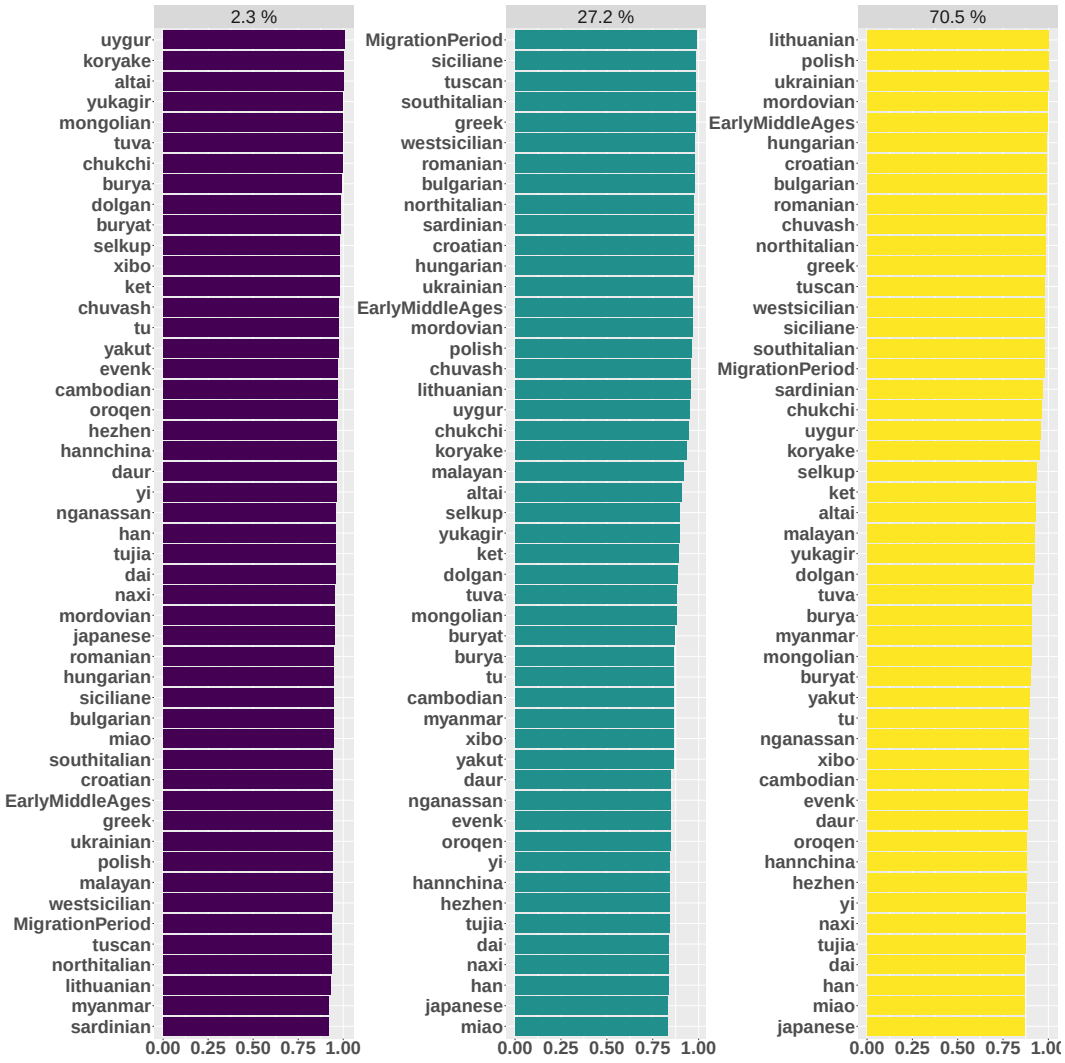


Figure 5.13: $1 - F_{st}$ between 3 inferred mixing sources for present-day Belorussians. Each panel represent a different mixing source. Each bar gives the value $1 - F_{st}$ between that samples population and the mixing source. Higher values of $1 - F_{st}$ suggest that source is well represented by a particular population.

Scandinavia to Southern Europe imply that the Migration Period was truly a period of Migration where individuals from distal ends of Europe lived among one another. In particular I inferred ancestry sources from Southern Europe and Scandinavia.

The results from the analysis of combined ancient and present-day genomes are consistent with those from Kushniarevich et al (2015) [150] 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 [165]. Similarly, I inferred that the Slavisiation of the Balkan peninsula doesn't extend beyond Croatia; the cluster of Croatian individuals only derives 1.2% of their ancestry from nearby Greek sources. However, admixture modeling suggested that Southern Slavs show signals of a historic admixture event where the minor source is related to present-day Mediterranean populations. An admixture event with a similar minor source is inferred in Migration period samples, albeit dated further in the past.

I recapitulated a previously described admixture event into not only present day Slavic speaking populations, but also Southern Europeans (e.g. North Italians). The source of this East-Asian admixture is closest to present-day Uygurs. However, the true ancient population that was repsonsible to transmitting East-Asian ancestry into Europe is yet to be determined. It seems likely that the ancestry was brought to Europe via an intermediate population containing East Asian ancestry, such as the Huns or Turkic peoples.

Chapter 6

General Conclusions

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

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

Paper	number of samples	Reference
Allentoft 2015	20	[166]
Antonio 2019	134	[42]
Broushaki 2016	1	[167]
Brunel 2020	58	[136]
Cassidy 2015	4	[168]
deBarrosDamgaard 2018a	34	[65]
deBarrosDamgaard 2018b	58	[169]
Gamba 2014	10	[132]
Gunther 2015	2	[170]
Hofmanova 2016	5	[125]
Jones 2015	2	[171]
Lazaridis 2014	1	[37]
Marchi 2020	4	[172]
Margaryan 20	442	[41]
Berger unpublished	14	NA
Olade 2014	1	[173]
Rivollat 20	101	[123]
Sanchez-Quinto 2019	7	[174]
Seguin-Orlando 2014	1	[175]
Veeramah 2018	1	[161]
Hofmanova unpublished	37	NA

A.1 Ancient reference dataset

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

The following steps were used to generate the data:

1. Each `.bam` was processed with `PicardTools ValidateBam` [68] 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 [53] (`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 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 [69]. 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 [75].

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 [68]
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 [118]

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 inds	data_collections
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
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,HGDP Transcriptome

(continued)

population_name	number of inds	data_collections
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 (GRCh37),Gambian Genome Variation Proje
Gambian Jola	100	Gambian Genome Variation Project (GRCh37),Gambian Genome Variation Proje
Gambian Mandinka	278	1000 Genomes
Gambian Mandinka,Gambian	2	1000 Genomes
Gambian Wolof	100	Gambian Genome Variation Project (GRCh37),Gambian Genome Variation Proje
Georgian	2	Simons Genome Diversity Project
Greek	2	Simons Genome Diversity Project
Gujarati	113	1000 Genomes
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

(continued)

population_name	number of inds	data_collections
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,HGDP Transcriptome
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,Human Genome Diversity Project
Mayan,Maya	2	Simons Genome Diversity Project,HGDP Transcriptome
Mbuti	13	Human Genome Diversity Project
Mbuti,Mbuti	2	Simons Genome Diversity Project,HGDP Transcriptome
Mende	126	1000 Genomes

(continued)

population_name	number of inds	data_collections
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,HGDP Transcriptome
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,Human Genome Diversity Project
Papuan,Papuan Sepik	6	Simons Genome Diversity Project,Human Genome Diversity Project
Pathan	23	Human Genome Diversity Project
Pathan,Pathan	1	Simons Genome Diversity Project,HGDP Transcriptome
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	HGDP Transcriptome,Human Genome Diversity Project
Sardinian	27	Human Genome Diversity Project
She	10	Human Genome Diversity Project

(continued)

population_name	number of inds	data_collections
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.2: 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` [71], setting `-pbwt 8` and keeping all other parameters as default. The 1000 Genomes was used as as reference (section [**?**]). 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.

Samples from Sawcer et al (2011) [176] (10299 individuals from 15 pops), Leslie et al 2015 [81] (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 [71] setting `-pbwt-depth 8`.
3. Remove duplicated SNPs using bcftools norm [118]
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.

Study	Population	n_inds
HB	abhkasian	20
HB	adygei	17
HB	altai	13
HB	armenian	35
HB	balkar	19
HB	balochi	24
HB	bantukenya	11
HB	bantusouthafrica	8
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

(continued)

Study	Population	n_inds
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
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

(continued)

Study	Population	n_inds
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
HB	maya	21
HB	mbutipygmy	13
HB	meena	1
HB	meghawal	1
HB	melanesian	10

(continued)

Study	Population	n_inds
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

(continued)

Study	Population	n_inds
HB	selkup	10
HB	she	10
HB	siciliane	10
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

(continued)

Study	Population	n_inds
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
MS	Norway	953
MS	Poland	58
MS	Spain	205
MS	Sweden	1212
MS	UK	1854
POBI	UK	2039

A breakdown of the POBI populations:

V1	n_inds
UK_Cheshire	33
UK_Cornwall_and_Isles_of_Scilly	90
UK_Cumbria	195
UK_Devon	73
UK_Dorset	37
UK_Dumfries_and_Galloway	42
UK_Durham	54
UK_Dyfed	55
UK_East_Riding_of_Yorkshire_Unitary_Authority	32
UK_East_Sussex	34
UK_Fife	59

(continued)

V1	n_inds
UK_Gloucestershire	70
UK_Gwent	31
UK_Gwynedd	76
UK_Hampshire	26
UK_Kent	50
UK_Leicestershire	66
UK_Lincolnshire	104
UK_Merseyside	47
UK_Norfolk	98
UK_North_Yorkshire	64
UK_Northamptonshire	37
UK_Northern_Ireland	44
UK_Northumberland	50
UK_Nottinghamshire	57
UK_Orkney_Islands	96
UK_Oxfordshire	77
UK_Somerset	17
UK_South_Yorkshire	77
UK_Staffordshire	28
UK_Suffolk	82
UK_Surrey	24
UK_Tyne_and_Wear	54
UK_West_Sussex	26
UK_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. This may confound results, becase.... [sam: not clear how to word this]. To avoid

this effect, we may perform a 'leave-one-out' painting, where each population is painted seperately, and a single individual from each other population is removed from the set of donors.

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

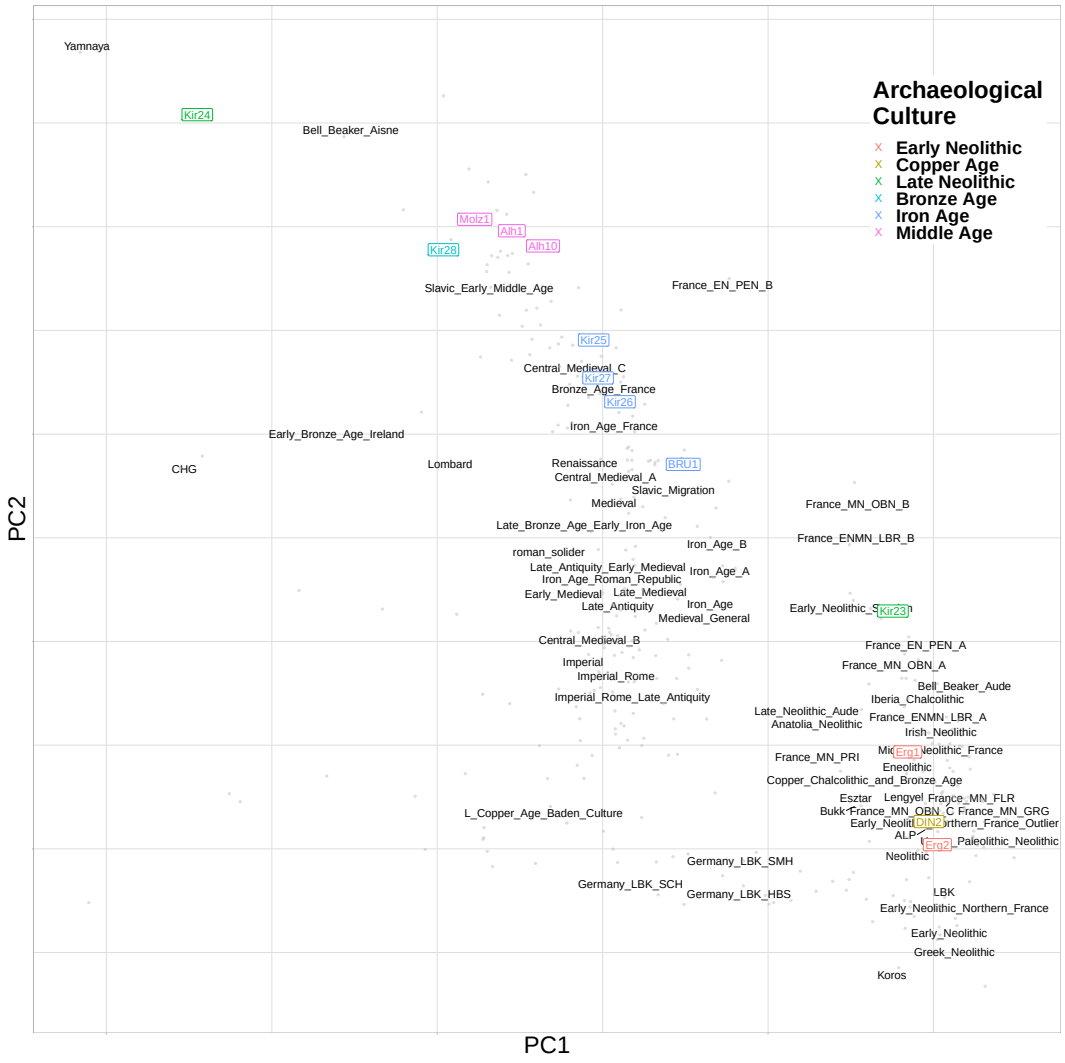


Figure D.1: Principle component analysis of genotype matrix using plink2. Grey points indicate principle component coordiantes for each sample. Black text indicated mean principle component coordinates for all individuals within that group. Coloured labels represent newly sequenced ancient samples.

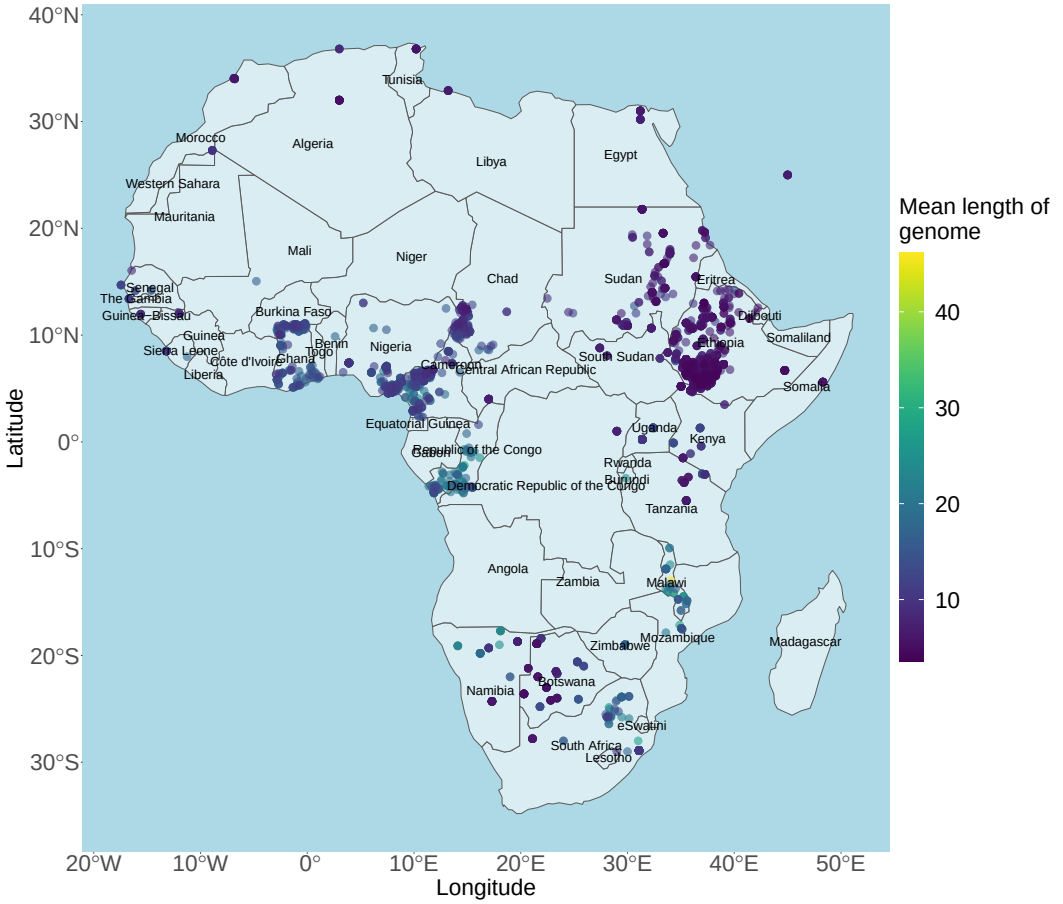


Figure D.2: Map of haplotype donation to U.K. Biobank individuals born in Brazil. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in Brazil.

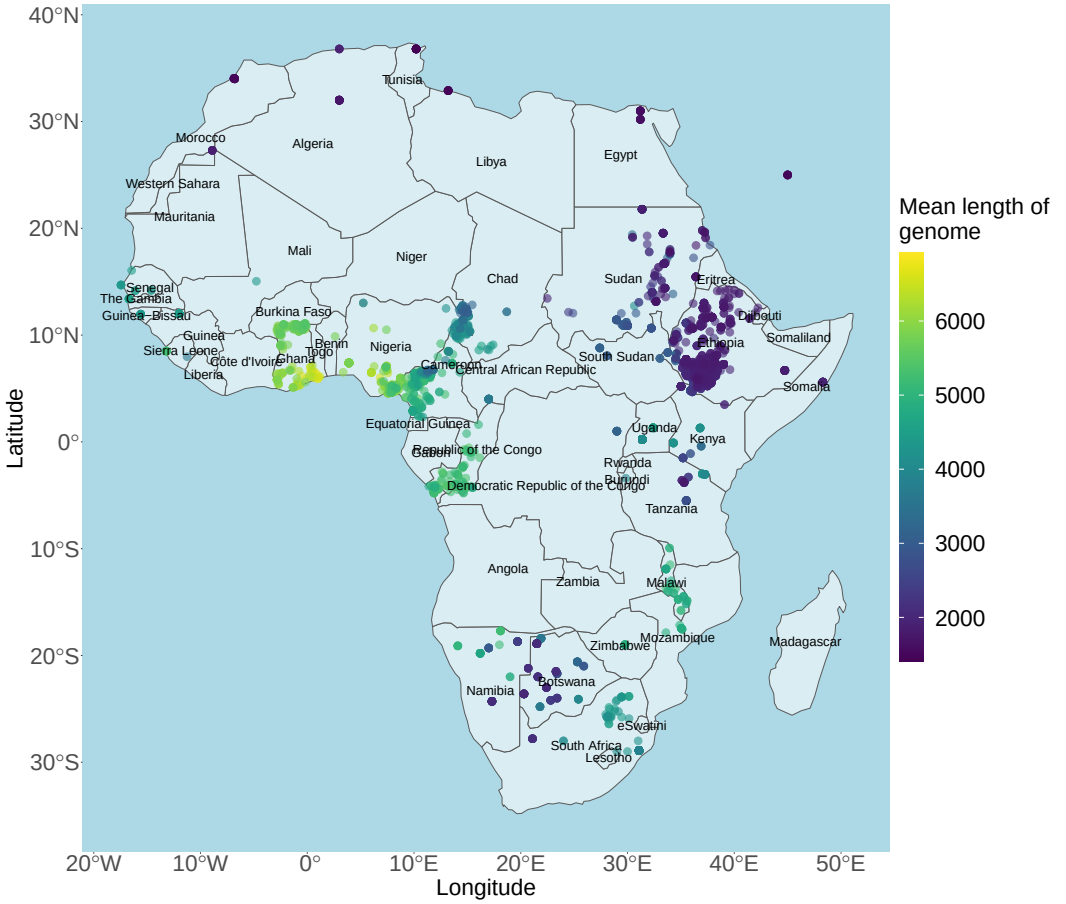


Figure D.3: PrMap of haplotype donation to U.K. Biobank individuals born in the Caribbean. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in the Caribbean.

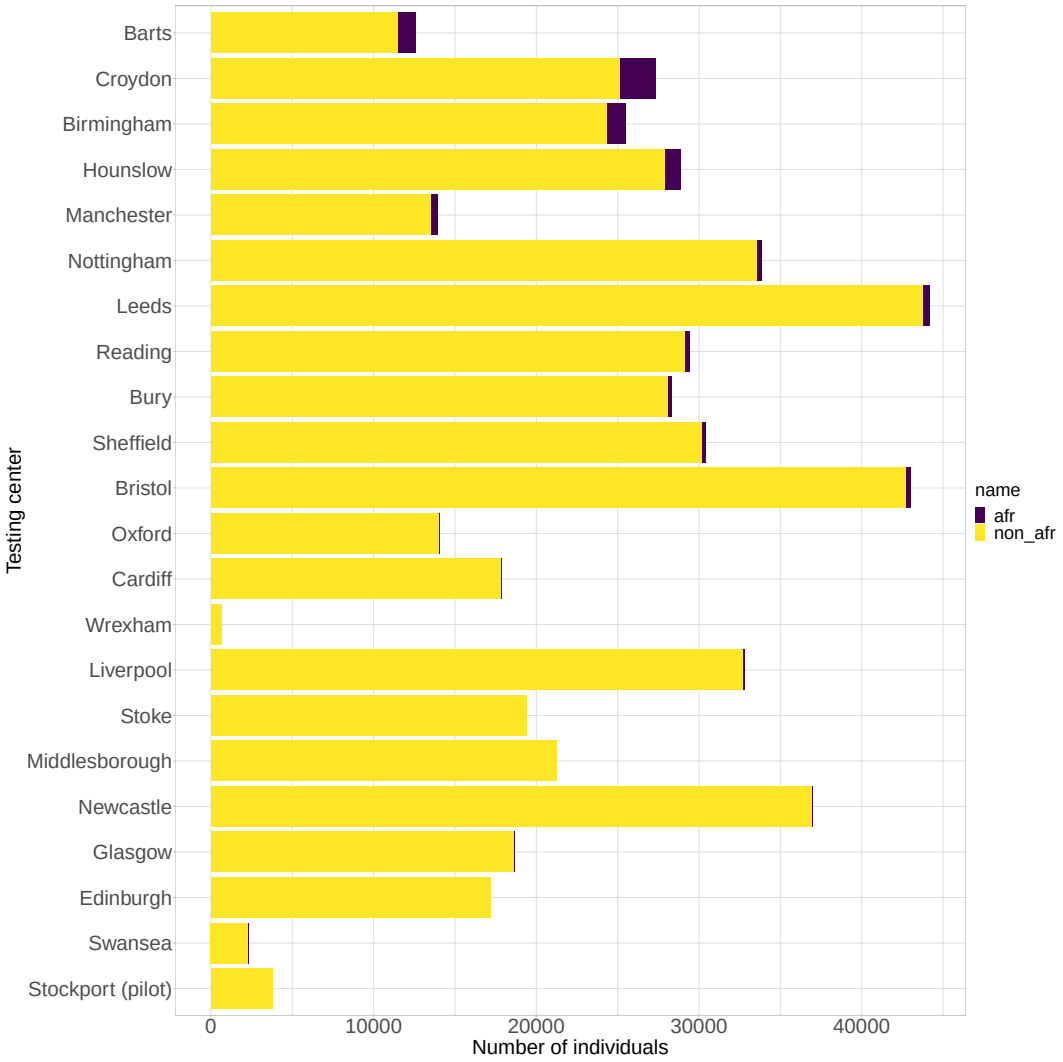


Figure D.4: Number of total individuals and proportion of total individuals who have at least 50% African ancestry by different testing centers. Centers ordered by proportion of individuals who have at least 50% African ancestry.

Appendix E

SOURCEFIND iteration test

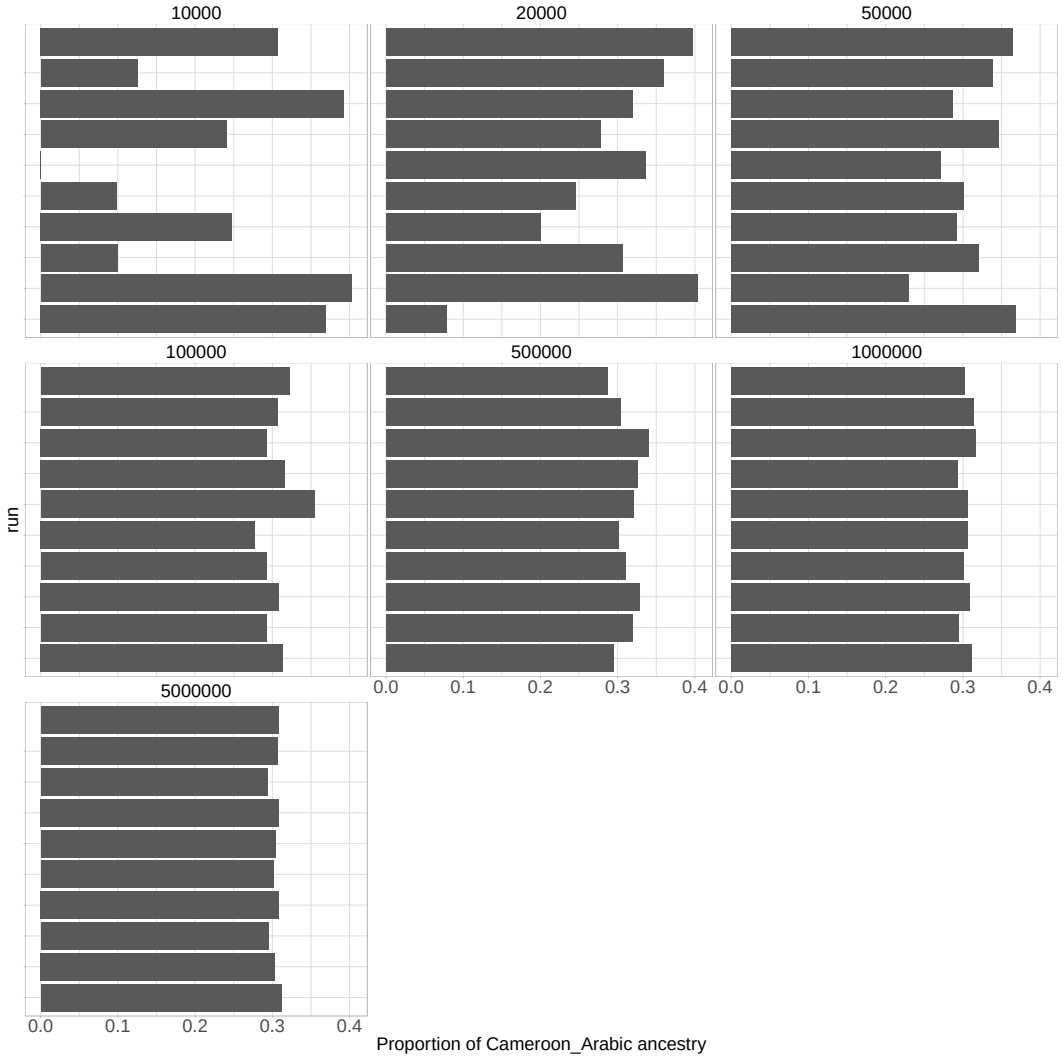


Figure E.1: Proportion of inferred Cameroon Arabic ancestry averaged across individuals from Cameroon Kanuri ethnic group. Each panel contains proportions for a different number of MCMC iterations. Within each panel, each bar is the proportion inferred from each of the 10 independent SOURCEFIND runs.

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

[5] Sarah A Tishkoff, Erin Dietzsch, William Speed, Andrew J Pakstis, Judith R Kidd, K Cheung, Batsheva Bonne-Tamir, A Silvana Santachiara-Benerecetti, Pedro Moral, Matthias Krings, et al. Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. *Science*, 271(5254):1380–1387, 1996.

[6] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, FL Yu, HM Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. 2003.

[7] Donald F Conrad, Mattias Jakobsson, Graham Coop, Xiaoquan Wen, Jeffrey D Wall, Noah A Rosenberg, and Jonathan K Pritchard. A

worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature genetics*, 38(11):1251–1260, 2006.

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

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

[10] Yun S Song. Na li and matthew stephens on modeling linkage disequilibrium. *Genetics*, 203(3):1005–1006, 2016.

[11] 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.

[12] 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.

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

[14] Mattias Jakobsson, Sonja W Scholz, Paul Scheet, J Raphael Gibbs, Jenna M VanLiere, Hon-Chung Fung, Zachary A Szpiech, James H Degnan, Kai Wang, Rita Guerreiro, et al. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451(7181):998–1003, 2008.

[15] 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.

[16] Garrett Hellenthal, George B.J. J. Busby, Gavin Band, James F. Wilson, Cristian Capelli, Daniel Falush, and Simon Myers. A Genetic Atlas of Human Admixture History. *Science*, 343(6172):747–751, 2014.

[17] Juan C. Chacon-Duque, Kaustubh Adhikari, Macarena Fuentes-Guajardo, Javier Mendoza-Revilla, Victor Acuna-Alonzo, Rodrigo Barquera Lozano, Mirsha Quinto-Sanchez, Jorge Gomez-Valdes, Paola Everardo Martinez, Hugo Villamil-Ramirez, Tabita Hunemeier, Virginia Ramallo, Caio C. Silva de Cerqueira, Malena Hurtado, Valeria Villegas, Vanessa Granja, Mercedes Villena, Rene Vasquez, Elena Llop, Jose R. Sandoval, Alberto A. Salazar-Granara, Maria-Laura Parolin, Karla Sandoval, Rosenda I. Penaloza-Espinosa, Hector Rangel-Villalobos, Cheryl Winkler, William Klitz, Claudio Bravi, Julio Molina, Daniel Corach, Ramiro Barrantes, Veronica Gomes, Carlos Resende, Leonor Gusmao, Antonio Amorim, Yali Xue, Jean-Michel Dugoujon, Pedro Moral, Rolando Gonzalez-Jose, Lavinia Schuler-Faccini, Francisco M. Salzano, Maria-Catira Bortolini, Samuel Canizales-Quinteros, Giovanni Poletti, Carla Gallo, Gabriel Bedoya, Francisco Rothhammer, David Balding, Garrett Hellenthal, and Andres Ruiz-Linares. Latin Americans show wide-spread Converso ancestry and the imprint of local Native ancestry on physical appearance. *Nature Communications*, page 252155, 2018.

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

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

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

[21] Anders Bergström, Shane A. McCarthy, Ruoyun Hui, Mohamed A. Almarri, Qasim Ayub, Petr Danecek, Yuan Chen, Sabine Felkel, Pille Hallast, Jack Kamm, Hélène Blanché, Jean-François Deleuze, Howard Cann, Swapan Mallick, David Reich, Manjinder S. Sandhu, Pontus Skoglund, Aylwyn Scally, Yali Xue, Richard Durbin, and Chris Tyler-Smith. Insights into human genetic variation and population history from 929 diverse genomes. *Science*, 367(6484):eaay5012, 2020.

[22] Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic structure of human populations. *science*, 298(5602):2381–2385, 2002.

[23] Sohini Ramachandran, Omkar Deshpande, Charles C Roseman, Noah A Rosenberg, Marcus W Feldman, and L Luca Cavalli-Sforza. 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.

[24] Anne M Bowcock, Andres Ruiz-Linares, James Tomfohrde, Eric Minch, Judith R Kidd, and L Luca Cavalli-Sforza. High resolution of human evolutionary trees with polymorphic microsatellites. *Nature*, 368(6470):455–457, 1994.

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

[26] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, 1000 Genomes Project, et al. Demographic history and rare

allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

[27] Timothy D O'Connor, Wenqing Fu, NHLBI GO Exome Sequencing Project, ESP Population Genetics, Emily Turner Statistical Analysis Working Group, Josyf C Mychaleckyj, Benjamin Logsdon, Paul Auer, Christopher S Carlson, Suzanne M Leal, Joshua D Smith, et al. Rare variation facilitates inferences of fine-scale population structure in humans. *Molecular biology and evolution*, 32(3):653–660, 2015.

[28] H.A. Green, R.E., Krause, J., Briggs, A., W., Maricic, T., Stenzel, U., Kircher, M., Patterson, N., Li, H., Zhai, W., Fritz, M.H., Hansen, N.F., Durand, E., Y., Malaspinas, A., Jensen, J., D., Marques-Bonet, T., Alkan, C., Prüfer, K., Meyer, M., Burbano. A Draft Sequence of the Neandertal Genome. *Science (New York, N.Y.)*, 328(5979):710–22, 2010.

[29] Nick Patterson, Priya Moorjani, Yontao Luo, Swapan Mallick, Nadin Rohland, Yiping Zhan, Teri Genschoreck, Teresa Webster, and David Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012.

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

[31] 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.

[32] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[33] Kendra A Sirak, Daniel M Fernandes, Mark Lipson, Swapan Mallick, Matthew Mah, Iñigo Olalde, Harald Ringbauer, Nadin Rohland, Carla S

Hadden, Éadaoin Harney, et al. Social stratification without genetic differentiation at the site of kulubnarti in christian period nubia. *bioRxiv*, 2021.

[34] 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.

[35] Rui Martiniano, Lara M. Cassidy, Ros Ó'Maoldúin, Russell McLaughlin, Nuno M. Silva, Licinio Manco, Daniel Fidalgo, Tania Pereira, Maria J. Coelho, Miguel Serra, Joachim Burger, Rui Parreira, Elena Moran, Antonio C. Valera, Eduardo Porfirio, Rui Boaventura, Ana M. Silva, Daniel G. Bradley, Maoldú In, Russell McLaughlin, Nuno M. Silva, Licinio Manco, Daniel Fidalgo, Tania Pereira, Maria J. Coelho, Miguel Serra, Joachim Burger, Rui Parreira, Elena Moran, Antonio C. Valera, Eduardo Porfirio, Rui Boaventura, Ana M. Silva, Daniel G. Bradley, Ros Ó'Maoldúin, Russell McLaughlin, Nuno M. Silva, Licinio Manco, Daniel Fidalgo, Tania Pereira, Maria J. Coelho, Miguel Serra, Joachim Burger, Rui Parreira, Elena Moran, Antonio C. Valera, Eduardo Porfirio, Rui Boaventura, Ana M. Silva, and Daniel G. Bradley. 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.

[36] Rui Martiniano, Erik Garrison, Eppie R Jones, Andrea Manica, and Richard Durbin. 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.

[37] Iosif Lazaridis, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, Joshua G. Schraiber, Sergi Castellano, Mark Lipson, Bonnie Berger, Christos Economou, Ruth Bollongino, Qiaomei Fu, Kirsten I. Bos, Susanne Nordenfelt, Heng Li,

Cesare De Filippo, Kay Prüfer, Susanna Sawyer, Cosimo Posth, Wolfgang Haak, Fredrik Hallgren, Elin Fornander, Nadin Rohland, Dominique Delsate, Michael Francken, Jean Michel Guinet, Joachim Wahl, George Ayodo, Hamza A. Babiker, Graciela Bailliet, Elena Balanovska, Oleg Balanovsky, Ramiro Barrantes, Gabriel Bedoya, Haim Ben-Ami, Judit Bene, Fouad Berrada, Claudio M. Bravi, Francesca Brisighelli, George B.J. J Busby, Francesco Cali, Mikhail Churnosov, David E.C. C Cole, Daniel Corach, Larissa Damba, George Van Driem, Stanislav Dryomov, Jean Michel Dugoujon, Sardana A. Fedorova, Irene Gallego Romero, Marina Gubina, Michael Hammer, Brenna M. Henn, Tor Hervig, Ugur Hodoglugil, Aashish R. Jha, Sena Karachanak-Yankova, Rita Khusainova, Elza Khusnutdinova, Rick Kittles, Toomas Kivisild, William Klitz, Vaidutis Kučinskas, Alena Kushniarevich, Leila Laredj, Sergey Litvinov, Theologos Loukidis, Robert W. Mahley, Béla Melegh, Ene Metspalu, Julio Molina, Joanna Mountain, Klemetti Näkkäläjärvi, Desislava Nesheva, Thomas Nyambo, Ludmila Osipova, Jüri Parik, Fedor Platonov, Olga Posukh, Valentino Romano, Francisco Rothhammer, Igor Rudan, Ruslan Ruizbakiev, Hovhannes Sahakyan, Antti Sajantila, Antonio Salas, Elena B. Starikovskaya, Ayele Tarekegn, Draga Toncheva, Shahlo Turdikulova, Ingrida Uktveryte, Olga Utevska, René Vasquez, Mercedes Villena, Mikhail Voevoda, Cheryl A. Winkler, Levon Yepiskoposyan, Pierre Zalloua, Tatijana Zemunik, Alan Cooper, Cristian Capelli, Mark G. Thomas, Andres Ruiz-Linares, Sarah A. Tishkoff, Lalji Singh, Kumarasamy Thangaraj, Richard Villems, David Comas, Rem Sukernik, Mait Metspalu, Matthias Meyer, Evan E. Eichler, Joachim Burger, Montgomery Slatkin, Svante Pääbo, Janet Kelso, David Reich, and Johannes Krause. Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature*, 513(7518):409–413, 2014.

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

[39] David Reich, Kumarasamy Thangaraj, Nick Patterson, Alkes L Price, and Lalji Singh. Reconstructing indian population history. *Nature*, 461(7263):489–494, 2009.

[40] Choongwon Jeong, Oleg Balanovsky, Elena Lukianova, Nurzhibek Kahbatkyzy, Pavel Flegontov, Valery Zaporozhchenko, Alexander Immel, Chuan-Chao Wang, Olzhas Ixan, Elmira Khussainova, et al. The genetic history of admixture across inner eurasia. *Nature ecology & evolution*, 3(6):966–976, 2019.

[41] Ashot Margaryan, Daniel J Lawson, Martin Sikora, Fernando Racimo, Simon Rasmussen, Ida Moltke, Lara M Cassidy, Emil Jørsboe, Andrés Ingason, Mikkel W Pedersen, et al. Population genomics of the viking world. *Nature*, 585(7825):390–396, 2020.

[42] Margaret L Antonio, Ziyue Gao, Hannah M Moots, Michaela Lucci, Francesca Candilio, Susanna Sawyer, Victoria Oberreiter, Diego Calderon, Katharina Devitofranceschi, Rachael C Aikens, et al. Ancient rome: a genetic crossroads of europe and the mediterranean. *Science*, 366(6466):708–714, 2019.

[43] Guy S. Jacobs, Georgi Hudjashov, Lauri Saag, Pradiptajati Kusuma, Chelzie C. Darusallam, Daniel J. Lawson, Mayukh Mondal, Luca Pagani, François-Xavier Ricaut, Mark Stoneking, Mait Metspalu, Herawati Sudoyo, J. Stephen Lansing, and Murray P. Cox. Multiple deeply divergent denisovan ancestries in papuans. *Cell*, 177(4):1010–1021.e32, 2019.

[44] João C Teixeira, Guy S Jacobs, Chris Stringer, Jonathan Tuke, Georgi Hudjashov, Gludhug A Purnomo, Herawati Sudoyo, Murray P Cox, Raymond Tobler, Chris SM Turney, 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.

[45] Yoshan Moodley, Andrea Brunelli, Silvia Ghirotto, Andrey Klyubin, Ayas S. Maady, William Tyne, Zilia Y. Muñoz-Ramirez, Zhemin Zhou, Andrea Manica, Bodo Linz, and Mark Achtman. Helicobacter pylori's historical journey through siberia and the americas. *Proceedings of the National Academy of Sciences*, 118(25), 2021.

[46] Ruoyun Hui, Eugenia D'Atanasio, Lara M Cassidy, Christiana L Scheib, and Toomas Kivisild. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Scientific reports*, 10(1):1–8, 2020.

[47] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 2010.

[48] Geraldine A. Van der Auwera, Mauricio O. Carneiro, Christopher Hartl, Ryan Poplin, Guillermo del Angel, Ami Levy-Moonshine, Tadeusz Jordan, Khalid Shakir, David Roazen, Joel Thibault, Eric Banks, Kiran V. Garimella, David Altshuler, Stacey Gabriel, and Mark A. DePristo. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Current Protocols in Bioinformatics*, 2013.

[49] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 2009.

[50] Ruiqiang Li, Yingrui Li, Xiaodong Fang, Huanming Yang, Jian Wang, Karsten Kristiansen, and Jun Wang. SNP detection for massively parallel whole-genome resequencing. *Genome Research*, 2009.

[51] Su Y. Kim, Kirk E. Lohmueller, Anders Albrechtsen, Yingrui Li, Thorfinn Korneliussen, Geng Tian, Niels Grarup, Tao Jiang, Gitte Andersen, Daniel Witte, Torben Jorgensen, Torben Hansen, Oluf Pedersen, Jun Wang, and

Rasmus Nielsen. Estimation of allele frequency and association mapping using next-generation sequencing data. *BMC Bioinformatics*, 2011.

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

[53] Vivian Link, Athanasios Kousathanas, Krishna Veeramah, Christian Sell, Amelie Scheu, and Daniel Wegmann. ATLAS: Analysis Tools for Low-depth and Ancient Samples. *bioRxiv*, page 105346, 2017.

[54] 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.

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

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

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

[58] Daniel Fernandes, Kendra Sirak, Mario Novak, John A Finarelli, John Byrne, Edward Connolly, Jeanette EL Carlsson, Edmondo Ferretti, Ron Pinhasi, and Jens Carlsson. The identification of a 1916 irish rebel: new approach for estimating relatedness from low coverage homozygous genomes. *Scientific reports*, 7(1):1–10, 2017.

[59] 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.

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

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

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

[63] 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.

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

[65] Peter de Barros Damgaard, Rui Martiniano, Jack Kamm, J. Víctor Moreno-Mayar, Guus Kroonen, Michaël Peyrot, Gojko Barjamovic, Simon Rasmussen, Claus Zacho, Nurbol Baimukhanov, Victor Zaibert, Victor Merz, Arjun Biddanda, Ilja Merz, Valeriy Loman, Valeriy Evdokimov, Emma Usmanova, Brian Hemphill, Andaine Seguin-Orlando, Fulya Eylem Yediay, Inam Ullah, Karl-Göran Sjögren, Katrine Højholt Iversen, Jeremy Choin, Constanza de la Fuente, Melissa Ilardo, Hannes Schroeder, Vyacheslav Moiseyev, Andrey Gromov, Andrei Polyakov, Sachihiro Omura, Süleyman Yücel Senyurt, Habib Ahmad, Catriona McKenzie, Ashot Margaryan, Abdul Hameed, Abdul Samad, Nazish Gul, Muhammad Hassan Khokhar, O. I. Goriunova, Vladimir I. Bazaliiskii, John Novembre, Andrzej W. Weber, Ludovic Orlando, Morten E. Allentoft, Rasmus Nielsen, Kristian Kristiansen, Martin Sikora, Alan K. Outram,

Richard Durbin, and Eske Willerslev. The first horse herders and the impact of early bronze age steppe expansions into asia. *Science*, 360(6396), 2018.

[66] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M. Slepchenko, Aleksei A. Bondarev, Philip L.F. Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare De Filippo, Matthias Meyer, Nicolas Zwyns, Domingo C. Salazar-García, Yaroslav V. Kuzmin, Susan G. Keates, Pavel A. Kosintsev, Dmitry I. Razhev, Michael P. Richards, Nikolai V. Peristov, Michael Lachmann, Katerina Douka, Thomas F.G. Higham, Montgomery Slatkin, Jean Jacques Hublin, David Reich, Janet Kelso, T. Bence Viola, and Svante Pääbo. Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature*, 2014.

[67] Torsten Günther, Helena Malmström, Emma M. Svensson, Ayça Omrak, Federico Sánchez-Quinto, Gülşah M. Kılınç, Maja Krzewińska, Gunilla Eriksson, Magdalena Fraser, Hanna Edlund, Arielle R. Munters, Alexandra Coutinho, Luciana G. Simões, Mário Vicente, Anders Sjölander, Berit Jansen Sellevold, Roger Jørgensen, Peter Claes, Mark D. Shriver, Cristina Valdiosera, Mihai G. Netea, Jan Apel, Kerstin Lidén, Birgitte Skar, Jan Storå, Anders Götherström, and Mattias Jakobsson. Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biology*, 2018.

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

[69] 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, Gonçalo R. Abecasis, David M. Altshuler, Richard M. Durbin, David R. Bentley, Aravinda Chakravarti, Andrew G. Clark, Peter Donnelly, Evan E. Eichler, Paul Flicek, Stacey B. Gabriel, Richard A. Gibbs, Eric D. Green,

Matthew E. Hurles, Bartha M. Knoppers, Jan O. Korbel, Eric S. Lander, Charles Lee, Hans Lehrach, Elaine R. Mardis, Gabor T. Marth, Gil A. McVean, Deborah A. Nickerson, Jeanette P. Schmidt, Stephen T. Sherry, Jun Wang, Richard K. Wilson, Eric Boerwinkle, Harsha Doddapaneni, Yi Han, Viktoriya Korchina, Christie Kovar, Sandra Lee, Donna Muzny, Jeffrey G. Reid, Yiming Zhu, Yuqi Chang, Qiang Feng, Xiaodong Fang, Xiaosen Guo, Min Jian, Hui Jiang, Xin Jin, Tianming Lan, Guoqing Li, Jingxiang Li, Yun Yingrui Li, Shengmao Liu, Xiaoming Liu, Yao Lu, Xuedi Ma, Meifang Tang, Bo Wang, Guangbiao Wang, Honglong Wu, Renhua Wu, Xun Xu, Ye Yin, Dandan Zhang, Wenwei Zhang, Jiao Zhao, Meiru Zhao, Xiaole Zheng, Namrata Gupta, Neda Gharani, Lorraine H. Toji, Norman P. Gerry, Alissa M. Resch, Jonathan Barker, Laura Clarke, Laurent Gil, Sarah E. Hunt, Gavin Kelman, Eugene Kulesha, Rasko Leinonen, William M. McLaren, Rajesh Radhakrishnan, Asier Roa, Dmitriy Smirnov, Richard E. Smith, Ian Streeter, Anja Thormann, Iliana Toneva, Brendan Vaughan, Xiangqun Zheng-Bradley, Russell Grocock, Sean Humphray, Terena James, Zoya Kingsbury, Ralf Sudbrak, Marcus W. Albrecht, Vyacheslav S. Amstislavskiy, Tatiana A. Borodina, Matthias Lienhard, Florian Mertes, Marc Sultan, Bernd Timmermann, Marie Laure Yaspo, Lucinda Fulton, Victor Ananiev, Zinaida Belaia, Dimitriy Beloslyudtsev, Nathan Bouk, Chao Chen, Deanna Church, Robert Cohen, Charles Cook, John Garner, Timothy Hefferon, Mikhail Kimelman, Chunlei Liu, John Lopez, Peter Meric, Chris O'Sullivan, Yuri Ostapchuk, Lon Phan, Sergiy Ponomarov, Valerie Schneider, Eugene Shekhtman, Karl Sirotkin, Douglas Slotta, Hua Zhang, Senduran Balasubramaniam, John Burton, Petr Danecek, Thomas M. Keane, Anja Kolb-Kokocinski, Shane McCarthy, James Stalker, Michael Quail, Christopher J. Davies, Jeremy Gollub, Teresa Webster, Brant Wong, Yiping Zhan, Christopher L. Campbell, Yu Kong, Anthony Marcketta, Fuli Yu, Lilian Antunes, Matthew Bainbridge, Aniko Sabo, Zhuoyi Huang, Lachlan J.M. Coin, Lin Fang,

Qibin Li, Zhenyu Li, Haoxiang Lin, Binghang Liu, Ruibang Luo, Haojing Shao, Yinlong Xie, Chen Ye, Chang Yu, Fan Zhang, Hancheng Zheng, Hongmei Zhu, Can Alkan, Elif Dal, Fatma Kahveci, Erik P. Garrison, Deniz Kural, Wan Ping Lee, Wen Fung Leong, Michael Stromberg, Alistair N. Ward, Jiantao Wu, Mengyao Zhang, Mark J. Daly, Mark A. DePristo, Robert E. Handsaker, Eric Banks, Gaurav Bhatia, Guillermo Del Angel, Giulio Genovese, Heng Li, Seva Kashin, Steven A. McCarroll, James C. Nemesh, Ryan E. Poplin, Seungtai C. Yoon, Jayon Lihm, Vladimir Makarov, Srikanth Gottipati, Alon Keinan, Juan L. Rodriguez-Flores, Tobias Rausch, Markus H. Fritz, Adrian M. Stütz, Kathryn Beal, Avik Datta, Javier Herrero, Graham R.S. Ritchie, Daniel Zerbino, Pardis C. Sabeti, Ilya Shlyakhter, Stephen F. Schaffner, Joseph Vitti, David N. Cooper, Edward V. Ball, Peter D. Stenson, Bret Barnes, Markus Bauer, R. Keira Cheetham, Anthony Cox, Michael Eberle, Scott Kahn, Lisa Murray, John Peden, Richard Shaw, Eimear E. Kenny, Mark A. Batzer, Miriam K. Konkel, Jerilyn A. Walker, Daniel G. MacArthur, Monkol Lek, Ralf Herwig, Li Ding, Daniel C. Koboldt, David Larson, Kai Kenny Ye, Simon Gravel, Anand Swaroop, Emily Chew, Tuuli Lappalainen, Yaniv Erlich, Melissa Gymrek, Thomas Frederick Willems, Jared T. Simpson, Mark D. Shriver, Jeffrey A. Rosenfeld, Carlos D. Bustamante, Stephen B. Montgomery, Francisco M. De La Vega, Jake K. Byrnes, Andrew W. Carroll, Marianne K. DeGorter, Phil Lacroute, Brian K. Maples, Alicia R. Martin, Andres Moreno-Estrada, Suyash S. Shringarpure, Fouad Zakharia, Eran Halperin, Yael Baran, Eliza Cerveira, Jaeho Hwang, Ankit Malhotra, Dariusz Plewczynski, Kamen Radew, Mallory Romanovitch, Chengsheng Zhang, Fiona C.L. Hyland, David W. Craig, Alexis Christoforides, Nils Homer, Tyler Izatt, Ahmet A. Kurdoglu, Shripad A. Sinari, Kevin Squire, Chunlin Xiao, Jonathan Sebat, Danny Antaki, Madhusudan Gujral, Amina Noor, Kai Kenny Ye, Esteban G. Burchard, Ryan D. Hernandez, Christopher R. Gignoux, David Haussler, Sol J. Katzman,

W. James Kent, Bryan Howie, Andres Ruiz-Linares, Emmanouil T. Dermitzakis, Scott E. Devine, Hyun Min Kang, Jeffrey M. Kidd, Tom Blackwell, Sean Caron, Wei Chen, Sarah Emery, Lars Fritsche, Christian Fuchsberger, Goo Jun, Bingshan Li, Robert Lyons, Chris Scheller, Carlo Sidore, Shiya Song, Elzbieta Sliwerska, Daniel Taliun, Adrian Tan, Ryan Welch, Mary Kate Wing, Xiaowei Zhan, Philip Awadalla, Alan Hodgkinson, Yun Yingrui Li, Xinghua Shi, Andrew Quitadamo, Gerton Lunter, Jonathan L. Marchini, Simon Myers, Claire Churchhouse, Olivier Delaneau, Anjali Gupta-Hinch, Warren Kretzschmar, Zamin Iqbal, Iain Mathieson, Androniki Menelaou, Andy Rimmer, Dionysia K. Xifara, Taras K. Oleksyk, Yunxin Yao Fu, Xiaoming Liu, Momiao Xiong, Lynn Jorde, David Witherspoon, Jinchuan Xing, Brian L. Browning, Sharon R. Browning, Fereydoun Hormozdiari, Peter H. Sudmant, Ekta Khurana, Chris Tyler-Smith, Cornelis A. Albers, Qasim Ayub, Yuan Chen, Vincenza Colonna, Luke Jostins, Klaudia Walter, Yali Xue, Mark B. Gerstein, Alexej Abyzov, Suganthi Balasubramanian, Jieming Chen, Declan Clarke, Yunxin Yao Fu, Arif O. Harmanci, Mike Jin, Donghoon Lee, Jeremy Liu, Xinmeng Jasmine Mu, Jing Zhang, Yan Yujun Zhang, Chris Hartl, Khalid Shakir, Jeremiah Degenhardt, Sascha Meiers, Benjamin Raeder, Francesco Paolo Casale, Oliver Stegle, Eric Wubbo Lameijer, Ira Hall, Vineet Bafna, Jacob Michaelson, Eugene J. Gardner, Ryan E. Mills, Gargi Dayama, Ken Chen, Xian Fan, Zechen Chong, Tenghui Chen, Mark J. Chaisson, John Huddleston, Maika Malig, Bradley J. Nelson, Nicholas F. Parrish, Ben Blackburne, Sarah J. Lindsay, Zemin Ning, Yan Yujun Zhang, Hugo Lam, Cristina Sisu, Danny Challis, Uday S. Evani, James Lu, Uma Nagaswamy, Jin Yu, Wangshen Li, Lukas Habegger, Haiyuan Yu, Fiona Cunningham, Ian Dunham, Kasper Lage, Jakob Berg Jespersen, Heiko Horn, Donghoon Kim, Rob Desalle, Apurva Narechania, Melissa A.Wilson Sayres, Fernando L. Mendez, G. David Poznik, Peter A. Underhill, David Mittelman, Ruby Banerjee, Maria Cerezo, Thomas W.

Fitzgerald, Sandra Louzada, Andrea Massaia, Fengtang Yang, Divya Kalra, Walker Hale, Xu Dan, Kathleen C. Barnes, Christine Beiswanger, Hongyu Cai, Hongzhi Cao, Brenna Henn, Danielle Jones, Jane S. Kaye, Alastair Kent, Angeliki Kerasidou, Rasika A. Mathias, Pilar N. Ossorio, Michael Parker, Charles N. Rotimi, Charmaine D. Royal, Karla Sandoval, Yeyang Su, Zhongming Tian, Sarah Tishkoff, Marc Via, Yuhong Wang, Huanming Yang, Ling Yang, Jiayong Zhu, Walter Bodmer, Gabriel Bedoya, Zhiming Cai, Yang Gao, Jiayou Chu, Leena Peltonen, Andres Garcia-Montero, Alberto Orfao, Julie Dutil, Juan C. Martinez-Cruzado, Rasika A. Mathias, Anselm Hennis, Harold Watson, Colin McKenzie, Firdausi Qadri, Regina LaRocque, Xiaoyan Deng, Danny Asogun, Onikepe Folarin, Christian Happi, Omonwunmi Omoniwa, Matt Stremlau, Ridhi Tariyal, Muminatou Jallow, Fatoumatta Sisay Joof, Tumani Corrah, Kirk Rockett, Dominic Kwiatkowski, Jaspal Kooner, Tran Tinh Hien, Sarah J. Dunstan, Nguyen ThuyHang, Richard Fonnie, Robert Garry, Lansana Kanneh, Lina Moses, John Schieffelin, Donald S. Grant, Carla Gallo, Giovanni Poletti, Danish Saleheen, Asif Rasheed, Lisa D. Brooks, Adam L. Felsenfeld, Jean E. McEwen, Yekaterina Vaydylevich, Audrey Duncanson, Michael Dunn, Jeffery A. Schloss, 1000 Genomes Project Consortium, Adam Auton, Lisa D. Brooks, Richard M. Durbin, Erik P. Garrison, Hyun Min Kang, Jan O. Korbel, Jonathan L. Marchini, Shane McCarthy, Gil A. McVean, and Gonçalo R. Abecasis. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[70] 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.

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

[72] Lucy Huang, Yun Li, Andrew B. Singleton, John A. Hardy, Gonçalo Abecasis, Noah A. Rosenberg, and Paul Scheet. Genotype-imputation accuracy across worldwide human populations. *The American Journal of Human Genetics*, 84(2):235–250, 2009.

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

[74] 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.

[75] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, et al. High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *bioRxiv*, 2021.

[76] 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.

[77] John G Cleary, Ross Braithwaite, Kurt Gaastra, Brian S Hilbush, Stuart Inglis, Sean A Irvine, Alan Jackson, Richard Littin, Sahar Nohzadeh-Malakshah, Mehul Rathod, 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.

[78] Wolfgang Haak, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, Susanne Nordenfelt, Eadaoin Harney, Kristin Stewardson, Qiaomei Fu, Alissa Mittnik, Eszter Bánffy,

Christos Economou, Michael Francken, Susanne Friederich, Rafael Garrido Pena, Fredrik Hallgren, Valery Khartanovich, Aleksandr Khokhlov, Michael Kunst, Pavel Kuznetsov, Harald Meller, Oleg Mochalov, Vayacheslav Moiseyev, Nicole Nicklisch, Sandra L. Pichler, Roberto Risch, Manuel A. Rojo Guerra, Christina Roth, Anna Szécsényi-Nagy, Joachim Wahl, Matthias Meyer, Johannes Krause, Dorcas Brown, David Anthony, Alan Cooper, Kurt Werner Alt, and David Reich. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, 2015.

[79] 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.

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

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

[82] Po-Ru Loh, Petr Danecek, Pier Francesco Palamara, Christian Fuchsberger, Yakir A Reshef, Hilary K Finucane, Sebastian Schoenherr, Lukas Forer, Shane McCarthy, Goncalo R Abecasis, et al. Reference-based phasing using the haplotype reference consortium panel. *Nature genetics*, 48(11):1443–1448, 2016.

[83] W Haak, P Forster, B Bramanti, S Matsumura, G Brandt, M Tänzer, R Villems, C Renfrew, D Gronenborn, K W Alt, and J Burger. Ancient

DNA from the first European farmer in 750-year-old Neolithic sites. *Science*, 310(November):1016–1019, 2005.

[84] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H. Sudmant, Cesare De Filippo, Heng Li, Swapan Mallick, Michael Dannemann, Qiaomei Fu, Martin Kircher, Martin Kuhlwilm, Michael Lachmann, Matthias Meyer, Matthias Ongyerth, Michael Siebauer, Christoph Theunert, Arti Tandon, Priya Moorjani, Joseph Pickrell, James C. Mullikin, Samuel H. Vohr, Richard E. Green, Ines Hellmann, Philip L.F. F Johnson, Hélène Blanche, Howard Cann, Jacob O. Kitzman, Jay Shendure, Evan E. Eichler, Ed S. Lein, Trygve E. Bakken, Liubov V. Golovanova, Vladimir B. Doronichev, Michael V. Shunkov, Anatoli P. Derevianko, Bence Viola, Montgomery Slatkin, David Reich, Janet Kelso, and Svante Pääbo. The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, 505(7481):43–49, 2014.

[85] 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.

[86] Augustine Kong, Gisli Masson, Michael L Frigge, Arnaldur Gylfason, Pasha Zusmanovich, Gudmar Thorleifsson, Pall I Olason, Andres Ingason, Stacy Steinberg, Thorunn Rafnar, et al. Detection of sharing by descent, long-range phasing and haplotype imputation. *Nature genetics*, 40(9):1068–1075, 2008.

[87] 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.

[88] Qiaomei Fu, Cosimo Posth, Mateja Hajdinjak, Martin Petr, Swapan Mallick, Daniel Fernandes, Anja Furtwängler, Wolfgang Haak, Matthias Meyer, Alissa Mittnik, Birgit Nickel, Alexander Peltzer, Nadin Rohland, Viviane Slon, Sahra Talamo, Iosif Lazaridis, Mark Lipson, Iain Mathieson, Stephan Schiffels, Pontus Skoglund, Anatoly P. Derevianko, Nikolai Drozdov, Vyacheslav Slavinsky, Alexander Tsybankov, Renata Grifoni Cremonesi, Francesco Mallegni, Bernard Gély, Eligio Vacca, Manuel R.González Morales, Lawrence G. Straus, Christine Neugebauer-Maresch, Maria Teschler-Nicola, Silviu Constantin, Oana Teodora Moldovan, Stefano Benazzi, Marco Peresani, Donato Coppola, Martina Lari, Stefano Ricci, Annamaria Ronchitelli, Frédérique Valentin, Corinne Thevenet, Kurt Wehrberger, Dan Grigorescu, Hélène Rougier, Isabelle Crevecoeur, Damien Flas, Patrick Semal, Marcello A. Mannino, Christophe Cupillard, Hervé Bocherens, Nicholas J. Conard, Katerina Harvati, Vyacheslav Moiseyev, Dorothée G. Drucker, Jiří Svoboda, Michael P. Richards, David Caramelli, Ron Pinhasi, Janet Kelso, Nick Patterson, Johannes Krause, Svante Pääbo, and David Reich. The genetic history of Ice Age Europe. *Nature*, 534(7606):200–205, 2016.

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

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

[92] 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.

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

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

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

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

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

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

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

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

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

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

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

[104] 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.

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

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

[107] Elena Bosch, Hafid Laayouni, Carlos Morcillo-Suarez, Ferran Casals, Andrés Moreno-Estrada, Anna Ferrer-Admetlla, Michelle Gardner, Araceli

Rosa, Arcadi Navarro, David Comas, et al. Decay of linkage disequilibrium within genes across hgdp-ceph human samples: most population isolates do not show increased ld. *BMC genomics*, 10(1):1–9, 2009.

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

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

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

[111] Lucy Van Dorp, Sara Lowes, Jonathan L Weigel, Naser Ansari-Pour, Saioa López, Javier Mendoza-Revilla, James A Robinson, Joseph Henrich, Mark G Thomas, Nathan Nunn, et al. Genetic legacy of state centralization in the kuba kingdom of the democratic republic of the congo. *Proceedings of the National Academy of Sciences*, 116(2):593–598, 2019.

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

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

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

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

[116] Nicholas J Conard. A female figurine from the basal aurignacian of hohle fels cave in southwestern germany. *Nature*, 459(7244):248–252, 2009.

[117] 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.

[118] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, and Richard Durbin. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[119] Hansi Weissensteiner, Dominic Pacher, Anita Kloss-Brandstätter, Lukas Forer, Günther Specht, Hans-Jürgen Bandelt, Florian Kronenberg, Antonio Salas, and Sebastian Schönherr. Haplogrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic acids research*, 44(W1):W58–W63, 2016.

[120] 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.

[121] Po-Ru Loh, Mark Lipson, Nick Patterson, Priya Moorjani, Joseph K Pickrell, David Reich, and Bonnie Berger. Inferring Admixture Histories of Human Populations Using Linkage Disequilibrium. *Genetics*, 193(4):1233–1254, 04 2013.

[122] 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.

[123] Maïté Rivollat, Choongwon Jeong, Stephan Schiffels, İşil Küçükkalıpçı, Marie-Hélène Pemonge, Adam Benjamin Rohrlach, Kurt W. Alt, Didier Binder, Susanne Friederich, Emmanuel Ghesquière, Detlef Gronenborn, Luc Laporte, Philippe Lefranc, Harald Meller, Hélène Réveillas, Eva Rosenstock, Stéphane Rottier, Chris Scarre, Ludovic Soler, Joachim Wahl, Johannes Krause, Marie-France Deguilloux, and Wolfgang Haak. Ancient genome-wide dna from france highlights the complexity of interactions between mesolithic hunter-gatherers and neolithic farmers. *Science Advances*, 6(22), 2020.

[124] Mark Lipson, Anna Szécsényi-Nagy, Swapan Mallick, Annamária Pósa, Balázs Stégmár, Victoria Keerl, Nadin Rohland, Kristin Stewardson, Matthew Ferry, Megan Michel, Jonas Oppenheimer, Nasreen Broomandkhoshbacht, Eadaoin Harney, Susanne Nordenfelt, Bastien Llamas, Balázs Mende Gusztáv, Kitti Köhler, Krisztián Oross, Mária Bondár, Tibor Marton, Anett Osztás, János Jakucs, Tibor Paluch, Ferenc Horváth, Piroska Csengeri, Judit Koós, Katalin Sebok, Alexandra Anders, Pál Raczky, Judit Regenye, Judit P. Barna, Szilvia Fábián, Gábor Serlegi, Zoltán Toldi, Emese Gyöngyvér Nagy, János Dani, Erika Molnár, György Pálfi, László Márk, Béla Melegh, Zsolt Bánfai, László Domboróczki, Javier Fernández-Eraso, José Antonio Mujika-Alustiza, Carmen Alonso Fernández, Javier Jiménez Echevarría, Ruth Bollongino, Jörg Orschiedt, Kerstin Schierhold, Harald Meller, Alan Cooper, Joachim Burger, Eszter Bánffy, Kurt W. Alt, Carles Lalueza-Fox, Wolfgang Haak, and David Reich. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. *Nature*, 551(7680):368–372, 2017.

[125] Zuzana Hofmanová, Susanne Kreutzer, Garrett Hellenthal, Christian Sell, Yoan Diekmann, David Díez-del Molino, Lucy van Dorp, Saioa López, Athanasios Kousathanas, Vivian Link, Karola Kirsanow, Lara M. Cassidy, Rui Martiniano, Melanie Strobel, Amelie Scheu, Kostas Kotsakis, Paul Halstead, Sevi Triantaphyllou, Nina Kyparissi-Apostolika,

Dushka Urem-Kotsou, Christina Ziota, Fotini Adaktylou, Shyamalika Gopalan, Dean M. Bobo, Laura Winkelbach, Jens Blöcher, Martina Unterländer, Christoph Leuenberger, Çiler Çilingiroğlu, Barbara Horejs, Fokke Gerritsen, Stephen J. Shennan, Daniel G. Bradley, Mathias Currat, Krishna R. Veeramah, Daniel Wegmann, Mark G. Thomas, Christina Papageorgopoulou, and Joachim Burger. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences*, 2016.

[126] Wolfgang Haak, Oleg Balanovsky, Juan J. Sanchez, Sergey Koshel, Valery Zaporozhchenko, Christina J. Adler, Clio S.I. I der Sarkissian, Guido Brandt, Carolin Schwarz, Nicole Nicklisch, Veit Dresely, Barbara Fritsch, Elena Balanovska, Richard Villems, Harald Meller, Kurt W. Alt, and Alan Cooper. Ancient DNA from European early Neolithic farmers reveals their near eastern affinities. *PLoS Biology*, 8(11), 2010.

[127] Wolfgang Haak, Peter Forster, Barbara Bramanti, Shuichi Matsumura, Guido Brandt, Marc Tänzer, Richard Villems, Colin Renfrew, Detlef Gronenborn, Kurt Werner Alt, et al. Ancient dna from the first european farmers in 7500-year-old neolithic sites. *Science*, 310(5750):1016–1018, 2005.

[128] Barbara Bramanti, Mark G Thomas, Wolfgang Haak, Martina Unterländer, Pia Jores, Kristiina Tambets, Indre Antanaitis-Jacobs, Miriam N Haidle, Rimantas Jankauskas, C-J Kind, et al. Genetic discontinuity between local hunter-gatherers and central europe's first farmers. *science*, 326(5949):137–140, 2009.

[129] Eva Fernández, Alejandro Pérez-Pérez, Cristina Gamba, Eva Prats, Pedro Cuesta, Josep Anfruns, Miquel Molist, Eduardo Arroyo-Pardo, and Daniel Turbón. 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.

[130] Iosif Lazaridis, Dani Nadel, Gary Rollefson, Deborah C. Merrett, Nadin Rohland, Swapan Mallick, Daniel Fernandes, Mario Novak, Beatriz Gamarra, Kendra Sirak, Sarah Connell, Kristin Stewardson, Eadaoin Harney, Qiaomei Fu, Gloria Gonzalez-Fortes, Eppie R. Jones, Songül Alpaslan Roodenberg, György Lengyel, Fanny Bocquentin, Boris Gasparian, Janet M. Monge, Michael Gregg, Vered Eshed, Ahuva Sivan Mizrahi, Christopher Meiklejohn, Fokke Gerritsen, Luminita Bejenaru, Matthias Blüher, Archie Campbell, Gianpiero Cavalleri, David Comas, Philippe Froguel, Edmund Gilbert, Shona M. Kerr, Peter Kovacs, Johannes Krause, Darren McGettigan, Michael Merrigan, D. Andrew Merriwether, Seamus O'Reilly, Martin B. Richards, Ornella Semino, Michel Shamoon-Pour, Gheorghe Stefanescu, Michael Stumvoll, Anke Tönjes, Antonio Torroni, James F. Wilson, Loic Yengo, Nelli A. Hovhannisyan, Nick Patterson, Ron Pinhasi, and David Reich. Genomic insights into the origin of farming in the ancient Near East. *Nature*, 536(7617):419–424, 2016.

[131] Iain Mathieson, Iosif Lazaridis, Nadin Rohland, Swapan Mallick, Nick Patterson, Songül Alpaslan Roodenberg, Eadaoin Harney, Kristin Stewardson, Daniel Fernandes, Mario Novak, Kendra Sirak, Cristina Gamba, Eppie R. Jones, Bastien Llamas, Stanislav Dryomov, Joseph Pickrell, Juan Luís Arsuaga, José María Bermúdez De Castro, Eudald Carbonell, Fokke Gerritsen, Aleksandr Khokhlov, Pavel Kuznetsov, Marina Lozano, Harald Meller, Oleg Mochalov, Vyacheslav Moiseyev, Manuel A.Rojo Guerra, Jacob Roodenberg, Josep Maria Vergès, Johannes Krause, Alan Cooper, Kurt W. Alt, Dorcas Brown, David Anthony, Carles Lalueza-Fox, Wolfgang Haak, Ron Pinhasi, and David Reich. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.

[132] Cristina Gamba, Eppie R. Jones, Matthew D. Teasdale, Russell L. McLaughlin, Gloria Gonzalez-Fortes, Valeria Mattiangeli, László Domboróczki, Ivett Kővári, Ildikó Pap, Alexandra Anders, Alasdair Whittle, János Dani, Pál Raczky, Thomas F. G. Higham, Michael Hofreiter,

Daniel G. Bradley, and Ron Pinhasi. Genome flux and stasis in a five millennium transect of European prehistory. *Nature Communications*, 5:5257, 2014.

[133] Mark Lipson, Anna Szécsényi-Nagy, Swapan Mallick, Annamária Pósa, Balázs Stégmár, Victoria Keerl, Nadin Rohland, Kristin Stewardson, Matthew Ferry, Megan Michel, Jonas Oppenheimer, Nasreen Broomandkhoshbacht, Eadaoin Harney, Susanne Nordenfelt, Bastien Llamas, Balázs Gusztáv Mende, Kitti Köhler, Krisztián Oross, Mária Bondár, Tibor Marton, Anett Osztás, János Jakucs, Tibor Paluch, Ferenc Horváth, Piroska Csengeri, Judit Koós, Katalin Sebok, Alexandra Anders, Pál Raczky, Judit Regenye, Judit P. Barna, Szilvia Fábián, Gábor Serlegi, Zoltán Toldi, Emese Gyöngyvér Nagy, János Dani, Erika Molnár, György Pálfi, László Márk, Béla Melegh, Zsolt Bánfai, Javier Fernández-Eraso, José Antonio Mujika-Alustiza, Carmen Alonso Fernández, Javier Jiménez Echevarría, Ruth Bollongino, Jörg Orschiedt, Kerstin Schierhold, Harald Meller, Alan Cooper, Joachim Burger, Eszter Bánffy, Kurt W. Alt, Carles Lalueza-Fox, Wolfgang Haak, and David Reich. Parallel ancient genomic transects reveal complex population history of early European farmers. *Nature*, page 114488, 2017.

[134] Fernando Racimo, Jessie Woodbridge, Ralph M. Fyfe, Martin Sikora, Karl-Göran Sjögren, Kristian Kristiansen, and Marc Vander Linden. The spatiotemporal spread of human migrations during the european holocene. *Proceedings of the National Academy of Sciences*, 117(16):8989–9000, 2020.

[135] Christine Keyser, Caroline Bouakaze, Eric Crubézy, Valery G Nikolaev, Daniel Montagnon, Tatiana Reis, and Bertrand Ludes. Ancient dna provides new insights into the history of south siberian kurgan people. *Human genetics*, 126(3):395–410, 2009.

[136] Samantha Brunel, E. Andrew Bennett, Laurent Cardin, Damien Garraud, Hélène Barrand Emam, Alexandre Beylier, Bruno Boulestin, Fanny Chenal, Elsa Ciesielski, Fabien Convertini, Bernard Dedet, Stéphanie Desbrosse-Degobertiere, Sophie Desenne, Jerôme Dubouloz, Henri Duday, Gilles Escalon, Véronique Fabre, Eric Gailledrat, Muriel Gandelin, Yves Gleize, Sébastien Goepfert, Jean Guilaine, Lamys Hachem, Michael Ilett, François Lambach, Florent Maziere, Bertrand Perrin, Suzanne Plouin, Estelle Pinard, Ivan Praud, Isabelle Richard, Vincent Riquier, Réjane Roure, Benoit Sendra, Corinne Thevenet, Sandrine Thiol, Elisabeth Vauquelin, Luc Vergnaud, Thierry Grange, Eva-Maria Geigl, and Melanie Pruvost. 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.

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

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

[139] Paul Fouracre, Rosamond McKitterick, David Abulafia, Timothy Reuter, David Edward Luscombe, CT Allmand, Michael CE Jones, Jonathan Riley-Smith, Michael Jones, et al. *The New Cambridge Medieval History: Volume 1, C. 500-c. 700*. Number 1. Cambridge University Press, 1995.

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

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

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

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

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

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

[146] Anna Juras, Miroslawa Dabert, Alena Kushniarevich, Helena Malmström, Maanasa Raghavan, Jakub Z. Kosicki, Ene Metspalu, Eske Willerslev, and Janusz Piontek. Ancient dna reveals matrilineal continuity in present-day poland over the last two millennia. *PLOS ONE*, 9(10):1–9, 10 2014.

[147] 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.

[148] 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.

[149] Cosimo Posth, Christoph Wißing, Keiko Kitagawa, Luca Pagani, Laura van Holstein, Fernando Racimo, Kurt Wehrberger, Nicholas J Conard, Claus Joachim Kind, Hervé Bocherens, et al. Deeply divergent archaic mitochondrial genome provides lower time boundary for african gene flow into neanderthals. *Nature communications*, 8(1):1–9, 2017.

[150] Alena Kushniarevich, Olga Utevska, Marina Chuhryaeva, Anastasia Agdzhoyan, Khadizhat Dibirova, Ingrida Uktveryte, Märt Möls, Lejla Mulahasanovic, Andrey Pshenichnov, Svetlana Frolova, Andrey Shanko, Ene Metspalu, Maere Reidla, Kristiina Tambets, Erika Tamm, Sergey Koshel, Valery Zaporozhchenko, Lubov Atramentova, Vaidutis Kučinskas,

Oleg Davydenko, Olga Goncharova, Irina Evseeva, Michail Churnosov, Elvira Pocheshchova, Bayazit Yunusbayev, Elza Khusnutdinova, Damir Marjanović, Pavao Rudan, Siiri Rootsi, Nick Yankovsky, Phillip Endicott, Alexei Kassian, Anna Dybo, The Genographic Consortium, Chris Tyler-Smith, Elena Balanovska, Mait Metspalu, Toomas Kivisild, Richard Villems, and Oleg Balanovsky. 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.

[151] Jiří Macháček, Robert Nedoma, Petr Dresler, Ilektra Schulz, Elias Lagonik, Stephen M. Johnson, Ludmila Kaňáková, Alena Slámová, Bastien Llamas, Daniel Wegmann, and Zuzana Hofmanová. 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.

[152] Vasili Pankratov, Sergei Litvinov, Alexei Kassian, Dzmitry Shulhin, Lieve Tchebotarev, Bayazit Yunusbayev, Märt Möls, Hovhannes Sahakyan, Levon Yepiskoposyan, Siiri Rootsi, 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.

[153] 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.

[154] Pengfei Qin, Ying Zhou, Haiyi Lou, Dongsheng Lu, Xiong Yang, Yuchen Wang, Li Jin, Yeun-Jun Chung, and Shuhua Xu. Quantitating and dating recent gene flow between european and east asian populations. *Scientific reports*, 5(1):1–8, 2015.

[155] Peter Ralph and Graham Coop. The geography of recent genetic ancestry across europe. *PLOS Biology*, 11(5):1–20, 05 2013.

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

[157] 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.

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

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

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

[161] Krishna R. Veeramah, Andreas Rott, Melanie Groß, Lucy van Dorp, Saioa López, Karola Kirsanow, Christian Sell, Jens Blöcher, Daniel Wegmann, Vivian Link, Zuzana Hofmanová, Joris Peters, Bernd Trautmann, Anja Gairhos, Jochen Haberstroh, Bernd Päffgen, Garrett Hellenthal, Brigitte Haas-Gebhard, Michaela Harbeck, and Joachim Burger. Population genomic analysis of elongated skulls reveals extensive female-biased immigration in Early Medieval Bavaria. *Proceedings of the National Academy of Sciences*, 2018.

[162] F Lotter. Völkerverschiebungen im ostalpen–mitteldonau–raum zwischen antike und mittelalter (365–600). *Gra Ergänzungsband*, 39, 2003.

[163] Iosif Lazaridis, Alissa Mittnik, Nick Patterson, Swapan Mallick, Nadin Rohland, Saskia Pfrengle, Anja Furtwängler, Alexander Peltzer, Cosimo Posth, Andonis Vasilakis, et al. Genetic origins of the minoans and mycenaeans. *Nature*, 548(7666):214–218, 2017.

[164] Garrett Hellenthal, Daniel Falush, Simon Myers, David Reich, George B.J. Busby, Mark Lipson, Cristian Capelli, and Nick Patterson. The Kalash Genetic Isolate? the Evidence for Recent Admixture. *American Journal of Human Genetics*, 98(2):396–397, 2016.

[165] Krishna R Veeramah, Anke Tönjes, Peter Kovacs, Arnd Gross, Daniel Wegmann, Patrick Geary, Daniela Gasperikova, Iwar Klimes, Markus Scholz, John Novembre, 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.

[166] Morten E. Allentoft, Martin Sikora, Karl Göran Sjögren, Simon Rasmussen, Morten Rasmussen, Jesper Stenderup, Peter B. Damgaard, Hannes Schroeder, Torbjörn Ahlström, Lasse Vinner, Anna Sapfo Malaspinas, Ashot Margaryan, Tom Higham, David Chivall, Niels Lynnerup, Lise Harvig, Justyna Baron, Philippe Della Casa, Paweł Dąbrowski, Paul R. Duffy, Alexander V. Ebel, Andrey Epimakhov, Karin Frei, Mirosław Furmanek, Tomasz Gralak, Andrey Gromov, Stanisław Gronkiewicz, Gisela Grupe, Tamás Hajdu, Radosław Jarysz, Valeri Khartanovich, Alexandr Khokhlov, Viktória Kiss, Jan Kolář, Aivar Kriiska, Irena Lasak, Cristina Longhi, George McGlynn, Algimantas Merkevicius, Inga Merkyte, Mait Metspalu, Ruzan Mkrtchyan, Vyacheslav Moiseyev, László Paja, György Pálfi, Dalia Pokutta, Łukasz Pospieszny, T. Douglas Price, Lehti Saag, Mikhail Sablin, Natalia Shishlina, Václav Smrčka, Vasilii I. Soenov, Vajk Szeverényi, Gusztáv Tóth, Synaru V. Trifanova, Liivi Varul, Magdolna Vicze, Levon Yepiskoposyan, Vladislav Zhitenev, Ludovic Orlando, Thomas Sicheritz-Pontén, Søren Brunak, Rasmus Nielsen, Kristian Kristiansen, and Eske Willerslev. Population genomics of Bronze Age Eurasia. *Nature*, 2015.

[167] Farnaz Broushaki, Mark G. Thomas, Vivian Link, Saioa López, Lucy van Dorp, Karola Kirsanow, Zuzana Hofmanová, Yoan Diekmann, Lara M.

Cassidy, David Díez-del Molino, Athanasios Kousathanas, Christian Sell, Harry K. Robson, Rui Martiniano, Jens Blöcher, Amelie Scheu, Susanne Kreutzer, Ruth Bollongino, Dean Bobo, Hossein Davoudi, Olivia Munoz, Mathias Currat, Kamyar Abdi, Fereidoun Biglari, Oliver E. Craig, Daniel G. Bradley, Stephen Shennan, Krishna R. Veeramah, Marjan Mashkour, Daniel Wegmann, Garrett Hellenthal, and Joachim Burger. Early Neolithic genomes from the eastern Fertile Crescent. *Science*, 2016.

[168] Lara M. Cassidy, Rui Martiniano, Eileen M. Murphy, Matthew D. Teasdale, James Mallory, Barrie Hartwell, and Daniel G. Bradley. 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.

[169] Peter de Barros Damgaard, Nina Marchi, Simon Rasmussen, Michaël Peyrot, Gabriel Renaud, Thorfinn Korneliussen, J Víctor Moreno-Mayar, Mikkel Winther Pedersen, Amy Goldberg, Emma Usmanova, et al. 137 ancient human genomes from across the eurasian steppes. *Nature*, 557(7705):369–374, 2018.

[170] Torsten Günther, Cristina Valdiosera, Helena Malmström, Irene Ureña, Ricardo Rodriguez-Varela, Óddny Osk Sverrisdóttir, Evangelia A Daskalaki, Pontus Skoglund, Thijessen Naidoo, Emma M Svensson, 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.

[171] Eppie R. Jones, Gloria Gonzalez-Fortes, Sarah Connell, Veronika Siska, Anders Eriksson, Rui Martiniano, Russell L. McLaughlin, Marcos Gallego Llorente, Lara M. Cassidy, Cristina Gamba, Tengiz Meshveliani, Ofer Bar-Yosef, Werner Müller, Anna Belfer-Cohen, Zinovi Matskevich, Nino Jakeli, Thomas F.G. Higham, Mathias Currat, David Lordkipanidze, Michael Hofreiter, Andrea Manica, Ron Pinhasi, and Daniel G. Bradley. Upper

Palaeolithic genomes reveal deep roots of modern Eurasians. *Nature Communications*, 6:1–8, 2015.

[172] Nina Marchi, Laura Winkelbach, Ilektra Schulz, Maxime Brami, Zuzana Hofmanová, Jens Blocher, Carlos S Reyna-Blanco, Yoan Diekmann, Alexandre Thiéry, Adamandia Kapopoulou, et al. The mixed genetic origin of the first farmers of europe. *bioRxiv*, 2020.

[173] Inigo Olalde, Morten E Allentoft, Federico Sánchez-Quinto, Gabriel Santpere, Charleston WK Chiang, Michael DeGiorgio, Javier Prado-Martinez, Juan Antonio Rodríguez, Simon Rasmussen, Javier Quilez, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old mesolithic european. *Nature*, 507(7491):225–228, 2014.

[174] Federico Sánchez-Quinto, Helena Malmström, Magdalena Fraser, Linus Girdland-Flink, Emma M Svensson, Luciana G Simões, Robert George, Nina Hollfelder, Göran Burenhult, Gordon Noble, 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.

[175] Andaine Seguin-Orlando, Thorfinn S. Korneliussen, Martin Sikora, Anna-sapfo Malaspinas, Andrea Manica, Ida Moltke, Michael Westaway, David Lambert, Valeri Khartanovich, Jeffrey D Wall, Philip R Nigst, and Robert A Foley. Genomic structure in Europeans dating back at least 36 , 200 years. *Science*, 346(6213):1113–1118, 2014.

[176] Stephen Sawcer, Garrett Hellenthal, Matti Pirinen, Chris C.A. Spencer, Nikolaos A. Patsopoulos, Loukas Moutsianas, Alexander Dilthey, Zhan Su, Colin Freeman, Sarah E. Hunt, Sarah Edkins, Emma Gray, David R. Booth, Simon C. Potter, An Goris, Gavin Band, Annette Bang Oturai, Amy Strange, Janna Saarela, Céline Bellenguez, Bertrand Fontaine, Matthew Gillman, Bernhard Hemmer, Rhian Gwilliam, Frauke

Zipp, Alagurevathi Jayakumar, Roland Martin, Stephen Leslie, Stanley Hawkins, Eleni Giannoulatou, Sandra D'Alfonso, Hannah Blackburn, Filippo Martinelli Boneschi, Jennifer Liddle, Hanne F. Harbo, Marc L. Perez, Anne Spurkland, Matthew J. Waller, Marcin P. Mycko, Michelle Ricketts, Manuel Comabella, Naomi Hammond, Ingrid Kockum, Owen T. McCann, Maria Ban, Pamela Whittaker, Anu Kemppinen, Paul Weston, Clive Hawkins, Sara Widaa, John Zajicek, Serge Dronov, Neil Robertson, Suzannah J. Bumpstead, Lisa F. Barcellos, Rathi Ravindrarajah, Roby Abraham, Lars Alfredsson, Kristin Ardlie, Cristin Aubin, Amie Baker, Katharine Baker, Sergio E. Baranzini, Laura Bergamaschi, Roberto Bergamaschi, Allan Bernstein, Achim Berthele, Mike Boggild, Jonathan P. Bradfield, David Brassat, Simon A. Broadley, Dorothea Buck, Helmut Butzkueven, Ruggero Capra, William M. Carroll, Paola Cavalla, Elisabeth G. Celius, Sabine Cepok, Rosetta Chiavacci, Françoise Clerget-Darpoux, Katleen Clysters, Giancarlo Comi, Mark Cossburn, Isabelle Cournu-Rebeix, Mathew B. Cox, Wendy Cozen, Bruce A.C. Cree, Anne H. Cross, Daniele Cusi, Mark J. Daly, Emma Davis, Paul I.W. De Bakker, Marc Debouverie, Marie Beatrice D'Hooghe, Katherine Dixon, Rita Dobosi, Bénédicte Dubois, David Ellinghaus, Irina Elovaara, Federica Esposito, Claire Fontenille, Simon Foote, Andre Franke, Daniela Galimberti, Angelo Ghezzi, Joseph Glessner, Refujia Gomez, Olivier Gout, Colin Graham, Struan F.A. Grant, Franca Rosa Guerini, Hakon Hakonarson, Per Hall, Anders Hamsten, Hans Peter Hartung, Rob N. Heard, Simon Heath, Jeremy Hobart, Muna Hoshi, Carmen Infante-Duarte, Gillian Ingram, Wendy Ingram, Talat Islam, Maja Jagodic, Michael Kabesch, Allan G. Kermode, Trevor J. Kilpatrick, Cecilia Kim, Norman Klopp, Keijo Koivisto, Malin Larsson, Mark Lathrop, Jeannette S. Lechner-Scott, Maurizio A. Leone, Virpi Leppä, Ulrika Liljedahl, Izaura Lima Bomfim, Robin R. Lincoln, Jenny Link, Jianjun Liu, Aslaug R. Lorentzen, Sara Lupoli, Fabio MacCiardi, Thomas MacK, Mark Marriott, Vitto-

rio Martinelli, Deborah Mason, Jacob L. McCauley, Frank Mentch, Inger Lise Mero, Tania Mihalova, Xavier Montalban, John Mottershead, Kjell Morten Myhr, Paola Naldi, William Ollier, Alison Page, Aarno Palotie, Jean Pelletier, Laura Piccio, Trevor Pickersgill, Fredrik Piehl, Susan Pobywajlo, Hong L. Quach, Patricia P. Ramsay, Mauri Reunanen, Richard Reynolds, John D. Rioux, Mariaemma Rodegher, Sabine Roesner, Justin P. Rubio, Ina Maria Rückert, Marco Salvetti, Erika Salvi, Adam Santaniello, Catherine A. Schaefer, Stefan Schreiber, Christian Schulze, Rodney J. Scott, Finn Sellebjerg, Krzysztof W. Selmaj, David Sexton, Ling Shen, Brigid Simms-Acuna, Sheila Skidmore, Patrick M.A. Sleiman, Cathrine Smestad, Per Soelberg Sørensen, Helle Bach Søndergaard, Jim Stankovich, Richard C. Strange, Anna Maija Sulonen, Emilie Sundqvist, Ann Christine Syvänen, Francesca Taddeo, Bruce Taylor, Jenefer M. Blackwell, Pentti Tienari, Elvira Bramon, Ayman Tourbah, Matthew A. Brown, Ewa Tronczynska, Juan P. Casas, Niall Tubridy, Aiden Corvin, Jane Vickery, Janusz Jankowski, Pablo Villoslada, Hugh S. Markus, Kai Wang, Christopher G. Mathew, James Wason, Colin N.A. Palmer, Erich Wichmann, Robert Plomin, Ernest Willoughby, Anna Rautanen, Juliane Winkelmann, Michael Wittig, Richard C. Trembath, Jacqueline Yaouanq, Ananth C. Viswanathan, Haitao Zhang, Nicholas W. Wood, Rebecca Zuvich, Panos Deloukas, Cordelia Langford, Audrey Duncanson, Jorge R. Oksenberg, Margaret A. Pericak-Vance, Jonathan L. Haines, Tomas Olsson, Jan Hillert, Adrian J. Ivinson, Philip L. De Jager, Leena Peltonen, Graeme J. Stewart, David A. Hafler, Stephen L. Hauser, Gil McVean, Peter Donnelly, and Alastair Compston. Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis, 2011.