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

October 20, 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

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

Haplotype-based methods are those which explicitly model Linkage Disequilibrium (LD) between neighbouring SNPs along a haplotype. 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. Interest in LD grew rapidly in the 1980s when its relevance to gene

mapping was realised. Although understanding LD is of critical important to natural selection, gene conversion, mutation and other forces that cause a change in allele-frequency [4], here I will focus on its application to inference of haplotype sharing between individuals.

The application of haplotype-based methods to research into human population structure inference was developed in the early 2000s [5]. Although the study only used a small number of SNPs relative to the number used today, it revealed insights which are still relevant today, such as the presence of long haplotype blocks in highly drifted Native American populations. Under the most basic model, individuals who share recent ancestry are expected to share long regions of alleles which are identical-by-descent (IBD). As recombination breaks haplotypes across generations, it is expected that individuals who share more recent common ancestry will share longer matching haplotypes. This means haplotype-based methods are especially powerful at detecting recent shared ancestry and admixture events within the last 500 years.

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?). The development of the Li and Stephens copying model (LSM) [6] was instrumental in the development of such methods [7] and provided an elegant solution to the increased complexity when using 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, imputation and haplotype estimation.

Perhaps the first paper to formalise haplotype-based approach for the study of population history was that of Hellenthal et al 2008 [8]. Building on the copying model proposed by Hellenthal et al, Lawson et al (2015) [9] created ChromoPainter, a model which explicitly accounts for linkage and variable recombination rates in order to find the genealogically closest haplotype to a

target in a set of references. 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 [10], and infer ancestry proportions [11].

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.

It has since become a mainstay of population genetic research, both in humans and other organisms. It has also been incorporated into many ancient DNA studies, which I will discuss in the next section.

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).

1.2 Methods used to analyse ancient DNA

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 [12–14] and Principle Component Analysis [15], can address a wide-range of questions pertaining, but not limited to, population structure, admixture, genetic similarity and population graphs.

The early studies of ancient DNA generally considered broad-scale questions about human history, such as the nature of human-archaic interactions or the spread of farming technology across Europe [16]. Due to the infancy of the field, sample sizes were small, as efficient methods to obtain genetic data from ancient samples, such as SNP capture arrays, had yet to be developed. Early studies tended to look for genetic differences between populations which diverged many generations ago. 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, with Loschbour being most similar to present-day North-East Europeans, but falling well outside the variation of present-day Europeans, and Stuttgart clustering with present-day individuals from Tuscany.

Such methods, which I will hereafter refer to as 'unlinked', as they assume a model of linkage equilibrium (i.e. each SNP is independent of another in the context of) between neighbouring SNPs, are useful for ancient DNA studies. Data is often represented in 'psuedo-haploid' format to avoid making diploid

[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

genotype calls at low coverage positions; true heterozygous positions may be mis-called as homozgyous at low coverages, as there is a probability of $p()$ that no reads mapping to a particular allele may be sampled. Representing variants psuedo-haploid is used widely in past and current ancient DNA studies. It is also simple to account for missing data in PCA and f-statistics.

However, 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. Include example here.

Another reason for the use of alternate methods to f_3 statistics is that they possibly display bias in the face of drifted populations.

The first use of chromopainter on ancient DNA was in a seminal paper from Lazaridis et al (2014) [16]. The two samples, Loschbour and Stuttgart, we of high coverage and therefore imputation was not used.

The first study to explicitly investigate the reliability of ChromoPainter on ancient DNA was Martiniano et al (2017) [17]. They first tested the accuracy of imputation on ancient DNA samples by downsampling high coverage ancient genomes and comparing imputed to full-coverage non-imputed genotypes. Overall, they found a good correspondence. They also found that filtering the SNPs based on posterior genotype probabilities had little effect.

As there is currently no way to allow missing positions ChromoPainter, imputation is necessary. Recently, Hui et al (2020) evaluated the accuracy of genotype imputation in ancient DNA samples [18].

As sample sizes used in ancient DNA studies has rapidly increased to >100, more analyses have incorporate haplotype-based methods. Each study typically carries out a small analysis to ensure that imputation in low-coverage ancient DNA samples is accurate. Antonio et al (2019) [19] analysed 127

ancient genomes of a mean coverage of 1x. To test imputation accuracy, they downsampled a single individual (NE1) to different levels of coverage and calculated the proportion of genotypes which matched the full coverage. However, this analysis was only performed on a single sample and the effect of imputation on the chromopainter process was not evaluated. 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 [20].

More recently, ChromoPainter has been used to study aspects of archaic hominin ancestry in present-day humans [21, 22].

1.3 Issues with low coverage

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. To mitigate this issue, many studies used 'pseudo-haploid' genotyping, where a diploid genotype is reduced to a haploid by randomly sampling an allele from all the reads which have been aligned to that position in the genome. Whilst the use of pseudo-haploid calling eradicates the need to call heterozygous genotypes at low coverage positions, it also necessarily reduces the amount of information present at each position in the genome.

I will discuss in more detail later the effect of coverage on other methods.

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 part 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 studies 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 [23], phasing and genotype imputation with GLIMPSE [24], ChromoPainter [9] analysis (copy-vector estimation and PCA) and SOURCEFIND ancestry component estimation [11]. Lastly, I will describe some of the existing issues pertaining to low coverage ancient DNA and several considered mitigation strategies.

2.2 Methods

2.2.1 Description of the ChromoPainter algorithm

ChromoPainter is a method designed to infer patterns of haplotype sharing between individuals [9]. The individuals being analysed are split into 'donor' and 'recipient' haplotypes. An individual may or may not be both a donor and recipient, but they cannot act as a donor to themselves. In diploid organisms such as humans and dogs, each individual thus consists of 2 haplotypes. It employs the widely-used Li and Stephens copying model [6] to model each recipient haplotype as a mosaic of haplotypes observed in the donor panel. 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, each with an allele x at position 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:

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

where $z_{dr} = 1$ if d and r both carry allele x, and otherwise $z_{!dr} = 0$, and θ is some pre-specified error likelihood, usually on the order of 0.001. [YOU HAVEN'T DEFINED r_x AND d_x; MAYBE JUST CHANGE TO $\Pr(r = x \mid$

$d = x$), WHICH SEEMS TO BE WHAT YOU MEAN?]

The transition probabilities (i.e. the probabilities of a change in r copying from one donor haplotype 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 because of 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 are 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 number of chunks 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.

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 position in the genome. 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 can 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 $G = 0 + 1$.

We also have a posterior genotype likelihood, in the form $GL(p_a, p_b, p_c)$, where p_a, p_b, and p_c are the posterior genotype probabilities. Dosage, D, is the expected total number of copies of the alternate allele given GL. D can be calculated as $p_b + [2 * p_c]$. We can calculate U, the uncertainty as $U = |G - D|$. Then, we can assign a probability to each allele; if the allele is `1` then the allele likelihood is simply $1 - U$ and if the allele is `0` then the allele likelihood is $0 + U$.

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 position p.

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

, where r_{xp} is the probability that r carries the alternate allele[DO YOU NEED r_{xp}; CAN YOU DO SOMETHING EASIER NOTATION-WISE? PREVIOUSLY YOU DIDN'T HAVE THE POSITION SUBSCRIPT p FOR r; NOT CLEAR WHY YOU USE IT NOW?], and d_{xp} is the probability that the donor carries the alternate allele. Note that above (3) reduces (1) if $d_{xp} = 0/1$ and if $r_{xp} = 0/1$ (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. Downsampling involves taking a high coverage genome and removing 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 the analysis on the full coverage and downsampled

genomes, and compared the results to determine how they are affected by coverage.

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) [25], (2) UstIshim (Siberian Upper Paleolithic hunter-gatherer) [26], (3) sf12 (Scandinavian Hunter-Gatherer) [27], (4) LBK (early European farmer from the Linearbandkeramik culture from Stuttgart, Germany) [16] and (5) Loschbour (an 8,000 year-old hunter-gatherer from Luxembourg) [16]. The samples were chosen due to their high original coverage and diversity of ancestries.

Each .bam file was processed using the atlas (version 1.0, commit f612f28) pipeline [23]
(`https://bitbucket.org/wegmannlab/atlas/wiki/Home`). First, each file was validated using ValidateSamFile command from PicardTools [28].

I downsampled each individual using the `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 atlas `estimatePMD` task. Recalibration parameters were then estimated using the 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 [29]. This was done to reduce the risk of calling false-positive non-polymorphic sites.

2.2.3 Generation of ancient samples

I also generated a set of ancient samples to use as donors in the ChromoPainter analysis (see section x).

This dataset consists of 918 other ancient samples from the literature. These samples were of variable coverage and chosen because they were relevant to studying the cultures discussed in chapters 4 and 5. These 918 consist of all samples from appendices A1, A2, A3, A4, and they were processed according to the stages outlined in appendix B.X, resulting in genotype and genotype likelihood information at the same 77,818,345 genome-wide positions. Each sample was processed in an identical way to the downsampled target individuals.

2.2.4 Imputation and phasing - GLIMPSE

Genotype refinement/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, or may not contain any information at some sites at all. Therefore, it can be helpful to use external information from a reference panel in order to improve the accuracy of genotype calls and reduce the impact of errors on downstream analyses. Given ChromoPainter uses haplotypes rather than genotype data, it is also necessary to phase the genotypes. Phasing refers to the process of determining which alleles were inherited together on the same chromosome. Imputation and phasing must be performed on all full coverage, downsampled individuals.

Three different characteristics are desirable for an imputation algorithm to be useful in this context. Firstly, to allow an input in the form of genotype likelihoods (or phred-Scale genotype likelihoods). 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, to emit posterior genotype-probabilities which, when accurately calibrated, give the probability that a particular genotype call is correct. This is crucial for our previous step 2.2.1.2 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-known that increasing the number of individuals used in imputation/phasing reference panels improves accuracy [24, 30–32].

Two programs, Beagle 4.0 [33] and GLIMPSE [24] 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 [24], 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 dataset (re-sequenced to 30x coverage), containing 3202 individuals from 26 worldwide populations [34]. A description of the processing of this reference dataset can be found in appendix ...

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) [35] to act as the samples for GLIMPSE to phase. Here, 'target' refers to the individuals being imputed/phased and 'reference' refers to the reference panel.

Following the glimpse tutorial (`https://odelaneau.github.io/GLIMPSE/tutorial_b38.html`), I first used `GLIMPSE_chunk` to split up each chromosome into chunks, keeping both `-window-size` and `-buffer-size` to 2,000,000, their

default settings. I used the b37 genetic map supplied by GLIMPSE for the `-map` argument. Across all chromosomes, this produced 936 chunks of an 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.

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 2nd target individual is imputed. This makes it necessary to avoid including the same sample, downsampled to different coverages, in the same set of targets for one imputation run, in order to avoid the confounding effect of allowing an individual to act as the reference to itself. For example, including Loschbour at 0.1x and 10x coverage could mean it imputed itself, a situation which would never occur in reality.

2.2.5 Estimating imputation sensitivity and specificity

I used rtg-tools-3.11 [36] 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:

$$sensitivity = \frac{TP_{baseline}}{TP_{baseline} + FN} \tag{2.3}$$

$$precision = \frac{TP_{call}}{TP_{call} + FP} \tag{2.4}$$

A "variant" is considered to be a SNP with a genotype that is either `0/1` or `1/1`, with $TP_{baseline}$ and TP_{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.

TP, 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 ($TP_{baseline}$) or downsampled ($TP_{baseline}$) 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', cl_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 cl_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 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) 918 ancient samples from the literature together, and did the same for the .vcfs containing the phased haplotypes. I combined the posterior genotype likelihoods with the phased alleles to generate allele likelihoods (described in section 2.2.1.2) 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 ancient samples downloaded from the literature as recipients.

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.

Unless otherwise specified, I compared copyvectors by calculating the r-squared between the full coverage copyvectors.

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.

Principle component analysis was performed on the chunklengths matrix using the IRLBA R library [37].

Including the same sample, downsampled to different levels, would confound the structure on the PCA (maybe explain a bit more what this means). Therefore, it is necessary to perform a slightly altered routine of plotting principle components.

1. Perform principal components on the subset the coancestry matrix containing only the reference ancients

2.2.8 SOURCEFIND

The chunklengths coancestry matrix produced by ChromoPainter contains information about the estimated length of genome a recipient most closely matches a given donor individual or population. However, incomplete lineage sorting, where alleles segregate in a way that is discordant to the 'true' phylogeny reflecting the orders in which populations split from one another, means that there are regions in the genome where a recipient individual most closely matches a reference individual that is not from the population that has split most recently from according to the 'true' phylogeny. This manifests itself as 'noise' in the coancestry matrix, where, for example, an individual from France copies non-zero amounts from African donors, despite not having any recent African ancestry through admixture. Furthermore, unequal donor population sizes may bias the aggregated amount copied to a given population. Accordingly, the values in the lengths coancestry matrix can be hard to interpret. For example, if we

take Mathieson and Scally's (2020) definition of ancestry proportions as "" [38], they should not be interpreted as admixture proportions.

Therefore, in order to account for differences in donor group size and to improve resolution in directly estimating ancestry proportions, it is necessary to run an additional step, SOURCEFIND [11]. Simulations have shown that SOURCEFIND ancestry proportions correspond well to simulated values. 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 even; 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 takes as input i) the chunklengths matrix described in section 2.2.6 and a parameter specification file. This input file allows for the definition of 'surrogate' individuals/populations. SOURCEFIND models each target copyvector as a linear mixture of copyvectors from the surrogate groups, inferring the proportion of ancestry that each surrogate group contributes to the target individual. 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. Ancestry proportions, credible intervals and chain mixing/convergence checking for each surrogate group were estimated using the CODA R library [39].

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 [16, 40]. 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 3 populations are highly genetically differentiated from one another.

For all runs of SOURCEFIND, I used 2,000,000 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 (reference to appendix?[YES, YOU SHOULD SHOW THIS.]). The rest of the parameters were left as default. Ancestry proportions were estimated by taking the mean proportion across all iterations.

2.3 Reducing SNP count

Xxcluding imputed SNPs which have a low probability of being imputed correctly or restricting analysis to non-imputed SNPs above a certain coverage may mitigate coverage-related bias.

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.

Previous studies showed it is possible to distinguish between individuals

from Devon and Cornwall using the fineSTRUCTURE algorithm, but not unlinked methods (ADMIXTURE [41]) [42]. fineSTRUCTURE mostly separated individuals from Devon and Cornwall into different clusters. Therefore, determining whether it is possible to distinguish between individuals from Devon and Cornwall acts as a good test case for reducing SNPs; how many SNPs can we remove before we lose the ability to distinguish between these two populations. Previous studies showed that it is possible to disinguish between individuals from Devon and Cornwall using ChromoPainter in linked mode, but not using unlinked methods (PCA, ADMIXTURE, ChromoPainter unlinked). Therefore, attempting to distinguish individuals from Devon and Cornwall is a useful test of whether there is still a benefit to be gained from using haplotype information.

I used the People of the British Isles (POBI) as a dataset to test this with. The original POBI dataset contains 2039 individuals from 33 populations from across England, Wales and Scotland, genotyped at 452 592 SNPs. Details of the data preparation for this dataset can be found in appendix xx.

I reduced the total number of SNPs down to a set of target levels, retaining a percentage of ra 0.2%, 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 100%. SNPs were randomly subsetted using the `shuf` unix command. 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 all 2039 POBI individuals as donors. I then combined the resulting chunklengths matrices across all chromosomes and combined copyvectors columns into donor groups.

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 procedure.

I took the Human Origins dataset (appendix A.19), containing 560,240 bi-allelic SNPs and selected 70,000 random SNPs to retain, removing all others using `bcftools`. I then submitted the reduced dataset to the Sanger Imputation Service (`https://www.sanger.ac.uk/tool/sanger-imputation-service/`). The Sanger Imputation Service uses Eagle2 [43] 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

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

I estimated the sensitivity (Fig. 2.1) and precision (Fig. 2.2) of genotype imputation using rtg-tools [36]. This compares genotype calls at each position in each downsampled individual after imputation to the same individual at full coverage without imputation.

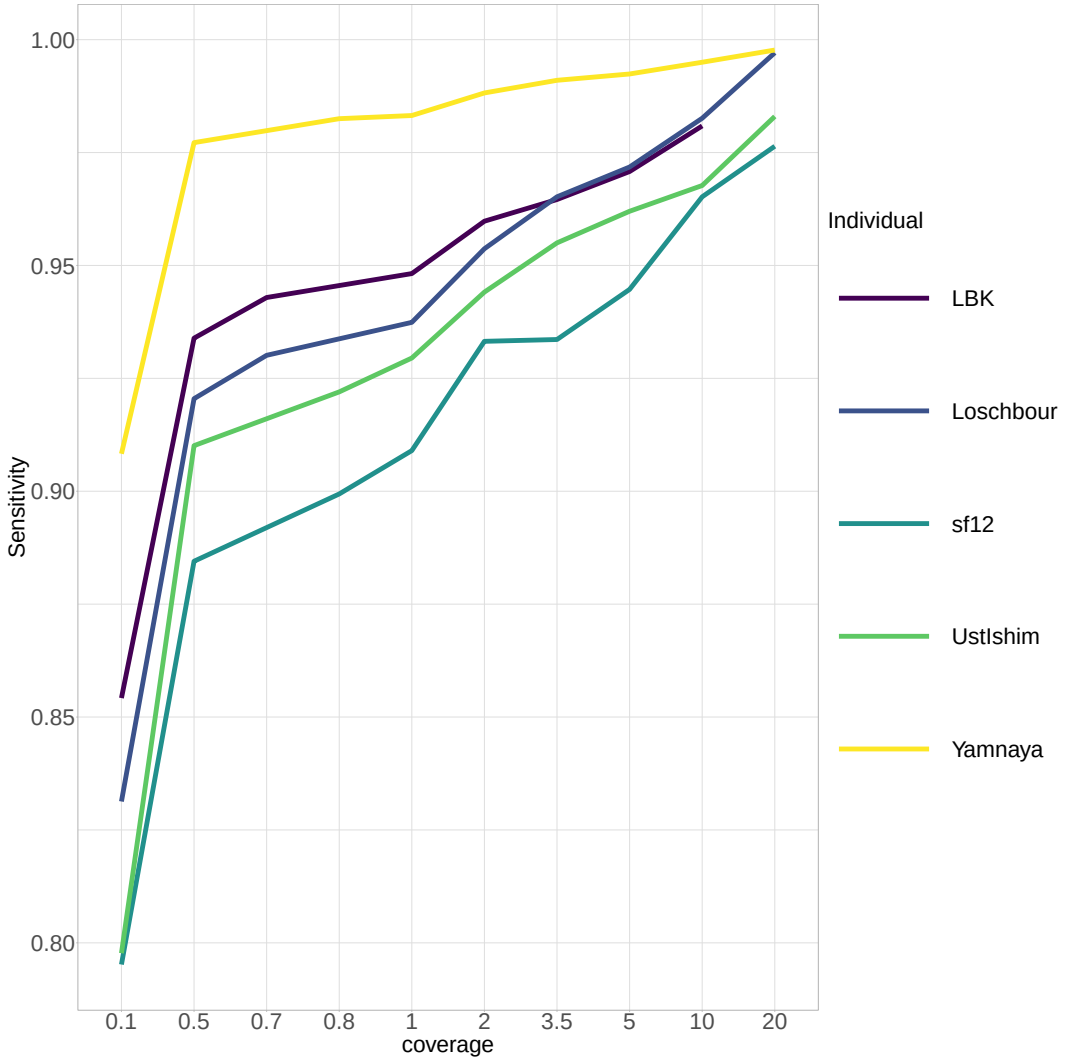


Figure 2.1: Sensitivity of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

As expected, both the overall sensitivity and precision of imputation fell with coverage, with a particularly sharp drop-off in both metrics observed at between 0.5x and 0.1x coverage.

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 contained more individuals who are genetically closer to Yamnaya relative to the other ancient individuals. This is consistent with present-day Europeans

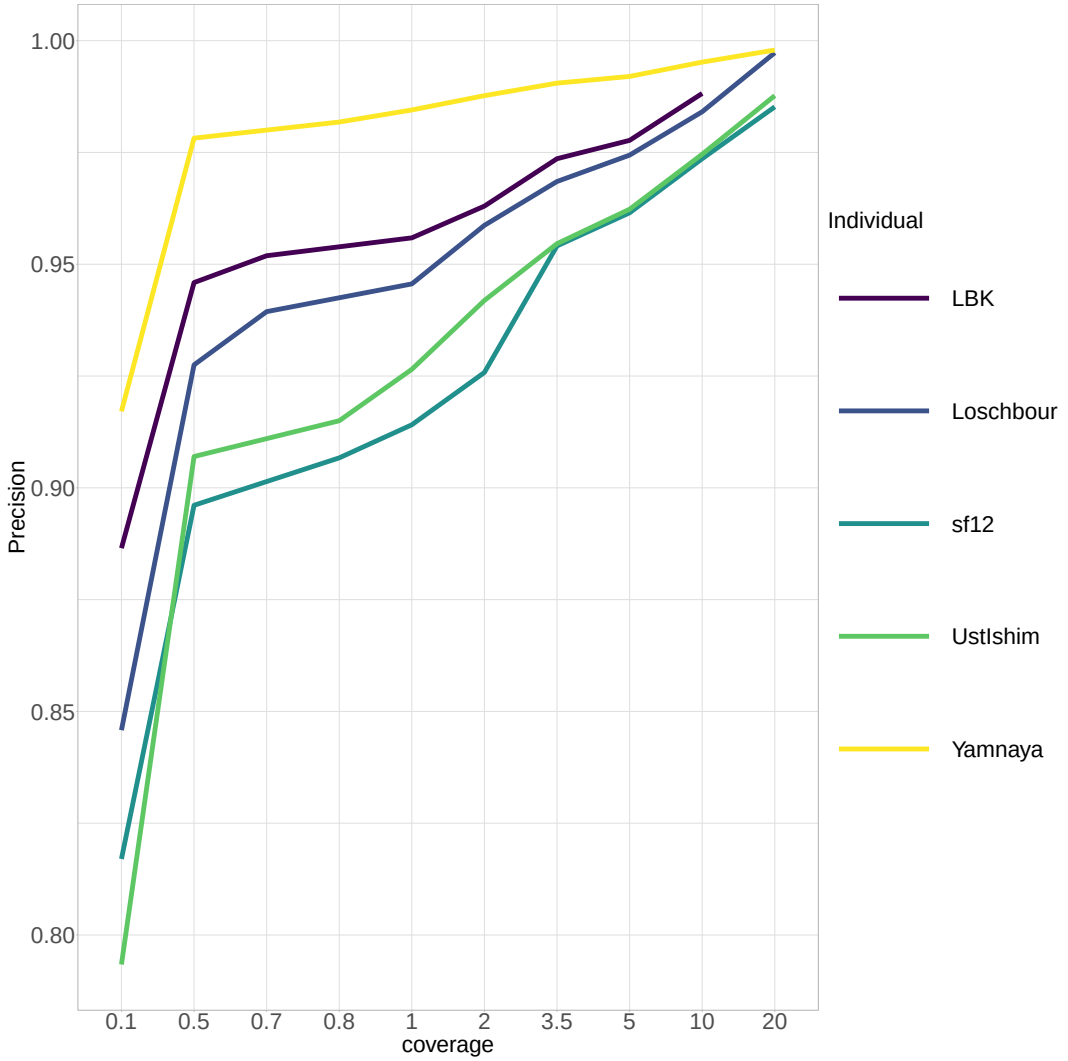


Figure 2.2: Precision of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

containing a high proportion of Yamnaya-like ancestry, relative to e.g. Hunter Gatherer ancestry [44]. 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 [30, 31]. On the other hand, the sample Ust'Ishim is known to have contributed very little genetic ancestry to present-day populations [45] and may therefore have fewer closely matching haplotypes in the reference panel, and a correspondingly lower imputation accuracy.

Imputation accuracy may also be related to demographic history. Populations which are known to have smaller effective population size, such as Western-Hunter Gathers, also contain longer tracts between individuals which are identical by descent (IBD) and fewer heterozygous positions. As imputation relies on matching IBD tracts between individuals, imputation accuracy increases where individuals share more IBD [46]. Additionally, switch-errors during the pre-phasing step of imputation may harm imputation accuracy, so a reduced density of heterozygous positions can improve accuracy.

2.5.2 Phasing accuracy

I 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 [24].

2.5.3 Validating posterior probability calibration

The genotype probabilities emitted by GLIMPSE correspond to 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.

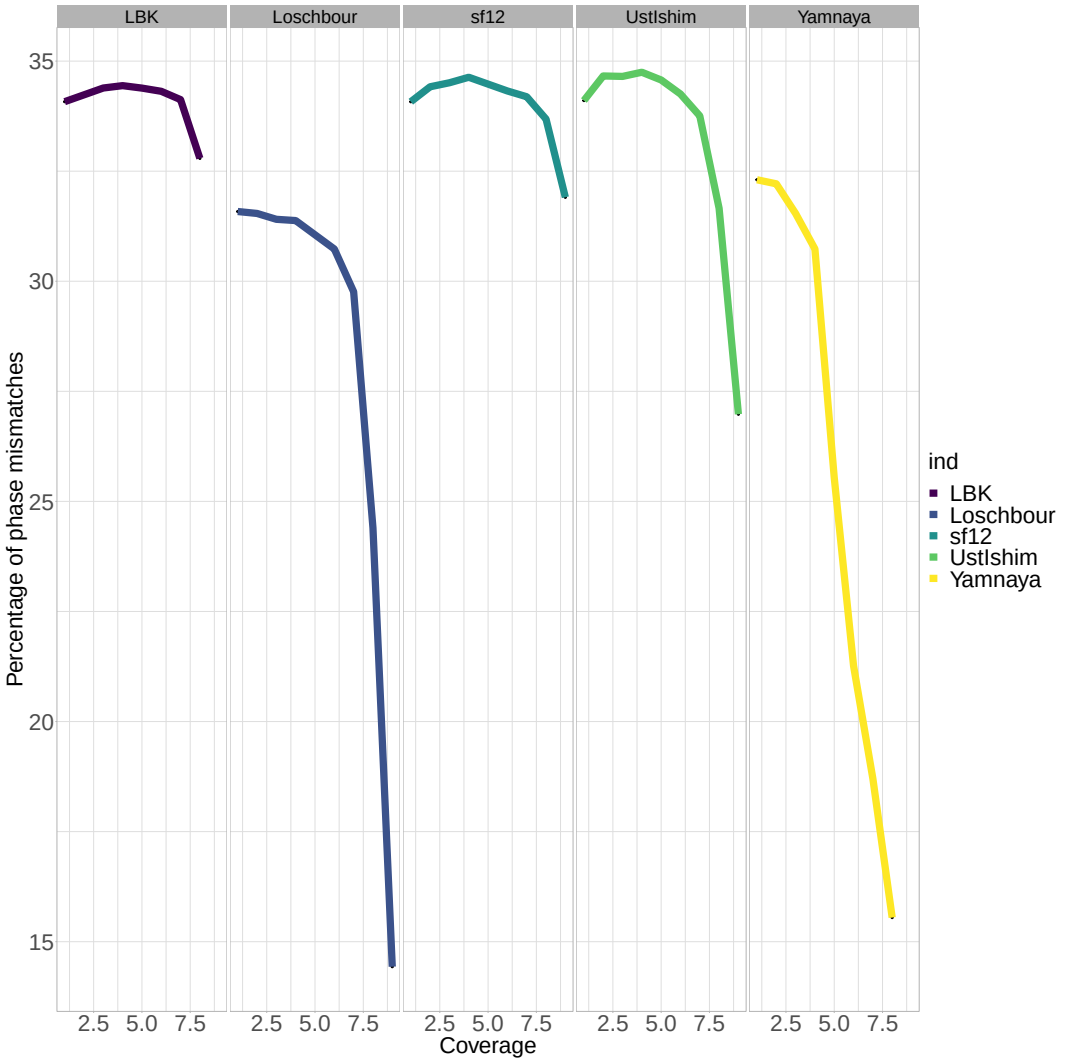


Figure 2.3: Percentage of phased genotypes which agree with the reference for each individual and each level of downsampling. Genotypes with phase deemed unresolvable by rtg-tools were excluded from the calculations. Note that these numbers are given as incorrect / (incorrect + correct - unresolved)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

I split the genome into 10,000 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, as we do.

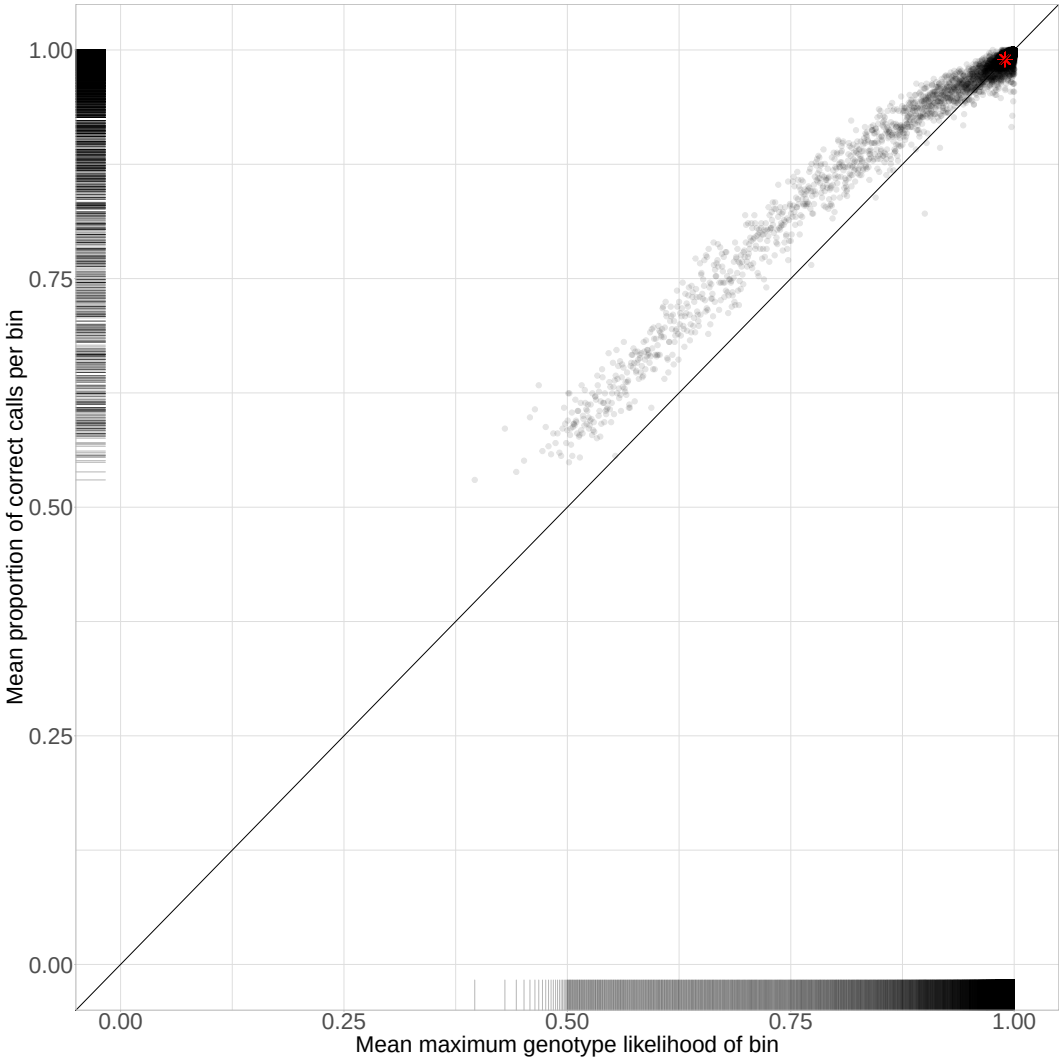


Figure 2.4: Relationship between genotype likelihood and probability of genotype call being correct for Yamnaya downsampled to 0.1x coverage. Genome binned by maximum posterior genotype likelihood and mean maximum posterior genotype likelihood (x-axis) and proportion of correct calls per bin (y-axis).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

The probabilities are slightly conservative, in that a majority of the points in

Fig. 2.4 are above the $y = x$ line. 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

I merged the dataset of downsampled individuals with the 'standard set' of ancient reference individuals (124 ancient samples > 2X coverage). I 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. Do do this, I estimated the 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

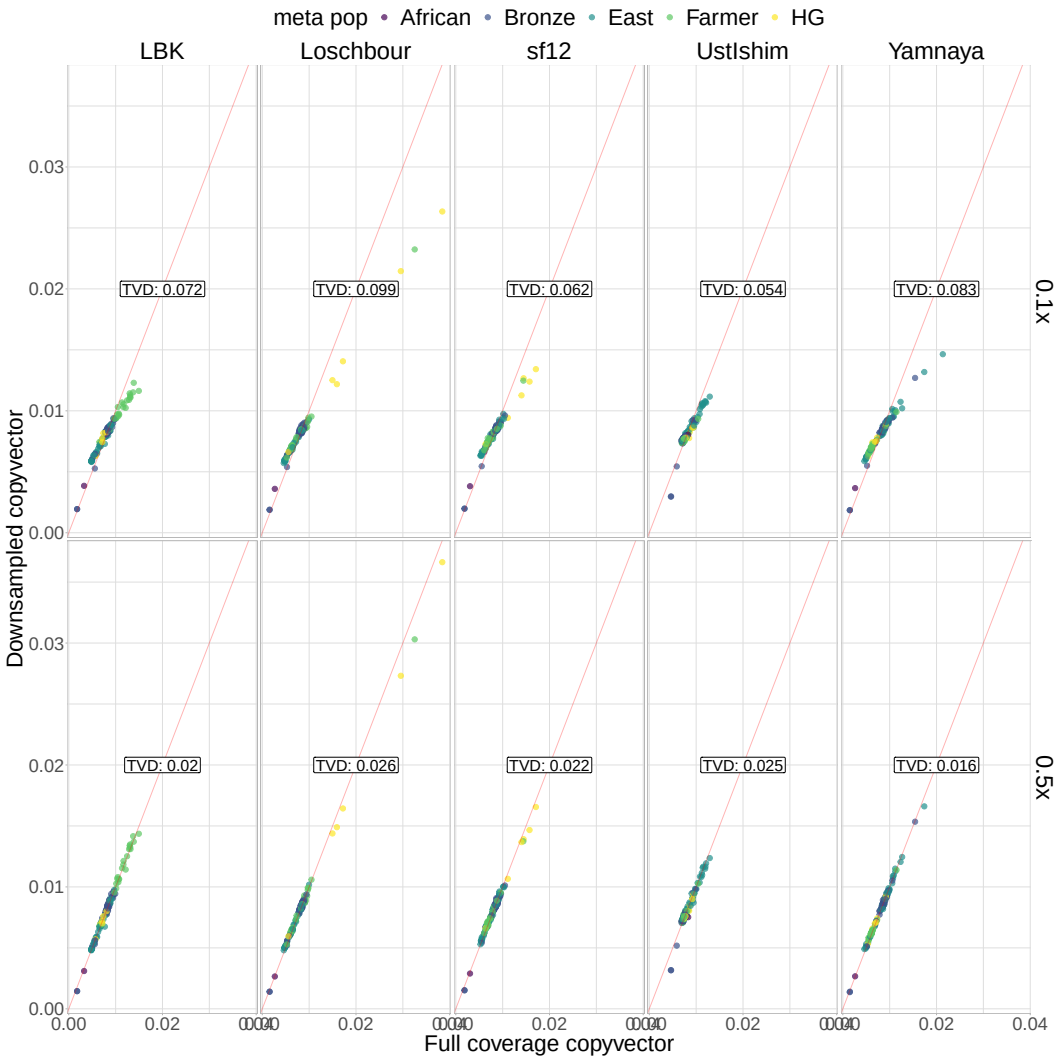


Figure 2.5: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to each of 125 ancient individuals (dots). Results are shown for 0.1x (top row) and 0.5x (bottom row) downsampled genomes. Points coloured by manual assignment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

coverage genome. Put in other words, the copyvectors at 0.1x are tending towards coming 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 also considered the effect of coverage on the copyvectors estimated when using modern individuals as donors (Fig. 2.7). 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 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. As a comparison, I painted each full coverage and down-sampled ancient against a set donor individuals from 26 present-day populations. 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 clumped together along 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

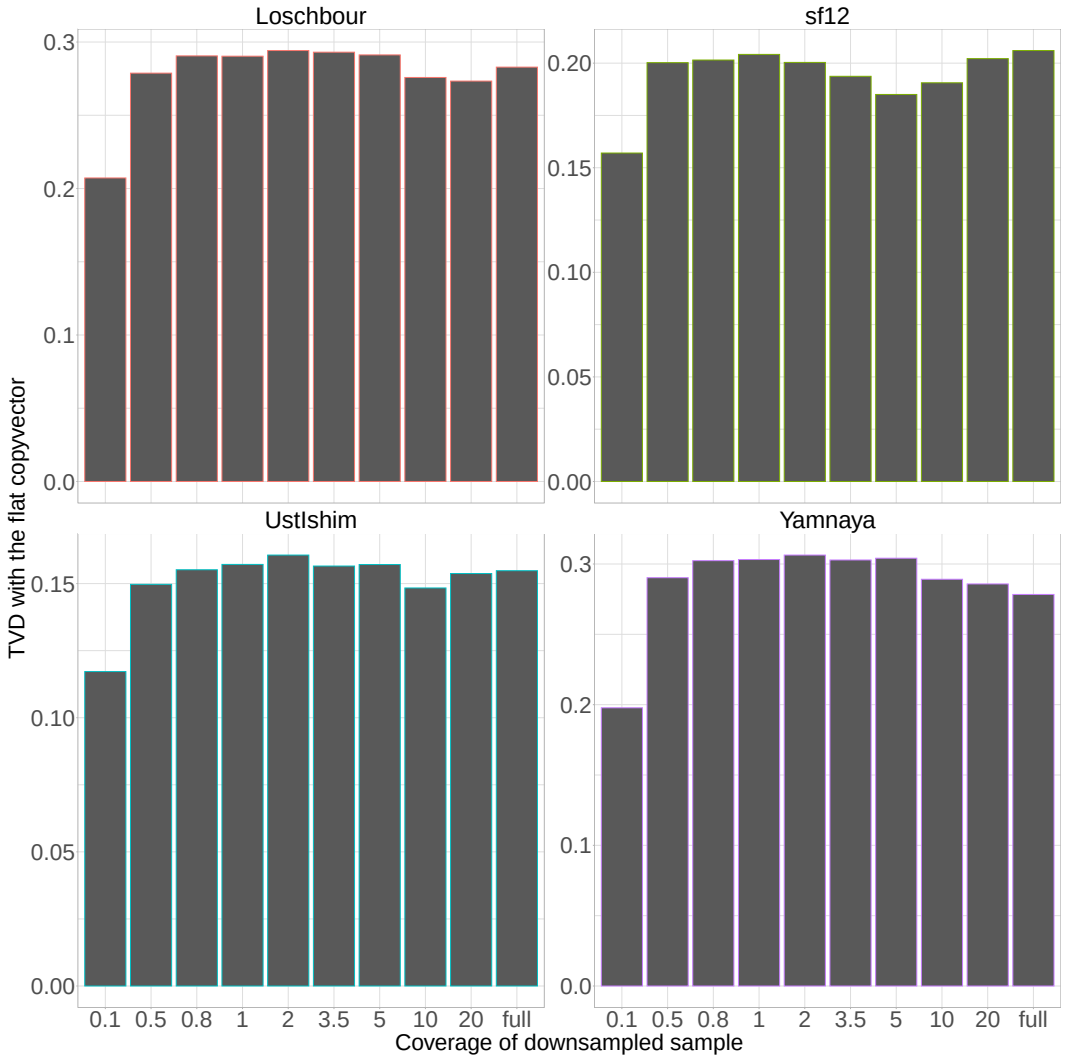


Figure 2.6: TVD (metric of copyvector dissimilarity between two individuals) between each downsampled ancient individual and a flat copyvector. Flat copyvector equivalent to a vector of length N where each element $= 1/N$.

LBK does. This is consistent with findings that UstIshim did not contribute ancestry towards present-day populations [26]. Accordingly, relatively less useful information is obtained from painting Ust'Ishim with a modern reference panel than LBK.

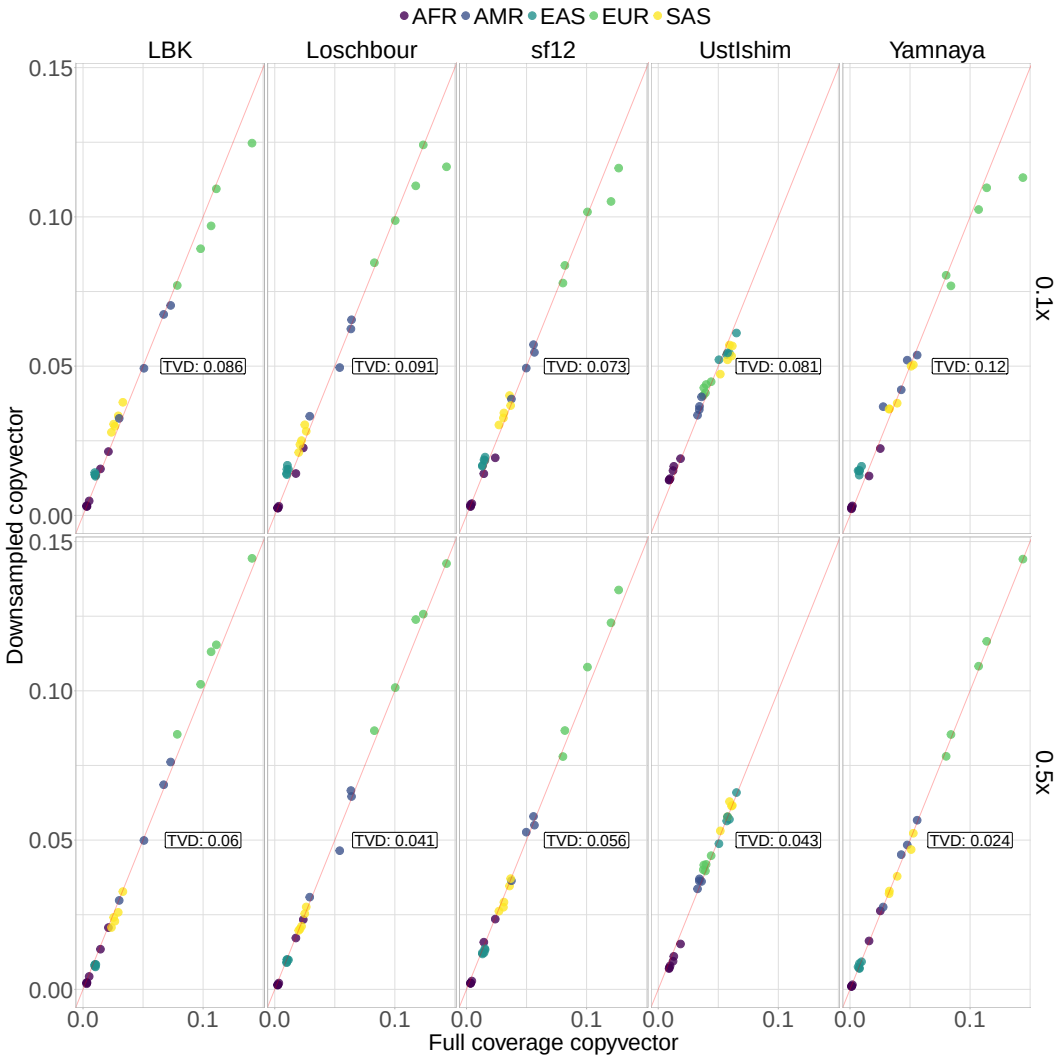


Figure 2.7: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to individuals from each of 26 present-day populations (dots). Red line is $y = x$.

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 a PCA on the genotype dosage matrix, which 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.

To avoid the confounding effect of having two almost identical individuals on a PCA (e.g. the 1x and 0.5x downsamples of the same individual), I performed PCA separately for each ancient individual (see methods section x for full details) (Fig. 2.8).

The position of the full coverage individuals are consistent with prior knowledge about their ancestry. 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.

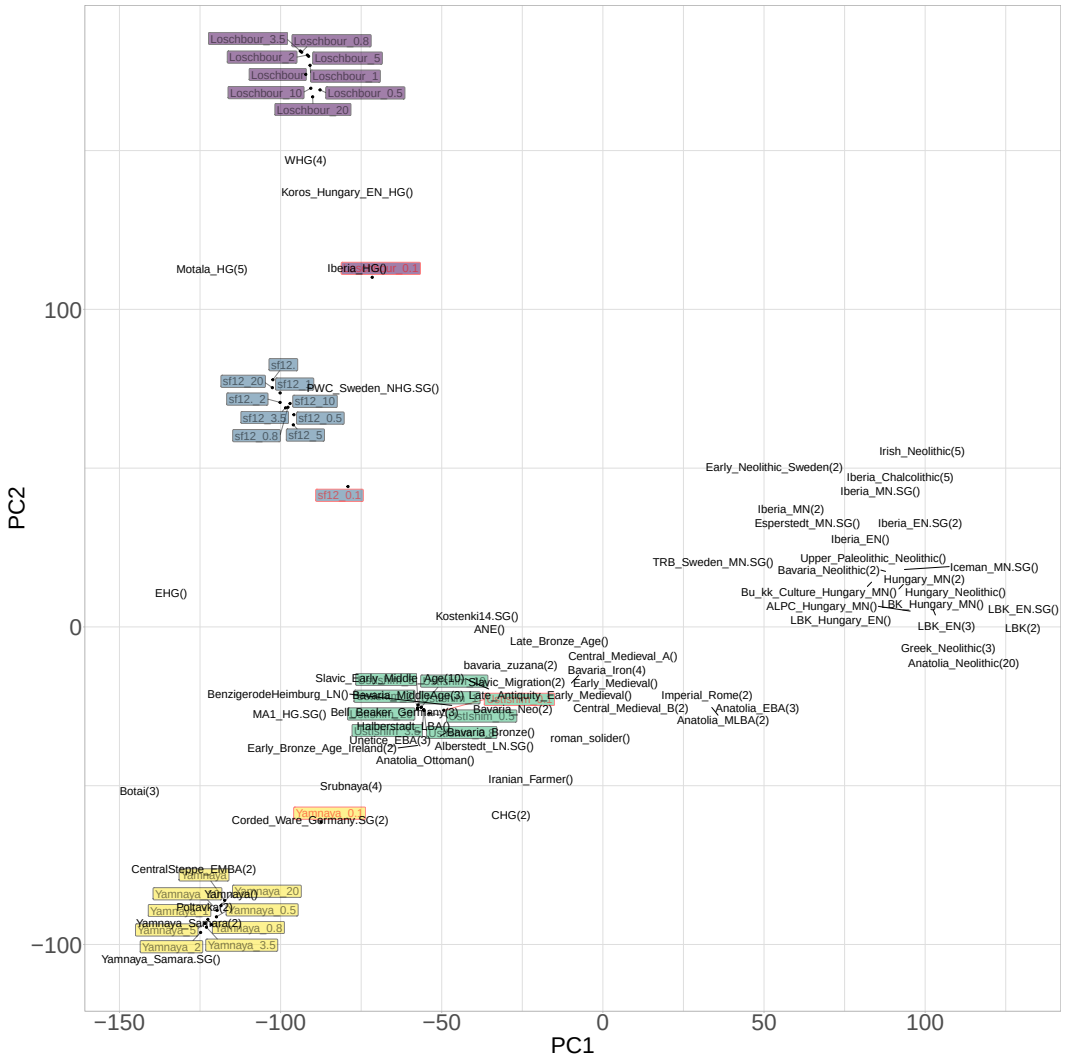


Figure 2.8: Principle component analysis (PCA) of downsampled, full coverage and downloaded ancient individuals generated from the linked chunklengths matrix. Full coverage and downsampled genomes of the same individual are coloured the same. Reference individuals are grouped into populations plotted 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. The chunklengths matrix contains information about the total length of genome one particular individual most closely matches to any other individual. However, this information is often noisy due to phenomena such as incomplete lineage sorting and variable donor group sizes. Therefore, it is often desirable to model out this noise and estimate ancestry proportions in each individual, which are cleaner and more interpretable than raw chunk lengths.

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).

The results suggest that SOURCEFIND estimates are robust down to 0.5-0.8x coverage. At 0.1x coverage, there is an increase in ancestry components that are not present in higher coverage samples, suggesting they are 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. A strength of SOURCEFIND is that many surrogate populations can be used; unlike qpAdm or ADMIXTURE, where a reduced set must be used (reword this and get some evidence)[YES – I THINK IN THEORY qpAdm

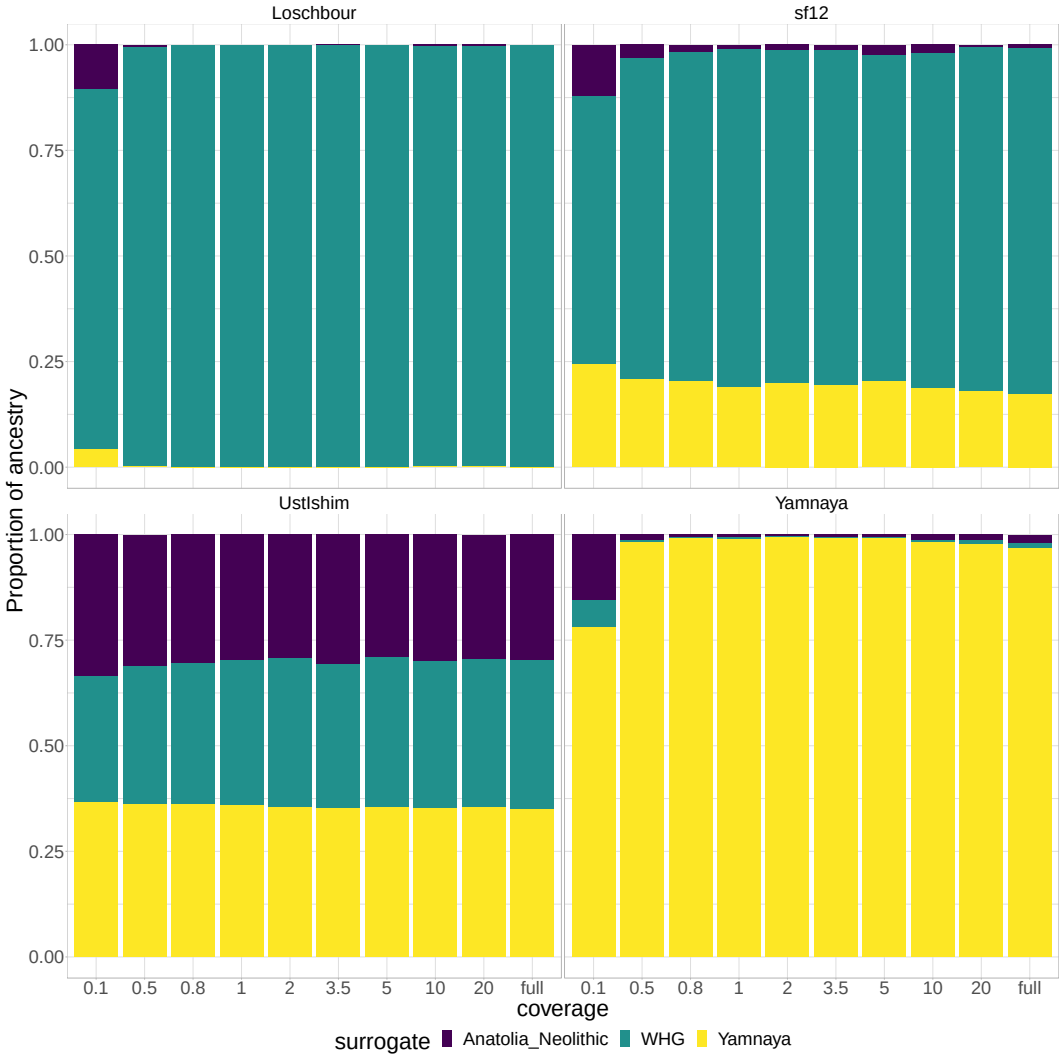


Figure 2.9: Each panel gives inferred recent ancestry sharing proportions for a different downsampled genome. Bars represent proportion of ancestry, coloured by different surrogates. Different coverages for the same individual are given within each panel. Three surrogates were used.

CAN USE MULTIPLE SURROGATES] (Fig. 2.10).

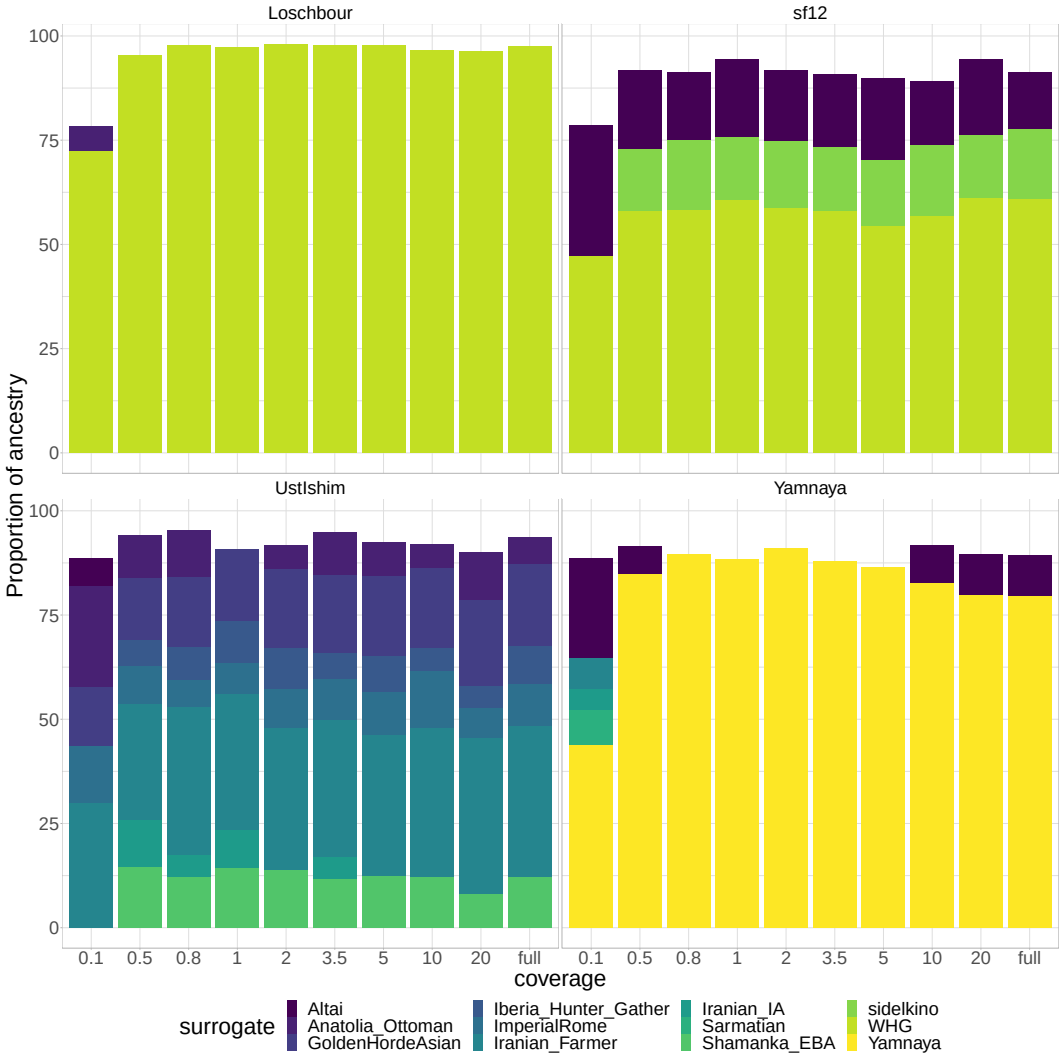


Figure 2.10: Each panel gives information for a different downsampled genome. Bars represent proportion of ancestry, coloured by different surrogates. Different coverages for the same individual are given within each panel. All 39 ancient surrogates were used. Only surrogates with more than 5% are shown. [ONE OPTION MIGHT BE TO COLOR ONLY {Ana_Neo,WHG,Yamnaya} AS IN FIG 2.9, THOUGH SF12 WILL BE ODD. CAN YOU SOMEHOW CATEGORISE THE SURROGATES INTO HOW THEY FALL INTO THOSE THREE FIG 2.9 CATEGORIES?]

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 [16, 47, 48]. It is therefore expected that

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

2.6 Issues and possible solutions for low coverage ancient DNA

The previous section detailed a particular drawback of performing ChromoPainter analysis on low coverage (less than 0.5x) ancient DNA samples. In particular, 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 is evidence that samples of this coverage cannot, at present, be reliably analysed using the 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 bias is introduced. By 'analysis pipeline', I refer to the stages of variant calling, imputation and phasing, and 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. 2.11). PCA projections are used when the target dataset, in this case downsampled ancients, contain variable levels of missing data.

The results show that there is no apparent coverage-related bias in the pre-

GLIMPSE PCA; the 0.1x samples do not substantially differ in their position from the other downsamples of the same individual. However, there is a degree of noise; for example, the LBK downsamples are spread over a small region on the PCA.

On the other hand, the 0.1x samples are clearly shifted to the centre 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. GLIMPSE appears to have removed some of the noise in the downsampled individuals. 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 ChromoPainter output. There is an increased amount of noise and evidence of coverage-related bias relative to the post-GLIMPSE genotype PCA. Fig. 2.11) 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 is reduced noise in the unlinked painting [THAT DOESN'T MAKE MUCH SENSE! WHAT IS TVD BETWEEN FULL-COVERAGE AND DOWN-SAMPLED $>$ 0.5X? ALSO WHAT HAPPENS AT HIGHER PCs?], suggesting discounting haplotype information may be a strategy to reduce coverage-related bias.

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, 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.

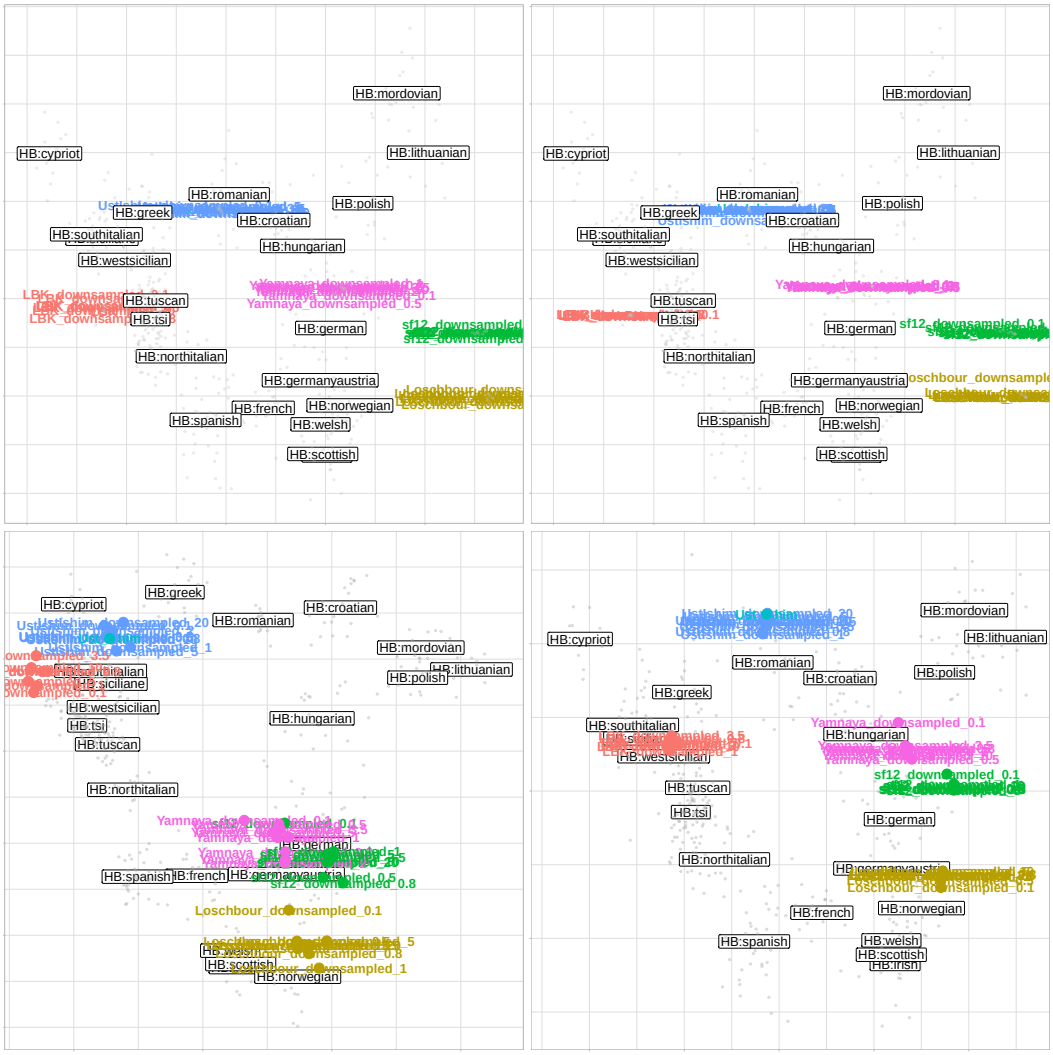


Figure 2.11: Principle Component Analysis. Top Left - pre-GLIMPSE genotypes. Top Right - post-GLIMPSE genotypes. Bottom Left - ChromoPainter Linked. Bottom-Right - ChromoPainter Unlinked. White labels correspond to the midpoint of all samples from that population, grey points correspond to modern individuals.

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. 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 XX SNPs[HOW MANY??] 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, or from populations closest to the target population[ISN'T THIS GOOD BIAS THOUGH?]. 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 [SAM: NOT SURE THIS IS THE BEST WAY TO DESCRIBE IT].

Therefore, we can compare the amount different donor groups donate under the imputed and non-imputed SNP set by plotting the mean amount donated by each population using imputed SNPs and non-imputed SNPs (Fig. 2.12)[THIS IS ACROSS ALL RECIPIENTS? ASSUMING THERE

ARE SIMILAR NUMBER OF SNPS IN BOTH ANALYSES, IT SEEMS UNCLEAR WHETHER THIS IS B/C HAPLOTYPES ARE CAPTURING MORE (CORRECT) INFORMATION WHEN USING IMPUTATION (E.G. B/C IMPUTATION FIXES INCORRECT CALLS). CAN YOU REDO WHILE EXCLUDING ALL EUROPEAN RECIPIENTS? YOU CAN JUST ADD A SENTENCE NOTING RESULTS DON'T CHANGE IF YOU DO SO]. The 20 populations that contribute most are either European or Jewish. Notably, the Haplotype Reference Consortium panel which was used to imputed 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 clear bias towards copying more from European populations when the data has been imputed using the HRC.

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 a generic error probability, the posterior genotype probability from GLIMPSE is accounted for in the emission probabilities of the copying model. The motivation behind this improvement 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

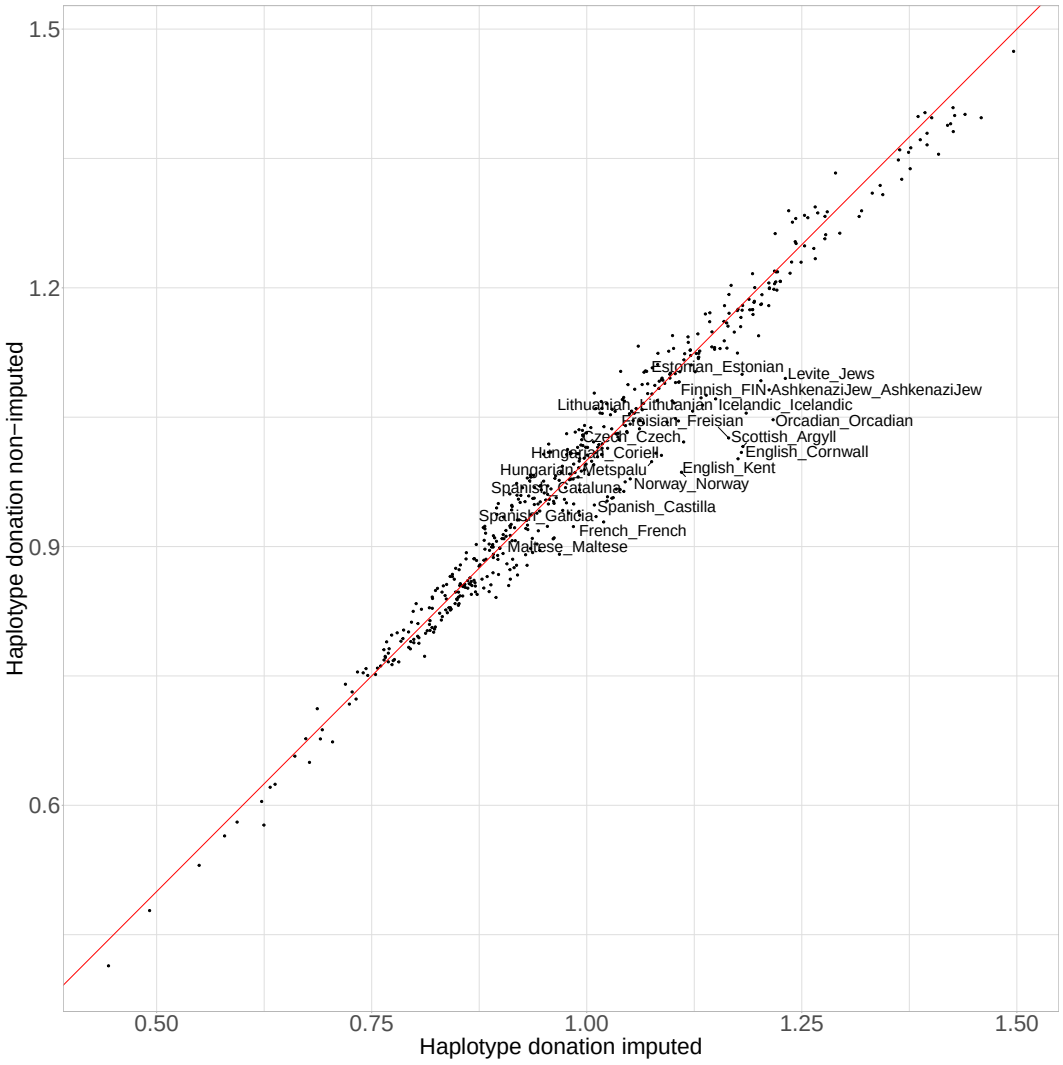


Figure 2.12: Comparison of the mean normalised cM donated by each donor population using the imputed and non-imputed SNP sets. The 20 populations with the largest different between imputed and non-imputed donation are highlighted.

in spirit to that of Viera et al (2016), who account for genotype likelihoods to infer inbreeding IBD tracts from low coverage sequencing data [49].

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. 2.13). 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 allele frequency) used for imputation is known to affect how accurately that variant can be imputed [18, 24, 30, 50]. 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

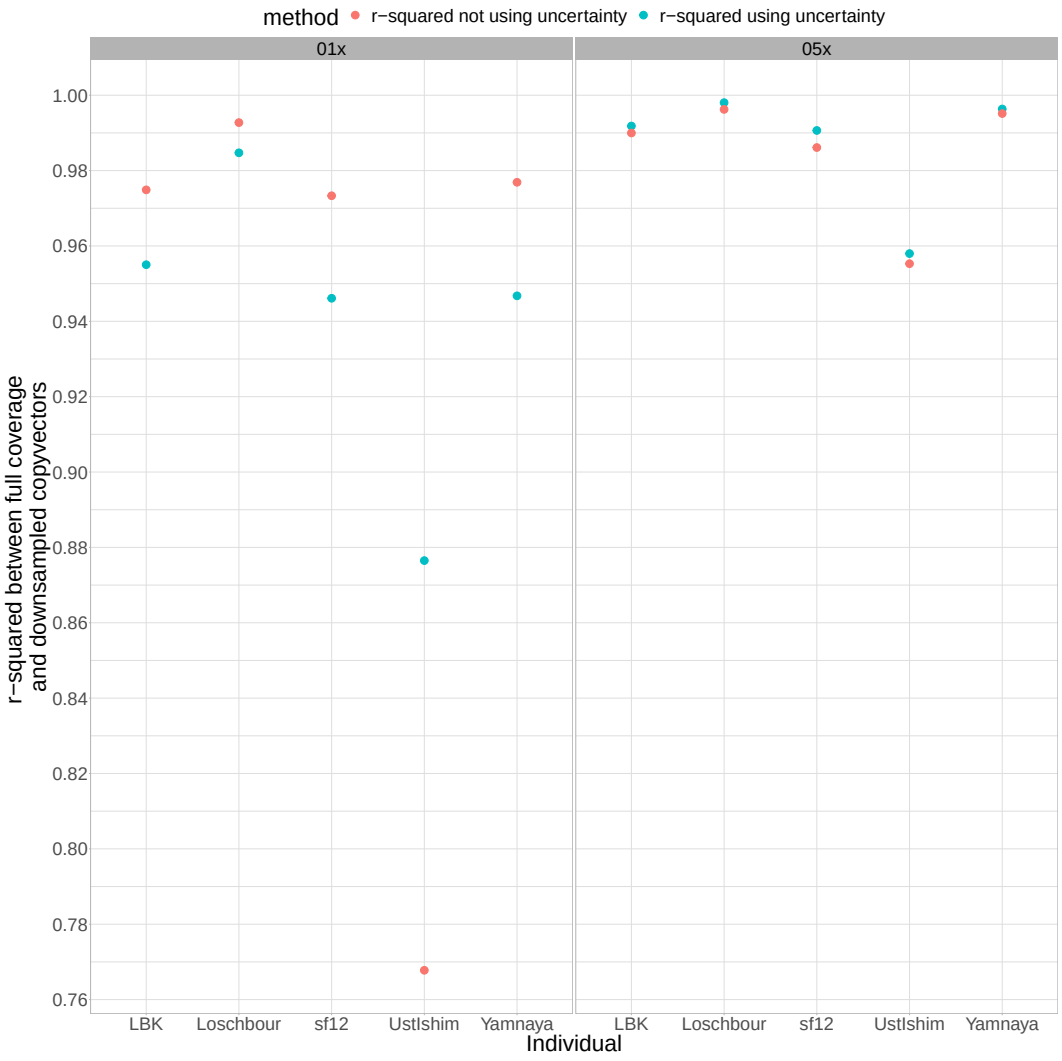


Figure 2.13: Comparison of performance of ChromoPainterV2 and ChromoPainterV2Uncertainty. Panels correspond to samples downsampled to 0.1x (left) and 0.5x (right). Points correspond to the r-squared between the downsampled individual and the same individual at full coverage. Red points are values obtained from ChromoPainterV2 and blue points are those obtained from ChromoPainterV2Uncertainty.

sample	uncertainty_01x	uncertainty_05x	standard_01x	standard_05x	raf_01x
LBK	0.989	0.996	0.989	0.996	0.979
Loschbour	0.998	0.999	0.998	0.999	0.992
sf12	0.989	0.995	0.989	0.995	0.974
Yamnaya	0.990	0.999	0.990	0.999	0.972
UstIshim	0.848	0.992	0.848	0.992	0.930

Table 2.1: Table of r-squared values between the copyvectors of full coverage and downsampled individuals. 'uncertainty' refers to ChromoPainterUncertainty, 'standard' refers to ChromoPainterV2, RAF refers to filtering SNPs with reference allele frequency (RAF) $0.1 > RAF$ or $RAF > 0.9$ and 'GP' refers to filtering $max(GP) >= 0.990$. [SAM: sorry this has gone off the edge - tried to fix it, but it was causing a lot of issues.[CAN USE longtable INSTEAD. OR USE "U" FOR "uncertainty", ETC.]]

SNPs with $0.1 > RAF$ or $RAF > 0.9$[SHOULD REMIND WHAT THE REFERENCE IS HERE, GIVEN YOU ARE NOW SWITCHING AGAIN FROM USING HRC], removing an average of 50,187 SNPs per individual. I then painted individuals downsampled to 0.1x and 0.5x using the standard set of 125 ancient donor individuals.

Comparing the r-squared and visually inspecting the relationship between the copvectors showed that this did not improve the 0.5x copyvectors (Fig. **??**[FIGURE MISSING?].

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.

2.7.3 Upweighting densely genotype regions of high coverage

The previous section[WHICH?] showed that imputation error was the likely cause of coverage related bias. Thus, excluding imputed SNPs which have a low probability of being imputed correctly or restricting analysis to non-imputed SNPs above a certain coverage may mitigate coverage-related bias.

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.

Previous studies showed it is possible to distinguish between individuals from Devon and Cornwall using the fineSTRUCTURE algorithm, but not unlinked methods (ADMIXTURE [41]) [42]. fineSTRUCTURE mostly separated individuals from Devon and Cornwall into different clusters. Therefore, determining whether it is possible to distinguish between individuals from Devon and Cornwall acts as a good test case for reducing SNPs; how many SNPs can we remove before we lose the ability to distinguish between these two populations.

First, to test the effect of reducing the number of SNPS, I painted individuals from Devon and Cornwall with all other POBI populations as donors, using the full set of SNPs (n=452,592) and a reduced set of SNPs, retaining from between 0.2% - 90% of the original number of SNPs (a full list of reduction levels and details of the painting procedure can be found in the methods section)[WHICH SECTION? HAVE YOU WRITTEN THIS, E.G. DESCRIBED LEAVE-ONE-OUT?].

Fig. 2.14 shows the effect of reducing the number of SNPs on the estimated copyvectors of individuals from Devon and Cornwall. At the lowest level of reducing SNPs (0.2% on the figure, corresponding to retaining 0.2% of the original number of SNPs), the black points, representing the mean amount individuals from Devon/Cornwall copy from a particular donor population, exactly align with the red points. The red points correspond to the sample size of that donor population, which is the prior expected amount of copying. Therefore, at this number of SNPs, there is almost no information in the copyvectors, which have regressed to the prior. On the other hand, at the highest level of reducing SNPs, 90%, the black points do not align with the red points.

From visual inspection, 20% is the level of reducing SNPs in both Devon and Cornwall whereby the copyvector appears qualitatively the same as the copyvector estimated from the full coverage of SNPs. In other words, there is roughly the same information with 20% of the original number of SNPs as there is with the full set of SNPs. Therefore, given populations of similar sizes in equivalent ancient individuals, and a starting number of approximately 450,000 SNPS, we could expect to reduce the total number of SNPs to 20% of the original number (approximately 90,000 SNPs) and still be able to distinguish between 2 populations which are as closely related to one another or less as Devon and Cornwall.

A second related question is whether it is better to paint small regions of known high-quality SNPs or a higher number of imputed SNPs. In other words, does the density of SNPs influence the accuracy of the estimated copyvectors? To test the effect of altering the density of SNPS, I compared the copvectors estimated from different SNP densities, but roughly the same total number of SNPs.

At 2% density there are 9,548 SNPs in total across 3000cM, giving a mean density of 1.38 SNPs/cM. On the other hand, at 90% density, there are 6,777

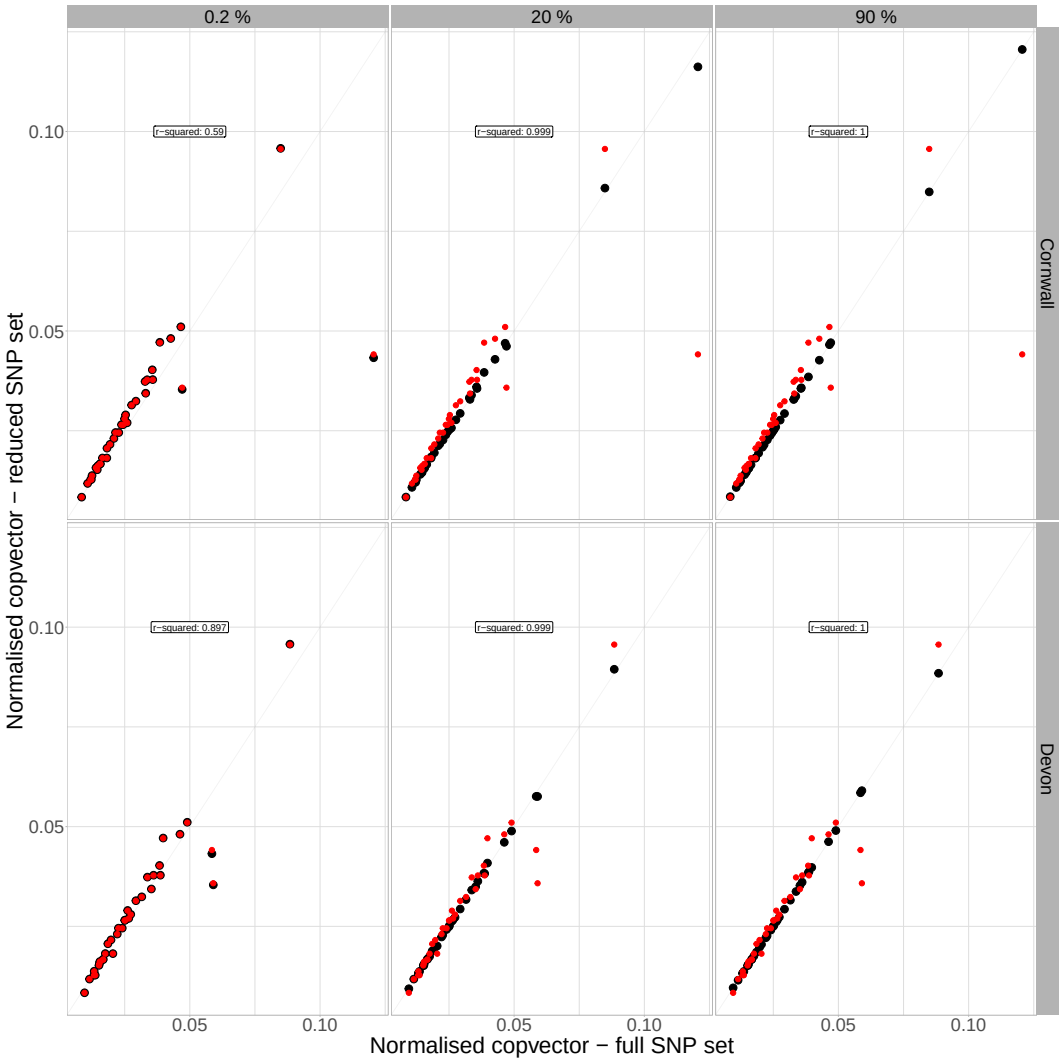


Figure 2.14: Relationship between copyvectors using reduced (y-axis) and full (x-axis) set of SNPs. Panels indicate different levels of reduced SNPs - the percentage corresponds to the percentage of the original number (n=452,592) of SNPs retained. Each black point is the amount that Devon/Cornwall copies from a particular POBI donor group. Each red point corresponds to the sample size (normalised to sum to one) for that[?? DO YOU MEAN "each"?] donor group using either the full set of SNPs (x-axis) and[?? DO YOU MEAN "or"?] the reduced set of SNPs (y-axis).[WHY DOES DONOR POP SAMPLE SIZE DEPEND ON THE NUMBER OF SNPS YOU USE?] Although only 3 levels are shown, the highest (90%), lowest (0.2%) and the lowest [???]ALSO, MAKE R-SQUARED MUCH LARGER.

SNPs across the 144.71cM long chromosome 22, giving a density of 46.83 SNPs/cM. Therefore, the SNPs on chromosome 22 are substantially more dense. Henceforth I will refer to the 0.02 density set of SNPs as the 'sparse' set and the 0.9 chromosome 22 set of SNPs as the 'dense' set.

I performed three different paintings, again, using all the POBI counties as donors and individuals from Devon and Cornwall as recipients. The first painting used all chromosomes and all SNPs at full density, to act as a 'truth set'. The second painting used the 'sparse' SNP set. The third painting used the 'dense' SNP set. For all three paintings, the mean copyvector for Devon and Cornwall was estimated by taking the mean copyvector across all individuals within a population. The motivation here is to see whether the 'dense' or 'sparse' set of SNPs was closer to the 'truth set'.

Fig. 2.15 displays the relative copyvector accuracy when using a dense and sparse set of SNPs. The r-squared value is higher using the denser set of SNPs, suggesting using denser SNPs can more accurately estimate the copy vector than the same number of more sparse SNPs. This also suggests that, if chosen well, a fraction of the original number of SNPs can recover a large amount of the original information.

However, the relationship between the 'dense' copyvector and full coverage copyvector is influenced by the number of individuals in the recipient population. Reducing the number of individuals within the recipient population (i.e. the number of individuals assigned to the population 'POBI:Cornwall' reduces the overall r-squared between the 'dense' and full coverage copyvectors Fig. 2.16. This may be because averaging across more individuals reduces the amount of noise.

With this information in mind, we can consider its use for ancient DNA. The previous section showed that imputed SNPs are the primary cause of coverage-related bias. Therefore, if we can find dense regions of non-imputed

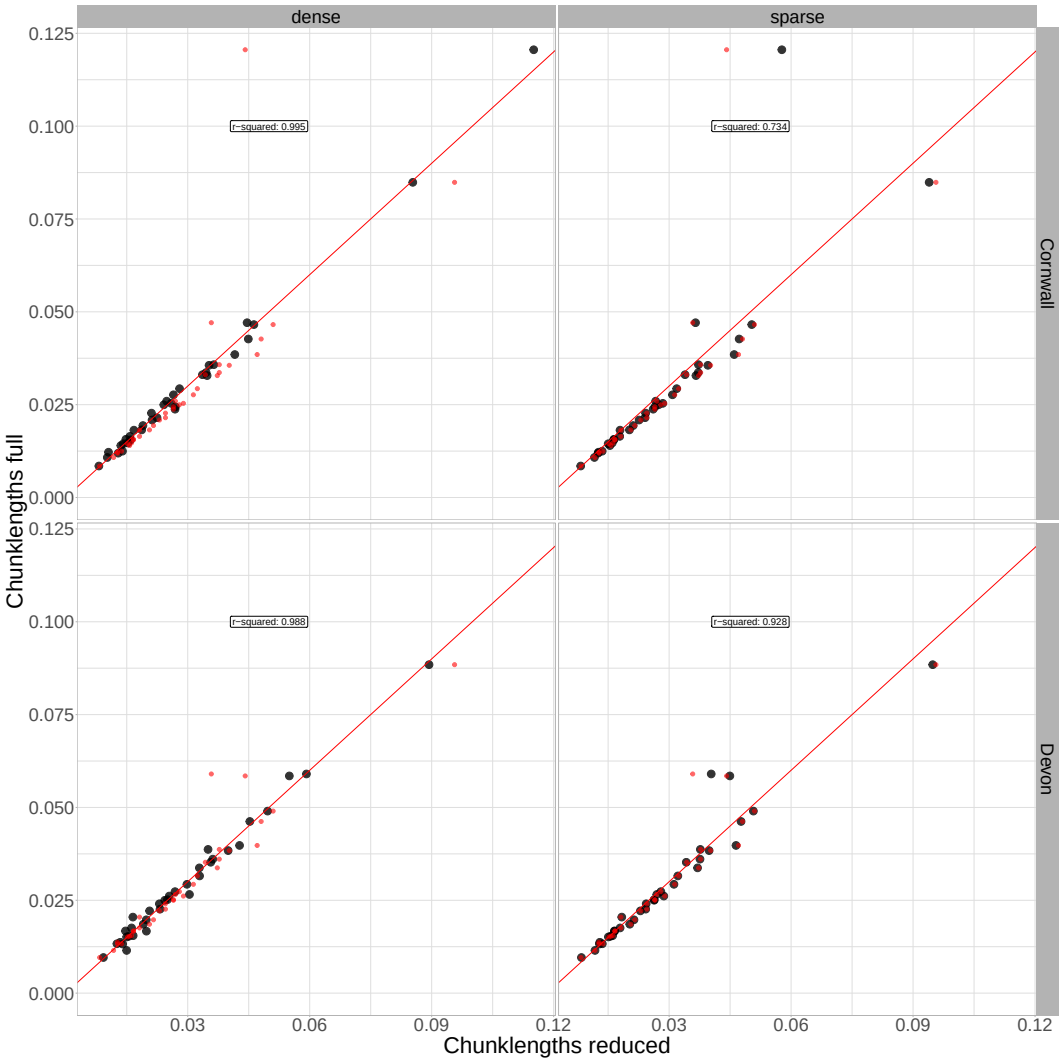


Figure 2.15: Relationship between copyvectors using reduced (y-axis) and full (x-axis) set of SNPs. Panels indicate different levels of SNP density. Copyvectors were estimated by averaging across all individuals within each population (black points). Also shown in red are the sample sizes of each donor population.[MAKE R-SQUARED VALUES LARGER. AGAIN IS IT POSSIBLE TO INCORPORATE UNLINKED FOR COMPARISON – PRESUMABLY 'SPARSE' SHOULD BE THE SAME AS UNLINKED?]

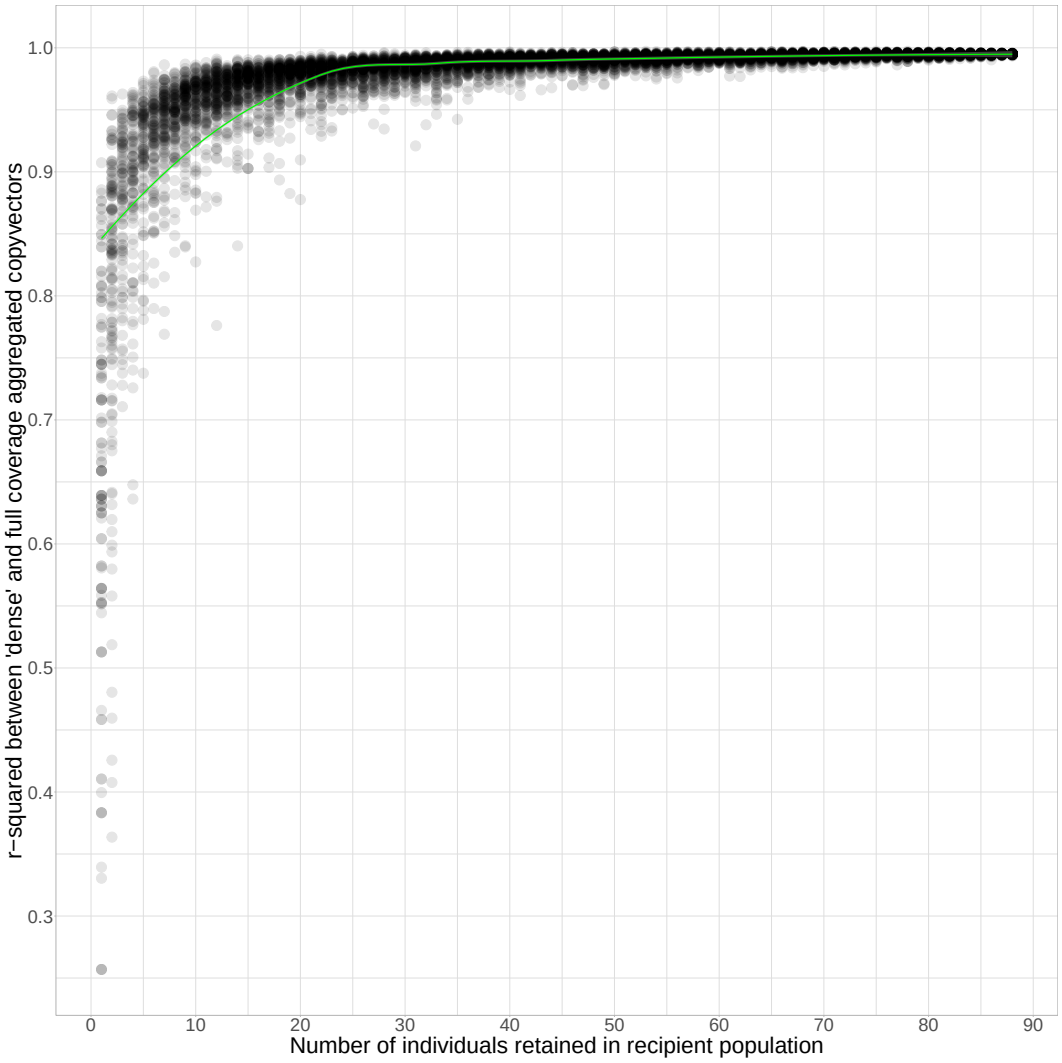


Figure 2.16: R-squared between the copyvectors estimated from 'dense' and full SNP sets (y-axis) using different sample sizes (x-axis). 'Dense' SNP set corresponds to the SNPs located on chromosome 22 at 0.9 density. Copyvectors were estimated by aggregating n randomly selected individuals within the population, corresponding to the x axis-value. Green line is local polynomial regression line. [TO BE CLEAR – FOR EACH X-AXIS VALUE, YOU ARE USING THE SAME INDIVIDUALS WHEN CALCULATING THE 'DENSE' AND FULL-COVERAGE COPY-VECTORS? ALSO SHOULD NOTE THIS IS FOR CORNWALL.]

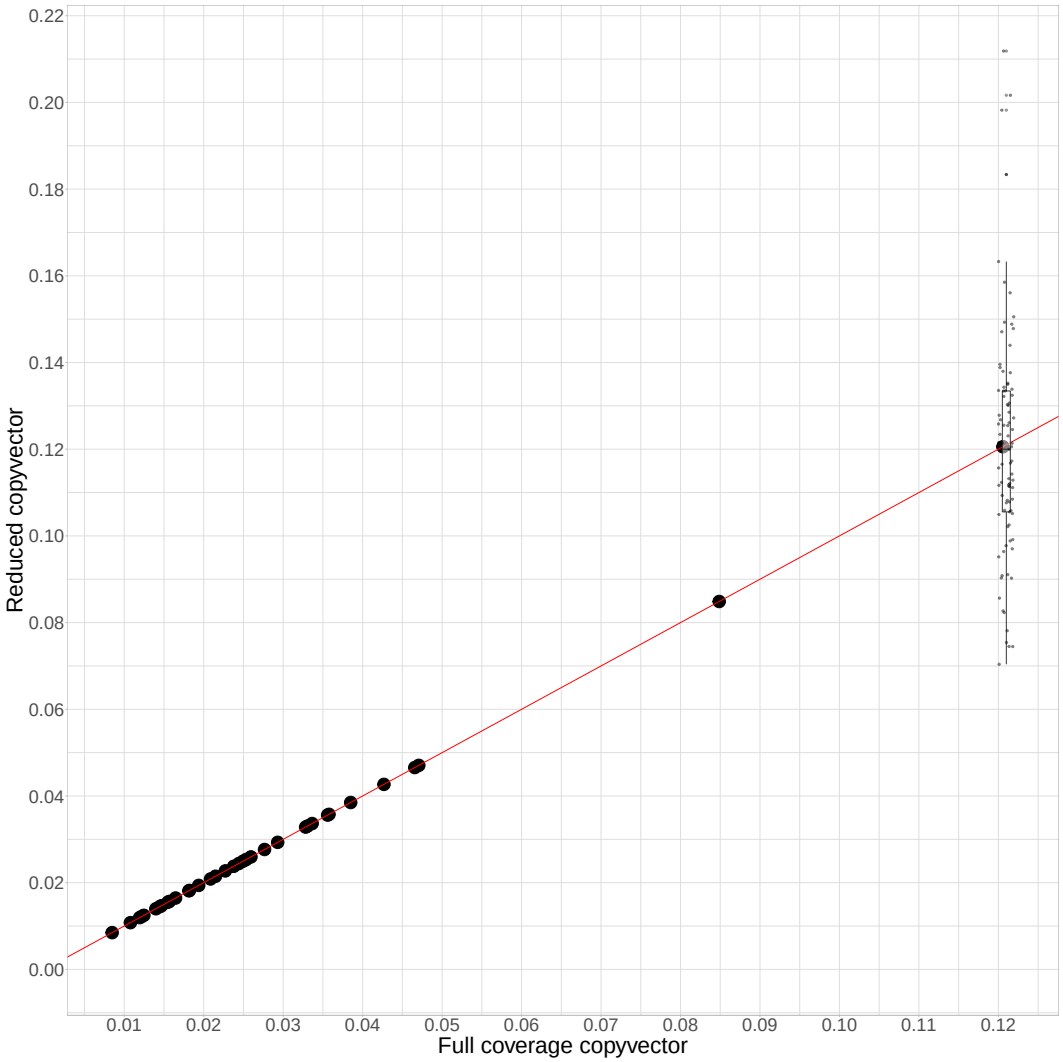


Figure 2.17: Aggregated copyvector for all individuals from Cornwall. Points correspond to the mean amount copied to different different POBI donor groups. Scattered points are the individual amounts each individual within the Cornwall group copies from the Cornwall donor group (i.e. self-copying).[AM I RIGHT THAT THIS MEANS THERE IS EFFECTIVELY no VARIABILITY IN THE AMOUNT OF MATCHING TO CORNWALL AMONG FULL-COVERAGE INDIVIDUALS? THAT IS VERY SURPRISING. OR IS THE X-AXIS JUST JITTERED? IF IT IS TRUE, THEN WE NEED TO TAKE THIS INTO ACCOUNT, BY LOOKING AT WHICH COVERAGE THE VARIABILITY IN MATCHING TO CORNWALL BECOMES VERY LARGE.]

genotypes at sufficient coverage, we may be able to recover a large amount of the haplotype information from low coverage (<0.5x) samples.

The previous section informed us that 2259 SNPs spaced across chromosome 22 was enough to infer structure between individuals from Devon and Cornwall. If we can calculate i) the average size of each chunk (cM) and ii) the average number of SNPs within each chunk, iii) the total number of such chunks, then we can search the genome of a downsampled individual for such chunks.

[DON'T THINK YOU NEED TO SAY HOW YOU CALCULATED IT – JUST SAY WHAT THE AVERAGE NUMBER OF CHUNKS ARE AND AVERAGE SNPS PER CHUNKS AND CHUNK SIZE (MAKING CLEAR WHAT SNP DENSITY YOU ARE TAKING ABOUT.)]First, to find the average number of chunks per individual, we take `(mean(rowSums(counts)) = 158)`. Secondly, to find the mean number of SNPs within each chunk, we calculate `(chr22 nsnps / mean chunks per ind = 14.3)`. Finally, to find the mean chunk size, we must calculate `(chr22 length cm / mean chunks per ind = 0.46cM)`.

Thus, we must search the genome of a downsampled individual for windows of size 0.46cM which contain at least 14.3 'good' SNPs. Ideally, we should find at least 158 of such windows. Table 2.1 displays the results of this search using the 5 genomes downsampled to 0.5x coverage.

min_depth	name	windows
1	a	371722
1	b	363921
1	c	370084
1	d	369058
1	e	377874
2	a	531
2	b	428
2	c	206
2	d	794
2	e	3235
3	a	0
3	b	0
3	c	0
3	d	0
3	e	0

Table 2.2: Number of approximately 0.46cM windows which contain at least 13 SNPs above the coverage specified in min_depth.[SEEMS THIS ISN'T COMPLETE? BUT WHY 0.5x RATHER THAN 0.1x?]

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 different studies sequencing or genotyping individuals from across the U.K (e.g. [42, 51–53]). These studies 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 [42, 54, 55].

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, 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" [56]. The primary reason for this is reasonable concerns over the confounding effect of population substructure within a cohort [57]; retaining a more genetically homogeneous cohort is one strategy to mitigate this.

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 [58–60]. The reasons for this is currently unclear, but it has been suggested that differences in LD structure may be the cause [61]. Ethnic minorities may therefore miss out on the advances in healthcare driven by large-scale genomic projects.

Understanding the population structure of ethnic minorities within the U.K. Biobank is an important step towards including a diversity of ancestries in GWAS. Zaidi and Mathieson (2020) [62] 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 can. Similarly, it has been shown (S.Hu, personal communication of unpublished data) that principle components did not correct for GWAS hits on birth location, whereas using a ChromoPainter coancestry matrix could.

Aside from the medical benefits, there is intrinsic value in studying the ancestry and population history of unerstudied minorities.

Recently, a large dataset, hereafter referred to as the 'Human Origins', has become available. 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. MOST OF WHAT YOU MENTION HERE ARE UNPUBLISHED DATA, SO YOU NEED TO DESCRIBE IT MORE FULLY. YOU SHOULD RE-WRITE THIS PARAGRAPH TO BE CLEAR ABOUT WHAT PUBLICATIONS SOME OF THE DATA CAME FROM AND WHAT THE NEW DATA ARE.] Therefore, this dataset is ideal for use as a reference panel to investigate the ancestry of ethnic minorities within the U.K. Biobank[THIS MAY BE CLEAR FOR AFRICA, BUT WHAT ABOUT E.ASIA AND S.ASIA? YOU SHOULD TALK ABOUT THESE REFERENCES AS WELL.]. In particular, given our newly acquired data come from parts of west Africa that may well represent sources of African ancestry among UK minority groups, I am interested in investigating the ancestry of individuals with recent African ancestry.

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 [30, 63]. 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.

3.2 Methods

3.2.1 U.K. Biobank data access and initial processing

The U.K. Biobank dataset I had access to contains extensive phenotype data for 488,378 individuals and X phenotype 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. I used plink1.9 [64] 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. Full details of imputation can be found in the paper of McCarthy et al (2016) [32]. The imputed data was downloaded and converted from `.bgen` to `.bcf` format using qctool2 (`https://www.well.ox.ac.U.K./~gav/qctool_v2/`).

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

3.2.2 ADMIXTURE analysis

I am primarily interested in using ChromoPainter [9] to explore the ancestry of ethnic minorities in the U.K. Biobank[SHOULD JUSTIFY WHY? I.E. REALLY YOU'RE LOOKING AT WHETHER WE GAIN ANYTHING FROM USING HAPLOTYPE INFORMATION]. 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 [41] 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` [64]. 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.

3.2.3 Data preparation - Human Origins

To determine the ancestry of U.K. Biobank individuals, I compared their SNP patterns to people 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 XX different ethnic groups from across Africa and XX world-wide groups

in total (Fig. 3.1. Full details of processing can be found in Appendix A.20 (Human Origins dataset).

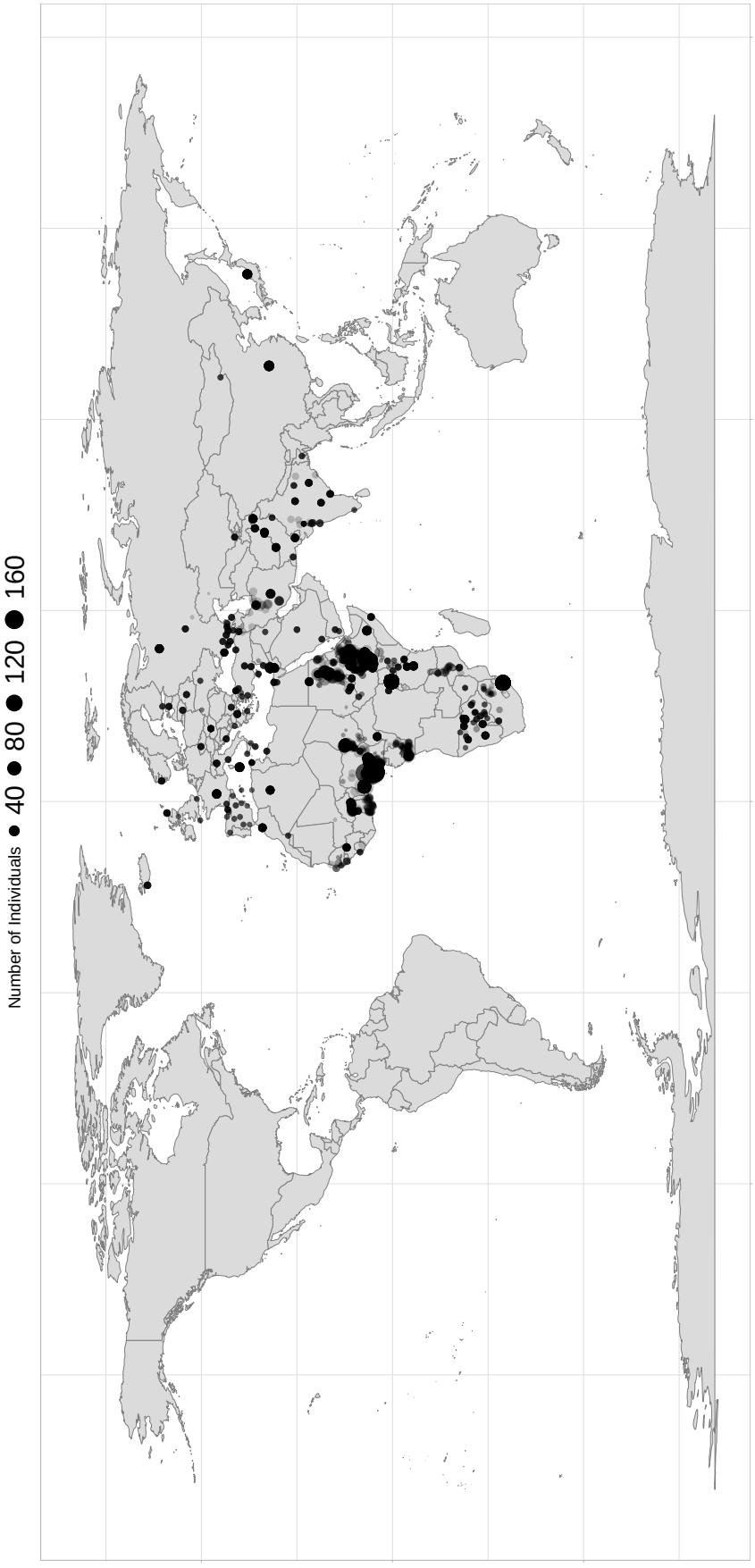
3.2.4 Data merge - non-imputed data and Human Origins

I merged the Human Origins and non-imputed U.K. Biobank datasets. I first identified which SNPs were common to both datasets and then subsetted both datasets so that they contained only those SNPs, resulting in a total of 65,749 SNPs. I identified reference allele mismatches between the U.K. Biobank and Human Origins using the gt-conform utilty from Beagle (`https://faculty.washington.edu/browning/conform-gt.html`) and removed any inconsistent positions.

I then merged the two datasets using `bcftools -merge`. I only retained U.K. Biobank individuals if they had 50% or more African ancestry, resulting in a total of 8476 U.K. Biobank and 5998 Human Origins individuals.

ChromoPainter requires genotypes to be phased prior to analysis. I phased the merged U.K. Biobank / Human Origins dataset using shapeit4 [30]. I set `-pbwt-depth 8`, using the supplied b37 genetic map and leaving all other parameters as default. The resulting `.bcf` of phased haplotypes was converted to ChromoPainter format using a custom script (`https://github.com/sahwa/vcf_to_ChromoPainter`).

[SUGGESTED CONDENSING OF THIS SECTION, WHICH DOESN'T REALLY NEED ITS OWN SECTION (IN GENERAL YOU SHOULD TRY TO SHORTEN WHAT YOU WRITE IN A SIMILAR MANNER):]I used `bcftools -merge` to merge 5,998 reference "Human Origins dataset" individuals with 8,476 UK Biobank participants that had $\geq$50% African ancestry, using the gt-conform utilty from Beagle (`https://faculty.washington.edu/browning/conform-gt.html`) to remove any inconsistent positions. This dataset con-


Number of Individuals
40
80
120
160

tained 65,749 non-imputed SNPs that overlap between the Human Origins and UK Biobank arrays. I phased these data with shapeit4 [30] using `-pbwt-depth 8`, the b37 genetic map and otherwise default parameters.

3.2.5 Data preparation - imputed data

I similarly merged the imputed UK Biobank data with the Human Origins refernce dataset at 525,566 SNPs that were typed in Human Origins, and phased these data with shapeit4 in the same manner.

3.2.6 Chromopainter

ChromoPainter is able to reconstruct the recent ancestry of U.K. Biobank individuals using a large dataset of reference populations. I performed two identical paintings which only differed based on whether they used the imputed or non-imputed dataset.

One alternative option would to have used PBWTpaint , as it is scalable to the size of a Biobank-scale dataset. However, I didn't use it because you can't use a reference panel (I think) and that would have meant a much larger painting.

Both paintings used all Human Origins samples as donors and all Human Origins and selected U.K. Biobank individuals with more than 50% African ancestry as recipients. This allowed me to characterise the ancestry of each of the selected U.K. Biobank individuals in terms of known Human Origins populations. For both the imputed and non-imputed data, each autosome was painted separately and the resulting files merged using chromocombine-0.0.4 (`https://people.maths.bris.ac.U.K./~madjl/finestructure-old/chromocombine.html`).

[CONDENSE THIS SECTION AND NEXT:]For each 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 [11] to infer the proportion of ancestry that each UK Biobank individual shares most recently with each of the XX Human Origin reference populations.[I ASSUME LEAVE-ONE-OUT WAS NOT DONE HERE? YOU SHOULD MENTION THIS, AND JUSTIFY WHY (THOUGH IDEALLY YOU SHOULD DO LEAVE-ONE-OUT TO COMPARE!).]

3.2.7 SOURCEFIND

I estimated ancestry proportions for each of the selected U.K. Biobank individuals using SOURCEFINDv2 [11]. I used the combined painting from the section above. I analysed each U.K. Biobank individual with more than 50% African ancestry seperately, using all Human Origins populations as surrogates. I left all parameters as default.

3.2.8 Imputation bias test

The imputed U.K. Biobank dataset was imputed using a reference panel containing the Haplotype Reference Consortium. Whilst this contains many European populations, it contains relatively few African individuals. Imputing variants in non-European individuals using a reference panel that is primarily composed of European individuals may lead to biased or inaccurate imputation. 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. To explicitly evaluate this possibility, I performed a test of the effect of imputation using the Human Origins dataset as a 'truthset'.[GIVEN THE LEAD UP TO THIS POINT, INTUITIVELY IT SEEMS YOU WOULD COMPARE IMPUTED VS NON-IMPUTED UK BIOBANK RESULT TO ASSESS THIS? I THINK YOU NEED A BIT MORE MOTIVATION – I.E.

THAT ONE ISSUE OF COMPARING IMPUTED VS NON-IMPUTED IS THAT LOSS OF POWER FROM THE MUCH FEWER SNPS IN NON-IMPUTED MAY OVER-RIDE ANY BIAS (OR AT LEAST MAKES IT DIFFICULT TO COMPARE). THUS YOU LOOKED AT HO DATA, WHERE YOU COULD USE THE SAME NUMBER OF SNPS FOR IMPUTED AND NON-IMPUTED.]

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 next 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[OKAY, THIS SUGGESTS YOU DIDN'T MATCH FOR SNPS?? YOU SHOULD MENTION THIS, BUT SAY THAT NUMBERS ARE SIMILAR, AND THAT THERE SHOULD BE ENOUGH INFORMATION IN 560K SNPS (OR INDEED THIS SLIGHTLY FAVORS THE IMPUTED MODEL IN TERMS OF POWER, WHICH SEEMS OKAY)]. [DO YOU NEED TO SAY "ALL-V-ALL"?]I phased the imputed and unimputed datasets separately using shapeitv4 [IS THIS TRUE?]. For each dataset, I then used CHROMOPAINTER to form each phased haploid of the XX Human Origins dataset individuals as a mosaic of all other phased Human Origins dataset haploids. [THEN DELETE THIS SENTENCE]I also performed an identical 'all-v-all' painting of the same Human Origins dataset, but using the original set of SNPs where none had been imputed.

Therefore, I had two co-ancestry matrices[HAVE YOU DEFINED THIS? PROBABLY GOOD TO REMIND THEM WHAT THIS IS EVEN IF YOU HAVE – E.G. "that contain the inferred proportion of DNA that each HO dataset individual shares most recently with each other HO dataset individual]

of identical donors and recipients, but one was generated using imputed SNPs and the other was generated using no imputed SNPs.

3.3 Results

[NOT CLEAR IT'S BEST TO DIVIDE INTO "METHODS" AND "RESULTS", GIVEN YOU PROVIDE MORE METHODS BELOW? IN GENERAL PROBABLY EASIEST (FOR EXAMINERS) TO HAVE SECTION HEADINGS DESCRIBE WHAT YOU'RE AIMING TO SHOW, AND THEN PUT BOTH METHODS AND RESULTS WITHIN THE SECTION.]

3.3.1 4% of U.K. Biobank individuals have at least 50% non-European ancestry

As performing ChromoPainter analysis on the approximately 500,000 individuals would be computationally unfeasible, I performed supervised ADMIXTURE on all 488,378 U.K. Biobank individuals, using $K = 4$ clusters that were defined using European (CEU), Gujarati, Han Chinese and Yoruban reference individuals, in order to identify individuals with at least 50% African ancestry. These individuals 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, it is important to 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 'African' as a broad

label which encompasses a large diversity of ancestries and ethnicities.

I chose to validate the ADMIXTURE results to ensure that there has not been any mixing of individual labels and that enough 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 assessed 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. 3.2), compared to 11 % British, 0.22% Han Chinese and 0.19% Gujarati.

However, there was substantial variation in the ancestry proportions for those self-identified as being either "Caribbean", "African" or "Black or Black British". Proportions of Yoruban and British ancestry ranged from 0.00001 to 1, Han Chinese from 0.00001 to 0.53 and Gujarati from 0.759 to 0.00001, 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.

3.3.2 To impute or not?

[AS I ALLUDE BELOW, IT MAY BE HELPFUL TO START THIS SECTION BY SAYING YOU WANT TO DETERMINE TWO THINGS: (1) IS THERE BIAS IN THE IMPUTED DATASET? (2) DO YOU LOSE POWER IN THE REDUCED DATASET? YOU CAN THEN PROCEED TO SHOW IMPUTED SEEMS TO BE BIASED (FIG 3.3), AND UNIMPUTED APPEARS TO BE MORE POWERFUL (BASED ON TVD AND SOURCEFIND) THAN IMPUTED (PERHAPS B/C OF THE BIAS) AND IS STILL MORE POWERFUL THAN USING UNLINKED.]

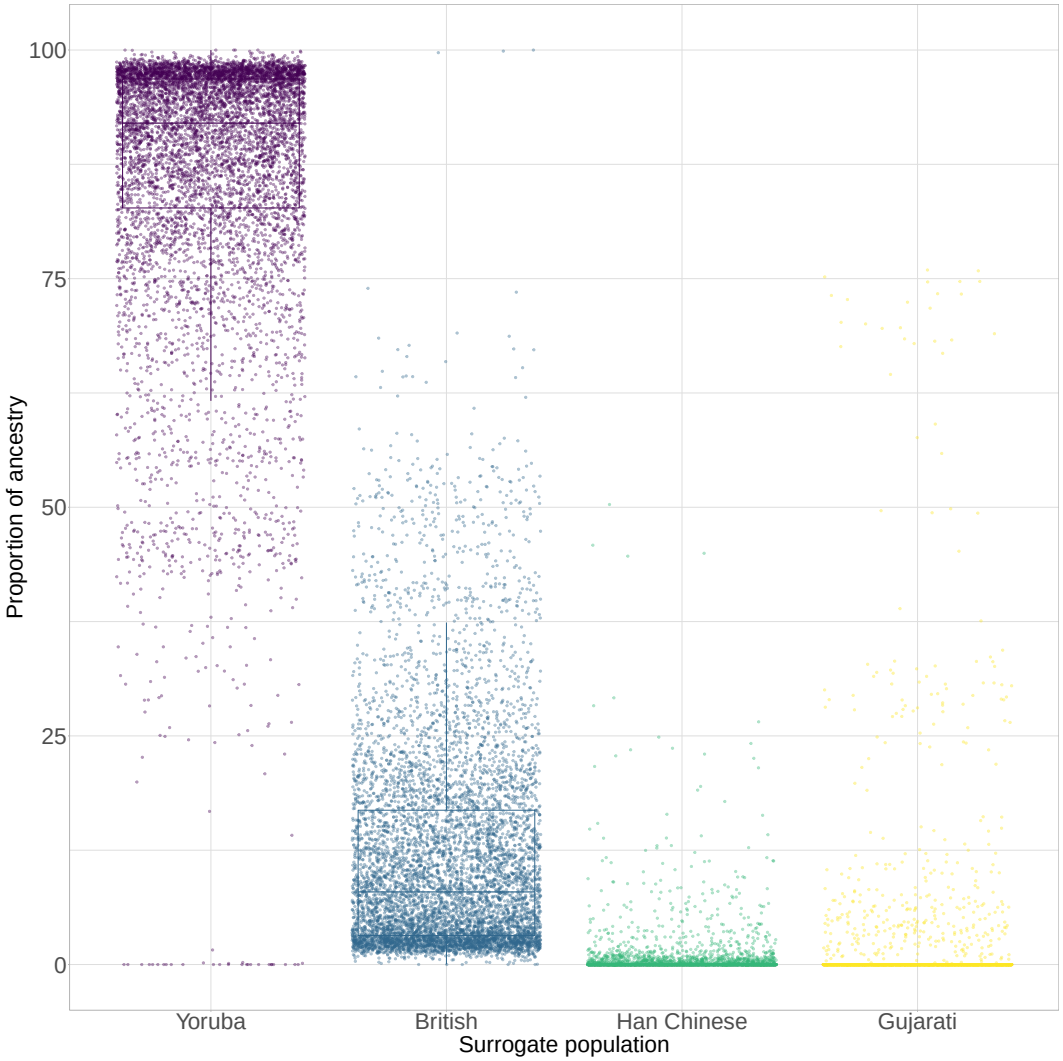


Figure 3.2: Ancestry proportions inferred from supervised Admixture run (k=4) for all individuals who self identified as being either "Caribbean", "African" or "Black or Black British"

In order to use the Human Origins dataset as a reference when inferring 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 ≈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 (REF – maybe https://bmcgenomics.biomedcentral.com/articles/10.1186/1471-2164-10-338 [SEE THEIR FIG 2, THOUGH THAT MAY BE ONLY WITHIN GENE REGIONS]), analysing 70K SNPs may substantially decrease any potential power gains from modelling 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 70K SNPs, including a high percentage of imputed SNPs may bias ancestry inference. Therefore, it is important to determine whether it is preferable to use fewer imputed SNPs or a larger number of imputed SNPs in common scenarios like this, where datasets using different SNP arrays are merged for joint analysis.

To assess whether imputation is better in this scenario, I painted Human Origins reference individuals using (i) the full XX genotyped SNPs, (ii) XX genotyped SNPs overlapping UK Biobank, and (iii) XX SNPs that include the XX genotyped SNPs and XX SNPs imputed using the HRC reference[IS THIS RIGHT?].

In particular I selected all ethinc groups from Nigerian, Cameroon and Ghana which had 5 or more individuals (n=51 pops, n=1203 individuals) and split each population randomly in half, into 'donors' and 'recipients'. painted each population using all other populations, using a 'leave-one-out' approach. I randomly assigned half of the individuals from each ethnic group to be 'donors' and half to be 'recipients' and painted each group of recipients using all other 'donors'. One way to test the information content of a painting is by testing whether individuals copy more from other individuals in their own populations than individuals from other populations[THIS SEEMS LOGICAL, BUT THIS DOESN'T APPEAR TO BE WHAT YOU DO – INSTEAD YOU USE TVD, WHICH DOES NOT NECESSARILY CAPTURE THIS? FOR THIS SECTION, I WONDER IF – INSTEAD OF TVD – IT MIGHT BE BETTER TO USE WHAT YOU DESCRIBE IN THIS SENTENCE?].

For both the imputed and non-imputed datasets, I calculated pairwise TVD between all individuals and then aggregated the values by population. TVD is a distance metric which is the sum of the absolute differences between two

copyvectors[NOT CLEAR WHAT THIS MEANS, UNLESS YOU DEFINE A COPY-VECTOR, I.E. IN TERMS OF WHAT EACH COMPONENT IS AND HOW MANY COMPONENTS (DONOR GROUPS) THERE ARE.]. When using the imputed data, 30% of the recipient populations had the lowest TVD with their own population, whereas the non-imputed data yielded a value of 48%.[NOT ENTIRELY CLEAR WHAT YOU DID HERE – I WAS THINKING YOU MIGHT COMPARE (ii) AND (iii) EACH TO (i) AS I'VE EDITED IT ABOVE, BUT PERHAPS THERE IS NO (i)? THIS MIGHT BE OKAY, BUT WHEN YOU SAY "LOWEST TO THEIR OWN", HOW MANY DID YOU COMPARE TO? AND WHICH POPULATIONS? ARE THEY OTHER ETHNIC GROUPS FROM THE SAME COUNTRY (AND HENCE A CHALLENGING PROBLEM)? WE ALSO NEED TO UNDERSTAND WHETHER YOU ARE GAINING ANYTHING FROM HAPLOTYPES WITH 70K SNPS COMPARED TO USING UNLINKED WITH 70K SNPS (USING THIS SAME TEST). YOU COULD ALSO LOOK AT THE NUMBER OF SNPS/CHUNK, ETC...]

Next, I aimed to mimic later analysis which aimed to assign individuals to ethnic groups using SOURCEFIND. I again selected all ethnic groups from Cameroon, Nigeria and Ghana which had at least 5 individuals[YOU DIDN'T GIVE THIS DETAIL BEFORE?], and split each group into two (randomly), again creating 'donor' and 'recipient' groups within each ethnic group. I then estimated SOURCEFIND ancestry proportions of all groups and identified the surrogate group to which they had the highest ancestry proportion from. When using the non-imputed data, 28/50 populations matched the most[MEANING? >50% OR JUST A MAJORITY OUT OF ALL GROUPS INCLUDED?] to the their own population, whereas only 16/50 matched when using imputed data. Both of these results suggest that using imputed data loses information relative to using only non-imputed SNPs.[AGAIN SHOULD COMPARE TO UNLINKED BY USING NNLS.]

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 relies on identifying reference haplotypes which are closest to the target haplotypes. However, if neither of the ethnic groups that the individuals derive ancestry from are present in the imputation reference panel, in theory, [ISN'T THIS WHAT YOU SAID IN THE PREVIOUS SENTENCE?]the missing variants should be imputed from populations in the reference panel which are most closely related to the target samples. [I'M NOT SURE HOW BEST TO WRITE THIS, BUT TWO CONCERNS ARE: (1) THE REFERENCE DOES NOT CONTAIN INDS FROM (OR DIFFERENTIALLY RELATED TO) BOTH TARGET GROUPS. IN PARTICULAR BOTH TARGET GROUPS, THOUGH DIFFERENT FROM EACH OTHER, MOST CLOSELY MATCH THE SAME POP INCLUDED IN THE REFERENCE. THEY THUS ARE BOTH IMPUTED TO LOOK LIKE THIS REFERENCE, REDUCING THE DIFFERENTIATION BETWEEN THEM. (2) LACK OF SNP DENSITY CAN EXACERBATE (1), AND – EVEN IF YOU HAD REFERENCE INDS THAT CAN DISTINGUISH BETWEEN THE TWO – YOU LOSE THE NECESSARY HAPLOTYPE INFORMATION TO DO THIS AND INSTEAD IMPUTE THEM (MAYBE SOMEWHAT RANDOMLY) TO LOOK THE SAME. I THINK YOU WANT TO GET THESE TWO POINTS ACROSS HERE.] In the case of the Haplotype Reference Consortium, the closest reference population to two African target samples may be the Yoruba from Nigeria, which is the only west African group in the reference[IS THIS TRUE? WANT TO EXPLAIN WHY YOU THINK YORUBA IS BEST]. [WOULD CHANGE THESE LAST TWO SENTENCES TO: If these issues are prevalent, then relative to unimputed individuals we expect imputed individuals to match more to donor populations related to those in the reference panel.]If missing are preferentially imputed from Yoruban individuals, then we would expect segments of the target samples to appear to me more 'Yoruba-like' than otherwise. Correspondingly, when painted using a

reference panel which includes, but is not exclusive to, those populations found in the imputation reference panel, we would expect the populations found in the imputation reference panel to donate more than if no imputation had taken place.

[I'M CONFUSED BY THIS – I THINK I ASSUMED AT OUR LAST MEETING THAT YOU WERE COMPARING WHAT I CALL (iii) ABOVE TO (i) HERE? BUT YOU'RE COMPARING (iii) TO (ii)? IF SO, HOW DO YOU KNOW THAT THE IMPUTED ISN'T BEING MORE ACCURATE B/C IT HAS MORE SNPS? I.E. MAYBE IT SHOULD BE MATCHING MORE TO THE REFERENCE POPS (THEY MAY BE GOOD SURROGATES)? TO CLARIFY – IS SOMETHING AKIN TO COMPARING (iii) TO (i) WHAT YOU DID IN (WHAT IS IN THIS VERSION) FIG 2.12? I THINK YOU WANT TO STICK WITH COMPARISONS LIKE THAT WHEN TESTING FOR BIAS (I.E. SNP DENSITY MUST BE MATCHED WHEN COMPARING IMPUTED VS NON-IMPUTED). IN PARTICULAR HERE YOU SHOULD REDO FIG 3.3 WHEN COMPARING (iii) TO (i); OTHERWISE YOU DON'T REALLY KNOW IF IT'S BIAS. TO GET AT LOSS OF POWER, YOUR PREVIOUS PARAGRAPHS SEEM TO DO THIS (IN PARTICULAR IT SUGGESTS THE BIAS MAY CAUSE A LOSS OF POWER IN (iii) RELATIVE TO (ii)), BUT I DON'T QUITE UNDERSTAND IT YET. AND YOU SHOULD COMPARE (ii) TO UNLINKED IN THE PREVIOUS PARAGRAPH, TO MAKE SURE WE ARE GAINING ANYTHING WITH HAPLOTYPES AT THAT SNP DENSITY.]Comparing the imputed and non-imputed coancestry matrices revealed biases consistent with the above expectation. If the coancestry matrix columns are combined into populations, then the sum of each column gives the total length of genome that population contributes to all recipient individuals in the dataset. Therefore, comparing the column sums between the imputed and non-imputed matrices informs us about which populations contribute more when using imputed compared to non-imputed SNPs. Fig 3.3 shows the amount of differential haplotype

donation on a per-population basis, with populations highlighted based on their presence or absence in the 1000 genomes dataset. It is clear that populations present in the 1000 genomes are primarily clustered towards the right hand side, rather than randomly distributed across figure [sam: is there a statistical test for this?][GIVEN OBSERVATIONS AREN'T INDEPENDENT (AS THEY HAVE TO SUM TO 1), I THINK RE-SAMPLING MIGHT BE BEST. I.E. SUM UP THE Y-AXIS VALUES IN FIG 3.3 FOR ALL 1KGP POPS; CALL THIS Y. IF YOU RE-SAMPLE X TIMES FROM THIS DISTRIBUTION IN FIG 3.3, WHERE X IS THE NUMBER OF 1KGP POPS, HOW OFTEN DO YOU THESE GIVE YOU A Y THAT IS GREATER THAN THAN THAT YOU OBSERVED. YOU CAN THEN REDO THIS WHEN COMPARING (i) TO (ii) TO HOPEFULLY SHOW THERE IS NO BIAS (THOUGH PERHAPS THERE WILL BE SOME BIAS IN (i)).]. This strongly suggests that imputation causes a bias towards those populations present in a reference panel.

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

3.3.3 The distribution of sub-continental African ancestry in the U.K. Biobank

Once I had decided to use the non-imputed dataset, I performed a painting using all Human Origins individuals as donors and all U.K. Biobank individuals with at least 50% African ancestry as recipients.[DIDN'T YOU PAINT BOTH IMPUTED AND UNIMPUTED? IF SO, IT'S WORTH COMPARING THEM.]

Principal component analysis on this matrix reveals the general structure of the selected individuals, alongside the reference populations (Fig. 3.4). Three clines can be observed; one of similarity to Southern African populations

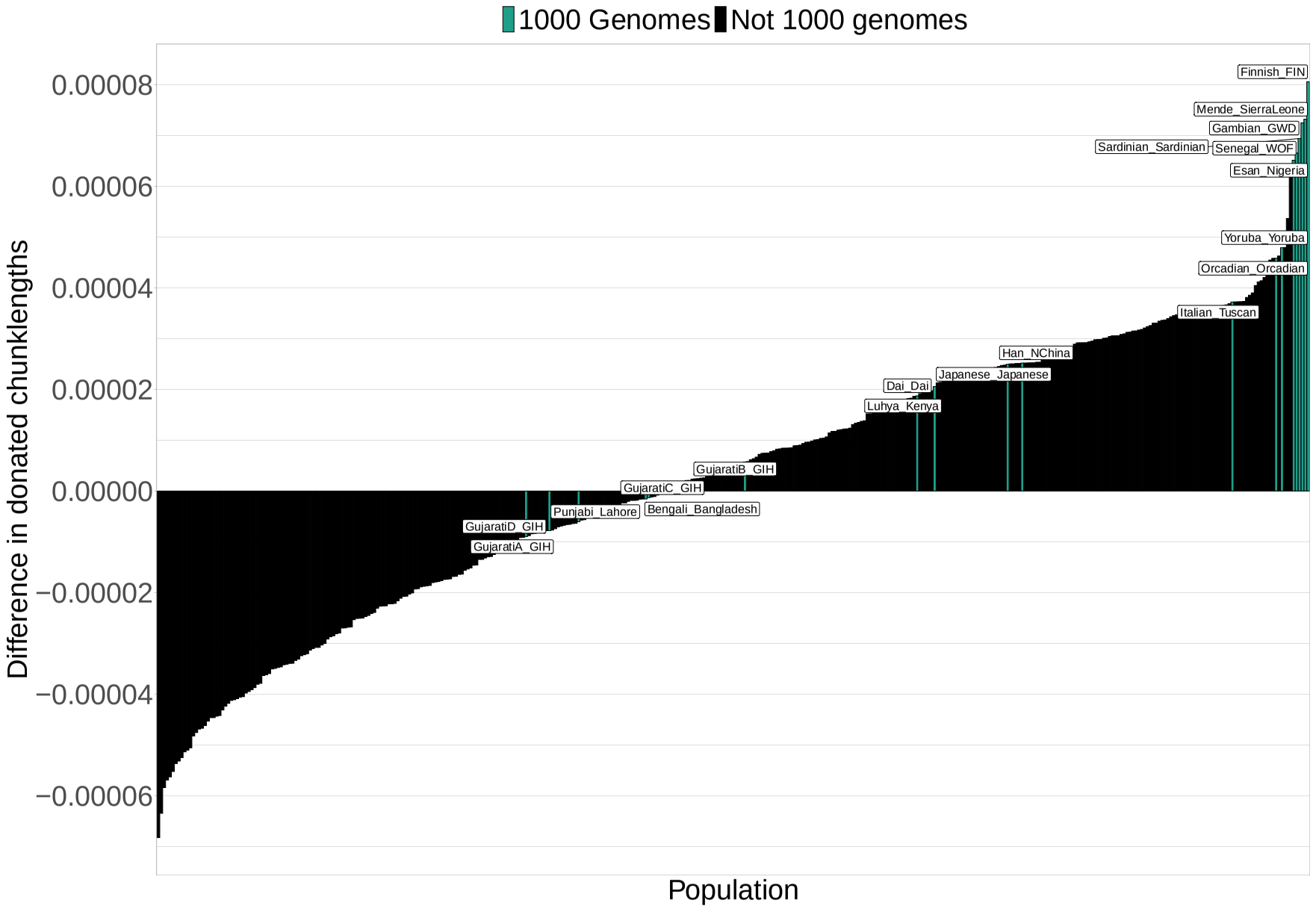


Figure 3.3: Differences in the amount donated by populations when using imputed and non-imputed data. Each vertical bar corresponds to a single population (N=395), with the height of the bar corresponding to the difference in haplotype donation, with positive values indicating increased donation and negative values indicating reduced donation. Bars coloured in green are also present in the 1000 genomes imputation panel and black bars are not present in the 1000 genomes.

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. A second cluster of UK Biobank individuals is located along the Southern African cline, close to the Bantu_SA label. 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.[BUT YOUR PREVIOUS SENTENCE MENTIONED CLOSENESS TO S.AFRICANS?? YOU NEED TO QUANTIFY HOW IT'S CLOSER TO W.AFRICANS THAN S.AFIRCANS.]

Aggregating the columns of the co-ancestry matrix by reference population and taking the sum 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. 3.5). Higher values correspond to more ancestry from that population in the U.K. Biobank sample.[THOUGH THIS CAN BE DUE (EVEN ENTIRELY) TO DONOR SAMPLE SIZE? YOU NEED TO NOTE THIS, AND DISCUSS WHY IT'S NOT A MAJOR ISSUE HERE.]

The map supports the findings from the PCA in Fig. 3.4; the populations with the largest contribution are those from West Africa (Fig. 3.5). In particular, populations from Ghana and Nigeria contribute the most to the ancestry of Biobank individuals. On the other hand, populations in East and North Africa contribute relatively little, with Southern / South-East Africa being approximately intermediate. This is consistent with two different historical events.

Figure 3.4: 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.

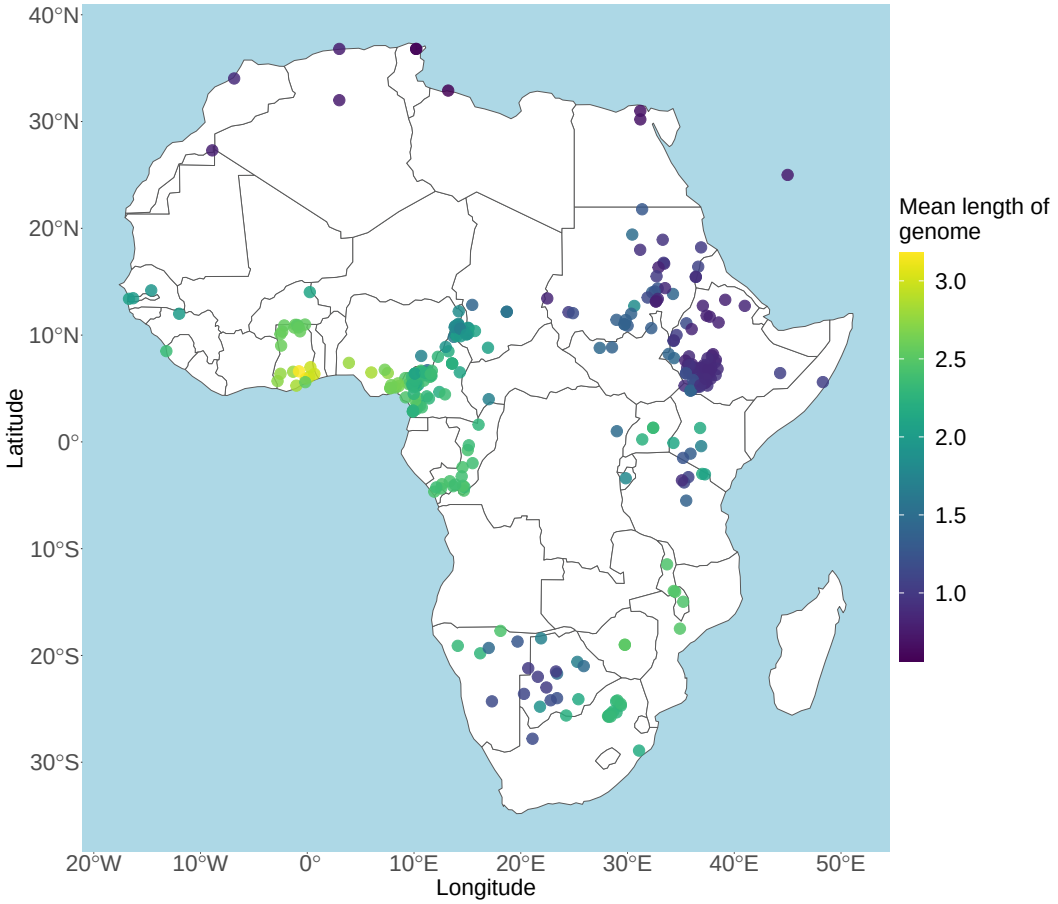


Figure 3.5: Map of haplotype donation to U.K. Biobank individuals. Each point represents a different African population. Colour corresponds to the mean length (cM) that populatation donated to all African U.K. Biobank individuals.

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 [65]. 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[REF], 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. 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 [reference needed].

[WOULD REMOVE]SOURCEFIND can provide a less noisy picture of the contribution of different ethnic groups to U.K. Biobank individuals. I analysed each U.K. Biobank individual with more than 50% African ancestry (n=8,476) using all Human Origins populations (n=535) as surrogates. [REPLACE WITH]: I used SOURCEFIND to infer the proportion of ancestry that each UK Biobank individual shares most recently with each of the 535 surrogate groups. [MAY WANT TO MENTION THIS WILL ACCOUNT FOR THE SAMPLE SIZE ISSUES IN FIG 3.5]]

Fig. 3.7 displays the 30 ethnic groups with the highest mean proportions of ancestry within the U.K. Biobank individuals. Yoruba was a clear standout for the most represented population; the mean proportion of Yoruba ancestry per individual was 40.0% 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 is. Interestingly, 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; and of the individuals born in the Caribbean, over half were assigned to the Yoruban ethnicity. 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. 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. 3.6). For example, Uganda is dominated by a single ethnic group because we only have reference data from a single group in Uganda. On the other hand, the individuals from Sudan are more evenly distributed across ethnicities, reflecting the wider array of reference populations from that country[OR THAT WE CAN'T DISTINGUISH THESE POPULATIONS – I THINK THERE ARE ONLY A FEW WE CAN, ACCORDING TO NANCY'S RESULTS. YOU SHOULD ASK HER FOR THIS.].

Some other patterns can be noted. Whilst many individuals have intermediate levels of ancestry from West African populations (e.g. Ghana_Fante or Yoruba_Yoruba), much fewer individuals have intermediate levels of Ethiopia_Somali ancestry (Fig. 3.7). This may be because Somali individuals tend to be relatively unadmixed relative to West African populations and hence can be modeled as a mixture of almost entirely Ethiopia_Somali ancestry. Another option is that we have relatively fewer surrogate populations from East Africa and so any individual with East African ancestry is likely to copy a lot from Ethiopia_Somali, whereas West African individuals may have their ancestry spread across the many surrogate groups from West Africa[YOU CAN TEST THIS – DO PEOPLE WITH WEST AFRICAN ANCESTRY TEND TO HAVE LESS OR MORE TOTAL AFRICAN ANCESTRY RELATIVE TO THOSE WITH SOMALI ANCESTRY? IT SEEMS ANOTHER EXPLANATION IS THAT PEOPLE WITH W.AFRICAN ANCESTRY ARE ADMIXED

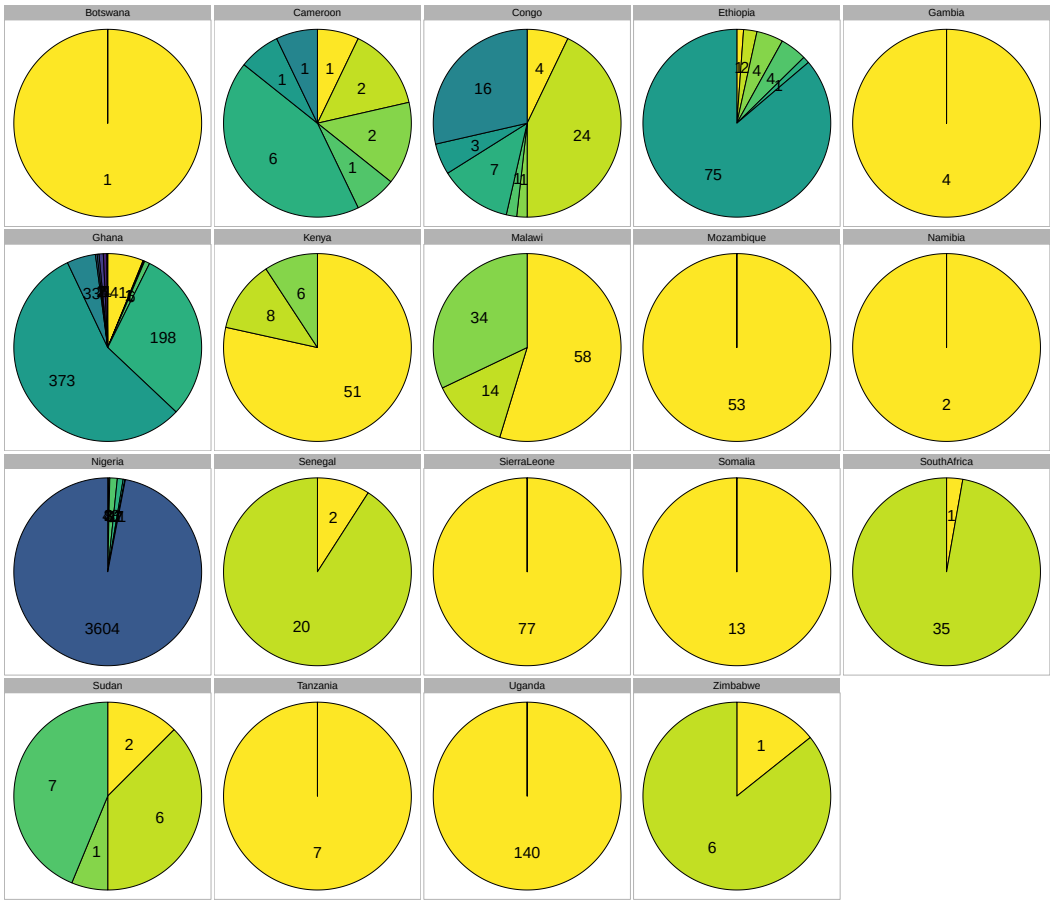


Figure 3.6: Variation of individuals assigned to different ethnic groups by country of assigned group. Each panel represents all individuals assigned to an ethnic group from that country, with proportions corresponding to different ethnic groups.

(E.G. WITH EUROPE, DUE TO SLAVE TRADE) WHILE SOMALIS ARE RECENT MIGRANTS AND HENCE UNADMIXED.].

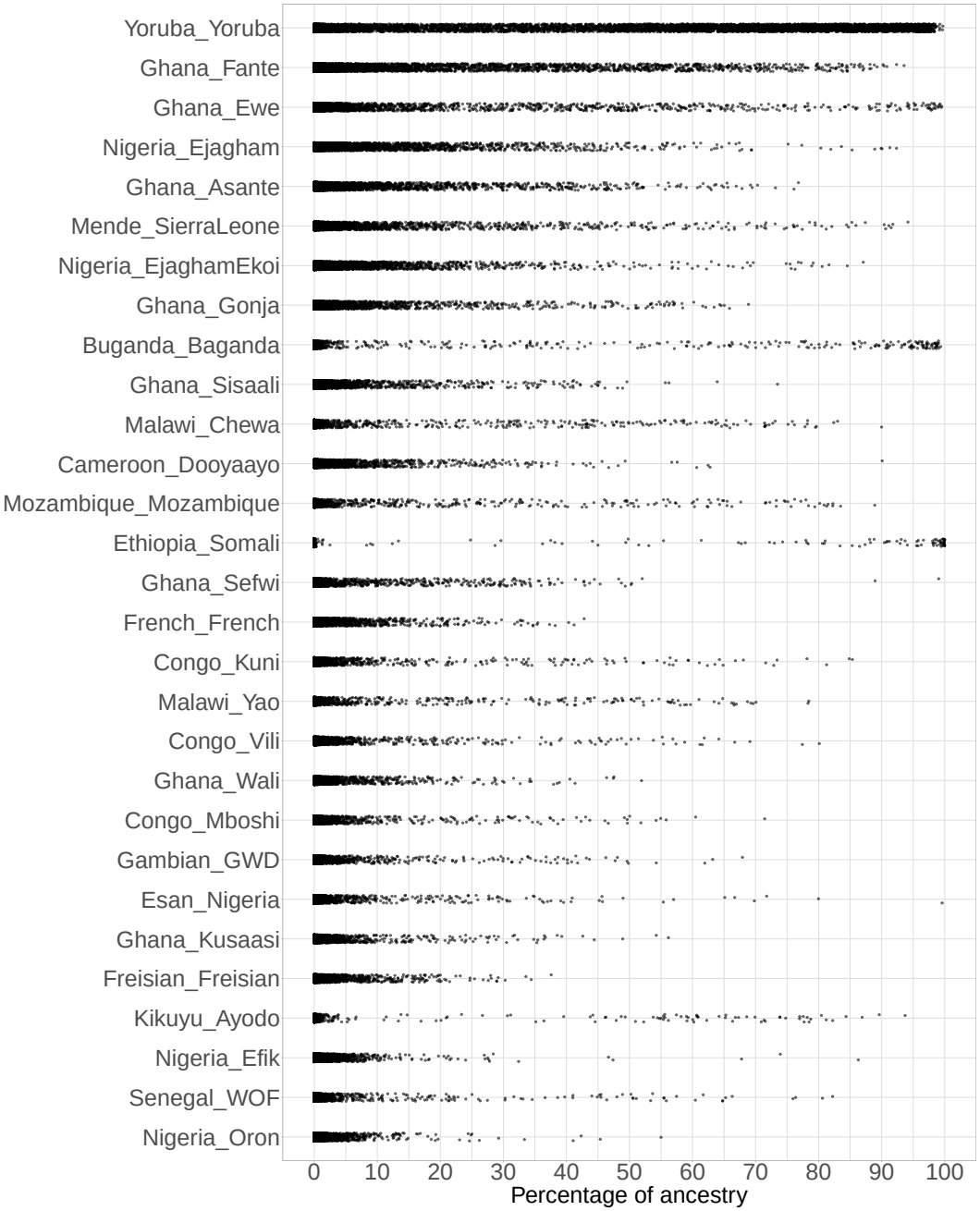


Figure 3.7: The 30 Human Origins populations which have the highest contribution to all U.K. Biobank individuals with at least 50% African Ancestry, based on SOURCEFIND analysis. Each row of points contains 8476 individuals and their position corresponds to the percentage of ancestry from that population.

3.3.4 Verifying painting accuracy

Given the total number of SNPs used in the analysis (n=65,727) is relatively low for studies modelling haplotype information, it is important to verify that the ChromoPainter results do not simply correspond to copying the most from the reference populations with the largest sample size. To test that the coancestry matrix contains relevant information about the ancestry of the retained U.K. Biobank individuals, I took advantage of the U.K. Biobank metadata and subsetted the coancestry matrix to contain only individuals who have data on birth location (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.[DO YOU NEED THIS? DOESN'T SOURCEFIND GET AROUND THE SAMPLE SIZE ISSUE? INSTEAD IT MAY BE WORTH LOOKING AT MEAN SF % MATCHING TO EACH AFRICAN SURROGATE VERSUS SAMPLE SIZE, AS HAVING A LARGER SAMPLE SIZE MAY INDICATE CAPTURING MORE DIVERSITY, WHICH IN TURN CAN LEAD TO MORE MATCHING (AS IT IS A MORE WIDELY REPRESENTATIVE SAMPLE).]

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 countries[BEST TO SAY "results"? OTHERWISE THIS LITERALLY IMPLIES SOME COUNTRIES ARE UNINTERESTING, WITHOUT PROVIDING CONTEXT.]. For example, there are

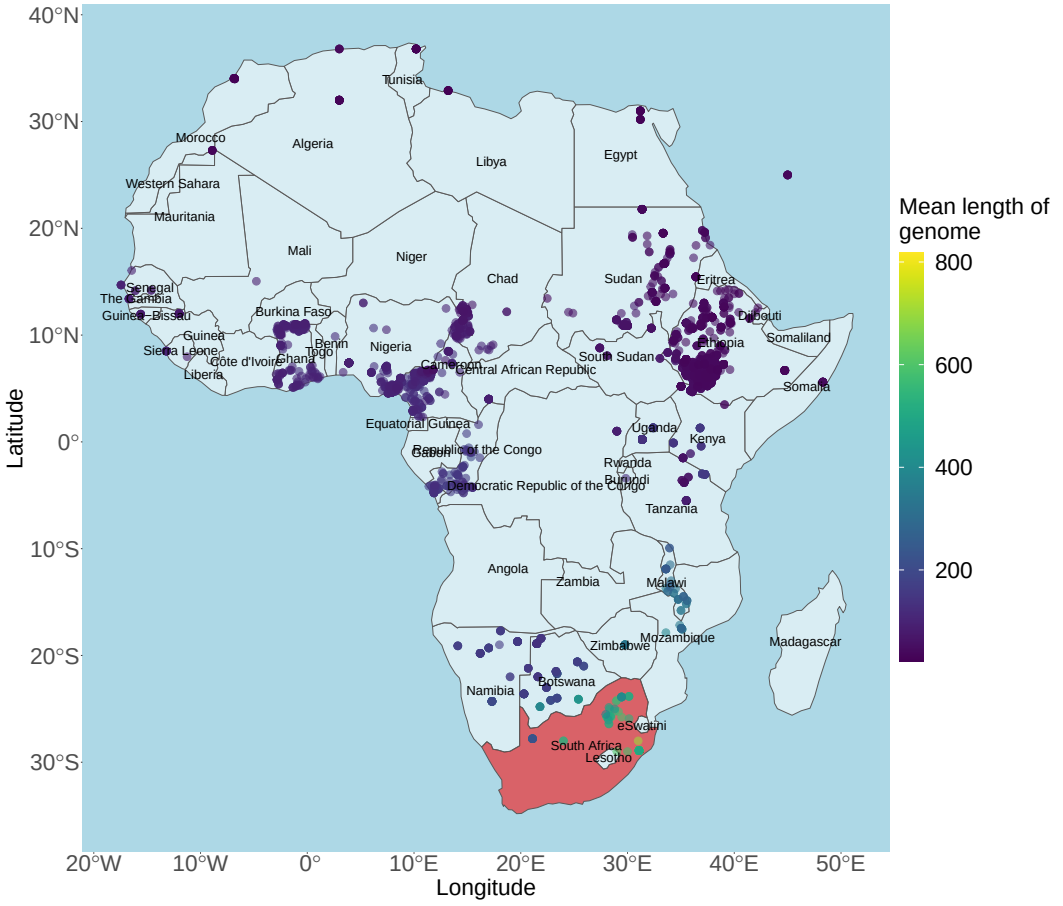


Figure 3.8: Map of haplotype donation to U.K. Biobank individuals born in South Africa. Each point represents one Human Origins population, coloured according to the summed amount of chunklengths that population donates to all U.K. Biobank individuals born in South Africa.

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 [FIGURE??], consistent with historical evidence [65]. Individuals born in Brazil have ancestry from further South, again consistent with historical evidence (citation needed). 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 more 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%[NOT CLEAR THIS IS BEST – WORTH HAVING A PLOT OF PREDICTION ACCURACY VS % THRESHOLD?] 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.

The overall accuracy at predicting birth location was 81.63%, suggesting there was substantial information within the coancestry matrix. For the countries where there was a large number of reference populations[NOT CLEAR THIS IS RELATED – GIVEN YOU SHOWED THAT NIGERIA MATCHES PRIMARILY TO ONLY ONE GROUP], such as Ghana and Nigeria, the prediction accuracy was high. For certain countries, the prediction accuracy was much lower. For example, Tanzania, which is only represented by a single reference population, had a prediction match of XX%. Zimbabwe had by far the lowest prediction accuracy (14%) out of countries with more than 100 individuals[DO YOU MEAN REF INDIVIDUALS OR UK 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,

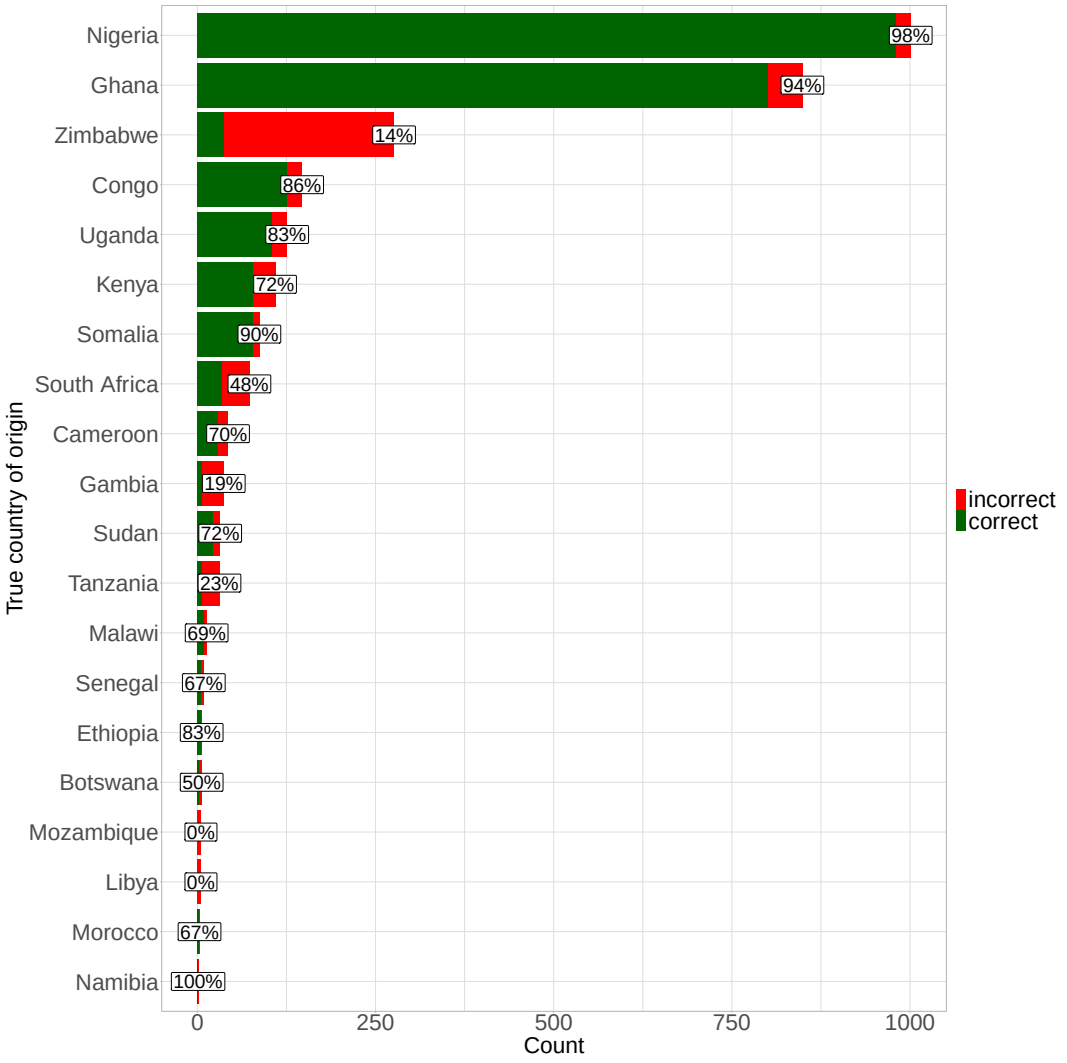


Figure 3.9: Correspondence of true birth country with estimated birth country. Each bar corresponds to a true birth country, with the length of the bar corresponding to the total number of people in our dataset born in that country. The green section corresponds to the total number of individuals where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

showing the possible weakness of this approach which aims to categorise individuals into a single country, as ethnic groups often transcend countries.

I also 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. This yielded an accuracy of 81.89%, a value almost identical to that obtained with the dataset containing approximately 70,000 non-imputed SNPs.[I THINK YOU WANT MORE HERE, AS THIS CONTRADICTS THE POINT YOU ARE TRYING TO MAKE ABOUT IMPUTATION BEING BAD. FOR EXAMPLE, "Despite my earlier results indicating that SOURCEFIND results are less accurate if using imputed data due to reference bias..." WOULD ALSO BE GOOD TO KNOW WHERE THEY DIFFER; FOR WHICH POPS IF AT ALL (E.G. MAYBE A SCATTERPLOT OF ACCURACIES UNDER THE TWO CASES)]

3.3.5 Patterns of African ancestry across the U.K.

The U.K. Biobank dataset also contains data on the testing centre that each individual registered at. I used this information to see 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 (Fig. 3.10). No clear pattern was apparent, other than Yoruba ancestry dominating most centres.

I also estimated the information entropy[HOW DID YOU CALCULATE THIS? DOES THIS INCLUDE EUROPEAN ANCESTRY?] of each centre based on the SOURCEFIND proportions, analogous to previous work done on European ancestry in UK Biobank individuals (S.Hu, personal communication). In this context, information entropy can be seen as analagous to the amount of diversity in different ancestries present in a particular centre. The two centres in the two largest metropolitan areas, London and Birmingham, have the largest amount of entropy[HOW MUCH SO? SHOULD HIGHLIGHT LONDON ON

THE FIG – IS CROYDON A MISTAKE?], suggesting that ethnic diversity is greater in larger cities.


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

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 [66] and musical instruments dated to 40kya [67], 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 a perfect 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. 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-hight 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 and represent an excellent opportunity to study a time transect.

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

I was interested in addressing the following questions:

1. **Second Neolithic immigration wave**. Do we observe genetic differences between the first and second wave of farmers that brought farming to the region technology from South-East Asia to Europe during the Early Neolithic?

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. **The Iron Age**. How do the samples from the Iron Age compare to the single sample from the Bronze Age,and both to central Europeans

4. **Iron Age - Middle Age** To what extent do the 'proto-Celtic' samples of the Iron Age show similarity to the later early 'Germanic' people of the middle ages.

5. 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.

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 Methods

4.2.1 Data generation

11 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). 6 of the samples were found in Cherry-Tree Cave in the Bavarian district of Forchheim (Fig. 4.1), 4 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.

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

Collaborators in University of Mainz performed DNA extraction, read alignment and variant calling.

4.2.3 Genotype imputation and phasing using GLIMPSE

ChromoPainter analysis requires that all samples contain i) phased genotypes and ii) no missing genotypes. GLIMPSE [24] is a software which is designed to

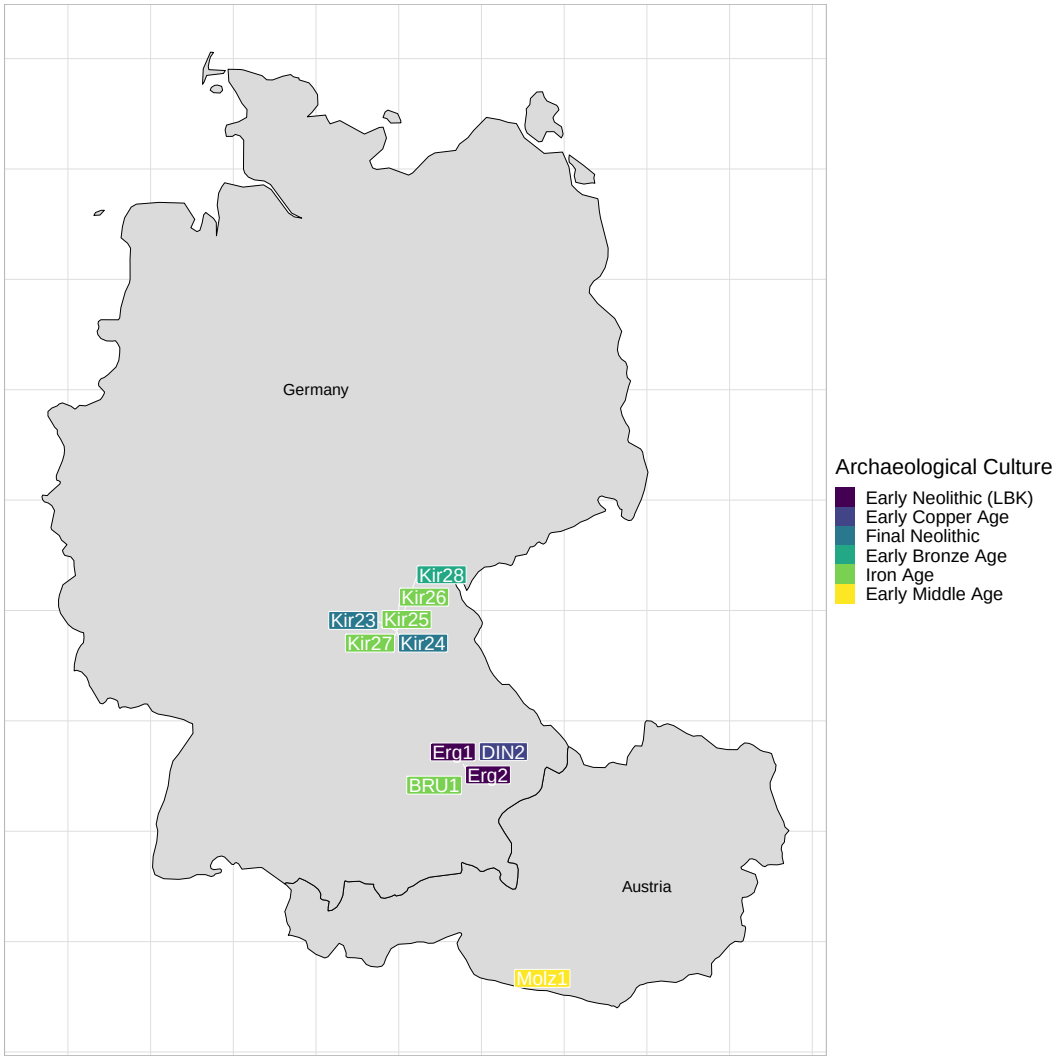


Figure 4.1: Map of newly sequenced ancient individuals, positioned according to where they were excavated. Colour on label corresponds to archaeological culture which they were found.

perform both phasing and imputation on low-coverage sequence data with the aid of a reference panel. GLIMPSE was chosen as it was the fastest phasing / imputation software available for low coverage sequence data avaliable at the time of writing [24]. The other possible option, Beagle4 [33], is too slow and memory inefficient to impute the large number combined number of individuals and SNPs present in this dataset.

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

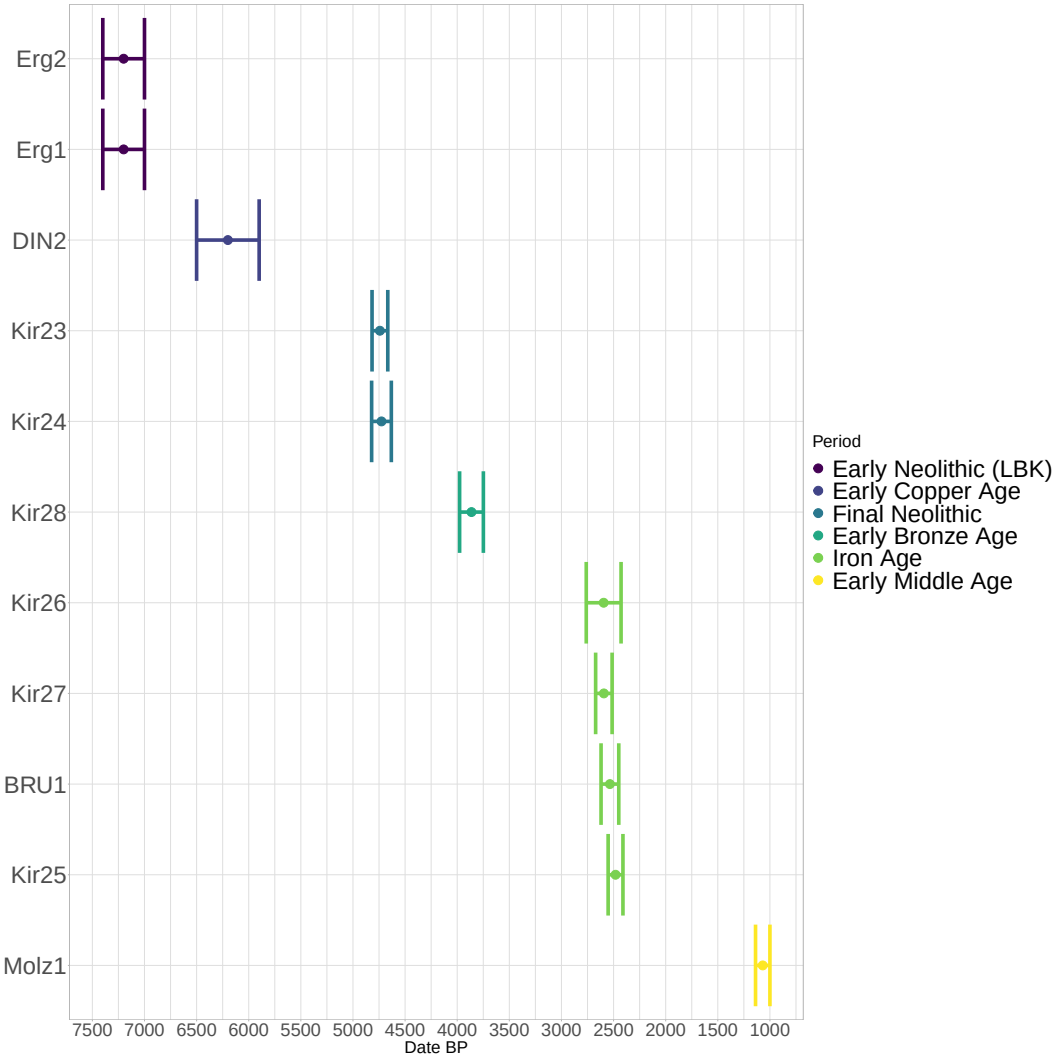


Figure 4.2: Estimated radiocarbon dates for each newly sequenced ancient individual, grouped by archaeological period.

likelihood data at 77,213,942 genome-wide SNPs. Data was then split into separate `.bcf` files for each chromosome and indexed using bcftools [68].

I followed the GLIMPSE [24] 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 [34], described in appendix A.5.

`GLIMPSE chunk` was used to split the present-day reference dataset into chunks. Default settings of `-window-size 2000000` and `-buffer-size`

200000 were used, generating a total of 936 regions genome wide. Splitting the genome into regions for imputation jointly maximises computational efficiency and accuracy [24]. For each region in turn, the target dataset, consisting of phred-scaled genotype likelihoods (PL), was imputed using `GLIMPSE phase` under default settings and the same reference panel. Here, 'impute' means filling in sporadic missing genotypes that are missing in a single individual (or sometimes more), rather than the whole-scale imputation of non-genotyped positions. `GLIMPSE ligate` was then used to concatenate the 936 imputed regions into 22 distinct chromosomes. Finally, `GLIMPSE sample` was used to generate phased haplotypes from the output of `GLIMPSE ligate` using default settings. 50,342,061 bi-allelic autosomal SNPs remained after phasing and imputation.

4.2.4 Determination of uniparental haplogroups

Haplogrep (`https://haplogrep.i-med.ac.at/`) was used to identify the mtDNA and y-chromosome haplogroups for each newly sequenced ancient samples [69] from the raw `.fastq` files.

4.2.5 Estimation sample-heterozygosity

The phased haplotypes from the output of GLIMPSE were used to estimate per-sample heterozygosity using the `plink2 -het` command.

4.2.6 IBD sharing

I used hap-IBD [70] to estimate IBD segments between all pairs of ancient individuals, using the phased output from GLIMPSE as input haplotypes, using the genetic maps supplied and leaving all parameters as default. I estimated IBD segments for each chromosome separately and summed their length segments between each pair of individuals across all chromosomes.

4.2.7 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 the identification 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 and the first two principle components plotted.

4.2.8 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 haplotype was compared to all other haplotypes in turn. 1.5x was somewhat arbitrarily chosen as a conservative threshold to reduce coverage related bias whilst still retaining a suitable number of individuals. This is the painting that can be used to perform fineSTRUCTURE clustering and tree building on ancient samples. Hereafter referred to as 'ancient' painting.

Principle Component Analysis was performed on the coancestry matrix of the 'ancients' painting using the `prcomp_irlba` function from the irlba R libary. 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 individuals at hand. 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

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

becomes impossible. Therefore, a process of trial and error is used to decide which individuals to include in a PCA. This process is somewhat subjective, but occasionally subjective methods are required when facing heterogeneous ancient DNA datasets.

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 4.2. Hereafter referred to as 'present-day painting'.

I selected the MS POBI HellBus dataset of present-day individuals, described in Appendix A.21, to co-analyse the newly sequenced ancient individuals with,

as it contains a large number of European individuals which are relevant to studying the history of Germany. Appendix A.21 describes the phasing and pre-processing performed on this dataset. Once phased, I converted it to ChromoPainter format using a custom script (`https://github.com/sahwa/vcf_to_chromopainter`). I then merged the ancient individuals, described in sections 4.2.3, with the MS POBI HellBus dataset. I retained all bi-allelic SNPs common to both datasets, resulting in 414,155 SNPs.

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 (v0.0.5) [9] clustering and tree building algorithm was applied to 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. 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 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`).

4.2.9 SOURCEFIND

The coancestry matrices outputted by ChromoPainter give informative but noisy estimates of the how much recent ancestry a given individual most closely shares with another individual. This noise is due to in part to incomplete lineage sorting. SOURCEFIND [11] is a method of de-noising the ancestry estimates outputted by ChromoPainter. In SOURCEFIND nomenclature, 'surrogates' are populations for which we wish to estimate proportions of ancestry of in our target population. For instance, in the most simple case, we may wish to estimate the proportion of 'African' and 'European' ancestry within an admixed target individual. There, we would use 2 surrogates, perhaps a population from Yoruba and a population from France. The surrogates need not be the 'true' mixing sources. It is important to to note that SOURCEFIND does not directly infer admixture events, in that it does not model the break down of admixture LD as does ALDER or GLOBETROTTER.

I performed 3 different SOURCEFIND analyses, each with a different set of surrogates.

Previous research suggests that almost all ancient Europeans (discounting particular paleolithic Hunter-Gather populations [48]) can be well modeled as differing amounts of Western Hunter Gatherers (WHG), Early Neolithic farmers from present-day Anatolia and Bronze Age pastoralists from the Eurasian Steppe (best represented by individuals from the Yamnaya culture [16]). Therefore, a simple approach to identifying the broad-scale ancestry of ancient individuals is to model them as a mixture of the three aforementioned populations. I performed SOURCEFIND analysis, using these 3 populations as surrogates.

I also estimated finer-scale ancestry proportions by modeling each newly sequenced ancient sample as a mixture of all ancient populations which are older than the target. I chose to use only older surrogate populations as modeling

an individual in terms of more recent populations can lead to results which are more difficult to directly interpret. I retained surrogate populations where the average age of samples within the population was older or within 100 years of the target individual. Accordingly, different newly sequenced ancient samples were analysed using different number of surrogate populations.

Finally, I wanted to estimate proportions of ancestry from present-day surrogates in ancient samples. For this analysis, I used the 'modern' painting and used all modern populations from the 'HellBus' dataset as surrogates, shown in Table 4.2.

For all of the three above analyses, I performed 3 independent SOURCEFIND runs. SOURCEFIND explores the parameter space of ancestry proportions using MCMC sampling. For each target, all 3 MCMC runs were then combined to form an MCMC list using the `coda` R library [39]. I used the `mcmc` function to jointly estimate ancestry proportions and empirical credible intervals for each target population.

4.2.10 MOSAIC admixture analysis

I inferred admixture events, dates and proportions using MOSAIC, a haplotype-based method [71]. MOSAIC was used because, unlike GLOBETROTTER [10], the 'painting' step and admixture inference step are combined into one, resulting in a simpler pipeline and more flexible assignment of different surrogates (i.e. the set of surrogates can be changed without repainting the samples). 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.

I performed 2 different kinds of admixture analysis. Firstly, I performed an 'ancient surrogates' analysis where the all ancient samples above 1.5x coverage were used as surrogates. I used the fineSTRUCTURE groupings to assign the

samples into surrogate population.

I then performed a 'present-day surrogates' analysis where a selected set of present-day populations were used to analyse both present-day Slavic populations and ancient Slavic populations. From prior experience, using these samples provided less-noisy results (due to larger population sample sizes and more homogenous populations), at the cost of reduced interpretability. By this, I mean that forming, for example, a 4,000 year-old Bronze Age sample as a mixture of present Europeans may be difficult to interpret; what does it mean if it is 70% French and 30% Tuscan? As the samples become more recent (i.e. into the Middle Ages), forming them as a mixture of moderns becomes more appropriate, since there has been relatively fewer large scale population movements separating the ancient and modern samples.

Phased `.vcf` files, the output from GLIMPSE, were converted to `.hap/.sample` files using `bcftools convert -hapsample`. The resulting `.hap/.sample` files were then converted to MOSAIC input using the provided script (`https://maths.ucd.ie/~mst/MOSAIC/convert_from_haps.R`). 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 when more than one sample was present. Otherwise, when there is a single target sample, it is not possible to obtain confidence intervals.

4.2.11 F-statistics

Many of the relevant samples in the literature were of either very low coverage (< 0.1) or genotyped on a capture array. Firstly, my in chapter 2 has shown that samples less than 0.5x cannot reliably be analysed using ChromoPainter. F-statistics (such as the f_3 admixture test or the f_4 branch test) are mostly robust to coverage related effects [72]. Therefore, using these methods allows

for the co-analysis of a much larger number of samples. In particular, there were many low-coverage samples from LBK cultures from Rivollat et al (2020) which would not have been suitable for use with ChromoPainter [73]. The dataset used to calculate F-statistics contained 942 ancient samples from 143 populations, described in appendices A1 to A18, from the literature and 2280 present-day individuals from 144 populations from the HellBus dataset.

I used Admixtools [13], implemented in admixtools2 R library (`https://uqrmaie1.github.io/admixtools`) to calculate several different f-statistics.

I converted imputed genotypes in `.vcf` format to `.ped`/`.map` format using plink. It has been shown that using imputed markers reduced reference bias relative to using pseudo-haploid markers [17]. Convertf (`https://github.com/argriffing/eigensoft/tree/master/CONVERTF`) from the Admixtools library was then used to convert `.ped`/`.map` files into Eigenstrat format suitable for use with Admixtools.

I used the f_4 branch test to test whether 2 populations form a clade to the exclusion of 2 other populations. For example, the $f_4(french, german; yoruba, mbuti)$ would give a Z score not significantly different to zero, given we would expect $french$ and $german$ to form a clade to the exclusion of $yoruba$ and $mbuti$. Exchanging $french$ with $yoruba$ would yield $|Z| > 3$.

f_3 in the form of $f_3(A, B; C)$ was used to either i) estimate the branch length between A and B after their divergence from C, or ii) to test whether C has been formed from an admixture event between A and B, based on if C has allele-frequencies which are intermediate between A and B.

qpAdm was used to infer ancestry proportions. I followed the steps (as closely as was possible given the samples avaliable to me) taken in Olalde et al (2018) with respect to outgroup selection, choosing the following populations/samples: *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 overview of genetic ancestry

To obtain a broad overview of the genetic ancestry of the newly sequenced samples in the context of a large number of previously published ancient samples, I performed a principle component analysis on the genotype matrix of selected ancient samples using plink2 (Fig. 4.3).

As expected, the samples from the Early Neolithic (approx 5200BC) and Copper Age (approx 4200BC) cluster with other samples from the European Neolithic. Previous studies have explained the pattern observed when Neolithic samples are plotted on a PCA [74]; the earliest Neolithic samples, from Anatolia and Greece, who are thought to be the source population from which all subsequent Neolithic farmers derive [16, 75–78], are usually positioned at the end of the cluster which is farthest away from the Bronze Age samples. This likely reflects the fact they are unadmixed and more drifted 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 [74] and acquired local hunter-gatherer ancestry. Accordingly, these samples are shifted away from the earlier Neolithic samples, 'north', towards the Bronze Age samples. I did not include hunter-gatherer samples in order to maximise the relevant space-efficiency of the PCA, but the later Neolithic samples would be shifted towards hunter-gatherer populations if present (see supplementary figure D.1). With this in mind, the position of Erg1, shifted north away from the contemporaneous sample Erg2, is suggestive of hunter-gatherer admixture.

The 2 samples from the Late Neolithic display substantial differences; one sample, Kir24, is positioned at the extreme top-left of the PCA, close to the Yamnaya type-specimen, indicating it shares very recent ancestry with the first Yamnaya individuals who carried 'Steppe-related' ancestry from central Asia to Europe. On the other hand, Kir23 clusters with samples from Neolithic Europe. Kir28, the single Bronze Age sample clusters primarily with other European Bronze Age samples, whereas the 4 samples from the Iron Age are shifted substantially towards the Neolithic cluster, consistent with the placement of other samples from the European Iron Age. Finally, the 3 samples from the Medieval period, Alh1, Alh10 and Molz1, cluster with the Bronze Age.

4.3.2 Early Neolithic

The 3 Early-Middle Neolithic samples all display strong affinity to Anatolian farmers, consistent with the prevailing theory that near-eastern farmers were responsible for the spread of farming across Europe, and that all Neolithic farmers share recent common ancestry [16, 76–78]. fineSTRUCTURE analysis grouped Erg1 with 2 samples from Upper Paleolithic/Neolithic Italy and DIN2 clusters with Neolithic samples from Germany, Greece, Anatolia and Hungary. Despite their age, the genetic variation of the Early Neolithic samples falls well within the variation of present-day individuals; when painted using modern individuals, the 3 Early Neolithic individuals cluster with present-day Italians, consistent with previous research [16, 40] (Fig. 5.7). Erg1 was assigned to mtDNA haplogroup K which has been found in Neolithic and pre-pottery sites across Europe [75, 79] and Western Asia [80, 81].

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 and around 500 years more recent, and is thought to potentially belong to a second wave of farmers who migrated along the Danube (J. Burger 2018,

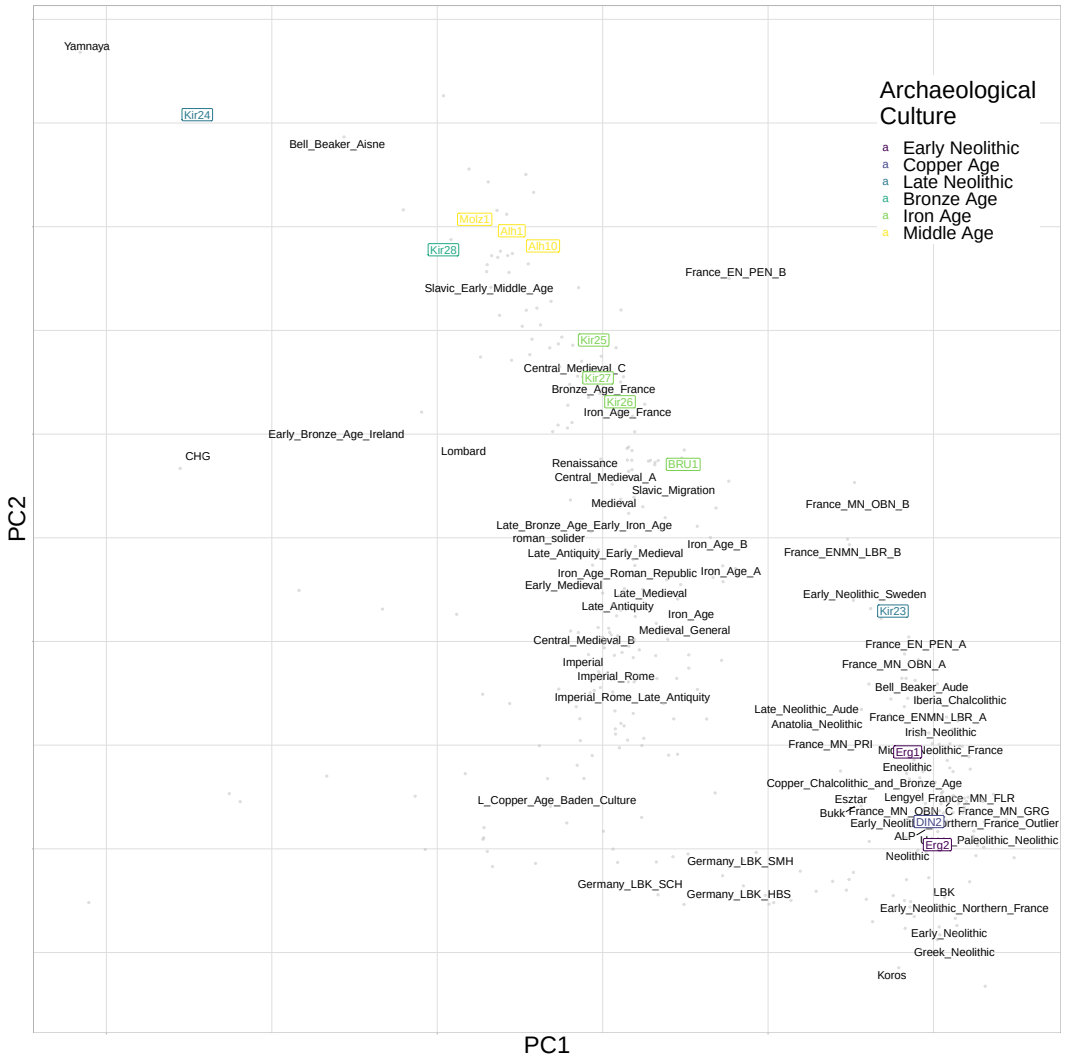


Figure 4.3: Principle component analysis of genotype matrix using plink2. Grey points indicate principle component coordiantes for each sample. Black text indicated mean principle component coordinates for all individuals within that group. Coloured labels represent newly sequenced ancient samples.

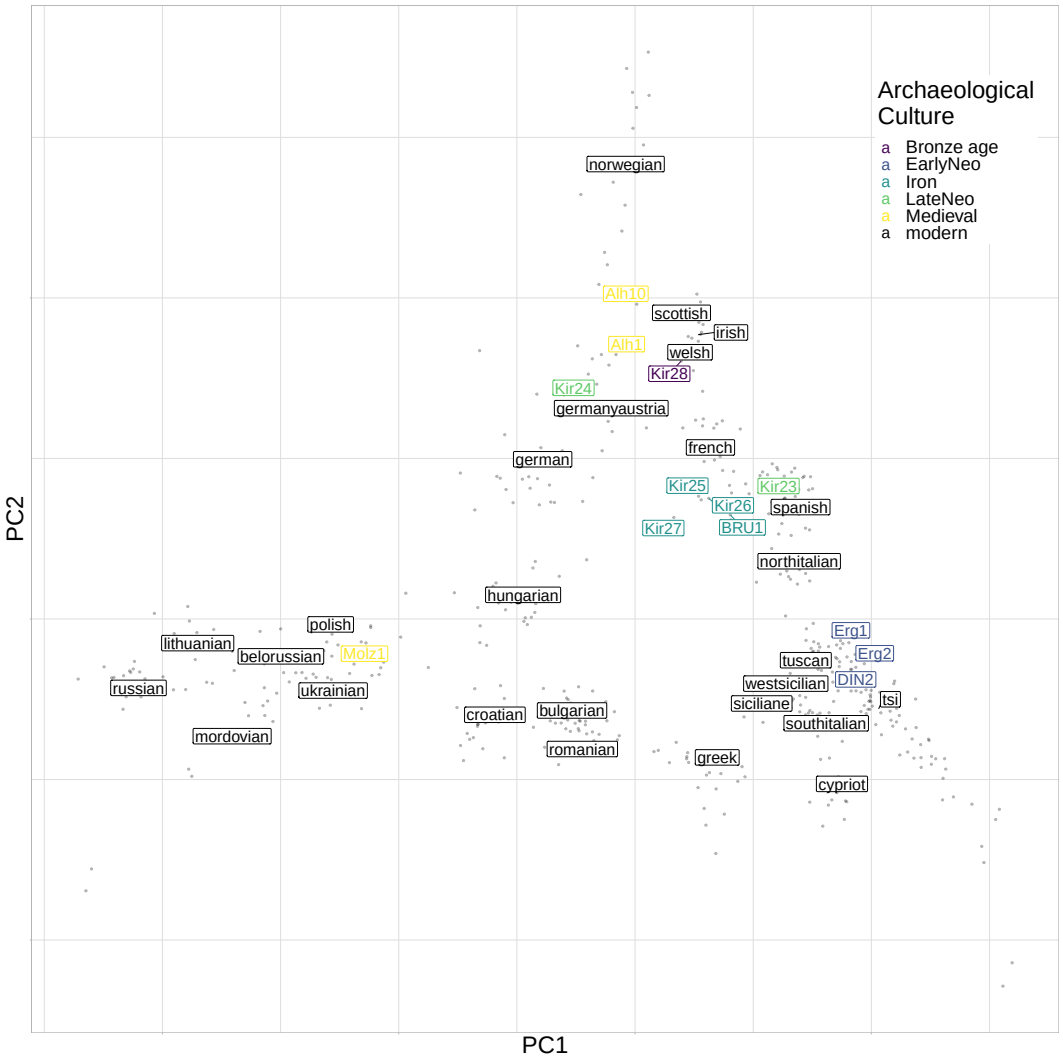


Figure 4.4: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations.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

personal communication). 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. It 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 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. However, 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 (Rivollat 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). I wanted to know which LBK population Erg1 and DIN2 were closest to. I found strong evidence ($|Z| = 7.97$) that Erg1 shared mored 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. 4.8).

4.3.3 Hunter-gather ancestry in Neolithic farmers

Prior research has shown that admixture occurred between newly arrived farming immigrants from Anatolia and local hunter-gatherers [74, 82]. 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. 4.9). 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.

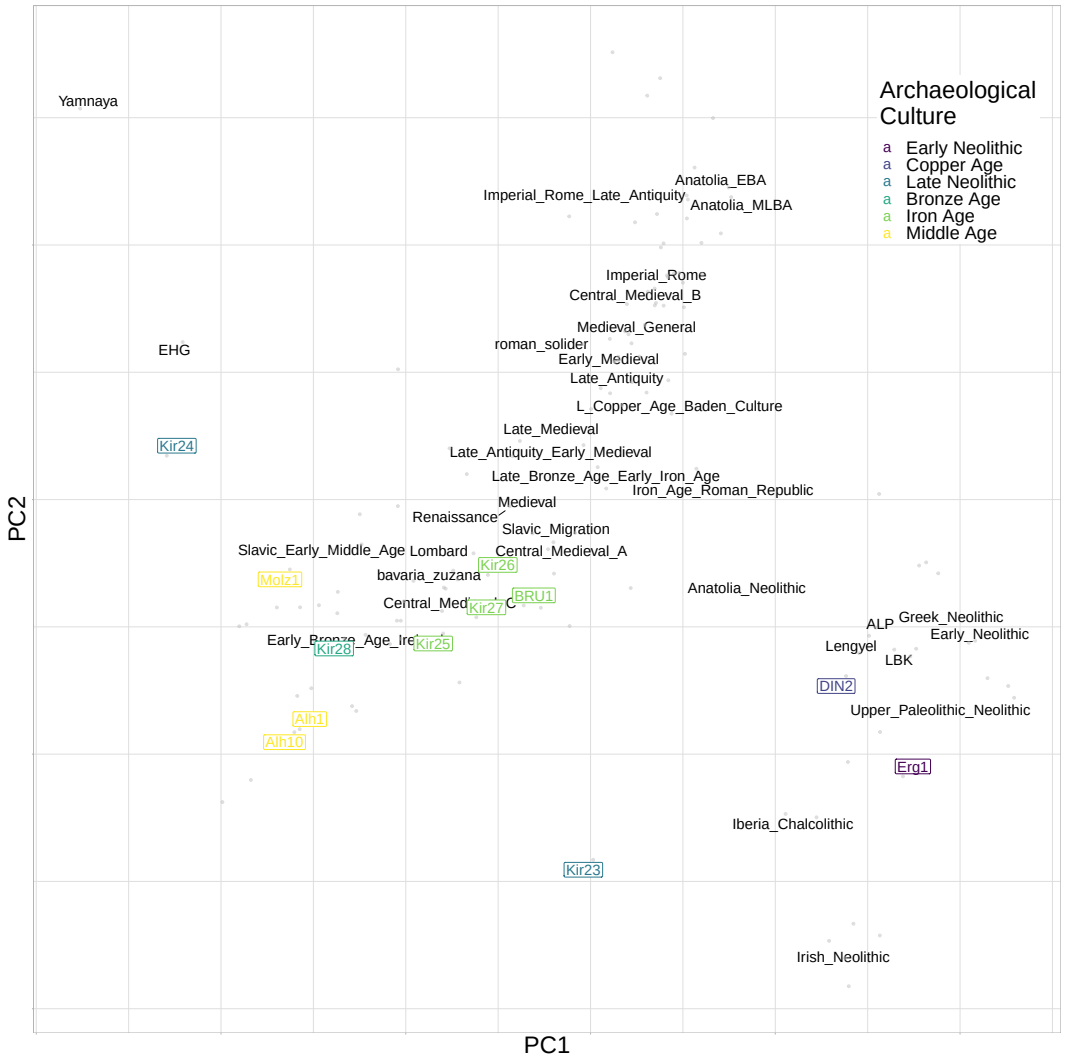


Figure 4.5: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using the finestructure library. Filled labels correspond to newly sequenced individuals and white labels correspond to reference populations. The position of each reference label is the mean PC coordinates of all individuals within that population

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, again implying it was part of a structured population with differential amounts of hunter-gatherer admixture.

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 the inferred proportion.

To determine the date of admixture between an Anatolian Farmer-like and WHG-like source into Erg1, I used MOSAIC [71], 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. 4.7). 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'

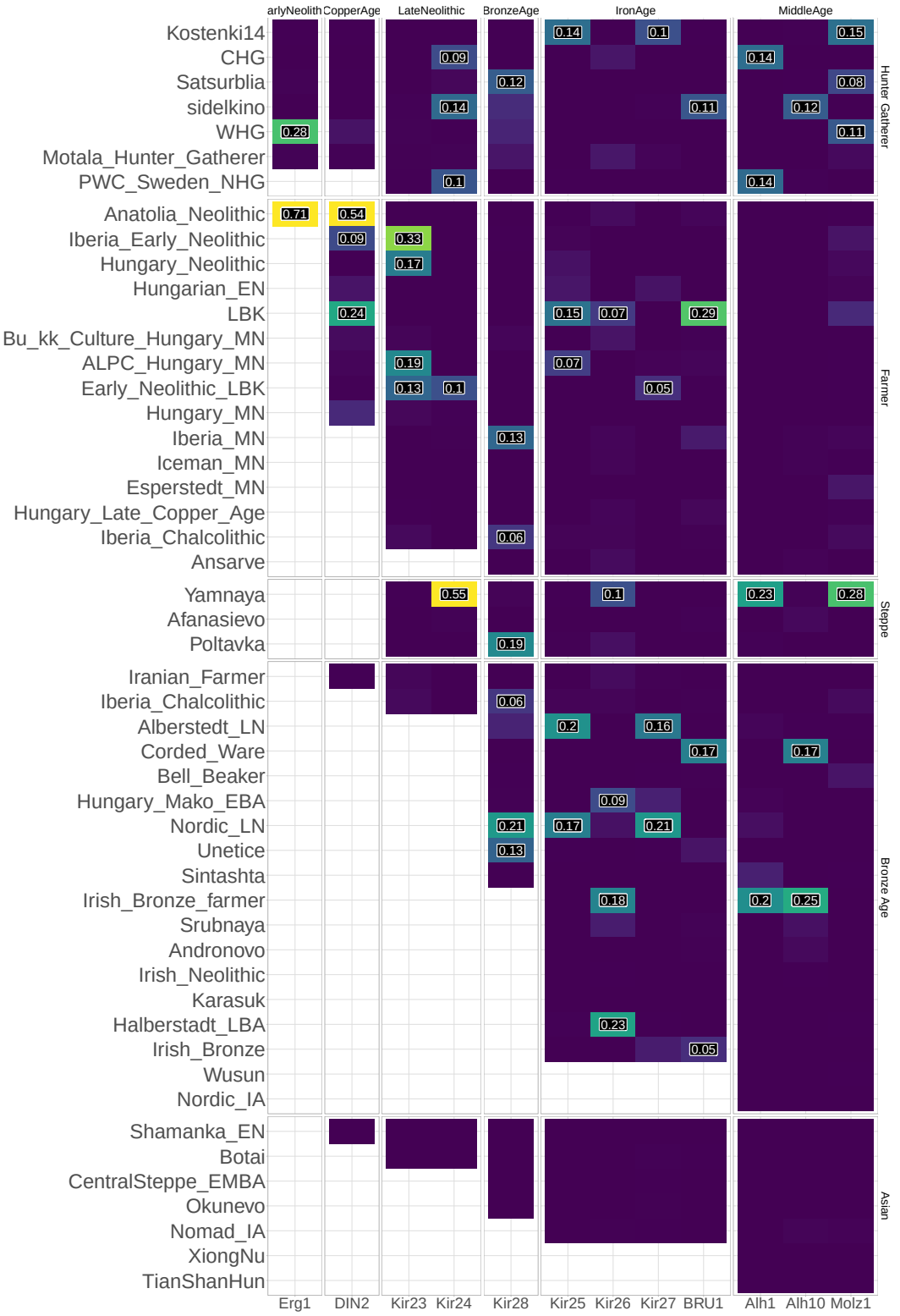


Figure 4.6: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture 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.

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.

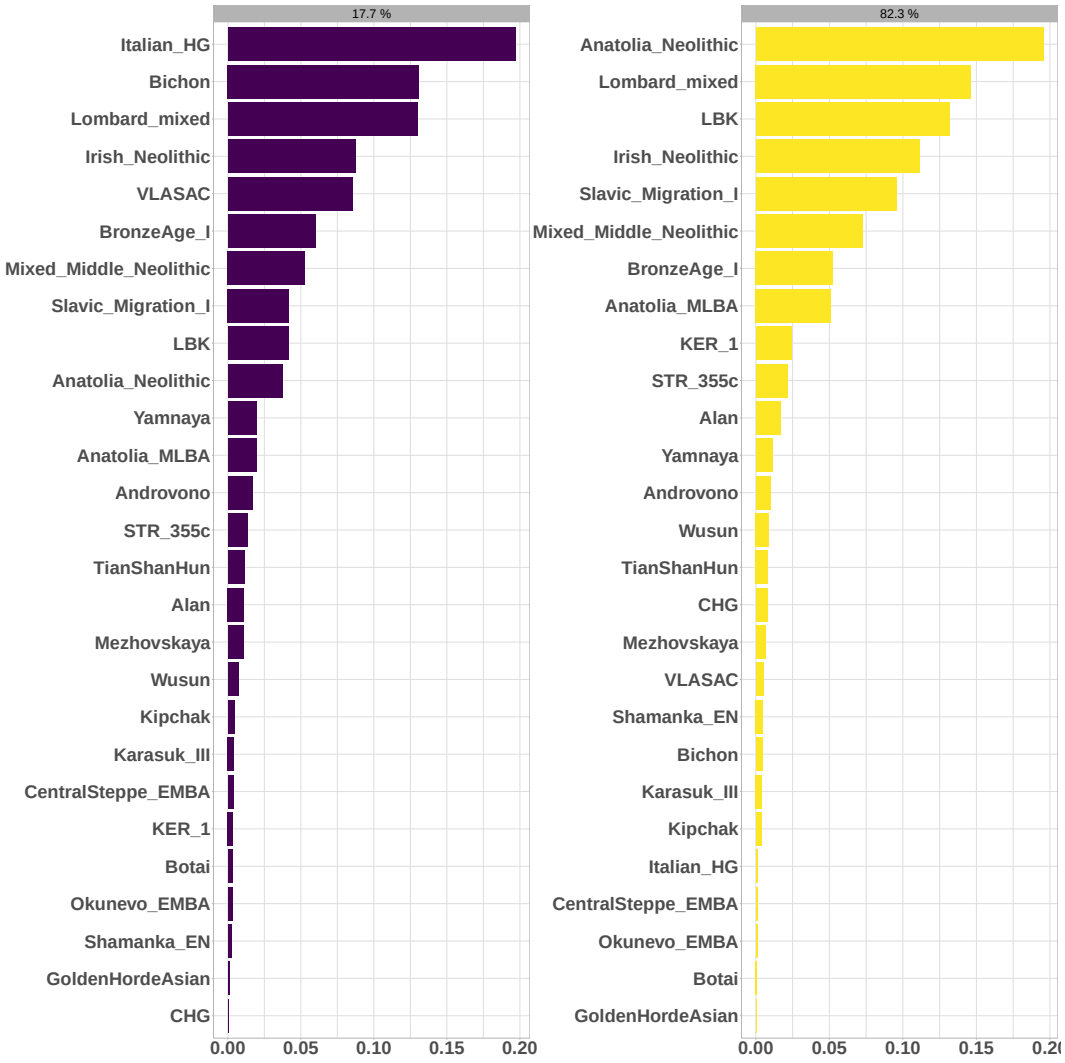


Figure 4.7: Copying matrix plot for sources in 2-way admixture event for Erg1. Each panel represents one of the 2 mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars within each panel represent the amount that mixing source copied from a particular population.

To confirm this admixture event, I performed an f_3 admixture test, which,

when significantly negative, provides unambiguous evidence of an admixture event [13]. 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 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 [83], 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. 4.8). On the other hand, Erg2, a sample which is contemporaneous and local to Erg1, showed no evidence of Hunter Gatherer admixture

4.3.4 Late Neolithic

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

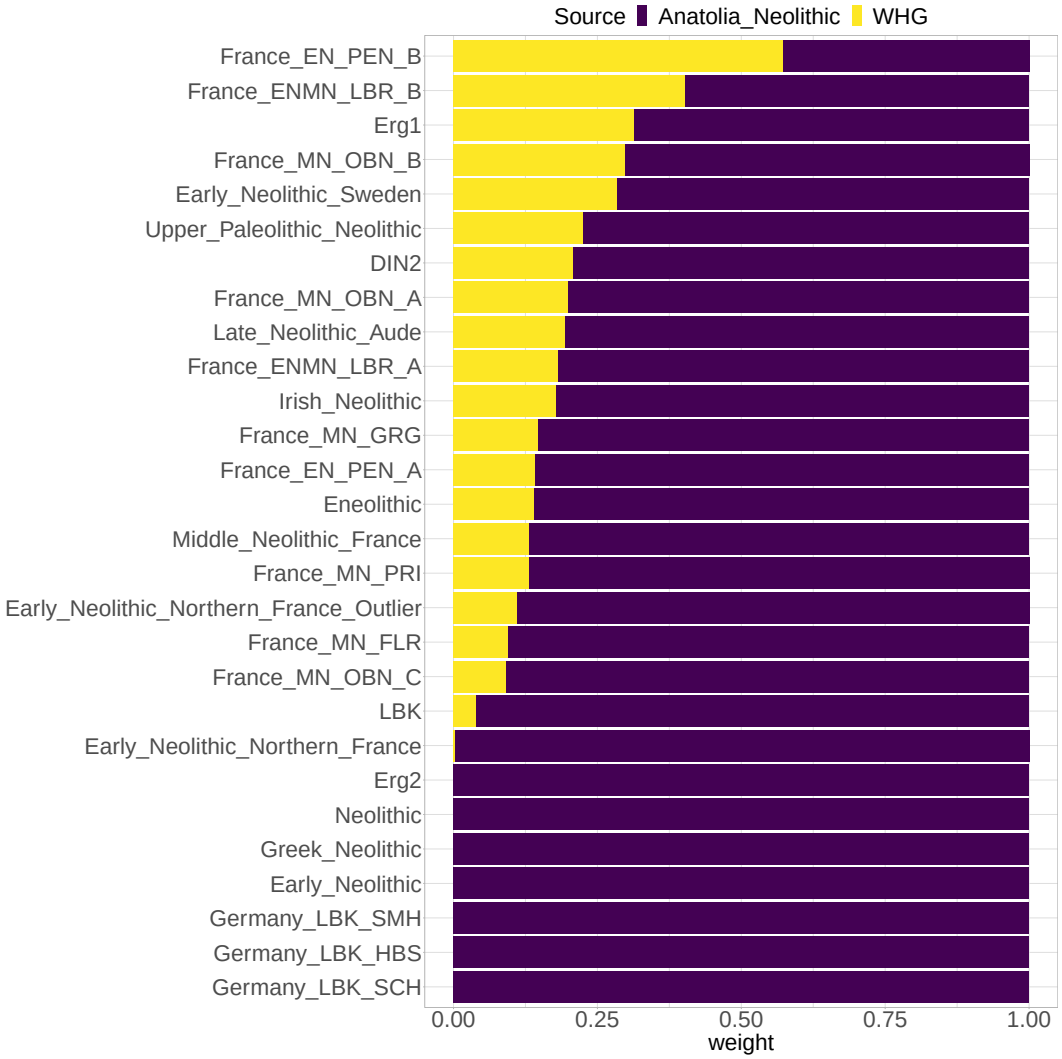


Figure 4.8: qpAdm ancestry proportion estimates for a selection of European Neolithic individuals. All individuals were modeled as a 2-way mixture between Anatolian Neolithic farmers and Western-Hunter Gatherers (WHG). Outgroups given in methods 4.2.9.

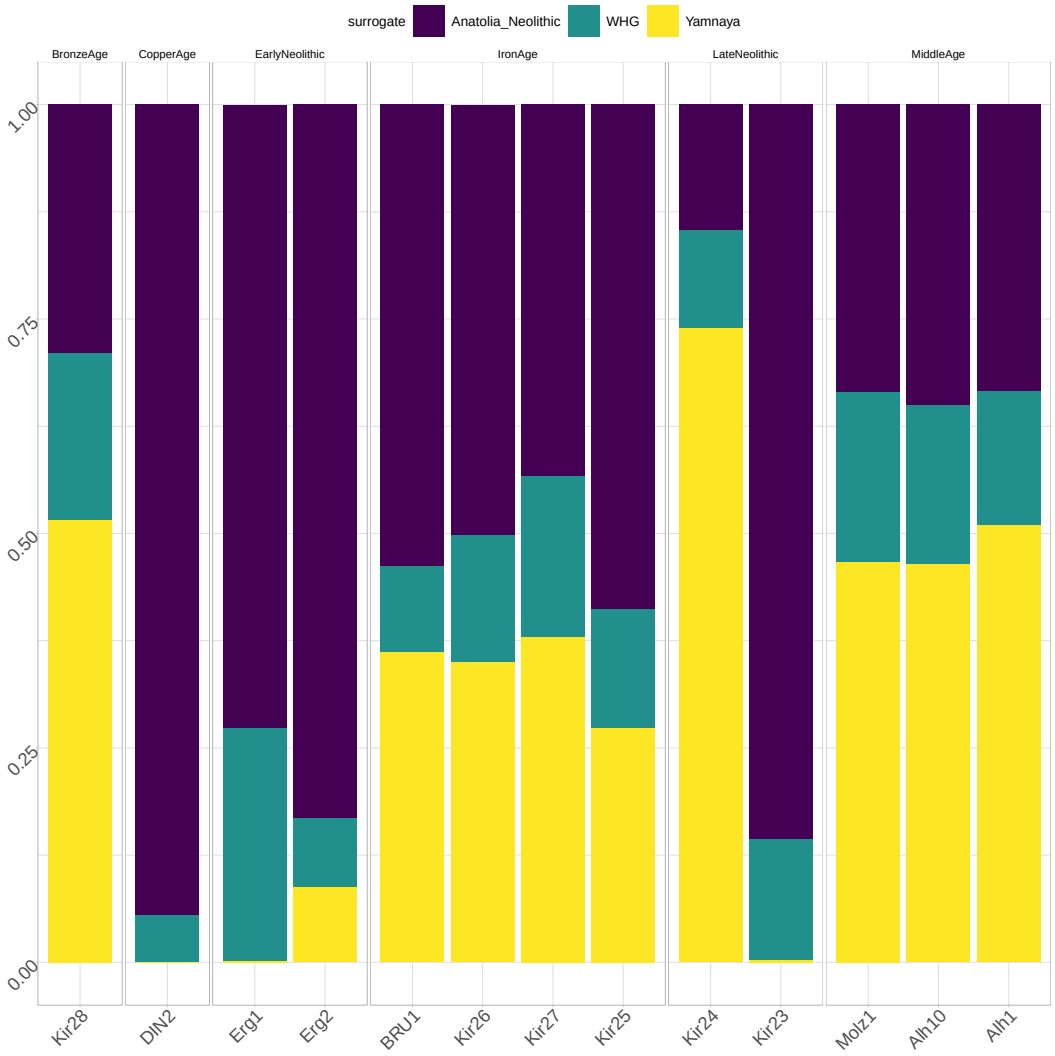


Figure 4.9: SOURCEFIND ancestry proportion estimates for all newly sequenced target samples (vertical columns). Target samples are grouped by archaeological age. Surrogate populations are represented as horizontal rows and also grouped into archaeological culture. Each target was modeled as a mixture 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.

(approx 4700 BP). Despite their temporal and spatial closeness, they show highly different ancestry profiles (Fig. 4.9).

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 [40]. These results support the findings of Allentoft (2015), who concluded that the Afanasievo 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 [84]. 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 [85]. 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 Bronze Age

The single Bronze Age individual, Kir28, is dated to approximately 4000BC. Kir28 is found in a fineSTRUCTURE group with other Central Europe Bronze Age samples; RISE150, a sample from the contemporaneous Bronze Age Unetice culture and Rathlin1, a sample from early Bronze Age Ireland. Thus, this sample is typical

qpAdm modeling showed that the Kir28 can be modeled as a mixture of the two previous Late Neolithic samples in roughly equal proportions, or a mixture of Bell Beakers and Irish Neolithic populations.

4.3.6 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 [86]. 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) [82]. 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 [87]. 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) [82], I found no evidence of East-Asian or East-Asian-like admixture (Fig. 4.6).

4.3.7 Modern day legacy of the Altheim and Molzbichl samples

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 individuals 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 4.10). 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).

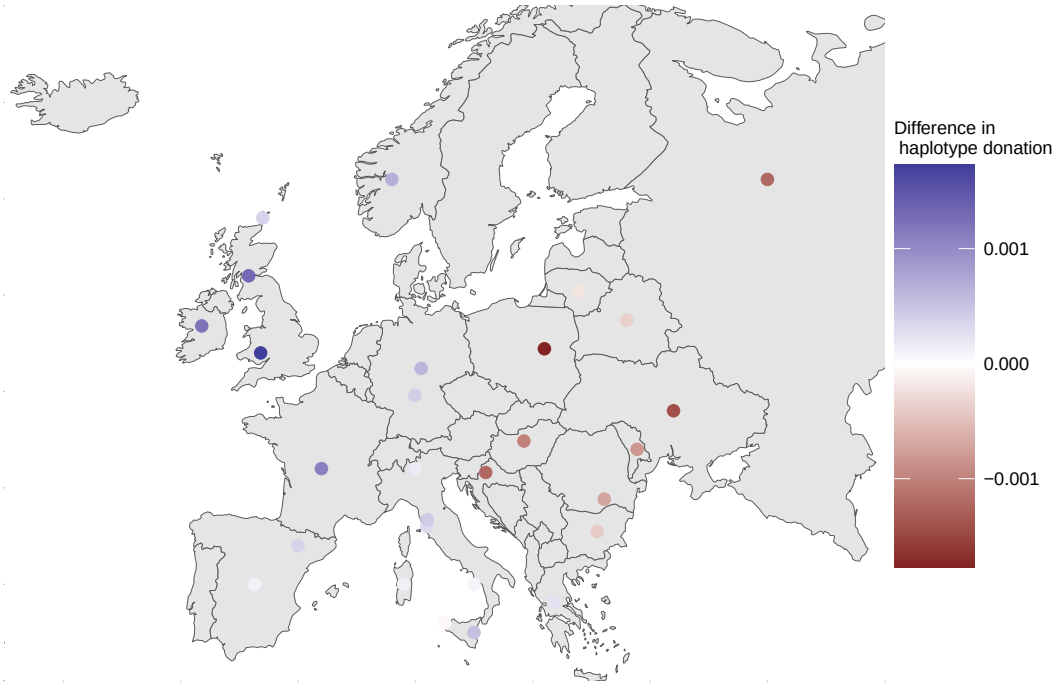


Figure 4.10: Differential haplotype-donation between Germanic and Slavic samples. Each coloured point is one present-day population. Points are coloured based on whether they donate relatively more to Germanic (blue) or Slavic (red) ancient samples.

4.3.8 Sample heterozygosity and homozygosity

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

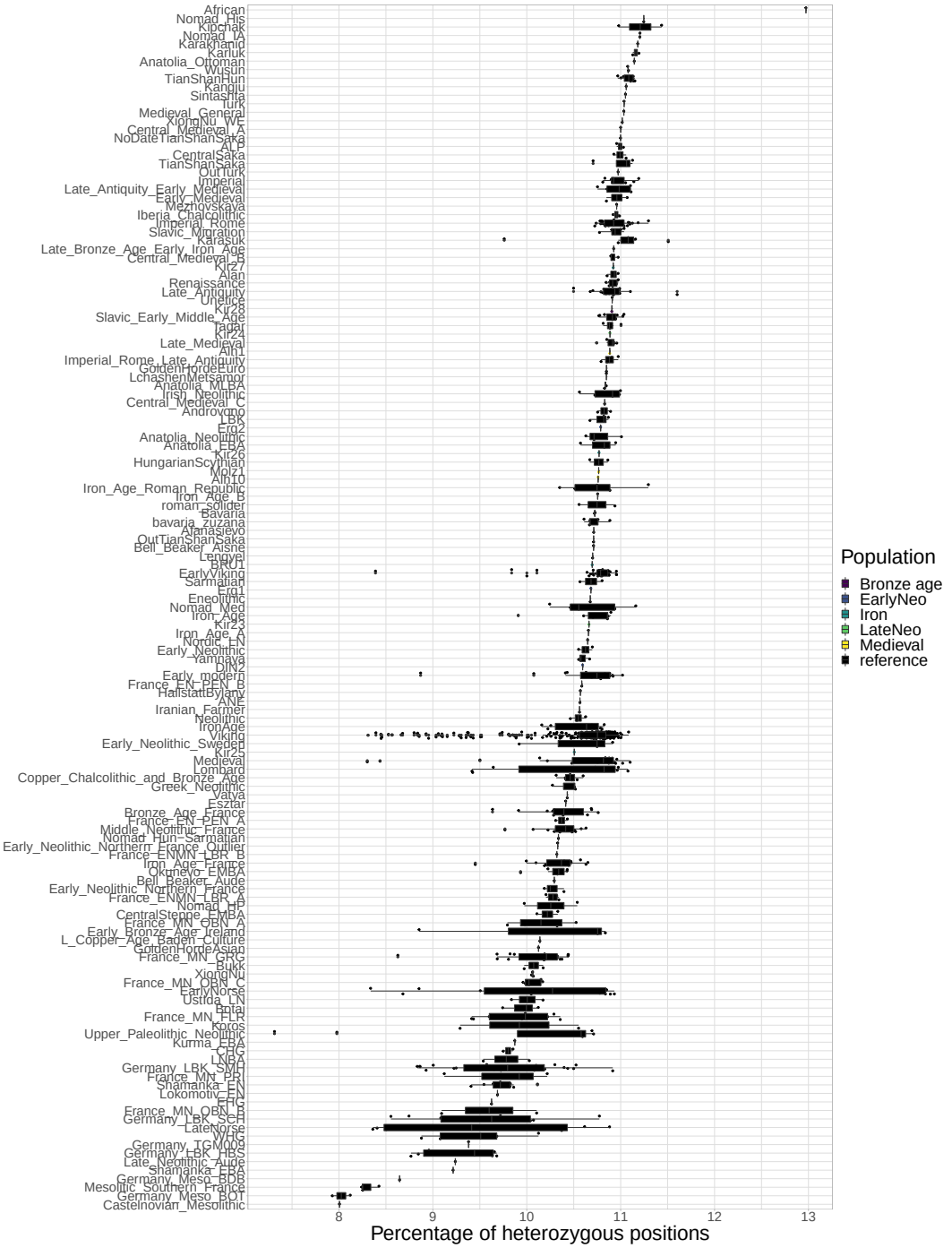


Figure 4.11

4.3.9 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 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 [88] and whose origin, although disputed, is thought to be Polesia (a marshy forested area straddling Poland, Belarus, Russiana and Ukraine) [89]. 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 [90]. 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 [91], 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 [88, 92–94]. It has been proposed that these migrations were key to forming the foundations of present-day Slavic (speaking) nations [88].

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) [95].

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 been relatively limited in their scope. Juras et al (2014) used uniparental

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

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 [96]. 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 [97] (although see [98]). For example, it is know that mtDNA and nuclear DNA may have different evolutionary histories and thus display discordant phylogenetic trees [99].

Kushniarevich et al (2015) [100] 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) [101] 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 [10, 71, 102–104]. 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 [105] 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) [10] 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 [106, 107]. 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, 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 12 samples are from the later Early Middle Ages (724 AD - 995 AD).

Apart from the age of the samples, the Migration Period and Early Middle Age samples can be differentiated by the style of pottery found in the burial grounds (Z. Hofmanová, personal communication).

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.

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 [24] 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 [34], described in appendix A.5.

GLIMPSE was chosen as it is the most efficient (in terms of time and memory) probabilistic phasing / imputation software available [24].

First, GLIMPSE chunk was used to split the present-day reference dataset

into chunks. Default settings of –window-size 2000000 and –buffer-size 200000 were used, generating a total of 936 regions genome wide. Splitting the genome into regions for imputation jointly maximises computational efficiency and accuracy. For each region in turn, the target dataset consisting of phred-scaled genotype likelihoods (PL) was imputed using GLIMPSE phase under default settings and the same reference panel. GLIMPSE ligate was then used to concatenate the 936 imputed regions into 22 distinct chromosomes. Finally, GLIMPSE sample was used to created phased haplotypes from the output of GLIMPSE ligate using default settings.

Previous sections (cite section) have shown that using a set of SNPs which are common yields results which are least affected by coverage related bias, likely in part because common variants are easier to impute. Therefore, I chose to filter the SNPs to those 477,416 those present in the HellBus dataset (A.20). Thus these SNPs were used for all subsequent analyses.

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 [30].

Appendix A.20 describes the initial filtering that was used to generate this

dataset. It was then phased using shapeit4 [30] 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 [13], implemented in Admixr R library [108] 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 [17]. 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 [9] 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.

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

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

The chunklengths / chunkcounts matrices outputted by ChromoPainter give informative but noisy estimates of the proportion of the genome a given individual most closely matches to another individual. However, these cannot be seen as true admixture fractions, due to phenomena such as incomplete lineage sorting. In order to infer ancestry proportions from the data, I ran the SOURCEFIND algorithm [11] on each cluster of individuals inferred from fineSTRUCTURE.

Each of the 47 clusters inferred by fineSTRUCTURE was analysed in turn, using the other 46 clusters to act as surrogates. 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 [39]. First, I ensured all chains had converged using the `autocorr.diag` function. I then used the `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 2 different (but similar) haplotype-based methods; MOSAIC [71] and GLOBETROTTER [10].

I performed 2 different kinds of admixture modeling. Firstly, I performed an 'ancient surrogates' model where the high coverage ancient samples described in the previous section were used as surrogates. I used the fineSTRUCTURE groupings to assign the samples. The results for these analysis were noisy and difficult to interpret, perhaps for several reasons, outlined in the discussion section.

I then performed a 'present-day surrogates' analysis where a selected set of

present-day populations were used to analyse both present-day Slavic populations and ancient Slavic populations. Using these samples provided cleaner results, at the cost of reduced interpretability.

Therefore, I decided to only use the present-day surrogate set unless otherwise stated. Phased .vcf files were converted to .hap/.sample files using `bcftools convert -hapsample`. The resulting .hap/.sample files were then converted to MOSAIC input using the provided script (`https://maths.ucd.ie/~mst/MOSAIC/convert_from_haps.R`).

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

To understand the ancestry of the newly sequenced ancient samples in the context of other ancient individuals, I performed a Principle Component Analysis (PCA) using plink2 (Fig. 5.2). This 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). The Early Middle Age samples are more clustered together, with all samples occupying the broad region containing European Iron Age samples. However, some samples, notably 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

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

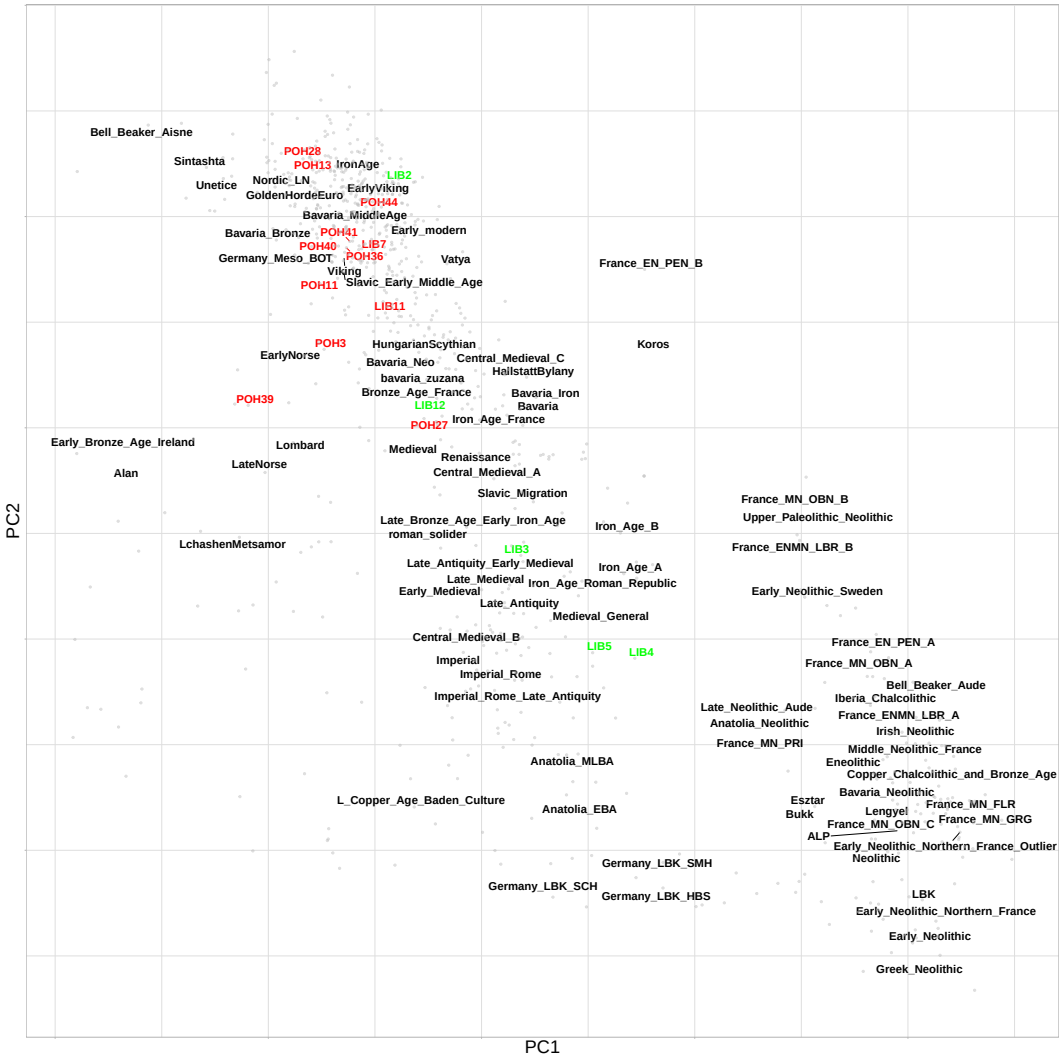


Figure 5.2: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using the plink2. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations. The position of each reference label is the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

assign them to genetically homogeneous groups.

The migration period consisted of whole-genomes from 5 individuals from Břeclav (Líbivá), Czech Republic, from 5 different burial sites, who had radiocarbon dates corresponding 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 [109,110]. 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, 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

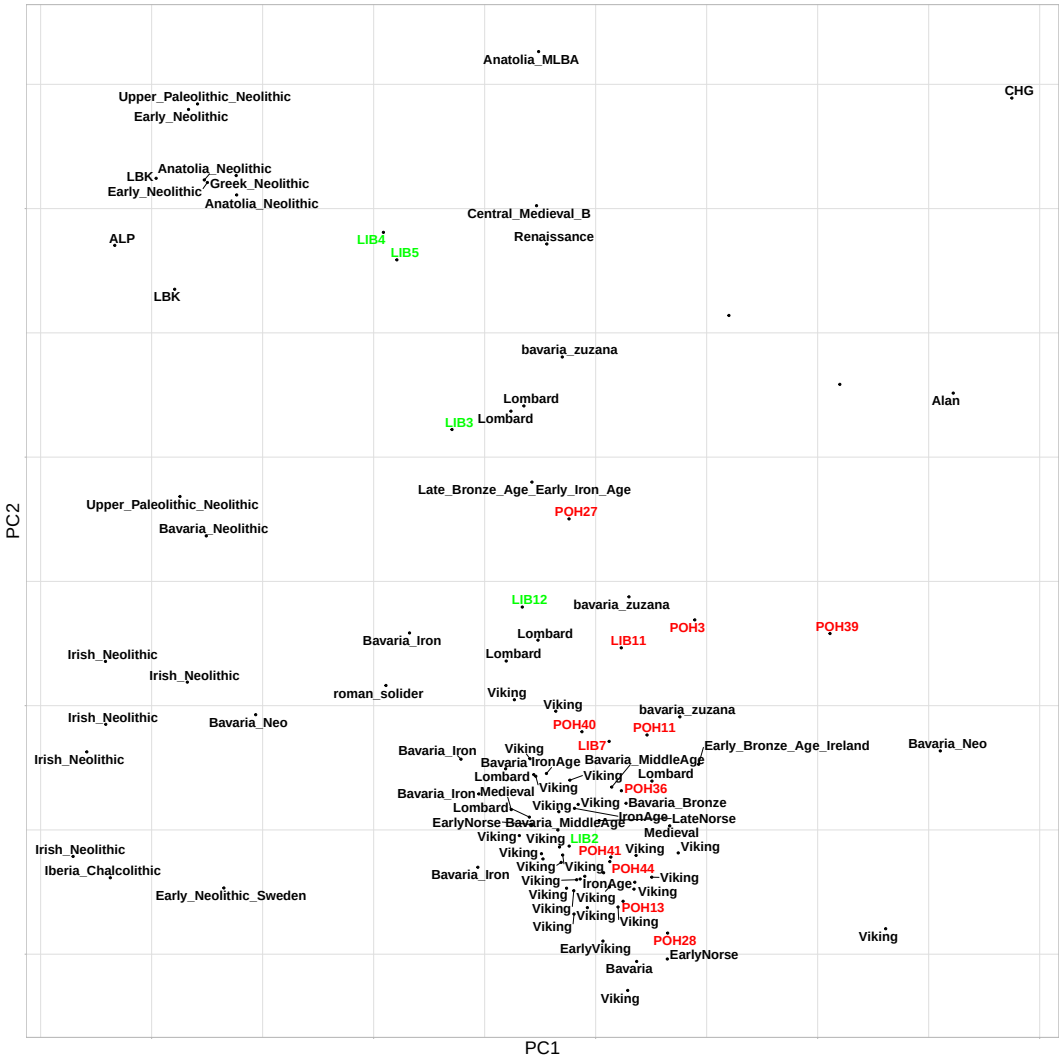


Figure 5.3: Principle component plot of newly sequenced ancient samples and reference ancient individuals performed using fineSTRUCTURE. Green labels correspond to Migration Era samples, red labels to Early Middle Age samples and black as reference populations.

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 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 deep 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) [111], where it was discovered 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 these samples 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 [112]. 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 Antiquty / Iron Age migrations to the region.

5.3.2 Early Middle Age Slavs

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 found 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' represented by Lombard 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 even 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 [10, 71]. 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 [42], ,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}

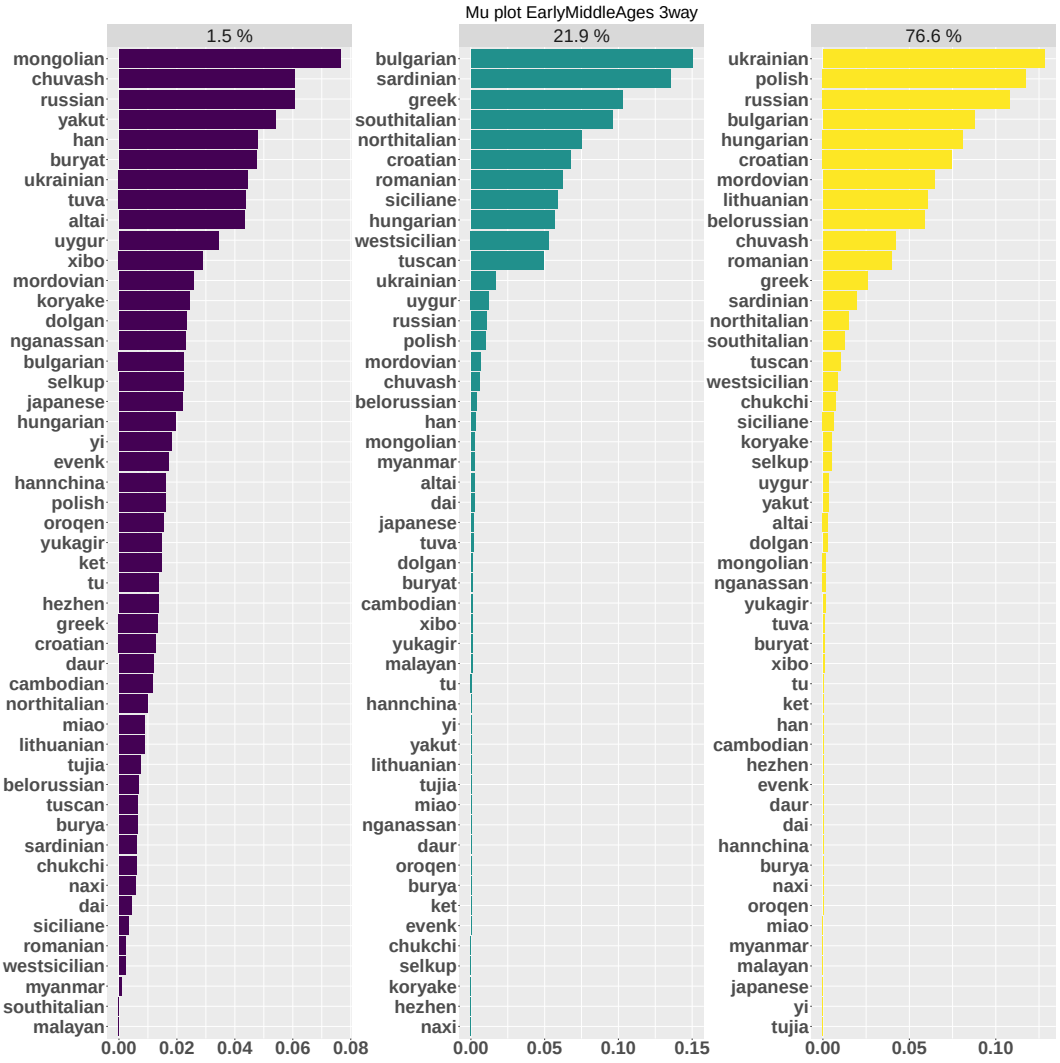


Figure 5.4: Copying matrix plot for sources in 3-way admixture event for Early Middle Age ancient Slavic samples. Each panel represents one of the 3 putative mixing sources. Labels above each panel gives the proportion that mixing source contributed to the Early Middle Age samples. Length of the bars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

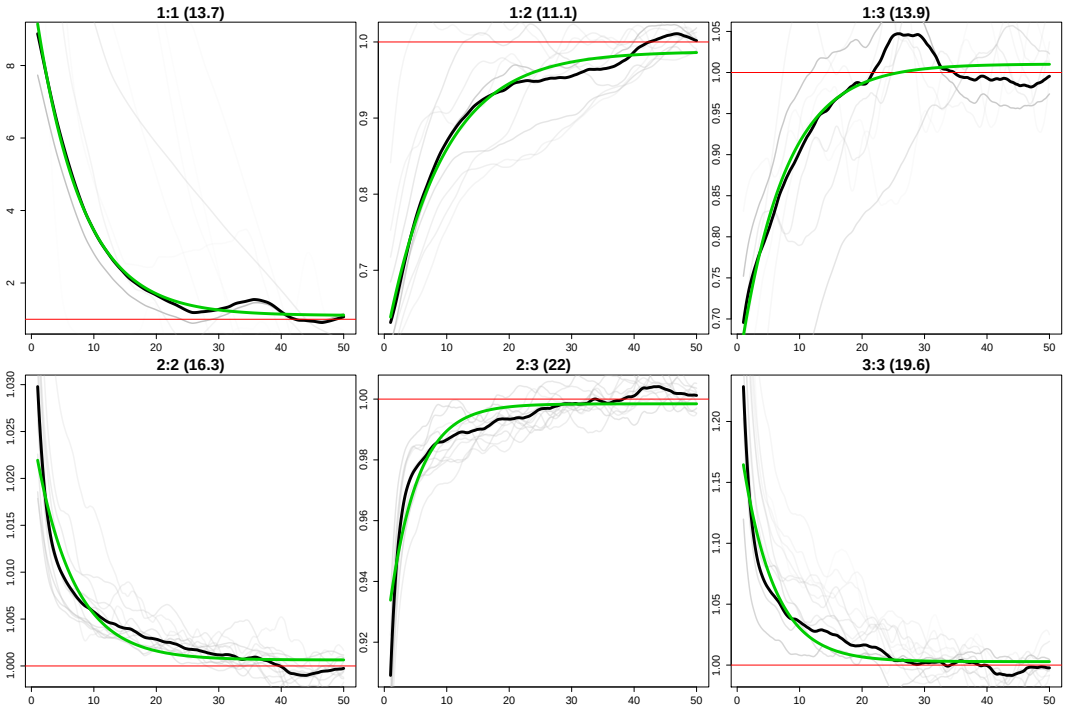


Figure 5.5: Inferred Coancestry Curves obtained from modeling Early Middle Age samples as a 3-way mixture of present-day individuals. Black lines are empirical coancestry curves across all target individuals, light grey are per individual, green is the fitted single-event coancestry curve. Note to self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

and C_{ema} across samples within each groups. Then, I calculated the empirical 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 ($p = 0.0013$).

This is in agreement with the fineSTRUCTURE results which grouped together all of the Early Middle Age samples together to the exclusion of the Migration Period samples.

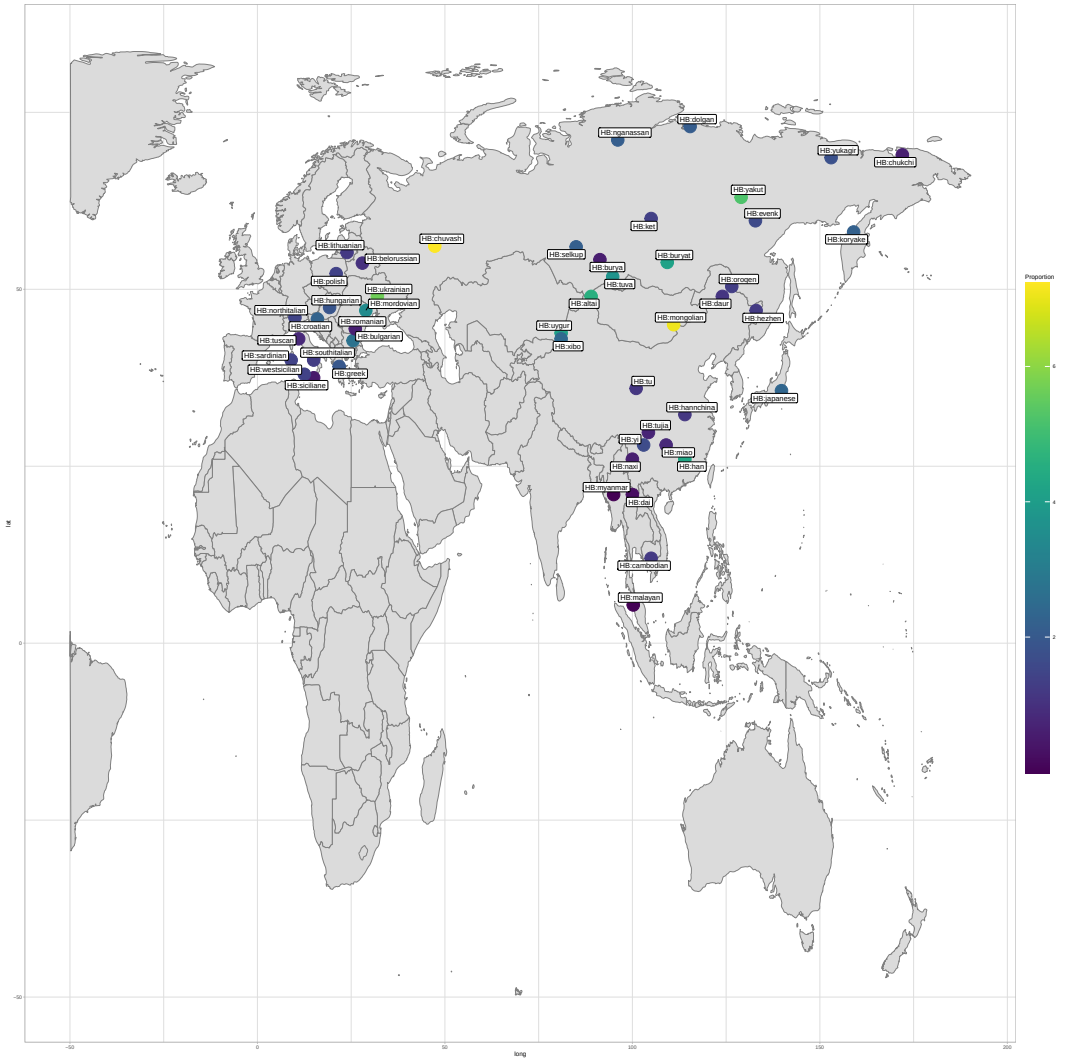


Figure 5.6: Distribution of East-Asian minor ancestry component in Early Middle Age samples.

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.

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.

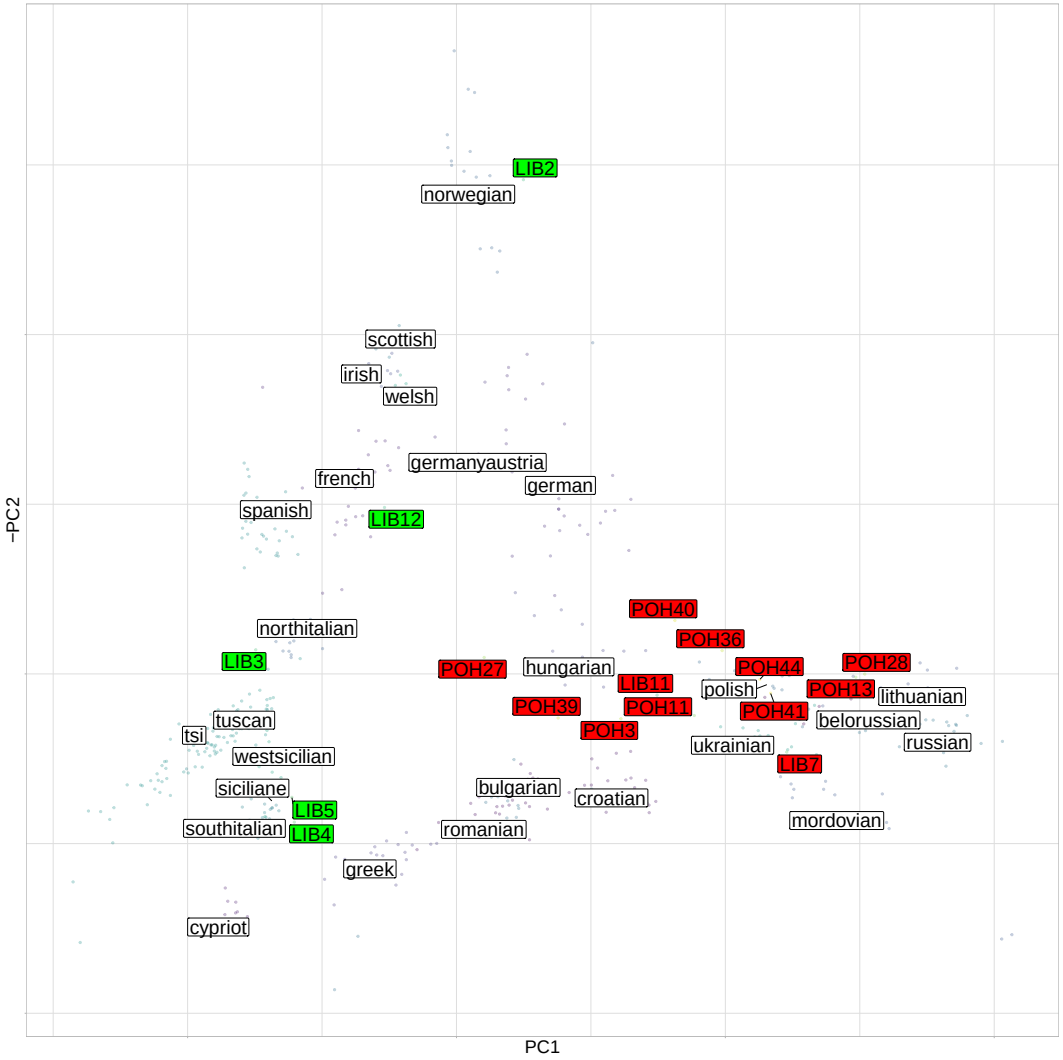


Figure 5.7: Principle component plot of newly sequenced ancient samples and reference modern individuals performed using the finestructure library. Green labels correspond to Migration Era samples, red labels correspond to Early Middle Age samples and white labels correspond to reference populations. 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 [113].

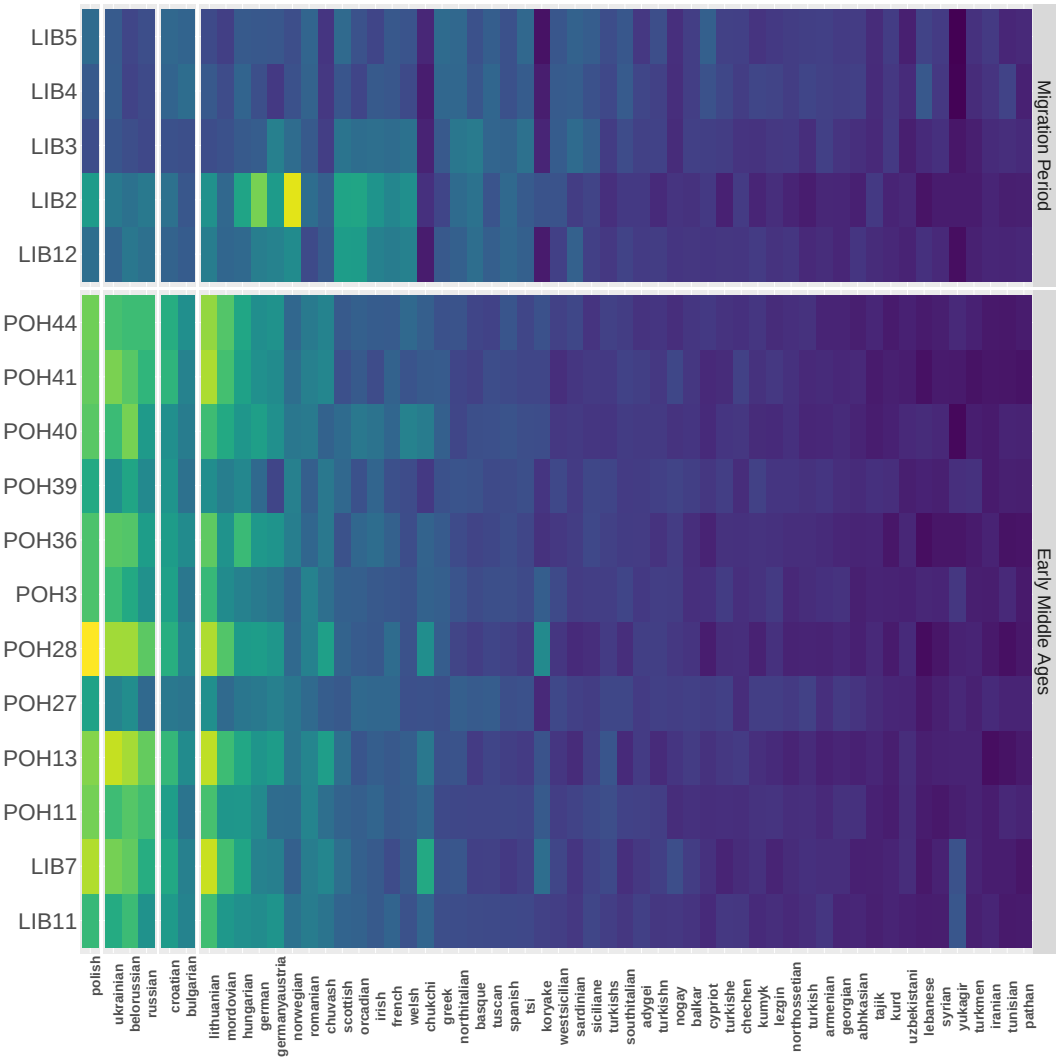


Figure 5.8: Raw chunklengths matrix from the 'present-day' painting. Rows correspond to different ancient recipient individuals, grouped into Migration Period and Early Middle Age period, and columns to different donor populations. Colour of cells corresponds to the total length of genome that a given 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).

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*

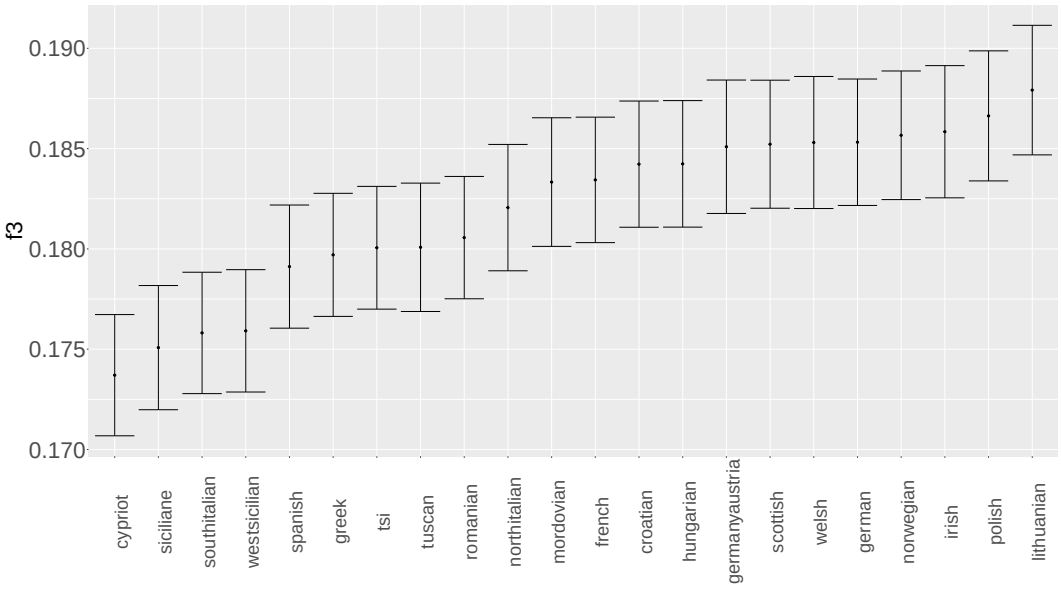


Figure 5.9: f_3 statistics in the form of $f_3(EMA, present{-}day; mbutipygmy)$, where *present-day* is different present-day European population. Error bars rerpesent $\pm * 2$ standard error.

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

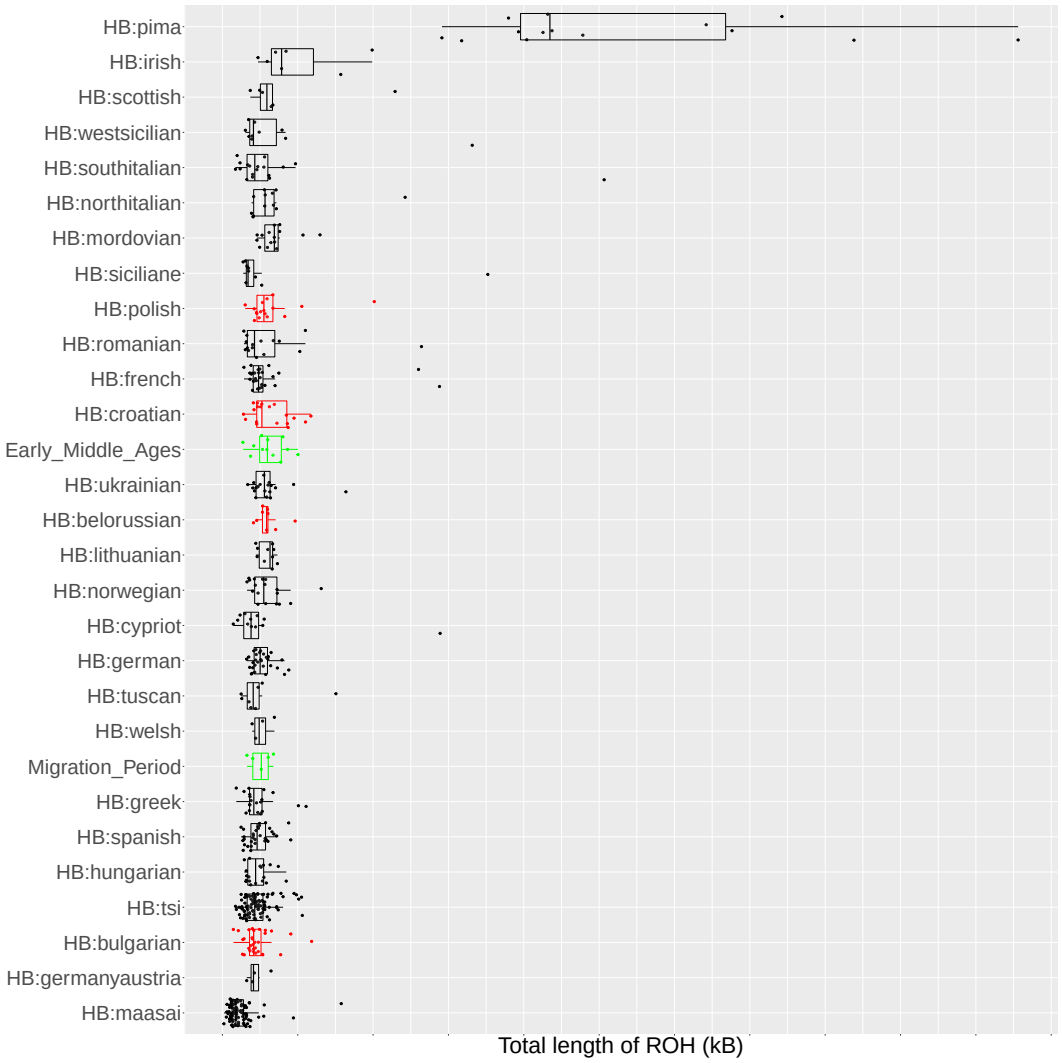


Figure 5.10: Total length of runs-of-homozygosity (ROH) in different present-day and ancient populations. Each point is the total length of ROH (kB) within an individual in that population. Points given jitter to aid visualisation. HB:pima and HB:masasai included to display extremes of ROH in different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

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 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.

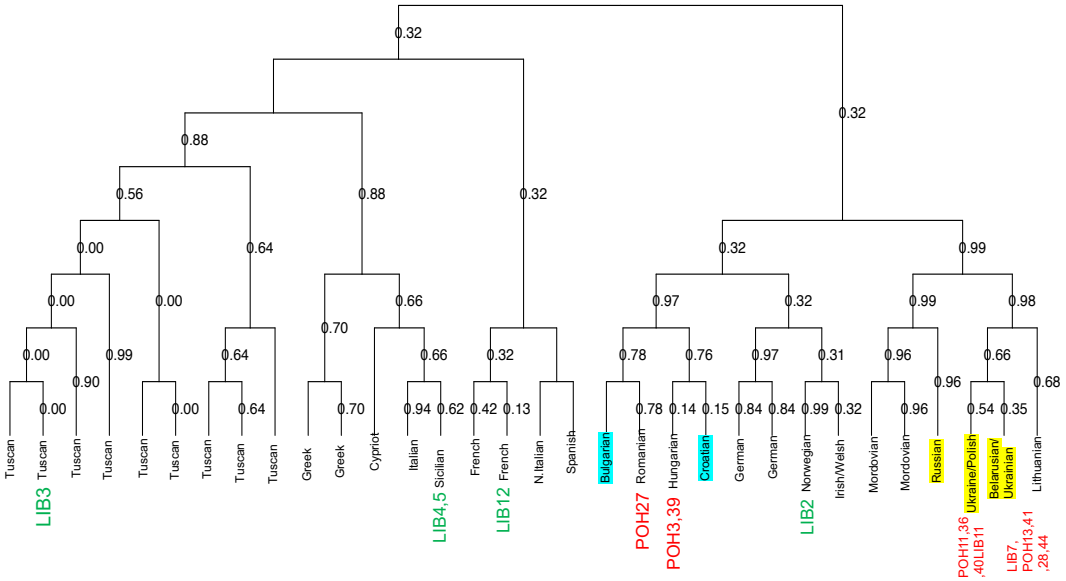


Figure 5.11: Population dendrogram generated by the fineSTRUCTURE tree building algorithm. Labeled tips refer to the primary population(s) represented in that clade. present-day non-Slavic populations shown in black. 'South-East' Slavs highlighted in cyan and 'North-West' Slavs highlighted in yellow. 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 [71, 114]. 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 [71] 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.

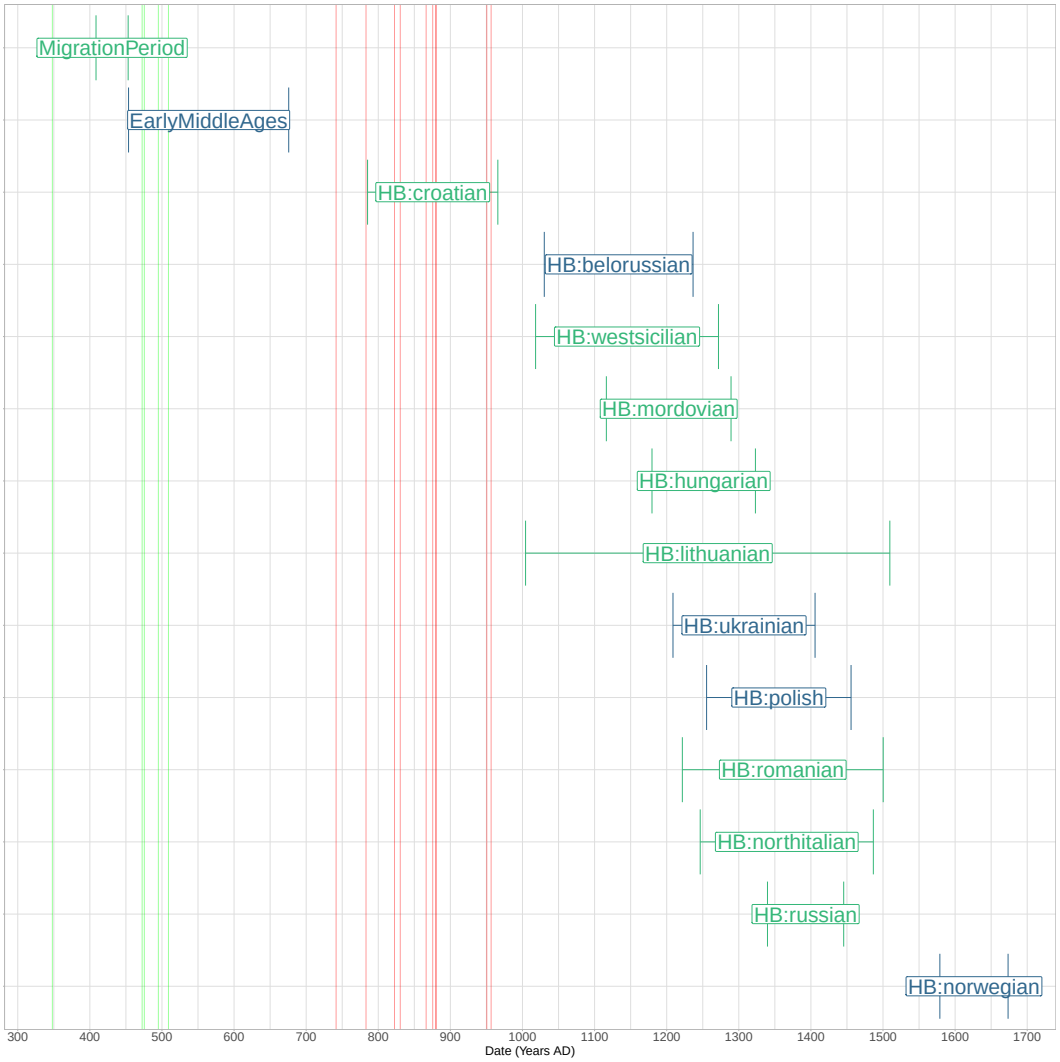


Figure 5.12: MOSAIC inferred 2-way admixture dates with bootstrapped 97.5% and 2.5% CI. Vertical green lines correspond to radiocarbon estimated dates of Migration Period samples and red lines equivalent for Early Middle Age samples. Estimated dates obtained by assuming an average generation time of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

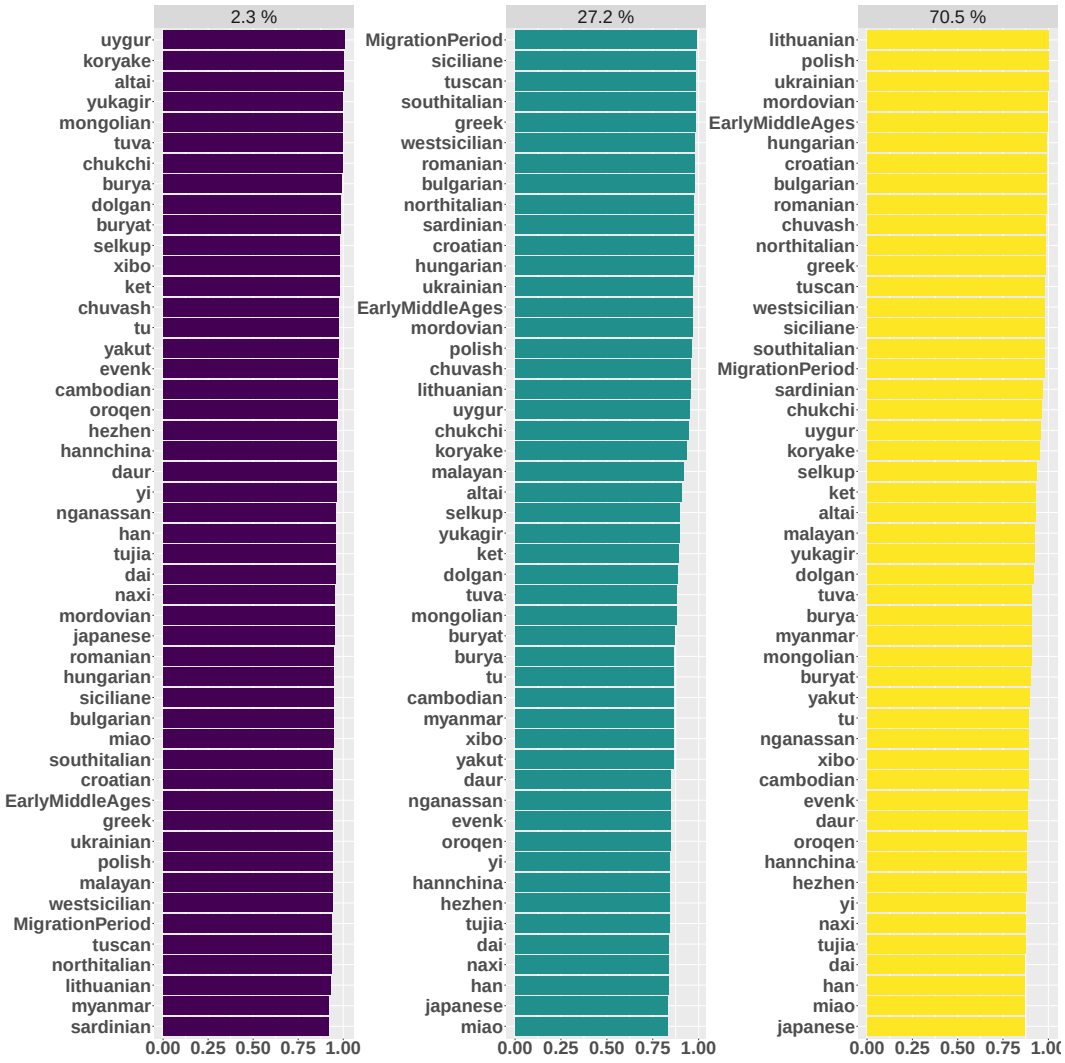


Figure 5.13: $1 - F_{st}$ between 3 inferred mixing sources for present-day Belorussians. Each panel represent a different mixing source. Each bar gives the value $1 - F_{st}$ between that samples population and the mixing source. Higher values of $1 - F_{st}$ suggest that source is well represented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

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 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) [100] 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 [115]. 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

A.1 Antonio et al 2019

Samples from Antonio et al (2019) [19]. Looking at the population genomics of ancient Rome.

This dataset consists of 134 shotgun-sequenced individuals from 12,000 years ago to the present day. Coverage ranges from 0.4x - 4x (median 1.05x).

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB32566.

A.2 Margaryan et al 2020

Samples from Margaryan et al 2020 [20], looking at the population genomics of Vikings.

This dataset consists of 442 shotgun-sequenced individuals from the Bronze Age to Medieval and Early Modern ages. Mean coverage of 1x. X individuals

did not have enough data to estimate recalibration parameters and were excluded from the rest of the analysis.

Aligned reads in .bam format were downloaded from the ENA under accression number PRJEB37976.

A.3 Rivollat et al 2020

Samples from Rivollat et al (2020) [73]. Looking at the population genomics of France and Germany.

This dataset consists of 101 captured individuals from 7000–3000 BCE. years ago to the present day.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB36208.

A.4 Brunel et al 2018

Samples from Brunel et al (2018) [86]. Looking at the population genomics of France.

This dataset consists of 52 shotgun sequenced individuals from Mesolithic to the Iron Age.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB36208.

A.5 Allentoft et al 2015

Samples from Allentoft et al 2015 [116]. Looking at the population genomics Bronze Age Eurasia.

This dataset consists of 20 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB9021.

A.6 Broushaki et al 2016

Samples from Broushaki et al 2016 [117]. Looking at the population genomics Bronze Age Eurasia.

This dataset consists of 1 shotgun sequenced individual.

Aligned reads in .bam format were downloaded from the ENA under accession number xxxx.

A.7 Cassidy et al 2016

Samples from Cassidy et al 2016 [118]. Looking at the population genomics of the Neolithic and Bronze Age migrations to Ireland.

This dataset consists of 4 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB11995.

A.8 de Barros Damgaard et al 2018a

The first horse herders and the impact of early Bronze Age steppe expansions into Asia.

Samples from de Barros Damgaard et al 2018a [25].

This dataset consists of 34 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB26349.

A.9 de Barros Damgaard et al 2018b

137 ancient human genomes from across the Eurasian steppes

Samples from de Barros Damgaard et al 2018b [119].

This dataset consists of 58 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB20658.

A.10 Gamba et al 2014

Genome flux and stasis in a five millennium transect of European prehistory

Samples from Gamba et al 2014 [82].

This dataset consists of 10 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the Sequence Read Archive (SRA) under the accession code SRP039766.

A.11 Gunther et al 2015

Genomic diversity and admixture differs for Stone-Age Scandinavian foragers and farmers

Samples from Gunther et al 2015 [120].

This dataset consists of 2 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under acces-

sion number PRJEB9783.

A.12 Hofmanová et al 2016

Early farmers from across Europe directly descended from Neolithic Aegeans

Samples from Hofmanová et al 2016 [75].

This dataset consists of 5 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB11848.

A.13 Jones et al 2015

Upper Palaeolithic genomes reveal deep roots of modern Eurasians

Samples from Jones et al 2015 [121].

This dataset consists of 2 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB11364.

A.14 Marchi et al 2020

The mixed genetic origin of the first farmers of Europe

Samples from Marchi et al 2020 [122].

This dataset consists of 4 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJEB11364.

A.15 Olalde et al 2014

Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European

Samples from Olalde et al 2014 [123].

This dataset consists of 1 shotgun sequenced individual.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJNA230689.

A.16 Sánchez-Quinto et al 2019

Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society

Samples from Sánchez-Quinto et al 2019 [124].

This dataset consists of 7 shotgun sequenced individuals.

Aligned reads in .bam format were downloaded from the ENA accession number PRJNA230689.

A.17 Seguin-Orlando et al 2014

Genomic structure in Europeans dating back at least 36,200 years

Samples from Seguin-Orlando et al 2014 [125].

This dataset consists of 1 shotgun sequenced individual.

Aligned reads in .bam format were downloaded from the ENA under accession number PRJNA230689.

A.18 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` [28] 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 [23] (`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 [29]. 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.19 30x 1000 genomes dataset

Samples from [34].

This dataset consists of 3,202 modern individuals from 26 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 [28]
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 [68]
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.

A.20 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.

A.20.1 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` [30], 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.21 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) [126] (10299 individuals from 15 pops), Leslie et al 2015 [42] (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 [30] setting `-pbwt-depth 8`.
3. Remove duplicated SNPs using bcftools norm [68]
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.

Appendix B

Another Appendix About Things

Some terms that are helpful to define

- linked
- unlinked
- all-v-all

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()`.

The final version of the thesis can be found at `https://github.com/sahwa/thesis`.

Appendix D

Supplementary figures

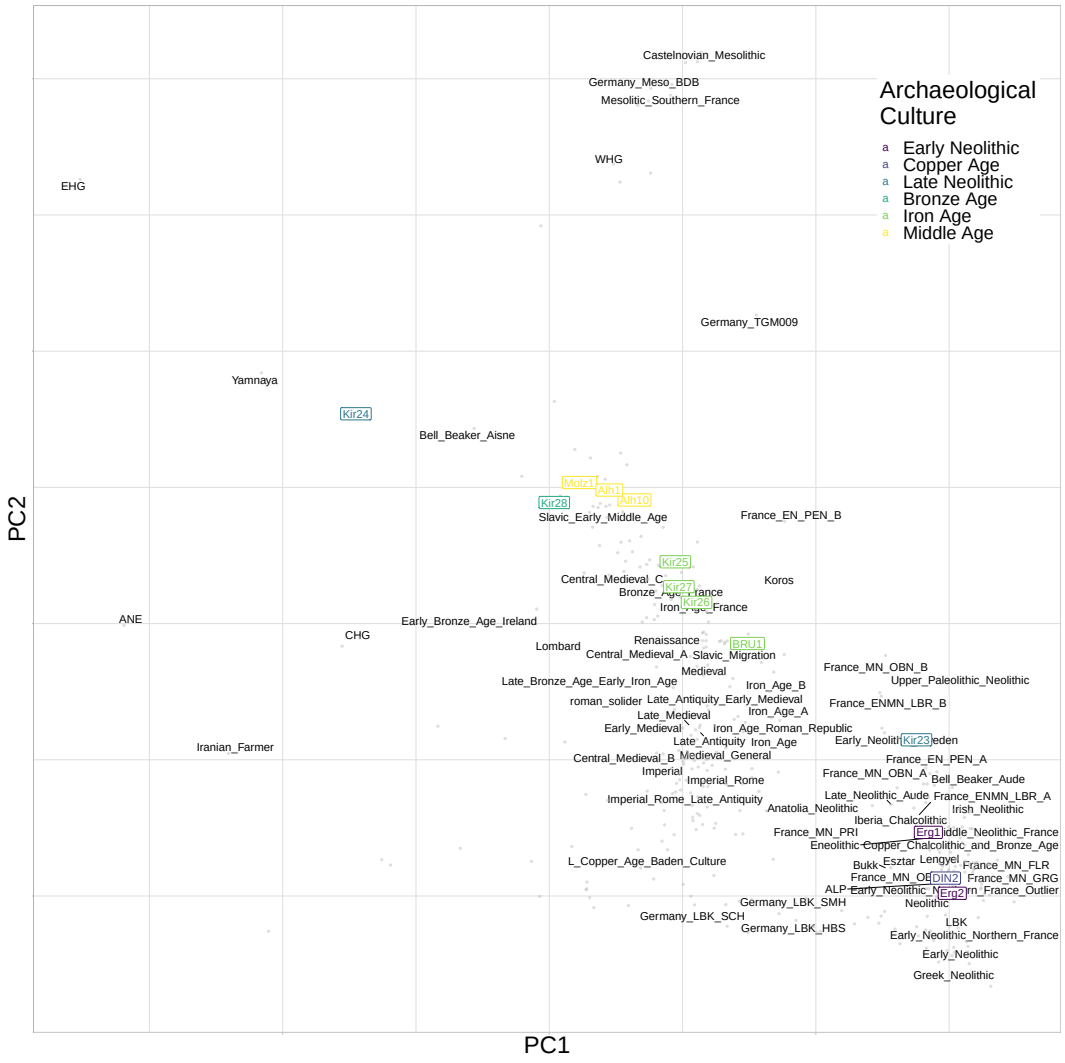


Figure D.1: Principle component analysis of genotype matrix using plink2. Grey points indicate principle component coordiantes for each sample. Black text indicated mean principle component coordinates for all individuals within that group. Coloured labels represent newly sequenced ancient samples.

Bibliography

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

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

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

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

[5] 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.

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

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

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

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

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

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] James Baglama and Lothar Reichel. Augmented implicitly restarted lanczos bidiagonalization methods. *SIAM Journal on Scientific Computing*, 27(1):19–42, 2005.

[38] Iain Mathieson and Aylwyn Scally. What is ancestry? *PLoS Genetics*, 16(3):e1008624, 2020.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

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

[51] 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.

[52] 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.

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

[54] 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.

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

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

[57] 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.

[58] 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.

[59] 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.

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

[61] 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.

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

[63] 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.

[64] 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 associ-

ation and population-based linkage analyses. *The American journal of human genetics*, 81(3):559–575, 2007.

[65] 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.

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

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

[83] 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.

[84] 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.

[85] 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.

[86] 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.

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

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

[89] 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.

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

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

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

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

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

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

[96] 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.

[97] 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.

[98] 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.

[99] 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.

[100] 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.

[101] 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.

[102] 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.

[103] 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.

[104] 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.

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

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

[107] 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.

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

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

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

[111] 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.

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

[113] 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.

[114] 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.

[115] 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.

[116] 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.

[117] 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.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

[123] 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.

[124] 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.

[125] 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.

[126] 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, Vittorio 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.