Chapter 1

ChromoPainter and ancient DNA

1.1 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 [1], phasing and genotype imputation with GLIMPSE [2], ChromoPainter [3] analysis (copy-vector estimation and PCA) and SOURCEFIND ancestry component estimation [4]. I will also describe some of the existing issues pertaining to low coverage ancient DNA and several considered mitigation strategies. Finally, I will simulate, using present-day samples, ancient samples with variable degrees of missing SNPs in order to determine whether ancient samples of a particular covererage have enough typed SNPs to retain haplotype information.

1.2 Methods

1.2.1 Description of the ChromoPainter algorithm

ChromoPainter is a method designed to infer patterns of haplotype sharing between individuals [3]. In diploid organisms such as humans and dogs, ignoring copy-number-variation, each genetic region of an individual is represented by two haplotypes. As input, ChromoPainter requires each individual's data to be phased into these two haplotypes, which refers to the

process of determining which alleles along a chromosome were inherited together from the same parent. Sampled individuals are split into 'donor' and 'recipient' haplotypes, and ChromoPainter employs the widely-used Li and Stephens copying model [5] to model each recipient haplotype as a mosaic of haplotypes observed in the donor panel. Typically (and throughout this thesis) an individual does not act as a donor to themself, e.g. one of the individual's two haplotypes can not act as a donor for the other haplotype. Unlike the original Li and Stephens model, which uses the product of approximate conditionals (PAC) likelihoods, ChromoPainter reconstructs each recipient haplotype as a mosaic of *all* other donor haplotypes. Here, the term 'copying' can be though of as a genealogical process where haplotypes are reconstructed using the genealogically closest haplotype. The copying model is implemented in the form of a Hidden Markov Model (HMM), with the observed states being the genotype data, and the hidden states being the 'nearest-neighbor' haplotype the recipient haplotype copies from. The emission probabilities are given as the probability of a recipient haplotype copying from a particular donor haplotype, given their respective genotypes. Consider a donor d and recipient r, carrying alleles x and y, respectively, at position (e.g. SNP) p. There are two possibilities - either the alleles match between the donor and recipient at p, or they do not. The probability of r copying from d is:

$$\Pr(r = x \mid d = y) = [(1 - \theta) * z_{dr}] + [\theta * z_{!dr}], \tag{1.1}$$

where $z_{dr} = 1$ if $x = y$ and $z_{!dr} = 0$ if $x \neq y$, and θ is the mutation probability.The mutation probability θ can be estimated using Watterson's estimator [6], or estimated using an iterative EM algorithm. Begin with an estimate of θ, usually Watterson's estimate, and at each iteration, replace the value of θ with:

$$\theta^* = \frac{\sum_{l=1}^{L}(\sum_{i=1}^{j} \alpha_{il}\beta_{il}I_{[h_{*l} \neq h_{il}}/P(D))}{L} \tag{1.2}$$

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

In ChromoPainterV2, the input genetic data comes in the form of genotype calls (i.e. 1/0, A/T/C/G). ChromoPainterV2 produces several different output files. The two which most used in this work are those appended with .chunklengths and .chunkcounts. In the chunklengths matrix, cl, the entry $cl_{d,r}$ gives the total expected proportion of haplotype

segments (defined as a contiguous set of SNPs copied from a single donor) that recipient r copies from donor d. Thus, higher values of $cl_{d,r}$ indicate that recipient r and donor d share more recent ancestry.

In this work, 'copyvector' is used to refer to the vector of chunklengths that a single recipient individuals copies from all donors. Throughout, I often define donors as populations, so that each element of the copy vector is the total amount of DNA that the recipient matches to all individuals from a given donor population.

1.2.1.1 Description of ChromoPainterV2Uncertainty

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

Consider the following example: we have a phased genotype in the form `0|1`, corresponding to the reference allele on the first haplotype and the alternative allele at the second haplotype. I define G as the sum of the genotypes at a SNP; in this case $G = 0 + 1 = 1$. As GLIMPSE provides hard genotype calls, G can be calculated directly.

We also have a posterior genotype likelihood, in the form $GL(p_0, p_1, p_2)$, where p_i is the posterior genotype probabilities of being genotype i. Dosage, D, is the expected total number of copies of the alternate allele given GL. D can be calculated as $p_1 + [2 * p_2]$. We can calculate U, the uncertainty as $U = |G - D|$. Then, we can assign a probability to each allele; if the is `1` then the allele likelihood is simply $1 - U$ and if the allele is `0` then the allele likelihood is $0 + U$.

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

$$\begin{aligned} p(r_x|d_x) = (1-\theta) * [r_x * d_x + (1-r_x) * (1-d_x)] \\ +\theta * [r_x * (1-d_x) + (1-r_x) * d_x] \end{aligned} \tag{1.3}$$

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

in the calls.

1.2.2 Generation of downsampled genomes

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

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

1. Yamnaya – Yamnaya Bronze Age steppe-pastoralist [7]
2. UstIshim – Siberian Upper Paleolithic hunter-gatherer [8]
3. sf12 – Scandinavian Hunter-Gatherer [9]
4. LBK – early European farmer from the Linearbandkeramik culture from Stuttgart, Germany [10]
5. Loschbour – 8,000 year-old hunter-gatherer from Luxembourg) [10]

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

Each original full-coverage `.bam` file was processed using the atlas (version 1.0, commit f612f28) pipeline [1]
(`https://bitbucket.org/wegmannlab/atlas/wiki/Home`). First, the validity (i.e. ensuring that each `.bam` file was not malformed in any way) using ValidateSamFile command from PicardTools [11]. atlas is a suite of software designed for processing low-coverage ancient DNA and was chosen following the reccomendation of Hui et al (2020) [12], as it explicitly accounts for post-mortem damage (PMD) patterns in ancient DNA. The most common form of PMD is C-deamination, which leads to a C->T transition on the affected strand and a G–>A transition on the complimentary strand.

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

For each full coverage and downsampled `.bam` file, I estimated post-mortem damage (PMD) patterns using the `atlas estimatePMD` task. Recalibration parameters were then estimated using the atlas `atlas recal` task. Finally, both the recalibration and PMD parameters were given to the `atlas callNEW` task which produces genotype calls and genotype likelihood estimates for each downsampled and full coverage `.bam`. For this stage, I made calls at the 77,818,345 genome-wide positions present in the phase 3 thousand genomes project [13]. This was done to reduce the risk of calling false-positive (i.e. falsely polymorphic) genotypes in the aDNA samples.

1.2.3 Generation of ancient samples

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

This dataset consists of 124 other ancient samples from the literature given in appendix section **??**. These samples were of variable coverage, ranging from 0.002-72x coverage, and chosen because of their previously reported relevance to understanding past ancestry patterns in European populations like those analysed in chapters 4 and 5. These 918 consist of all samples from appendices A1, A2, A3, A4, and they were processed in an identical way to the downsampled target individuals described in the previous section, other than they were not downsampled.

1.2.4 Imputation and phasing - GLIMPSE

Genotype imputation and phasing are two important steps for processing low-coverage ancient DNA. Low coverage (<1x) samples typically lack enough read information to make accurate genotype calls at most positions in the genome, often not containing any reads at several sites [14]. Therefore, it can be helpful to use external information from a high-coverage reference panel in order to improve the accuracy of genotype calls and phasing, and reduce the impact of errors on downstream analyses [2].

Three different characteristics are desirable for an imputation algorithm in this context. Firstly, it should take genotype likelihoods as input. This is because genotype likelihoods allow for flexible representation of the possible genotypes at a particular position, particularly when there may not be enough coverage to make a hard genotype call. Secondly, it should emit posterior genotype-probabilities which, when accurately calibrated, give the probability that a particular genotype call is correct. This is crucial for estimating uncertainty values, described in section 1.2.11, for including these genotype probabilities into the painting process. Thirdly, the algorithm must be able to complete in a reasonable running time when using

a large number of samples and high number of SNPs. Using a large number of densely positioned SNPs (e.g. such as the approximately 77 million identified in the 1000 Genomes Project) increases the useful linkage-disequilibrium information between each SNP, and it is well-established that increasing the number of individuals used in imputation/phasing reference panels improves accuracy [2, 15–17].

Two programs, Beagle 4.0 [18] and GLIMPSE [2] 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 [2], so I chose to use this algorithm for the imputation and phasing steps.

Phasing and imputation ideally requires a reference panel of high-coverage present-day individuals. I used the 1000 Genomes Project dataset re-sequenced to 30x average coverage, which contains 3202 individuals from 26 worldwide populations [19]. 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) [20] to act as the samples for GLIMPSE to phase. Here, 'target' refers to the individuals being imputed/phased and 'reference' refers to the reference panel.

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

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

GLIMPSE then imputed each chunk separately, using `GLIMPSE_phase` using the same 1000 genomes dataset as a reference and default settings. This stage both imputes missing

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

1.2.5 Estimating imputation sensitivity and specificity

I used rtg-tools-3.11 [21] and the `vcfeval` task to estimate the sensitivity and specificity of variant discovery in the downsampled individuals. Here, 'baseline' (i.e. the truthset) is defined as the genotype calls in the full coverage individual and the 'calls' as the genotype calls in the downsampled individual. Sensitivity and precision are defined as:

$$sensitivity = \frac{V_{call} - FP}{V_{call}} \tag{1.4}$$

$$precision = \frac{V_{baseline} - FN}{V_{baseline}} \tag{1.5}$$

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

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

1.2.6 ChromoPainter analysis

It is important to understand the effect of sequencing coverage on the accuracy of ChromoPainter copyvector estimation. A 'copyvector', c_r, is a vector of length D, where each entry gives the total length of genome that recipient individual r most closely matches to

each of the D donor individual/populations. I sometimes refer to 'normalised' copyvectors; this simply refers to where each entry of c_r is divided by the sum of all entries, scaling the copyvector to sum to 1.

I painted each downsampled and full coverage ancient individual using a set of 124 ancient individuals, hereafter referred to as the 'standard set', selected because they had a sequencing depth greater than 2x. I compared the copyvectors for the same individual at each level of downsampling. For example, I compared the copyvector of Yamnaya at 0.1x to the copyvector of the same Yamnaya sample at full coverage. A high correspondence, measured by r-squared for example, between the copyvectors of the full coverage and downsampled individual suggests less effect of coverage.

To prepare the data for ChromoPainter, I merged the .vcf containing the posterior genotype likelihoods of i) downsampled, ii) full coverage and iii) 124 ancient samples from the literature together, and did the same for the .vcfs containing the phased haplotypes. I combined the posterior genotype likelihoods with the phased alleles to generate allele likelihoods (described in section 1.2.1.1 in ChromoPainter-uncertainty format, in addition to per-position recombination rate files. This was performed for each chromosome in turn using my own script (`https://github.com/sahwa/vcf_to_ChromoPainter`).

I next used ChromoPainterUncertainty to perform the painting. I assigned the 'standard set' individuals as donors and all downsampled, full coverage and 124 ancient samples downloaded from the literature as recipients. The 124 ancient samples from the literature were included in order so that they can be used a surrogates in later SOURCEFIND analysis.

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

1.2.7 ChromoPainter Principle Component Analysis

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

1.2.8 SOURCEFIND

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

Therefore, to account for these issues when estimating ancestry proportions, it is necessary to run an additional step, SOURCEFIND [4]. Simulations have shown that SOURCEFIND ancestry proportions correspond well to simulated values [4]. The ancestry proportions produced by SOURCEFIND should be interpreted as the proportion of ancestry that each invidual/population shares most recently with each surrogate. This need not necessarily imply an admixture event; for instance, you might expect $France$ to have ancestry recently related to both *Germany* and *Spain* due to isolation-by-distance rather than admixture.

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

To test for the effect of coverage on the proportions estimated by SOURCEFIND, I performed two separate analyses, both using the downsampled and full coverage individuals as targets. The first uses three surrogate populations (Yamnaya, Western Hunter-Gatherer and Anatolia Neolithic Farmer), and the second uses an expanded list of 37 surrogate populations (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 [10, 22]. Note, this does not mean that there was not admixture from other sources, but that a majority of ancestry of ancient central Europeans can be derived from these sources. This stands to act

as a relatively 'easy' test case, since the three populations are highly genetically differentiated from one another.

For all runs of SOURCEFIND, I used 1,000,000 iterations, of which 50,000 were designated as burn-ins, and then samples were taken every 50 iterations. 2,000,000 iterations were chosen because my previous tests show that is the minimum necessary to provide reasonably confidence of convergence within reasonably running time (Appendix D.5). The rest of the parameters were left as default. Ancestry proportions, credible intervals and chain mixing/convergence checking for each surrogate group were estimated using the CODA R library [23].

1.3 Pre-post GLIMPSE and linked/unlinked PCA test

I wanted to determine at what stage of the analysis pipeline low coverage samples (0.1x) because significantly diverged from the other coverage samples when plotted on a PCA. For instance, it may be that the bias is introduced in the imputation stage. To test this, I performed a set four PCAs on all downsampled and equivalent full coverage samples and a set of present-day individuals shown in Table 1.1.

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

1. **Pre-GLIMPSE** Using the genotypes generated by atlas, but before imputation with GLIMPSE, I projected all downsampled ancients of all coverages onto the present-day populations using the eigenstrat library. [24].
2. **Post-GLIMPSE** Using the GLIMPSE generated imputed genotypes generated by atlas, I projected all downsampled ancients of all coverages onto the present-day populations using the eigenstrat library.
3. **ChromoPainter - unlinked** I performed an 'all-v-all' unlinked ChromoPainter painting, using all populations in Table 1.1.
4. **ChromoPainter - linked** I performed an 'all-v-all' unlinked ChromoPainter painting, using all populations in Table 1.1.

Bias present in PCA (2) but not (1) indicates it has been introduced in the imputation stage. Similarly, bias present in (4) but not (3) suggests that inlcuding linkage information introduces bias in low coverage samples.

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

Table 1.1: Population labels and sample sizes of populations included in the Pre-post GLIMPSE and linked/unlinked PCA test. All samples are from the Hellenthal and Busby dataset, described in **??**.

1.4 Reducing SNP count

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

However, reducing the total number and or density of SNPs used in a painting may reduce the accuracy of the estimated copyvectors. All other things being equal, there is less linkage information between two SNPs with are separated by a larger genetic distance. Therefore, it is necessary to precisely determine what effect reducing the number of SNPs has. In particular, we would like to know the minimum number and density of SNPs required to retain the advantages of haplotype-based methods over unlinked methods.

Using data from the People of the British Isles (POBI) project, previous work showed it is possible to distinguish between British individuals from neighboring counties Devon and Cornwall using the fineSTRUCTURE algorithm, but not using unlinked methods (ADMIXTURE [25]) [26]. Therefore, determining whether it is possible to distinguish between individuals from Devon and Cornwall acts as a good test case for reducing SNPs. In

particular I tested how many SNPs can we remove before we lose the ability to distinguish between these two populations.

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

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

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

1.5 Direct imputation test

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

I took the Human Origins dataset (appendix A.19), containing 560,240 bi-allelic SNPs and submitted the reduced dataset to the Sanger Imputation Service (`https://www.sanger.ac.uk/tool/sanger-imputation-service/`). The Sanger Imputation Service uses Eagle2 [27] and the Haplotype Reference Consortium as a reference to impute missing variants. Once the data had been imputed, I subsetted the data back to the original set of 560,240 SNPs. I therefore had a dataset which contained 70,000 non-imputed SNPs and 490,240 imputed SNPs. This is hereafter referred to as the 'imputed dataset'. 70,000 non-imputed SNPs was chosen because that is the number of SNPs which overlap between two datasets in Chapter 3 and thus represents a realistic case-study.

For both the imputed dataset and original Human Origins dataset, I performed an all-v-all painting and combined data across chromosomes. An 'all-v-all' painting is where

each individual is painted in turn by all other individuals, resulting in an n-by-n coancestry matrix, where n is the number of individuals analysed.

1.6 Results

1.6.1 Imputation accuracy

To estimate how accurately GLIMPSE imputes genotypes in ancient samples of differing coverages, I estimated the sensitivity (Fig. 1.1) and precision (Fig. 1.2) of genotype imputation using rtg-tools [21]. This approach compares genotype calls at each position in each downsampled individual after imputation to the same individual at full coverage without imputation.

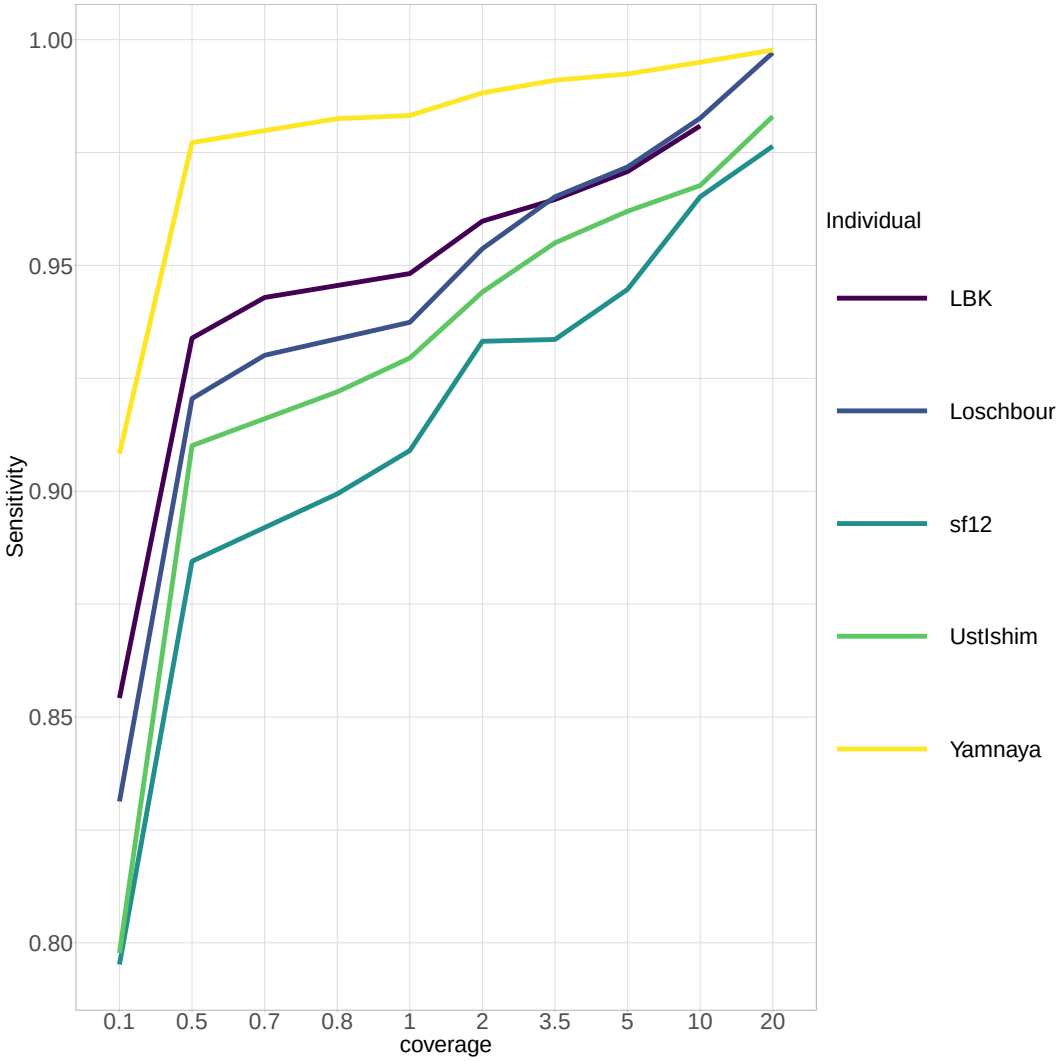

Figure 1.1: Sensitivity of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

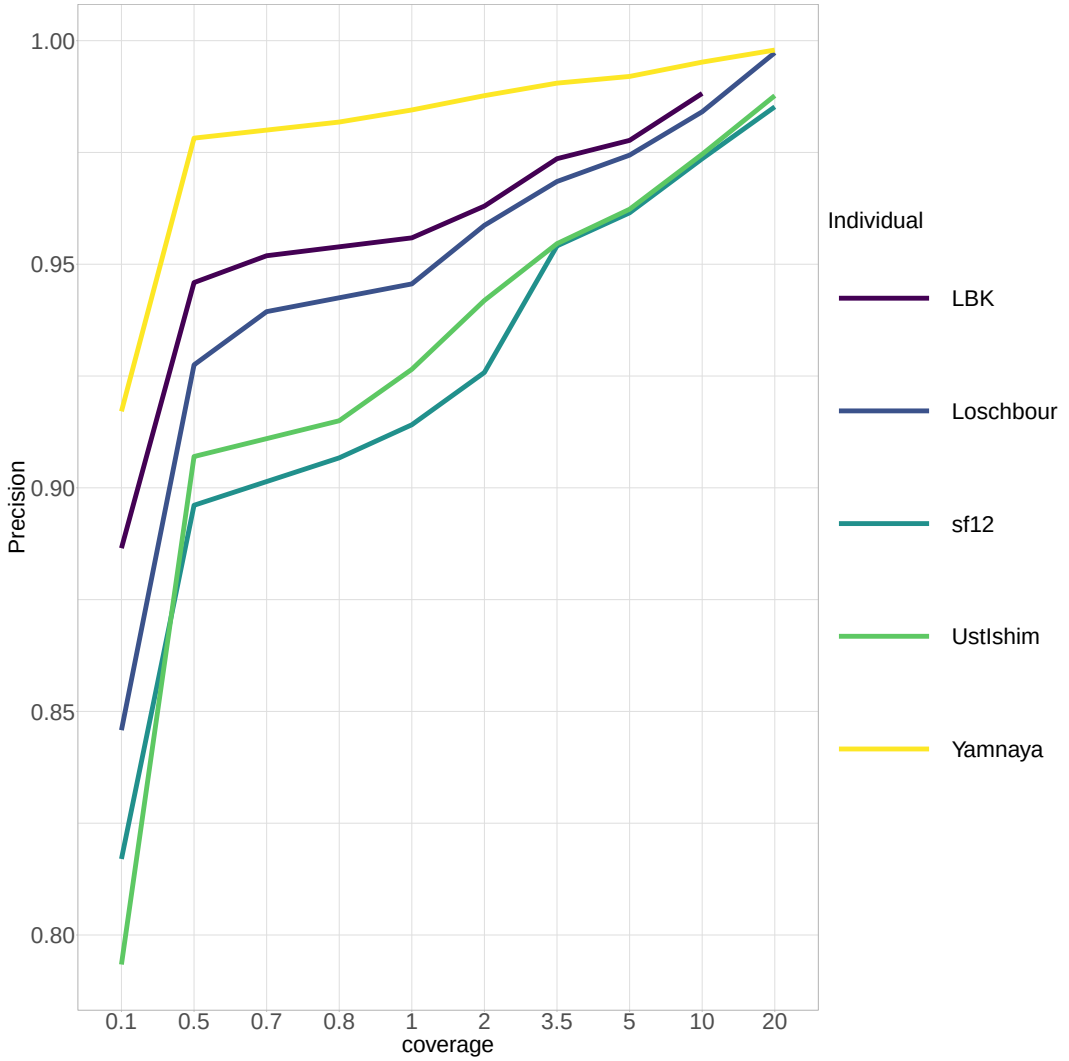

Figure 1.2: Precision of genotype calling at different coverages for different ancient individuals, assuming calls in the full coverage genome are correct, calculated using rtg-tools.

As expected, both the overall sensitivity and precision of imputation fell with coverage, with a particularly sharp drop-off in both metrics between 0.5x and 0.1x coverage. Whilst I did not investigate this, other studies have shown the probability of any one SNP in an sample being correctly imputed depends strongly on the frequency in the reference panel [2, 12]. In particular, alleles which are rare in the reference panel are less likely to be imputed correctly.

Different downsampled individuals differed in the precision and sensitivity of genotype imputation. At all coverages, Yamnaya had the both the highest sensitivity and precision. This may be because the imputation reference panel contains a high proportion of present-day Europeans, who have a relatively higher porportion of recent Yamnaya-like ancestry relative to e.g. Hunter Gatherer-like ancestry [28]. 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 [15, 16]. On the other hand, the sample Ust'Ishim is known to have contributed very little genetic ancestry to present-day populations [29] 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) [30] and fewer heterozygous positions. As imputation relies on matching IBD tracts between individuals, imputation accuracy increases where individuals share more IBD [31]. Additionally, switch-errors during the pre-phasing step of imputation may harm imputation accuracy, so a reduced density of heterozygous positions may result in increased accuracy.

1.6.2 Phasing accuracy

I also used rtg-tools to calculate the number of phased heterozygous genotypes where the downsampled individual has the same phasing as the full coverage individual (Fig 1.3). 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 [2].

1.6.3 Validating posterior probability calibration

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

I split the genome into 10,000 equally-sized bins according to $max(GL)$. For each bin, I calculated both the proportion of SNPs which were correctly imputed (i.e. that matched the same high coverage individual) and the mean $max(GL)$ (Fig. 1.4). 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.

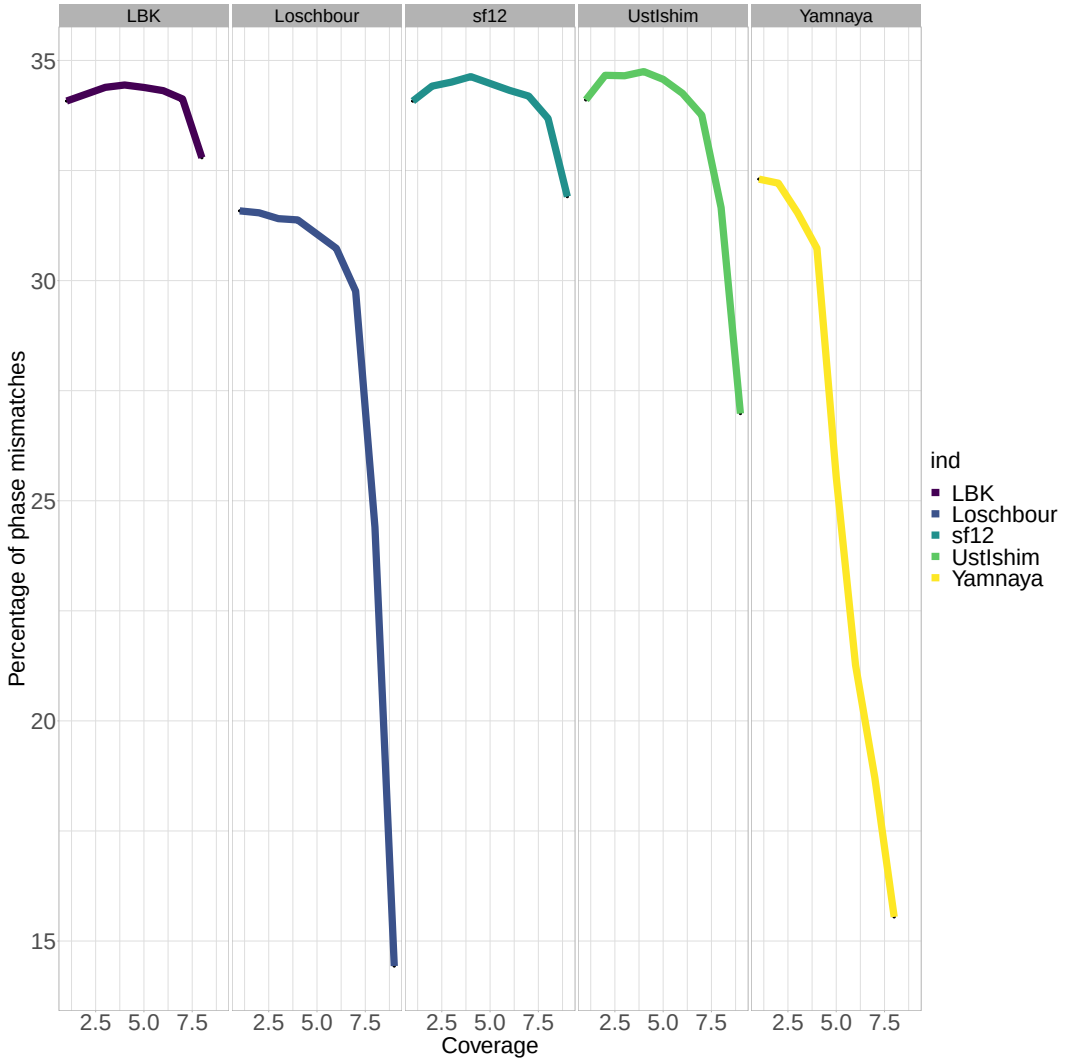

Figure 1.3: Percentage of phased genotypes which agree with the same full-coverage sample? for each individual and each level of downsampling. Genotypes with phase deemed unresolvable by rtg-tools were excluded from the calculations. Note that these numbers are given as incorrect / (incorrect + correct - unresolved) and so values are in part driven by the relative heterozygosity of each sample.

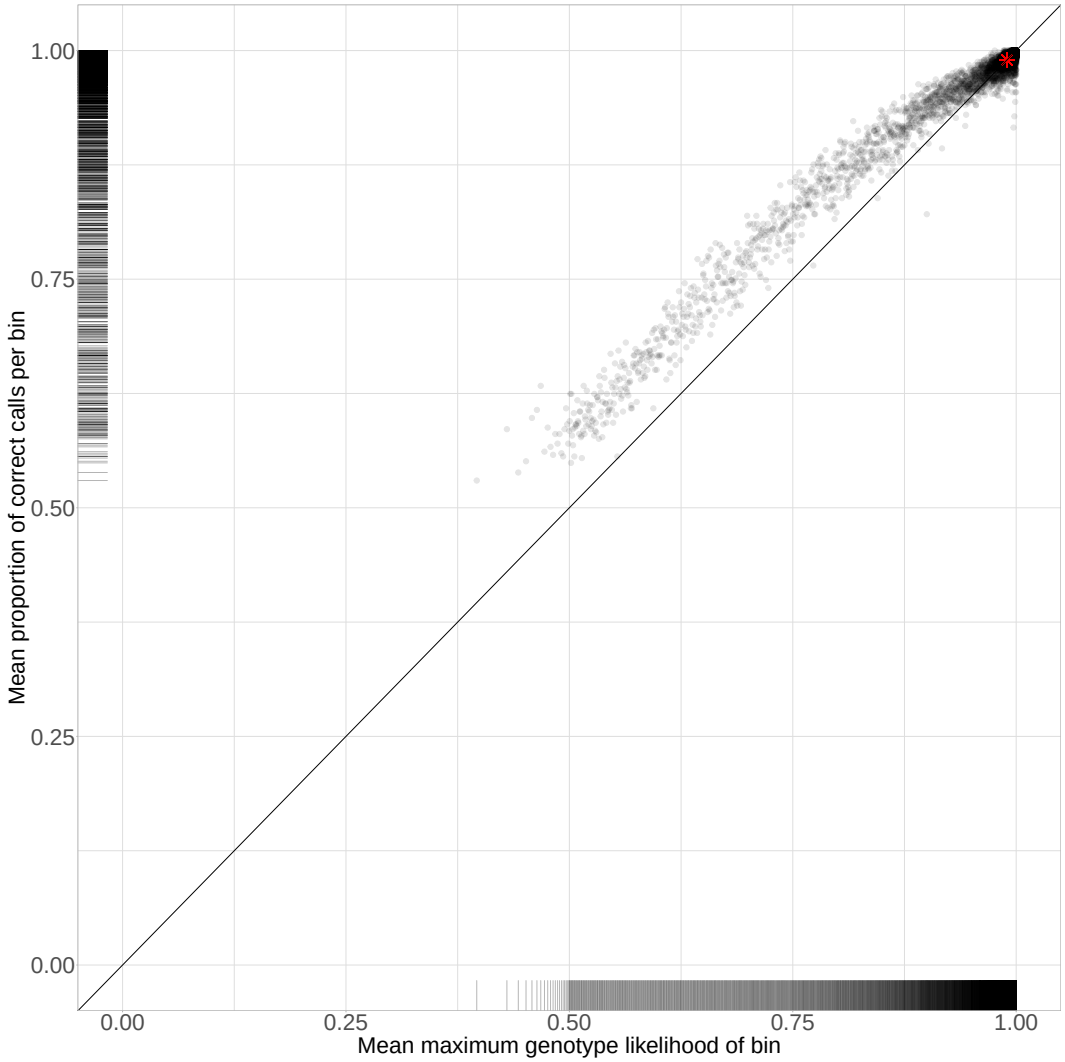

Figure 1.4: Relationship between genotype likelihood and probability of genotype call being correct for Yamnaya downsampled to 0.1x coverage. Genome binned by maximum posterior genotype likelihood and mean maximum posterior genotype likelihood (x-axis) and proportion of correct calls per bin (y-axis). Rugs on each margin show the distribution of x and y values. Black line is $y = x$.

The probabilities are well calibrated (r-squared = 0.981) and could therefore be useful for downstream analysis. It should be noted that they are slightly conservative, in that a majority of the points in Fig. 1.4 are above the line of equality. For example, the mean proportion of correct genotypes within all bins where $0.73 < max(GL) < 0.76$ was 82%. I performed the same analysis using different samples at different levels of coverage and the results were qualitatively similar (result omitted).

1.6.4 ChromoPainter analysis

To assess the impact of coverage on ChromoPainter analysis, I merged the dataset of downsampled individuals with the 'standard set' of ancient reference individuals (124 ancient samples > 2X coverage) and performed an 'all-v-all' painting of the merged dataset, which separately paints each individual as a recipient using all other individuals in the dataset as donors. The 'all-v-all' painting was necessary to paint the 124 'standard set' of individuals against one another so that they can act as surrogates in later SOURCEFIND analysis.

I was interested to see whether a downsampled individual and full coverage had similar copyvectors, or in other words, whether they matched similar amounts to the same donor individuals. To do this, I estimated r-squared between the copyvectors of the full coverage and downsampled individuals.

Fig. 1.5 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 coverage genome. Put in other words, the copyvectors at 0.1x are tending towards becoming more 'flat', or copying the same amount from each donor individual.

This can also be seen as 'regressing to the prior'. In this case, the prior is copying an equal amount to each donor individual. This can be visualised explicitly by calculating TVD between each downsampled genome and a flat prior, a vector of length D, where D is the total number of donor individuals and each element of D is equal to 1 / D (Fig. 1.6). This clearly shows the reduced TVD to the flat copyvector for the 0.1x individual relative to

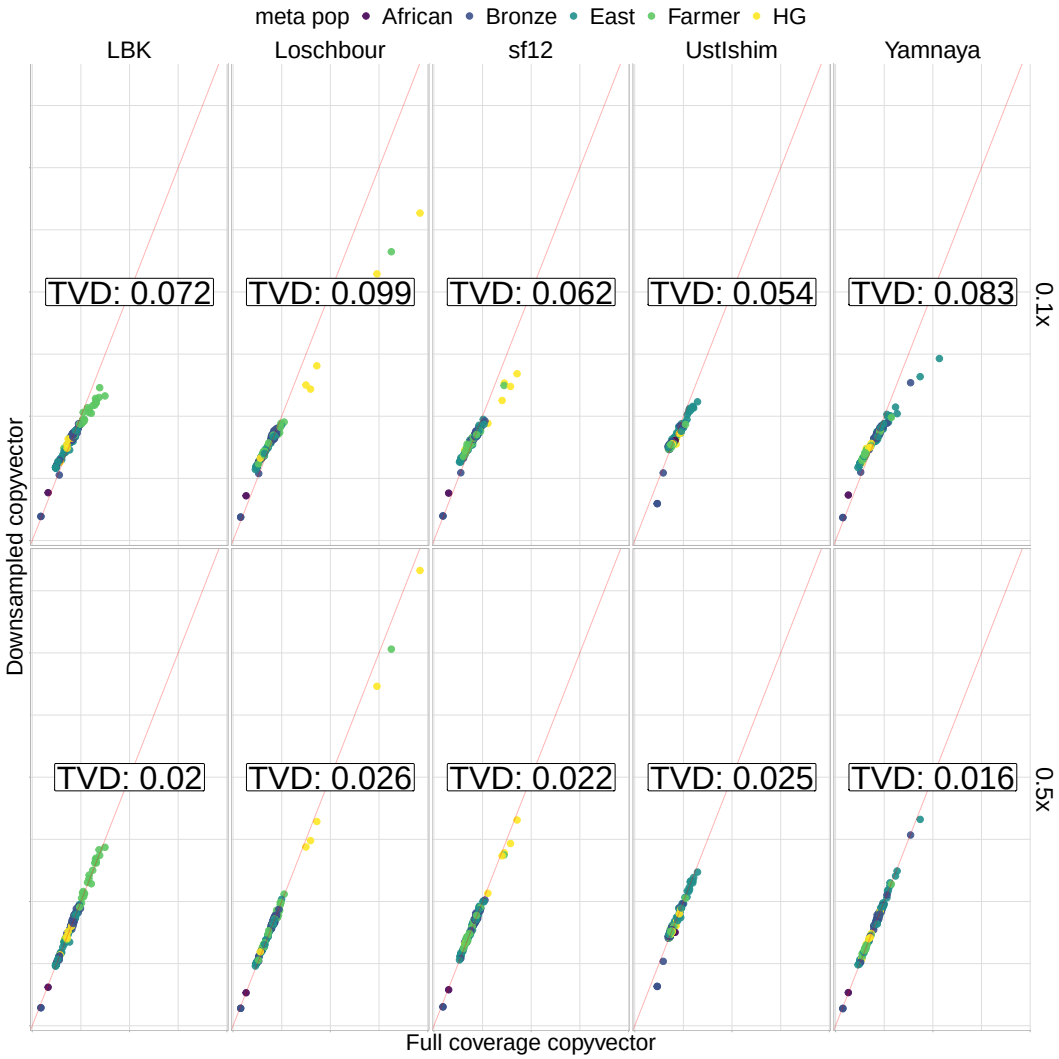

Figure 1.5: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to each of 125 ancient individuals (dots). Results are shown for 0.1x (top row) and 0.5x (bottom row) downsampled genomes. Points coloured by manual assignment to broad-scale populations. Red line is line of equality ($y = x$). x and y units are normalised copying values and thus removed for clarity.

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

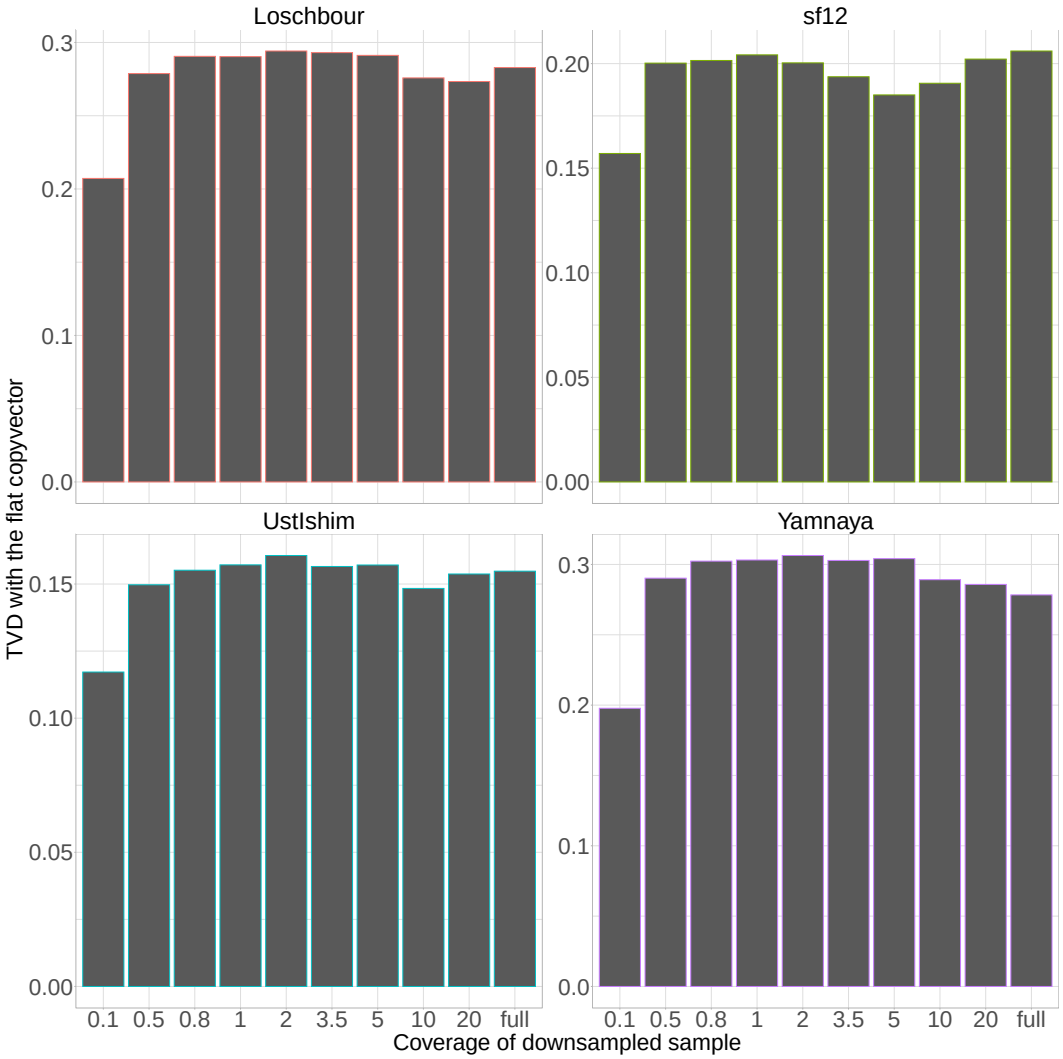

Figure 1.6: TVD (metric of copyvector dissimilarity between two individuals) between each downsampled ancient individual and a flat copyvector. Flat copyvector equivalent to a vector of length N where each element $= 1/N$.

I also considered the effect of coverage on the copyvectors estimated when using present-day individuals from the 1000 genomes project as donors (Fig. 1.7). Painting ancient samples using present-day donors is often useful, particularly with more recent ancient samples, as there may not be enough relevant ancient samples to paint the ancients with. I merged the downsampled and full coverage ancient individuals with the thousand genomes dataset (described in detail in appendix A.5). As was the case with the all-v-all ancients painting, the TVD between copyvectors was highest for the 0.1x individuals. However, the copyvectors show a strong correlation / low TVD for 0.5x individuals.

It should be noted that utility of painting different ancient individuals with a modern reference panel depends on the ancestry and age of the ancient sample. The spread of points along the $y = x$ line in Fig. 1.7 shows how much a particular ancient recipient preferentially copies more from particular modern population over others. LBK, for example, has points which are spread evenly across $y = x$, showing that they copy much more from some populations than others, suggesting modern populations are good for distinguishing this particular ancient sample. On the other hand, the points for Ust'Ishim are shrunk towards lower values of $y = x$, showing that the copyvector is relatively flat and that it does not preferentially copy from some populations to the same degree that LBK does. This is consistent with findings that UstIshim did not contribute ancestry towards present-day populations [8]. Accordingly, relatively less useful information is obtained from painting Ust'Ishim with a modern reference panel than LBK.

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

The position of the full coverage individuals are consistent with prior knowledge about their ancestry (Fig. 1.8). For example, Loschbour is positioned alongside other Hunter Gatherers, who are highly differentiated from the later Neolithic farmers and Bronze Age Europeans. sf12 clusters with the other Scandinavian Hunter Gatherers in the dataset. Yamnaya is differentiated from the group of Bronze Age individuals and situated close to individuals from the Poltavka and Srubnaya culture. LBK is located with other individuals from the early to middle Neolithic in central Europe. Consistent with sharing little ancestry with any group over another, UstIshim is positioned close to the central Bronze Age mass, where most of the individuals in the PCA are located.

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

Taken together, these data suggest minimal effect of coverage down to and including 0.5x mean depth. To my knowledge, no other study has evaluated the effect of coverage

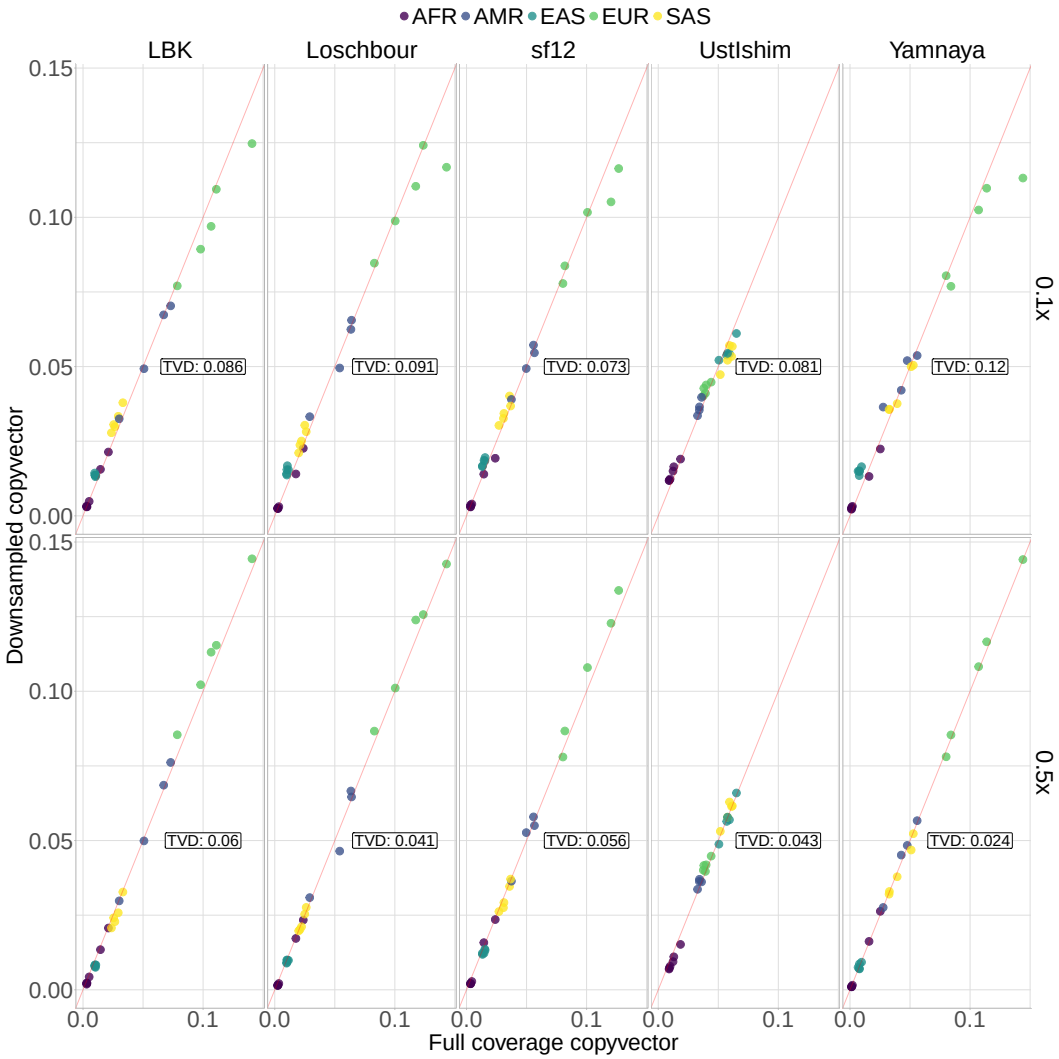

Figure 1.7: For five different samples (columns), the proportion of DNA that each downsampled (y-axis) or full coverage (x-axis) genome matches to individuals from each of 26 present-day populations (dots). Red line is $y = x$. x and y units are normalised copying values and thus removed for clarity.

on ChromoPainter analysis down to a coverage of 0.5x. Margaryan et al (2020) showed a minimal effect of coverage at 1x and that fineSTRUCTURE groupings, containing individuals as low as 0.1x coverage, were not driven by coverage [32].

1.6.5 SOURCEFIND

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

I began by considering three ancestral sources, or 'surrogates', fixed as Anatolia Neolithic,

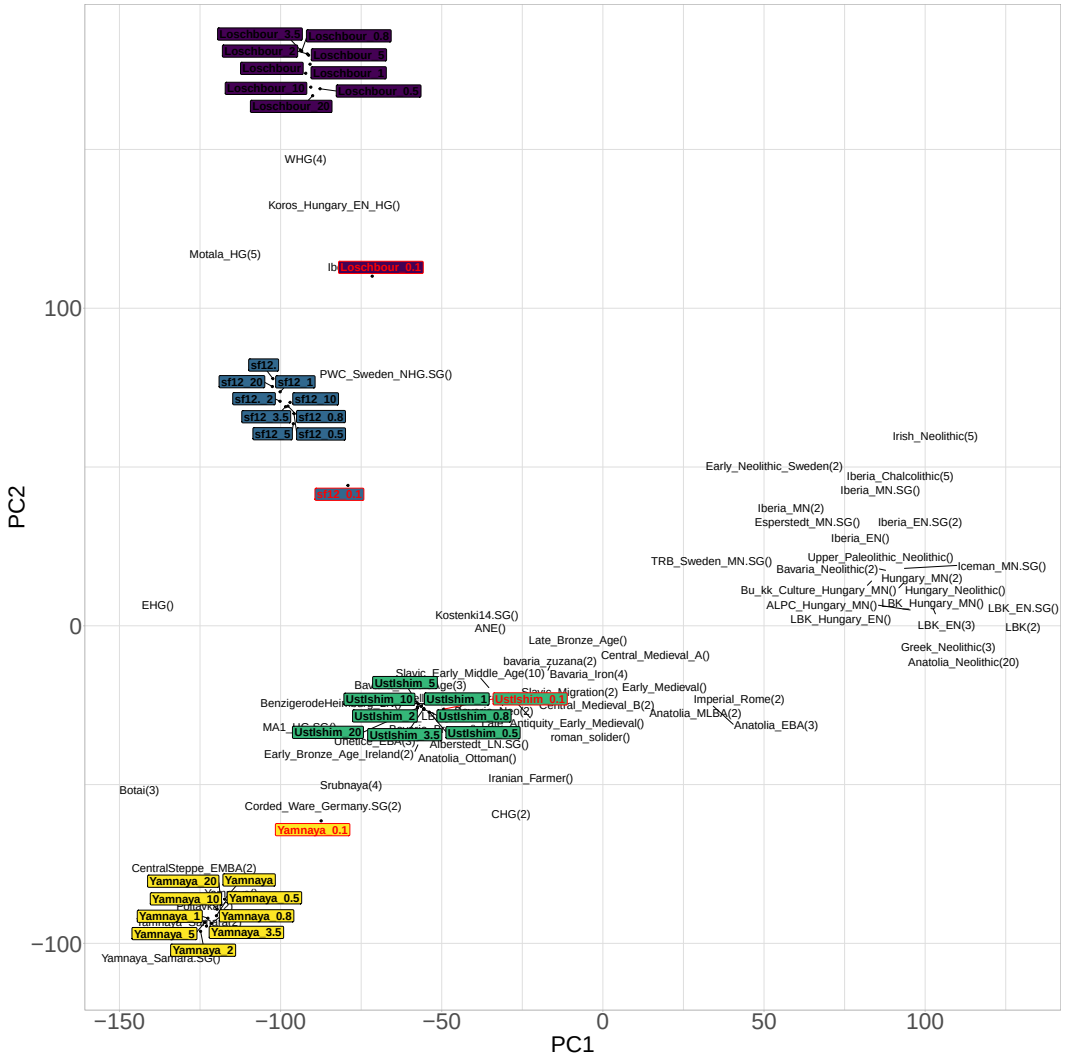

Figure 1.8: Principle component analysis (PCA) of downsampled, full coverage and downloaded ancient individuals generated from the linked chunklengths matrix. Full coverage and downsampled genomes of the same individual are coloured the same. Reference individuals are grouped into populations plotted 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.

Western Hunter-Gatherer and Yamnaya steppe pastoralist. I compared inferred proportions for the same individual across different levels of coverage (Fig. 1.9).

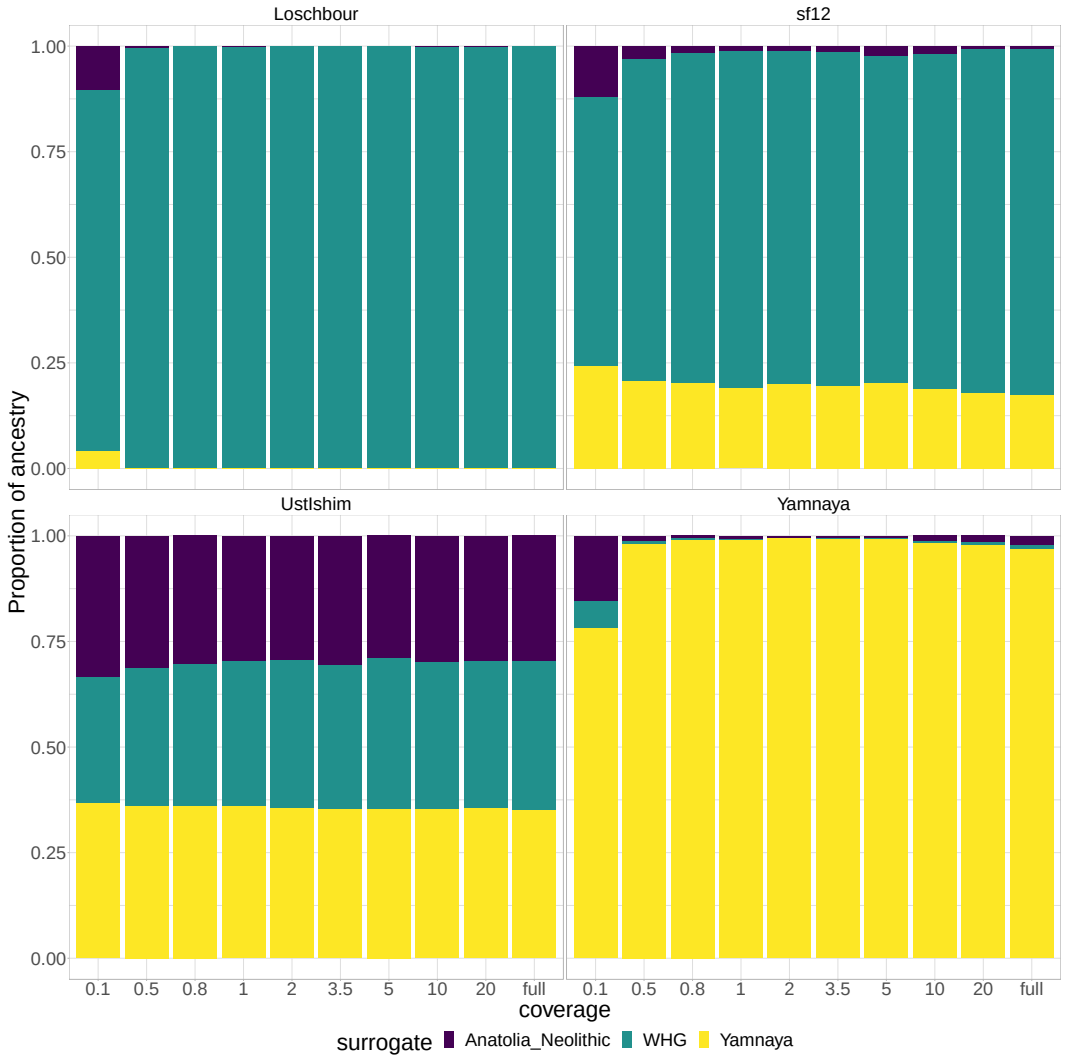

Figure 1.9: Each panel gives inferred recent ancestry sharing proportions for a different downsampled genome. Bars represent proportion of ancestry, coloured by different surrogates. Different coverages for the same individual are given within each panel. Three surrogates were used.

Consistent with previous the results, SOURCEFIND estimates appear to be are robust down to 0.5-0.8x coverage. At 0.1x coverage, there is an increase in ancestry components that are not present in higher coverage samples, suggesting they are artifacts caused by low coverage. For example, small components of Anatolia Neolithic and Yamnaya ancestry appear in Loschbour at 0.1x coverage, which are not present at any higher coverages. Above 0.5x coverage, the effect of coverage on estimated ancestry proportions appears to be marginal. For example, in sf12, the difference in the minor ancestry component of Anatolia Neolithic is, at most, 2.369%.

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

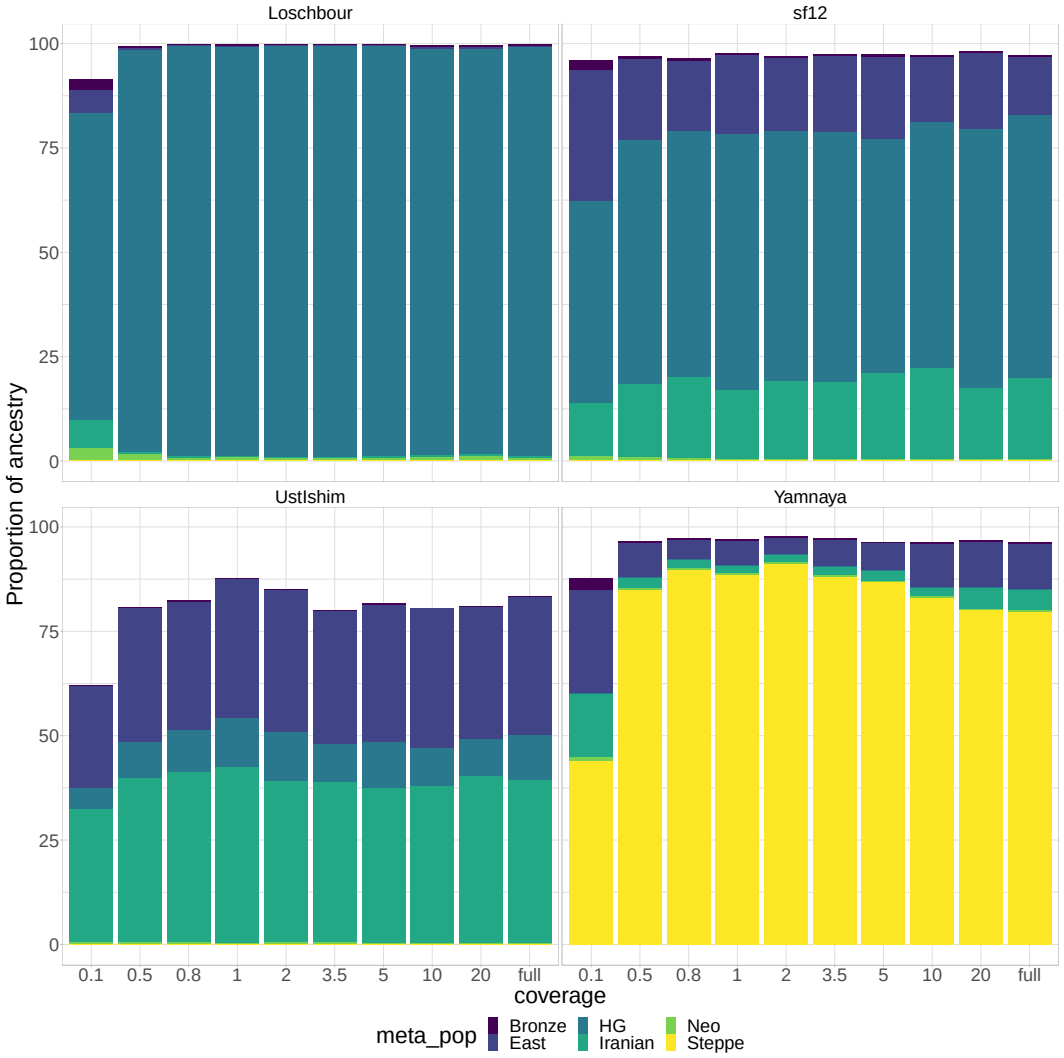

Figure 1.10: Each panel gives information for a different downsampled genome. Bars represent proportion of ancestry, coloured by different surrogates. Different coverages for the same individual are given within each panel. All 39 ancient surrogates were used. Only surrogates with more than 5% are shown. Ancient surrogates grouped into hand-assigned 'meta-populations' for visual clarity.

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 [10,33,34]. 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.

1.7 Issues and possible solutions for low coverage ancient DNA

The previous section outlined a drawback of performing ChromoPainter analysis on low coverage (<0.5x) ancient DNA samples; low coverage samples appear to be shifted towards the origin of a principle component analysis (PCA) relative to the same sample at higher coverage (Fig. 1.8). This is evident for the lowest coverage samples at 0.1x and suggests that samples of this coverage cannot be reliably analysed using current methodology.

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

1.7.1 PCA imputation test

To explicitly test at what stage the bias is introduced, I performed a set of principle component analyses on the downsampled data. First, I performed PCA projections of all downsampled ancient individuals onto a set of present-day European individuals (shown in Table 1.1) using i) pre-GLIMPSE genotypes and ii) post-GLIMPSE (imputed) genotypes (Fig. 1.12). PCA projections are used when the target dataset, in this case downsampled ancients, contain variable levels of missing data.

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

On the other hand, the 0.1x samples are clearly shifted to the centre of the post-GLIMPSE PCA, away from the full coverage individual and other downsamples. This suggests that coverage-related bias is being introduced in the imputation stage. At the same time, GLIMPSE appears to have removed some of the noise in the downsampled individuals of coverage ≥0.5x. For instance, the noise observed in the LBK samples in the pre-imputation PCA is substantially reduced and the samples cluster more tightly.

I also performed PCAs based upon an all-v-all ChromoPainter painting using the same set of present-day European samples (Table 1.1) and downsampled ancient individuals as previously, in both linked and unlinked modes. There is an increased amount of noise and evidence of coverage-related bias relative to the post-GLIMPSE genotype PCA. Fig.

1.12) displays the PCA for the same painting, but using the unlinked chunkcounts matrix. Comparing the linked and unlinked PCAs shows the effect of including linkage (i.e. haplotype information) on the amount of bias and noise across each sample. Per-sample, there appears to be reduced noise in the unlinked painting.

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

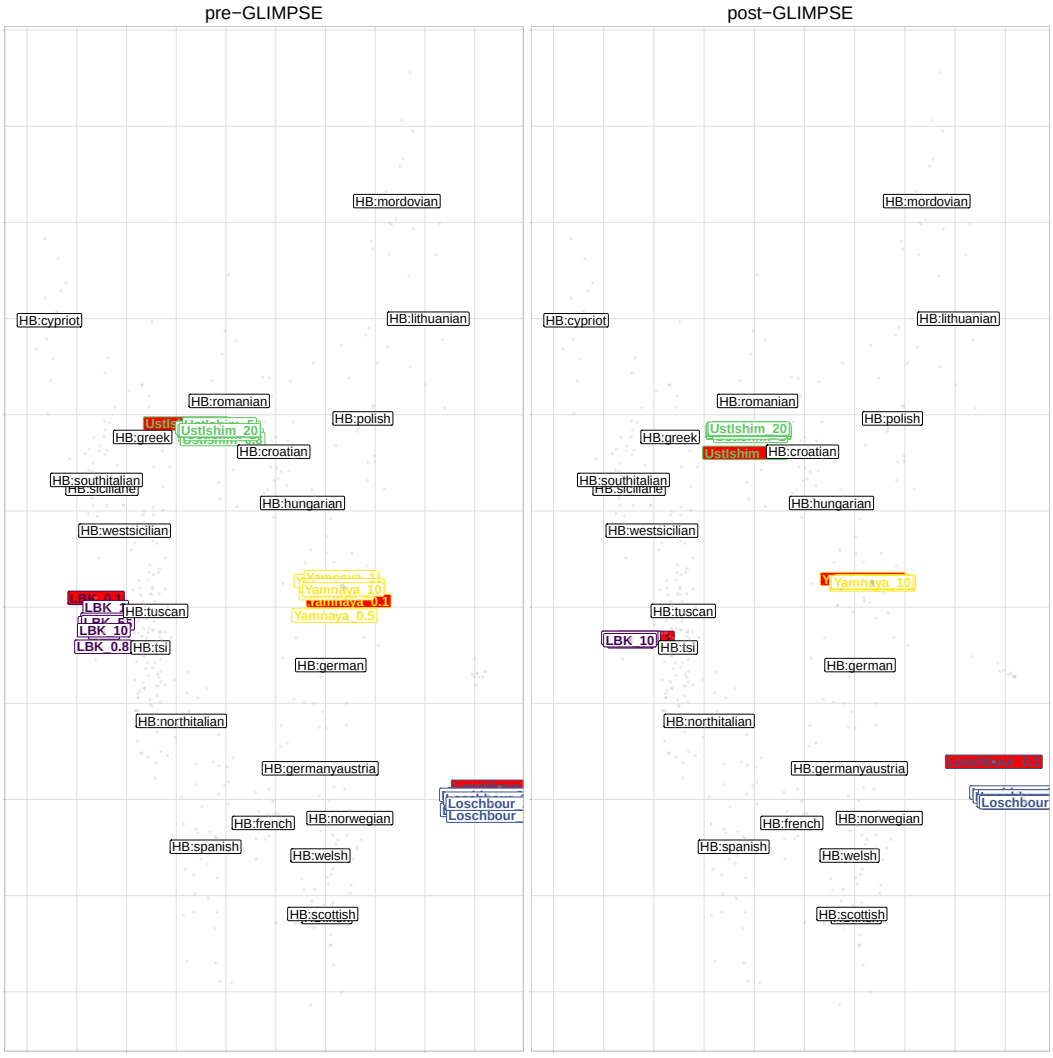

Figure 1.11: Principle Component Analysis. Left - pre-GLIMPSE genotypes. Right - post-GLIMPSE genotypes. White labels correspond to the midpoint of all samples from that population, grey points correspond to modern individuals.

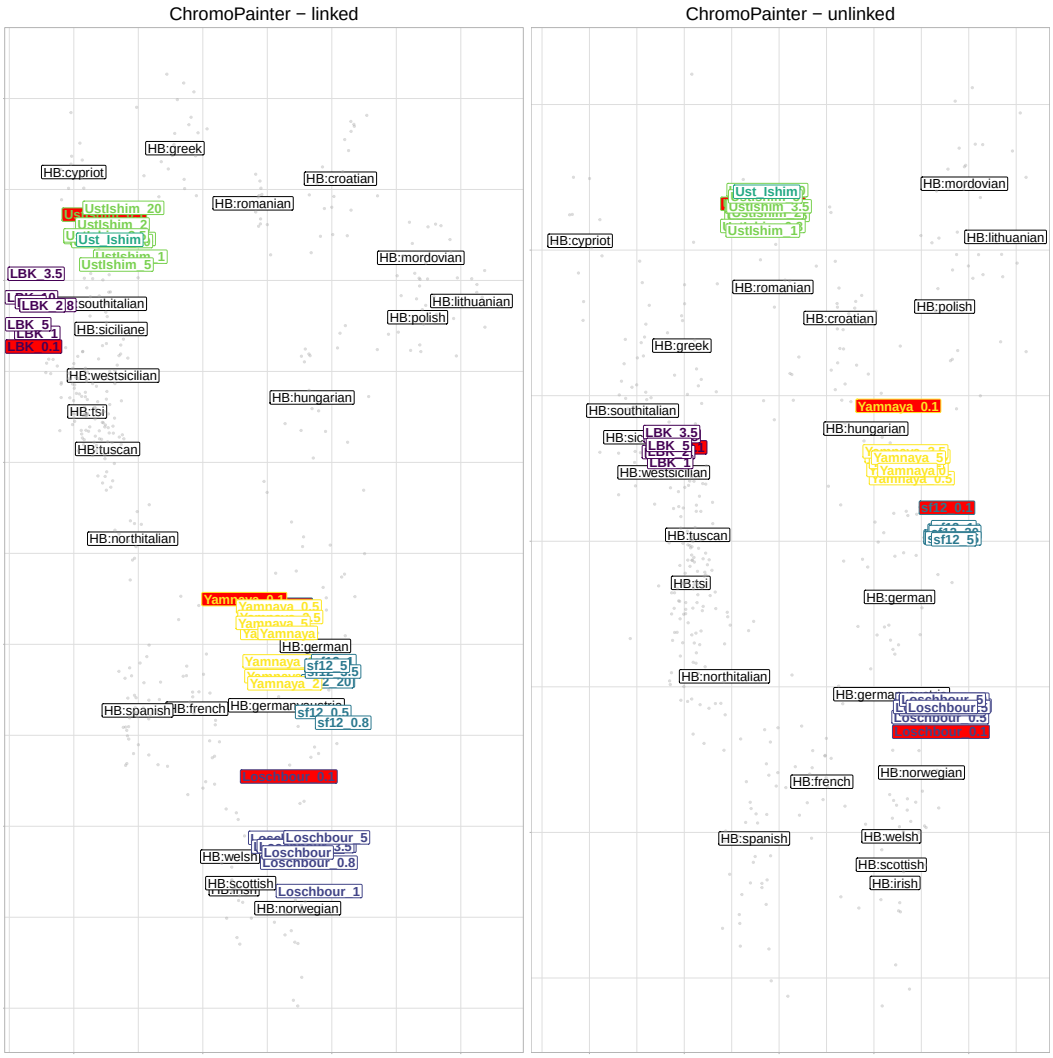

Figure 1.12: Left - ChromoPainter Linked. Right - ChromoPainter Unlinked. White labels correspond to the midpoint of all samples from that population, grey points correspond to modern individuals.

1.7.2 Direct imputation test

The previous section suggested that imputation plays a role in the introduction of coverage-related bias. However, it is not clear whether it is 'bias', i.e. towards the reference population used to assist imputation, or 'noise' due to random incorrect imputation. To directly test whether the effect of imputation is noise or bias, I used the Human Origins dataset (described in appendix A.19), containing the genotypes of 5998 present-day individuals from across the world, genotyped at 560,442 SNPs. I chose to use present-day samples because there is a larger total number of individuals and larger number of individuals per population, giving more power to detect any potential bias. Additionally, the populations in present-day samples are more homogenous and well-defined compared to ancient groups. I set all but

70,000 SNPs as missing and imputed missing positions using the HRC as a reference, in order to simulate a dataset where the majority of SNPs are imputed. I then performed an all-v-all painting of i) the original Human Origins dataset where none of the 560,442 SNPs had been imputed and ii) the simulated dataset where 430,000 SNPs had been imputed.

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

Therefore, we can compare the amount different donor groups donate under the dataset where none of the 560,442 SNPs had been imputed versus the dataset where 430,000 (86%) of these SNPs have been imputed by plotting the mean amount donated by each population using imputed SNPs and non-imputed SNPs (Fig. 1.13). The 20 populations that contribute most are either European or Jewish. Notably, the Haplotype Reference Consortium panel that was used to impute the data consists primarily of individuals of European descent. The two populations which are over-copied the most after imputation are two English populations from Kent and Cornwall. This suggests that there is a most likely a bias towards copying more from European populations when the data has been imputed using the HRC.

1.8 Solutions

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

1.8.1 Accounting for allele likelihoods

Section 2.2.1 describes an improvement to the ChromoPainter algorithm. Instead of assuming that each allele on a haplotype is correct with a probability $1 - \theta$, where θ represents an error probability, the posterior genotype probability from GLIMPSE is accounted for in the emission probabilities of the copying model. The motivation behind this update is that the uncertainty associated with genotype calls at low coverage is suitably propagated throughout the painting process, resulting in uncertain alleles contributing less towards the

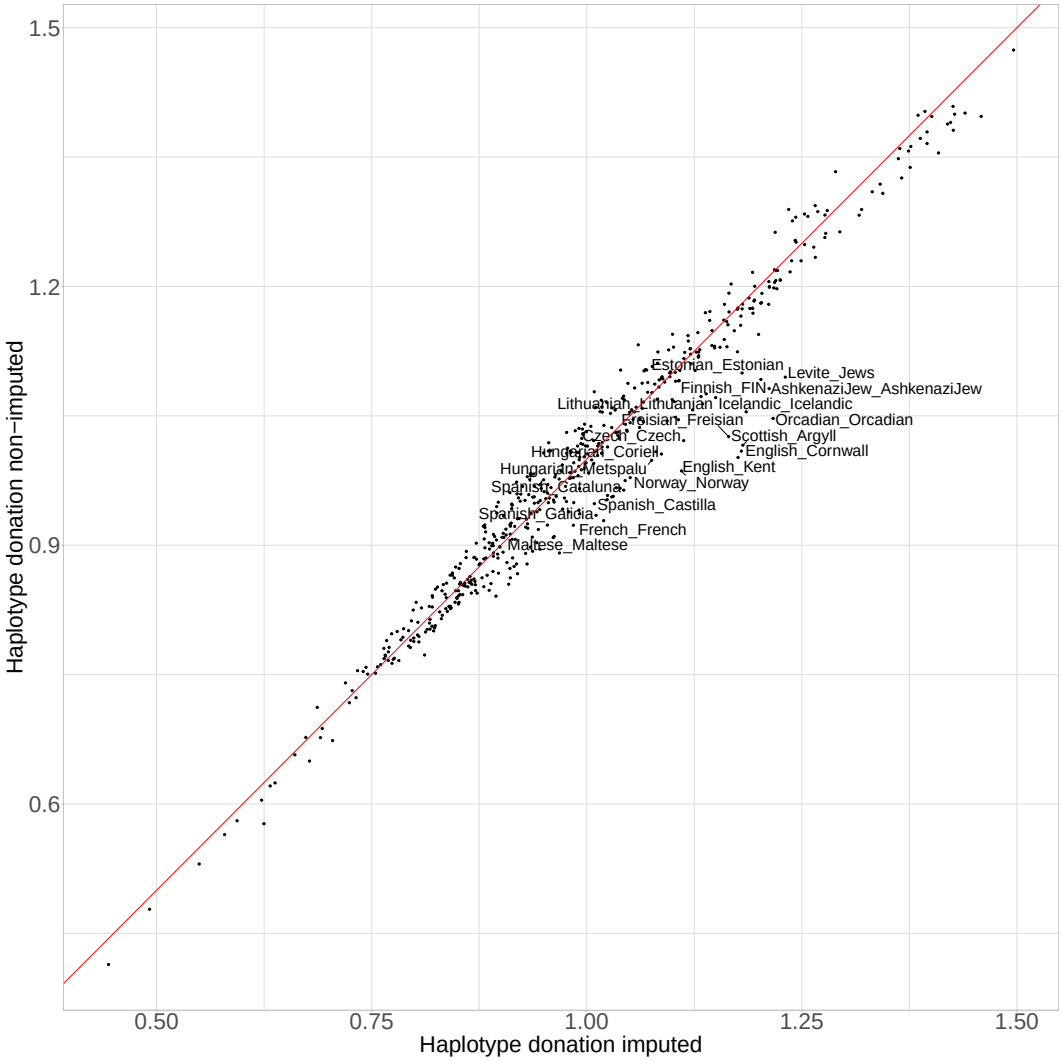

Figure 1.13: Comparison of the mean normalised cM donated by each donor population using the imputed and non-imputed SNP sets. The 20 populations with the largest different between imputed and non-imputed donation are highlighted.

expected copying values than more certain ones. This is similar in spirit to that of Viera et al (2016), who account for genotype likelihoods to infer inbreeding IBD tracts from low coverage sequencing data [35].

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. 1.15). 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 ≥0.5x, which has been shown to still capture some haplotype information, it does not help in cases of coverage of 0.1x where bias problems persist.

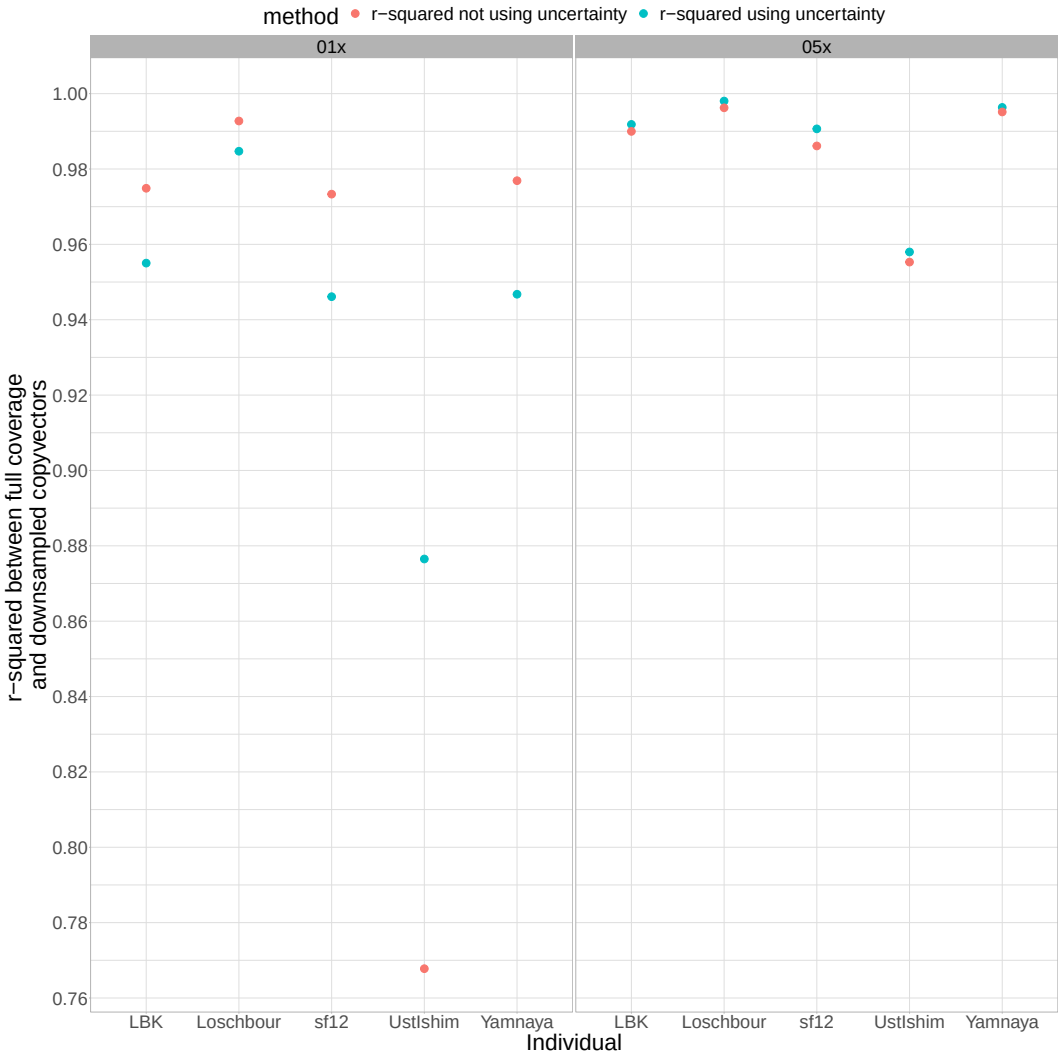

Figure 1.14: Comparison of performance of ChromoPainterV2 and ChromoPainterV2Uncertainty. Panels correspond to samples downsampled to 0.1x (left) and 0.5x (right). Points correspond to the r-squared between the downsampled individual and the same individual at full coverage. Red points are values obtained from ChromoPainterV2 and blue points are those obtained from ChromoPainterV2Uncertainty.

1.8.2 Filtering SNPs

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

The frequency of a particular variant in the reference panel (RAF - reference allele

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

Table 1.2: Table of r-squared values between the copyvectors of full coverage and downsampled individuals. 'u' refers to ChromoPainterUncertainty, 's' refers to ChromoPainterV2, 'r' refers to filtering SNPs with reference allele frequency (RAF) $0.1 > RAF$ or $RAF > 0.9$ and 'gp' refers to filtering by $max(GP) >= 0.990$.

frequency) used for imputation is known to affect how accurately that variant can be imputed [2, 12, 15, 36]. Specifically, we expect variants which are less frequent in the reference panel to be imputed at a lower accuracy than those which are more frequent. Therefore, removing variants with a low frequency in the reference panel may mitigate the coverage related bias by removing variants which have been incorrectly imputed. In other words, we want to retain the SNPs where both alleles are relatively common within the population.

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

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

I then chose to filter SNPs based on $max(GP)$ at each position. $max(GP)$ correspond to the accuracy with which a SNP has been imputed, with higher values reflecting a higher chance of that genotype being imputed correctly. For each individual downsampled to 0.5x, I only retained positions where the $max(GP) >= 0.990$. This resulted in a total of 348,852 SNPs for LBK, 339,949 for Loschbour, 315,075 for sf12, 308,961 for UstIshim and 386,484 for Yamnaya. Because different SNPs were removed from different individuals, each individual was painted separately. The same standard set of 124 ancient donors was used. Again, this did not improve the accuracy of the copyvectors.

1.8.3 Restricting analysis to non-imputed SNPs

Section 1.6.1 showed that imputation was the likely cause of coverage related bias. Thus, restricting ChromoPainter analysis to non-imputed SNPs above a certain coverage may mitigate such bias.

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

One case study to test whether a set of SNPs has enough linkage information is to determine whether it is possible to distinguish individuals born in Devon from those born in Cornwall. This has shown to be possible using the fineSTRUCTURE clustering algorithm using linkage information, but not using unlinked methods (ADMIXTURE [25]) [26]. Therefore, determining whether it is possible to distinguish between individuals from Devon and Cornwall acts as a test case for determining how many high-coverage SNPs would give sufficient SNP density to distinguish between these two populations.

To assess this, I painted individuals from Devon (n=73) and Cornwall (n=89) with all other POBI individuals as donors (n=2039), using the full set of SNPs (n=452,592). It is necessary to develop a classification score which quantifies to what degree it is possible to distinguish between individuals from Devon and Cornwall. For a classification score, I calculated the proportion of Cornwall individuals whose copy vector had a lower TVD with the mean copy vector of all other Cornwall individuals than with the mean copy vector of all Devon individuals. In other words, this asks whether the copyvector of an individual is closer to all other individuals from Devon or to Cornwall.

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

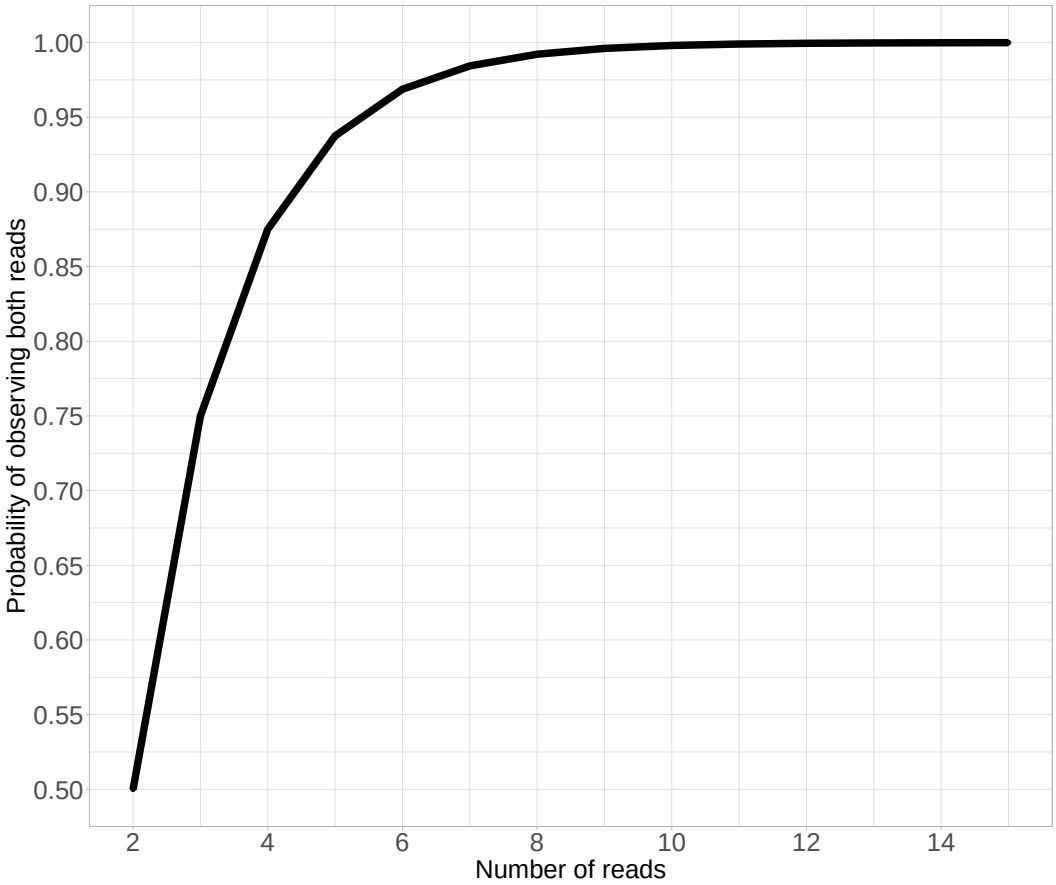

Figure 1.15: Probabilty of observing both reads given x reads at a given heterozygous positions and assuming equal probability of observing reference and non-reference alleles.

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

Table 1.3: Percentage of individuals correctly assigned to their population at different percentages of SNPs retained.

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

Table 1.4: Number of 250Kb, 500Kb or 1Mb windows required at different levels of SNP reduction to match the TVD assignment power of 500K fully genotyped SNPs for individuals in Devon and Cornwall. Note that the number of necessary 250kb and 500kb windows is roughly four and two times, respectively, the number of 1Mb windows, indicating the definition of window size makes little difference. ADD IN COLUMNS AFTER N_SNPS TO SAY WHATS THE NUMBER OF SNPS PER 500KB WINDOW

In my painting of 5998 world-wide samples on the Human Origins array (described in appendix section **??**), the average number of segments that forms a recipient genome is 9764 (range: 1437-18,963). Given a genome-wide size of ≈3000Mb, this implies that an average "chunk" size (in Mb) is 3000/9764 = 307.2 ≈500kb, where a "chunk" is a set of contiguous SNPs matched to a single donor. Therfore, for each of the 12 different numbers of genome-wide total SNPs used in my Devon/Cornwall analysis, I can calculate the average number of SNPs per 500kb chunk, and determine how many of these 500kb chunks are necessary to accurately distinguish individuals from Devon and Cornwall. To do so, for each reduced SNP percentage, I found the Cornwall/Devon classification score using only data from chromosome 22 (which has only W 500kb chunks), and using only chromsomes 21 and 22 (which has V 500kb chunks), etc, continuing until the classification scores were equivalent to that when analysing all 22 autosomes at all 452,592 SNPs. In this way, for each reduced SNP percentage, I found the number of 500kb chunks necessary to as accurately distinguish between Devon and Cornwall as in the case where we had analysed a full data set of 452,592 SNPs (Table **??**). I found results to be very similar to if chunk-size were instead defined as 250kb or 1Mb (Table **??**).

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

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

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

Table 1.5: Number of 250Kb, 500Kb or 1Mb windows required at different levels of SNP reduction to match the TVD assignment power of 500K fully genotyped SNPs for individuals in from Mandenka and Yoruba ethnic groups.

Fig 1.16 shows the mean number of 500Kb windows per individual with at least Y SNPs above Z coverage, with individuals being grouped into bins based on their mean coverage. Points are coloured yellow if, within the bin of coverage, samples have at least 2000 windows[NOTE WE CAN'T TELL HOW TO READ THIS WITHOUT KNOWING THE CORRESPONDING SNP DENSITIES IN TABLE 2.2? I.E. THAT 40K SNPS CORRESPONDS TO 6-7SNPS/WINDOW. CAN YOU PUT THIS INFO IN TABLE 2.2?]. Samples less than 0.5x do not have enough windows if the threshold for a 'good' SNPs is being covered by a single read. As it not possible to call a heterozygous position with only a single read, this suggests that there are not enough non-imputed SNPs with enough coverage to match the power seen in full coverage individuals.[NOT QUITE CLEAR IT'S THAT BAD – FOR 0.3-0.4x SAMPLES, THERE ARE 1000 SEGMENTS WITH >=10 SNPS, WHILE TABLE 2.3 SAYS 1565 SNPS WITH >=8.3 SNPS IS ENOUGH (AND THAT'S TO ACHIEVE FULL POWER; ONE QUESTION IS WHETHER YOU CAN REDUCE THIS; I.E. WOULD BE HELPFUL TO HAVE ANOTHER TABLE SHOWING THE CLASSIFICATION RATE YOU GET WITH 500). NOTE ALSO IT WOULD BE HELPFUL TO HAVE VALUES BETWEEN 50K AND 100K, GIVEN THE RANGE COVERED BY THE X-AXIS OF FIG 2.14] Indeed, even when there are 3 reads covering a site, there is still a 25% chance of not identifying a heterozygous position. Only the samples in the 2-5x coverage bin had enough windows when using a coverage threshold of 4 and 5 reads.

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

INSTEAD OF COMPARING TO FULL COVERAGE - COMPARED TO UNLINKED

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

1.8.4 Averaging across copyvectors

1.9 Discussion

Many of the analyses performed in this section only used a single target sample, as I did not identify a way to generate multiple downsampled individuals from the same population. For example, the SOURCEFIND analysis I performed used a single target downsample when estimating ancestry proportions. This differs from a typical ancient DNA analysis, such as those of Margaryan et al [32], where there may be up to 20 low coverage samples per population. This number may increase in the future as the technology to generate ancient DNA improves. Leveraging information across multiple samples from the same population would improve the accuracy of population-wide ancestry or admixture estimates, for example. Thus, he results presented in this section which used a single target individual may underestimate the ability to analyse low-coverage samples. It may be possible to accurately analyse 0.1x samples if there are multiple samples per population.

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

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

1.10 Summary of findings

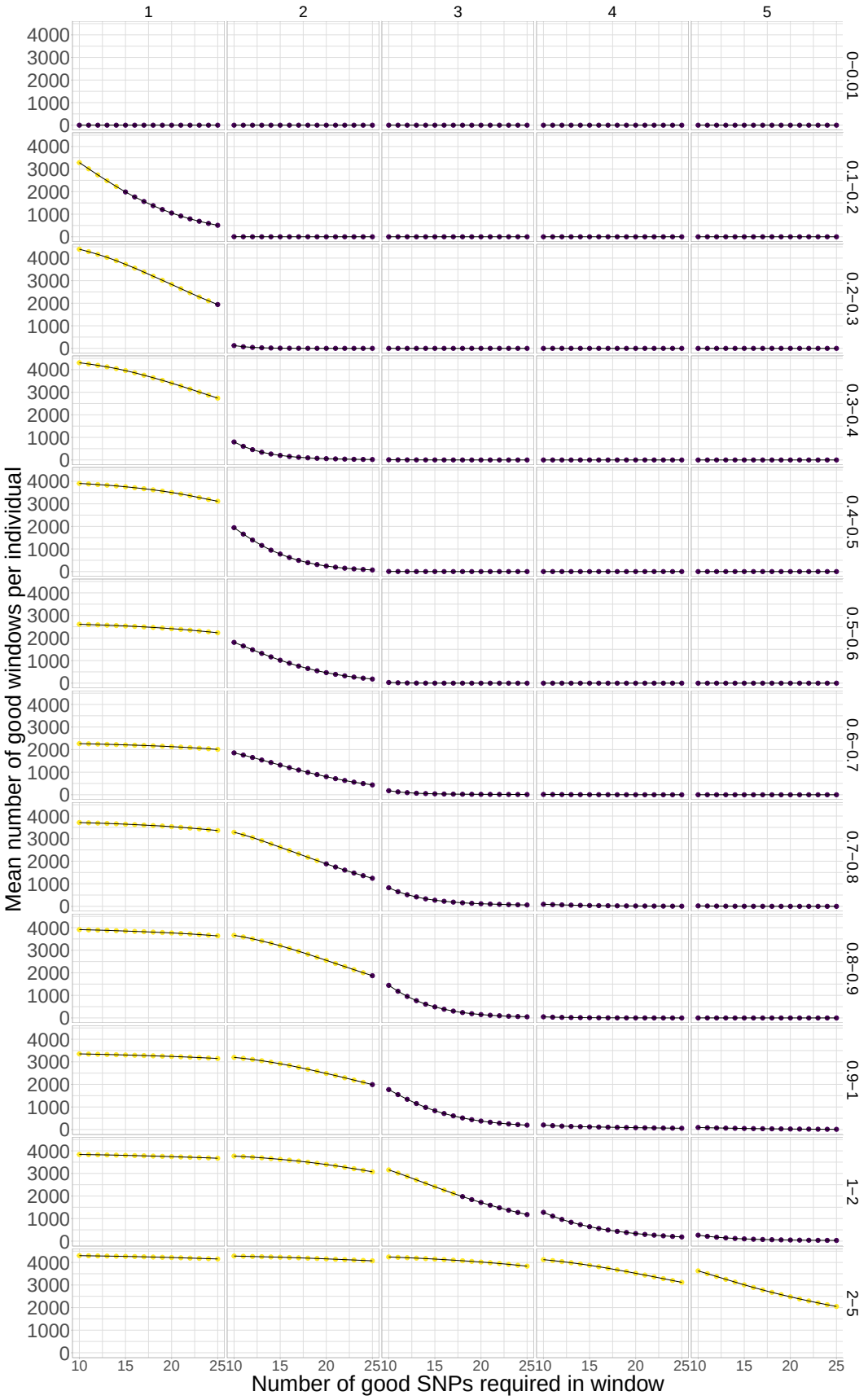

Figure 1.16: Mean number of 500Kb windows (y-axis) within the genome of each ancient individuals within a given range of coverages (rows) with at least Y SNPs (x-axis) above a particular coverage Z (columns)

Bibliography

[1] 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.

[2] 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.

[3] 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.

[4] 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.

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

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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 Nagu-

lapalli, et al. High coverage whole genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *bioRxiv*, 2021.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

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