Estimating allele frequencies, ancestry proportions and genotype likelihoods in the presence of mapping bias

Torsten Günther[1,2,*], Amy Goldberg[3] & Joshua G. Schraiber[4]

[1]Human Evolution, Department of Organismal Biology, Uppsala University, Uppsala, Sweden

[2]Science for Life Laboratory, Ancient DNA Unit, Department of Organismal Biology, Uppsala University, Uppsala, Swede

[3]Department of Evolutionary Anthropology, Duke University, USA

[4]Department of Quantitative and Computational Biology, University of Southern California, Los Angeles, USA

[*]Corresponding author: `torsten.gunther@ebc.uu.se`

Abstract

Population genomic analyses rely on an accurate and unbiased characterization of the genetic composition of the studied population. For short-read, high-throughput sequencing data, mapping sequencing reads to a linear reference genome can bias population genetic inference due to mismatches in reads carrying non-reference alleles. In this study, we investigate the impact of mapping bias on allele frequency estimates from pseudohaploid data, commonly used in ultra-low coverage ancient DNA sequencing. To mitigate mapping bias, we propose an empirical adjustment to genotype likelihoods. Using data from the 1000 Genomes Project, we find that our new method improves allele frequency estimation. To test a downstream application, we simulate ancient DNA data with realistic post-mortem damage to compare widely used methods for estimating ancestry proportions under different scenarios, including reference genome selection, population divergence, and sequencing depth. Our findings reveal that mapping bias can lead to differences in estimated admixture proportion of up to 4% depending on the reference population. However, the choice of method has a much stronger impact, with some methods showing differences of 10%. `qpAdm` appears to perform best at estimating simulated ancestry proportions, but it is sensitive to mapping bias and its applicability may vary across species due to its requirement for additional populations beyond the sources and target population. Our adjusted genotype likelihood approach largely mitigates the effect of mapping bias on genome-wide ancestry estimates from genotype likelihood-based tools. However, it cannot account for the bias introduced by the method itself or the noise in individual site allele frequency estimates due to low sequencing depth. Overall, our study provides valuable insights for obtaining more precise estimates of allele frequencies and ancestry proportions in empirical studies.

1 Introduction

A phenomenon gaining an increasing degree of attention in population genomics is mapping bias in re-sequencing studies employing short sequencing reads (Orlando et al., 2013; Gopalakrishnan et al., 2017; Günther and Nettelblad, 2019; Martiniano et al., 2020; Chen et al., 2021; Oliva et al., 2021; Prasad et al., 2022; Gopalakrishnan et al., 2022; Thorburn et al., 2023; Koptekin et al., 2025; Dolenz et al.,

2024). As most mapping approaches employ linear reference genomes, reads carrying the same allele as the reference will have fewer mismatches and higher mapping scores than reads carrying an alternative allele leading to some alternative reads being rejected. As a consequence, sequenced individuals may seem more similar to the reference genome (and hence, the individual/population/species it originates from) than they are in reality, biasing variant calling and downstream analysis. The effect of mapping bias is exacerbated in ancient DNA studies due to post-mortem DNA damage such as fragmentation and cytosine deamination to uracil (which is sequenced as thymine) (Orlando et al., 2021) which increases the chances of spurious mappings or rejected reads due to an excessive number of mismatches relative to the fragment length. The human reference genome is a mosaic sequence of multiple individuals from different continental ancestries (Green et al., 2010; Church et al., 2015). In most other species with an existing reference genome sequence, this genome represents a single individual from a certain population while for studies in species without a reference genome, researchers are limited to the genomes of related species. One consequence is that the sequence at a locus in the reference genome may either represent an ingroup or an outgroup relative to the other sequences studies in a population genomic analysis. It has been shown that this can bias estimates of heterozygosity, phylogenetic placement, assessment of gene flow, and population affinity (see e.g. Orlando et al., 2013; Heintzman et al., 2017; Gopalakrishnan et al., 2017; Günther and Nettelblad, 2019; van der Valk et al., 2020; Mathieson et al., 2020; Prasad et al., 2022). Notably, while mapping bias mostly manifests as bias in favor of the reference allele, it also exists as bias in favor of the alternative alelle, depending on the studied individual and the particular position in the genome (Günther and Nettelblad, 2019).

Different strategies have been proposed to mitigate or remove the effect of mapping bias. These include mapping to an outgroup species (Orlando et al., 2013), mapping to multiple genomes simultaneously (Huang et al., 2013; Chen et al., 2021), mapping to variation graphs (Martiniano et al., 2020), the use of an IUPAC reference genome (Oliva et al., 2021), masking variable sites (Koptekin et al., 2025) or filtering of "biased reads" (Günther and Nettelblad, 2019). All of these strategies have significant limitations, such as the exclusion of some precious sequencing reads (outgroup mapping or filtering) or requiring additional data that may not be available for all species prior to the particular study (variation graphs, IUPAC reference genomes, or mapping to multiple genomes). Therefore, it would be preferable to develop a strategy that uses the available sequencing reads and accounts for potential biases in downstream analyses. Genotype likelihoods (Nielsen et al., 2011) represent one promising approach that can be used with low- and medium-depth sequencing data (Lou et al., 2021). Instead of working with hard genotype calls at each position one can use $P(D|G)$, the probability of observing a set of sequencing reads D conditional on a true genotype G. Different approaches exist for calculating genotype likelihoods with the main aim of accounting for uncertainty due to random sampling of sequencing reads and sequencing error. Genotype likelihoods can be used in a wide range of potential applications for downstream analysis which include imputation (Rubinacci et al., 2021), estimation of admixture proportions (Skotte et al., 2013; Jørsboe et al., 2017; Meisner and Albrechtsen, 2018), principal component analysis (PCA, Meisner and Albrechtsen, 2018), relatedness analysis (Korneliussen and Moltke, 2015; Hanghøj et al., 2019; Nøhr et al., 2021), or to search for signals of selection (Korneliussen et al., 2013; Fumagalli et al., 2013). Many of these are available as part of the popular software package `ANGSD` (Korneliussen et al., 2014).

To render genotype likelihoods and their downstream applications more robust to the presence of mapping bias, we introduce a modified genotype likelihood, building off of the approach in Günther and Nettelblad (2019). We modify reads to carry both alleles at biallelic SNP positions to assess the distribution of mapping bias and to obtain an empirical quantification of the locus- and individual-specific mapping bias. We then calculate a modified genotype likelihood to account for mapping bias. The approach is similar to `snpAD` (Prüfer, 2018), with the contrast that we are using a set of pre-ascertained biallelic SNPs because our aim is not to call genotypes at all sites across the genome including potentially novel SNPs. Restricting to known biallelic SNPs is a common practice in the population genomic analysis of ancient DNA data as low-coverage and post-mortem damage usually limit the possibility of calling novel SNPs for most individuals (see e.g. Günther and Jakobsson, 2019),

and methods like `snpAD` are restricted to very few high quality, high coverage individuals (Prüfer, 2018). Instead, most studies resort to using pseudohaploid calls or genotype likelihoods at known variant sites (Günther and Jakobsson, 2019); using ascertained biallelic SNPs is particularly relevant when ancient DNA is enriched using a SNP capture array (Rohland et al., 2022). This choice also allows us to estimate mapping bias locus-specific rather than using one estimate across the full genome of the particular individual.

We examine two downstream applications of genetic data to determine the impact of mapping bias, and assess the ability of our corrected genotype likelihood to ameliorate issues with mapping bias. First, we look at a very high-level summary of genetic variation: allele frequencies. Because allele frequencies can be estimated from high-quality SNP array data, we can use them as a control and assess the impact of mapping bias and our corrected genotype likelihood in real short-read data.

Next, we examine the assignment of ancestry proportions. Most currently used methods trace their roots back to the software `STRUCTURE` (Pritchard et al., 2000; Falush et al., 2003, 2007; Hubisz et al., 2009), a model-based clustering approach modeling each individual's ancestry from K source populations (Pritchard-Stephens-Donnelley, or PSD, model). These source populations can be inferred from multi-individual data (unsupervised) or groups of individuals can be designated as sources (supervised). Popular implementations of this model differ in terms of input data (e.g. genotype calls or genotype likelihoods), optimization procedure and whether they implement a supervised and/or unsupervised approach (Table 1). In the ancient DNA field, f statistics (Patterson et al., 2012) and functions derived from them are fundamental to many studies due to their versatility, efficiency and their ability to work with pseudohaploid data, in which a random read is used to call haploid genotypes in low coverage individuals. Consequently, methods based on f statistics are also often used to estimate ancestry proportions in ancient DNA studies. One method that uses f statistics for supervised estimation of ancestry proportions is `qpAdm` (Haak et al., 2015; Harney et al., 2021). In addition to the source populations ("left" populations), a set of more distantly related "right" populations is needed for this approach. Ancestry proportions are then estimated from a set of f_4 statistics calculated between the target population and the "left" and "right" populations. We simulate sequencing data with realistic ancient DNA damage under a demographic model with recent gene flow (Figure 1) and then compare the different methods in their ability to recover the estimated admixture proportion and how sensitive they are to mapping bias.

2 Materials and Methods

2.1 Correcting genotype-likelihoods for mapping bias

Two versions of genotype likelihoods (Nielsen et al., 2011) were calculated for this study. First, we use the direct method as included in the original version of GATK (McKenna et al., 2010) and also implemented in `ANGSD` (Korneliussen et al., 2014). For a position ℓ covered by n reads, the genotype likelihood is defined as the probability for observing the bases $D_\ell = \{b_{\ell 1}, b_{\ell 2}, \ldots, b_{\ell n}\}$ if the true genotype is $A_1 A_2$:

$$P(D_\ell | G_\ell = A_1, A_2) = \prod_{i=1}^{n} P(b_{\ell i} | G_\ell = A_1, A_2) = \prod_{i=1}^{n} \frac{P(b_{\ell i}|A_1) + P(b_{\ell i}|A_2)}{2} \tag{1}$$

with

$$P(b_{\ell i}|A) = \begin{cases} 1 - e_{\ell i} & \text{if } b = A \\ \frac{e_{\ell i}}{3} & \text{if } b \neq A \end{cases}$$

where $e_{\ell i}$ is the probability of a sequencing error of read i at position ℓ, calculated from the phred scaled base quality score $Q_{\ell i}$, i.e. $e_{\ell i} = 10^{-Q_{\ell i}/10}$. The calculation of genotype likelihoods was implemented in Python 3 using the `pysam` library (`https://github.com/pysam-developers/pysam`), a wrapper

around `htslib` and the `samtools` package (Li et al., 2009), or by calling `samtools` mpileup and parsing the output in the Python script. Both corrected and default genotype likelihoods are calculated by the same Python script.

To quantify the impact of mapping bias, we restrict the following analysis to a list of pre-defined ascertained biallelic SNPs (list provided by the user) and modify each original read to carry the other allele at the SNP position, as in Günther and Nettelblad (2019). The modified reads are then remapped to the reference genome using the same mapping parameters. If there were no mapping bias, all modified reads would map to the same position as the unmodified original read. Consequently, when counting both original and modified reads together, we should observe half of our reads carrying the reference allele and the other half carrying the alternative allele at the SNP position. We can summarize the read balance at position ℓ as r_ℓ, which measures the proportion of reference alleles among all original and modified reads mapping to the position. Without mapping bias, we would observe $r_\ell = 0.5$. Under reference bias, we would observe $r_\ell > 0.5$ and under alternative bias $r_\ell < 0.5$. We can see r_ℓ as an empirical quantification of the locus- and individual-specific mapping bias. Similar to Prüfer (2018), we can then modify Equation 1 for heterozygous sites to

$$P(D_\ell|G_\ell = R_\ell, A_\ell) = \prod_{i=1}^{n} r_\ell P(b_{\ell i}|R_\ell) + (1 - r_\ell)P(b_{\ell i}|A_\ell) \tag{2}$$

where R_ℓ is the reference allele at position ℓ and A_ℓ is the alternative allele. Note that when $r_l \equiv \frac{1}{2}$, this recovers Equation 1. Genotype likelihood-based methods are tested with both genotype likelihood versions. All code used in this study can be found under https://github.com/tgue/refbias_GL

2.2 Empirical Data

To estimate the effect of mapping bias in empirical data we obtained low coverage BAM files for ten FIN (Finnish in Finland) individuals, ten JPT individuals (Japanese in Tokyo, Japan) and ten YRI (Yoruba in Ibadan, Nigeria) individuals from the 1000 Genomes project (mostly 2–4x coverage; Table S1) (Auton et al., 2015). We also downloaded Illumina Omni2.5M chip genotype calls for the same individuals (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz). The SNP data was filtered to restrict to sites without missing data in the 30 selected individuals, a minor allele frequency of at least 0.2 in the reduced dataset (considering individuals from all populations together), which makes it more likely that the SNPs are common in all populations and both over- and underestimation of allele frequencies could be observed. We also excluded A/T and C/G SNPs to avoid strand misidentification. Reads mapping to these positions were extracted from the BAM files using `samtools` (Li et al., 2009). To make the sequence data more similar to fragmented ancient DNA, each read was split into two halves at its mid-point and each sub-read was re-mapped separately. For mapping, we used `bwa aln` (Li and Durbin, 2009) and the non-default parameters -l 16500 (to avoid seeding), -n 0.01 and -o 2 (to allow for more gaps due to post-mortem damages and increased evolutionary distance to the reference) (Schubert et al., 2012; Oliva et al., 2021). Only reads with mapping qualities of 30 or higher were kept for further analysis.

Pseudohaploid genotypes were called with `ANGSD` v0.933 (Korneliussen et al., 2014) by randomly drawing one read per SNP with a minimum base quality of 30. This step was performed using `ANGSD` with the parameters -checkBamHeaders 0 (to deactivate checking the headers of the BAM files) -doHaploCall 1 (to sample a single base only) -doCounts 1 (needed to determine the most common base) -doGeno -4 (to not print genotypes) -doPost 2 (estimate the posterior genotype probability assuming a uniform prior, output files not used) -doPlink 2 (produce output in tfam/tped format) -minMapQ 30 (to set the minimum mapping quality) -minQ 30 (to set the minimum base quality) -doMajorMinor 1 (to infer major and minor from genotype likelihoods) -GL 2 (to calculate GATK genotype likelihood, output files not used) -domaf 1 (calculate allele frequencies with fixed major and minor alleles). This call also calculates genotype likelihoods in `ANGSD` but we used both default

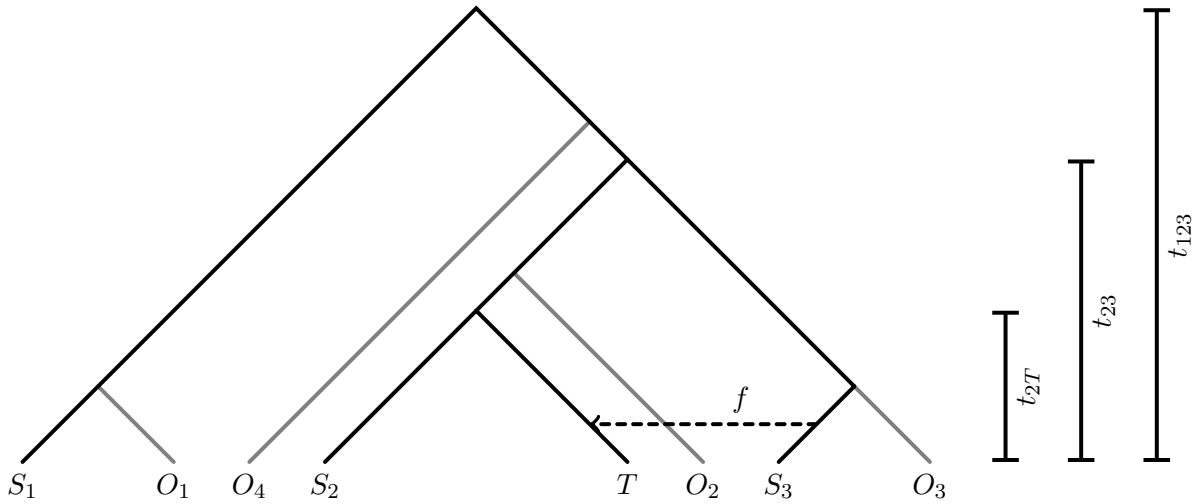

Figure 1: Illustration of the population relationships used in the simulations. Branch lengths are not to scale

and corrected likelihoods calculated from our own Python script to ensure consistency. Haplocall files were then converted to `Plink` format using `haploToPlink` distributed with `ANGSD` (Korneliussen et al., 2014). Only SNPs with the same two alleles in pseudohaploid and SNP chip data were included in all comparisons. Remapping of modified reads and genotype likelihood calculation were performed as described above. SNP calls from the genotyping array and pseudohaploid calls were converted to genotype likelihood files assuming no genotyping errors (i.e. the genotype likelihood of the observed genotype was set to 1.0, others to 0.0 whereas all three likelihoods were set to $\frac{1}{3}$ if data was missing for the site and individual). This allowed us to also estimate allele frequency estimates for this data with `ANGSD`. Allele frequencies were calculated from genotype likelihoods with `ANGSD` v0.933 (Korneliussen et al., 2014) using -doMaf 4 and the human reference as "ancestral" allele (-anc) in order to calculate the allele frequency of the reference alleles.

2.3 Simulation of genomic data

To test the methods while having control over the "true" admixture proportions, population histories were simulated using `msprime` v0.6.2 (Kelleher et al., 2016). We simulated a demographic history where a target population T receives a single pulse of admixture with proportion f from source $S3$ 50 generations ago. Furthermore, we simulated population $S1$ which forms an outgroup and population $S2$ which is closer to T than $S3$ to serve as second source for estimating ancestry proportions (Figure 1). Finally, we simulated populations $O1$, $O2$, $O3$, and $O4$ as populations not involved in the admixture events which split off internal branches of the tree to serve as "right" populations for `qpAdm` (Haak et al., 2015; Harney et al., 2021). Split times were scaled relative to the deepest split t_{123}: the split between $(S2, T)$ and $S3$, t_{23}, is set to $0.5 \times t_{123}$ while the split between T and $S2$ was set to $0.2 \times t_{123}$. To set t_{123}, we considered a value of 20,000 generations, approximately falling in the range of the split of all human populations (Schlebusch et al., 2017) or the Neanderthal-Denisovan split (Rogers et al., 2017) i.e. approximating the divergence between distant populations or sub-species, and 50,000 generations, corresponding to a comparison between closely related species. Mutation rate was set to 2.5×10^{-8} and recombination rate was set to 2×10^{-8}, which are both in the upper part of the ranges for mammals and vertebrates (Dumont and Payseur, 2008; Bergeron et al., 2023). The effective population size along all branches was 10,000, a value often considered for humans (Charlesworth, 2009). For each population, 21 diploid individuals (i.e. 42 haploid chromosomes) with 5 chromosome pairs of 20,000,000 bp (corresponding to a short mammalian chromosome arm, Lander et al. (2001)) each were simulated.

As `msprime` does not produce sequences but positions of derived alleles at each haploid chromosome,

we had to convert this information into a sequence. For each chromosome, a random ancestral sequence was generated with a GC content of 41% corresponding to the GC content of the human genome (Lander et al., 2001). Transversion polymorphisms were then placed along the sequence at the positions produced by the `msprime` simulations. The resulting sequences for each haploid chromosome were then stored as FASTA files. One of the 42 simulated sequences from populations $S1$, $S2$ and $S3$ were used as reference genomes. Out of the remaining sequences, pairs of FASTA files were then considered as diploid individuals and used as input for `gargammel` (Renaud et al., 2017) to serve as endogenous sequences for the simulation of next-generation sequencing data with ancient DNA damage. Data were simulated to mimic data generated with an Illumina HiSeq 2500 sequencing machine assuming the post-mortem damage pattern observed when sequencing Neandertals in Briggs et al. (2007). We simulated coverages of 0.5X and 2.0X. For each individual, fragment sizes followed a log-normal distribution with a location between 3.3 and 3.8 (randomly drawn per individual from a uniform distribution) and a scale of 0.2, corresponding to an average fragment length per individual between 27 and 46 bp. Fragments shorter than 30 bp were excluded. No contaminating sequences were simulated. Sequencing reads were then trimmed and merged with `AdapterRemoval` (Schubert et al., 2016). All reads (merged and the small proportion of unmerged) were then mapped to the haploid FASTA files representing reference genomes from the three populations ($S1$, $S2$ and $S3$) using `bwa aln` v0.7.17 (Li and Durbin, 2009) together with the commonly used non-default parameters -l 16500 (to avoid seeding), -n 0.01 and -o 2 (to allow for more gaps due to post-mortem damages and increased evolutionary distance to the reference) (Schubert et al., 2012; Oliva et al., 2021). BAM files were handled using `samtools` v1.5 (Li et al., 2009).

To ascertain SNPs, we avoided the effect of damage, sequencing errors and genotype callers, by identifying biallelic SNPs directly from the simulated genotypes, prior to the `gargammel` simulation of reads and mapping, and restricted to SNPs with a minimum allele frequency of 10% in the outgroup population $S1$. This mimics an ascertainment procedure in which SNPs are ascertained in an outgroup population, which may be common in many taxa. 100,000 SNPs were selected at random using `Plink` v1.90 (Chang et al., 2015) –thin-count. Genotype calling and downstream analysis were performed separately for the three reference genomes originating from populations $S1$, $S2$ and $S3$. Pseudohaploid calls were then generated for all individuals at these sites using `ANGSD` v0.917 (Korneliussen et al., 2014) by randomly sampling a single read per position with minimum base and mapping quality of at least 30. This step was performed using `ANGSD` with the parameters as described for the empirical data above and files were then converted to `Plink` format using `haploToPlink` distributed with `ANGSD` (Korneliussen et al., 2014). For downstream analyses, the set of SNPs was further restricted to sites with less than 50 % missing data and a minor allele frequency of at least 10% in $S1$, $S2$, $S3$ and T together. Binary and transposed `Plink` files were handled using `Plink` v1.90 (Chang et al., 2015). `convertf` (Patterson et al., 2006; Price et al., 2006) was used to convert between `Plink` and `EIGENSTRAT` file formats. `Plink` was also used for linkage disequilibrium (LD) pruning with parameters –indep-pairwise 200 25 0.7.

2.4 Estimating admixture proportions

We used four different approaches to estimate ancestry proportions in our target population T. In addition to differences in the underlying model and implementation, the tools differ in the type of their input data (genotype calls or genotype likelihoods) and whether their approaches are unsupervised and/or supervised (Table 1).

All software was set to estimate ancestry assuming two source populations. Unless stated otherwise, $S2$ and $S3$ were set as sources and T as the target population while no other individuals were included in when running the software. `ADMIXTURE` (Alexander et al., 2009; Alexander and Lange, 2011) is the only included method that has both a supervised (i.e. with pre-defined source populations) and an unsupervised mode. Both options were tested using the –haploid option without multithreading as the genotype calls were pseudo-haploid. For `qpAdm` (Haak et al., 2015; Harney et al., 2021), populations $O1$, $O2$, $O3$ and $O4$ served as "right" populations. `qpAdm` was run with the options allsnps: YES and

Table 1: Overview of the different tools used for ancestry estimation.

Method	Genotype calls	Genotype-likelihoods	Unsupervised	Supervised	Citation
ADMIXTURE	X	-	X	X	Alexander et al. (2009); Alexander and Lange (2011)
qpAdm	X	-	-	X	Haak et al. (2015); Harney et al. (2021)
NGSadmix	-	X	X	-	Skotte et al. (2013)
fastNGSadmix	-*	X	-	X	Jørsboe et al. (2017)

* source populations for fastNGSadmix can be either genotype calls or genotype likelihoods

details: YES. For fastNGSadmix (Jørsboe et al., 2017), allele frequencies in the source populations were estimated using NGSadmix (Skotte et al., 2013) with the option -printInfo 1. fastNGSadmix was then run to estimate ancestry per individual without bootstrapping. NGSadmix (Skotte et al., 2013) was run in default setting. The mean ancestry proportions across all individuals in the target population was used as an ancestry estimate for the entire population. In the case of unsupervised approaches, the clusters belonging to the source populations were identified as those where individuals from $S2$ or $S3$ showed more than 90 % estimated ancestry.

3 Results

3.1 Impact of mapping bias on allele frequency estimates in empirical data

We first tested the effect of mapping bias on allele frequency estimates in empirical data. We selected low to medium coverage (mostly between 2–4x coverage, except for one individual at 14x, Table S1) for ten individuals from each of three 1000 Genomes populations (FIN, JPT and YRI) from different continents. All individuals show an empirical bias towards the reference allele as indicated by average $r_L > 0.5$ (Tables S1 and S2). We used ANGSD to estimate allele frequencies from genotype likelihoods based on short-read NGS data (read lengths reduced to 36-54 bp to better resemble fragmented aDNA data) and compare them to allele frequencies estimated from the same individuals genotyped using a SNP array and pseudohaploid genotype data. In addition to fragmentation, deamination is a major factor contributing to mapping bias in ancient DNA due to the resulting excess of mismatches (Günther and Nettelblad, 2019; Martiniano et al., 2020), which we did not explore here. As the genotyping array does not involve a mapping step to a reference genome it should be less affected by mapping bias, we consider these estimates as "true" allele frequencies.

Overall, genotype likelihood-based point estimates of the allele frequencies tend towards more intermediate allele frequencies while pseudohaploid genotypes and "true" genotypes result in more alleles estimated to have low and high alternative allele frequency (Figure S1). In all tested populations, the default version of genotype likelihood calculation produced an allele frequency distribution slightly shifted towards lower non-reference allele frequency estimates compared to the corrected genotype likelihood (Paired Wilcoxon test $p < 2.2 \times 10^{-22}$ in all populations). Consistently, the per-site allele frequencies estimated from the corrected genotype likelihoods exhibit a slightly better correlation with the "true" frequencies (Table 2). Allele frequency estimates from pseudohaploid data display the best correlation with the "true" frequencies in all populations (Table 2).

Overall, the per-site differences between "true" allele frequencies and all frequencies estimated from NGS data (genotype-likelihoods and pseudohaploid) show a trend towards lower estimated non-reference alleles in the NGS data (Figure 2A-C), suggesting an impact of mapping bias. Outliers even reach a difference of up to -1.0. Interestingly, despite the overall closer concordance between the pseudohaploid allele frequency spectrum and the SNP array allele frequency spectrum, there is higher variation between pseudohaploid and true frequencies per-site (Figure 2A-C), suggesting that allele frequency estimates from pseudohaploid calls are relatively noisy but also relatively unbiased. A

Table 2: Pearson's correlation coefficients comparing different allele frequency estimates in the three empirical populations. 95% confidence intervals are shown in parentheses.

Population	True vs default GL	True vs. corrected GL	True vs. Pseudohaploid
FIN	0.8460 [0.8453, 0.8467]	0.8471 [0.8464, 0.8478]	0.8509 [0.8502, 0.8515]
YRI	0.8246 [0.8238, 0.8254]	0.8258 [0.8250, 0.8266]	0.8337 [0.8330, 0.8345]
JPT	0.8466 [0.8459, 0.8474]	0.8474 [0.8466, 0.8481]	0.8687 [0.8681, 0.8693]

consequence of the systematic over-estimation of the allele frequencies when using genotype likelihoods is that the population differentiation (here measured as f_2 statistic) is reduced compared to estimates from the SNP array or pseudohaploid genotype calls (Figure 2D-F). In Günther and Nettelblad (2019), we found that different parts of the human reference genome exhibit different types of mapping bias in the estimation of archaic ancestry which could be attributed to the fact that the human reference genome is a mosaic of different ancestries (Green et al., 2010; Church et al., 2015). Here, we do not find substantial differences in the allele frequency patterns between the different continental ancestries (Figures S2-S4).

3.2 Estimation of admixture proportions based on genotype calls in simulated data

We compare the accuracy of the different methods for estimating admixture proportion under a set of different population divergence times, sequencing depths, and with or without LD pruning of the SNP panel. Mapping to three different reference genomes, one from an outgroup ($S1$) and the two ingroups also representing the sources of the admixture event ($S2$ and $S3$), allows us to use $S1$ as a control which should not be affected by mapping bias and only other aspects of the data. We expect that mapping reads to one of the sources will cause a preference for reads carrying alleles from that population at heterozygous sites and, consequently, an overestimation of the ancestry proportion attributed to that population. The distance between the estimates when mapped to $S2$ or $S3$ (and their distances to the results when using $S1$) can then be seen as an estimate of the extent of mapping bias.

For most parts of this results section, we will focus on the scenario with an average sequencing depth of 0.5X where the deepest population split (t_{123}) was 50,000 generations ago and the split (t_{23}) between the relevant sources dating to 25,000 generations ago. Consequently, mapping the reads against a reference genome sequence from one or the other source would be equivalent to a study comparing (sub-)species where the reference genome originated from one of those populations. Results for other population divergences and sequencing depths are shown in Figures S5-S10.

We begin by assessing methods that require hard genotype calls, `ADMIXTURE` and `qpAdm`. For these approaches, we used single randomly drawn reads per individual and site to generate pseudo-haploid data in the target population. The popular implementation of the PSD (Pritchard et al., 2000) model working with SNP genotype calls, `ADMIXTURE` (Alexander et al., 2009; Alexander and Lange, 2011), has both supervised and unsupervised modes. Both modes show similar general patterns: low (10%) admixture proportions are estimated well while medium to high ($\geq$ 50%) admixture proportions are over-estimated (Figure 3). On the full SNP panel, the median estimated admixture proportion differs up to $\sim$ 4% when mapping to reference genomes representing either of the two sources ($S2$ or $S3$) while mapping to the outgroup reference genome ($S1$) results in estimates intermediate between the two (Data S1). LD pruning slightly reduces mapping bias and reduces the overestimation, at least for high (90%) admixture proportions. `qpAdm` (Haak et al., 2015; Harney et al., 2021), on the other hand, estimated all admixture proportions accurately when the outgroup ($S1$) was used for the reference genome sequence and when the full SNP panel was used. The median estimates of admixture differed up to 3% between mapping to reference genomes from one of the source populations ($S2$ or $S3$). Notably, LD pruning increased the noise of the `qpAdm` estimates (probably due to the reduced number of SNPs) and led to all admixture proportions being slightly underestimated (Figure 3).

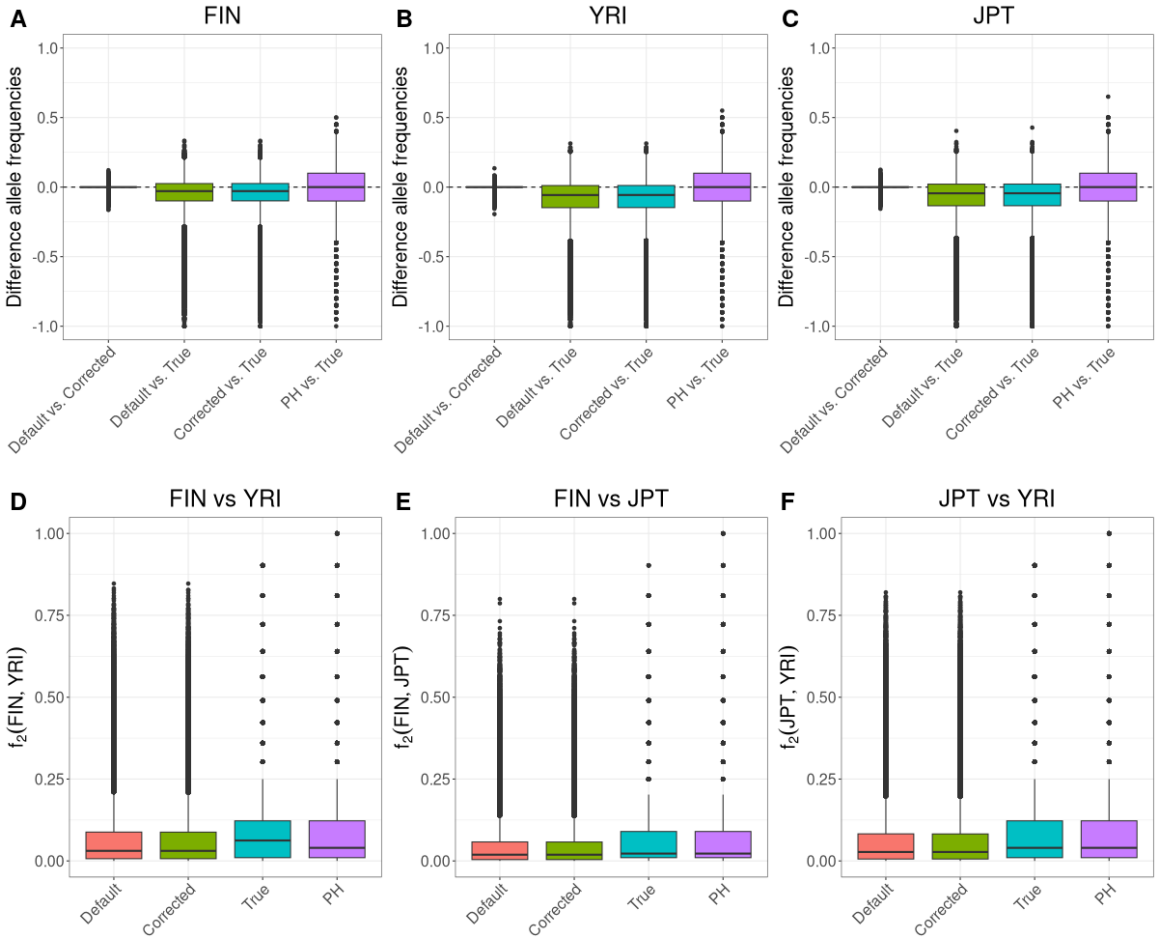

Figure 2: Differences in allele frequency estimates. Boxplots for the differences between default genotype likelihood-based estimates and corrected genotype likelihood-based estimates, default genotype likelihood-based estimates and SNP array-based estimates, corrected genotype likelihood-based estimates, pseudohaploid (PH) genotype-based and SNP array-based estimates (A) in the FIN population, (B) in the YRI population and (C) in the JPT population. (D-F) are showing boxplots of the pairwise per-site population differentiation (measured as f_2 statistic) for the four allele frequency estimates.

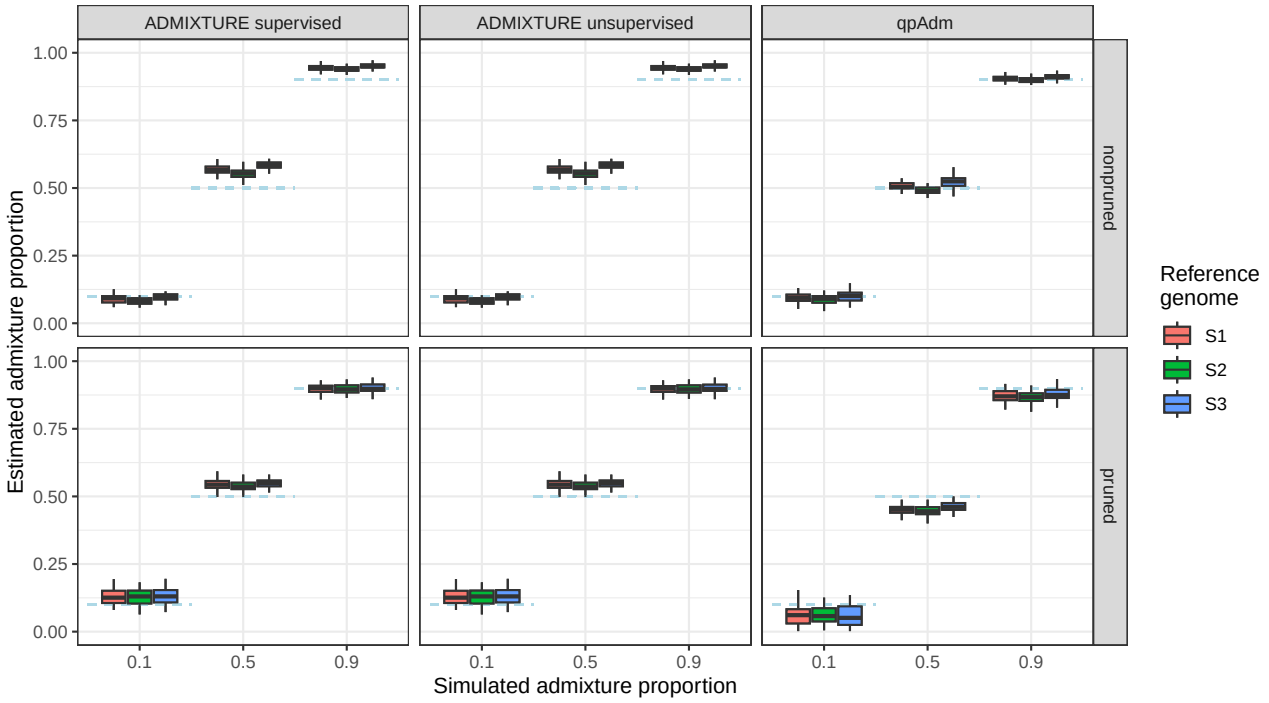

Figure 3: Simulation results for genotype call based methods using $t_{123} = 50000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

The extent of mapping bias decreases with lower population divergence between the sources across all methods (Figure S5), as mapping bias should correlate with distance to the reference genome sequence. Conversely, increasing sequencing depth mostly reduced noise but not mapping bias (Figures S6 and S9) as the genotype-based methods benefit from the increased number of SNPs but the genotype calls do not increase certainty when multiple reads are mapping to the same position.

3.3 Estimation of admixture proportions based on genotype likelihoods in simulated data

We next examined the performance of genotype-likelihood-based approaches to estimate admixture proportions. In principle, genotype likelihoods should be able to make better use of all of the data in ancient DNA, because more than a single random read can be used per site. Moreover, we are able to explicitly incorporate our mapping bias correction into the genotype likelihood. We compared the supervised `fastNGSadmix` (Jørsboe et al., 2017) to the unsupervised `NGSadmix` (Skotte et al., 2013). `fastNGSadmix` shows the highest level of overestimation of low to medium admixture proportions ($\leq 50\%$) among all tested approaches while high admixture proportions (90%) are estimated well (Figure 4). Mapping bias caused differences of up to $\sim 3\%$ in the admixture estimates when mapping to the different reference genomes. LD pruning enhances the overestimation of low admixture proportions while leading to an underestimation of high admixture proportions (Data S1). Notably, when employing the corrected genotype-likelihood the estimated admixture proportions when mapping to $S2$ or $S3$ are slightly more similar than with the default formula without correction, showing that the correction makes the genome-wide estimates less dependent on the reference sequence used for mapping while not fully removing the effect. The estimates when using the outgroup $S1$ as reference are slightly higher for high admixture proportions (90%). The results for `NGSadmix` show similar patterns to `ADMIXTURE` with a moderate overestimation of admixture proportions $\geq 50\%$ (Figure 4). Mapping bias caused differences of up to $\sim 4\%$ in the admixture estimates when mapping to the different reference genomes. After LD pruning, estimated admixture proportions for higher simulated values

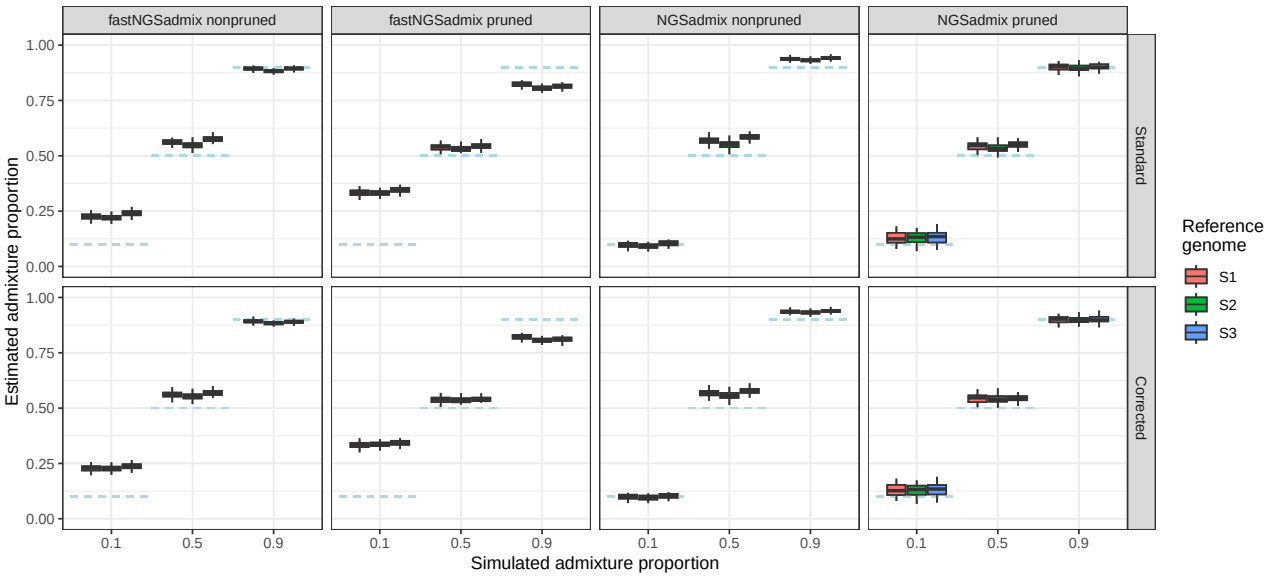

Figure 4: Simulation results for genotype likelihood based methods using $t_{123} = 50000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

were closer to the simulated values. Furthermore, employing the mapping bias corrected genotype-likelihoods made the estimated admixture proportions less dependent on the reference genome used during mapping, particularly when using `NGSadmix` in pruned data, where all three reference genomes produce nearly identical results. Notably, the extent of over-estimation for both methods seems to be somewhat negatively correlated with population divergence (Figures S7 and 4), i.e. increased distances between the source populations reduces the method bias. Further patterns are as expected: the extent of mapping bias is correlated with population divergence and increased sequencing depth reduces noise (Figures S7, 4, S8 and S10).

4 Discussion

We illustrate the impacts of mapping bias on downstream applications, such as allele frequency estimation and ancestry proportion estimation, and we introduced a new approach to recalibrate genotype likelihoods in the presence of mapping bias to alleviate its effects. The impact of mapping bias in our comparisons is small but pervasive suggesting that it can have an effect on the results of different types of analysis in empirical studies. In contrast to other approaches to alleviate mapping bias, such as employing pangenome variation graphs (Martiniano et al., 2020; Koptekin et al., 2025), it does not require establishing a separate pipeline. Instead, only reads mapping to a set of ascertained SNP positions need to be modified and remapped which only represents only a fraction of all reads and consequently will require a small proportion of the original mapping time. Our Python scripts used to calculate the genotype likelihoods could be optimized further, but this step is of minor computational costs compared to other parts of the general bioinformatic pipelines (∼1 minute per individual in the empirical data analysis for this study) in ancient DNA research. The corrected genotype likelihoods can then be directly used in downstream analyses using the same file structures and formats as other genotype likelihood-based approaches.

Increasing sample sizes in ancient DNA studies have motivated a number of studies aiming to detect selection in genome-wide scans or to investigate phenotypes in ancient populations (e.g. Mathieson et al., 2015; Cox et al., 2022; Klunk et al., 2022; Gopalakrishnan et al., 2022; Mathieson and Terhorst, 2022; Davy et al., 2023; Barton et al., 2025; Hui et al., 2024). Such investigations are potentially very sensitive to biases and uncertainties in genotype calls or allele frequencies at individual sites while

certain effects will average out for genome-wide estimates such as ancestry proportions. Concerns about certain biases and how to estimate allele frequencies have even reduced confidence in the results of some studies searching for loci under selection (Gopalakrishnan et al., 2022; Barton et al., 2025). Our results indicate that such concerns are valid as individual sites can show very strong deviations when allele frequencies are estimated from low-coverage sequencing data (Figure 2). This is due to a combination of effects, including mapping bias. Without high coverage data, genotype likelihood approaches without an allele frequency prior will naturally put some weight on all three potential genotypes at a site, ultimately collectively driving allele frequency to more intermediate values. The risk is then that most downstream analyses will treat the allele frequency point estimates at face value, potentially leading to both false positives and negatives. While our new approach to recalibrate genotype likelihoods reduces the number of outlier loci, there is still uncertainty in allele frequency estimates from low coverage data. Therefore, results heavily relying on allele frequency estimates or genotype calls at single loci from low-coverage sequencing data or even ancient DNA data need to be taken with a grain of salt.

The simulations in this study revealed a modest but noticeable effect of mapping bias on ancestry estimates as the difference between reference genomes never exceeded 5 percent. In particular, we found that mapping bias and method bias even counteract each other in certain cases, leading to better estimates of the admixture proportion when mapping to one of the sources (see also Akopyan et al., 2024; Yuan et al., 2024). The differences seen in our simulations are likely underestimates of what might occur in empirical studies, because real genomes are larger and more complex than what we used in the simulations. For instance, we simulated five 20 megabase long chromosomes for a 100 megabase genome, while mammalian genomes are one order of magnitude larger; the human genome is roughly 3 gigabases and the shortest human chromosome alone is ∼45 megabases long. Furthermore, the only added complexity when generating the random sequences was a GC content of 41%. Real genomes also experience more complex mutation events involving translocations and duplications, which, together with the increased length and the presence of repetitive elements, should increase mapping bias in empirical studies. Finally, the range of possible demographic histories including the relationships of targets and sources, the amount of drift, and the timing and number of gene flow events is impossible to explore in a simulation study. The restricted scenarios tested in this study should affect the quantitative results but the qualitative interpretation of mapping bias impacting ancestry estimates should extend beyond the specific model used in the simulations.

While the ancestry estimates depended slightly on the reference genome the reads were mapped to, they seemed more influenced by the choice of method or software. Methods differed by more than 10% in their ancestry estimates from the same source data. This highlights that other factors and biases play major roles in the performance of these methods. Depending on the method, the type of input data, and the implementation, they showed different sensitivities to e.g. linkage or the amount of missing data (which was on average ∼37% per SNP for the 0.5x and ∼3% for the 2.0x simulations). For non-pruned data, `qpAdm` performed best across all scenarios and did not show any method-specific bias in certain ranges of simulated admixture proportions. Multiple differences between the PSD and `qpAdm` methods may have contributed to the relative biases we observed. PSD models may propagate allele-frequency misestimation more than `qpAdm` because of their assumptions of linkage equilibrium and Hardy-Weinberg equilibrium. Indeed, we observed that LD pruning improved the performance of PSD models, but they are known to be sensitive to sample size and drift (e.g. Lawson et al., 2018; Toyama et al., 2020). More generally, because it is based on Patterson's f statistics (Patterson et al., 2012), `qpAdm` estimates ancestry from relative differences. If mapping bias affects all populations similarly, then their relative relationships remain more stable. In contrast, PSD models reconstruct exact allele frequencies for the putative source populations therefore emphasizing the impact of mapping bias. Finally, the ancestry proportions of PSD models are constrained to $[0, 1]$ which is not the case for `qpAdm`. Indeed, we see negative estimates in a small number of simulations (3 runs with 0.5X depth and 50,000 generations divergence). This (biologically unrealistic) flexibility of `qpAdm` compared to PSD models drives the mean estimated admixture admixture proportion down,

which may account for some of the reduction in upward method bias compared to the other methods.

Broadly speaking, our results support the common practice of using `qpAdm` in most human ancient DNA studies. However, the requirement of data from additional, "right" populations, may make it difficult to apply to many non-human species. Furthermore, `qpAdm` only works with genotype calls, so it is influenced by mapping bias in similar ways as `ADMIXTURE` and these methods cannot benefit from the newly introduced genotype likelihood estimation. We also need to note that we tested `qpAdm` under almost ideal settings in our simulations with left and right populations clearly separated and without gene flow between them. More thorough assessments of the performance of `qpAdm` can be found elsewhere (Harney et al., 2021; Yüncü et al., 2023). In our simulations, unsupervised PSD-model approaches (`ADMIXTURE`, `NGSadmix`) work as well as or even better than supervised PSD-model approaches (`ADMIXTURE`, `fastNGSadmix`) in estimating the ancestry proportions in the target population. `ADMIXTURE` and `NGSadmix` benefit from LD pruning while LD pruning increases the method bias for `fastNGSadmix` and introduces method bias for `qpAdm`.

Genotype likelihood-based methods for estimating ancestry proportions are not commonly used in human ancient DNA studies (but genotype likelihoods are popular as input for imputation pipelines). This may be surprising, because genotype-likelihood-based approaches are targeted at low coverage data, exactly as seen in ancient DNA studies. However, the definition of "low coverage" differs between fields. While most working with modern DNA would understand 2-4x as "low depth", the standards for ancient DNA researchers are typically much lower due to limited DNA preservation. Genotype likelihood methods perform much better with >1x coverage, an amount of data that is not within reach for most ancient DNA samples investigated so far (**?**). The large body of known, common polymorphic sites in human populations allows the use of pseudohaploid calls at those positions instead. Nonetheless, this study highlights that unsupervised methods employing genotype-likelihoods (`NGSadmix`) can reach similar accuracies as methods such as `ADMIXTURE` that require (pseudo-haploid) genotype calls. Moreover, methods that incorporate genotype likelihoods have the added benefit that the modified genotype likelihood estimation approach can be used to reduce the effect of mapping bias. Furthermore, if some samples in the dataset have >1x depth, genotype likelihood-based approaches will benefit from the additional data and provide more precise estimates of ancestry proportions while pseudo-haploid data will not gain any information from more than one read at a position. Finally, genotype likelihoods are very flexible and can be adjusted for many other aspects of the data. For example, variations of genotype likelihood estimators exist that incorporate the effect of post-mortem damage (Hofmanová et al., 2016; Link et al., 2017; Kousathanas et al., 2017) allowing use of all sequence data without filtering for potentially damaged sites or enzymatic repair of the damages in the wet lab.

As the main aim of this study was to show the general impact of mapping bias and introduce a modified genotype likelihood, we opted for a comparison of some of the most popular methods with a limited set of settings. This was done in part to limit the computational load of this study. We also decided to not set this up as a systematic assessment of different factors influencing mapping bias. The effects of fragmentation (shorter fragments increasing bias, Günther and Nettelblad, 2019), deamination damage (deamination increasing the number of mismatches and bias, Martiniano et al., 2020) and mapping algorithm/parameters (Dolenz et al., 2024) on mapping bias have been explored elsewhere. Our simulations were restricted to one mapping software (*bwa aln*) and the commonly used mapping quality threshold of 30. Mapping quality calculations differ substantially between tools and algorithms making their impact on mapping bias not directly comparable (Dolenz et al., 2024). For *bwa aln* (Li and Durbin, 2009), it has been suggested that a mapping quality threshold of 25 (the value assigned when the maximum number of mismatches is reached) reduces mapping bias (e.g. Martiniano et al., 2020; Dolenz et al., 2024), and we also see a reduction in mapping bias when using these thresholds (Figures S11-S14). Therefore, a general suggestion for users of *bwa aln* should be to use 25 as the mapping quality cutoff. However, many users are using other mappers (e.g. *bowtie*, Langmead and Salzberg, 2012) in their research, and adjusted genotype likelihoods allow correcting for mapping bias independent of the mapping software and its specifics in calculating mapping quality

values. Our results reiterate that mapping bias can skew results in studies using low-coverage data as is the case in most ancient DNA studies. Different strategies exist for mitigating these effects and we added a modified genotype likelihood approach to the population genomic toolkit. Nevertheless, none of these methods will be the ideal solution in all cases and they will not always fully remove the potential effect of mapping bias, making proper verification and critical presentation of all results crucial.

Acknowledgements

We thank Kay Prüfer for feedback on the preprint and Gabriel Renaud for making code for connecting `msprime` and `gargammel` available on Github. The computations were enabled by resources in projects SNIC 2017/7-259, SNIC 2018/8-6, SNIC 2021/2-17, SNIC 2022/22-874, NAISS 2023/22-883, sllstore2017087, UPPMAX 2023/2-30 and NAISS 2023/2-19 provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Uppmax, partially funded by Uppsala University and the Swedish Research Council through grant agreements no. 2022-06725 and no. 2018-05973.

Funding

TG was supported by grants from the Swedish Research Council Vetenskapsrådet (2017-05267) and Svenska Forskningsrådet Formas (2023-01381).

Conflict of interest disclosure

The authors declare they have no conflict of interest relating to the content of this article. Torsten Günther is a recommender for PCI Genomics and PCI Evolutionary Biology.

Data, script and code availability

Raw data for the boxplots can be found in Data S1. Code used in this study can be found under https://github.com/tgue/refbias_GL with a snapshot of the version used for this revision available on Zenodo (https://doi.org/10.5281/zenodo.14505750). Empirical data from the 1000 genomes project is available from their resources: SNP array data (http://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/hd_genotype_chip/ALL.chip.omni_broad_sanger_combined.20140818.snps.genotypes.vcf.gz) and low coverage sequencing data (https://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/).

References

M. Akopyan, M. Genchev, E. E. Armstrong, and J. A. Mooney. Divergent reference genomes compromise the reconstruction of demographic histories, selection scans, and population genetic summary statistics, Nov. 2024. URL https://www.biorxiv.org/content/10.1101/2024.11.26.625554v1. Pages: 2024.11.26.625554 Section: New Results.

D. H. Alexander and K. Lange. Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC Bioinformatics*, 12(1):246, June 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-246. URL https://doi.org/10.1186/1471-2105-12-246.

D. H. Alexander, J. Novembre, and K. Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009. ISSN 1088-9051. Number: 9 Publisher: Cold Spring Harbor Lab.

A. Auton, G. R. Abecasis, D. M. Altshuler, R. M. Durbin, G. R. Abecasis, D. R. Bentley, A. Chakravarti, A. G. Clark, P. Donnelly, E. E. Eichler, P. Flicek, S. B. Gabriel, R. A. Gibbs, E. D. Green, M. E. Hurles, B. M. Knoppers, J. O. Korbel, E. S. Lander, C. Lee, H. Lehrach, E. R. Mardis, G. T. Marth, G. A. McVean, D. A. Nickerson, J. P. Schmidt, S. T. Sherry, J. Wang, R. K. Wilson, R. A. Gibbs, E. Boerwinkle, H. Doddapaneni, Y. Han, V. Korchina, C. Kovar, S. Lee, D. Muzny, J. G. Reid, Y. Zhu, Y. Chang, Q. Feng, X. Fang, X. Guo, M. Jian, H. Jiang, X. Jin, T. Lan, G. Li, J. Li, Y. Li, S. Liu, X. Liu, Y. Lu, X. Ma, M. Tang, B. Wang, G. Wang, H. Wu, R. Wu, X. Xu, Y. Yin, D. Zhang, W. Zhang, J. Zhao, M. Zhao, X. Zheng, E. S. Lander, D. M. Altshuler, S. B. Gabriel, N. Gupta, N. Gharani, L. H. Toji, N. P. Gerry, A. M. Resch, P. Flicek, J. Barker, L. Clarke, L. Gil, S. E. Hunt, G. Kelman, E. Kulesha, R. Leinonen, W. M. McLaren, R. Radhakrishnan, A. Roa, D. Smirnov, R. E. Smith, I. Streeter, A. Thormann, I. Toneva, B. Vaughan, X. Zheng-Bradley, D. R. Bentley, R. Grocock, S. Humphray, T. James, Z. Kingsbury, H. Lehrach, R. Sudbrak, M. W. Albrecht, V. S. Amstislavskiy, T. A. Borodina, M. Lienhard, F. Mertes, M. Sultan, B. Timmermann, M.-L. Yaspo, E. R. Mardis, R. K. Wilson, L. Fulton, R. Fulton, S. T. Sherry, V. Ananiev, Z. Belaia, D. Beloslyudtsev, N. Bouk, C. Chen, D. Church, R. Cohen, C. Cook, J. Garner, T. Hefferon, M. Kimelman, C. Liu, J. Lopez, P. Meric, C. O'Sullivan, Y. Ostapchuk, L. Phan, S. Ponomarov, V. Schneider, E. Shekhtman, K. Sirotkin, D. Slotta, H. Zhang, G. A. McVean, R. M. Durbin, S. Balasubramaniam, J. Burton, P. Danecek, T. M. Keane, A. Kolb-Kokocinski, S. McCarthy, J. Stalker, M. Quail, J. P. Schmidt, C. J. Davies, J. Gollub, T. Webster, B. Wong, Y. Zhan, A. Auton, C. L. Campbell, Y. Kong, A. Marcketta, R. A. Gibbs, F. Yu, L. Antunes, M. Bainbridge, D. Muzny, A. Sabo, Z. Huang, J. Wang, L. J. M. Coin, L. Fang, X. Guo, X. Jin, G. Li, Q. Li, Y. Li, Z. Li, H. Lin, B. Liu, R. Luo, H. Shao, Y. Xie, C. Ye, C. Yu, F. Zhang, H. Zheng, H. Zhu, C. Alkan, E. Dal, F. Kahveci, G. T. Marth, E. P. Garrison, D. Kural, W.-P. Lee, W. Fung Leong, M. Stromberg, A. N. Ward, J. Wu, M. Zhang, M. J. Daly, M. A. DePristo, R. E. Handsaker, D. M. Altshuler, E. Banks, G. Bhatia, G. del Angel, S. B. Gabriel, G. Genovese, N. Gupta, H. Li, S. Kashin, E. S. Lander, S. A. McCarroll, J. C. Nemesh, R. E. Poplin, S. C. Yoon, J. Lihm, V. Makarov, A. G. Clark, S. Gottipati, A. Keinan, J. L. Rodriguez-Flores, J. O. Korbel, T. Rausch, M. H. Fritz, A. M. Stütz, P. Flicek, K. Beal, L. Clarke, A. Datta, J. Herrero, W. M. McLaren, G. R. S. Ritchie, R. E. Smith, D. Zerbino, X. Zheng-Bradley, P. C. Sabeti, I. Shlyakhter, S. F. Schaffner, J. Vitti, D. N. Cooper, E. V. Ball, P. D. Stenson, D. R. Bentley, M. Bauer, R. Keira Cheetham, A. Cox, M. Eberle, S. Humphray, S. Kahn, L. Murray, J. Peden, R. Shaw, E. E. Kenny, M. A. Batzer, M. K. Konkel, J. A. Walker, D. G. MacArthur, M. Lek, R. Sudbrak, V. S. Amstislavskiy, R. Herwig, E. R. Mardis, L. Ding, D. C. Koboldt, D. Larson, and K. Ye. A global reference for human genetic variation. *Nature*, 526(7571):68–74, Oct. 2015. ISSN 1476-4687. doi: 10.1038/nature15393. URL https://www.nature.com/articles/nature15393. Publisher: Nature Publishing Group.

A. R. Barton, C. G. Santander, P. Skoglund, I. Moltke, D. Reich, and I. Mathieson. Insufficient evidence for natural selection associated with the Black Death. *Nature*, 638(8051):E19–E22, Feb. 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08496-5. URL https://www.nature.com/articles/s41586-024-08496-5. Publisher: Nature Publishing Group.

L. A. Bergeron, S. Besenbacher, J. Zheng, P. Li, M. F. Bertelsen, B. Quintard, J. I. Hoffman, Z. Li, J. St. Leger, C. Shao, J. Stiller, M. T. P. Gilbert, M. H. Schierup, and G. Zhang. Evolution of the germline mutation rate across vertebrates. *Nature*, 615(7951):285–291, Mar. 2023. ISSN 1476-4687. doi: 10.1038/s41586-023-05752-y. URL https://www.nature.com/articles/s41586-023-05752-y. Publisher: Nature Publishing Group.

A. W. Briggs, U. Stenzel, P. L. Johnson, R. E. Green, J. Kelso, K. Prüfer, M. Meyer, J. Krause, M. T. Ronan, M. Lachmann, and others. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences*, 104(37):14616–14621, 2007.

C. C. Chang, C. C. Chow, L. C. Tellier, S. Vattikuti, S. M. Purcell, and J. J. Lee. Second-generation

PLINK: rising to the challenge of larger and richer datasets. *Gigascience*, 4(1):s13742–015, 2015. ISSN 2047-217X. Number: 1 Publisher: Oxford University Press.

B. Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, Mar. 2009. ISSN 1471-0064. doi: 10.1038/nrg2526. URL https://www.nature.com/articles/nrg2526. Publisher: Nature Publishing Group.

N.-C. Chen, B. Solomon, T. Mun, S. Iyer, and B. Langmead. Reference flow: reducing reference bias using multiple population genomes. *Genome Biology*, 22(1):8, Jan. 2021. ISSN 1474-760X. doi: 10.1186/s13059-020-02229-3. URL https://doi.org/10.1186/s13059-020-02229-3.

D. M. Church, V. A. Schneider, K. M. Steinberg, M. C. Schatz, A. R. Quinlan, C.-S. Chin, P. A. Kitts, B. Aken, G. T. Marth, M. M. Hoffman, J. Herrero, M. L. Z. Mendoza, R. Durbin, and P. Flicek. Extending reference assembly models. *Genome Biology*, 16(1):13, Jan. 2015. ISSN 1465-6906. doi: 10.1186/s13059-015-0587-3. URL https://doi.org/10.1186/s13059-015-0587-3.

S. L. Cox, H. M. Moots, J. T. Stock, A. Shbat, B. D. Bitarello, N. Nicklisch, K. W. Alt, W. Haak, E. Rosenstock, C. B. Ruff, and I. Mathieson. Predicting skeletal stature using ancient DNA. *American Journal of Biological Anthropology*, 177(1):162–174, 2022. ISSN 2692-7691. doi: 10.1002/ajpa.24426. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ajpa.24426. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ajpa.24426.

T. Davy, D. Ju, I. Mathieson, and P. Skoglund. Hunter-gatherer admixture facilitated natural selection in Neolithic European farmers. *Current Biology*, 33(7):1365–1371.e3, Apr. 2023. ISSN 0960-9822. doi: 10.1016/j.cub.2023.02.049. URL https://www.sciencedirect.com/science/article/pii/S0960982223001896.

S. Dolenz, T. van der Valk, C. Jin, J. Oppenheimer, M. B. Sharif, L. Orlando, B. Shapiro, L. Dalén, and P. D. Heintzman. Unravelling reference bias in ancient DNA datasets. *Bioinformatics*, 40(7): btae436, July 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae436. URL https://doi.org/10.1093/bioinformatics/btae436.

B. L. Dumont and B. A. Payseur. EVOLUTION OF THE GENOMIC RATE OF RECOMBINATION IN MAMMALS. *Evolution*, 62(2):276–294, Feb. 2008. ISSN 0014-3820. doi: 10.1111/j.1558-5646.2007.00278.x. URL https://doi.org/10.1111/j.1558-5646.2007.00278.x.

D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4):1567–1587, 2003. ISSN 0016-6731. Number: 4.

D. Falush, M. Stephens, and J. K. Pritchard. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Molecular ecology notes*, 7(4):574–578, 2007. ISSN 1471-8278. Number: 4.

M. Fumagalli, F. G. Vieira, T. S. Korneliussen, T. Linderoth, E. Huerta-Sánchez, A. Albrechtsen, and R. Nielsen. Quantifying Population Genetic Differentiation from Next-Generation Sequencing Data. *Genetics*, 195(3):979–992, Nov. 2013. ISSN 1943-2631. doi: 10.1534/genetics.113.154740. URL https://doi.org/10.1534/genetics.113.154740.

S. Gopalakrishnan, J. A. Samaniego Castruita, M.-H. S. Sinding, L. F. K. Kuderna, J. Räikkönen, B. Petersen, T. Sicheritz-Ponten, G. Larson, L. Orlando, T. Marques-Bonet, A. J. Hansen, L. Dalén, and M. T. P. Gilbert. The wolf reference genome sequence (Canis lupus lupus) and its implications for Canis spp. population genomics. *BMC Genomics*, 18:495, June 2017. ISSN 1471-2164. doi: 10.1186/s12864-017-3883-3. URL https://doi.org/10.1186/s12864-017-3883-3.

S. Gopalakrishnan, S. S. Ebenesersdóttir, I. K. C. Lundstrøm, G. Turner-Walker, K. H. S. Moore, P. Luisi, A. Margaryan, M. D. Martin, M. R. Ellegaard, Magnússon, Sigursson, S. Snorradóttir, D. N. Magnúsdóttir, J. E. Laffoon, L. van Dorp, X. Liu, I. Moltke, M. C. Ávila Arcos, J. G. Schraiber, S. Rasmussen, D. Juan, P. Gelabert, T. de Dios, A. K. Fotakis, M. Iraeta-Orbegozo, J. Vågene, S. D. Denham, A. Christophersen, H. K. Stenøien, F. G. Vieira, S. Liu, T. Günther, T. Kivisild, O. G. Moseng, B. Skar, C. Cheung, M. Sandoval-Velasco, N. Wales, H. Schroeder, P. F. Campos, V. B. Gumundsdóttir, T. Sicheritz-Ponten, B. Petersen, J. Halgunset, E. Gilbert, G. L. Cavalleri, E. Hovig, I. Kockum, T. Olsson, L. Alfredsson, T. F. Hansen, T. Werge, E. Willerslev, F. Balloux, T. Marques-Bonet, C. Lalueza-Fox, R. Nielsen, K. Stefánsson, A. Helgason, and M. T. P. Gilbert. The population genomic legacy of the second plague pandemic. *Current Biology*, 32 (21):4743–4751.e6, Nov. 2022. ISSN 0960-9822. doi: 10.1016/j.cub.2022.09.023. URL https://www.sciencedirect.com/science/article/pii/S0960982222014671.

R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, and M. H.-Y. Fritz. A draft sequence of the Neandertal genome. *science*, 328(5979):710–722, 2010. ISSN 0036-8075. Number: 5979 Publisher: American Association for the Advancement of Science.

T. Günther and M. Jakobsson. Population genomic analyses of DNA from ancient remains. In *Handbook of statistical genomics*, pages 295–324. John Wiley & Sons, 4th edition, 2019. ISBN 1-119-42914-5.

T. Günther and C. Nettelblad. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLOS Genetics*, 15(7):e1008302, July 2019. ISSN 1553-7404. doi: 10.1371/journal.pgen.1008302. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1008302.

W. Haak, I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, S. Nordenfelt, E. Harney, and K. Stewardson. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature*, 522(7555):207–211, 2015. ISSN 1476-4687. Number: 7555 Publisher: Nature Publishing Group.

K. Hanghøj, I. Moltke, P. A. Andersen, A. Manica, and T. S. Korneliussen. Fast and accurate relatedness estimation from high-throughput sequencing data in the presence of inbreeding. *GigaScience*, 8(5), May 2019. ISSN 2047-217X. doi: 10.1093/gigascience/giz034. URL https://doi.org/10.1093/gigascience/giz034.

E. Harney, N. Patterson, D. Reich, and J. Wakeley. Assessing the performance of qpAdm: a statistical tool for studying population admixture. *Genetics*, 217(4), Apr. 2021. ISSN 1943-2631. doi: 10.1093/genetics/iyaa045. URL https://doi.org/10.1093/genetics/iyaa045.

P. D. Heintzman, G. D. Zazula, R. D. MacPhee, E. Scott, J. A. Cahill, B. K. McHorse, J. D. Kapp, M. Stiller, M. J. Wooller, L. Orlando, J. Southon, D. G. Froese, and B. Shapiro. A new genus of horse from Pleistocene North America. *eLife*, 6, 2017. ISSN 2050-084X. doi: 10.7554/eLife.29944.

Z. Hofmanová, S. Kreutzer, G. Hellenthal, C. Sell, Y. Diekmann, D. Díez-del Molino, L. van Dorp, S. López, A. Kousathanas, V. Link, and others. Early farmers from across Europe directly descended from Neolithic Aegeans. *Proceedings of the National Academy of Sciences*, page 201523951, 2016.

L. Huang, V. Popic, and S. Batzoglou. Short read alignment with populations of genomes. *Bioinformatics*, 29(13):i361–i370, July 2013. ISSN 1367-4803. doi: 10.1093/bioinformatics/btt215. URL https://doi.org/10.1093/bioinformatics/btt215.

M. J. Hubisz, D. Falush, M. Stephens, and J. K. Pritchard. Inferring weak population structure with the assistance of sample group information. *Molecular ecology resources*, 9(5):1322–1332, 2009. ISSN 1755-098X. Number: 5.

R. Hui, C. L. Scheib, E. D'Atanasio, S. A. Inskip, C. Cessford, S. A. Biagini, A. W. Wohns, M. Q. Ali, S. J. Griffith, A. Solnik, H. Niinemäe, X. J. Ge, A. K. Rose, O. Beneker, T. C. O'Connell, J. E. Robb, and T. Kivisild. Genetic history of Cambridgeshire before and after the Black Death. *Science Advances*, 10(3):eadi5903, Jan. 2024. doi: 10.1126/sciadv.adi5903. URL https://www.science.org/doi/10.1126/sciadv.adi5903. Publisher: American Association for the Advancement of Science.

E. Jørsboe, K. Hanghøj, and A. Albrechtsen. fastNGSadmix: admixture proportions and principal component analysis of a single NGS sample. *Bioinformatics*, 33(19):3148–3150, 2017.

J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

J. Klunk, T. P. Vilgalys, C. E. Demeure, X. Cheng, M. Shiratori, J. Madej, R. Beau, D. Elli, M. I. Patino, R. Redfern, S. N. DeWitte, J. A. Gamble, J. L. Boldsen, A. Carmichael, N. Varlik, K. Eaton, J.-C. Grenier, G. B. Golding, A. Devault, J.-M. Rouillard, V. Yotova, R. Sindeaux, C. J. Ye, M. Bikaran, A. Dumaine, J. F. Brinkworth, D. Missiakas, G. A. Rouleau, M. Steinrücken, J. Pizarro-Cerdá, H. N. Poinar, and L. B. Barreiro. Evolution of immune genes is associated with the Black Death. *Nature*, 611(7935):312–319, Nov. 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05349-x. URL https://www.nature.com/articles/s41586-022-05349-x. Number: 7935 Publisher: Nature Publishing Group.

D. Koptekin, E. Yapar, K. B. Vural, E. Sağlıcan, N. E. Altınışık, A.-S. Malaspinas, C. Alkan, and M. Somel. Pre-processing of paleogenomes: mitigating reference bias and postmortem damage in ancient genome data. *Genome Biology*, 26(1):6, Jan. 2025. ISSN 1474-760X. doi: 10.1186/s13059-024-03462-w. URL https://doi.org/10.1186/s13059-024-03462-w.

T. S. Korneliussen and I. Moltke. NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics*, 31(24):4009–4011, 2015. ISSN 1460-2059. Number: 24 Publisher: Oxford University Press.

T. S. Korneliussen, I. Moltke, A. Albrechtsen, and R. Nielsen. Calculation of Tajima's D and other neutrality test statistics from low depth next-generation sequencing data. *BMC bioinformatics*, 14:289, Oct. 2013. ISSN 1471-2105. doi: 10.1186/1471-2105-14-289.

T. S. Korneliussen, A. Albrechtsen, and R. Nielsen. ANGSD: Analysis of Next Generation Sequencing Data. *BMC bioinformatics*, 15(1):356, 2014. ISSN 1471-2105. doi: 10.1186/s12859-014-0356-4.

A. Kousathanas, C. Leuenberger, V. Link, C. Sell, J. Burger, and D. Wegmann. Inferring Heterozygosity from Ancient and Low Coverage Genomes. *Genetics*, 205(1):317–332, Jan. 2017. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.116.189985. URL http://www.genetics.org/content/205/1/317.

E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson,

S. Wenning, T. Slezak, N. Doggett, J.-F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H.-C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. R. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. A. Smit, E. Stupka, J. Szustakowki, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S.-P. Yang, R.-F. Yeh, F. Collins, M. S. Guyer, J. Peterson, A. Felsenfeld, K. A. Wetterstrand, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, A. Patrinos, M. J. Morgan, International Human Genome Sequencing Consortium, C. f. G. R. Whitehead Institute for Biomedical Research, The Sanger Centre:, Washington University Genome Sequencing Center, US DOE Joint Genome Institute:, Baylor College of Medicine Human Genome Sequencing Center:, RIKEN Genomic Sciences Center:, Genoscope and CNRS UMR-8030:, I. o. M. B. Department of Genome Analysis, GTC Sequencing Center:, Beijing Genomics Institute/Human Genome Center:, T. I. f. S. B. Multimegabase Sequencing Center, Stanford Genome Technology Center:, University of Oklahoma's Advanced Center for Genome Technology:, Max Planck Institute for Molecular Genetics:, L. A. H. G. C. Cold Spring Harbor Laboratory, GBF—German Research Centre for Biotechnology:, a. i. i. l. u. o. h. *Genome Analysis Group (listed in alphabetical order, U. N. I. o. H. Scientific management: National Human Genome Research Institute, Stanford Human Genome Center:, University of Washington Genome Center:, K. U. S. o. M. Department of Molecular Biology, University of Texas Southwestern Medical Center at Dallas:, U. D. o. E. Office of Science, and The Wellcome Trust:. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb. 2001. ISSN 1476-4687. doi: 10.1038/35057062. URL https://www.nature.com/articles/35057062. Number: 6822 Publisher: Nature Publishing Group.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9(4): 357–359, Apr. 2012. ISSN 1548-7105. doi: 10.1038/nmeth.1923. URL https://www.nature.com/articles/nmeth.1923. Number: 4 Publisher: Nature Publishing Group.

D. J. Lawson, L. van Dorp, and D. Falush. A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots. *Nature Communications*, 9(1):3258, Aug. 2018. ISSN 2041-1723. doi: 10.1038/s41467-018-05257-7. URL https://www.nature.com/articles/s41467-018-05257-7. Number: 1 Publisher: Nature Publishing Group.

H. Li and R. Durbin. Fast and accurate short read alignment with Burrows–Wheeler transform. *bioinformatics*, 25(14):1754–1760, 2009. ISSN 1367-4803. Number: 14 Publisher: Oxford University Press.

H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. The Sequence Alignment/Map format and

SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, Aug. 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp352.

V. Link, A. Kousathanas, K. Veeramah, C. Sell, A. Scheu, and D. Wegmann. ATLAS: analysis tools for low-depth and ancient samples. *bioRxiv*, page 105346, 2017.

R. N. Lou, A. Jacobs, A. P. Wilder, and N. O. Therkildsen. A beginner's guide to low-coverage whole genome sequencing for population genomics. *Molecular Ecology*, 30(23):5966–5993, 2021. ISSN 1365-294X. doi: 10.1111/mec.16077. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.16077. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/mec.16077.

R. Martiniano, E. Garrison, E. R. Jones, A. Manica, and R. Durbin. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology*, 21(1):250, Sept. 2020. ISSN 1474-760X. doi: 10.1186/s13059-020-02160-7. URL https://doi.org/10.1186/s13059-020-02160-7.

I. Mathieson and J. Terhorst. Direct detection of natural selection in Bronze Age Britain. *Genome Research*, 32(11-12):2057–2067, Nov. 2022. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.276862.122. URL https://genome.cshlp.org/content/32/11-12/2057. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

I. Mathieson, I. Lazaridis, N. Rohland, S. Mallick, N. Patterson, S. A. Roodenberg, E. Harney, K. Stewardson, D. Fernandes, M. Novak, and others. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528(7583):499–503, 2015.

I. Mathieson, F. Abascal, L. Vinner, P. Skoglund, C. Pomilla, P. Mitchell, C. Arthur, D. Gurdasani, E. Willerslev, M. S. Sandhu, and G. Dewar. An Ancient Baboon Genome Demonstrates Long-Term Population Continuity in Southern Africa. *Genome Biology and Evolution*, 12(4):407–412, Apr. 2020. ISSN 1759-6653. doi: 10.1093/gbe/evaa019. URL https://doi.org/10.1093/gbe/evaa019.

A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, Sept. 2010. ISSN 1088-9051. doi: 10.1101/gr.107524.110. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2928508/.

J. Meisner and A. Albrechtsen. Inferring population structure and admixture proportions in low-depth NGS data. *Genetics*, 210(2):719–731, 2018. ISSN 1943-2631. Number: 2 Publisher: Oxford University Press.

R. Nielsen, J. S. Paul, A. Albrechtsen, and Y. S. Song. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics*, 12(6):443, 2011.

A. K. Nøhr, K. Hanghøj, G. Garcia-Erill, Z. Li, I. Moltke, and A. Albrechtsen. NGSremix: a software tool for estimating pairwise relatedness between admixed individuals from next-generation sequencing data. *G3*, (jkab174), May 2021. ISSN 2160-1836. doi: 10.1093/g3journal/jkab174. URL https://doi.org/10.1093/g3journal/jkab174.

A. Oliva, R. Tobler, A. Cooper, B. Llamas, and Y. Souilmi. Systematic benchmark of ancient DNA read mapping. *Briefings in Bioinformatics*, (bbab076), Apr. 2021. ISSN 1477-4054. doi: 10.1093/bib/bbab076. URL https://doi.org/10.1093/bib/bbab076.

L. Orlando, A. Ginolhac, G. Zhang, D. Froese, A. Albrechtsen, M. Stiller, M. Schubert, E. Cappellini, B. Petersen, I. Moltke, P. L. F. Johnson, M. Fumagalli, J. T. Vilstrup, M. Raghavan,

T. Korneliussen, A.-S. Malaspinas, J. Vogt, D. Szklarczyk, C. D. Kelstrup, J. Vinther, A. Dolocan, J. Stenderup, A. M. V. Velazquez, J. Cahill, M. Rasmussen, X. Wang, J. Min, G. D. Zazula, A. Seguin-Orlando, C. Mortensen, K. Magnussen, J. F. Thompson, J. Weinstock, K. Gregersen, K. H. Røed, V. Eisenmann, C. J. Rubin, D. C. Miller, D. F. Antczak, M. F. Bertelsen, S. Brunak, K. A. S. Al-Rasheid, O. Ryder, L. Andersson, J. Mundy, A. Krogh, M. T. P. Gilbert, K. Kjær, T. Sicheritz-Ponten, L. J. Jensen, J. V. Olsen, M. Hofreiter, R. Nielsen, B. Shapiro, J. Wang, and E. Willerslev. Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, 499(7456):74–78, July 2013. ISSN 1476-4687. doi: 10.1038/nature12323. URL https://www.nature.com/articles/nature12323. Bandiera_abtest: a Cg_type: Nature Research Journals Number: 7456 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Evolutionary genetics Subject_term_id: evolutionary-genetics.

L. Orlando, R. Allaby, P. Skoglund, C. Der Sarkissian, P. W. Stockhammer, M. C. Ávila Arcos, Q. Fu, J. Krause, E. Willerslev, A. C. Stone, and C. Warinner. Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1):1–26, Feb. 2021. ISSN 2662-8449. doi: 10.1038/s43586-020-00011-0. URL https://www.nature.com/articles/s43586-020-00011-0. Number: 1 Publisher: Nature Publishing Group.

N. Patterson, A. L. Price, and D. Reich. Population structure and eigenanalysis. *PLoS genetics*, 2 (12):e190, 2006. ISSN 1553-7390. Number: 12 Publisher: Public Library of Science San Francisco, USA.

N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich. Ancient admixture in human history. *Genetics*, 192(3):1065–1093, 2012. ISSN 1943-2631. Number: 3 Publisher: Oxford University Press.

A. Prasad, E. D. Lorenzen, and M. V. Westbury. Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. *Molecular Ecology Resources*, 22(1):45–55, 2022. ISSN 1755-0998. doi: 10.1111/1755-0998.13457. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13457. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13457.

A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006. ISSN 1546-1718. Number: 8 Publisher: Nature Publishing Group.

J. K. Pritchard, M. Stephens, and P. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000. ISSN 0016-6731. Number: 2.

K. Prüfer. snpAD: An ancient DNA genotype caller. *Bioinformatics*, 2018. doi: 10.1093/bioinformatics/bty507. URL https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/bty507/5042170.

G. Renaud, K. Hanghøj, E. Willerslev, and L. Orlando. gargammel: a sequence simulator for ancient DNA. *Bioinformatics*, 33(4):577–579, Feb. 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btw670. URL https://academic.oup.com/bioinformatics/article/33/4/577/2608651.

A. R. Rogers, R. J. Bohlender, and C. D. Huff. Early history of Neanderthals and Denisovans. *Proceedings of the National Academy of Sciences*, 114(37):9859–9863, Sept. 2017. doi: 10.1073/pnas.1706426114. URL https://www.pnas.org/doi/10.1073/pnas.1706426114. Publisher: Proceedings of the National Academy of Sciences.

N. Rohland, S. Mallick, M. Mah, R. Maier, N. Patterson, and D. Reich. Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs. *Genome Research*, 32(11-12): 2068–2078, Nov. 2022. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.276728.122. URL https:

//genome.cshlp.org/content/32/11-12/2068. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

S. Rubinacci, D. M. Ribeiro, R. J. Hofmeister, and O. Delaneau. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nature Genetics*, 53(1):120–126, Jan. 2021. ISSN 1546-1718. doi: 10.1038/s41588-020-00756-0. URL https://www.nature.com/articles/s41588-020-00756-0. Number: 1 Publisher: Nature Publishing Group.

C. M. Schlebusch, H. Malmström, T. Günther, P. Sjödin, A. Coutinho, H. Edlund, A. R. Munters, M. Vicente, M. Steyn, H. Soodyall, M. Lombard, and M. Jakobsson. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science*, 358(6363): 652–655, Nov. 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aao6266. URL http://science.sciencemag.org/content/358/6363/652.

M. Schubert, A. Ginolhac, S. Lindgreen, J. F. Thompson, K. A. AL-Rasheid, E. Willerslev, A. Krogh, and L. Orlando. Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, 13:178, May 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-178. URL https://doi.org/10.1186/1471-2164-13-178.

M. Schubert, S. Lindgreen, and L. Orlando. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC research notes*, 9(1):1–7, 2016. ISSN 1756-0500. Number: 1 Publisher: BioMed Central.

L. Skotte, T. S. Korneliussen, and A. Albrechtsen. Estimating individual admixture proportions from next generation sequencing data. *Genetics*, 195(3):693–702, 2013. ISSN 1943-2631. Number: 3 Publisher: Oxford University Press.

D.-M. J. Thorburn, K. Sagonas, M. Binzer-Panchal, F. J. J. Chain, P. G. D. Feulner, E. Bornberg-Bauer, T. B. H. Reusch, I. E. Samonte-Padilla, M. Milinski, T. L. Lenz, and C. Eizaguirre. Origin matters: Using a local reference genome improves measures in population genomics. *Molecular Ecology Resources*, 23(7):1706–1723, 2023. ISSN 1755-0998. doi: 10.1111/1755-0998.13838. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13838. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13838.

K. S. Toyama, P.-A. Crochet, and R. Leblois. Sampling schemes and drift can bias admixture proportions inferred by structure. *Molecular Ecology Resources*, 20(6):1769–1785, 2020. ISSN 1755-0998. doi: 10.1111/1755-0998.13234. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1755-0998.13234. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1755-0998.13234.

T. van der Valk, C. M. Gonda, H. Silegowa, S. Almanza, I. Sifuentes-Romero, T. B. Hart, J. A. Hart, K. M. Detwiler, and K. Guschanski. The Genome of the Endangered Dryas Monkey Provides New Insights into the Evolutionary History of the Vervets. *Molecular Biology and Evolution*, 37(1):183–194, Jan. 2020. ISSN 0737-4038. doi: 10.1093/molbev/msz213. URL https://doi.org/10.1093/molbev/msz213.

J. Yuan, J. Hu, W. Liu, S. Chen, F. Zhang, S. Wang, Z. Zhang, L. Wang, B. Xiao, F. Li, M. Hofreiter, X. Lai, M. V. Westbury, and G. Sheng. Camelus knoblochi genome reveals the complex evolutionary history of Old World camels. *Current Biology*, 34(11):2502–2508.e5, June 2024. ISSN 0960-9822. doi: 10.1016/j.cub.2024.04.050. URL https://www.cell.com/current-biology/abstract/S0960-9822(24)00524-4. Publisher: Elsevier.

E. Yüncü, U. Işıldak, M. P. Williams, C. D. Huber, L. A. Vyazov, P. Changmai, and P. Flegontov. False discovery rates of qpAdm-based screens for genetic admixture. *bioRxiv*, page 2023.04.25.538339, Apr.

2023. doi: 10.1101/2023.04.25.538339. URL https://www.biorxiv.org/content/10.1101/2023.04.25.538339v1. Section: New Results.

Supplementary Figures

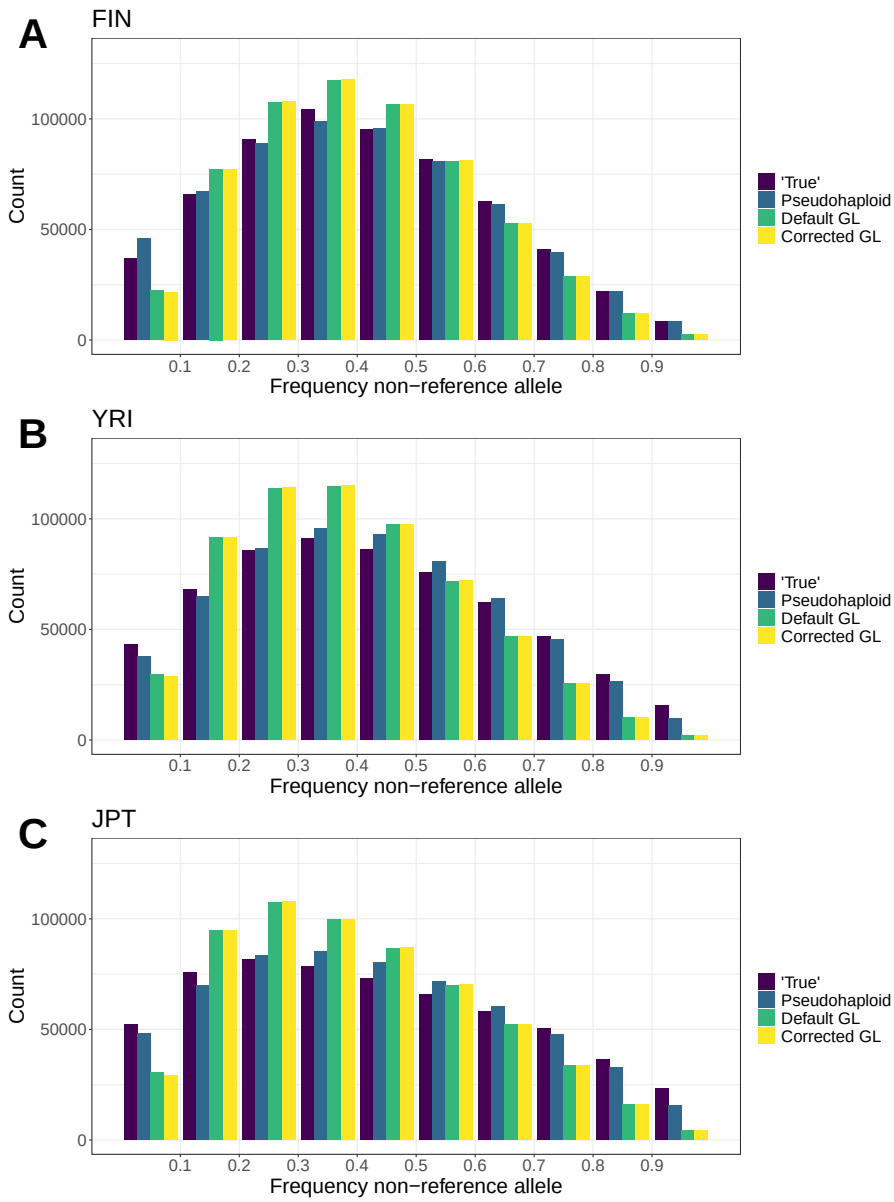

Figure S1: Binned spectrum of non-reference alleles in FIN (A), YRI (B) and JPT (C) for the four different estimation methods. Note that the specific ascertainment of common SNPs in the joint genotyping data contributes to the enrichment of variants with (true) intermediate frequencies.

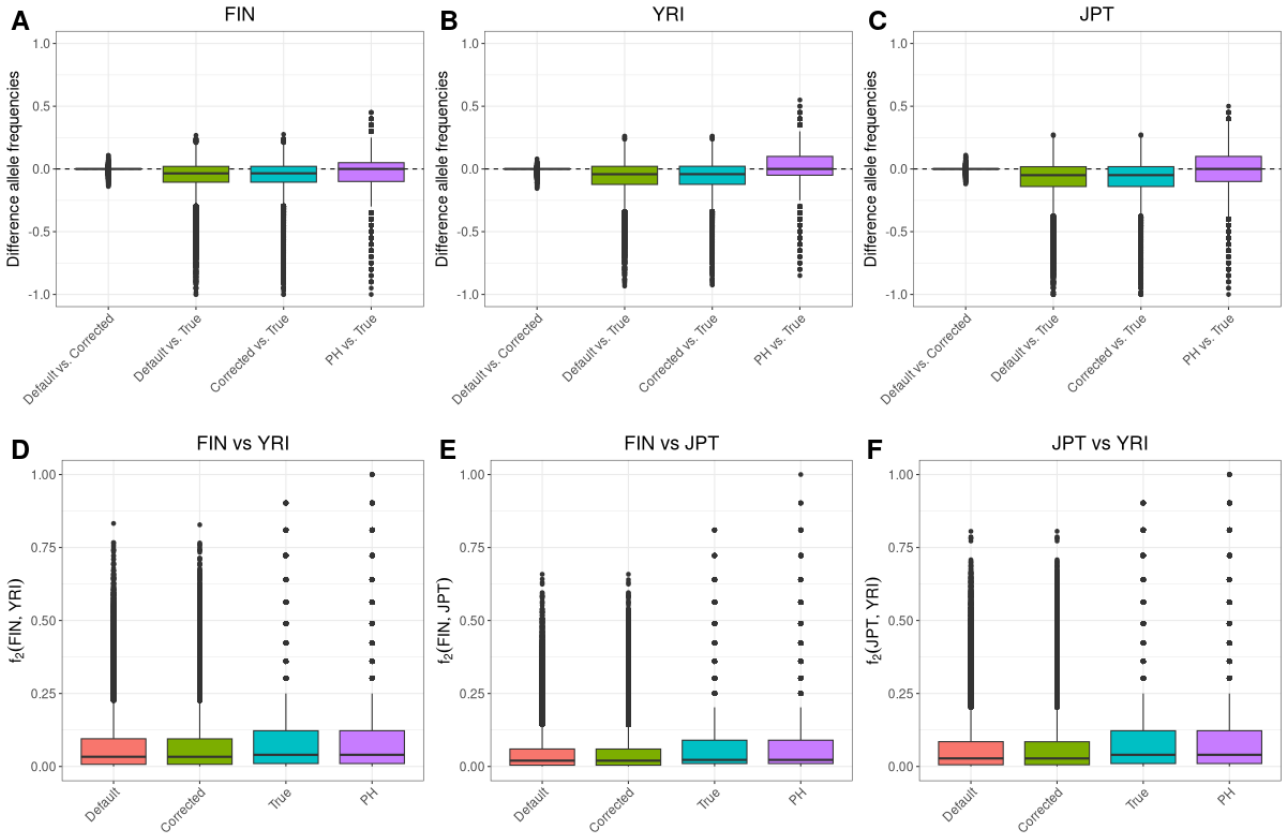

Figure S2: Differences in allele frequency estimates in the parts of the reference genome attributed to African ancestry. Boxplots for the differences between default genotype likelihood-based estimates and corrected genotype likelihood-based estimates, default genotype likelihood-based estimates and SNP array-based estimates, corrected genotype likelihood-based estimates, pseudohaploid (PH) genotype-based and SNP array-based estimates (A) in the FIN population and (B) in the YRI population. (C) is showing boxplots of the per-site population differentiation (measured as f_2 statistic) for the four allele frequency estimates.

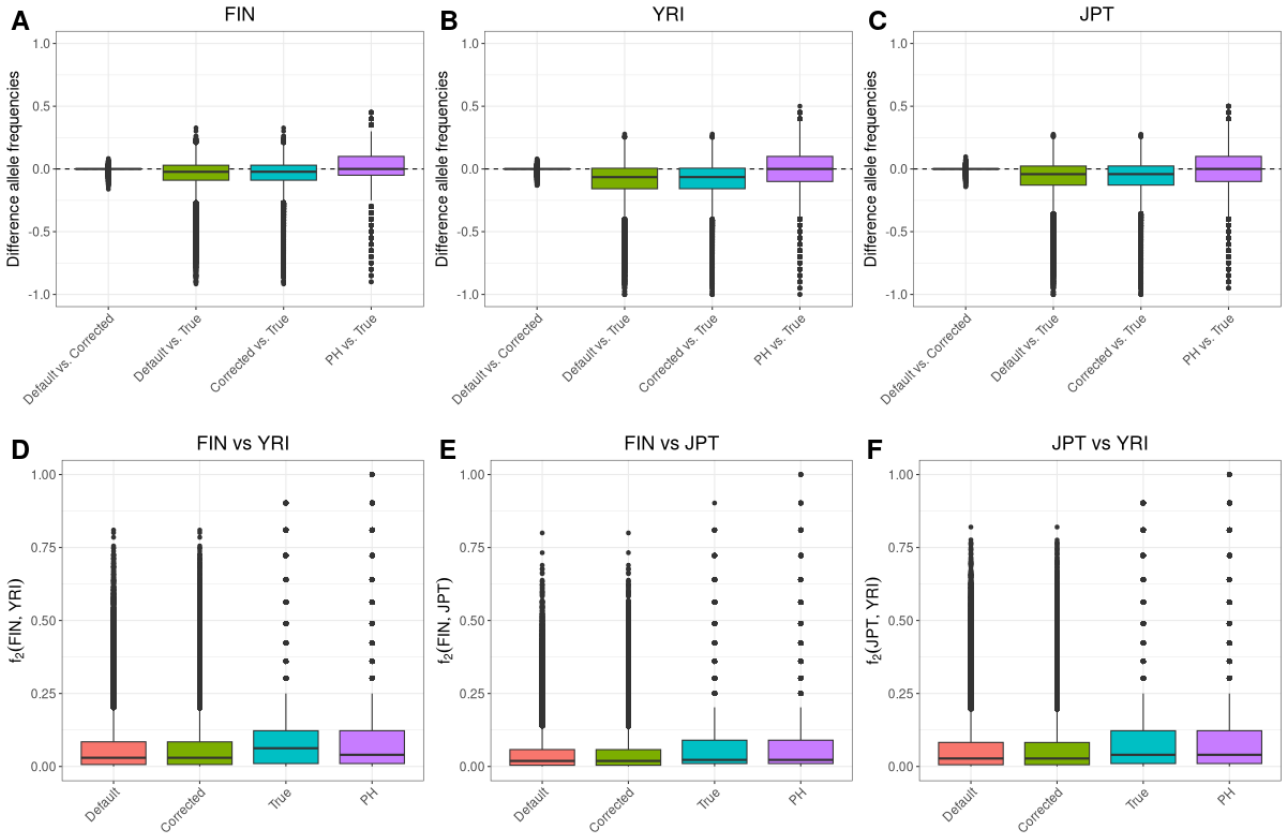

Figure S3: Differences in allele frequency estimates in the parts of the reference genome attributed to European ancestry. Boxplots for the differences between default genotype likelihood-based estimates and corrected genotype likelihood-based estimates, default genotype likelihood-based estimates and SNP array-based estimates, corrected genotype likelihood-based estimates, pseudohaploid (PH) genotype-based and SNP array-based estimates (A) in the FIN population and (B) in the YRI population. (C) is showing boxplots of the per-site population differentiation (measured as f_2 statistic) for the four allele frequency estimates.

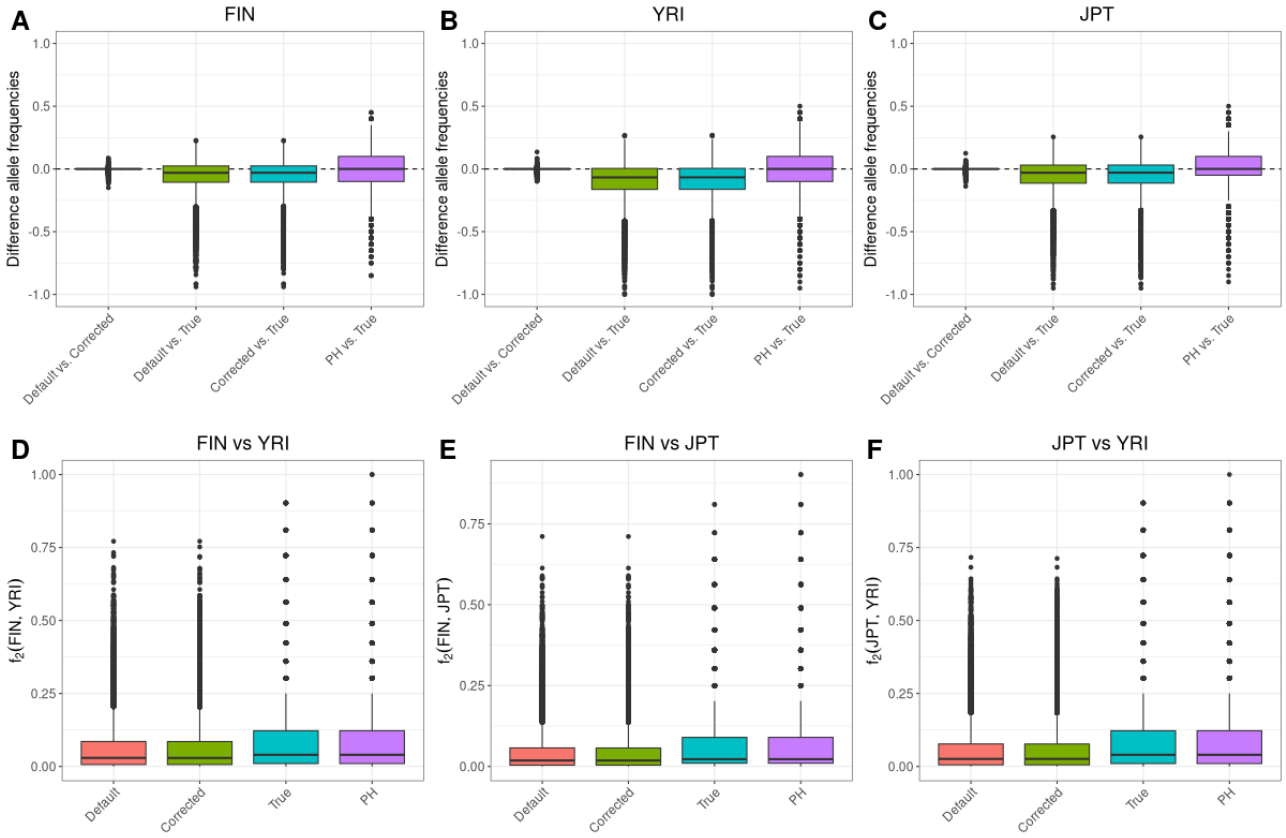

Figure S4: Differences in allele frequency estimates in the parts of the reference genome attributed to East Asian ancestry. Boxplots for the differences between default genotype likelihood-based estimates and corrected genotype likelihood-based estimates, default genotype likelihood-based estimates and SNP array-based estimates, corrected genotype likelihood-based estimates, pseudohaploid (PH) genotype-based and SNP array-based estimates (A) in the FIN population and (B) in the YRI population. (C) is showing boxplots of the per-site population differentiation (measured as f_2 statistic) for the four allele frequency estimates.

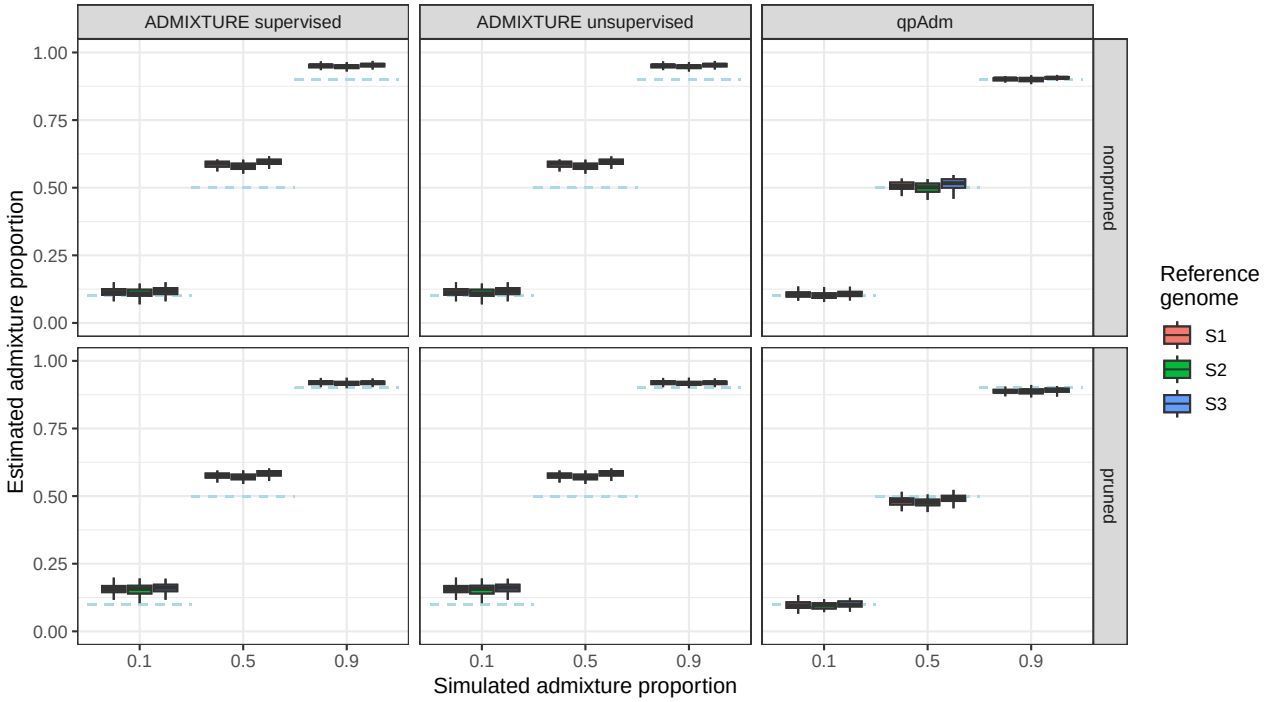

Figure S5: Simulation results for genotype call based methods using $t_{123} = 20000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

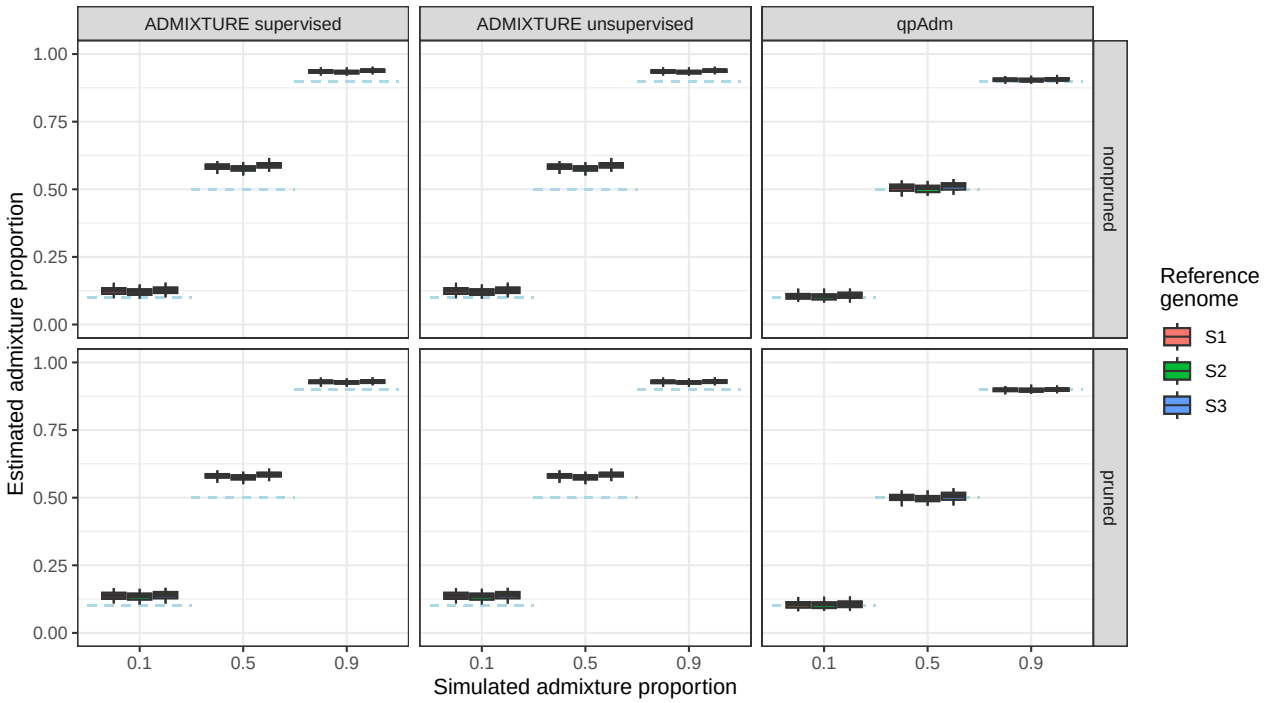

Figure S6: Simulation results for genotype call based methods using $t_{123} = 20000$ generations and a sequencing depth of 2.0X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

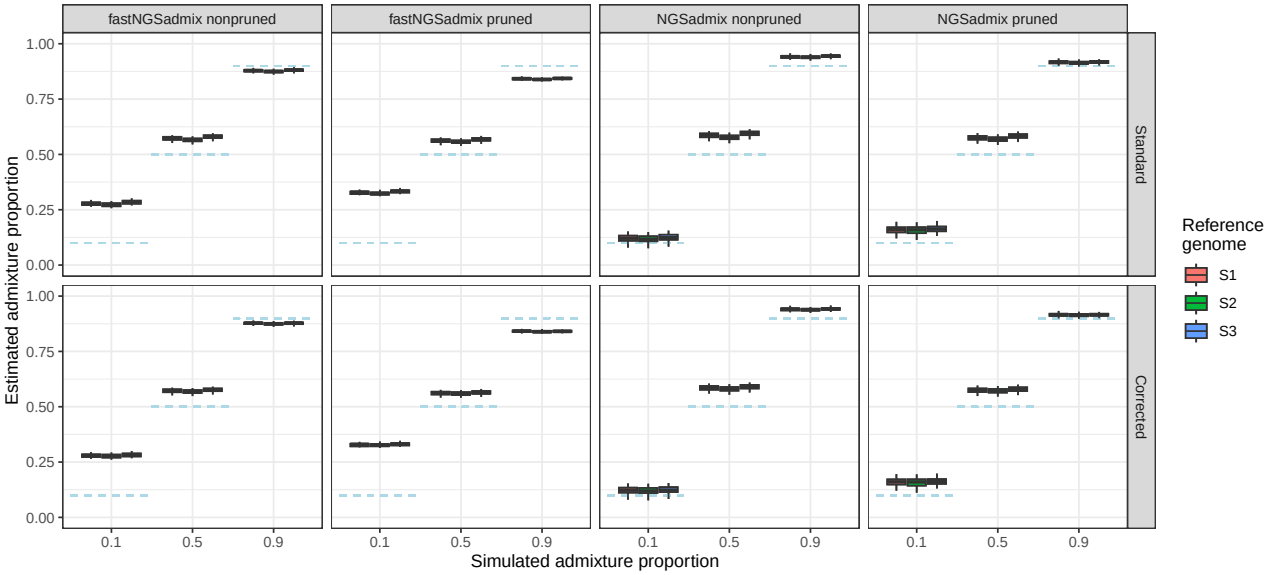

Figure S7: Simulation results for genotype likelihood based methods using $t_{123} = 20000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

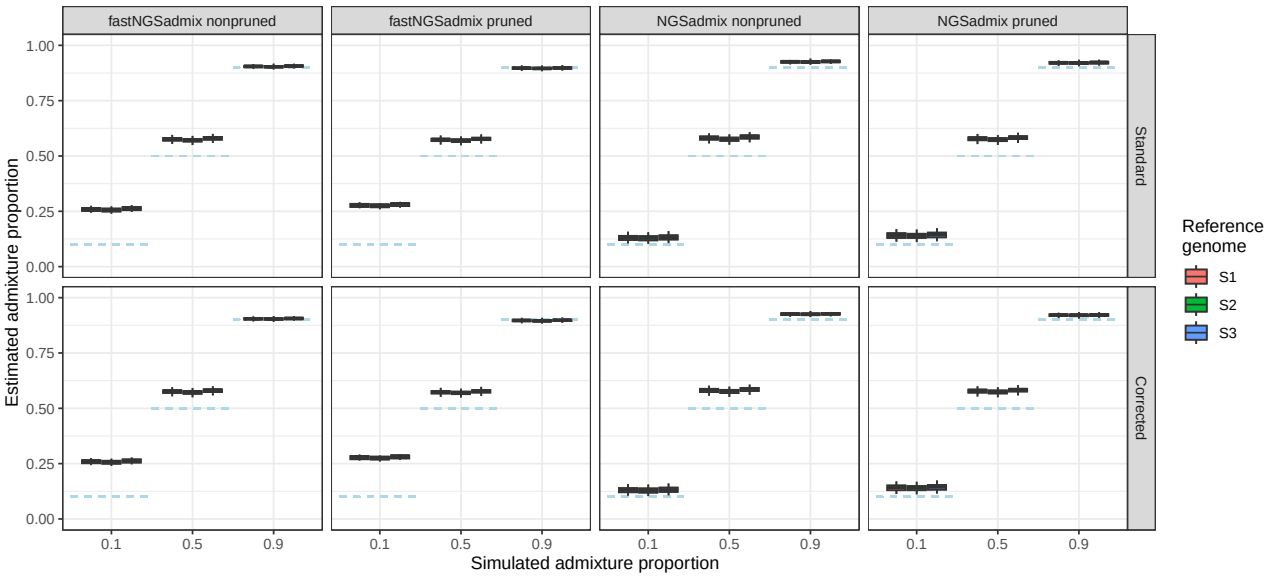

Figure S8: Simulation results for genotype likelihood based methods using $t_{123} = 20000$ generations and a sequencing depth of 2.0X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

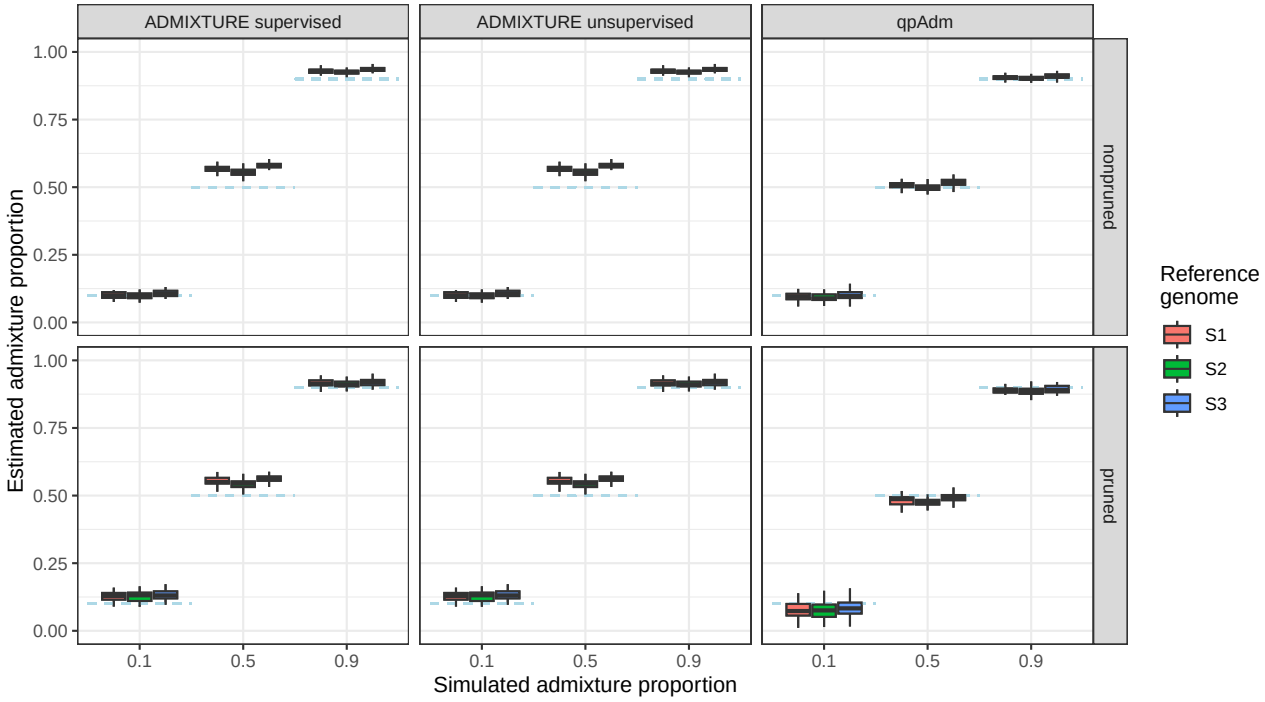

Figure S9: Simulation results for genotype call based methods using $t_{123} = 50000$ generations and a sequencing depth of 2.0X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

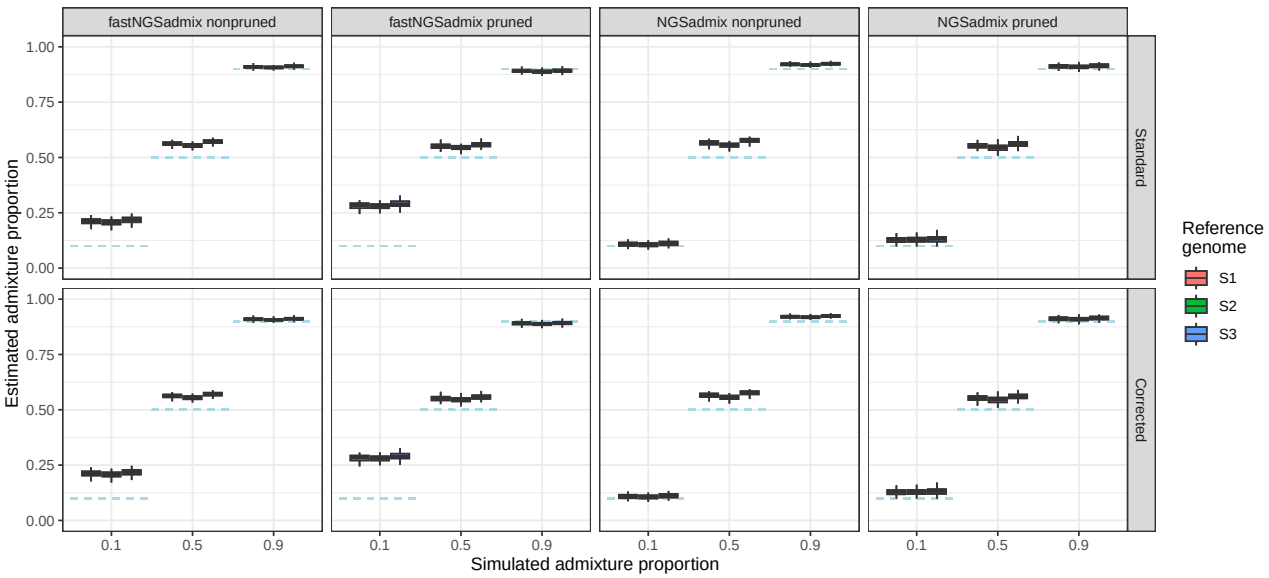

Figure S10: Simulation results for genotype likelihood based methods using $t_{123} = 50000$ generations and a sequencing depth of 2.0X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago.

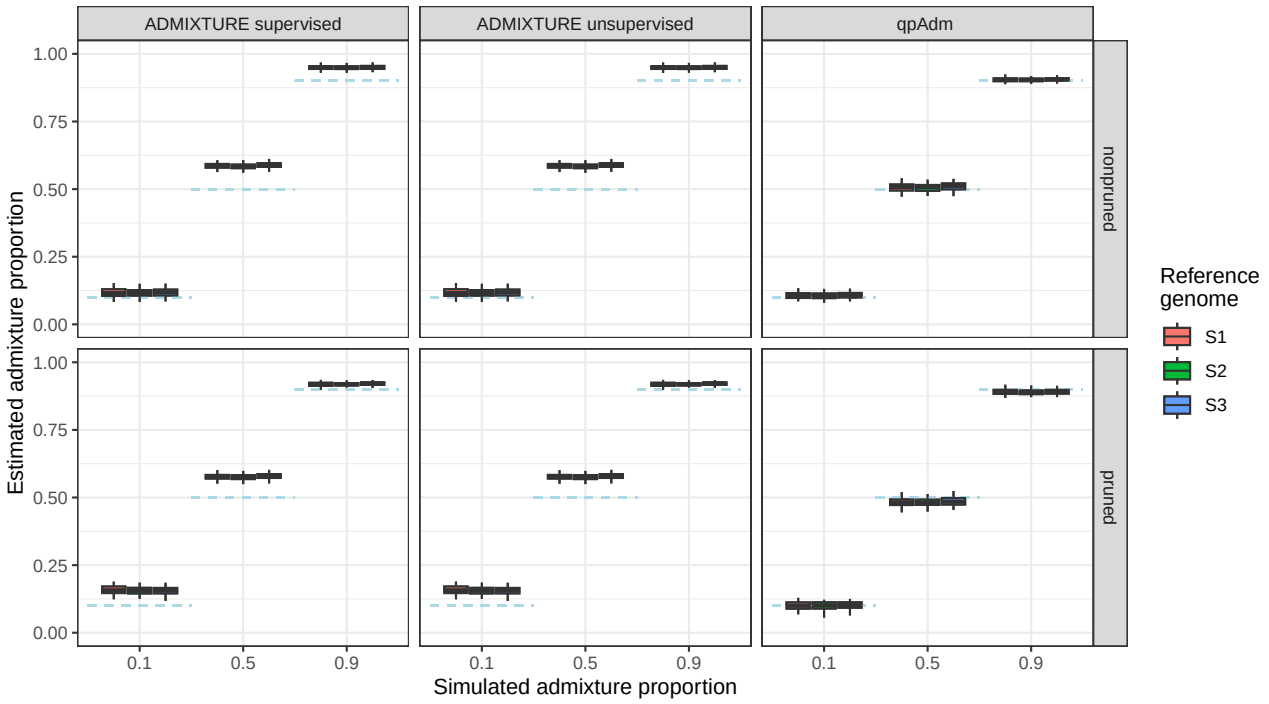

Figure S11: Simulation results for genotype call based methods using $t_{123} = 20000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago. For this run, the mapping quality threshold was set to 25 instead of 30 as in all other runs.

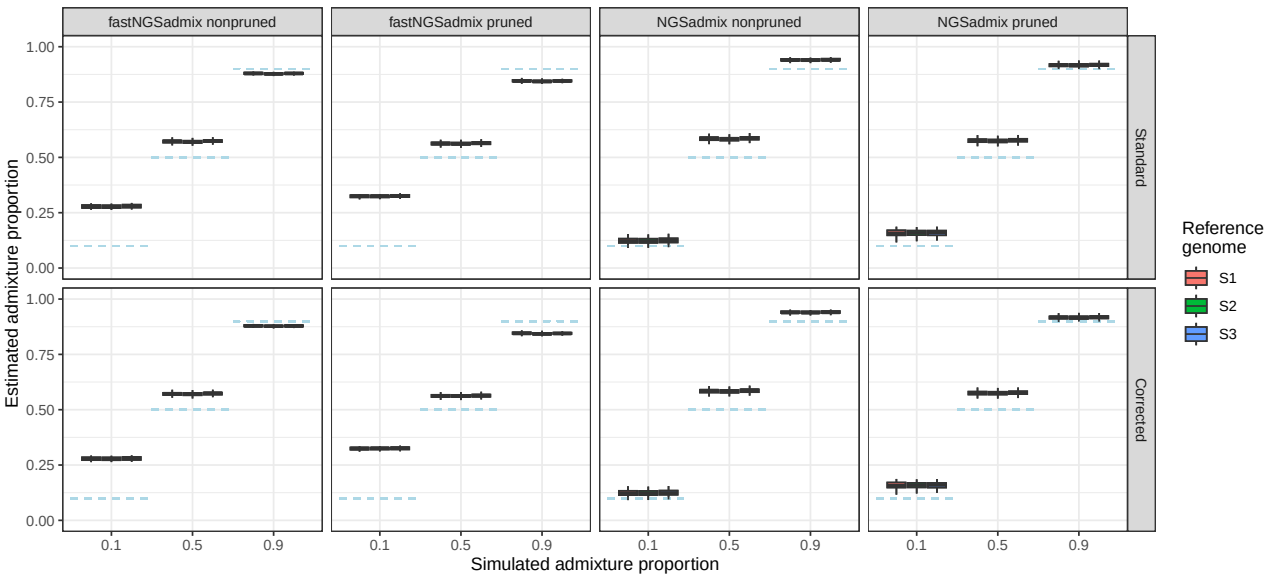

Figure S12: Simulation results for genotype likelihood based methods using $t_{123} = 20000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago. For this run, the mapping quality threshold was set to 25 instead of 30 as in all other runs.

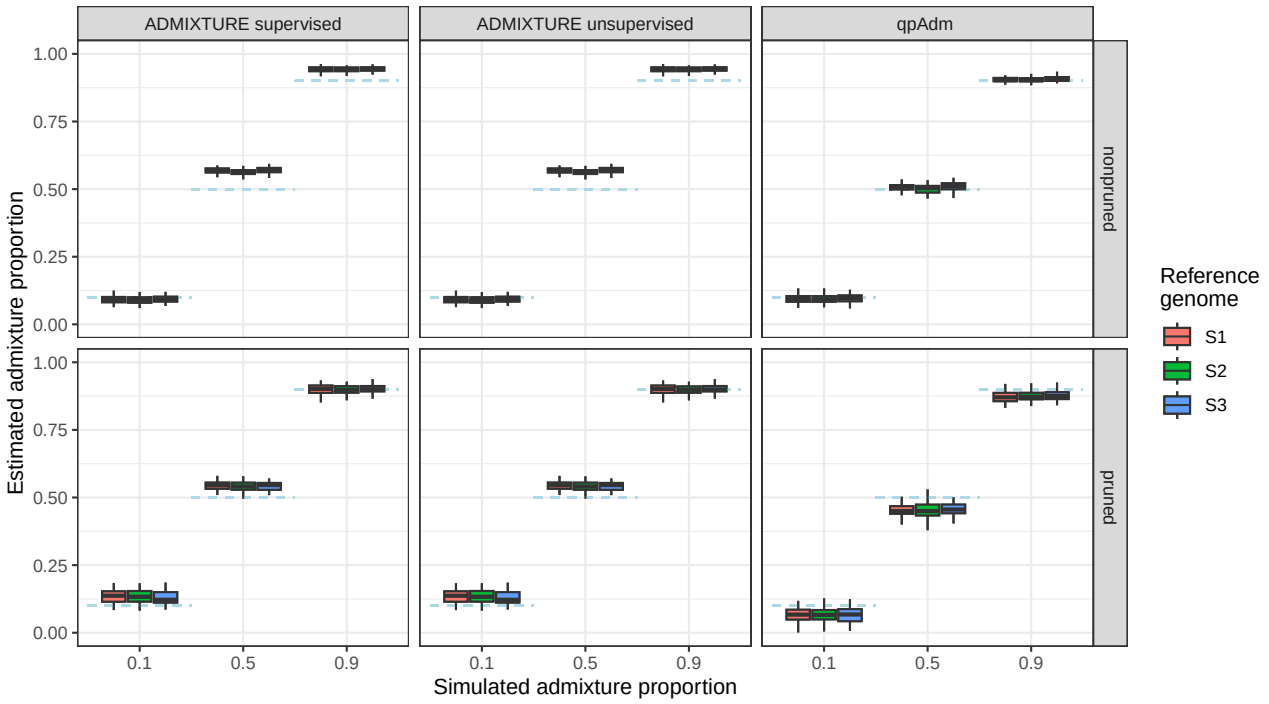

Figure S13: Simulation results for genotype call based methods using $t_{123} = 50000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago. For this run, the mapping quality threshold was set to 25 instead of 30 as in all other runs.

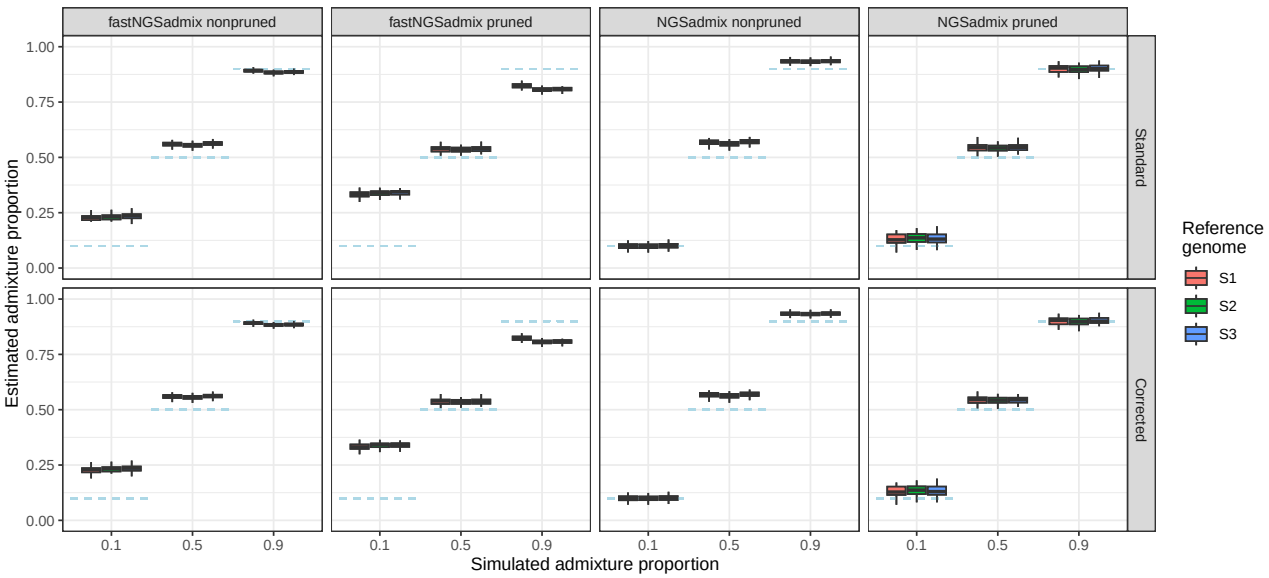

Figure S14: Simulation results for genotype likelihood based methods using $t_{123} = 50000$ generations and a sequencing depth of 0.5X. Dashed blue lines represent the simulated admixture proportions, i.e. the gene flow received from $S3$ 500 generations ago. For this run, the mapping quality threshold was set to 25 instead of 30 as in all other runs.

Supplementary Tables

Table S1: 1000 genomes individuals used for the analysis of empirical data.

Individual	Population	Autosomal sequencing depth	Average original read length	Average r_L
HG00171	FIN	3.12803	108	0.5031
HG00177	FIN	3.43327	108	0.5023
HG00189	FIN	3.48314	108	0.5026
HG00190	FIN	3.089	108	0.5023
HG00272	FIN	3.61242	108	0.5027
HG00277	FIN	3.86275	76	0.5052
HG00284	FIN	4.08807	76	0.5052
HG00323	FIN	2.80008	89.19	0.5035
HG00330	FIN	13.9648	90.22	0.5045
HG00380	FIN	3.45273	100	0.502
NA18961	JPT	3.48611	76	0.5067
NA18964	JPT	3.333	76	0.5052
NA18969	JPT	2.6653	100	0.5026
NA18970	JPT	4.47082	100	0.502
NA19009	JPT	3.94626	108	0.5033
NA19076	JPT	3.50604	108	0.5029
NA19080	JPT	3.84401	108	0.5055
NA19081	JPT	2.60827	108	0.5034
NA19082	JPT	3.58866	108	0.5018
NA19084	JPT	4.37475	108	0.5026
NA18520	YRI	3.99207	76	0.5057
NA18522	YRI	2.55368	76	0.5066
NA18853	YRI	2.56291	76	0.5099
NA18923	YRI	4.42742	100	0.5019
NA19116	YRI	3.03829	82.51	0.5056
NA19130	YRI	4.97799	76	0.5061
NA19197	YRI	4.19443	100	0.5021
NA19200	YRI	4.22902	100	0.502
NA19236	YRI	4.21535	76	0.5055
NA19248	YRI	4.24979	76	0.5058

Table S2: Average read balances for the 1000 genomes populations used for the analysis of empirical data.

Population	Average r_L
FIN	0.50334
JPT	0.5036
YRI	0.50512