Genome Biology

METHOD

Open Access

Pre-processing of paleogenomes: mitigating reference bias and postmortem damage in ancient genome data

Dilek Koptekin[1,2,3*†], Etka Yapar[1,4†], Kıvılcım Başak Vural[1†], Ekin Sağlıcan[1,5], N. Ezgi Altınışık[6], Anna-Sapfo Malaspinas[2,3], Can Alkan[7] and Mehmet Somel[1]

†Dilek Koptekin, Etka Yapar, and Kıvılcım Başak Vural contributed equally to this work.

*Correspondence: dilek.koptekin@metu.edu.tr

[1] Department of Biological Sciences, Middle East Technical University, Ankara, Turkey
[2] Department of Computational Biology, University of Lausanne, Lausanne, Switzerland
[3] Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland
[4] Department of Biology, Lund University, Lund, Sweden
[5] Department of Health Informatics, Graduate School of Informatics, Middle East Technical University, Ankara, Turkey
[6] Human-G Laboratory, Department of Anthropology, Hacettepe University, Beytepe, Ankara, Turkey
[7] Department of Computer Engineering, Bilkent University, Ankara, Turkey

Abstract

We investigate alternative strategies against reference bias and postmortem damage in low coverage paleogenomes. Compared to alignment to the linear reference genome, we show that masking known polymorphic sites and graph alignment effectively remove reference bias, but only starting from raw read files. We next study approaches to overcome postmortem damage: trimming, rescaling, and our newly developed algorithm, bamRefine (*github.com/etkayapar/bamRefine* and *zenodo.org/records/14234666*), masking reads only at positions possibly affected by PMD. We propose graph alignment coupled with bamRefine as a simple strategy to minimize data loss and bias, and urge the community to publish FASTQ files.

Keywords: Ancient DNA, Reference bias, Graph-reference genome, Post-mortem damage, Masking

Background

Ancient DNA (aDNA) has become today a major information source for studies of evolution or the human past. However, paleogenomic data has its specific challenges, being characterized by short fragment lengths, post-mortem damage (PMD) in the form of transitions at the ends of DNA molecules, and a low abundance of endogenous DNA resulting in low-coverage genomes. Standard aDNA data processing pipelines of low-coverage genomes typically involve (i) adapter trimming and merging of paired end reads, (ii) alignment of merged reads (fragments) to a linear reference genome, (iii) quality filtering of reads, (iv) modifications to the read data to avoid PMD confounding with true genetic variation, such as trimming or rescaling (or only using transversion polymorphisms), (v) genotyping or calculating genotype likelihoods at known polymorphic loci (as low coverage generally precludes de novo genotyping), (vi) pseudohaploidization, i.e., randomly choosing one allele per variant site (a strategy to overcome biases

related to heterogeneous coverage among studied genomes). Hence, genotyping of ancient genomes with low coverage is susceptible to various biases and shortcomings that can eventually lead to inaccurate interpretations of genetic relationships, population history, or evolutionary processes. We will tackle two of such issues in this study: reference bias, and biased and/or low-efficiency genotyping in the face of PMD.

Mapping biases against DNA reads that diverge from a reference genome had been noted earlier [1], but the reference bias phenomenon in ancient genomes was first coined and explained by Günther and Nettelblad [2]. These authors described how read alignment to a linear reference genome with low-coverage and short read-based sequencing data can lead to a higher frequency of reference allele calls over alternative allele calls at heterozygous sites when a 1:1 ratio would be expected. Reference bias arises due to the read mapping quality score calculation: reads with mismatches receive lower scores than perfectly matched reads. Hence, non-reference allele-carrying reads tend to be either unmapped or assigned lower mapping quality scores than the reference allele-carrying reads, and thus removed when filtering reads by a minimal mapping quality score [2, 3]. Consequently, reference allele-carrying reads are overrepresented in the aligned and filtered data.

Reference biases have been observed to impact population genetic and phylogenetic analyses of present-day taxa when evolutionarily distant linear reference genomes are used for alignment [4, 5]. Meanwhile, ancient DNA sequencing data is particularly prone to such bias, because when reads are short and/or have higher residual PMD, mismatches caused by alternative alleles can have a disproportionate impact on quality scores. The overrepresentation of reference allele-carrying reads may render ancient genome profiles more similar to the reference genome than they actually are. This effect can then lead to biased results in downstream inferences on phylogenetics, demographic history, or kinship.

Previous studies have suggested several methods to reduce reference bias in ancient DNA studies: (a) Statistically accounting for possible reference bias during variant calling [6], which can be effective but only on relatively high-coverage genomes; (b) aligning reads to a modified version of the linear reference genome, e.g., by representing both alleles or a third allele at polymorphic sites [2, 7, 8]; (c) modifying ancient reads at SNP sites by converting them to 'N' [2]; (d) using a graph reference genome that represents variants in large genomic variation datasets such as the 1000 Genomes Project for humans [3].

A second challenge in paleogenome data pre-processing involves ensuring that PMD on molecules does not impact inferred genotypes. One correction strategy is experimentally removing PMD after DNA extraction, most commonly using uracil-DNA glycosylase (UDG) treatment [9]. The majority of researchers who use UDG employ the half-UDG protocol, which still leaves a slight excess of transitions at molecule ends [10]. PMD may also be accounted for using post-alignment in silico approaches. One solution involves limiting analyses to transversion polymorphism sites, where allele frequencies will be largely unaffected by PMD [1] (only indirect effects are possible). However, using only transversions leads to the loss of approximately 60% of polymorphism data in humans and other mammals, as transition polymorphisms are about twice more numerous than transversions across the genome [11]. An alternative, and currently the

most prevalent method, is trimming, or masking the end of the reads in a BAM file. This involves changing bases at merged read (fragment) ends of a specific length to "N" and their quality to "!" (corresponding to zero in Phred + 33 encoding), e.g., using the tool *trimBAM* [12].

When trimming, most researchers remove 2–5 bases at read termini of half-UDG-treated libraries, or 8–10 bases of non-UDG-treated libraries [13–22]. This trimming process also leads to data loss, especially for the latter type of libraries. For instance, in a non-UDG-treated and paired-end library, 10 bp are masked from both ends ($2 \times 10 = 20$ bp in total) per standard 60 bp aDNA read, which means c.30% data loss. Other methods, such as *mapDamage* [23] and *ATLAS* [24], have attempted to reduce the effect of PMD by rescaling the base quality of possible PMD-driven misincorporations, but such approaches are potentially problematic as they could alter genotype frequencies, which has not yet been systematically investigated. Yet an alternative approach could be masking only PMD-sensitive regions on merged read ends, thus retaining more genetic information and enabling more comprehensive analysis of low-coverage ancient genomes.

We note that imputation using reference panels is another strategy increasingly being adopted in paleogenomics and can allow effective diploid genotyping in ancient genomes, including the removal of reference genome bias and PMD [25, 26]. However, accurate imputation requires at least modest coverage (e.g., $> 0.25 \times$ for Eurasian human shotgun genomes) and is thus not available to many poorly covered samples.

In this work, we study solutions to reduce the effect of reference bias and PMD on low-coverage genomes. We first investigate the degree of reference bias using linear mapping, mapping to a masked genome, and using a graph genome on simulated as well as empirical paleogenomic data of various types. We then study genotyping efficiency under PMD using standard trimming, rescaling base qualities with *mapDamage* or *ATLAS*, or masking merged read (fragment) ends that overlap with genomic positions that are sensitive to PMD-related false positive variant calls using our new algorithm, *bamRefine*. Our results show that using alternative reference genomes (either graph aligned or masked) together with *bamRefine* is a practical solution that results in accurate genotypes and reduces data loss. We also note the potential of the *ATLAS* tool for high accuracy, which is currently hindered by its low speed and difficult implementation.

Results

Simulated genomes: mapping to masked or graph genomes mitigates reference bias

We first simulated ancient human-like sequencing data to gauge reference bias under various alignment strategies. We used the human chromosome 1 (version hs37d5) reference sequence and 77,841 heterozygous sites chosen from a bi-allelic SNP set from the Turkish Genome Project dataset [27] (see Additional file 2: Table S1A-B). We created aDNA-like double-stranded genomes with the *gargammel* tool [28] such that reads would carry either allele at heterozygous sites with equal probability. We produced five such datasets with coverages from 0.05X to 10X with PMD damage with medium damage (35% at 5'C). We additionally generated a double-stranded genome with high damage (51% at 5'C), one with low damage (17% at 5'C), a half-UDG-treated library, and a single-stranded library with 10X coverage and with medium damage (35% at 5'C), where

damage extends into the center of the fragment (Additional file 1: Fig. S1, Additional file 2: Table S1A). We then aligned this data to reference genomes using three different strategies (see Methods for details): (i) the "LINEAR" strategy, which is the standard procedure of mapping to a linear reference genome using *bwa aln* with "*-l 16,500 -n 0.01 -o 2*"; (ii) the "MASKED" strategy, where, before alignment with *bwa aln*, we masked the linear reference genome sequence at variable positions to be genotyped by converting those bases to "N"; and (iii) the "GRAPH" strategy, where we used a graph reference genome (*SBG.Graph.B37.V6.rc6.vcf.gz*) representing both reference and alternative alleles at known polymorphic sites and used *GRAF aligner* for mapping [29]. We then randomly called pseudohaploid genotypes (a single allele sampled per diploid genotype) at the 77,841 heterozygous sites and calculated the alternative allele proportion. Because pseudohaploidization involves random sampling, we repeated these last steps 100 times to obtain point estimates.

In the absence of reference bias, we expect ~ 50% of pseudohaploid genotypes at heterozygous positions to represent the alternative allele. However, using the "LINEAR" strategy, we observed consistently lower match rates to the alternative allele across all coverages, i.e., reference bias (48.2–50.4%; on average ~ 1% lower than expected; binomial test $p < 0.0001$) (Fig. 1A, Additional file 3: Table S2A). This may appear small, but can lead to significant biases in demographic analyses, such as f_4 tests (see below), and create false positive results especially when studying closely related populations. We also note that the bias was slightly alleviated using MAPQ > 25 as mapping quality filter for "LINEAR" mapped reads (Fig. 1D,Additional file 1: Fig. S2).

Using the "MASKED" or "GRAPH" strategies, the total number of reads mapped and those passing the MAPQ > 30 were comparable to those using "LINEAR" (Additional file 1: Fig. S3). But in contrast to the "LINEAR" approach, we observed either slight or no bias towards either allele: the average fraction of alternative alleles was 50–50.3% with the former and 49.8–50% with the latter (Fig. 1A, Additional file 1: Fig. S4, Additional file 3: Table S2A). While alternative allele-carrying reads with more mismatches and of short length tended to be filtered out using "LINEAR," while trend was weaker using the "MASKED" strategy and absent using "GRAPH" (Additional file 1: Fig. S3). Overall, the deviations from 50% using "MASKED" or "GRAPH" were systematically lower than using the "LINEAR" strategy (Mann–Whitney *U* test $p < 0.0001$; Fig. 1A, Additional file 1: Fig. S5).

Published ancient genomes: reference bias mitigated using FASTQ files but not using filtered BAM files

We next studied reference bias in empirical paleogenomic data. For this, we started by collecting ten published genomes for which we could obtain raw data as FASTQ files [30–37] (Additional file 2: Table S1C). This was a random sample representing libraries derived from diverse geographic regions, of single- or double-stranded type, produced with or without UDG-treatment, shotgun-sequenced, whole-genome captured or 1240K SNP-enriched genomes had variable coverages, and originated from different laboratories (Additional file 2: Table S1C).

We first defined heterozygous sites for each ancient genome as those with 25–75% of reads representing the alternative allele, covered at least by 10X depth and no greater

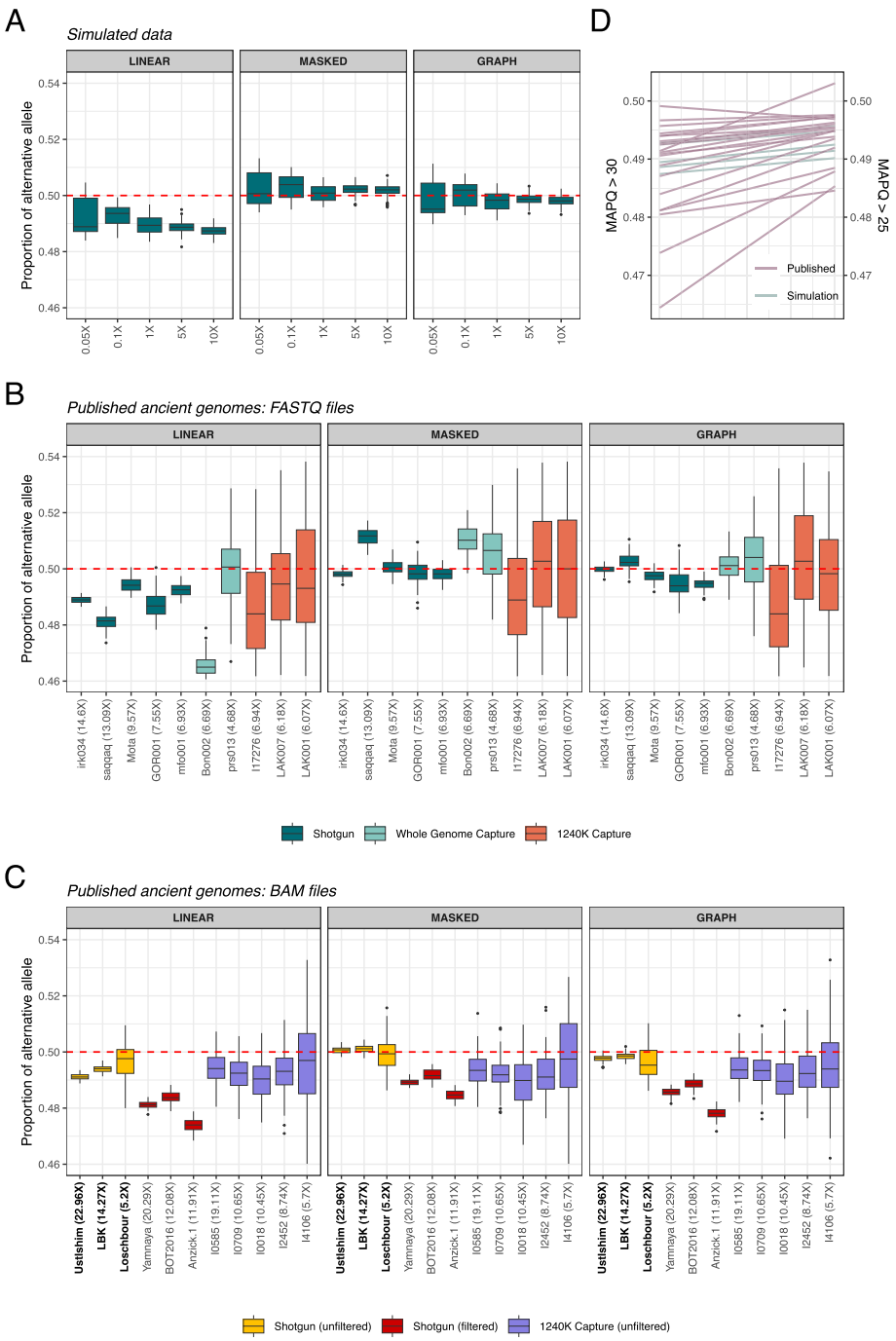

Fig. 1 Comparing reference bias under three different alignment strategies **A** using simulated aDNA-like genomes (see Additional file 4: Table S3A), **B** shotgun and 1240 K capture ancient genomes with available raw FASTQ files, and using published **C** shotgun and 1240 K capture ancient genomes with already processed BAM files. The plot shows the proportion of alternative alleles after randomly selecting one allele from heterozygote sites 100 times (panel **A**: 77,841 sites; panel **B**: 4658–422,046 sites; panel **C**: 2934–543,495) (see Additional file 4: Table S3B). The BAM files available without strict filtering (i.e., which included reads with MAPQ < 30) are shown in bold in panel **C** (Additional file 2: Table S1Cand Additional file 1: Fig. S7). We used reads with MAPQ > 30 for genotyping. For results using MAPQ > 25, see Additional file 1: Fig. S2. **D** Reference bias in simulated and published ancient genomes aligned with the "LINEAR" strategy and applying MAPQ > 30 (left axis) and MAPQ > 25 (right axis). The difference is significant in a Wilcoxon signed rank test ($p < 0.001$). In these comparisons, we did not apply any PMD-correction

than two times the genome mean coverage (Methods, Additional file 2: Table S1C). We mapped reads using the three strategies and randomly sampled reads 100 times at these presumed heterozygous sites. We found widespread reference bias among libraries using the "LINEAR" strategy, with the fraction of alternative alleles ranging between 46.4 and 49.9% for shotgun genomes (binomial test $p < 0.0001$) and 49.2–49.7% for 1240K enriched genomes (binomial test $p < 0.0001$) (Fig. 1B, Additional file 3: Table S2B). Consistent with the simulation results, the fraction of alternative alleles was ~ 50% when using either the "MASKED" (49.8–51.1% for shotgun genomes and 49.4–49.9% for 1240 K enriched genomes) or "GRAPH" strategies (49.4–50.4% for shotgun genomes and 49.4–50% for 1240K enriched genomes) (Mann–Whitney U test $p < 0.0001$; Fig. 1B, Additional file 3: TableS2B). However, we also noted slight differences between these two approaches: two genomes (mfo001, GOR001) processed using the "GRAPH" strategy still exhibited a bias against the alternative allele (~ 49.5%). The two other genomes (Bon002 and Saqqaq) processed using "MASKED" exhibited a weak but significant bias (~ 51%) towards the alternative allele ($p < 0.0001$), which was reduced when using MAPQ > 25 as mapping quality filter (Additional file 1: Fig. S2). Meanwhile, the three 1240K capture libraries available as FASTQ files also showed reference bias with "LINEAR" but little improvement using "MASKED" and "GRAPH." Still, because these were relatively lower coverage than shotgun data and showed high variation, it was more difficult to assess the performance of the three strategies on the 1240K data. Overall, although we lack an explanation for some of these inter-library variability patterns, we conclude that "MASKED" or "GRAPH" approaches both reduce the impact of reference bias on called ancient genotypes (Fig. 1B,Additional file 1:Fig.S6A, Additional file 3: Table S2B).

The majority of paleogenomes over the last decade have been published as processed BAM files rather than raw FASTQ files, where the former could be subject to irreversible reference bias introduced by mapping parameters as well as mapping quality filtering. To investigate this, we collected 11 additional paleogenomes available as BAM files [38–45] (Additional file 2: Table S1C). These included six shotgun-generated and five 1240K SNP-enriched genomes. Among the shotgun-generated genomes, the Ust-Ishim, LBK, and Loschbour BAM files were published without strict filtering (i.e., included reads with MAPQ < 30), while the rest had been quality filtered (all reads with MAPQ > 30). None of the 1240K SNP-enriched genomes had been subjected to strict filtering (Additional file 1: Fig. S7, Additional file 2: Table S1C).

We again remapped the reads and called pseudo-haploid genotypes using the three strategies. This revealed persistent reference bias for three shotgun-generated BAM files subjected to strict filtering, irrespective of the alignment strategy used (Fig. 1C, Additional file 1: Fig. S6B). In contrast, both the "MASKED" and "GRAPH" strategies significantly reduced reference bias in Ust-Ishim, LBK, and Loschbour, which were unfiltered (Fig. 1C, Additional file 1: Fig. S7, Additional file 2: Table S1C). This confirms the expectation that quality filtering of BAM files introduces irreversible reference bias.

Meanwhile, all five 1240K SNP-enriched BAM data showed the same level of reference bias irrespective of the alignment strategy used (the alternative allele on average ~ 0.7% lower than expected) (Fig. 1C, Additional file 1: Fig. S6, Additional file 3: Table S2B). Such bias appears independent of the mapping/filtering process and is likely attributable

to 1240K SNP capture favoring one allele over another at targeted SNPs, as reported recently [46, 47].

Trimming, rescaling, and refining as alternative PMD-correction approaches

We further investigated the performance of several approaches for PMD-correction on called genotypes: (A) trimming, i.e., the standard 2 or 10 bp masking of aligned reads using *trimBam*, (B) rescaling, which involves rescaling base qualities using *mapDamage2* [23] or *ATLAS* [24], and (C) refining, i.e., masking bases at the merged read (fragment) ends that overlap with variants sensitive to PMD-related genotyping errors using the new software we present here, *bamRefine.*

Our *bamRefine* algorithm was designed as a simple and fast alternative to rescaling and a more accurate alternative to trimming. It masks a user-defined number of bases from the 5′ end (similar to trimming) only if they overlap with variants that include a "C" allele, to prevent C- > T false-positives. In double-stranded libraries, it can also mask a user-defined number of bases from the 3' end only if they overlap with variants that include a "G" allele, to prevent G- > A false-positives. This approach also avoids biased genotyping due to PMD-induced C/G loss at transversion sites (e.g., C's being underrepresented at a C/A variant site due to PMD-induced C- > T transitions). *bamRefine* further avoids comprehensive data loss compared with using *trimBam*, as the latter involves masking extended regions at merged read ends for non-UDG-treated libraries (Methods).

We used the same simulation scheme using chromosome 1 polymorphisms as above, with the difference that here, along with the 77,841 heterozygous positions described earlier, we also genotyped 182,515 homozygous reference and 53,391 homozygous alternative positions, totalling 313,747 SNPs (Additional file 2: Table S1B). We generated aDNA-like read data at 10X coverage using *gargammel* [28], aligned these using either of the three mapping strategies ("LINEAR," "MASKED," and "GRAPH") and applied either of the four PMD-correction approaches: trimming using *trimBam*, rescaling with *mapDamage2* or *ATLAS*, and using *bamRefine*. We then called diploid genotypes at the 313,747 SNPs using *GATK HaplotypeCaller* [48] *for trimBam, mapDamage2* and *bamRefine*, and using the maximum likelihood method for genotype calling implemented in *ATLAS* [24] (see Methods). We then examined the missingness and error rates on these calls, comparing the four PMD-correction approaches with no PMD correction. This baseline is more appropriate than using transversions only, as most error derives from transitions.

Trimming causes data loss, mapDamage2 creates reference bias, and ATLAS has impractical run times

Irrespective of the mapping approach, trimming with *trimBam* exhibited the highest missingness (0.48–0.95%), followed by *bamRefine* (0.38–0.86%) and *mapDamage2* (0.26–0.64%), with the lowest rates attained by *ATLAS* (0.23–0.58%) (Fig. 2A, Additional file 1: Fig. S8). Trimming also showed the highest overall error rate among the four methods (2.42–2.48%) (Fig. 2A, Additional file 4: Table S3A). The bulk of these errors were caused by misassigning heterozygous sites as homozygous reference or homozygous alternative, due to sampling error (i.e., insufficient data to call

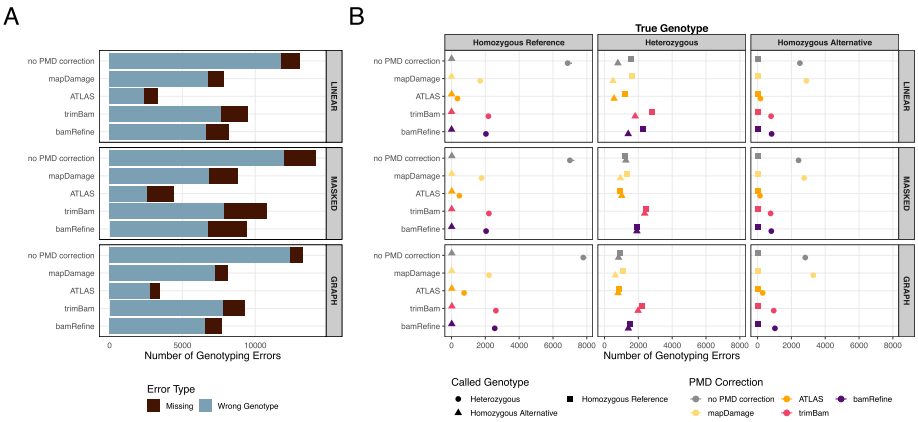

Fig. 2 Genotyping error under different PMD-correction approaches and mapping strategies. In both panels, the left side shows different PMD-correction approaches and the right side shows different mapping strategies. **A** Proportion of genotyping errors and missingness and **B** the frequency of the type of genotyping errors for each PMD-correction method for double-stranded 10X coverage simulated genomes with medium damage (35% at 5′C), calculated by comparing diploid calls with true genotypes. See also Additional file 1: Figs.S8-9 for simulations of libraries with different levels/patterns of PMD (half-UDG, high-level PMD, low-level PMD) as well as a simulated single-stranded library. See Additional file 1: Figs.S12-20 for distributions of different types of genotyping error

heterozygous sites) (Fig. 2B, Additional file 1: Fig. S9). *bamRefine* and *mapDamage2* had slightly lower error rates (2.07–2.15% and 2.14–2.31% respectively), while *ATLAS* had the lowest errors (0.73–0.87%). These error rates were higher (~ 7%, 6%, 5%, and 4% respectively) when repeating the analysis with 5X coverage data (Additional file 1: Fig. S10); this is expected as lower coverage elevates sampling error.

Despite the lower overall error rate and missingness using *mapDamage2*, closer inspection revealed that this approach suffers from significant reference bias. The majority of errors observed with *mapDamage2* were caused by favoring the reference allele in genotype calls during PMD-correction (Mann–Whitney U test $p < 0.0001$; Additional file 1:Fig.S11), leading to an overestimation of homozygous reference alleles and underestimation of homozygous alternative alleles (Fig. 2). In contrast, *trimBam* and *bamRefine* label genotypes incorrectly as homozygous reference or homozygous alternative at similar rates, and *ATLAS* introduces only subtle biases (Additional file 1: Figs.S12-20). The marked underestimation of alternative allele proportions in *mapDamage2* output (~ 47%) could also be observed when calling pseudohaploid genotypes at the 77,841 heterozygous sites (Fig. 3). *mapDamage2* bias persisted whichever alignment strategy was employed and it mostly involved transition sites (Additional file 1: Figs.S21-22).

Our results indicate *ATLAS* as having the highest performance with respect to minimal missing or wrong genotypes. Unfortunately, implementing the current *ATLAS* version is constrained by a number of practical issues: (a) Each sequencing library has to be individually calibrated. If a sample's libraries are not individually tagged with reading groups (RGs) but published jointly, as is usually the case [49], properly running *ATLAS* may not be possible. (b) *ATLAS* has an order of magnitude higher running times than the other tools tested. For example, it required 287 min, compared to only 21 min using *bamRefine* on the simulated 10X genome. (c) *ATLAS* also

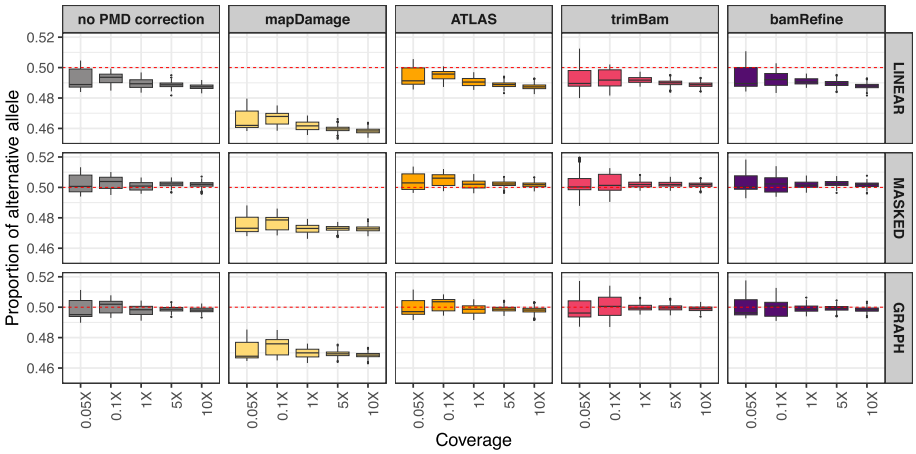

Fig. 3 Comparing reference bias in simulated ancient genomes with medium damage (35% at 5'C) aligned with different reference genomes and PMD-effects reduced with different approaches. The plot shows the proportion of alternative alleles after randomly selecting one allele from heterozygote sites 100 times (see also Additional file 1: Figs. S21-22). We used reads with MAPQ > 30 for genotyping. To compare results with MAPQ > 25, see Additional file 1: Fig. S23

has much heavier memory requirements compared to its alternatives. For instance, the published library GOR001 used more than 900 GB memory for recalibration and terminated with error, whereas memory requirements are minimal for tools such as *trimBam* or *bamRefine*. Under these circumstances, creating large datasets with *ATLAS* by combining published data from various different sources may be prohibitively difficult, despite the tool's high accuracy. For this same reason, we did not attempt to apply *ATLAS* on empirical data.

"GRAPH" or "MASKING" alignment followed by bamRefine is a practical solution against reference bias and PMD

We next investigated genotype accuracies among the five PMD-correction methods using the 10X simulated dataset. For this, we calculated the concordance rate (CR), the proportion of false negatives, the proportion of false positives, the non-reference true positive rate (NTPR), as well as recall (or sensitivity) and the F-score (Fig. 4), with the alternative allele as our pivot [50] (Additional file 1: Fig. S24).

This revealed a number of patterns. (a) Rescaling with *ATLAS* gives the best results for all eight statistics, including a much higher concordance as well as F-score (0.99) than all other three methods (< 0.975). Interestingly, *ATLAS* appeared to perform best using the "LINEAR" alignment and worst using "GRAPH." (b) *trimBam*, which involves aggressive masking 10 bp at read ends, leads to significant data loss (Additional file 1: Fig. S25) and the lowest F-scores (Fig. 4). (c) Overall, *bamRefine* and *mapDamage2* had similar overall performances; *bamRefine* had better recall but worse precision (Fig. 4).

The strong reference bias in *mapDamage2* (Fig. 3, Additional file 1: Figs. S11 and S13) renders it the least useful among the tested methods in our view. *trimBam* causes high data loss, while *ATLAS* has superb accuracy but is constrained by difficult and slow implementation. Overall, *bamRefine* combined with "GRAPH" alignment *emerges* as a highly practical approach, clean of bias and with decent F-scores (Figs. 3 and 4).

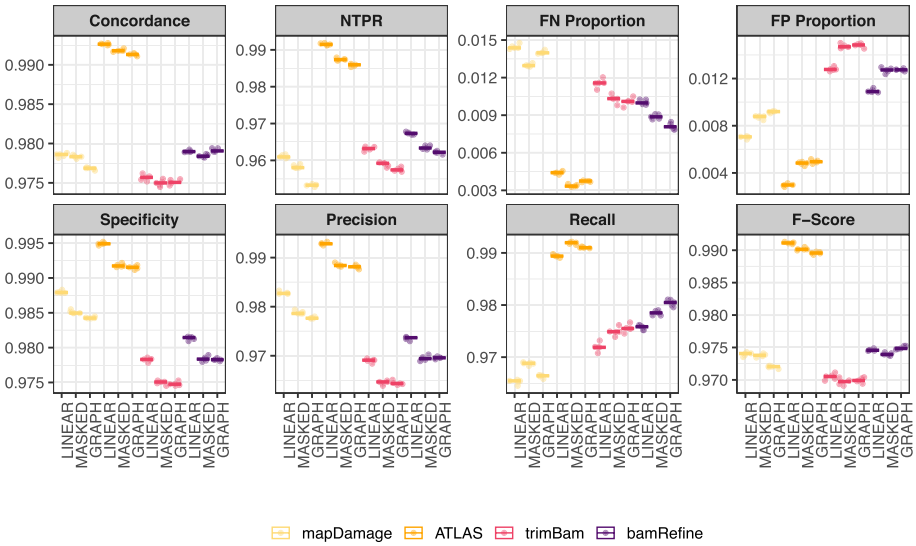

Fig. 4 PMD-correction performances of *mapDamage2*, *ATLAS*, *snpAD*, *trimBam*, and *bamRefine* on 10X simulated ancient genomes with medium damage (35% at 5'C), calculated by comparing diploid calls with true genotypes (Additional file 1: Fig. S24). "NTPR" stands for non-reference true positive rate, "FN" for false negatives, "FP" for false positives. We used reads with MAPQ > 30 for genotyping. To compare results with MAPQ > 25, see Additional file 1: Fig. S26

Finally, we applied the *trimBam* and *bamRefine* to the 21 published ancient genomes described earlier (Additional file 2: Table S1C, Additional file 1: Figs. 27–28). We did not use *ATLAS* for its difficulty in practical implementation and also did not use *mapDamage2* because of the strong bias it created. Consistent with the simulation results, when using shotgun FASTQ files (Fig. 5, Additional file 1: Fig. S29) and/or unfiltered BAM files (Additional file 1: Figs. S30-31) and mapping using the "GRAPH" or "MASKING" strategies, neither *trimBam* nor *bamRefine* led to reference bias (49.6–50.9% proportion of alternative allele) (Fig. 5, Additional file 1: Fig. S32, Additional file 3: Table S2B). The two tools had comparable error rates (Additional file 1: Figs. S32-33), while *trimBam* led up to 2% more data loss (as measured by the number of genotyped SNPs) than *bamRefine* (Additional file 4: Table S3B).

The impact of reference bias on measures of allele sharing

The overrepresentation of reference alleles in simulated and empirical aDNA libraries appeared modest, usually about 1%. However, such genome-wide bias could readily lead to statistically significant asymmetries in analyses such as f_4-statistics. To test this, we studied f_4-statistics of the form f_4*(Chimpanzee, Human Reference Genome; Ind1_MappingStrategy1, Ind1_MappingStrategy2)*. We found that the Human Reference Genome significantly shared more alleles with data processed using the "LINEAR" strategy ($|Z| > 3$) than using the "MASKED" and "GRAPH" strategies. This suggests that both "MASKED" and "GRAPH" strategies largely mitigate the reference bias that arises with the "LINEAR" strategy (Fig. 6A).

We further found that in 71% of comparisons, the Human Reference Genome shares more alleles with data processed using the "MASKED" strategy than the "GRAPH"

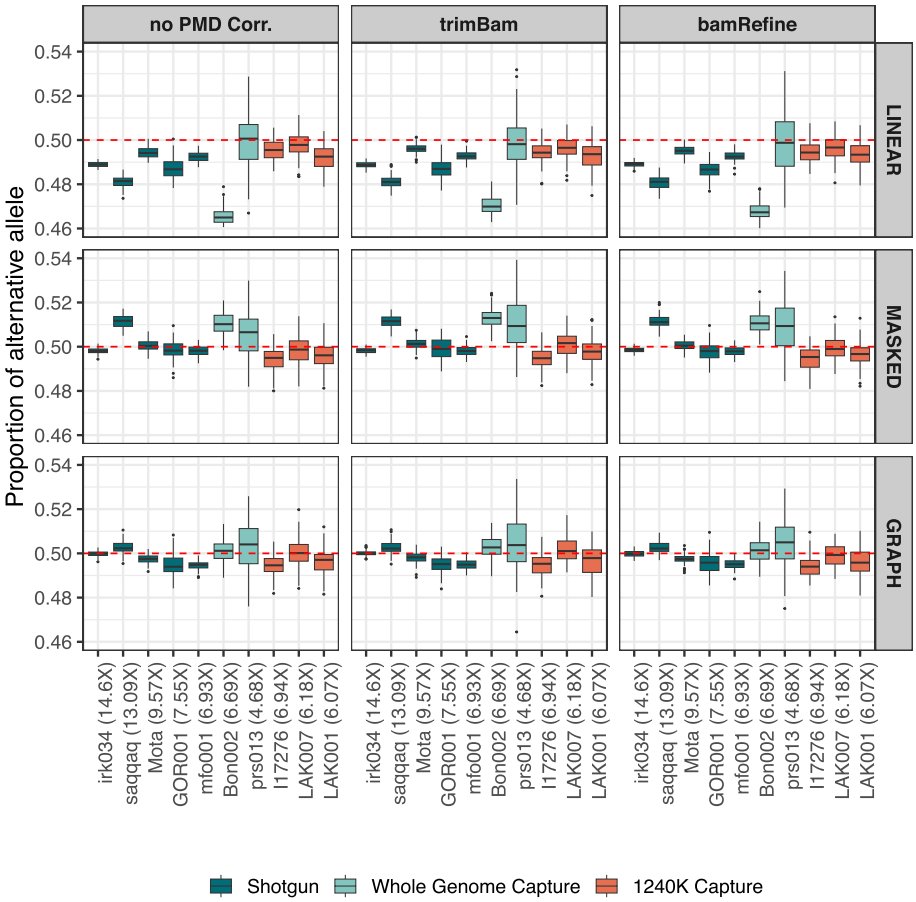

Fig. 5 Comparing reference bias in published ancient genomes that FASTQ file available aligned with different reference genomes and PMD-effects reduced with different approaches. The plot shows the proportion of alternative alleles after randomly selecting one allele from heterozygote sites 100 times. We used reads with MAPQ > 30 for genotyping. To compare results with MAPQ > 25, see Additional file 1: Fig. S29. (see also Additional file 1: Figs. S30-31)

strategy. This indicates that "GRAPH" is more effective in reducing reference bias, consistent with our earlier results (Fig. 6B).

Finally, we also compared if either *trimBam* or *bamRefine* showed additional bias in form f_4*(Chimpanzee, Human Reference Genome; Ind1_MappingStrategy1_TRIMMING, Ind1_MappingStrategy1_REFINING)* for three mapping strategies. All results were non-significant ($|Z| < 3$) (Fig. 6C).

Discussion and conclusion

Our results confirm a strong reference bias in ancient genomes that emerges when using linear reference genome alignment ("LINEAR"), which impacts downstream analyses such as f_4 tests. We also find that alignment to either a masked linear reference genome ("MASKED") or to a graph genome ("GRAPH") effectively reduces reference bias. This observation is consistent with previous findings [2, 3, 7, 8] and supports the feasibility of implementing these strategies for more accurate aDNA analysis. Compared to "MASKED," the "GRAPH" approach has higher F-scores when using *trimBam* or *bamRefine* and is also less affected by reference bias. These

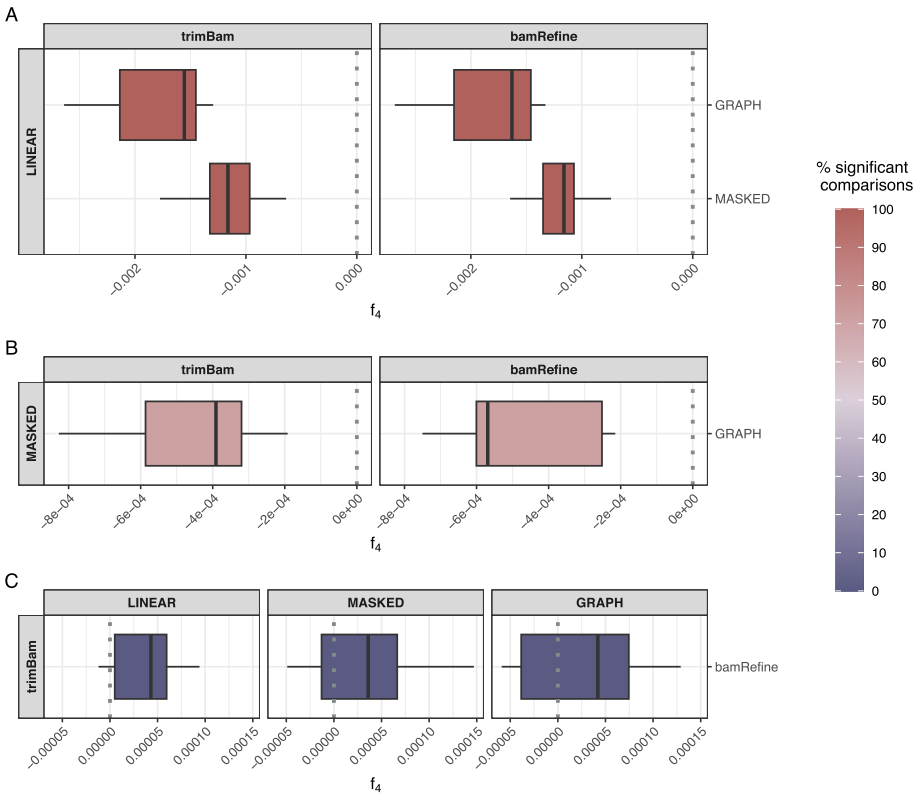

Fig. 6 Results from the model **A** f_4*(Chimpanzee, Human Reference Genome; Ind1_LINEAR, Ind1_MASKED/Ind1_GRAPH)*, **B** f_4*(Chimpanzee, Human Reference Genome; Ind1_MASKED, Ind1_GRAPH)* for both PMD correction strategies, and **C** f_4*(Chimpanzee, Human Reference Genome; Ind1_Mapping_Strategy1_TRIMMING, Ind1_Mapping_Strategy1_REFINING)* for all mapping strategies by using ancient genomes that FASTQ files available. The chimpanzee is the outgroup, and we are testing whether the Human Reference Genome is equally distant to the genetic data from the same individual but processed using alternative strategies (e.g., Ind1_LINEAR versus Ind1_MASKED). f_4 distributions shifted to the left or right suggest the reference genome shares more alleles with the genetic data processed using the methods indicated in the left or right, respectively. The color gradient from blue to red represents the fraction of comparisons that are nominally significant ($|Z| > 3$). See also Additional file 1: Fig. S34 for results when the genomes with just BAM files available are used

differences might be partly attributable to *bwa* converting "N" nucleotides (such as those created by masking) to one of four random bases before alignment. Such random conversion will remove reference bias on average and would not impact population genetic studies; however, it can lead to data loss, and also negatively impact studies of natural selection where accurate prediction of allele frequencies is critical. An alternative masking solution is to create two reference.fasta files that carry both alleles for alignment [e.g., [7]]. We further note that statistical correction of reference bias is also possible, as implemented in the software *snpAD* [6], but this requires high-coverage (> 10x) genomic data. Overall, the "MASKED" strategy appears as the simplest remedy for reference bias applicable with the standard *bwa aln* tool, while "GRAPH" can be described as a highly effective alternative. We also note that when "MASKED" or "GRAPH" strategies are not available (e.g., when a reference diversity panel is not present) using MAPQ > 25 (instead of the widely used MAPQ > 30) as quality filter can also reduce bias with the "LINEAR" strategy.

Despite their effectiveness on FASTQ data, neither "MASKED" nor "GRAPH" strategies can alleviate reference bias on paleogenome data published after mapping quality filtering. This outcome emphasizes the need for sharing all the raw data, such as the BAM files with all reads (including those with very low mapping quality), or even better, the raw FASTQ files (without merging paired end reads). Meanwhile, modified (e.g., trimmed or refined) BAM files should not be published to avoid technical effects that may be easily missed or may be difficult to trace back. Sharing raw data allows its long-term healthy reusability, avoiding possible batch effects created by data processing.

The reference bias in SNP-capture data appears inherent to the previously widely used Agilent 1240K platform [46] and, similar to filtered BAM data, cannot be corrected. Rohland and colleagues have recently suggested that another capture approach, the TWIST platform, is free of reference bias. Still, our results point to the risks introduced by experimental manipulation of ancient molecules. Imputation methods may partly help overcome such inherent biases [51], but imputation using modern-day haplotypes from specific populations may itself create new issues as, for instance, variants not present in present-day populations cannot be imputed; low-coverage genomes are also not imputable.

Overall, we believe the safest way forward for the community involves shotgun sequencing and full data sharing, as recently pointed out in a systematic study [49]. This can also allow new uses of paleogenomic data, such as copy number variation [52] or metagenomic analyses [53]. At the same time, population genetic analyses should be conducted in a bias-aware fashion to avoid false positives; e.g., by calculating *f*-statistics only using genomes with similar susceptibility to reference bias.

This study also introduced a new algorithm, *bamRefine*, for effective PMD-correction especially on non-UDG-treated libraries. Refining with *bamRefine* selectively masks only PMD-sensitive sites at merged read ends and makes a larger amount of genetic information available for genotyping than standard trimming, and it also is free from reference bias introduced by *mapDamage2*. In simulated and empirical datasets, the combination of "GRAPH" mapping and *bamRefine* yielded good results. *ATLAS* had the highest accuracy on simulated data, but its use was prohibitively slow and complicated. Streamlining the "*ATLAS*" algorithm [24] could be a worthwhile avenue for future work, as it involves the lowest data loss and highest accuracy. We also note that using UDG-treatment of aDNA is an alternative experimental solution used by a large number of laboratories.

Overall, these approaches offer promising solutions to overcome the challenges associated with aDNA analysis, extracting more information from the available data and enhancing our ability to accurately reconstruct the population history of past populations.

Methods

Simulating ancient genomes

We used chromosome 1 of the human reference genome (version hs37d5) as a template to generate the simulated ancient genome data. We chose bi-allelic SNPs on chromosome 1 of the individual 06A010111 of the Turkish Genome Project dataset [27], which consisted of 182,515 homozygous reference, 53,391 homozygous alternative, and 77,841 heterozygous positions (313,747 positions in total) (see Additional file 2: Table S1B).

We then inserted these into the chromosome 1 template with "*vcftools/vcf-consensus (v.0.1.6)*" [54].

We generated ancient DNA data using "*gargammel*" [28], using the template chromosome 1 data with polymorphism inserted. Five "*gargammel*" simulations were performed for five target coverages: 0.05X, 0.1X, 1X, 5X, and 10X. We used a normal distribution with a mean of 65 bp for the read size distribution.

We also generated additional simulation as 10X coverage for different levels of PMDs. The parameter "*-damage 0.024,0.36,0.0097,0.55*" was used to introduce PMD to simulated ancient genomes (see Additional file 1: Fig. S1). This parameter was adjusted as "*-damage 0.024,0.36,0.01455,0.825*" for introducing high-level damage, "*-damage 0.024,0.36,0.00485,0.275*" for introducing low-level damage, and "*-damage 0.024,0.8,0.0,0.55*" for simulating half-UDG protocol. Single-stranded library simulations were produced by "single" parameter of "*gargammel*." Deamination rate of single-stranded simulations was determined using empirical rates measured from single stranded library of the sample G31 from Koptekin et al. [30]. We generated data without bacterial or modern contamination using the "*-comp 0,0,1*" parameter.

Published ancient genomes

We selected 21 published ancient genomes, either shotgun-sequenced, whole-genome captured, or 1240K SNP-captured, all from human skeletal material originating from different geographic regions [30–45]. The coverage of samples ranges from low to medium coverage to high coverage. The dataset includes both damage-repaired and non-damage-repaired samples (see Additional file 2: Table S1C).

The raw FASTQ files were available for 10 out of 21 samples. Others were downloaded as BAM files and converted to FASTQ files using "*Picard SamToFastq (version 2.23.8)*" (http://broadinstitute.github.io/picard/). A number of FASTQ files were not publicly available (Additional file 2: Table S1C) and were provided by the research teams upon request.

Alignment strategies

We removed the residual adapter sequences in raw FASTQ files for each sample using the software "*Adapter Removal (version 2.3.1)*" [55] using "*–qualitybase 33 –gzip –trimns*" parameters. The reads in paired-end libraries were merged after removing residual adapter sequences, requiring at least 11 bp overlap between the pairs using the additional parameter "*–collapse –minalignmentlength 11*."

We aligned FASTQ files to three different reference genomes:

(i) Linear Reference Genome (version hs37d5): We used the program "*BWA aln/samse (version 0.7.15)*" [56] with parameters "*-n 0.01, -o 2*" and disabled the seed with "*-l 16,500*."

(ii) Masked Linear Reference Genome (masked version of hs37d5): We masked the positions we wanted to genotype on the linear reference genome using "*bedtools maskfasta (v. 2.29.1)*" [57] by converting the nucleotides to "N." After masking, we aligned samples using "*BWA aln/samse (version 0.7.15)*" [56] with the same parameters above.

(iii) Graph Reference Genome: We used a published graph genome version from Seven Bridges Inc. (*SBG.Graph.B37.V6.rc6.vcf.gz*), which included variants from 1000 Genomes (1000G) Phase 3 (with alternate allele frequency greater than 0.01) [58], the Simons Genome Diversity Panel (alternate allele occurrence of 10 or greater) [59], and other INDEL variant datasets. The pangenome construction is described in the supplement of Rakocevic and colleagues [29] under the section "Global Graph Reference" in detail. We used the "*GRAF tool (version 0.12.5)*" [29] to align the reads to this graph genome annotation together with the baseline reference genome GRCh37, using default parameters. For example for the Loschbour genome, we used the following command: "*sbg-aligner-latest -f SBG.Graph.B37.V7.dev2.fa -v SBG.Graph.B37.V7.dev2.vcf.gz -q Loschbour.single.fastq.gz -o Loschbour.single.bam –read_group_unit Loschbour –read_group_library Loschbour –read_group_id Loschbour –read_group_sample Loschbour -t 32*." See Rakocevic et al. [29] and https://www.sevenbridges.com/graph-genome-academic-release/ for more details on the algorithm.

After alignment, we removed PCR duplicates using "*FilterUniqueSAMCons.py*" [60] and removed reads < 35 bp, with > 10% mismatches to the reference genome, and with < 25 or < 30 mapping quality (MAPQ) from all BAM files. Note that the *GRAF* tool [29] outputs BAM files directly (not GAM files) and calculates mapping quality scores in similar fashion as *BWA*.

We also compared *GRAF* with another widely used graph alignment tool, *vg-giraffe* [61], for 10X coverage simulations. We used the chromosome 1 of the same graph genome (*SBG.Graph.B37*) with "vg *autoindex*" to construct the graph and the indexes to be used in the mapping of reads with *vg-giraffe* "*-z, -m, -d*" parameters together with "*-o BAM*" to get BAM file as an output (Additional file 1: Fig. S35).

We note that, we did not work on GAM (Graph Alignment / Map) file format since both graph aligners used here (*GRAF* and *vg-giraffe*) have an option to generate BAM file directly as an output. We prefer to get a BAM file as an output since the PMD-correction approaches we used here only work with BAM files at the moment.

We added read group information to all final BAM files by using "*picard AddOrReplaceReadGroups (version 2.23.8)*" (http://broadinstitute.github.io/picard/).

bamRefine

Here, we present a new variant-aware PMD-correction algorithm called *bamRefine*.

bamRefine masks reads only at positions that could be affected by PMD (e.g., a nucleotide overlapping with a C/G polymorphism at the 5′ end of a read), while it retains any transversions that would not be affected by PMD (e.g., a nucleotide overlapping with a T/G polymorphism at the 5′ end). It thus uses more information compared to standard trimming.

As an example, if a researcher will be using the 1240K SNP set, they will use the 1240K SNP set genomic positions as input for masking the BAM files with *bamRefine*. Since the PMD effect on the read sequence depends on the library type and sequence end, the masking of these sites will be performed depending on information whether the library is double- or single-stranded. For double-stranded libraries, PMD-induced

C- > T changes are expected to occur at 5′ ends whereas G- > A changes are expected at 3′ ends. For single-stranded library protocols, C- > T changes are expected at both ends albeit at different amounts. Accordingly, *bamRefine* masks select positions close to the read ends that (a) overlap with an input polymorphism list, (b) that are of certain type that would be affected by PMD-induced changes (e.g., C polymorphisms at 5′ ends), (c) that are close enough to read ends to be affected (within a user-defined lookup range).

bamRefine has a simple algorithm with the following steps: First, the variant list to be used in downstream analyses is parsed and classified into 5′ and 3′ "suspect" lists. The "suspect" lists are created according to the expected PMD artifact profile depending on the library preparation protocol. For example, for a double-stranded library, positions overlapping with variants involving a "C" allele would be masked if close to the 5′ tail, and those with a "G" allele if close to the 3′ tail. Meanwhile, positions overlapping with "T"/ "G" polymorphisms would *not* be masked if close to the 5′ tail.

Next, the BAM file is processed read by read, masking bases for a number of bases from each end that overlap with the respective end's suspect list. The lookup range within which the masking occurs from each end is to be provided by the user based on the PMD signature in the library. The program allows the 5′ and 3′ end lookup ranges to be asymmetrical to properly process reads from single-stranded library protocols. The

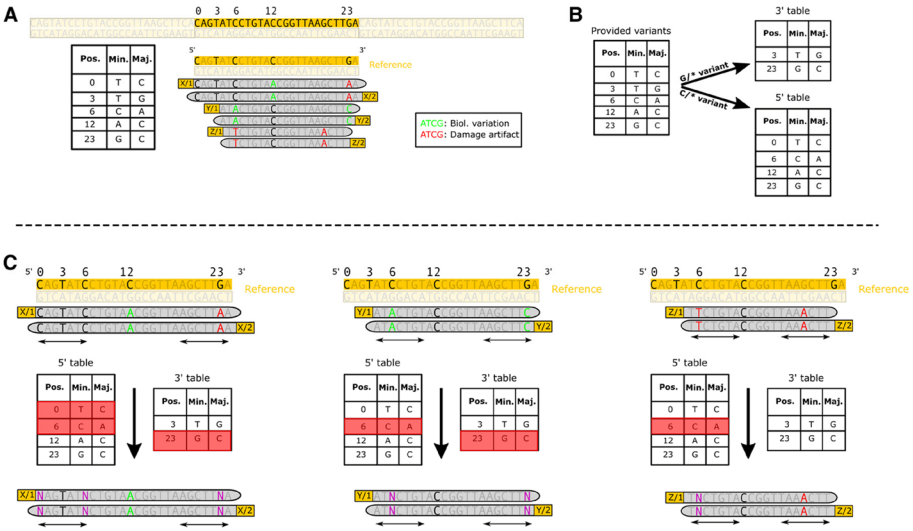

Fig. 7 Graphical representation of the *bamRefine* workflow. The graph describes a run using a BAM file with 6 reads from a double-stranded library, a variant list with 5 variants, and using the *bamRefine* parameter "*–pmd-length-threshold 7*" (i.e., a lookup range of 7 bp at each read tail). **A** A cartoon genome browser view of all the reads mapped to the genomic region, and the input variant list shown as a table. For didactic purposes, it is assumed on the figure that the real biological variation and those which are PMD artifacts are distinguishable and labelled differently. **B** The classification of the provided variant list into 5′ and 3′ suspects. Note that the variant at the position 23 ends up in both tables since it has both "C" and "G" alleles. **C** Masking of the three read pairs according to the specified options and input variant list. Masking happens regardless of the alleles the reads carry and only depends on whether a base within the lookup range overlaps with the variant table of the respective read tail. See the masking process for the first read as an example: it has two bases masked at the 5′ tail because those positions were included in the 5′ table, even though the read did not differ from the reference for those positions. Additionally, the 3rd position in the same read overlapping the T/G polymorphism is not masked and the genotype information is retained. This site overlapping the T/G polymorphism would have been lost from the read when using standard trimming, hence the advantage of *bamRefine*

masking is confined to the positions that overlap with the 5′/3′ variant tables within the user-specified lookup range from 5′/3′ ends of reads and is implemented regardless of the allele an individual read carries (Fig. 7).

Masking with *bamRefine* results in less data loss when compared to *trimming* the entire lookup range (e.g., T/G polymorphisms are not lost at 5′ tails of the reads for a double-stranded library). The choice of masking any position that could be affected (e.g., C/G transversions) also minimizes the PMD effect.

The job of flagging and masking positions of interest for each chromosome or contig in a BAM file is parallelized by multiprocessing, allowing the program to rapidly refine millions of reads. bamRefine adds a @PG header to modified BAM files to facilitate traceability.

More detailed information regarding usage and installation instructions can be found at https://github.com/etkayapar/bamRefine. bamRefine is also integrated to the Mapache ancient DNA pre-processing pipeline [62] (https://github.com/sneuensc/mapache).

PMD-correction strategies

We processed the data using three alternative strategies for avoiding the impact of PMD on genotypes.

(i) trimBAM: We applied trimming (clipping) to the sequencing data using the "*trimBam*" algorithm implemented in "*bamUtil (version 1.0.14)*" [12]. We trimmed (a) 10 bases from the ends of each read in non-damage repaired (non-UDG-treated) samples as well as in simulated ancient genomes, and (b) 2 bases from the ends of each read in damage repaired (UDG-treated) samples.

(ii) mapDamage: We applied rescaling to the sequencing data using the "*mapDamage2*" software [23]. We rescaled 10 bases from the ends of each read in simulated data using "*–rescale –seq-length 10*" parameters in non-damage repaired (non-UDG-treated) samples and (b) 4 bases (the lowest limit of mapDamage requires) from the ends of each read in damage repaired (UDG-treated) samples. We were unable to execute *mapDamage* analysis on UDG-treated published ancient samples, so we opted not to incorporate a *mapDamage* comparison in our analysis of published ancient data.

(iii) ATLAS: We applied the workflow that handles the estimation of recalibration and post-mortem damage (PMD) parameters by using "*ATLAS (version 0.9)*" [24]. First, we used default parameters "*task = PMD length = 50*" to estimate the extent of PMD in terms of position-specific. Then, for base quality score recalibration, we used *task = recal pmdFile = *PMD_input_Empiric.txt minDepth = 2*.

(iv) bamRefine: We applied refining to the sequencing data using "*bamRefine*." Similar to *trimming*, we refined (a) 10 bases (using "*–pmd-length-threshold 10*") from the ends of each read in non-damage repaired (non-UDG-treated) samples as well as the simulated ancient genomes, and (b) 2 bases (using "*–pmd-length-threshold 2*") from the ends of each read in damage repaired (UDG-treated) samples. Regardless of the samples being treated with UDG or not, we used our SNP dataset generated from the Turkish Genome Project for refining the simulated ancient reads (using

"*–snps* < *TGP-SNPS-FILE* > parameter") and the 1000G sub-Saharan African dataset for the empirical ancient reads (using "*–snps* < *AFR-SNPS-FILE* > " parameter).

We again caution that BAM files subject to any type of modification (filtering, as well as trimming, refining, or rescaling) should preferably not be published to facilitate reproducibility.

Dataset

In previous work, we had created a 1000 Genomes sub-Saharan African SNP diversity panel as a high-quality and relatively unbiased SNP dataset to use in demographic inference in Eurasian genomes [63]. The dataset includes 4,771,930 (4.7 M) bi-allelic autosomal SNPs ascertained in five sub-Saharan African populations in phase 3 of the 1000 Genomes project [58]. We used this dataset for genotyping the published ancient genomes included in the analysis.

Genotyping

We genotyped only targeted SNP positions; for simulated ancient genomes, these were the 313,747 positions defined from one individual of the TGP dataset [27], and for published ancient genomes these were the 4.7 M positions from 1000 Genomes sub-Saharan African dataset [63]. We called both diploid and pseudohaploid genotypes.

no PMD-correction and PMD-correction by using mapDamage, trimBam, and bamRefine

Diploid genotypes were obtained using "*GATK HaplotypeCaller (version 4.0.11.0)*" [48] by using the "*–min-base-quality-score 30, –minimum-mapping-quality 30, –genotyping-mode GENOTYPE_GIVEN_ALLELES, –output-mode EMIT_ALL_SITES*" parameters as well as the "*–alleles*" parameter to genotype the list of targeted SNP positions.

Pseudo-haploid genotypes were obtained by using "*pileupCaller (version 1.4.0)*" (https://github.com/stschiff/sequenceTools) by selecting one allele for each of the targeted SNP positions from the "*samtools mpileup*" [64] output file, which was generated by using the *"-R -B -q30 -Q30"* and the *"-l"* parameters to genotype the list of targeted SNP positions.

PMD-correction by using ATLAS

Diploid genotypes were obtained using "*ATLAS (version 0.9)*" [24] after modelling PMD with "*task* = *PMD length* = *50*" parameters (see above). Then we used empiric PMD estimation outputs for genotyping on maximum likelihood mode by using "*task* = *call method* = *MLE pmdFile* = **PMD_input_Empiric.txt recal* = **recalibrationEM.txt alleles* = *{a list of targeted SNP positions}" parameters.*

Pseudo-haploid genotypes were obtained by using "*ATLAS (version 0.9)*" [24] with the parameters *"task* = *call method* = *randomBase infoFields* = *DP pmdFile* = **PMD_input_Empiric.txt recal* = **recalibrationEM.txt alleles* = *{a list of targeted SNP positions}*." Here, this process was repeated 100 times for each sample.

Genotype concordance comparison

We assessed variant quality using 313,747 SNP positions, which included 182,515 homozygous reference (RR), 53,391 homozygous alternative (AA), and 77,841 heterozygous (RA) positions. These were used for evaluation. The SNPs inserted into simulated genomes used as the "True Genotypes" to empirically evaluate the accuracy of diploid SNP genotyping in 10X and 5X simulated genomes. For each simulated genome, we generated a 3×3 contingency table to capture all possible combinations of the three "Called Genotypes." We then estimated seven indices to evaluate genotype concordance: Concordance Rate (CR), False Negative Proportion (FN Proportion), False Positive Proportion (FP Proportion), Non-reference True Positive Rate (NTPR), recall (or sensitivity), specificity, and precision were calculated using the alternative allele as a reference, as illustrated in Additional file 1: Fig. S24 [50]. Additionally, we also calculated the F-score, which is the harmonic mean of precision and recall values, by the following formula: $2 \times$ [(Precision x Recall) / (Precision + Recall)].

f_4 statistics

We calculated f_4-statistics by using "*qpDstat (version: 980)*" algorithm implemented in "*AdmixTools (version 7.0.2)*" [65]. We used tests of the form f_4*(Human Reference Genome, Outgroup; Ind1_MappingStrategy1, Ind1_MappingStrategy2)* or f_4*(Human Reference Genome, Outgroup; Ind1_MappingStrategy1_PMDCorrectionStrategy1, Ind1_MappingStrategy1_PMDCorrectionStrategy2)* using the Chimpanzee Reference Genome (version panTro6) as an outgroup and with the "*f4mode: YES*" option. We used > 10,000 overlapping SNPs as cut-off for reporting f_4-test calculations.

Other statistical tests and visualization

We used the non-parametric Mann–Whitney *U* (MWU) test for testing for systematic differences between two groups in their average estimates (e.g., bias levels between "LINEAR" and "GRAPH" strategies). The data used here was not paired. It also included possible differences in variance among groups, for which reason we decided to choose a non-parametric two sample test.

We produced all graphs in R [66] after reading and manipulating data using "*gsheet*" [67] and "*tidyverse*" [68] packages. All figures were produced by using "*ggplot2*" [69] and its extension packages "*ggpubr*" [70], "*ggh4x*" [71], and "*ggpattern*" [72]. The multiple panel figures are combined by using the "*patchwork*" package [73]. In some figures, colors were assigned by using "*MetBrewer*" package [74].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-024-03462-w.

Additional file 1: This file contains Figures S1-35 to support the methodology and the main results.

Additional file 2: Table S1A: Sequencing statistics for simulated genomes. Table S1B: Number and characteristics of variants used to generate aDNA-like simulated genomes. The variants were chosen from a bi-allelic SNP set discovered in the individual 06A010111 of the Turkish Genome Project dataset (see Methods). Table S1C: Information on the published ancient genomes used in this study. "Sample ID": the genome ID used in the relevant publication. "Data type": whether the genome was shotgun-sequenced, or sequenced after whole-genome capture or 1240 K SNP capture. "Genome Coverage": genome-wide coverage for the samples generated by shotgun or whole-genome capture. "1240 K Coverage": coverage of targeted 1240 K SNPs for the samples generated by 1240 K capture. "Location": the country where the ancient genome material was retrieved. "Input file type": the published type of data available for analysis, i.e. FASTQ, BAM, or BAM (unfiltered). "PMD corr. bp": how many bases at the end of the reads to

PMD correction (rescaling, trimming). "Number of heterozygote SNPs": how many heterozygote SNPs are defined for each sample to use to calculate reference bias for Figs. 2B-D. These heterozygous positions were defined as positions where alternative allele frequencies were 25–75% and had a minimum of 10 reads.

Additional file 3: Table S2: The proportion of alternative alleles at heterozygous sites in aDNA-like simulated (Table S2A) and published ancient genomes (Table S2B). The proportions were calculated by randomly selecting one allele from 77,841 heterozygote sites in aDNA-like simulated genomes 100 times using pileupCaller. The "Alternative allele proportion" columns show the mean, minimum (min) and maximum (max) of the distribution of these proportions. "Coverage" indicates the depth-of-coverage. The "Alignment strategy" and "PMD correction strategy" columns indicate the methods used for alignment and PMD correction, respectively.

Additional file 4: Tables S3: Number of SNPs genotyped aDNA-like simulated (Table S3A) and published ancient genomes (Table S3B).

Additional file 5: Review history.

Acknowledgements
We thank all colleagues at the METU CompEvo and Hacettepe Human_G groups, and Torsten Günther and Anders Götherström for their support and helpful suggestions, and three anonymous reviewers for their constructive comments. We thank Anders Götherström, Mattias Jakobsson, and Carolina Bernhardsson for sharing the raw FASTQ files for the Bon002, prs013, mfo001, and irk034 genomes. We also thank Samuel Neuenschwander for implementing *bamRefine* in the *Mapache* pipeline.

Peer review information
Andrew Cosgrove was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions
Conceptualization, DK, EY, CA, MS; ***Methodology***, DK, EY, CA, MS***; Software, DK, EY, CA, MS; Formal Analysis***, DK, EY, KBV, ES, NEA, CA; ***Writing – Original Draft***, DK, EY, ASM, MS; ***Writing – Review & Editing***, KBV, ES, NEA, CA; ***Visualization***, DK, NEA.; ***Supervision***, ASM, CA, MS.

Funding
Wenner-Gren Foundation Dissertation Fieldwork grant (no. 9573 to DK), H2020 ERC Consolidator grant (no. 772390 NEOGENE to MS), and Swiss National Science Foundation (SNSF) project grant (no. PCEGP3_181251 to ASM).

Data availability
All samples of FASTQ/BAM files were downloaded from the online data repositories [30–45] using the accession numbers provided in the respective publications [13, 20, 62, 75–87], except for Bon002, prs013, mfo001, and irk034, for which raw FASTQ files were obtained from the corresponding authors of the relevant publications (see Additional file2: Table S1C). bamRefine is available at GitHub (https://github.com/etkayapar/bamrefine) [88], PyPI (https://pypi.org/project/bamrefine), and at Zenodo (https://doi.org/10.5281/zenodo.14234666) [89] under the BSD-3-Clause license. It is also integrated to the ***Mapache*** ancient DNA pre-processing pipeline (https://github.com/sneuensc/mapache) [62].

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 26 November 2023 Accepted: 16 December 2024
Published online: 09 January 2025

References

1. Prüfer K, Stenzel U, Hofreiter M, Pääbo S, Kelso J, Green RE. Computational challenges in the analysis of ancient DNA. Genome Biol. 2010;11:R47.
2. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. PLOS Genet. 2019;15:e1008302.
3. Martiniano R, Garrison E, Jones ER, Manica A, Durbin R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. Genome Biol. 2020;21:250.
4. Prasad A, Lorenzen ED, Westbury MV. Evaluating the role of reference-genome phylogenetic distance on evolutionary inference. Mol Ecol Resour. 2022;22:45–55.
5. Thorburn D-MJ, Sagonas K, Binzer-Panchal M, Chain FJJ, Feulner PGD, Bornberg-Bauer E, et al. Origin matters: Using a local reference genome improves measures in population genomics. Mol Ecol Resour. 2023;23:1706–23.
6. Prüfer K. snpAD: an ancient DNA genotype caller. Bioinformatics. 2018;34:4165–71.

7. Peyrégne S, Slon V, Mafessoni F, de Filippo C, Hajdinjak M, Nagel S, et al. Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. Sci Adv. 2019;5:eaaw5873.
8. Oliva A, Tobler R, Cooper A, Llamas B, Souilmi Y. Systematic benchmark of ancient DNA read mapping. Brief Bioinform. 2021;22:bbab076.
9. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010;38:e87.
10. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. Philos Trans R Soc B Biol Sci. 2015;370:20130624.
11. Bainbridge MN, Wang M, Wu Y, Newsham I, Muzny DM, Jefferies JL, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. Genome Biol. 2011;12:R68.
12. Jun G, Wing MK, Abecasis GR, Kang HM. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. Genome Res. 2015;25:918–25.
13. Koptekin D, Yüncü E, Rodríguez-Varela R, Altınışık NE, Psonis N, Kashuba N, et al. Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean. Curr Biol. 2023;33:41–57.e15.
14. Altınışık NE, Kazancı DD, Aydoğan A, Gemici HC, Erdal ÖD, Sarıaltun S, et al. A genomic snapshot of demographic and cultural dynamism in Upper Mesopotamia during the Neolithic Transition. Sci Adv. 2022;8:eabo3609.
15. Skourtanioti E, Ringbauer H, Gnecchi Ruscone GA, Bianco RA, Burri M, Freund C, et al. Ancient DNA reveals admixture history and endogamy in the prehistoric Aegean. Nat Ecol Evol. 2023;7:290–303.
16. Clemente F, Unterländer M, Dolgova O, Amorim CEG, Coroado-Santos F, Neuenschwander S, et al. The genomic history of the Aegean palatial civilizations. Cell. 2021;184:2565–2586.e21.
17. Daly KG, Mattiangeli V, Hare AJ, Davoudi H, Fathi H, Doost SB, et al. Herded and hunted goat genomes from the dawn of domestication in the Zagros Mountains. Proc Natl Acad Sci. 2021;118:e2100901118.
18. Mattila TM, Svensson EM, Juras A, Günther T, Kashuba N, Ala-Hulkko T, et al. Genetic continuity, isolation, and gene flow in Stone Age Central and Eastern Europe. Commun Biol. 2023;6:1–13.
19. Gelabert P, Sawyer S, Bergström A, Margaryan A, Collin TC, Meshveliani T, et al. Genome-scale sequencing and analysis of human, wolf, and bison DNA from 25,000-year-old sediment. Curr Biol. 2021;31:3564–3574.e9.
20. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528:499–503.
21. Bongers JL, Nakatsuka N, O'Shea C, Harper TK, Tantaleán H, Stanish C, et al. Integration of ancient DNA with transdisciplinary dataset finds strong support for Inca resettlement in the south Peruvian coast. Proc Natl Acad Sci. 2020;117:18359–68.
22. Rivollat M, Thomas A, Ghesquière E, Rohrlach AB, Späth E, Pemonge M-H, et al. Ancient DNA gives new insights into a Norman Neolithic monumental cemetery dedicated to male elites. Proc Natl Acad Sci. 2022;119:e2120786119.
23. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinformatics. 2013;29:1682–4.
24. Link V, Kousathanas A, Veeramah K, Sell C, Scheu A, Wegmann D. ATLAS: analysis tools for low-depth and ancient samples. bioRxiv. 2017:105346. Available from: https://www.biorxiv.org/content/10.1101/105346v2. Cited 2024 Dec 6.
25. da Mota BS, Rubinacci S, Cruz Dávalos DI, Amorim CEG, Sikora M, Johannsen NN, et al. Imputation of ancient human genomes. Nat Commun. 2023;14:3660.
26. Çubukcu H, Kılınç GM. Evaluation of genotype imputation using glimpse tools on low coverage ancient DNA. Mamm Genome. 2024;35:461–73.
27. Alkan C, Kavak P, Somel M, Gokcumen O, Ugurlu S, Saygi C, et al. Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe. Asia Africa BMC Genomics. 2014;15:963.
28. Renaud G, Hanghøj K, Willerslev E, Orlando L. gargammel: a sequence simulator for ancient DNA. Bioinformatics. 2017;33:577–9.
29. Rakocevic G, Semenyuk V, Lee W-P, Spencer J, Browning J, Johnson IJ, et al. Fast and accurate genomic analyses using genome graphs. Nat Genet. 2019;51:354–62.
30. Middle East Technical University. Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia. European Nucleotide Archive; 2022. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB51705.
31. Beijing Genome Institute. Ancient human genome sequence of an extinct Palaeo-Eskimo. National Library of Medicine; 2010. Available from: https://www.ncbi.nlm.nih.gov/sra?term=SRA010102.
32. UPPSALA UNIVERSITY. Southern African ancient genomes estimate modern human divergence to 350,000-260,000 years ago [Internet]. European Nucleotide Archive; 2017. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB22660.
33. Middle East Technical University. The demographic development of the first farmers and their expansion in anatolia and beyond. European Nucleotide Archive; 2016. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB14675.
34. Stockholm University, Hacettepe University. Human population dynamics and Yersinia pestis in ancient northeast Asia. European Nucleotide Archive; 2020. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB39378.
35. Max Planck Institute for Evolutionary Anthropology. The Anglo-Saxon migration and formation of the Early English Gene pool. European Nucleotide Archive; 2022. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB54899.
36. Smurfit Institute of Genetics. Ancient Ethiopian "'Mota"' genome. National Library of Medicine; 2015. Available from: https://www.ncbi.nlm.nih.gov/bioproject/PRJNA295861.
37. UPPSALA University. Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. European Nucleotide Archive; 2019. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB31045.
38. HARVARD MEDICAL SCHOOL. Eight thousand years of natural selection in Europe. European Nucleotide Archive; 2015. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB11450.

39. Max Planck Institute for Evolutionary Anthropology;MPI-EVA. The genome sequence of a 45,000-year-old modern human from Ust'-Ishim. European Nucleotide Archive; 2014. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB6622.
40. Reich Lab. Ancient human genomes suggest three ancestral populations for present-day Europeans. European Nucleotide Archive; 2014. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB6272.
41. Centre for geogenetics, natural history museum of Denmark. The first horse herders and the impact of early bronze age steppe expansions into Asia. European Nucleotide Archive; 2018. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB26349.
42. Technical University of Denmark, Department of Health Technology. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. 2014. Available from: https://services.healthtech.dtu.dk/suppl/clovis/.
43. HARVARD MEDICAL SCHOOL. Parallel paleogenomic transects reveal complex genetic history of early European farmers. European Nucleotide Archive; 2017. Available from: Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/PRJEB22629.
44. Harvard Medical School, Department of Genetics. The beaker phenomenon and the genomic transformation of Northwest Europe. European Nucleotide Archive; 2018. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB23635.
45. Harvard Medical School. Population Turnover in Remote Oceania Shortly After Initial Settlement. European Nucleotide Archive; 2018. Available from: https://www.ebi.ac.uk/ena/browser/view/PRJEB24938.
46. Rohland N, Mallick S, Mah M, Maier R, Patterson N, Reich D. Three assays for in-solution enrichment of ancient human DNA at more than a million SNPs. Genome Res. 2022;32:2068–78.
47. Davidson R, Williams MP, Roca-Rada X, Kassadjikova K, Tobler R, Fehren-Schmitz L, et al. Allelic bias when performing in-solution enrichment of ancient human DNA. Mol Ecol Resour. 2023;23:1823–40.
48. Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Auwera GAV der, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178. Available from: https://www.biorxiv.org/content/10.1101/201178v3. Cited 2024 Dec 6.
49. Bergström A. Improving data archiving practices in ancient genomics. Sci Data. 2024;11:754.
50. Kishikawa T, Momozawa Y, Ozeki T, Mushiroda T, Inohara H, Kamatani Y, et al. Empirical evaluation of variant calling accuracy using ultra-deep whole-genome sequencing data. Sci Rep. 2019;9:1784.
51. Ausmees K, Sanchez-Quinto F, Jakobsson M, Nettelblad C. An empirical evaluation of genotype imputation of ancient DNA. G3 (Bethesda). 2022;12:jkac089.
52. Söylev A, Çokoglu SS, Koptekin D, Alkan C, Somel M. CONGA: copy number variation genotyping in ancient genomes and low-coverage sequencing data. PLOS Comput Biol. 2022;18:e1010788.
53. Sikora M, Canteri E, Fernandez-Guerra A, Oskolkov N, Ågren R, Hansson L, et al. The landscape of ancient human pathogens in Eurasia from the Stone Age to historical times. bioRxiv. 2023:2023.10.06.561165. Available from: https://www.biorxiv.org/content/10.1101/2023.10.06.561165v1. Cited 2024 Dec 6.
54. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
55. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes. 2016;9:88.
56. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
57. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
58. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.
59. Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature. 2016;538:201–6.
60. Kircher M. Analysis of high-throughput ancient DNA sequencing data. In: Shapiro B, Hofreiter M, editors. Anc DNA Methods Protoc. Totowa, NJ: Humana Press; 2012. p. 197–228. https://doi.org/10.1007/978-1-61779-516-9_23. Cited 2024 Dec 6.
61. Sirén J, Monlong J, Chang X, Novak AM, Eizenga JM, Markello C, et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. Science. 2021;374:abg8871.
62. de Barros Damgaard P, Martiniano R, Kamm J, Moreno-Mayar JV, Kroonen G, Peyrot M, et al. The first horse herders and the impact of early bronze age steppe expansions into Asia. Science. 2018;360:eaar7711.
63. Koptekin D. Spatial and temporal heterogeneity in human mobility patterns in Holocene Southwest Asia and the East Mediterranean. Zenodo; 2022. Available from: https://zenodo.org/records/6377228. Cited 2024 Dec 7.
64. Danecek P, Bonfield JK, Liddle J, Marshall J, Ohan V, Pollard MO, et al. Twelve years of SAMtools and BCFtools. GigaScience. 2021;10:giab008.
65. Patterson N, Moorjani P, Luo Y, Mallick S, Rohland N, Zhan Y, et al. Ancient admixture in human history. Genetics. 2012;192:1065–93.
66. R Core Team. R: A Language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2023. Available from: https://www.R-project.org/.
67. Conway M. gsheet: Download Google Sheets Using Just the URL. 2020. Available from: https://cran.r-project.org/web/packages/gsheet/index.html. Cited 2024 Dec 6.
68. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. J Open Source Softw. 2019;4:1686.
69. Wickham H. ggplot2. Cham: Springer International Publishing; 2016. Available from: http://link.springer.com/10.1007/978-3-319-24277-4. Cited 2024 Dec 6.
70. Kassambara A. ggpubr: "ggplot2" Based Publication Ready Plots. 2023. Available from: https://cran.r-project.org/web/packages/ggpubr/index.html. Cited 2024 Dec 6.

71. Brand T van den. ggh4x: Hacks for "ggplot2". 2024. Available from: https://cran.rstudio.com/web/packages/ggh4x/index.html. Cited 2024 Dec 6.
72. FC M, Davis TL, authors ggplot2. ggpattern: "ggplot2" Pattern Geoms. 2024. Available from: https://cran.r-project.org/web/packages/ggpattern/index.html. Cited 2024 Dec 6.
73. Pedersen TL. patchwork: The Composer of Plots. 2024. Available from: https://cran.r-project.org/web/packages/patchwork/index.html. Cited 2024 Dec 6.
74. Mills BR. MetBrewer: Color Palettes Inspired by Works at the Metropolitan Museum of Art. 2022. Available from: https://cran.r-project.org/web/packages/MetBrewer/index.html. Cited 2024 Dec 6.
75. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463:757–62.
76. Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, et al. Population turnover in remote oceania shortly after initial settlement. Curr Biol. 2018;28:1157–1165.e7.
77. Lipson M, Szécsényi-Nagy A, Mallick S, Pósa A, Stégmár B, Keerl V, et al. Parallel palaeogenomic transects reveal complex genetic history of early European farmers. Nature. 2017;551:368–72.
78. Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, et al. The Beaker phenomenon and the genomic transformation of northwest Europe. Nature. 2018;555:190–6.
79. Gretzinger J, Sayer D, Justeau P, Altena E, Pala M, Dulias K, et al. The Anglo-Saxon migration and the formation of the early English gene pool. Nature. 2022;610:112–9.
80. Fu Q, Li H, Moorjani P, Jay F, Slepchenko SM, Bondarev AA, et al. Genome sequence of a 45,000-year-old modern human from western Siberia. Nature. 2014;514:445–9.
81. Kılınç GM, Omrak A, Özer F, Günther T, Büyükkarakaya AM, Bıçakçı E, et al. The demographic development of the first farmers in Anatolia. Curr Biol. 2016;26:2659–66.
82. Lazaridis I, Patterson N, Mittnik A, Renaud G, Mallick S, Kirsanow K, et al. Ancient human genomes suggest three ancestral populations for present-day Europeans. Nature. 2014;513:409–13.
83. Llorente MG, Jones ER, Eriksson A, Siska V, Arthur KW, Arthur JW, et al. Ancient Ethiopian genome reveals extensive Eurasian admixture in Eastern Africa. Science. 2015;350:820–2.
84. Kılınç GM, Kashuba N, Koptekin D, Bergfeldt N, Dönertaş HM, Rodríguez-Varela R, et al. Human population dynamics and Yersinia pestis in ancient northeast Asia. Sci Adv. 2021;7:eabc4587.
85. Rasmussen M, Anzick SL, Waters MR, Skoglund P, DeGiorgio M, Stafford TW, et al. The genome of a Late Pleistocene human from a Clovis burial site in western Montana. Nature. 2014;506:225–9.
86. Schlebusch CM, Malmström H, Günther T, Sjödin P, Coutinho A, Edlund H, et al. Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. Science. 2017;358:652–5.
87. Sánchez-Quinto F, Malmström H, Fraser M, Girdland-Flink L, Svensson EM, Simões LG, et al. Megalithic tombs in western and northern Neolithic Europe were linked to a kindred society. Proc Natl Acad Sci. 2019;116:9469–74.
88. Yapar E. etkayapar/bamRefine. 2024. Available from: https://github.com/etkayapar/bamRefine. Cited 2024 Dec 7.
89. Yapar E. etkayapar/bamRefine: v0.2.1. 2024. Available from: https://zenodo.org/records/14234666. Cited 2024 Dec 6.

Publisher's Note