scientific reports

Check for updates

OPEN

Assessing the impact of post-mortem damage and contamination on imputation performance in ancient DNA

Antonio Garrido Marques[1], Simone Rubinacci[2,3], Anna-Sapfo Malaspinas[1,4], Olivier Delaneau[5] & Bárbara Sousa da Mota[1,4]✉

Low-coverage imputation is becoming ever more present in ancient DNA (aDNA) studies. Imputation pipelines commonly used for present-day genomes have been shown to yield accurate results when applied to ancient genomes. However, *post-mortem* damage (PMD), in the form of C-to-T substitutions at the reads termini, and contamination with DNA from closely related species can potentially affect imputation performance in aDNA. In this study, we evaluated imputation performance (i) when using a genotype caller designed for aDNA, ATLAS, compared to bcftools, and (ii) when contamination is present. We evaluated imputation performance with principal component analyses and by calculating imputation error rates. With a particular focus on differently imputed sites, we found that using ATLAS prior to imputation substantially improved imputed genotypes for a very damaged ancient genome (42% PMD). Trimming the ends of the sequencing reads led to similar improvements in imputation accuracy. For the remaining genomes, ATLAS brought limited gains. Finally, to examine the effect of contamination on imputation, we added various amounts of reads from two present-day genomes to a previously downsampled high-coverage ancient genome. We observed that imputation accuracy drastically decreased for contamination rates above 5%. In conclusion, we recommend (i) accounting for PMD by either trimming sequencing reads or using a genotype caller such as ATLAS before imputing highly damaged genomes and (ii) only imputing genomes containing up to 5% of contamination.

Over the past decade, there has been a fast increase of sequenced ancient genomes[1]. The growing amount of ancient genetic data allows for a better understanding of the evolutionary history of humans and other species, as shown by a multitude of discoveries in recent years[2–5].

The field of ancient DNA (aDNA) research is accompanied by a unique set of challenges, namely, *post-mortem* damage (PMD) and contamination. PMD manifests as DNA fragmentation, and pronounced C-to-T deamination at the ends of DNA fragments[6], that accumulate over time due to the absence of DNA repair mechanisms following an organism's death. PMD can complicate sequencing and lead to erroneous genotype calls[7]. A strategy to mitigate C-to-T deamination is the application of a uracil-DNAglycosylase (UDG) and endonuclease VIII treatment during laboratory preparation[8,9]. UDG is an enzyme that recognizes and removes uracil residues arising from the deamination of cytosine[10]. The removal of uracil by UDG creates abasic sites, which are then recognized and cleaved by an endonuclease VIII. By actively targeting and removing uracil from aDNA, this treatment minimizes the incidence of post-mortem C-to-T deamination, thereby enhancing the accuracy of subsequent analyses.

Despite meticulous procedures intended to prevent contamination[11–13], it is hard to avoid and sources of contamination can range from unintentionally introduced human DNA during sample extraction or laboratory processing[14–16], to the persistent presence of DNA from microbial and environmental sources[17]. Microbial and environmental contamination are responsible for the underrepresentation of endogenous DNA. The competitive

[1]Department of Computational Biology, University of Lausanne, Lausanne, Switzerland. [2]Division of Genetics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [3]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [4]Swiss Institute of Bioinformatics, University of Lausanne, Lausanne, Switzerland. [5]Regeneron Genetics Center, Tarrytown, NY, USA. ✉email: barbara.mota@unil.ch

amplification of this non-target DNA results in the ancient DNA of interest to become overshadowed and to numerous genomic positions being either weakly covered or entirely devoid of sequencing reads. While we can practically remove all microbial and environmental contamination by mapping the sequenced reads, contamination with DNA from the same or closely-related species cannot be fully eliminated. Even at minimal levels, this type of contamination can skew downstream analyses, leading to potentially erroneous conclusions. As such, contamination estimation is a standard quality control step in aDNA studies. Genomes with a contamination level exceeding 5% (an adhoc threshold used in several studies, e.g., Allentoft et al.[18] and Nakatsuka et al.[19]) are often discarded, resulting in the loss of potentially informative data.

Both PMD and contamination contribute to reduced genome coverage. On the one hand, PMD produces shorter DNA fragments with reduced mappability, and the deaminations introduced by PMD can increase discrepancies relative to the reference genome, making the fragments harder to map[20,21]. On the other hand, microbial contaminant DNA tends to be several times more abundant than the endogenous DNA that is under study, resulting in low amounts of the latter and, consequently, low depth of coverage.

A standard approach to address the challenge of low coverage in ancient genomes is pseudo haploidization. This method involves selecting one single read to represent a haploid genotype at each genomic position[13]. However, this method carries inherent risks such as under-estimation of allele frequencies and may inadvertently introduce a bias towards the reference genome[22–24].

A different approach is genotype imputation. This is a method that makes use of statistical models to infer missing genotypes in a target sample, guided by a reference panel of haplotypes[25]. Given the nature of ancient genomes, one should apply low-coverage WGS (whole-genome sequencing) imputation methods[26–28]. In this case, genotype likelihoods (GLs) are the input data for imputation. These GLs are the probabilities of finding certain genotypes based on the available sequencing data, and can thus encode the uncertainty inherent to low-coverage ancient genomes.

While primarily used for present-day low-coverage genomes, recent studies have demonstrated the effectiveness of genotype imputation of ancient genomes[29–31]. In particular, imputing shotgun-sequenced non-African ancient genomes with at least 0.5 × coverage yields accurate results. For in-solution capture (1240 K[32–34]) sequenced genomes, which constitute the majority of publicly available ancient genomes, imputation accuracy tends to be lower and better results are attained at the capture sites, i.e., the 1240 K SNPs, and for coverages above 2 × at these sites[31]. Nevertheless, while this approach addresses the low-coverage problem in aDNA, it is unclear how contamination and PMD affect imputation and whether we could use imputation to tackle these challenges as well. Although it was shown that imputation can have a corrective effect on deaminated sites[31], it is not clear to what extent PMD affects imputation performance, and hence what the best practices are regarding pre-imputation PMD filtering. Some studies opted to exclude potentially deaminated positions before imputing ancient genomes[29,30]. But, since the PMD-inflicted SNPs are transitions that constitute around two thirds of all SNPs in humans[35], excluding such positions altogether leads to loss of information that can potentially affect the imputation output. Moreover, to our knowledge, there has been no research on the impact of contamination on imputation. Thus, a description of imputation impact on contaminated sites and vice-versa is lacking.

In this study, we aim to i) assess whether imputation performance can be improved by using a genotype caller specifically designed to account for PMD and ii) study the impact of contamination levels on imputation and vice-versa (i.e., can genotype imputation "correct" contaminated positions so as to retrieve the original genotypes?). To address these aims, we performed simulations to mimic the properties of ancient DNA. We achieved this by downsampling high-coverage ancient genomes to low coverage, and then performing imputation experiments. When investigating the effect of PMD, we compared two imputation pipelines: a combination of GLIMPSE[28] with GLs generated with either (i) bcftools[36], a widely used genotype caller, or (ii) ATLAS[37], a genotype caller designed to explicitly take into account PMD when generating the calls. To determine how contamination affects imputation, we artificially contaminated an ancient genome with varying amounts of sequencing reads coming from present-day genomes. In the following sections, we show that, while taking PMD into account prior to imputing has a noticeable impact in the presence of high deamination rates, it does not substantially improve imputation performance in other cases. Furthermore, we demonstrate that contamination can drastically reduce imputation accuracy, rendering imputed genotypes unreliable.

Results

Methodology for comparing imputation accuracy of ancient genomes using distinct genotype callers

To assess the impact of PMD on imputation accuracy, we imputed two sets of GLs that we called using either (i) ATLAS or (ii) bcftools, and compared these to validation datasets through non-reference discordance (NRD) estimates and principal component analysis (PCA), focusing on SNPs differently imputed across the two datasets. ATLAS, specifically designed for aDNA, infers PMD patterns and recalibrates base quality scores. In contrast, bcftools serves as a conventional genotype caller providing a comparative dataset. We performed these imputation experiments using seven high-coverage ancient genomes of diverse origins and varying PMD rates (Table S1, Supplementary Figures S1 and S2). We downsampled these genomes to emulate low-coverage (1x) conditions, and used the high-coverage genomes as validation (Fig. 1A). We generated three validation datasets: (i) by calling genotypes with ATLAS ("ATLAS validation"), (ii) by calling genotypes with bcftools ("bcftools validation"), and (iii) by intersecting the two previous genotype sets so as to only keep concordant sites between the two ("validation concordant"). However, we mostly relied on the validation concordant set due to inherent difficulties in establishing a ground truth for ancient genomes (Supplementary Figure S3). To better disentangle the effects of PMD from other factors affecting imputation performance, we also simulated low-coverage diploid chromosomes with gargammel[38] with varying amounts of PMD that we subsequently imputed. We used as a

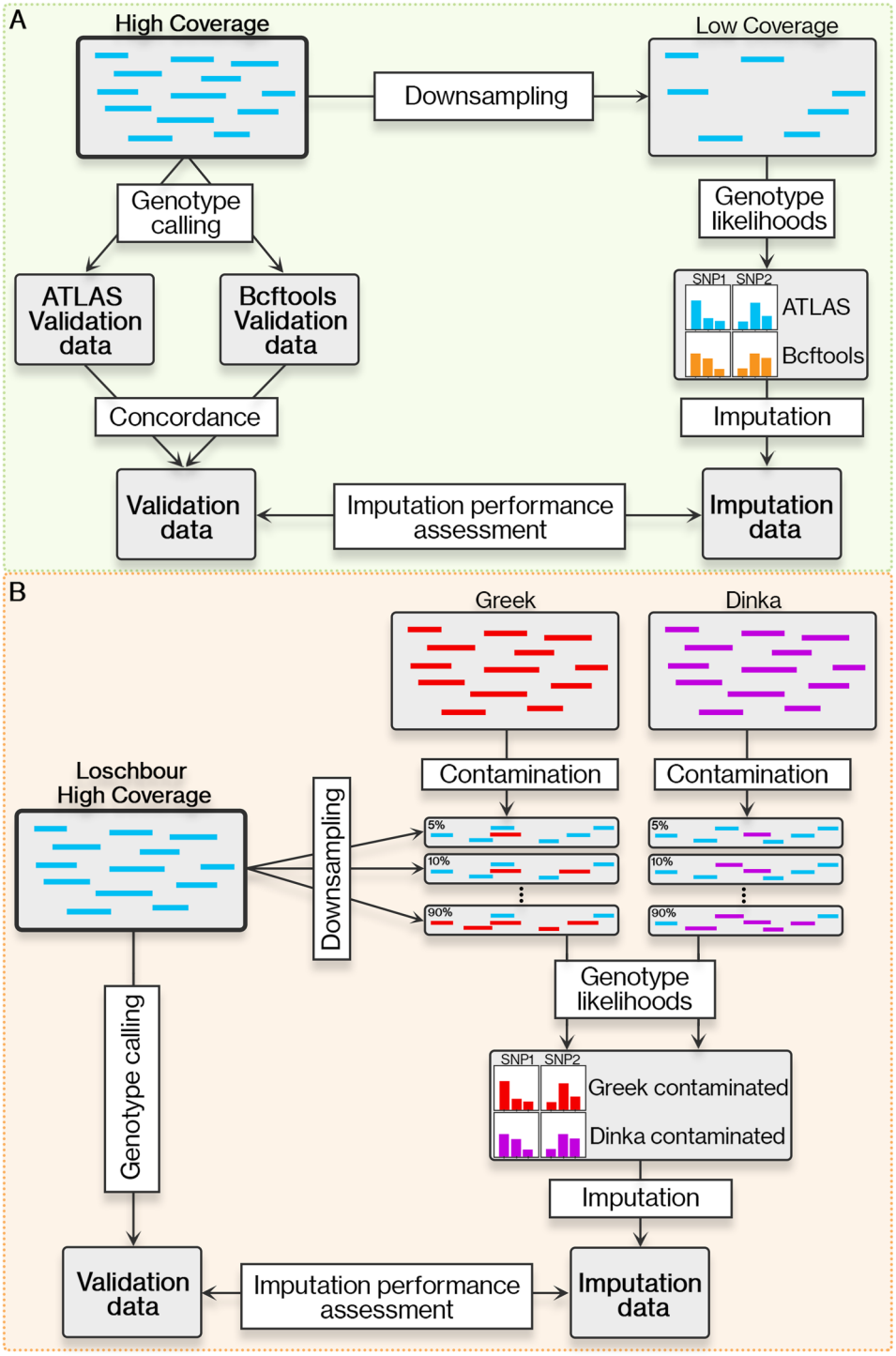

Figure 1. Schematic representation of the methodology. (**A**) Outline of the approach used to evaluate the effect of two different genotype callers on the imputation accuracy of ancient genomes. (**B**) Methodology followed to gauge the impact of contamination on genotype imputation of ancient genomes.

template a present-day genome from Turkey[24,39]. Afterwards, we performed genotype imputation on the two sets of genotypes likelihoods using GLIMPSE v1.1.1[28], a genotype imputation tool for low-coverage WGS data. To evaluate imputation performance, we compared our results with the three different validation datasets using the NRD metric, defined as NRD = (eRR + eRA + eAA) / (eRR + eRA + eAA + mRA + mAA), where 'e' and 'm' represent the number of errors and matches, respectively, while 'R' and 'A' stand for reference and alternative alleles, respectively. We calculated NRD on all the 1000 Genomes bi-allelic SNPs[40], on the intersection of the 1000 Genomes and 1240 K bi-allelic SNPs[32–34], and on a set of 2.8 M SNPs (see Methods section). Restricting

to the 2.8 M SNPs, we also performed PCA using smartPCA[41,42] to visualize and measure where imputed samples are positioned relative to their validation concordant set across the first 10 principal components (PCs). The PCA components were calculated using worldwide present-day genomes from SGDP[43] and we projected the ancient data onto these. For each ancient individual, we projected four datasets: (i) validation concordant, (ii) identically imputed genotypes when using the two GLs sets ("imputed concordant"), and the differently imputed genotypes across the two GLs datasets computed with (iii) ATLAS ("ATLAS imputed discordant") and (iv) bcftools ("bcftools imputed discordant"). Finally, for each individual sample, we calculated the Euclidean distance to their validation and adjusted by the corresponding eigenvalue of each PC with a focus on the subset of SNPs that are discordant.

Memory and time requirements of ATLAS and bcftools to generate genotype likelihoods

To assess the computational efficiency of genotype calling with either ATLAS or bcftools, we measured running times and memory usage when computing GLs for the 1× downsampled and the high-coverage genomes (Supplementary Figure S4). We observed that ATLAS required considerably longer running times. For chromosome 1 and the downsampled genomes, ATLAS took approximately 48 times more time than bcftools, averaging ~ 1.6 h per sample, while for the high-coverage genomes, ATLAS took ~ 5.9 h, that is, 17 times more time than bcftools. Genome-wide, ATLAS also required substantially more memory resources, using around 45 times the memory allocation of bcftools with an average of ~ 182 Gb per sample for the downsampled genomes, and ~ 3200 Gb for the high-coverage genomes (768 times more memory than bcftools for the same genomes). It should be noted, however, that generating GLs with ATLAS entails three steps instead of one, as is the case of bcftools: (i) splitting of single-end read groups according to read length, and merging of paired-end reads if present, (ii) generating empirical estimates for PMD, which are then used to refine the calling step, and (iii) genotype calling.

Impact of taking PMD into account on imputation accuracy

Accounting for PMD improves imputation performance in highly damaged individual samples

When restricting to identically imputed genotypes using ATLAS and bcftools GLs, that is, the imputed concordant set, we observed that these data were placed on the PCA space close to present-day individuals from the same continents, as expected. Furthermore, these datasets co-localised with their respective validation (validation concordant) (Fig. 2A). We found that the imputed ATLAS discordant data was closer to the validation for all individual samples except NE1[44] (9.6% PMD), as depicted in Fig. 2A, B. Different trends were observed when using the 1240 K SNPs (Supplementary Fig. S5), though. The biggest differences between imputed ATLAS and bcftools data were found for BOT2016[45] (20.8% PMD, 0.51% discordant sites), Sumidouro5[46] (42.0% PMD, 0.86% discordant sites), Yana[47] (13.9% PMD, 0.92% discordant sites), and ela01[48] (3.0% PMD, 1.67% discordant sites). This disparity is particularly noticeable for Sumidouro5, that exhibits the highest PMD rate (42.0%), and whose discordant imputed ATLAS genotypes are considerably closer to the validation. In the case of Yamnaya[45] (18.8% PMD, 0.26% discordance sites), the two discordant datasets were similarly close to their validation (weighted Euclidean distances of 0.40 vs. 0.42 for ATLAS and bcftools discordant data, respectively). In conclusion, the discordantly imputed bcftools GLs tend to be more different from the validation, and thus less accurate than the imputed ATLAS GLs. However, the proportions of the discordantly imputed SNPs were relatively low, ranging between 0.26% and 1.67%, indicating a high level of similarity between the two imputed datasets. A PCA of the same data without splitting the imputed datasets into concordantly and discordantly imputed SNPs can be found in Supplementary Figure S6.

An examination of the imputation error rates allows for a deeper understanding of how imputation is affected by distinct genotype callers. We computed NRD and other error rates using three validation sets obtained from the high-coverage genotypes: i) bcftools calls, ii) ATLAS calls and iii) validation concordant (intersection of i) and ii)). We computed NRD for three different sets of SNPs: (i) the same SNPs as in the abovementioned PCA analysis (2.8 M SNPs) (Fig. 2C–E, for more detailed error rates see Table S2), (ii) the intersection of the 1000 Genomes and 1240 K bi-allelic SNPs all sites (Supplementary Figure S5C–E), and (iii) all sites in the 1000 Genomes panel (Supplementary Figure S5F–H). At first sight, the (2.8 M SNPs) NRD results corroborate the previous observations regardless of the validation dataset. Imputation using ATLAS GLs resulted in lower NRD for ela01, Yana and Sumidouro5, while the opposite was seen for NE1. We caution that the NRD values calculated at all 1000 Genomes SNPs yielded the opposite trend for ela01. For SZ1[49], Yamnaya and BOT2016, the two imputed datasets had approximately the same NRD values, ~ 1.23%, ~ 0.85% and ~ 1.58%, respectively. While such a result was expected for SZ1, given that its libraries were UDG-treated and hence no gain in accuracy was expected with ATLAS, PMD is prevalent in the Yamnaya and the BOT2016 genomes (18.8% and 20.8% PMD, respectively) and we would expect ATLAS to have a considerable impact. Therefore, using one or the other genotype caller prior to imputation did not systematically affect imputation performance given PMD rates as expected. This suggests that there are other factors affecting GLs computation and/or imputation that ultimately impact imputation performance in a non-trivial way. Nonetheless, Sumidouro5 represents the clearest example of how using ATLAS to generate GLs prior to imputation can result in higher imputation accuracy, which is particularly visible when using the ATLAS validation dataset as the ground truth. In this case, imputed ATLAS GLs yielded 1.03% NRD, two times smaller than the NRD for imputed bcftools GLs (2.08%). Furthermore, in the case of Sumidouro5, we verified that combining trimming five base pairs at the reads ends with bcftools calling led to similar improvements in imputation performance as using ATLAS-generated genotype likelihoods. However, trimming 10 base pairs resulted in loss of information that negatively impacted imputation performance, with genome coverage dropping from 1× to 0.54× compared to 0.76× in the case of trimming five base pairs (Supplementary Figures S7, S8 and Table S3).

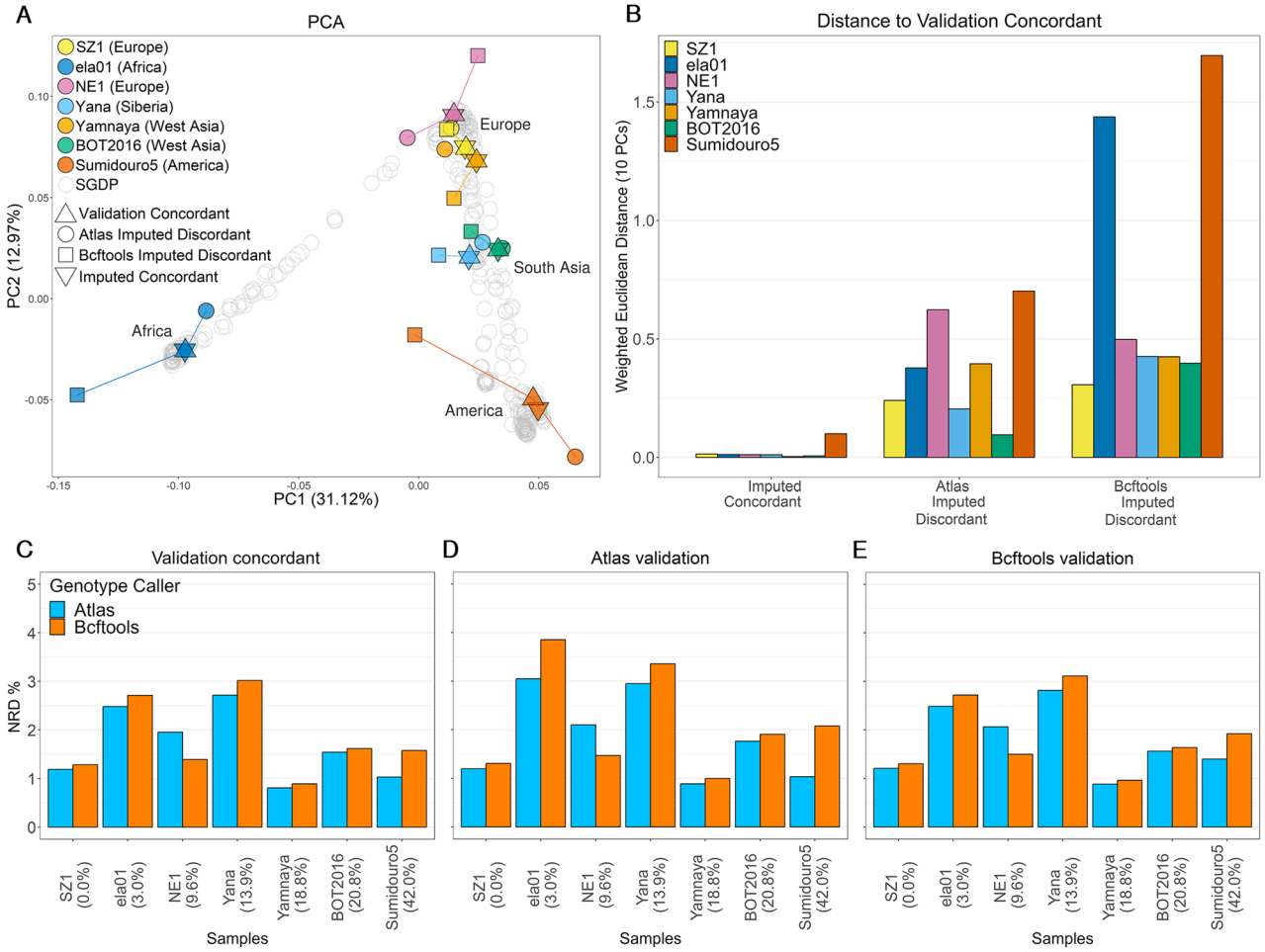

Figure 2. Effect of two different genotype callers on imputation accuracy. (**A**) Two first principal components of principal component analysis (PCA) of present-day genomes (SGDP). The genetic data of seven ancient individuals were projected onto it. For each individual, the triangle shape indicates the validation concordant dataset, the inverse triangle represents the genotypes that were equally imputed when using ATLAS and bcftools genotype likelihoods (GLs) ("Imputed concordant"), and the circles and square shapes correspond to the positions that were differently imputed when using ATLAS GLs ("ATLAS Imputed Discordant") and bcftools GLs ("Bcftools Imputed Discordant"), respectively. The continental origins of the individual samples are indicated on the legend. (**B**) Weighted Euclidean distances between the validation concordant and the three imputed datasets, for the seven ancient individuals. We calculated the Euclidean distances between the validation concordant set of each sample, and their concordant and discordant imputed genotypes. We measured the distances across the first 10 PCs, and subsequently weighted the distances by the eigenvalue of each PC. (**C**) Non-reference discordance (NRD) for the seven ancient individual samples when called with either ATLAS (blue) or bcftools (orange) prior to imputation. For each individual sample, the corresponding PMD values (first position at the 5' ends of the reads) is shown below the sample names. Samples are ordered by increasing PMD rates. NRD was assessed using three validation datasets: (**C**) validation concordant, (**D**) ATLAS validation and (**E**) bcftools. For NRD assessment on all variants, see Supplementary Figure S5F–H.

Transition substitutions proportions tend to increase in discordantly imputed variants

To determine whether the differences between the two imputed datasets arose from the fact that ATLAS takes into account PMD upon calculating GLs, we examined whether there was an overrepresentation of transitions, i.e., affected by PMD, among the discordantly imputed sites. The transition base change proportion in the imputation reference panel was 68.7% (Table S4). In comparison, the average transition proportion for concordantly imputed positions across all samples aligned with this expectation. At discordantly imputed positions, we found the proportion of transition sites to be similar to the overall transition rate for Yamnaya, NE1 and BOT2016. There was a clear increase in transition representation among discordantly imputed SNPs for Yana (71.1%) and Sumidouro5 (73.0%). Even though there was no clear correlation between PMD and transition numbers increase, the largest difference was again found for Sumidouro5. For this genome, there were 2.7 times more transitions than transversions among the discordantly imputed sites, compared to 2.2 times in all considered sites, further suggesting that ATLAS aids imputation in the presence of very high PMD rates. We observed similar trends of transition base change proportions when considering the intersection of the 1000 Genomes and 1240 K SNPs, and the 2.8 M SNPs used in PCA analyses (Tables S5 and S6).

Inherent difficulties in assessing imputation performance of African individual samples

While Sumidouro5 and ela01 have the largest discrepancies between their imputed datasets, the reasons underlying those differences are clearly distinct, given that ela01 has a much lower PMD rate (3%) and thus we would not expect imputation to considerably benefit from ATLAS GLs. Specifically, ela01's performance with ATLAS also stands out as shown by its Euclidean distance to the ground truth of 0.38, in contrast to 1.44 with bcftools (Fig. 2B), and reduced NRD values for ATLAS (2.48%) compared to bcftools (2.71%) with the validation concordant (Fig. 2C). However, the NRD calculated on all 1000 Genomes positions using the validation concordant showed a different trend. Here, bcftools's imputed GLs yielded a lower NRD of 4.06%, compared to 4.61% with ATLAS (Supplementary Figure S5F–H). The distinct characteristics of ela01 are further highlighted by its elevated discordance rate of 1.67% at the 2.8 million SNPs, nearly double that of most other samples (Table S7). Additionally, ela01 exhibits among the highest imputation error rates overall, irrespective of the genotype caller used (Table S2). Finally, ela01 stands out as having the lowest proportion of transition base changes in the discordant positions when considering all imputed SNPs (66.7% compared to 68.7% for all sites, Table S4). These findings suggest that evaluating imputation performance in African samples may be challenging due to the unique features and complexities of their genomic profiles and their ancestries being underrepresented in most reference panels.

Imputed simulated ancient genomes show that PMD has a negative but small impact on imputation performance regardless of the genotype likelihood calling strategy

Given the above-mentioned challenges in determining how PMD impacts imputation performance while using either ATLAS or bcftools GLs, we assessed imputation accuracy of imputed simulated ancient genomes with PMD varying between 0 and 50% (Supplementary Figure S9). We found that NRD increased with increasing PMD regardless of the genotype caller (Supplementary Figure S10 and Table S9). NRD increased from 4.97% at 0% PMD to 5.44% when using ATLAS GLs, while for bcftools NRD varied between 5.05% and 5.67%. For all PMD values, using ATLAS GLs led to smaller error rates, particularly for more damaged genomes, as expected (Supplementary Figure S10A). The difference in imputation accuracy between the most and the least damaged genomes was most pronounced at rare variants, that is, when MAF (minor allele frequency) was below 2% and 5% for ATLAS and bcftools GLs, respectively (Supplementary Figure S10B,C).

Assessing how contamination affects imputation performance

Methodology for assessing imputation's ability to rectify contamination in ancient genomes

To assess the potential of genotype imputation in mitigating contamination, we virtually contaminated the Loschbour genome, an 8000-year-old Western Hunter-Gatherer[50], with present-day human DNA from a European Greek and an African Dinka genomes from SGDP[43], as depicted in Fig. 1B. To ensure an accurate representation of the impact of contamination on a typical ancient sample, the Loschbour genome was initially downsampled to low coverage (1×). Then, based on the desired contamination levels ranging from 1 to 90%, we substituted a proportionate number of reads—drawn at random—from the ancient genome with reads from the modern genomes. We verified that the genome was contaminated as expected by estimating contamination on the X chromosome using contaminationX[51] (Supplementary Figure S11). Afterwards, we computed the genotype likelihoods on the two artificially-contaminated samples using bcftools, and imputed them with GLIMPSE. In order to assess to which degree imputation rectified contamination and allowed the retrieval of the high-coverage genotypes, we compared the contaminated samples with the original non-contaminated high-coverage Loschbour by calculating NRD and evaluating their relative positions on PCA whose PCs were calculated with present-day genomes. Finally, we verified whether contamination has a different signature at the haplotype level when compared to admixture. For that, we inferred local ancestry for imputed genomes that had been previously contaminated with Dinka DNA using three reference populations from the 1000 Genomes panel: CEU (Utah residents with Northern and Western European ancestry), LWK (Luhya in Webuye, Kenya) and YRI (Yoruba in Ibadan, Nigeria).

Imputation severely impacted by contamination

We projected the non-contaminated high-coverage Loschbour genome, the two present-day Greek and Dinka genomes, and the imputed 1× contaminated genomes onto the PCA built with the SGDP dataset to determine how far the latter were relative to the non-contaminated Lochbour genome. We observed that, as contamination increased, the imputed sample moved further away from the non-contaminated Loschbour (Fig. 3A). Particularly, in the case of Dinka-contaminated samples, it appears that, with an increasing contamination rate, the samples moved progressively closer to the genetic signature of the original Dinka individual. With increasing contamination by present-day Greek DNA, the samples similarly deviated further from the Loschbour original sample, although in a less linear trajectory. However, this trend became linear when restricting the analysis to only the Western Eurasian populations in SGDP (Fig. 3B), as well as at higher PCs (PC7 vs. PC8 in Supplementary Figure S12). The observed nonlinearity may be due to these dimensions not distinctly separating the Loschbour and Greek populations. In contrast, the Dinka contamination displays a linear trend as PC1 primarily distinguishes African versus non-African populations[52]. In addition, we verified the effect of contamination rate on the number of heterozygous sites. As expected, a maximum number of heterozygous sites was reached at around 50%-contamination rate, that is, 7.5% and 6.4% for the Dinka and the Greek contaminants, respectively, in comparison with 3.4% at 0% (imputed uncontaminated 1 × Loschbour genome) and 5.7% (Dinka) and 4.5% (Greek) at 100% contamination rates, respectively (Supplementary Figure S13).

Imputation of the uncontaminated Loschbour genome downsampled to 1 × resulted in an NRD of 2.7% (Fig. 3C). Between 0 and 5% contamination, i.e., the typical range of tolerated contamination in aDNA studies,

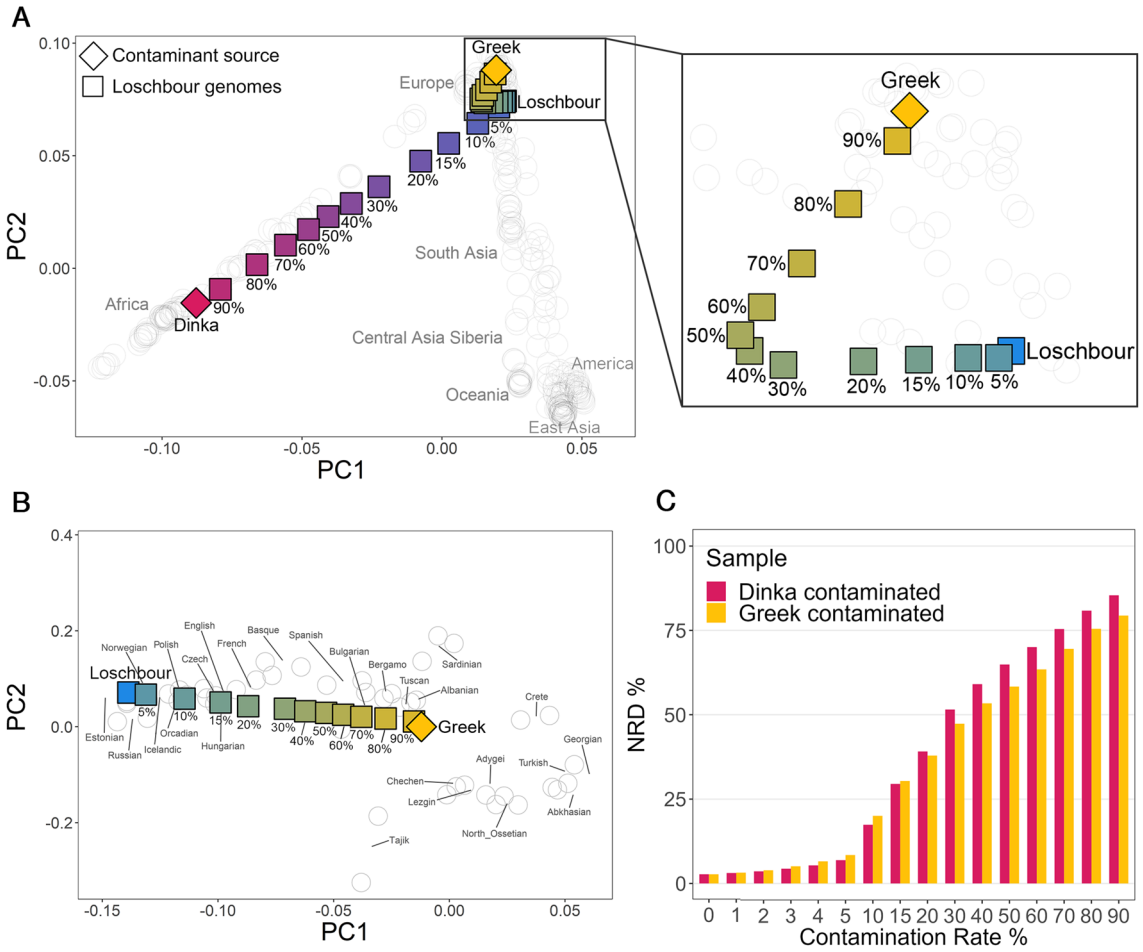

Figure 3. Effect of contamination on genotype imputation. Principal component analysis (PCA) of present-day genomes and projection of uncontaminated and imputed contaminated ancient genomes. The ancient downsampled (1×) Loschbour genome (light blue) was subjected to varying degrees of contamination, ranging from 1 to 90%, with DNA from a present-day Greek individual (yellow) and a present-day Dinka individual (red). (**A**) PCA of worldwide present-day populations (SGDP), represented by empty gray circles, and projection of all the imputed contaminated downsampled Loschbour genomes. (**B**) PCA of present-day Western Eurasians, with a focus on the Greek-DNA imputed contaminated Loschbour genomes. (**C**) Non-reference discordance (NRD) as a function of contamination levels, having the high-coverage Loschbour genome as ground truth. Yellow and red bars indicate Greek-contaminated and Dinka-contaminated Loschbour samples, respectively.

error rates remained relatively low, with NRD rising faster when the Greek genome was the contamination source: at 3% contamination, NRD was 5.1% and 4.4% when contaminating with Greek and Dinka DNA, respectively. However, for this range, contamination with Dinka DNA had a bigger impact on the PCA (Supplementary Figure S14). In the presence of 5% present-day Greek DNA contamination, NRD increased to 8.4%, which became 79.4% at 90% contamination. Similarly, samples contaminated with present-day Dinka DNA had an NRD of 6.9% at 5% and 85.4% at 90% contamination. The implications are that, as contamination levels rise, the reliability of imputed genotypes for downstream analyses deteriorates. Moreover, at higher contamination rates, contamination from present-day Dinka DNA seemed to have a more important detrimental impact on imputation accuracy relative to contamination from present-day Greek DNA. Specifically, beyond a contamination level of 20%, the NRD associated with Dinka contamination exceeded that of its Greek counterpart. Furthermore, squared Pearson correlation values (Supplementary Figure S15) highlighted that imputation of Greek-contaminated genomes outperformed Dinka-contaminated ones in imputation accuracy. This observation suggests that imputation performance was more negatively affected by higher levels of contamination coming from a more distantly related source. These results indicate that contamination substantially compromises genotype imputation accuracy, making it ineffective to rectify contamination in ancient genomes. We hypothesize that contaminant reads prevent the imputation model from copying from the most informative reference haplotypes, as contaminated genomes contain information from more than two haplotypes (four, in the case of a single contamination source).

Contamination leaves a distinct signal at the haplotype level

Admixture and contamination could leave similar genetic hallmarks, as implied by the findings that (i) similarity between imputed contaminated genomes and their contamination sources increased with contamination rate

and (ii) heterozygosity peaked at 50% contamination. When inferring local ancestry for imputed genomes contaminated with Dinka DNA, we found that haplotype tracts with an African origin were only inferred for contamination rates above 15%, while true admixed tracts can be found at much lower admixture levels (Supplementary Figure S16A, B). Furthermore, at 50% contamination rate, for instance, the tract length distribution was different from the expected distribution for 50% admixture levels, as exemplified with local ancestry inferred in an African American individual (Supplementary Figure S16C). In contrast to true admixture, we found an enrichment for short tracts across the genome (Supplementary Figure S16D, E), that can be attributed in part to imputation errors (Supplementary Figure S17). This excess in short tracts reflects that contamination affects all sites in an uniform fashion and not according to a recombination map as in the case of admixture.

Discussion

Genotype imputation has been increasingly employed in aDNA studies, unlocking downstream applications that were typically limited to high-coverage genomes, thus allowing to tackle even more questions. It is therefore essential to thoroughly understand its limitations and potential. Our study provides insights into how PMD, a common feature of ancient genomes, and contamination affect genotype imputation of ancient genomes. While we restricted our analyses to shotgun-sequenced ancient genomes, we expect that all conclusions we arrived at are equally valid for in-solution capture sequenced ancient genomes. Firstly, we found that taking PMD into account prior to imputation yielded unequivocally more accurate genotypes in the presence of very high PMD rates. Although imputation of ATLAS GLs led to overall lower imputation error rates, these constituted modest gains for most of the remaining genomes regardless of their PMD rates. This result suggests that imputation is robust to some level of damage. In fact, we found that PMD has a relatively small negative impact on imputation. As such, ATLAS's high computational needs and extended run times may not justify its benefits when ancient genomes are not highly damaged. We also found that trimming five base pairs at the reads' ends before using bcftools to generate GLs, which is common practice in aDNA studies, can be a more computationally efficient alternative to ATLAS to minimize the effect of PMD on imputation performance. Secondly, we showed that genomes with more than 5% contamination are inaccurately imputed. This is also the current threshold used to discard contaminated samples in aDNA studies. Moreover, imputation accuracy decreased with the increase of contamination levels in such a way that the imputed contaminated genomes were increasingly closer to contamination sources. Nonetheless, the contamination signature was clearly distinct from admixture at the haplotype level. This observation suggests that it is possible to detect and estimate nuclear DNA contamination using a framework built upon imputation and by exploiting disruption of linkage disequilibrium by contamination as in Nakatsuka et al[19]. Another potential future methodological development could be aimed at correcting PMD in ancient genome imputation, since we have shown before that imputation can correct deaminated sites. A full description of how and to what extent this correction works is however needed to that end. It also remains to be assessed how imputation varies across the genome, which can have implications for studies focused on particular genome regions, such as selection studies. Finally, a limitation of imputation not restricted to ancient genomes is the fact that present-day imputation reference panels do not contain private variation from under/non-represented populations. Such limitations can be mitigated by using more diverse reference panels and/or by assembling reference panels that include ancient genomes so as to recover variation that has been lost over time. However, high-coverage ancient genomes are the exception and their sample size is probably too small to improve imputation performance at private variants.

Methods

Datasets

Ancient genomes

We collected eight high-coverage ancient genomes from publicly available studies and reported them in Table S1. This table provides additional details about the origin of each sample, their coverage, and the average rate of PMD at the 5' read ends. These samples offer a global representation and display varying damage rates. Certain samples, as indicated in Table S1, underwent a UDG-treatment to reduce PMD. SZ1 underwent partial UDG-treatment[49]. Loschbour underwent partial UDG-treatment for three out of the four built libraries. The remaining library was not UDG-treated[50]. The PMD values for the remaining samples were either stated as reported in the original study or assessed by us. We selected seven ancient genomes for the first part of our experiment focusing on assessing how accounting for PMD impacts imputation accuracy, while the remaining genome, Loschbour, was used in the second part investigating the impact of contamination on imputation.

Simons Genome Diversity Project

We used the Simons Genome Diversity Project (SGDP)[43] dataset as a reference panel to perform PCA. This dataset includes genomic information from 300 individuals that come from 142 populations worldwide. In addition, we sampled reads from two SGDP genomes (a Dinka (B_Dinka-3) and a Greek individual (S_Greek-1)) to contaminate the Loschbour ancient genome.

1000 Genomes (imputation reference panel)

We used a version of the 1000 Genomes v5 phase 3 panel[40] containing 2504 genomes resequenced at 30x, that was phased using TOPMed[53], excluding singletons and only keeping biallelic sites. Furthermore, we lifted over the reference panel from b38 to hg19 using Picard liftoverVCF (https://gatk.broadinstitute.org/hc/en-us/articles/360037060932-LiftoverVcf-Picard-), with hg38ToHg19 chain from the University of California Santa Cruz liftOver tool (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/liftOver/).

PMD assessment

We used the bamdamage tool from bammds[54] to estimate PMD as the rate of C-to-T misincorporations at the 5' end first position. This tool enabled us to evaluate the PMD rate for 6 of the 7 samples. For the remaining sample (SZ1), we relied on the damage report from the published study (Table S1).

Downsampling

We downsampled all individual samples used in the study to 1× coverage using samtools[36] v1.10.

Genotype calling

We used bcftools[36] v1.10 and ATLAS[37] v0.9.9 to generate genotypes calls and genotype likelihoods at the 1000 Genomes bi-allelic SNPs, using the human reference genome humanG1Kv37 (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/). The genotype likelihoods (PL field) we generated from the downsampled genomes constitute the input data for imputation. We used genotype calls (GT field) from the high-coverage genomes to produce the validation datasets.

Bcftools

To generate genotype likelihoods from the downsampled genomes via bcftools, we used the *bcftools mpileup* command with parameters *-I*, *-E*, and *-a 'FORMAT/DP'* with the *–ignore-RG* flag enabled, followed by the *bcftools call* command with the *-Aim* and *-C* alleles parameters.

Before calling genotypes from the non-UDG treated high-coverage genomes using bcftools, we trimmed 10 base pairs from each end of the reads using BamUtil[55] v1.0.14 to reduce the impact of PMD. Then, we performed five additional quality control parameters:

1) We excluded positions presenting a minimum mapping quality of 30 (*-q 30*) and a minimum base quality of 20 (*-Q 20*), as well as reduced mapping quality for reads featuring an abundance of mismatches with the parameter *-C 50*.
2) We removed sites not included in the 1000 Genomes accessible genome strict mask[56].
3) We filtered out sites located in regions known to contain repeat elements[57].
4) We eliminated positions exhibiting extreme values of coverage depth. For the lower threshold, positions with depth less than one-third of the mean or less than eight (if the one-third mean coverage fell below eight) were filtered out. For the upper threshold, positions exceeding twice the mean depth of coverage were also excluded.
5) Finally, positions with a QUAL value below 30 ($QUAL < 30$) were also omitted.

ATLAS

Tailored for aDNA, ATLAS infers PMD patterns and uses that information upon calling variants allowing to choose between different variant callers. We opted to use the Maximum-Likelihood (MLE) caller, which works similarly to GATK[58], with the difference that the used genotype model accounts for PMD[59]. To call the genotypes with ATLAS, we followed the standard pipeline for the analysis of individual samples as described in https://bitbucket.org/wegmannlab/atlas/wiki/Home. Firstly, we split single-end reads groups according to length and merge paired-end reads if present (splitmerge step). Afterwards, we used BAMdiagnostics in order to determine the maximum read length to give as input to *splitmerge*, estimated the empirical post-mortem damage with *PMD*, and used the *MLE* caller to generate genotype likelihoods. We applied this process to evaluate genotype likelihoods for both downsampled and high-coverage genomes.

Due to ATLAS not listing positions that are not covered by any read, we manually added those missing positions. This was a necessary step, because GLIMPSE v1.1.1 does not impute unlisted positions even if they are present in the imputation reference panel.

Genotype imputation

To impute the downsampled genomes, we used GLIMPSE v1.1.1[28] as follows:

1. We used *GLIMPSE_chunk* to segment chromosomes into chunks ranging between 1–2 Mb in size, with an additional buffer region of 200 kb flanking each chunk.
2. We performed imputation on the chunks using *GLIMPSE_phase* with the default parameters and using the 1000 Genomes panel as a reference panel.
3. Finally, we used *GLIMPSE_ligate* to ligate the imputed chunks into complete chromosomes.

Imputation performance evaluation

In order to evaluate the imputation accuracy in both experiments, we used *GLIMPSE_concordance* from GLIMPSE v2[60]). The evaluation process was restricted to sites that had a minimum of eight reads and a genotype posterior probability of at least 0.9999. Using *GLIMPSE_concordance*, we calculated imputation errors and non-reference-discordance (NRD). NRD is defined as $NRD = (eRR + eRA + eAA)/(eRR + eRA + eAA + mRA + mAA)$ where e and m correspond to the number of errors and matches, respectively, while R and A stand for reference and alternative alleles, respectively. This approach is designed to exclude the number of correctly imputed homozygous reference allele sites (mRR), which are the majority, therefore emphasizing the significance and giving more weight to imputation errors at alternative allele sites. To validate our results, we used the concordant

dataset as validation in the first experiment, and in the second experiment, we used as validation the genotypes called with bcftools from the high-coverage non-contaminated Loschbour genome.

Principal component analysis (PCA)

Data preparation

We generated a set of 2.8 M SNPs that we used for PCA. These SNPs are a subset of the bi-allelic SNPs from the 1000 Genomes panel. We obtained this set of sites by first removing rare variants (MAF < 5%). Then, we retained only sites that are included in the 1000 Genomes accessible genome strict mask, and lastly, we filtered out SNPs known to be in repeat regions. This filtering resulted in a set of 2,814,418 SNPs, which we refer to as the 2.8 M SNP set.

Afterwards, we identified which positions differed based on whether genotypes were computed with bcftools or ATLAS. To do that, we intersected the imputed SNP sets from both tools so we could compare genotypes at each position, and check for agreement between the tools. Specifically, from both vcf files, we extracted the GT (genotype) columns using *bcftools query -f '[%GT]\n'*, we compared each position pairwise, and we determined if the imputed genotypes agreed or differed between the two sets with a python script. As for the validation samples, we similarly built validation datasets for each individual sample based on the positions concordantly identified between the two sets of genotypes called with bcftools and ATLAS on the high-coverage genomes. All SNP counts of concordant and discordant positions for both validation and imputed datasets are reported in Tables S7 and S8.

Performing PCA

For both PCA analyses, sample files in vcf format were first converted into the PLINK[61] format in order to subsequently convert into the required EIGENSOFT[41,42] format for smartPCA[41,42]. Principal components were calculated using the SGDP reference panel and all samples were projected onto the resulting components with the *lsqproject* parameter enabled. In the first experiment, we restricted the PCA to the 2.8 M SNP set. The projected samples included the validation concordant samples, which served as the ground truth, as well as the sets of positions differently and concordantly imputed when genotype likelihoods were computed with either bcftools and ATLAS. In the second experiment, we restricted the PCA to the 1240 K SNP set[32–34] and projected the Loschbour ancient sample, the two present-day Dinka and Greek samples, as well as the contaminated and then imputed genomes. We projected the samples onto the SGDP reference panel, but also on a subset restricted to only the Western Eurasian populations of SGDP.

Euclidean distance

To quantify the differences between the SNP sets, we calculated an Euclidean distance between them and the validation concordant of the respective sample. This distance was calculated across the first 10 PCs, and subsequently weighted by factoring the eigenvalue of each PC.

Simulation of ancient genomes

We simulated ancient genomes with varying levels of PMD using gargammel[38]. We used as input two fasta files containing information of SNPs from a present-day Turkish genome with identifier 06A010111 of the Turkish Genome Project dataset[24,39]. These fasta files were previously used in the study of Koptekin et al[24]. We specified the composition of the genomes with the parameter *-comp 0,0,1* to only include endogenous aDNA and ignore bacterial and modern human contamination. For the levels of damage, we specified the parameter *-damage 0.024, 0.36, 0.0097,X*, with X being the probability of cytosine deamination. We varied this value in order to have the desired PMD level as specified in Table S10. All tested genomes were simulated at 1 × coverage (*-c 1*), while the validation was generated at 30 × coverage (*-c 30*).

To map the simulated fastq paired-reads file from gargammel, we first removed the adapters and merged the paired reads with AdapterRemoval[62] with the parameters *–trimns –trimqualities –minlength 30 –collapse*. We then mapped the reads with BWA[63] using the command *bwa aln -l 1024*.

Contamination

In order to assess the impact of contamination on imputation, we virtually contaminated an ancient European genome known as the Loschbour man[50]. Modern human DNA samples (Table S1), one from an African Dinka individual and another from a Greek European individual, served as the sources of contaminant DNA. We first started by downsampling the Loschbour genome to 1 × coverage with samtools. Then, we computed and removed the corresponding proportion of reads from the ancient genome that would equate to the desired contamination level. As an illustration, for a contamination level of 10% with modern DNA, we initially removed 10% of the reads present in the ancient genome, then, we added an equivalent quantity of reads sourced from either the Dinka or Greek contaminant DNA.

Imputation of contaminated genomes

We imputed the ancient genomes with varying amounts of contamination following the above-mentioned methodology.

Contamination assessment

We assessed the level of contamination using contaminationX[51]. This tool, effective for low-coverage male samples, is based on the unique property of the X-chromosome in males (hemizygous), where typically

only one DNA sequence type is seen per cell. Any deviation from this, such as observing multiple alleles at a given site, indicates possible contamination, post-mortem damage, sequencing or mapping errors. We first estimated the allele counts using ANGSD[64] with the following parameters: *-b 5,000,000 -c 154,900,000 -k 1 -m 0.05 -d 3 -e 20*, and specified the correct reference allele frequency panel for the sample population with the parameter *-h HapMap_pop*. Then, we estimated the level of contamination with contaminationX, using the following parameters: *maxsites = 1000 nthr = 4* and *oneCns = 1*. We only estimated contamination up to 40% contamination rate, as beyond this, contaminationX tends to underestimate the actual contamination level as previously described[51].

Effect of contamination at the haplotype level

Before inferring local ancestry, we phased the genomes with SHAPEIT5[65]. We used the command *phase_common* while keeping rare variants (*–filter-maf 0*) and made use of the version of 1000 Genomes reference panel that we used before and the HapMap recombination genetic maps.

We inferred local ancestry with RFMix[66] v2.03-r0 and used subsets of the 1000 Genomes panel as a reference. For the imputed Dinka-contaminated 1 × Loschbour genomes, we had three reference populations (YRI, LWK and CEU), whereas we restricted to two reference populations (YRI and CEU) when inferring local ancestry on two African-American (labeled as ASW) genomes, NA19703 and NA20278, from the 1000 Genomes panel.

Text quality enhancing through the use of specialized AI tools

We used AI tools, specifically ChatGPT, to improve the quality of the text in the introduction section of our article.

Data availability

The eight ancient individual samples analyzed in this study are publicly available and were first described in the following studies: ela01[48] (https://doi.org/10.1126/science.aao6266); BOT2016 and Yamnaya[45] (https://doi.org/10.1126/science.aar7711); NE1[44] (https://doi.org/10.1038/ncomms6257); SZ1[49] (https://doi.org/10.1038/s41467-018-06024-4); Sumidouro5[46] (https://doi.org/10.1126/science.aav2621); Yana[47] (https://doi.org/10.1038/s41586-019-1279-z); and Loschbour[50] (https://doi.org/10.1038/nature13673). The SGDP reference panel[43] dataset in plink format can be found in https://sharehost.hms.harvard.edu/genetics/reich_lab/sgdp/variant_set/. We downloaded from Seven Bridges Cancer Genomics Cloud the bam files aligned to hg19 reference genome for two SGDP genomes, Dinka-3 and Greek-2. The 1000 Genomes v5 phase 3 panel resequenced to 30X coverage is available at the European Nucleotide Archive under accession code ERP114329 (https://www.ebi.ac.uk/ena/browser/view/PRJEB31736). The fasta files we used to simulate ancient genomes come from the study of Koptkin et al.[24] (https://doi.org/10.1101/2023.11.11.566695).

Code availability

The scripts used to generate the data and results presented in our study are provided in this github repository. https://github.com/TozeMarques/PMD_Contamination_Impact_on_aDNA_Imputation

Received: 19 December 2023; Accepted: 8 March 2024
Published online: 14 March 2024

References

1. Marciniak, S. & Perry, G. H. Harnessing ancient genomes to study the history of human adaptation. *Nat. Rev. Genet.* **18**, 659–674 (2017).
2. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
3. Reich, D. *et al.* Genetic history of an archaic hominin group from Denisova Cave in Siberia. *Nature* **468**, 1053–1060 (2010).
4. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat. Rev. Genet.* **20**, 323–340 (2019).
5. van der Valk, T. *et al.* Million-year-old DNA sheds light on the genomic history of mammoths. *Nature* **591**, 265–269 (2021).
6. Briggs, A. W. *et al.* Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. USA* **104**, 14616–14621 (2007).
7. Parks, M. & Lambert, D. Impacts of low coverage depths and post-mortem DNA damage on variant calling: A simulation study. *BMC Genomics* **16**, 19 (2015).
8. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
9. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20130624 (2015).
10. Lindahl, T., Ljungquist, S., Siegert, W., Nyberg, B. & Sperens, B. DNA N-glycosidases: Properties of uracil-DNA glycosidase from Escherichia coli. *J. Biol. Chem.* **252**, 3286–3294 (1977).
11. Boessenkool, S. *et al.* Combining bleach and mild predigestion improves ancient DNA recovery from bones. *Mol. Ecol. Resour.* **17**, 742–751 (2017).
12. Fulton, T. L. & Shapiro, B. Setting up an ancient DNA laboratory. In *Ancient DNA: Methods and Protocols* (eds. Shapiro, B. et al.) 1–13 (Springer, 2019).
13. Orlando, L. *et al.* Ancient DNA analysis. *Nat. Rev. Methods Primers* **1**, (2021).
14. Llamas, B. *et al.* From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR: Sci. Technol. Archaeol. Res.* **3**, 1–14 (2017).
15. Sampietro, M. L. *et al.* Tracking down human contamination in ancient human teeth. *Mol. Biol. Evol.* **23**, 1801–1807 (2006).
16. Peyrégne, S. & Prüfer, K. Present-day DNA contamination in ancient DNA datasets. *Bioessays* **42**, e2000081 (2020).
17. Der Sarkissian, C. *et al.* Shotgun microbial profiling of fossil remains. *Mol. Ecol.* **23**, 1780–1798 (2014).
18. Allentoft, M. E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).

19. Nakatsuka, N. *et al.* ContamLD: Estimation of ancient nuclear DNA contamination using breakdown of linkage disequilibrium. *Genome Biol.* **21**, 199 (2020).
20. Pääbo, S. Ancient DNA: Extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. USA* **86**, 1939–1943 (1989).
21. Ginolhac, A., Rasmussen, M., Gilbert, M. T. P., Willerslev, E. & Orlando, L. mapDamage: Testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27**, 2153–2155 (2011).
22. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, e1008302 (2019).
23. Martiniano, R., Garrison, E., Jones, E. R., 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.* **21**, 250 (2020).
24. Koptekin, D. *et al.* Pre-processing of paleogenomes: Mitigating reference bias and postmortem damage in ancient genome data. *bioRxiv* (2023) https://doi.org/10.1101/2023.11.11.566695.
25. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10**, 387–406 (2009).
26. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* **5**, e1000529 (2009).
27. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
28. Rubinacci, S., Ribeiro, D. M., Hofmeister, R. J. & Delaneau, O. Efficient phasing and imputation of low-coverage sequencing data using large reference panels. *Nat. Genet.* **53**, 120–126 (2021).
29. Hui, R., D'Atanasio, E., Cassidy, L. M., Scheib, C. L. & Kivisild, T. Evaluating genotype imputation pipeline for ultra-low coverage ancient genomes. *Sci. Rep.* **10**, 18542 (2020).
30. Ausmees, K., Sanchez-Quinto, F., Jakobsson, M., & Nettelblad, C. An empirical evaluation of genotype imputation of ancient DNA. *G3* **12**, (2022).
31. Sousa da Mota, B. *et al.* Imputation of ancient human genomes. *Nat. Commun.* **14**, 3660 (2023).
32. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
33. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
34. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
35. Collins, D. W. & Jukes, T. H. Rates of transition and transversion in coding sequences since the human-rodent divergence. *Genomics* **20**, 386–396 (1994).
36. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
37. Link, V. *et al. ATLAS: Analysis tools for low-depth and ancient samples*. 105346 https://www.biorxiv.org/content/https://doi.org/10.1101/105346v2 (2017) doi:https://doi.org/10.1101/105346.
38. Renaud, G., Hanghøj, K., Willerslev, E. & Orlando, L. gargammel: A sequence simulator for ancient DNA. *Bioinformatics* **33**, 577–579 (2017).
39. Alkan, C. *et al.* Whole genome sequencing of Turkish genomes reveals functional private alleles and impact of genetic interactions with Europe, Asia and Africa. *BMC Genomics* **15**, 963 (2014).
40. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
41. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
42. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
43. Mallick, S. *et al.* The Simons genome diversity project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
44. Gamba, C. *et al.* Genome flux and stasis in a five millennium transect of European prehistory. *Nat. Commun.* **5**, 5257 (2014).
45. de Barros Damgaard, P. *et al.* The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, (2018).
46. Moreno-Mayar, J. V. *et al.* Early human dispersals within the Americas. *Science* **362**, eaav2621 (2018).
47. Sikora, M. *et al.* The population history of northeastern Siberia since the Pleistocene. *Nature* **570**, 182–188 (2019).
48. Schlebusch, C. M. *et al.* Southern African ancient genomes estimate modern human divergence to 350,000 to 260,000 years ago. *Science* **358**, 652–655 (2017).
49. Amorim, C. E. G. *et al.* Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat. Commun.* **9**, 3547 (2018).
50. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
51. Moreno-Mayar, J. V. *et al.* A likelihood method for estimating present-day human contamination in ancient male samples using low-depth X-chromosome data. *Bioinformatics* **36**, 828–841 (2020).
52. McVean, G. A genealogical interpretation of principal components analysis. *PLoS Genet.* **5**, e1000686 (2009).
53. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
54. Malaspinas, A.-S. *et al.* bammds: A tool for assessing the ancestry of low-depth whole-genome data using multidimensional scaling (MDS). *Bioinformatics* **30**, 2962–2964 (2014).
55. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).
56. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
57. Karolchik, D. *et al.* The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, D493-496 (2004).
58. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
59. Hofmanová, Z. *et al.* Early farmers from across Europe directly descended from Neolithic Aegeans. *Proc. Natl. Acad. Sci. USA* **113**, 6886–6891 (2016).
60. Rubinacci, S., Hofmeister, R., da Mota, B. S. & Delaneau, O. *Imputation of low-coverage sequencing data from 150,119 UK Biobank genomes* (2022) doi:https://doi.org/10.1101/2022.11.28.518213.
61. Purcell, S. *et al.* PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
62. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
63. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
64. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of next generation sequencing data. *BMC Bioinform.* **15**, 356 (2014).
65. Hofmeister, R. J., Ribeiro, D. M., Rubinacci, S. & Delaneau, O. Accurate rare variant phasing of whole-genome and whole-exome sequencing data in the UK Biobank. *Nat. Genet.* **55**, 1243–1249 (2023).
66. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference. *Am. J. Hum. Genet.* **93**, 278–288 (2013).

Acknowledgements
We want to thank Dilek Koptekin, Théo Cavinato and Samuel Neuenschwander for fruitful discussions. A.G.M. and B.S.d.M. were supported by a Swiss National Science Foundation (SNSF) project grant (PP00P3_176977) to O.D., and by an SFNS grant (PCEGP3_181251) and a European Research Council grant (CAMERA 679330) to A.-S.M. S.R. was funded by an SNSF project grant (PP00P3_176977) to O.D. and by an SNSF Postdoc.Mobility project grant (P500PB_211106).

Author contributions
B.S.d.M., S.R. and O.D. conceptualized and designed the study. A.G.M. and B.S.d.M. conducted the experiments. A.G.M. and B.S.d.M. wrote the manuscript with input from O.D, A.-S.M and S.R. This work has been supervised by B.S.d.M. All authors interpreted the results and reviewed the final manuscript.

Competing interests
Olivier Delaneau is a current employee of Regeneron Genetics Center which is a subsidiary of Regeneron Pharmaceuticals. The remaining authors declare no conflicts of interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-56584-3.

Correspondence and requests for materials should be addressed to B.S.d.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.