Benchmarking software tools for trimming adapters and merging next-generation sequencing data for ancient DNA

Annette Lien[1], Leonardo Pestana Legori[2], Louis Kraft[1], Peter Wad Sackett[1] and Gabriel Renaud[1]*

[1]Department of Health Technology, Section for Bioinformatics, Technical University of Denmark, Kongens Lyngby, Denmark, [2]University of Debrecen, Debrecen, Hungary
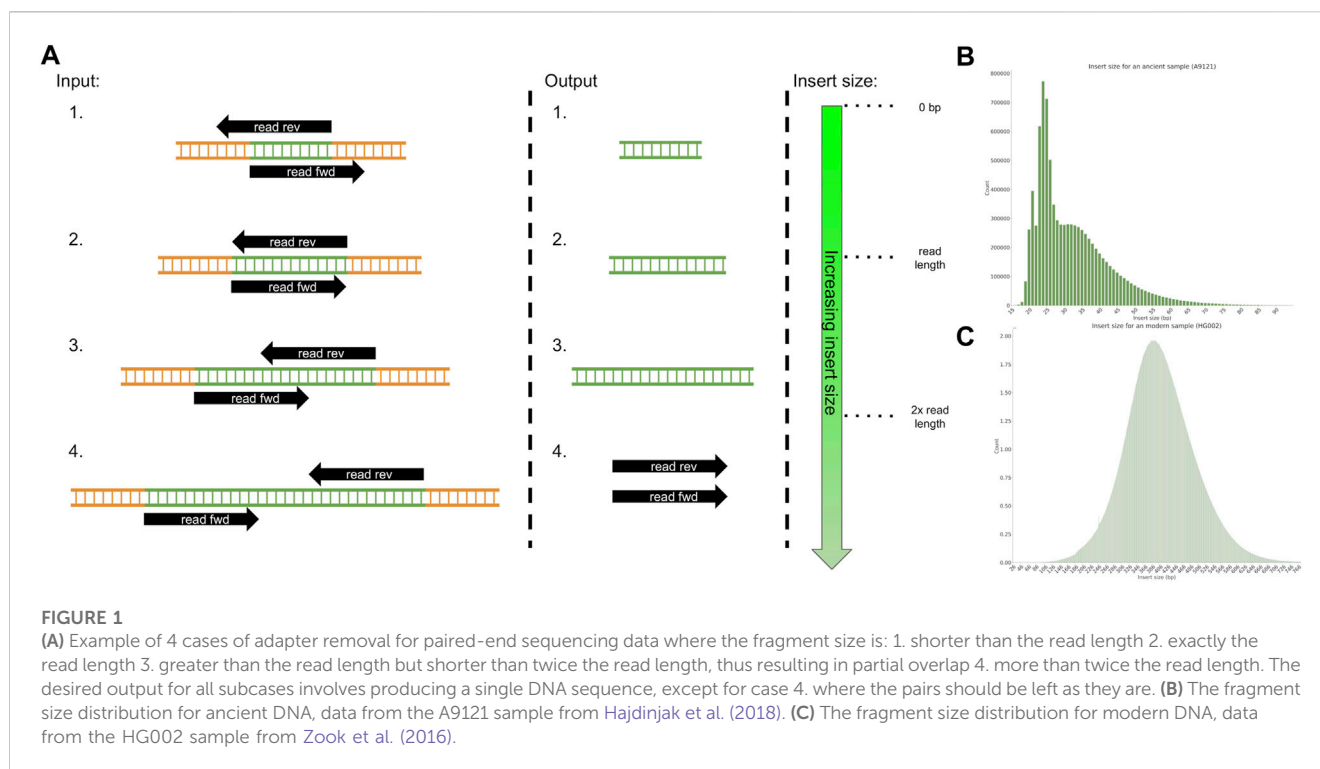
Ancient DNA is highly degraded, resulting in very short sequences. Reads generated with modern high-throughput sequencing machines are generally longer than ancient DNA molecules, therefore the reads often contain some portion of the sequencing adaptors. It is crucial to remove those adaptors, as they can interfere with downstream analysis. Furthermore, overlapping portions when DNA has been read forward and backward (paired-end) can be merged to correct sequencing errors and improve read quality. Several tools have been developed for adapter trimming and read merging, however, no one has attempted to evaluate their accuracy and evaluate their potential impact on downstream analyses. Through the simulation of sequencing data, seven commonly used tools were analyzed in their ability to reconstruct ancient DNA sequences through read merging. The analyzed tools exhibit notable differences in their abilities to correct sequence errors and identify the correct read overlap, but the most substantial difference is observed in their ability to calculate quality scores for merged bases. Selecting the most appropriate tool for a given project depends on several factors, although some tools such as fastp have some shortcomings, whereas others like leeHom outperform the other tools in most aspects. While the choice of tool did not result in a measurable difference when analyzing population genetics using principal component analysis, it is important to note that downstream analyses that are sensitive to wrongly merged reads or that rely on quality scores can be significantly impacted by the choice of tool.

KEYWORDS

next-generation sequencing, ancient DNA, read merging, adapter trimming, read processing, benchmarking

1 Introduction

Next-generation sequencing (NGS) has ushered in a new era of genomics by enabling researchers to sequence DNA at an unprecedented rate, leading to a significant increase in the number of available genome references. Illumina is currently the NGS platform with the highest market share. Library preparation for Illumina sequencing involves three main steps: DNA fragmentation, the addition of adaptors to bind to the flowcell (i.e., the solid medium to allow sequencing), and the amplification of DNA templates for sequencing. DNA templates are then sequenced through the synthesis of complementary DNA strands and optical base calling. Reversible terminator nucleotides are used, with one nucleotide incorporated at a time while

FIGURE 1
(A) Example of 4 cases of adapter removal for paired-end sequencing data where the fragment size is: 1. shorter than the read length 2. exactly the read length 3. greater than the read length but shorter than twice the read length, thus resulting in partial overlap 4. more than twice the read length. The desired output for all subcases involves producing a single DNA sequence, except for case 4. where the pairs should be left as they are. **(B)** The fragment size distribution for ancient DNA, data from the A9121 sample from Hajdinjak et al. (2018). **(C)** The fragment size distribution for modern DNA, data from the HG002 sample from Zook et al. (2016).

capturing its fluorescent signal through high-precision optical imaging. This process is repeated for a specific number of cycles, resulting in read lengths typically ranging from 75 to 250 base pairs (bp). An additional round of sequencing can be performed on the same molecules where the reads are reverse complemented, bound to the flowcell and the nucleotides are read from the other end and the other strand. Reads produced that way are called paired-end reads, the first read is called the *forward* read whereas the second is the *reverse* read. The basecaller is the software used to call nucleotides from raw intensity values. However, due to certain factors, nucleotides can be misread and the basecaller therefore also provides quality scores for each base call, indicating the probability of a sequencing error for that base Liu et al. (2012); Mardis (2008).

One instance where the fragmentation step is not required is working with DNA extracted from fossils, sediments, or museum samples, otherwise called ancient DNA (aDNA). This DNA is naturally degraded and therefore extracted fragments can be on average very short (< 50 bp). While other idiosyncrasies of aDNA including contamination and chemical damage create specific challenges for downstream analyses, degradation of the DNA molecule needs to be addressed in the earliest stages of the bioinformatics processing Orlando et al. (2021).

While removing lingering sequencing adapters from sequencing reads is important for modern DNA as it can lead to mismappings or misassemblies, it is a crucial step for aDNA. The short fragment size entails that the vast majority of templates will be shorter than the read length and therefore have lingering sequencing adapters, often for the majority of the read (see Figure 1). Furthermore, if the reads are paired-end, after performing adapter trimming, both the remaining forward and reverse reads should be the same length and the reverse complement of each other as it is the same molecule. An additional bioinformatics step is to merge both the remaining

forward and reverse reads into a single sequence, as this can correct potential sequencing errors and lead to more accurate data. If paired-end reads are produced and if the length of the template is larger than the read length but still shorter than twice the read length, then both reads should partially overlap (see Figure 1). Similarly, both reads can be merged into a single one, where corrections can be applied to the overlapping portion to mitigate sequencing errors. For aDNA, this bioinformatics step can be construed as reconstructing the original aDNA fragment from either single-end or paired-end reads.

Several tools have been developed to reconstruct aDNA fragments from sequencing data. These include AdapterRemoval Schubert et al. (2016), bbmerge Bushnell et al. (2017), ClipAndMerge Peltzer et al. (2016), fastp Chen et al. (2018), leeHom Renaud et al. (2014), SeqPrep John (2016), and adna-trim Li (2018). The aforementioned programs have been used in several studies. They use different algorithms, handle mismatches between both reads differently, and produce different quality scores. In aDNA studies, often previously published samples from different groups are merged together to answer biologically relevant questions. It is unknown however if different adaptor trimming and read merging strategies lead to batch effects.

In this study, we benchmark different tools which have been designed and used in aDNA studies to reconstruct aDNA from sequencing data. More specifically, using simulated NGS data, we evaluated how these tools perform at various simulated lengths of the template, how robust they are to sequencing errors, and how they compute consensus bases in the presence of matches/mismatches between 2 nucleotides. Finally, we sought to evaluate if using different trimming programs had any impact on the sample placement within a principal component analysis (PCA), which is a very common analysis in aDNA studies.

Tool	References	Version	Description	Non-standard parameters	Examples of aDNA studies
AdapterRemoval	Schubert et al. (2016)	v2.3.2	Modified Needleman-Wunsch algorithm for finding the best alignment between paired-end reads. Quality scores serve as position specific scoring matrix. The most likely sequence is chosen	–collapse	8; 25; 6
				–minlength 1	
				–minalignmentlength 10	
				–qualitymax 93	
bbmerge	Bushnell et al. (2017)	v38.91	Adapter trimming and read merging by overlap detection using multiple heuristics controlled by several constants	mininsert = 1	40; 41; 14
				mininsert0 = 1	
				minoverlap = 10	
				minoverlap0 = 1	
ClipAndMerge	Peltzer et al. (2016)	v1.7.8	Smith–Waterman algorithm for read pair alignment. Using error rate cut-offs for detecting adaptors	-l 1 -p 10	16; 37; 17
fastp	Chen et al. (2018)	v0.21.1	Adapter trimming and read merging by finding the overlap of each read pair using cut-offs. Bases that fall out of the overlapped regions are considered adapters	–merge	39; 30; 43
				–disable_length_filtering	
				–overlap_len_require 10	
leeHom	Renaud et al. (2014)	v1.2.17	Adapter trimming and read merging using Bayesian maximum *a posteriori* probability	–ancientdna	4; 42; 36
SeqPrep	John (2016)	v1.3.2	If adapters are detected using cut-off values, the two reads are forcefully merged, otherwise quality score aware cut-offs are used for merging reads	-L 1	13; 28; 10
				-o 10	
adna-trim	Li (2018)	v1.3-r106	Adapter trimming if length threshold is fulfilled. Base with higher quality score is chosen or difference of quality scores in case of mismatch	-l 1	22; 35; 38
				-o 10	

2 Materials and methods

2.1 Generating simulated data

For a detailed description of the generation of simulated data, see Supporting Material. In brief, we generated simulated aDNA data using the telomere-to-telomere assembly of a human genome T2T-CHM13 and raw sequencing reads for two individuals from the National Institute of Standards and Technology's Genome in a Bottle (GIAB) project Nurk et al. (2022); Zook et al. (2016). We simulated DNA fragments and paired-end reads using gargammel Renaud et al. (2017) and ART Huang et al. (2012), with varying sequence length distributions, error rates, and Phred quality scores.

2.2 Trimming and merging

We compared seven different tools for adapter clipping and paired-end read merging: AdapterRemoval Schubert et al. (2016), bbmerge Bushnell et al. (2017), ClipAndMerge Peltzer et al. (2016), fastp Chen et al. (2018), leeHom Renaud et al. (2014), SeqPrep (John, 2016), and seqtk/adna-trim (Li, 2018; Li, 2019). Only leeHom and seqtk/adna-trim were specifically designed for aDNA analysis, although the merging algorithm of AdapterRemoval is also specifically claimed to be suited for aDNA processing. To enable a representative comparison, we standardized some parameters across the tools. Specifically, the minimum number of overlapping bases was set to 10 and length

filtering was disabled across all tools except leeHom, which instead uses a specific parameter for aDNA processing. Detailed descriptions of the parameters used for each tool and results with default minimum overlap and length filtering settings can be found in the Supporting Material.

2.3 Evaluating base quality score accuracies

We evaluated the accuracies of the merged per-base Phred quality scores by comparing the Phred values to the actual error rates. To calculate the confidence intervals for the observed error rates, we used the Clopper-Pearson method Clopper and Pearson (1934). We then used the inverted ranges of the Phred-scaled confidence intervals as weights to calculate a weighted coefficient of determination (R^2) for each tool. For a detailed description of this process, see Supporting Material.

2.4 Population analysis

We compared our aDNA data to the genetic profiles of Eurasian populations from the Affymetrix Human Origins array Lazaridis et al. (2014). Genotyping was performed using BWA mem Li and Durbin (2010), followed by subsampling and pseudohaploid calling with samtools Danecek et al. (2021) and PileupCaller Schiffels (2022). Principal component analysis (PCA) was conducted with

smartpca Galinsky and Mah (2022). Within cluster variance was calculated to estimate the influence of the choice of trimming and merging tools on PCA results, more details on its calculation can be found in the Supporting Material.

2.5 Measuring runtime and memory usage

Runtime and memory usage were measured for each trimming and merging tool using snakemake's Köster and Rahmann (2012) benchmarking functionality. For more details on the machine and memory measurement, please see the Supporting Material.

2.6 Code availability

Code and workflows are available at https://github.com/liannette/mergingtools_benchmark.

Please refer to the supplement for the complete methodology, including details on reference datasets, genotyping, PCA, within cluster variance calculation, and runtime and memory usage measurement.

3 Results

3.1 Merge rate and sequence accuracy

Here, the tools are evaluated regarding their efficacy in reconstructing the sequences of DNA insert molecules from raw paired-end reads. To quantify the differences between the sequence of a merged read and the original DNA insert sequence, the edit distance (also known as Levenshtein distance) is used, as it measures the number of single-character edits required to transform one sequence into another. In other words, the edit distance indicates how many errors are present in the merged read, with an edit distance of 0 indicating that the merged read perfectly reconstructs the sequence of the DNA insert, which is the preferred outcome. The errors in the merged reads originate either from sequencing errors, or from misidentification of the correct positions for adapter trimming or read merging, although adapter trimming is generally straightforward, making it a negligible error source. Substitution errors that occurred during sequencing in overlapping regions of raw paired-end reads may be corrected through merging. Low edit distances indicate the presence of substitution errors which have not been corrected, whereas edit distances larger than 25 can be attributed to larger deletions due to incorrect merging. In addition to measuring the edit distance, we also analyzed the tools regarding false-negative and false-positive merging. A false negative is defined as a read that should have been merged but was not, which leads to undesirable data loss. Conversely, we have simulated fragment sizes of 1,000 bp that should not have been merged, as the reads do not contain any overlap. However, if they were merged, such sequences would be labeled as a false positive. False-positive merging creates

incorrectly merged reads, which contaminates the data and is especially prominent if the sequenced library contains a lot of long DNA insert molecule.

The accuracy of each tool in reconstructing the sequence of the DNA insert molecules for different insert lengths is depicted in Figure 2.

3.1.1 False-negative merging

To reiterate, paired-end reads that have a sufficient read overlap for merging, but remain unmerged, are classified as false negatives. As the minimum read overlap of merging was standardized to 10 bp (see 2.2), it is not surprising that none of the tools are able to merge reads from inserts with a length exceeding 240 bp. But interestingly, AdapterRemoval, leeHom and SeqPrep are able to merge reads from extremely short inserts that have a length of less than 10 bp, despite them being shorter than the minimum read overlap. Among insert lengths ranging from 10 to 240 bases, high rates of false-negative merging were found for ClipAndMerge and SeqPrep, particularly for insert lengths that surpass the read length. Unmerged reads were also observed across all insert lengths for bbmerge and, to a much lesser degree, for seqtk/adna-trim. For bbmerge, the merge rate is highest when the overlap is maximized, which is the case for inserts the size of the read length. Additionally, seqtk/adna-trim was incapable of merging reads from inserts with a length of 125 bp, equivalent to the read length, which might be due to a bug in the software. False-negative merging behavior was almost non-existent in AdapterRemoval, fastp, and leeHom.

3.1.2 False positive merging

In the event that non-overlapping paired-end reads are merged, it is deemed a false-positive. With a read length of 125 bp, the paired-end reads that originate from DNA inserts of 250 bp or more lack any overlapping bases. False-positive merging rates for each tool were calculated as the percentage of merged reads for DNA inserts of 1,000 bp length, and are presented both in the grey bars next to all figures in Figure 2 and numerically in Table 1. Among the analyzed tools, AdapterRemoval had the highest false-positive merging rate with over 6 percent, while bbmerge has the lowest false-positive merging rate with less than 0.1 percent. Some tools also incorrectly merge paired-end reads, even when there is adequate overlap for correct merging. Merged reads with an edit distance exceeding 25 for DNA inserts with lengths ranging from 10 to 240 bp are considered to be a result of this. This negative behavior is most pronounced for AdapterRemoval, where the amount of incorrectly merged reads increases with the insert size, ultimately reaching a maximum of over 3 percent for an insert size of 240 bp, indicating that AdapterRemoval is prone to misidentify correct overlap positions. The same can also be observed for ClipAndMerge, fastp and leeHom, but at much lower rates.

3.2 Robustness to sequencing errors

This experiment evaluates the ability of each tool in reconstructing the sequences of DNA inserts following different empirical insert length distributions and their robustness to higher rates of sequencing errors. The latter was achieved through quality
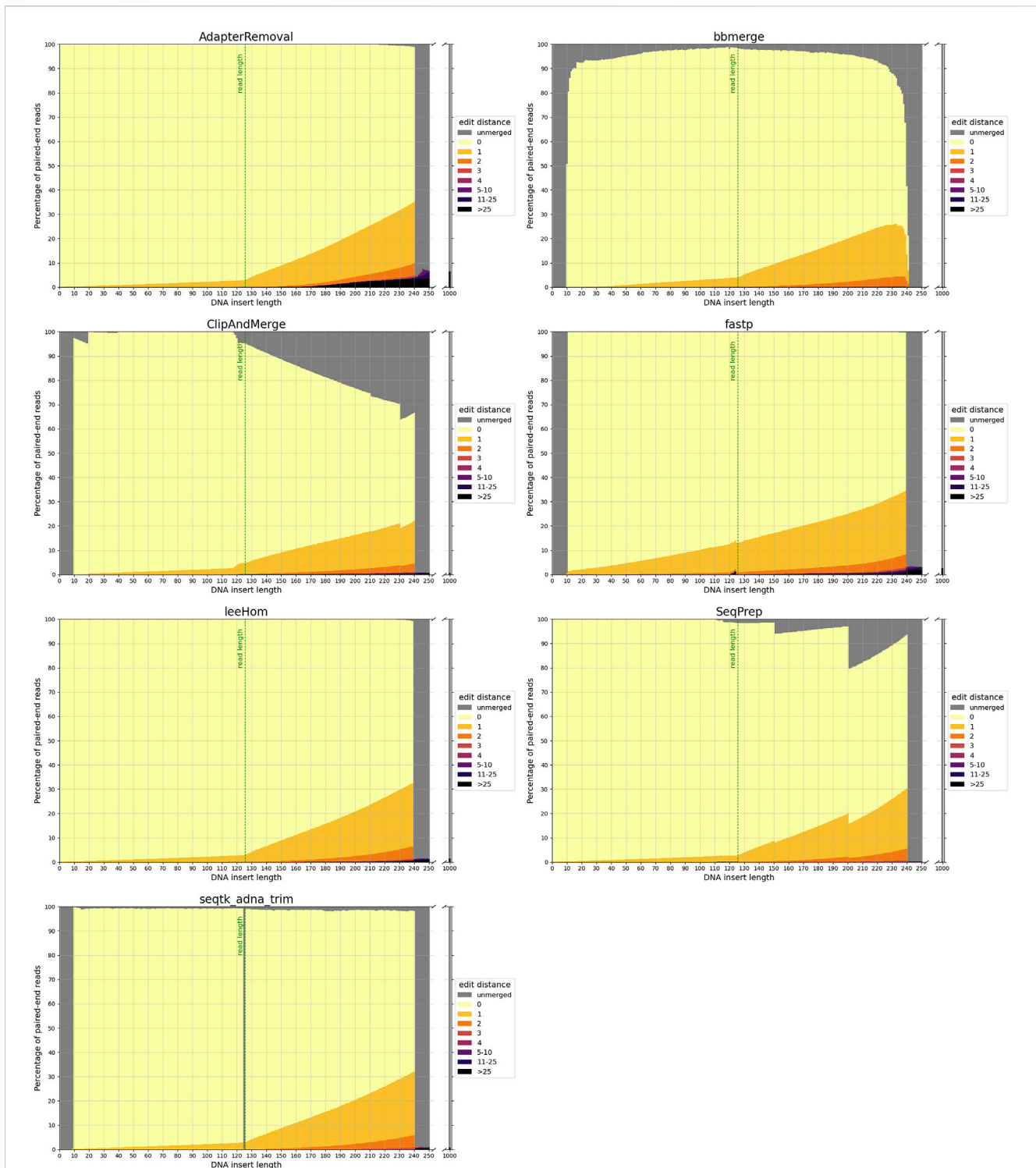
FIGURE 2
Sequence accuracy of merged reads for different insert sizes. Paired-end reads of DNA inserts with different lengths were simulated and then trimmed and merged with the different tools. High edit distances (in darker colors) correspond to high sequence dissimilarity between the merged read and the DNA insert, and unmerged reads are colored in gray. The green line at 125 bp marks the read length. Paired-end reads of inserts with a length up to this value overlap completely. For higher insert lengths, the read overlap decreases, and the proportion of bases that are unique to only one of the paired-end reads increases. For an insert length of 250 bp and higher, the paired-end reads do not overlap.

shift adjustments of the sequencing simulator (further details can be found in the Supplementary Material). The results are illustrated in Figure 3.

Three aDNA molecule length distributions (A9180, Vi33.19 and chagyrskaya8) and one distribution of modern, cfDNA were used for this experiment. Histograms of the four insert sizes distributions are

TABLE 1 False positive merging rate, calculated as the percentage of merged paired-end reads with no overlap.

Tool	Estimated false-positive merging rate [%]
AdapterRemoval	6.54
bbmerge	0.0832
ClipAndMerge	0.597
Fastp	2.65
leeHom	1.44
SeqPrep	0.192
seqtk/adna-trim	0.814

depicted in Figure 4. Among the analyzed distributions, A9180 exhibits the shortest insert sizes, predominantly ranging between 10 and 30 bp. It is also the only distribution containing lengths under 30 bp. For Vi33.19 and chagyrskaya8, insert sizes between 40 and 50 bp are most frequent, however, chagyrskaya8 contains also a substantial quantity of lengths exceeding 125 bp, which is not the case for Vi33.19. Most cfDNA molecules span between 125 and 225 bp, with a secondary, smaller distribution at 275–375 bp. At a read length of 125 bp, the vast majority of paired-end reads from aDNA can potentially be merged, whereas for cfDNA, a fraction of reads will remain unmerged due to the lack of overlap.

First, the three aDNA insert size distributions will be examined in detail. At lower sequencing error rates (quality shifts 0 to −10), most tools exhibit similar performance, though some tools fail to merge some reads. This behavior is particularly pronounced for bbmerge. The merged reads with the most errors are generated by fastp. Especially once the sequencing quality falls below a specific threshold (quality shift −9), can an abrupt increase in not-perfectly reconstructed sequences be observed for fastp. AdapterRemoval and leeHom display the highest merge rates and perform almost identically for the aDNA distributions. SeqPrep does not merge some reads at high sequence error rates, but this behavior acts as a sort of filter for reads with a high sequence dissimilarity. In summary, AdapterRemoval and leeHom demonstrate the most robust behavior, meaning that they are capable of merging reads at high sequencing error rates.

Looking at the results for the cfDNA insert size distribution, we see that a significant number of reads remain unmerged across all tools, even at low error rates. This can be attributed to the presence of DNA molecules longer than 250 bp in cfDNA, resulting in paired-end reads with no overlap. ClipAndMerge, however, stands out as it has a higher amount of unmerged reads compared to the other tools. The overall higher amount of longer DNA molecules in cfDNA also accounts for the elevated number of merged reads with high edit distances compared to aDNA, as this not only results in more nucleotides per merged read, but also a smaller read overlap. For bbmerge, fastp and SeqPrep, the merge rate declines dramatically at high sequencing error rates (quality shift of −10 or lower), indicating that these tools are more sensitive to sequencing errors in the overlap. The highest overall merging rate is again exhibited by AdapterRemoval and leeHom. However, of all the tools, AdapterRemoval has the largest number of reads with an edit

distance higher than 25 across all quality shifts, indicating that those reads are incorrectly merged.

3.2.1 Correcting sequencing errors through merging

The probability of at least one sequencing error occurring within the insert sequence increases with the number of bases in the insert. As a consequence, the percentage of not perfectly reconstructed sequences tends to rise with increasing insert length. Nonetheless, a distinction can be made between paired-end reads with full and partial overlaps. For inserts with a length shorter or equal than the read length, 125 bp, every base pair of the insert is sequenced twice, and as such, is present in both reads. Sequencing errors at any position can, in theory, be corrected by using information from the other read. For inserts that are longer than the read length, some base pairs are, however, only sequenced in one of the reads, making it impossible to correct sequencing errors in those positions. For DNA inserts lengths shorter than the read length, the reads merged with fastp contain significantly more reads with sequencing errors compared to the other tools. For all the analyzed tools, except fastp, the number of reads with edit distances greater than zero increases with the insert length at a slower rate for lengths up to 125 bp, compared to insert lengths exceeding 125 bp.

3.3 Per-base merging behavior

In the overlapping region of a paired-end read, each nucleotide base and corresponding Phred quality score is present twice, once on the forward and once on the reverse read. This information is collapsed into a single base and quality score when merging the reads. This experiment evaluates what values each of the tools outputs for any combination of base and quality scores on the forwards and reverse reads.

3.3.1 Matching bases

The heatmaps in Figures 5, 6 visualize the per-base merging behaviors when the base is identical on both the forward and the reverse read. The merging of the base is unambiguous in this case, but the strategies for merging of Phred quality scores vary among the analyzed tools. Some tools generate a new Phred quality score that exceeds either value from the paired-end read, while others select the greater Phred quality score from either read, and some simply select the Phred quality score from the forward read. Sequencing the same base on both reads increases the confidence in the accuracy of that base, thus, the most accurate approach is to increase the Phred quality score in the merged read. Among the analyzed tools, AdapterRemoval, leeHom, and SeqPrep exhibit this behavior. LeeHom and SeqPrep can generate Phred quality scores of up to 60, while AdapterRemoval can even generate values up to 93, although this requires specifying a parameter, as the default maximum Phred value of AdapterRemoval is 41. In contrast, ClipAndMerge and seqtk/adna-trim simply select the greater Phred quality score. The merging behavior of bbmerge lies between the two aforementioned approaches, although it is more similar to the second one. The most simple behavior is displayed by fastp, which just selects the Phred quality score of the forward read and ignores the reverse read entirely.
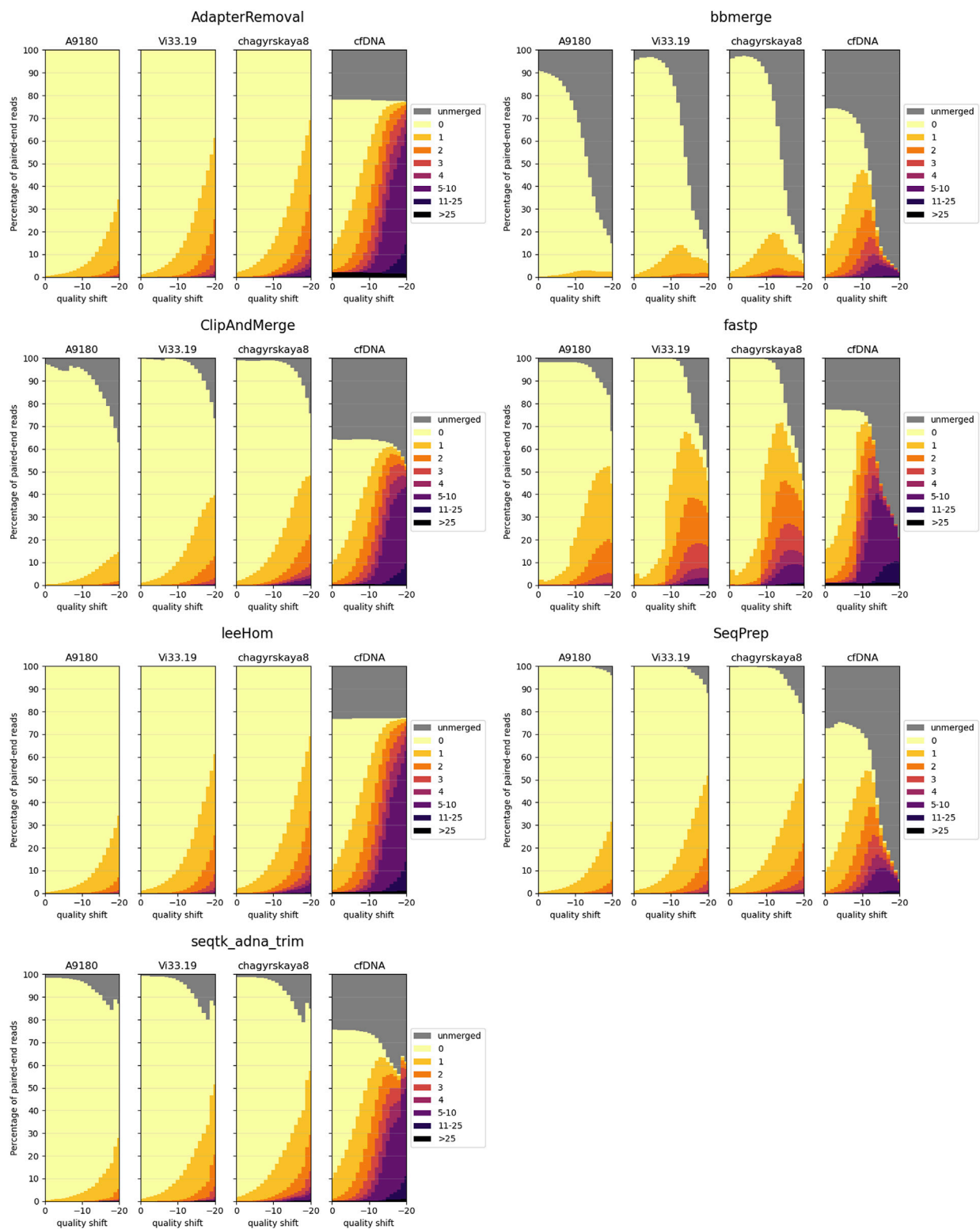
FIGURE 3
Sequence accuracy of merged reads for different insert size distributions. Paired-end reads of DNA inserts with different length distributions were simulated and then trimmed and merged with the different tools. A9189, Vi33.19 and chagyrskaya8 refer to insert size distributions of aDNA, whereas cfDNA represents the insert size distribution of cell-free DNA. High edit distances (in darker colors) correspond to high sequence dissimilarity between the merged read and the DNA insert.

FIGURE 4
Empirical insert length distributions. To evaluate the tools' ability to reconstruct the sequence of fragmented DNA molecules, DNA inserts following specific length distributions were simulated. In addition to the three aDNA insert length distributions A9180, Vi33.19 and chagyrskaya8, which were generated from different Neanderthal bone samples, also one insert size distribution was used, which was obtained from a sample of cfDNA.
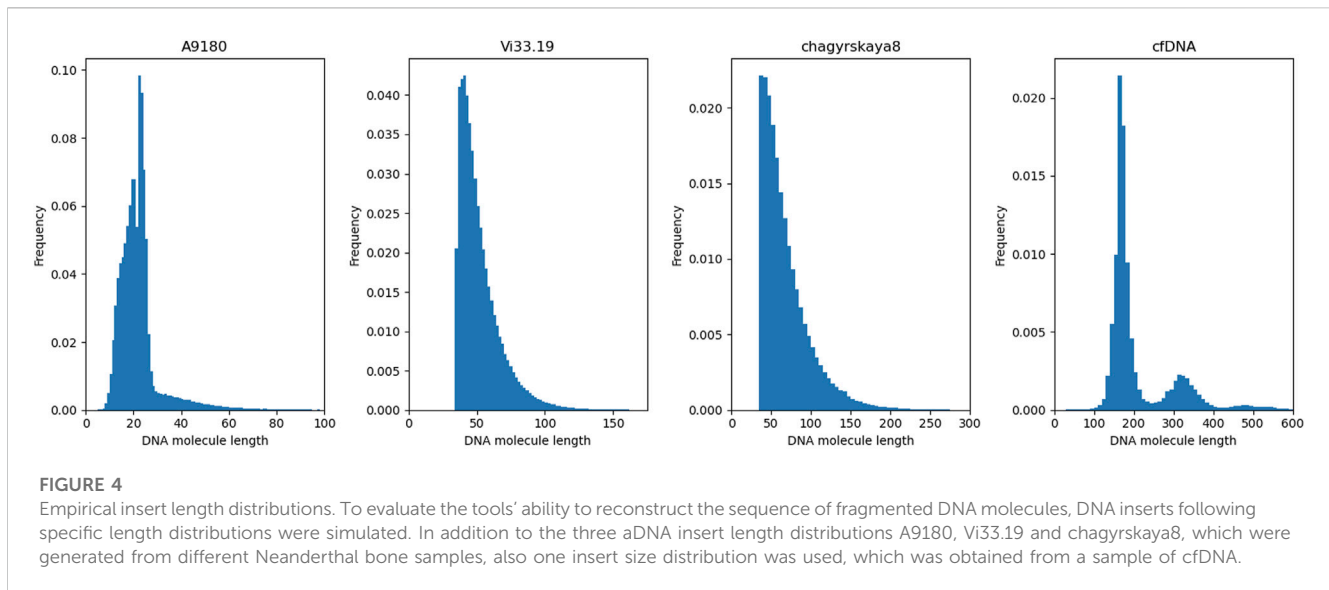
3.3.2 Mismatching bases

The heatmaps featured in Figures 7, 8 illustrate the per-base merging behavior of the analyzed tools when the nucleotide base differs between the forward and the reverse read. Again, the tools exhibit distinct behavior patterns in these situations. When a mismatch occurs, the confidence in the accuracy of the merged base should be reduced, particularly when the quality scores of both reads are similar. AdapterRemoval, bbmerge, leeHom, SeqPrep, and seqtk/adna-trim adopt this approach, selecting the base with the highest Phred quality score and using the quality score information from the other read to appropriately adjust the Phred quality score of the merged read. Conversely, ClipAndMerge selects both the base and quality score from the read with the highest Phred quality score, without incorporating information from the other read to modify the Phred quality score in the merged read. fastp, on the other hand, generally obtains the base and quality score from the forward read, except when the quality score is below 15 on the forward read and at least 30 on the reverse read, in which case fastp derives the nucleotide and quality score from the reverse read. This behavior is surprising and indicates that fastp is not as capable as the other tools to reduce sequencing errors.

In the case of all tools except fastp, an interesting scenario arises when these tools must choose between bases with equal Phred scores during the read merging process. In such instances, AdapterRemoval and leeHom make a random selection between the two bases, while ClipAndMerge and SeqPrep opt for the base from the forward read, and seqtk/adna-trim selects from the reverse read. Notably, bbmerge stands out as the only tool that generates an N in this situation.
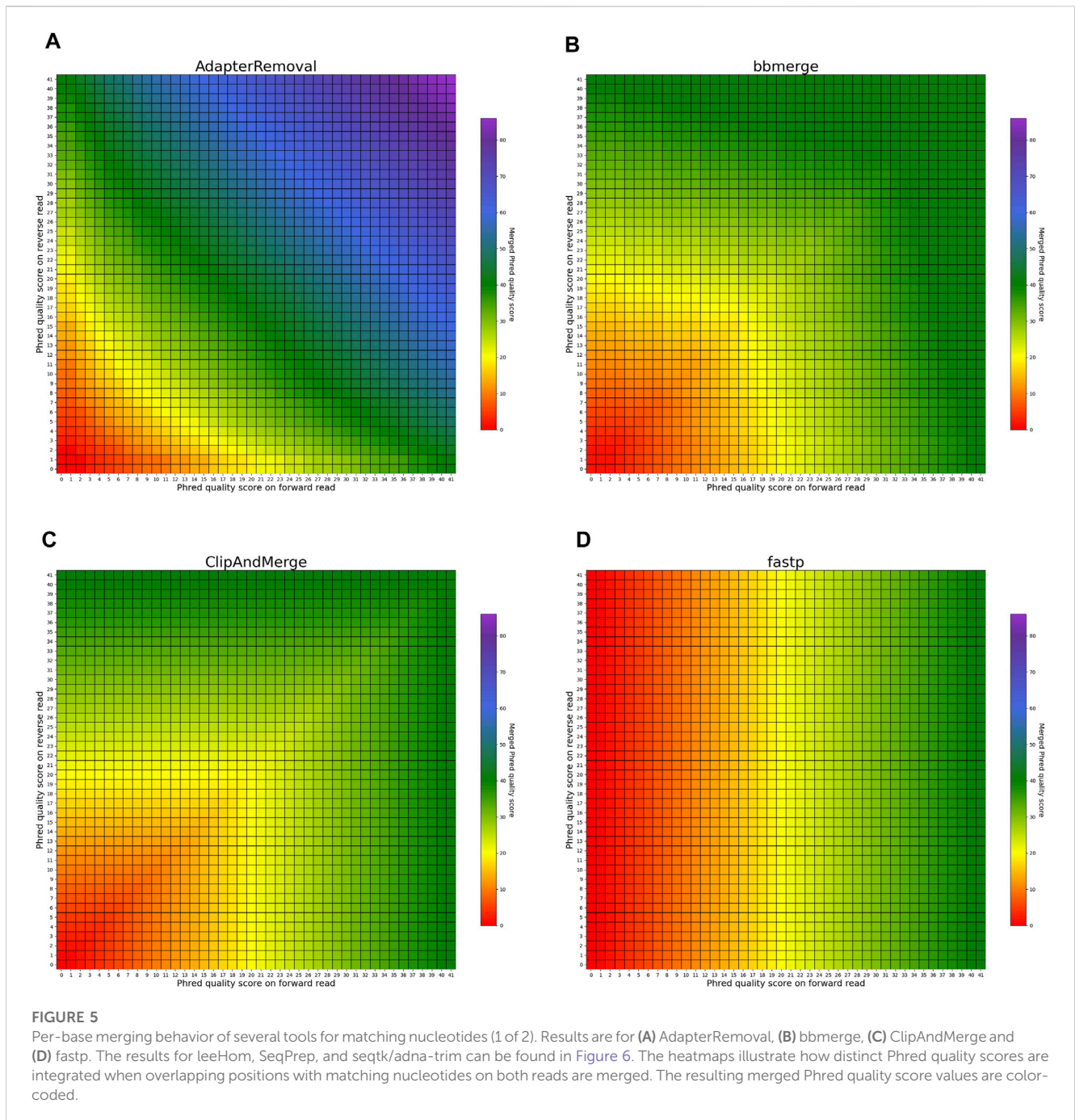
3.4 Phred quality score accuracy

In this experiment, the Phred quality score accuracy of the tools is analyzed by measuring how close the theoretical error probabilities of the Phred quality scores are to the observable error rates. Figure 9 illustrates the Phred-scaled observed error rates for each merged Phred value, combining the data from five

datasets utilizing different quality shifts. Figures displaying the results for each quality shift separately can be found in the Supporting Material (Supplementary Figures S2–S8). To indicate how well the Phred values of the observed error rate fit the merged Phred values overall, the weighed coefficient of determination, also called R^2, was calculated for each tool from the combined data, using the inverse range of the confidence intervals as weights to reduce the influence of values with large uncertainty. The weighed R^2 for each of the analyzed tools, whose values can range from 0 to 1, with 1 indicating a perfect fit, can be found in Table 2.

The worst fit between the observed error rates and merged Phred values, as indicated by the R^2 value, is seen for seqtk/adna-trim, followed by bbmerge. For both tools, the Phred values of the observed error rates are often higher than the merged Phred values, which means that the actual error rates of merged bases are lower than indicated by their Phred quality scores. For leeHom and SeqPrep, the actual error rate for the (maximum) merged Phred value of 61 is lower than expected. Furthermore, for leeHom and AdapterRemoval more errors were observed for some Phred values around 14 and 25. However, this is only the case at very high sequencing error rates (see Supplementary Results), which is not typically the case with most real data. For fastp, the Phred value calibration is nearly perfect, except for Phred values lower than 15. The best R^2 value is displayed by ClipAndMerge, indicating a near perfect fit of the Phred calibration.

3.5 Principal component analysis for population analysis

To examine the potential impact of the different tools, a PCA was performed. Reads of an Ashkenazi Jewish individual (HG002) and a Han Chinese individual (HG005), modified to follow an aDNA insert length distribution, were preprocessed with each of the tools and then genotyped at typical aDNA genome coverages. For each tool and genome coverage, ten samples were generated. Figure 10 shows the first two principal components calculated from the genotype data of present-day Eurasian individuals, on which the

FIGURE 5
Per-base merging behavior of several tools for matching nucleotides (1 of 2). Results are for **(A)** AdapterRemoval, **(B)** bbmerge, **(C)** ClipAndMerge and **(D)** fastp. The results for leeHom, SeqPrep, and seqtk/adna-trim can be found in Figure 6. The heatmaps illustrate how distinct Phred quality scores are integrated when overlapping positions with matching nucleotides on both reads are merged. The resulting merged Phred quality score values are color-coded.
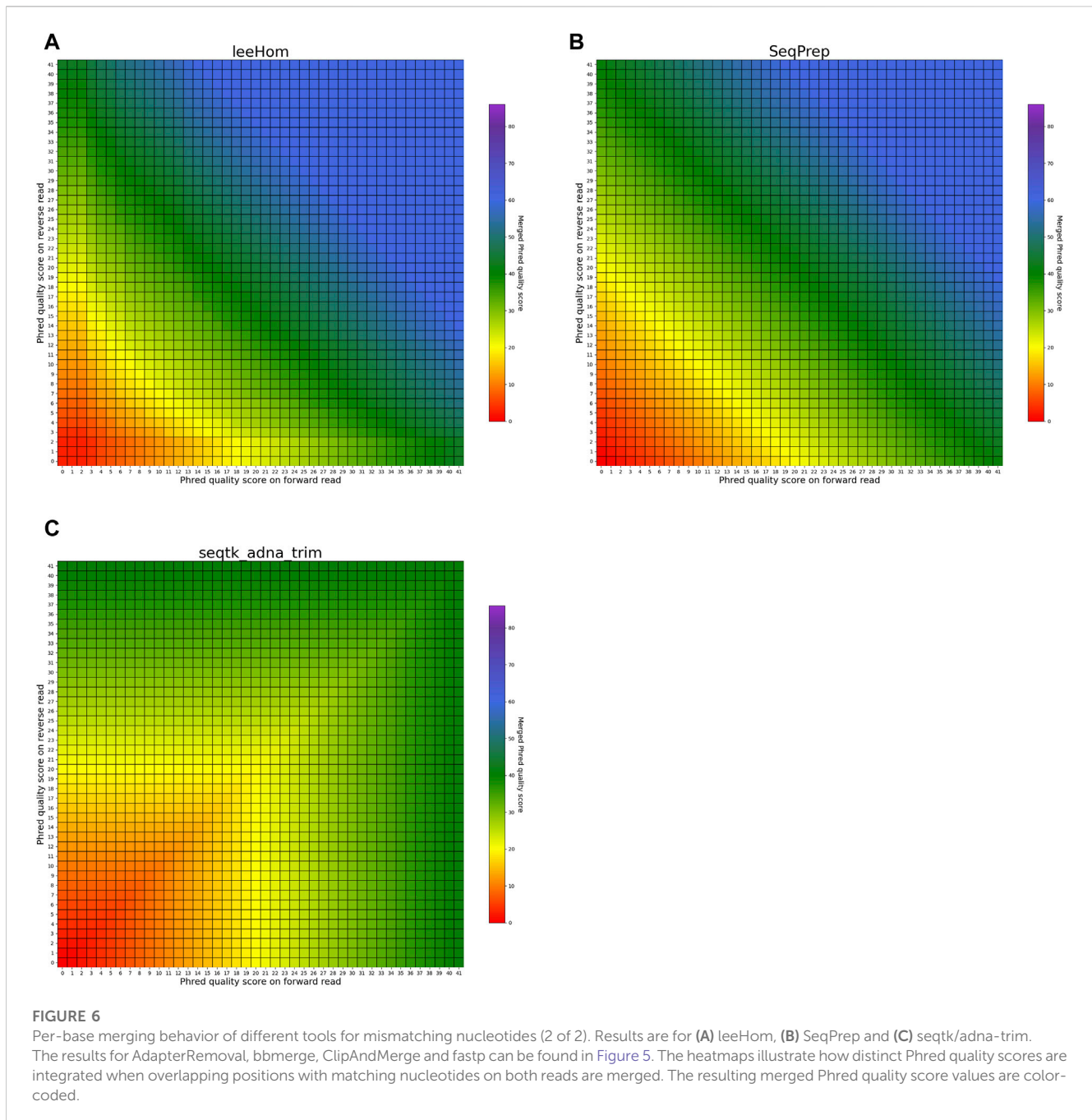
HG002 and HG005 samples of approximately 0.25X coverage were projected. All the HG002 and HG005 samples project in the same area as their respective populations, irrespective of the tool used for adapter trimming and read merging. Figures for higher coverages, including zoomed-in plots, in which the samples belonging to the different tools can be distinguished, can be found in the supplement (Supplementary Figures S15–S19). To estimate the accuracy of each tool when performing a PCA, the variability between the samples was assessed by calculating the within-cluster variance using the coordinates on the first ten principal components. A small within cluster variance indicates that the projections of the samples lie close to each other in the PCA space, which indicates robust results. As seen in Figure 11, the within cluster variance does not significantly

differ between the tools, in other words, the choice of tool did not impact the accuracy of the performed PCA.

3.6 Runtime and memory usage

The runtime and memory usage of each adapter trimming and read merging tool were measured while processing 1 million reads following an aDNA insert size distribution. The tools differ substantially in their resource consumption, shown in Table 3. When comparing the results, it has to be taken into account that ClipAndMerge was utilizing three threads, whereas all other tools were run single-threaded for this experiment.

FIGURE 6
Per-base merging behavior of different tools for mismatching nucleotides (2 of 2). Results are for **(A)** leeHom, **(B)** SeqPrep and **(C)** seqtk/adna-trim. The results for AdapterRemoval, bbmerge, ClipAndMerge and fastp can be found in Figure 5. The heatmaps illustrate how distinct Phred quality scores are integrated when overlapping positions with matching nucleotides on both reads are merged. The resulting merged Phred quality score values are color-coded.
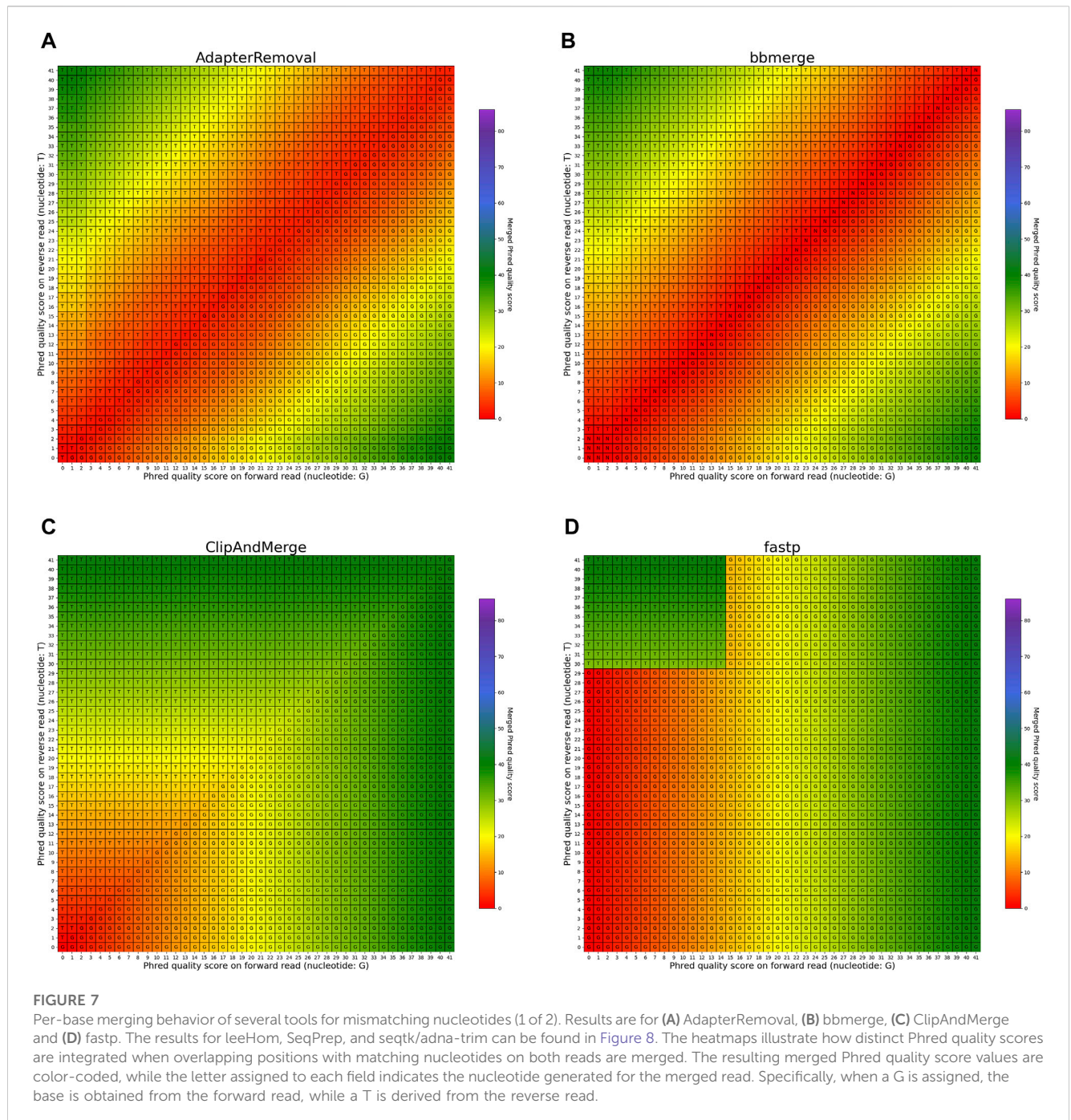
Among the tools, fastp stands out as having the highest memory usage among the analyzed tools, followed by ClipAndMerge and bbmerge. The tool with the lowest memory usage is SeqPrep, although it has also by far the longest runtime, almost 4-fold higher than any other tool. A comparably long runtime was also observed for leeHom and ClipAndMerge, despite the latter used three threads. The tool characterized by the shortest runtime is seqtk/adna-trim, followed by fastp, bbmerge and AdapterRemoval. Other considerations related to runtime and memory are the option of piping commands, which not possible with AdapterRemoval, among the possibility of multithreading, which is not supported by SeqPrep, whereas ClipAndMerge number of used threads is fixed at three.
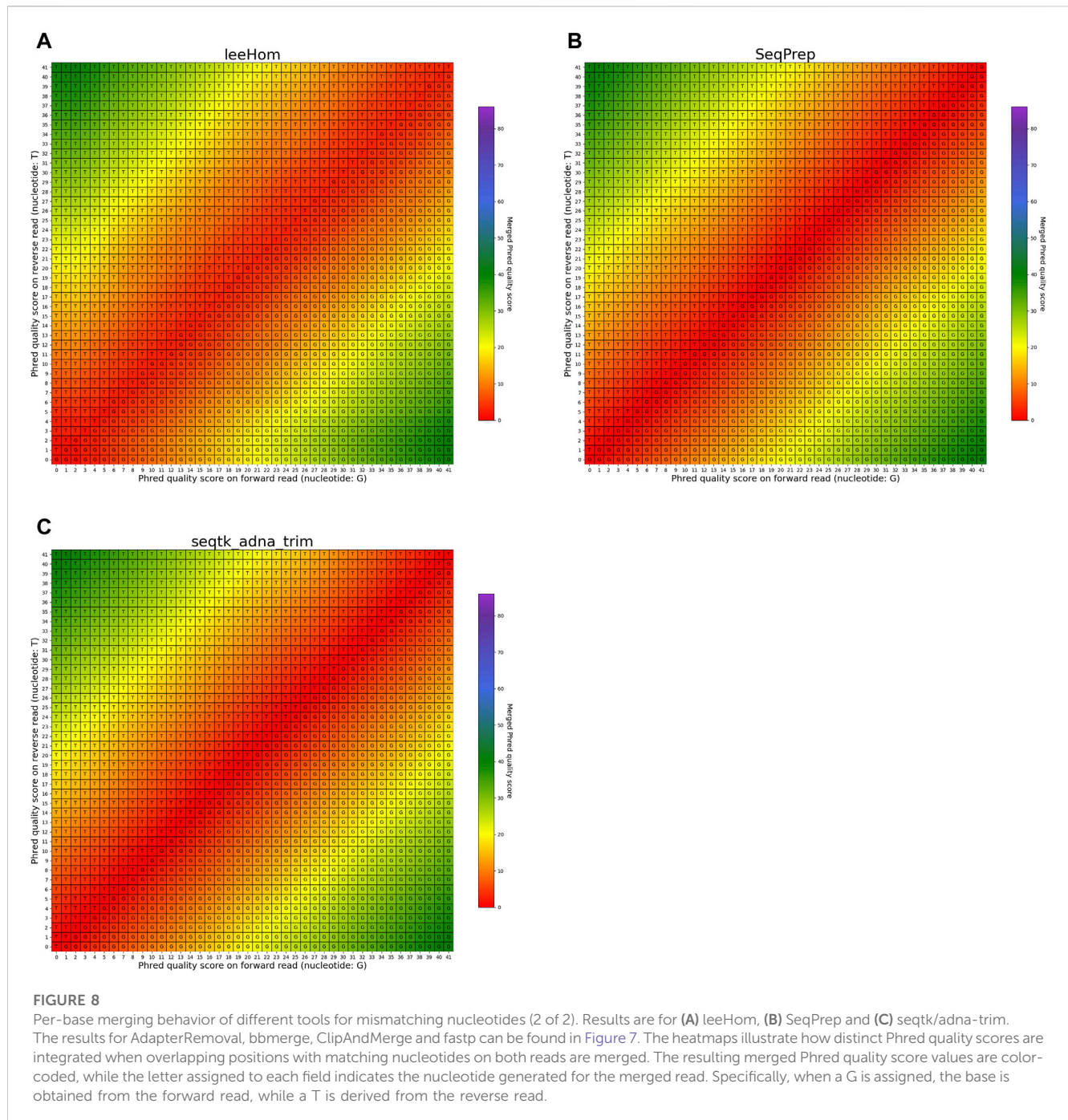
4 Discussion

4.1 Merged Phred value calculation and accuracy

In order to provide the most accurate Phred quality score for a merged base, the information about a nucleotide match/mismatch should be incorporated by the tools. The confidence in a base call being correct should increase if the same base is called on both reads in an overlapping part of a paired-end read, and decrease if a different base is called in the other read. fastp and ClipAndMerge are considered the least capable among the examined tools in improving read quality, as they always adopt

FIGURE 7
Per-base merging behavior of several tools for mismatching nucleotides (1 of 2). Results are for **(A)** AdapterRemoval, **(B)** bbmerge, **(C)** ClipAndMerge and **(D)** fastp. The results for leeHom, SeqPrep, and seqtk/adna-trim can be found in Figure 8. The heatmaps illustrate how distinct Phred quality scores are integrated when overlapping positions with matching nucleotides on both reads are merged. The resulting merged Phred quality score values are color-coded, while the letter assigned to each field indicates the nucleotide generated for the merged read. Specifically, when a G is assigned, the base is obtained from the forward read, while a T is derived from the reverse read.

the Phred value from the selected nucleotide instead of calculating a new quality score for a merged base, even when there is a nucleotide mismatch. However, it should be noted that ClipAndMerge generates quality scores that almost perfectly correlate with the actual error rates. It might therefore be relevant for niche applications that depend on highly accurate Phred value calibrations. In contrast, seqtk/adna-trim, and similarly bbmerge, generate a decreased quality score when mismatching nucleotides are merged, but neither of the tools calculates an increased Phred value when merging matching nucleotides. This ambiguous strategy also leads to the least accurate Phred calibrations. The most advanced per-base merging strategies are employed by

AdapterRemoval, leeHom, and SeqPrep, as those tools always calculate a new Phred quality score for the merged reads, taking the quality scores from both reads and whether the bases match or not into account. To summarize, AdapterRemoval, leeHom and SeqPrep are the most suitable tools for downstream applications that rely on accurate base quality scores, one example being Bayesian variant calling methods. If the base quality scores are only used to identify sequencing errors, for example, to perform quality trimming, bbmerge and seqtk/adna-trim offer sufficient functionality. On the other hand, it is advisable to avoid using fastp and ClipAndMerge for any application that uses base quality scores.
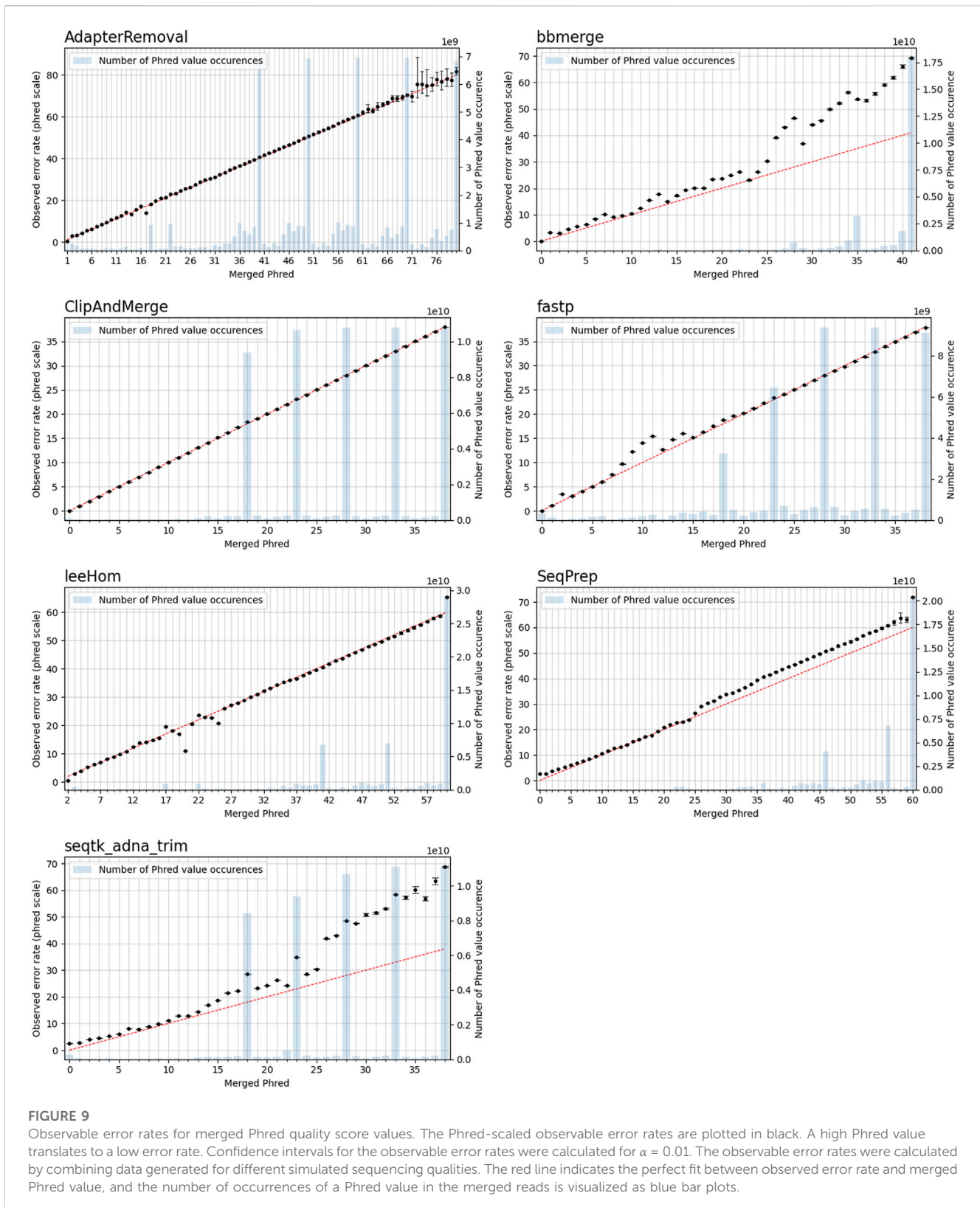
FIGURE 8
Per-base merging behavior of different tools for mismatching nucleotides (2 of 2). Results are for **(A)** leeHom, **(B)** SeqPrep and **(C)** seqtk/adna-trim. The results for AdapterRemoval, bbmerge, ClipAndMerge and fastp can be found in Figure 7. The heatmaps illustrate how distinct Phred quality scores are integrated when overlapping positions with matching nucleotides on both reads are merged. The resulting merged Phred quality score values are color-coded, while the letter assigned to each field indicates the nucleotide generated for the merged read. Specifically, when a G is assigned, the base is obtained from the forward read, while a T is derived from the reverse read.

4.2 Correcting sequencing errors

Almost all the analyzed tools select the nucleotide with the highest quality score when merging. Only fastp chooses in some cases the base with the lower Phred value, limiting its ability to correct sequencing errors. Specifically, fastp corrects sequencing errors on the forward read only when a high Phred value is found on the reverse read, meaning that sequencing errors on the forward read are not corrected when the sequencing quality is below a certain threshold. Completely overlapping reads merged with fastp contain more sequencing errors compared to the other tools. Furthermore, all tools, except for fastp, exhibit differing rates of sequencing error increase depending on the insert length. This difference is especially apparent when comparing completely overlapping reads to partially overlapping reads, which indicates that some sequencing errors are indeed corrected in the overlapping parts. Therefore, it is recommended to avoid fastp for processing aDNA data, as it retains more sequencing errors in the merged reads as any of the other tools.

4.3 Incorrect merging and not-merging

In general, there is a trade-off between incorrectly merged reads and wrongly unmerged reads. For some downstream analyses,

FIGURE 9
Observable error rates for merged Phred quality score values. The Phred-scaled observable error rates are plotted in black. A high Phred value translates to a low error rate. Confidence intervals for the observable error rates were calculated for $\alpha = 0.01$. The observable error rates were calculated by combining data generated for different simulated sequencing qualities. The red line indicates the perfect fit between observed error rate and merged Phred value, and the number of occurrences of a Phred value in the merged reads is visualized as blue bar plots.

incorrectly merged reads would not have a big impact, therefore obtaining the highest possible merge rate is desired. One example would be alignment to a reference genome, where the incorrectly merged reads would either most likely not align at all or result in low alignment scores, making it easy to filter them out. For other

applications on the other hand, such as *de novo* genome assembly, incorrectly merged reads pose a big challenge, and generating fewer reads, but with a higher quality, is preferred.

Among all tools, leeHom and AdapterRemoval have the highest merge rate. Both tools are able to merge almost all reads with

TABLE 2 Weighted R^2 values indicating the error rate accuracy of merged Phred quality scores, calculated by combining data of different sequencing qualities.

Tool	R^2
AdapterRemoval	0.9954
bbmerge	0.9277
ClipAndMerge	0.9999
Fastp	0.9943
leeHom	0.9725
SeqPrep	0.9707
seqtk/adna-trim	0.8032

sufficient overlap, even at high sequencing error rates. However, AdapterRemoval has the highest tendency to incorrectly merge reads among all tools, often merging non-overlapping paired-end reads and misidentifying correct overlap positions. A similar tendency is seen for fastp. Generally, incorrect merging behavior is primarily present at insert lengths close to double the read length, when the read overlap is small. As aDNA molecules are typically shorter, this aspect has not a big influence when reconstructing the sequences of DNA molecules following the insert size distributions of aDNA. However, the effect is clearly visible when working with DNA insert sizes matching those of cfDNA. A similar effect was observed for ClipAndMerge, but with wrongly unmerged reads (false-negatives) instead of incorrectly merged reads. ClipAndMerge has especially low merge rates for insert sizes between the read length and double the read length (125–250 bp), which is a range that encompasses a significant proportion of cfDNA, but not of aDNA molecules. If the goal is to minimize the number of incorrectly merged reads and sequencing errors, bbmerge would be the best choice. However, bbmerge's high false-negative merging rate, especially at higher sequencing error rates, makes it a suboptimal choice when working with aDNA, where data is already scarce, or any other analysis where it is desired to merge as many reads as possible. The differences in the merge behavior can be explained by the different strategies employed by the tools to detect a read overlap. Most tools use a two-step approach, first using local alignment to identify the adapter sequences and then performing read merging using threshold values for mismatches or edit distance in the overlap. Some tools also consider the quality scores, ClipAndMerge, for example, ignores mismatches if one base has a low quality score, and bbmerge determines the overlap position based on multiple heuristics controlled by predefined constants. Those cut-off values or constants determine how sensitive or robust a tool is to false-positive and false-negative merging. The user is often able to modify those values, although it can be expected that in the majority of cases the default values are used. The only tool that does not depend on arbitrary cut-offs is leeHom, which employs a Bayesian maximum *a posteriori* algorithm, performing adapter trimming and read merging in
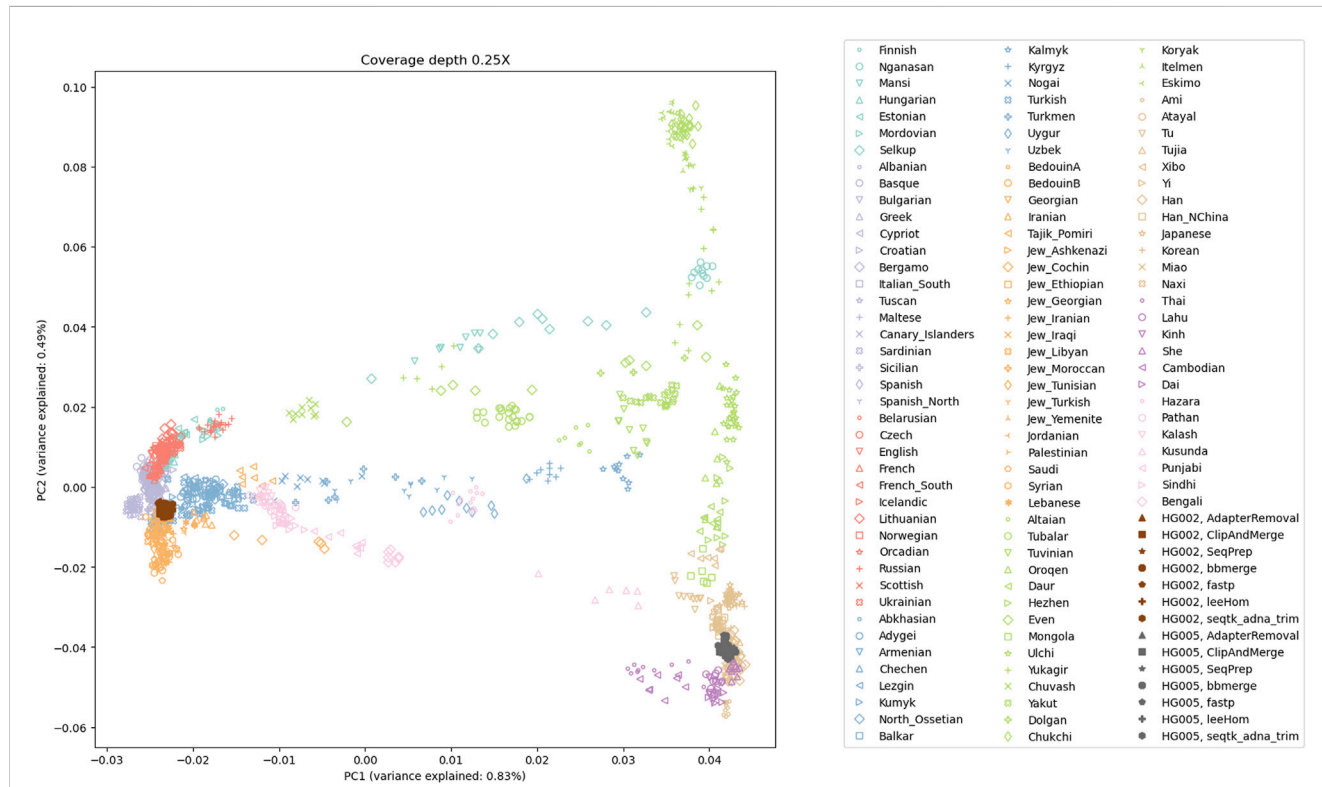

FIGURE 10
The first two principal components of a PCA of present–day Eurasian individuals; with projected low coverage (0.25X) samples for HG002 and HG005. The projections of all samples for the Ashkenazi Jewish individual (HG002) and the Han Chinese individual (HG005) fall within proximity to their respective populations.
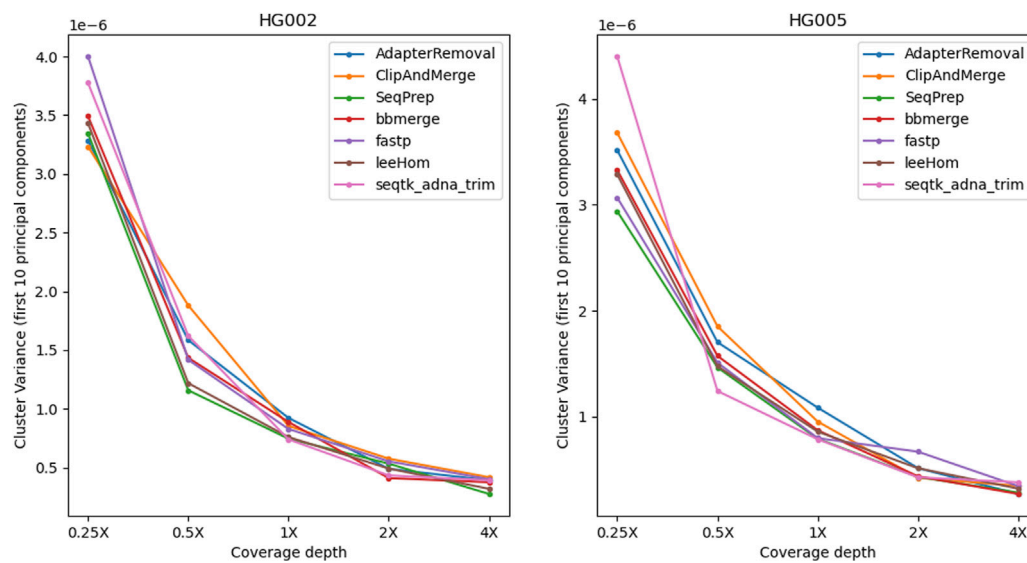
FIGURE 11
Within-cluster variance of samples projected onto the PCA. For each tool, the alignment was subsampled to simulate different genome coverage depths, before genotyping the samples and projecting them onto the PCA. To estimate the accuracy of the tools in reproducing the same results, the process has been repeated ten times for each coverage and tool. The variability between the PCA projections was assessed using the first ten principal components to calculate the within-cluster variance, where a high value indicates a high variability. Overall, the within-cluster variance is comparable across the tools.

TABLE 3 Runtime and memory usage of the analyzed adapter trimming and merging software tools.

Tool	Utilized threads	Adaptable thread number	Pipeable	Runtime [s]	Memory usage [MB]
AdapterRemoval	1	Yes	No	14.6 ± 0.2	25.56 ± 00.08
bbmerge	1	Yes	Yes	37.5 ± 1.0	315.75 ± 39.20
ClipAndMerge	3	No	Yes	172.0 ± 3.7	704.89 ± 09.77
Fastp	1	Yes	Yes	34.5 ± 0.3	1,009.18 ± 03.89
leeHom	1	Yes	Yes	164.3 ± 0.5	73.80 ± 01.11
SeqPrep	1	No	Yes	633.6 ± 2.1	1.26 ± 00.01
seqtk/adna-trim	1	Yes	Yes	16.5 ± 1.3	70.41 ± 00.01

one step. Given a prior belief on the fragment length distribution, leeHom uses the quality scores to calculate the error probability for each overlap position, including no overlap, and then chooses the case with the highest posterior probability. In this study, the prior distribution was adapted using the–ancientdna parameter, whereas at default settings a uniform prior is applied.

Overall, leeHom's merging algorithm outperforms the other approaches. Even at high rates of sequencing errors, leeHom is able to correctly merge nearly all overlapping paired-end reads while introducing only low amounts of contamination in the form of false-positively merged reads. A good alternative to leeHom is seqtk/adna-trim. Although it fails to merge a small percentage of reads, it produces fewer incorrectly merged reads than leeHom. However, bbmerge is the most suitable tool when the presence of incorrectly merged reads is very problematic for further analyses, such as *de novo* genome assembly. ClipAndMerge and SeqPrep are not recommended due to their reduced ability to merge paired-end

reads with a smaller overlap, whereas AdapterRemoval and fastp should be avoided as they have a tendency to incorrectly identify the correct overlap positions and display a high false-positive merging rate.

4.4 Effect on downstream analyses

To investigate the potential impact of different tools on downstream analyses, a PCA based on genome-wide SNP data was performed, a commonly employed analysis for determining ancestry. Among the tested tools, no measurable differences in the results were observed, suggesting that reads processed with any of the tools can be combined without introducing batch effects. It is important to note, however, that the workflow is to a high degree robust against incorrectly merged reads and imprecise base quality scores, which are key differentiating factors among the tools.

TABLE 4 Summary of the performance of the tools for adapter trimming and read merging.

Tool	Merge rate	Sequencing error correction	Phred values calculation	Runtime	Memory usage	Suitability for aDNA analysis
AdaperRemoval	- - highest rate of incorrect merging	+ chooses the base with the highest quality score	++ considers both mismatch and match	++ shortest runtime, but not pipeable	+	partly recommended, prone to misidentify correct overlap for longer fragments
bbmerge	+ lowest incorrect merging rate, but fails to merge some reads and sensitive to seq. errors in overlap	+ chooses the base with the highest quality score	+ considers only mismatch	+	-	only recommended if downstream analyses are highly sensitive to incorrectly merged reads
ClipAndMerge	- reduced ability to merge reads from longer templates	+ chooses the base with the highest quality score	- no recalculation of the quality score	- fixed thread number of 3	-	not recommended, inferior quality score calculation and low merge rate for longer fragments
fastp	- - high rate of incorrect merging	- biased to bases from the forward read	- no recalculation of the quality score	+	- -	should be avoided, due to inferior sequence correction capabilities
leeHom	++ best merge rate, but misidentifies the overlap in some cases	+ chooses the base with the highest quality score	++ considers both mismatch and match	-	+	overall recommended for most applications
SeqPrep	- reduced ability to merge reads from longer templates	+ chooses the base with the highest quality score	++ considers both mismatch and match	- - multithreading not supported	++	partly recommended, low merge rate for longer fragments
adna-trim	+ low rate of incorrect merging, but fails to merge sometimes	+ chooses the base with the highest quality score	+ considers only mismatch	++	+	suitable, not able to merge all reads, but low incorrect merging rate

Incorrectly merged reads have no negative effect if they fail to align to the reference genome, and pseudo-haploid genotyping was performed by randomly selecting a read for each SNP position, irrespective of the base quality scores. It is therefore possible that the choice of the tools have a greater influence on downstream analyses that are more sensitive to either of those two factors.

4.5 Conclusion

We find that the analyzed tools exhibit notable differences in their abilities to correct sequence errors and identify the correct read overlap, but the most substantial difference is observed in their ability to calculate quality scores for merged bases. In this regard, we find that fastp might not be suitable for aDNA as it displays inferior sequence correction capabilities, whereas leeHom outperforms the other tools in most aspects except for runtime. Not only does leeHom apply the most statistically correct method when calculating merged Phred values, it also correctly identifies the read overlap in most cases and is robust against sequencing errors. AdapterRemoval offers favorable quality score predictions, but often misidentifies the correct read overlap and has a high rate of false positives. bbmerge seems sensitive to sequencing errors, and its predicted quality scores suffer from poor correlation to their actual error rates. However, it is the recommended tool in cases where it is preferred to have a reduced amount of merged reads in favor of minimal incorrectly-merged reads. adna-trim offers good accuracy and the best speed, but seems to overestimate the error rate of merged portions. ClipAndMerge and SeqPrep seem to have built-in cutoffs that limit their ability to accurately infer longer templates and lower robustness to sequencing errors. Furthermore, ClipAndMerge's calculation of quality scores does not incorporate mismatch information and should be avoided if the identification of sequencing errors is important for further analysis steps. Some tools are characterized by especially large resource usage, with SeqPrep having an exceptionally large runtime, and fastp having an exceptionally large memory usage. Choosing the appropriate trimming and merging tool for a given project depends on the time and memory that is available, the length of the DNA library molecules, the sequencing quality and what downstream analysis will be performed. Table 4 provides an overview of the performance of the tools against each criterion, and the overall suitability for its use with aDNA. Although the analyzed tools differ profoundly in their algorithms and the resulting ability to reconstruct aDNA sequences, they had no measurable effect when analyzing population genetics using PCA. However, the choice of merging tool may have a bigger impact on downstream analyses that are sensitive to incorrectly merged reads or that rely on Phred quality scores, as those are the main aspects in which the tools differ from each other. This should be tested in future experiments.

4.6 Permission to reuse and copyright

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author.

Author contributions

Funding

Acknowledgments

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest. GR is the programmer behind leeHom. However, AL and LPL did not treat this program any differently from the others.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2023.1260486/full#supplementary-material

References

Bushnell, B., Rood, J., and Singer, E. (2017). BBMerge–accurate paired shotgun read merging via overlap. *PloS ONE* 12, e0185056. doi:10.1371/journal.pone.0185056

Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* 34, i884–i890. doi:10.1093/bioinformatics/bty560

Clopper, C. J., and Pearson, E. S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika* 26, 404–413. doi:10.1093/biomet/26.4.404

Cohen, P., Bacilieri, R., Ramos-Madrigal, J., Privman, E., Boaretto, E., Weber, A., et al. (2023). Ancient DNA from a lost Negev Highlands desert grape reveals a Late Antiquity wine lineage. *Proc. Natl. Acad. Sci.* 120, e2213563120. doi:10.1073/pnas.2213563120

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., et al. (2021). Twelve years of SAMtools and BCFtools. *GigaScience* 10, giab008. doi:10.1093/gigascience/giab008

Der Sarkissian, C., Möller, P., Hofman, C. A., Ilsøe, P., Rick, T. C., Schiøtte, T., et al. (2020). Unveiling the ecological applications of ancient DNA from mollusk shells. *Front. Ecol. Evol.* 8, 37. doi:10.3389/fevo.2020.00037

Galinsky, K., and Mah, M. (2022). *EIG: eigen tools by nick patterson and alkes price lab*. Available at: https://github.com/DReichLab/EIG/releases/tag/v8.0.0.

Gopalakrishnan, S., Ebenesersdottir, S. S., Lundstrøm, I. K., Turner-Walker, G., Moore, K. H., Luisi, P., et al. (2022). The population genomic legacy of the second plague pandemic. *Curr. Biol.* 32, 4743–4751.e6. doi:10.1016/j.cub.2022.09.023

Hajdinjak, M., Fu, Q., Hübner, A., Petr, M., Mafessoni, F., Grote, S., et al. (2018). Reconstructing the genetic history of late Neanderthals. *Nature* 555, 652–656. doi:10.1038/nature26151

Harney, É., May, H., Shalem, D., Rohland, N., Mallick, S., Lazaridis, I., et al. (2018). Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation. *Nat. Commun.* 9, 3336. doi:10.1038/s41467-018-05649-9

Huang, W., Li, L., Myers, J. R., and Marth, G. T. (2012). ART: a next-generation sequencing read simulator. *Bioinformatics* 28, 593–594. doi:10.1093/bioinformatics/btr708

John, J. S. (2016). *SeqPrep: tool for stripping adaptors and/or merging paired reads with overlap into single reads*. Available at: https://github.com/jstjohn/SeqPrep/releases/tag/v1.2.

Kehlmaier, C., Barlow, A., Hastings, A. K., Vamberger, M., Paijmans, J. L., Steadman, D. W., et al. (2017). Tropical ancient DNA reveals relationships of the extinct Bahamian giant tortoise Chelonoidis alburyorum. *Proc. R. Soc. B Biol. Sci.* 284, 20162235. doi:10.1098/rspb.2016.2235

Kehlmaier, C., Graciá, E., Ali, J. R., Campbell, P. D., Chapman, S. D., Deepak, V., et al. (2023). Ancient DNA elucidates the lost world of western Indian Ocean giant tortoises and reveals a new extinct species from Madagascar. *Sci. Adv.* 9, eabq2574. doi:10.1126/sciadv.abq2574

Köster, J., and Rahmann, S. (2012). Snakemake-a scalable bioinformatics workflow engine. *Bioinformatics* 28, 2520–2522. doi:10.1093/bioinformatics/bts480

Krause-Kyora, B., Nutsua, M., Boehme, L., Pierini, F., Pedersen, D. D., Kornell, S.-C., et al. (2018a). Ancient DNA study reveals HLA susceptibility locus for leprosy in medieval Europeans. *Nat. Commun.* 9, 1569. doi:10.1038/s41467-018-03857-x

Krause-Kyora, B., Susat, J., Key, F. M., Kühnert, D., Bosse, E., Immel, A., et al. (2018b). Neolithic and medieval virus genomes reveal complex evolution of hepatitis B. *eLife* 7, e36666. doi:10.7554/elife.36666

Lazaridis, I., Patterson, N., Mittnik, A., Renaud, G., Mallick, S., Kirsanow, K., et al. (2014). Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 513, 409–413. doi:10.1038/nature13673

Li, H. (2018). *Seqtk: toolkit for processing sequences in FASTA/Q formats*. Available at: https://github.com/lh3/seqtk/releases/tag/v1.3.

Li, H. (2019). *adna: processing WGS aDNA data using the ReichLab protocol*. Available at: https://github.com/DReichLab/adna/tree/c80e94aba2df837310bfc29d39f1f93461fe2e71.

Li, H., and Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics* 26, 589–595. doi:10.1093/bioinformatics/btp698

Lipson, M., Sawchuk, E. A., Thompson, J. C., Oppenheimer, J., Tryon, C. A., Ranhorn, K. L., et al. (2022). Ancient DNA and deep population structure in sub-Saharan African foragers. *Nature* 603, 290–296. doi:10.1038/s41586-022-04430-9

Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., et al. (2012). Comparison of next-generation sequencing systems. *J. Biomed. Biotechnol.* 2012, 1–11. doi:10.1155/2012/251364

Mardis, E. R. (2008). The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24, 133–141. doi:10.1016/j.tig.2007.12.007

Margaryan, A., Lawson, D. J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., et al. (2020). Population genomics of the Viking world. *Nature* 585, 390–396. doi:10.1038/s41586-020-2688-8

Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A. V., Mikheenko, A., et al. (2022). The complete sequence of a human genome. *Science* 376, 44–53. doi:10.1126/science.abj6987

Orlando, L., Allaby, R., Skoglund, P., Der Sarkissian, C., Stockhammer, P. W., Ávila-Arcos, M. C., et al. (2021). Ancient DNA analysis. *Nat. Rev. Methods Prim.* 1, 14. doi:10.1038/s43586-020-00011-0

Ozga, A. T., Nieves-Colón, M. A., Honap, T. P., Sankaranarayanan, K., Hofman, C. A., Milner, G. R., et al. (2016). Successful enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus. *Am. J. Phys. Anthropol.* 160, 220–228. doi:10.1002/ajpa.22960

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., et al. (2016). EAGER: efficient ancient genome reconstruction. *Genome Biol.* 17, 60–14. doi:10.1186/s13059-016-0918-z

Pérez, V., Liu, Y., Hengst, M. B., and Weyrich, L. S. (2022). A case study for the recovery of authentic microbial ancient DNA from soil samples. *Microorganisms* 10, 1623. doi:10.3390/microorganisms10081623

Renaud, G., Hanghøj, K., Willerslev, E., and Orlando, L. (2017). gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 33, 577–579. doi:10.1093/bioinformatics/btw670

Renaud, G., Stenzel, U., and Kelso, J. (2014). leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res.* 42, e141. doi:10.1093/nar/gku699

Schiffels, S. (2022). *sequenceTools*. Available at: https://github.com/stschiff/sequenceTools/releases/tag/v1.5.2.

Schubert, M., Lindgreen, S., and Orlando, L. (2016). AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. notes* 9, 88. doi:10.1186/s13104-016-1900-2

Sirak, K., Fernandes, D., Cheronet, O., Harney, E., Mah, M., Mallick, S., et al. (2020). Human auditory ossicles as an alternative optimal source of ancient DNA. *Genome Res.* 30, 427–436. doi:10.1101/gr.260141.119

Slon, V., Clark, J. L., Friesem, D. E., Orbach, M., Porat, N., Meyer, M., et al. (2022). Extended longevity of DNA preservation in levantine paleolithic sediments, sefunim cave, Israel. *Sci. Rep.* 12, 14528. doi:10.1038/s41598-022-17399-2

Spyrou, M. A., Tukhbatova, R. I., Wang, C.-C., Valtueña, A. A., Lankapalli, A. K., Kondrashin, V. V., et al. (2018). Analysis of 3800-year-old *Yersinia pestis* genomes suggests Bronze Age origin for bubonic plague. *Nat. Commun.* 9, 2234. doi:10.1038/s41467-018-04550-9

Teschler-Nicola, M., Fernandes, D., Händel, M., Einwögerer, T., Simon, U., Neugebauer-Maresch, C., et al. (2020). Ancient DNA reveals monozygotic newborn twins from the Upper Palaeolithic. *Commun. Biol.* 3, 650. doi:10.1038/s42003-020-01372-8

Westbury, M. V., and Lorenzen, E. D. (2022). Iteratively mapping ancient DNA to reconstruct highly divergent mitochondrial genomes: an evaluation of software, parameters and bait reference. *Methods Ecol. Evol.* 13, 2419–2428. doi:10.1111/2041-210x.13990

Weyrich, L. S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., et al. (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* 544, 357–361. doi:10.1038/nature21674

Wibowo, M. C., Yang, Z., Borry, M., Hübner, A., Huang, K. D., Tierney, B. T., et al. (2021). Reconstruction of ancient microbial genomes from the human gut. *Nature* 594, 234–239. doi:10.1038/s41586-021-03532-0

Zavala, E. I., Thomas, J. T., Sturk-Andreaggi, K., Daniels-Higginbotham, J., Meyers, K. K., Barrit-Ross, S., et al. (2022). Ancient DNA methods improve forensic DNA profiling of Korean War and World War II unknowns. *Genes.* 13, 129. doi:10.3390/genes13010129

Zimmermann, H. H., Stoof-Leichsenring, K. R., Dinkel, V., Harms, L., Schulte, L., Hütt, M.-T., et al. (2023). Marine ecosystem shifts with deglacial sea-ice loss inferred from ancient DNA shotgun sequencing. *Nat. Commun.* 14, 1650. doi:10.1038/s41467-023-36845-x

Zook, J. M., Catoe, D., McDaniel, J., Vang, L., Spies, N., Sidow, A., et al. (2016). Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data* 3, 160025–160026. doi:10.1038/sdata.2016.25