Removing reference bias in ancient DNA data analysis by mapping to a sequence variation graph

Rui Martiniano[1*], Erik Garrison[2,3*], Eppie R. Jones[4], Andrea Manica[4], Richard Durbin[1,2#]

September 26, 2019

1. Department of Genetics, University of Cambridge, Cambridge CB3 0DH, UK
2. Wellcome Sanger Institute, Cambridge CB10 1SA, UK
3. Genomics Institute, University of California, Santa Cruz, CA 95064, USA
4. Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK
* Equal contribution
# Correspondence to richard.durbin@gen.cam.ac.uk

Abstract

Background: During the last decade, the analysis of ancient DNA (aDNA) sequence has become a powerful tool for the study of past human populations. However, the degraded nature of aDNA means that aDNA sequencing reads are short, single-ended and frequently mutated by post-mortem chemical modifications. All these features decrease read mapping accuracy and increase reference bias, in which reads containing non-reference alleles are less likely to be mapped than those containing reference alleles. Recently, alternative approaches for read mapping and genetic variation analysis have been developed that replace the linear reference by a variation graph which includes all the alternative variants at each genetic locus. Here, we evaluate the use of variation graph software `vg` to avoid reference bias for ancient DNA.

Results: We used `vg` to align multiple previously published aDNA samples to a variation graph containing 1000 Genome Project variants, and compared these with the same data aligned with `bwa` to the human linear reference genome. We show that use of `vg` leads to a much more balanced allelic representation at polymorphic sites and better variant detection in comparison with `bwa`, especially in the presence of post-mortem changes, effectively removing reference bias. A recently published approach that filters `bwa` alignments using modified reads also removes bias, but has lower sensitivity than `vg`.

Conclusions: Our findings demonstrate that aligning aDNA sequences to variation graphs allows recovering a higher fraction of non-reference variation and effectively mitigates the impact of reference bias in population genetics analyses using aDNA, while retaining mapping sensitivity.

Keywords: ancient DNA, variation graph, sequence alignment, reference bias.

Background

In suitable conditions, DNA can survive for tens or even hundreds of thousands of years *ex vivo*, providing a unique window into the past history of life [1]. Since the initial application of high-throughput sequencing to ancient human remains [2], the number of aDNA samples with available sequence data has been increasing at a fast pace, and currently over 2000 ancient samples have been published [3]. These studies have provided insights into past population history and allow direct tests of hypotheses raised in archaeology, anthropology and linguistics [4, 5].

However, aDNA sequence analysis poses several significant challenges. The amount of DNA available is limited, and often only a small fraction is endogenous, coming from the target individual, with the rest originating from microbial contamination [6]. Read lengths are limited by the degradation of DNA due to taphonomic processes and subsequent environmental exposure. Post-mortem damage (PMD) of the DNA occurs at a high rate, introducing mismatches in the tails of the short DNA molecules, which are frequently in a single-stranded and relatively unprotected state. This manifests mostly as the conversion of cytosine to uracil, but also can lead to depurination [1]. Ancient DNA may be treated with uracil-DNA-glycosylase (UDG) and endonuclease VIII to fully [7] or partially [8] remove uracil residues and abasic sites, leaving undamaged portions of the DNA fragments intact. However, this process results in a reduction of read length and library depth, which is disadvantageous. Furthermore, a number of unique and irreplaceable samples were sequenced prior to the adoption of UDG treatment. Taking all these factors into account, ancient DNA data is generally of low coverage, short length and high intrinsic error rate.

The typical workflow for ancient DNA data processing starts with the alignment of sequencing reads to a linear reference genome, which contains only the reference allele at polymorphic sites.

Reads containing the alternate allele are less likely to map than reads containing the reference allele, creating a potentially strong bias against non-reference variation, which can have a significant effect on population genetic inference and implications for many aDNA studies [9, 10]. For example, a standard approach to genotyping is to generate pseudo-haploid calls by selecting a random read crossing each variable site. However, because of reference bias, at heterozygous sites reads containing the reference allele compose the majority of reads, resulting in a more frequent sampling of the reference allele than the alternate one.

There have been previous attempts to mitigate the effects of reference bias and low coverage in aDNA, such as by implementing a model of reference bias in genotyping [11], or by working with genotype likelihoods throughout all downstream population genetic analyses [12]. A recently proposed approach modifies both the reference genome and the aDNA sequencing reads in order to account for alternate alleles at polymorphic sites[10]. The authors show that by taking into account non-reference variation in the alignment process, reference bias can be substantially reduced. However, a limitation of this approach is that it has only considered biallelic single nucleotide polymorphisms (SNPs). Therefore, non-reference alleles at insertion and deletion (indel) loci are not accounted for, despite there being hundreds of thousands of non-reference indels in a typical human genome [13], and these having a greater affect on read mapping than SNPs [14].

An alternative way to improve read mapping and avoid reference bias is to map reads to a sequence graph that represents both reference and alternate alleles at known variable sites[15]. However, the application of this approach to ancient DNA data has not yet been examined. We recently introduced the variation graph (`vg`) software[14], and in Figure 1 show an example of how `vg` can recover the alignment of short aDNA reads to alternate alleles. Here we apply `vg` and `bwa` systematically to map both simulated data and 34 previously published ancient human DNA samples, and demonstrate that mapping with `vg` can effectively reduce reference bias for ancient DNA samples in comparison to `bwa`. Furthermore `vg` increases sensitivity for detection of variation in aDNA, unlike read modification methods.

Results

Evaluating reference bias in aDNA using simulation

First we used simulation to examine the impact of post-mortem deamination (PMD) in `vg` and `bwa` read alignment, including assessments after applying read modification as in reference [10]. We generated all possible 50 bp reads spanning variant sites on chromosome 11 of the Human Origins SNP panel [16, 17], which contains a set of SNPs designed to be highly informative about the genetic diversity in human populations. In half of the simulated reads the SNP position was mutated to the alternate allele. Different levels of ancient DNA PMD estimated in 102 ancient genomes from [18] were introduced in the reads using `gargammel` [19].

We generated a variation graph (1000GP graph) with variants identified as part of the phase 3 of the 1000 Genomes Project [13] above 0.1% MAF (minor allele frequency), to be used for read mapping with `vg`. We then mapped simulated sequences back to the 1000GP graph or GRCh37 linear genome using `vg map` and `bwa aln` respectively, and filtered the resulting alignments for those above mapping quality 30 for `bwa` aligned reads and mapping quality 50 for `vg` (see Supplementary Figure S1 and Methods for details). The reason for using different mapping quality thresholds is that mapping qualities are estimated differently in `bwa` and `vg`, and have different ranges: `bwa aln`'s maximum values are capped at 37 and `vg`'s at 60.

At high levels of simulated PMD, alignment with `bwa` against the linear reference prevents the observation of non-reference alleles in a large fraction of cases (Figure 2). This effect is notable at deamination rates as low as 10%, and with 30% deamination the rate of alignment to non-reference alleles is reduced by nearly 15% relative to the total. In contrast there is no such reduction for `vg map`. These observations are maintained across a range of different mapping quality thresholds (Supplementary Figure S2). Given that we simulated the same number of reads at each SNP site, one with the reference allele and the other with the alternate, we would expect a 1:1 ratio of alternate reads to reference reads in the final alignment. However, because of reference bias, this ratio is on average 0.9335 in `bwa` but essentially 1 in `vg` (0.9995), supporting that `vg` alignment is not affected by reference bias in the same way as `bwa`.

To discern whether these differences are due to the use of a variation graph or the `vg` mapper, we also aligned simulated reads with `vg` to the human linear reference genome GRCh37 ('vg linear') and compared the results obtained with `vg` alignments to the 1000GP graph (Supplementary Figures S2 and S3). In the `vg` alignment to the linear reference, the fraction of reads containing the reference allele that are aligned remains constant at increasing rates of deamination, while, similarly to `bwa`, the percentage of aligned reads with the alternate allele drops as deamination increases. We also applied the read modification protocol of Günther et al.[10] to our `bwa` mapping data. In this case, the bias is removed, but at the cost of a substantial decrease in sensitivity for reads containing the reference as well as alternate alleles (Supplementary Figure S3).

Considering that `vg` mapped 2-18% more reads than `bwa`, we next examined the error rates of the various alignment strategies. We considered a given read to be correctly mapped if there was an exact match between the genomic coordinates from which it had been simulated and the ones for a major part of the alignment, taking into account any offsets introduced by insertions, deletions and soft clips. (Soft clipping is the masking of a number of bases at the end of the read where they appear to be diverging significantly from the reference; this is done by read aligners to avoid misalignments around insertions and deletions, or problems with chimaeric sequences.) In `vg` graph alignments, 0.00012% of reference allele reads and 0.00025% of the alternate allele reads were incorrectly mapped. In `bwa` alignments, we observed that 0.00002% and 0.00024% of reads containing the reference and the alternate allele, respectively, were incorrectly mapped. With `vg` alignment to the linear reference sequence, reads containing the reference allele were mapped with similar accuracy to that observed in `vg` graph (0.00013%), but the error in the alignment of reads with the alternate allele was an order of magnitude higher (0.00110%) (Table S1, Supplementary Figure S4). The `bwa` read modification approach only removes excess reference-allele reads, so does not change the false positive rates for reads containing alternative alleles.

Error rates in all three of `vg` graph, `vg` linear and `bwa` were positively correlated with the amount of deamination (Supplementary Figure S5). There appears to be a qualitative difference between the types of errors made by `vg` and `bwa`, in that `bwa` makes more scattered errors that are typically long range, whereas `vg` tends to make clusters of errors at nearby locations. Unsurprisingly, the majority of errors (≈70%) made by both methods occur in regions of reduced mappability (Supplementary Figure S6).

Together, the results of our analysis of simulated data demonstrate that the high degree of reference bias in ancient DNA read alignment is mitigated at known sites by aligning against a variation graph with `vg`. Although read modification in `bwa` also removes bias, it does this at the cost of decreasing sensitivity for reads containing the reference allele, whereas `vg` increases the sensitivity for reads containing the alternate allele. This increase in `vg`'s sensitivity in mapping reads containing the alternate allele is achieved at comparable error rates to those observed with

bwa, although there is a slight decrease in accuracy in mapping the reference allele.

Aligning ancient samples to the 1000GP variation graph

To evaluate whether the improvements seen in the previous section mitigate reference bias in real ancient DNA data, we selected 34 previously published ancient DNA samples (Table 1), including Iron Age, Roman, and Anglo-Saxon period samples shotgun sequenced to low-medium coverage [20, 21], high-coverage Yamnaya and Botai culture individuals [22], and target captured samples from South America [23]. These are representative of the different types of data produced in the field of aDNA, as they are of variable genomic coverage, they were generated as part of SNP array target capture or whole-genome shotgun sequencing experiments, and were subject to different enzymatic treatments.

First we evaluated the effect of using vg on standard quality control metrics used in aDNA analysis, using mapping quality threshold 50 for vg and 30 for bwa as above, except where stated otherwise. When using ANGSD to estimate sample contamination from the X chromosome of male samples, vg gave similar but marginally increased values (mean 0.95%, range 0.30-3.36%) compared to bwa (0.87%, range 0.26-3.32%) (Supplementary Figure S7). In terms of total endogenous DNA percentage, vg gave slightly lower percentages (Table S2), though as we will see below, more reads are mapped to alternate alleles. Finally, reads mapped with vg continue to show terminal deamination damage, which is used as a standard diagnostic for the presence of true ancient DNA, as seen in mapDamage [24] plots, although levels are slightly reduced (Supplementary Figure S8). We attribute this reduction to differences in softclipping by the vg algorithm, which follows bwa mem not bwa aln.

To investigate the effect of using vg and a variation graph on genetic variant calling and genotyping, we focused on the Yamnaya sample from reference [22], which provides approximately 20-fold coverage of the genome, thus allowing us to compare results to confident genotype calls, and to downsample to explore behaviour at different sequencing depths. We called variants on the full depth sample using bcftools [25] for both vg and bwa alignments, and used these callsets as ground truth. Looking at high quality heterozygous transversion sites, vg has an alternate allele mapped read fraction of 0.4925 compared to bwa of 0.4742. As expected, this difference was entirely due to mapping to previously identified 1000 Genomes Project sites present in the graph: new sites not in the graph showed no difference between the methods (Supplementary Figure S9). The restriction to transversions for this analysis is a standard approach in aDNA analysis to control for noise created by deamination damage, which generates apparent transitions.

We next measured our ability to recover the heterozygous variants in the full coverage set at lower coverage levels. As seen in Figure 3, when calling using bcftools, bwa aln recovers fewer heterozygous SNPs than vg map alignment to the 1000GP graph at all coverage levels. For example at 4x coverage, vg map recovers ≈13% more heterozygotes as a fraction of the total. Filtering bwa alignments using read modification reduces sensitivity still further. We note that if pseudo-haploid calls were made by selecting a random spanning read as is often done in aDNA analysis [26], then the allele imbalance described in the previous paragraph will directly lead to undercalling of alternate alleles.

Population genetics analyses

In order to evaluate the consequences of reference bias, we applied the ABBA-BABA test of phylogenetic tree topology based on the D-statistic of population relationship [27, 16]. When estimating D-statistics of the form D(vg, bwa; GRCh37, Chimp), a deviation from zero indicates an excess of shared alleles between `bwa`- or `vg`- aligned samples and the GRCh37 reference genome. Our results based on pseudo-haploid random-allele calls, summarized in Figure 4a, show negative D-statistics for all but a handful of samples, consistent with `bwa` calls being closer to the reference than `vg` calls. The positive D scores obtained were for samples with very low coverage, and therefore their D-statistic result is less reliable, as demonstrated by the wide error bars for these samples. Indeed, we observe some variation for these samples in the Supplementary Material, where we included 5 replicates per ancient sample to account for the randomness in pseudo-haploid genotype generation (Supplementary Figures S10-S13). When we applied the read modification approach to the `bwa`-mapped data, we saw no consistent deviation from zero in D(vg, bwa-modreads; GRCh37, Chimp), as expected from our earlier results (Supplementary Figure S14).

The bias in D statistic between `vg` and `bwa` is strongest for shorter reads, as seen when we stratify the data by read length as in Figure 4b (Supplementary Figures S15 and S16), consistent with a previously reported increase in reference bias for shorter reads[10]. It is also stronger at higher levels of deamination in our simulated data (Supplementary Figures S17 and S18). Together these results indicate that `vg` offers an advantage in mapping both shorter fragments and also fragments containing high rates of deamination.

We also investigated the effect of `vg` or `bwa` alignment on Principal Component Analysis (PCA), another widely used analysis technique in the field of aDNA. Restricting this analysis to samples from Europe and West/Central Asia, we projected the ancient samples and the reference genome onto a PCA plot derived from modern samples. We observe modest differences between the positions of `vg` and `bwa` aligned samples, but these are not conclusive in terms of the direction of the bias (Supplementary Figures S19 and S20). For example, the `bwa` processed Botai sample appears to be slightly closer to the reference than its `vg` aligned equivalent, while the opposite pattern is observed for the Yamnaya sample. Given the variability in our PCA results, it is not possible to make strong conclusions about the effects of removing reference bias on PCA projection.

Discussion

The analysis of highly fragmented and damaged ancient DNA sequence data is challenging and subject to reference bias, leading to a relative under-representation of alternate alleles at polymorphic sites. The consequences of this in downstream analysis can be real but quite subtle, as has been noted before [10], and we have seen in our results. Here we have shown that `vg` can be used to effectively remove this reference bias, especially in the presence of post-mortem damage and when dealing with short sequences, both inherent characteristics of aDNA data.

Although other methods have recently been introduced to address the same problem[10], these have the consequence of reducing sensitivity of mapping of reads containing the reference allele, rather than increasing the sensitivity of reads containing the alternate allele. In contrast mapping with `vg` increases sensitivity, and this is achieved while maintaining specificity of mapping, and avoiding increased error rate. Therefore the `vg` approach retains more information, which can be important at low read coverage (Figures 3 and Supplementary Figure S3).

One complication of our analysis is that mapping qualities are not directly comparable between

`vg` and `bwa` aln. Because of this, we presented comparisons between `vg` and `bwa` at different mapping quality filter thresholds. More generally, we note that different mapping programs and parameters, and different procedures for data pre-processing such as adapter trimming, or the imposition of a minimum read length threshold prior alignment, will all affect how ancient samples compare to each other. For any given analysis, it is important to standardize these settings and to remap all ancient samples in the same way to reduce spurious findings.

One possible concern with the use of `vg` as proposed is that it depends on a reference graph constructed from present day human variation. For modern human samples from the last 40,000 years, this is not a major issue, since almost all common variation is shared with extant populations on that time frame, although there will be increased private rare variation in very divergent samples such as Ust-Ishim [28]. However this approach would not be appropriate for samples from archaic populations such as Neanderthals and Denisovans, for which we do not yet have substantial collections of genetic variation. Introgressed material from archaic humans within modern humans can provide a partial source of information on genetic variation in those parts of the genome where it persists, but for graph alignment approaches to work effectively across the whole genome these samples we will have to wait until sufficient archaic genomes have been sequenced to high depth to enable construction of a representative archaic variation graph. A related advantage of working with graph genomes is that as multiple independently assembled human genomes are added into the reference variation graph, we will be able to assign ancient DNA sequence to human sequences not in the current reference graph, which are currently hidden from standard analyses of ancient DNA.

Methods

Datasets and sequence data processing

In order to compare read mapping between `vg` and `bwa`, we compiled a dataset of sequencing reads from previously published ancient individuals (Table 1). Adapter trimming was done with AdapterRemoval [29] for [22] and [30] for [20]. Unaligned fastq data from the other two datasets [21, 23] were already provided with trimmed adapters. We aligned trimmed reads to the human linear reference genome (hs37d5) using `bwa` aln [25] with parameters -l1024 -n 0.02 [31], keeping bases with quality above or equal to 15. We constructed the index file for `vg` [14] with hs37d5 and variants from the 1000 Genomes Project phase 3 dataset [13] above 0.1% MAF. In total, the graph contained 27,485,419 SNPs, 2,662,263 indels and 4,753 other small complex variants. Trimmed reads were aligned to the variation graph using `vg` map with parameters '–surject-to bam' and '-k 15'. Duplicate reads were removed with sambamba markdup [32] using the '–remove-duplicates' parameter. BAM files were subsequently filtered with samtools view [25], selecting reads with different mapping qualities thresholds (`bwa` and `vg`: mapQ > 0; ≥ 30; `vg` only: ≥ 50; ≥ 60). The reason for using different mapping quality thresholds is that `bwa` uses a different mapping quality estimation process with maximum around 37 than `vg` with maximum 60. Coverage was estimated with qualimap [33] bamqc utility. We present read number, endogenous DNA content and coverage for samples aligned with `vg` and `bwa` in Table S2.

Simulations

We simulated all possible reads overlapping chromosome 11 SNPs in the Human Origins dataset [16, 17]. In half of the simulated reads, the alternate allele was introduced. We then added different levels of deamination into simulated reads using `gargammel` [19], based on empirically estimated post-mortem damage in a dataset of 102 ancient genomes [18]. We aligned these simulated reads to the 1000GP graph with `vg` or to the linear human reference genome (GRCh37) with `bwa` and `vg` (here referred to as 'vg linear'). Read mapping with `vg` took approximately four (2.12-7.57) times longer than with `bwa`. The resulting alignments were sorted with sambamba sort, converted to bam with samtools view, and filtered with different mapping qualities thresholds (`bwa` and `vg`: mapQ > 0; ≥ 30; `vg` only: ≥ 50; ≥ 60). We estimated read alignment accuracy by comparing the genome coordinate from where each read originates and the coordinate obtained after mapping, accounting for offsets between these caused by softclips, deletions and insertions. Circos plots were plotted using [34].

Authenticity and contamination estimates

Post-mortem deamination plots were generated with mapDamage v2 [24], sampling one million reads per sample. We estimated X-chromosome contamination in all male samples with ANGSD [35], with the parameters '-r X:5000000-154900000 -doCounts 1 -iCounts 1 -minQ 20' and using polymorphic sites identified in the HapMap Project.

Variant calling and population genetics analyses

For population genetics analyses, we used the Human Origins dataset distributed with Lazaridis et al. [36]. In order to estimate D-statistics and Principal Component Analyses, we generated

pileups for each individual [25] at 1233553 SNPs from the Human Origins dataset using samtools mpileup, imposing a minimum base quality filter of q20. We note that pileups were generated from `bwa` aligned bam files filtered for mapping quality 30 and `vg` for mapping quality 50. We generated pseudo-haploid genotypes by randomly sampling one allele at each SNP site and converted resulting pseudo-haploid genotypes to PLINK format using PLINK 1.9 [37]. These were subsequently merged with the Chimp and Href (the human reference genome) samples from the Human Origins dataset and converted to eigenstrat format using convertf. We estimated D-statistics with qpDstat [16], passing the parameter 'printsd: YES' to obtain standard deviation estimates.

For Principal Component Analysis, we first filtered the Human Origins dataset, removing variants with minor allele frequency below 0.02 and genotyping missingness of 0.05, and selecting West Eurasian individuals. We merged this dataset with the pseudo-haploid genotypes belonging to the ancient samples as described above and ran smartpca [38, 39], restricting the analysis to transversion SNPs, using the parameters 'lsqproject: YES' to project ancient samples into the PCA coordinates estimated with present-day populations, 'killr2: YES' to exclude SNPs in high linkage disequilibrium (r2thresh: 0.2) and performing two iterations for outlier removal (numoutlieriter: 2).

Downsampling experiment

We downsampled `bwa` and `vg` alignments belonging to the high-coverage Yamnaya individual from 1 to 10x using samtools. We then called 1,054,447 biallelic snps present in the 1000 Genomes chr21 VCF from all alignments using bcftools v. 1.8, requiring a base quality of at least 20. From the resulting variant calls, we kept only biallelic SNPs and selected heterozygous genotypes. We removed potential deamination SNPs and excluded variant calls with quality score below 30. Finally, we estimated the proportion of variants correctly recovered by comparing the genotypes obtained from the downsampled alignments with those obtained at full coverage. Comparison with the read modification method was done by modifying the downsampled and full coverage `bwa`-aligned reads with the 1000 Genomes SNP alleles and calling variants as described above.

Alternate allele support and allele balance

In order to compare alternate allele support between `vg` and `bwa` alignments, we called chromosome 1 SNPs from the Yamnaya alignemnts with bcftools. We then filtered these by variant quality greater or equal than 30, with depth of coverage above 8, and selected heterozygous variants. From these genotype calls, we obtained reference and alternate allelic depth and compared alternate allele support between the `vg` and `bwa` aligned sample.

Comparison with an alternative method for reducing reference bias

We compared `vg` with the workflow proposed by [10] to reduce reference bias. The following method was applied to both real and simulated data. First, for each `bwa`-aligned sample, we selected reads overlapping with the Human Origins SNPs. We then modified the allele in these reads using the 'modify_read_alternative.py' script, distributed with [10], and remapped them with `bwa` to GRCh37 as described above. We then kept the original reads which mapped to the the same location of the modified reads with 'filter_sam_startpos_dict.py'. We estimated D-statistics from the resulting filtered alignments as described above.

List of abbreviations

aDNA - ancient DNA
MAF - minor allele frequency
PCA - Principal Component Analysis
UDG - uracil-DNA-glycosylase
PMD - Post-mortem damage
1000GP - 1000 Genomes Project
mapQ - mapping quality
OLS - ordinary least squares
bp - basepairs

Figures and Tables

Table 1: Datasets analysed in the present study.

Dataset	Number of individuals	Genomic coverage	Treatment	Type	Region
Damgaard et al. 2018	2	11.24-18.95x	untreated	whole-genome shotgun sequencing	Eurasian Steppe
Martiniano et al. 2016	9	0.54-1.63x	untreated	whole-genome shotgun sequencing	England
Shiffels et al., 2016	11	0.47-7.86x	partial UDG/USER	whole-genome shotgun sequencing	England
Posth et al. 2018	13	0.02-0.40x	partial UDG	target capture	Americas

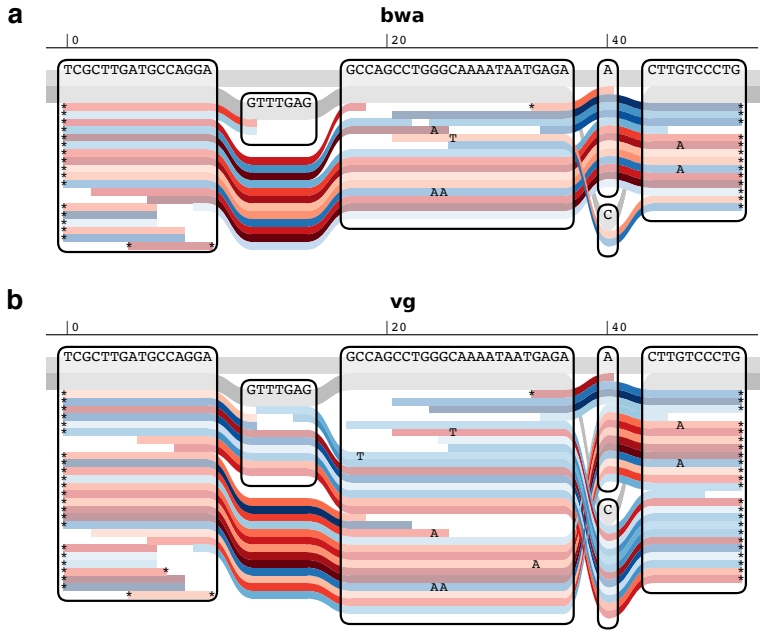

Fig. 1: Sequence tube maps [40] of a small region of the human genome with aDNA reads from the Yamnaya individual aligned with A) `bwa aln` to a linear reference sequence and B) `vg map` to a graph containing 1000 Genomes variants. The individual is heterozygous for both an indel (GTTTGAG/-) and a SNP (A/C) in this region, with insertion and alternate allele on the same haplotype. The two underlying haplotypes in this region are coloured in grey, and red and blue lines indicate forward and reverse reads respectively. None of the 6 reads across the insertion and only 2 of 12 reads across the SNP were mapped by `bwa`. Reads were locally realigned with `vg map` to the graph for the purpose of visualization.

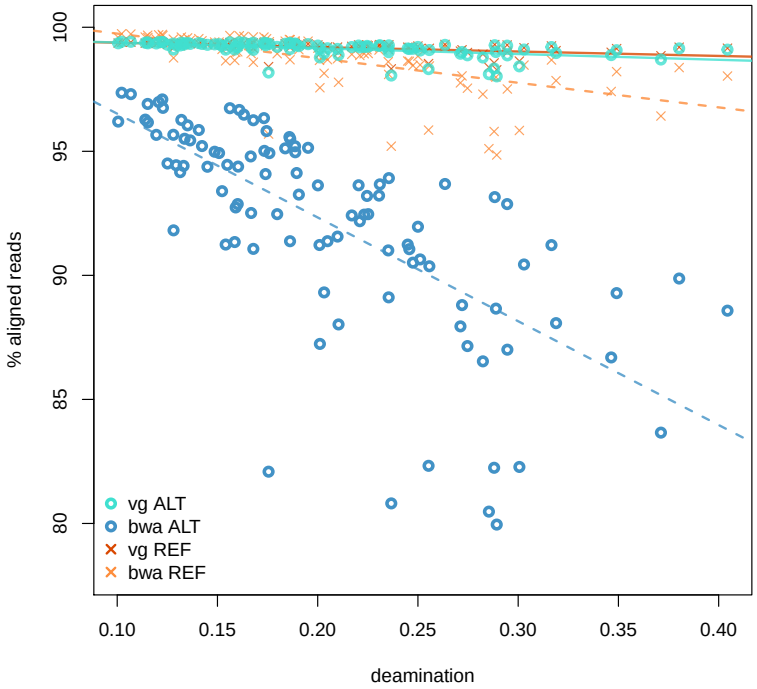

Fig. 2: Comparing `bwa aln` (mapping quality 30) and `vg map` (mapping quality 50) performance when aligning reads simulated from chromosome 11 of the Human Origins panel. Lines represent ordinary least squares (OLS) regression results for the allele/aligner conditions corresponding to their colors.

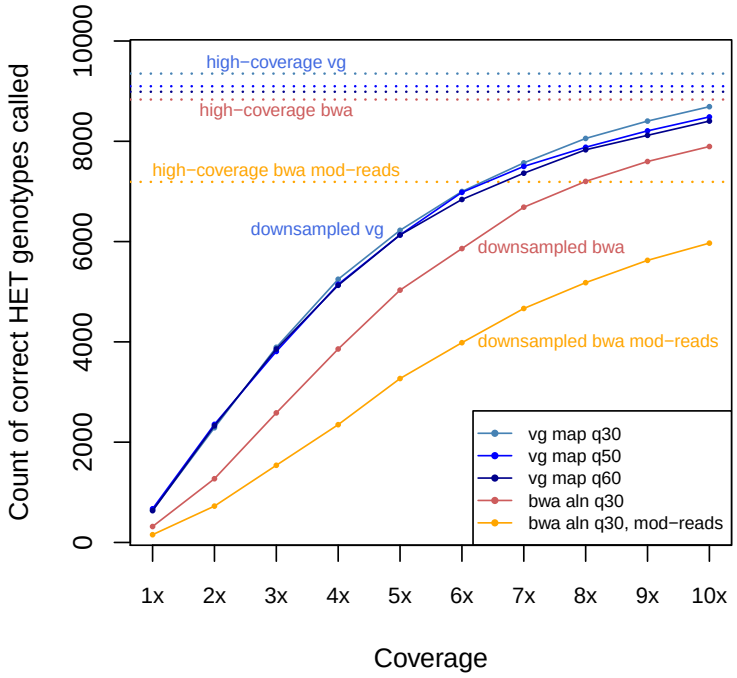

Fig. 3: The comparative effect of downsampling on heterozygous variant calling following `bwa aln` and `vg map` alignment of reads from the ancient Yamnaya sample [22], including post-processing of `bwa aln` with the modified read filter [10].

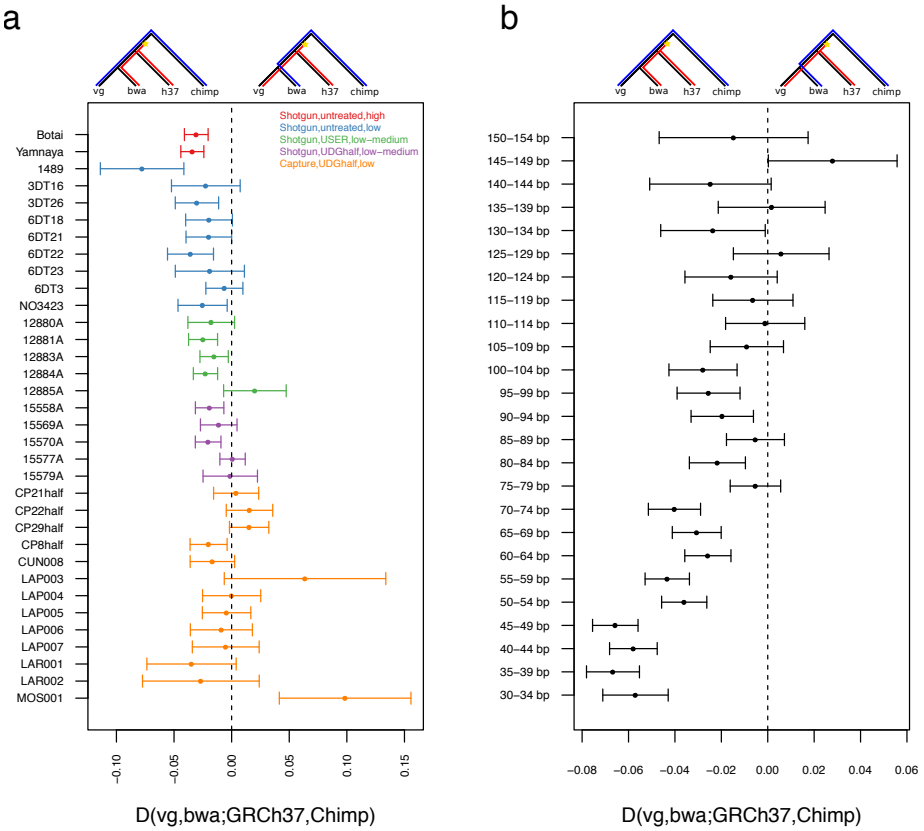

Fig. 4: D-statistic based ABBA-BABA tests of reference bias in aDNA. a) D-statistics of the form D(vg, bwa; GRCh37, Chimp) estimated for 34 ancient individuals, using transversion SNPs only. The figure label indicates the type of sequencing, enzymatic treatment and genomic coverage for each sample. b) D-statistic for the ancient Yamnaya sample stratified by read length. In both panels, negative D results indicate excess allele sharing between `bwa` alignments of the samples and GRCh37 (the ABBA pattern), while positive values indicate an excess of sharing between `vg` alignments and GRCh37 (the BABA pattern). The two patterns are illustrated by the trees above the panels, where A alleles are given as blue lines on the tree and B alleles are shown in red.

Additional Files

Supplementary Figures and Tables

Acknowledgements

We thank Cosimo Posth and Stephan Schiffels for assistance in obtaining raw sequence data, Christiana Lyn Scheib and Toomas Kivisild for sharing their results on reference bias in ancient DNA and encouraging this study, and Lara Cassidy for helpful discussions.

Funding

RM was supported by an EMBO Long-Term Fellowship (No. ALTF 133-2017), EG by a Wellcome Sanger Institute studentship under grant WT206194, and RD by Wellcome grant WT207492. ERJ was supported by a Herchel Smith Fellowship and AM by the ERC Consolidator Grant 647787 (LocalAdaptation).

Author information

Affiliations

Department of Genetics, University of Cambridge, Cambridge CB3 0DH, UK.
Rui Martiniano
Richard Durbin

Wellcome Sanger Institute, Cambridge CB10 1SA, UK.
Erik Garrison
Richard Durbin

Genomics Institute, University of California, Santa Cruz, CA 95064, USA.
Erik Garrison

Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, UK.
Eppie R. Jones
Andrea Manica

Contributions

RD conceived the work, all authors designed the experiments, analysed and interpreted the data, RM, EG and RD wrote the manuscript. All authors read and approved the final manuscript.

Corresponding author

Richard Durbin (rd109@cam.ac.uk)

Ethics approval and consent to participate

All the sequencing read data analyzed in this work was previously published.

Availability of data and material

Simulated raw sequence data generated in this study can be downloaded at Zenodo (DOI: 10.5281/zenodo.3416364). Raw sequence data for the samples analysed in this study can be found in the European Nucleotide Archive, accessions PRJEB25389 [22], PRJEB11004 [20], ERP003900 and ERP006581 [21]. FASTQ files belonging to the South American individuals were provided by the authors of [23].

Competing interests

There are no competing interests.

Consent for publication

Not applicable.

References

[1] Jesse Dabney, Matthias Meyer, and Svante Pääbo. "Ancient DNA damage". In: *Cold Spring Harbor Perspectives in Biology* (2013), a012567.

[2] Morten Rasmussen et al. "Ancient human genome sequence of an extinct Palaeo-Eskimo". In: *Nature* 463.7282 (2010), p. 757.

[3] Katherine Brunson and David Reich. "The Promise of Paleogenomics Beyond Our Own Species". In: *Trends in Genetics* (2019).

[4] Torsten Günther and Mattias Jakobsson. "Genes mirror migrations and cultures in prehistoric Europe—a population genomic perspective". In: *Current Opinion in Genetics & Development* 41 (2016), pp. 115–123.

[5] Pontus Skoglund and Iain Mathieson. "Ancient genomics of modern humans: the first decade". In: *Annual review of genomics and human genetics* 19 (2018), pp. 381–404.

[6] Mark Stoneking and Johannes Krause. "Learning about human population history from ancient and modern genomes". In: *Nature Reviews Genetics* 12.9 (2011), p. 603.

[7] Adrian W Briggs et al. "Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA". In: *Nucleic Acids Research* 38.6 (2009), e87–e87.

[8] Nadin Rohland et al. "Partial uracil–DNA–glycosylase treatment for screening of ancient DNA". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 370.1660 (2015), p. 20130624.

[9] Kay Prüfer et al. "Computational challenges in the analysis of ancient DNA". In: *Genome Biology* 11.5 (May 2010), R47. ISSN: 1474-760X. DOI: `10.1186/gb-2010-11-5-r47`. URL: `https://doi.org/10.1186/gb-2010-11-5-r47`.

[10] Torsten Günther and Carl Nettelblad. "The presence and impact of reference bias on population genomic studies of prehistoric human populations". In: *PLoS genetics* 15.7 (2019), e1008302.

[11] Kay Prüfer. "snpAD: an ancient DNA genotype caller". In: *Bioinformatics* (2018). DOI: `10.1093/bioinformatics/bty507`.

[12] Vivian Link et al. "ATLAS: analysis tools for low-depth and ancient samples". In: *bioRxiv:105346* (2017).

[13] 1000 Genomes Project Consortium et al. "A global reference for human genetic variation". In: *Nature* 526.7571 (2015), pp. 68–74.

[14] Erik Garrison et al. "Variation graph toolkit improves read mapping by representing genetic variation in the reference". In: *Nature Biotechnology* 36.9 (2018), pp. 875–879.

[15] B. Paten et al. "Genome graphs and the evolution of genome inference". In: *Genome Research* 27.5 (2017), pp. 665–676.

[16] N. Patterson et al. "Ancient Admixture in Human History". In: *Genetics* 192.3 (2012), pp. 1065–1093.

[17] Iosif Lazaridis et al. "Ancient human genomes suggest three ancestral populations for present-day Europeans". In: *Nature* 513.7518 (2014), pp. 409–413.

[18] Morten E Allentoft et al. "Population genomics of bronze age Eurasia". In: *Nature* 522.7555 (2015), pp. 167–172.

[19] Gabriel Renaud et al. "gargammel: a sequence simulator for ancient DNA". In: *Bioinformatics* 33.4 (2016), pp. 577–579.

[20] Rui Martiniano et al. "Genomic signals of migration and continuity in Britain before the Anglo-Saxons". In: *Nature Communications* 7 (2016), p. 10326.

[21] Stephan Schiffels et al. "Iron age and Anglo-Saxon genomes from East England reveal British migration history". In: *Nature Communications* 7 (2016), p. 10408.

[22] Peter de Barros Damgaard et al. "The first horse herders and the impact of early Bronze Age steppe expansions into Asia". In: *Science* 360.6396 (2018), pp. 1422–1442.

[23] Cosimo Posth et al. "Reconstructing the deep population history of Central and South America". In: *Cell* 175.5 (2018), pp. 1185–1197.

[24] Hákon Jónsson et al. "mapDamage2. 0: fast approximate Bayesian estimates of ancient DNA damage parameters". In: *Bioinformatics* 29.13 (2013), pp. 1682–1684.

[25] Heng Li et al. "The sequence alignment/map format and SAMtools". In: *Bioinformatics* 25.16 (2009), pp. 2078–2079.

[26] Pontus Skoglund et al. "Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe". In: *Science* 336.6080 (2012), pp. 466–469.

[27] Richard E Green et al. "A draft sequence of the Neandertal genome". In: *Science* 328.5979 (2010), pp. 710–722.

[28] Qiaomei Fu et al. "Genome sequence of a 45,000-year-old modern human from western Siberia". In: *Nature* 514.7523 (2014), p. 445.

[29] Mikkel Schubert, Stinus Lindgreen, and Ludovic Orlando. "AdapterRemoval v2: rapid adapter trimming, identification, and read merging". In: *BMC research notes* 9.1 (2016), p. 88.

[30] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads". In: *EMBnet. journal* 17.1 (2011), pp. 10–12.

[31] Mikkel Schubert et al. "Improving ancient DNA read mapping against modern reference genomes". In: *BMC genomics* 13.1 (2012), p. 178.

[32] Artem Tarasov et al. "Sambamba: fast processing of NGS alignment formats". In: *Bioinformatics* 31.12 (2015), pp. 2032–2034.

[33] Fernando García-Alcalde et al. "Qualimap: evaluating next-generation sequencing alignment data". In: *Bioinformatics* 28.20 (2012), pp. 2678–2679.

[34] Zuguang Gu et al. "circlize implements and enhances circular visualization in R". In: *Bioinformatics* 30.19 (2014), pp. 2811–2812.

[35] Thorfinn Sand Korneliussen, Anders Albrechtsen, and Rasmus Nielsen. "ANGSD: analysis of next generation sequencing data". In: *BMC bioinformatics* 15.1 (2014), p. 356.

[36] Iosif Lazaridis et al. "Genomic insights into the origin of farming in the ancient Near East". In: *Nature* 536.7617 (2016), p. 419.

[37] Christopher C Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *Gigascience* 4.1 (2015), p. 7.

[38] Alkes L Price et al. "Principal components analysis corrects for stratification in genome-wide association studies". In: *Nature genetics* 38.8 (2006), p. 904.

[39] Nick Patterson, Alkes L Price, and David Reich. "Population structure and eigenanalysis". In: *PLoS genetics* 2.12 (2006), e190.

[40] Wolfgang Beyer et al. "Sequence tube maps: making graph genomes intuitive to commuters". In: *Bioinformatics* (2019).