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

Rui Martiniano*, Erik Garrison*, Eppie R. Jones, Andrea Manica, Richard Durbin

*Equal contribution; contact:rm890@cam.ac.uk

UNIVERSITY OF CAMBRIDGE

Abstract

Background: During the last decade, the analysis of ancient DNA (aDNA) 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. 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 and simulated data 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.

Materials and Methods

Table 1: Datasets analysed in the present study.

Dataset	Number of individuals	Genomic coverage	Region
Damgaard et al. 2018	2	11.24-18.95x	Eurasian Steppe
Martiniano et al. 2016	9	0.54-1.63x	England
Shiffels et al., 2016	11	0.47-7.86x	England
Posth et al. 2018	13	0.02-0.40x	Americas

To compare `bwa` and and `vg`, we analysed both real data (Table 1) and simulated data. For the simulations, we generated all possible reads overlapping chromosome 11 SNPs in the Human Origins dataset, introducing the REF and ALT alleles in equal proportions, and different levels of deamination based on empirical data. Sequencing reads were aligned with `bwa` to the linear reference genome and with `vg` to the 1000 GP variation graph.

References

[1] Erik Garrison et al. *Nature Biotechnology*, 36(9):875–879, 2018.

[2] 1000 Genomes Project. *Nature*, 491(7422):56–65, 2012.

[3] Li and Durbin. *Bioinformatics*, 25(14):1754–1760, 2009.

[4] Posth et al. *Cell*, 175(5):1185–1197, 2018.

[5] Damgaard et al. *Science*, 360(6396):1422–1442, 2018.

[6] Schiffels et al. *Nature Communications*, 7:10408, 2016.

[7] Martiniano et al. *Nature Communications*, 7:10326, 2016.

[8] Günther and Nettelblad. *PLoS genetics*, 15(7):e1008302, 2019.

Representative example

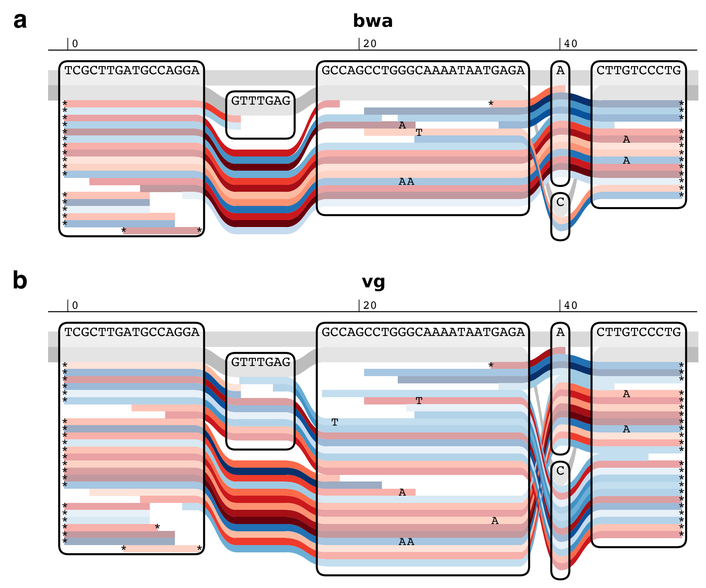

Figure 1: Sequence tube maps of a small region of the human genome with aDNA reads from the Yamnaya individual aligned with A) `bwa aln` to a linear reference sequence and B) `vg map` to a graph containing 1000 Genomes variants.

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`.

Downsampling experiment.

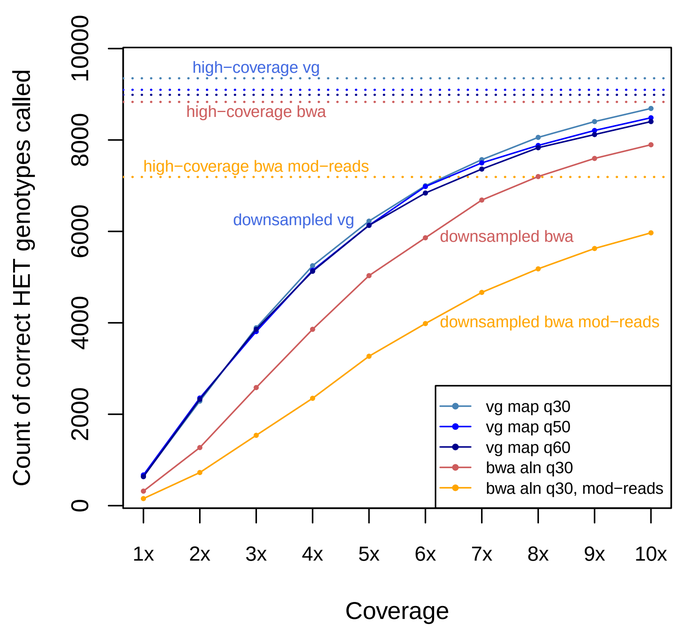

Figure 3: The comparative effect of downsampling on heterozygous variant calling following `bwa aln` and `vg map` alignment of reads from the ancient Yamnaya sample, including post-processing of `bwa aln` with the modified read filter.

We next measured our ability to recover the heterozygous variants in the full coverage set at lower coverage levels. `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 (Torsten et al., 2019) reduces sensitivity still further.

Alignment accuracy

Table 2. `vg` graph mapping to the graph (q50) and `bwa` (q30) have comparable read mapping accuracy. `vg` linear mapping is less accurate.

mapping	mean error REF (%)			mean error ALT (%)			mean error All (%)		
quality threshold	vg	vg linear	bwa	vg	vg linear	bwa	vg	vg linear	bwa
q>=30	0.00015	0.00017	0.00002	0.00037	0.00162	0.00024	0.00026	0.00089	0.00012
q>=50	0.00012	0.00013	-	0.00025	0.00110	-	0.00019	0.00061	-
q>=60	0.00011	0.00012	-	0.00021	0.00097	-	0.00016	0.00054	-

Simulated ancient DNA data

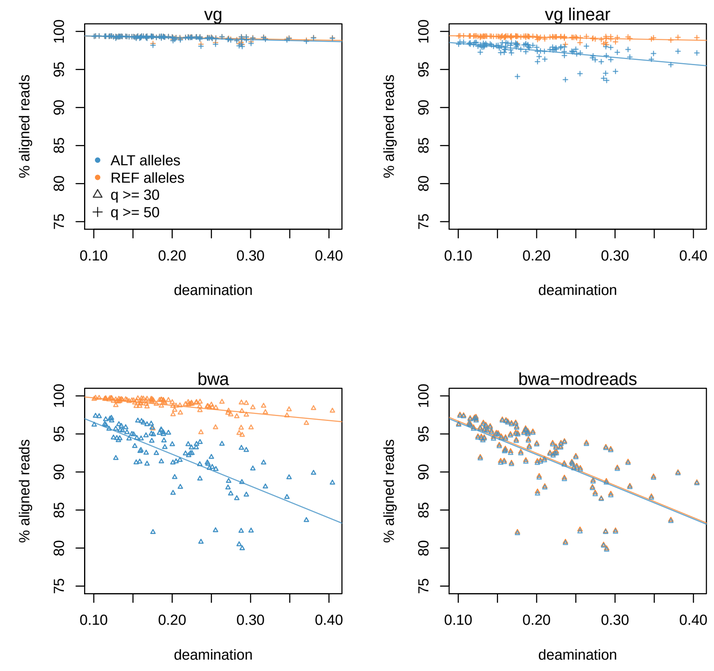

Figure 2: Comparison of the percentage of mapped reads in simulated data between bwa, bwa-modreads) (mapping quality filter q≥30), vg graph and vg linear reference (q≥50).

At high levels of simulated damage, alignment with `bwa` against the linear reference prevents the observation of non-reference alleles in a large fraction of cases. In contrast there is no such reduction for `vg map`.
In the `vg` alignment to the linear reference, similarly to `bwa`, the percentage of aligned reads with the alternate allele drops as deamination increases.
We also applied the read modification protocol of Gunther et al. 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.

D-statistics

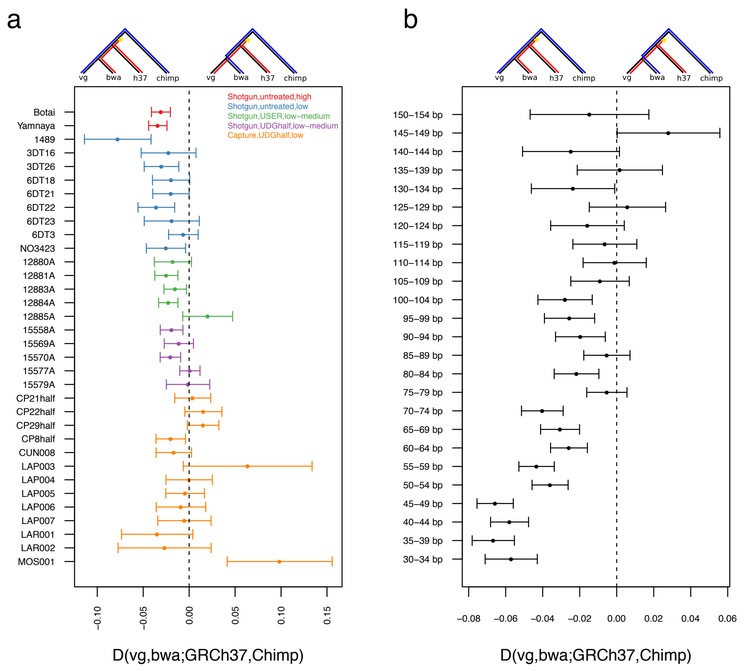

Figure 4: *D*-statistic of the form *D*(vg, bwa; GRCh37, Chimp) to test reference bias in aDNA. a) *D*-statistics estimated for 34 ancient individuals, using transversion SNPs only. The figure label indicates the type of sequencing, enzymatic treatment and genomic coverage for each sample. b) *D*-statistic for the ancient Yamnaya sample stratified by read length.

Our results show negative *D*-statistics for all but a handful of samples, consistent with `bwa` calls being closer to the reference than `vg` calls. 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 and as previously reported (Gunther, 2019).