QUANTIFYING AND REDUCING BIASES IN PALEOGENOMIC RESEARCH

A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

Adrien Oliva

The University of Adelaide
Faculty of Sciences
School of Biological Sciences
Department of Ecology and Evolutionary Biology

April 2022

Table of Contents

Thesis abstract

The emergence of high-throughput DNA sequencing technologies has enabled the sequencing of genomes at unprecedented rates and low costs. In parallel, paleogenomics research, which extends the study of ancient DNA molecules to whole genomes, has led to a number of transformative discoveries in evolutionary biology, environmental sciences, and even medicine. However, ancient DNA has a number of properties that make it challenging to investigate, including short fragment length, contamination, and damage, which are often exacerbated at genomic scales. It is now clear that numerous biases are pervasive in paleogenomics investigations, and their influence on downstream inferences is undeniable. In this thesis, I present results from a series of interrelated empirical studies that investigate issues related to reference bias and reproducibility in paleogenomics, providing the relevant historical and technical background in the introduction.

In Chapter 1 of this thesis, I benchmark a range of short read alignment methods and algorithms available to paleogenomicists and quantify the impact of reference bias on downstream inferences. I show that the current standard alignment method in the paleogenomics field, i.e., using the *BWA-aln* software with specific settings developed during the early stages of paleogenomics, is still one of the best available tools for minimising the impact of reference bias. However, reference bias can be decreased even further when using *NovoAlign* software and an augmented version of the linear reference that incorporates known variants using IUPAC characters.

In Chapter 2, I extend this investigation to include the recently developed variation graph methods to paleogenomic datasets, and assess its impact on a series of contentious population genetics inferences when compared to the two best performing traditional (linear) alignment methods identified in chapter 1. Consistent with the results from chapter 1, the added variation captured by variation graphs make them less susceptible to reference bias than linear alignments (including IUPAC augmented methods). I also show that changes in bioinformatic parameters and sample choice can lead to subtle but significant differences in statistical inferences that could impact interpretations.

Therefore, in the third chapter, I emphasise the importance of reproducibility in paleogenomic research, and make a series of recommendations regarding the minimum reported information required across all key steps of data processing and analyses to ensure reproducibility of paleogenomic results.

Finally, in the discussion chapter, I summarise my findings and discuss their implications for the field of paleogenomics as well as potential directions for future research. Ultimately, this knowledge should help improve the reliability and robustness of paleogenomic inferences, leading to an improved understanding of population history and evolutionary phenomena.

Thesis declaration

I certify that this work contains no material which has been accepted for the award of any other degree or diploma in my name in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text.

In addition, I certify that no part of this work will, in the future, be used in a submission in my name for any other degree or diploma in any university or other tertiary institution without the prior approval of the University of Adelaide and where applicable, any partner institution responsible for the joint award of this degree.

The author acknowledges that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I give permission for the digital version of my thesis to be made available on the web, via the University's digital research repository, the Library Search and also through web search engines, unless permission has been granted by the University to restrict access for a period of time.

Adrien Oliva

Date: 06/04/2022

Publications

Journals articles

Oliva A, Tobler R, Cooper A, Llamas B, Souilmi Y. Systematic benchmark of ancient DNA read mapping. Brief Bioinform. 2021 Sep 2;22(5):bbab076. https://doi.org/10.1093/bib/bbab076. PMID: 33834210.

Oliva, A., Tobler, R., Llamas, B. and Souilmi, Y. (2021), Additional evaluations show that specific *BWA-aln* settings still outperform *BWA-mem* for ancient DNA data alignment. Ecol Evol, 11: 18743-18748. https://doi.org/10.1002/ece3.8297

Nobody exists on purpose. Nobody belongs anywhere. Everybody's gonna die. Come watch TV?

Morty Smith (in Rick and Morty, "Rixty Minutes")

Acknowledgments

First, I would like to thank my supervisors, Dr. Souilmi, Dr. Tobler and Assoc. Prof. Llamas for their support during my candidature. Thanks Yassine for your technical help and for teaching me that precious basketball move. Thanks Raymond to be the best host in Adelaide; I have still so many skills to learn from you, but at least I taught you some stuffs on BL3. Finally, merci Bastien for your help in so many ways and for being supportive at any time.

I acknowledge all members of ACAD and I have enjoyed being part of the lab, which allowed me to discover how a world class research group works.
I have received the support of a large number of people during my PhD and I would like to address them a message in their own languages.

Pour commencer, je remercie les seuls qui ont été là pour m'accompagner durant ces centaines de nuits blanches : le J, Dams, PNL, Ter Ser Shem K-wayIno et évidemment Ours Samplus.

Je remercie également l'authentique Iron Man, Thibault, que je copie tous les jours un peu plus (sauf les poissons... pas fan des poissons), pour toutes ces discussions au labo et pour ce magnifique voyage solo à Cairns.

Vielen Dank für Ihre Unterstützung und unsere Gespräche im Laufe der Jahre. Sie sind einer meiner engsten Freunde geworden und ich danke Ihnen für Ihre Hilfe.

អរគុណអូនសម្រាប់ការគាំទ្របង យកចិត្តទុកដាក់ថែបង
ធ្វើឱ្យបងបានស្វែងយល់ពីវប្បធម៌របស់ប្រទេសអូន
និងរៀបចំការណាត់ជួបរបស់យើងទាំងពី
និង អរគុណអូនដែលធ្វើឱ្យបងសប្បាយចិត្តក្នុងអំឡុងឆ្នាំដ៏លំបាកនេះ!
យ៉ាងហោចណាស់ឥឡូវនេះ
ទាំងពីរអ្នកយើងដឹងថាអូនបានផ្លាស់ប្តូរបងច្រើនជាងបងទៅទៀត។

Je remercie les vrais qui sont présents depuis 10 ans maintenant, Bastien, Cyril, Damien et Guillaume, vous êtes le S. Depuis les soirées raclette sur le parking, ou bien devant la machine à sous sur BL, aux soirées LoL avec un adc en carton (c’est dur de cs quand même), jusqu’au TP de bio où l’on a étudié́ la structure physique d’une flaque d’eau.
Merci pour toutes ces soirées/ matinées passées ensemble à jouer ou discuter à en pleurer de rire (sorry Ben for waking you up late/early, that’s their fault) !
Merci à Damien pour m’avoir supporter et aider lors d’un projet commun, qui, admettons-le, ne verra jamais le jour… t’facon t’es QLF tu le sais.

Bobby, je sais que tu ne pourras jamais lire tout ça, que ce soit en français ou en anglais, mais je ne me vois pas écrire cette section sans parler de toi. Tu m’as rendu heureux depuis le premier jour et tu as eu un plus grand impact sur ma vie que n’importe qui.

Petite cacededi, à Corentin, mon frère (A.K.A Crococo tmtc)... Merci d’avoir pris soin de la famille ... et oui je te dois sûrement masse thune avec tous les cadeaux que j’ai jamais payés. Fier de toi, mais il faut que tu arrêtes de rêver, je ne viendrai pas au Canada ; vous allez venir en Australie tkt.

Enfin, merci à toute la famille et particulièrement mes parents pour votre soutien infaillible durant toutes ces années compliquées à l’autre bout du monde. Merci de nous avoir supportés et aimés coco et moi depuis toujours. Même si aujourd’hui la famille est dispersée entre le Canada, la France et l’Australie, demain, on sera tous réunis. Je pourrai écrire des paragraphes entiers, mais je ne pourrai jamais assez-vous remercier pour tout ce que vous avez fait pour nous, Je Vous Aime.

Introduction

Population genetics and the study of genetic variation

The molecular basis of inheritance in nearly all living organisms is a large composite molecule called deoxyribonucleic acid (or DNA), with the total DNA in a living organism referred to as a genome. The unique structure of DNA facilitates its replication, though this process is imperfect, leading to the accumulation of novel mutations in daughter copies over serial replication events. These mutations serve as the basis for phenotypic differences amongst individuals within the same population and, over extended periods, the more distinctive differences that distinguish different species. By enumerating mutational differences between genomes we can reconstruct the genealogical connections within populations of the same species and older phylogenetic relationships between species.

The field of population genetics studies how this genetic diversity accumulates and evolves over time [1], typically by using patterns of genetic variation in one or more populations to identify historical modes of evolutionary change, which include migration [2] and gene flow [3], genetic drift [4] and natural selection [5].

Since its inception at the turn of the 20th century [6,7], population genetics theories have been continually extended and refined [8–17] culminating in a sophisticated theoretical and empirical discipline that is a central pillar of modern evolutionary biology. Population genetics has become increasingly critical for understanding the history of our own species, with some significant impacts in present-day healthcare.

Indeed, Werner Kalow was the first in 1982 to draw attention to the existence of a heterogeneity of response to drug treatments between human populations [18]. Since then, multiple examples of differences in drug responses between individuals with different genetic ancestries have been reported in the literature [19–22]. Whether these differences are in the effectiveness of drugs or in the incidence of their side effects, it is now accepted that ancestry plays an important role in the variability of response to drugs [23,24] [25]. Knowing an individual's ancestry can inform about their likelihood of responding favourably to a drug, although it cannot predict it with certainty [26]. In June 2005, the FDA granted authorisation for a first "ethnic" drug, BiDil® [27], developed to treat heart failure specifically in African Americans. Despite the initial unconvincing results [28], a second and improved clinical trial led to FDA approval, paving the way towards personalised medicine [29].

Limits of DNA studies in present-day populations

Ongoing work has shown that reconstructing past population history from genomic data sourced solely from present-day populations may be insufficient to recover key demographic and evolutionary events, as some past events may be masked by subsequent changes or are not uniquely identifiable relative to other possible causes [30–32].

Genetic variation in modern genomes also have temporal boundaries [33,34] on their historical informativeness, which is reached when all the variants in a given population ‘coalesce’ (i.e., the time to the most recent common ancestor for a given DNA sequence [35]), and genetic variation from past lineages is often not represented in present-day genomes [36]. These factors complicate inferences of population genetic history and have led researchers to turn to studies of ancient genomes where possible to recover lost genetic lineages and help resolve competing scenarios that are not distinguishable in modern genomes.

The power of ancient DNA and the emergence of paleogenomics

Since the first ancient DNA (aDNA) molecules were recovered and analysed in the mid-1980s [37,38], aDNA research has undergone a series of technological and methodological breakthroughs on both the laboratory and computational fronts. These advances have ushered in the current paleogenomic era in the past decade, marking the transition of aDNA research from studies involving a few genetic sequences [39,40] to computationally intensive analyses involving hundreds to thousands of individual genome-wide datasets [41,42]. The paleogenomic era has revolutionised archaeological and paleontological research, providing new insights into the demographic and adaptive history of humans [31,43–47] and other species that leave a fossil record [48,49].

By accessing genetic information that can be accurately dated and contextualised with other historical and paleoenvironmental information, the study of aDNA provides researchers access to evolutionary and demographic information that may no longer be accessible due to extinction or other events (e.g., bottlenecks or interbreeding with other populations) [50–52]. Accordingly, aDNA has the potential to reveal lost genetic variation and distinguish between competing causal hypotheses [53–55] and has allowed researchers to illuminate the causes of extinction for several species [51,56–58]. In the case of human populations, the large paleogenomic datasets now available provide access to the genetic diversity of populations across time and space [59], allowing researchers to trace population migrations that resulted in the current distribution, and to support, or in some cases refute, conclusions drawn from the study of modern human DNA.

Issues arising from paleogenomic research

While undoubtedly powerful, paleogenomics is not without its pitfalls, some of which are the focus of this thesis. The majority of these problems relate to the specific proprieties of aDNA that are not apparent in modern DNA. Following the death of an organism, cellular processes maintaining DNA integrity cease to function and a succession of chemical and enzymatic reactions are activated, causing permanent alterations in DNA sequences. Those known to impact population genetic analyses of aDNA molecules are briefly summarised below.

Low DNA yields and fragmentation

In the absence of active maintenance, DNA is a fragile molecule with a limited lifespan influenced by various environmental factors , such as humidity, pH, salinity, temperature, and also microbial activity [60]. For instance, cold environments (e.g., permafrost) preserve DNA better and increase the likelihood of recovering older samples [61], which is why the majority of published genomes are from samples sourced from temperate or polar environments [62].

The continuous post-mortem degradation of DNA molecules means that only small quantities of DNA can be extracted from human remains or other repositories (such as sediments and coprolites), and the resulting fragments are typically minuscule (a few hundred bases at most). The first paleogenomic investigation of ancient DNA characteristics discovered that most of the exogenous extracted DNA (aged from

4,000 to 13,000 years) had deteriorated into fragments of 40-500 bp in length, leaving very little material to work with [63]. It has been estimated that DNA can survive for a few hundred thousand years up to a million years [64, 65].

DNA damage

Processes of chemical degradation of the DNA molecule are compensated by specialised repair mechanisms in living cells; however after death, DNA damage accumulates indefinitely [66], though the complex range of environmental and biological factors underlying the damage process means that there is no strong correlation between damage accumulation with time from death in general [67–69]

Several different DNA damage processes are known [65], including enzymatic activity triggered by DNAses (autolysis) [70] and the action of the microbes that decompose the corpse. Fragmentation without a "clean break" results in uneven ends with single-stranded "overhangs" at the ends of molecules that are sensitive to chemical reactions, such as hydrolysis and oxidation. Oxidation is the primary source of deterioration that can result in mutation, while hydrolysis leads to the loss of certain bases. Once a base is lost, the DNA molecule becomes extremely fragile and therefore has a very high risk of breaking.

Specific strategies have been developed that improve the mapping of highly damaged aDNA reads; for example, trimming the ends of reads to remove deaminated sites [71] and through the use of tools like *BWA-aln* (backtrack algorithm) [72] that have settings that allow a higher number of gaps and mismatches between the reference and the reads, allowing diverging reads to match more easily.

These degradation and damage processes result in the characteristic overrepresentation of purines (adenine and guanine) at positions immediately preceding or following strand breaks [66,73]. The accumulation of C-to-T and G-to-A substitutions at the 5' and 3’ ends of aDNA fragments, respectively, is another hallmark of aDNA damage from double-stranded aDNA libraries [66,73] (or just C-to-T substitutions at both ends in the case of single-stranded libraries [74]).

Awareness of these specific damage patterns has led to the creation of novel mitigation strategies, including laboratory methods (UDG treatment [75,76]) and a variety of in silico treatments such as the removal of bases at the ends of sequencing reads [75] (bamUtils trimBam [77]), base quality score recalibration of likely damaged positions (mapDamage [66] and ATLAS [78]), and the extraction of ancient reads identified by their damage profile (PMDtools [79]). Further, these DNA damage profiles are also widely used to authenticate aDNA molecules (e.g., DamageProfiler [80] and mapDamage [66]).

Contamination

An additional complication of ancient samples is that a significant proportion of the extracted DNA is not endogenous to that sample [81,82]. Instead, the extracted DNA often contains exogenous contamination from microorganisms from the surrounding environment [83], cross-contamination from other ancient samples, DNA from modern organisms occupying the environment the sample was collected from, and DNA from humans (e.g., handling archaeologists, museum curators) [79,84].

Even though several guidelines have been published [73,79,81,82,85–87] outlining suggestions to reduce the possibility of contamination while handling aDNA materials, some level of human contamination is an unavoidable reality of paleogenomics research as samples are handled by several individuals between death and laboratory analysis. With the help of specific software [66,79,88–91] and by analysing fragment length and aDNA damage patterns, a thorough verification of the sequences generated can help assess the authenticity of aDNA and differentiate contamination from the endogenous genetic information.

Reference bias in population genomics

Population genomic research has been enabled by the development of high throughput short-read sequencing technologies, which have allowed individual genomes to be reconstructed by aligning vast numbers of short reads to a previously assembled linear (i.e., haploid) reference sequence. These genomes can then be used to quantify genetic diversity within and between populations and to reconstruct the evolutionary and demographic history of different taxa. Notably, the haploid nature of linear reference genomes means that all individual genomes (including the donor) will carry alleles not present in the reference genome. Accordingly, read alignment against linear reference genomes inevitably leads to a well known statistical artefact where reads carrying alleles that differ from the reference have a reduced capacity to correctly align, resulting in an overrepresentation of reference alleles in alignments [92] leads to inaccurate haplotype and genotype calls, with more genetically divergent individuals being more heavily impacted. Crucially, reference bias can potentially lead to biassed downstream population genomic inferences and subsequent interpretations regarding historical demographic changes, selection patterns, and migratory movements [93–95].

Many of the unique proprieties of aDNA, including those outlined above, may make it particularly sensitive to reference bias. However, despite its potential to negatively impact paleogenomic research, reference bias has received little empirical investigation in an explicit paleogenomic context to date. In the remainder of the introduction, I discuss the potential for reference bias in paleogenomic studies and introduce new approaches that capture population genetic variation in reference genomes, which may help mitigate this bias. Finally, I close this section by briefly introducing the content of the subsequent empirical chapters.

Impact of reference bias on paleogenomics studies

Ancient DNA exhibits a range of features that likely make reference bias more problematic than for present-day samples. One major factor is that aDNA reads tend to be extremely short, frequently shorter than 50 bp, due to post mortem degradation and fragmentation of DNA molecules [67,96]. Shorter reads display a stronger reference bias than longer sequences – presumably because the relative effect of an alternate allele increases as read lengths shorten – and read length may be the main driver of this bias [94]. Reference bias may be further exacerbated by the low yields of endogenous DNA that are typical for paleogenomic samples [97] and also by post-mortem damage imposing an additional penalty that makes reads carrying true alternate alleles less likely to map.

The low quantity of endogenous DNA also means that it is rarely possible to have high redundancy at each locus with aDNA compared to modern DNA – it is standard to have a coverage depth below 1X for aDNA – making the determination of the two possible alleles in a diploid sample particularly difficult. Consequently, most studies involving aDNA use pseudo-haploid variant calls, i.e., rather than trying to call both alleles at each variable site, a single allele is randomly sampled per site from mapped read(s) covering the position of interest [94,98]. Accordingly, reference bias increases the chance that pseudo-haploid calls will favour the reference allele.

Finally, ancient samples may contain genetic sequences that are no longer present in modern populations – which may have been lost or changed through mutation or recombination [99] – and therefore are missing in the linear reference. Altogether, the unique combination of missing and error-prone data, along with low endogenous DNA yields, make aDNA studies particularly vulnerable to reference bias. Indeed, systematic underestimation of genetic variation in aDNA samples could partly explain why a recent study reported low genetic diversity for several prehistoric human groups [100]. Despite the likelihood of reference bias in aDNA studies, the extent of reference biases in paleogenomics studies, and its impact on our current understanding of human history, remains unclear.

Incorporating genetic variation into reference genomes

All linear reference genomes are a haploid representation of a particular species, including those for diploid organisms. For our own species, the reference genome [101] is a haploid consensus sequence that was built from a small group of individuals [102] with African and European backgrounds [103]. The majority of the reference (92%) is based on the consensus sequence of eight individuals, with 70% of the reference coming from a single individual with admixed European and African ancestry [104,105] . Hence, reference bias is a particular issue for individuals from populations not represented in the reference, which include DNA from ancient specimens that likely include variants that no longer segregate in modern human populations. Nowadays, reference bias is a pervasive feature of genomic studies in humans and other organisms. To help ameliorate problems associated with reference bias, a handful of approaches have emerged that incorporate population genetic variation into the reference genomes. For humans, the latest version of the reference genome (GRCh38) contains several alternative sequences of highly variable regions that are too complicated to be adequately represented in a linear sequence, such as the HLA region [106]. Because these alternative tracks increase the diversity represented in particular genomic regions, mappability in these regions is improved [107] which could result in a reduced impact of reference bias locally.

Another possible way to increase the information present in a linear reference, is to use the IUPAC character (International Union of Pure and Applied Chemistry). The IUPAC alphabet is a 16-character code that allows for ambiguous nucleic acid classification. The code can represent states with single nucleic acid requirements (A, G, C, T/U) or states with ambiguity among 2, 3, or 4 potential nucleic acid states. Even though this approach can be useful and decrease the impact of reference bias, only a handful of software are able to use IUPAC codes in their alignment process (*NovoAlign* http://www.novocraft.com/products/novoalign, *BBmap* [108]).

Although the addition of alternative [109] (alt alleles) and decoy sequences in linear reference genomes have improved mapping and resulted in some reduction in off-target alignment [110,111] and reference bias [112], we have likely reached the practical and theoretical limits of how well population genetic variation can be captured in a linear reference, prompting the development of new graph-like representations, i.e., variation graphs.

Pangenomes and variation graphs

The limitations of incorporating population genomic diversity within a single linear data structure [113], have led researchers to turn to graph representation as a potential solution [114]. Graph representations (graphs) counter the key weakness of linear reference-based alignment caused by the absence of representative genetic diversity by using a structure capable of containing population-level genetic variation. These graph structures are called pangenome graphs and/or variation graphs [115–117].

Pangenome graphs are built from a representative set of whole-genome assemblies, composed of a “core genome” that contains haplotypes present in all individuals, and the “variable or accessory genome”, which contains haplotypes that are unique to a subset of individuals [118]. While graph-like representations of population genetic variation will undoubtedly be useful to research and science in general, the complexity of the task and the lack of suitable methods mean that only a handful currently exist [119]. Pangenomes for bacterial species make up most of those currently available [120–127], with only a few having been published for humans [95,116,128,129].

Variation graphs on the other hand, are built using a reference genome sequence and a set of known genetic variants, where variations form “bubbles" on the genetically homogeneous background sequence. The graphs can then be updated as new data are produced and genetic variants discovered, facilitating population genomic analyses by capturing and representing much of the segregating genetic variation in a given species.

Crucially, by incorporating population genetic variation directly into the reference structure, more accurate alignments can be achieved through the use of variation-aware alignment methods. In particular, pangenome and variation graphs are able to directly incorporate complex forms of genetic variation such as structural variants (i.e. insertions, deletions, inversions, translocations and duplications) – which are typically difficult to incorporate within a conventional linear reference – allowing accurate differentiation between sequencing errors and true structural variants in reads. Furthermore, graph-based alignment methods use an iterative updating process and a separate realignment step near indels, thereby improving read alignment and facilitating more robust detection of new variants relative to conventional linear genome alignment [95,130]. Notably, despite the increased complexity of graphs relative to linear reference, graph-based alignment methods achieve similar memory usage and running times to linear approaches, in part by making use of computational efficient *De Bruijn* algorithm [131] to collapse repeated sequences [132].

By increasing the accuracy of alignment and variant detection and calling [130,133], genome graphs offer a powerful new method to reduce reference bias and improve the accuracy of genomic research. However, despite evidence pointing to graph-based approaches having several advantages over alignments to linear reference, only a handful of studies have attempted to employ this new approach in human paleogenomic research [95], even though software and suitable reference genomes are now available [132,134–137].

My PhD

Paleogenomics is a relatively young discipline that accordingly faces several novel technical challenges, with some key issues being highlighted above. In the following empirical chapters, I seek to address specific issues related to reference bias and reproducibility in paleogenomic research and make several recommendations in the hope of improving these fundamental aspects in future research. My thesis project is structured into the following three interrelated chapters.

Determining the best linear aligner for paleogenomic datasets

In my first chapter, I focus on accurately quantifying the impact of reference bias on paleogenomic datasets using different alignment methods. It has been shown that most population genetics studies are affected by reference bias [94], leading to biassed variant calls [138] and impacting downstream statistical analysis [139] However, the most recent study of aligner performance on aDNA reads context dates back to 2012 [71]the early stages of paleogenomics. I investigate the impact of reference bias on paleogenomic datasets by undertaking a systematic benchmark of the most widely-used mapping tools currently in use on simulated datasets that capture common aDNA contingencies, including ultra-short reads, DNA damage, and contamination, and make updated recommendations for future paleogenomic projects.

Impact of graph-based methods on population genetic inferences

Whether paleogenomic inferences can be further improved by aligning reads to reference graphs forms the focus of the second chapter. I compare the impact of graph-based and linear alignment methods on statistical inferences involving ancient samples that have been used to support surprising or contentious population genetic relationships [71,140]. I show that graph based alignments tend to produce more conservative statistical inferences that are consistent with recent observations that they are less affected by reference bias than alignments to linear reference genomes [95]. Notably, changes in alignment method resulted in reversals in the statistical significance of particular results, which was also observed when altering samples used in the analyses, highlighting the potential for non-trivial impacts following subtle changes in paleogenomic data processing and analysis and prompting the final chapter of this thesis.

Defining reproducibility standards for paleogenomic research

Reproducibility is an integral component of any research enterprise that promotes more reliability in science [141]. Worryingly, recent attempts to replicate results in several scientific fields have failed [142–145] – including some that were published in leading research journals such as *Science* or *Nature* [146] – which has led some researchers to proclaim that a general "reproducibility crisis" may now exist [147].

As a relatively young but rapidly growing and technically complex field, paleogenomics is likely to also be susceptible to inadequate reporting practices. Accordingly, in my third chapter, I make a series of recommendations regarding the minimum information needed to ensure full reproducibility in future paleogenomic research.

Chapter 1: Benchmarking of mapping tools for ancient DNA

Systematic Benchmark of ancient DNA read mapping

Statement of Authorship

Title of Paper	Systematic benchmark of ancient DNA read mapping	
Publication Status	☒ Published □ Submitted for Publication	□ Accepted for Publication □ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Oliva A, Tobler R, Cooper A, Llamas B, Souilmi Y. *Systematic benchmark of ancient DNA read mapping*. Brief Bioinform. 2021 Sep 2;22(5):bbab076. doi: 10.1093/bib/bbab076. PMID: 33834210.	

Principal Author

Name of Principal Author (Candidate)	Adrien Oliva		
Contribution to the Paper	Designed the experiment, performed data processing, analysed and interpreted the data, wrote the manuscript		
Overall percentage (%)	80		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	02/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Raymond Tobler		
Contribution to the Paper	Analysed the data, contributed and edited the manuscript		
Signature		Date	02/03/2022

Name of Co-Author	Alan J. Cooper		
Contribution to the Paper	Critically evaluated the manuscript		
Signature		Date	3/3/2022

Name of Co-Author	Bastien Llamas		
Contribution to the Paper	Contributed and edited the manuscript		
Signature		Date	02/03/2022

Name of Co-Author	Yassine Souilmi		
Contribution to the Paper	Conceptualised, contributed and edited the manuscript		
Signature		Date	02/03/2022

Briefings in Bioinformatics, 22(5), 2021, 1–12

https://doi.org/10.1093/bib/bbab076
Method Review

Systematic benchmark of ancient DNA read mapping

Adrien Oliva, Raymond Tobler, Alan Cooper, Bastien Llamas and Yassine Souilmi

Corresponding author: Yassine Souilmi, Australian Centre for Ancient DNA, University of Adelaide, Australia. Tel: +61883135565; Fax: +6183134364; E-mail: yassine.souilmi@adelaide.edu.au

Abstract

The current standard practice for assembling individual genomes involves mapping millions of short DNA sequences (also known as DNA 'reads') against a pre-constructed reference genome. Mapping vast amounts of short reads in a timely manner is a computationally challenging task that inevitably produces artefacts, including biases against alleles not found in the reference genome. This reference bias and other mapping artefacts are expected to be exacerbated in ancient DNA (aDNA) studies, which rely on the analysis of low quantities of damaged and very short DNA fragments (∼30–80 bp). Nevertheless, the current gold-standard mapping strategies for aDNA studies have effectively remained unchanged for nearly a decade, during which time new software has emerged. In this study, we used simulated aDNA reads from three different human populations to benchmark the performance of 30 distinct mapping strategies implemented across four different read mapping software—*BWA-aln*, *BWA-mem*, *NovoAlign* and *Bowtie2*—and quantified the impact of reference bias in downstream population genetic analyses. We show that specific *NovoAlign, BWA-aln* and *BWA-mem* parameterizations achieve high mapping precision with low levels of reference bias, particularly after filtering out reads with low mapping qualities. However, unbiased *NovoAlign* results required the use of an IUPAC reference genome. While relevant only to aDNA projects where reference population data are available, the benefit of using an IUPAC reference demonstrates the value of incorporating population genetic information into the aDNA mapping process, echoing recent results based on graph genome representations.

Key words: ancient DNA; alignment; reference bias; benchmarking

INTRODUCTION

Genomic data are playing an increasingly prominent role in contemporary evolutionary biology and biomedical research. This has been spurred by the development of high-throughput DNA sequencing (HTS) technologies [1, 2] and the resulting reduction in sequencing costs, leading to exponential growth in the amount of genetic data available. To date, the most cost-effective and widely used approach for generating standard genomic data is to break up genomes into large numbers of short DNA fragments, which are subsequently sequenced to produce 'reads' around 100–150 nucleotides (nt) in length. The genome sequence is then reconstituted by aligning (or 'mapping') these

Adrien Oliva is a PhD candidate at the Australian Centre for Ancient DNA at the University of Adelaide. He researches methods for reducing analytical bias in ancient DNA studies.
Raymond Tobler is a population geneticist and Australian Research Council DECRA Fellow based at the Australian Centre for Ancient DNA at the University of Adelaide. He works on human evolutionary history in Island South East Asia, New Guinea and Australia.
Alan Cooper is a recent Australian Research Council Laureate Fellow specializing in ancient DNA, evolutionary biology and the impacts of environmental change through time.
Bastien Llamas is an Australian Research Council Future Fellow and an Associate Professor based at the Australian Centre for Ancient DNA at the University of Adelaide. He specializes in the analysis of ancient DNA extracted from past humans and extinct animals to address outstanding questions in evolutionary biology.
Yassine Souilmi is an evolutionary genomicist at the Australian Centre for Ancient DNA at the University of Adelaide. He specializes in utilizing large-scale genomic analyses to address evolutionary and medical questions using modern and ancient DNA.

Submitted: 28 July 2020; **Received (in revised form):** 5 January 2021

reads to a previously constructed reference genome from the relevant species (or, alternatively, another closely related species depending on availability) using purpose-built DNA mapping software, with two particularly popular options being *BWA* [3] and *Bowtie2* [4]. In the case of ancient DNA (aDNA; retrieved from paleontological, archaeological and museum specimens) the degraded nature of the DNA means that most fragments are <150 nt long, obviating the need for further fragmentation.

Crucially, accurate mapping of the DNA reads is essential for proper identification of genetic variants that are often the focus of investigation in medical or population genetic studies, e.g. the discovery of alleles that contribute to disease, or determination of genetic markers that help define ancestry and population affiliations. However, the DNA alignment step is known to be error-prone [5] and can also result in biases that are related to the properties of the reads and specific characteristics of the reference genome [6]. In particular, reads carrying one or more alleles that are not present in the reference genome are known to be less likely to map than reads carrying the reference allele, which results in a bias against calling non-reference (i.e. alternate) alleles [7, 8]. This problem, which is commonly referred to as reference bias, becomes more problematic for DNA from individuals that are highly diverged from the individual(s) used to create the reference genome and is known to lead to incorrect inferences in downstream genetic analyses [9, 10].

Errors in DNA alignment and reference bias are expected to be particularly problematic for the analysis of aDNA, a rapidly growing research field that introduces new challenges for read-mapping algorithms that are absent from modern genetic datasets. Because of postmortem DNA degradation, DNA fragments recovered from ancient specimens are much shorter (resulting in the majority of reads being between 30 and 80 nt [11]) than those from artificially fragmented modern genomes, and also typically contain damage-modified sequences and high levels of DNA contamination from other organisms (mainly microbes). These factors hinder the alignment of aDNA reads to the reference genome and result in much lower genomic coverage relative to modern samples for a given sequencing effort—i.e. a lower amount of unique reads mapping to each nucleotide in the reference genome—making aDNA highly susceptible to mapping errors, miscalled alleles and reference bias [6].

Despite these problems, only a handful of studies have examined the performance of different alignment tools for aDNA reads, with the current widely adopted standard mapping practices in aDNA research established nearly a decade ago [12]. However, this key benchmarking study only tested different options from a single alignment tool—*BWA-aln* (Burrows-Wheeler Aligner [3])—and one of the tested sequencing platforms is now obsolete (Helicos True Single-Molecule DNA Sequencing platforms) [13]. Today, improved Illumina short-read sequencing laboratory methods have become standard in the aDNA field, including treatments that partly or fully remove damaged nucleotides caused by deamination of unmethylated cytosine sites [14]. Additionally, several alternate alignment software are now available that might improve aDNA mapping [15]. Importantly, the optimal choice of aligner may depend on characteristics of the genetic data—such as fragment length, damage profiles and the treatment of deaminated sites—which have yet to be formally investigated in the context of aDNA mapping. Recent work has examined the impact of reference bias in empirical ancient human datasets and its implications on downstream population genetic analyses and how reference bias in aDNA studies can be corrected using new genome graph-based alignment software. However, there has still not been a systematic investigation of these questions that includes multiple mapping software and inbuilt parameter settings using standard linear reference genomes [6, 16, 17]. Furthermore, alignment to linear reference genomes is likely to remain the standard approach for several more years while a standardized coordinate system and annotations are developed for genome graphs [18, 19].

Here, we provide a comprehensive systematic evaluation of the alignment performance and reference bias for multiple DNA aligners in an aDNA context, using simulated aDNA reads modelled on empirical data from human populations from three different continental regions (i.e. Africa, East Asia and Europe). The performance of four commonly used software tools i.e. *BWA-aln*, *BWA-mem*, *Bowtie2* and *NovoAlign* are evaluated with 30 different parameterizations by measuring read alignment precision and reference bias at simulated polymorphic loci. We compare performance both with and without treatments that remove damaged nucleotides that are now commonplace in aDNA studies, and also using standard filtering strategies based on read alignment qualities. Additionally, we examine if alignment performance is improved when using an augmented reference genome that incorporates information about previously identified polymorphic sites, by replacing the reference nucleotide with the IUPAC (International Union of Pure and Applied Chemistry) character that captures diversity at the locus (see Methods). Lastly, we assess the impact of mapping strategy on downstream population genetic analyses. Our results reveal that all methods exhibit some degree of reference bias when aligning short aDNA reads, and we provide recommendations to optimize alignment performance and minimize reference bias at SNP-based levels and in downstream analyses.

METHODS

Data simulation

To provide an empirical basis for examining the performance of DNA aligners we generated a set of simulated DNA fragments and used them in a benchmarking pipeline. As humans are the most widely studied species in aDNA research, we chose to limit our study to simulated data based on human population genetic datasets. Nonetheless, our findings should be relevant to other species commonly studied with aDNA (typically mammalian megafauna).

We used *Mitty* (https://github.com/sbg/Mitty) to simulate reads of 250 nt in length from chromosome 22 of the human reference genome version GRCh37. Reads were simulated using the '1kg-pcr-free.pkl' model included in *Mitty* to incorporate sequencing errors that follow an Illumina profile (*TrueSeq*; see Supplementary Methods). To determine the impact of read-lengths pertinent in ancient DNA studies on mapping performance, we trimmed the reads to lengths ranging between 30 and 120 nt, such that there were exactly 100 000 reads falling within each consecutive 10 nt bin between 30 and 120 nt (i.e. 30–39 nt; 40–49 nt; ...; 100–109 nt; 110–119 nt), and 900 000 simulated reads in total.

Variant sites were introduced into the simulated reads using polymorphic sites identified in the human 1000 Genomes Project (1000 GP) [20]. Specifically, we used *Mitty* to incorporate genetic variation at single nucleotide polymorphic sites (SNPs) and indels (insertions and deletions [21]) obtained from the following three samples from the 1000 GP [20]: *HG00119* (British in England and Scotland; GBR; labelled Europe in this study), *NA19471*

(Luhya in Webuye, Kenya; LWK; Africa in this study), *HG00513* (Han Chinese in Beijing, China; CHB; East Asia in this study). We chose individuals from three different continents to quantify the potential impact of reference bias on mapping.

aDNA damage incorporation

After generating the simulated reads, we used the 'deamSim' function from *gargammel* [22] to add aDNA damage to each read according to the empirical damage profile from the ~7000 year old 'La Braña' specimen [22, 23] (16% nuclear C > T misincorporation level [24]). Because of the large amounts of DNA damage arising from cytosine deamination, most aDNA studies now include specific treatments that repair or remove some or all deaminated cytosine sites in aDNA fragments to reduce artefacts in downstream analyses. Accordingly, we explored the impact of the two most widely used treatments, which use uracil DNA glycosylase (UDG) to remove uracil bases resulting from cytosine deamination, on alignment performance. We simulated reads that mimic the read sequence modifications specific to Full-UDG (*F-UDG*) and Half-UDG (*H-UDG*) treatments, which result in the partial replacement of deaminated cytosines [14, 25] and compared the results to reads with the full damage spectrum. To simulate the impact of Half-UDG treatment on aDNA reads, we used *gargammel* [22] to model the partial replacement of deaminated cytosines based on an empirical profile produced from this treatment (i.e. individual 'SAMEA4843644' [26]). For the Full-UDG treatment, we trimmed the five terminal nucleotides from either end of each read to mimic the removal of terminal regions and the corresponding fragment shortening characteristic of this treatment. Because our analytical results were consistent across the three samples for untreated reads (see results), we only performed the UDG treatments on the East Asian dataset (See Supplementary Methods).

DNA aligners tested

To provide a broad assessment of potential aDNA alignment strategies we tested *BWA-aln* [3], which is currently the most widely used alignment software in aDNA studies, and three other aligners that are widely used in modern DNA research but for which aDNA performance has not been formally explored in relation to multiple alternate aligners under standardized benchmarking criteria, namely, *BWA-mem* (version 0.7.15) [27], *Bowtie2* (version 2.2.9) [4] and *NovoAlign* (version v3.09.01; http://www.novocraft.com/products/novoalign) (Table S1).

BWA-aln, also known as *BWA-backtrack* [3], is the current standard software used to map aDNA reads, using parameter settings outlined in Schubert *et al.* [12]. We include all of the *BWA-aln* parameter settings used in the benchmarking study of Schubert and colleagues, and include a handful of additional parameter combinations that have also been reported in the aDNA field [28, 29], which incorporate alternate gap penalties and mismatch allowances (Table S1).

We also evaluate the performance of *BWA-mem* [27] using the default parameter settings. *BWA-mem* is significantly faster than *BWA-aln* but the current implementation is optimized for reads longer than 70 bp and so it is not expected to be optimal for many aDNA studies; however, it is of general interest due to the large speed improvements relative to *BWA-aln*.

Bowtie2 [4] is another popular aligner that was recently shown to perform well with ancient DNA reads under specific circumstances [15, 30]. By default, *Bowtie2* performs an end-to-end (i.e. 'global') alignment that forces the software to avoid soft-clipping of reads (i.e. masking of terminal portions of the reads that do not align to the genome). In addition to this default option, we also tested a local alignment option, which might be useful for aDNA mapping as DNA damage is more likely to accumulate at the terminal ends of reads. We also tested the 'very sensitive' option of *Bowtie2* for both global and local alignment methods, and included another local-alignment implementation that was shown to improve aDNA alignments in a recent empirical study (*BWT5*; see Table S1 for parameters; [30]). Although most *Bowtie2* preset options are designed to improve speed while maintaining high accuracy across a range of variables, those with the 'sensitive' option sacrifice speed for improved precision and accuracy.

Finally, we used the proprietary DNA aligner *NovoAlign* developed by Novocraft Technologies. While *NovoAlign* requires a licence to access some features, a single-core version is freely available to researchers. We tested *NovoAlign* using the default parameters and several additional parameter combinations that are outlined in Table S1.

IUPAC implementation

A recent study revealed that alignment of ancient DNA reads to a genome graph results in reduced reference bias compared to *BWA-aln* alignments to a linear reference genome [6]. Hence, we tested if the introduction of polymorphism information in the linear reference could improve mapping performance and decrease reference bias, by rerunning the alignment on a modified reference genome where all single nucleotide polymorphic sites identified in the 1000 GP (phase 3) were converted to the corresponding IUPAC character. IUPAC-modified reference genomes contain redundant nucleotides at known polymorphic positions that allow for two or more alleles to map to these genome positions without incurring a penalty [31]. The IUPAC alignments were limited to a subset of the *NovoAlign* strategies, as this is the only aligner that currently supports IUPAC implementations (note that *NovoAlign* only supports the inclusion of SNP variants in IUPAC-modified reference genomes, prohibiting any testing of the impact of polymorphic indels that were present in our samples).

Post-alignment processing

Following read mapping, we applied a set of post-alignment processing steps that are standard for the aDNA field [32]. Briefly, we used the *MarkDuplicates* function in *Picard* Tools (https://broadinstitute.github.io/picard) to identify all duplicated reads and then realigned reads around known indels using *GATK indelrealigner* (version 3.7) [33]. The final set of processed reads was analysed through comparison with their true mapping locations. We note that the assessment of alignments took soft clipping (i.e. removal of one or both terminal read ends in the alignment) into account wherever this occurred.

Quantifying aligner performance

Precision

For all software parameterizations and each dataset, three classes of read alignments were defined and used to calculate statistical indices: (1) true positives (*TP*): simulated reads that mapped to the correct genomic location; (2) false positives (*FP*): simulated reads that mapped to the incorrect genomic location; and (3) false negatives (*FN*): simulated reads that did not map. The best performing software is expected to have the majority of reads correctly mapped to the reference while minimizing the amount of incorrect and unmapped reads. Accordingly, we measured the precision (i.e. $\frac{TP}{TP+FP}$) of each parameterization and contrasted this against the proportion of reads that mapped overall (see Supplementary Methods).

SNP-based reference bias

For each sample, we also quantified the level of reference bias across all SNPs by contrasting the alignment precision for reads carrying an alternate versus those carrying reference alleles at each polymorphic site. Specifically, for each sample, we determined all reads that overlapped a known SNP conditional on there being at least one alternate allele for a given SNP in the simulated reads for that sample. We then calculated the precision for reads carrying the alternate allele across all polymorphic SNPs, and subtracted this from the precision of reads that carried the reference allele at these positions. Accordingly, positive values are a direct aggregate measure of reference bias summarized across all SNPs in each sample.

Testing the impact of read filters on performance

Following standard approaches in modern aDNA studies, all precision and SNP-based reference bias measurements were made after excluding reads with mapping quality of 0 ($mapQ > 0$; i.e. removing any reads with ambiguous mapping), and additionally after excluding reads with mapping quality below 25 ($mapQ \geq 25$). The application of mapping filters is a common practice in genomic studies to improve alignment performance, and results in the removal of both unmapped reads (*FN*) and a subset of mapped reads with *mapQ* below the prescribed threshold (ideally removing a higher proportion of *FP* than *TP* reads). We chose the $mapQ \geq 25$ cut-off as it has been shown to improve the quality of read alignments in previous aDNA studies that use the *BWA-aln* algorithm [34]—though we caution that the maximum *mapQ* value varies widely across the alignment software tested here (i.e. 37 for BWA-aln, 42 for *Bowtie2*, 60 for *BWA-mem* and 70 for *Novoalign*), such that this *mapQ* value may not be as meaningful for other software tools.

Additionally, we also tested the impact of applying mappability filters and masking known CpG islands on SNP-based reference bias. Studies of *BWA-aln* alignments to the human reference genome have revealed considerable variation across the genome in the mapping accuracy, leading to the availability of files that demarcate regions of low mappability that can be masked from analyses. Similarly, CpG islands were recently shown to impact the alignment precision in aDNA studies, presumably because such regions are highly methylated and therefore prone to read deamination and damage-related mapping artefacts [15]. Files with mappable regions and CpG islands were obtained from *BWA-aln* developer Heng Li's blog (https://lh3.github.io/) and from UCSC genome browser 'CpG island' table (https://genome.ucsc.edu/cgi-bin/hgTables), respectively.

Measuring the impact of reference bias on downstream analyses

To test the impact of reference bias on downstream population genetic analyses in each mapping strategy, we called individual variants (see next section) for each individual dataset (NA19471 from Africa, HG00119 from Europe and HG00513 from East Asia) and then used these variants in two commonly applied population genetic summary statistics: the D-statistic and Principal Component Analysis (PCA). We quantified the *D*-statistic under the following topology $D(Y, X = Truth; Ref, Chimp)$, where *X* is the true set of variants for that individual, Y is the set of calls inferred from the simulated data set, *Ref* is human reference genome used in the alignment (GRCh37) and *Chimp* is the chimpanzee reference genome. Notably, reference bias is expected to result in significantly positive *D*-statistic scores, which results from an excess of allele sharing between the inferred simulated calls and the human reference genome relative to the truth set [35]. All *D*-statistic estimates were computed using admixr [38].

For the PCA test, we used the popular *SmartPCA* ordination method from the *EIGENSOFT* package [36] to compute the sample loadings for both simulated and true datasets for each sample, and then measured the Euclidean distance (d_{sim} and d_{truth}, for samples and truth sets, respectively) of each to the reference sequence in a 10-dimensional PCA space. Subsequently, we quantified the reference bias as follows:

$$\frac{d_{\text{truth}} - d_{\text{sim}}}{d_{\text{truth}}}$$

Accordingly, positive values indicate the relative shift towards the reference in PCA space caused by reference bias for a given simulated data set.

Variant calling

To call variant positions we followed the standard pseudo-haploid calling process applied in aDNA studies which uses the *mpileup* function from *SAMtools* [37] and *pileupCaller* from *sequenceTools* (https://github.com/stschiff/sequenceTools). The pseudo-haploidization step calls the allele for each predefined variant by determining the identity of the homologous variant in a randomly sampled read. In order to capture the uncertainty introduced by the random sampling step we repeated the pseudo-haploidization call three times for each software parameterization and sampled individual.

For each alignment method and population we downsampled the set of aligned reads to match the empirical read-length distribution of aligned reads provided by *gargammel* (i.e. Motala ancient specimen [23]). This ensures that our calls are based on aligned reads with lengths that are representative of empirical ancient datasets and reduces the possibility that the calls are unduly influenced by alleles carried on longer reads. Each downsampled subset of reads were filtered for mapping quality using the same two filters used to measure alignment performance (i.e. $mapQ > 0$ and $mapQ \geq 25$), resulting in a final set of high-quality mapped reads that were used to assess reference bias.

Aligner specificity for common DNA contaminants

DNA extracts from ancient specimens typically contain large amounts of additional DNA from contaminant organisms that end up being sequenced along with the host DNA. Accordingly, we examined the specificity (i.e. in this context the proportion of unmapped reads) of each aligner for aligning contaminant reads. To this end, we used *gargammel* [22] to simulate 10 000 single-end bacterial reads modelled upon the bacterial reads recovered from Kostenski 14 [39] (as recommended in *gargammel*), and also aligned a random sample of 10 000 dog reads taken from a recent publication (i.e. sample Basenji2/SAMN03366708 [40]) to

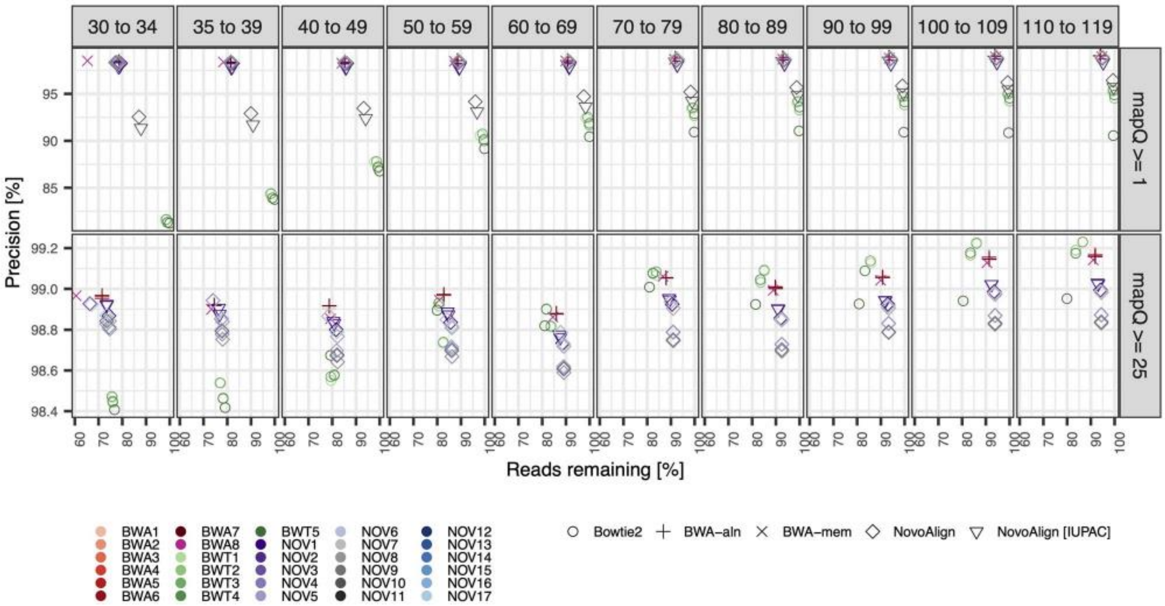


Figure 1. Alignment precision relative to read length and two mapping quality filters for the simulated East Asian sample. We assessed the precision of 30 parameterizations (different colours) for four different alignment software, including an IUPAC-based alignment for a subset of the *NovoAlign* parameterizations (different shapes). The x-axis measures the number of reads remaining after applying the specific mapping quality filter, which results in the removal of all reads below the required mapping along with all unmapped reads. Results were similar for the European and African samples and are shown in the Figures S1 and S2, respectively.

examine the alignment specificity of potential contamination derived from a more closely related commensal species.

Computational resources

We used a standard computational configuration for all aligner parameterizations applied in this study. Specifically, all analyses were performed on the University of Adelaide's Phoenix cluster and used 16GB of memory and 8 cores. To compare the running time for the different aligners we report the 'core-hours' for each alignment, which refers to the number of processor units (cores) used to run a simulation multiplied by the duration of the job in hours. For instance, 1 core hour could comprise 1 h on a single core, 30 min each on 2 cores, and so on.

RESULTS

Mapping precision

Mapping precision was estimated for sets of reads falling into successive 10 nt read length bins, with the exception that the smallest reads were split into two 5 nt length bins (i.e. 30–34 nt and 35–39 nt), since many aDNA studies use either 30 nt or 35 nt as the lower read length threshold in their analyses. For all software parameterizations, we observe that longer reads were consistently more likely to map to their true genomic position with fewer *FPs* than shorter reads, as expected (Figure 1 and Table S2), and showed qualitatively similar patterns across the three different human samples (Figures S1 and S2). Notable differences in performance were observed between the different software tested, though different parameterizations of the same software had highly similar performance overall (i.e. more variation was observed between the different aligners than across the different parameterizations of each aligner). In particular, when excluding reads with *mapQ* of 0, the four *Bowtie2* parameterizations had consistently lower precision and a higher proportion of mapped reads across all read length bins compared to the three other tested aligners, with the differences between the software becoming more exaggerated for shorter reads (i.e. *Bowtie2* having ∼15% lower precision while mapping ∼20% more reads of 30–34 nt length; Figure 1 and Figures S1 and S2). In contrast, *BWA-aln*, *BWA-mem* and *NovoAlign* aligners showed highly consistent results across each of the read-length bins when excluding reads with *mapQ* of 0, albeit with three notable exceptions: (1) *BWA-mem* had a lower proportion of reads mapped in the 30–34 nt read length bin but otherwise had comparable performance with *BWA-aln* across other read lengths; (2) the *NovoAlign* multiple-alignment parameterization (*NOV10*) behaved more like *Bowtie2*—i.e. comparatively lower precision and more mapped reads—particularly for longer reads; (3) there were slight improvements in alignment precision for the *NovoAlign* methods when using the reference genome with IUPAC characters (i.e. incorporating information on polymorphic sites from the 1000 GP reference populations) relative to alignments to the unmodified reference.

When reads with mapping qualities less than 25 were removed, as is common practice for aDNA studies using *BWA-aln* software [34], all methods exhibited a boost in precision at the cost of mapping less reads overall (Figure 1 and Figures S1 and S2). Using this more stringent mapping quality filter resulted in the precision becoming relatively uniform across the different alignment software and also across read length bins, with precision ranging between 98.4 and 99.3% for all read lengths and parameterizations. Overall, the four *Bowtie2* parameterizations showed the biggest gains in precision after applying the *mapQ* ≥25 filter, particularly for shorter reads, with precision increasing more than 15% for reads less than 40 nt in

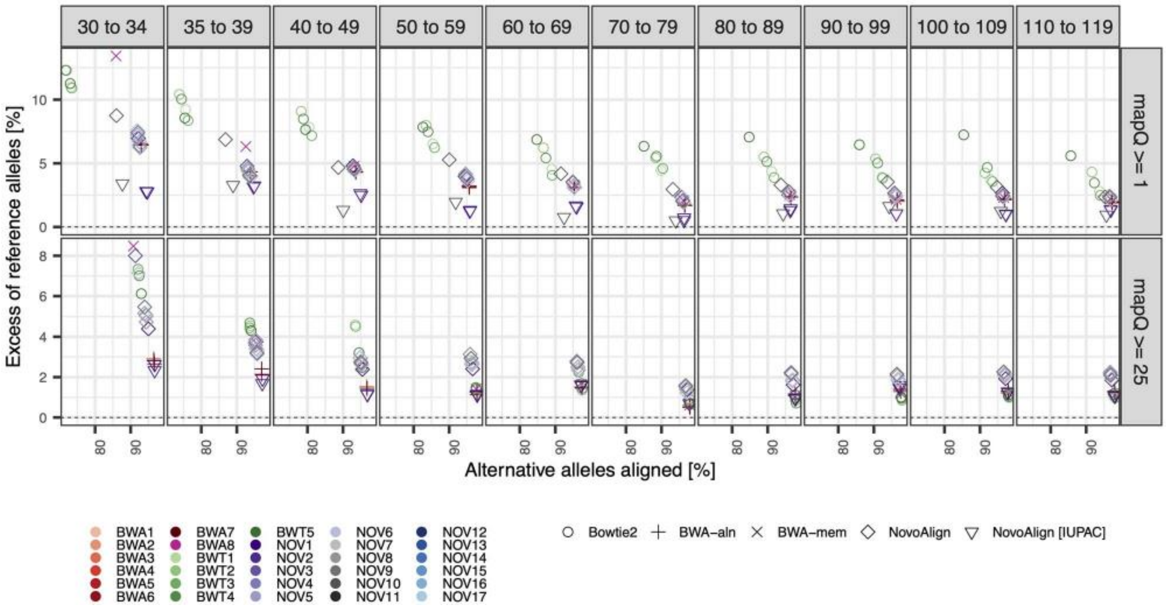


Figure 2. SNP-based reference bias relative to read length and two mapping quality filters for the simulated East Asian sample. We measured the degree of reference bias by evaluating the difference in sensitivity between alternate and reference alleles across all SNPs covered by at least one alternate allele across successive read length bins. All other parameters are identical to those shown in Figure 1. Notably, the reference bias patterns were broadly consistent with sensitivity patterns shown in Figure 1, and were also similar for European and African samples (see Figures S3 and S4, respectively).

length (compared to <2% gain for parameterizations of the other aligners for these read lengths on average).

SNP-based reference bias

The level of reference bias closely complemented the alignment precision results, with all strategies showing a consistent positive bias across all read length bins that increased steadily as reads become shorter—for reads with *mapQ* > 0, bias ranged between ∼2 and 14% for the reads between 30 and 34 nt, reducing to between ∼0.5 and 5% for reads between 100 and 119 nt (Figure 2 and Figures S3 and S4). Reference bias was further reduced after applying the *mapQ* ≥25 filter, with the largest gains occurring for the most biased strategies and the shortest reads (e.g. reducing ∼4–6% for *BWA-aln*, *BWA-mem* and *Bowtie2* for reads between 30 and 34 nt), but application of the quality filter did not completely remove bias for any strategies in any read length bin (Figure 2 and Figures S3 and S4). Notably, the *NovoAlign* methods that aligned reads to the IUPAC-augmented genome were the least biased strategies across all read lengths and easily outperformed *NovoAlign* strategies aligned to the standard reference, though BWA-aln also showed comparable performance with IUPAC-aligned *NovoAlign* strategies after applying the *mapQ* ≥25 filter (<3% reference bias across all read length bins; Figure 2 and Figures S3 and S4).

UDG treatments

The reads simulated to reflect either a full or half-UDG treatment (*F-UDG* and *H-UDG*, respectively) resulted in additional improvements in mapping precision (Figures S5 and S6) and reductions in reference bias relative to untreated reads (Figures S7 and S8). The reductions in reference bias were most marked for *F-UDG*-treated reads of shorter lengths (*F-UDG*-mediated reductions in reference bias >3% for reads between 30 and 34 nt relative to untreated reads, cf. <1% reductions for *F-UDG*-treated reads of the same length; Figures S7 and S8). The positive impacts of UDG treatment and mapping quality filters on reference bias tended to aggregate, with the lowest overall reference biases being observed amongst *F-UDG*-treated reads that were filtered for *mapQ* ≥25. Notably, while both UDG treatments tended to reduce the reference bias, the total number of mapped reads was comparable with non-UDG treatments (Figure S6) and the reduction in reference bias appeared to be caused by a lower proportion of reference alleles being correctly aligned (rather than an increase in proportion of mapped alternate alleles; Figure S8). In contrast, applying the *mapQ* ≥25 filter resulted in a sizable decrease in the number of mapped reads across all treatments, and reductions in reference bias were largely driven by an inflation in proportion of alternate alleles being aligned within this smaller set of reads.

Filtering based on mappability and CpG islands

Additional reductions in reference bias were achieved when removing reads in regions of low mappability (∼1% reduction for standard *NovoAlign* and *Bowtie2*, and between 0.2 and 0.6% for other strategies across untreated and UDG-treated reads), but were negligible when masking CpG islands (<0.1% reduction for all strategies and read treatments; Figure 3). Nonetheless, reference bias was consistently inflated in CpG islands relative to other genomic regions, but this did not result in a meaningful reduction in reference bias since only ∼1.5% of polymorphic alleles fell within CpG islands for each sample. In contrast, an order of magnitude more SNPs (∼15%) are situated within low mappability regions, whereby filtering these regions provided a small but consistent reduction in bias on top of that gained from UDG treatments and mapping quality filtering.

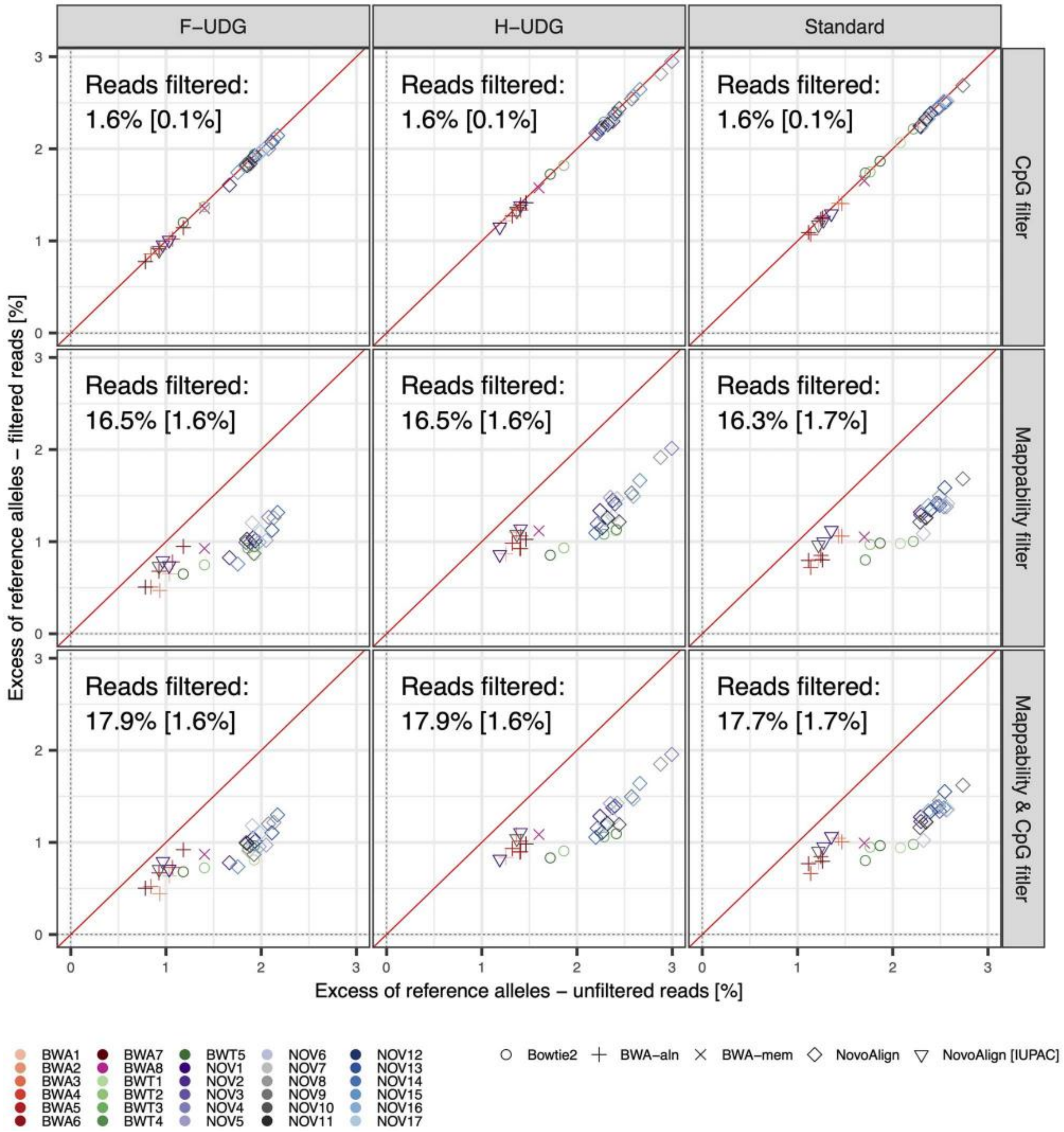


Figure 3. Impact of filtering CpG islands and low mappability regions on SNP-based reference bias. Small but consistent reductions in the reference bias were observed when masking regions with low mappability from alignments (~1%), though little change was observed in bias when masking CpG islands (<0.1%). The differences in efficacy of the two masking strategies was largely due to the latter overlapping far fewer SNPs overall (~1.5% for CpG islands versus ~17% for low mappability regions; mean percentage of filtered reads shown in each panel, standard deviation shown in square brackets). Reference bias was measured after downsampling simulated reads to match the distribution of read lengths in an empirical ancient sample (see Methods).

Impact of reference bias on population genetic analyses

To assess if the reference bias inherent in the tested alignment strategies can lead to errors in common downstream population genetic analyses, we used the *D*-statistic [41] to test if the simulated reads shared an excess of reference alleles relative to the truth set (i.e. the genotypes from the relevant population sample), using a Chimpanzee genome as an outgroup to polarize the alleles [6]. Using this topology, software parameterizations showing *D* values with a *Z*-score greater than 3 indicate a significant excess of shared alleles with the reference genome, implying that the software is impacted by reference bias.

The different strategies showed varying susceptibility to reference bias in a manner that was consistent with the inferred

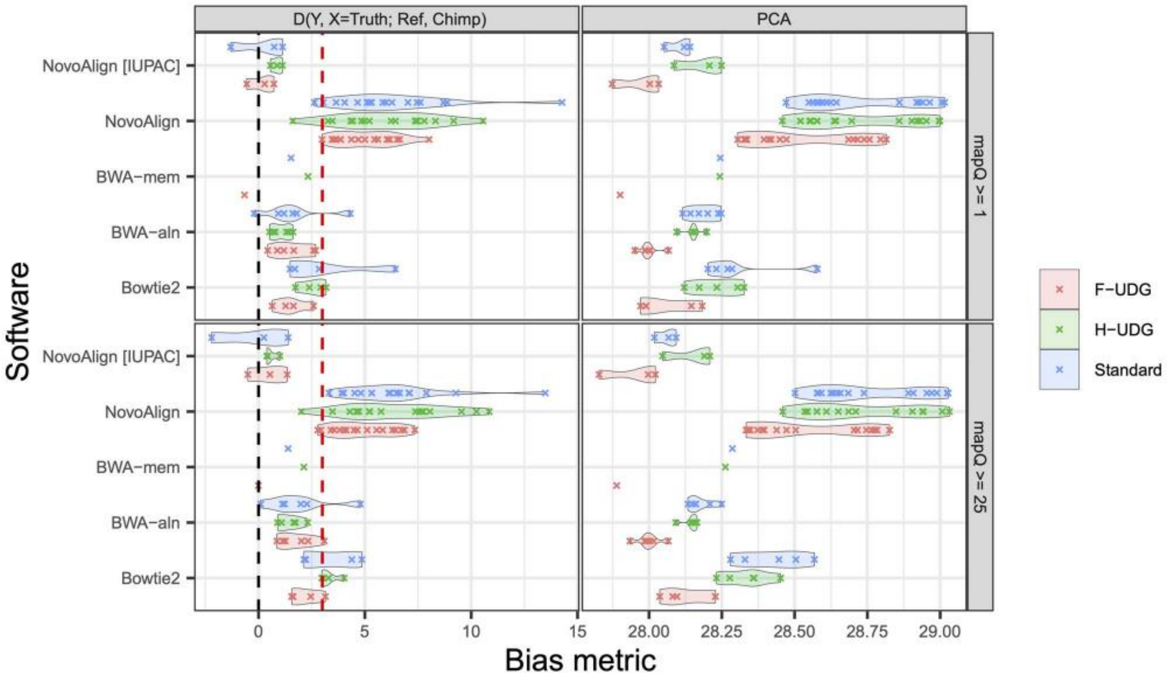


Figure 4. The degree of reference bias across the different alignment software evaluated using common population genetic analyses. Pseudo-haploid variant calls were made for all strategies, after filtering for mapping quality (different row panels) and downsampling reads to have a length distribution matching an ancient sample, and the *D*-statistic (right panels) and PCA computed (left panels). The pseudo-haploidization process was replicated three times for each parameterization and the average results are shown here. For both metrics, unbiased alignments have a value of 0. For the *D*-statistic, the *Z*-scores are plotted and positive values above 3 (red line) imply a significant excess of shared alleles between the simulations and the reference genome relative to the truth, indicative of reference bias. The PCA metric measures the reduced distance between simulations and the reference relative to expectations, with larger values indicating more reference bias (values are multiplied by 100; see Methods). Figure S9 shows the results for untreated reads in all three samples.

SNP-based reference bias Figure S10, with the *BWA-aln* and *BWA-mem* parameterizations having only a handful of significantly biased tests, while the *Bowtie2* parameterizations exhibited moderate levels of bias and most *NovoAlign* parameterizations were consistently impacted by reference bias (Figure 4 and Figures S9–S11). Interestingly, the simulated East Asian sample alignments were more prone to reference bias on average—a pattern that was also evident in the SNP-based reference bias analyses (Figure 2 and Figures S3 and S4). This may be a consequence of the reference genome assembly being mostly a composite of African and European ancestries—∼42% of the assembly is derived from a single donor of African American descent capturing a mixture of African and European ancestries—with only a minor amount of East Asian DNA being incorporated [42]. Despite our SNP-based estimates of reference bias consistently decreasing when applying the mapQ ≥25 filter, similar improvements were not evident for the *D* statistic when using the more stringent mapping quality filter (with bias actually becoming slightly exacerbated for the *Bowtie2* methods). Nonetheless, the marked reference bias exhibited by *NovoAlign* is almost entirely overcome when aligning to a reference that includes IUPAC characters, which agrees with the large reduction in bias observed when comparing standard and IUPAC NovoAlign mappings in the SNP-based metric (Figure 4 and Figures S9–S11).

Considerable reference bias was also evident in the PCA analysis (Figure 4 and Figure S12), with simulated data clustering by sample and being ∼28–29% closer to the reference genome in 10-dimensional PCA space than the associated truth set. In agreement with our previous results, *BWA-aln* and the *NovoAlign* strategies that used the IUPAC reference showed the least bias overall amongst the aligners. Similarly, *F-UDG* treated reads tended to have less reference bias than *H-UDG* and untreated reads and the mapQ ≥25 filters did not lead to appreciable changes in bias, other than an apparent increase in bias for *Bowtie2* alignments that was also observed for the *D* statistic. Notably, the SNP-based measure of reference bias actually improved under Bowtie2 when applying a more stringent filter, but at the cost of aligning considerably fewer reads (∼10–20% reduction for reads <60 nt, versus ∼2–5% reduction for nearly all other aligners; Figure S8). This suggests that the increased reference bias observed at these summary statistics for *Bowtie2* might be due to the exacerbation of reference bias at some SNPs—despite the average decrease across all SNPs—due to the large reduction in coverage.

Impact of DNA contaminants

We measured the specificity (i.e. the proportion of unmapped reads) of several of the better performing strategies across the different aligners to quantify how readily they mapped simulated aDNA from bacterial taxa, a common contaminant in aDNA studies, along with empirical aDNA derived from domestic dogs to quantify the impact of potential contamination coming from a species more closely related to the host DNA source. For the bacterial reads, all tested strategies had consistent high specificities that approached 100%, with *BWT4* performing slightly worse with a specificity around 95%. Results were more variable for

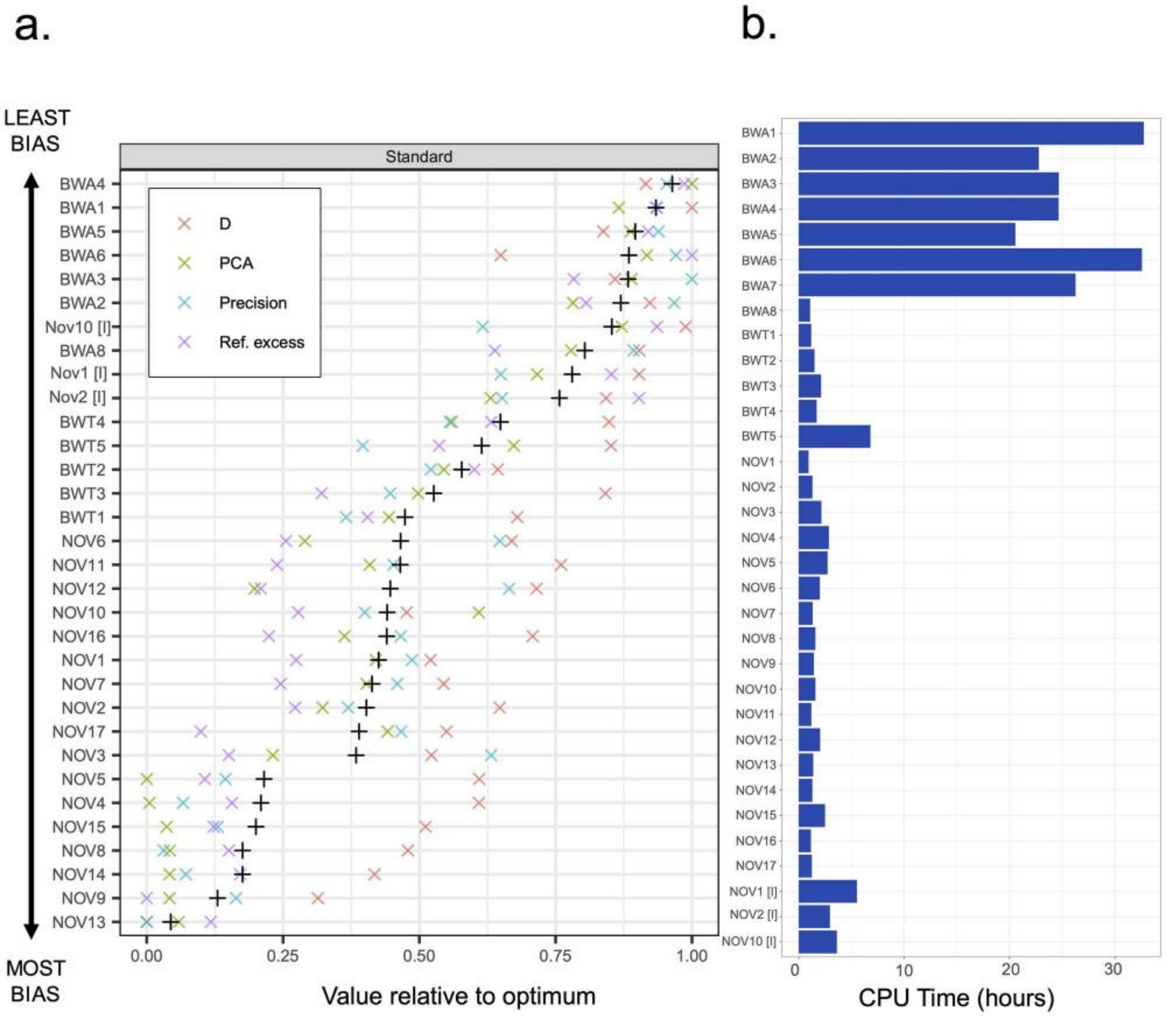


Figure 5. Summary of mapping performance and reference bias. a. Each strategy is ranked according to the mean value across the four different metrics assessed in this study, i.e. precision, SNP-based reference bias, D-statistic, and PCA (mean = black plus sign and individual metrics = different coloured crosses; see key). Each metric is scaled to take a value between 0 and 1 by dividing the metric value by the range across each metric. This plot represents the results for the East Asian dataset using the untreated reads; however, the patterns are consistent across the both UDG treatments, (see Figures S13 and S14). Note that *BWA7* had results that differed markedly from other strategies for some metrics and consequently is not represented in panel a to avoid distorting the scale for these metrics. b. Average execution time for each of the strategies for the three samples, based on 1.5 million reads (see SI for details). In both panels, strategies with the suffix '[I]' were aligned to the IUPAC-modified reference genome.

the dog reads where the two tested *BWA-aln* (BWA1 and BWA2) strategies had specificities above 99% and *BWT3* and all tested *NovoAlign* strategies performed slightly worse with specificities around 95%, whereas *BWA-mem* (i.e. *BWA8*) and *BWT4* performed poorly with specificities of 70 and 60%, respectively. This result is likely the consequence of both *BWA-mem* and *BWT4* being the only tested strategies that use local alignment, which is a more permissive alignment algorithm than the end-to-end read alignment adopted in the other strategies.

Software running times

Execution times varied widely amongst the different aligners, with the *NovoAlign BWA-mem* and *Bowtie2* parameterizations running in ~1 CPU-hour, considerably faster than the *BWA-aln* parameterizations which consumed 10 times more CPU-hours (9–11 CPU-hours) on average (Figure 5b and Table S2).

DISCUSSION

The choice of genomic alignment software and parameters is often contingent upon the properties of the genetic material being mapped. For aDNA studies, the short and degraded nature of the genetic material poses a novel challenge for sequence alignment algorithms as the software was developed for longer undamaged DNA fragments from modern specimens.

Our results are consistent with this expectation and demonstrate that in general alignment becomes more difficult as reads become shorter, with decreased precision and fewer mapped reads at lengths typical for aDNA. Moreover, our results

Table 1. Summary of mapping performance and recommendations

Rank	Software	Pros	Cons	Best Options
1=	NovoAlign IUPAC	– Large number of mapped reads – High alignment precision and low reference bias – Fast execution	– Requires variation information from target populations – Some options are proprietary – Requires a license to unlock multi-threading and additional functionality	NOV1 NOV2
1=	BWA aln	– Large number of mapped reads – High alignment precision and low bias	– Slow execution	BWA1 BWA2
3	BWA mem	– Large number of mapped reads – High alignment precision and low bias – Fast execution	– Optimized for reads $\geq$70 bp	BWA8
4=	NovoAlign	– Large number of mapped reads – Fast execution	– Significant reference bias	NOV2
4=	Bowtie2	– High precision if using stringent mapQ filter – Fast execution	– Significant data loss when using stringent mapQ filter – Additional bias in analyses when using stringent mapQ filter – Low sensitivity for contaminants	BWT3 BWT4 BWT5

show that reference bias tends to increase with declining read length—i.e. as reads become shorter, those carrying alternate alleles become increasingly less likely to correctly align relative to reads bearing reference alleles. Importantly, the application of mapping quality filters improves alignment precision and reduces reference bias of the short read alignments to levels only slightly below that of longer reads. The application of UDG treatments resulted in further improvements in mapping precision and reference bias reduction, though these gains were less substantial for the half-UDG treatment, which is currently the most widely adopted pre-treatment strategy in contemporary aDNA studies. Notably, the two UDG treatments did not result in a large reduction the total number of aligned reads, whereas the application of the mapQ $\geq$25 filter resulted in a marked reduction in the number of aligned reads that disproportionately affected shorter reads (i.e. a 7 to 10% reduction in the number of mapped $<$40 nt reads for *BWA-mem*, *BWA-aln* and *NovoAlign*, and a $>$ 20% reduction for *Bowtie2* for the same length reads; Figures S6 and S8).

Filtering regions based on CpG islands and low mappability offer further reductions to in reference bias, though improvements from the former are more limited due to comprising only a small proportion of genome. While mappability filters can reduce bias by an additional 1% on top of improvements offered by mapping quality filters and UDG treatments, this is expected to have a limited impact on human aDNA studies that restrict their analyses to a set $\sim$1 million SNPs (assayed using a set of proprietary probes), as $>$99% of these SNPs lie in regions of high mappability. Nonetheless, both CpG and mappability filters could be important for reducing biases in future aDNA studies that use full genome sequencing.

Recommendations and conclusions

The trade-off between the quality and quantity of mapped reads is ultimately important in how it impacts downstream analyses. In this study, we show that while filtering aDNA reads based on a minimum mapping quality of 25 can improve precision and reduce reference bias on an aggregate SNP-based level, it does not completely eradicate reference bias in downstream analyses, and may even exacerbate this effect (e.g. with the tested *Bowtie2* parameterizations). Accordingly, to help guide researchers in choosing alignment strategies for aDNA studies, we provide a visual summary of our core results (Figure 5a and Figures S13 and S14) and a table summary (Table 1) that outline the overall performance across the four different metrics assessed in this study (i.e. precision, SNP-based reference bias, *D* statistic, PCA) along with the execution time (Figure 5b).

For the parameterizations tested here we find most of the *BWA-aln* and two of the IUPAC-based *NovoAlign* parameterizations (including the free default implementation) offer consistently high precision and negligible levels of bias across the two different mapping quality filters and for untreated (Figure 5b) and UDG-treated reads (Figures S13 and S14), suggesting that these methods should generally provide robust results in aDNA studies. Notably, the *NovoAlign* parameterizations ran at least 10 times faster than *BWA-aln* counterparts (Figure 5b), providing a significant speed advantage. However, free use of *NovoAlign* is currently restricted to the default parameterization in single-threaded mode (*NOV1* in the current study), limiting exploration of other potential improvements, and reference bias-free implementation requires the creation of a suitable IUPAC-augmented reference genome. The generation of an IUPAC reference requires the input of variant positions based on polymorphisms from a genetic reference panel, which may not be available for many species. When multiple panels are available it also remains unclear how the choice of reference panel, along with decisions on which SNPs to use, impacts the final quality of the alignments. In the present study, we have incorporated all known variants from the 1000 GP, whereby every SNP in our three samples is represented in the IUPAC reference. Consequently, our results likely represent the minimum reference bias that is achievable for *NovoAlign* using an IUPAC reference, and the results in aDNA studies may not reach similarly low levels.

BWA-mem shows similar levels of performance to both *BWA-aln* and *NovoAlign* and has rapid run times comparable to the latter. However, *BWA-mem* is not optimized for reads shorter than 70 nt which may be prohibitive for highly degraded samples, and also has comparatively low sensitivity for DNA coming from closely related organisms which may further impact analyses if such contamination is present but not properly accounted for. Taken together, if time is not a limiting factor for analyses, our results indicate the widely adopted *BWA-aln* parameterizations remain the best choice for most aDNA studies. *BWA-mem* is a good alternative option in cases where time constraints apply and DNA contamination from closely related taxa is not present or can be accounted for, providing that the majority of reads have lengths that exceed 70 nt, otherwise the free *NovoAlign* parameterization could be used if a suitable IUPAC reference is available or can be generated.

The IUPAC results in the present study highlight the potential for improvements when incorporating known genetic variation into the alignment process, which was recently demonstrated through the use of graph genome representation in aDNA alignments (outperforming standard *BWA-aln* strategies; [43]). Variation graphs and other graph pangenomes offer a promising alternative mapping strategy to reduce reference bias and increase alignment precision [16]. However, graph genome technologies remain in an early stage of development and currently lack a universally adopted graph coordinate system [44, 45]. A community-wide effort is required to port all existing data to a new coordinate space, along with annotations and associated metadata. In the meantime, linear reference genomes will likely remain the default alignment strategy for aDNA studies, although polymorphism-augmented linear reference genomes clearly merit further research as this technology matures.

Finally, post-alignment processing steps also have a documented effect on variant calling and are sensitive to the choice of mapping software and depth of read coverage [32]. Commonly used post-alignment procedures such as Genome Analysis Toolkit (GATK) [33], ANGSD [46] or more purpose-built toolkits such as ATLAS are used inconsistently across the aDNA field. Accordingly, benchmarking studies are sorely needed to document the efficacy of all processing steps for aDNA to provide a more complete and in-depth understanding of aDNA bioinformatic pipelines.

Key Points

- We systematically evaluated the performance of 30 read mapping strategies using simulated short ancient DNA sequencing reads against the human reference genome.
- All strategies exhibited some level of reference bias, with *BWA-aln* and *NovoAlign* paired with an IUPAC reference performing the best (least amount of bias).
- When accounting for runtime (CPU cost) *NovoAlign* with IUPAC reference offers the best compromise.

Supplementary Data

Supplementary data are available online at *Briefings in Bioinformatics*.

Acknowledgements

We are grateful to Joshua Schmidt, Thomas Litster and two anonymous reviewers for their invaluable suggestions that greatly improved the rigour and scope of our study. We thank the Novocraft Technologies team for providing access to their proprietary NovoAlign software during this project and for helpful discussions on implementation and data interpretation.

Funding

This work was supported by the Australian Research Council [ARC DP190103705]. AO was supported by an ARC PhD Scholarship [ARC IN180100017]. RT was supported by an ARC DECRA Fellowship [ARC DE190101069]. AC was supported by an ARC Laureate Fellowship [ARC FL140100260]. BL was supported by an ARC Future Fellowship [ARC FT170100448].

Data availability

The scripts used to create the used datasets in this study are available in the github repository at: https://github.com/AdrienOliva/Benchmark-aDNA-Mapping.

References

1. Bao S, Jiang R, Kwan W, *et al.* Evaluation of next-generation sequencing software in mapping and assembly. *J Hum Genet* 2011;**56**:406–14.
2. Morozova O, Marra MA. Applications of next-generation sequencing technologies in functional genomics. *Genomics* 2008;**92**:255–64.
3. Li H, Durbin R. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics* 2009;**25**:1754–60.
4. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. *Nat Methods* 2012;**9**(4):357–9.
5. Ma X, Shao Y, Tian L, *et al.* Analysis of error profiles in deep next-generation sequencing data. *Genome Biol* 2019;**20**:50.
6. Günther T, Nettelblad C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet* 2019;**15**:e1008302.
7. Brandt DYC, Aguiar VRC, Bitarello BD, *et al.* Mapping bias overestimates reference allele frequencies at the HLA genes in the 1000 genomes project phase I data. *G3* 2015;**5**:931–41.
8. Ros-Freixedes R, Battagin M, Johnsson M, *et al.* Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genet Sel Evol* 2018;**50**:64.
9. Chen L, Liu P, Evans TC Jr, *et al.* DNA damage is a pervasive cause of sequencing errors, directly confounding variant identification. *Science* 2017; **355**:752–6
10. Nielsen R, Paul JS, Albrechtsen A, *et al.* Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* 2011;**12**:443–51.
11. Orlando L, Gilbert MTP, Willerslev E. Reconstructing ancient genomes and epigenomes. *Nat Rev Genet* 2015;**16**:395–408.
12. Schubert M, Ginolhac A, Lindgreen S, *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* 2012;**13**:178.
13. Ginolhac A, Vilstrup J, Stenderup J, *et al.* Improving the performance of true single molecule sequencing for ancient DNA. *BMC Genomics* 2012;**13**:177.
14. Rohland N, Harney E, Mallick S, *et al.* Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos Trans R Soc Lond B Biol Sci* 2015;**370**:20130624.
15. Poullet M, Orlando L. Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes. *Front Ecol Evol* 2020;**8**:105.
16. Martiniano R, Garrison E, Jones ER, *et al.* Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biol* 2020;**21**:782755.
17. Peyrégne S, Slon V, Mafessoni F, *et al.* Nuclear DNA from two early Neandertals reveals 80,000 years of genetic continuity in Europe. *Sci Adv* 2019;**5**:eaaw5873.
18. Rand KD, Grytten I, Nederbragt AJ, *et al.* Coordinates and intervals in graph-based reference genomes. *BMC Bioinformatics* 2017;**18**:263.

19. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol* 2020;**21**: 265.
20. Consortium T. 1000 GP, the 1000 genomes project Consortium. A global reference for human genetic variation. *Nature* 2015;**526**:68–74.
21. Mullaney JM, Mills RE, Pittard WS, *et al.* Small insertions and deletions (INDELs) in human genomes. *Hum Mol Genet* 2010;**19**:R131–6.
22. Renaud G, Hanghøj K, Willerslev E, *et al.* Gargammel: a sequence simulator for ancient DNA. *Bioinformatics* 2017;**33**:577–9.
23. Lazaridis I, Patterson N, Mittnik A, *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* 2014;**513**:409–13.
24. Olalde I, Allentoft ME, Sánchez-Quinto F, *et al.* Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. *Nature* 2014;**507**:225–8.
25. Briggs AW, Stenzel U, Meyer M, *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res* 2010;**38**:e87.
26. Harney É, May H, Shalem D, *et al.* Ancient DNA from chalcolithic Israel reveals the role of population mixture in cultural transformation. *Nat Commun* 2018;**9**:3336.
27. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* 2013;
28. Prüfer K, Racimo F, Patterson N, *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* 2014;**505**:43–9.
29. Taron UH, Lell M, Barlow A, *et al.* Testing of alignment parameters for ancient samples: evaluating and optimizing mapaping parameters for ancient samples using the TAPAS tool. *Genes* 2018;**9**:157.
30. Cahill JA, Heintzman PD, Harris K, *et al.* Genomic evidence of widespread admixture from polar bears into Brown bears during the last ice age. *Mol Biol Evol* 2018;**35**: 1120–9.
31. Cornish-Bowden A. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* 1985;**13**:3021–30.
32. Tian S, Yan H, Kalmbach M, *et al.* Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics* 2016;**17**:403.
33. Van der Auwera GA, Carneiro MO, Hartl C, *et al.* From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. *Curr Protoc Bioinformatics* 2013;**43**:11.10.1–33.
34. Slon V, Mafessoni F, Vernot B, *et al.* The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature* 2018;**561**:113–6.
35. Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 2005;**437**:69–87.
36. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. *PLoS Genet* 2006;**2**:e190.
37. Li H, Handsaker B, Wysoker A, *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;**25**:2078–9.
38. Petr M, Vernot B. Kelso J. admixr - R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics* 2019;**35**:3194–5.
39. Seguin-Orlando A, Korneliussen TS, Sikora M, *et al.* Paleogenomics. Genomic structure in Europeans dating back at least 36,200 years. *Science* 2014;**346**:1113–8.
40. Plassais J, Kim J, Davis BW, *et al.* Whole genome sequencing of canids reveals genomic regions under selection and variants influencing morphology. *Nat Commun* 2019;**10**:1489.
41. Durand EY, Patterson N, Reich D, *et al.* Testing for ancient admixture between closely related populations. *Mol Biol Evol* 2011;**28**:2239–52.
42. Green RE, Krause J, Briggs AW, *et al.* A draft sequence of the Neandertal genome. *Science* 2010;**328**:710–22.
43. Valenzuela D, Norri T, Välimäki N, *et al.* Towards pangenome read alignment to improve variation calling. *BMC Genomics* 2018;**19**:87.
44. Li H, Feng X, Chu C. The design and construction of reference pangenome graphs with minigraph. *Genome Biol.* 2020;**21**:265.
45. Paten B, Novak AM, Eizenga JM, *et al.* Genome graphs and the evolution of genome inference. *Genome Res* 2017;**27**:665–76.
46. Korneliussen TS, Albrechtsen A, Nielsen R. ANGSD: analysis of next generation sequencing data. *BMC Bioinformatics* 2014;**15**:356.

Additional evaluation shows that specific BWA-aln setting still outperform BWA-mem for ancient DNA data alignment

Statement of Authorship

Title of Paper	Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment
Publication Status	☒ Published □ Accepted for Publication □ Submitted for Publication □ Unpublished and Unsubmitted work written in manuscript style
Publication Details	Oliva, A., Tobler, R., Llamas, B. and Souilmi, Y. (2021), *Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment*. Ecol Evol, 11: 18743-18748. https://doi.org/10.1002/ece3.8297

Principal Author

Name of Principal Author (Candidate)	Adrien Oliva		
Contribution to the Paper	Designed the experiment, performed data processing, analysed and interpreted the data, wrote the manuscript		
Overall percentage (%)	80		
Certification:	This paper reports on original research I conducted during the period of my Higher Degree by Research candidature and is not subject to any obligations or contractual agreements with a third party that would constrain its inclusion in this thesis. I am the primary author of this paper.		
Signature		Date	02/03/2022

Co-Author Contributions

By signing the Statement of Authorship, each author certifies that:

i. the candidate's stated contribution to the publication is accurate (as detailed above);

ii. permission is granted for the candidate in include the publication in the thesis; and

iii. the sum of all co-author contributions is equal to 100% less the candidate's stated contribution.

Name of Co-Author	Bastien Llamas		
Contribution to the Paper	Contributed and edited the manuscript		
Signature		Date	02/03/2022

Name of Co-Author	Raymond Tobler		
Contribution to the Paper	Contributed and edited the manuscript		
Signature		Date	02/03/2022

Name of Co-Author	Yassine Souilmi		
Contribution to the Paper	Contributed and edited the manuscript		
Signature		Date	02/03/2022

Received: 7 August 2021 | Revised: 7 October 2021 | Accepted: 15 October 2021
DOI: 10.1002/ece3.8297

Ecology and Evolution Open Access WILEY

LETTER TO THE EDITOR

Additional evaluations show that specific *BWA-aln* settings still outperform *BWA-mem* for ancient DNA data alignment

1 | INTRODUCTION

Xu et al. (2021) recently suggested a new parameterization of *BWA-mem* (Li, 2013) as an alternative to the current standard *BWA-aln* (Li & Durbin, 2009) to align ancient DNA sequencing data. The authors tested several combinations of the *-k* and *-r* parameters to optimize *BWA-mem's* performance with degraded and contaminated ancient DNA samples. They report that using *BWA-mem* with *-k* 19 *-r* 2.5 parameters results in a mapping efficiency comparable to *BWA-aln* with *-l* 1024 *-n* 0.03 (i.e., a derivation of the standard parameters used in ancient DNA studies; (Schubert et al., 2012)), while achieving significantly faster run times.

We recently performed a systematic benchmark of four mapping software (i.e., *BWA-aln*, *BWA-mem*, *NovoAlign* (http://www.novocraft.com/products/novoalign), and *Bowtie2* (Langmead & Salzberg, 2012)) for ancient DNA sequencing data and quantified their precision, accuracy, specificity, and impact on reference bias (Oliva et al., 2021). Notably, while multiple parameterizations were tested for *BWA-aln*, *NovoAlign*, and *Bowtie2*, we only tested *BWA-mem* with default parameters.

Here, we use the alignment performance metrics from Oliva et al. to directly compare the recommended *BWA-mem* parameterization reported in Xu et al. with the best performing alignment methods determined in the Oliva et al. benchmarks, and we make recommendations based on the results.

2 | METHODS

We investigated the alignment performance of the parameterization recommended by Xu et al. (2021), that is, *-k* 19 and *-r* 2.5 (hereafter called BWA9) against several of the best performing strategies identified in Oliva et al. (namely, BWA1, BWA2, BWA8, Novo1IUPAC, Novo2IUPAC, and Novo2, see Table 1 for parameter settings).

Following the analytical framework of Oliva et al., our benchmark is based on simulated reads (including fragmentation, damage, and sequencing errors typical for ancient DNA samples; see (Oliva et al., 2021)) that were generated for each of the following three samples from the 1000 Genome Project dataset (1000 Genomes Project Consortium et al., 2015), each coming from a distinct population, and were aligned to reference genome GRCh37:

- Individual *HG00119* from the British in England and Scotland population; GBR; labeled Europe in this study.
- Individual *NA19471* from the Luhya population in Webuye, Kenya; LWK; labeled Africa in this study.
- Individual *HG00513* from the Han Chinese population in Beijing, China; CHB; labeled East Asia in this study.

In addition to quantifying read alignment precision (i.e., the proportion of correctly aligned reads relative to all aligned reads) and proportion of aligned reads (i.e., the fraction of aligned reads relative to the total number of simulated reads) for each strategy, we tested the specificity (i.e., the fraction of unmapped reads) of these strategies for two sets of potential contaminants—that is, bacterial and dog reads—that were also used in Oliva et al. (2021).

3 | RESULTS

BWA9 had a slight improvement in the proportion of total reads aligned relative to *BWA-mem* using default settings (BWA8), but this came at the cost of consistently lower precision (Figure 1, Figures A1 and A2). These precision differences are particularly accentuated for reads between 30 and 60bp, the range of read lengths that is

TABLE 1 Different alignment parameterizations tested. To simplify comparisons with the results reported in Oliva et al. (2021), we reuse the alignment strategy labels from that study

Strategy	Software	Parameterization
BWA1	*BWA-aln*	-l 1024 -n 0.01 -o 2
BWA2	*BWA-aln*	-l 1024
BWA8	*BWA-mem*	default
BWA9	*BWA-mem*	-k 19 -r 2.5
Novo1IUPAC	*NovoAlign*	-k
Novo2(IUPAC)[a]	*NovoAlign*	default

[a]Used with and without the IUPAC reference (Novo2 and NovoIUPAC).

typical of ancient DNA. As demonstrated here and in more detail in our recent alignment software benchmark (Oliva et al., 2021), *BWA-aln* (BWA1 and BWA2) is the most precise alignment method among the tested strategies, having moderately higher precision relative to *BWA-mem* for shorter reads while mapping a much higher percentage of reads overall (Oliva et al., 2021; van der Valk et al., 2021).

When comparing specificity against potential contaminants, BWA9 has a near-identical specificity to the default *BWA-mem* parameterization (BWA8) for dog reads, and slightly poorer specificity when testing with bacterial reads, but both parameterizations perform considerably worse than the tested *NovoAlign* (Novo1IUPAC, Novo2IUPAC, and Novo2) and *BWA-aln* (BWA1 and BWA2) strategies for dog reads (Figure 2).

Finally, comparing running times of the two *BWA-mem* parameterizations for each of the three simulated human datasets showed that BWA9 is slightly quicker than BWA8 (Figure 3), confirming the results of ref. (Xu et al., 2021).

4 | CONCLUSION

Xu et al. (2021) report that *BWA-mem* produces alignment results that are comparable to a derivation of *BWA-aln* widely used in the ancient DNA field. Consequently, they recommend the use of a specific non-default *BWA-mem* parameterization for ancient DNA studies because of its superior runtime relative to *BWA-aln*. However, we find that this parameterization actually decreases alignment precision relative to *BWA-mem* using default settings for sequencing reads shorter than 70 bases, which are particularly abundant in ancient DNA samples. Moreover, *BWA-mem* is consistently outperformed by *BWA-aln* under the tested parameterizations for both precision and the proportion of reads mapped, and also had greatly improved specificity when the DNA contamination came from a phylogenetically related organism (i.e., a dog in the present study). Crucially, Oliva et al. have demonstrated that improvements in these alignment metrics are also complemented by a reduction in

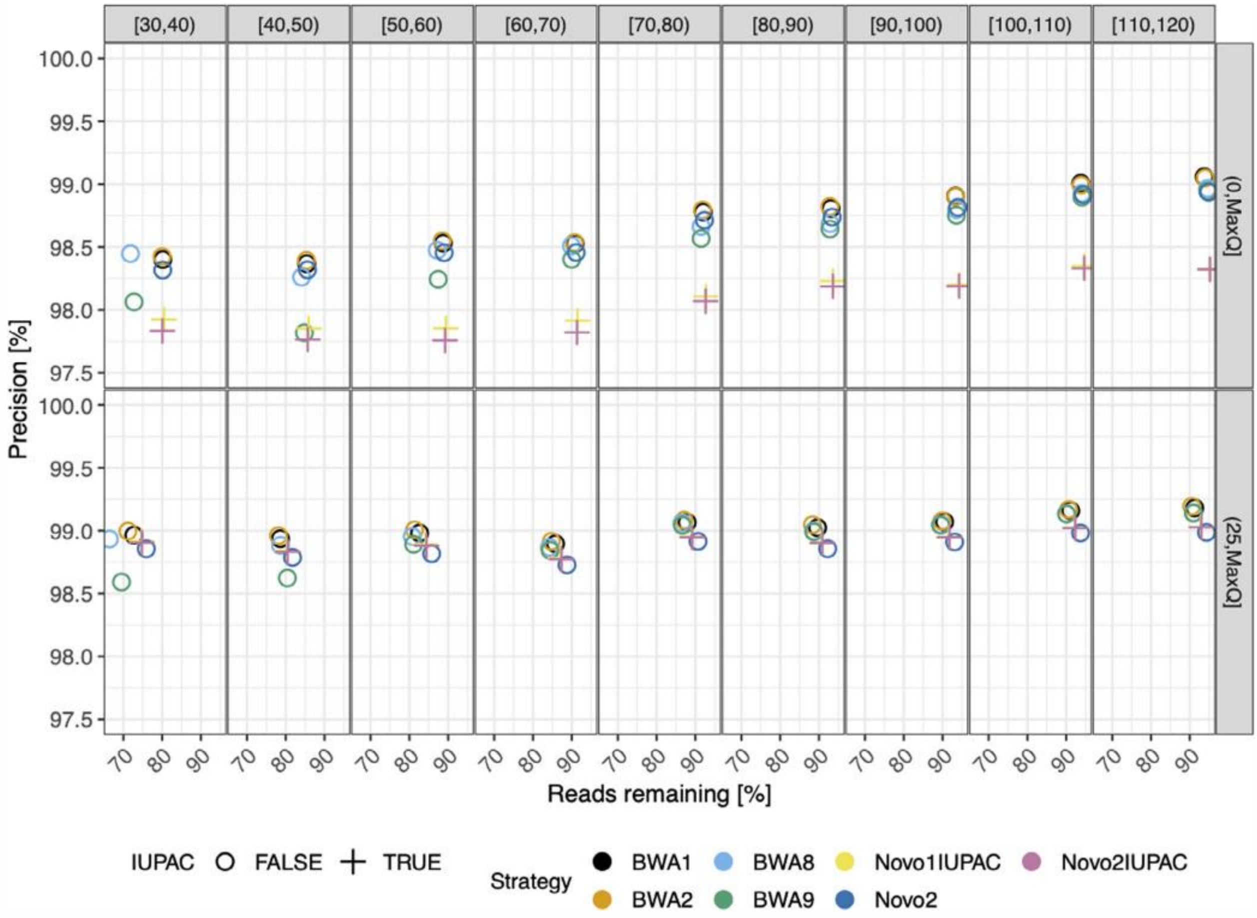


FIGURE 1 Alignment precision relative to read length and mapping quality for the simulated East Asian sample. Results are shown for seven parameterizations of four different alignment software, including an IUPAC reference-based alignment for a subset of the *NovoAlign* parameterizations (see key). BWA9 is the *BWA-mem* strategy recommended by Xu et al. (2021), with parameter details for the other strategies provided in Table 1. The panels in each row show results after applying the specific mapping quality filter, which results in the removal of all reads below the required mapping quality. Results were similar for the simulated European and African samples and are shown in Figures A1 and A2, respectively

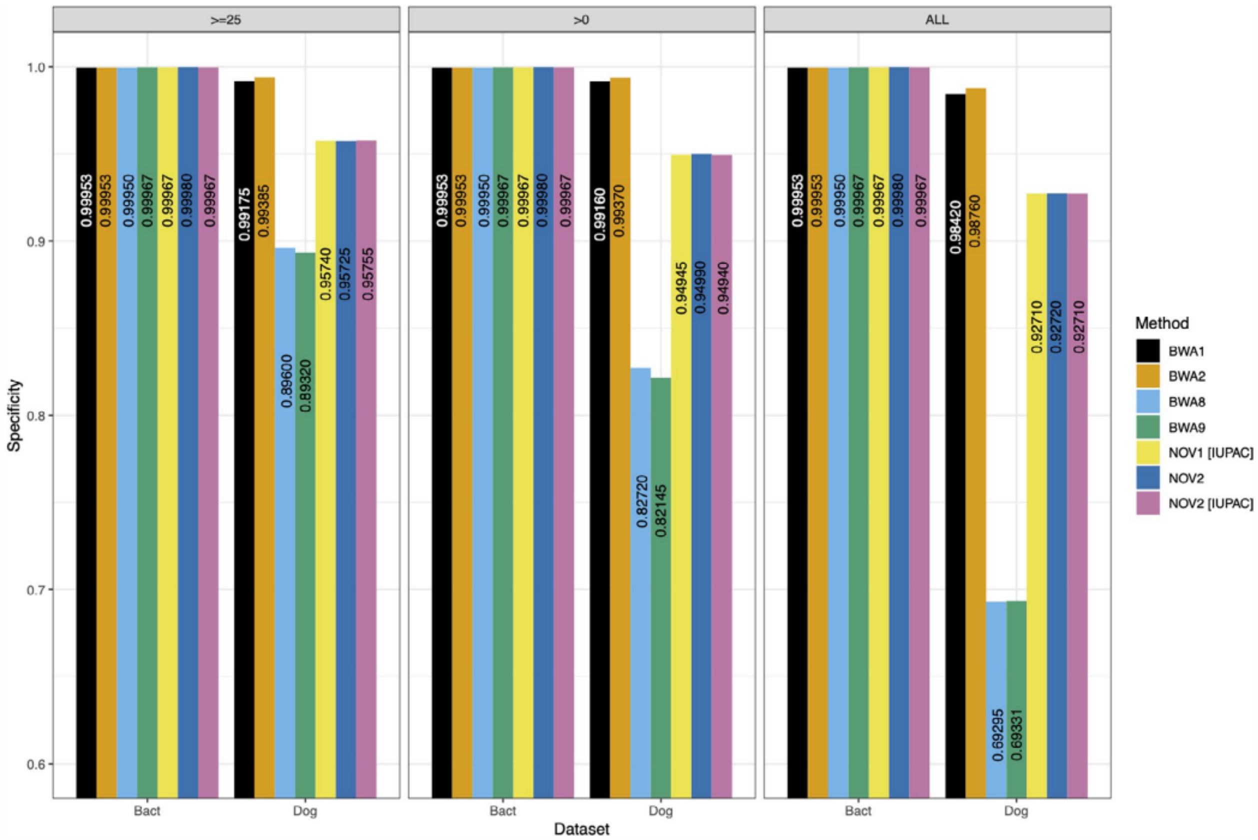


FIGURE 2 Specificity of all tested alignment methods. Bacterial and dog reads were aligned to the GRCh37 reference using the seven tested parameterizations of four different alignment software, including an IUPAC reference-based alignment for a subset of the *NovoAlign* parameterizations (see key). The specificity corresponds to the number of unmapped reads, with higher values being better

reference genome bias—an alignment-related bias resulting from the preferential mapping of alleles on the reference genome (relative to alternate alleles) that can inflate false positives and is particularly problematic for ancient DNA studies.

The recommendations of Xu et al. (2021) are based on the lack of statistical differences between the alignment performance of *BWA-mem* and *BWA-aln* evaluated using a repeated-measures ANOVA approach, whereby they recommend *BWA-mem* because of its faster execution. However, when re-examining the alignment performance results reported in supplementary table 4 of Xu et al. (2021) we find that *BWA-aln* maintains a small but consistent advantage over *BWA-mem* across different levels of contamination for both tested alignment metrics (see Figure A3)—a result that is consistent with the findings in the present study using complementary metrics. Indeed, the lack of statistical support for the difference between the two alignment algorithms most likely results from the effect size being small relative to the variance observed across the tested replicates used in the Xu et al. study (see Figure 2 in Xu et al., 2021), leading to insufficient power to detect these differences.

Taken together, our results indicate that the *BWA-aln* strategies tested here provide a small but consistent improvement over the *BWA-mem* parameterization recommended by Xu et al. (2021) for simulated aDNA read datasets when evaluated using the complimentary sets of metrics employed in Xu et al. (2021) and the present study. Importantly, while the differences between the two alignment methods are relatively small (on the order of 0.1–0.5%; see Figure 1), they are sufficient to inflate reference bias in downstream analyses that can negatively impact inferences (Oliva et al., 2021).

Accordingly, despite having improved run times, we do not recommend that *BWA-mem* be prioritized over *BWA-aln* for research using short reads—such as ancient DNA, cell-free DNA, and forensic research fields. If run time is an issue for researchers, we recommend the use of *NovoAlign* using the free default parameterization, so long as an appropriate IUPAC reference can be generated. Readers interested in a more detailed discussion of these issues are directed to refs. (Oliva et al., 2021; Poullet & Orlando, 2020; Schubert et al., 2012; van der Valk et al., 2021) for recent benchmarks of different alignment strategies using short reads.

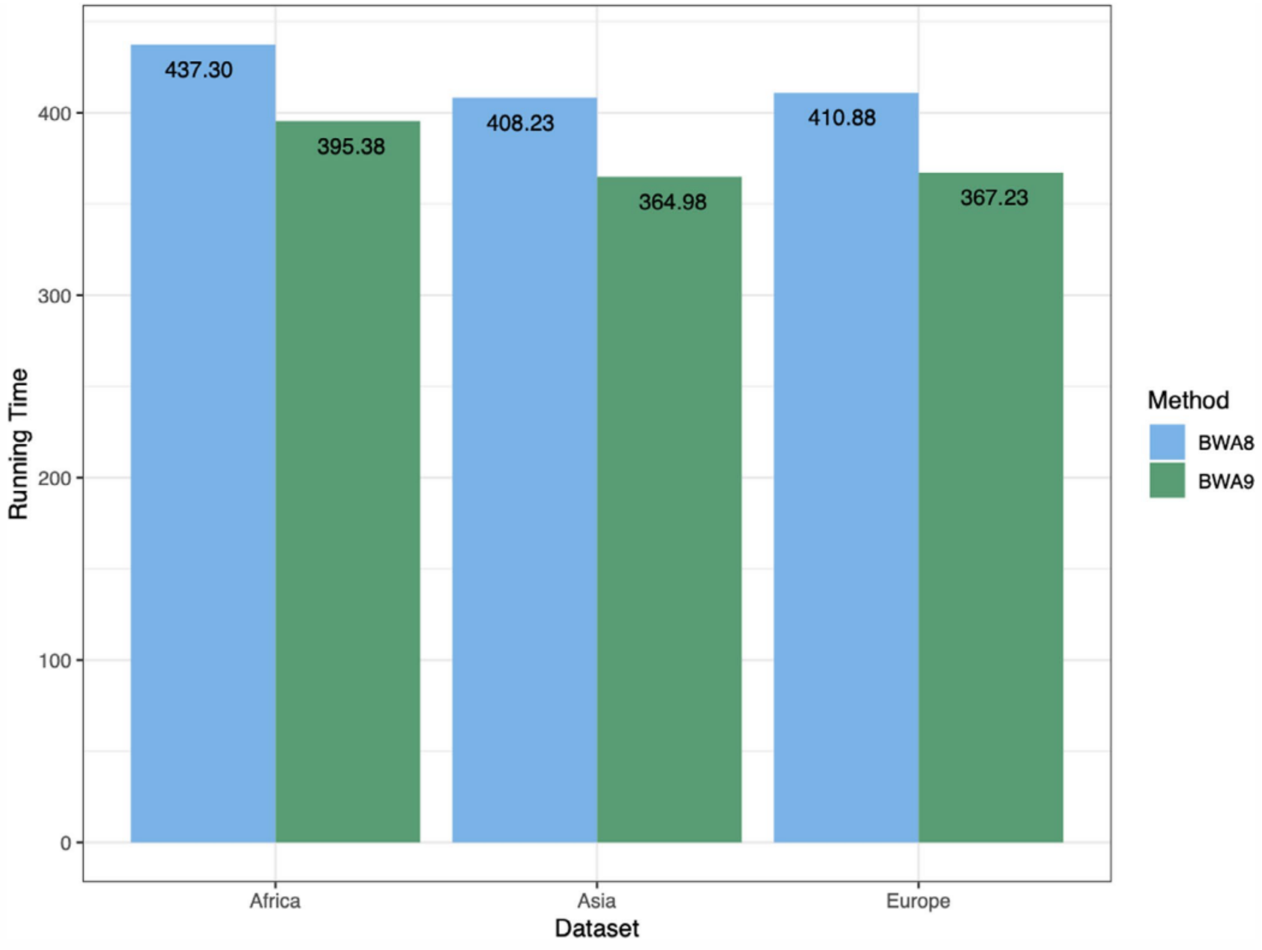


FIGURE 3 Execution time for each of the *BWA-mem* strategies. The execution time (walltime) in seconds of BWA8 (default parameters) and BWA9 (Xu et al., 2021 parameterization; -k 19 -r 2.5) based on 1.5 million simulated reads

ACKNOWLEDGMENTS

We thank Stefan Prost and an anonymous reviewer for their feedback and valuable comments. We thank the Novocraft Technologies team for providing access to their proprietary *NovoAlign*.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Adrien Oliva: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Writing-original draft (equal). **Raymond Tobler:** Supervision (equal); Writing-review & editing (equal). **Bastien Llamas:** Supervision (equal); Writing-review & editing (equal). **Yassine Souilmi:** Conceptualization (equal); Supervision (equal); Writing-review & editing (equal).

FUNDING INFORMATION

This work was supported by the Australian Research Council [ARC DP190103705], AO was supported by an ARC PhD Scholarship [ARC IN180100017], RT was supported by an ARC DECRA Fellowship [ARC DE190101069], BL was supported by an ARC Future Fellowship [ARC FT170100448]

DATA AVAILABILITY STATEMENT

The scripts used to create the used datasets in this study are available in the github repository at: https://github.com/AdrienOliva/Benchmark-aDNA-Mapping.

Adrien Oliva[1]
Raymond Tobler[1,2]
Bastien Llamas[1,2,3]
Yassine Souilmi[1,2,3]

[1]*Australian Centre for Ancient DNA, School of Biological Sciences, Faculty of Sciences, The University of Adelaide, Adelaide, South Australia, Australia*
[2]*The Environment Institute, Faculty of Sciences, The University of Adelaide, Adelaide, South Australia, Australia*
[3]*National Centre for Indigenous Genomics, Australian National University, Canberra, Australian Capital Territory, Australia*

Correspondence

Yassine Souilmi, Australian Centre for Ancient DNA, School of Biological Sciences, Faculty of Sciences, The University of Adelaide, Adelaide SA 5005, Australia.
Email: yassine.souilmi@adelaide.edu.au

ORCID

Adrien Oliva https://orcid.org/0000-0001-9950-7220
Raymond Tobler https://orcid.org/0000-0002-4603-1473
Bastien Llamas https://orcid.org/0000-0002-5550-9176
Yassine Souilmi https://orcid.org/0000-0001-7543-4864

REFERENCES

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*, 68–74. https://doi.org/10.1038/nature15393

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*, 357–359. https://doi.org/10.1038/nmeth.1923

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. arXiv [q-bio.GN]. https://arxiv.org/abs/1303.3997v2

Oliva, A., Tobler, R., Cooper, A., Llamas, B., & Souilmi, Y. (2021). Systematic benchmark of ancient DNA read mapping. *Briefings in Bioinformatics*, *22*, bbab076. https://doi.org/10.1093/bib/bbab076

Poullet, M., & Orlando, L. (2020). Assessing DNA Sequence alignment methods for characterizing ancient genomes and methylomes. *Frontiers in Ecology and Evolution*, *8*, 105. https://doi.org/10.3389/fevo.2020.00105

Schubert, M., Ginolhac, A., Lindgreen, S., Thompson, J. F., Al-Rasheid, K. A. S., Willerslev, E., Krogh, A., & Orlando, L. (2012). Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics*, *13*, 178. https://doi.org/10.1186/1471-2164-13-178

van der Valk, T., Pečnerová, P., Díez-del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., … Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, *591*, 265–269. https://doi.org/10.1038/s41586-021-03224-9

Xu, W., Lin, Y., Zhao, K., Li, H., Tian, Y., Ngatia, J. N., Ma, Y., Sahu, S. K., Guo, H., Guo, X., Xu, Y. C., Liu, H., Kristiansen, K., Lan, T., & Zhou, X. (2021). An efficient pipeline for ancient DNA mapping and recovery of endogenous ancient DNA from whole-genome sequencing data. *Ecology and Evolution*, *11*, 390–401. https://doi.org/10.1002/ece3.7056

APPENDIX

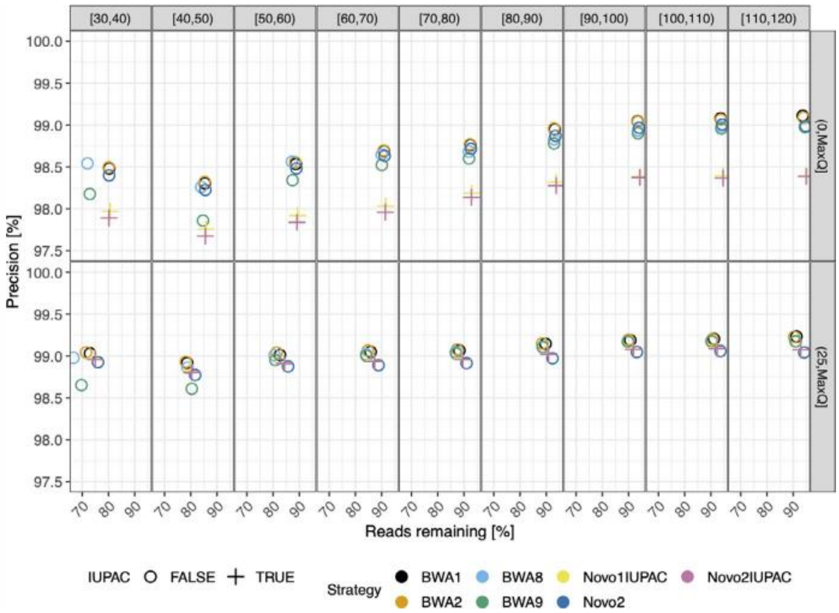


FIGURE A1 Alignment precision relative to read length and mapping quality for the simulated European sample. See Figure 1

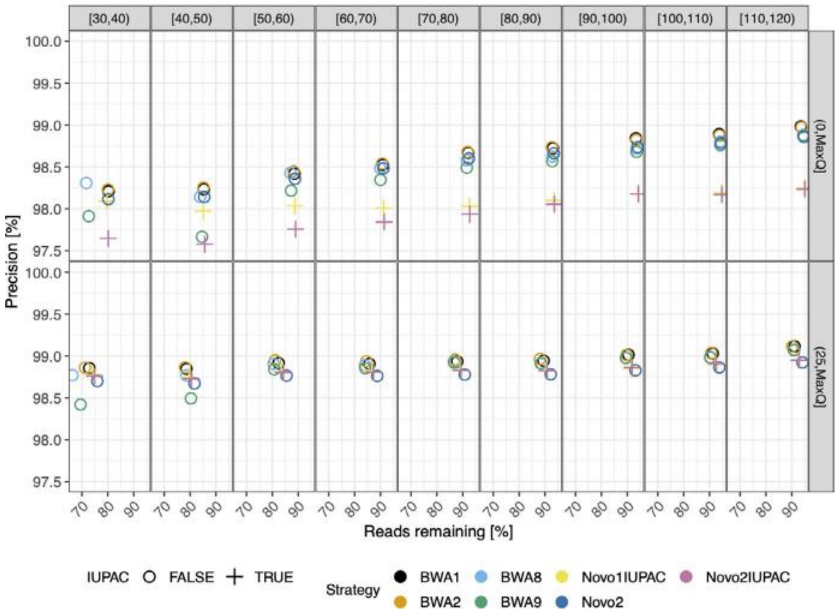


FIGURE A2 Alignment precision relative to read length and mapping quality for the simulated African sample. See Figure 1

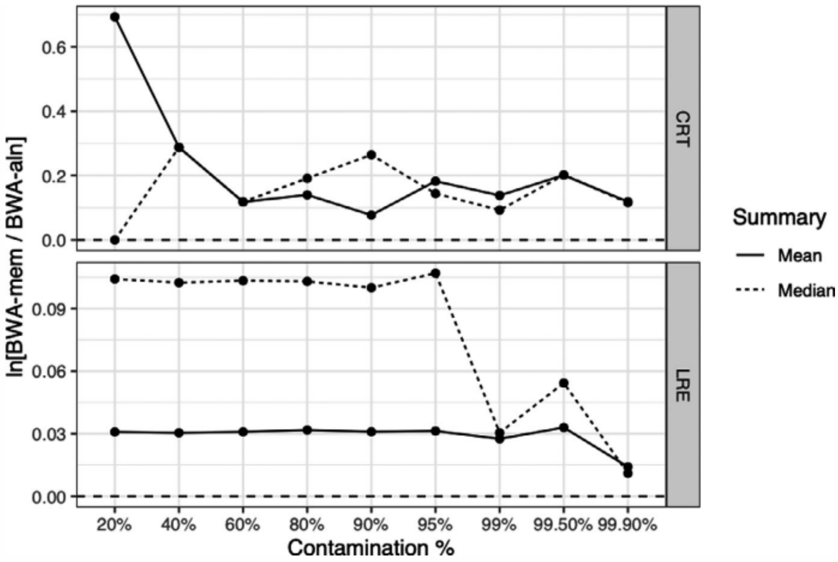


FIGURE A3 Summary of alignment performance of *BWA-mem* relative to *BWA-aln* across increasing levels of contamination from results reported in Xu et al. supplementary table 4. Xu et al. (2021) summarise alignment results using two statistics: (1) the contamination rate after treatment (CRT; top panel), which measures the proportion of aligned contaminants relative to all contaminant reads, and (2) the loss rate of endogenous DNA (LRE; bottom panel), which records the proportion of unmapped endogenous reads relative to all endogenous reads. Notably, the reported mean and median values for these both metrics are consistently higher for *BWA-mem* relative to *BWA-aln* – as shown by the natural logarithm of the ratio of *BWA-mem* to *BWA-aln* being consistently above 0 (dashed line) for both metrics – indicating that *BWA-mem* tends to map more contaminant reads and less endogenous reads than *BWA-aln* across all tested contamination levels, whereby *BWA-mem* has poorer overall performance

Chapter 2: Impact of variation graphs on population genetic inferences.

Introduction

In recent years, developments in high throughput sequencing and improvements in ancient DNA methodologies have facilitated powerful paleogenomic investigations of the human population genetic history [41,148,149], including the migrations that led to the worldwide distribution of contemporary populations [150–155]. These studies have been instrumental in supporting the dispersal of anatomically modern humans from Africa into Eurasia around 50-60 kya (i.e., the Out-of-Africa migration) and subsequent migrations involving descendant populations that gave rise to all modern Non-African populations [156,157].

Paleogenomic research has also revealed several more contentious findings that remain the topic of ongoing discussion [68,158–161]. While Europe and Eurasia have been the major focus of paleogenomic investigations [162], one intriguing example is the apparent genetic relationship between specific indigenous populations in the Americas and modern Australo-Melanesian groups. This signal was first highlighted in 2015 and again in four subsequent publications that included a number of new populations, suggesting that the signal may be quite widely dispersed [163–165]. However, another recent study [166] did not find any evidence or a genetic relationship between Australasians in any of the ancient South Americans when testing a subset of these populations, and suggested that the signal may be an artefact [167].

The properties of aDNA, such as short DNA fragments, limited coverage and the existence of damage in the recovered material [60,66] can impose subtle biases that reduce confidence in conclusions drawn from paleogenomic studies. One particularly prevalent technical problem is "reference bias," which arises because reads with non-reference alleles are less likely to map than those carrying reference alleles, leading to an overrepresentation of reference alleles in downstream analysis. Reference bias is particularly problematic in paleogenomic research and may impact downstream analysis if not accounted for (see Chapter 1).

A recent study highlighted the potential of variation graphs to help mitigate potential reference biases in paleogenomic studies [95] however, very little is otherwise known about the performance of graph-based methods relative to standard linear alignment techniques in an aDNA context. In this chapter, I compare a recent graph-based alignment approach to the two linear aligners with the lowest reference bias and best statistical performance from Chapter 1 of this thesis; i.e., *BWA-aln* and *NovoAlign.* Specifically, I test the impact of different alignment methods on the inference of *D*-statistics that have been used to support specific population genetic relationships that have been regarded as contentious [71,140] by some researchers.

Methods

Datasets

In this section, I outline the specific population genetic relationships that will be examined (also see Table 1).

Dated 40 thousand years old, the *Tianyuan* specimen [160] sampled from a cave in the the North-East of China, has been found to be more closely related to present East Asians than Europeans and is thought to represent the earliest ancestor of East Asians that is not shared by modern Europeans. However, *Tianyuan* displays genetic affinities with the West European *Goyet* specimen (Belgium, dated at 35ka) but not other Palaeolithic European samples, implying that the separation between early Europeans and early Asians was not a single split and that this genetic separation must have been earlier than 40ka [159].

The linkage of an individual approximately 24,000-year-old, from *Mal'ta* in south-central Siberia will be used to verify multiple findings. First, it has been discovered that the *Mal'ta* population is basal to present-day Western Eurasian; it was also observed that it was close to the root of most Native Americans without having a close affinity with East Asians [158].

Then, the *Yana* population, a Siberian community of 31 thousand years old (31ka) [161] is also part of that study and its relationship with Europeans and South Americans will be tested. Indeed, it has been found that this population have genetic similarities with present day Northern Eurasian and American, unlike the *Tianyuan* population.

Relationships of an Eastearn European population, represented by the *Vestonice16* [*159*] individual (30 ka), will also be used. Indeed, despite the fact that cultural similarity was found between the *Vestonice* and *Mal'ta* population, no genetic connection was observed between the European and the Siberian population [159].

One of the most intriguing and interesting relation that will be examined is the evidence of excess affinity of the present day South American (*Surui* and *Karitiana*) to Australasian populations found in *Skoglund et al.* [164] To investigate those relationships, different present day populations will be used such as the *Surui* population [164] (Brazil), *Karitiana* population [74,168] (Brazil), the *Papuan* population [164,169] (New Guinea), *Mixe* population [164,169] (Mexico) and *Mbuti* population [168,169] (Congo) as an outgroup.

Another relationship to verify is the relationship of the *Tianyuan* population toward the *Mixe* and *Surui* populations. It has been shown that the *Tianyuan* and *Mixe* populations are more closely related to each other than to the *Surui* population.

All the aforementioned ancient and present day individuals are reported in the 1240K (v44.3) dataset (https://reich.hms.harvard.edu/) and in the SGDP dataset [168]; see Table S1 for more details about the samples within each population.

In this chapter, I will verify multiple tests (see Table 1), shown to be marginally significant in previous studies [74,158–161,164,168,169], using a subset of mapping software strategies that are less prone to reference bias.

Variation graph creation

The variation graph was built with the *Variation-graph* "vg" [132] software (version vg.1.27.1) and the *hs37d5* linear-reference as a starting point. Variants from 1000 GP phase 3 [106] are then integrated, and variants with less than 0.1% Minor Allele Frequency (MAF) are filtered out, as indicated by *Martiniano et al.* [95],

Following methods in *Oliva et al.*[140], an IUPAC linear reference using the IUPAC is created using the function called "*IUPAC"* in *NovoAlign* (version v3.09.01) toolkit: "*NovoUtil*". The same linear reference is used to create the variation graph and the IUPAC linear reference *hs37d5*. The same variants and thresholds have been used for both references. However, only Single Nucleotide Polymorphisms (SNPs) have been used to create this IUPAC reference.

Pre-alignment processing

To remove any remaining adapters and obtain homogeneous data, all collected data was run through the *fastp* software (https://github.com/OpenGene/fastp) (version 0.20.1). Reads less than 30 base-pairs (bp), as reported in my first chapter, are not worth studying since they introduce more noise than longer reads. To get more solid and reliable findings, any reads shorter than 30bp are then ignored (*--length required 30*) as anything shorter would be unreliable to utilise [140]. Following the methodologies employed in many research in the field, the default base quality (baseQ) threshold is modified from 15 to 20 [160] (-q 20). These various threshold numbers are important as they provide limits that exclude undesired data from our dataset, such as exogenous data, reads that are too short, reads with poor quality, and so on. These unwanted data might provide wrong findings by mapping the incorrect place in the genome and then studying non-real variants. Those threshold values must also consider other factors such as avoiding discarding potentially useful endogenous data by being too strict; especially in aDNA case where there is poor coverage.

Read alignment

Sequencing reads for all individuals used in this study were mapped using the following three separate mapping methodologies:

1. **Using *BWA-aln* [72] (version 0.7.15) alignment software and the *hs37d5* as the linear reference.** Multiple studies [71,140] have found a specific set of parameters (called *BWA1* [140] in chapter 1) that has the best performance for ancient DNA datasets. The parameters are as follows: ***-l 1024*** to disable the seeding, ***-n 0.01*** to reduce the edit distance (i.e., allow less mismatches) and ***-o 2*** for the open gaps parameter.
2. **Using *NovoAlign* (http://www.novocraft.com/products/novoalign) and the IUPAC-augmented *hs37d5* linear reference (see “Reference creation”).** All individuals are aligned using the default *NovoAlign* option that is available for free public use. When using an IUPAC reference [140], the default version of *NovoAlign* has been demonstrated to perform equally well as the “Base Quality calibration” option (**-k**) that is recommended for use with an IUPAC reference.

3. **Using the *Variation-graph* "vg" [132] software and the previously built variation graph (see "Reference creation").** The function "*vg map"* is used to map reads using the default parameters. To allow direct comparison with other alignment tools, the resulting graph alignment is surjected (i.e., collapsing the graph alignment to a conventional linear reference) onto a linear reference using the *--surject* option.

All the different mapping strategies have been processed on the HPC1 cluster from the University of Adelaide. The cluster consists of 48 skylake nodes for a total of 1920 physical cores; hyperthreading is enabled by default, this resulted in up to 3840 logical cpus being available.

Post-alignment processing

The preceding alignment can be improved by undergoing multiple steps of post-processing that can mitigate the impact of various biases [170–173]. PCR duplicates are eliminated using *MarkDuplicates* from Picard Tools (version 2.21.8; https://broadinstitute.github.io/picard). The GATK (version 3.7) [174] function *IndelRealigner* is then used to realign readings around known indels.

SNP sets and pseudo-haploid calls

Researchers employ the pseudo-haploid calls approach to account for the scarcity of data typical in the aDNA studies, which causes poor coverage and complicates standard genotype calls. Aligned data from all examined samples were processed with the same pseudo-haploid call approach using *Samtools* (v1.9) and *sequenceTools* (v1.2.2). Because the poor quality bases are already filtered using *fastp* (see "Pre-alignment processing"), *BaseQ* filtering is deactivated (**-B**). Following results from chapter 1 (published [140,175]), only reads with a mapping quality (*mapQ*) greater than 0 were saved (**-q 1**). Pseudo-haploid calls were made at positions covered by the 1240k SNP variants set (https://reich.hms.harvard.edu/), which is currently the most widely used variant set in human paleogenomics.

Results

For each of the tested population quartets, *D*-statistics were estimated for the three different approaches (i.e., *BWA-aln*, *NovoAlign*, *vg*; collectively called “*Retested”* results) and compared to the original results (i.e., the statistic reported in the original publications, henceforth called "*Original*" results). While none of the *Retested* results were identical with the *Original* results (which is not unexpected, as one or more alignment, processing, or variant calling steps likely differ between the *Retested* and *Original* results), they were all strongly correlated with the original data (Figure 1) and each other (Pearson’s *r* between 0.996 and 0.998 across all pairwise comparisons). This indicates that the *Retested* *D*-statistics do not systematically differ from the *Original* results; however, the different alignment methods may have more subtle impacts on inference that could impact interpretations.

To investigate the direct impact of alignment method choice on the *D*-statistics, I ranked the absolute *D*-statistic for all three methods separately across each of the 33 evaluated population quartets (standardised *D*-statistics, i.e., *Z*-scores, being used in each case). By comparing the distribution of these ranks across all methods, I tested if any of the three tested alignment methods were consistently more or less conservative in assigning *D*-statistics (noting that absolute values were prefered as the sign [i.e., -ve or +ve] is informative about which right and left populations share an excess of alleles but not the strength of the relationship). The results show that *D*-statistics from *VG* aligned reads tend to be the most conservative overall, and *BWA-aln* being the least conservative of the three alignment methods (i.e., each tending to have either the smallest or largest absolute *D*-values, respectively, across the tested quartets), with *NovoAlign* having intermediate values (Figure 2). Notably, these results are consistent with recent findings from *Martiniano et al.* [95] , which showed that surjected (i.e., collapsing the graph alignment to a conventional linear reference) *vg* alignments tended to have less reference bias than widely used *BWA-aln* parameterisations, and also results from chapter 1 of this thesis which demonstrated that *NovoAlign* is also less biassed than *BWA-aln*. These results are consistent with reference bias leading to anticonservative *D*-statistics, potentially through the resulting subtle excess of alleles shared by pairs of left and right populations that have more reference bias than other two populations in the quartet (see Discussion).

Reproducibility and the impact of sample choice on D-statistic sensitivity

In general the differences amongst the three different alignment methods was subtle; however in some cases this resulted in differences in statistical significance (i.e., *D0, D9.2, D10.2, D12.3*; Figure 3). Further, none of the *Retested BWA-aln* results were identical to the *Original* results, despite sharing identical alignment software for many cases, which also resulted in changes in significance for some population quartets (Figure 3). The latter case suggests that changes in post-alignment processing and variant calling parameters may also lead to downstream changes in statistical inferences that may ultimately impact interpretations. This highlights the importance of accurately reporting all steps in alignment and bioinformatic pipelines to allow accurate reproduction of results (see Chapter 3).

To investigate the potential impact of subjective researcher choices on inference, I also tested if changes in the samples chosen for a particular population could also produce statistically meaningful differences in D-statistic values. The set of samples used for a particular population in paleogenomic studies might differ between studies – for instance, only a subset may be chosen based on specific genomic properties (e.g., sequencing coverage, SNP presence) – potentially leading to different results, particularly when sample sizes are small or genetically diverse. This highlights the wider issue of reproducibility, which is the focus of chapter 3 of this thesis; however, in the present chapter I concentrate on the impact of excluding particular samples on D-statistics. Specifically, I investigate sample choice on tests involving the *Karitiana*, *Mixe*, and *Papuan* populations, by recomputing the D-statistics using different combinations of samples. Because the *Karitiana* and *Mixe* have small sample sizes (n = 4 and 3, respectively), all combinations are explored that result from the removal of a single individual, with ten different combinations being tested for the larger *Papuan* dataset (n = 16) that resulted from the removal of two randomly chosen individuals.

The results confirm that different sample choices can alter the significance of *D*-statistic tests (Figure 4). In general, the impact of sample removal varied considerably across populations (but was reasonably consistent across tests), with the variability in *Z*-scores produced by downsampling amongst the *Mixe* group being surprisingly similar to the much larger *Papuan* group (post removal *n* = 2 *Karitiana* and 14 for *Papuans*) and much less variable than *Karitiana*, despite the latter having a similar number of samples (post removal *n* = 3). This suggests that *Mixe* samples were likely reasonably genetically homogeneous relative to *Karitiana* samples, resulting in more consistent *Z*-score estimates following downsampling. Nonetheless, sample removal in the *Mixe* group still had a notable impact in the form of higher variability of the *D*-

statistic (i.e., the denominator used to calculate the *Z*-score), which resulted in consistently lower scores than when all three samples are included.

The net result of sample removal across all three groups was that the initially significant *Z*-score for the *f4(Mbuti, Papuan; Mixe, Karitiana)* quartet became insignificant for all cases of the downsampled *Mixe* group and for one case of the downsampled *Karitiana* group, with a single case of the opposite scenario (i.e., reversing an initially insignificant *Z*-score) observed for the *f4(Mbuti, Tianyuan; Mixe, Karitiana)* quartet when downsampling the *Karitiana* group*.* In contrast, no cases of *Z*-score significance reversal were observed for downsampled *Papuans*. The consistency in *Z*-scores across the downsampled *Papuan* groups and similarity with the initial *Z*-score estimated from the full set of samples, suggest that this set was likely sufficiently large and genetically homogeneous to buffer the effects of sample removal.

Discussion & Conclusion

Our comparison of graph-based and two linear alignment strategies on *D*-statistic inferences emphasise the utility of aligning aDNA reads against a reference that captures population genetic variation. The most widely used alignment tool in paleogenomic research at present, *BWA-aln*, was the least conservative overall, with *VG* tending to have the most conservative *D*-statistics of the three tested aligners and *NovoAlign* having intermediate results. These results echo the patterns in reference bias reported by Martiniano and colleagues [95] and also in chapter 1 of this thesis, which show that alignments on a linear reference using *BWA-aln* produce more reference bias than alignments to genome graphs using *VG* and to IUPAC augmented reference genomes using *NovoAlign*. These results suggest that *D*-statistics may be biassed toward more extreme results (i.e., larger absolute values) when using aligners that are more prone to reference bias.

To understand what might cause this phenomenon, we note that the degree of reference bias shows a strong dependency on sample origin for reads less than 50bp in length (i.e., the majority of aDNA reads), with bias being lowest in Europeans, intermediate in Africans, and highest in Asians, in reflection to their divergence from the linear reference genome (e.g., reference bias ~ 3%, 4% and 5%, respectively, for reads <40bp; see chapter 1). Accordingly, inferring *D*-statistics for population quartets where either the left or right populations (or both) differ in their reference bias induces artefactual allele sharing between the left and right populations, potentially inflating the overall magnitude of the statistic. This should be true in general providing that alternate left and right populations do not share a historical true gene-flow event, which should drown out any artifactual signal, though further work is needed to understand how this manifests under different population contexts.

While the effects tend to be subtle (*BWA-aln* produced *D*-statistics around 5% larger in magnitude than *VG* on average), these changes occasionally resulted in differences in significance. Similarly, removal of specific samples from one of the four populations used in the *D*-statistic test can also result in reversals of significance. In the latter case the impact of sample choice was most evident when the total number of samples for a particular population was small, resulting either in large variability amongst scores that depended on which samples were used, or a consistent reduction in the resulting *Z*-score when genetic homogeneity was high amongst all samples in a particular group. These results highlight the impacts that different sample choices can have when populations are represented by a few samples, a reasonably common occurrence in paleogenetic studies.

In conclusion, reference graphs, and to a lesser extent IUPAC augmented linear references, appear to offer improved robustness for inferring population genetic relationships using the *D*-statistic than standard linear alignments, and this may well extend to other widely used paleogenomic analyses. Importantly, while the differences across alignment methods tend to be subtle, they are sometimes sufficient to result in reversals in significance, impacting interpretations when the results are taken at face value. Accordingly, it is the researchers responsibility to accurately report all processing and analytical choices in their studies as these choices can produce inferences that may support different conclusions. More often than not, researchers follow the correct procedure and do not rely on a single D-statistic comparison to reach a big conclusion, especially when the finding pertains to a group with just a few samples.

Table 1:

All relationships tested with the corresponding test ID. Tests where the significance has changed are highlighted in bold text.

Test ID	Corresponding *D*-statistics computed
D0	**D(GoyetQ116-1, Vestonice16; Tianyuan, Mbuti)**
D1	D(Tianyuan, Mal'ta; GoyetQ116-1, Mbuti)
D2	D(Tianyuan, GoyetQ116-1; Mal'ta, Mbuti)
D3	D(GoyetQ116-1, Mal'ta; Tianyuan, Mbuti)
D4.1	D(GoyetQ116-1, Surui; Tianyuan, Mbuti)
D4.2	D(Mal'ta, Surui; Tianyuan, Mbuti)
D5.1	D(Tianyuan, GoyetQ116-1; Surui, Mbuti)
D5.2	D(Tianyuan, Mal'ta; Surui, Mbuti)
D6.1	D(Tianyuan, Surui; GoyetQ116-1, Mbuti)
D6.2	D(Tianyuan, Surui; Mal'ta, Mbuti)
D7	D(Mal'ta, Tianyuan; GoyetQ116-1, Mbuti)
D8.1	D(Tianyuan, GoyetQ116-1; Papuan, Mbuti)
D8.2	D(Tianyuan, Mal'ta; Papuan, Mbuti)
D9.1	D(GoyetQ116-1, Papuan; Tianyuan, Mbuti)
D9.2	**D(Mal'ta, Papuan; Tianyuan, Mbuti)**
D10.1	D(Tianyuan, Papuan; GoyetQ116-1, Mbuti)
D10.2	**D(Tianyuan, Papuan; Mal'ta, Mbuti)**
D11.1	D(Yana1, Tianyuan; Surui, Mbuti)
D11.2	D(Yana1, GoyetQ116-1; Surui, Mbuti)
D11.3	D(Yana1, Mal'ta; Surui, Mbuti)
D12.1	D(Yana1, Surui; Tianyuan, Mbuti)
D12.2	D(Yana1, Surui; GoyetQ116-1, Mbuti)
D12.3	**D(Yana1, Surui; Mal'ta, Mbuti)**
D13.1	D(Tianyuan, Surui; Yana1, Mbuti)
D13.2	D(GoyetQ116-1, Surui; Yana1, Mbuti)
D13.3	D(Mal'ta, Surui; Yana1, Mbuti)
D14	D(Surui, Mixe; Tianyuan, Mbuti)

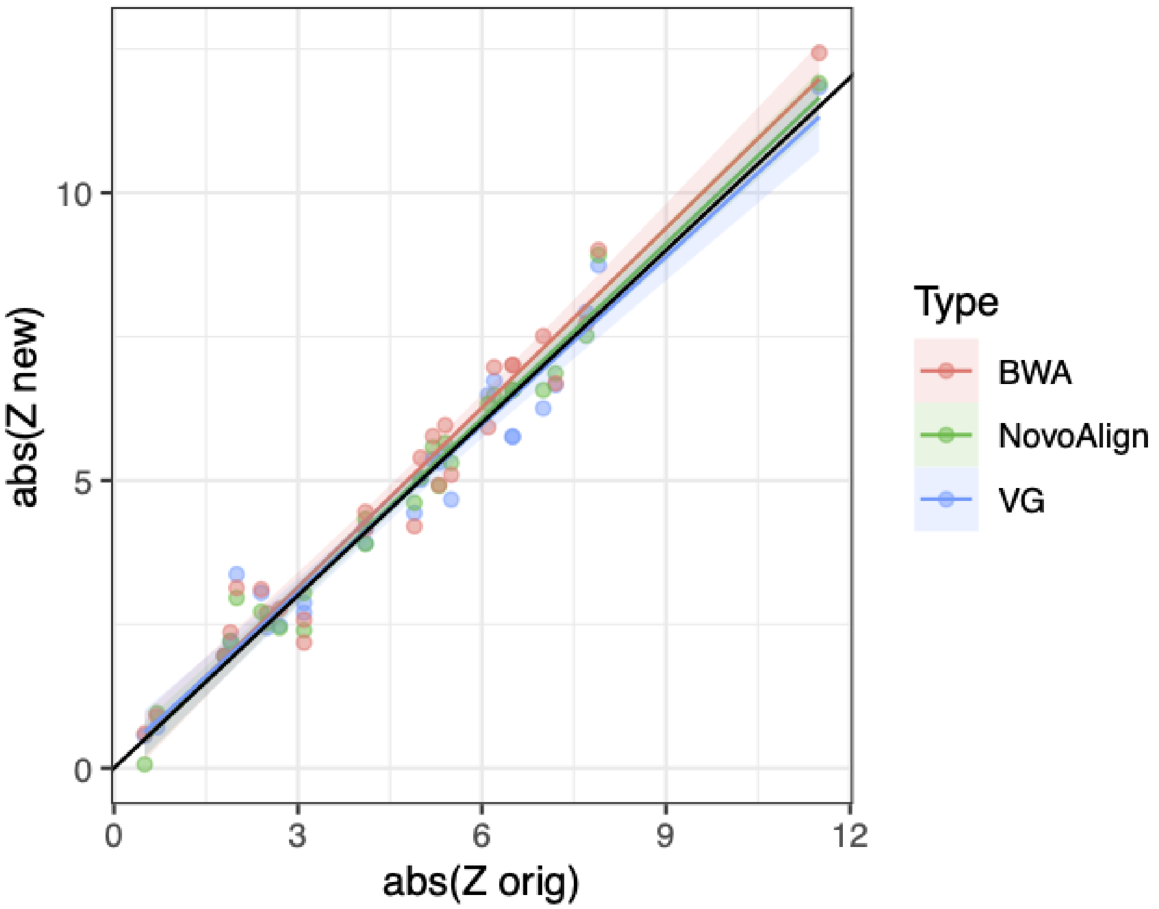


Figure 1. Relationships between the Original and Retested D-statistics.

Z-scores were calculated for each population quartet and their absolute values regressed against the reported values. As expected, strong positive linear correlations exist between *Retested* and *Original* values (>0.996 in each case, with similar correlations amongst the *Retested* results), though there were no cases where scores were perfectly replicated due to differences in processing, sample choice, or other unidentified factors.

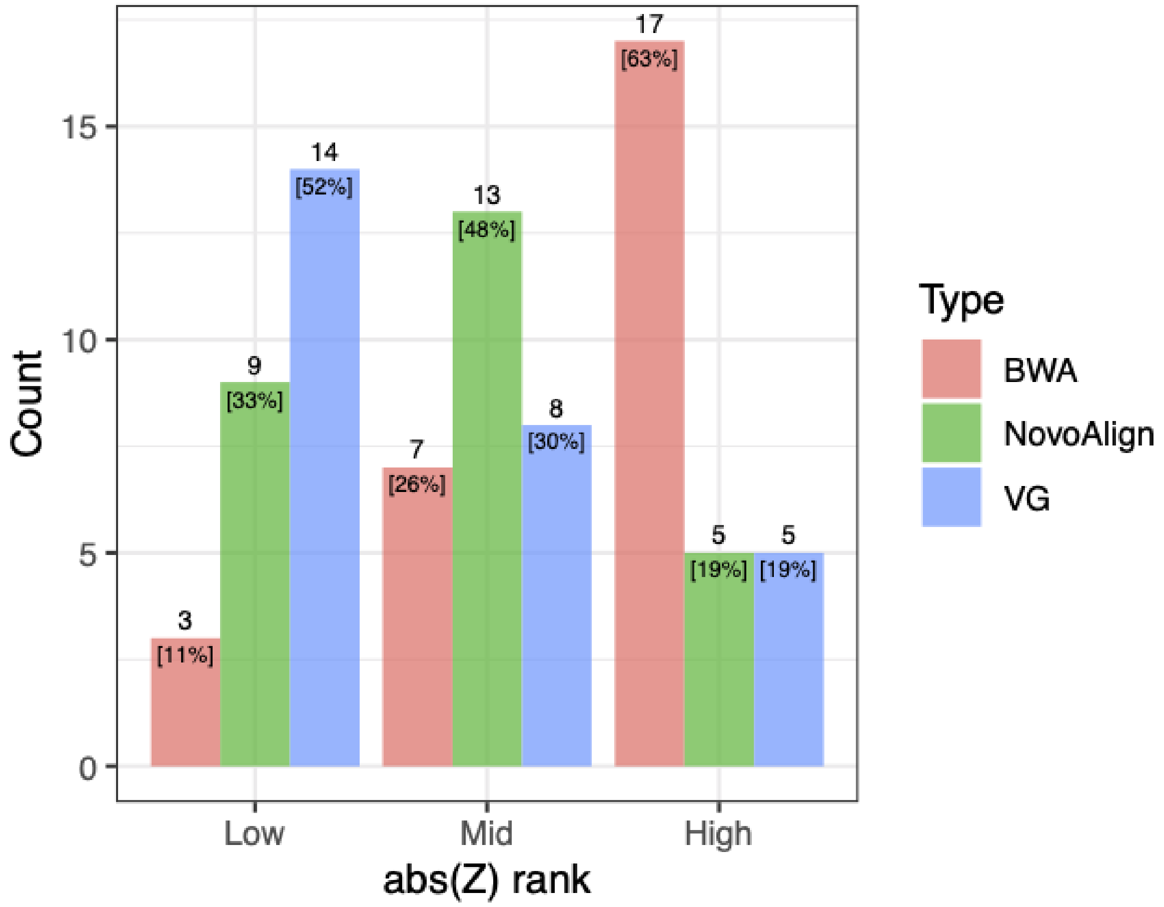


Figure 2. Augmenting reference genomes with population genetic variation leads to conservative *D*-statistics. For each population quartet, the *Z*-scores were computed for each alignment method and then ranked according to their magnitude (i.e., absolute value; Low = minimum rank, Mid = intermediate rank, High = maximum rank). The *vg* method tended to have the most conservative *Z*-scores, followed by *NovoAlign* (using an IUPAC augmented reference) and *BWA-aln*. Numbers and percentages at the top of each bar indicate the count and relative frequency of a given rank for each alignment method. Note that *Novoalign* and *vg* had equal lowest rank for one population quartet and were both assigned to Mid rank in this instance.

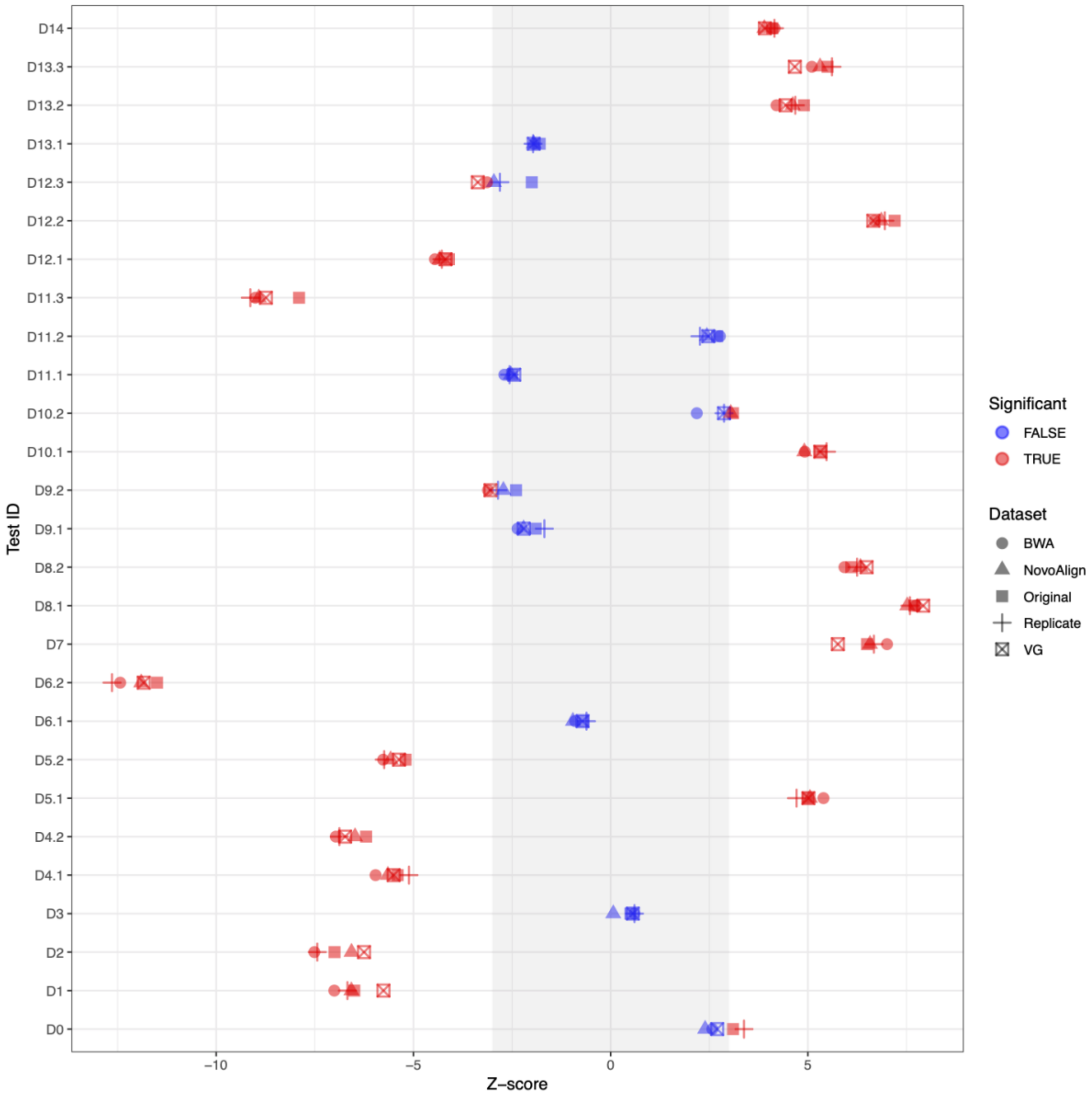


Figure 3. *Z*-score of all *D-statistics* calculated using the 1240K dataset across the different alignment software evaluated. The Z-scores values above or below 3 (grey zone) imply a significant excess of shared alleles between the populations of interest in that specific test. The colour of the dot does also signal this excess of shared alleles (red). The shape of each dot represents each new software in the Retested dataset *(BWA-aln*, *NovoAlign*, *vg*), the Original results (reported results) or Replicate (results calculated using the AADR dataset v44.3).

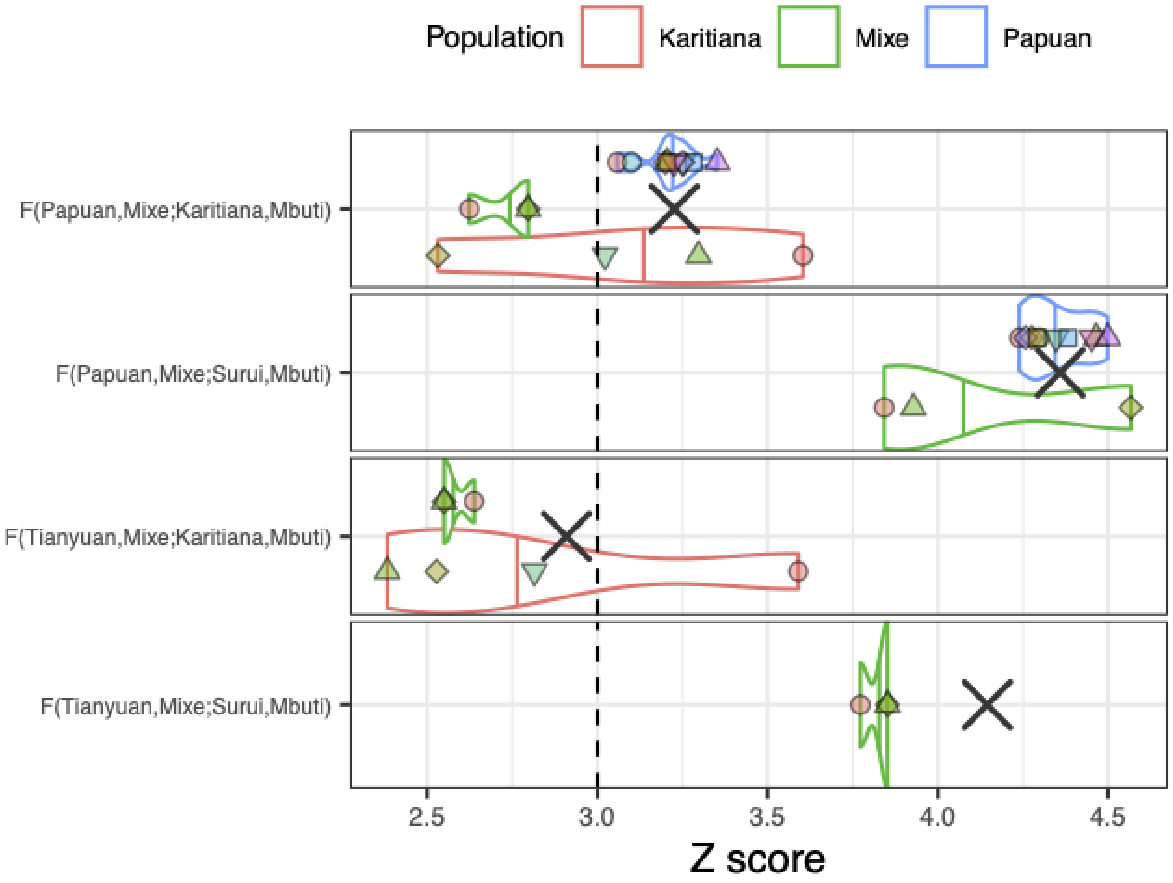


Figure 4. F4 statistics calculated by subsetting individuals from different populations using the 1240K dataset. For the F4-score, the significance is computed as it is for the *D*-statistics in previous plots: values above 3 (dotted line) imply a significant excess of shared alleles between the first population listed from each of the left and right pairs. The colour of each violin represents the downsampled population, and each distinct excluded individual(s) within that population is represented by a different shape. The results obtained without downsampling any population are represented by the black cross for each test.

Chapter 3: Fundamental steps towards reproducible paleogenomic studies

Introduction

The powerful historical lens provided by paleogenomic research has led to new interpretations of the past—which can have far-reaching implications for present-day societies and institutions (e.g., politics [176], policymaking [177], health [178,179], wildlife conservation [56]) — emphasising the need for paleogenomicists to implement and adhere to ethical, accurate, and reproducible research standards. However, while attempts have been made by paleogenomicists to define baseline ethical principles for future human paleogenomic research [96,180–182] , explicit guidelines pertaining to minimum standards for reproducible aDNA data analyses to ensure FAIR (i.e., findable, accessible, interoperable, and reusable [183]) research in the paleogenomic era have yet to emerge.

To date, publications on reproducible aDNA research have been limited to recommendations for aDNA fieldwork and laboratory best practices [81,82,85,86,184,185]. While the need for a complementary set of computational best practices for aDNA sequence data has previously been recognised [96,186], existing guidelines focus on modern DNA sequence data [187–191] and only a few [44,96] consider practicalities particular to aDNA data processing and analysis. In this chapter I identify five key areas in aDNA data processing and analysis where improvements are needed and provide recommendations for best practises in data handling and reporting for each of these steps: 1) DNA library preparation, 2) sequence data processing, 3) variant calling, 4) genetic dataset merging, and 5) population assignment and population genetic analyses. The main focus is to elaborate and encourage the adoption of specific paleogenomic research practices that ensure reproducibility and facilitate automated aDNA reprocessing and meta-analysis (see Table 1 for a concise summary).

Table 1:

<table>
<tr><td colspan="2"><i>For each of the following steps, it is recommended that:</i>
● <i>The software used is clearly specified and the exact version reported</i>
● <i>All parameters used must be reported, including default parameters</i>
● <i>Custom scripts should be version controlled, and the exact version used provided</i>
● For stochastic systems, seeds should be provided if available
● <i>Output files provided as Supplementary Information should be reported in a machine-readable format (log files, reports, tables, etc) to facilitate reproducibility and enable meta-analyses</i>
● <i>README file(s) documenting the rationale behind the parameterization used should be provided</i></td></tr>
<tr><th>Step</th><th>Report</th></tr>
<tr><td><b>Sequencing reads processing</b></td><td>● Raw deconvoluted fastQ files
● All details about filtering steps</td></tr>
<tr><td><b>Reads alignment</b></td><td>● Reference genome used
● BAM files</td></tr>
<tr><td><b>Alignment enhancing</b></td><td>● Report the steps applied (trimming, duplicates removal, indel realignment, etc) and the cut-off values used</td></tr>
<tr><td><b>Filtering on mapped reads</b></td><td>● Report any filtering applied (read length cut-off, mapping quality threshold, unmapped reads, secondary alignments, etc)</td></tr>
<tr><td><b>Variant calling</b></td><td>● Make the genotype calls publicly available where ethically appropriate
● When necessary, report the genotyped positions, the exact reference system coordinates, and the alleles targeted</td></tr>
<tr><td><b>Merging datasets</b></td><td>● Report the number of SNPs before <u>and</u> after merging for each population
● If new scripts are used to merge dataset, make them publicly available (e.g., GitHub, GitLab)
● Provide a table with individuals within each population studied</td></tr>
<tr><td><b>Population genetic tests</b></td><td>For each test performed, the following metrics should be reported:
● Standard error
● <i>D-score / f-score</i>
● <i>Z-score</i> and significance threshold
● Number of informative SNPs
● Set of chromosomes used
● All the metrics should be provided in a <i><u>parsable</u></i> log file.</td></tr>
</table>

DNA preparation and sequencing

In paleogenomic research, samples are prepared using different molecular techniques tailored to the particulars of each sample. These techniques range from different extraction protocols [192–196] that depend on the sampled tissue (e.g., bone, tooth, hair), preservation conditions [197], enzymatic treatment of postmortem damage [66], and the amount of endogenous DNA contained in the sample. Subtle differences in sample preparation and subsequent sequencing strategies often require tweaking of downstream data processing steps to account for the known technical biases. While knowledge of these steps are not needed to reproduce results when working from the resulting sequence data, they are essential for motivating appropriate processing and analytical choices, and therefore all steps should be accurately reported.

DNA sequence data processing

Pre-alignment processing

Processing of demultiplexed aDNA reads typically involves multiple steps to improve subsequent sequence alignment and ultimately mitigate biases in downstream analyses. Standard pre-alignment processing steps can include minimum and maximum read length thresholds (*fastp* [198], *seqkit* [198,199]), removal of adaptor sequences and barcodes (e.g., *AdapterRemoval2* [200], *fastp* [198] , *leeHom* [201]), trimming poly-A or poly-G sequences from read termini (*GATK* [202], *fastp* [198]) , base quality rescaling [66], and quality-based read filtering (*fastp* [198], *seqkit* [198,199]), to exclude potentially problematic reads. Because these initial processing steps impact all subsequent processing steps and analyses, accurate documentation of the exact steps performed – including the software and their versions, and parameter settings – is of the utmost importance. Ideally, all pre-alignment processing steps would be performed using a flexible and repeatable bioinformatic pipeline (e.g., *EAGER* [91], *ATLAS* [78], *Paleomix* [203]) and the configuration files would be provided to allow exact replication of these steps from the demultiplexed raw reads.

Additionally, because technological developments will likely mean that new trimming and filtering methods will eventually emerge and become standard in paleogenomic research, it is recommended to make the unfiltered demultiplexed reads available (such as refs. [204] and [205]), thereby avoiding potential scenarios where obsolete software are no longer available or usable. Demultiplexed reads can be made available on multiple platforms such as the *Sequence Read Archive* (SRA) or the *European Nucleotide Archive* (ENA).

Read alignment and post-alignment processing

The next step in standard paleogenomics projects involves alignment of suitably processed reads to a species-specific reference genome. As multiple choices of reference genome exist for many species, it is essential to report the version of the reference used for read alignment. Similarly, multiple software tools are available to align DNA sequence reads to the chosen reference (e.g., *BWA-mem* [206], *BWA-aln* [72], *Bowtie2* [207], *NovoAlign* http://novocraft.com/), each having parameters that can be manually altered by users. Different alignment tools and parameterisation choices are known [71,140,175] to impact read alignments and can also alter downstream analysis, making full reporting of all options essential. Multiple benchmarks for aDNA read alignment have been published that provide guidelines to choose the software and parameterisations best suited to the users' requirements [71,140,208].

Post-alignment filtering and processing steps are another standard component of aDNA bioinformatic workflows that are used to mitigate biases and improve analytical performance [170,171]. Several filtering criteria are commonly used, including minimum mapping quality thresholds (*MapQ*), removal of unmapped reads, and managing secondary read alignments. Post-alignment processing of aDNA alignment files typically involves detection and removal of duplicated reads (*Picard MarkDuplicates* [http://broadinstitute.github.io/picard/], *SAMtools rmdup* [209], *DeDup* [91], *Biohazard-Tools* [https://github.com/mpieva/biohazard-tools]), indel realignment (*GATK* [202], *BCFtools* [210],), and base quality recalibration (*ATLAS* [78], *mapDamage* [66]). Repeating these steps requires that all software, options, and threshold values are reported.

We advocate that researchers make demultiplexed raw reads available. However, we acknowledge that read mapping is a tedious process that may not need repeating. In this case, read alignment files and relevant details about processing steps must be provided.

Genotyping and Variant Calling

Due to poor sequencing coverage, most paleogenomic studies apply pseudo-haploidisation techniques to identify variants for subsequent analyses [211]. Importantly, pseudo-haploidisation involves the random sampling of a single allele at predesignated diploid sites to create the pseudo-haploid output state. Accordingly, reproducing identical allelic states requires either fixing the random seed employed in the sampling process prior to allele sampling, or identifying it after sampling, and reporting this seed. Currently, the software available for generating pseudo-haploid samples, FrAnTK [212], *sequenceTools pileupCaller* (https://github.com/stschiff/sequenceTools), and Consensify [213], do not allow users to fix the random seed, whereby it is recommended that this functionality is included in new software and added to existing versions to facilitate further reproducibility.

When sufficient read data is available, diploid genotypes can be called using numerous suitable software[214]. In either case, the researcher's chosen variant calling software and settings need to be reported. Version information is also essential, as the outputs of certain software – such as sequenceTools (github.com/stschiff/sequenceTools), or PLINK [215] – are known to vary depending on their version.

Variant calling on a genome-wide scale (i.e., over all positions covered by the alignment) is currently rare in paleogenomic studies, with calling more commonly performed on a predesignated subset of polymorphic loci obtained from SNP capture [31]. At present, the most widely used loci come from the 1240k DNA capture assay (https://reich.hms.harvard.edu/), though extended probe sets have recently become available (e.g., [160]). Regardless of the option chosen, positional information about the variant positions called from ancient genomes needs to be reported.

Most importantly, the called variants should be made publicly available using either the *Eigenstrat* (https://reich.hms.harvard.edu/software/InputFileFormats) or the *VCF* format. This is particularly important given that the pseudo-haploidisation process currently cannot be replicated without the provision of the random seed, making the reproduction of identical pseudo-haploid calls from alignment files effectively impossible at present [216–218].

Variant data merging

A common step in paleogenomic research is to merge the data produced during the study with publicly available ancient and/or modern genetic datasets. While multiple variant file merging software currently exists, the *mergeit* function from the *Eigensoft* package [219,220] is perhaps the most widely used approach. Certain parameter choices are known to alter *mergeit* outputs – for instance, using the *strandcheck* option can result in different numbers of variants in the resulting merged file – emphasising the need for these options to be recorded along with the location of remaining variants (similar to ref. [221]). Further, researchers should report software and parameterisations used in any additional steps taken to manipulate or convert the merged variant files (e.g., conversion from *Eigenstrat* to *PACKEDPED* and *VCF* formats using *PLINK* [215]). If bespoke scripts are used to convert from one format to another, these scripts should also be made available in public repositories (e.g., GitHub, Zenodo, figshare, OSF).

Recently, aggregated datasets containing ancient and modern human genomes have been made available (i.e., Allen Ancient DNA Resource, AADR; https://reich.hms.harvard.edu/allen-ancient-dna-resource-aadr-downloadable-genotypes-present-day-and-ancient-dna-data), with separate datasets available for two widely used SNP sets (i.e., 1240k SNPs assay and ~600k SNPs from the Human Origins Array). Because each dataset is periodically updated with new and resequenced ancient and modern samples, it is imperative that researchers who utilise the AADR or similar datasets record the SNP set and version number used. Regardless of the source of the datasets used in the variant file merging, researchers should report the set(s) of variants resulting from the data merging and manipulation steps following the standards outlined in the variant calling section. Crucially, researchers should also report which samples were included in the merging process: sample reporting remains a reasonably uncommon practice in paleogenomic research (see refs. [222–225]), despite the potential for sample choice to produce different sets of merged variants and alter outcomes in downstream analyses.

Population genetic tests

Paleogenomic research draws heavily upon a handful of population genetic tools to investigate past genetic and demographic history. Some of the most widely used statistical frameworks are the *f*-statistics [17,103,226], which use covariance in allele frequencies to perform specific hypothesis tests concerning phylogenetic relationships amongst two, thee, or four populations (*f2*, *f3*, and *f4*/*D-statistics*, respectively). These basic statistics are often combined to perform more sophisticated tests on genetic relationships across larger phylogenies (e.g., *qpAdm*, *qpGraph*). Despite their ubiquity in paleogenomics, most publications currently only report the tested topology and the standardised results (i.e., *Z-score*) obtained from each test. To ensure reproducibility, the following information should be supplied for *each* test:

- Standard error
- Unstandardised statistics (i.e., raw *f/D-statistics*)
- *Z*-score and significance thresholds
- Number of informative SNPs
- Set of chromosomes used (e.g., autosomes only, autosomes and X chromosome)
- Additional filtering of original statistics

Complete reporting of this information is still uncommon in paleogenomics (e.g., [221]), with previously published papers often providing standardised *f*-statistics but neglecting to report the standard errors of each test (which are necessary to recalculate the original *f*-statistics; e.g., [227]). A simple and robust means for reporting results from *f*-statistics and related analyses (e.g., qpWave, qpAdm, qpGraph) is through the provision of parameter files (i.e., *parfiles*) in cases where command line implementations were used (e.g., Admixtools [17]), or by reporting the specific commands if an alternative software implementation was used (such as AdmixR [228]) along with the software version used. The same standards also apply to any additional statistical analyses performed on *f*-statistics. As word limits and table size/numbers can be a constraint for published results, a viable alternative is to provide *parsable* log files in a public repository (e.g., Dryad, figshare, github) that include the recommended information on each test used.

Conclusion

Every choice made by researchers during an empirical study can have an impact on downstream analyses, with even seemingly benign choices potentially having subtle impacts that are not always readily predictable and that can hamper reproducibility if not accurately reported. In this chapter, I have attempted to outline the minimum information required for the fledgling field of paleogenomics. Ultimately, reproducibility is effectively crippled by incomplete metadata [229,230] ideally should be provided in a machine-readable format so that it can be easily integrated in standard computational pipelines [203,231]. As clearly emphasised in the FAIR principles [232–234], the provision of accurate and complete metadata remains the cornerstone to conducting reproducible research, and ensuring a more robust review process in all empirical disciplines, including paleogenomics [232–234]

Although the focus of this work is compliance with the FAIR principles, we acknowledge that paleogenomic research should embed the CARE (Collective benefit, Authority to control, Responsibility, Ethics) principles for Indigenous data governance where relevant [235]. Our proposed guidelines are compatible with both the FAIR and CARE principles of data management.

Overview of the thesis

The study of ancient DNA provides a new lens through which to understand the evolution of humans and other species, and has been used to shed new light on their diets [236,237], historical demography and migrations [238] and health [236,239], amongst many other areas. Paleogenomics research ultimately depends on parsing genetic information preserved in aDNA molecules; however, aDNA is prone to multiple biases that can impede accurate recovery of genetic variation and potentially impact downstream analytical results. In order for paleogenomic researchers to generate robust results, these biases must be mitigated through either the validation of existing methods or by developing new approaches.

Altogether, this thesis sought to gain insight into three keys areas as follows:

1. Identify the best current linear mapper for ancient DNA datasets and assess the influence of reference bias on downstream analyses.
2. Characterise the impact of graph-based methods on population genetics tests compared to conventional linear approaches.
3. Establish FAIR [183] reproducibility standards and protocols in paleogenomic research.

In this section, I briefly summarise the key results from each chapter of this thesis, focusing on the major contributions to paleogenomic research. This is followed by a general discussion that synthesises these results, and outlines potential limitations and future research directions.

Summary and main findings from Chapter 1

In this chapter I presented a detailed evaluation of the current linear mapping software available for use in paleogenomic research and provide the first systematic benchmark of aDNA alignment performance since *Schubert et al.* [71] in 2012. Specifically, I provided performance benchmarks for the standard alignment approach, *BWA-aln*, along with three additional tools (*BWA-aln*, *BWA-mem*, *NovoAlign*, and *Bowtie2*), that have not previously been scrutinised in an aDNA context.

By evaluating a wide range of parameterisations, I show that *BWA-aln* remains the best performing software in terms of reducing bias in population genetic inferences, closely followed by the default NovoAlign implementation using an IUPAC augmented reference. Notably, the latter approach had slightly less reference bias than BWA-aln for short reads, though these subtle differences did not manifest in noticeable differences in the downstream population genetic inferences.

Additionally, following a recent publication recommending *BWA-mem* over *BWA-aln* in paleogenomics studies [240], I reanalysed *BWA-mem* performance under a much wider range of parameterisations and confirm that *BWA-aln* remains the superior choice for aligning aDNA reads.

Summary and main findings from Chapter 2

In this chapter I explored the comparative performance of the current state of the art in variation-aware alignment methods, i.e. using the vg aligner and a pangenome graph, against the two optimal methods from chapter 1: i.e. NovoAlign using an IUPAC augmented reference and the current gold standard linear alignment method, BWA aln.

My results show that alignment to variation-aware reference genomes results in more conservative estimates of the widely used f4 population statistic. This is consistent with reductions in reference bias resulting from the use of variation augmented reference genomes relative to standard linear alignments, as reported in chapter 1 and another recent study [95]. Moreover, while these differences are subtle, they occasionally reverse the significance of some contentious population genetics tests, with similar significance reversals being observed when using different subsets of samples for one of the four populations used in an f4 test. These results highlight the sensitivity of marginal f4 results upon sample choice and subtle changes in reference bias caused by choice of alignment method, which could potentially lead to different interpretations when taken at face value.

Summary and main findings from Chapter 3

Although reproducibility is a pillar of the scientific method, many research disciplines are currently undergoing a "replication crisis" [241,242] that is marked by the inability to reproduce original results. This lack of reproducibility has drawn the robustness and validity of initial findings into question and ultimately erodes public confidence in empirical research.

As public awareness of the problem has grown, paleogenomic researchers have made recommendations for best practices in the field and laboratory, but this is still lacking for computational and analytical aspects of the field. In an attempt to fill this gap, in my final chapter I outlined a set of requirements pertaining to the minimum information that must be reported in order to achieve fully reproducible paleogenomic results in accordance with the FAIR (findable, accessible, interoperable, and reusable [183]) definitions.

Discussion

Pervasive presence of reference bias and future directions

Despite the development of some standardised protocols for the use of aDNA [243], limitations and biases are still present and may have a significant impact on the interpretation of results and conclusions of these studies. These biases can arise at any time during sampling, sample preparation, primary data processing, and analysis. My first chapter [140,175] combines two published studies which show that reference bias is particularly strong in reads <50bp, which make up the bulk of ancient samples [94], suggesting that this bias is pervasive in paleogenomics studies.

Intriguingly, I also observed that including known variation as IUPAC characters within the linear reference reduced the reference bias when using *NovoAlign* software (the only tested software that can handle IUPAC characters). Notably, the current gold standard aligner used in paleogenomics, BWA-aln, does not support the use of IUPAC augmented reference genomes, but I anticipate that adding this functionality would reduce reference to levels beyond those reported in chapter 1. My result from chapter 2 and another recent study [95] indicates that aDNA alignments to variation graphs is likely to reduce reference bias beyond what is possible with augmented linear references due to their ability to represent more complex variation (e.g., INDELs). Accordingly, an important next step would be to broaden the systematic benchmark of chapter 1 by systematically analysing alignments of simulated aDNA reads using a variety of pangenome graphs and variation graph methods (e.g. vg [132], GraphTyper [244], Gramtools [245], PanVC [136], Seven Bridges' Graph Genome Pipeline, etc). Because variation graphs are a fusion of a linear reference and several variations, graphs with distinct sets of variants could also be investigated to assess the impact of the quantity and general properties of genetic information incorporated into the reference graph. In particular, introducing more variants opens up more "paths" for read alignments and therefore greater potential for false positives [246,247] (i.e., reads mapping to the incorrect positions). Hence, key aspects to investigate in the future would be: 1) verifying the optimal number of variants to include in a variation graph); and 2) determining if there is an optimal level of variation to include (i.e., if introduction of more variants does not lead to further reductions in bias or produces more noise than genuine alignments).

Sampling biases in population genetics

Paleogenomics can provide new insights about human history through the study of aDNA material. In addition to providing novel insights into the evolution of our species and the origin of modern human populations, it may also provide useful information for the development of personalised medicine, by helping to unravel the complex genetic ancestries captured in contemporary genomes. One of the most widely used tools to understand ancestry and population relationships in paleogenomics are the suite of f statistics, which use patterns of allele sharing to identify phylogenetic relationships and historical admixture events. However, we still know relatively little about the sensitivity of the f statistics to different choices made by researchers during data processing and analysis. In chapter 2 of this thesis, I show that choice of alignment method or sample composition can reverse the significance of an f4 test (Figure 4 in Chapter 2). The impact of sample choice was particularly problematic when the overall number of samples used for a particular population was also small and where these individuals are divergent. Given that small sample sizes are a standard feature of paleogenomic studies – where it is common for one sample to be used as a proxy for an ancient population – these results serve to caution researchers about making overly strong statements based on results pertaining to one or a few samples. Moreover, there is a need for more systematic investigation of the factors that can impact f-statistics, including those that researchers often have some choice in (such as sample selection), to make practitioners aware of potential pitfalls and to define a set of best practices where possible.

Standards in paleogenomic research

Until now, standardised reporting standards and minimum information requirements are lacking in paleogenomic research, with different studies using various formats to report their findings. Researchers in the field recognise the dire need for guidelines [44,96]. In my third chapter, I addressed and discussed in depth all of the minimum required information that must be reported in order to ensure replication of paleogenomic research. I also suggested the formats to be used when publishing the results. Notably, I proposed reporting results in a machine-readable format, including log files, and I summarised all of the information in a simple table as a convenient checklist before submitting a new manuscript.

To have a genuine influence in the field, the suggestions must be accepted and adopted by a substantial portion of the community, even if the actual suggestions might be altered. I intend to publish the suggested guidelines as FAIRa protocol [183]. This would greatly benefit paleogenomic research by allowing automated aDNA reprocessing and thereby producing more reliable and robust findings.

Additional actions can be taken to increase repeatability in the field even further. Extensive benchmarking of the different tools should be performed in order to study many aDNA-related stages that have never been thoroughly studied (e.g., comprehensive post- and pre-processing steps, variant calling and genotyping, variant filtering). Having a clearer and better understanding of each phase, as well as the best available tools and options, would help to determine the best parameters for specific aDNA datasets. While a few projects, such as *Eager* [91], are embraced as a computational methodology pipeline by the whole paleogenomics community, this has not been accomplished as of yet. New standards should arise in order for the community to eventually utilise the same pipeline, eliminating the bias that occurs when data from multiple sources is integrated and studied together.

Conclusion

Each chapter in this thesis has tackled a key issue in obtaining robust and reproducible results in paleogenomic research. I have demonstrated that reference bias has a clear impact on statistical procedures that are widely used in contemporary paleogenomic research, and have made a series of recommendations intended to assist researchers to reduce its impact in future studies. By getting insight on biases that influence the current studies and by investigating the advantages of using graphs-based methods, my thesis provides knowledge to improve the available methodological toolkit for researchers. While graphs-based methods are still under active development [248], the benefits of using such methods with aDNA are already promising [95]. Results show that it helps to overcome the shortcomings of current methods and lessen the influence of biases.

While improving the outcomes of all future research in the field would improve our understanding of past migrations and human history, graph-based methods are capable of dramatically improving variant calling [95] and can lead to biomedical insights that might inform medical research.

References

1. Charlesworth, B. & Charlesworth, D. Population genetics from 1966 to 2016. *Heredity* **118**, 2–9 (2017).
2. Lenormand, T. Gene flow and the limits to natural selection. *Trends Ecol. Evol.* **17**, 183–189 (2002).
3. Futuyma, D. J. & Moreno, G. The evolution of ecological specialization. *Annu. Rev. Ecol. Syst.* **19**, 207–233 (1988).
4. Wright, S. Evolution in Mendelian Populations. *Genetics* **16**, 97–159 (1931).
5. Darwin, C. On the origin of species, 1859. (2016).
6. Hardy, G. H. MENDELIAN PROPORTIONS IN A MIXED POPULATION. *Science* **28**, 49–50 (1908).
7. Weinberg, W. Über den Nachweis der Vererbung beim Menschen. *Wuertt. Ver. vaterl. Natkd. 64,* (1908).
8. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* **52**, 399–433 (1919).
9. Fisher, R. A. & Others. 024: On the Dominance Ratio. (1922).
10. Haldane, J. B. S. A Mathematical Theory of Natural and Artificial Selection, Part V: Selection and Mutation. *Math. Proc. Cambridge Philos. Soc.* **23**, 838–844 (1927).
11. Fisher, R. A. The Evolution of Dominance in Certain Polymorphic Species. *Am. Nat.* **64**, 385–406 (1930).
12. Schiller, F. C. S. HALDANE, JBS-The Causes of Evolution. *Mind* **41**, (1932).
13. Theodosius, D. Genetics and the Origin of Species. (1937).
14. Huxley, J. & Others. Evolution. The modern synthesis. *Evolution. The Modern Synthesis.* (1942).
15. Harris, H. C. Genetics of Man Enzyme polymorphisms in man. *Proceedings of the Royal Society of London. Series B. Biological Sciences* **164**, 298–310 (1966).

16. Provine, W. B. *The Origins of Theoretical Population Genetics: With a New Afterword*. (University of Chicago Press, 2001).
17. Patterson, N. *et al.* Ancient admixture in human history. *Genetics* **192**, 1065–1093 (2012).
18. Kalow, W. Ethnic differences in drug metabolism. *Clin. Pharmacokinet.* **7**, 373–400 (1982).
19. Wilson, J. F. *et al.* Population genetic structure of variable drug response. *Nat. Genet.* **29**, 265–269 (2001).
20. Tate, S. K. & Goldstein, D. B. Will tomorrow's medicines work for everyone? *Nat. Genet.* **36**, S34–42 (2004).
21. Tishkoff, S. A. & Kidd, K. K. Implications of biogeography of human populations for 'race' and medicine. *Nat. Genet.* **36**, S21–7 (2004).
22. Muszkat, M. Interethnic differences in drug response: the contribution of genetic variability in beta adrenergic receptor and cytochrome P4502C9. *Clin. Pharmacol. Ther.* **82**, 215–218 (2007).
23. Chung, W.-H. *et al.* Medical genetics: a marker for Stevens-Johnson syndrome. *Nature* **428**, 486 (2004).
24. Chung, W.-H., Hung, S.-I. & Chen, Y.-T. Genetic predisposition of life-threatening antiepileptic-induced skin reactions. *Expert Opin. Drug Saf.* **9**, 15–21 (2010).
25. Mancinelli, L. M. *et al.* The pharmacokinetics and metabolic disposition of tacrolimus: a comparison across ethnic groups. *Clin. Pharmacol. Ther.* **69**, 24–31 (2001).
26. Pena, S. D. J. *et al.* The genomic ancestry of individuals from different geographical regions of Brazil is more uniform than expected. *PLoS One* **6**, e17063 (2011).
27. Brody, H. & Hunt, L. M. BiDil: assessing a race-based pharmaceutical. *Ann. Fam. Med.* **4**, 556–560 (2006).
28. Sankar, P. & Kahn, J. BiDil: race medicine or race marketing? *Health Aff.* **Suppl Web Exclusives**, W5–455–63 (2005).
29. Krimsky, S. The short life of a race drug. *Lancet* **379**, 114–115 (2012).

30. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).

31. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).

32. Schiffels, S. *et al.* Iron Age and Anglo-Saxon genomes from East England reveal British migration history. *Nat. Commun.* **7**, 10408 (2016).

33. Helgason, A., Sigureth ardóttir, S., Gulcher, J. R., Ward, R. & Stefánsson, K. mtDNA and the origin of the Icelanders: deciphering signals of recent population history. *Am. J. Hum. Genet.* **66**, 999–1016 (2000).

34. Morelli, L. *et al.* Frequency distribution of mitochondrial DNA haplogroups in Corsica and Sardinia. *Hum. Biol.* **72**, 585–595 (2000).

35. Hudson, R. R. & Others. Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology* **7**, 44 (1990).

36. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).

37. Higuchi, R., Bowman, B., Freiberger, M., Ryder, O. A. & Wilson, A. C. DNA sequences from the quagga, an extinct member of the horse family. *Nature* **312**, 282–284 (1984).

38. Pääbo, S. Molecular cloning of Ancient Egyptian mummy DNA. *Nature* **314**, 644–645 (1985).

39. Green, R. E. *et al.* Analysis of one million base pairs of Neanderthal DNA. *Nature* **444**, 330–336 (2006).

40. Fu, Q. *et al.* DNA analysis of an early modern human from Tianyuan Cave, China. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 2223–2227 (2013).

41. Narasimhan, V. M. *et al.* The formation of human populations in South and Central Asia. *Science* **365**, (2019).

42. Margaryan, A. *et al.* Population genomics of the Viking world. *Nature* **585**, 390–396 (2020).

43. Ermini, L., Der Sarkissian, C., Willerslev, E. & Orlando, L. Major transitions in human

evolution revisited: a tribute to ancient DNA. *J. Hum. Evol.* **79**, 4–20 (2015).

44. Der Sarkissian, C. *et al.* Ancient genomics. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20130387 (2015).
45. Refoyo-Martínez, A. *et al.* Identifying loci under positive selection in complex population histories. *Genome Res.* **29**, 1506–1520 (2019).
46. Souilmi, Y. *et al.* Ancient human genomes reveal a hidden history of strong selection in Eurasia. *Cold Spring Harbor Laboratory* 2020.04.01.021006 (2020) doi:10.1101/2020.04.01.021006.
47. Tobler, R. *et al.* Genetic and climatic factors in the dispersal of Anatomically Modern Humans Out of Africa. *Research Square* (2021) doi:10.21203/rs.3.rs-800178/v1.
48. Palkopoulou, E. *et al.* Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Curr. Biol.* **25**, 1395–1400 (2015).
49. Barlow, A. *et al.* Partial genomic survival of cave bears in living brown bears. *Nat Ecol Evol* **2**, 1563–1570 (2018).
50. Haile, J. *et al.* Ancient DNA reveals late survival of mammoth and horse in interior Alaska. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 22352–22357 (2009).
51. Orlando, L. & Cooper, A. Using Ancient DNA to Understand Evolutionary and Ecological Processes. *Annu. Rev. Ecol. Evol. Syst.* **45**, 573–598 (2014).
52. Brüniche-Olsen, A. *et al.* Ancient DNA tracks the mainland extinction and island survival of the Tasmanian devil. *J. Biogeogr.* **45**, 963–976 (2018).
53. Chan, Y. L., Lacey, E. A., Pearson, O. P. & Hadly, E. A. Ancient DNA reveals Holocene loss of genetic diversity in a South American rodent. *Biol. Lett.* **1**, 423–426 (2005).
54. Lippold, S. *et al.* Discovery of lost diversity of paternal horse lineages using ancient DNA. *Nat. Commun.* **2**, 450 (2011).
55. Sheng, G.-L. *et al.* Ancient DNA from Giant Panda (Ailuropoda melanoleuca) of South-Western China Reveals Genetic Diversity Loss during the Holocene. *Genes* **9**, (2018).
56. Leonard, J. A. Ancient DNA applications for wildlife conservation. *Mol. Ecol.* **17**, 4186–4196 (2008).

57. Willerslev, E. *et al.* Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* **506**, 47–51 (2014).
58. Thomas, J. E. *et al.* Demographic reconstruction from ancient DNA supports rapid extinction of the great auk. *Elife* **8**, (2019).
59. Leonardi, M. *et al.* Evolutionary Patterns and Processes: Lessons from Ancient DNA. *Syst. Biol.* **66**, e1–e29 (2017).
60. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harb. Perspect. Biol.* **5**, (2013).
61. Hofreiter, M. *et al.* The future of ancient DNA: Technical advances and conceptual shifts. *Bioessays* **37**, 284–293 (2015).
62. Bollongino, R., Tresset, A. & Vigne, J.-D. Environment and excavation: Pre-lab impacts on ancient DNA analyses. *C. R. Palevol* **7**, 91–98 (2008).
63. Pääbo, S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. *Proc. Natl. Acad. Sci. U. S. A.* **86**, 1939–1943 (1989).
64. Callaway, E. Million-year-old mammoth genomes shatter record for oldest ancient DNA. *Nature* **590**, 537–538 (2021).
65. Lindahl, T. Instability and decay of the primary structure of DNA. *Nature* **362**, 709–715 (1993).
66. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29**, 1682–1684 (2013).
67. Sawyer, S., Krause, J., Guschanski, K., Savolainen, V. & Pääbo, S. Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One* **7**, e34131 (2012).
68. Orlando, L. *et al.* Recalibrating Equus evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**, 74–78 (2013).
69. Wagner, S. *et al.* High-Throughput DNA sequencing of ancient wood. *Mol. Ecol.* **27**, 1138–1154 (2018).

70. Yang, D. Y. & Watt, K. Contamination controls when preparing archaeological remains for ancient DNA analysis. *J. Archaeol. Sci.* **32**, 331–336 (2005).

71. Schubert, M. *et al.* Improving ancient DNA read mapping against modern reference genomes. *BMC Genomics* **13**, 178 (2012).

72. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

73. Krause, J. *et al.* A complete mtDNA genome of an early modern human from Kostenki, Russia. *Curr. Biol.* **20**, 231–236 (2010).

74. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).

75. Briggs, A. W. *et al.* Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).

76. Rohland, N., Harney, E., Mallick, S., Nordenfelt, S. & Reich, D. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **370**, 20130624 (2015).

77. Jun, G., Wing, M. K., Abecasis, G. R. & Kang, H. M. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Res.* **25**, 918–925 (2015).

78. Link, V. *et al.* ATLAS: Analysis Tools for Low-depth and Ancient Samples. *bioRxiv* 105346 (2017) doi:10.1101/105346.

79. Skoglund, P. *et al.* Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 2229–2234 (2014).

80. Neukamm, J., Peltzer, A. & Nieselt, K. DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab190.

81. Willerslev, E. & Cooper, A. Ancient DNA. *Proc. Biol. Sci.* **272**, 3–16 (2005).

82. Gilbert, M. T. P., Bandelt, H.-J., Hofreiter, M. & Barnes, I. Assessing ancient DNA studies. *Trends Ecol. Evol.* **20**, 541–544 (2005).

83. Poinar, H. N. *et al.* Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. *Science* **311**, 392–394 (2006).
84. Peyrégne, S. & Prüfer, K. Present-Day DNA Contamination in Ancient DNA Datasets. *Bioessays* **42**, e2000081 (2020).
85. Pääbo, S. Of bears, conservation genetics, and the value of time travel. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 1320–1321 (2000).
86. Cooper, A. Ancient DNA: Do It Right or Not at All. *Science* vol. 289 1139b–1139 (2000).
87. Renaud, G., Schubert, M., Sawyer, S. & Orlando, L. Authentication and Assessment of Contamination in Ancient DNA. in *Ancient DNA: Methods and Protocols* (eds. Shapiro, B. et al.) 163–194 (Springer New York, 2019).
88. Helgason, A. *et al.* A statistical approach to identify ancient template DNA. *J. Mol. Evol.* **65**, 92–102 (2007).
89. Korneliussen, T. S., Albrechtsen, A. & Nielsen, R. ANGSD: Analysis of Next Generation Sequencing Data. *BMC Bioinformatics* **15**, 356 (2014).
90. Renaud, G., Slon, V., Duggan, A. T. & Kelso, J. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biol.* **16**, 224 (2015).
91. Peltzer, A. *et al.* EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
92. Brandt, D. Y. C. *et al.* Mapping Bias Overestimates Reference Allele Frequencies at the HLA Genes in the 1000 Genomes Project Phase I Data. *G3* **5**, 931–941 (2015).
93. Lister, T. Exploring Graph-Based mapping as a means to reduce reference bias, in Genetic and Evolution. (The University of Adelaide, 2018).
94. Günther, T. & Nettelblad, C. The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genet.* **15**, e1008302 (2019).
95. Martiniano, R., Garrison, E., Jones, E. R., Manica, A. & Durbin, R. Removing reference bias and improving indel calling in ancient DNA data analysis by mapping to a sequence variation graph. *Genome Biology 21* 782755 (2020) doi:10.1101/782755.

96. Orlando, L. *et al.* Ancient DNA analysis. *Nature Reviews Methods Primers* **1**, 1–26 (2021).
97. Ros-Freixedes, R. *et al.* Impact of index hopping and bias towards the reference allele on accuracy of genotype calls from low-coverage sequencing. *Genet. Sel. Evol.* **50**, 64 (2018).
98. Martiniano, R. *et al.* The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods. *PLoS Genet.* **13**, e1006852 (2017).
99. Llamas, B., Willerslev, E. & Orlando, L. Human evolution: a tale from ancient genomes. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **372**, (2017).
100. Skoglund, P. *et al.* Origins and genetic legacy of Neolithic farmers and hunter-gatherers in Europe. *Science* **336**, 466–469 (2012).
101. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
102. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
103. Reich, D., Thangaraj, K., Patterson, N., Price, A. L. & Singh, L. Reconstructing Indian population history. *Nature* **461**, 489–494 (2009).
104. Reich, D. *et al.* Reduced neutrophil count in people of African descent is due to a regulatory variant in the Duffy antigen receptor for chemokines gene. *PLoS Genet.* **5**, e1000360 (2009).
105. Kowal, E. & Llamas, B. Race in a genome: long read sequencing, ethnicity-specific reference genomes and the shifting horizon of race. *J. Anthropol. Sci.* **96**, 91–106 (2019).
106. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
107. Guo, Y. *et al.* Improvements and impacts of GRCh38 human reference on high

throughput sequencing data analysis. *Genomics* **109**, 83–90 (2017).

108. Bushnell, B. BBMap: a fast, accurate, splice-aware aligner. (2014).

109. Church, D. M. *et al.* Modernizing reference genome assemblies. *PLoS Biol.* **9**, e1001091 (2011).

110. Genovese, G. *et al.* Using population admixture to help complete maps of the human genome. *Nat. Genet.* **45**, 406–14, 414e1–2 (2013).

111. Church, D. M. *et al.* Extending reference assembly models. *Genome Biol.* **16**, 13 (2015).

112. Serhat Tetikol, H. *et al.* Population-specific genome graphs improve high-throughput sequencing data analysis: A case study on the Pan-African genome. *bioRxiv* 2021.03.19.436173 (2021) doi:10.1101/2021.03.19.436173.

113. Dilthey, A., Cox, C., Iqbal, Z., Nelson, M. R. & McVean, G. Improved genome inference in the MHC using a population reference graph. *Nat. Genet.* **47**, 682–688 (2015).

114. Miga, K. H. & Wang, T. The Need for a Human Pangenome Reference Sequence. *Annu. Rev. Genomics Hum. Genet.* **22**, 81–102 (2021).

115. Tetz, V. V. The pangenome concept: a unifying view of genetic information. *Med. Sci. Monit.* **11**, HY24–9 (2005).

116. Li, R. *et al.* Building the sequence map of the human pan-genome. *Nat. Biotechnol.* **28**, 57–63 (2010).

117. Sirén, J., Välimäki, N. & Mäkinen, V. Indexing Graphs for Path Queries with Applications in Genome Research. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **11**, 375–388 (2014).

118. Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Brief. Bioinform.* **19**, 118–135 (2018).

119. Llamas, B. *et al.* A strategy for building and using a human reference pangenome. *F1000Res.* **8**, 1751 (2019).

120. Tettelin, H. & Masignani, V. Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial ‘pan-genome’. *Proceedings of the* (2005).

121. Hiller, N. L. *et al.* Comparative genomic analyses of seventeen Streptococcus

pneumoniae strains: insights into the pneumococcal supragenome. *J. Bacteriol.* **189**, 8186–8195 (2007).

122. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11**, 472–477 (2008).
123. Donati, C. *et al.* Structure and dynamics of the pan-genome of Streptococcus pneumoniae and closely related species. *Genome Biol.* **11**, R107 (2010).
124. Eppinger, M. *et al.* Genome sequence of the deep-rooted Yersinia pestis strain Angola reveals new insights into the evolution and pangenome of the plague bacterium. *J. Bacteriol.* **192**, 1685–1699 (2010).
125. Mira, A., Martín-Cuadrado, A. B., D'Auria, G. & Rodríguez-Valera, F. The bacterial pan-genome:a new paradigm in microbiology. *Int. Microbiol.* **13**, 45–57 (2010).
126. Mongodin, E. F. *et al.* Inter- and intra-specific pan-genomes of Borrelia burgdorferi sensu lato: genome stability and adaptive radiation. *BMC Genomics* **14**, 693 (2013).
127. Gordienko, E. N., Kazanov, M. D. & Gelfand, M. S. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J. Bacteriol.* **195**, 2786–2792 (2013).
128. Sherman, R. M. *et al.* Assembly of a pan-genome from deep sequencing of 910 humans of African descent. *Nat. Genet.* **51**, 30–35 (2019).
129. Li, Q. *et al.* Building a Chinese pan-genome of 486 individuals. *Commun Biol* **4**, 1016 (2021).
130. Chen, S. *et al.* Paragraph: a graph-based structural variant genotyper for short-read sequence data. *Genome Biol.* **20**, 291 (2019).
131. Pevzner, P. A., Tang, H. & Waterman, M. S. An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 9748–9753 (2001).
132. Garrison, E. *et al.* Variation graph toolkit improves read mapping by representing genetic variation in the reference. *Nat. Biotechnol.* **36**, 875–879 (2018).
133. Gordon, S. P. *et al.* Extensive gene content variation in the Brachypodium distachyon pan-genome correlates with population structure. *Nat. Commun.* **8**, 2184 (2017).

134. Schneeberger, K. *et al.* Simultaneous alignment of short reads against multiple genomes. *Genome Biol.* **10**, R98 (2009).

135. Maciuca, S., del Ojo Elias, C., McVean, G. & Iqbal, Z. A Natural Encoding of Genetic Variation in a Burrows-Wheeler Transform to Enable Mapping and Genome Inference. in *Algorithms in Bioinformatics* 222–233 (Springer International Publishing, 2016).

136. Valenzuela, D., Norri, T., Välimäki, N., Pitkänen, E. & Mäkinen, V. Towards pan-genome read alignment to improve variation calling. *BMC Genomics* **19**, 87 (2018).

137. Rakocevic, G. *et al.* Fast and accurate genomic analyses using genome graphs. *Nat. Genet.* **51**, 354–362 (2019).

138. Ros-Freixedes, R., Reixach, J., Tor, M. & Estany, J. Expected genetic response for oleic acid content in pork. *J. Anim. Sci.* **90**, 4230–4238 (2012).

139. Bryc, K., Patterson, N. & Reich, D. A novel approach to estimating heterozygosity from low-coverage genome sequence. *Genetics* **195**, 553–561 (2013).

140. Oliva, A., Tobler, R., Cooper, A., Llamas, B. & Souilmi, Y. Systematic benchmark of ancient DNA read mapping. *Brief. Bioinform.* (2021) doi:10.1093/bib/bbab076.

141. Atmanspacher, H. & Maasen, S. *Reproducibility: Principles, Problems, Practices, and Prospects*. (John Wiley & Sons, 2016).

142. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).

143. Open Science Collaboration. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

144. Freedman, L. P., Cockburn, I. M. & Simcoe, T. S. The Economics of Reproducibility in Preclinical Research. *PLoS Biol.* **13**, e1002165 (2015).

145. Hinsen, K. & Rougier, N. Challenge to test reproducibility of old computer code. *Nature* **574**, 634 (2019).

146. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav* **2**, 637–644 (2018).

147. Baker, M. 1,500 scientists lift the lid on reproducibility. *Nature* **533**, 452–454 (2016).

148. de Barros Damgaard, P. *et al.* The first horse herders and the impact of early Bronze Age steppe expansions into Asia. *Science* **360**, (2018).

149. Damgaard, P. de B. *et al.* 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369–374 (2018).

150. Quintana-Murci, L. *et al.* Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nat. Genet.* **23**, 437–441 (1999).

151. Richards, M. *et al.* Tracing European founder lineages in the Near Eastern mtDNA pool. *Am. J. Hum. Genet.* **67**, 1251–1276 (2000).

152. Fernandes, V. *et al.* The Arabian cradle: mitochondrial relicts of the first steps along the southern route out of Africa. *Am. J. Hum. Genet.* **90**, 347–355 (2012).

153. Pagani, L. *et al.* Tracing the route of modern humans out of Africa by using 225 human genome sequences from Ethiopians and Egyptians. *Am. J. Hum. Genet.* **96**, 986–991 (2015).

154. Walsh, S. *et al.* Positive selection in admixed populations from Ethiopia. *BMC Genet.* **21**, 108 (2020).

155. Haber, M. *et al.* A Genetic History of the Near East from an aDNA Time Course Sampling Eight Points in the Past 4,000 Years. *Am. J. Hum. Genet.* **107**, 149–157 (2020).

156. Lazaridis, I. *et al.* Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).

157. López, S., van Dorp, L. & Hellenthal, G. Human Dispersal Out of Africa: A Lasting Debate. *Evol. Bioinform. Online* **11**, 57–68 (2015).

158. Raghavan, M. *et al.* Upper Palaeolithic Siberian genome reveals dual ancestry of Native Americans. *Nature* **505**, 87–91 (2014).

159. Fu, Q. *et al.* The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).

160. Yang, M. A. *et al.* 40,000-Year-Old Individual from Asia Provides Insight into Early Population Structure in Eurasia. *Curr. Biol.* **27**, 3202–3208.e9 (2017).

161. Sikora, M. *et al.* The population history of northeastern Siberia since the Pleistocene.

Nature **570**, 182–188 (2019).

162. Olalde, I. & Posth, C. Latest trends in archaeogenetic research of west Eurasians. *Curr. Opin. Genet. Dev.* **62**, 36–43 (2020).
163. Raghavan, M. *et al.* POPULATION GENETICS. Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
164. Skoglund, P. *et al.* Genetic evidence for two founding populations of the Americas. *Nature* **525**, 104–108 (2015).
165. Moreno-Mayar, J. V. *et al.* Early human dispersals within the Americas. *Science* **362**, (2018).
166. Posth, C. *et al.* Reconstructing the Deep Population History of Central and South America. *Cell* **175**, 1185–1197.e22 (2018).
167. Castro, M. A. *et al.* Deep genetic affinity between coastal Pacific and Amazonian natives evidenced by Australasian ancestry. *Proceedings of the National Academy of Sciences* **118**, (2021).
168. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).
169. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2013).
170. Tian, S., Yan, H., Kalmbach, M. & Slager, S. L. Impact of post-alignment processing in variant discovery from whole exome data. *BMC Bioinformatics* **17**, 403 (2016).
171. Fages, A. *et al.* Tracking Five Millennia of Horse Management with Extensive Ancient Genome Time Series. *Cell* **177**, 1419–1435.e31 (2019).
172. Borges, M. G., Moraes, H. T. de, Rocha, C. de S. & Lopes-Cendes, I. The impact of post-alignment processing procedures on whole-exome sequencing data. *Genet. Mol. Biol.* **43**, e20200047 (2020).
173. Tourdot, R. W., Brunette, G. J., Pinto, R. A. & Zhang, C.-Z. Determination of complete chromosomal haplotypes by bulk DNA sequencing. *Genome Biol.* **22**, 139 (2021).
174. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the

Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).

175. Oliva, A., Tobler, R., Llamas, B. & Souilmi, Y. Additional evaluations show that specific BWA-aln settings still outperform BWA-mem for ancient DNA data alignment. *Ecol. Evol.* **11**, 18743–18748 (2021).

176. Wolinsky, H. Ancient DNA and contemporary politics: The analysis of ancient DNA challenges long-held beliefs about identity and history with potential for political abuse. *EMBO Rep.* **20**, e49507 (2019).

177. Jensen, E. L. *et al.* Ancient and historical DNA in conservation policy. *Trends Ecol. Evol.* (2022) doi:10.1016/j.tree.2021.12.010.

178. Frieman, C. J. & Hofmann, D. Present pasts in the archaeology of genetics, identity, and migration in Europe: a critical essay. *World Archaeol.* **51**, 528–545 (2019).

179. Benton, M. L. *et al.* The influence of evolutionary history on human health and disease. *Nat. Rev. Genet.* **22**, 269–283 (2021).

180. Prendergast, M. E. & Sawchuk, E. Boots on the ground in Africa's ancient DNA 'revolution': archaeological perspectives on ethics and best practices. *Antiquity* **92**, 803–815 (2018).

181. Wagner, J. K. *et al.* Fostering Responsible Research on Ancient DNA. *Am. J. Hum. Genet.* **107**, 183–195 (2020).

182. Alpaslan-Roodenberg, S. *et al.* Ethics of DNA research on human remains: five globally applicable guidelines. *Nature* (2021) doi:10.1038/s41586-021-04008-x.

183. Corpas, M., Kovalevskaya, N. V., McMurray, A. & Nielsen, F. G. G. A FAIR guide for data providers to maximise sharing of human genomic data. *PLoS Comput. Biol.* **14**, e1005873 (2018).

184. Fulton, T. L. Setting up an ancient DNA laboratory. *Methods Mol. Biol.* **840**, 1–11 (2012).

185. Knapp, M., Clarke, A. C., Ann Horsburgh, K. & Matisoo-Smith, E. A. Setting the stage – Building and working in an ancient DNA laboratory. *Annals of Anatomy - Anatomischer*

Anzeiger vol. 194 3–6 (2012).

186. Llamas, B. *et al.* From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR: Science & Technology of Archaeological Research* **3**, 1–14 (2017).

187. Sandve, G. K., Nekrutenko, A., Taylor, J. & Hovig, E. Ten simple rules for reproducible computational research. *PLoS Comput. Biol.* **9**, e1003285 (2013).

188. Rule, A. *et al.* Ten simple rules for writing and sharing computational analyses in Jupyter Notebooks. *PLoS Comput. Biol.* **15**, e1007007 (2019).

189. Riginos, C. *et al.* Building a global genomics observatory: Using GEOME (the Genomic Observatories Metadatabase) to expedite and improve deposition and retrieval of genetic data and metadata for biodiversity research. *Mol. Ecol. Resour.* **20**, 1458–1469 (2020).

190. Nüst, D. *et al.* Ten simple rules for writing Dockerfiles for reproducible data science. *PLoS Comput. Biol.* **16**, e1008316 (2020).

191. Ma, L. *et al.* NPARS-A Novel Approach to Address Accuracy and Reproducibility in Genomic Data Science. *Front Big Data* **4**, 725095 (2021).

192. Rohland, N. & Hofreiter, M. Ancient DNA extraction from bones and teeth. *Nat. Protoc.* **2**, 1756–1762 (2007).

193. Kistler, L. Ancient DNA Extraction from Plants. in *Ancient DNA: Methods and Protocols* (eds. Shapiro, B. & Hofreiter, M.) 71–79 (Humana Press, 2012).

194. Sirak, K. A. *et al.* A minimally-invasive method for sampling human petrous bones from the cranial base for ancient DNA analysis. *Biotechniques* **62**, 283–289 (2017).

195. Patzold, F., Zilli, A. & Hundsdoerfer, A. K. Advantages of an easy-to-use DNA extraction method for minimal-destructive analysis of collection specimens. *PLoS One* **15**, e0235222 (2020).

196. Sugita, N. *et al.* Non-destructive DNA extraction from herbarium specimens: a method particularly suitable for plants with small and fragile leaves. *J. Plant Res.* **133**, 133–141 (2020).

197. Hansen, H. B. *et al.* Comparing Ancient DNA Preservation in Petrous Bone and Tooth Cementum. *PLoS One* **12**, e0170940 (2017).
198. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
199. Shen, W., Le, S., Li, Y. & Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS One* **11**, e0163962 (2016).
200. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9**, 88 (2016).
201. Renaud, G., Stenzel, U. & Kelso, J. leeHom: adaptor trimming and merging for Illumina sequencing reads. *Nucleic Acids Res.* **42**, e141 (2014).
202. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
203. Schubert, M. *et al.* Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nat. Protoc.* **9**, 1056–1082 (2014).
204. Järve, M. *et al.* Shifts in the Genetic Landscape of the Western Eurasian Steppe Associated with the Beginning and End of the Scythian Dominance. *Curr. Biol.* **29**, 2430–2441.e10 (2019).
205. Günther, T. *et al.* Population genomics of Mesolithic Scandinavia: Investigating early postglacial migration routes and high-latitude adaptation. *PLoS Biol.* **16**, e2003703 (2018).
206. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
207. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. Nat Meth. 2012; 9 (4): 357--9.
208. Cahill, J. A. *et al.* Genomic Evidence of Widespread Admixture from Polar Bears into Brown Bears during the Last Ice Age. *Mol. Biol. Evol.* **35**, 1120–1129 (2018).
209. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**,

2078–2079 (2009).

210. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, (2021).

211. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).

212. Moreno-Mayar, J. V. FrAnTK: a Frequency-based Analysis ToolKit for efficient exploration of allele sharing patterns in present-day and ancient genomic datasets. *G3* (2021) doi:10.1093/g3journal/jkab357.

213. Barlow, A., Hartmann, S., Gonzalez, J., Hofreiter, M. & Paijmans, J. L. A. Consensify: A Method for Generating Pseudohaploid Genome Sequences from Palaeogenomic Datasets with Reduced Error Rates. *Genes* **11**, (2020).

214. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).

215. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

216. Novak, M. *et al.* Genome-wide analysis of nearly all the victims of a 6200 year old massacre. *PLoS One* **16**, e0247332 (2021).

217. Harney, É. *et al.* A minimally destructive protocol for DNA extraction from ancient teeth. *Genome Res.* **31**, 472–483 (2021).

218. Lipson, M. *et al.* Three Phases of Ancient Migration Shaped the Ancestry of Human Populations in Vanuatu. *Curr. Biol.* **30**, 4846–4856.e6 (2020).

219. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

220. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).

221. Popović, D. *et al.* Ancient genomes reveal long-range influence of the pre-Columbian culture and site of Tiwanaku. *Sci Adv* **7**, eabg7261 (2021).

222. Cassidy, L. M. *et al.* Neolithic and Bronze Age migration to Ireland and establishment of the insular Atlantic genome. *Proc. Natl. Acad. Sci. U. S. A.* **113**, 368–373 (2016).

223. Amorim, C. E. G. *et al.* Understanding 6th-century barbarian social organization and migration through paleogenomics. *Nat. Commun.* **9**, 3547 (2018).

224. Antonio, M. L. *et al.* Ancient Rome: A genetic crossroads of Europe and the Mediterranean. *Science* **366**, 708–714 (2019).

225. Brace, S. *et al.* Ancient genomes indicate population replacement in Early Neolithic Britain. *Nat Ecol Evol* **3**, 765–771 (2019).

226. Joseph, T. A. & Pe'er, I. Inference of Population Structure from Time-Series Genotype Data. *Am. J. Hum. Genet.* **105**, 317–333 (2019).

227. Bongers, J. L. *et al.* Integration of ancient DNA with transdisciplinary dataset finds strong support for Inca resettlement in the south Peruvian coast. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 18359–18368 (2020).

228. Petr, M., Vernot, B. & Kelso, J. admixr-R package for reproducible analyses using ADMIXTOOLS. *Bioinformatics* **35**, 3194–3195 (2019).

229. FAIRsharing Team. FAIRsharing record for: FAIR Metrics - Metadata Longevity. (2018) doi:10.25504/FAIRSHARING.A2W4NZ.

230. FAIRsharing Team. FAIRsharing record for: FAIR Metrics - Machine-readability of metadata. (2018) doi:10.25504/FAIRSHARING.ZTR3N9.

231. Fellows Yates, J. A. *et al.* Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *Cold Spring Harbor Laboratory* 2020.06.11.145615 (2021) doi:10.1101/2020.06.11.145615.

232. FAIRsharing Team. FAIRsharing record for: Minimum information for reporting Next Generation Sequencing genotyping. (2018) doi:10.25504/FAIRSHARING.JW7RQ3.

233. FAIRsharing Team. FAIRsharing record for: Minimum Information About a Bioinformatics investigation. (2018) doi:10.25504/FAIRSHARING.28YEC8.

234. Leipzig, J., Nüst, D., Hoyt, C. T., Ram, K. & Greenberg, J. The role of metadata in reproducible computational research. *Patterns (N Y)* **2**, 100322 (2021).

235. Carroll, S. R., Herczog, E., Hudson, M., Russell, K. & Stall, S. Operationalizing the CARE and FAIR Principles for Indigenous data futures. *Sci Data* **8**, 108 (2021).

236. Warinner, C. *et al.* Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46**, 336–344 (2014).

237. Hendy, J. *et al.* Proteomic evidence of dietary sources in ancient dental calculus. *Proc. Biol. Sci.* **285**, (2018).

238. Roberts, P. *et al.* Calling all archaeologists: guidelines for terminology, methodology, data handling, and reporting when undertaking and reviewing stable isotope applications in archaeology. *Rapid Commun. Mass Spectrom.* **32**, 361–372 (2018).

239. Reitsema, L. & Holder, S. Stable isotope analysis and the study of human stress, disease, and nutrition. *Bioarchaeology int.* **2**, 63–74 (2018).

240. Xu, W. *et al.* An efficient pipeline for ancient DNA mapping and recovery of endogenous ancient DNA from whole-genome sequencing data. *Ecol. Evol.* **11**, 390–401 (2021).

241. Shrout, P. E. & Rodgers, J. L. Psychology, Science, and Knowledge Construction: Broadening Perspectives from the Replication Crisis. *Annu. Rev. Psychol.* **69**, 487–510 (2018).

242. Loken, E. & Gelman, A. Measurement error and the replication crisis. *Science* **355**, 584–585 (2017).

243. A Fellows Yates, J. *et al.* A-Z of ancient DNA protocols for shotgun Illumina Next Generation Sequencing v1. *protocols.io* (2020) doi:10.17504/protocols.io.bj8nkrve.

244. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).

245. Maciuca, S., del Ojo Elias, C., McVean, G. & Iqbal, Z. A natural encoding of genetic variation in a Burrows-Wheeler Transform to enable mapping and genome inference. *bioRxiv* 059170 (2016) doi:10.1101/059170.

246. Ribeiro, A. *et al.* An investigation of causes of false positive single nucleotide polymorphisms using simulated reads from a small eukaryote genome. *BMC Bioinformatics* **16**, 382 (2015).

247. Duan, J. *et al.* Detection of False-Positive Deletions from the Database of Genomic Variants. *Biomed Res. Int.* **2019**, 8420547 (2019).

248. Garrison, E. & Guarracino, A. Unbiased pangenome graphs. *bioRxiv* 2022.02.14.480413 (2022) doi:10.1101/2022.02.14.480413.

Appendix 1: Supplementary materials for Chapter 1

Supplementary Information for

Systematic Benchmark of Ancient DNA Read Mapping

Adrien Oliva[1], Raymond Tobler[1], Alan Cooper[2], Bastien Llamas[1,3], Yassine Souilmi[1,3,*]

[1]Australian Centre for Ancient DNA, School of Biological Sciences, The University of Adelaide, South Australia, 5005, Australia

[2]South Australian Museum, Adelaide, SA 5005, Australia

[3]The Environment Institute, The University of Adelaide, South Australia, 5005, Australia

*Corresponding Author: YS yassine.souilmi@adelaide.edu.au, Ph: +61 8 8313 5565, Fax: +61 8 8313 4364.

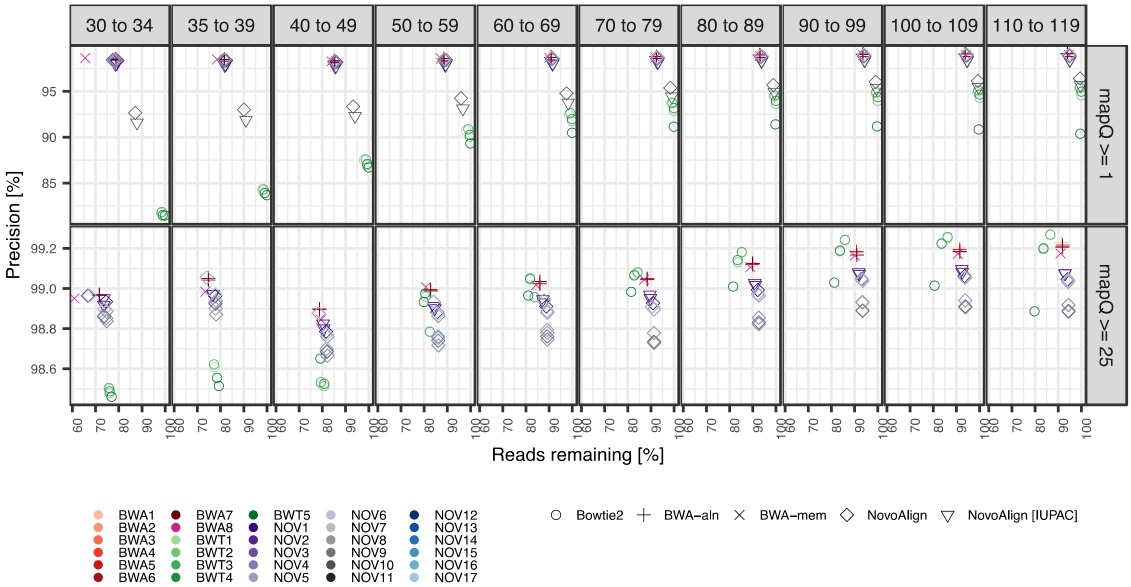
This PDF file includes:

- Supplementary Figures S1 to S14.
- Supplementary Table S1 and S2 captions.
- Supplementary Text.

Tables S1 and S2 are provided as separate excel files.

Table of Contents

Supplementary Figures



<u>Figure S1:</u> Alignment precision relative to read length and both mapping quality for the simulated European sample. We assessed the precision of 30 paramerizations (different colors) for four different alignment software, including an IUPAC-based alignment for a subset of the *NovoAlign* parameterizations (different shapes). The x-axis measures the number of reads remaining after applying the specific mapping quality filter, which results in the removal of all reads below the required mapping along with all unmapped reads.

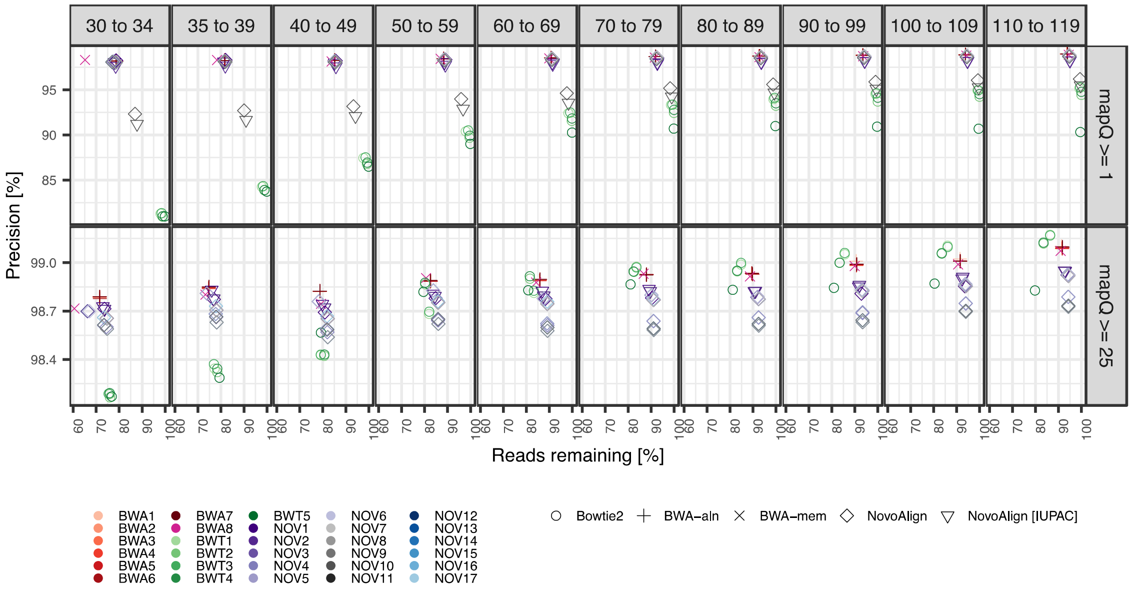
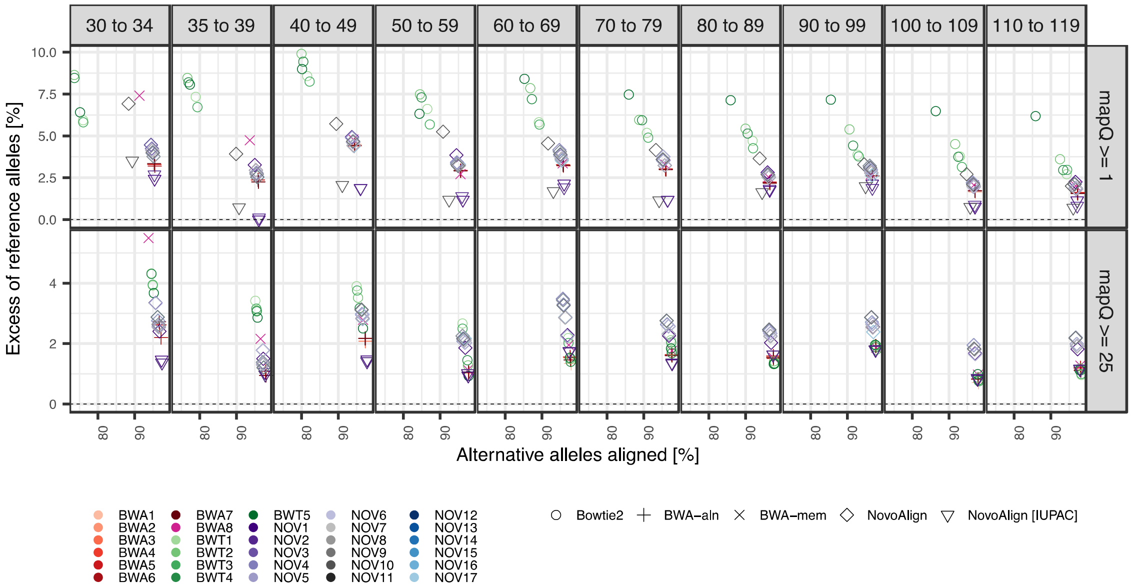
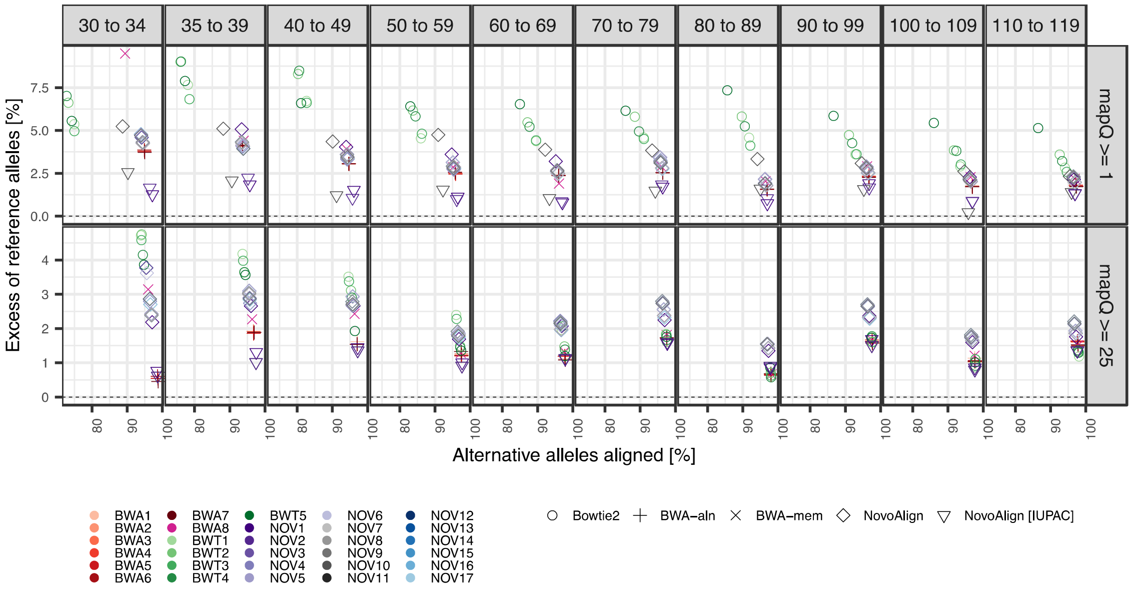


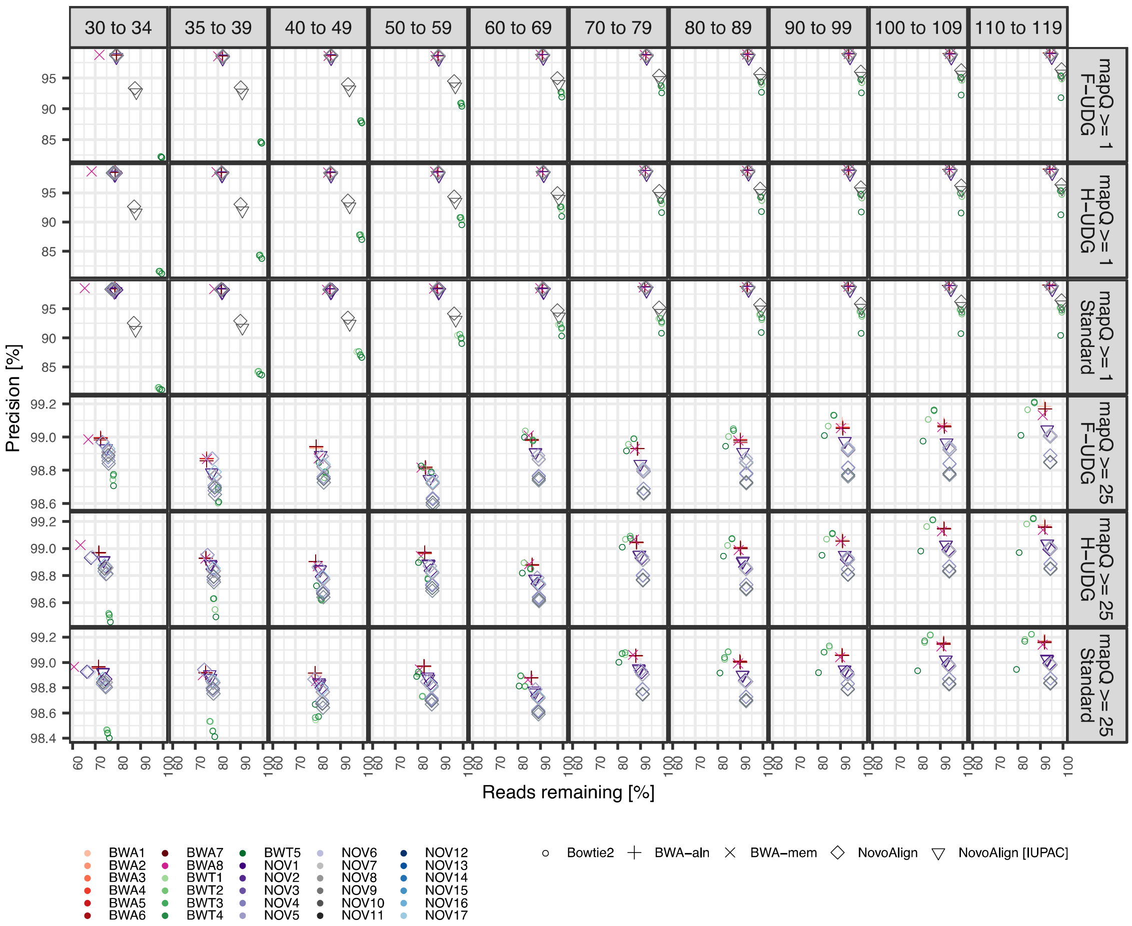
Figure S2: Alignment precision relative to read length and both mapping quality for the simulated African sample. We assessed the precision of 30 paramerizations (different colors) for four different alignment software, including an IUPAC-based alignment for a subset of the *NovoAlign* parameterizations (different shapes). The x-axis measures the number of reads remaining after applying the specific mapping quality filter, which results in the removal of all reads below the required mapping along with all unmapped reads.



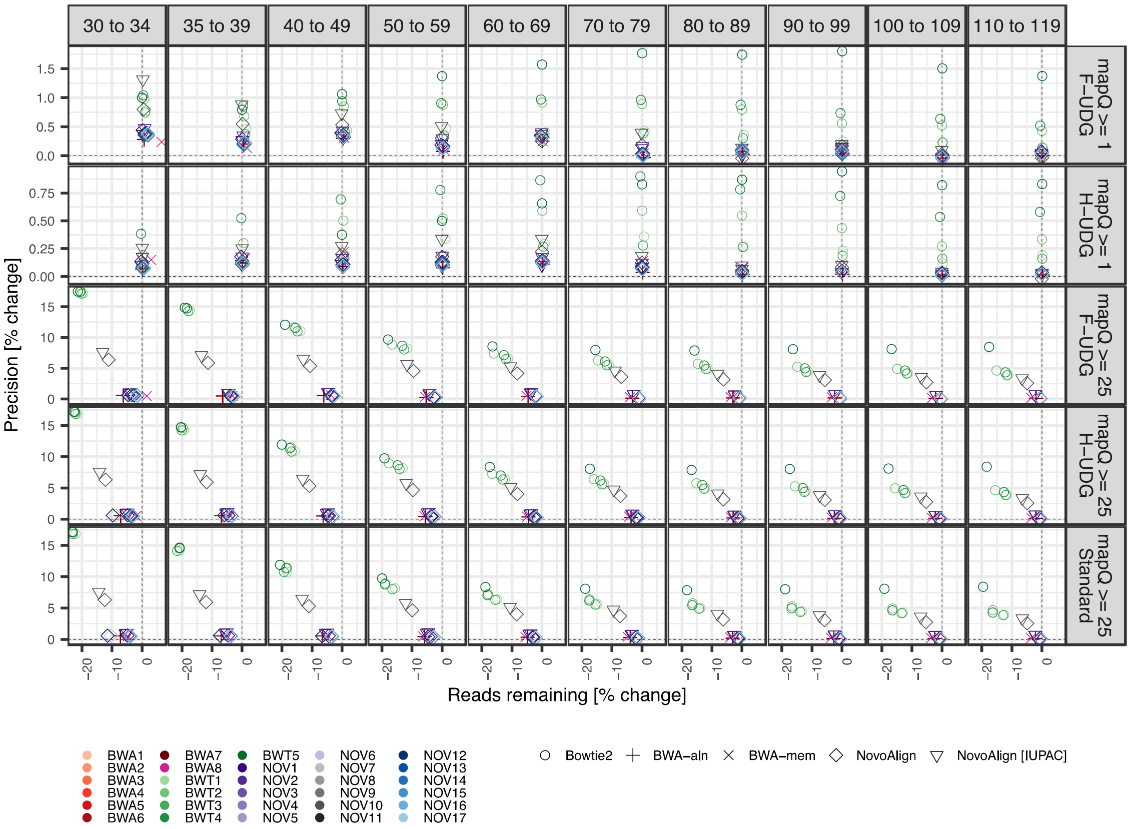
<u>Figure S3:</u> SNP-based reference bias relative to read length and both mapping quality for the simulated European sample. We measured the degree of reference bias by evaluating the difference in sensitivity between alternate and reference alleles across all SNPs covered by at least one alternate allele across successive read length bins.



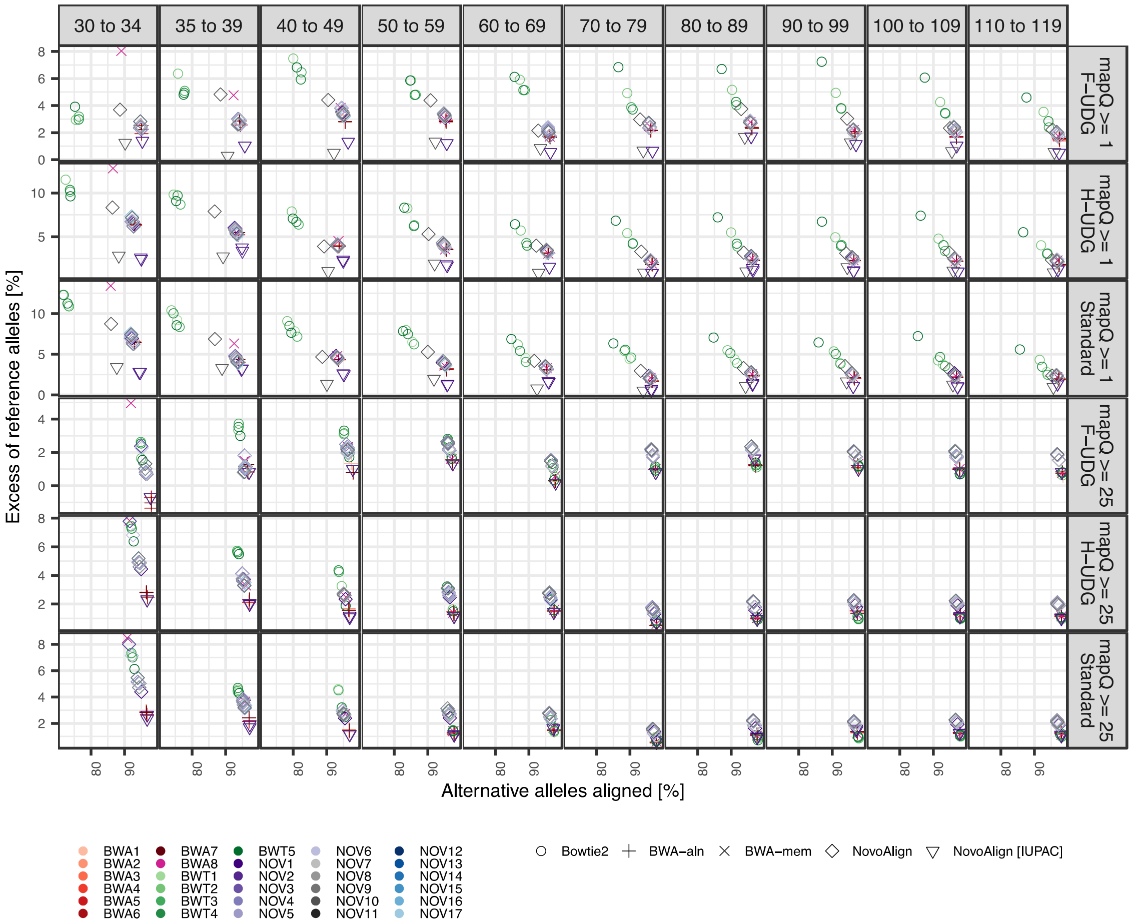
<u>**Figure S4:**</u> **SNP-based reference bias relative to read length and both mapping quality for the simulated African sample.** We measured the degree of reference bias by evaluating the difference in sensitivity between alternate and reference alleles across all SNPs covered by at least one alternate allele across successive read length bins.



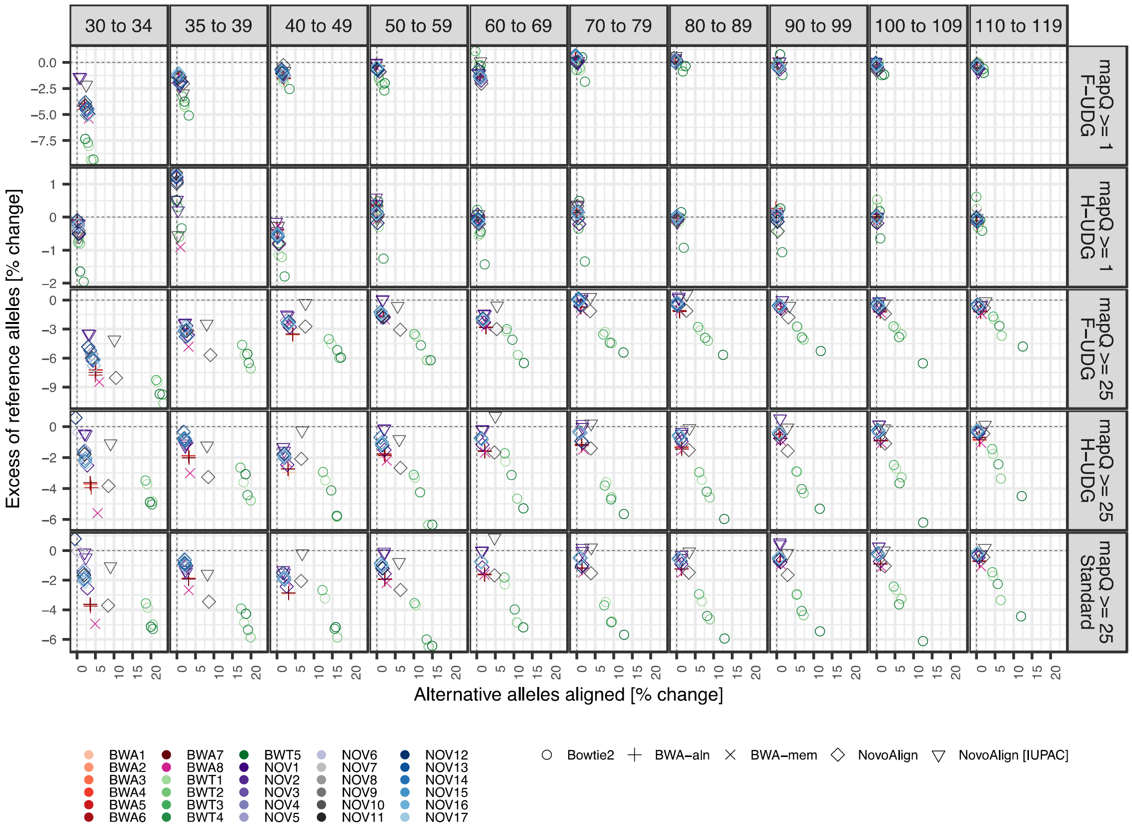
<u>Figure S5:</u> Precision and percentage of reads remaining for UDG treated and untreated reads for the simulated East Asian sample. Similar to Figure 1, but showing the results after simulating to commonly applied DNA pre-treatments that result in partial (Half-UDG; H-UDG) or complete (Full-UDG; F-UDG) removal of misspecified nucleotides caused by deaminated cytosine sites, along with results for untreated reads (Standard).



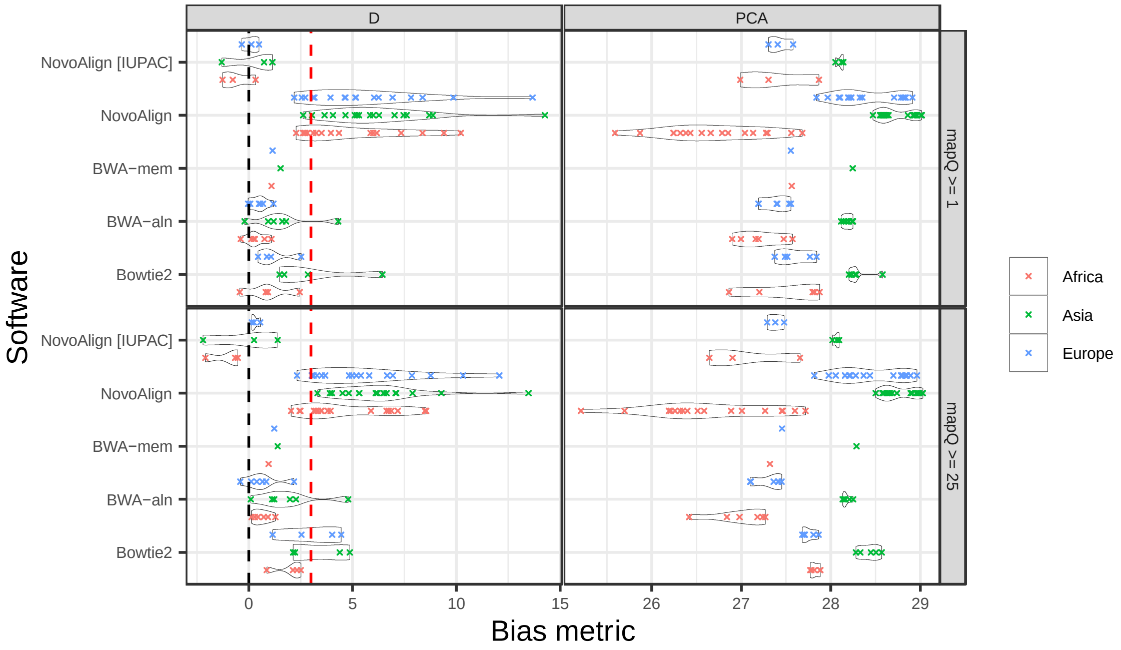
<u>**Figure S6:**</u> **Change in alignment precision when increasing the minimum mapping quality threshold from ≥1 to ≥25 and using UDG pre-treatments for the East Asian sample.** Percentage change in precision and the proportion of reads remaining relative to the untreated reads with mapping quality ≥ 1. Absolute values are shown in Figure S5.



<u>Figure S7:</u> SNP-based reference bias for UDG treated and untreated reads for the East Asian sample. Similar to Figure 2 but showing the results after simulating to commonly applied DNA pre-treatments that result in partial (Half-UDG; H-UDG) or complete (Full-UDG; F-UDG) removal of misspecified nucleotides caused by deaminated cytosine sites, along with results for untreated reads (Standard).



<u>**Figure S8:**</u> **Change in SNP-based reference bias when increasing the minimum mapping quality threshold from ≥1 to ≥25 and using UDG pre-treatments for the East Asian sample.** The percentage change in precision and the proportion of reads remaining relative to the untreated reads with mapping quality ≥ 1. Absolute values are shown in Figure S7.



<u>Figure S9:</u> The degree of reference bias across different aligners using common population genetic analyses for the three simulated samples. Pseudo-haploid variant calls were made for all strategies, after filtering for mapping quality (different row panels) and downsampling reads to have a length distribution matching an ancient sample, and the *D*-statistic (right panels) and PCA computed (left panels). The pseudo-haploidisation process was replicated 3 times for each parameterisation and the average results are shown here. For both metrics, unbiased alignments have a value of 0. For the *D*-statistic, the *Z*-scores are plotted and positive values above 3 (red line) imply a significant excess of shared alleles between the simulations and the reference genome relative to the truth, indicative of reference bias. The PCA metric measures the reduced distance between simulations and the reference relative to expectations, with larger values indicating more reference bias (values are multiplied by 100; see Methods). Figure S9 shows the results for untreated reads in all three samples.

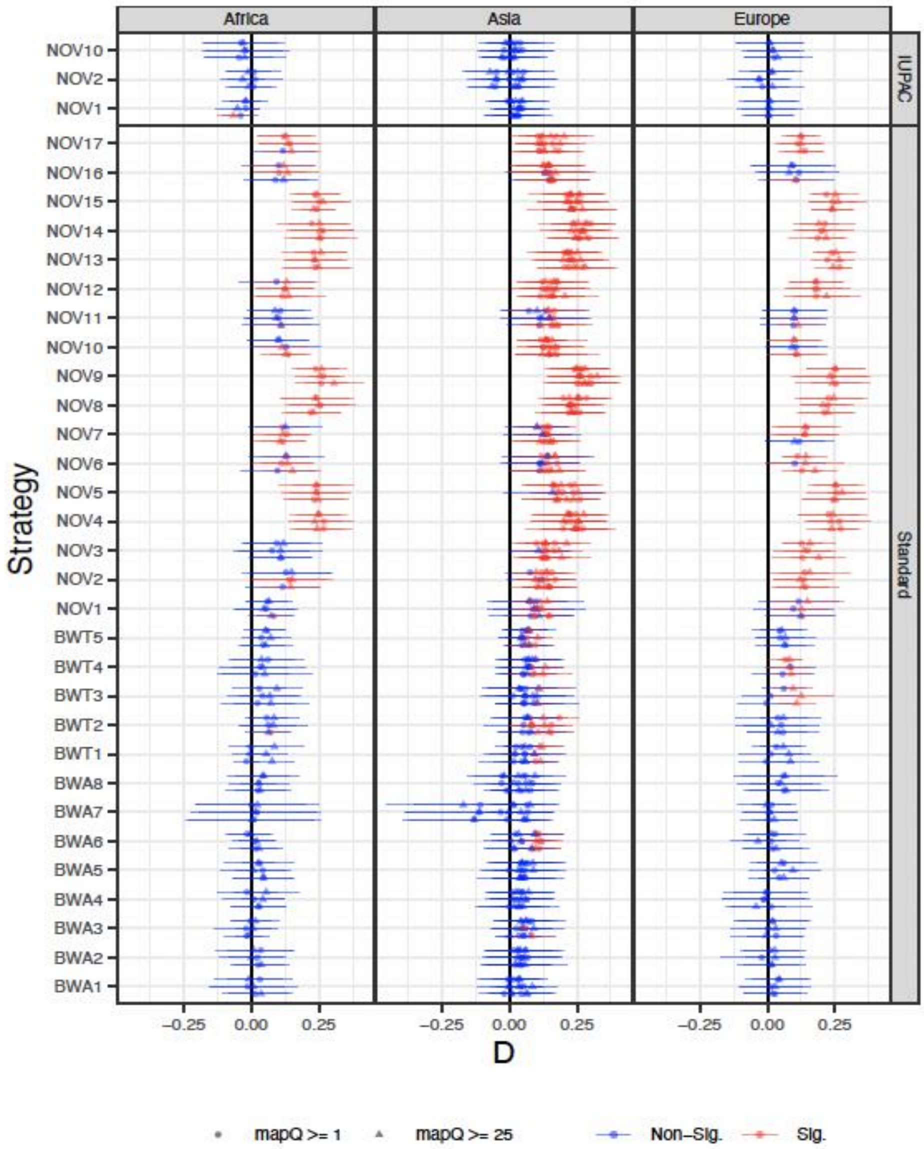
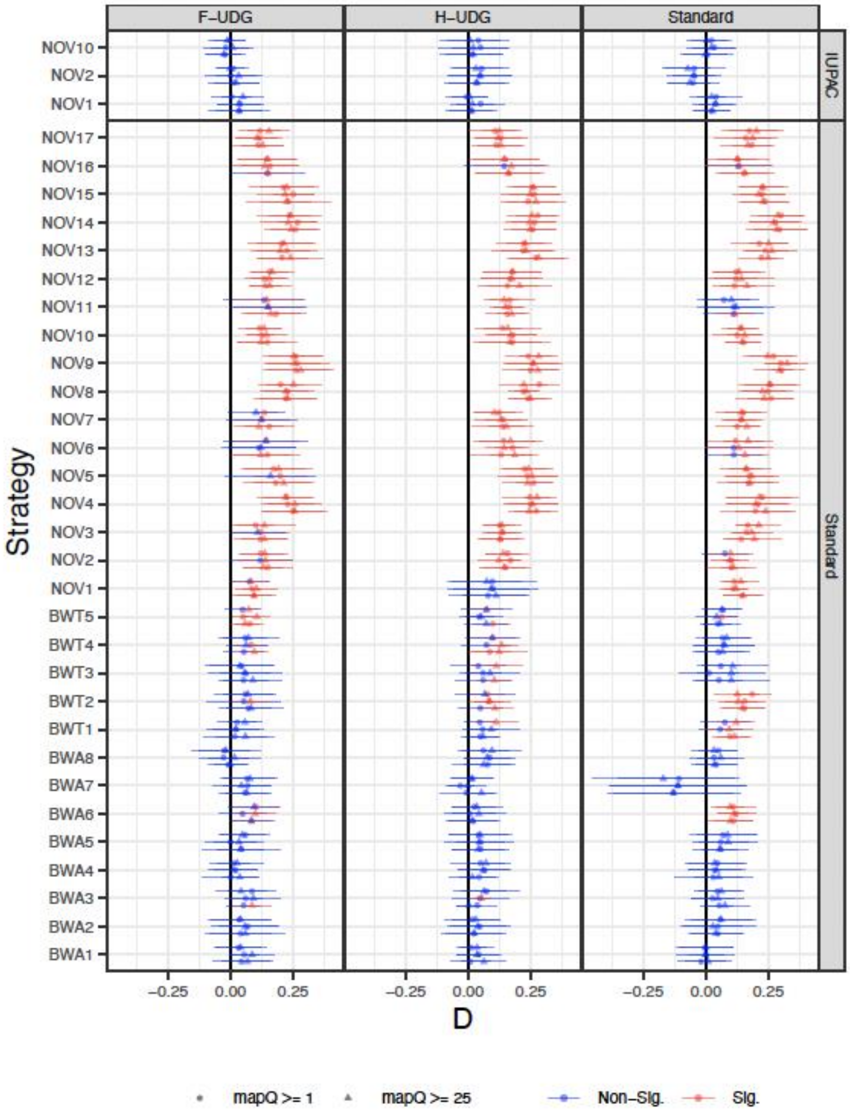
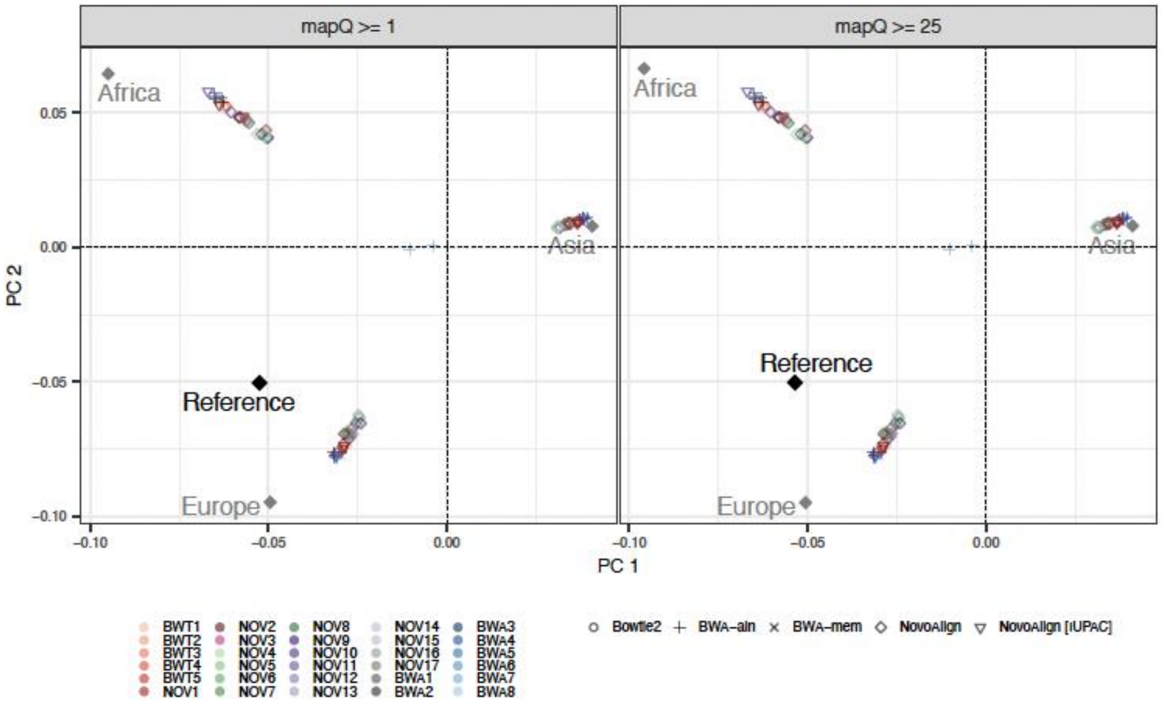


Figure S10: Reference bias across the different alignment software evaluated using the *D*-statistic and in three different samples. Three pseudo-haploid variant calls were made for each of the parameterizations, after filtering for mapping quality (circles = mapQ >= 1; triangles = mapQ ≥ 25, see key) and downsampling reads to have a length distribution matching an ancient sample, and the computed *D*-statistic (see Methods). Point estimates and bootstrap error intervals were estimated, and two-sided significance tests performed (D-statistic ≠ 0). Significant positive values imply an excess of shared alleles between the variant calls from the

simulations and the reference genome relative to the true calls, which indicates that the alignment favored reads carrying the reference allele (i.e., reference bias).



<u>Figure S11:</u> Reference bias across the different alignment software evaluated using the *D*-statistic for UDG-treated and untreated reads. The same as Figure S10 but evaluating the impact of UDG-based pre-treatments on *D*-statistic for the East Asian sample. See Figure S10 for more details.



<u>Figure S12:</u> PCA showing the relationships of untreated reads for the three samples using different mapping quality filters with respect to their true mappings. The first two PCA dimensions showing the position of simulated samples aligned using different strategies (see key), relative to the true position of each sample (grey diamonds) and the reference genome (black diamond). The simulated samples are pulled towards the reference sample relative to their expected PCA position, indicative of reference bias. The positions in the first two dimension are relatively invariant to mapping quality.

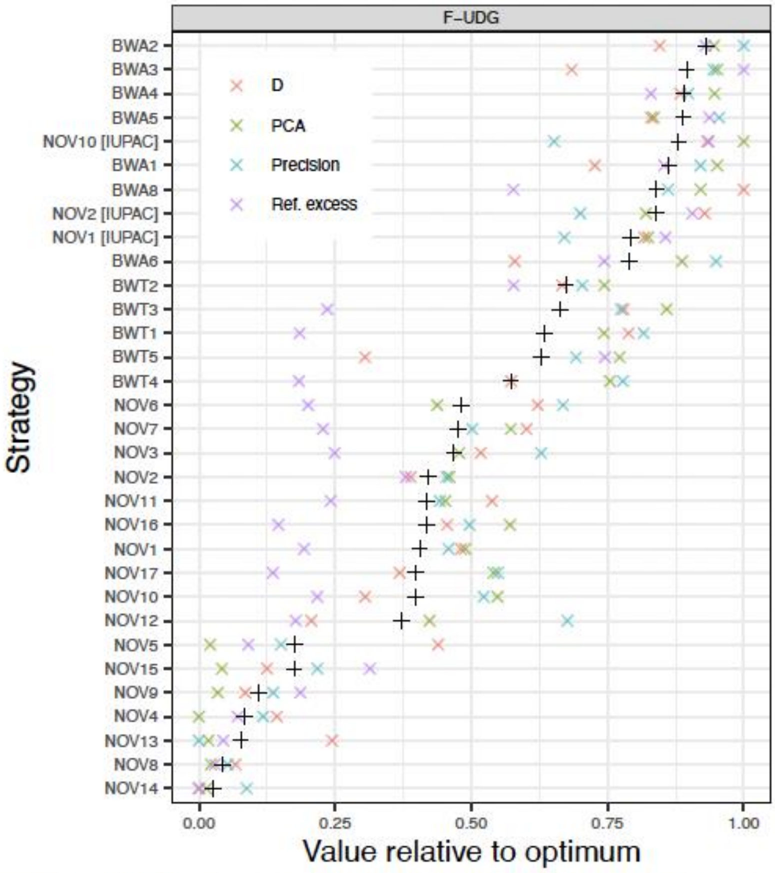


Figure S13: **Summary of mapping performance and reference bias for the Full-UDG dataset.** Each strategy is ranked according to the mean value across the four different metrics assessed in this study, i.e. precision, SNP-based reference bias, *D*-statistic, and PCA (mean = black plus sign and individual metrics = different coloured crosses; see key). Each metric is scaled to take a value between 0 and 1 by dividing the metric value by the range across each metric. This plot represents the results for the East Asian dataset using the reads treated with F-UDG.

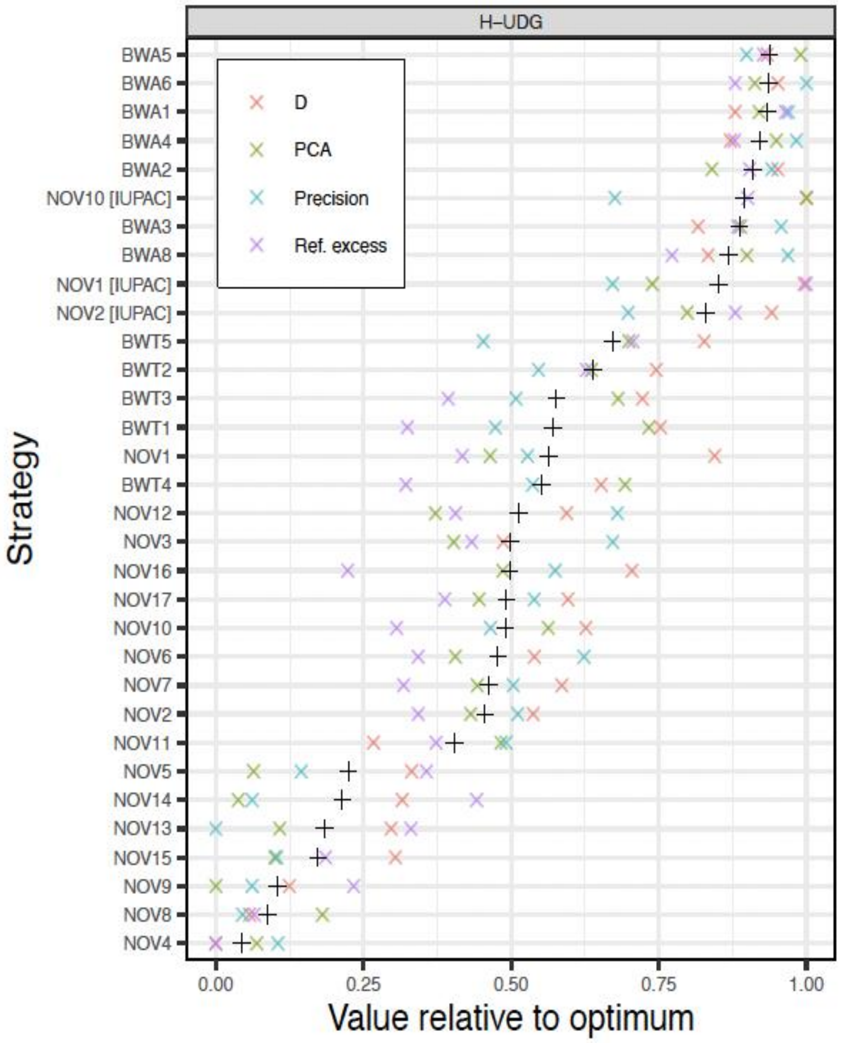


Figure S14: **Summary of mapping performance and reference bias for the Half-UDG dataset.** Each strategy is ranked according to the mean value across the four different metrics assessed in this study, i.e. precision, SNP-based reference bias, *D*-statistic, and PCA (mean = black plus sign and individual metrics = different coloured crosses; see key). Each metric is scaled to take a value between 0 and 1 by dividing the metric value by the range across each metric. This plot represents the results for the East Asian dataset using the reads treated with H-UDG.

Supplementary Tables

Table S1 contains the details of each option tested in this study. The shortened names for each method are represented in the first column, using the prefixes *BWT* for *Bowtie2, BWA* for *Burrows-Wheeler Aligner* and *Nov* for *NovoAlign.* Only the options marked with an "*" have were used with the IUPAC reference. When using the IUPAC reference only the results from *NovoAlign* are reported in the main text as others software don't have a built-in support for IUPAC characters in the reference, and both *BWA and Bowtie2* consider all IUPAC characters as ambiguous (i.e. Ns).

Table S2 contain several statistics for each dataset. For each of them we reported:

- Type = reference used, either "Standard" or IUPAC (using IUPAC reference).
- Country = which population dataset used.
- Strategy = short name of the aligners and options used (see Table S1).
- True Positive (TP) = total amount of reads mapping the correct position.
- False Positive (FP) = total amount of reads mapping the wrong position.
- False Negative (FN) = total amount of reads declared as unmapped.
- Total reads = total amount of reads mapped by the software in this dataset. We only look at a total of 900,000 reads, but some of options allows reads to map multiple locations and have higher numbers (reads mapping multiple locations have been removed before any analysis).
- TP rate % = percentage of reads mapping the correct position.
- FP rate % = percentage of reads mapping the wrong position.
- FN rate % = percentage of reads declared as unmapped.
- False Discovery Rate (FDR) = $\frac{FP}{FP+TP}$

- Precision = $\frac{TP}{TP+FP}$
- Sensitivity = $\frac{TP}{TP+FN}$
- Accuracy = $\frac{TP}{ALL\ READS}$
- CPU-Time (hours) = running time of the mapping for 1,5 million reads.
- Coverage = average coverage of our mapping reads on chromosome 22, computed using the following formula $\frac{N \times L}{G}$ where *N* is the number of reads; *L* the average read length and *G* the size of our genome (here the length of Ch22).

Supplementary text

1. Other datasets

Simulated reads were initially generated between 20 bp to 170 bp, resulting in 1.5 million reads in total. In subsequent analyses we focused on reads between 30 to 120bp, as these are more reflective of read distributions in ancient DNA datasets and reads lengths above 100bp show largely consistent patterns. The mapping time reported in our study is based on the alignment of all 1.5 million reads.

To explore the effect of deamination, two datasets have been created on the East Asian dataset only (the most biased with the main dataset).

a. Full-UDG

To simulate Full-UDG treated reads, 5bp of the undamaged reads were cut on each sides using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html). The reads have then been treated as the main reads (See Methods).

b. Half-UDG

To simulate Full-UDG treated reads, reads have been created using Mitty (https://github.com/sbg/Mitty) but using a custom damage profile. This damage profile is generated using data from the individual “SAMEA4843644” [1]. Then the reads have been treated as the main reads (See Methods).

2. Alignment Statistics

We used alignment precision to compare the performance of different aligners, which measures is the ratio of reads mapped at the correct location to the total reads mapped (which includes true positives and false positive reads; Eq. 1):

$$\text{Precision}= \frac{TP}{TP+FP}$$

To capture the impact of quality filtering the number of mapped reads, we also calculated the proportion of mapped reads that remained after removing reads below the minimum quality threshold (Eq. 2):

$$\text{Proportion mapped} = \frac{TP_{filt}+FP_{filt}}{TP_{all}+FP_{all}+FN}$$

Where the *filt* subscript implies the reads passing filtering, and the *all* subscript indicates all reads.

To compute the specificity of each software for the bacterial and the dog dataset, we computed the proportion of True Negative reads correctly set as unmapped (Eq. 3):

$$\text{Specificity} = \frac{TN_{mapped}}{TN_{all}}$$

References

1. Harney, É., et al., *Ancient DNA from Chalcolithic Israel reveals the role of population mixture in cultural transformation.* Nature Communications, 2018. **9**(1): p. 3336.

Table S1:

Software and parameters

Method name	Software	Option
BWT1	Bowtie2	None
BWT2	Bowtie2	-local
BWT3	Bowtie2	--very-sensitive
BWT4	Bowtie2	--very-sensitive-local
BWT5	Bowtie2	-local -N 1 --mp 4
BWA1	BWA aln	-l 1024 -n 0.01 -o 2
BWA2	BWA aln	-l 1024
BWA3	BWA aln	-l 1024 -i0
BWA4	BWA aln	-l 1024 -o 2
BWA5	BWA aln	-l 1024 -o 3
BWA6	BWA aln	-l 1024 -n 0.02
BWA7	BWA aln	-l 1024 -n 0.03
BWA8	BWA mem	None
Novo1*	NovAlign	-k
Novo2*	NovAlign	None
Novo3	NovAlign	-x 2
Novo4	NovAlign	-x 2 --matchreward 3
Novo5	NovAlign	-x 2 --matchreward 3 -t 15,3
Novo6	NovAlign	-x 2 -t 15,3
Novo7	NovAlign	-c 8
Novo8	NovAlign	--matchreward 3
Novo9	NovAlign	--matchreward 3 -t 15,3
Novo10*	NovAlign	-r Random
Novo11	NovAlign	-p 5,20
Novo12	NovAlign	-p 5,20 -x 2
Novo13	NovAlign	-p 5,20 --matchreward 3
Novo14	NovAlign	-p 5,20 --matchreward 3 -t 15,3
Novo15	NovAlign	-p 5,20 --matchreward 3 -t 15,3 -x 2
Novo16	NovAlign	-p 5,20 -t 15,3
Novo17	NovAlign	-t 15,3

* Used with the IUPAC reference

Table S2:

Core statistics measured for each dataset.

For each dataset we report:

- Type: reference genome used, can be: “Standard”, IUPAC (standard dataset using IUPAC reference), F-UDG (using IUPAC reference when specified), H-UDG (using IUPAC reference when specified).
- Country: the country of origin for each of the three samples.
- Strategy: short name of the aligners and options used (see Table S1).
- True Positive (TP): total amount of reads mapping to the correct position.
- False Positive (FP): total amount of reads mapping to the wrong position.
- False Negative (FN): total amount of reads declared as unmapped.
- Total Reads: total amount of reads mapped by the software in this dataset. Note that some options include reads that map to multiple locations and therefore have multiple entries. These reads were removed from all analyses.
- TP rate %: percentage of reads mapping to the correct position.
- FP rate %: percentage of reads mapping to the wrong position.
- FN rate %: percentage of reads declared as unmapped.
- False Discovery Rate (FDR): FP / (FP + TP)
- Precision: TP / (TP + FP)
- Sensitivity: TP / (TP + FN)
- Accuracy: TP / ALL READS

- CPU-Time (hours): running time to align the full set of 1.5 million reads (see comment in section 1 of SI).
- Coverage: average coverage across aligned reads computed using $N \times LG$ where N is the number of reads; L the average read length and G the size of the genome (i.e. the length of Ch22 in the context of this study)

Type	Country	Strategy	True Positive (TP)	False Positive (FP)	False Negative (FN)	Total.Reads	TP %	FP %	FN %	False Discovery Rate (FDR)	Precision	Accuracy	CPU-time (hours)	Coverage
Standard	Africa	Nov1	804247	12497	83256	900000	0.89361	0.01389	0.09251	0.01530	0.98470	0.89361	1.64	2.52850
Standard	Africa	Nov2	802631	11962	85407	900000	0.89181	0.01329	0.09490	0.01468	0.98532	0.89181	3.84	2.52907
Standard	Africa	Nov3	800983	12142	86875	900000	0.88998	0.01349	0.09653	0.01493	0.98507	0.88998	3.70	2.52574
Standard	Africa	Nov4	801684	12905	85411	900000	0.89076	0.01434	0.09490	0.01584	0.98416	0.89076	8.24	2.52910
Standard	Africa	Nov5	800588	12811	86601	900000	0.88954	0.01423	0.09622	0.01575	0.98425	0.88954	8.07	2.52765
Standard	Africa	Nov6	800764	12132	87104	900000	0.88974	0.01348	0.09678	0.01492	0.98508	0.88974	3.62	2.52508
Standard	Africa	Nov7	802631	11962	85407	900000	0.89181	0.01329	0.09490	0.01468	0.98532	0.89181	3.42	2.52906
Standard	Africa	Nov8	801334	13173	85493	900000	0.89037	0.01464	0.09499	0.01617	0.98383	0.89037	4.76	2.52929
Standard	Africa	Nov9	800150	12967	86883	900000	0.88906	0.01441	0.09654	0.01595	0.98405	0.88906	4.33	2.52473
Standard	Africa	Nov10	827020	72565	415	900000	0.91891	0.08063	0.00046	0.08066	0.91934	0.91891	4.39	2.63710
Standard	Africa	Nov11	802631	11962	85407	900000	0.89181	0.01329	0.09490	0.01468	0.98532	0.89181	3.04	2.52907
Standard	Africa	Nov12	800983	12142	86875	900000	0.88998	0.01349	0.09653	0.01493	0.98507	0.88998	3.30	2.52574
Standard	Africa	Nov13	801334	13173	85493	900000	0.89037	0.01464	0.09499	0.01617	0.98383	0.89037	3.70	2.52929
Standard	Africa	Nov14	800150	12967	86883	900000	0.88906	0.01441	0.09654	0.01595	0.98405	0.88906	3.72	2.52473
Standard	Africa	Nov15	800588	12811	86601	900000	0.88954	0.01423	0.09622	0.01575	0.98425	0.88954	6.88	2.52765
Standard	Africa	Nov16	801452	11851	86697	900000	0.89050	0.01317	0.09633	0.01457	0.98543	0.89050	2.85	2.52747
Standard	Africa	Nov17	801452	11851	86697	900000	0.89050	0.01317	0.09633	0.01457	0.98543	0.89050	3.38	2.45747
Standard	Africa	BWA1	824266	72117	3617	900000	0.91585	0.08013	0.00402	0.08045	0.91955	0.91585	87.17	1.88902
Standard	Africa	BWA2	823236	72177	4587	900000	0.91471	0.08020	0.00510	0.08061	0.91939	0.91471	79.44	2.02573
Standard	Africa	BWA3	823330	72121	4549	900000	0.91481	0.08013	0.00505	0.08054	0.91946	0.91481	88.75	2.02281
Standard	Africa	BWA4	823291	72123	4586	900000	0.91477	0.08014	0.00510	0.08055	0.91945	0.91477	80.60	2.00378
Standard	Africa	BWA5	823272	72134	4594	900000	0.91475	0.08015	0.00510	0.08056	0.91944	0.91475	67.76	2.00356
Standard	Africa	BWA6	824774	72140	3086	900000	0.91642	0.08016	0.00343	0.08043	0.91957	0.91642	107.63	1.98019
Standard	Africa	BWA7	824293	71936	3771	900000	0.91588	0.07993	0.00419	0.08027	0.91973	0.91588	92.06	0.40307
Standard	Africa	BWA8	816542	74672	9727	900941	0.90632	0.08288	0.01080	0.08379	0.91621	0.90632	1.36	2.59275

Standard	Africa	BWT1	813452	78169	8379	900000	0.90384	0.08685	0.00931	0.08767	0.91233	0.90384	1.23	2.58649
Standard	Africa	BWT2	816079	81367	2554	900000	0.90675	0.09041	0.00284	0.09067	0.90933	0.90675	1.46	2.59575
Standard	Africa	BWT3	819257	77046	3697	900000	0.91029	0.08561	0.00411	0.08596	0.91404	0.91029	2.10	2.60700
Standard	Africa	BWT4	818912	79057	2031	900000	0.90990	0.08784	0.00226	0.08804	0.91196	0.90990	2.08	2.60898
Standard	Africa	BWT5	799986	98852	1162	900000	0.88887	0.10984	0.00129	0.10998	0.89002	0.88887	7.24	2.59899
Standard	Asia	Nov1	805759	11638	82603	900000	0.89529	0.01293	0.09178	0.01424	0.98576	0.89529	0.94	2.53670
Standard	Asia	Nov2	803667	11133	85200	900000	0.89296	0.01237	0.09467	0.01366	0.98634	0.89296	0.81	2.53305
Standard	Asia	Nov3	802128	11262	86610	900000	0.89125	0.01251	0.09623	0.01385	0.98615	0.89125	2.15	2.53143
Standard	Asia	Nov4	802819	11961	85220	900000	0.89202	0.01329	0.09469	0.01468	0.98532	0.89202	1.54	2.53304
Standard	Asia	Nov5	801802	11860	86338	900000	0.89089	0.01318	0.09593	0.01458	0.98542	0.89089	1.64	2.53171
Standard	Asia	Nov6	801919	11250	86831	900000	0.89102	0.01250	0.09648	0.01383	0.98617	0.89102	1.96	2.53079
Standard	Asia	Nov7	803666	11133	85201	900000	0.89296	0.01237	0.09467	0.01366	0.98634	0.89296	0.82	2.53305
Standard	Asia	Nov8	802571	12174	85255	900000	0.89175	0.01353	0.09473	0.01494	0.98506	0.89175	0.93	2.53301
Standard	Asia	Nov9	801457	11997	86546	900000	0.89051	0.01333	0.09616	0.01475	0.98525	0.89051	0.80	2.53145
Standard	Asia	Nov10	828384	71242	374	900000	0.92043	0.07916	0.00042	0.07919	0.92081	0.92043	1.34	2.63913
Standard	Asia	Nov11	803667	11133	85200	900000	0.89296	0.01237	0.09467	0.01366	0.98634	0.89296	0.86	2.53305
Standard	Asia	Nov12	802128	11262	86610	900000	0.89125	0.01251	0.09623	0.01385	0.98615	0.89125	2.12	2.53143
Standard	Asia	Nov13	802571	12174	85255	900000	0.89175	0.01353	0.09473	0.01494	0.98506	0.89175	0.78	2.53301
Standard	Asia	Nov14	801457	11997	86546	900000	0.89051	0.01333	0.09616	0.01475	0.98525	0.89051	0.79	2.53145
Standard	Asia	Nov15	801802	11860	86338	900000	0.89089	0.01318	0.09593	0.01458	0.98542	0.89089	1.51	2.53171
Standard	Asia	Nov16	802564	11026	86410	900000	0.89174	0.01225	0.09601	0.01355	0.98645	0.89174	0.70	2.53166
Standard	Asia	Nov17	802564	11026	86410	900000	0.89174	0.01225	0.09601	0.01355	0.98645	0.89174	0.76	2.53166
Standard	Asia	BWA1	825416	71350	3234	900000	0.91713	0.07928	0.00359	0.07956	0.92044	0.91713	20.20	2.53632
Standard	Asia	BWA2	824827	71070	4103	900000	0.91647	0.07897	0.00456	0.07933	0.92067	0.91647	9.19	2.60712
Standard	Asia	BWA3	824675	71273	4052	900000	0.91631	0.07919	0.00450	0.07955	0.92045	0.91631	10.10	2.60654
Standard	Asia	BWA4	824846	71028	4126	900000	0.91650	0.07892	0.00458	0.07928	0.92072	0.91650	11.77	2.60493
Standard	Asia	BWA5	824968	70891	4141	900000	0.91663	0.07877	0.00460	0.07913	0.92087	0.91663	9.96	2.60468

Standard	Asia	BWA6	825979	71255	2766	900000	0.91775	0.07917	0.00307	0.07942	0.92058	0.91775	14.50	2.59084
Standard	Asia	BWA7	825444	71202	3354	900000	0.91716	0.07911	0.00373	0.07941	0.92059	0.91716	12.64	0.89690
Standard	Asia	BWA8	818164	72516	9455	900135	0.90893	0.08056	0.01050	0.08142	0.91858	0.90893	1.18	2.59588
Standard	Asia	BWT1	815018	77145	7837	900000	0.90558	0.08572	0.00871	0.08647	0.91353	0.90558	1.34	2.60498
Standard	Asia	BWT2	817506	80095	2399	900000	0.90834	0.08899	0.00267	0.08923	0.91077	0.90834	1.46	2.60854
Standard	Asia	BWT3	820767	75817	3416	900000	0.91196	0.08424	0.00380	0.08456	0.91544	0.91196	2.15	2.61920
Standard	Asia	BWT4	820079	78010	1911	900000	0.91120	0.08668	0.00212	0.08686	0.91314	0.91120	2.02	2.61859
Standard	Asia	BWT5	801478	97334	1188	900000	0.89053	0.10815	0.00132	0.10829	0.89171	0.89053	6.88	2.60131
Standard	Europe	Nov1	806817	11057	82126	900000	0.89646	0.01229	0.09125	0.01352	0.98648	0.89646	0.79	2.53702
Standard	Europe	Nov2	804628	10575	84797	900000	0.89403	0.01175	0.09422	0.01297	0.98703	0.89403	0.86	2.53341
Standard	Europe	Nov3	803187	10667	86146	900000	0.89243	0.01185	0.09572	0.01311	0.98689	0.89243	1.74	2.53189
Standard	Europe	Nov4	803783	11401	84816	900000	0.89309	0.01267	0.09424	0.01399	0.98601	0.89309	1.76	2.53342
Standard	Europe	Nov5	802828	11305	85867	900000	0.89203	0.01256	0.09541	0.01389	0.98611	0.89203	1.49	2.53213
Standard	Europe	Nov6	802987	10658	86355	900000	0.89221	0.01184	0.09595	0.01310	0.98690	0.89221	1.68	2.53126
Standard	Europe	Nov7	804628	10574	84798	900000	0.89403	0.01175	0.09422	0.01297	0.98703	0.89403	0.77	2.53341
Standard	Europe	Nov8	803527	11605	84868	900000	0.89281	0.01289	0.09430	0.01424	0.98576	0.89281	0.78	2.53335
Standard	Europe	Nov9	802491	11449	86060	900000	0.89166	0.01272	0.09562	0.01407	0.98593	0.89166	0.69	2.53190
Standard	Europe	Nov10	828823	70818	359	900000	0.92091	0.07869	0.00040	0.07872	0.92128	0.92091	1.30	2.63911
Standard	Europe	Nov11	804628	10575	84797	900000	0.89403	0.01175	0.09422	0.01297	0.98703	0.89403	0.74	2.53341
Standard	Europe	Nov12	803187	10667	86146	900000	0.89243	0.01185	0.09572	0.01311	0.98689	0.89243	1.77	2.53189
Standard	Europe	Nov13	803527	11605	84868	900000	0.89281	0.01289	0.09430	0.01424	0.98576	0.89281	0.85	2.53335
Standard	Europe	Nov14	802491	11449	86060	900000	0.89166	0.01272	0.09562	0.01407	0.98593	0.89166	0.68	2.53190
Standard	Europe	Nov15	802828	11305	85867	900000	0.89203	0.01256	0.09541	0.01389	0.98611	0.89203	1.64	2.53213
Standard	Europe	Nov16	803601	10467	85932	900000	0.89289	0.01163	0.09548	0.01286	0.98714	0.89289	0.87	2.53206
Standard	Europe	Nov17	803601	10467	85932	900000	0.89289	0.01163	0.09548	0.01286	0.98714	0.89289	0.70	2.53206
Standard	Europe	BWA1	826303	70514	3183	900000	0.91811	0.07835	0.00354	0.07863	0.92137	0.91811	18.60	2.53920
Standard	Europe	BWA2	825793	70349	3858	900000	0.91755	0.07817	0.00429	0.07850	0.92150	0.91755	8.64	2.60807

Standard	Europe	BWA3	826146	70030	3824	900000	0.91794	0.07781	0.00425	0.07814	0.92186	0.91794	9.36	2.60778
Standard	Europe	BWA4	825976	70175	3849	900000	0.91775	0.07797	0.00428	0.07831	0.92169	0.91775	12.71	2.60678
Standard	Europe	BWA5	825906	70243	3851	900000	0.91767	0.07805	0.00428	0.07838	0.92162	0.91767	8.89	2.60628
Standard	Europe	BWA6	826789	70589	2622	900000	0.91865	0.07843	0.00291	0.07866	0.92134	0.91865	18.14	2.59225
Standard	Europe	BWA7	826538	70365	3097	900000	0.91838	0.07818	0.00344	0.07845	0.92155	0.91838	10.57	2.60683
Standard	Europe	BWA8	818770	72013	9339	900122	0.90962	0.08000	0.01038	0.08084	0.91916	0.90962	0.88	2.59636
Standard	Europe	BWT1	816064	76198	7738	900000	0.90674	0.08466	0.00860	0.08540	0.91460	0.90674	1.17	2.60571
Standard	Europe	BWT2	818489	79144	2367	900000	0.90943	0.08794	0.00263	0.08817	0.91183	0.90943	1.72	2.60982
Standard	Europe	BWT3	821815	74793	3392	900000	0.91313	0.08310	0.00377	0.08342	0.91658	0.91313	2.12	2.61991
Standard	Europe	BWT4	821349	76759	1892	900000	0.91261	0.08529	0.00210	0.08547	0.91453	0.91261	2.02	2.61994
Standard	Europe	BWT5	802304	96512	1184	900000	0.89145	0.10724	0.00132	0.10738	0.89262	0.89145	6.66	2.60994
IUPAC	Africa	Nov1	799333	15717	84950	900000	0.88815	0.01746	0.09439	0.01928	0.98072	0.88815	12.93	2.52299
IUPAC	Africa	Nov2	799875	16876	83249	900000	0.88875	0.01875	0.09250	0.02066	0.97934	0.88875	5.30	2.52728
IUPAC	Africa	Nov10	823318	76310	372	900000	0.91480	0.08479	0.00041	0.08482	0.91518	0.91480	6.14	2.62880
IUPAC	Africa	BWA2	750875	94783	54342	900000	0.83431	0.10531	0.06038	0.11208	0.88792	0.83431	58.64	1.75051
IUPAC	Africa	BWA7	759379	95644	44977	900000	0.84375	0.10627	0.04997	0.11186	0.88814	0.84375	97.19	1.72641
IUPAC	Africa	BWA8	761223	108932	32686	902841	0.84314	0.12065	0.03620	0.12519	0.87481	0.84314	1.81	2.51351
IUPAC	Africa	BWT1	719686	110911	69403	900000	0.79965	0.12323	0.07711	0.13353	0.86647	0.79965	1.43	2.39948
IUPAC	Africa	BWT3	760350	105064	34586	900000	0.84483	0.11674	0.03843	0.12140	0.87860	0.84483	2.34	2.50033
IUPAC	Asia	Nov1	802162	15800	82038	900000	0.89129	0.01756	0.09115	0.01932	0.98068	0.89129	1.79	2.53259
IUPAC	Asia	Nov2	800642	16084	83274	900000	0.88960	0.01787	0.09253	0.01969	0.98031	0.88960	1.92	2.53035
IUPAC	Asia	Nov10	824484	75174	342	900000	0.91609	0.08353	0.00038	0.08356	0.91644	0.91609	2.22	2.63105
IUPAC	Asia	BWA2	751265	94671	54064	900000	0.83474	0.10519	0.06007	0.11191	0.88809	0.83474	31.38	2.31488
IUPAC	Asia	BWA7	760048	95202	44750	900000	0.84450	0.10578	0.04972	0.11131	0.88869	0.84450	38.32	2.28406
IUPAC	Asia	BWA8	761949	106495	32816	901260	0.84543	0.11816	0.03641	0.12263	0.87737	0.84543	1.90	2.51444
IUPAC	Asia	BWT1	720220	110378	69402	900000	0.80024	0.12264	0.07711	0.13289	0.86711	0.80024	1.15	2.43433
IUPAC	Asia	BWT3	761220	104451	34329	900000	0.84580	0.11606	0.03814	0.12066	0.87934	0.84580	2.20	2.51699

IUPAC	Europe	Nov1	802952	15229	81819	900000	0.89217	0.01692	0.09091	0.01861	0.98139	0.89217	1.82	2.53217
IUPAC	Europe	Nov2	801425	15498	83077	900000	0.89047	0.01722	0.09231	0.01897	0.98103	0.89047	1.63	2.52996
IUPAC	Europe	Nov10	825112	74531	357	900000	0.91679	0.08281	0.00040	0.08285	0.91715	0.91679	2.45	2.63087
IUPAC	Europe	BWA2	752098	93910	53992	900000	0.83566	0.10434	0.05999	0.11100	0.88900	0.83566	33.48	2.31491
IUPAC	Europe	BWA7	760996	94340	44664	900000	0.84555	0.10482	0.04963	0.11030	0.88970	0.84555	39.54	2.28395
IUPAC	Europe	BWA8	762360	106260	32679	901299	0.84585	0.11790	0.03626	0.12233	0.87767	0.84585	1.65	2.51455
IUPAC	Europe	BWT1	721042	109497	69461	900000	0.80116	0.12166	0.07718	0.13184	0.86816	0.80116	1.15	2.43433
IUPAC	Europe	BWT3	761844	103653	34503	900000	0.84649	0.11517	0.03834	0.11976	0.88024	0.84649	2.44	2.51686
H-UDG	Asia	Nov1	805566	10820	83614	900000	0.89507	0.01202	0.09290	0.01325	0.98675	0.89507	0.95	2.53663
H-UDG	Asia	Nov2	804939	10603	84458	900000	0.89438	0.01178	0.09384	0.01300	0.98700	0.89438	0.82	2.53490
H-UDG	Asia	Nov3	804230	10600	85170	900000	0.89359	0.01178	0.09463	0.01301	0.98699	0.89359	1.69	2.53410
H-UDG	Asia	Nov4	804110	11421	84469	900000	0.89346	0.01269	0.09385	0.01400	0.98600	0.89346	1.55	2.53490
H-UDG	Asia	Nov5	803560	11342	85098	900000	0.89284	0.01260	0.09455	0.01392	0.98608	0.89284	1.40	2.53401
H-UDG	Asia	Nov6	804113	10595	85292	900000	0.89346	0.01177	0.09477	0.01300	0.98700	0.89346	1.60	2.53360
H-UDG	Asia	Nov7	804939	10603	84458	900000	0.89438	0.01178	0.09384	0.01300	0.98700	0.89438	0.78	2.53490
H-UDG	Asia	Nov8	803841	11638	84521	900000	0.89316	0.01293	0.09391	0.01427	0.98573	0.89316	0.77	2.53483
H-UDG	Asia	Nov9	803218	11497	85285	900000	0.89246	0.01277	0.09476	0.01411	0.98589	0.89246	0.73	2.53376
H-UDG	Asia	Nov10	829487	70267	246	900000	0.92165	0.07807	0.00027	0.07810	0.92190	0.92165	1.38	2.64081
H-UDG	Asia	Nov11	804939	10603	84458	900000	0.89438	0.01178	0.09384	0.01300	0.98700	0.89438	0.74	2.53490
H-UDG	Asia	Nov12	804230	10600	85170	900000	0.89359	0.01178	0.09463	0.01301	0.98699	0.89359	1.66	2.53410
H-UDG	Asia	Nov13	803841	11638	84521	900000	0.89316	0.01293	0.09391	0.01427	0.98573	0.89316	0.77	2.53483
H-UDG	Asia	Nov14	803218	11497	85285	900000	0.89246	0.01277	0.09476	0.01411	0.98589	0.89246	0.69	2.53376
H-UDG	Asia	Nov15	803560	11342	85098	900000	0.89284	0.01260	0.09455	0.01392	0.98608	0.89284	1.37	2.53401
H-UDG	Asia	Nov16	804321	10520	85159	900000	0.89369	0.01169	0.09462	0.01291	0.98709	0.89369	0.68	2.53396
H-UDG	Asia	Nov17	804321	10520	85159	900000	0.89369	0.01169	0.09462	0.01291	0.98709	0.89369	0.72	2.53396
H-UDG	Asia	BWA1	826645	70497	2858	900000	0.91849	0.07833	0.00318	0.07858	0.92142	0.91849	26.18	2.54904
H-UDG	Asia	BWA2	827356	70127	2517	900000	0.91928	0.07792	0.00280	0.07814	0.92186	0.91928	12.05	2.61544

H-UDG	Asia	BWA3	827565	69939	2496	900000	0.91952	0.07771	0.00277	0.07793	0.92207	0.91952	9.46	2.61544
H-UDG	Asia	BWA4	827270	70166	2564	900000	0.91919	0.07796	0.00285	0.07818	0.92182	0.91919	10.15	2.61346
H-UDG	Asia	BWA5	827306	70117	2577	900000	0.91923	0.07791	0.00286	0.07813	0.92187	0.91923	10.19	2.61356
H-UDG	Asia	BWA6	827554	70355	2091	900000	0.91950	0.07817	0.00232	0.07835	0.92165	0.91950	14.41	2.59748
H-UDG	Asia	BWA7	827612	70172	2216	900000	0.91957	0.07797	0.00246	0.07816	0.92184	0.91957	9.25	2.61289
H-UDG	Asia	BWA8	823012	70415	6682	900109	0.91435	0.07823	0.00742	0.07881	0.92119	0.91435	1.01	2.60083
H-UDG	Asia	BWT1	822310	73789	3901	900000	0.91368	0.08199	0.00433	0.08234	0.91766	0.91368	1.06	2.61788
H-UDG	Asia	BWT2	822757	76089	1154	900000	0.91417	0.08454	0.00128	0.08465	0.91535	0.91417	1.55	2.61831
H-UDG	Asia	BWT3	825323	72872	1805	900000	0.91703	0.08097	0.00201	0.08113	0.91887	0.91703	2.10	2.62706
H-UDG	Asia	BWT4	822310	73789	3901	900000	0.91368	0.08199	0.00433	0.08234	0.91766	0.91368	1.15	2.61788
H-UDG	Asia	BWT5	807777	91282	941	900000	0.89753	0.10142	0.00105	0.10153	0.89847	0.89753	7.00	2.50742
F-UDG	Asia	Nov1	683262	8887	207851	900000	0.75918	0.00987	0.23095	0.01284	0.98716	0.75918	0.70	2.23520
F-UDG	Asia	Nov2	683280	8886	207834	900000	0.75920	0.00987	0.23093	0.01284	0.98716	0.75920	0.65	2.23522
F-UDG	Asia	Nov3	683117	8847	208036	900000	0.75902	0.00983	0.23115	0.01279	0.98721	0.75902	1.36	2.23495
F-UDG	Asia	Nov4	682683	9499	207818	900000	0.75854	0.01055	0.23091	0.01372	0.98628	0.75854	1.19	2.23523
F-UDG	Asia	Nov5	682210	9408	208382	900000	0.75801	0.01045	0.23154	0.01360	0.98640	0.75801	0.97	2.23490
F-UDG	Asia	Nov6	682731	8823	208446	900000	0.75859	0.00980	0.23161	0.01276	0.98724	0.75859	1.08	2.23475
F-UDG	Asia	Nov7	683280	8886	207834	900000	0.75920	0.00987	0.23093	0.01284	0.98716	0.75920	0.75	2.23522
F-UDG	Asia	Nov8	682458	9652	207890	900000	0.75829	0.01072	0.23099	0.01395	0.98605	0.75829	0.60	2.23513
F-UDG	Asia	Nov9	681932	9507	208561	900000	0.75770	0.01056	0.23173	0.01375	0.98625	0.75770	0.55	2.23462
F-UDG	Asia	Nov10	707305	80997	111698	900000	0.78589	0.09000	0.12411	0.10275	0.89725	0.78589	1.26	2.33151
F-UDG	Asia	Nov11	683280	8886	207834	900000	0.75920	0.00987	0.23093	0.01284	0.98716	0.75920	0.58	2.23522
F-UDG	Asia	Nov12	683117	8847	208036	900000	0.75902	0.00983	0.23115	0.01279	0.98721	0.75902	1.20	2.23495
F-UDG	Asia	Nov13	682458	9652	207890	900000	0.75829	0.01072	0.23099	0.01395	0.98605	0.75829	0.65	2.23513
F-UDG	Asia	Nov14	681932	9507	208561	900000	0.75770	0.01056	0.23173	0.01375	0.98625	0.75770	0.54	2.23462
F-UDG	Asia	Nov15	682210	9408	208382	900000	0.75801	0.01045	0.23154	0.01360	0.98640	0.75801	0.98	2.23490
F-UDG	Asia	Nov16	682742	8791	208467	900000	0.75860	0.00977	0.23163	0.01271	0.98729	0.75860	0.60	2.23482

F-UDG	Asia	Nov17	682742	8791	208467	900000	0.75860	0.00977	0.23163	0.01271	0.98729	0.75860	0.58	2.23482
F-UDG	Asia	BWA1	735956	162838	1206	900000	0.81773	0.18093	0.00134	0.18117	0.81883	0.81773	11.48	2.28662
F-UDG	Asia	BWA2	736254	162461	1285	900000	0.81806	0.18051	0.00143	0.18077	0.81923	0.81806	4.47	2.32699
F-UDG	Asia	BWA3	736252	162489	1259	900000	0.81806	0.18054	0.00140	0.18080	0.81920	0.81806	5.70	2.32653
F-UDG	Asia	BWA4	736161	162553	1286	900000	0.81796	0.18061	0.00143	0.18087	0.81913	0.81796	8.01	2.32600
F-UDG	Asia	BWA5	736074	162640	1286	900000	0.81786	0.18071	0.00143	0.18097	0.81903	0.81786	5.92	2.32587
F-UDG	Asia	BWA6	736246	162794	960	900000	0.81805	0.18088	0.00107	0.18108	0.81892	0.81805	8.11	2.31960
F-UDG	Asia	BWA7	736112	162772	1116	900000	0.81790	0.18086	0.00124	0.18108	0.81892	0.81790	6.74	2.32350
F-UDG	Asia	BWA8	634290	61229	204579	900098	0.70469	0.06802	0.22729	0.08803	0.91197	0.70469	0.93	2.29153
F-UDG	Asia	BWT1	733774	163910	2316	900000	0.81530	0.18212	0.00257	0.18259	0.81741	0.81530	1.04	2.32245
F-UDG	Asia	BWT2	733763	144878	21359	900000	0.81529	0.16098	0.02373	0.16489	0.83511	0.81529	1.21	2.32370
F-UDG	Asia	BWT3	735212	163524	1264	900000	0.81690	0.18169	0.00140	0.18195	0.81805	0.81690	2.07	2.32919
F-UDG	Asia	BWT4	733774	163910	2316	900000	0.81530	0.18212	0.00257	0.18259	0.81741	0.81530	1.12	2.32245
F-UDG	Asia	BWT5	729352	150174	20474	900000	0.81039	0.16686	0.02275	0.17074	0.82926	0.81039	6.17	2.24445
H-UDG IUPAC	Asia	Nov1	802312	14010	83678	900000	0.89146	0.01557	0.09298	0.01716	0.98284	0.89146	0.78	2.23495
H-UDG IUPAC	Asia	Nov10	802313	68545	29142	900000	0.89146	0.07616	0.32380	0.07690	0.92310	0.91423	0.69	2.31278
H-UDG IUPAC	Asia	Nov2	801477	14001	84522	900000	0.89053	0.01556	0.09391	0.01717	0.98283	0.89053	1.03	2.27875
F-UDG IUPAC	Asia	Nov1	746812	12573	140615	900000	0.82979	0.01397	0.15624	0.01656	0.98344	0.82979	0.57	2.24415
F-UDG IUPAC	Asia	Nov10	767477	65613	66910	900000	0.85275	0.07290	0.07434	0.07876	0.92124	0.85275	0.61	2.23522
F-UDG IUPAC	Asia	Nov2	746819	12573	140608	900000	0.82980	0.01397	0.15623	0.01656	0.98344	0.82980	0.98	2.25511

Appendix 2: Supplementary materials for Chapter 2

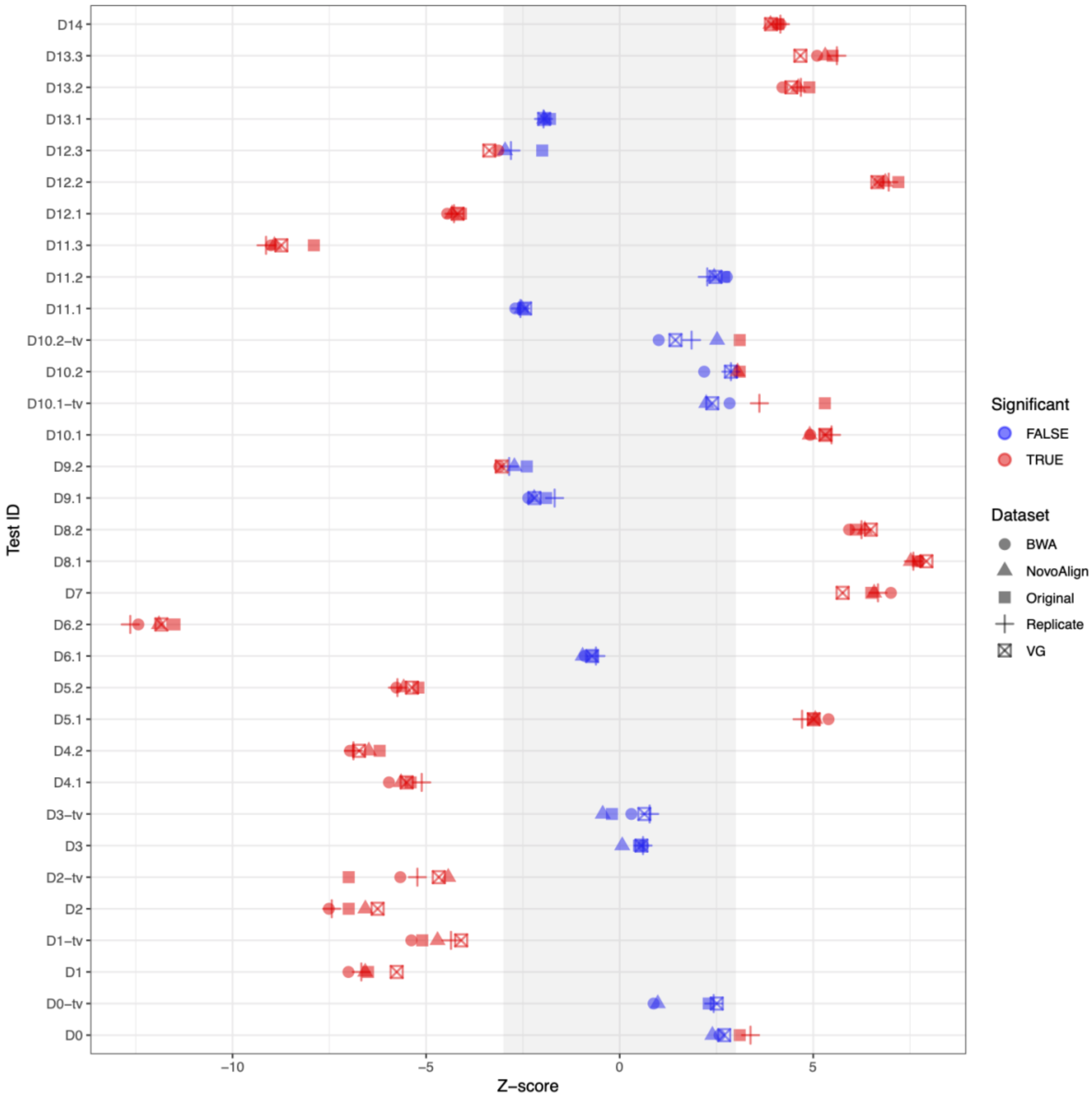


Figure S1. *Z*-score of all *D-statistics* calculated across the different alignment software evaluated. For the *D*-statistic, the *Z*-scores are plotted and values above or below 3 (grey zone) imply a significant excess of shared alleles between the populations of interest in that specific test. The colour of the dot does also signal this excess of shared alleles (red). The shape of each dot represents a software in the *Retested* dataset (*BWA-aln*, *NovoAlign*, *vg*) , the *Original* (i.e., the published results) or *Replicate* (results calculated using the 1240K dataset, v44.3).

TableS1:

Individuals, from the 1240K (v44.3) dataset, included in each population.

Individual	Label	Population	Source
HGDP00540	S_Papuan-2.DG	Papuan	SkoglundNature2015 [161]
HGDP00541	S_Papuan-3.DG	Papuan	SkoglundNature2015 [161]
HGDP00542	S_Papuan-9.DG	Papuan	SkoglundNature2015 [161]
HGDP00543	S_Papuan-4.DG	Papuan	SkoglundNature2015 [161]
HGDP00545	S_Papuan-5.DG	Papuan	SkoglundNature2015 [161]
HGDP00546	B_Papuan-15.DG	Papuan	PrueferNature2013 [166]
HGDP00547	S_Papuan-6.DG	Papuan	SkoglundNature2015 [161]
HGDP00548	S_Papuan-7.DG	Papuan	SkoglundNature2015 [161]
HGDP00549	S_Papuan-8.DG	Papuan	SkoglundNature2015 [161]
HGDP00550	S_Papuan-1.DG	Papuan	SkoglundNature2015 [161]
HGDP00551	A_Papuan-16.DG	Papuan	MeyerScience2012 [72]
HGDP00552	S_Papuan-13.DG	Papuan	SkoglundNature2015 [161]
HGDP00553	S_Papuan-10.DG	Papuan	SkoglundNature2015 [161]
HGDP00554	S_Papuan-14.DG	Papuan	SkoglundNature2015 [161]

HGDP00555	S_Papuan-11.DG	Papuan	SkoglundNature2015 [161]
HGDP00556	S_Papuan-12.DG	Papuan	SkoglundNature2015 [161]
mixe0007	B_Mixe-1.DG	Mixe	PrueferNature2013 [166]
S_Mixe-2	S_Mixe-2.DG	Mixe	SkoglundNature2015 [161]
S_Mixe-3	S_Mixe-3.DG	Mixe	SkoglundNature2015 [161]
HGDP00998	A_Karitiana-4.DG	Karitiana	MeyerScience2012 [72]
HGDP01012	B_Karitiana-1.DG	Karitiana	MeyerScience2012 [72]
HGDP01015	S_Karitiana-3.DG	Karitiana	MallickNature2016 [165]
HGDP01018	S_Karitiana-2.DG	Karitiana	MallickNature2016 [165]
HGDP00449	S_Mbuti-3.DG	Mbuti	MallickNature2016 [165]
HGDP00982	B_Mbuti-4.DG	Mbuti	PrueferNature2013 [166]
HGDP00476	S_Mbuti-2.DG	Mbuti	MallickNature2016 [165]
HGDP00474	S_Mbuti-1.DG	Mbuti	MallickNature2016 [165]
Yana1	Yana_old.SG	Yana	SikoraNature2019 [158]
Tianyuan	Tianyuan	Tianyuan	YangCurrentBiology2017 [157]
MA1	MA1.SG	Mal’ta	RaghavanNature2013 [155]

HGDP00846	S_Surui-1.DG	Surui	SkoglundNature2015 [161]
HGDP00852	S_Surui-2.DG	Surui	SkoglundNature2015 [161]
Vestonice16	Vestonice16	Vestonice	FuNature2016[156]

Table S2:

Details on the *D*-statistics computed for each individual. The data is merged using the 1240K dataset (v44.3). Note that for the "Original" dataset, some information is unknown in the initial publication and then the values are set at 0.

Test ID	Dataset	W	X	Y	Z	D	stderr	Zscore	BABA	ABBA	nsnps
D14	Original	Surui	Mixe.DG	Tianyuan	Mbuti.DG	0.019	0	4.1	0	0	0
D13.1	Original	Tianyuan	Surui	Yana	Mbuti.DG	0	0	-1.8	0	0	0
D13.2	Original	Goyet	Surui	Yana	Mbuti.DG	0	0	4.9	0	0	0
D13.3	Original	Malta	Surui	Yana	Mbuti.DG	0	0	5.5	0	0	0
D12.1	Original	Yana	Surui	Tianyuan	Mbuti.DG	0	0	-4.1	0	0	0
D12.2	Original	Yana	Surui	Goyet	Mbuti.DG	0	0	7.2	0	0	0
D12.3	Original	Yana	Surui	Malta	Mbuti.DG	0	0	-2	0	0	0
D11.1	Original	Yana	Tianyuan	Surui	Mbuti.DG	0	0	-2.5	0	0	0
D11.2	Original	Yana	Goyet	Surui	Mbuti.DG	0	0	2.7	0	0	0
D11.3	Original	Yana	Malta	Surui	Mbuti.DG	0	0	-7.9	0	0	0
D10.1	Original	Tianyuan	Papuan	Goyet	Mbuti.DG	0	0	5.3	0	0	0

D10.2	Original	Tianyuan	Papuan	Malta	Mbuti.DG	0	0	3.1	0	0	0
D9.1	Original	Goyet	Papuan	Tianyuan	Mbuti.DG	0	0	-1.9	0	0	0
D9.2	Original	Malta	Papuan	Tianyuan	Mbuti.DG	0	0	-2.4	0	0	0
D8.1	Original	Tianyuan	Goyet	Papuan	Mbuti.DG	0	0	7.7	0	0	0
D8.2	Original	Tianyuan	Malta	Papuan	Mbuti.DG	0	0	6.1	0	0	0
D7	Original	Malta	Tianyuan	Goyet	Mbuti.DG	0	0	6.5	0	0	0
D6.1	Original	Tianyuan	Surui	Goyet	Mbuti.DG	0	0	-0.7	0	0	0
D6.2	Original	Tianyuan	Surui	Malta	Mbuti.DG	0	0	-11.5	0	0	0
D5.1	Original	Tianyuan	Goyet	Surui	Mbuti.DG	0	0	5	0	0	0
D5.2	Original	Tianyuan	Malta	Surui	Mbuti.DG	0	0	-5.2	0	0	0
D4.1	Original	Goyet	Surui	Tianyuan	Mbuti.DG	0	0	-5.4	0	0	0
D4.2	Original	Malta	Surui	Tianyuan	Mbuti.DG	0	0	-6.2	0	0	0
D3	Original	Goyet	Malta	Tianyuan	Mbuti.DG	0.004	0	0.5	0	0	0
D2	Original	Tianyuan	Goyet	Malta	Mbuti.DG	-0.05	0	-7	0	0	0
D1	Original	Tianyuan	Malta	Goyet	Mbuti.DG	-0.05	0	-6.5	0	0	0
D0	Original	Goyet	Vestonice16	Tianyuan	Mbuti.DG	0.02	0	3.1	0	0	0

D14	Replicate	Surui	Mixe.DG	Tianyuan	Mbuti.DG	0.0186	0.004478	4.152	38311	36913	869113
D13.1	Replicate	Tianyuan	Surui	Yana	Mbuti.DG	-0.0115	0.005865	-1.967	46212	47291	869082
D13.2	Replicate	Goyet	Surui	Yana	Mbuti.DG	0.0314	0.006717	4.681	43806	41136	751911
D13.3	Replicate	Malta	Surui	Yana	Mbuti.DG	0.0357	0.006357	5.611	43736	40724	787389
D12.1	Replicate	Yana	Surui	Tianyuan	Mbuti.DG	-0.0267	0.006235	-4.28	46212	48746	869082
D12.2	Replicate	Yana	Surui	Goyet	Mbuti.DG	0.045	0.006478	6.95	43806	40031	751911
D12.3	Replicate	Yana	Surui	Malta	Mbuti.DG	-0.0185	0.006594	-2.809	43736	45387	787389
D11.1	Replicate	Yana	Tianyuan	Surui	Mbuti.DG	-0.0151	0.005908	-2.564	47291	48746	869082
D11.2	Replicate	Yana	Goyet	Surui	Mbuti.DG	0.0136	0.006012	2.263	41136	40031	751911
D11.3	Replicate	Yana	Malta	Surui	Mbuti.DG	-0.0542	0.005929	-9.135	40724	45387	787389
D10.1	Replicate	Tianyuan	Papuan	Goyet	Mbuti.DG	0.0332	0.006077	5.472	37531	35116	669848
D10.2	Replicate	Tianyuan	Papuan	Malta	Mbuti.DG	0.017	0.005908	2.874	34209	33067	626240
D9.1	Replicate	Goyet	Papuan	Tianyuan	Mbuti.DG	-0.0106	0.006278	-1.681	37531	38332	669848
D9.2	Replicate	Malta	Papuan	Tianyuan	Mbuti.DG	-0.0179	0.006255	-2.855	34209	35453	626240
D8.1	Replicate	Tianyuan	Goyet	Papuan	Mbuti.DG	0.0438	0.005771	7.587	38332	35116	669848
D8.2	Replicate	Tianyuan	Malta	Papuan	Mbuti.DG	0.0348	0.005575	6.247	35453	33067	626240

D7	Replicate	Malta	Tianyuan	Goyet	Mbuti.DG	0.0482	0.007224	6.673	28589	25960	487477
D6.1	Replicate	Tianyuan	Surui	Goyet	Mbuti.DG	-0.0037	0.006103	-0.613	36024	36295	669740
D6.2	Replicate	Tianyuan	Surui	Malta	Mbuti.DG	-0.078	0.00617	-12.642	31292	36586	626135
D5.1	Replicate	Tianyuan	Goyet	Surui	Mbuti.DG	0.0296	0.006277	4.712	38507	36295	669740
D5.2	Replicate	Tianyuan	Malta	Surui	Mbuti.DG	-0.0353	0.006155	-5.741	34089	36586	626135
D4.1	Replicate	Goyet	Surui	Tianyuan	Mbuti.DG	-0.0333	0.006514	-5.114	36024	38507	669740
D4.2	Replicate	Malta	Surui	Tianyuan	Mbuti.DG	-0.0428	0.006219	-6.878	31292	34089	626135
D3	Replicate	Goyet	Malta	Tianyuan	Mbuti.DG	0.0043	0.0072	0.603	25960	25735	487477
D2	Replicate	Tianyuan	Goyet	Malta	Mbuti.DG	-0.0525	0.007063	-7.438	25735	28589	487477
D1	Replicate	Tianyuan	Malta	Goyet	Mbuti.DG	-0.0482	0.007224	-6.673	25960	28589	487477
D0	Replicate	Goyet	Vestonice16.DG	Tianyuan	Mbuti.DG	0.0244	0.007217	3.38	28752	27383	545484
D14	VG	Surui_DR	Mixe_DR	Tianyuan	Mbuti_DR	0.018	0.004593	3.91	37262	35947	844522
D13.1	VG	Tianyuan	Surui_DR	Yana	Mbuti_DR	-0.0122	0.006252	-1.951	44901	46010	842486
D13.2	VG	Goyet	Surui_DR	Yana	Mbuti_DR	0.029	0.006544	4.435	44689	42169	766868
D13.3	VG	Malta	Surui_DR	Yana	Mbuti_DR	0.0298	0.00638	4.669	45515	42882	818891
D12.1	VG	Yana	Surui_DR	Tianyuan	Mbuti_DR	-0.0265	0.006335	-4.186	44901	47347	842486

D12.2	VG	Yana	Surui_DR	Goyet	Mbuti_DR	0.0438	0.006581	6.659	44689	40937	766868
D12.3	VG	Yana	Surui_DR	Malta	Mbuti_DR	-0.0221	0.006573	-3.367	45515	47576	818891
D11.1	VG	Yana	Tianyuan	Surui_DR	Mbuti_DR	-0.0143	0.005864	-2.444	46010	47347	842486
D11.2	VG	Yana	Goyet	Surui_DR	Mbuti_DR	0.0148	0.005989	2.476	42169	40937	766868
D11.3	VG	Yana	Malta	Surui_DR	Mbuti_DR	-0.0519	0.005936	-8.741	42882	47576	818891
D10.1	VG	Tianyuan	Papuan_DR	Goyet	Mbuti_DR	0.0317	0.005975	5.313	37055	34774	662413
D10.2	VG	Tianyuan	Papuan_DR	Malta	Mbuti_DR	0.0173	0.006026	2.874	34690	33509	633052
D9.1	VG	Goyet	Papuan_DR	Tianyuan	Mbuti_DR	-0.0138	0.00624	-2.204	37055	38088	662413
D9.2	VG	Malta	Papuan_DR	Tianyuan	Mbuti_DR	-0.0199	0.006514	-3.047	34690	36095	633052
D8.1	VG	Tianyuan	Goyet	Papuan_DR	Mbuti_DR	0.0455	0.005741	7.922	38088	34774	662413
D8.2	VG	Tianyuan	Malta	Papuan_DR	Mbuti_DR	0.0372	0.005728	6.487	36095	33509	633052
D7	VG	Malta	Tianyuan	Goyet	Mbuti_DR	0.042	0.0073	5.76	29344	26976	501064
D6.1	VG	Tianyuan	Surui_DR	Goyet	Mbuti_DR	-0.0044	0.006104	-0.713	35638	35950	662330
D6.2	VG	Tianyuan	Surui_DR	Malta	Mbuti_DR	-0.0761	0.006432	-11.838	31839	37087	632965
D5.1	VG	Tianyuan	Goyet	Surui_DR	Mbuti_DR	0.0308	0.006138	5.013	38232	35950	662330
D5.2	VG	Tianyuan	Malta	Surui_DR	Mbuti_DR	-0.0337	0.006283	-5.365	34668	37087	632965

D4.1	VG	Goyet	Surui_DR	Tianyuan	Mbuti_DR	-0.0351	0.006382	-5.503	35638	38232	662330
D4.2	VG	Malta	Surui_DR	Tianyuan	Mbuti_DR	-0.0425	0.006319	-6.732	31839	34668	632965
D3	VG	Goyet	Malta	Tianyuan	Mbuti_DR	0.0041	0.007376	0.562	26976	26753	501064
D2	VG	Tianyuan	Goyet	Malta	Mbuti_DR	-0.0462	0.007387	-6.253	26753	29344	501064
D1	VG	Tianyuan	Malta	Goyet	Mbuti_DR	-0.042	0.0073	-5.76	26976	29344	501064
D0	VG	Goyet	Vestonice	Tianyuan	Mbuti_DR	0.0195	0.007224	2.7	28751	27651	543549
D14	NovoAlign	Surui_DR	Mixe_DR	Tianyuan	Mbuti_DR	0.0176	0.00451	3.896	36920	35646	837556
D13.1	NovoAlign	Tianyuan	Surui_DR	Yana	Mbuti_DR	-0.0123	0.006292	-1.951	44481	45587	836316
D13.2	NovoAlign	Goyet	Surui_DR	Yana	Mbuti_DR	0.0315	0.006825	4.614	44189	41491	759077
D13.3	NovoAlign	Malta	Surui_DR	Yana	Mbuti_DR	0.0347	0.006529	5.309	44821	41817	806555
D12.1	NovoAlign	Yana	Surui_DR	Tianyuan	Mbuti_DR	-0.0273	0.006297	-4.332	44481	46976	836316
D12.2	NovoAlign	Yana	Surui_DR	Goyet	Mbuti_DR	0.0461	0.006717	6.861	44189	40295	759077
D12.3	NovoAlign	Yana	Surui_DR	Malta	Mbuti_DR	-0.0195	0.006588	-2.96	44821	46604	806555
D11.1	NovoAlign	Yana	Tianyuan	Surui_DR	Mbuti_DR	-0.015	0.005932	-2.53	45587	46976	836316
D11.2	NovoAlign	Yana	Goyet	Surui_DR	Mbuti_DR	0.0146	0.005992	2.44	41491	40295	759077
D11.3	NovoAlign	Yana	Malta	Surui_DR	Mbuti_DR	-0.0541	0.00607	-8.919	41817	46604	806555

D10.1	NovoAlign	Tianyuan	Papuan_DR	Goyet	Mbuti_DR	0.0293	0.005964	4.909	36591	34509	657166
D10.2	NovoAlign	Tianyuan	Papuan_DR	Malta	Mbuti_DR	0.0183	0.005999	3.045	34319	33088	625551
D9.1	NovoAlign	Goyet	Papuan_DR	Tianyuan	Mbuti_DR	-0.0138	0.006242	-2.208	36591	37614	657166
D9.2	NovoAlign	Malta	Papuan_DR	Tianyuan	Mbuti_DR	-0.0177	0.006502	-2.72	34319	35555	625551
D8.1	NovoAlign	Tianyuan	Goyet	Papuan_DR	Mbuti_DR	0.043	0.005725	7.519	37614	34509	657166
D8.2	NovoAlign	Tianyuan	Malta	Papuan_DR	Mbuti_DR	0.0359	0.005676	6.332	35555	33088	625551
D7	NovoAlign	Malta	Tianyuan	Goyet	Mbuti_DR	0.0486	0.007399	6.574	29045	26350	495233
D6.1	NovoAlign	Tianyuan	Surui_DR	Goyet	Mbuti_DR	-0.0058	0.006012	-0.957	35252	35660	657085
D6.2	NovoAlign	Tianyuan	Surui_DR	Malta	Mbuti_DR	-0.0765	0.00643	-11.902	31430	36640	625468
D5.1	NovoAlign	Tianyuan	Goyet	Surui_DR	Mbuti_DR	0.0305	0.00603	5.053	37902	35660	657085
D5.2	NovoAlign	Tianyuan	Malta	Surui_DR	Mbuti_DR	-0.0356	0.006376	-5.579	34123	36640	625468
D4.1	NovoAlign	Goyet	Surui_DR	Tianyuan	Mbuti_DR	-0.0362	0.006411	-5.649	35252	37902	657085
D4.2	NovoAlign	Malta	Surui_DR	Tianyuan	Mbuti_DR	-0.0411	0.006337	-6.481	31430	34123	625468
D3	NovoAlign	Goyet	Malta	Tianyuan	Mbuti_DR	5.00E-04	0.007383	0.066	26350	26324	495233
D2	NovoAlign	Tianyuan	Goyet	Malta	Mbuti_DR	-0.0491	0.007478	-6.57	26324	29045	495233
D1	NovoAlign	Tianyuan	Malta	Goyet	Mbuti_DR	-0.0486	0.007399	-6.574	26350	29045	495233

D0	NovoAlign	Goyet	Vestonice	Tianyuan	Mbuti_DR	0.0171	0.007121	2.395	28416	27462	539815
D14	BWA	Surui_DR	Mixe_DR	Tianyuan	Mbuti_DR	0.0187	0.004497	4.157	37306	35937	844526
D13.1	BWA	Tianyuan	Surui_DR	Yana	Mbuti_DR	-0.0125	0.006397	-1.953	44724	45856	842387
D13.2	BWA	Goyet	Surui_DR	Yana	Mbuti_DR	0.0281	0.006676	4.204	44615	42179	766889
D13.3	BWA	Malta	Surui_DR	Yana	Mbuti_DR	0.0336	0.006585	5.1	44846	41932	807832
D12.1	BWA	Yana	Surui_DR	Tianyuan	Mbuti_DR	-0.0282	0.00633	-4.455	44724	47320	842387
D12.2	BWA	Yana	Surui_DR	Goyet	Mbuti_DR	0.0442	0.00661	6.688	44615	40838	766889
D12.3	BWA	Yana	Surui_DR	Malta	Mbuti_DR	-0.0208	0.006653	-3.133	44846	46755	807832
D11.1	BWA	Yana	Tianyuan	Surui_DR	Mbuti_DR	-0.0157	0.005831	-2.696	45856	47320	842387
D11.2	BWA	Yana	Goyet	Surui_DR	Mbuti_DR	0.0162	0.005837	2.768	42179	40838	766889
D11.3	BWA	Yana	Malta	Surui_DR	Mbuti_DR	-0.0544	0.006038	-9.007	41932	46755	807832
D10.1	BWA	Tianyuan	Papuan_DR	Goyet	Mbuti_DR	0.0294	0.005979	4.922	36963	34850	662556
D10.2	BWA	Tianyuan	Papuan_DR	Malta	Mbuti_DR	0.0133	0.006093	2.183	34146	33250	624993
D9.1	BWA	Goyet	Papuan_DR	Tianyuan	Mbuti_DR	-0.0147	0.006216	-2.361	36963	38064	662556
D9.2	BWA	Malta	Papuan_DR	Tianyuan	Mbuti_DR	-0.0202	0.006502	-3.109	34146	35555	624993
D8.1	BWA	Tianyuan	Goyet	Papuan_DR	Mbuti_DR	0.0441	0.0057	7.735	38064	34850	662556

D8.2	BWA	Tianyuan	Malta	Papuan_DR	Mbuti_DR	0.0335	0.005655	5.926	35555	33250	624993
D7	BWA	Malta	Tianyuan	Goyet	Mbuti_DR	0.0518	0.007398	7.006	29219	26340	494662
D6.1	BWA	Tianyuan	Surui_DR	Goyet	Mbuti_DR	-0.0054	0.005991	-0.9	35601	35987	662475
D6.2	BWA	Tianyuan	Surui_DR	Malta	Mbuti_DR	-0.0811	0.006524	-12.434	31223	36736	624911
D5.1	BWA	Tianyuan	Goyet	Surui_DR	Mbuti_DR	0.0322	0.005972	5.395	38383	35987	662475
D5.2	BWA	Tianyuan	Malta	Surui_DR	Mbuti_DR	-0.0366	0.006349	-5.768	34140	36736	624911
D4.1	BWA	Goyet	Surui_DR	Tianyuan	Mbuti_DR	-0.0376	0.00631	-5.96	35601	38383	662475
D4.2	BWA	Malta	Surui_DR	Tianyuan	Mbuti_DR	-0.0446	0.006404	-6.968	31223	34140	624911
D3	BWA	Goyet	Malta	Tianyuan	Mbuti_DR	0.0044	0.007391	0.602	26340	26106	494662
D2	BWA	Tianyuan	Goyet	Malta	Mbuti_DR	-0.0563	0.00749	-7.511	26106	29219	494662
D1	BWA	Tianyuan	Malta	Goyet	Mbuti_DR	-0.0518	0.007398	-7.006	26340	29219	494662
D0	BWA	Goyet	Vestonice	Tianyuan	Mbuti_DR	0.0181	0.007003	2.578	28721	27703	543697

Table S3:

Details of the *f4-statistics* run in the substitutions tests. Those tests are run using the Replicate dataset (i.e., the 1240k shared genotype dataset).

ID	Individual removed	W	X	Y	Z	f4	stderr	Zscore	BABA	ABBA	nsnps
F4-5	B_Karitiana-1.DG (K1)	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001267	0.000353	3.59	38171	37069	869242
F4-4	B_Karitiana-1.DG (K1)	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	B_Karitiana-1.DG (K1)	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000957	0.000266	3.604	49451	48378	1121318
F4-2	B_Karitiana-1.DG (K1)	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001365	0.000313	4.359	49695	48164	1121148
F4-1	B_Karitiana-1.DG (K1)	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Karitiana-2.DG (K2)	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.000913	0.000361	2.528	38109	37316	869243
F4-4	S_Karitiana-2.DG (K2)	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Karitiana-2.DG (K2)	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000671	0.000265	2.532	49366	48613	1121323
F4-2	S_Karitiana-2.DG (K2)	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001365	0.000313	4.359	49695	48164	1121148
F4-1	S_Karitiana-2.DG (K2)	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202

F4-5	S_Karitiana-3.DG (K3)	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.00087	0.000365	2.383	38026	37270	869241
F4-4	S_Karitiana-3.DG (K3)	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Karitiana-3.DG (K3)	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000907	0.000275	3.297	49430	48413	1121324
F4-2	S_Karitiana-3.DG (K3)	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001365	0.000313	4.359	49695	48164	1121148
F4-1	S_Karitiana-3.DG (K3)	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG .SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	A_Karitiana-4.DG (K4)	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001009	0.000358	2.815	38020	37143	869208
F4-4	A_Karitiana-4.DG (K4)	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	A_Karitiana-4.DG (K4)	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000827	0.000274	3.022	49315	48389	1121282
F4-2	A_Karitiana-4.DG (K4)	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001365	0.000313	4.359	49695	48164	1121148
F4-1	A_Karitiana-4.DG (K4)	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG .SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	B_Mixe-1.DG (M1)	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001	0.000379	2.638	38094	37224	869136
F4-4	B_Mixe-1.DG (M1)	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001594	0.000423	3.772	38244	36859	869012
F4-3	B_Mixe-1.DG (M1)	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000765	0.000292	2.624	49400	48542	1121171
F4-2	B_Mixe-1.DG (M1)	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001294	0.000337	3.842	49630	48179	1120997

F4-1	B_Mixe-1.DG (M1)	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Mixe-2.DG (M2)	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.000969	0.00038	2.55	37995	37152	869100
F4-4	S_Mixe-2.DG (M2)	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001567	0.000407	3.852	38293	36931	868975
F4-3	S_Mixe-2.DG (M2)	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000779	0.000279	2.796	49254	48380	1121130
F4-2	S_Mixe-2.DG (M2)	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001497	0.000328	4.566	49808	48130	1121026
F4-1	S_Mixe-2.DG (M2)	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Mixe-3.DG (M3)	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.000969	0.00038	2.55	37995	37152	869100
F4-4	S_Mixe-3.DG (M3)	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001567	0.000407	3.852	38293	36931	868975
F4-3	S_Mixe-3.DG (M3)	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000779	0.000279	2.796	49254	48380	1121130
F4-2	S_Mixe-3.DG (M3)	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001305	0.000332	3.928	49626	48163	1120958
F4-1	S_Mixe-3.DG (M3)	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-1.DG & S_Papuan-2.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-1.DG & S_Papuan-2.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-1.DG & S_Papuan-2.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000798	0.000261	3.058	49367	48472	1121329
F4-2	S_Papuan-1.DG &	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001329	0.000313	4.239	49673	48184	1121148

	S_Papuan-2.DG										
F4-1	S_Papuan-1.DG & S_Papuan-2.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-1.DG & S_Papuan-9.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-1.DG & S_Papuan-9.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-1.DG & S_Papuan-9.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000836	0.000261	3.2	49388	48450	1121330
F4-2	S_Papuan-1.DG & S_Papuan-9.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001341	0.000314	4.276	49681	48177	1121148
F4-1	S_Papuan-1.DG & S_Papuan-9.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-10.DG & S_Papuan-11.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-10.DG & S_Papuan-11.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-10.DG & S_Papuan-11.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000842	0.000263	3.205	49397	48453	1121330
F4-2	S_Papuan-10.DG & S_Papuan-11.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001401	0.000314	4.465	49713	48143	1121148
F4-1	S_Papuan-10.DG & S_Papuan-11.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-13.DG & S_Papuan-14.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245

F4-4	S_Papuan-13.DG & S_Papuan-14.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-13.DG & S_Papuan-14.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000848	0.000261	3.251	49393	48442	1121330
F4-2	S_Papuan-13.DG & S_Papuan-14.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001358	0.000313	4.347	49688	48165	1121148
F4-1	S_Papuan-13.DG & S_Papuan-14.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	B_Papuan-15.DG & A_Papuan-16.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	B_Papuan-15.DG & A_Papuan-16.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	B_Papuan-15.DG & A_Papuan-16.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000812	0.000262	3.101	49363	48453	1121300
F4-2	B_Papuan-15.DG & A_Papuan-16.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001353	0.000315	4.298	49679	48161	1121119
F4-1	B_Papuan-15.DG & A_Papuan-16.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-3.DG & S_Papuan-8.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-3.DG & S_Papuan-8.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-3.DG & S_Papuan-8.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000855	0.00026	3.283	49408	48450	1121330
F4-2	S_Papuan-3.DG & S_Papuan-8.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001376	0.000314	4.382	49709	48166	1121148

F4-1	S_Papuan-3.DG & S_Papuan-8.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-4.DG & S_Papuan-6.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-4.DG & S_Papuan-6.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-4.DG & S_Papuan-6.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000846	0.00026	3.251	49392	48443	1121330
F4-2	S_Papuan-4.DG & S_Papuan-6.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001332	0.000313	4.257	49677	48184	1121148
F4-1	S_Papuan-4.DG & S_Papuan-6.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-5.DG & S_Papuan-10.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-5.DG & S_Papuan-10.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-5.DG & S_Papuan-10.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000877	0.000262	3.352	49422	48438	1121330
F4-2	S_Papuan-5.DG & S_Papuan-10.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001413	0.000314	4.499	49731	48147	1121148
F4-1	S_Papuan-5.DG & S_Papuan-10.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-7.DG & S_Papuan-12.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-7.DG & S_Papuan-12.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117

F4-3	S_Papuan-7.DG & S_Papuan-12.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000845	0.000262	3.225	49390	48443	1121330
F4-2	S_Papuan-7.DG & S_Papuan-12.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001401	0.000315	4.452	49711	48140	1121148
F4-1	S_Papuan-7.DG & S_Papuan-12.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202
F4-5	S_Papuan-8.DG & S_Papuan-9.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Karitiana.DG	0.001014	0.000349	2.91	38082	37200	869245
F4-4	S_Papuan-8.DG & S_Papuan-9.DG	Mbuti.DG	China_Tianyuan	Mixe.DG	Surui.DG	0.001609	0.000388	4.146	38312	36913	869117
F4-3	S_Papuan-8.DG & S_Papuan-9.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Karitiana.DG	0.000835	0.000261	3.204	49395	48459	1121330
F4-2	S_Papuan-8.DG & S_Papuan-9.DG	Mbuti.DG	Papuan.DG	Mixe.DG	Surui.DG	0.001345	0.000313	4.292	49689	48181	1121148
F4-1	S_Papuan-8.DG & S_Papuan-9.DG	Mbuti.DG	Yana_UP.SG	Russia_MA1_HG.SG	China_Tianyuan	-0.004736	0.000745	-6.358	33370	36336	626202