New and Refined Tools and Guidelines to Expand the Scope and Improve the Reproducibility of Palaeomicrobiological Research

Raphael A. Eisenhofer

Australian Centre for Ancient DNA
School of Biological Sciences
Faculty of Sciences
University of Adelaide
Australia

Advisers: Dr Laura S. Weyrich,
Prof. Alan Cooper,
Prof. Keith Dobney (University of Liverpool, UK)

Thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy

May 2018

Table of Contents

Thesis abstract

Microorganisms vastly outnumber animals and play key roles in our planet's biosphere. Recent advances in technology and computational tools have made it possible to study the great diversity of microorganisms on Earth rapidly and efficiently. A large fraction of this research has focused on the microbial communities that inhabit the human body—the human microbiota—which account for more than half of the cells we carry and collectively possess >100-fold more genes than the human genome. This research has discovered key coevolutionary relationships between the host and microbiota, many of which have been shaped through human history to the benefit of both partners. Evidence is mounting that disruptions to these microbial communities and to the relationship between the host and microbiota (dysbioses) can have a drastic effect on human health. There is also evidence that recent changes in human societies, such as antibiotic use and exposure to bioactive chemicals, have promoted dysbiosis to the detriment of human wellbeing. Thus, there is great interest in studying human microbiota that existed prior to these recent changes, with the hope of providing insight into the evolution of the human microbiota and informing modern medical strategies and the development of new therapies.

The recent finding that ancient microbial DNA is preserved in human dental calculus (calcified dental plaque) offers us the ability to investigate how oral microbiota have changed through human history. Additionally, further advances in DNA sequencing technology and laboratory methods have made it possible to rapidly process ancient specimens and to obtain large quantities of ancient microbial data. However, our ability to analyse this data has not caught up with the speed at which we generate it, and there are many analytical challenges and pitfalls that stand in the way of realising the full potential of ancient microbial DNA studies.

This thesis aims to develop and improve methods for analysing ancient microbial DNA and to identify and highlight challenges and pitfalls present in the field in order to increase the quality of future research. Initially, I propose a novel approach of using ancient microbial DNA in dental calculus as a proxy for determining past human migrations. Next, I develop and propose criteria to improve research standards in low-biomass microbiota research. I then assess

how characteristics of ancient DNA impact our ability to determine the composition of past microbiota and develop new analytical strategies and methods to improve current taxonomic identification approaches. I also generate and authenticate high-quality data for 132 new ancient dental calculus samples from the Asia-Pacific region, and develop and test two new methods to analyse this data and determine if oral microbiota can be used to infer past human migration and demographic history. Finally, I critically review and respond to three questionable palaeomicrobiological studies with the hope that future researchers, reviewers, and editors will learn from the issues highlighted. Ultimately, this thesis highlights and constructively addresses key pitfalls of palaeomicrobiological research and pushes the field closer to realising its potential.

Thesis Declaration

I, Raphael Eisenhofer, 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.

I give consent to this copy of my thesis when deposited in the University Library, being made available for loan and photocopying, subject to the provisions of the Copyright Act 1968.

I acknowledge that copyright of published works contained within this thesis resides with the copyright holder(s) of those works.

I also 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.

I acknowledge the support I have received for my research through the provision of an Australian Government Research Training Program Scholarship.

Raphael Eisenhofer: ' Date: /4/18

Publications

Journal articles

Eisenhofer, R.; Cooper, A.; Weyrich, L. S., 2016: Isolating Viable Ancient Bacteria: What You Put In Is What You Get Out. *Genome Announcements*., 4.

Eisenhofer, R.; Cooper, A.; Weyrich, L. S., 2017: Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential. *FEMS Microbiology Ecology*., 93.

Eisenhofer, R.; Anderson, A.; Dobney, K.; Cooper, A.; Weyrich, L. S., 2017: Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events. *The Journal of Island and Coastal Archaeology*., 0, 1–14.

Eisenhofer, R.; Weyrich, L. S., 2018: Proper Authentication of Ancient DNA Is Still Essential. *Genes*., 9, 122.

Grants and awards

2016: $1,500 — Royal Society of South Australia Small Research Grants: “Using the human microbiota to trace the final stages of Polynesian settlement”

Scientific outreach

2016: Blog posts for the Australian Centre for Ancient DNA scientific outreach program:
https://acadadelaide.wordpress.com/2016/04/08/micro-magicians-make-the-cheese/
https://acadadelaide.wordpress.com/2016/10/10/gmos-not-just-for-dinner/

2017: Australia Society of Microbiology (ASM) Communication Ambassador

2017: Guest teaching session at Nuriootpa High School

2017: Oral presentation to ~150 high school students taking part in the STEM Explorers program at the South Australian Museum

2017: Oral presentation about ancient DNA to Great Southern Rail visitors at the South Australian Museum

2017: Development and execution of a South Australian Museum Night Hack event

Acknowledgements

I owe a great deal to the many people who helped me during my candidature, both directly and indirectly.

Firstly, I would like to thank my principal supervisor Laura Weyrich, who spent her time and energy training me to become a scientist. In these past three years, you have pushed me to better myself personally and professionally, and helped to realise my goals. I appreciate everything you've done for me and look forward to working with you in the future. I would also like to thank Alan Cooper and Keith Dobney for their supervision, stimulating discussions, assistance, and opportunities given to me throughout my candidature.

Many thanks to the wonderful people at the Australian Centre for Ancient DNA (both past and present), who have all enriched my life and contributed to my development and learning. You've all been extremely welcoming and have made my life better both at work and outside.

A huge shout out to members of the metagenomics team both past and present: Laura, Andrew, Luis, Emily, Muslih, Yichen, Caitlin, Matilda, Michael, and Thibault. Special thanks to Andrew for training me in the tram-barn and for his friendship and support outside of work.

Special thanks to Kieren and Graham for the many hiking trips, bike rides, craft beers, board games, and great times had throughout my candidature.

Maria Lekis, thank you for your professional advice during my candidature and the many laughs. Thank you for also keeping ACAD together by giving us an earful when needed and all the administrative things behind the scenes you do to help us as a research group.

To my family in New Zealand, thank you for all of your support, especially to my mother Deanna, father Patrick, and stepfather and dear friend Marcel.

Lastly, and by no means least, I am eternally grateful to my partner, Matilda Handsley-Davis. Matilda, thank you for providing such joy to my life and for helping me to become a better person. Your constant love, support, and feedback have been invaluable to me during my candidature and life, and I look forward to spending my future with you.

Introduction

Overview

There is growing interest in studying how human microbiota have changed through time and how these changes might have influenced human health. Technological and methodological advances have accelerated the research of ancient dental plaque microbiota preserved in human dental calculus. However, our ability to realise the potential of this research rests on identifying the pitfalls present and developing new tools with which to analyse the burgeoning data. In this introduction, I briefly discuss the history of the ancient DNA field and highlight some of its difficulties, specifically the characteristics of ancient DNA and DNA contamination. I then describe the human microbiota and their recently associated links to human disease, before focusing specifically on the oral microbiota. I go on to review current knowledge of the dynamics of the dental plaque microbiota, highlighting the strong degree of vertical inheritance of this community and the syntrophic partnerships between its members. I then discuss how disruptions to these communities and to interactions with the human host can lead to oral diseases. Finally, I review the nascent field of palaeomicrobiology—focusing on dental calculus as a source of ancient human microbiota—and highlight how analytical techniques are critical to realising the prospects and dodging the pitfalls of this new field.

The difficulties of working with ancient DNA

A brief history and scope

Ancient DNA research provides direct measurements of genetic material from past organisms through the recovery of preserved DNA. The field of ancient DNA was kick-started with the characterisation of mitochondrial DNA obtained from a museum specimen of a quagga (extinct zebra) [1]. Since then, next-generation DNA sequencing (NGS) has revolutionised the field, unlocking the ability to investigate whole ancient genomes [2] and microbial communities [3]. Ancient DNA analysis has been successfully applied to a wide range of organisms, including mammoths [4], dogs [5], horses [6], bovids [7], rats [8], chickens [9], Tasmanian tigers [10], humans [2], Neanderthals [11], and microorganisms [3,12]. This research has allowed for the direct study of evolution through time, shedding light onto past population demographics and extinctions, genomic adaptation, pathogen evolution, and past population migrations. To date,

the oldest genome successfully recovered and analysed is ~700,000 years old, obtained from a horse bone buried in Yukon permafrost [6]. Current theoretical estimates place the limit of recoverable DNA to ~1 million years [13]. Ultimately, the ability to successfully obtain and analyse ancient DNA depends on the magnitude of its degradation through time.

Characteristics of ancient DNA

A distinctive feature of ancient DNA is its degraded nature. DNA repair mechanisms cease to function at the time of death, resulting in the fragmentation and chemical modification of DNA over time. An initial study into the properties of ancient DNA found that almost all extracted DNA from samples aged 4 to 13,000 years was degraded into fragments 40-500 bp (base pairs) in length [14]. *In vitro* experiments using modern DNA suggested that a major cause of DNA fragmentation is hydrolytic depurination followed by β-elimination, resulting in single-stranded breaks [15,16]. This was later supported by ancient DNA research which found an overrepresentation of purines (adenine and guanine) at the 5' ends of ancient DNA [17], which would be observed if hydrolytic depurination was causing DNA fragmentation. Furthermore, a recent methodological advancement examining single-stranded DNA molecules allowed for the finding that purines are also overrepresented at the 3' ends of ancient DNA fragments [18]. These studies provide a mechanistic basis for why we observe short DNA fragment length distributions from ancient samples (Figure 1A).

Chemical modifications can occur within these short DNA fragments and result in two major forms sequence modification: blocking and miscoding lesions [19]. Blocking lesions are DNA modifications that prevent the movement of polymerases along the template strand, preventing DNA amplification and sequencing. These blocking lesions can result from nucleotide modifications or cross-links within and/or between different DNA fragments or other molecules [19]. In contrast, miscoding lesions do not obstruct DNA polymerases but instead result in the incorporation of incorrect nucleotide. The most common miscoding lesion results from the hydrolytic deamination of cytosine to uracil, which results in the incorporation of adenine in place of guanine when read by most DNA polymerases [20]. This misincorporation results in the observed C to T or G to A substitutions (depending on the strand sequenced) characteristic of ancient DNA (Figure 1B). These substitutions are primarily observed at the ends of ancient DNA molecules and are likely due to the accelerated rate of cytosine deamination in single-stranded DNA, reflecting the occurrence of single-stranded overhangs at the end of ancient DNA molecules [21]. While posing a difficulty in certain bioinformatic analyses such as phylogenetic reconstruction, this characteristic misincorporation at the ends of ancient DNA

molecules has led to the development of a fundamental bioinformatic tool for assessing the authenticity of ancient DNA (Figure 1B) [22].

The rate of DNA degradation is not constant and depends on myriad factors, such as the environment (*e.g.* temperature, humidity, water flow, salinity, pH) and host physiology (histones, metabolic environment, cell wall, etc.). Recent studies into the dynamics of DNA degradation found that while cytosine deamination correlates with age of the sample, DNA fragmentation does not [23,24]. Ultimately, ancient DNA degradation results in small concentrations of highly fragmented and modified DNA (Figure 1), which require special laboratory protocols and care to successfully extract and analyse.

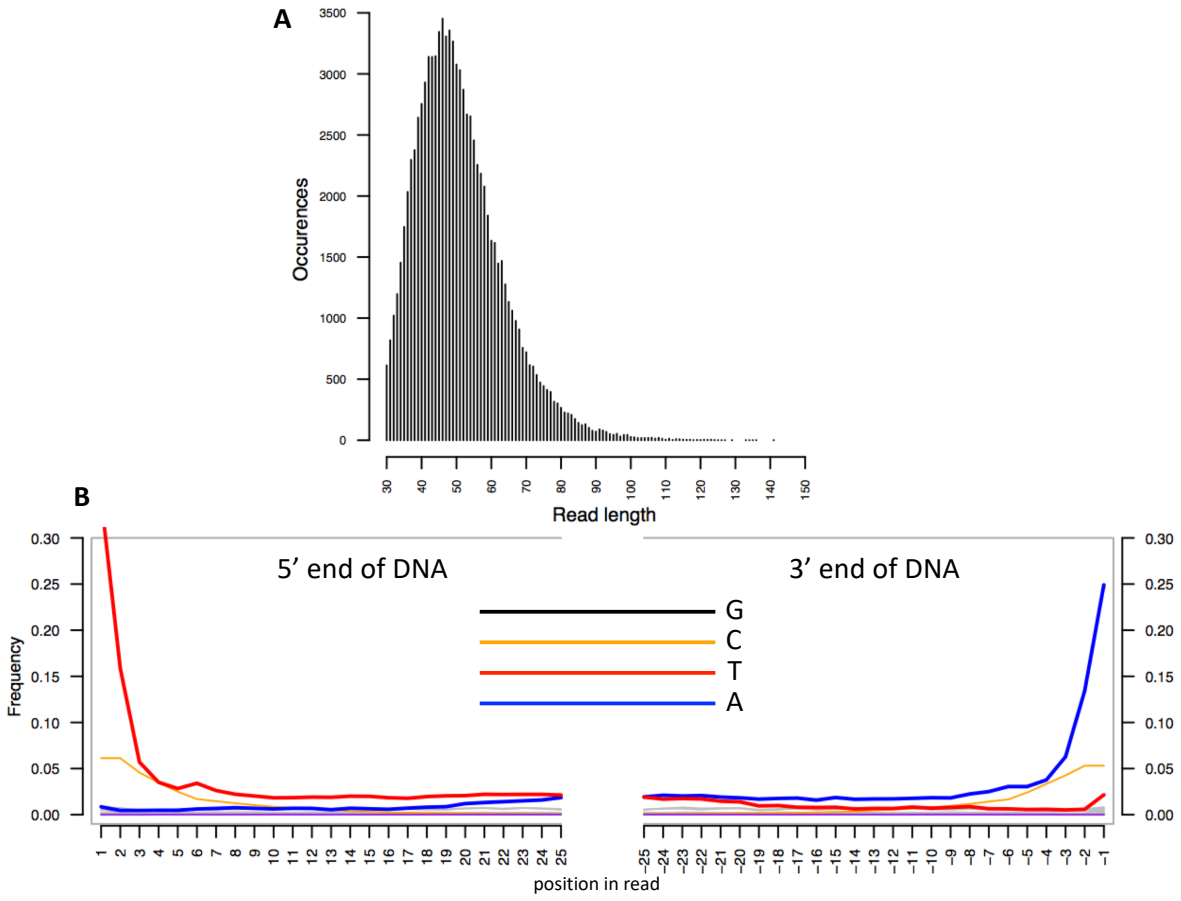


Figure 1. Characteristics of ancient DNA. (A) Ancient DNA typically has a short log-normal fragment length distribution due to post-mortem fragmentation. (B) Cytosine deamination results in the elevated frequencies of observed C-to-T and G-to-A substitutions at the 5' and 3' end of sequenced molecules, respectively. Importantly, these characteristics allow for an assessment of the authenticity of presumed ancient DNA. Data is from a ~3,000-year-old Vanuatu dental calculus sample, and the figure was made using MapDamage 2.0 [25].

Controlling for DNA contamination

The high sensitivity of PCR (polymerase chain reaction) coupled with Next-Generation Sequencing allows for minute quantities of DNA to be analysed. As such, these technologies have ushered in a new age of ancient DNA research. While sensitive, these tools are not specific to the preserved DNA from an ancient sample (endogenous DNA). Modern DNA that is not from the ancient sample (exogenous DNA) is typically less damaged, present in higher abundance, and can be preferentially amplified. Therefore, exogenous DNA can result in erroneous conclusions. For example, modern human DNA introduced by researchers can confound ancient hominin DNA studies [26], and modern microorganisms (which cover most surfaces – including humans) can critically influence palaeomicrobiological studies [27–29]. Typical sources of exogenous DNA that can affect ancient DNA studies include soil microorganisms (if the sample was buried), museum curators, the laboratory environment, laboratory reagents, air-borne microorganisms, and researchers [30,31].

There are multiple strategies to try and reduce the impacts of such exogenous contamination within ancient DNA research [31,32]. First, proper handling of ancient specimens at the point of sampling or museum curation is essential to reduce the crossover of human and microbial DNA from people into samples [31]. This involves the use of adequate physical barriers (*e.g.* latex gloves, long-sleeved clothes, and face-masks) to prevent transfer between person and sample. Once collected, ancient samples should be examined in an ultra-clean, dedicated ancient DNA facility. The facility should be physically isolated from modern molecular biology laboratories, which typically contain high concentrations of DNA from PCR reactions, and DNA amplification and sequencing should be performed in a separate laboratory to prevent the crossover of PCR products into ancient samples [32]. There should be positive air pressure to prevent the introduction of air from outside the facility, coupled with HEPA filtered ventilation. Laboratory users should enter via dedicated entry rooms, where they don sterile, full-body suits, gloves, boots, and face-masks to limit the introduction of modern DNA from their persons. Laboratory surfaces should be regularly cleaned with ≥3% sodium hypochlorite (bleach) and irradiated with ultra-violet (UV) bulbs, both of which act to destroy or limit the amplification of exogenous DNA [33]. Finally, work should be done in still-air hoods to limit cross-contamination between samples.

Second, decontamination of ancient samples by removal of exogenous DNA (both modern and ancient) from the surfaces is required. For example, the exterior of the sample could be physically removed, or the sample could be soaked in bleach to remove exterior DNA

contamination. These methods have been shown to improve the recovery of endogenous ancient DNA [34,3,12], and are valuable techniques in further reducing the burden of exogenous DNA.
Third, proper implementation of negative controls is essential to monitor modern exogenous DNA introduced into ancient samples [31,32,35]. The inclusion of extraction blank controls (empty tubes without sample DNA) in every DNA extraction is required to capture exogenous DNA present in the laboratory environment (researchers, reagents, etc.) [30,36,37]. These controls should be treated like ordinary samples and taken all the way through to sequencing [47–58][28,29].

Finally, comparison of sequences found in ancient samples to those in negative controls and assessment of ancient DNA damage patterns is necessary to minimise the impact of contamination on analyses and conclusions. To assess whether exogenous laboratory contamination is confounding the interpretation of ancient data, the presumed ancient data must be compared to sequenced extraction blank controls. For example, if microbial DNA belonging to a particular species is found in both ancient samples and negative controls, subtractive filtering (*i.e.* removal of that species from further analyses) or other forms of authentication must be employed. An important technique for ancient DNA authentication relies on the characteristics of ancient DNA damage to discern modern from ancient DNA, specifically, the fragmented nature of ancient DNA and the elevated frequency of observed cytosine deamination [22].

Apart from these stringent techniques and strategies to reduce and monitor the introduction of exogenous DNA into ancient samples, the molecular methods used to analyse ancient samples are similar to those employed in modern molecular work, albeit with some optimisations to DNA extraction and library preparation to improve the recovery of short and damaged molecules [38–40]. The end result is DNA that can be sent for sequencing on a high-throughput DNA sequencing machine, such as the Illumina HiSeq or NextSeq, which typically yield billions of DNA sequences. Overall, technological advancements in DNA sequencing coupled with careful control of contamination has expanded the spectrum of organisms able to be studied by researchers in an ancient context. For instance, it is now possible to study ancient microbial communities, especially those associated with ancient humans [3,12]. Such research can expand our understanding of host-microbial evolution, past human health, and even ancient human migrations.

The human microbiota

In 1985 the development of an efficient, culture-free method of investigating microbial communities opened the door for researchers to survey the amazing diversity of microorganisms on Earth [41]. This method was built upon pioneering work by Woese and Fox on the 16S ribosomal RNA gene, which, being an essential subunit of the ribosome is present in all prokaryotes [42]. The combination of conserved regions from which to design universal primers and the presence of hypervariable regions renders the 16S rRNA gene a useful marker for phylogenetic analysis of prokaryotes, allowing researchers to determine the evolutionary relationships between microorganisms and to assign taxonomy to them using databases such as SILVA [43] or Greengenes [44]. This powerful amplicon method of determining "who's there" in a microbial sample, combined with the massive throughput and reduced cost of DNA sequencing has unlocked a burgeoning new field of research into the microbial communities (microbiota) inhabiting diverse environments.

One habitat of great interest is the human body, which is occupied by more microbial than human cells [45]. The human body contains >1,000 microbial species spread across different body sites that each support distinct microbial communities due to differences in physicochemical properties of each site [46,47]. Some of these communities can play important roles in maintaining immune development and homeostasis [48], preventing pathogen invasion and colonisation [49], and assisting in host digestion of food [50]. Importantly, some of these functions can be achieved by groups of different microbes, *i.e.* while individuals may differ in their microbial compositions, functional redundancy can exist between them [51,52]. However, dysregulation of these functions through maladaptation or imbalance is now recognised as an important factor in human disease [53,54]. Indeed, changes in these microbial communities have been associated with myriad human diseases, including obesity [55,56], type I diabetes [57], asthma [58], and depression [59,60].

Oral microbiota

While the bulk of human microbiota research has been on the gut communities [61], the oral cavity is also critically important in understanding human health due to its links with oral health [62]. The oral cavity has several distinct microbial habitats, such as teeth, gingiva (gums), soft tissue (cheeks), tongue, and saliva. These sites are different from each other due to physicochemical factors, such as oxygen concentration, availability of specific surface moieties for microbial adhesion, saliva flow rates, abrasion, and epithelial shedding. These heterogeneous ecological habitats support the growth of distinct microbiota, both at the community [63,47,64] and strain level [65,66].

Dynamics of dental plaque microbiota

Dental plaque is a microbial biofilm that forms on teeth both above and below the gingiva. Host proteins adhere to tooth enamel and contain binding sites for the adhesins of bacteria—especially *Actinomyces* and *Streptococcus*—allowing for primary colonisation of the teeth [67]. This is followed by further coaggregation of other microbial taxa and the formation of complex microbial biofilms [68,69]. Several studies suggest that vertical transmission, either through direct transmission from parents to offspring or shared environment, plays a major role in the acquisition and establishment of these communities [70–76], with related factors such as host genetics [77,78] and early establishment of immunological tolerance [79] also likely playing important roles.

While there is scant literature on the long-term stability of the plaque microbiota [80], studies looking at other oral sites (mainly saliva) suggest a large degree of stability through time [46,73,81]. The reason for this perceived stability is understudied, and little is known about the factors that influence plaque microbial communities. A potential reason for the long-term stability of plaque microbiota through time is the establishment of syntrophic relationships between microorganisms to harness nutrients provided by the host. Saliva is constantly being produced by the host and contains a rich mixture of glycoproteins and enzymes. Mucins, which are the major host salivary protein secreted, are rich in sugars that require complex metabolic partnerships between microbes to process [82,83]. Recent findings support this idea that symbiotic partnerships are an important factor for plaque microbiota [69]. Welch *et al*. found that the 13 most abundant and prevalent genera (present in at least 90% of samples) identified in plaque samples accounted for 85% of their sequencing data. These authors used FISH (Fluorescent In-Situ Hybridisation) to fluorescently label these 'core' plaque genera, allowing for spatial visualisation of these communities. Strikingly, they observed co-localisation of producers and consumers of metabolites, and functional niche separation (*e.g.* anaerobes in the centre of the structure, aerobes towards the extremities) [69]. Ultimately, plaque microbial communities appear not be a random assortment of environmental taxa, but a multispecies consortium characterised by a high degree of communication, partnership, and interdependence [84] (Figure 2). The fact that teeth are the only non-shedding microbiota-associated body site allows more time for co-evolution between plaque microbiota and may aid in the development of these highly cooperative partnerships. This also extends to the development of relationships between plaque microbiota and the host, which should favour communities positive to the host's health [85]. Indeed, it is being increasingly recognised that disruptions in the relationship between host and plaque microbiota can result in dental diseases [86–89].

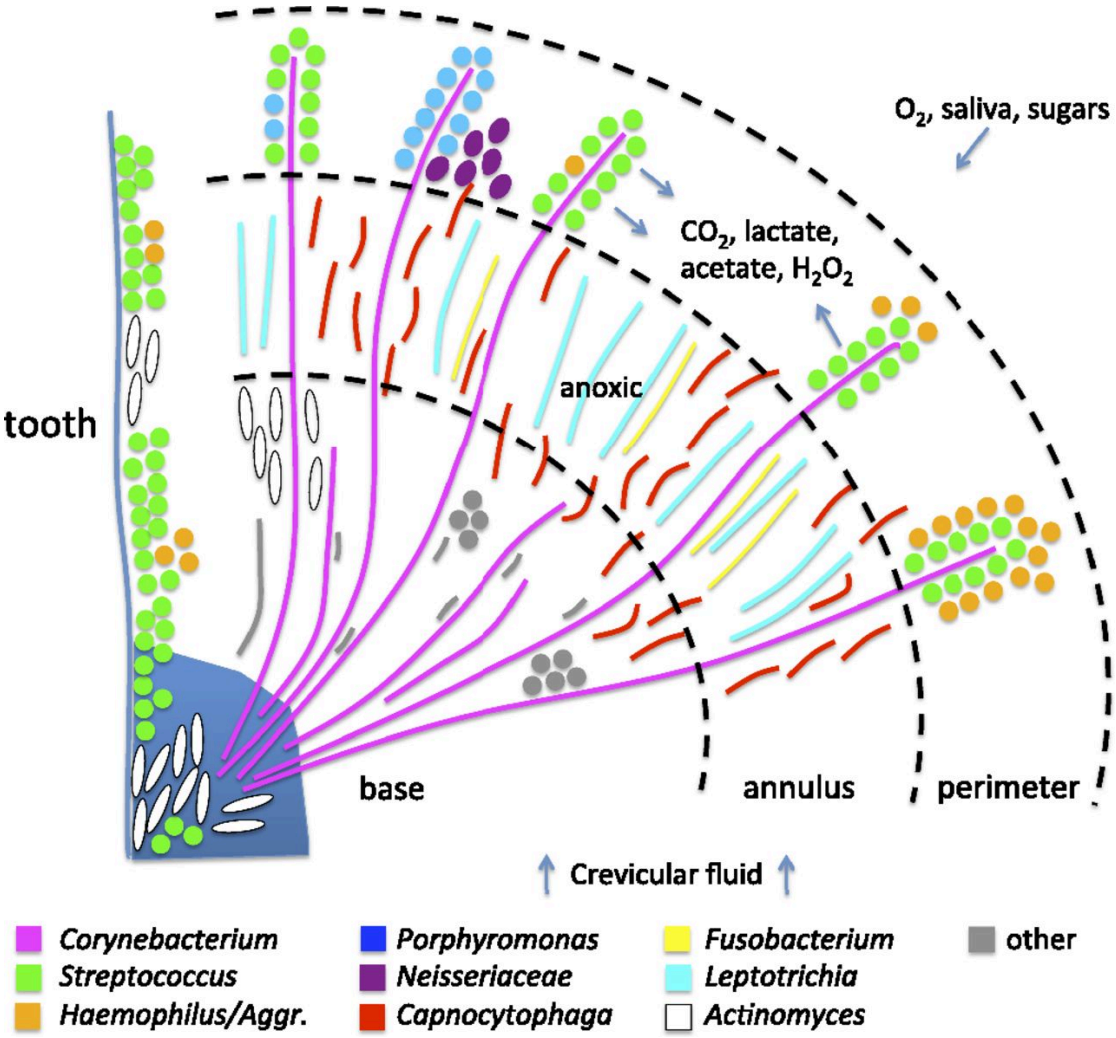


Figure 2. Hypothesised relationships between bacteria identified in complex plaque structures from the 2016 Welch *et al.* study. Figure reproduced from [69].

Dental plaque microbiota in relation to disease

According to a 2004-2006 nationwide dental health survey, the prevalence of dental caries (tooth decay) and periodontal disease in Australians over 15 years of age was 12.8% and 22.9%, respectively [90], with Indigenous populations suffering disproportionately [91]. The total expenditure on dental services in Australia for 2012-2013 was $8.7 billion [92]; however, this figure does not account for indirect economic and societal costs such as losses in productivity and the impacts of oral diseases on systemic health [93]. Dental caries is the destruction of teeth by acids that result from the fermentation of dietary carbohydrates by bacteria [94]. *Streptococcus mutans* is a bacterium that has long been touted as the causative agent of dental caries [95]; however, recent high-throughput molecular studies have challenged the simplicity of this notion, and instead, support a more complex, polymicrobial aetiology for dental caries [96]. This shift in thinking from a handful of pathogenic microorganisms to a complex

polymicrobial involvement has also been seen in periodontal disease. Periodontal disease refers to the inflammation of tissues surrounding the tooth (*i.e.* gums), which can result in the formation of pockets or gaps between the tooth and its surrounding gingiva (gum). Severe forms of periodontal disease can result in loss of bone supporting the tooth, resulting in the loosening or even loss of the tooth. The association between plaque microbiota and periodontal disease has long been appreciated [97], with research in the late 80s and early 90s attributing it three specific bacteria — the "red complex" (*Tannerella forsythia, Treponema denticola, and Porphyromonas gingivalis*) [98]. As with dental caries, the aetiology of periodontal disease is increasingly recognised as more complex [99,100], with ecological interactions within plaque microbiota [101,86] and breakdown in the relationship between the microbiota and host (*e.g.* immune intolerance) being key contributing factors [89,102,103]. The ability to study how these intricate relationships have evolved with humans through history could offer important insights to aid in the development of new treatments for these polymicrobial diseases.

Studying ancient human microbiota: prospects and pitfalls

The brief yet controversial history of palaeomicrobiology

Palaeomicrobiology—the study of ancient microorganisms—is a rapidly growing area of research that has the potential to enhance our understanding of microbial evolution [104], host-microbiota interactions [12], past human interactions and movements [105], and the emergence and evolution of human pathogens [106,107]. However, in the short history of the field, these prospects have been tarnished by numerous claims [108–119] that have been questioned [120–128,27–29] on the basis of insufficient evidence, controls, and authenticity. As with other ancient DNA research, NGS has expanded the scope of palaeomicrobiological research by allowing researchers to rapidly profile microbial communities and their functions and to reconstruct whole genomes [12,104]. NGS has also enhanced researchers' ability to assess the authenticity of their own findings [25,35], as well as contentious claims made by others [28,29]. Another catalyst critical to the recent advancement of palaeomicrobiology was the discovery that ancient dental calculus (calcified dental plaque) is a robust reservoir of ancient microbial DNA [129].

Dental calculus: ushering in a new age of palaeomicrobiology

Dental plaque undergoes periodic mineralisation events which incrementally trap the resident microbiota in a hard calcium phosphate matrix called dental calculus [130] (Figure 3). This occurs throughout the lifetime of an individual. Coupled with the absence of modern dentistry practices in past populations, this lifelong process allows for recovery of calculus on human skulls hundreds or even thousands of years old [104]. The mineralisation of plaque to calculus also acts to protect the endogenous microbial DNA from external contamination, rendering calculus an ideal substrate for palaeomicrobiological research. Dobney and Brothwell were the first to observe microorganisms trapped in archaeological dental calculus over 30 years ago [131,132], but it was not known whether ancient microbial DNA was preserved within calculus until a 2011 study by Preus *et al.* [129]. This was quickly followed by the first community-level analysis of microorganisms within dental calculus [3], leveraging recent advances in NGS and 16S rRNA amplicon techniques [133]. A subsequent study by Warinner *et al.* was the first to characterise the protein functions associated within dental calculus and to reconstruct the genome of a putative periodontal pathogen [12]. In the most recently published study of ancient dental calculus by Weyrich *et al.,* the authors were able to reconstruct a ~49,000-year-old oral archaeal genome (*Methanobrevibacter oralis*) from a Neanderthal dental calculus sample [104].

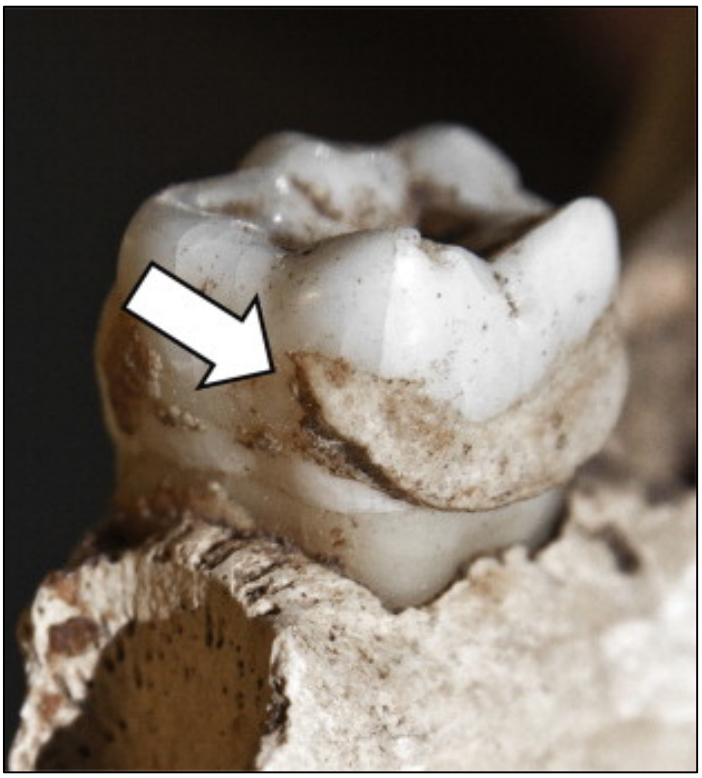


Figure 3. Example of ancient dental calculus on a molar from a Medieval specimen, York, U.K. Figure reproduced from [134]

Collectively, these landmark studies have demonstrated the feasibility of obtaining high-resolution human microbiota data from ancient dental calculus and have ushered in a new age of palaeomicrobiology. It is now possible to investigate how oral microbiota have changed through human history, both at the community and the individual genome level [3,12,104]. Such research is poised to shed light upon host-microbiota co-evolution and how this relationship influences dental diseases. Ancient dental calculus research could also provide valuable bioarchaeological information about past human cultures. If plaque communities are labile to host dietary habits, such changes could be used to infer differences in diet both within and between cultures. Recent research suggests that dental calculus can trap dietary items, potentially providing new insights into the behaviour and diet of past cultures [104]. Additionally, given the putative stability of plaque microbiota, past human migration, movements, and interactions between cultures could be measured using genomic information from microorganisms within ancient dental calculus, following an approach similar to that which has already been used with modern microorganisms [135,136] (this concept is proposed and discussed in Chapter I). Underpinning and crucial to realising these prospects of ancient dental calculus research are the methods used to analyse ancient microbial DNA.

Analysing ancient microbial DNA

While technological and methodological advances now allow for the recovery and authentication of vast amounts of ancient microbial DNA from ancient dental calculus, the bioinformatic tools used to analyse such data are still in their infancy. Initial attempts at classifying microbial communities in ancient dental calculus [3] used the 16S rRNA gene amplicon technique [133]. However, Ziesemer *et al.* later found that this method is inappropriate when applied to ancient DNA [137]. Specifically, the amplicon sizes targeted by 16S primers are typically larger than the short fragment sizes of ancient DNA; hence, longer contaminant DNA can be preferentially amplified. Critically, it was found that differences between microbial taxa in the length of the regions amplified lead to differential amplification success for different taxa, and therefore to biased representation of communities [137]. Since the publication of these findings, shotgun metagenomic sequencing—the nonspecific recovery of all DNA in a sample—has been established as the standard for classifying ancient microbial communities [12,35,104,137].

The shift from a single amplicon to nonspecific, shotgun metagenomic information containing millions or billions of base pairs is a major computational and analytical challenge for the field. Currently, the two main strategies for analysing shotgun metagenomic are *de novo* assembly

and reference-based alignment. The goal of *de novo* assembly is to find overlaps in DNA fragments and merge the fragments to build larger sequences (contigs), and potentially whole genomes. Tools commonly used for this approach in modern metagenomic studies include MetaVelvet [138] and GroopM [139]. An advantage of *de novo* assembly is that it allows for the discovery of new genomic information or microorganisms that are not currently in reference databases [140]. However, the extremely short length of ancient DNA renders *de novo* assembly unfeasible for palaeomicrobiology, as ultra-short reads are computationally difficult to merge into longer contigs. This is illustrated in Warinner *et al.* [12] whereby the mean and maximum contig length for two deeply sequenced ancient specimens were short, being ~200 and 12,000 base pairs, respectively.

Reference-based alignment works by matching (aligning) DNA fragments within a sample to reference sequences or genomes in a database. Commonly used tools for this approach include BLAST (Basic Local Alignment Search Tool) [141], Bowtie2 [142], and BWA (Burrows-Wheeler Aligner) [143]. Reference-based alignment can reliably align reads ~30 bp in length making it applicable to palaeomicrobiological data (I demonstrate this in Chapter III). However, the constant growth of reference databases, coupled with larger DNA sequencing outputs, substantially expands the number of possible alignments between sequences, which drastically increases computation time and the resources required for such analysis. This is especially an issue in metagenomic research where researchers are interested in surveying the large diversity of microorganisms, leading to databases with tens of thousands of reference genomes. To combat this issue, new algorithms, such as MALT [144], MetaPhlAn [145], and KRAKEN [146], have been developed to reduce the computational burden. Of these tools, MALT has been shown to offer great performance for classifying microbial communities in palaeomicrobiological studies [104,147]. However, a key issue for the reference-based alignment strategy is the relatively small breadth of microbial diversity captured in whole-genome reference databases, which is yet to catch up with larger 16S databases [148,149]. The impact of this problem on the reconstruction of ancient microbial communities is yet to be investigated, although it is likely substantial due to microbial 'extinctions' in past human microbiota and the focus of modern microbial genome reconstructions on healthy individuals of 'Western' ancestry [66].

Overall, technological and methodological advances have improved our ability to generate ancient microbial DNA, but further improvements and assessments of analytical techniques are essential to expand both the scope and quality of palaeomicrobiological research. Such advances will allow us to realise novel prospects of palaeomicrobiology and help in avoiding its pitfalls.

Thesis overview:

This thesis contains six chapters and three appendix chapters that build upon ideas and address issues identified in this introductory section. The overarching theme of this thesis is the development and critical assessment of analytical techniques for analysing ancient microbial DNA. This thesis also calls into question recent palaeomicrobiological studies that lacked appropriate experimental controls and made claims unsubstantiated by evidence. Ultimately, this thesis seeks to expand the scope of palaeomicrobiology and improve the quality and reproducibility of future research.

Chapter I: *Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events*

In this first chapter, I propose the use of microbial DNA in ancient dental calculus as a proxy for past human movements and explore the advantages of such an approach to enhance our understanding of past human demographic histories. I propose that the best location to test this idea is in the Pacific Islands, so I review the current archaeological, linguistic, and genetic evidence for the peopling of the Pacific and highlight the difficulties of determining past, rapid human settlements in this area using currently available tools.

Chapter II: *Contamination in low-biomass microbiome studies: issues and recommendations*

This chapter, while not focusing on palaeomicrobiology, seeks to share practices and authentication criteria used within the field of ancient DNA with modern low-biomass microbiota research. Low-biomass microbiota research suffers from similar pitfalls to palaeomicrobiology, including low concentrations of endogenous DNA, DNA contamination, and the presence of controversial studies lacking appropriate controls. In this chapter, I review the state of low-biomass microbiota research, highlighting the issues of reproducibility and the lack of an agreed-upon set of standards unifying the field. With the help of experts in modern microbiota research, I develop and propose a set of authentication criteria for low-biomass microbiota studies to help researchers, reviewers, and editors improve the quality and reproducibility of future research.

Chapter III: *Assessing alignment-based taxonomic classification of ancient microbial DNA*

While palaeomicrobiologists have converged on reference-based alignment for the reconstruction of ancient microbial communities, there has not yet been a thorough examination of how the characteristics of ancient DNA impact such analysis. In this chapter, I use simulated and real data to perform an in-depth assessment of how DNA fragment length, deamination,

divergence, and missing reference sequences influence alignment-based methods. By constructing and using the largest and most diverse reference database to date for reference-alignment, I also examine the extent of missing microbial diversity when reconstructing the taxonomic composition of ancient microbiota, and perform a reanalysis of a previously published study. Finally, I use these findings to provide clear recommendations for other researchers that aim to use alignment-based methods in future palaeomicrobiological research.

Chapter IV: *Development and validation of a complementary approach to reconstruct ancient microbial communities*

While the use of 16S rRNA gene fragments from metagenomes has been previously used to reconstruct ancient microbial communities, there is yet to be a robust assessment this approach when applied to ancient DNA. In Chapter IV, I develop and assess a new technique for reconstructing ancient microbial communities using hybridisation enrichment of 16S rRNA gene fragments from metagenomes. Using both simulated and real data, I also assess the influence of the characteristics of ancient DNA on the quality of taxonomic assignment of 16S rRNA gene fragments.

Chapter V: *Palaeomicrobiology of the Pacific: unlocking a high-resolution proxy for past human movements*

In this chapter, I test if oral microbiota can be used to examine ancient human migrations by generating and authenticating high-quality data for 117 ancient Asia-Pacific dental calculus samples. I compare and contrast this data to modern plaque samples, explore the microbial community structure between islands, and develop and test two different approaches for using dental calculus as a proxy for past human migration. Building upon the concepts highlighted in Chapter I of this thesis, this is the first study to demonstrate that microbial DNA in ancient dental calculus can act as a high-resolution proxy for past human movement.

Chapter VI: *Insights into the demographic history of Japan using ancient oral microbiota*

How cultural and population admixtures influence human microbiota remain to be understood. Ancient DNA allows for the measurement of microbiota states prior to and post major demographic changes, such as what happened in ancient Japan with the admixture of Jomon hunter-gatherers with agriculturalists from mainland Asia. In this chapter, I compare and contrast the microbiota of Jomon hunter-gatherers with later Edo period agriculturalists, exploring how diet, disease, and culture can alter or confound ancient human microbiota

studies. Using genomic information, I also investigate whether bacterial lineages can be lost due to such changes.

Appendices I-III:

Isolating viable ancient bacteria: what you put in is what you get out

Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential

Proper Authentication of Ancient DNA Is Still Essential

The field of ancient DNA has been fraught with controversial studies lacking the appropriate procedures and controls. While recent technological and methodological improvements have made it easier to authenticate ancient DNA, there are still studies being published that lack sufficient scientific rigour. In these three replies to such studies, I call into question to the validity of the claims made by these authors, and alert other researchers, reviewers, and editors to the various pitfalls of palaeomicrobiological research.

References

1. Higuchi R, Bowman B, Freiberger M, Ryder OA, Wilson AC. DNA sequences from the quagga, an extinct member of the horse family. Nature. 1984;312:282–4.

2. Rasmussen M, Li Y, Lindgreen S, Pedersen JS, Albrechtsen A, Moltke I, et al. Ancient human genome sequence of an extinct Palaeo-Eskimo. Nature. 2010;463:757–62.

3. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45:450–5.

4. Haile J, Froese DG, MacPhee RDE, Roberts RG, Arnold LJ, Reyes AV, et al. Ancient DNA reveals late survival of mammoth and horse in interior Alaska. Proc Natl Acad Sci. 2009;106:22352–7.

5. Thalmann O, Shapiro B, Cui P, Schuenemann VJ, Sawyer SK, Greenfield DL, et al. Complete Mitochondrial Genomes of Ancient Canids Suggest a European Origin of Domestic Dogs. Science. 2013;342:871–4.

6. Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, et al. Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. Nature. 2013;499:74–8.

7. Soubrier J, Gower G, Chen K, Richards SM, Llamas B, Mitchell KJ, et al. Early cave art and ancient DNA record the origin of European bison. Nat Commun. 2016;7:13158.

8. Matisoo-Smith E, Robins JH. Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat. Proc Natl Acad Sci U S A. 2004;101:9167–72.
9. Thomson VA, Lebrasseur O, Austin JJ, Hunt TL, Burney DA, Denham T, et al. Using ancient DNA to study the origins and dispersal of ancestral Polynesian chickens across the Pacific. Proc Natl Acad Sci. 2014;111:4826–4831.
10. Feigin CY, Newton AH, Doronina L, Schmitz J, Hipsley CA, Mitchell KJ, et al. Genome of the Tasmanian tiger provides insights into the evolution and demography of an extinct marsupial carnivore. Nat Ecol Evol. 2018;2:182–92.
11. Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A Draft Sequence of the Neandertal Genome. Science. 2010;328:710–22.
12. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336–44.
13. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc R Soc Lond B Biol Sci. 2012;279:4724–33.
14. Pääbo S. Ancient DNA: extraction, characterization, molecular cloning, and enzymatic amplification. Proc Natl Acad Sci U S A. 1989;86:1939–43.
15. Lindahl T, Andersson A. Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. Biochemistry (Mosc). 1972;11:3618–23.
16. Lindahl T, Nyberg B. Rate of depurination of native deoxyribonucleic acid. Biochemistry (Mosc). 1972;11:3610–8.
17. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci. 2007;104:14616–21.
18. Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, et al. A high-coverage genome sequence from an archaic Denisovan individual. Science. 2012;338:222–6.
19. Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. Cold Spring Harb Perspect Biol. 2013;a012567.
20. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S. DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res. 2001;29:4793–9.
21. Lindahl T. Instability and decay of the primary structure of DNA. Nature. 1993;362:709–15.
22. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics. 2011;27:2153–5.

23. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLOS ONE. 2012;7:e34131.

24. Kistler L, Ware R, Smith O, Collins M, Allaby RG. A new model for ancient DNA decay based on paleogenomic meta-analysis. Nucleic Acids Res. 2017;45:6310–20.

25. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinforma Oxf Engl. 2013;29:1682–4.

26. Wall JD, Kim SK. Inconsistencies in Neanderthal genomic DNA sequences. PLoS Genet. 2007;3:1862–6.

27. Eisenhofer R, Cooper A, Weyrich LS. Isolating Viable Ancient Bacteria: What You Put In Is What You Get Out. Genome Announc [Internet]. 2016 [cited 2017 Feb 8];4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5000818/

28. Eisenhofer R, Cooper A, Weyrich LS. Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential. FEMS Microbiol Ecol [Internet]. 2017 [cited 2017 Jun 27];93. Available from: https://academic.oup.com/femsec/article/93/5/fix042/3089752/Reply-to-Santiago-Rodriguez-et-al-proper

29. Eisenhofer R, Weyrich LS. Proper Authentication of Ancient DNA Is Still Essential. Genes. 2018;9:122.

30. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.

31. Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W. From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR Sci Technol Archaeol Res. 2017;3:1–14.

32. Cooper A, Poinar HN. Ancient DNA: Do It Right or Not at All. Science. 2000;289:1139–1139.

33. Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S, et al. Decontamination of MDA reagents for single cell whole genome amplification. PloS One. 2011;6:e26161.

34. Kemp BM, Smith DG. Use of bleach to eliminate contaminating DNA from the surface of bones and teeth. Forensic Sci Int. 2005;154:53–61.

35. Warinner C, Herbig A, Mann A, Yates JAF, Weiß CL, Burbano HA, et al. A Robust Framework for Microbial Archaeology. Annu Rev Genomics Hum Genet. 2017;18:null.

36. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8:24.
37. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome. 2016;4:29.
38. Rohland N, Hofreiter M. Ancient DNA extraction from bones and teeth. Nat Protoc. 2007;2:1756–62.
39. Glocke I, Meyer M. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Res. 2017;27:1230–7.
40. Gansauge M-T, Meyer M. Single-stranded DNA library preparation for the sequencing of ancient or damaged DNA. Nat Protoc. 2013;8:737–48.
41. Lane DJ, Pace B, Olsen GJ, Stahl DA, Sogin ML, Pace NR. Rapid determination of 16S ribosomal RNA sequences for phylogenetic analyses. Proc Natl Acad Sci. 1985;82:6955–9.
42. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci. 1977;74:5088–90.
43. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
44. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a Chimera-Checked 16S rRNA Gene Database and Workbench Compatible with ARB. Appl Environ Microbiol. 2006;72:5069–72.
45. Sender R, Fuchs S, Milo R. Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans. Cell. 2016;164:337–40.
46. Costello EK, Lauber CL, Hamady M, Fierer N, Gordon JI, Knight R. Bacterial Community Variation in Human Body Habitats Across Space and Time. Science. 2009;326:1694–7.
47. Consortium THMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
48. Belkaid Y, Hand TW. Role of the Microbiota in Immunity and Inflammation. Cell. 2014;157:121–41.
49. Buffie CG, Pamer EG. Microbiota-mediated colonization resistance against intestinal pathogens. Nat Rev Immunol. 2013;13:790–801.
50. Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I, et al. Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr. 2018;57:1–24.

51. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R. Diversity, stability and resilience of the human gut microbiota. Nature. 2012;489:220–30.
52. Jorth P, Turner KH, Gumus P, Nizam N, Buduneli N, Whiteley M. Metatranscriptomics of the Human Oral Microbiome during Health and Disease. mBio. 2014;5:e01012-14.
53. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13:260–70.
54. Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ. Dysbiosis of the gut microbiota in disease. Microb Ecol Health Dis [Internet]. 2015 [cited 2018 Mar 18];26. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315779/
55. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006;444:1027–131.
56. Ravussin Y, Koren O, Spor A, LeDuc C, Gutman R, Stombaugh J, et al. Responses of Gut Microbiota to Diet Composition and Weight Loss in Lean and Obese Mice. Obesity. 2012;20:738–47.
57. Brown CT, Davis-Richardson AG, Giongo A, Gano KA, Crabb DB, Mukherjee N, et al. Gut Microbiome Metagenomics Analysis Suggests a Functional Model for the Development of Autoimmunity for Type 1 Diabetes. PLOS ONE. 2011;6:e25792.
58. Chen Y, Blaser MJ. Inverse Associations of Helicobacter pylori With Asthma and Allergy. Arch Intern Med. 2007;167:821–7.
59. Yano JM, Yu K, Donaldson GP, Shastri GG, Ann P, Ma L, et al. Indigenous Bacteria from the Gut Microbiota Regulate Host Serotonin Biosynthesis. Cell. 2015;161:264–76.
60. Sampson TR, Mazmanian SK. Control of Brain Development, Function, and Behavior by the Microbiome. Cell Host Microbe. 2015;17:565–76.
61. Lloyd-Price J, Abu-Ali G, Huttenhower C. The healthy human microbiome. Genome Med. 2016;8:51.
62. Wade WG. The oral microbiome in health and disease. Pharmacol Res. 2013;69:137–43.
63. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W-H, et al. The Human Oral Microbiome. J Bacteriol. 2010;192:5002–17.
64. Xu X, He J, Xue J, Wang Y, Li K, Zhang K, et al. Oral cavity contains distinct niches with dynamic microbial communities. Environ Microbiol. 2015;17:699–710.
65. Donati C, Zolfo M, Albanese D, Truong DT, Asnicar F, Iebba V, et al. Uncovering oral Neisseria tropism and persistence using metagenomic sequencing. Nat Microbiol. 2016;1:16070.

66. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.
67. Kolenbrander PE, London J. Adhere today, here tomorrow: oral bacterial adherence. J Bacteriol. 1993;175:3247–52.
68. Kolenbrander Paul E., Palmer Robert J., Rickard Alexander H., Jakubovics Nicholas S., Chalmers Natalia I., Diaz Patricia I. Bacterial interactions and successions during plaque development. Periodontol 2000. 2006;42:47–79.
69. Welch JLM, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci. 2016;113:E791–800.
70. Okada M, Hayashi F, Soda Y, Zhong X, Miura K, Kozai K. Intra-familial distribution of nine putative periodontopathogens in dental plaque samples analyzed by PCR. J Oral Sci. 2004;46:149–56.
71. Li Y, Ismail AI, Ge Y, Tellez M, Sohn W. Similarity of Bacterial Populations in Saliva from African-American Mother-Child Dyads. J Clin Microbiol. 2007;45:3082–5.
72. Corby PM, Bretz WA, Hart TC, Schork NJ, Wessel J, Lyons-Weiler J, et al. Heritability of Oral Microbial Species in Caries-Active and Caries-Free Twins. Twin Res Hum Genet. 2007;10:821–828.
73. Stahringer SS, Clemente JC, Corley RP, Hewitt J, Knights D, Walters WA, et al. Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. Genome Res. 2012;22:2146–52.
74. Ebersole JL, Holt SC, Delaney JE. Acquisition of Oral Microbes and Associated Systemic Responses of Newborn Nonhuman Primates. Clin Vaccine Immunol CVI. 2014;21:21–8.
75. Shaw L, Ribeiro ALR, Levine AP, Pontikos N, Balloux F, Segal AW, et al. The Human Salivary Microbiome Is Shaped by Shared Environment Rather than Genetics: Evidence from a Large Family of Closely Related Individuals. mBio. 2017;8:e01237-17.
76. Mason MR, Chambers S, Dabdoub SM, Thikkurissy S, Kumar PS. Characterizing oral microbial communities across dentition states and colonization niches. Microbiome. 2018;6:67.
77. Gomez A, Espinoza JL, Harkins DM, Leong P, Saffery R, Bockmann M, et al. Host Genetic Control of the Oral Microbiome in Health and Disease. Cell Host Microbe. 2017;22:269-278.e3.
78. Demmitt BA, Corley RP, Huibregtse BM, Keller MC, Hewitt JK, McQueen MB, et al. Genetic influences on the human oral microbiome. BMC Genomics. 2017;18:659.
79. Chistiakov DA, Bobryshev YV, Kozarov E, Sobenin IA, Orekhov AN. Intestinal mucosal tolerance and impact of gut microbiota to mucosal tolerance. Front Microbiol [Internet]. 2015

[cited 2018 Mar 25];5. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4292724/

80. Utter DR, Mark Welch JL, Borisy GG. Individuality, Stability, and Variability of the Plaque Microbiome. Front Microbiol [Internet]. 2016 [cited 2016 Jul 8];7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4840391/

81. Cameron SJS, Huws S, Hegarty MJ, Smith DPM, Mur LAJ. The human salivary microbiome exhibits temporal stability in bacterial diversity. FEMS Microbiol Ecol. 2015;fiv091.

82. Bradshaw DJ, Homer KA, Marsh PD, Beighton D. Metabolic cooperation in oral microbial communities during growth on mucin. Microbiol Read Engl. 1994;140 (Pt 12):3407–12.

83. Wickström C, Herzberg MC, Beighton D, Svensäter G. Proteolytic degradation of human salivary MUC5B by dental biofilms. Microbiology. 2009;155:2866–72.

84. Kolenbrander PE, Jr RJP, Periasamy S, Jakubovics NS. Oral multispecies biofilm development and the key role of cell–cell distance. Nat Rev Microbiol. 2010;8:471.

85. B.T. Rosier, P.D. Marsh, A. Mira. Resilience of the Oral Microbiota in Health: Mechanisms That Prevent Dysbiosis. J Dent Res. 2017;0022034517742139.

86. Marsh PD. Are dental diseases examples of ecological catastrophes? Microbiology. 2003;149:279–94.

87. Socransky SS, Haffajee AD. Periodontal microbial ecology. Periodontol 2000. 2005;38:135–87.

88. Marsh PD, Zaura E. Dental biofilm: ecological interactions in health and disease. J Clin Periodontol. 2017;44:S12–22.

89. Mark Bartold P, Van Dyke TE. Host modulation: controlling the inflammation to control the infection. Periodontol 2000. 2017;75:317–29.

90. Slade GD, Roberts-Thomson KF, Ellershaw AE. Australia's dental generations: the National Survey of Adult Oral Health 2004–06. Dental statistics and research series no. 34. Cat. no. DEN 165. Canberra: AIHW. 2007.

91. Tiwari T, Jamieson L, Broughton J, Lawrence HP, Batliner TS, Arantes R, et al. Reducing Indigenous Oral Health Inequalities: A Review from 5 Nations. J Dent Res. 2018;22034518763605.

92. Australian Institute of Health and Welfare. Oral health and dental care in Australia: key facts and figures 2015. 2015.

93. Listl S, Galloway J, Mossey PA, Marcenes W. Global Economic Impact of Dental Diseases. J Dent Res. 2015;94:1355–61.

94. Selwitz RH, Ismail AI, Pitts NB. Dental caries. The Lancet. 2007;369:51–9.

95. Loesche WJ, Rowan J, Straffon LH, Loos PJ. Association of Streptococcus mutans with human dental decay. Infect Immun. 1975;11:1252–60.
96. Simón-Soro A, Mira A. Solving the etiology of dental caries. Trends Microbiol. 2015;23:76–82.
97. Meyer KF. The Present Status of Dental Bacteriology. J Natl Dent Assoc. 1917;4:966–96.
98. Socransky S s., Haffajee A d., Cugini M a., Smith C, Kent RL. Microbial complexes in subgingival plaque. J Clin Periodontol. 1998;25:134–44.
99. Hajishengallis G, Lamont RJ. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. Mol Oral Microbiol. 2012;27:409–19.
100. Rosier BT, De Jager M, Zaura E, Krom BP. Historical and contemporary hypotheses on the development of oral diseases: are we there yet? Front Cell Infect Microbiol [Internet]. 2014 [cited 2015 May 22];4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4100321/
101. Marsh PD. Microbial ecology of dental plaque and its significance in health and disease. Adv Dent Res. 1994;8:263–71.
102. Darveau RP. Periodontitis: a polymicrobial disruption of host homeostasis. Nat Rev Microbiol. 2010;8:481.
103. Meyle J, Chapple I. Molecular aspects of the pathogenesis of periodontitis. Periodontol 2000. 2015;69:7–17.
104. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature. 2017;544:357–61.
105. Eisenhofer R, Anderson A, Dobney K, Cooper A, Weyrich LS. Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events. J Isl Coast Archaeol. 2017;0:1–14.
106. McNally A, Thomson NR, Reuter S, Wren BW. “Add, stir and reduce”: Yersinia spp. as model bacteria for pathogen evolution. Nat Rev Microbiol. 2016;14:177–90.
107. Achtman M. How old are bacterial pathogens? Proc R Soc B. 2016;283:20160990.
108. Cano RJ, Borucki MK. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science. 1995;268:1060–4.
109. Drancourt M, Aboudharam G, Signoli M, Dutour O, Raoult D. Detection of 400-year-old Yersinia pestis DNA in human dental pulp: An approach to the diagnosis of ancient septicemia. Proc Natl Acad Sci. 1998;95:12637–40.

110. Vreeland RH, Rosenzweig WD, Powers DW. Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. Nature. 2000;407:897–900.
111. Raoult D, Aboudharam G, Crubézy E, Larrouy G, Ludes B, Drancourt M. Molecular identification by "suicide PCR" of Yersinia pestis as the agent of Medieval Black Death. Proc Natl Acad Sci. 2000;97:12800–3.
112. Raoult D, Drancourt M. Cause of Black Death. Lancet Infect Dis. 2002;2:459.
113. Drancourt M, Raoult D. Molecular detection of Yersinia pestis in dental pulp. Microbiology. 2004;150:263–4.
114. Drancourt M, Roux V, Dang LV, Tran-Hung L, Castex D, Chenal-Francisque V, et al. Genotyping, Orientalis-like Yersinia pestis, and Plague Pandemics. Emerg Infect Dis. 2004;10:1585–92.
115. Papagrigorakis MJ, Yapijakis C, Synodinos PN, Baziotopoulou-Valavani E. DNA examination of ancient dental pulp incriminates typhoid fever as a probable cause of the Plague of Athens. Int J Infect Dis. 2006;10:206–14.
116. Santiago-Rodriguez TM, Patrício AR, Rivera JI, Coradin M, Gonzalez A, Tirado G, et al. luxS in bacteria isolated from 25- to 40-million-year-old amber. FEMS Microbiol Lett. 2014;350:117–24.
117. Goncharov A, Grigorjev S, Karaseva A, Kolodzhieva V, Azarov D, Akhremenko Y, et al. Draft Genome Sequence of Enterococcus faecium Strain 58m, Isolated from Intestinal Tract Content of a Woolly Mammoth, Mammuthus primigenius. Genome Announc [Internet]. 2016 [cited 2018 Apr 3];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751320/
118. Santiago-Rodriguez TM, Fornaciari G, Luciani S, Dowd SE, Toranzos GA, Marota I, et al. Taxonomic and predicted metabolic profiles of the human gut microbiome in pre-Columbian mummies. FEMS Microbiol Ecol. 2016;92.
119. Santiago-Rodriguez TM, Fornaciari G, Luciani S, Toranzos GA, Marota I, Giuffra V, et al. Gut Microbiome and Putative Resistome of Inca and Italian Nobility Mummies. Genes. 2017;8:310.
120. Austin JJ, Ross AJ, Smith AB, Fortey RA, Thomas RH. Problems of reproducibility – does geologically ancient DNA survive in amber–preserved insects? Proc R Soc Lond B Biol Sci. 1997;264:467–74.
121. Gutiérrez G, Marín A. The most ancient DNA recovered from an amber-preserved specimen may not be as ancient as it seems. Mol Biol Evol. 1998;15:926–9.
122. Hazen RM, Roedder E. Biogeology. How old are bacteria from the Permian age? Nature. 2001;411:155–6.

123. Nickle DC, Learn GH, Rain MW, Mullins JI, Mittler JE. Curiously Modern DNA for a ``250 Million-Year-Old'' Bacterium. J Mol Evol. 2002;54:134–7.
124. Gilbert MTP, Cuccui J, White W, Lynnerup N, Titball RW, Cooper A, et al. Absence of Yersinia pestis-specific DNA in human teeth from five European excavations of putative plague victims. Microbiology. 2004;150:341–54.
125. Gilbert MTP, Cuccui J, White W, Lynnerup N, Titball RW, Cooper A, et al. Response to Drancourt and Raoult. Microbiology. 2004;150:264–5.
126. Vergnaud G. Yersinia pestis Genotyping. Emerg Infect Dis. 2005;11:1317–9.
127. Shapiro B, Rambaut A, Gilbert MTP. No proof that typhoid caused the Plague of Athens (a reply to Papagrigorakis et al.). Int J Infect Dis. 2006;10:334–5.
128. Weyrich LS, Llamas B, Cooper A. Reply to Santiago-Rodriguez et al.: Was luxS really isolated from 25- to 40-million-year-old bacteria? FEMS Microbiol Lett. 2014;353:85–6.
129. Preus HR, Marvik OJ, Selvig KA, Bennike P. Ancient bacterial DNA (aDNA) in dental calculus from archaeological human remains. J Archaeol Sci. 2011;38:1827–31.
130. Schroeder HE, Shanley D. Formation and Inhibition of Dental Calculus. J Periodontol. 1969;40:643–6.
131. Dobney K, Brothwell D. Dental calculus: Its relevance to ancient diet and oral ecology. Teeth Anthropol. 1986;291:55–81.
132. Dobney K, Brothwell D. A scanning electron microscope study of archaeological dental calculus [Internet]. British Archaeological Reports; 1988 [cited 2017 Oct 8]. Available from: http://www.bcin.ca/Interface/openbcin.cgi?submit=submit&Chinkey=110366
133. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6:1621–1624.
134. Weyrich LS, Dobney K, Cooper A. Ancient DNA analysis of dental calculus. J Hum Evol. 2015;79:119–24.
135. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of Human Migrations in Helicobacter pylori Populations. Science. 2003;299:1582–5.
136. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu J-Y, et al. The Peopling of the Pacific from a Bacterial Perspective. Science. 2009;323:527–30.
137. Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci Rep. 2015;5:16498.
138. Afiahayati, Sato K, Sakakibara Y. MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Res. 2015;22:69–77.

139. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ [Internet]. 2014 [cited 2017 Oct 13];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4183954/

140. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533.

141. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.

142. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.

143. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

144. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv. 2016;050559.

145. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12:902–3.

146. Wood DE, Salzberg SL. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. 2014;15:R46.

147. Vågene ÅJ, Herbig A, Campana MG, García NMR, Warinner C, Sabin S, et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. Nat Ecol Evol. 2018;1.

148. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499:431–7.

149. Choi J, Yang F, Stepanauskas R, Cardenas E, Garoutte A, Williams R, et al. Strategies to improve reference databases for soil microbiomes. ISME J. 2017;11:829–34.

Chapter I

Ancient Microbial DNA in Dental Calculus: A New Method for Studying Rapid Human Migration Events

Statement of Authorship

Title of Paper	Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events
Publication Status	☑ Published ☐ Accepted for Publication ☐ Submitted for Publication ☐ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Manuscript published

Principal Author

Name of Principal Author (Candidate)	Raphael Eisenhofer
Contribution to the Paper	Reviewed the literature, wrote, and edited the manuscript
Overall percentage (%)	60%
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 11/4/18

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	Atholl Anderson
Contribution to the Paper	Co-wrote and edited the manuscript
Signature	Date 7/3/2018

Name of Co-Author	Keith Dobney		
Contribution to the Paper	Edited the manuscript		
Signature		Date	02/04/18

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Edited the manuscript		
Signature		Date	17/04/18

Name of Co-Author	Laura S. Weyrich		
Contribution to the Paper	Edited the manuscript		
Signature		Date	9/5/18

The Journal of Island and Coastal Archaeology, 0:1–14, 2017

ISSN: 1556-4894 print / 1556-1828 online
DOI: 10.1080/15564894.2017.1382620

Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events

Raphael Eisenhofer [1], Atholl Anderson,[2] Keith Dobney,[3] Alan Cooper [1], and Laura S. Weyrich [1]

[1]*Australian Centre for Ancient DNA, University of Adelaide, Adelaide, Australia*
[2]*Department of Archaeology & Natural History, Australian National University, Canberra, Australia*
[3]*Department of Archaeology Classics and Egyptology, University of Liverpool, Liverpool, UK*

ABSTRACT

Ancient human migrations provide the critical genetic background to historical and contemporary human demographic patterns. However, our ability to infer past human migration events, especially those that occurred over rapid timescales, is often limited. A key example is the peopling of Polynesia, where the timing is relatively well defined, but the exact routes taken during the final stages and the source populations are not. Here, we discuss the technical limitations of current methods for inferring rapid human migration events, using the final stages of Polynesian migration as an example. We also introduce a promising new proxy method to infer human migrations—patterns of bacterial evolution within ancient dental calculus (calcified plaque). While we focus on Polynesia, this method should be applicable to other past migrations, enhancing our understanding of human prehistory and revealing the crucial events that shaped it.

Keywords commensal models, ancient DNA, archaeogenetics, Polynesia, Pacific settlement

INTRODUCTION

Determining past human migrations (defined here as the movement by people from one place to another with the intention of settling) are of great cultural interest and significance to many people, and can provide a sense of identity and connectedness to one's culture. The peopling of East Polynesia ~AD 1000–1300 was the last major

Received 4 October 2016; accepted 8 September 2017.
Address correspondence to Raphael Eisenhofer, Department of Genetics & Evolution, University of Adelaide, North Terrace, Adelaide, 5005 Australia. E-mail: raph.eisenhofer@gmail.com
Color versions of one or more figures in this article are available online at www.tandfonline.com/uica.

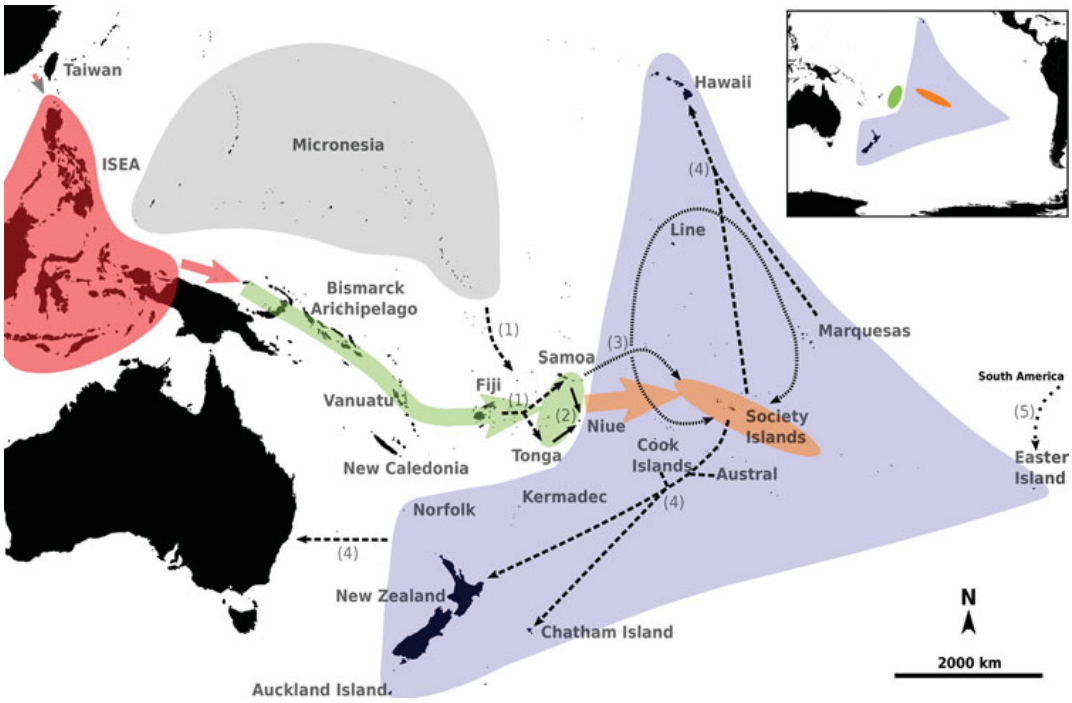


Figure 1. *General pattern of the colonization of Polynesia, and unresolved questions. The origin of populations ancestral to Polynesians in Island Southeast Asia (ISEA—red shaded area). Movement eastwards to the Bismarck Archipelago and development of Lapita culture ~1400 BC (large red arrow). Further movement eastwards by Lapita culture settling as far as Samoa and Tonga ~800 BC (green arrow and shaded area marking eastern-most extent of Lapita material culture). A ~1,800-year pause before migration into the Society Islands ~1000–1100 AD (orange arrow and shaded area). Rapid settlement of the remaining islands in the Polynesian Triangle (blue shaded area) by ~1300 AD. The numbered, differently patterned black arrows are specific to sections in the discussion highlighting unanswered questions (i.e., sections 1–5). Top-right inset illustrates the extensive size of Polynesia.*

human migration before the modern era, and one of the most geographically extensive. East Polynesia covers an area larger than North America (Figure 1), and the colonization of its islands involved the longest maritime migrations in pre-modern history (i.e., before AD 1500). Studies of material culture, linguistics, seafaring and climatic simulation, and genetics have disclosed much about the tempo and directions of migration in the western Pacific, notably of the Lapita culture that extended out to Tonga and Samoa by ~800 BC. Despite this, similar research has shown relatively little about the peopling of East Polynesia, as the timespan involved is considerably shorter.

The chronological data for East Polynesia remains unsettled, as some archaeologists argue for initial colonization of Central East Polynesia (Society, Cooks, Marquesas, western Tuamotus) or perhaps Hawaii by the late first millennium AD (Athens et al. 2014), while others argue that colonization in these archipelagos began in the twelfth century and had reached all the marginal islands, including Easter Island and New Zealand, by the late thirteenth century (Wilmshurst et al. 2011). Either way, these migrations seem to have been episodic, with an earlier influx to Central East Polynesia, a later movement to Marginal East Polynesia (Hawaii, Easter Island), and a

subsequent expansion into the outlying islands around New Zealand. The brevity of these migration pulses, each <200 years in duration, makes it difficult to reconstruct their internal structure, order, and the directions of the movements involved using current methods (archaeological material culture, linguistics, and human genetics).

A new means to infer rapid human migration events has been created through the study of the diverse communities of bacteria (microbiota) contained within the human body. These communities contain thousands of bacterial species (Hooper and Gordon 2001) that are acquired from birth via both vertical inheritance (i.e., transmission from parents to offspring) and common environment (diet, cohabitation, etc.) (Goodrich et al. 2016; Song et al. 2013; Tims et al. 2013). In addition to the strong element of vertical inheritance, these bacteria replicate quickly—a characteristic that renders the microbiota a promising proxy to infer rapid human migrations. The recent discovery that archaeological dental calculus (calcified plaque) contains ancient microbial DNA (Adler et al. 2013; Preus et al. 2011; Warinner et al. 2014) offers researchers a powerful new tool for reconstructing ancient or historical human movements that remain undocumented. Here, we summarize the existing narrative of Polynesian migrations, noting the issues that remain both unresolved and undetectable by current methods, and explain the merits of a microbial genetics approach.

POLYNESIAN MIGRATION

The origins of populations ancestral to Polynesians have been traced back to Island Southeast Asia (ISEA), on the basis of linguistic, archaeological, and genetic evidence (Anderson 2016; Anderson et al. 2014). Genetic and archaeological evidence suggest a Taiwanese contribution to modern Oceanic populations, along with significant contributions from ISEA and older groups resident in the western Pacific (Papuan)—although recent evidence suggests that Papuan genetic contributions occurred after the initial wave of settlement in western Polynesia (Skoglund et al. 2016). All of the Austronesian languages spoken outside Taiwan belong to the Malayo-Polynesian group, which extends from Madagascar through ISEA and across the Pacific to Easter Island and New Zealand. However, Malayo-Polynesian is not documented as occurring in Taiwan (Blust 2009:740) and the possibility that it entered ISEA by a different route cannot be ruled out. Further, all of the components of the so-called "transported landscape" of Oceanic populations (Kirch 2000), which included root and tree crops and commensal and domestic animals, originate in ISEA or mainland SEA rather than Taiwan.

Moving eastwards, the development of the Lapita culture ~1400 BC in the Bismarck Archipelago was a major driving process in the peopling of Remote Oceania. The Lapita culture was the first to cross the boundary between Near and Remote Oceania (a 400 km stretch of open ocean) to Vanuatu ~1000–1200 BC (Bedford et al. 2006), and moved as far as Samoa and Tonga by ~800 BC (Rieth and Hunt 2008), marking the easternmost edge of Lapita material culture, as Lapita sites further east have not been discovered. Post-Lapita migration reached the outlying islands of West Polynesia, such as Rotuma, Niue, and Pukapuka 200 BC–AD 1 (Anderson et al. 2014:25), followed much later by colonization of the Society Islands ~AD 1000–1100 (Wilmshurst et al. 2011), and possibly others in central East Polynesia as distant as Mangareva. Some 100–200 years later (AD ~1200), further eastward migration occurred out to the marginal islands of East Polynesia: Hawaii, Easter Island, then southwards to New Zealand and its outlying islands (i.e., Norfolk Island, Kermadecs, Chathams, Subantarctic islands) (Wilmshurst et al. 2011).

Although some aspects of the chronology of these migration episodes remain uncertain, it is apparent that the general

pattern of initial colonization within Polynesia is well known (Figure 1). However, much less is known about the particular origins of the migrating populations. Some clues are evident in material culture—especially where pottery is included, but pottery was not produced in East Polynesia. Many early East Polynesian artefact assemblages contain types of adzes, fish hooks, and ornaments that are so widely shared, they provide little indication of the sequence of historical relationships amongst them (Kirch 2000:243–244). Within Central East Polynesia, some specific connections can be established by source identification of basalt adzes, notably those from Marquesan and Samoan sources (Weisler 1997), but two artefacts of tropical marine shell that reached New Zealand cannot be attributed to a specific origin (Anderson et al. 2014; Davidson et al. 2011), and no other specific connections can be drawn between the central and marginal archipelagos. Historical linguistics are similarly constrained. For example, on Captain Cook's first voyage to the Pacific, Tupaia (a navigator from Raiatea in the Society Islands) was able to act as a translator during contacts with Maori in New Zealand (Cook 2003) because Maori and Tahitian languages had remained mutually intelligible to a large extent despite 500 years of separation; this was also true of Cook Islands Maori, Tahitian-Hawaiian, and so on. However, there was also some regional clustering within East Polynesian languages, which centered upon the Society and Cook Islands to the west and the Marquesas and southeastern Tuamotus-Mangareva to the east, but the extent of difference is insufficient to provide more than a general indication of the routes of migration. For example, it is clear that the Maori language and much of Maori esoteric culture belongs to the Tahitic group, but it has been impossible to use linguistic data to show that migrations to New Zealand originated in the Society Islands, rather than the Cook Islands or the Australs. In short, a means to more finely discriminate human population origins is needed, especially where migrations occurred within short timescales.

Human Genetics

To date, most human DNA studies have not directly addressed the sequence of colonization within Polynesia and have instead focused on questions about the earlier phases of Austronesian migration and the extent of admixture between Austronesians and Near Oceanians (Kayser et al. 2008; Lipson et al. 2014; Skoglund et al. 2016; Soares et al. 2011; Xu et al. 2012). Migration within Polynesia, especially East Polynesia, is difficult to study with genetics because the initial founding populations were likely small and underwent drastic and successive bottlenecks during each island migration event, further reducing genetic variation (Murray-McIntosh et al. 1998). Back migrations and modern-day admixture also likely rendered genetic diversity more homogeneous, resulting in a loss of dispersal signals. Genetic diversity in Polynesia is also likely to have been further reduced by disease epidemics introduced by early European explorers and colonists, as well as the spread of these diseases in more recent history (e.g., 1918 Influenza epidemic) (Kirch and Rallu 2007). As a result, the few studies that have measured modern genetic diversity in East Polynesia found limited genetic variation (Benton et al. 2012; Murray-McIntosh et al. 1998; Whyte et al. 2005). It is possible, however, that future studies examining larger amounts of nuclear DNA through SNP (Single-Nucleotide-Polymorphism) panels or whole genome sequencing may provide greater resolution.

Ancient DNA from human skeletal remains has been used to determine historical human migrations around the globe (Allentoft et al. 2015; Haak et al. 2015; Lazaridis et al. 2014). A major advantage of using ancient DNA is that it circumvents the influence of subsequent back-migration and contemporary genetic admixture on the migratory signal—providing key data both in time and space. However, there are a number of limitations when examining ancient human DNA in Polynesia. Warm, humid climates dominate throughout Polynesia and are known to result in poor DNA preservation (Allentoft et al. 2012). Ethical issues

involved with performing destructive sampling and analysis of human remains has also restricted the application of ancient human DNA analysis in the region. Despite this, Deguilloux et al. (2011) were able to obtain mtDNA from five ~500-year-old samples located in the Gambier Islands, and identified novel polymorphisms in a Near Oceania-associated haplogroup (Q1) and the Polynesian Motif mtDNA haplogroup. In addition, Knapp et al. (2012) applied next-generation sequencing to ~700-year-old human remains in the Wairau Bar site of New Zealand. From the 19 samples screened, DNA was successfully obtained from only 4 samples, from which full mitochondrial genomes were obtained and novel Polynesian mtDNA variation identified. Recently, Skoglund et al. obtained genome-wide ancient DNA data from three individuals from a Lapita site in Teouma (Vanuatu) and one individual from Tonga (Skoglund et al. 2016). Their results suggest that the first wave of humans into Remote Oceania had little to no Papuan ancestry, contrasting the ~25% Papuan genetic contribution found in modern Oceanic populations, and suggesting later population movements introduced Papuan ancestry to Remote Oceania. However, the small number of available ancient samples in Polynesia has limited the use of ancient human genetic data for testing hypotheses regarding later migration routes.

Genetics of Animal Proxies

Polynesians are known to have transported a number of domestic animal taxa on their voyages: dogs (*Canis lupus familiaris*), pigs (*Sus scrofa*), and chickens (*Gallus gallus*), as well as the commensal rodent—the Pacific rat (*Rattus exulans*). The DNA of these animals can be used as a proxy for human migration, with the added bonus that the marine dispersal abilities of these animals were generally poor, limiting natural migration from blurring the human migratory signal (for a review see Storey et al. 2013). Modern and ancient DNA from these animals have been used as a proxy for human migration into the Pacific. Mitochondrial DNA has been used to trace the dispersal of dogs to Polynesia via a southwestern route through Indonesia (Oskarsson et al. 2011). Ancient mitochondrial DNA from 14 dogs found at the Wairau Bar site identified a small number of haplogroups, suggesting limited genetic variation in the founding population (Greig et al. 2015). Genetic evidence for pig dispersal to Polynesia mirrors that of dogs (Larson et al. 2007, 2010), which is concordant with geometric morphometric analyses of teeth and bones (Dobney et al. 2008). The Pacific rat has been traced back as far as Flores in Indonesia using ancient and modern mitochondrial DNA (Thomson et al. 2014a), with eastwards dispersal to Polynesia (Matisoo-Smith and Robins 2004). Studies using modern and ancient mitochondrial DNA from chickens suggest an origin in the Philippines with movement eastwards to Polynesia (Dancause et al. 2011; Thomson et al. 2014b).

These studies have contributed substantially to our understanding of the earlier phases of migration in Oceania, but have yet to be applied to East Polynesia. This may be due to several limitations. Pig, chicken, and dog are distributed patchily in East Polynesia, and only the Pacific rat (*R. exulans*) occurred on nearly all islands (Anderson 2009). Again, modern DNA analyses suffer from contemporary admixture, including through European introductions. Ancient DNA analyses of these animals share many of the same limitations as for human specimens, albeit with fewer ethical considerations. As with humans, the increasing and extreme genetic bottlenecks that these organisms experienced through successive island colonization, coupled with the short and rapid timescale of migrations, prevent resolution of the routes taken during the final phases of Polynesian migration. Future recovery of more and better-preserved samples—coupled with genome-scale DNA analyses—may yet provide an improved reconstruction of events.

While constraints in current research approaches to understanding migrations in East Polynesia might well be overcome with future technical advances in existing fields of study, the most promising alternative for the moment is the study of microbial DNA.

Modern Microbial DNA

The human microbiota contains >100 × more genes than the human genome (Hooper and Gordon 2001), which—in combination with the typically fast generation time of bacteria—should provide more information about migration events than can be obtained from human or animal proxy species. While microorganisms have been used to trace large-scale human migrations, they have yet to be tested on rapid migratory events. For example, *Helicobacter pylori* is a bacterium that lives in the stomach of most individuals and is vertically transmitted within families (Rocha et al. 2003). It accompanied humans out of Africa and has thus been used as a proxy for global prehistoric human migration (Falush 2003). *H. pylori* phylogenies have also been used to explore the peopling of Oceania (Moodley et al. 2009), supporting other lines of evidence for two distinct migrations into the Pacific—an earlier one by Australians/Near Oceanians, and a later one by Malayo-Polynesian speaking people (Lapita).

Biogeographic signatures have also been obtained from bacteria within the modern human oral microbiota. Using a single protein-coding gene, *gtf* from *Streptococcus oralis*, Henne et al. (2014) were able to detect remarkable geographic resolution, especially considering its small size (330 bp). Due to the falling costs of DNA sequencing, future studies combining multiple informative genes from many bacterial species may provide the necessary resolution to reconstruct rapid human migrations. However, there are several limitations to using modern microbial DNA for this purpose. For *H. pylori*, sampling involves performing a stomach biopsy on a living individual—an invasive procedure that precludes extensive sampling. While microbial DNA likely has the resolution required to infer rapid human migratory events, there is the potential issue of modern genetic admixture reducing geographic signals. Consequently, a common, robust source of ancient human-associated microbial DNA is needed, and this has now been identified on the teeth of our long-dead ancestors.

ANCIENT DENTAL CALCULUS AS A NEW MEANS OF TRACKING HUMAN MIGRATIONS

Dental plaque is a dense, complex living microbial community that adheres to and grows on the surface of teeth. During the life of an individual, calcium phosphate ions from saliva cause this soft plaque to undergo periodic mineralization events, trapping lower layers in an extremely hard, calcified deposit called dental calculus. The prevalence and robust nature of dental calculus makes it common on the teeth of archaeological human remains. The first definitive observation of calcified bacteria trapped within archaeological dental calculus occurred nearly 30 years ago (Dobney and Brothwell 1986), which led to subsequent research over the next 8 years (Dobney and Brothwell 1988; Dobney 1994). It was not until the application of new scientific techniques in archaeology some 13 years later that actual bacterial DNA preserved within ancient dental calculus was observed for the first time via gold-labeled antibody transmission electron microscopy (Preus et al. 2011). This was further verified by the successful extraction, amplification, and sequencing of bacterial DNA from ancient dental calculus (De La Fuente et al. 2013). The first community-level analyses of ancient dental calculus detected changes in human oral microbiota communities likely correlating with dietary changes over 8,000 years, from early agriculturalists to the Industrial Revolution (Adler et al. 2013). An increase in the carriage of tooth decay-associated bacteria was also observed through time, likely reflecting the increasing availability of carbohydrates. The composition of the human oral microbiota appeared relatively distinct to each culture and geographic region, highlighting the potential power of microbiota DNA preserved within dental calculus to provide an ancient genetic signal of cultural affinity (Adler

et al. 2013). Subsequent studies have reconstructed full genomes of oral pathogens such as *Tannerella forsythia* from medieval specimens (Warinner et al 2014). Collectively, these pioneering studies illustrate that high-resolution investigation of ancient oral microbiota, both at the community and individual bacterium level has the power to provide novel views of human bio-cultural evolution. Protein sequencing has also been applied to ancient calculus, showing that bacterial functions (e.g., virulence factors) and their interactions with the human host (e.g., immune proteins) are obtainable from ancient dental calculus (Warinner et al. 2014). A further key finding was that DNA from ancient calculus includes signals from food sources such as plants and animals, demonstrating the ability to obtain dietary information from ancient human populations, providing further information to delineate past human lifeways (Warinner et al. 2014; Weyrich et al. 2017).

While the ethics involved in analyzing ancient human DNA can be extremely sensitive, ancient dental calculus is almost entirely microbial in origin (>99.9%), with very limited amounts of human DNA. In addition, dental calculus can be easily removed from teeth, avoiding the destruction of human remains. These two factors ensure ancient calculus can be examined with minimal alteration to valuable museum specimens, potentially allowing large numbers of samples to be collected and analyzed. Practically speaking, ancient dental calculus contains a much higher concentration of DNA than bone (Warinner et al. 2014), which increases the odds of successfully obtaining DNA from poorly preserved ancient specimens. Importantly, the use of oral bacterial DNA in ancient dental calculus also solves the issue of modern genetic admixtures confounding human migration signal. In addition, the oral microbiota exhibits a strong degree of vertical inheritance (Corby et al. 2007; Ebersole et al. 2014; Li et al. 2007; Okada et al. 2004), making it a suitable proxy for human migration. Finally, the large amount of genetic material found in the oral microbiota, coupled with the rapid generation time in bacteria, should provide the resolution required for discerning rapid human migratory events. These characteristics render ancient human dental calculus a promising new means of studying past human population movements.

Potential Methods for Analyzing Oral Bacteria for Human Migrations

To track human migrations using ancient dental calculus, genetic mutations within bacterial species must be identified, as bacterial community structure would likely provide insufficient resolution. To examine species and strain level resolution within ancient calculus specimens, several techniques could be applied. The most promising centers around a recently published method called StrainPhlAn (Truong et al. 2017). StrainPhlAn uses species-specific marker genes (Truong et al. 2015) to measure strain-level genetic diversity from metagenomic samples. Briefly, DNA reads from a metagenomic sample are mapped to species-specific markers, and consensus sequences of the dominant strains are constructed. For each species, the consensus sequences are then aligned, concatenated, and used as input for maximum likelihood phylogenetic analysis. This program provides the ability to investigate genetic differences with high resolution between bacterial species from different samples, which may be sufficient to infer past rapid human migrations. In addition, genome level information from strains could be similarly obtained; strain level resolution of commensal species has been recently observed, suggesting that this approach may be feasible for future studies (Weyrich et al. 2017). In either approach, different strains would need to be identified and assessed individually; then, DNA from the dominant strain or multiple strains can be compared and assessed through time as a marker for human migration.

To ensure this method can be utilized, shared bacterial taxa must be identified within each of the samples of interest. This should be a limited issue in dental calculus,

as research suggests that many species are shared between individuals within populations, reflecting a 'core' calculus microbiota (Welch et al. 2016). The oral microbiome has also been shown to be the one of the most stable and conserved human microbiomes to date (Utter et al. 2016), and many oral strains have been shown to be conserved across hominin species (Weyrich et al. 2017), indicating that shared microbial taxa may be available for this type of analyses across human populations. Strain heterogeneity within a sample might confound basic phylogenetic analysis in some cases, but it is likely that the sample will be dominated by a single strain (Truong et al. 2017), which could be used to infer migratory processes. For example, (Truong et al. 2017) found that for 1,500 deeply sequenced gut metagenomes, a single strain typically dominated—even through time—for >70% of the species analyzed, indicating that this approach is plausible within diverse, mixed strain metagenomes.

Another important consideration is the sequencing coverage of the genetic loci to be analyzed (i.e., how many DNA sequences can be obtained for each genetic locus). Insufficient coverage would likely result in the inability to determine if a nucleotide change is due to stochastic effects (sequencing artefacts, DNA damage, etc.), or if a mixed signal is resulting from strain-level polymorphisms. Obtaining sufficient coverage is especially challenging for ancient DNA due to its fragmented and damaged nature (Dabney et al. 2013). However, one possible solution is to improve the coverage of genetic loci of interest using hybridization-enrichment (Maricic et al. 2010), whereby specific DNA sequences are captured prior to sequencing. This technique drastically lowers the cost of sequencing and increases the coverage obtained, especially for ancient DNA (Schuenemann et al. 2011; Skoglund et al. 2016). Finally, while reference bias towards well-studied strains may make strain-level identification challenging, (Truong et al. 2017) were able to reconstruct strain-level genetic diversities 10-fold higher than were previously available, demonstrating StrainPhlAn's ability to accommodate incomplete reference data.

A potential issue for using bacteria as a proxy for human migration is horizontal gene transfer (HGT). Bacteria are known to transfer genetic material between each other—even among distantly related species—and such transfers could confound phylogenetic analyses. To circumvent this issue, one can target genetic loci that are conserved, single-copy, and show no or low levels of horizontal gene transfer. Such genes are typically involved in essential informational processes and are parts of large, complex systems (e.g., transcription, translation, tRNA synthetases) (Rivera et al. 1998), and are thought to be recalcitrant to horizontal transfer (Hao and Golding 2008; Jain et al. 1999). The practical use of such loci for bacterial phylogenetic analyses has been previously demonstrated in the literature (Ankenbrand and Keller 2016; Darling et al. 2014; Wu et al. 2013; Wu and Scott 2012). To ensure that these genetic loci are not horizontally transferred, software is available that can detect horizontal gene transfer/recombination in bacterial genomes (Croucher et al. 2015; Martin et al. 2015). One can use these tools to identify putatively horizontally-transferred loci, and once identified, these loci can be discarded from phylogenetic analyses.

To summarize, the approach and tools required for employing dental calculus as a high-resolution proxy of past human migrations are available and feasible, and so is the prospect of being able to infer rapid human migration events from such data.

DISCUSSION AND CONCLUSIONS

As we continue to investigate the origins of migrating populations in Polynesia, it is pertinent to ask which migration events to consider, and why these specific events are of interest to Polynesian prehistory. Below, we have listed five issues for which analysis of ancient microbial DNA could provide critical insights into Polynesian prehistory.

1. The initial colonization of West Polynesia, and questions about possible later migrations from Micronesia (Addison and Matisoo-Smith 2010; Davidson 2012), constitute an important unresolved issue. The radiocarbon chronology of colonization sites indicates migration from Fiji, colonized about 900 BC, to Tonga and Samoa by 800 BC (Burley et al. 2010; Petchey 2001). In contrast, the relative frequency of Lapita sites suggests a linear migration through Tonga (where Lapita sites are scarcer to the north) and Samoa (where there is only one Lapita site) (Clark and Anderson 2009). However, it is also possible that Fiji was the direct source of migrants for both Samoa and Tonga, in which case a progressive decay of the migratory pulse would require an alternative explanation (e.g., Burley et al. 2011; Rieth et al. 2008).
2. On the same theme, the later colonization of Niue remains unknown. Niue is closer to Tonga than Samoa, and the Niuean language is closer to Tongan. However, Samoan features are strongly evident in the place names and oral traditions. In addition, there is some evidence that Niue was colonized from East Polynesia (Walter and Anderson 2002). Was the initial source population from Tonga or Samoa, and when did East Polynesian influence begin? Identification of either Tonga or Samoa as the source has potentially significant implications for the later colonization of East Polynesia.
3. Samoa is commonly referred to as the most likely source of the original migrants to East Polynesia, with suggested voyages through the northern and southern Cook Islands, through the northern Cook Islands to the Society Islands, or perhaps on several routes during the colonizing period (Kirch 2000; Kirch and Green 2001; Montenegro et al. 2014, 2016; Wilson 2012). However, there is a strong linguistic argument for initial migration into East Polynesia through the northern atolls, including the Phoenix and Line Islands, suggesting that a more likely entry to East Polynesia would have been through the Marquesas (Wilson 2012). The different scenarios imply quite different histories of colonization, including the development and spread of language, material culture, and seafaring capabilities.
4. Within East Polynesia, there are several cases where alternative source populations of initial migrants are possible, and sorting out the alternatives will help to refine issues of origin, relatedness, and interaction in the region. Hawaii may have been colonized from the Marquesas (Kirch 2000:291), but it is possible that some colonists came from Tahiti, as marked by a Tahiti cosmogony in Hawaii (Marck 2000:230). Similarly, the colonization of New Zealand was from the Tahitic (western) region of central East Polynesia, but its specific origin remains unknown. Was it directly from the Society Islands? Was it from the Southern Cook Islands or from the Austral Islands? Was it a single population dispersal from one particular island or archipelago or a general movement from across the Tahitic region? These differences can help inform the causes of migration (i.e., dispersals triggered by specific events mentioned in oral traditions, or a broader event horizon or process). The Moriori people of the Chatham Islands that lie east of New Zealand have linguistic and material culture parallels with New Zealand Maori (Sutton 1980), but their oral traditions are somewhat different (Shand 1911). Modern Moriori prefer the possibility that their ancestors arrived by direct passages from central East Polynesia, and the distance to Chatham Island and New Zealand from the southern Cook Islands is almost exactly the same.
5. Accumulating data are once more raising the question about whether some earlier migrants to Easter Island had Amerindian origins. The oldest archaeology that appears on Easter Island clearly has East Polynesian origins,

but research in genetics (Moreno-Mayar 2014) and in the origins of monumental architecture (Martinsson-Wallin et al. 2013) has suggested a New World influence soon after initial colonization.

All of these and many other questions about the prehistory of Polynesia exist primarily because our current methodologies do not have sufficient temporal discrimination to trace the movement of people between islands at a centennial scale resolution. Such difficulties are not unique to Polynesia and are observed elsewhere, including the Caribbean and within the American continents (Fitzpatrick 2015; Hofman et al. 2008; Keegan and Hofman 2017). The study of ancient microbial DNA within dental calculus, as outlined here, presents a powerful new tool to identify human cultural signals, track bacterial genome evolution, and ultimately reconstruct human migration patterns. Analysis of the microbiota in dental calculus will provide unprecedented opportunities to trace human movements, thereby enhancing our understanding of human prehistory.

ORCID

Raphael Eisenhofer http://orcid.org/0000-0002-3843-0749
Alan Cooper http://orcid.org/0000-0002-7738-7851
Laura S. Weyrich http://orcid.org/0000-0001-5243-4634

REFERENCES

Addison, D. J., and E. Matisoo-Smith. 2010. Rethinking Polynesians origins: A West-Polynesia Triple-I Model. *Archaeology in Oceania* 45: 1–12.

Adler, C. J., K. Dobney, L. S. Weyrich, J. Kaidonis, A. W. Walker, W. Haak, C. J. A. Bradshaw, et al. 2013. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics* 45:450–455.

Allentoft, M. E., M. Collins, D. Harker, J. Haile, C. L. Oskam, M. L. Hale, P. F. Campos, et al. 2012. The half-life of DNA in bone: Measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society of London B: Biological Sciences* 279:4724–4733.

Allentoft, M. E. M. Sikora, K. G. Sjögren, S. Rasmussen, M. Rasmussen, J. Stenderup, P. B. Damgaard, H. Schroeder, et al. 2015. Population genomics of Bronze Age Eurasia. *Nature* 522:167–172.

Anderson, A. 2009. The rat and the octopus: Initial human colonization and the prehistoric introduction of domestic animals to Remote Oceania. *Biological Invasions* 11: 1503–1519.

Anderson, A. 2016. *The First Migration: Maori origins 3000 BC - AD 1450*. Wellington: BWB Texts.

Anderson, A., J. Binney, A. Harris, Auckland War Memorial Museum; Stout Trust. 2014. *Tangata Whenua: An Illustrated History*. Wellington, New Zealand: Bridget Williams Books.

Ankenbrand, M. J., and A. Keller. 2016. bcgTree: Automatized phylogenetic tree building from bacterial core genomes. *Genome*: 1–9.

Athens, J. S., T. Rieth, and T. Dye. 2014. A paleoenvironmental and archaeological model-based age estimate for the colonization of Hawai'i. *American Antiquity* 79:144–155.

Bedford, S., M. Spriggs, and R. Regenvanu. 2006. The Teouma Lapita site and the early human settlement of the Pacific Islands. *Antiquity* 80:812–828.

Benton, M., D. Macartney-Coxson, D. Eccles, L. Griffiths, G. Chambers, and R. Lea. 2012. Complete mitochondrial genome sequencing reveals novel haplotypes in a Polynesian population. *PLOS ONE* 7:e35026.

Blust, R. A. 2009. The Austronesian languages Vol. 602. Pacific Linguistics: Research School of Pacific and Asian Studies, Australian National University.

Burley, D., A. Barton, W. R. Dickinson, S. P. Connaughton, and K. Taché. 2010. Nukuleka as a founder colony for West Polynesian settlement: New insights from recent excavations. *Journal of Pacific Archaeology* 1:128–144.

Burley, D. V., P. J. Sheppard, and M. Simonin. 2011. Tongan and Samoan volcanic glass: pXRF analysis and implications for constructs of ancestral Polynesian society. *Journal of Archaeological Science* 38:2625–2632.

Clark, G., and A. Anderson. 2009. *Colonisation and Culture Change in the Early Prehistory of Fiji. The Early Prehsitory of Fiji. Terra Australis*. Canberra: ANU E-Press.

Cook, J. 2003, June 22. *Captain Cook's Journal During His First Voyage Round the World Made in H. M. Bark "Endeavour" 1768–71*. eBooks@Adelaide: The University of Adelaide.

Corby, P. M., W. A. Bretz, T. C. Hart, N. J. Schork, J. Wessel, J. Lyons-Weiler, and B. J. Paster. 2007. Heritability of oral microbial species in caries-active and caries-free twins. *Twin Research and Human Genetics* 10: 821–828.

Croucher, N. J., A. J. Page, T. R. Connor, A. J. Delaney, J. A. Keane, S. D. Bentley, J. Parkhill, and S. R. Harris. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* 43:e15–e15.

Dabney, J., M. Meyer, and S. Pääbo. 2013. Ancient DNA damage. *Cold Spring Harbor Perspectives in Biology*: a012567.

Dancause, K. N., M. G. Vilar, R. Steffy, and J. K. Lum. 2011. Characterizing genetic diversity of contemporary Pacific chickens using mitochondrial DNA analyses. *PLOS ONE* 6:e16843.

Darling, A. E., G. Jospin, E. Lowe, F. A. M. Iv, H. M. Bik, and J. A. Eisen. 2014. PhyloSift: Phylogenetic analysis of genomes and metagenomes. *PeerJ* 2:e243.

Davidson, J., A. Findlater, R. Fyfe, J. MacDonald, and B. Marshall. 2011. Connections with Hawaiki: The evidence of a shell tool from Wairau Bar, Marlborough, New Zealand. *Journal of Pacific Archaeology* 2:93–102.

Davidson, J. M. 2012. Intrusion, integration and innovation on small and not-so-small islands with particular reference to Samoa. *Archaeology in Oceania* 47:1–13.

Deguilloux, M. F., M. H. Pemonge, V. Dubut, S. Hughes, C. Hänni, L. Chollet, E. Conte, and P. Murail. 2011. Human ancient and extant mtDNA from the Gambier Islands (French polynesia): Evidence for an early Melanesian maternal contribution and new perspectives into the settlement of Easternmost Polynesia. *American Journal of Physical Anthropology* 144:248–257.

De La Fuente, C., S. Flores, and M. Moraga. 2013. DNA from human ancient bacteria: A novel source of genetic evidence from archaeological dental calculus. *Archaeometry* 55: 767–778.

Dobney, K., 1994. Study of the dental calculus. *The Jewish burial ground at Jewbury*.

Dobney, K., and D. Brothwell. 1986. Dental calculus: Its relevance to ancient diet and oral ecology. *Teeth and Anthropology* 291:55–81.

Dobney, K., and D. Brothwell. 1988. A scanning electron microscope study of archaeological dental calculus. *British Archaeological Reports*.

Dobney, K., T. Cucchi, and G. Larson. 2008. The pigs of Island Southeast Asia and the Pacific: New evidence for taxonomic status and human-mediated dispersal. *Asian Perspectives*.

Ebersole, J. L., S. C. Holt, and J. E. Delaney. 2014. Acquisition of oral microbes and associated systemic responses of newborn nonhuman primates. *Clinical and Vaccine Immunology: CVI* 21:21–28.

Falush, D., T. Wirth, B. Linz, J. K. Pritchard, M. Stephens, M. Kidd, and M. J. Blaser, et al. 2003. Traces of human migrations in *Helicobacter pylori* populations. *Science* 299: 1582–1585.

Fitzpatrick, S. M. 2015. The Pre-Columbian Caribbean: Colonization, population dispersal, and island adaptations. *PaleoAmerica* 1: 305–331.

Goodrich, J. K., E. R. Davenport, M. Beaumont, M. A. Jackson, R. Knight, C. Ober, T. D. Spector, J. T. Bell, A. G. Clark, and R. E. Ley. 2016. Genetic determinants of the gut microbiome in UK twins. *Cell Host & Microbe* 19:731–743.

Greig, K., J. Boocock, S. Prost, K. A. Horsburgh, C. Jacomb, R. Walter, and E. Matisoo-Smith. 2015. Complete mitochondrial genomes of New Zealand's first dogs. *PLOS ONE* 10:e0138536.

Haak, W., I. Lazaridis, N. Patterson, N. Rohland, S. Mallick, B. Llamas, G. Brandt, et al. 2015. Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* 522:207–211.

Hao, W., and G. B. Golding. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics* 9:235.

Henne, K., J. Li, M. Stoneking, O. Kessler, H. Schilling, A. Sonanini, G. Conrads, and H. P. Horz. 2014. Global analysis of saliva as a source of bacterial genes for insights into human population structure and migration studies. *BMC Evolutionary Biology* 14:190.

Hofman, C. L., A. J. Bright, M. L. P. Hoogland, and W. F. Keegan. 2008. Attractive ideas, desirable goods: Examining the Late Ceramic Age relationships between Greater and Lesser Antillean societies. *The Journal of Island and Coastal Archaeology* 3:17–34.

Hooper, L. V., and J. I. Gordon. 2001. Commensal host-bacterial relationships in the gut. *Science* 292:1115–1118.

Jain, R., M. C. Rivera, and J. A. Lake. 1999. Horizontal gene transfer among genomes: The complexity hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 96:3801–3806.

Kayser, M., O. Lao, K. Saar, S. Brauer, X. Wang, P. Nürnberg, R. J. Trent, and M. Stoneking. 2008. Genome-wide analysis indicates more

Asian than Melanesian ancestry of Polynesians. *American Journal of Human Genetics* 82:194–198.

Keegan, W. F., and C. L. Hofman. 2017. *The Caribbean before Columbus*. Oxford; New York: Oxford University Press.

Kirch, P. V. 2000. On the road of the winds: An archaeological history of the Pacific islands before European contact, University of California Press.

Kirch, P. V., and R. C. Green. 2001. *Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology*. Cambridge: Cambridge University Press.

Kirch, P. V., and J. L. Rallu (eds.). 2007. *The Growth and Collapse of Pacific Island Societies: Archaeological and Demographic Perspectives*. Honolulu: University of Hawai'i Press.

Knapp, M., K. A. Horsburgh, S. Prost, J. A. Stanton, H. R. Buckley, R. K. Walter, and E. A. Matisoo-Smith, 2012. Complete mitochondrial DNA genome sequences from the first New Zealanders. *Proceedings of the National Academy of Sciences of the United States of America* 109:18350–18354.

Larson, G., T. Cucchi, M. Fujita, E. Matisoo-Smith, J. Robins, A. Anderson, B. Rolett, et al. 2007. Phylogeny and ancient DNA of *Sus* provides insights into neolithic expansion in Island Southeast Asia and Oceania. *Proceedings of the National Academy of Sciences* 104:4834–4839.

Larson, G., R. Liu, X. Zhao, J. Yuan, D. Fuller, L. Barton, K. Dobney, et al. 2010. Patterns of East Asian pig domestication, migration, and turnover revealed by modern and ancient DNA. *Proceedings of the National Academy of Sciences of the United States of America* 107:7686–7691.

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

Li, Y., A. I. Ismail, Y. Ge, M. Tellez, and W. Sohn. 2007. Similarity of bacterial populations in saliva from African-American mother-child dyads. *Journal of Clinical Microbiology* 45:3082–3085.

Lipson, M., P. R. Loh, N. Patterson, P. Moorjani, Y. C. Ko, M. Stoneking, B. Berger, and D. Reich. 2014. Reconstructing Austronesian population history in Island Southeast Asia. *Nature Communications* 5.

Marck, J. C. 2000. *Topics in Polynesian Language and Culture History*. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Maricic, T., M. Whitten, and S. Pääbo. 2010. Multiplexed DNA Sequence capture of mitochondrial genomes using PCR products. *PLOS ONE* 5:e14004.

Martin, D. P., B. Murrell, M. Golden, A. Khoosal, and B. Muhire. 2015. RDP4: Detection and analysis of recombination patterns in virus genomes. *Virus Evolution* 1:vev003.

Martinsson-Wallin, H., P. Wallin, A. Anderson, and R. Solsvik. 2013. Chronogeographic variation in initial East Polynesian construction of monumental ceremonial sites. *Journal of Island and Coastal Archaeology* 8: 405–421.

Matisoo-Smith, E., and J. H. Robins. 2004. Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat. *Proceedings of the National Academy of Sciences of the United States of America* 101:9167–9172.

Montenegro, A., R. T. Callaghan, and S. M. Fitzpatrick. 2014. From west to east: Environmental influences on the rate and pathways of Polynesian colonization. *The Holocene* 24: 242–256.

Montenegro, Á., R. T. Callaghan, and S. M. Fitzpatrick. 2016. Using seafaring simulations and shortest-hop trajectories to model the prehistoric colonization of Remote Oceania. *Proceedings of the National Academy of Sciences* 113:12685–12690.

Moodley, Y., B. Linz, Y. Yamaoka, H. M. Windsor, S. Breurec, J. Y. Wu, A. Maady, et al. 2009. The peopling of the Pacific from a bacterial perspective. *Science* 323:527–530.

Moreno-Mayar, J. V., S. Rasmussen, A. Seguin-Orlando, M. Rasmussen, M. Liang, S. T. Flåm, B. A Lie, G. D. Gilfillan, R. Nielsen, E. Thorsby, E. Willerslev, and A.S. Malaspinas. 2014. Genome-wide Ancestry Patterns in Rapanui Suggest Pre-European Admixture with Native Americans. *Current Biology* 24:2518–2525.

Murray-McIntosh, R. P., B. J. Scrimshaw, P. J. Hatfield, and D. Penny. 1998. Testing migration patterns and estimating founding population size in Polynesia by using human mtDNA sequences. *Proceedings of the National Academy of Sciences of the United States of America* 95:9047–9052.

Okada, M., F. Hayashi, Y. Soda, X. Zhong, K. Miura, and K. Kozai. 2004. Intra-familial distribution of nine putative periodontopathogens in dental plaque samples analyzed by PCR. *Journal of Oral Science* 46:149–156.

Oskarsson, M. C. R., C. F. C. Klütsch, U. Boonyaprakob, A. Wilton, Y. Tanabe, and P. Savolainen. 2011. Mitochondrial DNA data indicate an introduction through Mainland Southeast Asia for Australian dingoes and Polynesian domestic dogs. *Proceedings of the Royal Society of London B: Biological Sciences*: rspb20111395.

Petchey, F. 2001. Radiocarbon determinations from the Mulifanua Lapita site, Upolu, western Samoa. *Radiocarbon* 43(1):63–68.

Preus, H. R., O. J. Marvik, K. A. Selvig, and P. Bennike. 2011. Ancient bacterial DNA (aDNA) in dental calculus from archaeological human remains. *Journal of Archaeological Science* 38:1827–1831.

Rieth, T. M., and T. L. Hunt. 2008. A radiocarbon chronology for Sāmoan prehistory. *Journal of Archaeological Science* 35:1901–1927.

Rieth, T. M., A. E. Morrison, and D. J. Addison. 2008. The temporal and spatial patterning of the initial settlement of Sāmoa. *Journal of Island and Coastal Archaeology* 3:214–239.

Rivera, M. C., R. Jain, J. E. Moore, and J. A. Lake. 1998. Genomic evidence for two functionally distinct gene classes. *Proceedings of the National Academy of Sciences* 95: 6239–6244.

Rocha, G. A., A. M. C. Rocha, L. D. Silva, A. Santos, A. C. D. Bocewicz, R. de, M. Queiroz, J. Bethony, A. Gazzinelli, R. Corrêa-Oliveira, and D. M. M. Queiroz. 2003. Transmission of *Helicobacter pylori* infection in families of preschool-aged children from Minas Gerais:Brazil. *Tropical Medicine & International Health: TM & IH* 8:987–991.

Schuenemann, V. J., K. Bos, S. DeWitte, S. Schmedes, J. Jamieson, A. Mittnik, S. Forrest, et al. 2011. Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proceedings of the National Academy of Sciences* 108:E746–E752.

Shand, A. 1911. *The Moriori People of the Chatham Islands: Their History and Traditions / by the Late Alexander Shand*. Memoirs of the Polynesian Society; v. 2. Wellington: Polynesian Society of New Zealand.

Skoglund, P., C. Posth, K. Sirak, M. Spriggs, F. Valentin, S. Bedford, G. R. Clark, et al. 2016. Genomic insights into the peopling of the Southwest Pacific. *Nature* advance online publication.

Soares, P., T. Rito, J. Trejaut, M. Mormina, C. Hill, E. Tinkler-Hundal, M. Braid, et al. 2011. Ancient voyaging and Polynesian origins. *American Journal of Human Genetics* 88:239–247.

Song, S. J., C. Lauber, E. K. Costello, C. A. Lozupone, G. Humphrey, D. Berg-Lyons, J. G. Caporaso, et al. 2013. Cohabiting family members share microbiota with one another and with their dogs. *eLife* 2:e00458.

Storey, A. A., A. C. Clarke, T. Ladefoged, J. Robins, and E. Matisoo-Smith. 2013. DNA and Pacific commensal models: Applications, construction, limitations, and future prospects. *Journal of Island and Coastal Archaeology* 8:37–65.

Sutton, D. G. 1980. A culture and history of the Chatham Islands. *Journal of the Polynesian Society* 89:67–93.

Thomson, V. A., K. P. Aplin, A. Cooper, S. Hisheh, H. Suzuki, I. Maryanto, G. Yap, and S. C. Donnellan. 2014a. Molecular genetic evidence for the place of origin of the Pacific rat, Rattus exulans. *PLOS ONE* 9:e91356.

Thomson, V. A., O. Lebrasseur, J. J. Austin, T. L. Hunt, D. A. Burney, T. Denham, N. J. Rawlence, et al. 2014b. Using ancient DNA to study the origins and dispersal of ancestral Polynesian chickens across the Pacific. *Proceedings of the National Academy of Sciences* 111: 4826–4831.

Tims, S., C. Derom, D. M. Jonkers, R. Vlietinck, W. H. Saris, M. Kleerebezem, W. M. de Vos, and E. G. Zoetendal. 2013. Microbiota conservation and BMI signatures in adult monozygotic twins. *ISME Journal* 7:707–717.

Truong, D. T., E. A. Franzosa, T. L. Tickle, M. Scholz, G. Weingart, E. Pasolli, A. Tett, C. Huttenhower, and N. Segata. 2015. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods* 12:902–903.

Truong, D. T., A. Tett, E. Pasolli, C. Huttenhower, and N. Segata. 2017. Microbial strain-level population structure and genetic diversity from metagenomes. *Genome Research* 27:626–638.

Utter, D. R., J. L. Mark Welch, and G. G. Borisy. 2016. Individuality, stability, and variability of the plaque microbiome. *Frontiers in Microbiology* 7.

Walter, R. K., and A. Anderson. 2002. *The Archaeology of Niue Island, West Polynesia*. Honolulu: Bishop Museum Press.

Warinner, C., J. F. M. Rodrigues, R. Vyas, C. Trachsel, N. Shved, J. Grossmann, A. Radini, et al. 2014. Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics* 46:336–344.

Weisler, M. I. 1997. Prehistoric long-distance interaction at the margins of Oceania. In *Prehistoric Long-distance Interaction in Oceania: An Interdisciplinary Approach*:

149–172. Dunedin: New Zealand Archaeological Association Monograph 21.

Welch, J. L. M., B. J. Rossetti, C. W. Rieken, F. E. Dewhirst, and G. G. Borisy. 2016. Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences* 113:E791–E800.

Weyrich, L. S., S. Duchene, J. Soubrier, L. Arriola, B. Llamas, J. Breen, A. G. Morris, et al. 2017. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* 544, 357–361.

Whyte, A. L. H., S. J. Marshall, and G. K. Chambers. 2005. Human evolution in Polynesia. *Human Biology* 77:157–177.

Wilmshurst, J. M., T. L. Hunt, C. P. Lipo, and A. J. Anderson. 2011. High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. *Proceedings of the National Academy of Sciences* 108: 1815–1820.

Wilson, W. H. 2012. Whence the East Polynesians?: Further linguistic evidence for a northern outlier source. *Oceanic Linguistics* 51:289–359.

Wu, D., G. Jospin, and J. A. Eisen. 2013. Systematic identification of gene families for use as "markers" for phylogenetic and phylogeny-driven ecological studies of bacteria and archaea and their major subgroups. *PLOS ONE* 8:e77033.

Wu, M., and A. J. Scott. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. *Bioinformatics* 28:1033–1034.

Xu, S., I. Pugach, M. Stoneking, M. Kayser, L. Jin, and T. H. P. A. S. Consortium. 2012. Genetic dating indicates that the Asian–Papuan admixture through Eastern Indonesia corresponds to the Austronesian expansion. *Proceedings of the National Academy of Sciences* 109: 4574–4579.

Chapter II

—

Contamination in low-biomass microbiome studies: issues and recommendations

Statement of Authorship

Title of Paper	Contamination in low-biomass microbiome studies: issues and recommendations
Publication Status	☐ Published ☐ Accepted for Publication ☑ Submitted for Publication
Publication Details	Submitted to *Microbiome*.

Principal Author

Name of Principal Author (Candidate)	Raphael Eisenhofer		
Contribution to the Paper	Reviewed the literature, wrote, and edited the manuscript.		
Overall percentage (%)	60%		
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	10/05/18

Co-Author Contributions

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

i. the candidate's statec 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	Clarisse Marotz		
Contribution to the Paper	Reviewed the literature and edited the manuscript		
Signature		Date	27/3/2018

Name of Co-Author	Jeremiah J. Minich		
Contribution to the Paper	Reviewed the literature and edited the manuscript		
Signature		Date	3/4/2018

Name of Co-Author	Alan Cooper		
Contribution to the Paper	Reviewed the literature and edited the manuscript		
Signature		Date	17/04/18

Name of Co-Author	Rob Knight		
Contribution to the Paper	Reviewed the literature and edited the manuscript		
Signature		Date	26/3/18

Name of Co-Author	Laura S. Weyrich		
Contribution to the Paper	Reviewed the literature and edited the manuscript		
Signature		Date	9/5/18

Contamination in low-biomass microbiome studies: issues and recommendations

Authors: Raphael Eisenhofer[1], Jeremiah J. Minich[2], Clarisse Marotz[3], Alan Cooper[1], Rob Knight[3,4,5], and Laura S. Weyrich[1]

Author affiliations:

1. Australian Centre for Ancient DNA, University of Adelaide, Australia
2. Marine Biology Research Division, Scripps Institution of Oceanography, La Jolla, California, USA
3. Department of Pediatrics, University of California San Diego, La Jolla, USA
4. Center for Microbiome Innovation, University of California San Diego, La Jolla, USA
5. Department of Computer Science and Engineering, University of California San Diego, La Jolla, California, USA

Abstract

Next-generation sequencing approaches in microbiome research have reshaped the way we view the world, allowing researchers to survey microbial communities, their genomes, and their functions with higher sensitivity than ever before. However, this sensitivity is a double-edged sword, as these tools also efficiently detect contaminant DNA (DNA from sources other than the samples under study) and cross-contamination (DNA exchange between samples). Contaminant DNA and cross-contamination can confound the interpretations of microbiome data by generating false signals, which have been recently, and repeatedly, interpreted as meaningful findings. Therefore, there is urgent need for the field to integrate key controls into microbiome research to improve the integrity of microbiome studies. Here, we review how contaminant DNA and cross-contamination arise within microbiome studies and discuss the negative impacts they can have in microbiome research, especially during the analysis of low-biomass samples. We then identify several key measures that researchers can implement to reduce the impacts of contaminant DNA and cross-contamination during microbiome research. We put forth a set of minimal experimental criteria to improve the validity of future low-biomass research, and we package our recommendations into a novel checklist to help scientists, reviewers, and editors improve microbiome research, called the '**RIDE**' checklist – **R**eport methodology, **I**nclude controls, **D**etermine the limit of detection, and **E**xplore the impacts of contamination in downstream analysis. We hope that these criteria will help improve the scientific integrity of future microbiome research.

Background

The completion of the Human Microbiome Project in 2017 [1] was a major landmark in **microbiome** research. This research field has the potential to create novel therapies for human disease, aid in environmental conservation, improve agricultural outputs, understand our ancestor's lifestyles, and identify criminals in forensic casework, amongst many other areas [2–6].

Amplification-based methods that target hypervariable regions (*e.g.* PCR amplification of the 16S ribosomal RNA (rRNA) gene) account for the majority of current studies exploring **microbiota** because of their speed and inexpensive cost [7]. Shotgun sequencing has also become more popular in recent years due to decreasing DNA sequencing costs and the ability to obtain both species-level taxonomic resolution and functional genomic information. Both of these approaches rapidly illuminate unculturable microorganisms and allow researchers to compare and contrast microbial communities in diverse environments, including the human body, subglacial Antarctic lakes, NASA's space equipment, deep-sea hydrothermal vents, extinct hominids, and coral reefs [5,8–12].

Despite their benefits, the molecular methods used to investigate microbial communities have key limitations, namely non-proportional target amplification and the inclusion of **contamination**. While tools to address non-proportional target amplification have been developed [13–15], strategies to limit contamination are less appreciated. Several studies have documented the routine amplification of contamination and its impacts on biological interpretations [16–24], but there is still no systematic requirement to examine or report contamination within microbiota or microbiome (hereby referred to as microbiome) studies. Here, we highlight how contamination has negatively impacted microbiome research, especially when assessing low-biomass samples, and provide several recommendations to minimize the effects of contamination in future research.

Main text

What are contaminants in microbiome studies, where do they arise, and are they static?

Two key types of contamination can arise in microbiome studies: **contaminant DNA** and **cross-contamination**. Contaminant DNA can originate from many sources despite the utmost care in sample collection and preparation, including the sampling and laboratory environments [25–27], researchers, plastic consumables [28], nucleic acid extraction kits [5,19,23,24,29–32], laboratory reagents, including PCR mastermixes [16–18,33–36], and cross-contamination from other samples and sequencing runs [37,38]. To date, over 30 common contaminant taxa have been identified in **DNA extraction blank controls** and **no-template controls** across multiple studies. For example, Salter *et al.* found that several contaminant taxa were shared in blank controls across multiple studies, laboratories, and DNA extraction methods [19]. These widespread contaminant taxa appear to originate from common sources (*e.g.* kit and reagent manufacturing, human commensals on lab personnel, or thrive within laboratory environments). Despite the identification of some common contaminants, the types and abundance of contaminant taxa vary between extraction kits and laboratories [5,19,23,24] and even through time within the same laboratory [76: Weyrich et al. in-review].

Cross-contamination is another challenge during microbiome sample processing and includes the transfer of primary sample DNA, barcodes, or amplicons from neighboring wells or tubes to create "batch effects" [39]. Cross-contamination can occur at multiple steps throughout sample processing: sample DNA can be accidentally transferred during initial sample processing and placement into tubes or plates [40], from aerosolization during pipetting, or during plate cover removal [41]. Barcode cross-contamination may also occur when incorrect neighboring barcodes 'jump' into sample wells or tubes — a phenomenon known as 'tag switching' [42]. Finally, cross-contamination can also occur on the sequencing instrument from barcode sequencing errors, residual amplicons from past sequencing runs, or "barcode hopping," where some platforms mismatch indexing reads to sequencing reads. Overall, both contaminant DNA and cross-contamination are dynamic and need to be consistently and routinely monitored.

Which sample types are most affected?

The impacts of contaminant DNA and cross-contamination can vary between samples according to their levels of microbial biomass. If one assumes an equal and stable background contaminant profile, the biomass in a sample can be determined by comparing the quantity of DNA in samples (*e.g.* quantitative PCR of 16S rRNA amplicons) to that in blank controls [23]. Samples of 'high-biomass' (*e.g.* gut or soil) will contain substantially more DNA than blank controls, while 'low-biomass' samples will contain DNA levels similar to or lower than blank controls and include glacial ice, air, rocks, the built environment, placenta, and blood. Lower levels of DNA within low-biomass samples allow contaminant DNA and cross-contamination (*e.g.* from high-biomass samples processed simultaneously) to easily outcompete and dominate the biological signal within samples [19,23,24,43].

How does contaminant DNA influence microbiome studies?

The amount and composition of contaminant DNA and cross-contamination can vary through time and location, generating signals within low-biomass samples that can be easily perceived as biological; this concept is illustrated in Figure 1. Numerous studies have described contaminant DNA and demonstrated how it can skew results, including those in published low-biomass studies [19,23,24]. For example, >95% of the taxonomic composition in a *Salmonella bongori* culture diluted to ~1,000 cell was revealed to be contamination using both amplicon and shotgun DNA sequencing [19]. Another study found that infant nasopharyngeal swabs clustered according to the DNA extraction kit lot number, demonstrating that contaminant taxa introduced during DNA extraction were driving the observed signal [19]. A comparison of low-biomass placental samples with blank controls, saliva, and vaginal swabs revealed that 16S rRNA sequences in placental samples could not be distinguished from those in blank controls [23]. Lastly, an analysis of peripheral blood and submucosal tissue samples demonstrated that 99% and 95% of the respective identified sequences corresponded to contaminant taxa [24]. The impacts of contaminant DNA and cross-contamination are not limited to these 'whistle-blower' studies and have likely impacted each and every low-biomass study published to date. Even if controls and low-biomass samples can be distinguished using beta-diversity analyses (*e.g.* a PCoA plot of unweighted UniFrac distances), measures of alpha (within-sample)

diversity are easily inflated in microbiome studies due to contaminant DNA and cross-contamination. Together, these findings demonstrate that contaminant DNA and cross-contamination can have a severe impact on low-biomass microbiota studies and will continue to pose a demonstrable threat to the integrity of the field if left unaddressed.

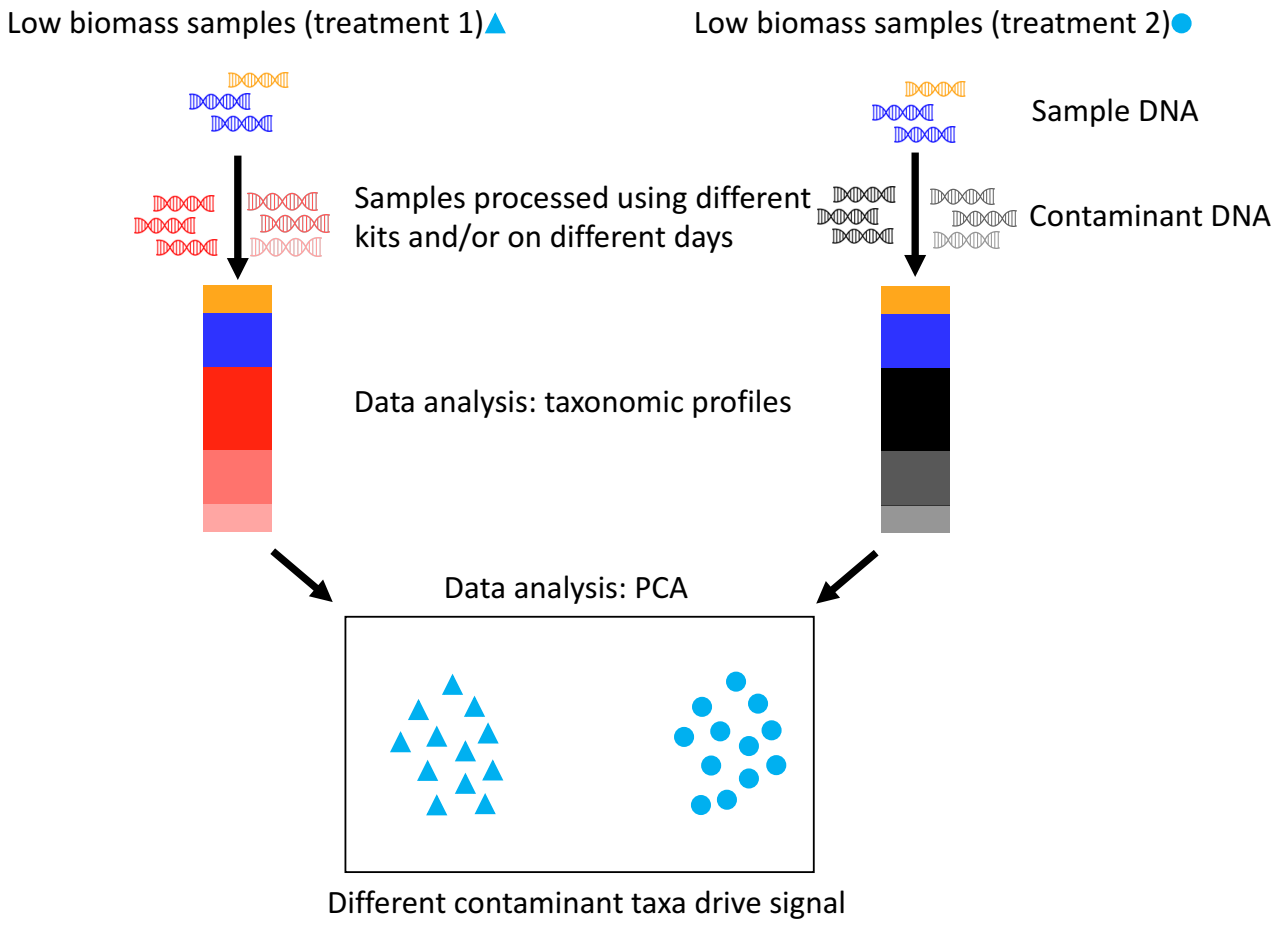


Figure 1: Illustration of how contaminant DNA can influence interpretations of low-biomass microbiome data.

Both treatment groups (triangle vs. circle) of low-biomass samples are not different in microbial composition (sample DNA colors are same, blue and orange). However, differences in contaminant DNA (in this case, red vs. black) drive the signal, leading to the conclusion that the treatment groups have different microbial compositions. Proper randomization of sample collection/processing would eliminate this artifact.

How has DNA contamination already impacted the microbiome research field?

The failure to include controls to assess DNA contaminants and cross-contamination has resulted in several controversial studies. For example, a recent study identified a distinct microbial community within human placenta without publishing appropriate controls [44]. Bacterial DNA contribution from maternal blood was raised as an issue [45], and no evidence for a distinct placental microbiota was found when placental samples were compared with blank controls in a follow-up study [23]. A recent, comprehensive review concluded that current evidence does not support the notion that the human placenta harbors a distinct microbiota [46]. Nevertheless, the initial publication [44] spurred several subsequent studies [47–50] on the 'placental microbiota'; all lacked appropriate controls and further perpetuated the notion that the placenta harbors a distinct microbiota. In addition to the placenta, there has been a recent surge of other low-biomass microbiota studies, especially in clinical medicine, and include investigations of the microbial components of brain tissue [51], breast tissue [52,53], nipple aspirate fluid [54], intrauterine samples [55], and seminal fluid [56]. None of these studies included appropriate controls or an assessment of contaminant taxa and cross-contamination in their findings. Unsurprisingly, each of these studies identified common contaminant taxa from commercial extraction kits and molecular reagents as the taxa driving the observed biological signals. In addition, the studies failed to examine the limit of detection using their methodology – the critical first step when exploring low-biomass communities. While it is possible that these are true biological signals, it is also possible that they arise from contaminant DNA, and additional experiments should be included to determine if such microbial DNA originates from living cells as opposed to contaminant DNA [57]. Together, these studies highlight the desperate need for the field to recognize and adhere to a minimum set of experimental criteria to ensure valid and reproducible findings.

How can we mitigate the impacts of contaminant DNA?

To control for contaminant DNA and cross-contamination in low-biomass microbiome studies, there are several measures that need to be taken to 1.) reduce all types of contamination and experimental bias, 2.) monitor and identify contaminant sources, and 3.) recognize and mitigate

the effects of contaminant DNA and cross-contamination during analysis (Figure 2). We briefly provide suggestions for each approach, put forth minimum guidelines, and establish the new '**RIDE**' checklist to help researchers, editors, and reviewers manage the effects of contamination in future microbiome research (**Box 1**).

Box 1: For authors, reviewers, and editors, the 'RIDE' minimum standards checklist for performing/reviewing low-biomass microbiome studies.

- ✓ **R**eport experimental design and approaches used to reduce and assess the contributions of contamination.
- ✓ **I**nclude controls to assess contaminant DNA and cross-contamination: *s*ampling blanks, DNA extraction blanks, and no-template amplification controls at a ratio of at least 1 per 12 biological samples; and 12-24 mock community positive extraction controls and positive amplification controls.
- ✓ **D**etermine the limit of detection using comparisons to controls.
- ✓ **E**xplore the contaminant taxa within each study and report its impacts on the interpretation of biological samples.

1.) Reduce experimental bias and contamination during sampling and processing.

Simple measures during sample collection and processing can be used to limit the introduction of contaminant DNA and cross-contamination and minimize their downstream effects. First, randomizing samples and treatments (*i.e.* collecting or processing samples from different treatments together) is an important experimental design consideration to prevent erroneous conclusions arising from batch effects or day-to-day variation of contaminant DNA (Figure 1). In addition, the same researcher, reagents, robots, and equipment should be used to process all of the samples in a specific study, if possible. To specifically avoid contaminant DNA, there are several key considerations. Samples should be collected in the cleanest available environment, and personnel should wear protective clothing and equipment to cover all exposed human surfaces (*i.e.* lab coats, face masks, hair nets, sleeves, and clean disposable gloves). Ideally, trained researchers with protective clothing should also process the samples in an isolated, low-contaminant, controlled environment (*e.g.* still-air cabinet or laminar-flow hood) where surfaces and equipment are treated with a ≥3% bleach solution and ultraviolet radiation to minimize environmental contaminant DNA [58]. Samples should be processed using certified sterile or DNA-free consumables, including reagents, lab ware, and sampling

equipment. As consumables labeled 'DNA free' typically contain degraded microbial DNA [36], consumables with hard surfaces, such as plastic tubes and pipettes, can be decontaminated using ethylene oxide treatment [28], and reagents can be decontaminated by UV treatment that is optimized for each reagent (*i.e.* UV irradiation can destroy enzyme function) [59]. Ideally, a physically isolated workstation should also be used to aliquot stock reagents to limit contamination [60]. To minimize cross-contamination, there are additional steps to consider. Library preparation should be performed in a separate room from DNA extraction to minimize contamination from highly-amplified products (*i.e.* pre-PCR work should be physically isolated from post-PCR work). It is also important to perform the recommended bleach and maintenance washes in the DNA sequencer between sequencing runs, as this can reduce run-to-run cross-contamination in Illumina MiSeq studies by 100-fold (from 0.01% to 0.0001%; Illumina correspondence).

Minimum guidelines: Different sample types or treatments should be randomized and not processed independently. Researchers should wear disposable lab gloves, face masks, and avoid exposed skin to reduce the introduction of contaminant DNA into the samples. As many procedures as possible (*e.g.* sample transfer, DNA extraction, library preparation, and sequencing) should be performed in a cleaned, isolated working environment with appropriately treated equipment and consumables.

2.) Include controls from sampling to sequencing.

Several types of controls should be included in every analysis to monitor contaminant DNA and assess the levels of cross-contamination between samples. These controls include both negative controls to monitor background levels of contaminant DNA: (1) sampling blank controls, (2) DNA extraction blank controls, and (3) no-template amplification negative controls; and two types of positive controls to determine the limit of detection and ensure cross-contamination does not drive the results of the study: (4) DNA extraction positive controls and (5) amplification positive controls.

Negative controls

Three types of negative controls are minimally required to allow adequate monitoring of contaminants throughout sample handling and processing and provide the ability to detect when and how contaminants are introduced into biological samples. Each type of negative control

should be included at a minimum rate of 1:12, control to biological samples; for larger studies (>100 samples), 8 of each negative control type should be minimally required per study. (1) **Sampling blank controls** allow for detection of contaminant DNA introduced during the sampling procedure, including items used to collect the sample, such as swabs, gauze, or drills, and any reagents or preservatives used to store or transport the samples (*e.g.* media, alcohol, or RNA stabilizer). Material analyzed in sampling blanks should be collected in the same room and at the same time as biological samples and should undergo the same laboratory treatment as the biological samples, from collection to sequencing. (2) **DNA extraction blank controls** monitor the contaminant DNA content in extraction kits, molecular reagents, and the laboratory environment through the DNA extraction process and, as above, should be processed alongside the biological samples from extraction to sequencing. (3) **No-template amplification controls** can monitor contaminant DNA present in reagents and the laboratory environment during library preparation and sequencing. All negative controls provide a semi-quantitative estimate of background contaminants and allow researchers to identify contaminants that can be used in downstream subtractive analyses. Finally, it should be noted that blank controls can contain too little DNA to be effectively processed. In these cases, the use of known carrier DNA in blank controls can help to efficiently amplify contaminants [61].

Positive Controls

Two types of positive controls should be included to determine the limit of detection and provide insight into the effects of cross-contamination during extraction, library preparation, and sequencing. Both positive control types should be included at a minimum rate of 1:12, control to biological samples; in larger studies, at least 12 total positive controls should be included per study. (4) **DNA extraction positive controls** monitor DNA extraction efficiency, determine the level of detection, and examine levels of cross-contamination during DNA extraction. To include a DNA extraction positive control, a serial dilution of a known cell type(s) (*e.g.* 1, 10, 100, and 1000 cells) should be extracted alongside samples and span the expected limit of detection of the assay (see Kathroseq below) [62]. Ideally, researchers should use a commercially available mixed community, such as the Zymo mock community (Zymo, D6300), as this enables standardization across different laboratories. Researchers can also consider including a range of positive titration spike-ins into liquid samples, such as blood, urine, or mucus, to evaluate the efficiency of extraction and the limit of detection, which is important as many sample types have inhibitors or chemicals that can increase the limit of detection. The bottom line is to use a positive control of known concentration that is relevant to your study and experimental questions. (5) The last recommended control is a **positive**

amplification control, which is again a titration of DNA from known organism type(s) to be processed solely during the library preparation stage. This control enables a detection limit to be established for library preparation. Critically, both positive control types can be used to calculate the limit of detection within the laboratory techniques used and the levels of cross-contamination using novel bioinformatic approaches [62]. For example, Katharoseq utilizes differences in amplification efficiencies of true positives compared to negatives to mathematically determine a limit of detection by calculating cutoff scores to guide sample exclusion. In doing so, cross-contamination can also be evaluated, as positive controls from DNA extractions should be different from those used in library preparation.

Controls samples often produce libraries of lower quantity and quality, but this should not prevent the control samples from being sequencing. Libraries should be quantified (*i.e.* using a PicoGreen or Qubit assay for amplicon studies or a TapeStation or BioAnalyzer for shotgun sequencing) and pooled at equal molarity (*e.g.* X ng per observed fragment lengths per sample). If amplified control samples contain significantly lower amounts of DNA compared to biological samples, they should be included in sequencing pools by pooling the controls at a certain maximum volume (*e.g.* 20 µl of each control). In addition, amplified biological samples with low amounts of DNA can be pooled at this same maximum volume as controls (*e.g.* 20 µl) [62]. Alternatively, all samples and controls can be pooled at equal volumes; however, this approach requires deeper sequencing because the higher-biomass samples will dominate the DNA sequencing effort. For highly contentious sample types and claims (*e.g.* placenta), reproducibility across labs is highly recommended.

Minimum guidelines: On average, all negative controls must be included at a ratio of 1:12, control to samples for smaller studies, or a minimum of 8 negative controls for studies with >100 samples. All positive controls should be included at a ratio of 1:12, control to samples, for smaller studies, or a minimum of 12 positive controls for studies that contain >100 samples. Controls must be processed alongside samples to account for contamination and should not be processed separately.

3.) Critically assess and report contributions of contamination during analysis.

The impacts of contaminant taxa must be assessed in the final analysis and interpretation of the data. Three different strategies currently exist to assess the impacts of contamination in microbiome datasets: (1) compare controls to biological samples; (2) filter contaminants from

biological samples; and (3) use predictive modeling to identify putative contaminants. Each method varies in its stringency and application.

(1) Comparisons of biological samples to the controls can be used to assess the limit of detection and the variety of contaminant taxa. The first step is to assess the limit of detection in the DNA extraction and library preparation methodologies [62]. For example, the limit of detection for an amplicon study can be assessed by comparing the number of DNA sequences in positive and negative controls to that in the biological samples to calculate a sample exclusion value (*e.g.* $K_{1/2}$ value). Samples with fewer reads than the exclusion value should be discarded [62]. Once a limit of detection is established, researchers can then compare the taxa identified in the negative controls (sampling blank, extraction blank, and no-template controls) with those from the biological samples, as this provides a list of taxa within biological samples that may have arisen from contaminant DNA. This is especially important to ensure that the significant differences in taxa abundances between sample types or treatments are not driven by contaminant taxa. If negative controls were not included in the study, taxa that drive differences between samples or treatments should be individually scrutinized and compared to previously published contaminant taxa lists (Table 1). We provide the largest, current list of contaminant taxa shared across multiple 16S rRNA-based studies within this publication for this purpose (Table 1).

(2) A thorough, yet conservative approach can be taken to filter contaminant taxa from biological samples, removing all taxa found within negative controls from the biological samples or all contaminant taxa from reported sources (Table 1). While an extremely conservative approach, it effectively eliminates the effects of contaminant taxa and most cross-contamination from the downstream data set. However, DNA present in controls can also arise from biological samples via cross-contamination, so this analysis will also likely remove some biological signal that corresponds to the highest biomass organisms. While unlikely and typically rare, it is also possible that contaminant taxa could be truly present in a biological specimen (*e.g.* in a rare infection in an immunocompromised hospital patient), but this would need to be verified using a different approach, such as Fluorescent In-Situ Hybridization (FISH) [63,64]. Filtering contaminants from data sets can be a useful tool for rapid, preliminary assessment or in specific scenarios that require a very conservative evaluation.

(3) Bioinformatic modeling has been developed to estimate the source and proportions of contaminant taxa within biological samples. For example, SourceTracker analysis uses

Bayesian modeling to estimate the proportion of potential contaminant taxa from a data set [65]. To do this, the blank controls can serve as contaminant 'sources' and the biological samples as 'sinks' to estimate the origin and abundance of contaminant taxa within biological samples. Subsequently, the relative contributions of contaminant DNA within the samples can be factored into downstream analysis and data interpretation. However, it should be stressed that sufficient cross-contamination can confound SourceTracker analysis.

Minimum guidelines: Biological samples must be compared to blank controls or known contaminant resources, and taxa identified as contaminants and biological samples must be reported. The approach taken to identify and minimize the effects of contaminant DNA during analysis should be clearly reported to enhance reproducibility and allow such approaches to be critically evaluated by others.

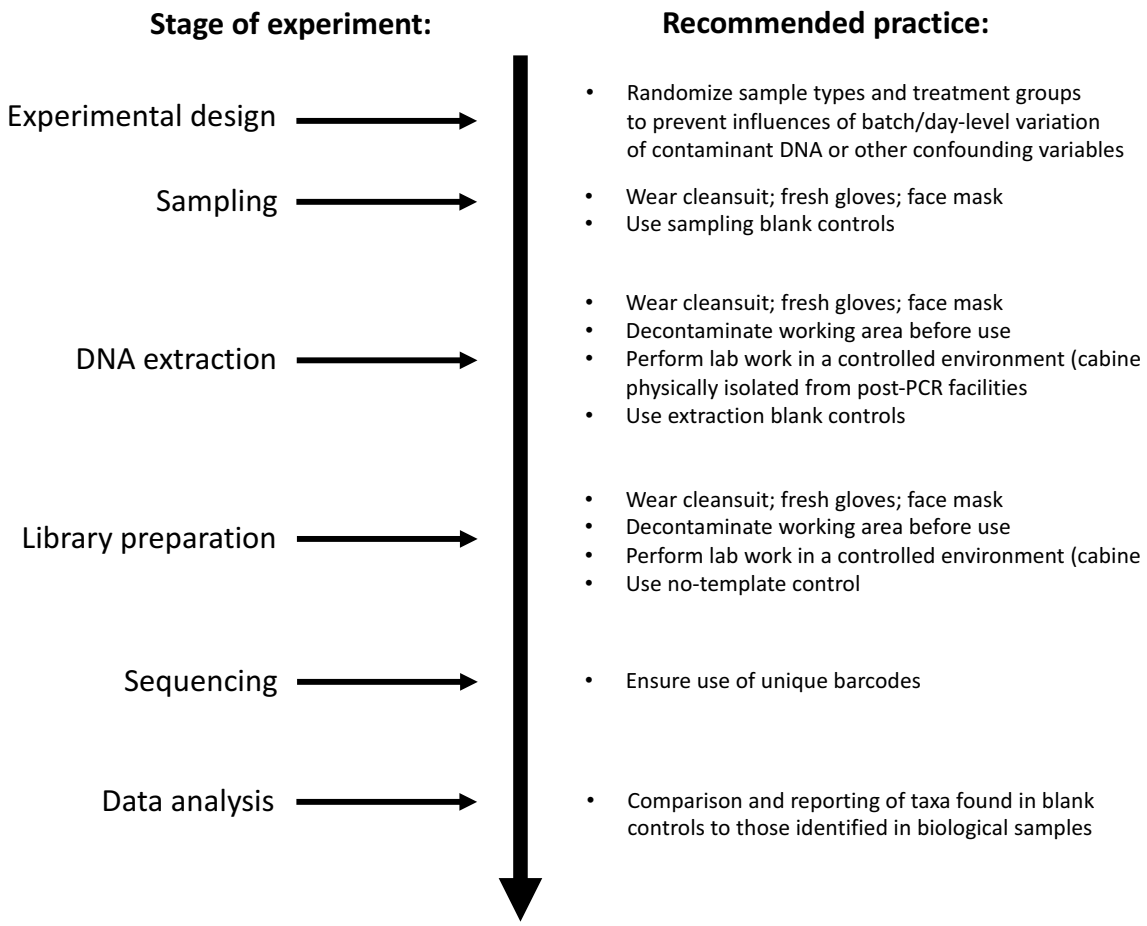


Figure 2: Flowchart of methods to minimize influence of contaminant DNA in low-biomass samples.

Conclusions

Microbiome research holds great promise for multiple fields, but methodological pitfalls can easily undermine the progress and reputation of this developing research area. Therefore, these pitfalls must be recognized and explicitly addressed at each phase of the scientific process by researchers, reviewers, and editors alike. Here, we present the '**RIDE**' checklist for contaminant assessment to be applied across a wide-range of disciplines interested in exploring the microbial communities in low-biomass samples (see **Box 1** for our '**RIDE**' minimum standards checklist). Failure to take these caveats into account is likely to waste valuable time and money and erode the credibility of microbiome research. The current situation is similar in many ways to the methodological issues in ancient DNA research recognized over 20 years ago. A series of high-profile publications based on PCR amplification of short sequences were used to support remarkable findings, including the reported recovery of DNA more than 40 million years old [66–68] – well beyond the theoretical limit of DNA survival of around one million years [69]. Although these findings were heavily criticized by other ancient DNA researchers [70–74] and are now recognized as erroneous, these publications nevertheless damaged the credibility of the ancient DNA field. As a direct result, a set of ancient DNA authentication criteria was formulated and widely adopted [60]. These standards, improved techniques, and greater attention to the issue of contaminant DNA dramatically improved the credibility of ancient DNA research. In microbiome research, similar standards need to be established to improve scientific integrity and secure the credibility of such research. It is important to note that the minimum set of guidelines and the 'RIDE' checklist that we propose (**Box 1**) will not guarantee that all contamination can be accounted for or removed, nor will it provide a solution for every contaminant problem. New methodologies will likely only improve our ability to detect and quantify contamination in low-biomass samples. As new methods and analyses for microbiome analysis are also developed, novel solutions to account for contaminant DNA and cross-contamination will need also to be established. In the meantime, it is imperative that low-biomass research generates sufficient control data and that researchers develop and maintain a critical mindset when dealing with low-biomass microbiome samples. In this regard, we hope that the guidelines introduced in this article will help authors, reviewers, and editors monitor and protect the future of the microbiome field.

Declarations

Acknowledgments

We would like to thank Alan W. Walker, Bastien Llamas, Jessica L. Metcalf, Kieren J. Mitchell, and Matilda Handsley-Davis for their feedback and suggestions.

Funding

LSW and RE were funded from the Australian Research Council grants: DECRA (DE150101574) and ARC Centre of Excellence CABAH (CE170100015).

Author contributions

RE, JJM, and LSW wrote the paper. All authors read, edited, and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Ethics approval

Not applicable.

Availability of data and material

Not applicable.

Consent for publication

Not applicable.

Glossary

Contamination: An umbrella term encompassing both contaminant DNA and cross-contamination (see below).

Contaminant DNA: DNA from sources other than the sample(s) under study (*e.g* DNA from reagents or researchers performing laboratory work).

Cross-contamination: DNA exchange between samples within a study (*e.g.* accidental movement of DNA between different sample tubes during DNA extraction).

DNA extraction blank control: A negative control consisting of an empty tube/well that is processed alongside biological samples during DNA extraction and allows for the detection of contaminant DNA introduced during DNA extraction.

DNA extraction positive control: A positive control consisting of serially diluted cells of known type(s) that is processed alongside biological samples during DNA extraction and allows for determination of the limit of detection, monitoring of extraction efficiency, and quantification of cross-contamination during DNA extraction.

Microbiome: The microorganisms of a specific habitat, their genomes, and the surrounding environmental conditions [75].

Microbiota: The assemblage of microorganisms present in a defined environment [75].

No-template amplification control: A negative control made by preparing an amplification or library preparation reaction without input template (*i.e.* sample DNA) that is processed alongside biological samples and allows for the detection contaminant DNA during library preparation/PCR amplification.

Positive amplification control: A positive control consisting of serially diluted DNA from known organism type(s) that are processed alongside biological samples during amplification or library preparation and allows for determination of the limit of detection, monitoring of

library preparation efficiency, and quantification of cross-contamination during library preparation.

RIDE: Report methodology, Include controls, Determine the limit of detection, and Explore the impacts of contamination in downstream analysis. Minimum standards checklist for low-biomass microbiome studies.

Sampling blank control: A negative control consisting of an empty tube that is processed alongside the collection of biological samples. Allows for the detection of contaminant DNA introduced during the sampling procedure (*e.g.* airborne, swabs, preservatives).

Table 1: Contaminant taxa previously identified in negative controls

Taxon	Study found
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Actinomycetaceae;g__Actinomyces	23,24,76
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Corynebacteriaceae;g__Corynebacterium	19,24,
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Arthrobacter	19,24,
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Micrococcaceae;g__Rothia	23,24,
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Actinomycetales;f__Propionibacteriaceae;g__Propionibacterium	23,19,24,
Bacteria;p__Actinobacteria;c__Actinobacteria;o__Coriobacteriales;f__Coriobacteriaceae;g__Atopobium	23,24,
Bacteria;p__Bacteroidetes;c__[Saprospirae];o__[Saprospirales];f__Chitinophagaceae;g__Sediminibacterium	23,76
Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Porphyromonadaceae;g__Porphyromonas	23,24,
Bacteria;p__Bacteroidetes;c__Bacteroidia;o__Bacteroidales;f__Prevotellaceae;g__Prevotella	23,24,
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__[Weeksellaceae];g__Chryseobacterium	19,76
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__Capnocytophaga	23,24,
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__Chryseobacterium	19,24,
Bacteria;p__Bacteroidetes;c__Flavobacteriia;o__Flavobacteriales;f__Flavobacteriaceae;g__Flavobacterium	23,19,21,76
Bacteria;p__Bacteroidetes;c__Sphingobacteriia;o__Sphingobacteriales;f__Sphingobacteriaceae;g__Pedobacter	19,76
Bacteria;p__Candidate Phylum TM7;c__unclassifiedTM7;o__unclassifiedTM7;f__unclassifiedTM7;g__unclassifiedTM7	23,24,
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Bacillus	19,24,76
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Bacillaceae;g__Geobacillus	24,76
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Paenibacillaceae;g__Brevibacillus	19,24,
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Paenibacillaceae;g__Paenibacillus	19,24,76
Bacteria;p__Firmicutes;c__Bacilli;o__Bacillales;f__Staphylococcaceae;g__Staphylococcus	24,76
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Aerococcaceae;g__Abiotrophia	19,24,
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Carnobacteriaceae;g__Granulicatella	23,24,
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Enterococcaceae;g__Enterococcus	23,24,76
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Lactobacillaceae;g__Lactobacillus	23,24,76
Bacteria;p__Firmicutes;c__Bacilli;o__Lactobacillales;f__Streptococcaceae;g__Streptococcus	23,19,24,76
Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Clostridiaceae;g__Clostridium	24,76
Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Lachnospiraceae;g__Coprococcus	23,24,
Bacteria;p__Firmicutes;c__Clostridia;o__Clostridiales;f__Peptoniphilaceae;g__Anaerococcus	23,24,
Bacteria;p__Firmicutes;c__Negativicutes;o__Selenomonadales;f__Veillonellaceae;g__Dialister	23,24,
Bacteria;p__Firmicutes;c__Negativicutes;o__Selenomonadales;f__Veillonellaceae;g__Megasphaera	23,24,
Bacteria;p__Firmicutes;c__Negativicutes;o__Selenomonadales;f__Veillonellaceae;g__Veillonella	23,24,
Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Fusobacteriaceae;g__Fusobacterium	23,24,
Bacteria;p__Fusobacteria;c__Fusobacteriia;o__Fusobacteriales;f__Leptotrichiaceae;g__Leptotrichia	23,24,
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae;g__Afipia	19,24,
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Bradyrhizobiaceae;g__Bradyrhizobium	19,24,21,76
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Hyphomicrobiaceae;g__Devosia	19,76
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Methylobacteriaceae;g__Methylobacterium	23,19,18,76
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Phyllobacteriaceae;g__Mesorhizobium	19,76
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhizobiales;f__Phyllobacteriaceae;g__Phyllobacterium	19,24,
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Methylobacteriaceae;g__Methylobacterium	19,24,
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodobacterales;f__Phyllobacteriaceae;g__Phyllobacterium	19,24,
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rhodospirillales;f__Acetobacteraceae;g__Roseomonas	19,24,
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Novosphingobium	19,76
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingobium	19,76
Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Sphingomonadales;f__Sphingomonadaceae;g__Sphingomonas	19,21,18,76
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Alcaligenaceae;g__Achromobacter	21,76
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Burkholderiaceae;g__Burkholderia	19,24,76
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Comamonas	19,24,18,76
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Curvibacter	19,24,
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Comamonadaceae;g__Pelomonas	19,24,
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Cupriavidus	19,18,76
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Herbaspirillum	19,24,
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Janthinobacterium	19,24,
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Massilia	19,24,
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Oxalobacter	19,24,
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Burkholderiales;f__Oxalobacteraceae;g__Ralstonia	19,21,18,17,76
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__kingella	19,24,
Bacteria;p__Proteobacteria;c__Betaproteobacteria;o__Neisseriales;f__Neisseriaceae;g__Neisseria	23,24,
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Enterobacteriales;f__Enterobacteriaceae;g__Escherichia	19,24,
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pasteurellales;f__Pasteurellaceae;g__Haemophilus	23,24,
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Acinetobacter	23,19,18,16,76
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Moraxellaceae;g__Enhydrobacter	19,24,76
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Pseudomonadales;f__Pseudomonadaceae;g__Pseudomonas	19,24,21,17,76
Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Xanthomonadales;f__Xanthomonadaceae;g__Stenotrophomonas	19,24,21,18,17,16,76

References

1. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.

2. Kassam Z, Lee CH, Yuan Y, Hunt RH. Fecal Microbiota Transplantation for Clostridium difficile Infection: Systematic Review and Meta-Analysis. Am J Gastroenterol. 2013;108:500–8.

3. Bahrndorff S, Alemu T, Alemneh T, Lund Nielsen J. The Microbiome of Animals: Implications for Conservation Biology [Internet]. Int. J. Genomics. 2016 [cited 2017 Jul 27]. Available from: https://www.hindawi.com/journals/ijg/2016/5304028/

4. Sessitsch A, Mitter B. 21st century agriculture: integration of plant microbiomes for improved crop production and food security. Microb Biotechnol. 2015;8:32–3.

5. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature. 2017;544:357–61.

6. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Forensic identification using skin bacterial communities. Proc Natl Acad Sci U S A. 2010;107:6477–81.

7. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6:1621–1624.

8. Consortium THMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

9. Christner BC, Priscu JC, Achberger AM, Barbante C, Carter SP, Christianson K, et al. A microbial ecosystem beneath the West Antarctic ice sheet. Nature. 2014;512:310–3.

10. Checinska A, Probst AJ, Vaishampayan P, White JR, Kumar D, Stepanov VG, et al. Microbiomes of the dust particles collected from the International Space Station and Spacecraft Assembly Facilities. Microbiome. 2015;3:50.

11. Anantharaman K, Breier JA, Dick GJ. Metagenomic resolution of microbial functions in deep-sea hydrothermal plumes across the Eastern Lau Spreading Center. ISME J. 2016;10:225–39.

12. Kelly LW, Williams GJ, Barott KL, Carlson CA, Dinsdale EA, Edwards RA, et al. Local genomic adaptation of coral reef-associated microbiomes to gradients of natural variability and anthropogenic stressors. Proc Natl Acad Sci. 2014;111:10227–32.

13. Lozupone C, Knight R. UniFrac: a New Phylogenetic Method for Comparing Microbial Communities. Appl Environ Microbiol. 2005;71:8228–35.

14. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, et al. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 2011;12:R18.
15. Gagic D, Maclean PH, Li D, Attwood GT, Moon CD. Improving the genetic representation of rare taxa within complex microbial communities using DNA normalization methods. Mol Ecol Resour. 2014;n/a-n/a.
16. Tanner MA, Goebel BM, Dojka MA, Pace NR. Specific Ribosomal DNA Sequences from Diverse Environmental Settings Correlate with Experimental Contaminants. Appl Environ Microbiol. 1998;64:3110–3.
17. Grahn N, Olofsson M, Ellnebo-Svedlund K, Monstein HJ, Jonasson J. Identification of mixed bacterial DNA contamination in broad-range PCR amplification of 16S rDNA V1 and V3 variable regions by pyrosequencing of cloned amplicons. FEMS Microbiol Lett. 2003;219:87–91.
18. Barton HA, Taylor NM, Lubbers BR, Pemberton AC. DNA extraction from low-biomass carbonate rock: An improved method with reduced contamination and the low-biomass contaminant database. J Microbiol Methods. 2006;66:21–31.
19. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
20. Lusk RW. Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data. PLOS ONE. 2014;9:e110808.
21. Laurence M, Hatzis C, Brash DE. Common Contaminants in Next-Generation Sequencing That Hinder Discovery of Low-Abundance Microbes. PLOS ONE. 2014;9:e97876.
22. Adams RI, Bateman AC, Bik HM, Meadow JF. Microbiota of the indoor environment: a meta-analysis. Microbiome. 2015;3:49.
23. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome. 2016;4:29.
24. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8:24.
25. Willerslev E, Hansen AJ, Poinar HN. Isolation of nucleic acids and cultures from fossil ice and permafrost. Trends Ecol Evol. 2004;19:141–7.

26. Witt N, Rodger G, Vandesompele J, Benes V, Zumla A, Rook GA, et al. An Assessment of Air As a Source of DNA Contamination Encountered When Performing PCR. J Biomol Tech JBT. 2009;20:236–40.
27. Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W. From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR Sci Technol Archaeol Res. 2017;3:1–14.
28. Motley ST, Picuri JM, Crowder CD, Minich JJ, Hofstadler SA, Eshoo MW. Improved multiple displacement amplification (iMDA) and ultraclean reagents. BMC Genomics. 2014;15:443.
29. Fierer N, Hamady M, Lauber CL, Knight R. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. Proc Natl Acad Sci. 2008;105:17994–9.
30. Dunn RR, Fierer N, Henley JB, Leff JW, Menninger HL. Home Life: Factors Structuring the Bacterial Diversity Found within and between Homes. PLOS ONE. 2013;8:e64133.
31. Naccache SN, Greninger AL, Lee D, Coffey LL, Phan T, Rein-Weston A, et al. The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. J Virol. 2013;87:11966–77.
32. Adams RI, Miletto M, Lindow SE, Taylor JW, Bruns TD. Airborne Bacterial Communities in Residences: Similarities and Differences with Fungi. PLOS ONE. 2014;9:e91283.
33. McFeters GA, Broadaway SC, Pyle BH, Egozy Y. Distribution of bacteria within operating laboratory water purification systems. Appl Environ Microbiol. 1993;59:1410–5.
34. Nogami T, Ohto T, Kawaguchi O, Zaitsu Y, Sasaki S. Estimation of bacterial contamination in ultrapure water: application of the anti-DNA antibody. Anal Chem. 1998;70:5296–301.
35. McAlister MB, Kulakov LA, O'Hanlon JF, Larkin MJ, Ogden KL. Survival and nutritional requirements of three bacteria isolated from ultrapure water. J Ind Microbiol Biotechnol. 2002;29:75–82.
36. Shen H, Rogelj S, Kieft TL. Sensitive, real-time PCR detects low-levels of contamination by Legionella pneumophila in commercial reagents. Mol Cell Probes. 2006;20:147–53.
37. Seitz V, Schaper S, Dröge A, Lenze D, Hummel M, Hennig S. A new method to prevent carry-over contaminations in two-step PCR NGS library preparations. Nucleic Acids Res. 2015;43:e135.
38. Ballenghien M, Faivre N, Galtier N. Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol. 2017;15:25.
39. Nguyen NH, Smith D, Peay K, Kennedy P. Parsing ecological signal from noise in next generation amplicon sequencing. New Phytol. 2015;205:1389–93.

40. Tamariz J, Voynarovska K, Prinz M, Caragine T. The application of ultraviolet irradiation to exogenous sources of DNA in plasticware and water for the amplification of low copy number DNA. J Forensic Sci. 2006;51:790–4.
41. Joung YS, Ge Z, Buie CR. Bioaerosol generation by raindrops on soil. Nat Commun. 2017;8:14668.
42. Carlsen T, Aas AB, Lindner D, Vrålstad T, Schumacher T, Kauserud H. Don't make a mista(g)ke: is tag switching an overlooked source of error in amplicon pyrosequencing studies? Fungal Ecol. 2012;5:747–9.
43. Eisenhofer R, Cooper A, Weyrich LS. Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential. FEMS Microbiol Ecol [Internet]. 2017 [cited 2017 Jun 27];93. Available from: https://academic.oup.com/femsec/article/93/5/fix042/3089752/Reply-to-Santiago-Rodriguez-et-al-proper
44. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. Sci Transl Med. 2014;6:237ra65.
45. Kliman HJ. Comment on "The placenta harbors a unique microbiome." Sci Transl Med. 2014;6:254le4-254le4.
46. Perez-Muñoz ME, Arrieta M-C, Ramer-Tait AE, Walter J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. Microbiome. 2017;5:48.
47. Antony KM, Ma J, Mitchell KB, Racusin DA, Versalovic J, Aagaard K. The preterm placental microbiome varies in association with excess maternal gestational weight gain. Am J Obstet Gynecol. 2015;212:653.e1-653.e16.
48. Zheng J, Xiao X, Zhang Q, Mao L, Yu M, Xu J. The Placental Microbiome Varies in Association with Low Birth Weight in Full-Term Neonates. Nutrients. 2015;7:6924–37.
49. Amarasekara R, Jayasekara RW, Senanayake H, Dissanayake VHW. Microbiome of the placenta in pre-eclampsia supports the role of bacteria in the multifactorial cause of pre-eclampsia. J Obstet Gynaecol Res. 2015;41:662–9.
50. Bassols J, Serino M, Carreras-Badosa G, Burcelin R, Blasco-Baque V, Lopez-Bermejo A, et al. Gestational diabetes is associated with changes in placental microbiota and microbiome. Pediatr Res. 2016;80:777–84.
51. Branton WG, Ellestad KK, Maingat F, Wheatley BM, Rud E, Warren RL, et al. Brain Microbial Populations in HIV/AIDS: α-Proteobacteria Predominate Independent of Host Immune Status. PLOS ONE. 2013;8:e54673.

52. Xuan C, Shamonki JM, Chung A, DiNome ML, Chung M, Sieling PA, et al. Microbial Dysbiosis Is Associated with Human Breast Cancer. PLOS ONE. 2014;9:e83744.
53. Hieken TJ, Chen J, Hoskin TL, Walther-Antonio M, Johnson S, Ramaker S, et al. The Microbiome of Aseptically Collected Human Breast Tissue in Benign and Malignant Disease. Sci Rep. 2016;6:30751.
54. Chan AA, Bashir M, Rivas MN, Duvall K, Sieling PA, Pieber TR, et al. Characterization of the microbiome of nipple aspirate fluid of breast cancer survivors. Sci Rep. 2016;6:28061.
55. Fang R-L, Chen L-X, Shu W-S, Yao S-Z, Wang S-W, Chen Y-Q. Barcoded sequencing reveals diverse intrauterine microbiomes in patients suffering with endometrial polyps. Am J Transl Res. 2016;8:1581–92.
56. Javurek AB, Spollen WG, Ali AMM, Johnson SA, Lubahn DB, Bivens NJ, et al. Discovery of a Novel Seminal Fluid Microbiome and Influence of Estrogen Receptor Alpha Genetic Status. Sci Rep [Internet]. 2016 [cited 2016 Nov 23];6. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4789797/
57. Vaishampayan P, Probst AJ, La Duc MT, Bargoma E, Benardini JN, Andersen GL, et al. New perspectives on viable microbial communities in low-biomass cleanroom environments. ISME J. 2013;7:312–24.
58. Champlot S, Berthelot C, Pruvost M, Bennett EA, Grange T, Geigl E-M. An Efficient Multistrategy DNA Decontamination Procedure of PCR Reagents for Hypersensitive PCR Applications. PLoS ONE. 2010;5:e13042.
59. Woyke T, Sczyrba A, Lee J, Rinke C, Tighe D, Clingenpeel S, et al. Decontamination of MDA reagents for single cell whole genome amplification. PloS One. 2011;6:e26161.
60. Cooper A, Poinar HN. Ancient DNA: Do It Right or Not at All. Science. 2000;289:1139–1139.
61. Xu Z, Zhang F, Xu B, Tan J, Li S, Jin L. Improving the sensitivity of negative controls in ancient DNA extractions. ELECTROPHORESIS. 2009;30:1282–5.
62. Minich JJ, Zhu Q, Janssen S, Hendrickson R, Amir A, Vetter R, et al. KatharoSeq Enables High-Throughput Microbiome Analysis from Low-Biomass Samples. mSystems. 2018;3:e00218-17.
63. Russell JH, Keiler KC. RNA Visualization in Bacteria by Fluorescence In Situ Hybridization. Bact Regul RNA [Internet]. Humana Press, Totowa, NJ; 2012 [cited 2018 Feb 20]. p. 87–95. Available from: https://link.springer.com/protocol/10.1007/978-1-61779-949-5_7

64. Kostic AD, Gevers D, Pedamallu CS, Michaud M, Duke F, Earl AM, et al. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Res. 2012;22:292–8.

65. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of experimental contamination in microbiome studies. Genome Biol. 2014;15:564.

66. Cano RJ, Poinar HN, Pieniazek NJ, Acra A, Poinar GO. Amplification and sequencing of DNA from a 120–135-million-year-old weevil. Nature. 1993;363:536–8.

67. Woodward, Weyand NJ, Bunnell M. DNA sequence from Cretaceous period bone fragments. Science. 1994;266:1229–32.

68. Cano RJ, Borucki MK. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science. 1995;268:1060–4.

69. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc R Soc Lond B Biol Sci. 2012;279:4724–33.

70. Austin JJ, Ross AJ, Smith AB, Fortey RA, Thomas RH. Problems of reproducibility – does geologically ancient DNA survive in amber–preserved insects? Proc R Soc Lond B Biol Sci. 1997;264:467–74.

71. Hedges SB, Schweitzer MH. Detecting dinosaur DNA. Science. 1995;268:1191–2.

72. Henikoff S. Detecting dinosaur DNA. Science. 1995;268:1192–1192.

73. Zischler H, Hoss M, Handt O, Haeseler A von, Kuyl A van der, Goudsmit J. Detecting dinosaur DNA. Science. 1995;268:1192–3.

74. Yousten AA, Rippere KE. DNA similarity analysis of a putative ancient bacterial isolate obtained from amber. FEMS Microbiol Lett. 1997;152:345–7.

75. Marchesi JR, Ravel J. The vocabulary of microbiome research: a proposal. Microbiome. 2015;3:31.

76. Weyrich L. S. *et al.* Laboratory contamination over time during low-biomass sample analysis. *In Review*.

Chapter III

—

Assessing alignment-based taxonomic classification of ancient microbial DNA

Statement of Authorship

Title of Paper	Assessing alignment-based taxonomic classification of ancient microbial DNA
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted work written in manuscript style
Publication Details	In preparation for submission to *Molecular Ecology Resources*.

Principal Author

Name of Principal Author (Candidate)	Raphael Eisenhofer
Contribution to the Paper	Designed the experiment. Performed all bioinformatic analyses. Wrote the manuscript.
Overall percentage (%)	85%
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 10/5/18

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	Laura S. Weyrich
Contribution to the Paper	funding, design, editing
Signature	Date 9/5/18

Assessing alignment-based taxonomic classification of ancient microbial DNA

Authors: Raphael Eisenhofer[1] & Laura S Weyrich[1]

Affiliations: [1]Australian Centre for Ancient DNA, University of Adelaide, Australia

Abstract

The field of palaeomicrobiology—the study of ancient microorganisms—is rapidly growing due to recent methodological and technological advancements. It is now possible to obtain vast quantities of DNA data from ancient specimens in a high-throughput manner and use this information to investigate the dynamics of past microbial communities. However, knowledge is currently limited about how the characteristics of ancient DNA influence our ability to assign taxonomy (*i.e.* determine who's there) in ancient metagenomic samples. Here, using both simulated and published metagenomic data, we investigate how ancient DNA characteristics affect alignment-based taxonomic classification. We find that nucleotide-to-protein alignments are currently unsuitable for ancient metagenomic data as they are unable to assign reads shorter than 60 base pairs, which is the typical length of ancient DNA. We determine that deamination (a form of ancient DNA damage) and random sequence substitutions corresponding to 100,000 years of evolution minimally impact alignment-based classification. We also test four different reference databases and find that database choice is an important factor to consider for alignment-based taxonomic classification. Finally, we perform a reanalysis of previously published ancient dental calculus data, increasing the number of reads assigned taxonomy by an average of 64.2-fold, and identifying taxa previously unidentified in the study. Overall, this study enhances our understanding and ability to assign taxonomy to ancient microorganisms and provides recommendations for future palaeomicrobiological studies.

Introduction

Palaeomicrobiology—the study of ancient microorganisms—is a rapidly growing field of research. Like modern microbiology [1,2], palaeomicrobiology has witnessed a renaissance with the development of high-throughput sequencing technology [3,4]. The study of ancient microorganisms has the potential to shed light on a range of topics, such as the evolution of the human microbiota [5,6], adaptation and spread of ancient pathogens [7–9], the reconstruction of human migrations and interactions [10–12], and climate change [13].

Palaeomicrobiology is especially challenging because ancient DNA is fragmented, damaged, and mixed with the DNA of contaminant microorganisms. Ancient DNA is highly fragmented due to the post-mortem cessation of DNA repair, resulting in short fragments of lengths typically shorter than 100 bp [14,15]. These short fragments are also subjected to chemical modifications (*e.g.* deamination), which yields an increased rate of observed cytosine to thymine, and guanine to adenine substitutions at the 5' and 3' ends of the sequenced DNA molecules, respectively [15]. Finally, contamination of ancient DNA with modern microbial DNA is a serious issue which must be mitigated with expensive ultra-clean laboratories, rigorous decontamination, and the extensive use of extraction blank and no-template negative controls [16–18]. Collectively, these factors influence the choice of molecular techniques [19] and bioinformatic tools used for taxonomic classification of ancient microbial DNA [6].

Identifying the microbial species present within an ancient sample, *i.e.* taxonomic classification, is a standard first step in palaeomicrobiology studies [6]. Initially, targeted amplification of the 16S ribosomal RNA encoding gene was used to discover which microbes were present in ancient samples [5], as is routinely done in modern microbiota studies seeking to characterize microbial communities [1,20]. However, these targeted regions are often longer than the typical fragment length of ancient DNA and can contain length polymorphisms which bias the taxonomic reconstruction of ancient metagenomes [19]. Considering these findings, the palaeomicrobiology field has converged on shotgun sequencing as the best-practice approach to reproducibly identify microbial species within ancient samples. While more expensive than the targeted PCR approach, shotgun sequencing also provides genomic and functional information which can be used to reconstruct ancient microbial genomes, observe functional changes through time, and identify non-prokaryotic information within samples [6,9].

Methods for analysing shotgun sequencing data broadly fall into two categories: assembly-based and alignment-based. Assembly-based techniques involve merging overlapping DNA fragments into longer sequences, with the goal of assembling whole genomes. Such techniques have been successful in generating new genomes from modern metagenomic samples [21,22]. However, the short, damaged nature of ancient DNA renders assembly-based techniques intractable for palaeomicrobiology. Alignment-based techniques involve the alignment of DNA fragments to previously characterized reference sequences using alignment algorithms such as Bowtie2 or the Burrows-Wheeler Aligner (BWA) [23,24]. Commonly used alignment-based methods include: MetaPhlAn [25], MG-RAST [26], DIAMOND [27], and MALT (MEGAN alignment tool) [28]. A recent study benchmarked these tools and found that MALT performed better for short, fragmented DNA [6]. MALT is

an alignment-based tool which allows researchers to query DNA sequences against reference databases using a method similar to BLAST (Basic Local Alignment Search Tool) [29], albeit >100 times faster [28]. MALT can either align nucleotide sequences to nucleotide databases (MALTn) or nucleotide to amino acid databases by translating the DNA prior to alignments (MALTx). A potential advantage to using amino acid alignments for palaeomicrobiology is the greater sequence conservation of peptides due to codon redundancy. This property may help smooth over small changes occurring in DNA sequence over time, allowing ancient sequences to be more easily aligned to modern references. However, the already short nature of ancient DNA yields even shorter amino acid sequences (*e.g.* 60 bp DNA translated = 20 amino acid sequence), which may not provide a sufficiently high alignment score for taxonomic classification [30,31]. Additionally, DNA damage can result in errors during in silico translation, further lowering alignment scores. To date, there has been no formal testing of nucleotide versus amino acid alignments for taxonomically classifying short reads typical of ancient DNA.

Here, using both simulated and published ancient DNA data, we test how characteristics of ancient DNA influence alignment-based taxonomic classification. We demonstrate that the BLASTx approach is inappropriate for the alignment of ancient DNA and show that deamination minimally impacts alignment-based taxonomic classification. We also show that reference database choice is an important consideration when attempting to reconstruct ancient microbial communities, and perform an extensive reanalysis of previously published shotgun DNA sequences from ancient dental calculus.

Methods

Data used in this study

To test MALT parameters on real ancient data, collapsed reads from a recent ancient metagenomic dataset were downloaded from OAGR (Online Ancient Genome Repository) https://www.oagr.org.au/experiment/view/65/ [6].

Reference sequences and databases

For the analysis of simulated metagenomes, we downloaded complete bacterial genomes from the NCBI Assembly (6,896 total as of 17th May 2017). Additionally, the coding sequences (CDS) and translated coding sequences were downloaded for these complete bacterial genomes on the same date. These three sources of sequences were used to construct different MALT databases: (MALTn-genome — full genomic sequences, MALTn-CDS — nucleotide coding

sequencing from these genomes, and MALTx — translated coding sequences from these genomes).

For the analysis of previously published dental calculus data, we used sequences from the four following databases:
(1) 2014nr — containing the 2014 protein BLAST database, (downloaded 11th November 2014), used in [6].
(2) 2017nt — the 2017 nucleotide BLAST database (downloaded 6th June 2017).
(3) HOMD — genomic sequences from the Human Oral Microbiome Database (downloaded July 2017 — "All human oral microbial genomes, total: 1,362").
(4) RefSeqGCS — RefSeq genomic sequences from bacterial and archaeal entries (Complete-, Chromosome-, and Scaffold-level assemblies downloaded from NCBI Assembly, Archaeal 349, Bacterial 47,347, total: 47,696).

Simulated metagenome construction

Gargammel [32] was used to generate simulated ancient metagenomes. Bacterial genomic sequences selected from the NCBI assembly were assigned abundances (representing a typical dental plaque community (Table S1)) and then fragmented into five metagenomes containing either strict 30, 50, 70, 90 bp (base pair) fragments, or an empirical ancient DNA fragment length distribution (--loc 4, --scale 0.3 in Gargammel) (Figure S1) (Figure 1). Each fragmented simulated metagenome had 1.5 million sequences. To benchmark the influence of deamination on taxonomic classification, these five simulated metagenomes were then deaminated using Gargammel with the following [33] parameters: nick frequency=0.03, length of overhanging ends (geometric parameter)=0.25, probability of deamination in double-stranded parts=0.01, probability of deamination in single-stranded parts=0.1 for light deamination (10% δ_s), or 0.5 for heavy deamination (50% δ_s). Additionally, a real Mapdamage profile from the LaBrana sample [32] was simulated using Gargammel for the moderate deamination (~20% δ_s). Overall, this resulted in a total of 20 different simulated metagenomes: (five different fragment lengths, 30, 50, 70, 90, and empirical) multiplied by (four different deamination magnitudes 0% δ_s, 10% δ_s, 20% δ_s, and 50% δ_s) = 20 (Metagenome 1-20, Table S2).

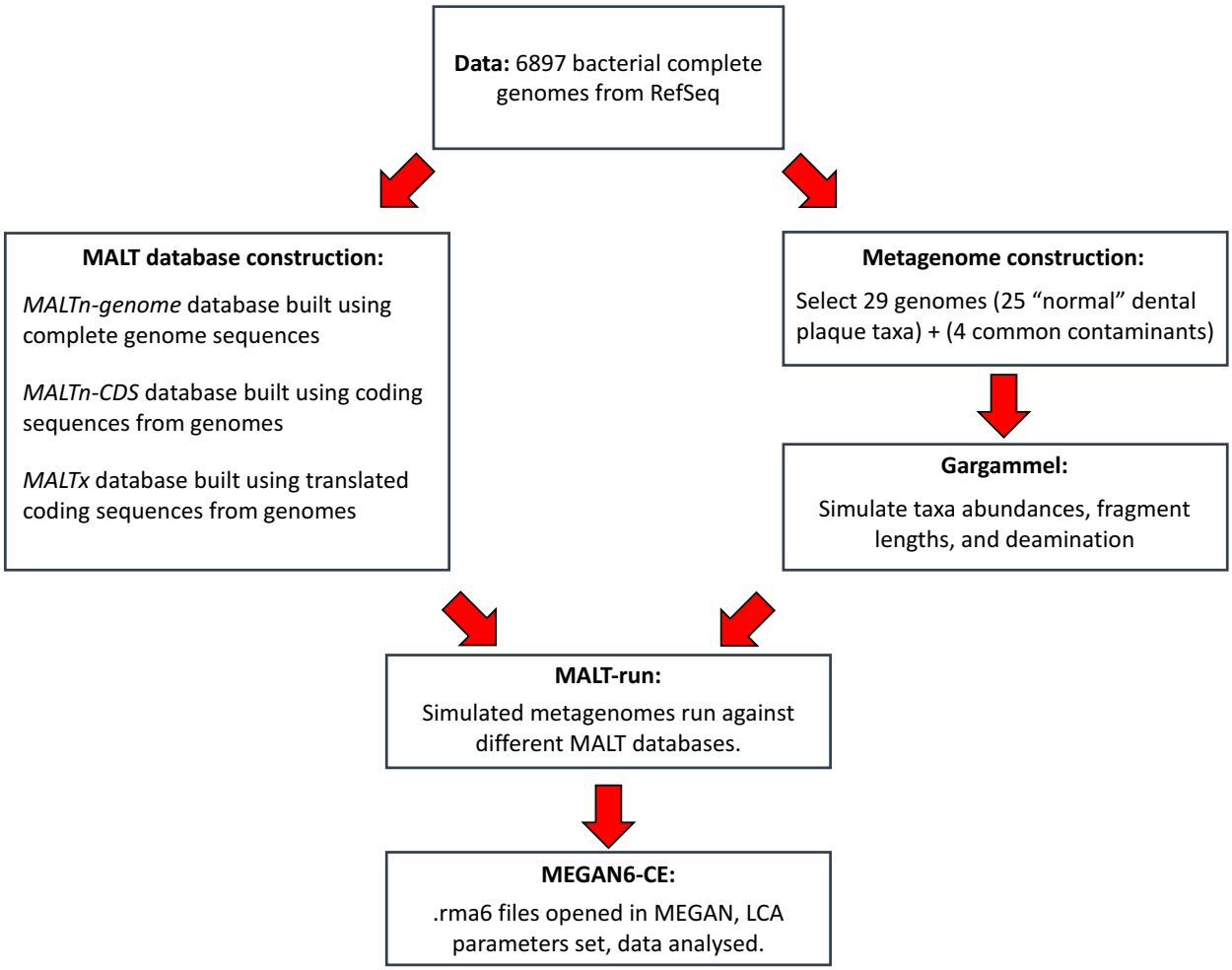


Figure 1 General overview of simulated data construction and analysis.

Generation of divergent sequences

Nucleotide substitution rates are known to differ between different species of bacteria. Proper modelling of bacterial genome evolution is a difficult task. Here we apply a simplified approach which ignores insertions and deletions and instead focused on creating a worst-case scenario for benchmarking the effects of nucleotide substitutions on taxonomic classification. We chose a rate of 10^{-7} substitutions per site per year, representing the mean of known rates for bacterial genomes [34]. We assumed an average bacterial genome size of 3 million bp, thus 10^{-7} * 3,000,000 = 0.3 substitutions per genome per year. Scaling up for multiple years yielded: 10,000 years = 3,000 substitutions (0.1% of genome), 30,000 substitutions (1% of genome), and 300,000 substitutions (10% of genome). We used these numbers to randomly mutate (substitutions only) the bacterial genomes using EMBOSS msbar [35]. These 'mutated' genomes were then used as input for Gargammel as above and deaminated using the heavy deamination magnitude (50% δ_s) (Metagenome 21-23, Table S2).

MALT/MEGAN analysis

MALT-build v 0.3.8 was used on the reference sequences mentioned above with the default parameters. MALT-run v 0.3.8 was used to align the simulated and real data against the different databases using default settings and outputting BLASTtext files. The resulting BLASTtext files were converted to RMA6 files using the MEGAN tool blast2rma, and were then imported and analyzed in MEGAN CE V6.8.13 [36]. The Weighted LCA algorithm was applied to the imported RMA files, as suggested [36]. For analysis of the published ancient DNA data, we used default LCA parameters, with the following exceptions: minimum support percent filter of 0.1% was applied to remove poorly supported assignments (*i.e.* taxonomic assignments require at least 0.1% of the total reads to be considered), and the minimum expected value (E-value) was set to 0.01. Little research has been done regarding the effect of LCA parameters on taxonomic classification, and such research deserves its own study. Regardless, the parameters chosen for this study represent a conservative approach. PCoA plots were generated within MEGAN using Euclidean distances between samples.

Statistical analysis

Divergence between predicted and simulated abundances was calculated using log-odds scores: log odds = $\log_2$(predicted abundance/simulated abundance), and the Pearson correlation coefficient.

UPGMA tree

The UPGMA tree was constructed by exporting the distance matrix from MEGAN6 and importing it into SplitsTree4 [37].

Data availability

Simulated metagenomes and the genomic sequences used to build the metagenomes are available here: XXXX.

Results

Nucleotide-to-nucleotide alignments classify shorter DNA sequences

To compare the alignment performance of nucleotide-to-nucleotide (MALTn) and nucleotide-to-protein (MALTx) alignments at different read lengths, we used simulated metagenomes that contained DNA fragmented with strict length cut-offs (30bp, 50bp, 70bp, 90bp), as well as a log-normal read length distribution commonly found in ancient DNA data (subsequently

referred to as "empirical") (Figure S1). MALTx was unable to align sequences from the 30 and 50 bp simulated metagenomes and could only align 33% of sequences from the 70 bp simulated metagenome (Table 1). Using the empirical fragment length distribution metagenome, MALTn-CDS (coding sequences only) classified 5.55-fold more total sequences than MALTx (MALTn-CDS vs. MALTx) (Figure 2). Nucleotide alignments including non-coding sequences (MALTn-genome) were able to classify 6.25-fold more total sequences than MALTx (6.93 and 9.93-fold more sequences at the genus and species level, respectively) (Figure 2).

Table 1 Percentages of total reads assigned at different taxonomic levels with different read length cut-offs Weighted-LCA settings

Fragment length	Reads assigned total	Reads assigned genus	Reads assigned species
30bp_MALTn-Genome	100	100	97
30bp_MALTn-CDS	86	86	83
30bp_MALTx	**0**	**0**	**0**
50bp_MALTn-Genome	100	100	98
50bp_MALTn-CDS	88	88	86
50bp_MALTx	**0**	**0**	**0**
70bp_MALTn-Genome	100	100	98
70bp_MALTn-CDS	90	90	88
70bp_MALTx	33	31	25
90bp_MALTn-Genome	100	100	98
90bp_MALTn-CDS	91	91	89
90bp_MALTx	82	75	55
Empirical_MALTn-Genome	99	98	97
Empirical_MALTn-CDS	87	87	86
Empirical_MALTx	16	14	10

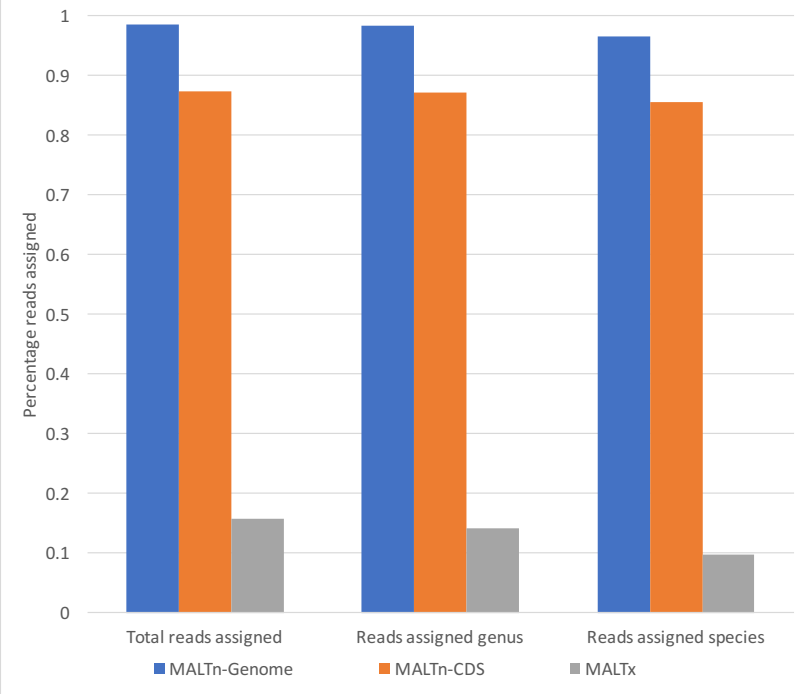


Figure 2. Percentage of reads assigned taxonomy using simulated metagenomes of empirical ancient DNA fragment length against different MALT databases.

We then tested if our findings held true on real ancient dental calculus data using previously published samples from [6], and found nucleotide-to-nucleotide alignments resulted in 7.43-fold more sequences being assigned than nucleotide-to-protein alignments (MALTn-CDS vs. MALTx, Table S3; Figure S2). When including non-coding sequences, this improvement increased to 8.62-fold (10.74 and 13.07-fold at the genus, and species level, respectively) (MALTn-genome vs. MALTx, Table S3; Figure S2), corroborating our simulated metagenome findings. These results suggest that nucleotide-to-amino acid alignments are inappropriate for the alignment and classification of short DNA fragments typical of ancient DNA, and that non-coding sequences in prokaryotic genomes contain information useful for taxonomic classification.

MALTn taxonomic classifications are more accurate than MALTx

While MALTn can classify substantially more sequences than MALTx, the accuracy of these assignments has not yet been examined. We tested the accuracy of these assignments by comparing them to the "ground truth" (*i.e.* what was put into the simulated metagenomes). Overall, MALTn performed well, almost perfectly recapitulating the input simulated metagenome of empirical ancient DNA fragment length (0.998; Pearson correlation; -0.48 sum of log-odds scores between MALTn-CDS and actual metagenome) (Figure 3). Even though sequences below 50 bp were not classified, MALTx also performed well, albeit with poorer abundance predictions (0.943; Pearson correlation and -6.66 sum of log-odds scores between MALTx and actual metagenome) (Figure 3). MALTx also misclassified more sequences *i.e.* assigned sequences to taxa not used for constructing the simulated metagenome. At the species level, 2.4% of assigned sequences using MALTx were misclassified, resulting in 24 taxa being falsely predicted. Whereas only 0.29% of sequences were misclassified using MALTn-CDS with 11 taxa being falsely predicted (Table S4). Additionally, MALTn maintained accuracy classifying sequences as short as 30bp (Figures S3 & S4)

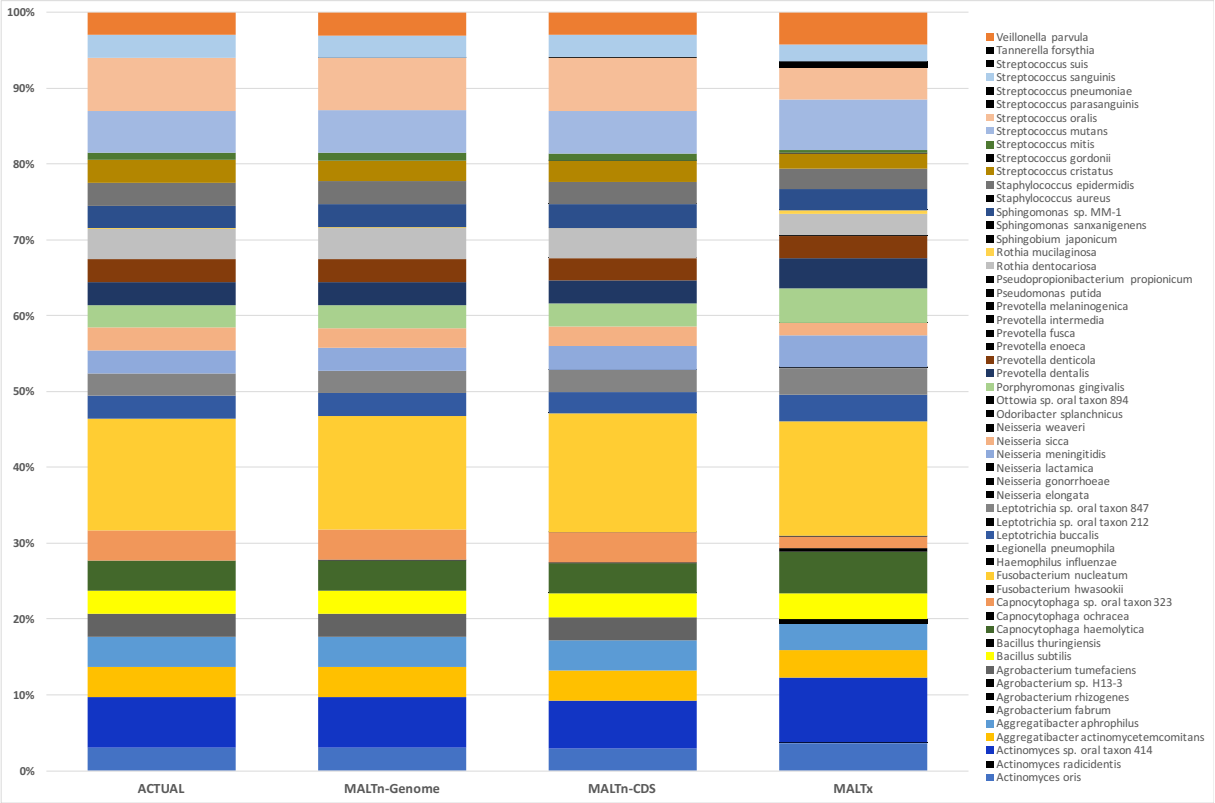


Figure 3. Species level taxonomic classification of empirical fragment length simulated metagenome. Species coloured black were not used as input for constructing the simulated metagenomes, representing misclassifications.

Lastly, the addition of non-coding sequences to the reference database had very little effect on the accuracy of taxonomic classifications, as the MALTn-genome classifications were almost identical to MALTn-CDS (0.999; Pearson correlation between MALTn-genome and MALTn-CDS) (Figure 3); however, more misclassifications at the species level were identified using MALTn-CDS (11 species for MALTn-CDS vs. 2 species for MALTn-genome). Overall, these results suggest that MALTn classifications are more accurate than MALTx both in providing fewer misclassifications, and better abundance predictions. Additionally, it appears that including non-coding information in reference databases (*e.g.* MALTn-genome) substantially reduces misclassifications.

Deamination minimally affects alignment-based classification

We next wanted to test the effects of deamination (a commonly observed form of ancient DNA damage) on alignment-based taxonomic classification. We tested three scenarios: light deamination 10% δ_s (deamination rate on single-stranded overhangs), moderate deamination ~20% δ_s, and heavy deamination 50% δ_s (Table 2). Heavy deamination did not substantially impact the number of sequences assigned for the empirical ancient DNA fragment length distribution metagenome when using MALTn (0.9% loss of sequences assigned at the species level for and MALTn-genome, 1.1% for MALTn-CDS) (Table 2). As expected, lower

magnitudes of deamination had an even smaller impact (Table 2). We also assessed the impacts of heavy deamination on the assignment of DNA sequences of different lengths. Shorter (30bp) sequences were more affected for nucleotide alignments (9.53% loss of sequences assigned at the species level for MALTn-genome, 8.41% for MALTn-CDS), but this loss was not observed for sequences longer than 50bp (Tables S5-S7). Heavy deamination did not substantially increase the percentage of misclassifications at the species level (0.06% to 0.07% for MALTn-genome, 0.29% to 0.30% for MALTn-CDS and 2.42% to 2.48% MALTx). Deamination also did not substantially affect taxonomic composition (Figures S5-S7). Overall, these results suggest that deamination appears to minimally affect alignment-based taxonomic classification.

Table 2. Effects of deamination on taxonomic classification of typical ancient DNA read-length distribution

Fragment length	Reads assigned total (%)	Reads assigned genus (%)	Reads assigned species (%)
MALTn-genome_0δs	98.6	98.4	96.6
MALTn-genome_10δs	98.4	98.2	96.5
MALTn-genome_20δs	98.5	98.3	96.5
MALTn-genome_50δs	97.7	97.5	95.7
MALTn-CDS_0δs	87.4	87.1	85.5
MALTn-CDS_10δs	87.2	86.9	85.3
MALTn-CDS_20δs	87.2	86.9	85.3
MALTn-CDS_50δs	86.5	86.2	84.6
MALTx_0δs	15.8	14.2	9.7
MALTx_10δs	15.2	13.7	9.4
MALTx_20δs	15.0	13.6	9.2
MALTx_50δs	14.5	13.1	8.9

The influence of sequence divergence on taxonomic classification

The effects of sequence divergence on alignment-based taxonomic classification have not yet been explored. To this end, we created divergent simulated metagenomes by introducing random substitution mutations into the same reference genomes used in the above experiments. We chose three different divergence magnitudes: 0.1% sequence divergence (equating to roughly 10ky (thousand years) of evolution), 1% (100ky), and 10% (1,000ky), allowing us to examine the worst-case impacts of sequence divergence on taxonomic classification. Overall, MALTn-genome, MALTn-CDS, and MALTx were able to effectively assign taxonomy with minimal loss of alignments (~1%) at 0.1% and 1% sequence divergence (Figure 4). At 10% divergence, the influence of divergence was more pronounced, as the percentage of sequences not assigned taxonomy increased from 2.28% to 25.1% for MALTn-genome, 13.48% to 35.7%

for MALTn-CDS, and 85.45% to 95.4% for MALTx. Even with the loss of sequences assigned with 10% divergence, the taxonomic classifications and abundances remained relatively stable (Figures S8 & S9), although protein alignments were more effected — 0.944 Pearson correlation coefficient between 1,000ky composition and actual simulated metagenome composition for MALTn-genome, 0.944 for MALTn-CDS, and 0.825 for MALTx. As expected, shorter sequences were more affected by sequence divergence and deamination (Figure S10). Overall, our simulations suggest that random sequence divergence of up to 100,000 years may minimally affect alignment-based classification.

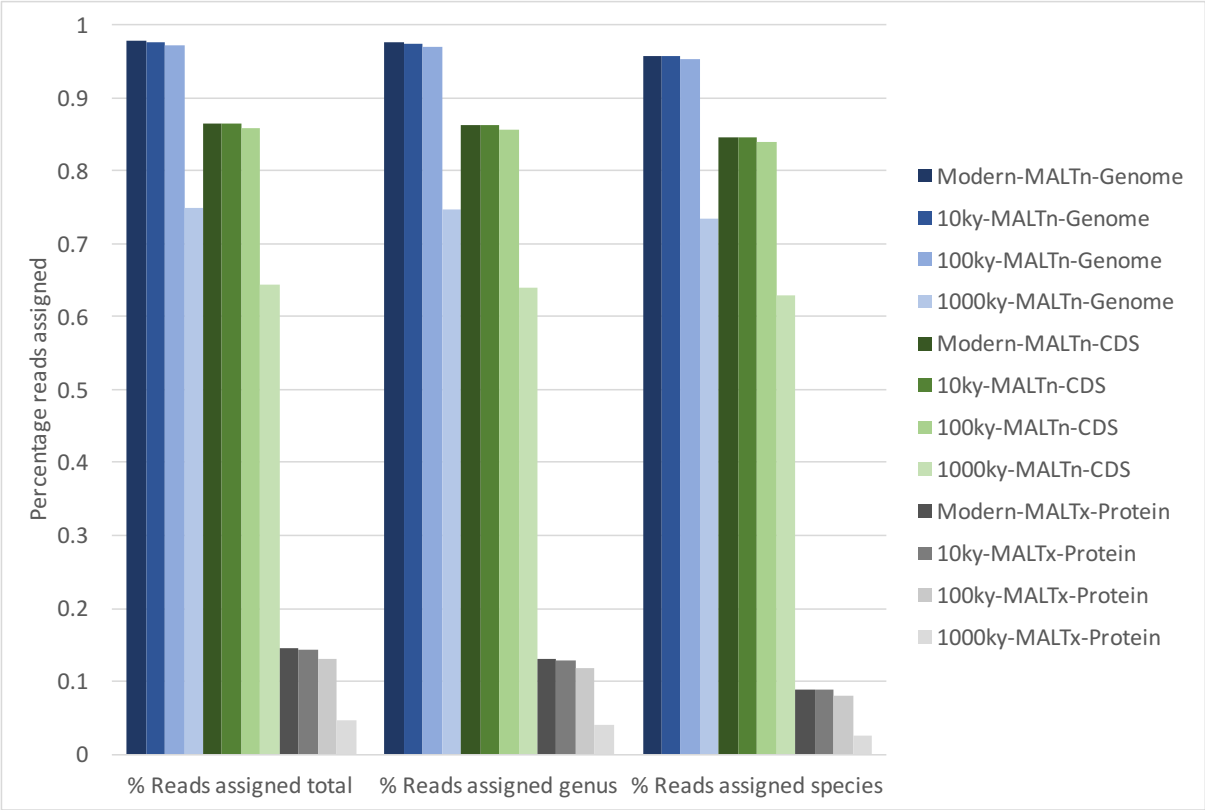


Figure 4. Percentage of reads assigned taxonomy using divergent and deaminated simulated metagenomes of typical ancient DNA fragment length.

Reference database choice strongly influences taxonomic classification

Because alignment-based methods are highly reliant on reference sequences available in databases, we next sought to test the influence of database choice on taxonomic classification of ancient microbial DNA. To this end, we constructed four different reference databases from different sources:

(1) 2014nr: This contains the 2014 non-redundant protein BLAST database, which was used in a recent palaeomicrobiology publication by Weyrich *et al.* [6], and represents the example of a database used with the MALTx method.

(2) 2017nt: The 2017 nucleotide BLAST database, this is the default for BLAST searches on the NCBI website, and does not include draft genome assemblies (chromosomes, scaffolds, or contigs).
(3) HOMD: genomic sequences from the Human Oral Microbiome Database. This is a curated nucleotide database comprised of oral-associated microbial species, and includes all genome assembly levels (complete, chromosome, scaffold, and contig).
(4) RefSeqGCS: Genomic sequences of bacterial and archaeal assemblies from the NCBI Assembly. This database includes complete, chromosome, and scaffold level genome assembly levels (contigs could not be included due to size/memory constraints). It also contains substantially more entries 47,696 vs. 1,362 for HOMD, with a broader diversity of entries (*i.e.* not primarily oral taxa).
To test the effects of these different databases on the taxonomic classification of real data palaeomicrobiological data, we aligned the reads from four previously published, deeply sequenced dental calculus samples (three ancient, one modern) [6] against the four databases mentioned above. As expected, the MALTx approach using the 2014nr database assigned the least number of reads taxonomy (Figure 5; Figure S11), while the MALTn approach using the RefSeqGCS database assigned the most sequences. In addition, the highest percentage of sequences assigned was from the modern dental calculus sample, with the RefSeqGCS database assigning the most reads taxonomy (80.8%) to this sample (Figure 5). In the ancient samples, the highest number of species were identified when sequences were aligned to the HOMD (Table 3), while higher numbers of genera were classified when comparing ancient sequences to the RefSeqGCS and 2017nt databases (Table 3). The higher number of species observed in the HOMD could be due to either cross-mapping from environmental taxa (as it contains few soil/environmental genomes), or a higher diversity of oral-specific assemblies.

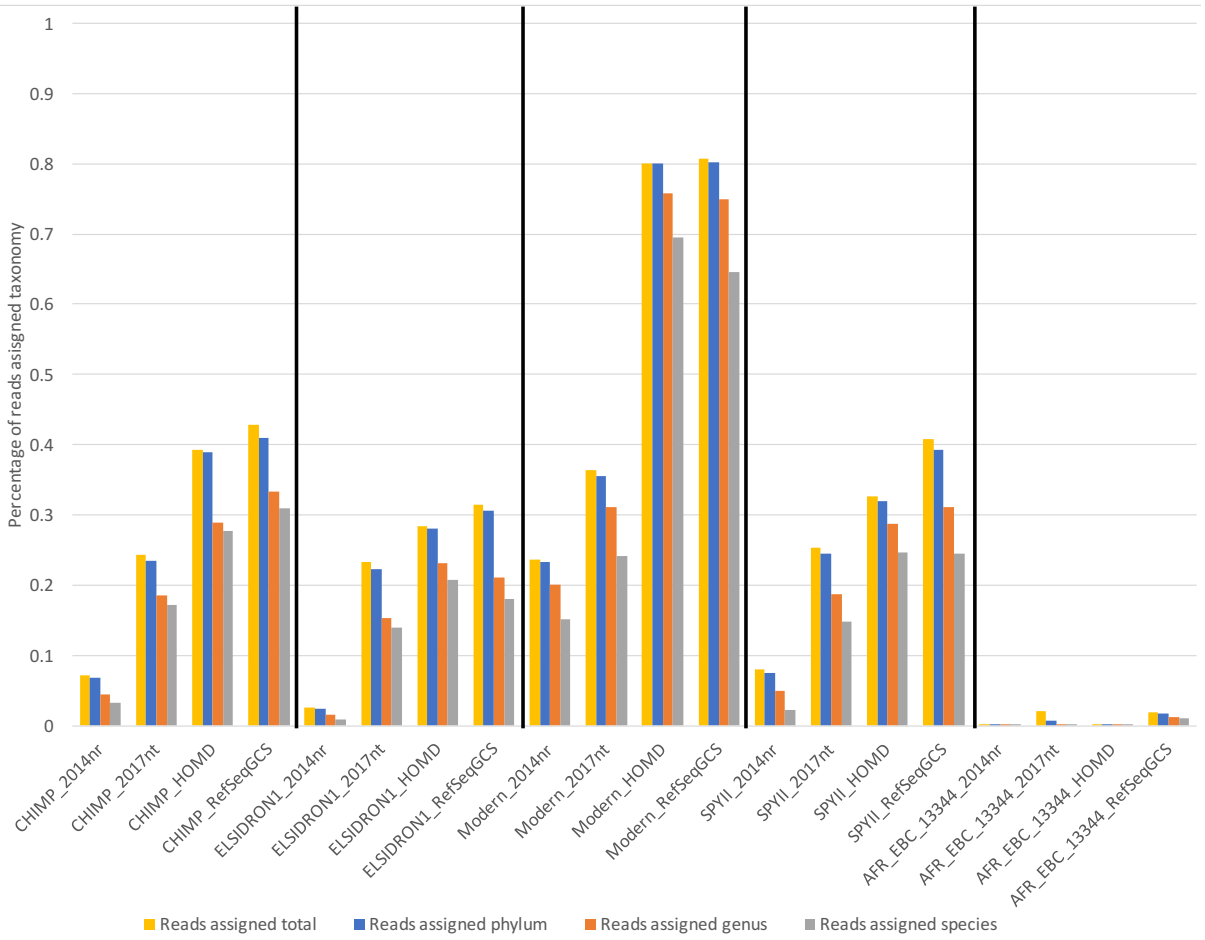


Figure 5. Percentage of reads assigned taxonomy for deeply sequenced published data from Weyrich 2017 *et al*. Clustered columns represent samples analysed using different reference databases. Colours indicate specificity of assignments.

Species identified in the analysis were also markedly impacted by the database used (Figures S12-S15; Table S8). When assessing the dominant taxa (>1% abundance), important, high-abundance oral taxa that have scaffold-level assemblies were not identified using the 2017nt BLAST database. These include *Actinomyces dentalis, Bacteriodetes sp. oral taxon 274, Capnocytophaga granulosa, Corynebacterium matruchotii, Methanobrevibacter oralis, Prevotella sp. oral taxon 317, and Pseudoramibacter alactolyticus*. This is a likely reason for the 2017nt performing worse regarding the percentage of total reads assigned taxonomy (27.3%) when compared to the HOMD (45.1%) and RefSeqGCS (48.9%) (Figure 5), and highlights the importance of including scaffold-level assemblies for taxonomic classification. Overall, the RefSeqGCS database assigned the most reads taxonomy, and contained the most diverse selection of reference genomes. Therefore, we chose the RefSeqGCS for subsequent reanalysis of published dental calculus samples.

Reanalysis of published dental calculus data with nucleotide alignment

To further test the performance of the RefSeqGCS database, we included more ancient dental calculus samples (total of n=24) from the same study use above [6]. We found that MALTn substantially increased the number of reads assigned taxonomy, especially for samples with

short fragment length distributions (Table S9). We found an average 64.2-fold (minimum 1.5, maximum 525.5) increase in number of reads successfully assigned taxonomy when using MALTn against the RefSeqGCS versus MALTx against the 2014nr (Table S9). Despite the increase in reads assigned using MALTn, the average percentage of unassigned reads remained relatively high 58.2% (minimum 19.4%, maximum 95.6%), although this was substantially lower than MALTx (average 94.2%, minimum 58.3%, maximum 99.8%). Regarding taxonomic assignment, species level composition was more affected by database choice/alignment method, showing a clearer separation using Principal Component Analysis (PCoA) (Figure 6A & 6B). Finally, changes in taxonomic classification appeared to alter the grouping of samples originally reported in the original paper [6] (Figure 7). The previously reported 'Forager-gatherers' and 'Hunter-gatherers' clades disappeared and became intermixed, whereas the 'Ancient agriculturalists' clade remained intact. However, further research is needed to confirm this. Overall, these findings suggest that it will be important to revisit previously published datasets as reference databases become larger and analytical techniques are improved.

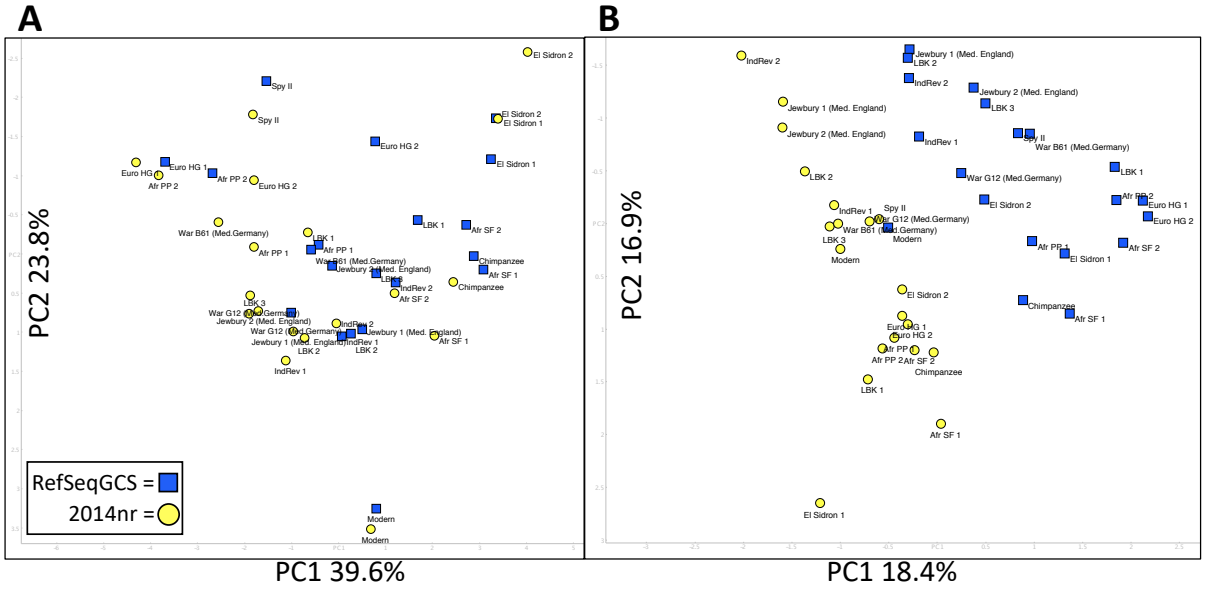


Figure 6. PCoA of Euclidean distances of microbial communities between samples. (A) genus level. (B) species level. Yellow circles represent communities classified using nucleotide-to-protein alignment against the 2014nr BLAST database, and purple squares represent communities classified using nucleotide-to-nucleotide alignment against the RefSeqGCS data.

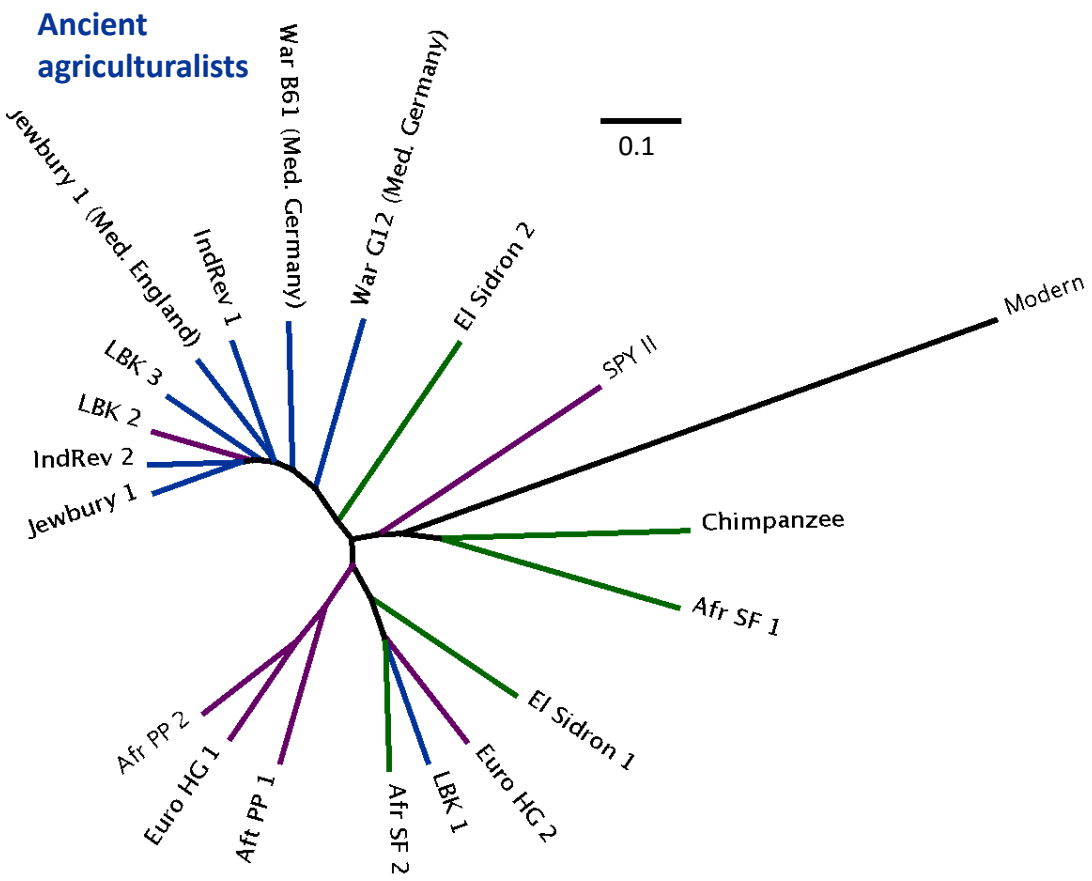


Figure 7. UPGMA tree constructed using Bray-Curtis dissimilarity of microbial composition between samples from Weyrich 2017 *et al.* Branch scale bar represents Bray-Curtis dissimilarity between samples.

Discussion

Studying past microorganisms relies on our ability to reliably assign taxonomy. The field of palaeomicrobiology is in its infancy, and to our knowledge, there has not yet been a thorough study of how the characteristics of ancient DNA affect taxonomic assignments. To this end, we sought to provide such a resource for palaeomicrobiological researchers, and suggestions moving forward.

Nucleotide-to-nucleotide versus nucleotide-to-protein alignments for palaeomicrobiology

We evaluated the performance of both nucleotide-to-nucleotide and nucleotide-to-protein alignments for taxonomic classification, and found that sequences shorter than ~60 bp could not be aligned using a the nucleotide-to-protein approach. This limits the feasibility of nucleotide-to-protein alignments for palaeomicrobiological studies given that ancient DNA reads are typically shorter than 60bp. The likely reason for the poorer performance of nucleotide-to-protein is that nucleotide translation reduces alignment length by ~66.6% (e.g. a

60bp nucleotide sequence = a 20 aa protein sequence), yielding a lower bit (alignment) score. Given that the default bit-score threshold for MALT is 50, most short reads would struggle to obtain a sufficient score to pass filtering. Additionally, amino acid scoring matrices influence the final score of the alignment; the default MALTx scoring matrix is BLOSUM62 which optimized for longer sequences [31]. The inability to align short reads may also bias taxonomic composition towards modern environmental and laboratory contaminant taxa, whose reads are typically longer. Additionally, it has been suggested that mycobacterial cell walls may protect DNA from hydrolytic damage-induced fragmentation and result in longer average read lengths for some mycobacterial taxa [38]. This would result in an overrepresentation of taxa that have longer DNA fragments if using protein alignments.

Despite the 5.55-fold loss of reads assigned using nucleotide-to-protein alignments, the taxonomic classifications were relatively similar to the nucleotide alignments for the simulated dataset. However, nucleotide-to-nucleotide alignments clearly outperformed nucleotide-to-protein alignments in terms of number of reads assigned, and the lower rate of misclassifications. These misclassifications primarily resulted from the lack of non-coding sequences in the protein and CDS nucleotide databases, with misclassifications being supported by sequences that were derived from non-coding genes in the simulated inputs (*e.g.* tRNA, rRNA etc.). Recent estimates from 2,671 complete bacterial genomes place the average percentage of non-coding DNA at 12% [39], this represents a non-trivial amount of information that should be harnessed when using reference-based taxonomic alignment. Finally, we also demonstrated nucleotide-to-nucleotide alignments can faithfully recapitulate simulated taxonomic composition using reads as short as 30bp, highlighting the applicability of nucleotide-to-nucleotide alignments for ultra-short fragments typical of palaeomicrobiological studies.

Pending further optimization to nucleotide-to-protein alignment methods we recommend using a nucleotide-to-nucleotide alignment approach for taxonomic classification of short length ancient DNA, and the inclusion of non-coding information in reference databases to reduce potential misclassification and to increase the amount of information used in alignments.

Characteristics of ancient DNA that influence taxonomic classification

In this study, we tested the impacts of deamination on shotgun metagenomic taxonomic classifications. We demonstrated that high levels of cytosine deamination (50% δs) did not substantially impact taxonomic classification in longer sequences; however, we observed a loss of ~15% of the species level classifications when analyzing 30 bp DNA sequences. This

suggests that the use of uracil-DNA-glycosylase (UDG) [40] — an enzyme that cleaves deaminated cytosines to reduce the rate of ancient DNA errors — may not be required for microbial taxonomic classifications of ancient remains, as this also reduces the total number of sequences that can be analyzed. Additionally, treatment with UDG — either full or partial [41] — substantially reduces a key source of ancient DNA authentication, which is critical in palaeomicrobiological studies to verify ancient taxa from modern contamination. The lack of such authentication in palaeomicrobiological research has already led to contentious claims [42,16,18]. Given the minimal impact of deamination on alignment-based taxonomic classification, and the importance of deamination as a measure of ancient DNA authenticity, we recommend against the use of UDG for future palaeomicrobiological studies that focus on alignment-based classification.

Sequence divergence is another characteristic of ancient DNA that can render taxonomic classification difficult. We tested three substitution-based sequence divergence simulations, and found that rates of sequence divergence corresponding to <100,000 years unlikely to alter palaeomicrobiological classifications. A substantial reduction in the number of identified sequences was observed for samples with sequence divergence simulated at one million years (~20% loss of reads assigned taxonomy). However, this is at the theoretical limit of DNA preservation [14], and is thus unlikely to hamper most palaeomicrobiological studies. We also found that shorter sequences were more affected by sequence divergence and deamination, and this can intuitively be explained by the reduction in raw alignment score due to mismatches to the reference. As such, the use of new molecular techniques to obtain even shorter DNA fragments (*e.g.* <25 bp [43]) may prove especially difficult to classify taxonomically given the combined effects of sequence divergence and deamination.

The influence of insertions, deletions, and recombination would have additional impacts on sequence divergence that were not tested here but would likely further hinder taxonomic classifications. Future simulations accounting for differences in synonymous/non-synonymous mutations may give amino acid alignments the upper-hand given the excess synonymous mutations observed due to purifying selection [44], although amino acid alignment scoring would still have to be optimized to deal with short DNA fragments. Additionally, future studies simulating the effects of insertions, deletions, and recombination on taxonomic classification are warranted.

Overall, we found that alignment-based taxonomic classification appears robust against magnitudes of random nucleotide substitution that could be observed in ancient DNA <100,000 years old. We also demonstrated that shorter fragments of DNA are more affected by nucleotide sequence divergence and deamination.

Reference database strongly influences alignment-based taxonomic classification

We found that database choice had a major impact on both the number of reads that were assigned taxonomy and the taxa classified. The 2017nt BLAST database performed poorly compared to the HOMD and RefSeqGCS, assigning on average 33% fewer reads taxonomy and lacking numerous key oral taxa. This is likely because the 2017nt BLAST database does not contain draft, unfinished bacterial genomes assemblies, which is a major limitation for ancient dental calculus research given that some important oral taxa currently have only chromosome or scaffold-level assemblies, such as *Acintomyces dentalis, Bacteroidetes sp. oral taxon 274*, *Capnocytophaga granulosa, Corynebacterium matruchotii, Eikenella corrodens, Lautropia mirabilis, Methanobrevibacter oralis,* numerous *Prevotella species, Pseudoramibacter alactolyticus, Slackia exigua,* and *Treponema socranskii*. While the HOMD database contained substantially fewer reference sequences compared to the RefSeqGCS (1,362 vs. 47,696, respectively), it performed comparably regarding the number of reads assigned from ancient dental calculus samples. However, we don't recommend the HOMD database alone for taxonomic classification of ancient dental calculus, as it does not contain many environmental or laboratory contaminant taxa that are typically present in ancient samples. These environmental and laboratory contaminant taxa allow for the quantification of contamination, and competitive alignment — which can prevent false positive assignments. Overall, given the larger diversity of the RefSeqGCS database, and its ability to classify the most reads taxonomy, we would recommend it over the others tested for future palaeomicrobiological studies. However, further work is needed to assess and curate the quality of reference assemblies — especially of scaffold-level and below — to ensure reliable and accurate alignment-based taxonomic classification [45]. There is also scope for a concerted effort by palaeomicrobiological researchers to work together in constructing a curated, regularly updated reference database. This could help foster reproducibility and set a standard for future work in the field — similar to what has been accomplished by the HOMD for oral microbiome studies [46].

Reanalysis of previously published data

We also performed a reanalysis of previously published ancient dental calculus data from Weyrich *et al.* [6] to test if our in-silico findings were true for real data, explore the proportion of sequences currently classifiable, and to see whether the relationships between samples changed when using the RefSeqGCS database. Nucleotide alignment against the RefSeqGCS database performed considerably better compared to protein alignment against the 2014nr, with

an average 64.2-fold increase in the number of reads assigned taxonomy. As expected, this increase was higher for samples with shorter mean fragment lengths and highlights the importance of using nucleotide-to-nucleotide alignments to allow fair comparisons between samples of different mean fragment lengths.

Despite the substantial increase in the number of reads aligned, the average number of reads that did not have any alignment was 58.2%. When compared to the latest extension to the human microbiome project where the average number of reads without alignment was ~25% for 265 supragingival plaque samples [47], this suggests that reference bias exists for ancient calculus samples. This is not likely due to methodological differences between studies, as the modern calculus sample we analyzed in this study (European descent) had a similar percentage of its reads without alignment (19.4%). One hypothesis for this finding is that modern reference databases are missing a large number of oral microorganisms that were present in historical and ancient humans. Additionally, given that most modern microbiome studies and microbial genomes assembled are from European/American individuals [2,47], current reference databases are likely missing oral microbial diversity from other human populations. Another possibility for this finding is DNA contamination of the dental calculus samples with ancient or modern soil microorganisms that do not currently have reference sequences. While we limited our reference database to prokaryotes to save space, the percentage of eukaryotic DNA in ancient dental calculus and modern dental plaque is low (<0.3%) [9] and is therefore unlikely to be driving this high percentage of unaligned reads. Clearly, further research is needed to investigate the large proportion of 'microbial dark matter' present in ancient samples, in the meantime, we suggest that researchers report the percentage of unaligned reads per sample for greater transparency.

We determined that choice of database and alignment method influenced the resulting taxonomic composition of samples, especially at the species level, where the major split on PC1 was associated with database/alignment method. This could be due to both an increase in the diversity captured by the RefSeqGCS, and the ability to align a larger proportion of the data using nucleotide-to-nucleotide alignments. Indeed, we were able to identify species not found by the previous study, and which could be the subject of future research. Finally, differences were also observed in the relationships between samples, with two of the three large groupings previously identified disappearing, with the exception of the 'Ancient agriculturalists' group. However, further research is needed to confirm this finding, as it could be a result of different LCA parameters used between studies (default in [6], more stringent in this paper).

Overall, we demonstrated the importance of revisiting previously published data with new reference databases and methods. This will be increasingly important as new genome

assemblies are obtained in the future, which could reduce the proportion of microbial 'dark matter' and yield new insights into previously published datasets.

Future challenges for taxonomic classification of ancient DNA

The ability to investigate ancient microbial communities can shed valuable insights into microbial evolution, climate change, human migration, and the evolution of the human microbiota. However, there are analytical challenges that hinder our understanding of ancient microbial communities, and we list some below.

Database sizes are a limitation for the currently implemented algorithms in MALT, as MALT uses large amounts of memory (*e.g.* >1 TB of RAM) when comparing sequences to the 2017nt and RefSeqGCS databases, and these requirements will increase as more genomes are added to databases in the near future. A possible solution may be better database curation, *e.g.* through deduplication of the same strain with multiple entries, which could be accomplished using a sequence similarity clustering-based approach. Additionally, future algorithmic refinements in database compression may alleviate this issue. Ultimately, database choice is an essential facet of alignment-based taxonomic classification, and we urge researchers to carefully consider the pros and cons of different databases and how they can affect their findings. Additionally, databases are a fluid issue; as more reference sequences are generated, reanalysis of palaeomicrobiological datasets will be important to reassess past interpretations and findings.

Our current inability to assign taxonomy to >50% of DNA sequences in ancient dental calculus samples is a major issue, and we need new methods to identify these reads. A potential approach could be to *de-novo* assemble genomes from these ancient samples and use these as reference sequences for further alignment-based taxonomic classification. Such tools currently exist [21], but their performance on short and degraded ancient DNA is yet to be determined. An alternative and complementary approach is to obtain more high-quality reference genome from modern samples, including from non-European individuals. Until we can comfortably assign a higher proportion of ancient DNA reads taxonomy, we recommend that palaeomicrobiological researchers report the percentage of unassigned reads when classifying taxonomy.

Finally, this paper did not investigate eukaryotic or viral classification in ancient metagenomes, instead focusing on prokaryotes which account for >99% of DNA in ancient dental calculus [3,6]. The inclusion of eukaryotic and viral sequences within the prokaryotic RefSeqGCS database was not possible due to size hardware constraints even with a sophisticated computer server containing 1.5 TB of RAM. Given that a recent paper identified

eukaryotic DNA putatively present in ancient dental calculus using the nucleotide-to-protein alignment approach [6], a future reanalysis using the more applicable nucleotide-to-nucleotide approach is warranted.

Overall, we hope that this paper is a useful resource for palaeomicrobiological researchers and that future studies will tackle the issues highlighted. Only through the development and improvement of analytical techniques will the full potential of palaeomicrobiology be realized.

References

1. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6:1621–1624.

2. Consortium THMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

3. Warinner C, Speller C, Collins MJ. A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. Philos Trans R Soc B Biol Sci. 2014;370:20130376–20130376.

4. Weyrich LS, Dobney K, Cooper A. Ancient DNA analysis of dental calculus. J Hum Evol. 2015;79:119–24.

5. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45:450–5.

6. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature. 2017;544:357–61.

7. Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, et al. A draft genome of Yersinia pestis from victims of the Black Death. Nature. 2011;478:506–10.

8. Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, et al. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. Nature. 2014;514:494–7.

9. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336–44.

10. Dominguez-Bello MG, Blaser MJ. The Human Microbiota as a Marker for Migrations of Individuals and Populations. Annu Rev Anthropol. 2011;40:451–74.
11. Maixner F, Krause-Kyora B, Turaev D, Herbig A, Hoopmann MR, Hallows JL, et al. The 5300-year-old Helicobacter pylori genome of the Iceman. Science. 2016;351:162–5.
12. Eisenhofer R, Anderson A, Dobney K, Cooper A, Weyrich LS. Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events. J Isl Coast Archaeol. 2017;0:1–14.
13. Frisia S, Weyrich LS, Hellstrom J, Borsato A, Golledge NR, Anesio AM, et al. The influence of Antarctic subglacial volcanism on the global iron cycle during the Last Glacial Maximum. Nat Commun. 2017;8:15425.
14. Allentoft ME, Collins M, Harker D, Haile J, Oskam CL, Hale ML, et al. The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. Proc R Soc Lond B Biol Sci. 2012;279:4724–33.
15. Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. Cold Spring Harb Perspect Biol. 2013;a012567.
16. Eisenhofer R, Cooper A, Weyrich LS. Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential. FEMS Microbiol Ecol [Internet]. 2017 [cited 2017 Jun 27];93. Available from: https://academic.oup.com/femsec/article/93/5/fix042/3089752/Reply-to-Santiago-Rodriguez-et-al-proper
17. Llamas B, Valverde G, Fehren-Schmitz L, Weyrich LS, Cooper A, Haak W. From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. STAR Sci Technol Archaeol Res. 2017;3:1–14.
18. Eisenhofer R, Weyrich LS. Proper Authentication of Ancient DNA Is Still Essential. Genes. 2018;9:122.
19. Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci Rep. 2015;5:16498.
20. Gilbert JA, Jansson JK, Knight R. The Earth Microbiome project: successes and aspirations. BMC Biol. 2014;12:69.
21. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ [Internet]. 2014 [cited 2017 Oct 13];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4183954/

22. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533.
23. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
24. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9:357–9.
25. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12:902–3.
26. Meyer F, Paarmann D, D'Souza M, Olson R, Glass E, Kubal M, et al. The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9:386.
27. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
28. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv. 2016;050559.
29. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
30. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17:377–86.
31. Pearson WR. Selecting the Right Similarity-Scoring Matrix. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al. 2013;43:3.5.1-3.5.9.
32. Renaud G, Hanghøj K, Willerslev E, Orlando L. gargammel: a sequence simulator for ancient DNA. Bioinformatics. 2017;33:577–9.
33. Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci. 2007;104:14616–21.
34. Duchêne S, Holt KE, Weill F-X, Le Hello S, Hawkey J, Edwards DJ, et al. Genome-scale rates of evolutionary change in bacteria. Microb Genomics [Internet]. 2016 [cited 2017 Jul 26];2. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5320706/
35. Rice P, Longden I, Bleasby A. EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. 2000;16:276–7.
36. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Comput Biol. 2016;12:e1004957.

37. Huson DH, Bryant D. Application of Phylogenetic Networks in Evolutionary Studies. Mol Biol Evol. 2006;23:254–67.
38. Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jäger G, Bos KI, et al. Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. Science. 2013;341:179–83.
39. Land M, Hauser L, Jun S-R, Nookaew I, Leuze MR, Ahn T-H, et al. Insights from 20 years of bacterial genome sequencing. Funct Integr Genomics. 2015;15:141–61.
40. Briggs AW, Stenzel U, Meyer M, Krause J, Kircher M, Pääbo S. Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. Nucleic Acids Res. 2010;38:e87.
41. Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. Partial uracil–DNA–glycosylase treatment for screening of ancient DNA. Philos Trans R Soc B Biol Sci [Internet]. 2015 [cited 2017 Sep 21];370. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4275898/
42. Weyrich LS, Llamas B, Cooper A. Reply to Santiago-Rodriguez et al.: Was luxS really isolated from 25- to 40-million-year-old bacteria? FEMS Microbiol Lett. 2014;353:85–6.
43. Glocke I, Meyer M. Extending the spectrum of DNA sequences retrieved from ancient bones and teeth. Genome Res. 2017;27:1230–7.
44. Ochman H. Neutral Mutations and Neutral Substitutions in Bacterial Genomes. Mol Biol Evol. 2003;20:2091–6.
45. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. 2015;25:1043–55.
46. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database [Internet]. 2010 [cited 2018 Feb 9];2010. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baq013/405450
47. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.

Supplementary figures and tables

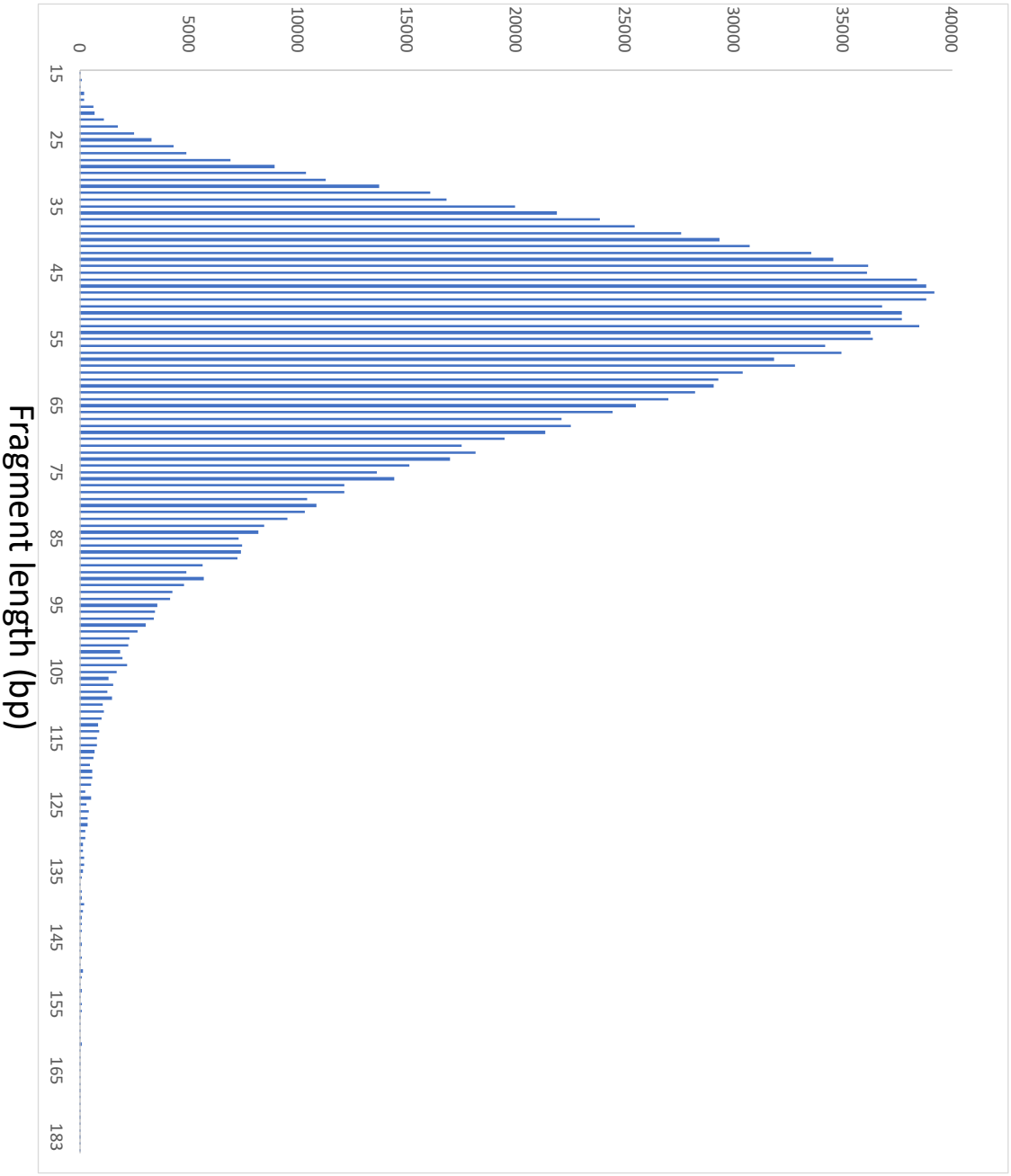


Figure S1. Read length distribution of simulated metagenome with commonly observed fragment length distribution of ancient DNA datasets

Table S1. Simulated metagenome overview, “normal” plaque community based on Mark-Welsh et al. 2016

Taxon	fna filename:	Abundance	% Abundance	Genus	% Abundance Genus
Actinomyces oris strain T14V	GCF_001553935.1_ASM155393v1_cds_from_genomic.fna	0.03	3	Actinomyces	
Actinomyces sp. oral taxon 414 strain F0588	GCF_001278845.1_ASM127884v1_cds_from_genomic.fna	0.067	6.7	Actinomyces	9.7
Aggregatibacter actinomycetemcomitans strain 624	GCF_001594265.1_ASM159426v1_cds_from_genomic.fna	0.04	4	Aggregatibacter	
Aggregatibacter aphrophilus strain W10433	GCF_001262035.1_ASM126203v1_cds_from_genomic.fna	0.04	4	Aggregatibacter	8
Agrobacterium tumefaciens strain A	GCF_000971565.1_ASM97156v1_cds_from_genomic.fna	0.03	3	Agrobacterium	3
Bacillus subtilis BSn5	GCF_000186745.1_ASM18674v1_cds_from_genomic.fna	0.03	3	Bacillus	3
Capnocytophaga haemolytica strain CCUG 32990	GCF_001553545.1_ASM155354v1_cds_from_genomic.fna	0.04	4	Capnocytophaga	
Capnocytophaga sp. oral taxon 323 strain F0383	GCF_001278825.1_ASM127882v1_cds_from_genomic.fna	0.04	4	Capnocytophaga	8
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	GCF_000007325.1_ASM732v1_cds_from_genomic.fna	0.1	10	Fusobacterium	
Fusobacterium nucleatum subsp. polymorphum strain ChDC F306	GCF_001433955.1_ASM143395v1_cds_from_genomic.fna	0.04	4	Fusobacterium	
Fusobacterium nucleatum subsp. vincentii 3_1_36A2	GCF_000162235.2_ASM16223v2_cds_from_genomic.fna	0.007	0.7	Fusobacterium	14.7
Leptotrichia buccalis DSM 113	GCF_000023905.1_ASM2390v1_cds_from_genomic.fna	0.03	3	Leptotrichia	
Leptotrichia sp. oral taxon 847	GCF_001553645.1_ASM155364v1_cds_from_genomic.fna	0.03	3	Leptotrichia	6
Neisseria meningitidis MC58 chromosome	GCF_000008805.1_ASM880v1_cds_from_genomic.fna	0.03	3	Neisseria	
Neisseria sicca strain FDAARGOS_2	GCF_002073715.1_ASM207371v1_cds_from_genomic.fna	0.03	3	Neisseria	6
Porphyromonas gingivalis ATCC 33277 DNA	GCF_000010505.1_ASM1050v1_cds_from_genomic.fna	0.03	3	Porphyromonas	3
Prevotella dentalis DSM 3688	GCF_000242335.1_ASM24233v3_cds_from_genomic.fna	0.03	3	Prevotella	
Prevotella denticola F0289	GCF_000193395.1_ASM19339v1_cds_from_genomic.fna	0.03	3	Prevotella	6
Rothia dentocariosa ATCC 17931	GCF_000164695.2_ASM16469v2_cds_from_genomic.fna	0.04	4	Rothia	
Rothia mucilaginosa DNA complete genome strain: NUM-Rm6536	GCF_001548235.1_ASM154823v1_cds_from_genomic.fna	0.001	0.1	Rothia	4.1
Sphingomonas sp. MM-1	GCF_000347675.2_ASM34767v2_cds_from_genomic.fna	0.03	3	Sphingomonas	3
Staphylococcus epidermidis ATCC 12228	GCF_000007645.1_ASM764v1_cds_from_genomic.fna	0.03	3	Staphylococcus	3
Streptococcus cristatus AS 1.3089	GCF_000385925.1_ASM38592v1_cds_from_genomic.fna	0.03	3	Streptococcus	
Streptococcus mitis B6	GCF_000027165.1_ASM2716v1_cds_from_genomic.fna	0.01	1	Streptococcus	
Streptococcus mutans NN202DNA	GCF_000091645.1_ASM9164v1_cds_from_genomic.fna	0.005	0.5	Streptococcus	
Streptococcus mutans UA159 chromosome	GCF_000007465.2_ASM746v2_cds_from_genomic.fna	0.05	5	Streptococcus	
Streptococcus oralis Uo5	GCF_000253155.1_ASM25315v1_cds_from_genomic.fna	0.07	7	Streptococcus	
Streptococcus sanguinis SK36	GCF_000014205.1_ASM1420v1_cds_from_genomic.fna	0.03	3	Streptococcus	19.5
Veillonella parvula DSM 2008	GCF_000024945.1_ASM2494v1_cds_from_genomic.fna	0.03	3	Veillonella	3
	TOTAL:	1	100		100

Table S2. Characteristics of simulated metagenomes used in this study

	Fragment length (BP)	Deamination (%ss)	Divergence (% nucleotides)
Metagenome1	30	0	0
Metagenome2	30	0.1	0
Metagenome3	30	0.5	0
Metagenome4	30	Empirical	0
Metagenome5	50	0	0
Metagenome6	50	0.1	0
Metagenome7	50	0.5	0
Metagenome8	50	Empirical	0
Metagenome9	70	0	0
Metagenome10	70	0.1	0
Metagenome11	70	0.5	0
Metagenome12	70	Empirical	0
Metagenome13	90	0	0
Metagenome14	90	0.1	0
Metagenome15	90	0.5	0
Metagenome16	90	Empirical	0
Metagenome17	Empirical	0	0
Metagenome18	Empirical	0.1	0
Metagenome19	Empirical	0.5	0
Metagenome20	Empirical	Empirical	0
Metagenome21	Empirical	0.5	0.1
Metagenome22	Empirical	0.5	1
Metagenome23	Empirical	0.5	10

Table S3. Comparisons of fold increases of reads assigned between databases

Comparison:	Fold increase total	Fold increase genus	Fold increase species
MALTn-Genome vs MALTn-CDS	1.16	1.15	1.14
MALTn-Genome vs MALTx	8.62	10.74	13.07
MALTn-CDS vs MALTx	7.43	9.37	11.42

Figure S2. Average percentage of reads assigned across all ancient dental calculus samples analyzed using Genome, CDS, and Protein databases

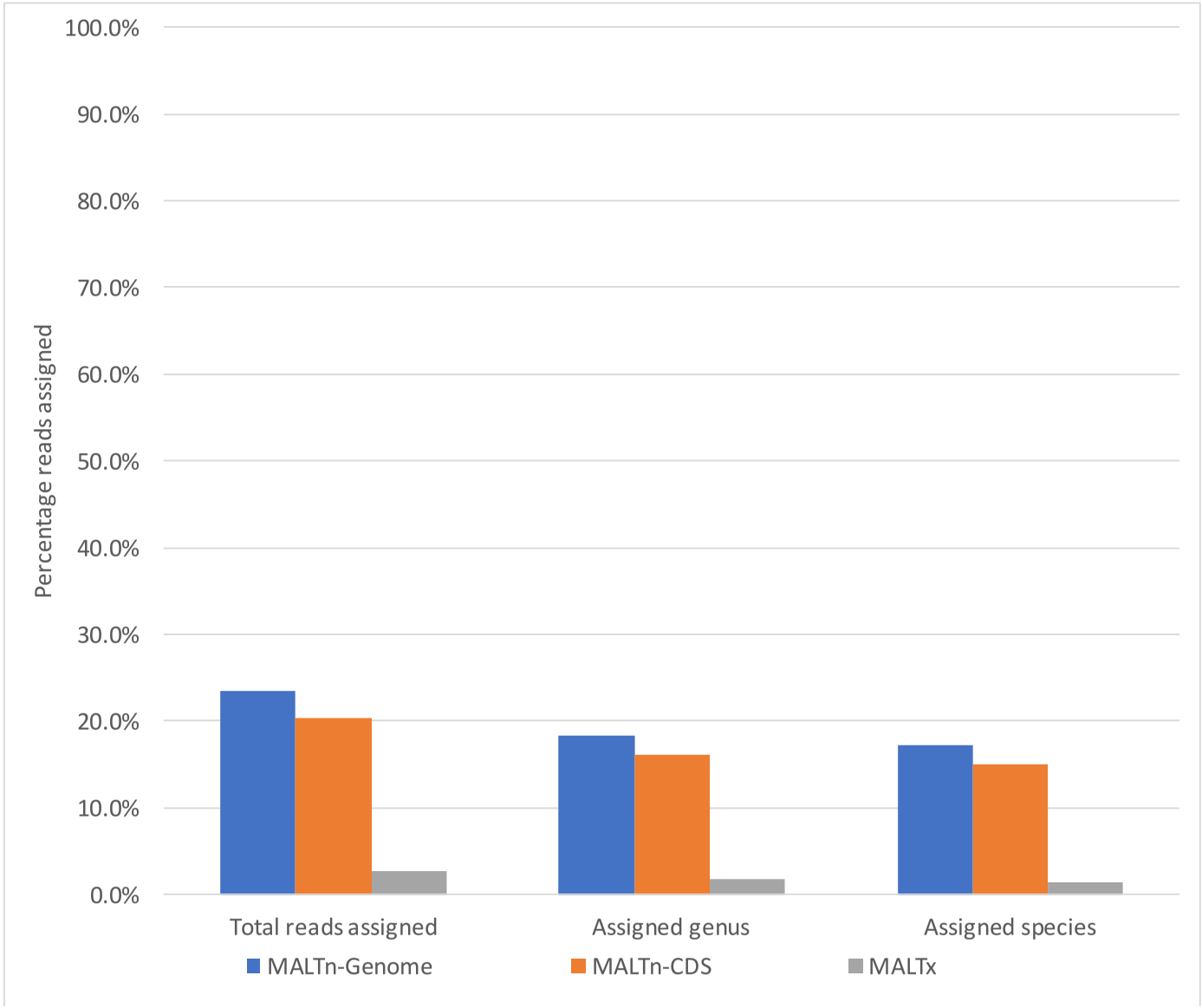


Table S4. False-positive taxa classified using different MALT databases

MALTx	MALTn-CDS	MALTn-Genome
Actinomyces radicidentis	Agrobacterium fabrum	Capnocytophaga ochracea
Agrobacterium fabrum	Bacillus thuringiensis	Streptococcus parasanguinis
Agrobacterium rhizogenes	Capnocytophaga ochracea	
Agrobacterium sp. H13-3	Fusobacterium hwasookii	
Capnocytophaga ochracea	Haemophilus influenzae	
Fusobacterium hwasookii	Legionella pneumophila	
Leptotrichia sp. oral taxon 212	Neisseria gonorrhoeae	
Neisseria elongata	Neisseria lactamica	
Neisseria gonorrhoeae	Pseudomonas putida	
Neisseria lactamica	Staphylococcus aureus	
Neisseria weaveri	Streptococcus gordonii	
Odoribacter splanchnicus		
Ottowia sp. oral taxon 894		
Prevotella enoeca		
Prevotella fusca		
Prevotella intermedia		
Prevotella melaninogenica		
Pseudopropionibacterium propionicum		
Sphingobium japonicum		
Sphingomonas sanxanigenens		
Streptococcus gordonii		
Streptococcus pneumoniae		
Streptococcus suis		
Tannerella forsythia		

Figure S3. Genus-level taxonomic assignments of simulated metagenomes

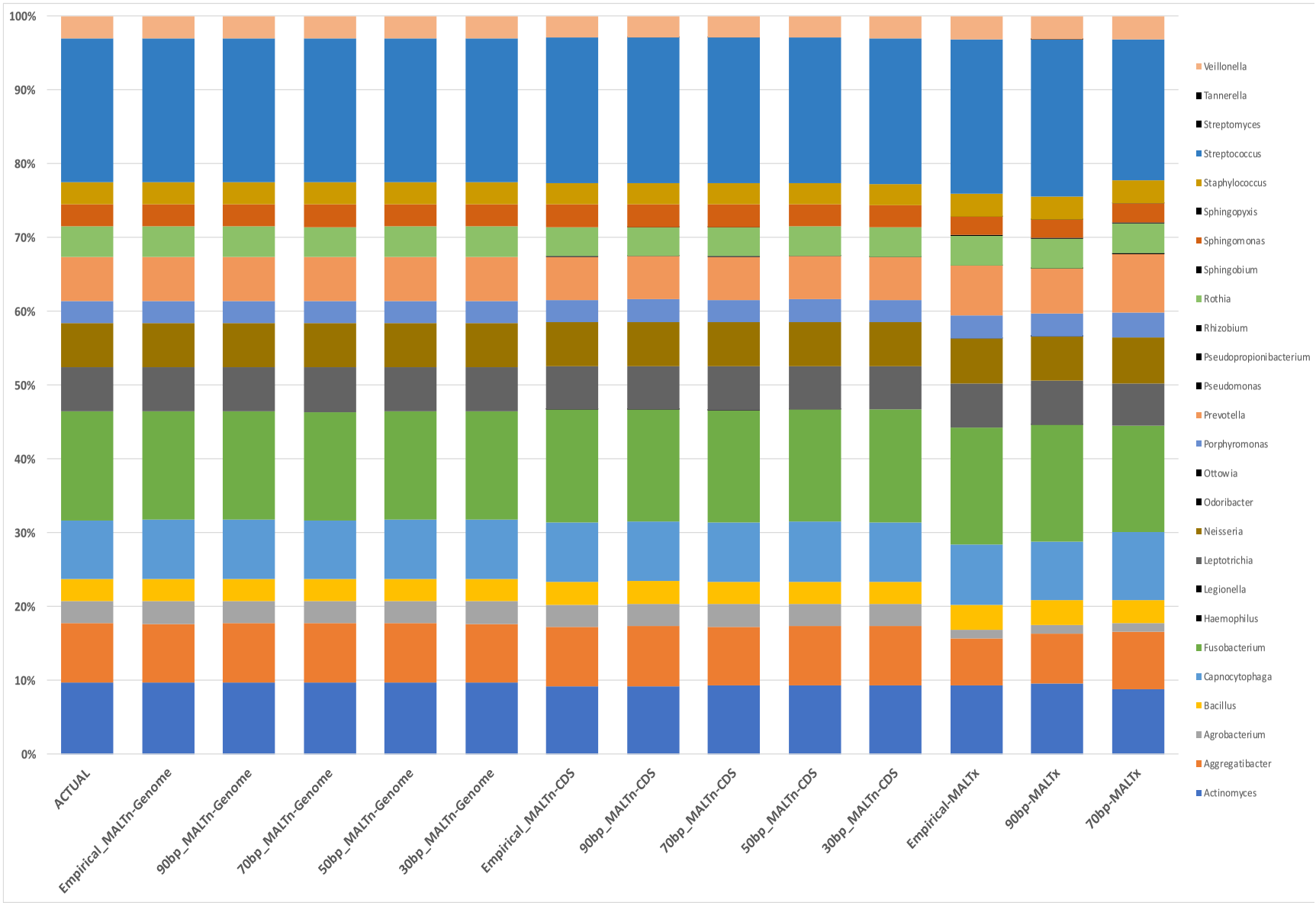


Figure S4. Species-level taxonomic assignments of simulated metagenomes

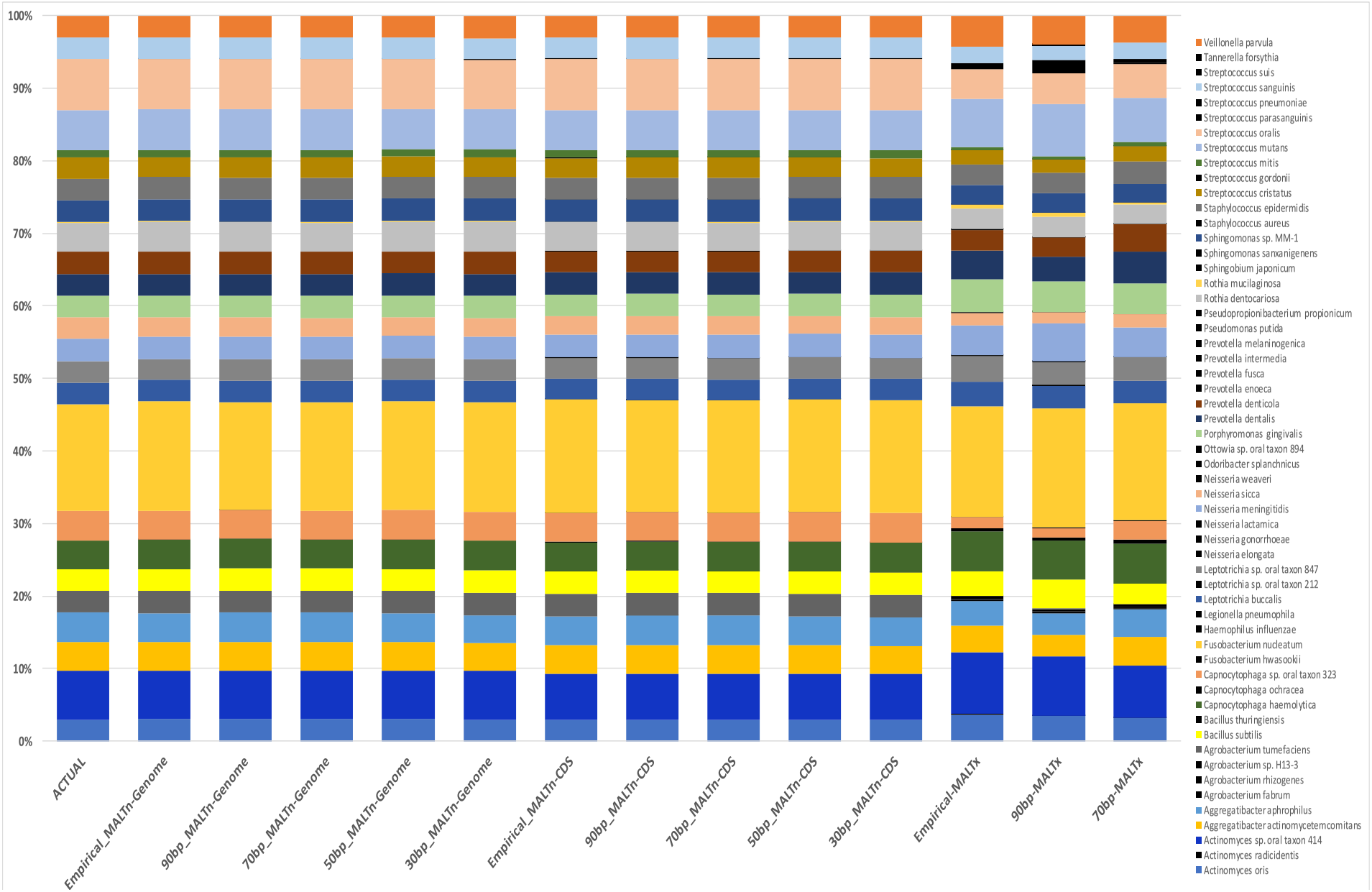


Table S5. Influence of deamination on MALTn-genome alignments

Fragment length	Reads assigned total	Reads assigned genus	Reads assigned species
30bp_MALTn-genome_0%D	99.97%	96.04%	74.55%
30bp_MALTn-genome_10%D	98.11%	94.65%	74.37%
30bp_MALTn-genome_50%D	84.21%	81.55%	65.02%
30bp_MALTn-genome_20%	98.54%	95.06%	74.78%
50bp_MALTn-genome_0%D	99.92%	97.93%	77.97%
50bp_MALTn-genome_10%D	99.94%	98.00%	78.51%
50bp_MALTn-genome_50%D	99.93%	98.03%	78.98%
50bp_MALTn-genome_20%	99.93%	98.02%	78.57%
70bp_MALTn-genome_0%D	99.95%	98.63%	82.77%
70bp_MALTn-genome_10%D	99.96%	98.63%	82.64%
70bp_MALTn-genome_50%D	99.97%	98.61%	82.32%
70bp_MALTn-genome_20%	99.97%	98.64%	82.58%
90bp_MALTn-genome_0%D	99.97%	98.75%	82.89%
90bp_MALTn-genome_10%D	99.98%	98.74%	82.80%
90bp_MALTn-genome_50%D	99.99%	98.74%	82.68%
90bp_MALTn-genome_20%	99.98%	98.77%	82.93%
Emp_MALTn-genome_0%D	98.62%	96.91%	79.30%
Emp_MALTn-genome_10%D	98.44%	96.70%	79.11%
Emp_MALTn-genome_50%D	97.72%	95.93%	78.33%
Emp_MALTn-genome_20%	98.48%	96.72%	79.05%

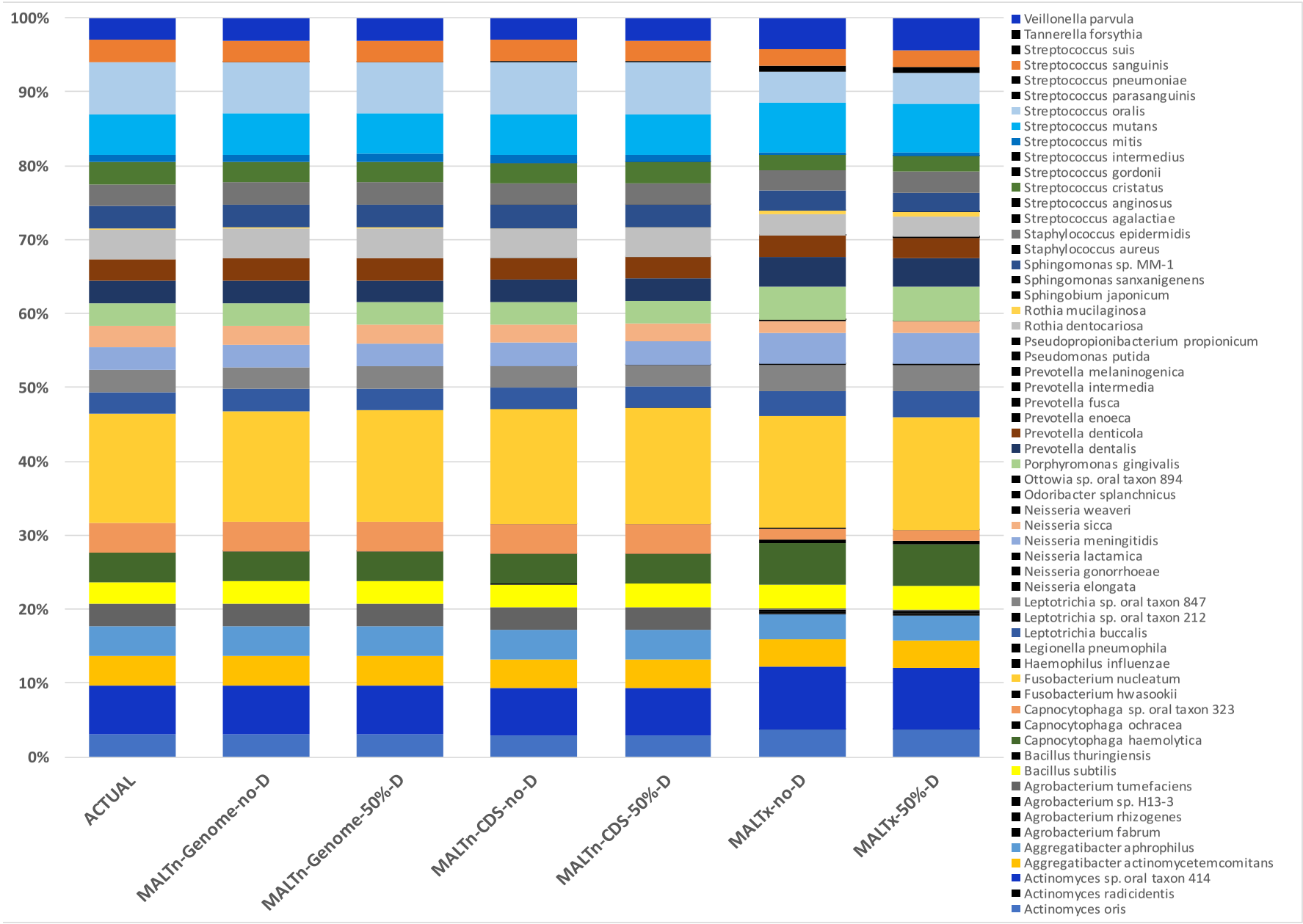
Table S6. Influence of deamination on MALTn-CDS alignments

Fragment length	Reads assigned total	Reads assigned genus	Reads assigned species
30bp_MALTn-CDS_0%D	85.80%	82.66%	63.70%
30bp_MALTn-CDS_10%D	84.19%	81.36%	63.52%
30bp_MALTn-CDS_50%D	71.81%	69.70%	55.29%
30bp_MALTn-CDS_20%	84.57%	81.74%	63.90%
50bp_MALTn-CDS_0%D	88.07%	86.48%	68.38%
50bp_MALTn-CDS_10%D	88.01%	86.48%	68.83%
50bp_MALTn-CDS_50%D	87.96%	86.45%	69.23%
50bp_MALTn-CDS_20%	88.02%	86.48%	68.88%
70bp_MALTn-CDS_0%D	89.84%	88.67%	73.88%
70bp_MALTn-CDS_10%D	89.79%	88.62%	73.73%
70bp_MALTn-CDS_50%D	89.78%	88.60%	73.45%
70bp_MALTn-CDS_20%	89.77%	88.60%	73.66%
90bp_MALTn-CDS_0%D	91.26%	90.07%	75.02%
90bp_MALTn-CDS_10%D	91.22%	90.03%	74.90%
90bp_MALTn-CDS_50%D	91.12%	89.93%	74.71%
90bp_MALTn-CDS_20%	91.18%	90.00%	75.01%
Emp_MALTn-CDS_0%D	87.44%	86.01%	69.95%
Emp_MALTn-CDS_10%D	87.22%	85.78%	69.75%
Emp_MALTn-CDS_50%D	86.52%	85.04%	69.01%
Emp_MALTn-CDS_20%	87.21%	85.75%	69.66%

Table S7. Influence of deamination on MALTx alignments

Fragment length	Reads assigned total	Reads assigned genus	Reads assigned species
30bp_MALTx_0%D	0.00%	0.00%	0.00%
30bp_MALTx_10%D	0.00%	0.00%	0.00%
30bp_MALTx_50%D	0.00%	0.00%	0.00%
30bp_MALTx_20%	0.00%	0.00%	0.00%
50bp_MALTx_0%D	0.00%	0.00%	0.00%
50bp_MALTx_10%D	0.00%	0.00%	0.00%
50bp_MALTx_50%D	0.00%	0.00%	0.00%
50bp_MALTx_20%	0.00%	0.00%	0.00%
70bp_MALTx_0%D	33.02%	29.83%	20.51%
70bp_MALTx_10%D	29.94%	27.08%	18.64%
70bp_MALTx_50%D	26.04%	23.58%	16.24%
70bp_MALTx_20%	29.21%	26.40%	18.22%
90bp_MALTx_0%D	82.17%	70.68%	40.14%
90bp_MALTx_10%D	81.82%	70.34%	39.94%
90bp_MALTx_50%D	81.70%	70.21%	39.86%
90bp_MALTx_20%	81.51%	70.09%	39.80%
Emp_MALTx_0%D	15.77%	13.84%	8.46%
Emp_MALTx_10%D	15.23%	13.37%	8.18%
Emp_MALTx_50%D	14.55%	12.77%	7.80%
Emp_MALTx_20%	15.04%	13.22%	8.06%

Figure S5. Influence of heavy deamination on taxonomic assignment at species level using empirical ancient DNA fragment length distribution metagenome



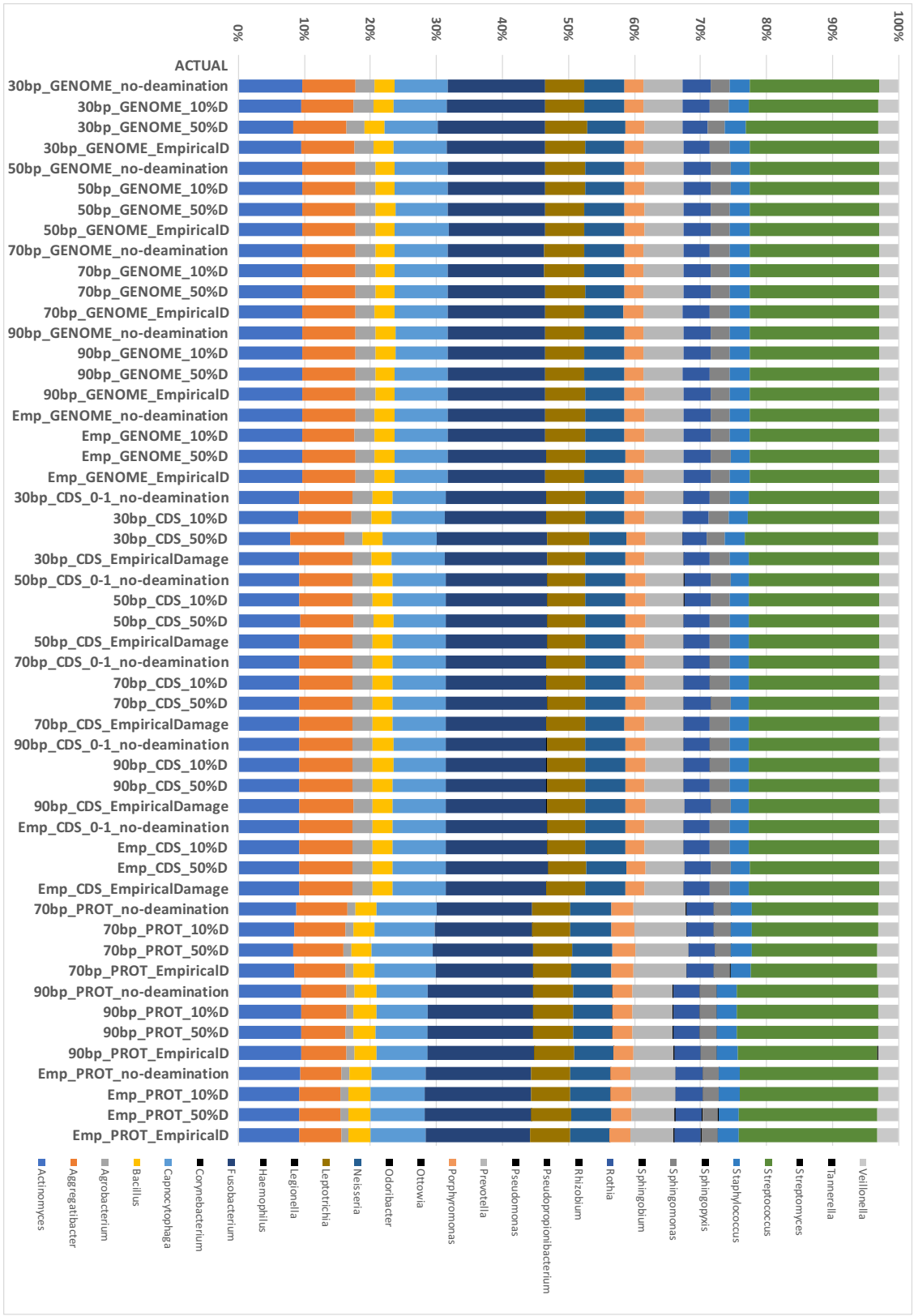


Figure S6. Influence of deamination on taxonomic assignment at genus level for all read lengths

Figure S7. Influence of deamination on taxonomic assignment at species level for all read lengths

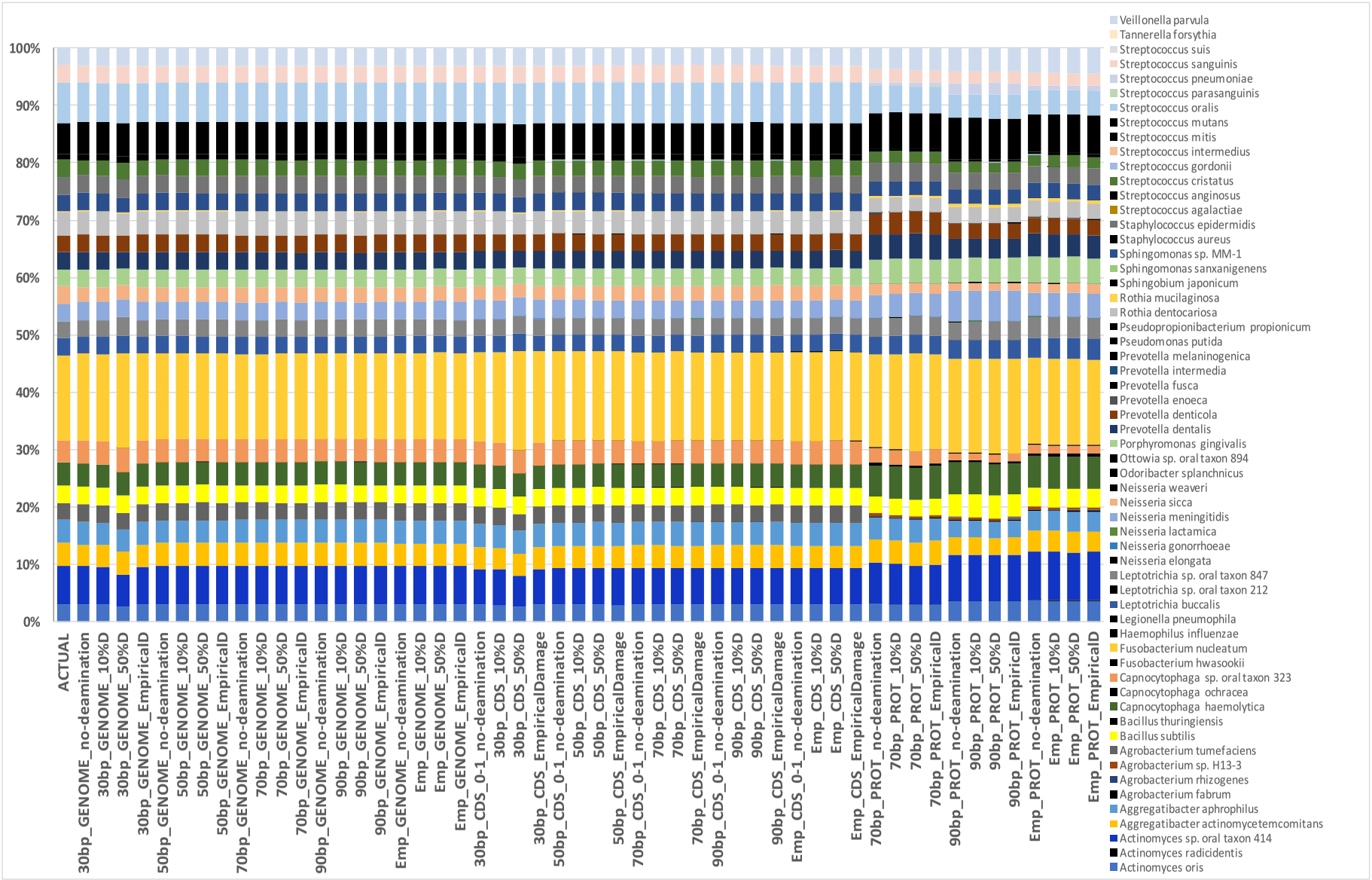


Figure S8. Influence of divergence and heavy deamination on taxonomic classification at genus level on empirical ancient DNA fragment length distribution metagenome

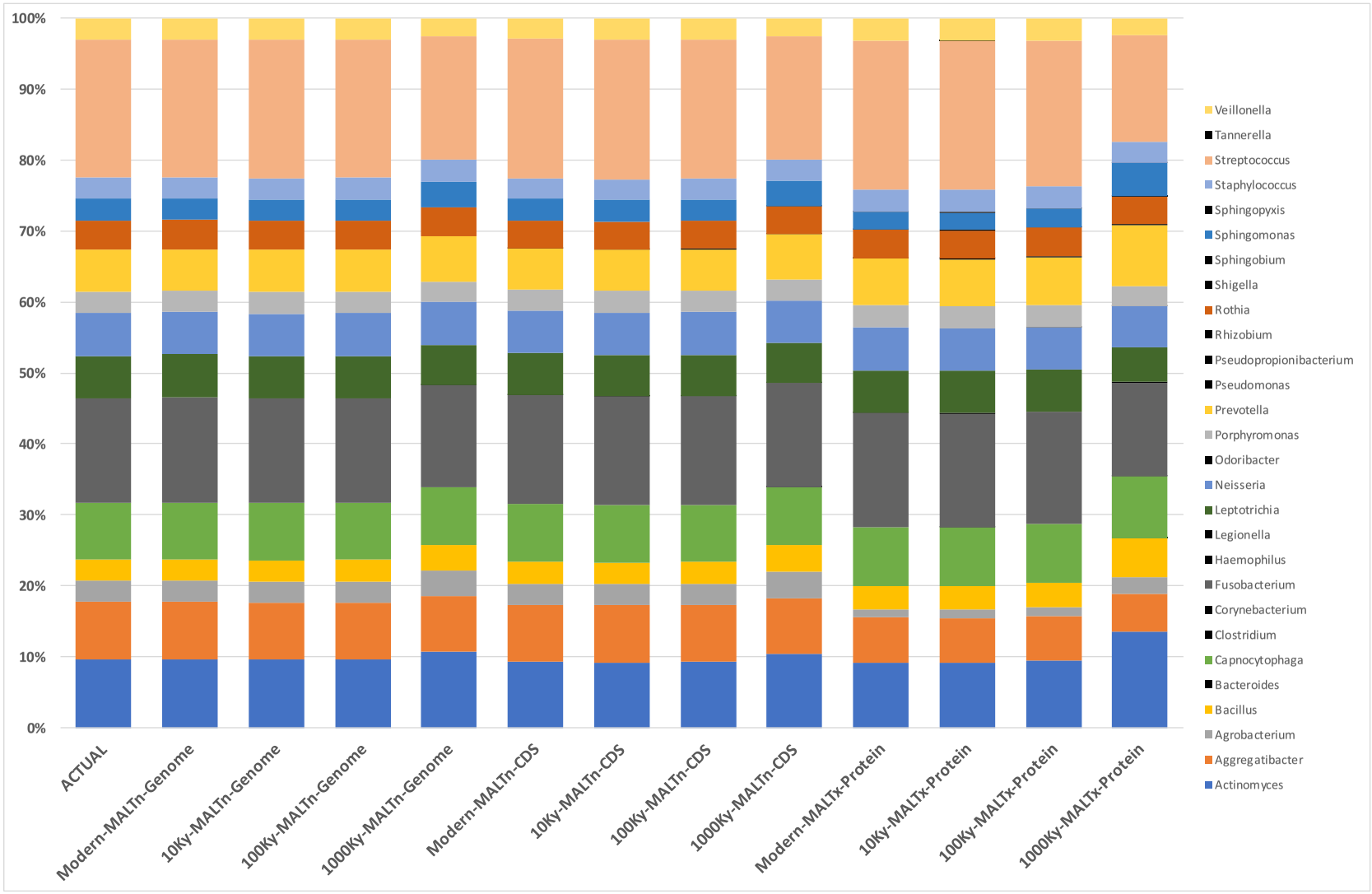


Figure S9. Influence of divergence and heavy deamination on taxonomic classification at species level on empirical ancient DNA fragment length distribution metagenome

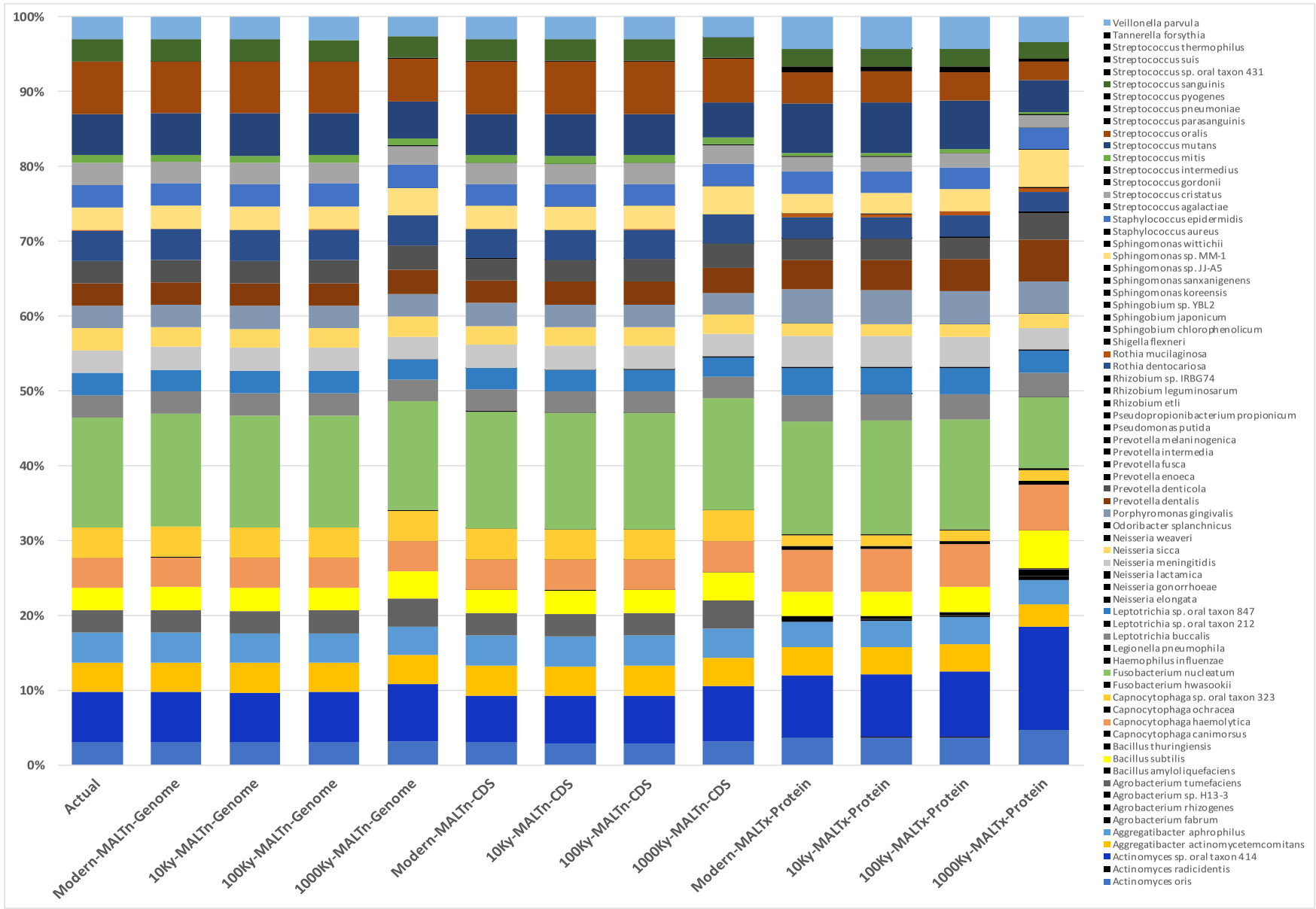


Figure S10. Read length distribution of simulated metagenome, MALTn-genome aligned reads, and unaligned reads for 1,000ky simulation

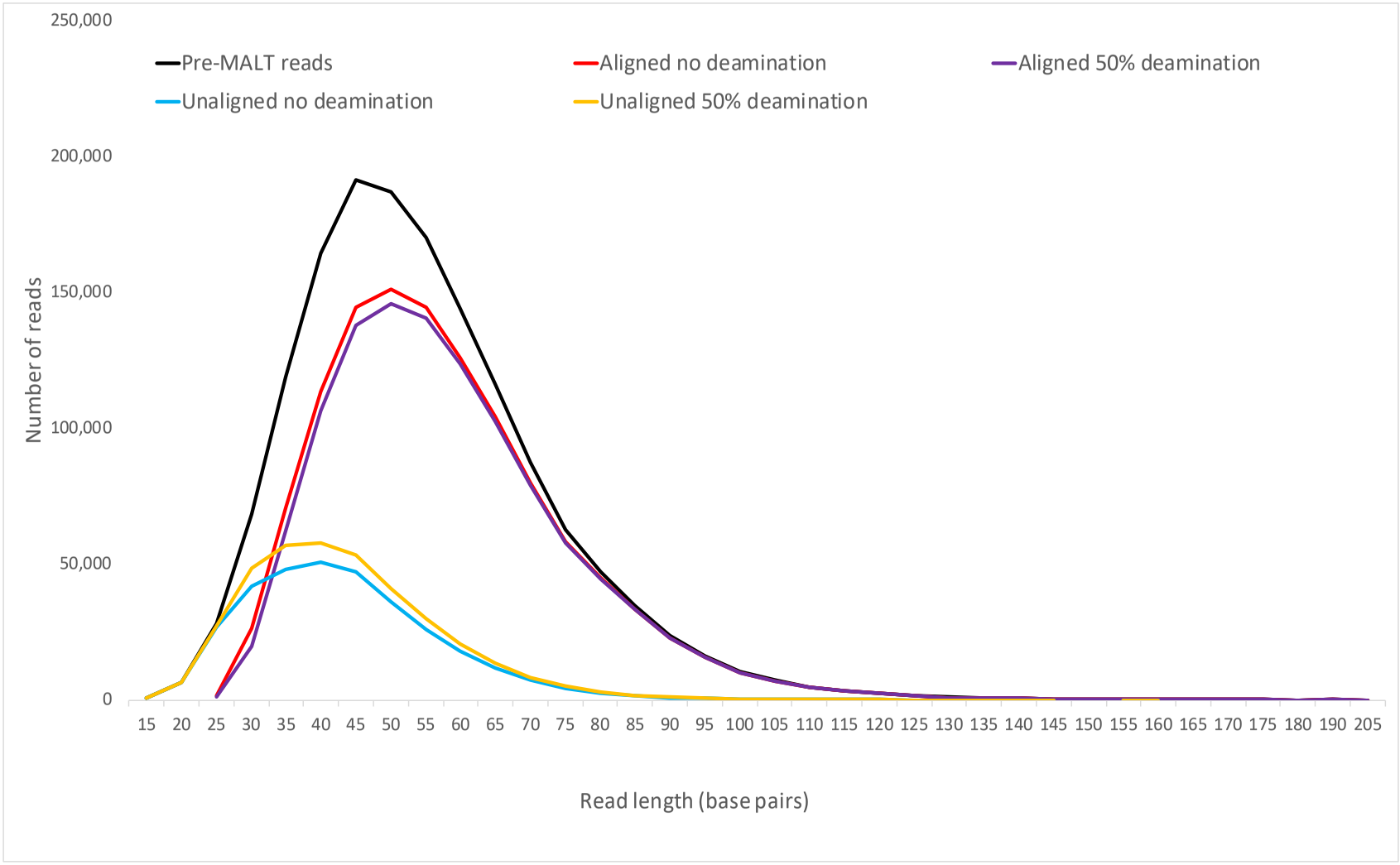


Figure S11. Average percentage of reads assigned bacterial or archaeal taxonomy to four deeply sequenced ancient samples using different databases

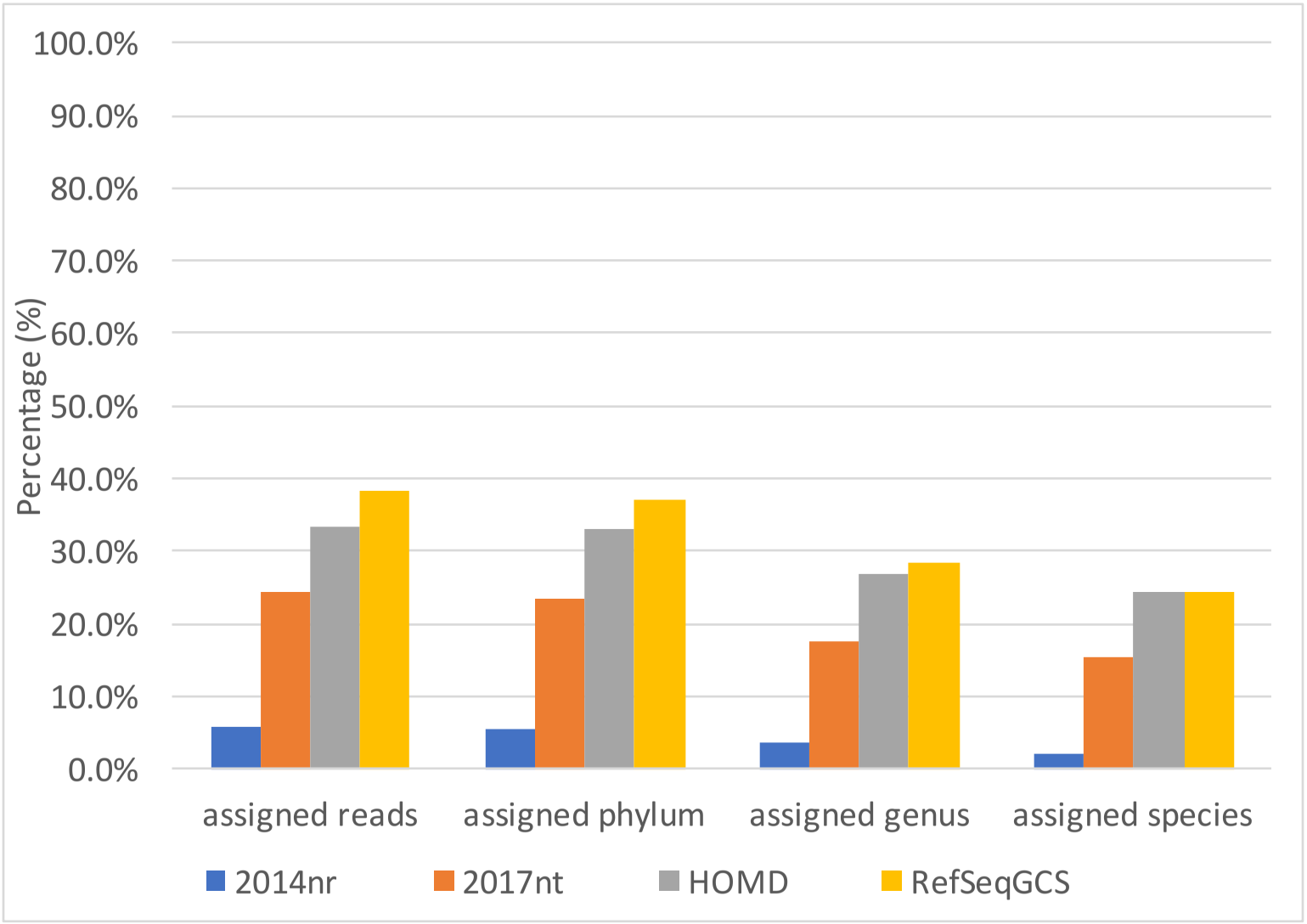
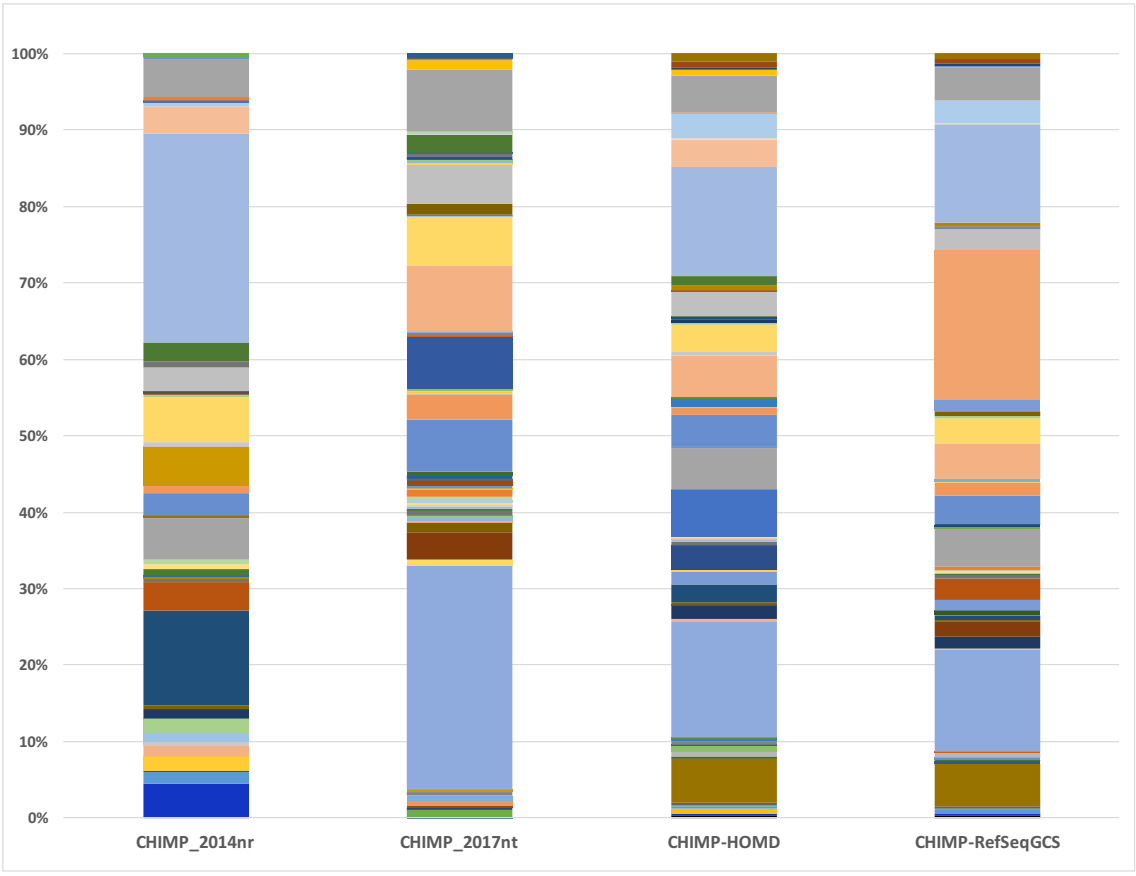
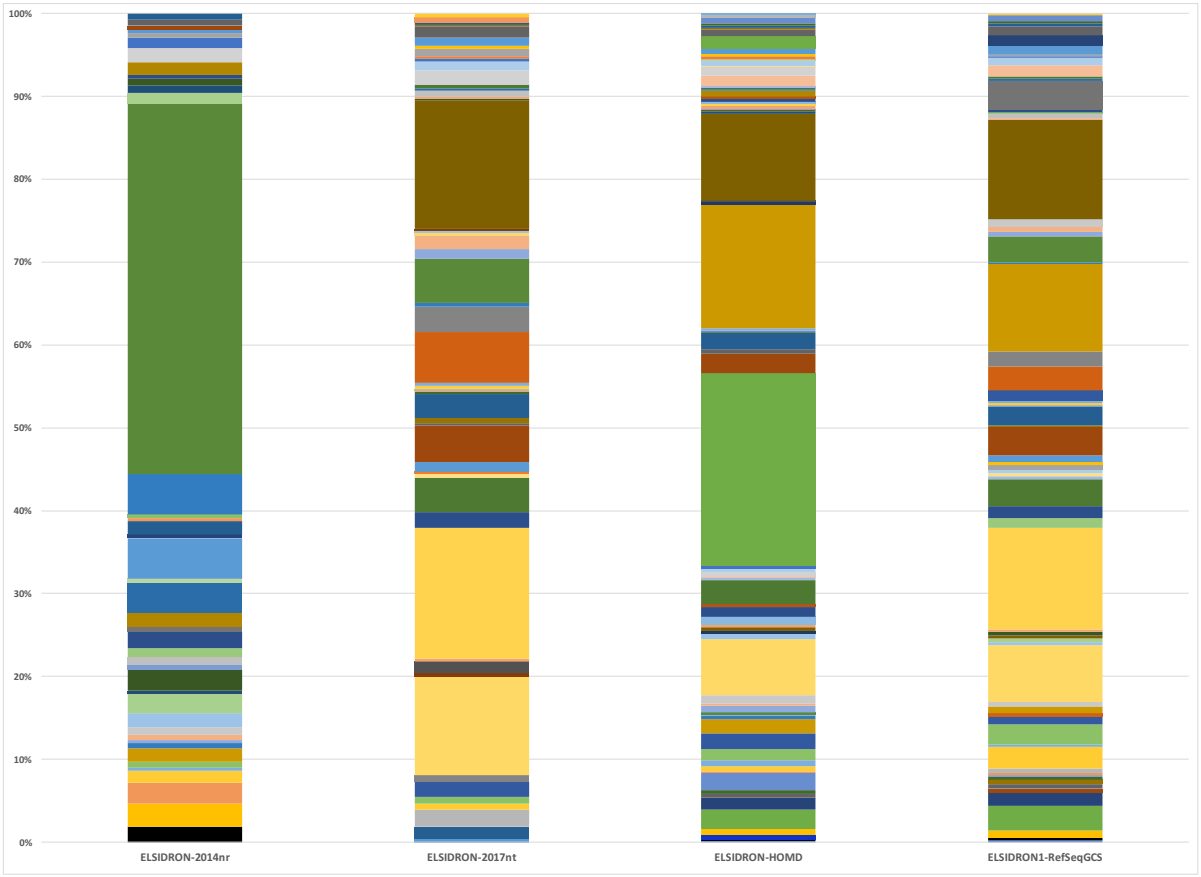


Figure 12: Species-level classification of the Chimpanzee sample



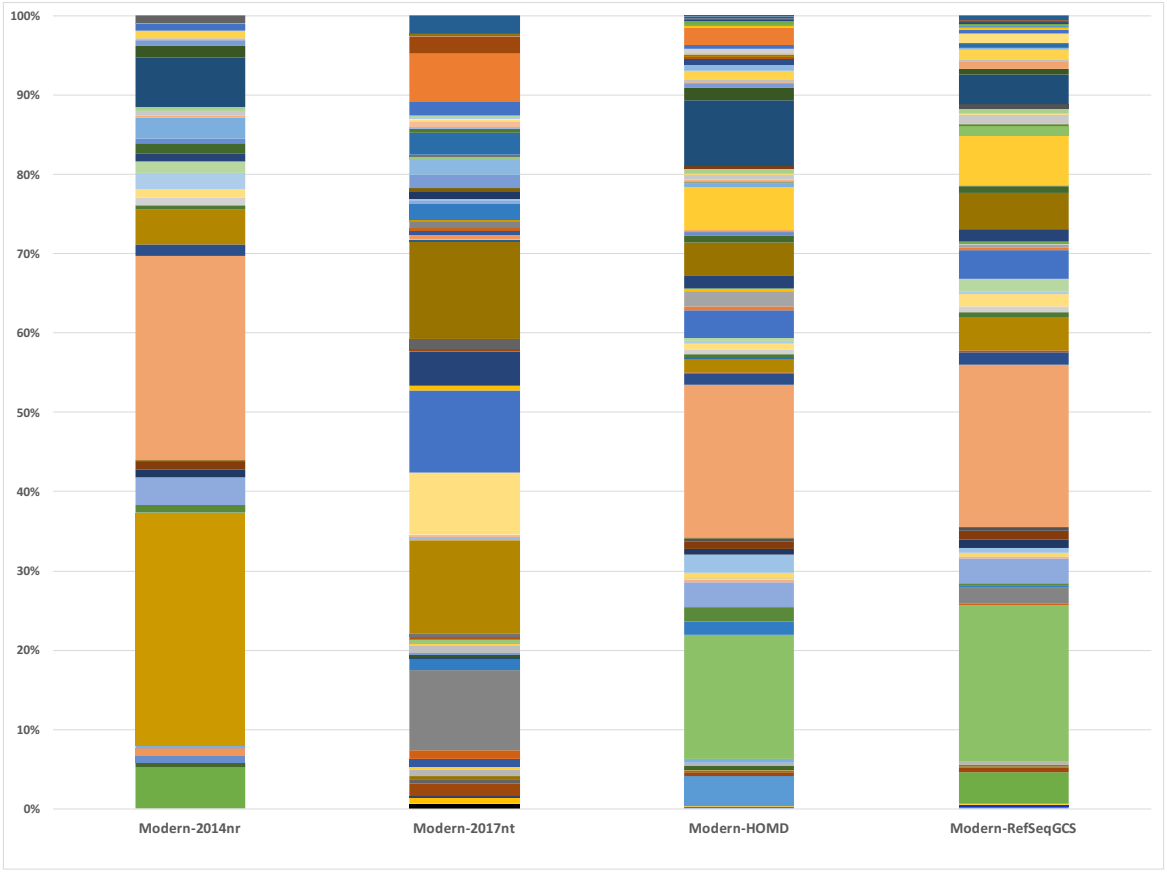
[Clostridium] saccharolyticum
[Eubacterium] brachy
[Eubacterium] infirmum
[Eubacterium] nodatum
[Eubacterium] saphenum
[Eubacterium] sulci
Acidipropionibacterium acidipropionici
Acinetobacter junii
Actinomyces cardiffensis
Actinomyces dentalis
Actinomyces georgiae
Actinomyces gerencseriae
Actinomyces glycerinitolerans
Actinomyces hongkongensis
Actinomyces israelii
Actinomyces massiliensis
Actinomyces meyeri
Actinomyces odontolyticus
Actinomyces oris
Actinomyces provencensis
Actinomyces radicidentis
Actinomyces sp. Chiba101
Actinomyces sp. oral taxon 172
Actinomyces sp. oral taxon 180
Actinomyces sp. oral taxon 414
Actinomyces sp. oral taxon 448
Actinomyces sp. oral taxon 849
Actinomyces succiniciruminis
Actinomyces turicensis
Actinomyces urogenitalis
Anaeroglobus geminatus
Anaerolineaceae bacterium oral taxon 439
Arsenicicoccus sp. oral taxon 190
Atopobium parvulum
Atopobium rimae
Atopobium sp. HMSC064B08
Atopobium sp. oral taxon 199
Bacteroides cellulosilyticus
Bacteroides fragilis
Bacteroides pyogenes
Bacteroides salanitronis
Bacteroides thetaiotaomicron
Bacteroidetes bacterium oral taxon 274
Bacteroidetes oral taxon 274
Campylobacter gracilis
candidate division TM7 single-cell isolate TM7a
candidate division TM7 single-cell isolate TM7c
Candidatus Methanomethylophilus alvus
Candidatus Saccharibacteria oral taxon TM7x
Chloroflexi bacterium oral taxon 439
Clostridium sp. SY8519
Comamonas testosteroni
Coriobacteriaceae bacterium 68-1-3
Desulfobulbus propionicus
Desulfobulbus sp. oral taxon 041
Desulfomicrobium orale
Dialister invisus
Dialister pneumosintes
Eggerthella lenta
Eggerthia catenaformis
Escherichia coli
Eubacterium limosum
Faecalibacterium prausnitzii
Faecalitalea cylindroides
Filifactor alocis
Flavonifractor plautii
Fretibacterium fastidiosum
Fusobacterium nucleatum
Gemella sp. oral taxon 928
Intestinimonas butyriciproducens
Lachnoanaerobaculum saburreum
Lachnoclostridium phocaeense
Lawsonella clevelandensis
Libanicoccus massiliensis
Megasphaera elsdenii
Mogibacterium sp. CM50
Mogibacterium timidum
Olsenella profusa
Olsenella sp. Marseille-P2300
Olsenella sp. oral taxon 807
Olsenella sp. oral taxon 809
Olsenella uli
Olsenella umbonata
Oribacterium sp. oral taxon 078
Oribacterium sp. oral taxon 108
Oscillibacter valericigenes
Parabacteroides merdae
Parvimonas micra
Parvimonas sp. oral taxon 110
Parvimonas sp. oral taxon 393
Peptostreptococcaceae bacterium oral taxon 113
Phocaeicola abscessus
Porphyromonas gingivalis
Prevotella dentalis
Prevotella denticola
Prevotella enoeca
Prevotella intermedia
Propionibacterium acidifaciens
Propionibacterium freudenreichii
Propionibacterium sp. oral taxon 192
Pseudomonas putida
Pseudopropionibacterium propionicum
Pseudoramibacter alactolyticus
Pyramidobacter piscolens
Selenomonas sp. oral taxon 136
Shuttleworthia satelles
Slackia exigua
Slackia heliotrinireducens
Slackia sp. CM382
Solobacterium moorei
Tannerella forsythia
Tannerella sp. oral taxon HOT-286
Thermoplasmatales archaeon BRNA1
Treponema denticola
Treponema lecithinolyticum
Treponema maltophilum
Treponema putidum
Treponema socranskii
uncultured bacterium

Figure 13: Species-level classification of the El Sidron1 Neanderthal



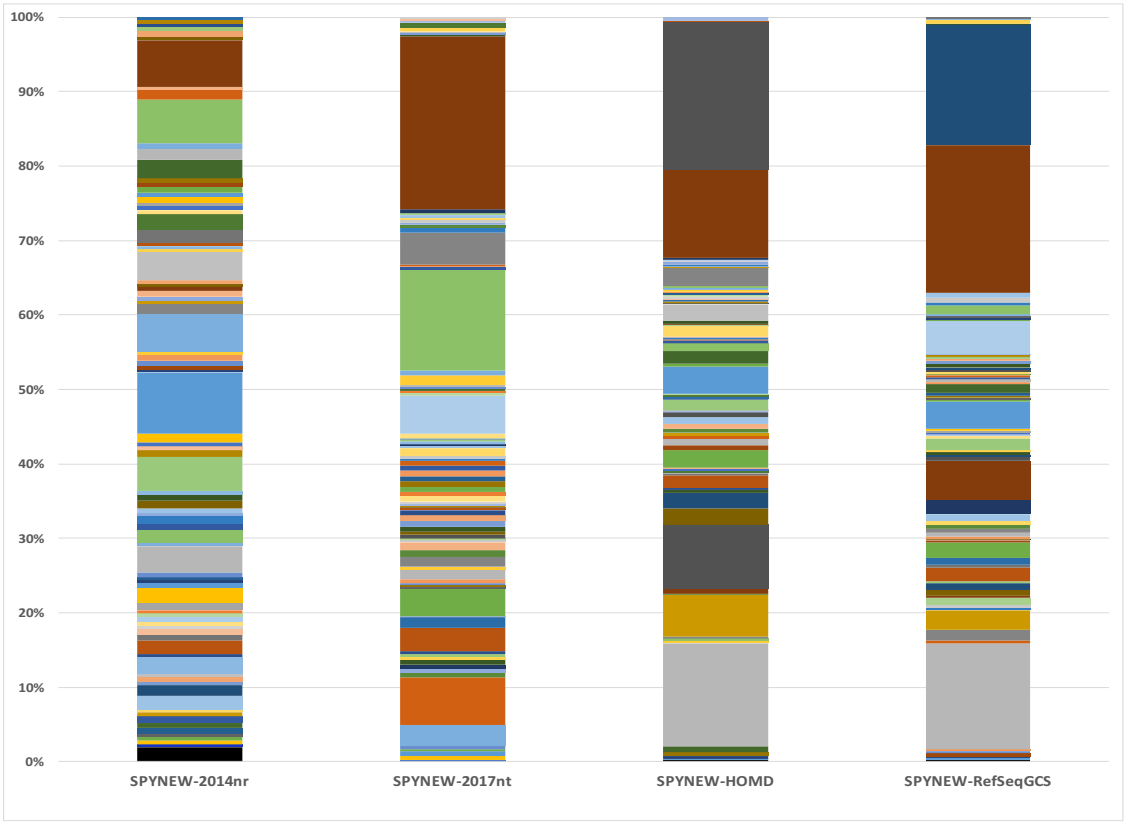
- [Eubacterium] brachy
- [Eubacterium] infirmum
- [Eubacterium] nodatum
- [Eubacterium] saphenum
- [Eubacterium] sulci
- Actinomyces cardiffensis
- Actinomyces dentalis
- Actinomyces georgiae
- Actinomyces gerencseriae
- Actinomyces glycerinitolerans
- Actinomyces hongkongensis
- Actinomyces israelii
- Actinomyces johnsonii
- Actinomyces massiliensis
- Actinomyces meyeri
- Actinomyces naeslundii
- Actinomyces odontolyticus
- Actinomyces oris
- Actinomyces radicidentis
- Actinomyces slackii
- Actinomyces sp. Chiba101
- Actinomyces sp. oral taxon 170
- Actinomyces sp. oral taxon 171
- Actinomyces sp. oral taxon 172
- Actinomyces sp. oral taxon 175
- Actinomyces sp. oral taxon 178
- Actinomyces sp. oral taxon 180
- Actinomyces sp. oral taxon 414
- Actinomyces sp. oral taxon 448
- Actinomyces sp. oral taxon 849
- Actinomyces sp. oral taxon 877
- Actinomyces sp. pika_114
- Actinomyces succiniciruminis
- Actinomyces timonensis
- Actinomyces turicensis
- Actinomyces urogenitalis
- Actinomyces viscosus
- Aggregatibacter aphrophilus
- Anaerolinea thermophila
- Anaerolineaceae bacterium oral taxon 439
- Bacteroidetes bacterium oral taxon 274
- Bacteroidetes oral taxon 274
- Campylobacter gracilis
- Campylobacter showae
- candidate division TM7 genomosp. GTL1
- candidate division TM7 single-cell isolate TM7a
- candidate division TM7 single-cell isolate TM7c
- Candidatus Saccharibacteria oral taxon TM
- Cardiobacterium valvarum
- Catonella morbi
- Chloroflexi bacterium oral taxon 439
- Coriobacteriaceae bacterium 68-1-3
- Corynebacterium matruchotii
- Cryptobacterium curtum
- Delftia acidovorans
- Denitrobacterium detoxificans
- Desulfobulbus elongatus
- Desulfobulbus mediterraneus
- Desulfobulbus propionicus
- Desulfobulbus sp. oral taxon 041
- Desulfomicrobium baculatum
- Desulfomicrobium orale
- Eggerthella lenta
- Filifactor alocis
- Fretibacterium fastidiosum
- Fusobacterium nucleatum
- Gemella morbillorum
- Gordonibacter pamelaeae
- Gordonibacter sp. Marseille-P2775
- Leptotrichia buccalis
- Leptotrichia sp. oral taxon 212
- Methanobacterium paludis
- Methanobrevibacter arboriphilus
- Methanobrevibacter millerae
- Methanobrevibacter olleyae
- Methanobrevibacter oralis
- Methanobrevibacter ruminantium
- Methanobrevibacter smithii
- Methanobrevibacter sp. AbM4
- Methanobrevibacter sp. YE315
- Methanobrevibacter wolinii
- Methanosphaera stadtmanae
- Methanothermus fervidus
- Mogibacterium sp. CM50
- Mogibacterium timidum
- Ndongobacter massiliensis
- Olsenella profusa
- Olsenella sp. oral taxon 807
- Olsenella sp. oral taxon 809
- Olsenella uli
- Oribacterium sp. oral taxon 078
- Ottowia sp. oral taxon 894
- Parvimonas micra
- Parvimonas sp. oral taxon 110
- Parvimonas sp. oral taxon 393
- Peptoanaerobacter stomatis
- Peptoniphilus sp. oral taxon 386
- Peptoniphilus sp. oral taxon 836
- Peptostreptococcaceae bacterium oral taxon 113
- Peptostreptococcus stomatis
- Porphyromonas gingivalis
- Prevotella intermedia
- Propionibacterium acidifaciens
- Propionibacterium sp. oral taxon 192
- Pseudopropionibacterium propionicum
- Selenomonas sp. oral taxon 126
- Selenomonas sputigena
- Slackia exigua
- Slackia heliotrinireducens
- Streptococcus anginosus
- Streptococcus cristatus
- Streptococcus gordonii
- Streptococcus sanguinis
- Streptococcus sinensis
- Streptococcus sp. DD04
- Subdoligranulum sp. 4_3_54A2FAA
- Tannerella forsythia
- Tannerella sp. oral taxon HOT-286
- Treponema denticola
- Treponema maltophilum
- Treponema socranskii
- uncultured bacterium
- Variovorax paradoxus
- Veillonella parvula
- Veillonella sp. AS16

Figure 14: Species-level classification of the modern sample



Abiotrophia defectiva
Actinomyces hongkongensis
Actinomyces naeslundii
Actinomyces sp. oral taxon 414
Bacteroidetes bacterium oral taxon 274
Bacteroidetes oral taxon 274
Barnesiella viscericola
Campylobacter concisus
Campylobacter curvus
Campylobacter gracilis
Campylobacter rectus
Campylobacter showae
candidate division TM7 single-cell isolate TM7a
candidate division TM7 single-cell isolate TM7c
Candidatus Saccharibacteria oral taxon TM7x
Capnocytophaga canimorsus
Capnocytophaga gingivalis
Capnocytophaga granulosa
Capnocytophaga haemolytica
Capnocytophaga ochracea
Capnocytophaga sp. ChDC OS43
Capnocytophaga sp. CM59
Capnocytophaga sp. oral taxon 323
Capnocytophaga sp. oral taxon 326
Capnocytophaga sp. oral taxon 329
Capnocytophaga sp. oral taxon 332
Capnocytophaga sp. oral taxon 336
Capnocytophaga sp. oral taxon 338
Capnocytophaga sp. oral taxon 863
Capnocytophaga sputigena
Cardiobacterium hominis
Cardiobacterium valvarum
Catonella morbi
Centipeda periodontii
Corynebacterium aquilae
Corynebacterium diphtheriae
Corynebacterium glutamicum
Corynebacterium matruchotii
Corynebacterium mustelae
Corynebacterium pseudotuberculosis
Corynebacterium ulcerans
Corynebacterium vitaeruminis
Eikenella corrodens
Fretibacterium fastidiosum
Fusobacterium hwasookii
Fusobacterium nucleatum
Fusobacterium sp. oral taxon 370
Gemella morbillorum
Gemella sp. oral taxon 928
Haemophilus parainfluenzae
Lautropia mirabilis
Leptotrichia buccalis
Leptotrichia goodfellowii
Leptotrichia hofstadii
Leptotrichia sp. oral taxon 212
Leptotrichia sp. oral taxon 215
Leptotrichia sp. oral taxon 225
Leptotrichia sp. oral taxon 847
Leptotrichia trevisanii
Leptotrichia wadei
Neisseria elongata
Neisseria meningitidis
Neisseria sicca
Ottowia sp. oral taxon 894
Porphyromonas asaccharolytica
Porphyromonas catoniae
Porphyromonas endodontalis
Porphyromonas gingivalis
Porphyromonas sp. KLE 1280
Porphyromonas sp. oral taxon 278
Porphyromonas sp. oral taxon 279
Prevotella conceptionensis
Prevotella dentalis
Prevotella denticola
Prevotella enoeca
Prevotella fusca
Prevotella intermedia
Prevotella loescheii
Prevotella melaninogenica
Prevotella micans
Prevotella nigrescens
Prevotella oris
Prevotella ruminicola
Prevotella saccharolytica
Prevotella scopos
Prevotella shahii
Prevotella sp. HMSC073D09
Prevotella sp. oral taxon 299
Prevotella sp. oral taxon 317
Prevotella sp. oral taxon 472
Pseudopropionibacterium propionicum
Selenomonas artemidis
Selenomonas infelix
Selenomonas noxia
Selenomonas sp. oral taxon 126
Selenomonas sp. oral taxon 136
Selenomonas sp. oral taxon 137
Selenomonas sp. oral taxon 138
Selenomonas sp. oral taxon 478
Selenomonas sp. oral taxon 892
Selenomonas sp. oral taxon 920
Selenomonas sputigena
Streptococcus cristatus
Streptococcus gordonii
Streptococcus mitis
Streptococcus oralis
Streptococcus pneumoniae
Streptococcus sp. NPS 308
Tannerella forsythia
Tannerella sp. oral taxon HOT-286
Treponema denticola
Treponema lecithinolyticum
Treponema maltophilum
Treponema medium
Treponema socranskii
Treponema sp. OMZ 838
Treponema vincentii
uncultured bacterium
Veillonella parvula
Veillonella sp. oral taxon 158

Figure 15: Species-level classification of the Spy II Neanderthal



[Clostridium] cellulolyticum
[Eubacterium] infirmum
[Eubacterium] saphenum
[Eubacterium] siraeum
[Eubacterium] sulci
[Ruminococcus] torques
Abiotrophia defectiva
Abiotrophia sp. HMSC24B09
Acetonema longum
Achromobacter xylosoxidans
Acidipropionibacterium acidipropionici
Acinetobacter baumannii
Acinetobacter junii
Acinetobacter venetianus
Actinomyces cardiffensis
Actinomyces georgiae
Actinomyces hongkongensis
Actinomyces johnsonii
Actinomyces massiliensis
Actinomyces meyeri
Actinomyces naeslundii
Actinomyces odontolyticus
Actinomyces oris
Actinomyces radicidentis
Actinomyces radingae
Actinomyces sp. HMSC035G02
Actinomyces sp. HPA0247
Actinomyces sp. ICM39
Actinomyces sp. ICM58
Actinomyces sp. Marseille-P2825
Actinomyces sp. Marseille-P2985
Actinomyces sp. oral taxon 170
Actinomyces sp. oral taxon 171
Actinomyces sp. oral taxon 172
Actinomyces sp. oral taxon 180
Actinomyces sp. oral taxon 414
Actinomyces sp. oral taxon 448
Actinomyces sp. oral taxon 848
Actinomyces sp. oral taxon 849
Actinomyces succiniciruminis
Actinomyces turicensis
Adlercreutzia equolifaciens
Aggregatibacter actinomycetemcomitans
Aggregatibacter aphrophilus
Aggregatibacter segnis
Aggregatibacter sp. oral taxon 458
Anaerolineaceae bacterium oral taxon 439
Arsenicicoccus sp. oral taxon 190
Arthrobacter saudimassiliensis
Atopobium rimae
Azorhizobium caulinodans
Bacillus coagulans
Bifidobacterium thermophilum
Brachybacterium faecium
Bradyrhizobium elkanii
Campylobacter showae
candidate division TM7 single-cell isolate TM7a
candidate division TM7 single-cell isolate TM7c
Candidatus Koribacter versatilis
Candidatus Saccharibacteria oral taxon TM7x
Capnocytophaga sp. oral taxon 324
Catonella morbi
Christensenella massiliensis
Chryseobacterium indologenes
Clostridium sp. BNL1100
Clostridium sp. Marseille-P3244
Clostridium sp. SY8519
Comamonas aquatica
Comamonas testosteroni
Coriobacteriaceae bacterium 68-1-3
Corynebacterium durum
Cryptobacterium curtum
Cutibacterium acnes
Delftia acidovorans
Denitrobacterium detoxificans
Desulfobulbus sp. oral taxon 041
Desulfomicrobium baculatum
Desulfomicrobium orale
Dialister succinatiphilus
Eggerthella lenta
Eggerthella sp. YY7918
Eggerthellaceae bacterium AT8
Eikenella corrodens
Enterobacter cloacae
Enterococcus faecalis
Enterococcus faecium
Escherichia coli
Ethanoligenens harbinense
Eubacterium callanderi
Faecalibacterium prausnitzii
Filifactor alocis
Flavonifractor plautii
Gemella bergeri
Gemella cuniculi
Gemella haemolysans
Gemella morbillorum
Gemella sp. oral taxon 928
Gemmata sp. SH-PL17
Gordonia polyisoprenivorans
Gordonibacter pamelaeae
Granulicatella adiacens
Haemophilus haemolyticus
Haemophilus influenzae
Helicobacter pylori
Histophilus somni
Intestinimonas butyriciproducens
Intestinimonas massiliensis
Johnsonella ignava
Lachnoanaerobaculum saburreum
Lachnoclostridium phocaeense
Lachnospiraceae bacterium 1_1_57FAA
Lachnospiraceae bacterium oral taxon 500
Lautropia mirabilis
Leptotrichia sp. oral taxon 212
Marvinbryantia formatexigens
Methanobacterium sp. Maddingley MBC34
Methanobrevibacter arboriphilus
Methanobrevibacter millerae
Methanobrevibacter olleyae
Methanobrevibacter oralis
Methanobrevibacter ruminantium
Methanobrevibacter smithii
Methanobrevibacter wolinii
Methylobacterium nodulans
Mogibacterium sp. CM50
Mogibacterium timidum
Neisseria elongata
Neisseria sicca
Neisseria sp. oral taxon 014
Nitrospira defluvii
Olsenella sp. oral taxon 807
Oribacterium asaccharolyticum
Oribacterium sinus
Oribacterium sp. oral taxon 078
Oscillibacter valericigenes
Ottowia sp. oral taxon 894
Paenibacillus polymyxa
Parascardovia denticolens
Parvimonas micra
Parvimonas sp. oral taxon 393
Pediococcus acidilactici
Pelosinus fermentans
Peptoanaerobacter stomatis
Peptoniphilus sp. oral taxon 386
Peptostreptococcaceae bacterium oral taxon 113
Peptostreptococcus anaerobius
Peptostreptococcus stomatis
Photorhabdus luminescens
Polaromonas naphthalenivorans
Polaromonas sp. JS666
Porphyromonas gingivalis
Prevotella intermedia
Propionibacterium freudenreichii
Propionibacterium sp. oral taxon 192
Propionibacterium sp. oral taxon 193
Pseudoflavonifractor capillosus
Pseudoflavonifractor sp. Marseille-P3106
Pseudomonas aeruginosa
Pseudomonas fluorescens
Pseudomonas putida
Pseudomonas tolaasii
Pseudopropionibacterium propionicum
Pseudoxanthomonas suwonensis
Ramlibacter tataouinensis
Riemerella anatipestifer
Roseburia intestinalis
Roseburia inulinivorans
Ruminococcaceae bacterium D16
Ruminococcus flavefaciens
Ruminococcus sp. SR1/5
Sanguibacter keddieii
Schlesneria paludicola
Slackia exigua
Slackia heliotrinireducens
Staphylococcus aureus
Stenotrophomonas maltophilia
Streptococcus agalactiae
Streptococcus anginosus
Streptococcus constellatus
Streptococcus cristatus
Streptococcus dysgalactiae
Streptococcus equi
Streptococcus gordonii
Streptococcus infantis
Streptococcus intermedius
Streptococcus marmotae
Streptococcus mitis
Streptococcus mutans
Streptococcus oralis
Streptococcus parasanguinis
Streptococcus pneumoniae
Streptococcus pyogenes
Streptococcus salivarius
Streptococcus sanguinis
Streptococcus sinensis
Streptococcus sp. 2_1_36FAA
Streptococcus sp. DD04
Streptococcus sp. I-G2
Streptococcus sp. NPS 308
Streptococcus sp. oral taxon 056
Streptococcus sp. oral taxon 431
Streptococcus suis
Subdoligranulum sp. 4_3_54A2FAA
Syntrophothermus lipocalidus
Tannerella forsythia
Tannerella sp. oral taxon HOT-286
Thermoanaerobacter siderophilus
Treponema caldarium
Treponema phagedenis
uncultured bacterium
Variovorax paradoxus
Variovorax sp. PAMC 28711
Veillonella parvula

Table S8. Species assignments specific to databases

2014nr-specific	2017nt-specific	HOMD-specific	RefSeqGCS-specific
Candidatus Koribacter versatilis	uncultured bacterium	Bacteroides pyogenes	Porphyromonas sp. KLE 1280
Parabacteroides merdae	Bacteroides cellulosilyticus	Prevotella shahii	Prevotella conceptionensis
Capnocytophaga sp. CM59	Bacteroides fragilis	Capnocytophaga sp. oral taxon 336	Prevotella sp. HMSC073D09
Capnocytophaga sp. oral taxon 324	Bacteroides salanitronis	Capnocytophaga sputigena	Phocaeicola abscessus
Riemerella anatipestifer	Bacteroides thetaiotaomicron	Bacteroidetes bacterium oral taxon 274	Leptotrichia trevisanii
Nitrospira defluvii	Barnesiella viscericola	Fusobacterium sp. oral taxon 370	Desulfobulbus elongatus
Methylobacterium nodulans	Porphyromonas asaccharolytica	Bradyrhizobium elkanii	Desulfobulbus mediterraneus
Azorhizobium caulinodans	Prevotella dentalis	Achromobacter xylosoxidans	Acinetobacter venetianus
Neisseria sp. oral taxon 014	Prevotella denticola	Delftia acidovorans	Actinomyces glycerinitolerans
Desulfomicrobium baculatum	Prevotella fusca	Desulfobulbus sp. oral taxon 041	Actinomyces provencensis
Helicobacter pylori	Prevotella melaninogenica	Campylobacter rectus	Actinomyces slackii
Photorhabdus luminescens	Prevotella ruminicola	Aggregatibacter sp. oral taxon 458	Actinomyces sp. HMSC035G02
Pseudoxanthomonas suwonensis	Prevotella scopos	Haemophilus haemolyticus	Actinomyces sp. HPA0247
Schlesneria paludicola	Prevotella sp. oral taxon 299	Pseudomonas aeruginosa	Actinomyces sp. Marseille-P2825
Treponema caldarium	Capnocytophaga canimorsus	Pseudomonas fluorescens	Sanguibacter keddieii
Treponema phagedenis	Capnocytophaga haemolytica	Stenotrophomonas maltophilia	Atopobium sp. HMSC064B08
Actinomyces sp. ICM39	Chryseobacterium indologenes	Actinomyces sp. oral taxon 877	Eggerthellaceae bacterium AT8
Actinomyces sp. ICM58	Fusobacterium hwasookii	Parascardovia denticolens	Gemella cuniculi
Actinomyces sp. oral taxon 848	Ramlibacter tataouinensis	Arsenicicoccus sp. oral taxon 190	Abiotrophia sp. HMSC24B09
Actinomyces turicensis	Variovorax sp. PAMC 28711	Propionibacterium acidifaciens	Enterococcus faecalis
Actinomyces viscosus	Neisseria meningitidis	Olsenella profusa	Enterococcus faecium
Bifidobacterium thermophilum	Enterobacter cloacae	Chloroflexi bacterium oral taxon 439	Pediococcus acidilactici
Corynebacterium durum	Haemophilus influenzae	Gemella bergeri	Streptococcus sp. DD04
Brachybacterium faecium	Haemophilus parainfluenzae	Granulicatella adiacens	Clostridium sp. Marseille-P3244
Cutibacterium acnes	Histophilus somni	Streptococcus infantis	Eubacterium callanderi
Cryptobacterium curtum	Gemmata sp. SH-PL17	Streptococcus sinensis	Oribacterium asaccharolyticum
Slackia sp. CM382	Treponema putidum	Mogibacterium timidum	Lachnospiraceae bacterium 1_1_57FAA

2014nr-specific	2017nt-specific	HOMD-specific	RefSeqGCS-specific
Anaerolinea thermophila	Actinomyces hongkongensis	Johnsonella ignava	Peptoanaerobacter stomatis
Bacillus coagulans	Actinomyces radingae	Oribacterium sp. oral taxon 108	Peptostreptococcaceae bacterium oral taxon 113
Gemella haemolysans	Actinomyces sp. Chiba101	Selenomonas sp. oral taxon 137	Ruminococcus flavefaciens
Streptococcus mutans	Actinomyces sp. Marseille-P2985	Selenomonas sp. oral taxon 138	Intestinimonas massiliensis
Streptococcus sp. 2_1_36FAA	Actinomyces sp. pika_114	Selenomonas sp. oral taxon 892	Pseudoflavonifractor sp. Marseille-P3106
Streptococcus sp. oral taxon 056	Actinomyces succiniciruminis	Veillonella sp. AS16	Thermoanaerobacter siderophilus
Clostridium sp. BNL1100	Corynebacterium aquilae	Veillonella sp. oral taxon 158	Eggerthia catenaformis
Mogibacterium sp. CM50	Corynebacterium diphtheriae	Parvimonas sp. oral taxon 110	Methanobrevibacter arboriphilus
Marvinbryantia formatexigens	Corynebacterium glutamicum	Peptoniphilus sp. oral taxon 836	Methanobrevibacter wolinii
Oribacterium sinus	Corynebacterium mustelae		
Roseburia intestinalis	Corynebacterium pseudotuberculosis		
Roseburia inulinivorans	Corynebacterium ulcerans		
Ruminococcus sp. SR1/5	Corynebacterium vitaeruminis		
Syntrophothermus lipocalidus	Gordonia polyisoprenivorans		
Pseudoflavonifractor capillosus	Lawsonella clevelandensis		
Faecalitalea cylindroides	Arthrobacter saudimassiliensis		
Acetonema longum	Libanicoccus massiliensis		
Pelosinus fermentans	Olsenella sp. Marseille-P2300		
Dialister succinatiphilus	Olsenella umbonata		
candidate division TM7 genomosp. GTL1	Eggerthella sp. YY7918		
candidate division TM7 single-cell isolate TM7a	Gordonibacter sp. Marseille-P2775		
candidate division TM7 single-cell isolate TM7c	Paenibacillus polymyxa		
Methanobacterium paludis	Staphylococcus aureus		
Methanobacterium sp. Maddingley MBC34	Streptococcus dysgalactiae		
	Streptococcus marmotae		
	Streptococcus parasanguinis		
	Streptococcus pyogenes		
	Streptococcus sp. NPS 308		

2014nr-specific	2017nt-specific	HOMD-specific	RefSeqGCS-specific
	Streptococcus sp. oral taxon 431		
	Christensenella massiliensis		
	Eubacterium limosum		
	Lachnoclostridium phocaeense		
	Flavonifractor plautii		
	Selenomonas sp. oral taxon 136		
	Selenomonas sp. oral taxon 478		
	Dialister pneumosintes		
	Megasphaera elsdenii		
	Ndongobacter massiliensis		
	Methanosphaera stadtmanae		
	Methanothermus fervidus		

Table S9. Sequencing and alignment statistics for the reanalysis of previously published ancient dental calculus

Sample/database	Mean read length	# total reads	# reads assigned taxonomy	% total reads assigned taxonomy	% total reads unassigned	Fold-increase in number of reads assigned over 2014nr
A11105_AfrSudan1-2014nr	41	5,246	98	1.9%	98.1%	
A11105_AfrSudan1-RefSeqGCS	41	5,246	1,688	32.2%	67.8%	17.2
A11106_AfrSudan2-2014nr	34	81,155	202	0.2%	99.8%	
A11106_AfrSudan2-RefSeqGCS	34	81,155	3,545	4.4%	95.6%	17.5
A12014_EuroHG1-2014nr	53	145,907	2,188	1.5%	98.5%	
A12014_EuroHG1-RefSeqGCS	53	145,907	57,071	39.1%	60.9%	26.1
A12017_EuroHG2-2014nr	51	93,208	1,184	1.3%	98.7%	
A12017_EuroHG2-RefSeqGCS	51	93,208	35,762	38.4%	61.6%	30.2
A12824_LBK3-2014nr	48	39,463	330	0.8%	99.2%	
A12824_LBK3-RefSeqGCS	48	39,463	17,716	44.9%	55.1%	53.7
A12826_LBK1-2014nr	47	125,570	781	0.6%	99.4%	
A12826_LBK1-RefSeqGCS	47	125,570	45,990	36.6%	63.4%	58.9
A12829_LBK2-2014nr	47	158,406	1,351	0.9%	99.1%	
A12829_LBK2-RefSeqGCS	47	158,406	78,905	49.8%	50.2%	58.4
A13204_AfrPP2-2014nr	43	821,114	1,894	0.2%	99.8%	
A13204_AfrPP2-RefSeqGCS	43	821,114	233,981	28.5%	71.5%	123.5
A13209_AfrSF3-2014nr	41	3,391,497	2,256	0.1%	99.9%	
A13209_AfrSF3-RefSeqGCS	41	3,391,497	1,185,540	35.0%	65.0%	525.5
A13210-AfrSF4-2014nr	44	308,202	532	0.2%	99.8%	
A13210-AfrSF4-RefSeqGCS	44	308,202	69,405	22.5%	77.5%	130.5
A13213_AfrPP1-2014nr	44	2,642,915	7,146	0.3%	99.7%	
A13213_AfrPP1-RefSeqGCS	44	2,642,915	853,865	32.3%	67.7%	119.5
A13232_IndRev1-2014nr	49	116,542	2,892	2.5%	97.5%	
A13232_IndRev1-RefSeqGCS	49	116,542	62,908	54.0%	46.0%	21.8
A13234_IndRev2-2014nr	44	11,464,274	85,625	0.7%	99.3%	
A13234_IndRev2-RefSeqGCS	44	11,464,274	6,080,790	53.0%	47.0%	71.0
A13344_ET11_EBC-2014nr	30	238,422	65	0.03%	100.0%	
A13344_ET11_EBC-RefSeqGCS	30	238,422	4,570	1.9%	98.1%	70.3
A8812_JewBury1-2014nr	51	51,926	875	1.7%	98.3%	
A8812_JewBury1-RefSeqGCS	51	51,926	27,054	52.1%	47.9%	30.9
A8824_JewBury2-2014nr	50	71,125	929	1.3%	98.7%	
A8824_JewBury2-RefSeqGCS	50	71,125	36,979	52.0%	48.0%	39.8
AFR8-EBC-2014nr	39	9,076	55	0.6%	99.4%	
AFR8-EBC-RefSeqGCS	39	9,076	6,034	66.5%	33.5%	109.7
CHIMP_2014nr	57	17,575,167	1,266,014	7.2%	92.8%	
CHIMP-RefSeqGCS	57	17,575,167	7,526,697	42.8%	57.2%	5.9
ELSIDRON1-2014nr	57	50,238,935	1,333,986	2.7%	97.3%	
ELSIDRON1-RefSeqGCS	57	50,238,935	15,771,444	31.4%	68.6%	11.8
ELSIDRON2-2014nr	60	48,231,792	3,171,401	6.6%	93.4%	
ELSIDRON2-RefSeqGCS	60	48,231,792	19,216,338	39.8%	60.2%	6.1
Modern-2014nr	67	29,469,839	6,991,350	23.7%	76.3%	
Modern-RefSeqGCS	67	29,469,839	23,754,716	80.6%	19.4%	3.4
SPYNEW-2014nr	61	4,041,681	324,843	8.0%	92.0%	
SPYNEW-RefSeqGCS	61	4,041,681	1,649,560	40.8%	59.2%	5.1
War-B61-Med.Germany-2014nr	82	13,260,566	4,544,233	34.3%	65.7%	
War-B61-Med.Germany-RefSeqGCS	82	13,260,566	8,196,949	61.8%	38.2%	1.8
War-G12-Med.Germany-2014nr	88	8,999,409	3,754,002	41.7%	58.3%	
War-G12-Med.Germany-RefSeqGCS	88	8,999,409	5,693,022	63.3%	36.7%	1.5
AVERAGE 2014nr	51	7,982,560	895,593	5.8%	94.2%	
AVERAGE RefSeqGCS	51	7,982,560	3,775,439	41.8%	58.2%	**64.2**

Chapter IV

—

Development and validation of a complementary approach to reconstruct ancient microbial communities

Statement of Authorship

Title of Paper	Development and validation of a complementary approach to reconstruct ancient microbial communities
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Unpublished and unsubmitted work written in manuscript style.

Principal Author

Name of Principal Author (Candidate)	Raphael Eisenhofer
Contribution to the Paper	Co-designed the RNA probe set. Designed the experiments. Performed all the laboratory work and data analysis. Wrote the manuscript.
Overall percentage (%)	85%
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 10/5/18

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	Laura S. Weyrich
Contribution to the Paper	fundings, helped design study, editing
Signature	Date 9/5/18

Development and validation of a complementary approach to reconstruct ancient microbial communities

Authors: Raphael Eisenhofer[1] & Laura S. Weyrich[1]
Affiliations: [1]Australian Centre for Ancient DNA, University of Adelaide, Australia

Abstract

The taxonomic characterisation of ancient microbial communities (microbiota) is a key step in the rapidly growing field of palaeomicrobiology. To date, PCR amplification of the 16S rRNA gene is the most commonly used technique for taxonomic classification in modern microbiota studies, but recent research has found that this method has severe biases when applied to ancient microbial DNA. Untargeted amplification methods such as shotgun metagenomic sequencing offer a less biased approach for reconstructing ancient microbial communities, allowing for the use of 16S rRNA gene fragments within an ancient sample for taxonomic classification. However, 16S rRNA gene fragments only make up a small proportion of DNA sequences in a metagenome (<0.05%), making it costly to sample the extent of microbial diversity present. Additionally, it is not known how the characteristics of ancient DNA influence the quality of taxonomic assignment of 16S rRNA gene fragments. Here, we develop, test, and apply a hybridisation enrichment technique to selectively target 16S rRNA gene fragments from untargeted shotgun libraries. Using this method, we increase the sequencing of 16S rRNA gene fragments from ancient metagenomic samples by 334-fold over unenriched libraries, allowing us to empirically examine the extent of 16S rRNA gene diversity present in ancient samples. Using simulated data, we also investigate the influence of ancient DNA characteristics on the taxonomic assignment of 16S rRNA gene fragments. This study validates the use of 16S rRNA gene fragments for the reconstruction of ancient microbial communities and offers a method that can complement existing approaches for studying ancient microbiota.

Introduction

Research into the microbial communities (microbiota) inhabiting the human body has intensified in the past decade due to these communities' associations with human health and disease [1–3]. Given the long-term association and co-evolution of these microbial communities with humans throughout evolution [4], understanding past human microbiota and how they have changed through time may offer important medical insights [5]. The discovery that dental calculus is an excellent source of ancient human-associated microbial DNA [6,7], coupled with recent advances in DNA sequencing and laboratory techniques, have made it possible study human microbiota through time using ancient DNA [6–8]. These studies also have the potential to improve our understanding of microbial evolution [7], the lifeways of our ancestors [8], and even past demographic events [9]. Critical to our understanding of these past microbial communities are the techniques available for us to classify them (to find out who's there).

In modern microbiota studies, amplification of hypervariable 16S rRNA gene regions is most frequently used for taxonomic classification of microbial communities [10]. This marker is present in all prokaryotes, possessing both conserved regions that can be used to design broad-specificity primers and hypervariable regions that contain phylogenetic information useful in determining relationships between bacteria and archaea [11]. However, recent research has found that targeted amplification of the 16S rRNA gene is not a valid approach for the taxonomic classification of ancient DNA [8,12]. This is because the regions targeted in 16S amplification are longer than the typically short fragment lengths of ancient DNA, which can lead to longer modern contaminant DNA being preferentially amplified [8,12]. Furthermore, variability in length of the 16S region targeted between taxa can lead to biased amplification, confounding microbial composition estimates [8,12]. The alignment-based approach of shotgun metagenomic data analysis is less biased for taxonomic classification of ancient microbiota compared to the amplicon-based 16S approach [8,12], and is now the standard in palaeomicrobiological research. However, this approach is more expensive than the amplicon-based 16S approach and requires more computational resources in order to align sequences against reference databases containing genomic information. For example, a dedicated computer server with 1,500 GB of RAM was required for aligning shotgun metagenomic data against a database containing 47,696 prokaryotic reference genomes [13]. It has also been demonstrated that the shotgun metagenomic alignment method is strongly dependent on the diversity and availability of references genomes in databases, which can limit the taxonomic reconstruction of ancient microbial communities [13]. A recent study demonstrated that on average ~60% of DNA sequences in ancient dental calculus samples could

not be assigned taxonomy when using the shotgun metagenomic alignment approach, suggesting incomplete sampling of prokaryotic genome diversity in current reference databases [13]. This is a critical issue, as we are potentially missing a large proportion of microbial diversity present in ancient samples.

A possible solution to this issue is to analyse 16S rRNA gene fragments present in metagenomic shotgun data [8]. This both avoids the biases inherent in the specific amplification of the 16S rRNA gene and could allow for use of the phylogenetic information within the 16S rRNA gene to classify sequences even if they do not have direct matches in a reference database. This approach could also leverage the greater diversity present in 16S rRNA gene databases compared to genome databases. For example, the latest SILVA database SSU (small subunit) Ref NR (non-redundant) 132 contains 695,171 16S rRNA gene sequences (even after clustering at 99% sequence identity), compared to the 47,696 reference genomes tested in [13]. Another advantage to this approach is that it does not require dedicated, RAM-heavy computer servers capable of handling the larger genomic reference databases. While this approach has been used in previous palaeomicrobiology studies [8,12,14], there has not yet been a robust assessment of this technique for reconstructing ancient microbial communities. Given the untargeted nature of metagenomic shotgun sequencing, 16S rRNA gene fragments make up only a small fraction of the total data generated (<0.05% [15]). Therefore, sequencing a sample at a depth of 1,000,000 reads may provide fewer than 1,000 16S rRNA gene fragments, which may not represent the total microbial diversity present a sample. Furthermore, the influence of ancient DNA characteristics such as short fragment lengths and deamination (a common damage-induced substitution) on the ability to assign taxonomy to 16S rRNA gene fragments has not been formally investigated.

Here, we develop a new experimental approach to selectively enrich 16S rRNA gene fragments from ancient shotgun metagenomic libraries to obtain high sampling depth suitable for empirical determination of the extent of 16S rRNA gene diversity present in an ancient metagenomic sample. We create and benchmark a bioinformatic pipeline for the analysis of short 16S rRNA gene fragments and investigate the influence of ancient DNA characteristics on taxonomic assignments in 16S rRNA data sets. Finally, we compare our new approach to the traditional shotgun sequencing and whole-genome alignment method to better verify current methods in palaeomicrobiological analysis.

Methods

16S rRNA gene RNA probe design

Full-length 16S rRNA genes were obtained database from all species identified from a recent ancient dental calculus study [8] from the Ribosome Database Project (RDP) [16] n=285. To further increase diversity, 16S rRNA genes were also taken from species in the Human Oral Microbiome Database (HOMD) [17] if they did not match species identified by Weyrich *et al.* This added an additional 285 sequences, yielding a total of 570 full-length 16S rRNA genes to be used for probe design (See supplementary file X for a list of taxa). RepeatMasker [18] was used to remove any simple or low complexity repeats from the sequences. 80 bp (base pair) RNA baits with 20 bp probe spacing and 4x tiling density were synthesized by Arbor Biosciences (formerly MyBaits). Baits were collapsed if they had fewer than 11 mismatches between each other, yielding a total of 19,634 80 bp RNA baits (Supplementary file X). While the probe design was based on 16S rRNA genes from ancient and modern oral taxa, these probes should also capture microbial diversity that was not used as input for the probe sequences.

Hybridization enrichment

Shotgun libraries generated in a previous study were used for hybridization enrichment [8]. Briefly, each library was created by amplifying existing library DNA in four 25 µL PCR reactions (each containing: 13.625 µL dH_2O, 2.5 µL 10X AmpliTaq Gold Buffer, 2.5 µL of 25 mM $MgCl_2$, 0.625 µL of 10mM dNTPs, 2.5 µL of 10 µM forward and reverse primer, 0.25 µL AmpliTaq Gold, 3 µL template DNA) with the following PCR conditions: denaturing at 94°C for 12 minutes before 13 cycles of (30 seconds denaturing at 94°C, 30 seconds annealing at 60°C, 45 seconds extension at 72°C) followed by a final extension of 10 minutes at 72°C. PCR amplifications were pooled and then cleaned using AMPure XP beads to reach 100 ng of DNA for input into hybridization enrichment. A modified version 3 of the MyBaits protocol was used for hybridization enrichment, whereby the input RNA bait concentration was reduced to 25% of the recommended amount, and custom oligonucleotides were used to block the P5/P7 library adapters. For the enrichment efficiency test, samples were enriched at either 55°C or 65°C for 40 hours; otherwise, samples were enriched at 65°C for 40 hours. Post-capture, the streptavidin beads were washed three times using Wash Buffer 2 (MyBaits v3 manual) and resuspended in PCR mastermix (as above) before amplification with 13 cycles of PCR. Amplified libraries were cleaned with AMPure, quantified with an Agilent TapeStation, pooled at equimolar concentrations, and sequenced on the Illumina HiSeq X Ten platform (2 x 150 bp).

DNA Sequencing and 16S rRNA gene enrichment analysis

The resulting sequencing data was converted into fastq format using Illumina's bcl2fastq software, before being trimmed and demultiplexed using AdapterRemoval2 based on unique P5/P7 barcode combinations [19]. To reduce the computational time of downstream analyses, seqtk (https://github.com/lh3/seqtk) was used to randomly subsample libraries to 100,000 reads each. To filter 16S rRNA gene fragments from samples, SortMeRNA [20] was used with the default SILVA bacterial and archaeal 16S rRNA databases (SILVA 119, 95% clustered). To assign taxonomic identifications to the putative 16S rRNA gene fragments, we constructed a BLAST database [21] using the SILVA 128 NR99 16S rRNA reference database [22] and aligned the putative 16S rRNA gene fragments to the database using default parameters, with the following exceptions: –evalue = 0.01 for added stringency and –outfmt = 0 to allow for import into MEGAN. To test the impact of missing reference sequences on taxonomic classification, we also created a filtered SILVA database by removing reference sequences belonging to members of our simulated dataset using the Filterbyname.sh script from BBtools (https://jgi.doe.gov/data-and-tools/bbtools/). Briefly, grep was used to create a list of taxa to remove from the reference sequences in the SILVA database using Filterbyname.sh (Table S1). Lastly, BLAST outputs were then imported into MEGAN CE 6.10.10 [23] using the "Import from BLAST…" option with the synonyms mapping file (SSURef_NR99_128_tax_silva_to_NCBI_synonyms.map.gz) obtained from the MEGAN community website: (http://ab.inf.uni-tuebingen.de/data/software/megan6/download/).

Generating simulated dataset with ancient DNA characteristics

To test the 16S rRNA gene fragment analysis pipeline, we created a simulated dataset using Gargammel [24]. 16S rRNA genes from 19 phylogenetically diverse prokaryotic species were obtained from the SILVA 128 NR99 16S rRNA database [22] and randomly fragmented to create 100,000 16S rRNA gene fragments fitting a log-normal ancient DNA fragment-length distribution (Figure S1) (gargammel -n 100000 --loc 4 --scale 0.3). Varying levels of cytosine deamination (ancient DNA damage) were simulated using deamSim from gargammel on our simulated dataset to create three different levels of simulated single-stranded overhang deamination: 10% (low) deamination (-damage 0.03,0.25,0.01,0.1); empirical (moderate) deamination from a published mapDamage profile [25] (-mapdamage examplesMapDamage/results_LaBrana/misincorporation.txt); and 50% (high) deamination (-damage 0.03,0.25,0.01,0.5). Finally, the resulting simulated 16S rRNA gene datasets were

aligned against the SILVA 128 NR99 database and the filtered SILVA database using BLAST as described above.

Taxonomic classification of unenriched shotgun metagenomic data

Unenriched shotgun metagenomic data were aligned against a reference database containing 47,696 bacterial and archaeal genomes [13] using MALTn [26] with default parameters and outputting BLAST text files. The resulting BLAST text files were converted into RMA6 using the blast2rma script in MEGAN.

Comparison of 16S rRNA gene and shotgun metagenomics datasets

The LCA parameters used for the shotgun metagenomic data were: bitscore=50, E-value=0.01, minsupp=0.1 (*i.e.* a taxonomic assignment requires at least 0.1% of reads to pass), and the weighted LCA algorithm (80%) as suggested in [23]. LCA parameters used for the 16S enrichment method were the same, except that the naïve LCA algorithm (80%) was used since the weighted algorithm resulted in false-positive taxonomic assignments of the simulated 16S rRNA gene fragments. Genera identified in the extraction blank controls were removed from biological samples in MEGAN. Classifications at each taxonomic level were selected in MEGAN and exported as a text file, before being collated in an excel spreadsheet. Neighbour joining trees were constructed from distance matrices generated in MEGAN.

Statistical analyses

To compare beta-diversity in the two datasets, Euclidean distances exported from MEGAN were ordinated by Principal Components Analysis using STAMP [27]. For QIIME analyses [28], genus level assignments in MEGAN were exported as a BIOM table into QIIME v1.9.1 and rarefied to the sample with the lowest number of genus-level assignments (576 for 16S enrichment-temperature tests, 10,715 for the 16S vs. UEWGA comparisons). Alpha diversity measures and rarefaction curves were calculated using the core_diversity.py script. Distances matrices were created using the beta_diversity.py script and ANOSIM/PERMANOVA tests were computed on these distance matrices using the compare_categories.py script with 999 permutations.

Results

Assessing taxonomic assignment of simulated ancient 16S rRNA gene fragments

To test if we could accurately classify short 16S rRNA gene fragments, we created a simulated dataset containing 19 phylogenetically diverse prokaryotic species (Table S1). The 16S rRNA gene from each species was selected and randomly fragmented to fit a commonly observed ancient DNA fragment length distribution (*e.g.* a mean length of 50bp; Figure S1). Taxonomy was classified by mapping DNA fragments to the SILVA 16S rRNA database (NR 99, release 128) using BLAST nucleotide alignment, following by applying the lowest common ancestor (LCA) algorithm in MEGAN. We found that 98.4% of the total DNA fragments could be assigned taxonomy, and of these, 53.2% were assigned to the genus level, and only 3.5% to the species level (Table 1). The resulting genus-level taxonomic composition matched that of the input sequences with one exception: *Yersinia* 16S rRNA gene fragments could not be assigned to the genus *Yersinia* and were pushed up to the order Enterobacterales (Figure 1B). The 1.6% of fragments that could not be aligned were extremely short (between 15-30 bp) (Figure S2), suggesting that accurate classifications are limited to at least 30bp for 16S rRNA fragments, as previously observed for whole-genome alignments [13].

Table 1. Alignment statistics for simulated data

Percentage of total reads assigned:	Total	Domain	Phylum	Class	Order	Family	Genus	Species
BLASTN-MEGAN	98.4%	4.5%	3.7%	11.1%	4.8%	17.0%	53.2%	3.5%
BLASTN-MEGAN-SpeciesExclusion	98.3%	4.6%	3.9%	11.1%	4.9%	17.1%	55.4%	0.7%
BLASTN-MEGAN-GenusExclusion	97.7%	**9.6%**	**6.7%**	**12.0%**	**7.4%**	**18.8%**	**41.2%**	0.2%
BLASTN-MEGAN-10%Deamination	97.8%	4.9%	3.8%	11.2%	4.8%	16.9%	52.7%	3.4%
BLASTN-MEGAN-LaBrana-Deamination	97.8%	5.1%	3.8%	11.1%	4.8%	16.9%	52.6%	3.4%
BLASTN-MEGAN-50%Deamination	96.9%	5.5%	3.7%	11.2%	4.9%	16.8%	52.1%	3.4%

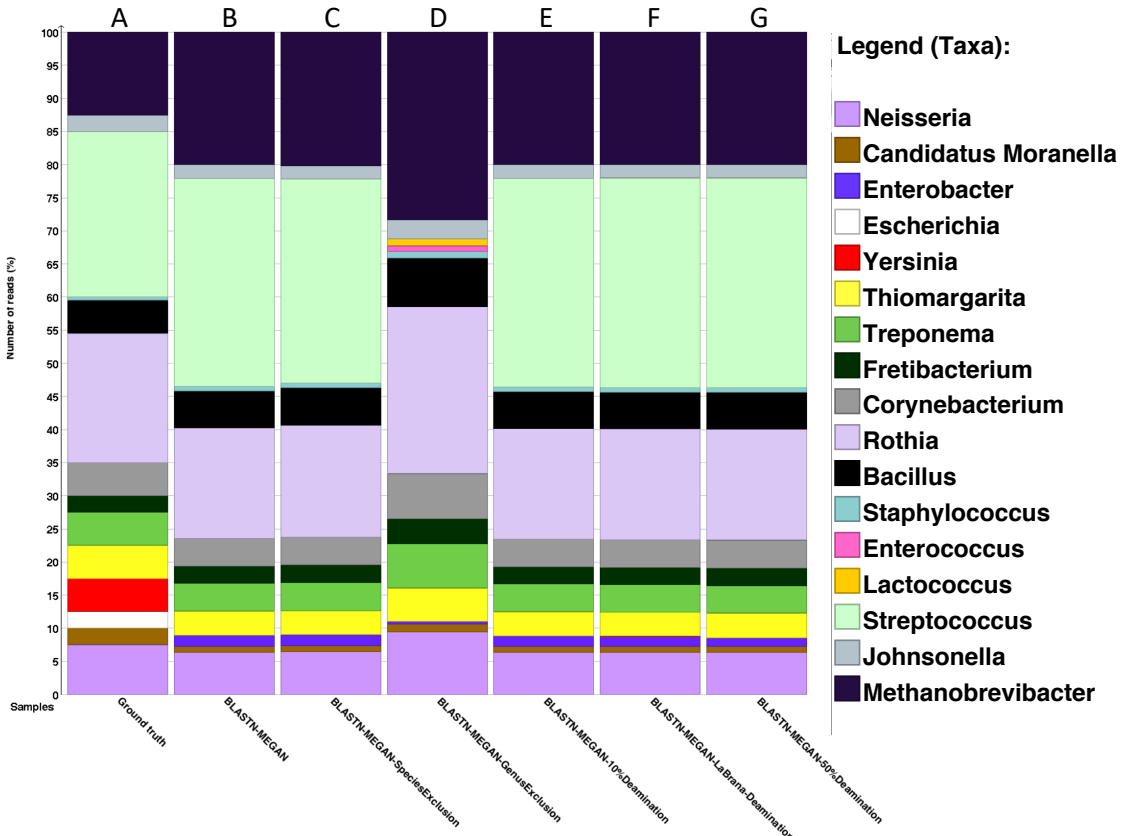


Figure 1. Recapitulation of genus-level taxonomic composition from simulated metagenome. (A) Actual abundance and structure of simulated Community. (B) Reconstruction of community structure using MALT/MEGAN. (C),(D) Results of genus and species exclusion test. (E)-(G) Results different rates of simulated cytosine deamination on metagenomic reconstruction.

Assessing the impact of missing reference sequences on taxonomic classification

To test the impact of missing reference sequences on our ability to taxonomically classify 16S rRNA gene fragments, we performed a species and genus exclusion experiment whereby we removed reference sequences corresponding to species or genera present within the simulated dataset from the SILVA database (Table S2), and then repeated BLAST alignments using this modified database. Sequences originating from excluded species could be classified to their respective genera (Figure 1C, Table 1). Similarly, removal of all reference sequences attributed to the *Streptococcus* genus (which accounts for 25% of the simulated dataset) resulted in a 12% reduction in assignments to the genus-level (53.2% vs. 41.2%), and an associated increase in sequences at higher taxonomic ranks (Figure 1D, Table 1). Importantly the total number of aligned reads did not substantially decrease (97.7% vs. 98.4% for non-exclusion), suggesting that sequence conservation and phylogenetic signal within the 16S gene allowed for placement of these sequences higher in the taxonomy, rather than discarding them outright.

Influence of DNA damage on short 16S sequence classification

To test if cytosine deamination (a common form of ancient DNA damage) influences taxonomic classification of short 16S rRNA gene fragments, we simulated three levels of sequence

deamination on our simulated dataset representing low (10%), moderate (simulated from a real mapDamage profile from an ancient specimen; ~30% (LaBrana [29]), and high (50%). We found that deamination did not have a notable impact on taxonomic classification, as there was no loss of taxonomic assignments or misclassifications (Figure 1E-1G). Only a 1.5% loss of total reads assigned taxonomy was observed when assessing the highest level of deamination (50%) compared to the non-deaminated simulated dataset (Table 1). Overall, these findings suggest that deamination does not meaningfully hinder the ability to classify short, ancient 16S DNA fragments and that the method developed here could be used to examine taxonomic diversity even in highly degraded ancient samples.

Optimising hybridization enrichment efficiency

We designed RNA baits to capture 16S rRNA gene fragments from a diverse range of microbial taxa (see methods). To optimize the hybridization efficiency (proportion of on-target sequences vs. non-target sequences) of our RNA bait set, we tested two different hybridization temperatures, 55°C and 65°C. We found that enrichment at 65°C yielded the highest on-target enrichment of 16S rRNA gene fragments (average 53.42% putative 16S rRNA gene fragments). This was more than a two-fold increase over the 55°C enrichment (average 26.17% putative 16S rRNA gene fragments), and a 334-fold increase over the un-enriched shotgun libraries (average 0.16% putative 16S rRNA gene fragments) (Table 2). We found that increasing hybridisation temperature selected for longer mean DNA fragment lengths (unenriched = 56 bp, 55°C = 69 bp, 65°C = 79 bp), but did not find changes in the mean GC content between temperatures (Table 2). This enrichment of longer DNA fragments by higher temperatures could be explained by shorter DNA fragments not having sufficient Watson-Crick hydrogen bonding sites to remain attached to probes at higher temperatures, resulting in the preferential binding and enrichment of longer DNA fragments.

Table 2. Enrichment statistics for hybridization temperature test

Sample	Total reads	Putative 16S reads	% putative 16S reads	%duplicates	%GC	avg_sequence_length
A13204_AfrPP2_un-enriched	1,750,000	2,767	0.16%	3.52%	49	54
A13204_AfrPP2_55°C	1,790,862	509,206	28.43%	80.54%	49	69
A13204_AfrPP2_65°C	1,649,694	854,834	51.82%	85.81%	50	77
A13208_AfrSF2_un-enriched	2,000,000	3,531	0.18%	1.34%	49	58
A13208_AfrSF2_55°C	2,105,596	539,978	25.64%	52.98%	48	74
A13208_AfrSF2_65°C	1,897,804	1,114,041	58.70%	71.82%	50	86
A13209_AfrSF3_un-enriched	2,050,000	2,265	0.11%	0.89%	53	51
A13209_AfrSF3_55°C	2,185,137	513,106	23.48%	34.97%	51	62
A13209_AfrSF3_65°C	1,869,330	948,471	50.74%	52.66%	51	68
A13213_AfrPP1_un-enriched	2,350,000	4,361	0.19%	1.84%	50	60
A13213_AfrPP1_55°C	2,725,464	738,998	27.11%	43.60%	49	73
A13213_AfrPP1_65°C	2,015,235	1,056,566	52.43%	55.66%	50	83
Average un-enriched	2,037,500	3,231	**0.16%**	1.90%	50	56
Average 55°C	2,201,765	575,322	**26.17%**	53.02%	49	69
Average 65°C	1,858,016	993,478	**53.42%**	66.49%	50	79

Given that hybridization temperature was found to select for longer DNA sequences we next tested to see if this influenced the taxonomic composition of samples. Due to constraints in the alignment speed of BLAST, each sample from each treatment (55°C, 65°C, unenriched) was randomly subsampled to 100,000 reads. After alignment and classification using the SILVA database, genus level assignments were exported from MEGAN into QIIME 1.9.1 and rarefied to the depth of the sample with the fewest genus level assignments (576 assignments). There were no significant differences in alpha diversity measures (Shannon; observed-species) between treatment types (Figure S2A & S2B; *p*-values >0.05). Alpha-rarefaction curves indicated that most diversity was sampled, with as few as 550 genus level 16S rRNA gene fragment assignments for all treatments (Figure S2C & S2D). Differences in microbial community composition were also compared using PCA (Principal Components Analysis) of Euclidean distances (Figure 2). We found clustering based on sample rather than hybridization temperature, suggesting that temperature minimally affected the overall taxonomic composition. This was corroborated by statistical analysis using ANOSIM and PERMANOVA on different distance metrics (Table 3; Treatment *p*-values=>0.05). Additionally, taxonomic bar plots representing the sum of samples per treatment were nearly identical in both abundance and diversity (Figure S4).

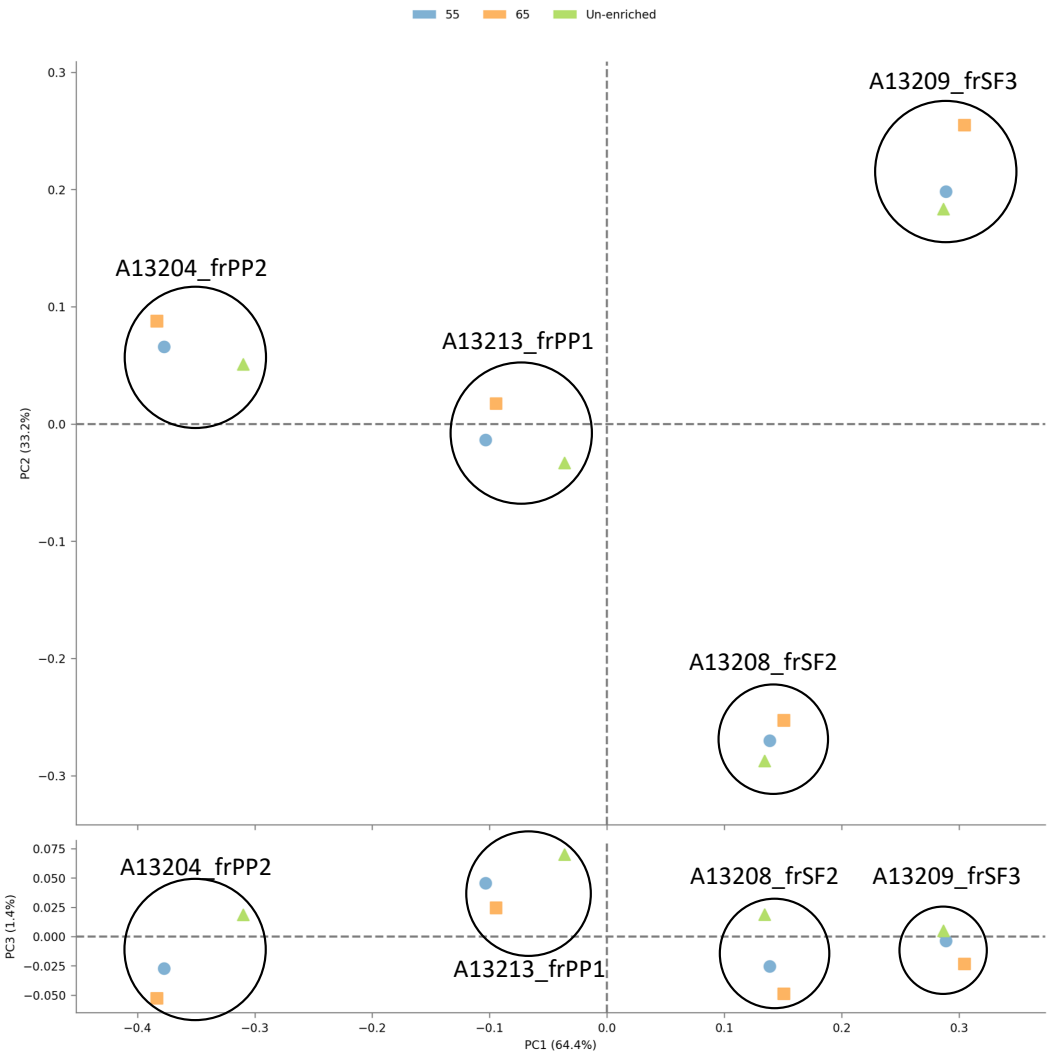


Figure 2. PCA of Euclidean distances of microbial composition between samples. Principal Components 1-2, and 1-3 are shown. Enrichment method is indicated by different coloured shapes. Circle outlines represent groupings of the same sample.

Table 3. Statistical assessment of beta diversity metrics by treatment method or sample

Beta-diversity metric	Variable	Statistical method	*R*	F statistic	*p*-value
Binary-Jaccard	Treatment	ANOSIM	-0.202		0.943
Binary-Jaccard	Treatment	PERMANOVA		0.327	0.962
Binary-Jaccard	Sample	ANOSIM	0.976		**0.001**
Binary-Jaccard	Sample	PERMANOVA		11.658	**0.001**
Bray-Curtis	Treatment	ANOSIM	-0.218		0.955
Bray-Curtis	Treatment	PERMANOVA		0.160	0.962
Bray-Curtis	Sample	ANOSIM	1.000		**0.001**
Bray-Curtis	Sample	PERMANOVA		44.460	**0.001**
Binary-Euclidean	Treatment	ANOSIM	-0.192		0.930
Binary-Euclidean	Treatment	PERMANOVA		0.431	0.957
Binary-Euclidean	Sample	ANOSIM	0.959		**0.001**
Binary-Euclidean	Sample	PERMANOVA		6.567	**0.001**

Regarding taxonomic assignments specific to hybridization temperature, both enrichment temperatures allowed for the detection of *Corynebacterium* (a genus commonly found in human dental plaque) that was not identified in the unenriched samples. Additionally, the 65°C treatment was the only one to detect *Pseudoramibacter* (an oral taxon previously identified in ancient dental calculus and modern studies [13,30]) (Figure S5). Taxa specific to the unenriched treatment include: *Bacteroides, Microbacter, Parabacteriodes, Desulfoplanes, Pseudomonas, Merismopedia, Blautia, Acholeplasma, and Candidatus Phytoplasma* (Figure S5). None of these taxa are commonly found in the oral cavity [17] and likely represent laboratory or reagent contamination. It is possible that they were identified here because the unenriched libraries contain fewer 16S rRNA gene assignments compared to the enriched libraries, the minimum percentage of reads needed for taxonomic assignment (minimum support percent 0.1%) is lower (*e.g.* the minimum support percent for 3,000 assignments is [3,000*0.001 = 3], versus minimum support percent for 30,000 [30,000*0.001 = 30]). Therefore, contaminant DNA present in the library at low levels may have been more likely to be assigned taxonomy in the unenriched libraries.

Overall, these results suggest that a hybridization temperature of 65°C substantially increases the on-target enrichment of 16S rRNA gene fragments with minimal observable taxonomic dropout or bias. Therefore, a hybridization temperature of 65°C was used for subsequent enrichments.

Comparison of 16S enrichment method to shotgun method for taxonomic classification

We next sought to compare taxonomic classifications of both 16S rRNA gene enrichment and unenriched shotgun metagenomic methods using previously published ancient dental calculus data. We enriched 11 additional dental calculus samples from [8] at 65°C, bringing the total number of samples to 15 (including four samples from the previous section) (Table S2). To normalize the number of sequences present in both methods, we randomly subsampled to equal sequencing depths prior to any analysis (Table S2 & S3). For the 16S enrichment samples, we observed an average of 58% on-target enrichment of 16S rRNA gene fragments, and an increase in mean read length from 48 for unenriched to 73 for enriched (Table S2), supporting our previous results.

Because DNA contamination from reagents and the laboratory environment can influence taxonomic analyses (especially for ancient samples) we removed genera identified in the extraction blank controls (EBCs) (Table S4) from the biological samples. For all 28 biological samples (14 from each treatment), 0.2% of reads were removed by filtering in the shotgun data set, while 3.12% were removed from enriched 16S rRNA gene sequences. This

suggests that 16S rRNA gene enriched data is more prone to the impacts of modern contaminant DNA.

We next assessed the specificity of the taxonomic assignments for both methods (Table 4). As expected, the shotgun method was more specific (average of 64.8% of assigned reads at the species level versus 1.8% for the 16S enrichment method). However, on average, the shotgun method could not align 59% of reads, whereas only 0.5% could not be aligned for the 16S enrichment method (Table 4). Given that both methods represent different approaches and use different reference databases, we next tested if there were genus identifications specific to each method. The 16S enrichment method classified 20 genera that were not present in the shotgun method, although these assignments each had a mean abundance <1% (Table 5). The shotgun method had 18 genus identifications not present in the 16S enrichment method, and all but three of these each had a mean abundance of <1% (Table 5). Comparing these assignments to taxa found in the Human Oral Microbiome Database (HOMD), 9/20 were putatively oral for the 16S enrichment method, with 14/18 for the shotgun method. Importantly, the 16S enrichment method was unable to detect the common oral genus *Tannerella*, which was present at a mean abundance of 3.42% in the shotgun method. Additionally, the putative oral genus *Olsenella* was also not detected in the 16S enrichment method and was at a mean abundance of 7.85% in the shotgun method.

Table 4. Mean proportions of taxonomic assignments at given taxonomic ranks for 16S-enrichment and unenriched shotgun approaches

Method	Total # reads	Total reads used for alignment	# reads aligned	# reads not aligned	% unaligned	% assigned domain	% assigned phylum	% assigned class	% assigned order	% assigned family	% assigned genus	% assigned species
Average 16S-enrichment	94,019	53,905	52,823	430	0.8%	6.4%	4.6%	9.3%	18.3%	21.0%	35.9%	1.8%
Average Shotgun	94,019	94,019	37,856	56,102	59.0%	1.7%	2.1%	2.7%	3.2%	3.9%	20.8%	64.8%

Table 5. Mean abundance of method-specific taxonomic assignments. Bolded Names represent taxa present in the Human Oral Microbiome Database

Taxon specific to 16S	Mean abundance	Taxon specific to shotgun	Mean abundance
Peptococcus	0.90%	**Olsenella**	**7.85%**
Desulfobulbus	0.85%	**Tannerella**	**3.42%**
Mogibacterium	0.77%	**Parvimonas**	1.19%
Peptostreptococcus	0.41%	Dialister	0.48%
Fastidiosipila	0.39%	**Eikenella**	0.45%
Bergeyella	0.23%	**Atopobium**	0.38%
Brachymonas	0.22%	**Slackia**	0.35%
Fusibacter	0.19%	**Kingella**	0.27%
Pelospora	0.16%	Clostridium	0.24%
Flavobacterium	0.15%	**Peptoniphilus**	0.17%
Roseburia	0.14%	**Anaeroglobus**	0.12%
Nocardioides	0.13%	**Granulicatella**	0.08%
Paenibacillus	0.08%	**Solobacterium**	0.06%
Desulfovibrio	0.06%	**Oribacterium**	0.05%
Polaromonas	0.06%	**Eggerthia**	0.05%
Candidatus Tammella	0.05%	**Stomatobaculum**	0.04%
Peptoclostridium	0.04%	Chlorobium	0.04%
Acetobacterium	0.04%	Peptoanaerobacter	0.02%
Pyramidobacter	0.02%		
Verrucomicrobium	0.02%		
Sum of mean abundances	4.92%		15.28%

We next tested to see if these differences in genus level assignments between methods influenced the relationships between samples. We found that samples classified by sample were more similar to one another than samples classified by method (ANOSIM of Binary Euclidean distances R=0.179 for method, and R=0.516 by sample; p-value for both <0.05) (see Table 6 for other distance metrics and statistical tests). Visualising both non-abundance-weighted and abundance-weighted compositional distances between samples on neighbour joining trees, we found similar groupings of samples despite the method used (Figure 3; Figure S6). This suggests that despite differences in the genera identified in each method, the relationships between ancient samples based on microbial composition is similar.

Table 6. Statistical assessment of beta-diversity metrics by method or sample

Beta-diversity metric	Variable	Statistical test	*R*	F statistic	*p*-value
Binary-Jaccard	Method	ANOSIM	0.442		**0.001**
Binary-Jaccard	Method	PERMANOVA		7.519	**0.001**
Binary-Jaccard	Sample	ANOSIM	0.236		**0.025**
Binary-Jaccard	Sample	PERMANOVA		1.428	**0.028**
Bray-Curtis	Method	ANOSIM	0.315		**0.001**
Bray-Curtis	Method	PERMANOVA		5.596	**0.001**
Bray-Curtis	Sample	ANOSIM	0.589		**0.001**
Bray-Curtis	Sample	PERMANOVA		2.800	**0.001**
Binary-Euclidean	Method	ANOSIM	0.179		**0.005**
Binary-Euclidean	Method	PERMANOVA		3.711	**0.015**
Binary-Euclidean	Sample	ANOSIM	0.516		**0.001**
Binary-Euclidean	Sample	PERMANOVA		3.274	**0.001**

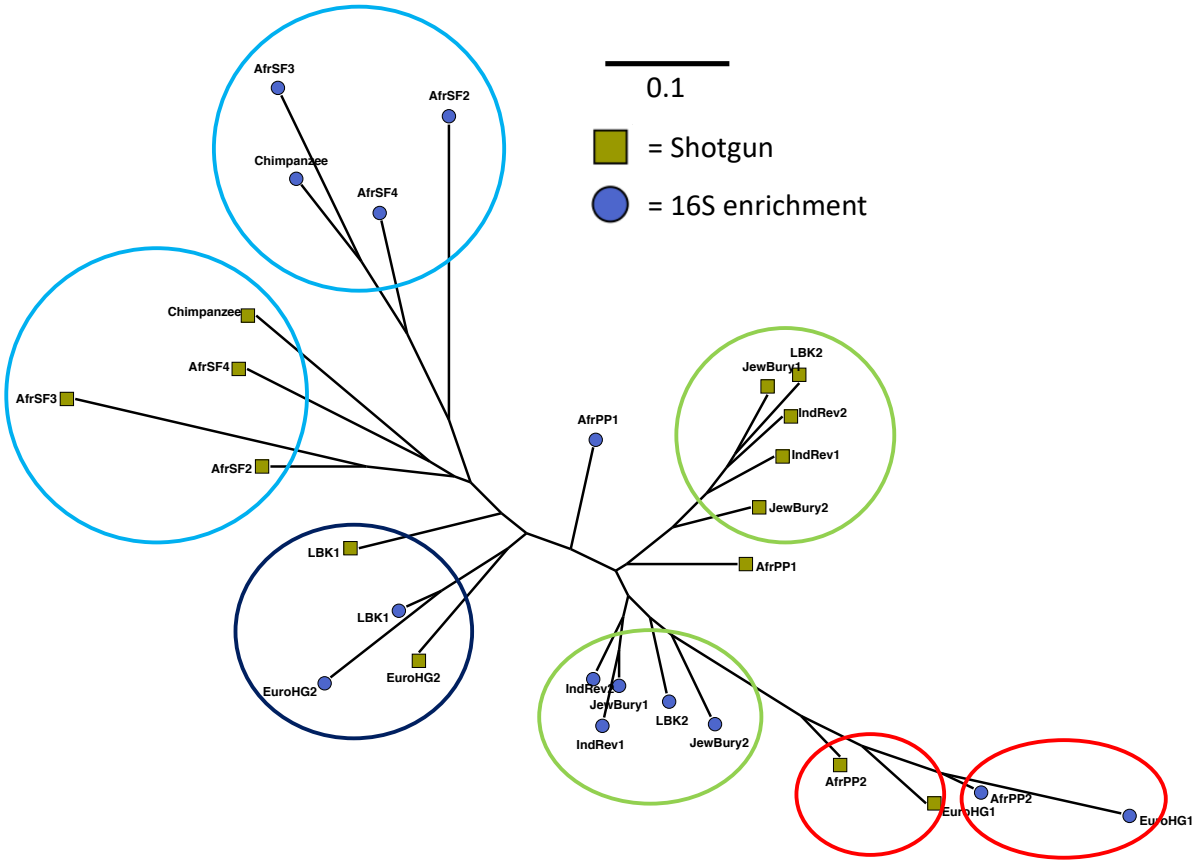


Figure 3. Neighbour Joining tree of genus-level Euclidean distances between samples. Coloured outlines represent similar groupings of samples.

Finally, to compare the cost associated with reconstructing microbial composition, we compared the effectiveness of obtaining genus or species level assignments of the aligned sequences in both methods given equal sequencing depth. Given equal sequencing depth for each method, the 16S enrichment method assigned an average of 21.1% of sequences a taxonomic classification to the genus and species level, while the shotgun method assigned 34.5% of the sequences to a similar level. This represents a 39% increase in the number of reads assigned at the genus or species level for shotgun over the 16S enrichment method, despite the increase in the percentage of reads assigned for the 16S enrichment method (56% for 16S enrichment, 41% for shotgun). This suggests that while the 16S enrichment method can align and assign a higher percentage of sequences, these assignments are not as specific as the shotgun method, and given the extra costs associated with the enrichment method, the shotgun method is currently the most cost-effective means of classifying genus or species level taxonomy in highly degraded ancient samples.

Discussion

Reconstructing ancient microbial communities is a challenging endeavour. Currently, there is a lack of studies comparing different methods for doing so. By using both in-silico data and developing a new technique to obtain deep sampling of 16S rRNA fragments from metagenomes, this study is the first to assess the influence of ancient DNA characteristics and sampling depth on the reconstruction of ancient microbial communities using 16S rRNA gene fragments. While we found that this technique was not as efficient at obtaining genus or species level resolution as the shotgun metagenomics approach, the overall community composition between the methods was similar. Therefore, analysis of 16S rRNA gene fragments could be used to complement community reconstruction using the shotgun metagenomics approach.

Using simulated data, this study was the first to formally test the influence of ancient DNA characteristics on the alignment and classification of 16S rRNA gene fragments, a technique previously used in the field [8,12,14]. We found that short 16S rRNA gene fragments did not contain sufficient information for species-level resolution, limiting their use for species-level classification of ancient DNA. This supports previous findings in modern microbiome research whereby longer lengths of the 16S rRNA can provide more specific taxonomic information [31,32]. We could classify 98.4% of all simulated reads, with 53.2% of these being placed at the genus level. The resulting taxonomic classifications had no false-positive taxonomic assignments, and almost perfectly recapitulated the true community. One exception to this was an issue classifying *E. coli* and *Y. pestis* 16S rRNA gene fragments, which has been noticed before [33–35] and may be due to insufficient resolution of the 16S rRNA gene to discriminate between these two species. Regardless, 16S rRNA gene fragments from these species were placed at the family level, so not all taxonomic information was lost. By performing a reference sequence exclusion experiment, we found that our analytical method was robust to missing reference sequences in databases, which is currently a major issue for shotgun metagenomic methods [13]. We also demonstrated that deamination does not meaningfully impact classification of short 16S rRNA gene fragments and that fragments as short as 30 bp could be assigned, supporting a previous assessment of alignment-based methods for ancient DNA [13]. Overall, these simulations suggest that the alignment and classification of 16S rRNA fragments is robust to the characteristics of ancient DNA, allowing for its use in the reconstruction of ancient microbial communities in highly degraded samples.

To empirically test our simulated findings, we developed a highly efficient hybridisation enrichment method to obtain deep sampling of 16S rRNA gene fragments from shotgun metagenomic libraries. For the 65°C hybridisation enrichment temperature tested, an average

of 58% of sequenced DNA reads were 16S rRNA gene fragments, which represents a 334-fold increase over the same unenriched libraries. We also found that enriched libraries had longer read lengths than unenriched, which could be explained by shorter DNA fragments not having sufficient Watson-Crick hydrogen bonding sites to remain attached to probes at higher temperatures. Despite this different in read length, we found no major influence of enrichment and enrichment temperature on observed taxonomic composition, suggesting little bias when using enriched 16S rRNA gene fragments compared to unenriched libraries.

Because 16S rRNA gene fragments make up a small fraction of reads obtained by untargeted shotgun metagenomic sequencing (~0.1%), the ability to detect rare taxa using this method is limited. By developing an unbiased method to provide deep sampling of 16S rRNA gene fragments, our hybridisation enrichment technique also allowed us to empirically test the extent of 16S rRNA gene diversity in metagenomic samples. While this method allowed for the detection of oral taxa not present in unenriched libraries, the overall taxonomic composition was similar despite the large difference in the number of putative 16S rRNA gene fragments (average of 3,231 for unenriched, 993,478 for enriched). This suggests that most 16S rRNA microbial diversity is captured from unenriched shotgun metagenomic sequencing with as few as 3,000 16S rRNA gene fragments, corresponding to ~2,000,000 unenriched shotgun metagenomic sequences. This empirical verification adds further support to the microbial diversity captured in previous palaeomicrobiology studies that used unenriched 16S rRNA gene fragments to classify ancient microbial communities [8,12,14].

To our knowledge, this is also the first study to empirically test the specificity of taxonomic classifications derived from 16S rRNA gene fragment alignments and whole-genome alignments using ancient DNA. We found that the shotgun whole-genome alignment method has greater specificity than 16S rRNA gene fragment alignments, supporting in-silico findings here and in a previous study [13], as well as in modern microbiome research [34,36]. While there were putatively oral taxa identified specific to either method, the shotgun approach possessed a higher proportion of oral to contaminant taxa and was uniquely able to detect two relatively highly prevalent and abundant oral genera commonly found in dental calculus, *Tannerella* and *Olsenella* [13]. Our findings, both in-silico and empirical, suggest that the shotgun approach is the best suited for future palaeomicrobiological using dental calculus seeking to obtain specific microbial classifications. However, the analysis of 16S rRNA gene fragments could be advantageous for studying sample types that are less well-characterised (*e.g.* sediment), as higher-level classifications could be obtained using the phylogenetic information obtained in the 16S rRNA gene.

While we found differences in the specificity of assignments and taxa identified between methods, the overall taxonomic compositions classified were similar. It has been previously demonstrated that taxonomic composition derived from metagenomic 16S rRNA gene fragments is similar to taxonomic composition derived from shotgun libraries [9]. Our study validated and expanded on this work by using more samples (14 versus 6) and performing in-silico validation of the method used for 16S rRNA gene fragment analysis. We also used nucleotide-to-nucleotide alignments for shotgun alignments, which have been demonstrated to be more accurate and robust to the characteristics of ancient DNA than nucleotide-to-protein alignments [13]. Overall, our findings suggest that while there are differences between ancient microbial community reconstruction using either 16S rRNA gene fragments or shotgun whole-genome alignments, the overall genus-level taxonomic compositions and the resulting relationships between samples are similar. This suggests that the analysis of 16S rRNA gene fragments from shotgun metagenomic data can be used as a complementary approach to support relationships between samples found using shotgun whole-genome alignment methods.

In summary, using both in-silico and empirical data, this study validates the use of 16S rRNA gene fragments for the reconstruction of ancient microbial communities. While less specific that the shotgun whole-genome alignment approach, this method could be used to add further support to community reconstructions and the relationships between samples. Additionally, this method would be suitable for surveying microbial diversity in ancient sample types that do not have robust modern sampling of reference genomes. Our findings provide a valuable resource for future studies seeking to reconstruct ancient microbial communities.

References

1. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project [Internet]. Nature. 2007 [cited 2018 Apr 23]. Available from: https://www.nature.com/articles/nature06244

2. Ley RE. Obesity and the human microbiome. Curr Opin Gastroenterol. 2010;26:5–11.

3. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.

4. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, et al. Cospeciation of gut microbiota with hominids. Science. 2016;353:380–2.

5. Weyrich LS. Evolution of the Human Microbiome and Impacts on Human Health, Infectious Disease, and Hominid Evolution. Reticul Evol [Internet]. Springer, Cham; 2015 [cited 2018

Mar 26]. p. 231–53. Available from: https://link.springer.com/chapter/10.1007/978-3-319-16345-1_9
6. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45:450–5.
7. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336–44.
8. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature. 2017;544:357–61.
9. Eisenhofer R, Anderson A, Dobney K, Cooper A, Weyrich LS. Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events. J Isl Coast Archaeol. 2017;0:1–14.
10. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, et al. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. 2012;6:1621–1624.
11. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. Proc Natl Acad Sci. 1977;74:5088–90.
12. Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci Rep. 2015;5:16498.
13. Eisenhofer R, Weyrich LS. Assessing alignment-based taxonomic classification of ancient microbial DNA. (Chapter III) in preparation. 2018;
14. Velsko IM, Overmyer KA, Speller C, Klaus L, Collins MJ, Loe L, et al. The dental calculus metabolome in modern and historic samples. Metabolomics [Internet]. 2017 [cited 2018 Apr 26];13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5626792/
15. Guo J, Cole JR, Zhang Q, Brown CT, Tiedje JM. Microbial Community Analysis with Ribosomal Gene Fragments from Shotgun Metagenomes. Appl Environ Microbiol. 2016;82:157–66.
16. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Res. 2014;42:D633–42.
17. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and

genomic information. Database [Internet]. 2010 [cited 2018 Feb 9];2010. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baq013/405450
18. Smit A, Hubley R, Green P. RepeatMasker Open-4.0 [Internet]. 2013. Available from: http://www.repeatmasker.org
19. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes [Internet]. 2016 [cited 2018 Feb 26];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751634/
20. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinforma Oxf Engl. 2012;28:3211–7.
21. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
22. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. 2013;41:D590–6.
23. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Comput Biol. 2016;12:e1004957.
24. Renaud G, Hanghøj K, Willerslev E, Orlando L. gargammel: a sequence simulator for ancient DNA. Bioinformatics. 2017;33:577–9.
25. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinforma Oxf Engl. 2013;29:1682–4.
26. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv. 2016;050559.
27. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics. 2014;30:3123–4.
28. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.
29. Olalde I, Allentoft ME, Sánchez-Quinto F, Santpere G, Chiang CWK, DeGiorgio M, et al. Derived immune and ancestral pigmentation alleles in a 7,000-year-old Mesolithic European. Nature. 2014;507:225–8.
30. Siqueira JF, Rôças IN. Pseudoramibacter alactolyticus in Primary Endodontic Infections. J Endod. 2003;29:735–8.

31. Martínez-Porchas M, Villalpando-Canchola E, Vargas-Albores F. Significant loss of sensitivity and specificity in the taxonomic classification occurs when short 16S rRNA gene sequences are used. Heliyon [Internet]. 2016 [cited 2018 Apr 26];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5037269/

32. Fuks G, Elgart M, Amir A, Zeisel A, Turnbaugh PJ, Soen Y, et al. Combining 16S rRNA gene variable regions enables high-resolution microbial community profiling. Microbiome. 2018;6:17.

33. Chakravorty S, Helb D, Burday M, Connell N, Alland D. A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. J Microbiol Methods. 2007;69:330–9.

34. Jovel J, Patterson J, Wang W, Hotte N, O'Keefe S, Mitchel T, et al. Characterization of the Gut Microbiome Using 16S or Shotgun Metagenomics. Evol Genomic Microbiol. 2016;459.

35. Warinner C, Herbig A, Mann A, Yates JAF, Weiß CL, Burbano HA, et al. A Robust Framework for Microbial Archaeology. Annu Rev Genomics Hum Genet. 2017;18:null.

36. Ranjan R, Rani A, Metwally A, McGee HS, Perkins DL. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. Biochem Biophys Res Commun. 2016;469:967–77.

Supplementary figures and tables

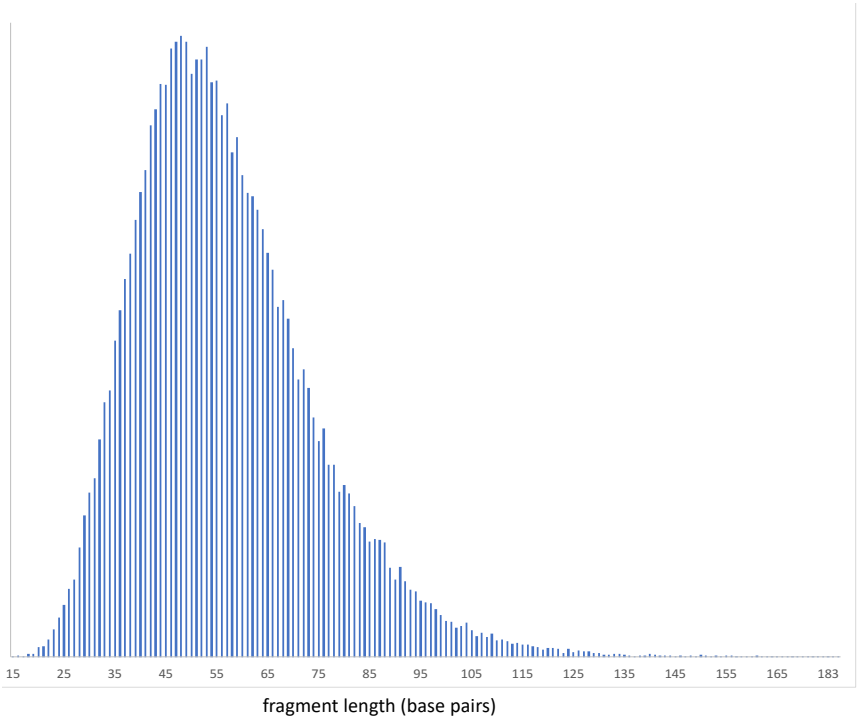


Figure S1. Read length distribution of simulated metagenome

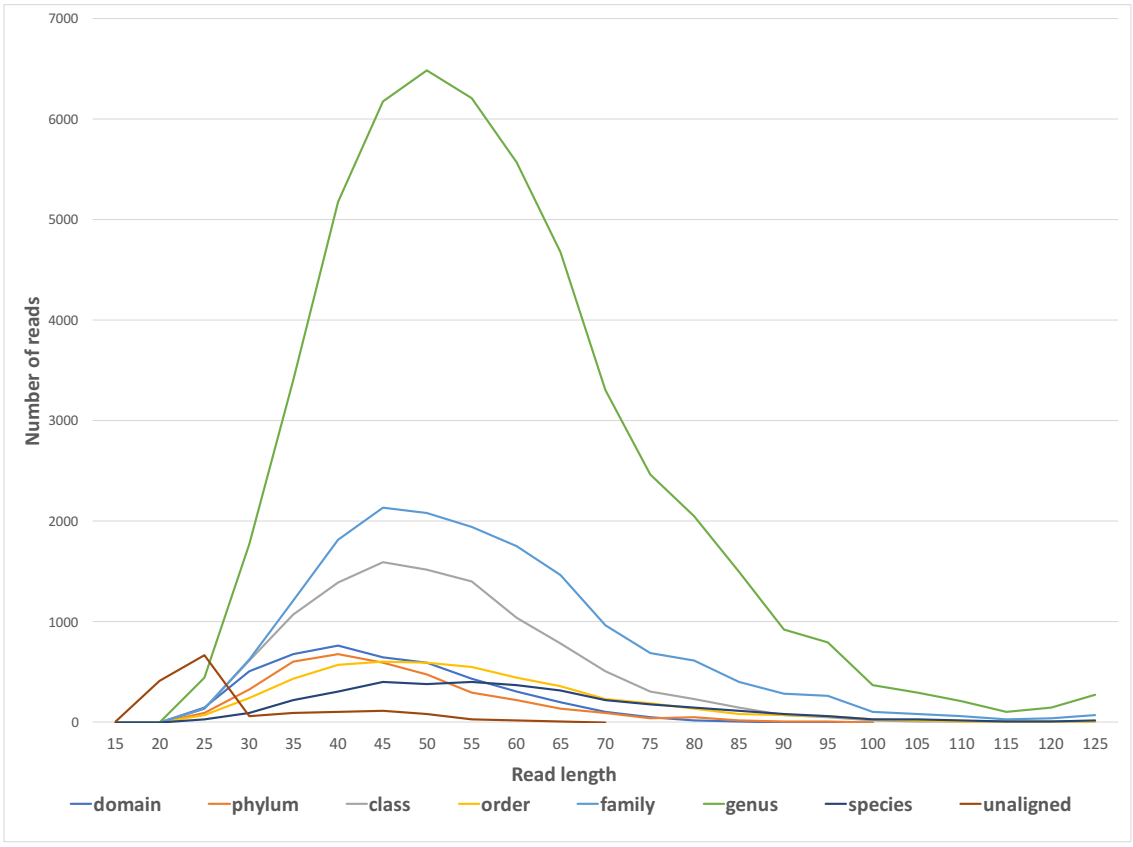


Figure S2. Read length distribution of aligned reads at different taxonomic ranks

Table. S1. Simulated 16S metagenome composition

Taxon	Abundance	
Bacillus_anthracis	5.0%	
Corynebacterium_matruchotii	2.5%	
Corynebacterium_durum	2.5%	Removed for species exclusion experiment
Escherichia_coli	2.5%	
Fretibacterium_fastidiosum	2.5%	
Johnsonella_ignava	2.5%	Removed for species exclusion experiment
Methanobrevibacter_oralis	12.5%	
Moranella_endobium	2.5%	
Neissera_sicca	7.5%	Removed for species exclusion experiment
Rothia_aeria	19.5%	
Staphylococcus_aureus	0.5%	
Streptococcus_mitis	4.0%	Removed for genus exclusion experiment
Streptococcus_mutans	1.0%	Removed for genus exclusion experiment
Streptococcus_oralis	5.0%	Removed for genus exclusion experiment
Streptococcus_sanguinis	5.0%	Removed for genus exclusion experiment
Streptococcus_suis	10.0%	Removed for species and genus exclusion experiment
Thiomargarita_namibiensis	5.0%	
Treponema_denticola	5.0%	
Yersinia_pestis	5.0%	

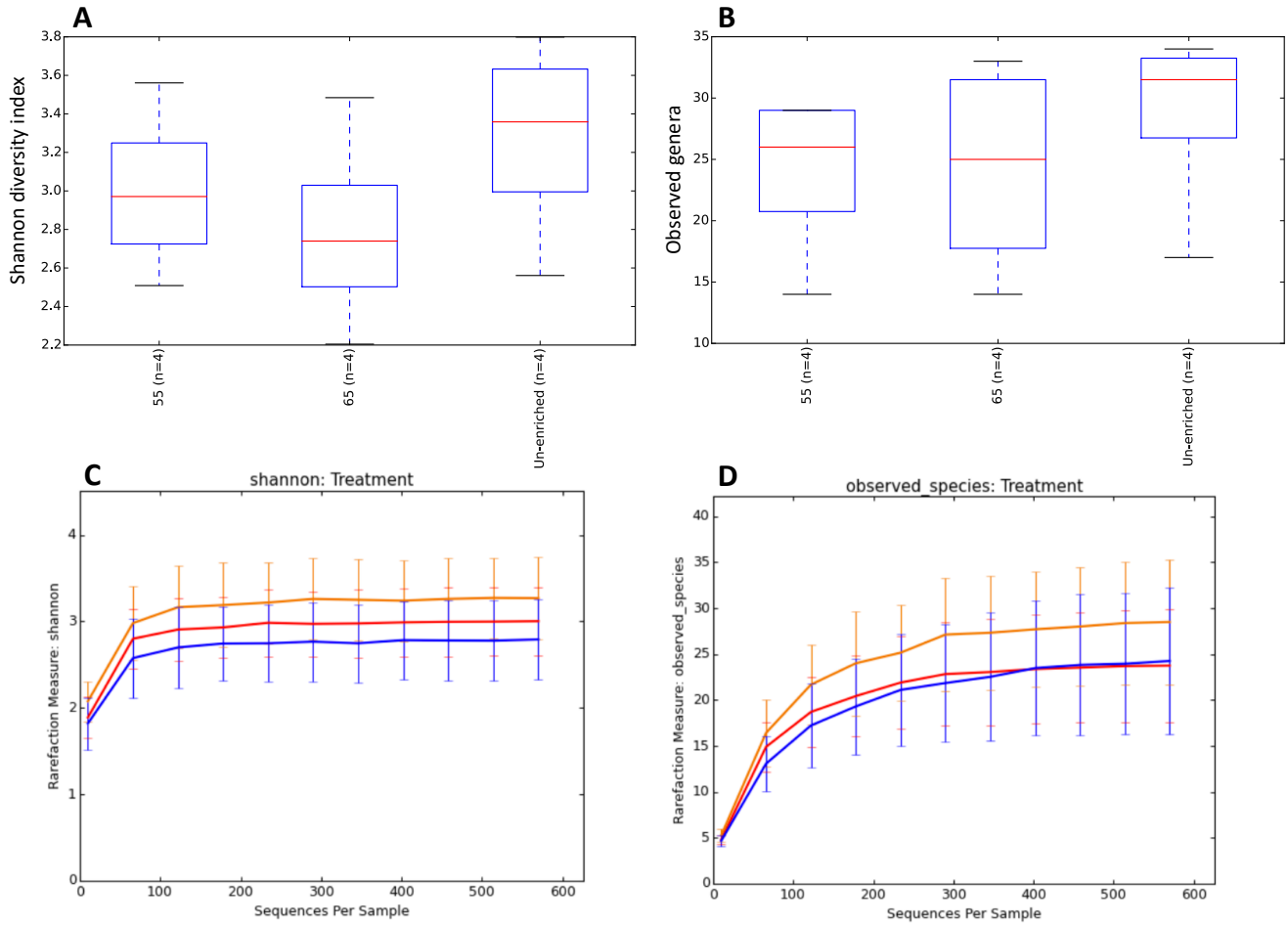


Figure S2. Influence of enrichment treatment on microbial alpha diversity. (A) Shannon diversity index. (B) Observed genera. (C) Alpha rarefaction curve of Shannon diversity index. (D) same as C, but using the observed genera metric.

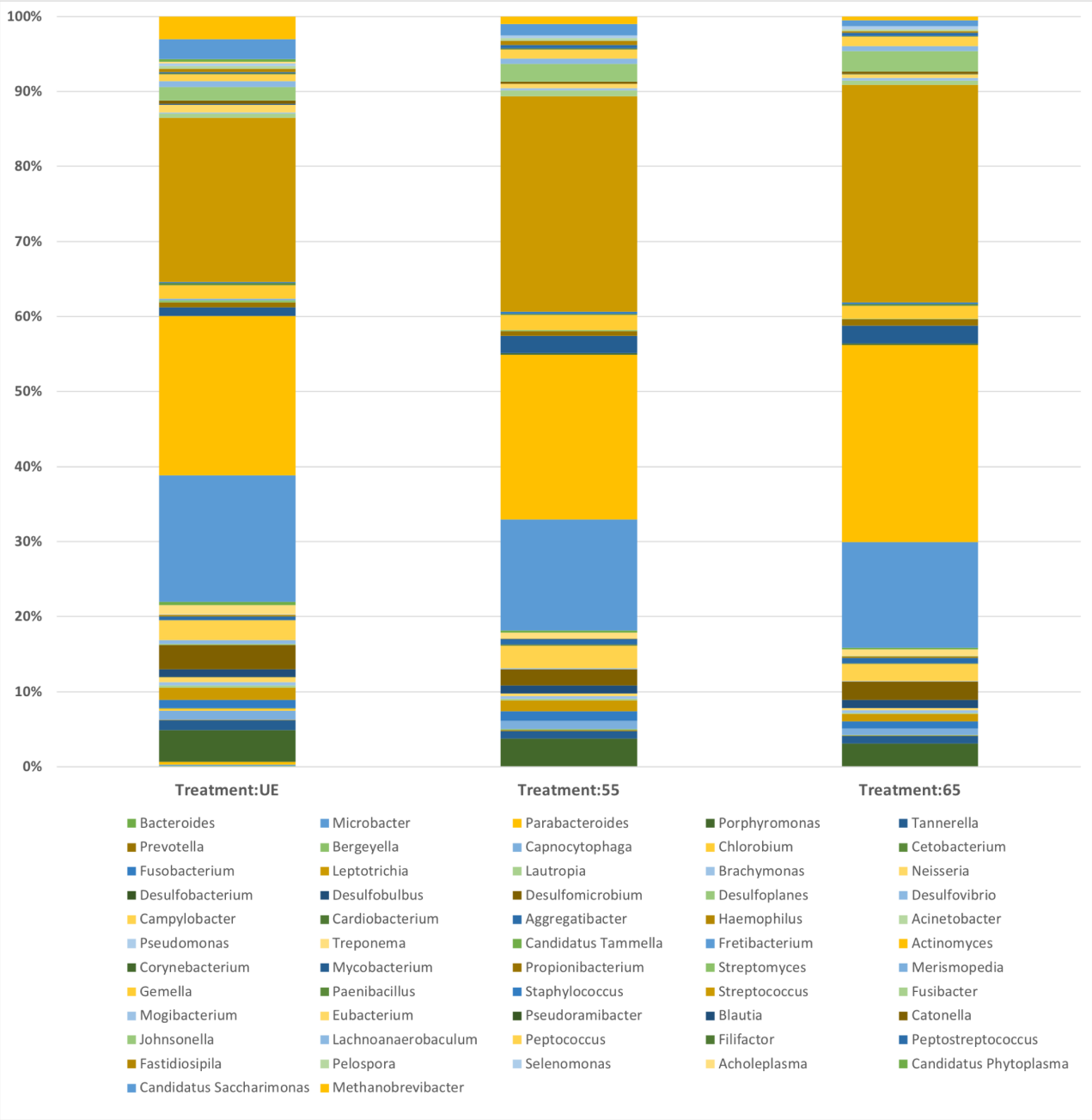


Figure S4. Taxonomic composition of samples collapsed by enrichment treatment type

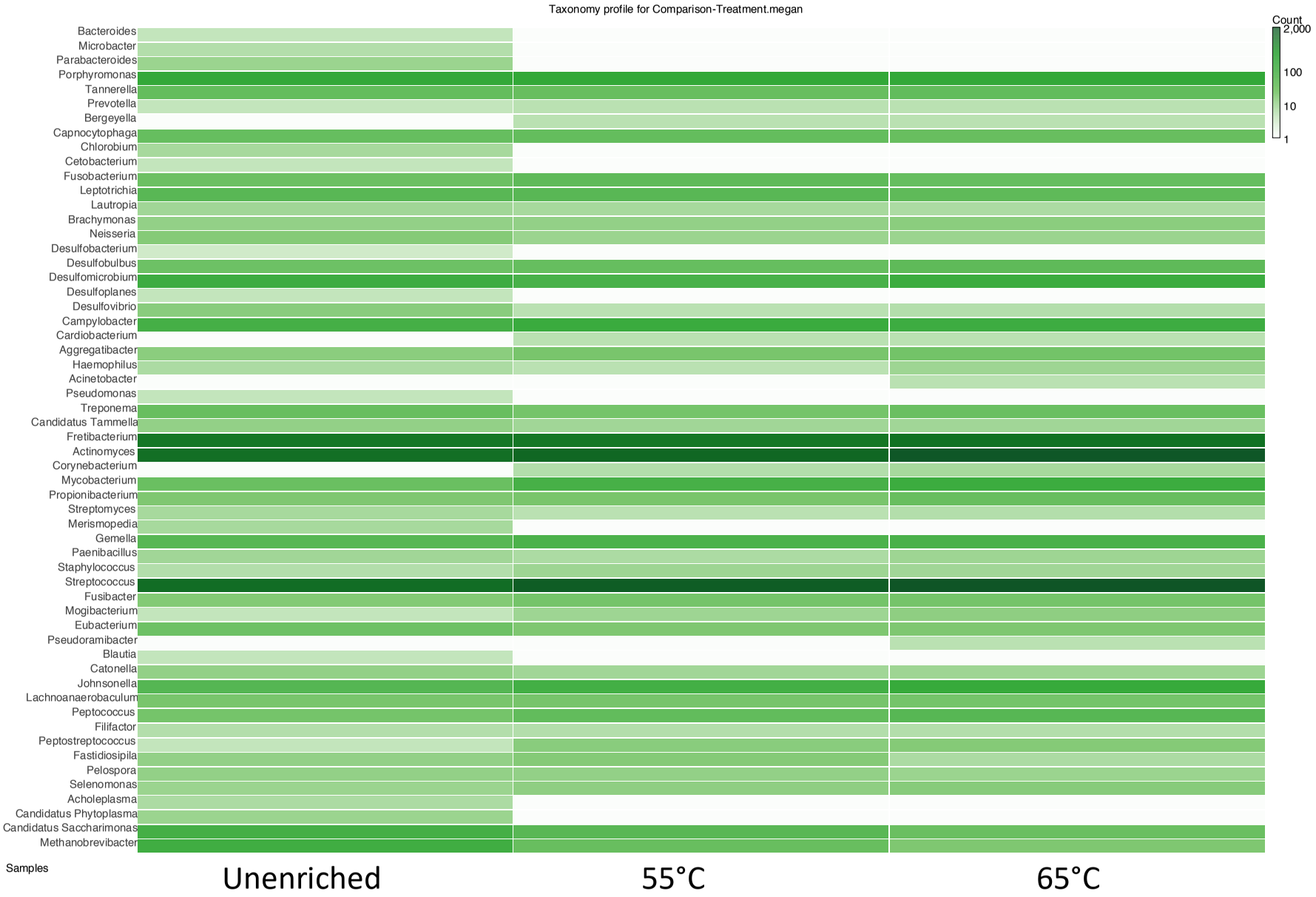


Figure S5. Heat map of genus-level assignments collapsed by enrichment treatment

Table S2. Enrichment statistics for subsampled, 16S-enriched samples

Sample	total sequences	putative 16S fragments	% putative 16S fragments	%duplicates	%GC	avg_sequence_length
A12014_EuroHG1-16S-Enriched	100,000	68,369	68.4%	91.5%	49%	81
A12017_EuroHG2-16S-Enriched	93,208	64,593	69.3%	65.7%	52%	80
A12826_LBK1-16S-Enriched	100,000	54,794	54.8%	28.3%	52%	67
A12829_LBK2-16S-Enriched	100,000	52,971	53.0%	8.9%	51%	68
A12873_Chimp-16S-Enriched	100,000	47,463	47.5%	91.5%	54%	50
A13204_AfrPP2-16S-Enriched	100,000	52,010	52.0%	43.7%	49%	71
A13208_AfrSF2-16S-Enriched	100,000	59,146	59.1%	20.8%	50%	78
A13209_AfrSF3-16S-Enriched	100,000	51,305	51.3%	10.6%	51%	63
A13210-AfrSF4-16S-Enriched	100,000	56,234	56.2%	87.5%	44%	65
A13213_AfrPP1-16S-Enriched	100,000	52,847	52.8%	10.4%	50%	76
A13232_IndRev1-16S-Enriched	100,000	48,456	48.5%	21.7%	50%	88
A13234_IndRev2-16S-Enriched	100,000	49,201	49.2%	9.3%	52%	70
A8812_JewBury1-16S-Enriched	51,926	36,720	70.7%	21.3%	50%	75
A8824_JewBury2-16S-Enriched	71,125	60,560	85.1%	54.3%	51%	85
AFR8_EBC-16S-Enriched	9,076	5,388	59.4%	97.0%	52%	86
Average all	88,356	50,670	58%	44%	50%	73
Average excluding EBCs	94,019	53,905	57%	40%	50%	73

Table S3. Sequence statistics for subsampled, unenriched shotgun samples

Sample	total_sequences	%duplicates	%GC	avg_sequence_length
A12014_EuroHG1-UnEnriched	100,000	22.1	44	53
A12017_EuroHG2-UnEnriched	93,208	2.0	47	51
A12826_LBK1-UnEnriched	100,000	1.1	51	47
A12829_LBK2-UnEnriched	100,000	0.6	52	47
A12873-Chimpanzee-UnEnriched	100,000	14.9	53	57
A13204_AfrPP2-UnEnriched	100,000	2.3	42	43
A13208_AfrSF2-UnEnriched	100,000	0.1	49	53
A13209_AfrSF3-UnEnriched	100,000	0.2	46	41
A13210-AfrSF4-UnEnriched	100,000	12.3	43	44
A13213_AfrPP1-UnEnriched	100,000	0.2	43	44
A13232_IndRev1-UnEnriched	100,000	1.6	50	49
A13234_IndRev2-UnEnriched	100,000	0.1	56	45
A8812_JewBury1-UnEnriched	51,926	0.8	52	51
A8824_JewBury2-UnEnriched	71,125	1.0	51	50
AFR8-EBC-UnEnriched	9,076	68.2	45	39
Average all	88,356	8.5	48.3	48
Average excluding EBC	94,019	4.2	48.5	48

Table. S4. Number of reads assigned to genera in extraction blank controls

Genus	AFR8_EBC-16S	AFR8-EBC-Shotgun
Pedobacter	0	89
Paracoccus	99	0
Comamonas	393	25
Enterobacter	7	13
Serratia	0	20
Acinetobacter	4624	559
Pseudomonas	110	85
Stenotrophomonas	10	0
Mycobacterium	0	13
Brachybacterium	0	9
Staphylococcus	0	491
Enterococcus	0	4215

Table S5. Reads filtered from extraction blank controls per sample

Sample	# reads assigned genus in sample	# reads removed by filtering	% removed
A8812_JewBury1-Shotgun-Unenriched	23,037	0	0.00%
A8824_JewBury2-Shotgun-Unenriched	31,609	0	0.00%
A12014_EuroHG1-Shotgun-Unenriched	28,822	0	0.00%
A12017_EuroHG2-Shotgun-Unenriched	24,780	0	0.00%
A12826_LBK1-Shotgun-Unenriched	27,502	0	0.00%
A12829_LBK2-Shotgun-Unenriched	43,629	0	0.00%
A12873_Chimp-Shotgun-Unenriched	31,218	0	0.00%
A13204_AfrPP2-Shotgun-Unenriched	21,421	0	0.00%
A13208_AfrSF2-Shotgun-Unenriched	13,808	0	0.00%
A13209_AfrSF3-Shotgun-Unenriched	20,463	804	3.93%
A13210-AfrSF4-Shotgun-Unenriched	15,631	0	0.00%
A13213_AfrPP1-Shotgun-Unenriched	24,296	0	0.00%
A13232_IndRev1-Shotgun-Unenriched	45,252	0	0.00%
A13234_IndRev2-Shotgun-Unenriched	47,189	0	0.00%
A8812_JewBury1-16S-Enriched	16,459	455	2.76%
A8824_JewBury2-16S-Enriched	28,868	486	1.68%
A12014_EuroHG1-16S-Enriched	27,346	199	0.73%
A12017_EuroHG2-16S-Enriched	26,710	777	2.91%
A12826_LBK1-16S-Enriched	17,777	829	4.66%
A12829_LBK2-16S-Enriched	20,952	727	3.47%
A12873_Chimp-16S-Enriched	11,654	175	1.50%
A13204_AfrPP2-16S-Enriched	18,018	386	2.14%
A13208_AfrSF2-16S-Enriched	10,556	107	1.01%
A13209_AfrSF3-16S-Enriched	12,359	816	6.60%
A13210-AfrSF4-16S-Enriched	15,648	2,134	13.64%
A13213_AfrPP1-16S-Enriched	21,135	297	1.41%
A13232_IndRev1-16S-Enriched	24,607	464	1.89%
A13234_IndRev2-16S-Enriched	21,030	679	3.23%
Average Shotgun-Unenriched	28,476	57	0.20%
Average 16S-Enriched	19,509	609	3.12%
Average All samples	23,992	333	1.39%

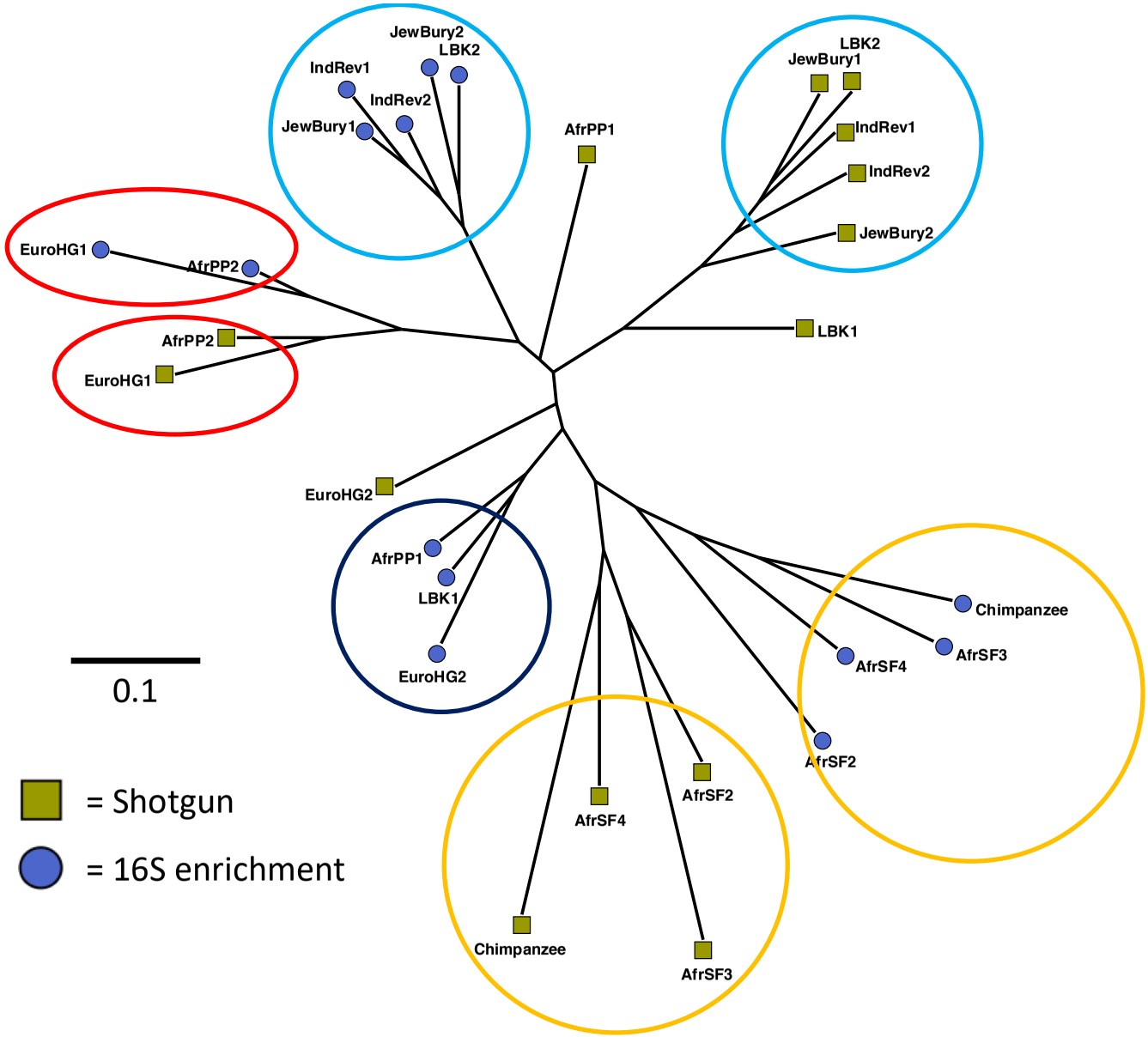


Figure S6. NJ tree of genus-level Bray-Curtis distances between samples

Chapter V

—

Palaeomicrobiology of the Pacific: unlocking a high-resolution proxy for past human movements

Statement of Authorship

Title of Paper	Palaeomicrobiology of the Pacific: unlocking a high-resolution proxy for past human movements
Publication Status	☐ Published ☐ Accepted for Publication ☐ Submitted for Publication ☑ Unpublished and Unsubmitted w ork w ritten in manuscript style
Publication Details	Unpublished and unsubmitted work written in manuscript style.

Principal Author

Name of Principal Author (Candidate)	Raphael Eisenhofer		
Contribution to the Paper	Performed all the laboratory procedures. Analysed the data and interpreted the results. 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	10/5/18

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	Alan Cooper		
Contribution to the Paper	Collected the samples and edited the manuscript.		
Signature		Date	23/April/2018

Name of Co-Author	Laura S. Weyrich		
Contribution to the Paper	collected samples, obtained funding, collected metadata, designed study, edit manuscript		
Signature		Date	9/5/18

Palaeomicrobiology of the Pacific: unlocking a high-resolution proxy for past human movements

Authors: Raphael Eisenhofer[1], Alan Cooper[1], and Laura S. Weyrich[1]

Affiliations:

1: Australian Centre for Ancient DNA, School of Biological Sciences, University of Adelaide, South Australia

Abstract

The peopling of the Pacific represents the greatest feat of maritime settlement in human history, requiring long-distance seafaring technology to reach thousands of islands spread across 30% of the Earth's surface. While the timing of human settlement in the Pacific is well defined, the source populations and the routes taken during the settlement of Polynesia are not. Here, we generate and authenticate ancient oral microbial DNA data for 117 ancient and historical human dental calculus samples from across the Pacific. Using this data, we develop and test two methods for inferring past human movements using ancient microbial DNA preserved in dental calculus. We also compare and contrast oral microbial communities (microbiota) from these samples to the oral microbiota of modern individuals from the Human Microbiome Project. Finally, we explore the influence of oral diseases on the oral microbiota composition of ancient humans. This study expands our understanding of global oral microbiota diversity and is the first to demonstrate that microbial DNA preserved in ancient dental calculus can be used to investigate past human movements.

Main text

Roughly 3,000 years before present (BP) people of the Lapita culture were the first to settle Remote Oceania—the previously unoccupied region from Vanuatu eastwards (Figure 1; orange and yellow shaded areas) [1,2]. These early Lapita, otherwise known as the First Remote Oceanians, derived almost all of their ancestry from East Asian/ISEA (Island South-East Asia) populations [3] and maintained fleets of boats capable of long-distance seafaring, rapidly setting up colonies as far as Samoa and Tonga (Figure 1) [4]. However, there is no evidence of Lapita culture reaching further eastwards than Tonga and Samoa despite their long-term occupation and seafaring abilities, raising questions about the last 1,000 years of human movement and settlement in Eastern Polynesia. The earliest evidence for the initial settlement of East Polynesia is 1,000-800 BP [5,6], but the routes taken to settle these islands are still highly contentious. Some argue for eastern movement from Samoa [5,7,8], while others suggest that Eastern Polynesia was initially settled via Marquesas from the central northern outliers in the Solomon Islands [9–11].

Our ability to infer rapid, recent past human movements is limited by a variety of factors, including resolution (*e.g.* rate of evolution), modern day or post-settlement admixture events that confound past demographic signals, and ethical issues relating to the analysis of human DNA from indigenous cultures [12]. A solution to this is to examine human proxies that have faster generation times which can yield higher rates of mutation (*e.g.* chickens [13]; rats [14]; or microorganisms [15]). Calcified dental plaque (calculus) is a robust microbial biofilm that preserves human-associated oral microorganisms within the archaeological record, providing an unprecedented opportunity to study past human diet, health, and culture [16,17]. Because oral microorganisms exhibit a strong degree of vertical inheritance [18–22], it could be possible to use microbial DNA preserved in ancient dental calculus as a high-resolution proxy of past human movement [12]. Here, we sequence microbial DNA preserved within 130 ancient and historic dental calculus specimens from 16 geographic areas throughout Island South East Asia (ISEA) and the Pacific Islands (Figure 1; Table S1) and, for the first time, use the evolutionary history of oral microorganisms as a proxy to reconstruct past human movements in the Pacific.

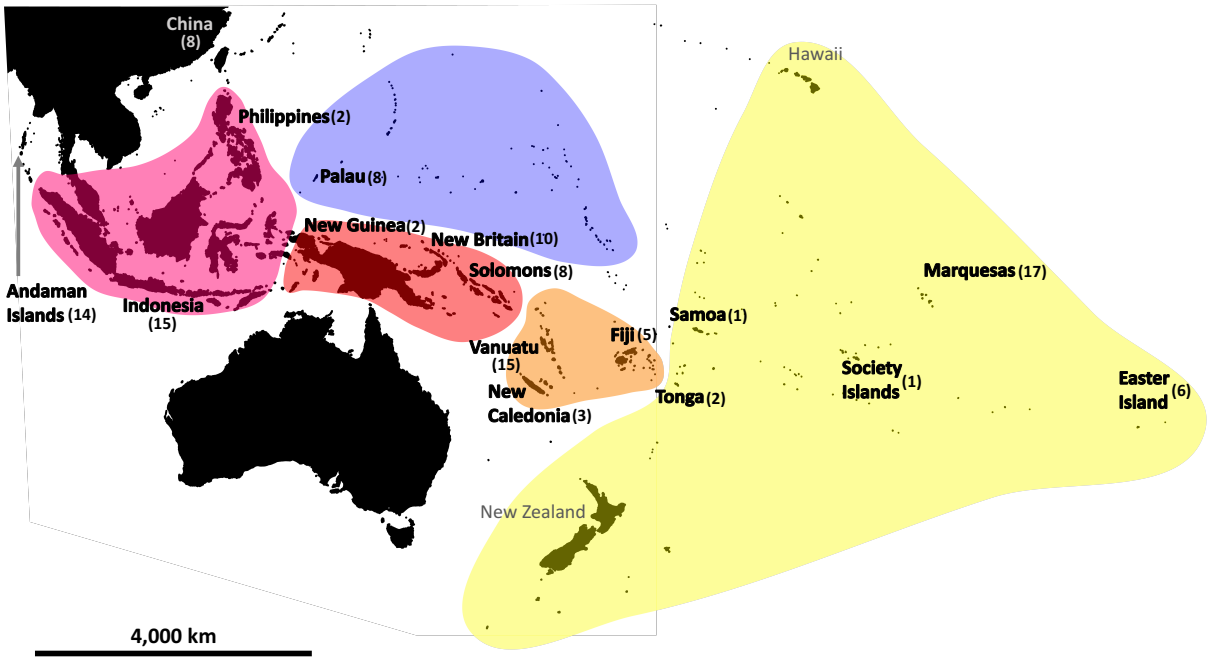


Figure 1. Map of the Pacific ocean illustrating geographic areas and islands where ancient dental calculus were obtained. Shaded colours represent broader geographic regions, and correspond to colours in the phylogenetic tree in Figure 3.

Historic (n=120) dental calculus samples collected from the Asia Pacific region in the 1800s (prior to large-scale European interactions in the region) were obtained from museums, and 10 ancient dental calculus samples were obtained from archaeological sites in Vanuatu (Teouma Lapita; 2,940-2,710 BP [23]) and Palau (Chelechol ra Orrak; 3000-1700 BP [24]). We sequenced these ancient and historic calculus samples to an average depth of 2.2 million sequences (±1.3 million) and obtained data for 210 modern supragingival plaque samples from the human microbiome project (HMP)[25], which we subsampled to 1.5 million sequences each. We classified taxonomic composition for all samples using MALTn [26] against a database containing 47,696 bacterial/archaeal genome assemblies [27]. We then applied a conservative approach and removed any microbial taxa identified in extraction blank controls (EBCs) from the dental calculus samples (Supplementary Note 1); post-filtering, 117 of the ancient dental calculus and all modern plaque samples exhibited a strong oral microbiome signal (Supplementary Note 1 & SI figure 1).

Given the widespread geographic sampling in study, we focused our taxonomic composition analyses on genus level classifications to allow for better comparisons between cultures (Supplementary Note 2). Of the 137 microbial genera that we identified in historic and ancient dental calculus samples, a set of seven genera were widespread (present in >75% of samples) and had the highest mean abundances ranging from 3.7%—23.6% (Figure 2A). These

seven genera accounted for 73.1% of all taxonomic assignments (Figure 2B) and each possessed DNA damage patterns typical of authentic ancient DNA (Figures S3-16). For the modern HMP plaque samples, a similar trend was observed, with the top seven genera being present in >95% of samples, and having the highest mean abundances ranging from 5.3—20.3% (Figure 2A). These genera also accounted for 71.1% of all taxonomic assignments (Figure 2B), which is in accord with previous modern plaque studies [28,29]. Interestingly, the only shared genus in the top seven genera for both datasets is *Actinomyces*, which could suggest large differences in the abundances potentially due to diet [17], disease, or geographic isolation. Additionally, issues with classifying ancient microorganisms due to missing reference genomes [27], or unknown taphonomic issues pertaining to ancient DNA degradation (*i.e.* preferential degradation of some taxa over others) could also explain this finding.

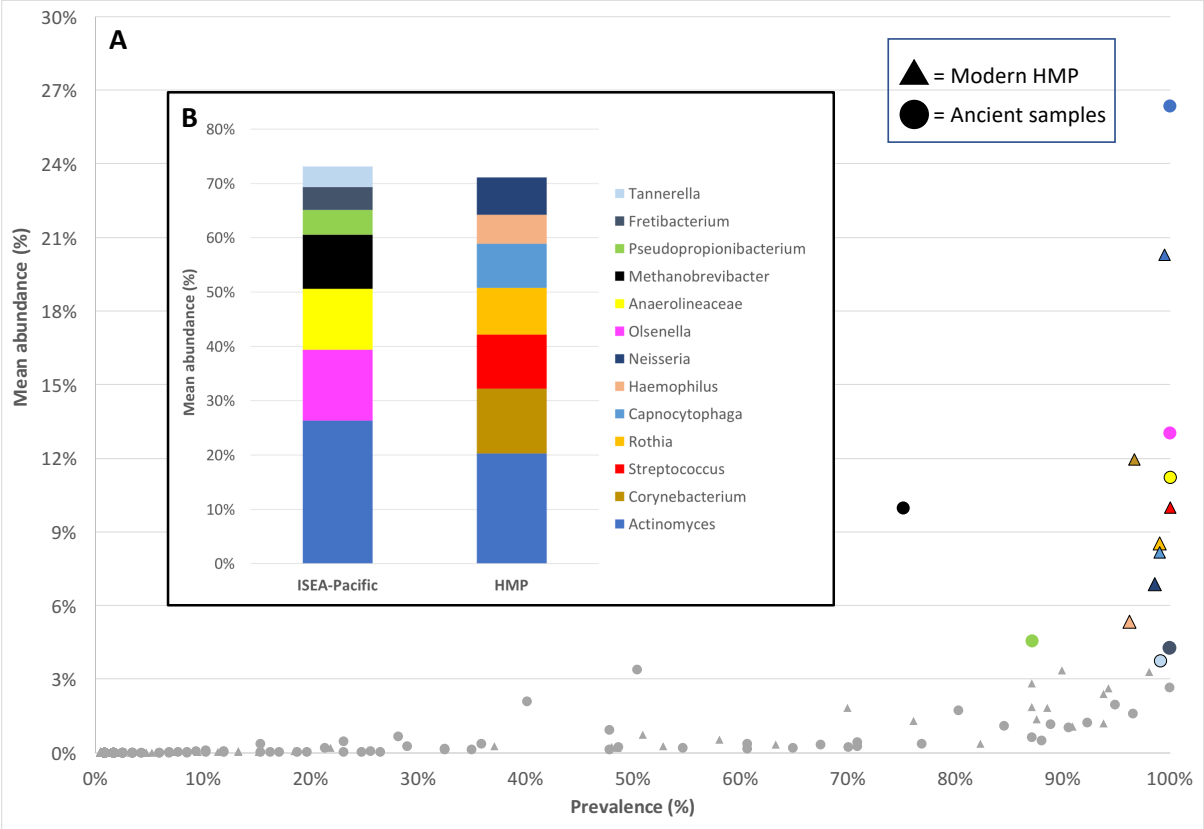


Figure 2. Core genera identified in ancient ISEA/Pacific samples. (A) Prevalence and abundance plot of core genera identified in ancient dental calculus and modern HMP plaque samples. (B) Total mean abundance of core genera across both datasets. The seven top genera account for >70% of the classified sequencing data.

We hypothesised that the different human cultures sampled would possess distinct oral microbiomes due to geographic isolation, diet, or cultural practices. After controlling for tooth sampled (*e.g.* molar, incisor), which is known to influence oral microbiome composition [30,31], we were left with 74 molar dental calculus samples (Supplementary Note 2). While we did not find statistically significant differences in alpha (within group) diversity between the different islands and geographic regions tested, we found small but statistically significant

differences in community composition by island and geographic region for both abundance-weighted and unweighted distance metrics (PERMANOVA, ANOSIM p-values <0.005) (Table S3) (Supplementary Note 2). This could suggest that individuals and cultures from different geographic regions possess distinct oral microbiomes.

We next explored whether evidence of dental disease (caries, periodontal disease) influenced microbial composition in our dataset (Supplementary Note 2). Differences in microbial composition between samples grouped by evidence of dental caries or periodontal disease were not statistically significant (Supplementary Note 2). However, we found six specific oral genera that were only present in the ancient dental calculus samples but not HMP plaque samples: *Anaerolineaceae, Methanobrevibacter, Pseudoramibacter, Desulfomicrobium, Desulfobulbus,* and *Slackia*. These genera are present in the HOMD (Human Oral Microbiome Database) and have been associated with periodontal disease [32–36]. The absence of these periodontal disease-associated taxa in modern HMP plaque samples is expected as these samples were collected from healthy individuals [25]. We found that the abundance of *Methanobrevibacter* appeared to correlate with the axis of greatest variation in Principal coordinates analysis (PCoA) (Figure S24A) (Supplementary Note 2). This clustering was not observed using the unweighted Binary Jaccard metric (Figure S24B), suggesting that the abundance of a periodontal-associated *Methanobrevibacter* alone was enough to influence abundance-weighted distance metrics in ancient microbiome analysis. Therefore, periodontal disease could be a confounding factor when analysing ancient microbial communities and should be evaluated in future studies using larger datasets with more robust and even sampling.

Given that taxonomic composition is unlikely to be informative regarding past human movements, we next focused on genomic analyses of the seven 'core' genera we identified in ancient samples as these represent the best candidates for vertically inherited microbial species—a trait necessary for phylogenetic analysis and inference of past host movements. We developed and investigated two approaches. The first methodology involved the hybridization enrichment of specific, phylogenetically informative microbial genes to examine phylogenetic relationships between samples; many of the relationships identified using this method had poor phylogenetic support and are unsupported by current evidence (Supplementary Note 4; Figures S36-48). The second approach involved the analysis of low-coverage whole bacterial genomes (Supplementary Note 3). Because low coverage phylogenetic reconstruction can be prone to the influence of cytosine deamination (an ancient DNA damage-based substitution), we used transversions substitutions only. Of these seven genera, *Anaerolineaceae sp. oral taxon 439* produced a robust whole genome phylogeny possessing strong node support values, and

phylogenetic relationships that corroborate the available evidence on past human movements in the Pacific (Figure 3) [3,37,38].

We found two major clades, one containing samples from ISEA, China, and the Andaman Islands (Figure 3; pink), and the other containing all of the Pacific samples. The first split within the Pacific clade is between the ~3,000-year-old Lapita (Teouma) and 3,000-1,700-year-old Palau samples (Figure 3; blue), and the rest of the Pacific samples (Figure 3; red, orange, yellow). The close affinity of the Teouma Lapita sample (~3,000 BP) with the Palau samples is in line with current evidence for early occupation of Palau ~3,300-3,000 BP [39,40], with archaeological evidence pointing towards ISEA (especially the Philippines) as a likely source of settlement [41]. The placement of the Teouma Lapita sample (representing the First Remote Oceanians) outside of the Papuan clade (red) supports previous ancient human DNA findings that the First Remote Oceanians had little, if any, Papuan ancestry and possessed highest affinity to East Asian/Taiwan and populations from the ISEA [3]. Additionally, the placement of the Teouma Lapita sample outside of the Vanuatu and Polynesian clade (Figure 3; orange and yellow) is also supported by recent ancient human DNA and linguistic evidence for later waves of Papuan movement and admixture with First Remote Oceanians [37,38]. This may explain the observed loss of the ancient First Remote Oceanian *Anaerolineaceae* lineage from the more recent Vanuatu and Polynesian samples (Figure 3).

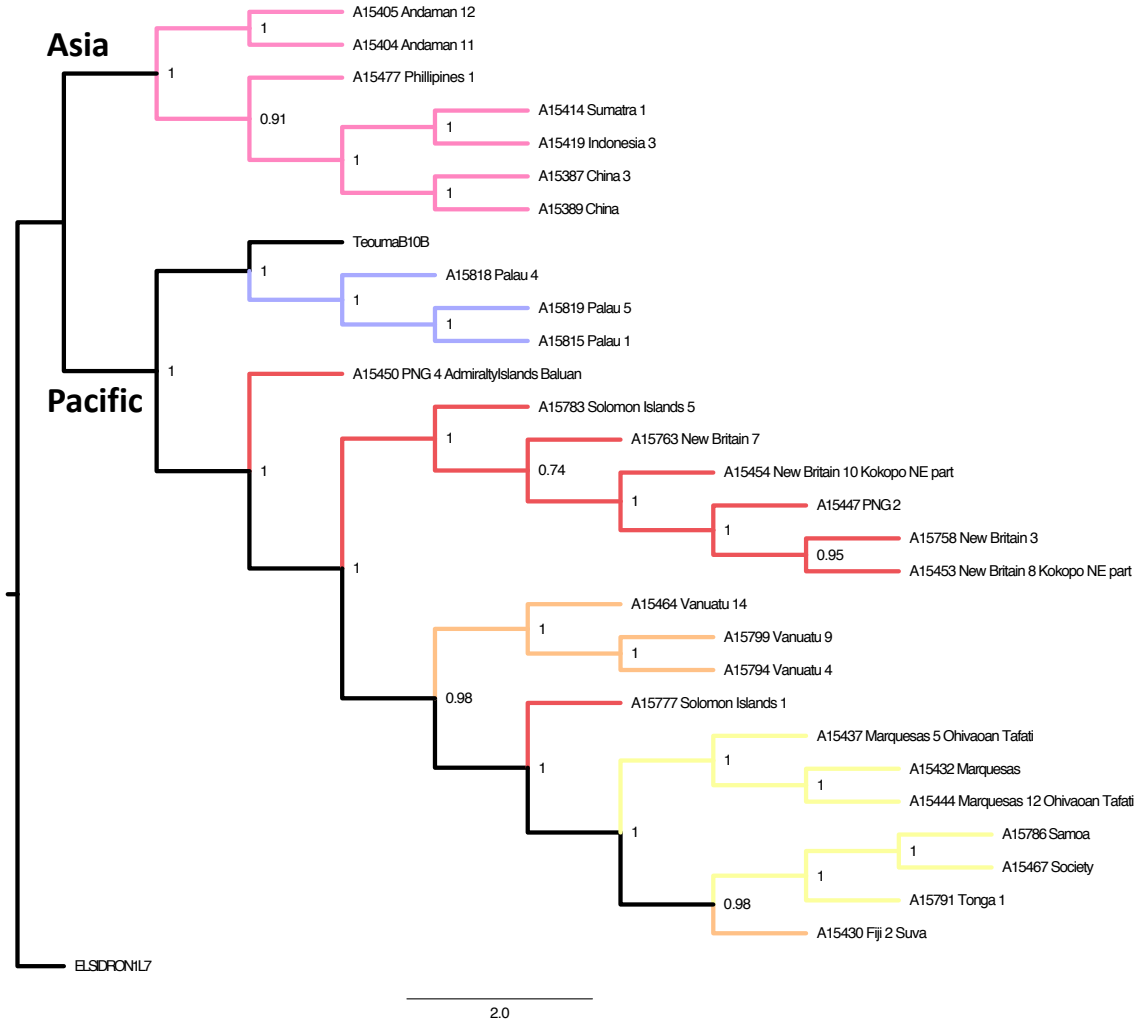


Figure 3. Transversions-only maximum composite likelihood neighbour joining tree. Node values indicate percentage of support for 1,000 bootstrap replicates. Branch colours correspond to geographic regions highlighted in Figure 1.

Our data also suggests that Eastern Polynesia was settled by an initial movement from the northern Solomon outliers to the Marquesas [9–11], rather than eastwards movement from Samoa [5,7,8]. This is supported by the placement of a Solomon Island sample outside of the Polynesian clade (yellow), coupled with a major split between the Marquesan and other Polynesian samples (Figure 3; yellow). Overall, these relationships are consistent with current evidence of human settlement in the region, and future sampling (both spatially and temporally) will undoubtedly reveal more about the demographic history of the Pacific.

While the *Anaerolineaceae* genome appeared to provide a robust phylogeny, the other species tested yielded relationships not currently supported by current evidence (Figure S26-34). There are numerous potential reasons for this. First, the sequencing effort in our study may be too shallow, resulting in missing data and making it difficult to align whole-genome sequences (Supplementary Note 3). Indeed, when we attempted to call consensus sequences of our *Anaerolineaceae* genomes with a minimum depth of 3, the resulting whole-genome alignments were severely truncated and upon visual inspection, poorly aligned. In contrast, our *Anaerolineaceae* consensus sequences called with no minimum depth cut-off yielded a robust

genome alignment and phylogenetic tree, suggesting that missing data was interfering with our attempts at phylogenetic reconstruction. While missing data likely contributed, it is insufficient to explain why phylogenies produced by species with higher abundance and therefore more sequence coverage (*Olsenella* and *Actinomyces* — Figure 2) failed to recapitulate past human movements. Other factors include cross-mapping of DNA reads due to inappropriate reference genome choice, or the biology of the microorganisms used. Given that we could only align an average of 51.7% of ancient DNA reads to modern reference genomes (Supplementary Note 2) [27], we are potentially unable to identify microbial species in our ancient data. This could be an issue for the phylogenetic reconstruction of some microbial species, as we cannot perform competitive mapping (Supplementary Note 3) to prevent cross-mapping of reads from other species—a situation that would violate phylogenetic inference [42]. The biology of the microorganisms used could also explain our findings. For example, some microbial species exhibit more homologous recombination and horizontal gene transfer than others [43], which would further hamper phylogenetic inference [44,45]. Additionally, differences in the degree of heritability and stability through time of microbial species could also influence their use phylogenetically. Clearly, further research is needed to understand why some oral species are not suitable for inferring past human movements, and to increase the number of species available for future studies. Until then, it appears that *Anaerolineaceae sp. oral taxon 439* is a viable proxy for determining past human movements.

In conclusion, this study is the first to examine the oral microbial communities of ancient and historic ISEA and Pacific peoples. The high-quality and authentic data that we generated will be useful for learning more about the history and culture of these regions. We are also the first to demonstrate that microbial DNA in ancient dental calculus can be used as a proxy for past human demographic history, and our results lend further support to current evidence about the peopling of the Pacific. Future studies incorporating deeper sequencing depth and greater temporal/spatial sampling should allow new insights into the peopling of East Polynesia. Furthermore, the method employed here could also be used to shed light on the past demographic histories of other cultures around the world.

Materials and methods

Sample collection

Historical dental calculus samples were collected from the Musée de l'Homme in Paris, France, and the Natural History Museum of London, U.K. Ancient Lapita samples were obtained from the Teouma site in Vanuatu [1], and ancient Palau samples were obtained from the Chelechol

ra Orrak site [39]. All samples were collected by individuals wearing face-masks and gloves, and dental picks were decontaminated between samples. Samples were stored in sterile plastic bags and transported at room temperature to the University of Adelaide, where they were refrigerated at 4°C until DNA extraction.

Sample processing and DNA extraction

All sample processing and molecular biology procedures prior to PCR amplification were carried out at the Australian Centre for Ancient DNA facility at the University of Adelaide. Experiments were performed within UV-treated, 3% bleach cleaned, still-air hoods located in isolated, still-air rooms to limit the introduction of modern contaminant DNA. Dental calculus samples were decontaminated to minimise environmental contamination by UV-irradiation for 15 minutes on each side, following by soaking in 2 ml of 5% bleach for 3 minutes, rinsing in 90% ethanol for 1 minute, and drying at room temperature for two minutes. Immediately post-decontamination, dental calculus samples were crushed on the side of plastic tubes with sterile tweezers, and DNA extracted using an in-house silica-based method described previously [46]. Because of the highly degraded nature of the two Teouma Vanuatu samples (B10B and B10C), DNA was extracted using a different method to obtain shorter DNA fragments (Method Y [47]). Extraction blank and no-template library controls were included alongside all samples and were sequenced.

DNA library preparation and sequencing

Shotgun metagenomic libraries were constructed as previously described [48], using unique combinations of 7-bp forward and reverse barcodes [48]. Thirteen cycles of amplification were completed with P5/P7 barcoded adapters, followed by an additional thirteen cycles for the addition of GAII-index and sequencing primers. Metagenomic shotgun libraries were cleaned using Ampure XP, quantified using an Agilent TapeStation, and pooled at equimolar concentrations prior to sequencing on the Illumina HiSeq X Ten platform (2 x 150 bp). Shotgun metagenomic libraries were constructed as previously described [48], using unique combinations of 7-bp forward and reverse barcodes [48]. Thirteen cycles of amplification were completed with P5/P7 barcoded adapters, followed by an additional thirteen cycles for the addition of GAII-index and sequencing primers. Metagenomic shotgun libraries were cleaned using Ampure XP, quantified using an Agilent TapeStation, and pooled at equimolar concentrations prior to sequencing on the Illumina NextSeq and X10 platforms.

Hybridization enrichment

Thirty-two samples including an extraction blank control (SI table S9) were chosen for hybridization enrichment using custom-designed RNA baits (Supplementary Note 4). All samples were amplified in five individual reactions to reduce clonality, cleaned, and pooled to obtain 100 ng of DNA for input into enrichments. The MyBaits protocol (v3) was used for hybridization enrichment with minor alterations, whereby the input RNA bait concentration was reduced to 25% of the recommended amount and custom oligonucleotides were used to block the barcoded P5/P7 adapters. Hybridization time was 40 hours with the following conditions: 65°C start temperature followed by a lowering of 1°C per hour to 55°C; and 30 hours at 55°C. Post-capture, the beads were washed three times using Wash Buffer 2 (MyBaits) and resuspended in PCR mastermix for on-bead PCR. To further reduce clonality, the number of PCR cycles needed for each library to reach plateau was estimated for each sample using qPCR with primers targeting the barcoded adapters. Each library was then amplified using the determined number of cycles, cleaned, pooled, and sequenced on the Illumina HiSeq X Ten platform (2 x 150 bp).

Data processing and microbiome composition analyses

The resulting data were converted into fastq format using Illumina's bcl2fastq software, before being trimmed and demultiplexed using AdapterRemoval 2 based on unique P5/P7 barcode combinations, (minimum length 25 bp, 1 barcode mismatch, trim Ns) [49]. Taxonomic composition was determined using MEGAN Alignment Tool (MALT) [26], whereby DNA reads from samples were aligned against a database containing 47,696 archaeal and bacterial genome assemblies from the NCBI Assembly database [27]. The resulting BLAST-text files were converted into RMA files via the blast2rma script included in the program MEGAN [50], with the following Last Common Ancestor (LCA) parameters: Weighted-LCA=80%, minimum bitscore=42, minimum E-value=0.01, minimum support percent=0.1. Samples were assessed for ancient DNA authenticity by comparison to extraction blank controls and by estimation of cytosine deamination using mapDamage [51] (Supplementary Note 1). Samples that passing authentication criteria were normalised to equal read depth (187,293) in MEGAN before PCoA ordination of Bray-Curtis dissimilarities. Genus level assignments were also exported from MEGAN into QIIME 1.9.1 [52] for taxonomic and statistical analyses (Supplementary Note 2). PCoA plots were constructed using PhyloToAST [53].

Genomic and phylogenetic analysis

We used the previously published El Sidron 1 Neanderthal dental calculus data as an outgroup for our phylogenetic analyses [17]. Genomic sequences were assembled by mapping to

reference genomes using the PALEOMIX [54] pipeline with the BWA-MEM aligner [55]. The resulting BAM files were imported into Geneious (v. 10.2.3) [56], and consensus sequences were called using a 75% consensus threshold (Supplementary Note 3; Figure S25). Consensus sequences were then aligned using the Mauve Genome aligner with default settings [57]. Gblocks [58] with default settings was used to clean alignments before aligned sequences were imported into MEGA7 [59]. In MEGA7, phylogenetic reconstruction was performed using the Neighbour-Joining method of evolutionary distances computed using the Maximum Composite Likelihood method [60]. Trees shown are bootstrap consensus trees from 1,000 replicates. The rate variation among sites was modelled with a gamma distribution (shape parameter = 1), and the differences in the composition bias among sequences were considered in evolutionary comparisons [61]. All positions with less than 75% site coverage were eliminated (*e.g.* fewer than 25% alignment gaps, missing data, and ambiguous bases were allowed at any position). Details about the analysis of the hybridization-enriched genes can be found in Supplementary Note 4.

Supplementary note 1: Subtractive filtering of laboratory contaminants and authentication of ancient DNA

Subtractive filtering of laboratory contaminants

Samples with low concentrations of DNA are more vulnerable to the effects of DNA contamination from the laboratory environment [62–64]. While ancient dental calculus has been shown to contain relatively high concentrations of DNA (tens to hundreds of ng mg^{-1})[65], there is variation between samples that are likely due to differences in preservation that results from exposure to different microenvironments (*e.g.* heat, moisture). Therefore, it is expected that some ancient dental calculus samples will be more affected by laboratory contamination than others. Figure S1 clearly illustrates this concept, with some ancient dental calculus samples clustering closer to the laboratory extraction blank controls than others.

To remove taxa derived from laboratory contamination from ancient dental calculus samples, we employed subtractive filtering, whereby genera or species identified in the extraction blank controls were removed from our ancient dental calculus samples. While conservative, this approach is currently the most effective way of reducing the effects of laboratory contamination on microbiome analyses. If the ancient dental calculus samples had fewer than 50,000 reads assigned at the genus-level post filtering and had fewer than 20 genus-

level taxonomic assignments, they were removed from the analysis. These criteria resulted in 13 of the 130 samples being removed from further analyses (blue circles in Figure S1).

Testing for authentic ancient DNA with mapDamage

Authentication of ancient DNA is an essential criterion of paleomicrobiology research [66,67]. While sample decontamination and subtractive filtering can mitigate DNA contamination introduced during sampling and laboratory procedures, respectively, another necessary form of authentication is the assessment of ancient DNA damage patterns. Ancient DNA is typically short due to post-mortem DNA fragmentation and usually presents as a log-normal distribution (Figure S2A). In addition, cytosine deamination of ancient DNA results in an observed increase in cytosine to thymine and guanine to adenine substitutions at the 5' and 3' ends of ancient molecules, respectively (Figure S2B) [68]. Modern DNA does not show this pattern of increased cytosine deamination at the termini and is typically longer than ancient molecules.

To assess the authenticity of our ancient DNA, we used PALEOMIX [54] to map the DNA reads from our ancient samples to the reference genomes of the highest abundance species from the core genera we identified (Figure 2; Table S2) and to estimate the level of cytosine deamination using mapDamage [51]. These figures illustrate damage patterns typical of authentic ancient DNA for each of the seven species tested, with increasing levels of substitutions for nucleotides closer to the ends of the sequenced molecules (Figures S3-16). The percentage of cytosine deamination at the terminal ends ranged from 0.43—46.43% (mean 8.03%) with an average of 6.5% for the historic and 23.8% for the ~3,000-year-old Palau and Teouma samples. This is to be expected, as age has been previously identified as a factor that correlates with increased cytosine deamination rate [69,70]. As a control we used mapDamage on the 210 modern plaque samples, using *Actinomyces sp oral taxon 414* as the reference (as of the top seven genera per dataset, this was the only genus shared). As expected, we did not observe evidence of ancient DNA cytosine deamination for the modern samples. The average rate of observed cytosine deamination at the terminal positions was of 1.58% (Figures S17 & S18), and this did not reach above background rate (*e.g.* Figure S19). We did observe an elevated rate of T-to-C substitutions, which could be due to the library preparation technique or sequence trimming used in the HMP study (Figure S19).

Overall, the subtractive filtering of taxa found in EBCs from our samples together with the mapDamage analyses support the authenticity of the ancient DNA in our samples.

Supplementary note 2: Microbiome composition analysis of ISEA-Pacific dental calculus

The analysis of microbial compositions is a complex problem – especially for ancient communities, where detailed information about host health and lifestyle are limited, taphonomic processes can alter initial community structure [71], and the problem of missing reference genomes is more pronounced [27]. We chose to focus on genus classifications instead of species for microbial composition analyses as database bias (due to missing reference genomes) can result in most reads being assigned at the genus level [72], which would not be used in a species level analysis. This is likely to disproportionately affect the ancient/historical samples analysed here as most modern oral microbial reference genomes have been obtained from 'Western' individuals.

Controlling for tooth type and tooth surface

Tooth type and surface have been shown to have a small but significant effect on plaque microbial composition in modern [30,73] and ancient studies [73]. Failure to account for such difference can hinder the ability to detect differences between samples based on other metadata. Given that our samples contain a mixture of tooth types (molar=74, incisor=7, canine=11, Premolar=22, Unknown=2), and tooth surfaces (buccal=56, lingual=43, interproximal=2, distal=4, unknown=13), we tested to see if these factors were driving variation in our dataset.

We exported genus level assignments from MEGAN into QIIME 1.9.1 [52], and rarefied (normalised) the resulting OTU table to a depth of 66,538 genus level assignments per sample — corresponding to the depth of the sample with the lowest number of assignments. Alpha diversity (within sample) and beta diversity (between sample differences) were calculated on the rarefied table using the core_diversity_analyses.py script. To measure alpha diversity, we calculated both observed_species (number of genus assignments per sample—richness), and Shannon's diversity index, which takes into account both richness and abundance. For Beta diversity, we calculated both Bray-Curtis (abundance-weighted), and Binary Jaccard (non-abundance-weighted) distance metrics. The resulting distance matrices were converted into principal components and were visualised using PhyloToAST [53].

Alpha diversity between tooth types were not significantly different (non-parametric t-test, 999 Monte Carlo permutations; *p*-values >0.05) (Figure S20). We observed no clustering of samples by PCoA of both distance metrics (Figure S21 & S22). To formally test if tooth type and tooth surface influence the microbial composition between our samples, we used the nonparametric statistical methods PERMANOVA and ANOSIM in QIIME 1.9.1 with 1000

permutations. To reduce stochasticity and improve statistical power, metadata groups with <5 samples were removed from the analyses. We found no statistical significance for difference in microbial composition between tooth type and tooth surface (Table S2; *p*-values >0.05).

While we observed no statistically significant effect of tooth type or tooth surface on our dataset, the studies which found such differences had greater power to detect differences both in terms of sample size and measuring individuals from a similar culture/geographic region [30,31,73]. Given that our dataset contains small sample sizes of individuals from different cultures and geographic regions, tooth type and tooth surface still likely contributes to some variation between samples that we could not detect. Therefore, further microbial composition analyses were performed on a subset of the data containing just molars (n=74). We chose not to further subdivide the dataset based on tooth surface as this would have further depleted our statistical power, and it has been demonstrated that tooth type drives more variation than tooth surface [30,31,73].

Testing for cultural factors that influence microbiome composition

We hypothesized that each island/region would possess a unique microbial composition brought about due to differences in lifestyle practices, diet, and geographic isolation. We first tested to see if there were any differences in microbial composition between the different islands/regions. Statistical tests were only performed for islands/regions that had five or more samples (Andaman=10, China=7, Indonesia=7, Marquesas=14, Papua New Guinea=9, Vanuatu=13). We did not find any differences in alpha diversity between island/region (observed_species and Shannon *p*-values >0.05), and did not observe clustering by PCoA of beta diversity metrics in the first three axes (Figure S23). However, we found small but statistically significant differences in community composition by island/region for both abundance-weighted and unweighted distance metrics (PERMANOVA, ANOSIM *p*-values <0.005) (Table S3). To test if differential abundance of specific genera were driving this difference we used the non-parametric Kruskal-Wallis test with Bonferroni multiple test correction. We found three genera (*Desulfomicrobium*, *Asaccharospora, Pseudopropionibacterium*) that were differentially abundant between groups (Table S4; Bonferroni-corrected *p*-values <0.05).

Desulfomicrobium was found to be significantly lower in abundance in the Vanuatu samples (Table S4), and is a genus of anaerobic, sulphate-reducing bacteria that have been isolated from periodontal pockets and are putatively involved in periodontal disease [32]. *Desulfomicrobium* was not found in the modern HMP (Human Microbiome Project) plaque samples.

Asaccharospora was only identified in samples from China (Table S4), and is a genus of spore-forming bacteria that are unable to ferment sugars which was isolated from the gastro-intestinal track of a rat [74]. The low mean abundance of this genus (48 reads) combined with its unlikely lifestyle for the oral cavity and origin suggest that it is noise or contamination.

Pseudopropionibacterium was found to be in higher abundance in samples from Indonesia (mean 6,580) and lowest in samples from Vanuatu (mean 794) (Table S4). This genus is a commonly observed member of the plaque microbiota in modern samples from the HMP (mean abundance: 1.36%, mean prevalence 87.6%). This genus is typically anaerobic and produces propionic acid from glucose.

Overall, we found a small, but statistically significant difference in microbial community composition between islands/regions. However, only two putative oral genera were found to be differentially abundant between groups: *Pseudopropionibacterium,* which is a commonly found member of plaque microbiota, and *Desulfomicrobium,* which is associated with periodontal disease and was not found in HMP plaque samples. These findings suggest that there are only minor differences in microbial composition between islands/regions. However, given that recent studies have found that inter-personal variation in oral microbiota are large [29,72], future studies with larger sample sizes will be needed to verify these findings.

Testing for influence of oral disease on microbial community composition

Oral diseases, such as dental caries and periodontal disease can influence the microbial composition of plaque samples [75–77]. Therefore, we tested the impacts of classified oral disease state on microbial community structure. Evidence of dental caries (Yes=7, No=67) or periodontal disease (Yes=24 No=50) were not statistically significant factors for microbial composition (ANOSIM/PERMANOVA p-values >0.05; Table S5). This could be explained by recent evidence supporting both caries and periodontal disease are polymicrobial, *i.e.* not caused by a handful of pathogens or a specific community type [78,79]. Therefore, different microbial assemblages across the different cultures we measured may result in different microbial communities for periodontal disease in each culture—something we are unable to test due to our sampling effort or reference database bias [27]. Another possibility is that our classifications of periodontal disease was wrong, as there are difficulties in making paleopathological assessments of periodontal disease for ancient skulls [80,81]. It is also possible that some of the ancient individuals measured could have had conditions supporting periodontal disease-associated taxa (*e.g.* gingivitis) [82,83] without reabsorption of the jaw, leaving no trace of such disease for classification. Additionally, some of our samples were taken from isolated teeth, and could not have their periodontal status classified.

Several species present in ancient microbiome studies [84] are linked with periodontal disease in modern populations, such as *Methanobrevibacter oralis* [34,85]. We therefore classified each sample based on the percentage of reads assigned to *Methanobrevibacter* (0%=20, 1-5%=35, 5-10%=7, 10-20+%=12), and used these as labels for PCoA plots. We found that the abundance of *Methanobrevibacter* appeared to correlate with the axis of greatest variation (Figure S24A). This clustering was not observed using the unweighted Binary Jaccard metric (Figure S24B), suggesting that the abundance of a periodontal-associated *Methanobrevibacter* alone was enough to influence abundance-weighted distance metrics in ancient microbiome analysis. Therefore, periodontal disease could be a confounding factor when analysing ancient microbial communities and should be evaluated in future studies using larger datasets with more robust sampling and classification methods.

Microbial functional analyses

A recent study has demonstrated that nucleotide-to-protein alignments are currently unable to assign short (<60 bp) DNA sequences. Given that current analytical tools for assigning functional information rely on nucleotide-to-protein alignments, and the mean fragment lengths of our samples vary (42-110 bp), such analysis would be severely biased towards samples with longer fragment length distributions and consequently confound results. We therefore decided not to explore the functions associated within these samples. Development of software to account for this issue will allow future studies to explore the functions associated within these samples.

Paucity of reference genomes may hinder ability to classify and analyse microbial community composition

Finally, alignment-based taxonomic classification of ancient microbial communities suffers from missing reference genomes in databases [27], so we tested to see what proportion of reads in our samples could be aligned to reference genomes. An average of 51.7% (± 7.1%) of reads could be aligned for our dataset, suggesting that we are potentially a missing large proportion of microbial community. Future reanalysis of our data with better bioinformatic techniques and the addition of more microbial reference genomes may allow for greater discrimination of how culture and lifestyle may influence the microbial community composition of these samples.

Supplementary note 3: Mapping and whole-genome phylogenetic analyses

Reference-based mapping

To test if phylogenetic reconstruction of microbial genomes from dental calculus could be used to infer the past movements of humans, we reconstructed microbial genomes using a reference-based mapping approach. All of the genera, except *Actinomyces*, had most of their reads assigned to a single species within that genera (Table S6). *Actinomyces* is known to contain many oral species that can co-occur within the same individual [86]. In our data, we identified 28 species of *Actinomyces*, of which the 7 most abundant accounted for 97.4% all reads assigned to the genus (Table S7). Phylogenetic reconstruction can be confounded by the mapping of reads between closely related species [72], as we use single nucleotide polymorphisms (SNPs) in these reads to reconstruct the phylogenetic history of these species. If DNA sequences map to multiple genomes due to insufficient stringency of mapping algorithms or high sequence similarity (*e.g.* closely relates species), the phylogenetically informative history of a species is confounded by the signatures from mismapped reads. To account for such cross-mapping, we used a technique called competitive mapping, whereby reference genomes from closely related species are placed in the same fasta file, and the DNA reads are mapped against this concatenated file [42]. Using this approach, reads that map to multiple *Actinomyces* species (and are likely shared) will have a lower mapping quality score, and the sequences can be discarded from the final alignment. In the case of *Actinomyces*, we merged the reference genomes of the top 7 most abundant *Actinomyces* species into a single file and used PALEOMIX as described in the methods. As only *Actinomyces oralis* and *Actinomyces sp. oral taxon 414* have complete genomes (the others being scaffold-level assemblies), we merged the scaffolds for a given species in Geneious (adding 250 base pairs of N's between each scaffold) to create pseudo-complete genomes and makes downstream genome alignment feasible. Post competitive mapping, only *Actinomyces. sp. oral taxon 414, Actinomyces dentalis, and Actinomyces israelii* (the three most abundant species) had sufficient data for consensus calling and phylogenetic reconstruction.

Consensus calling

Once all of our reads were mapped to the their specific reference genomes, we called consensus sequences from these alignments. Our data has relatively low sequencing effort — average of 2.2 million reads (± 1.3 million) per sample. Therefore, applying a strict consensus call whereby

a minimum read depth of 3 is required for downstream use would result too much missing data (N's) (Figure S25B), making whole-genome alignment extremely challenging—as these alignment algorithms were designed for high quality modern data. We instead performed consensus calling without a minimum depth requirement, but with a 75% consensus requirement (*i.e.* for genomic regions with multiple reads aligning, a SNP is only called if at least 75% of these reads agree with each other — see the illustration in Figure S25). The 75% consensus calling requirement accounts for strain-level variation, for which there is strong evidence against [87,88], these studies found that typically a single bacterial strain dominates through time for a given individual. This consensus requirement also controls for the influence of cytosine deamination as this preferentially occurs at the ends of molecules, and therefore reduces the probability of deamination influencing the alignments and phylogenetic inference (Figure S25A). The consensus requirement only works for regions where adequate read depth is obtained and given that some regions of our genomes will only be covered by a single read, deamination could influence phylogenetic reconstruction (Figure S25A). Therefore, we chose to take a conservative approach by only using transversions for phylogenetic reconstruction (see below), which, while lowering phylogenetic resolution, is the more conservative approach. Future benchmarking will be useful to determine the influence of cytosine deamination on phylogenetic reconstruction. Finally, because missing data is a major issue for genome alignments, our BAM files (alignments) were only used if they had a mean coverage >1 (for unique reads). Additionally, BAM files were inspected visually to determine the evenness of coverage, this is important as a mean coverage of 1 could be observed if only a small region of the genome is covered with high depth.

Whole-genome alignment and trimming

Once our consensus sequences were obtained, we performed whole-genome alignments using the Mauve algorithm [57] with the default settings. Inspection and editing of ambiguous or poor alignments by eye is typically done for short alignments (<50,000 nucleotides); however, when using whole bacterial genomes (averaging 3,000,000 nucleotides), visual inspection is not feasible. Therefore, we used Gblocks [58] to eliminate poorly aligned positions in an automated and reproducible fashion. Phylogenetic inference was performed on these 'cleaned' alignments using transversion substitutions only to account for the influence of cytosine deamination on phylogenetic reconstruction. The resulting phylogenies—apart from *Anaerolineaceae sp. oral taxon 439*— (Figures S26-34) do not recapitulate the established relationships and past movement patterns between geographical regions, and the potential reasons for this are discussed in the main text.

Supplementary note 4: Hybridization enrichment of microbial phylogenetic markers: design, use, and analyses

Overview of approach and incentive

Ancient dental calculus contains many microbial species that results in a lot of genomic information, and therefore, obtaining sufficient sequencing coverage and depth for phylogenetic reconstruction is expensive. To address this issue, we designed RNA probes that are complementary to putatively phylogenetically informative genes from 11 different, highly prevalent oral bacterial species (Table S8). These probes allow for selective enrichment of these genes, therefore decreasing DNA sequencing effort required and also allowing for the analysis of genomic information from highly prevalent species that have low abundance. While these species may not be highly abundant, their widespread prevalence across individuals suggests that they may be evolutionarily conserved. This would ultimately expand the spectrum of microbial species surveyed for their ability to inform us about past human movements.

RNA probe design

We chose three different gene types for design of our RNA probes: (1) clade-specific markers (CSMs); (2) single-copy core genes (SCCGs); and (3) Multi-Locus Sequencing Typing (MLST) loci.

(1) Clade specific markers:

Clade-specific markers are coding sequences (CDSs) that are unique to a given clade or species. This factor makes them a promising target for enrichment and phylogenetic analysis, as there is be expected to be minimal cross-mapping between species. Clade-specific markers have been previously identified by comparing thousands of microbial genomes to each other, and these markers are used in the popular taxonomic-profiling program MetaPhlAn [89,90], and for strain-level analysis in the program StrainPhlAn [87]. Given the limited space on our RNA probe set (40,000 unique probes), we randomly selected over 50 clade-specific markers from four bacterial species that were found to be highly prevalent in our ancient dental calculus samples (*Fretibacterium fastidiosum, Pseudopropionibacterium propionicum, Pseudoramibacter alactolyticus,* and *Tannerella forsythia*).

(2) Single-copy core genes:

Single-copy core genes are genes that have critical housekeeping functions (transcriptions, translation, etc.) and are therefore present in most microbial species. While these genes are highly conserved—containing motifs that are under purifying selection—they are sufficiently different enough from each other between species to be distinguishable. These genes are typically parts of complex interactive networks, and as a result, are thought to be recalcitrant to horizontal transfer [91,92], which can confound phylogenetic reconstruction.

Phylogenetic reconstruction using such genes has been previously demonstrated [93–95], highlighting their viability for such an approach. We chose eight single-copy core genes: GyrA (DNA gyrase A), GryB (DNA Gyrase B), DnaG (DNA Grimase), SecA (Secretase A), RpoB (DNA-directed RNA Polymerase-Beta), DnaN (DNA Polymerase III subunit-beta), RplB (Ribosomal protein L2), and RecA (Recombination protein A) from 10 highly prevalent bacterial species identified in our dataset.

(3) MLST loci:

Multi-Locus Sequencing Typing (MLST) is a technique that was originally developed for typing closely related pathogenic bacteria for epidemiological strain tracking [96]. The MLST approach uses SNPs or alleles found in multiple hypervariable regions of 5-7 housekeeping genes. This technique has been previously used on *Helicobacter pylori* (a bacterium that lives in the stomach) for inferring past human population movements around the world [97,98]. Given that MLST schemas are typically only developed for pathogenic bacteria, the only bacterium widely prevalent in our dataset that had a MLST schema available was the periodontal diseases-associated *Porphyromonas gingivalis*. Using these three classes of genes, we created 325 unique RNA probes targeting genes from 10 different species of bacteria (Table S8). These RNA probes were used to enrich 32 ancient dental calculus samples, as described in the methods.

Bioinformatic analysis of enriched data

To map the reads from our enriched libraries to the genes targeted in our RNA probes, we used PALEOMIX as described in the methods.

We accounted for cross-mapping of reads from different species by using competitive mapping, whereby reads were mapped against a fasta file containing all genes used in the probes (Supplementary Note 3). For equal sequencing effort of enriched vs unenriched, we obtained an average 27.8-fold (± 26.8) enrichment of genes targeted in our probe set (Table S9). The average number of raw hits (including PCR duplicates) for unenriched libraries was 4,146 (S.D. 2,807), and 475,172 (± 245,718) for enriched libraries. The average number of unique hits (with

PCR duplicated removed) was 3,814 (± 2,842) for unenriched libraries, and 128,368 (170,674) for enriched libraries. Therefore, while the enrichment was successful, there was a substantial increase in clonality (PCR duplicates) for enriched libraries (average 71% S.D. 30.2%) versus unenriched libraries (12.3% ± 17.3%). Overall, the enrichments were successful in selectively capturing the target genes.

BAM file quality filtering

To increase the quality of phylogenetic inference, we removed BAM files that had less than 90% of the gene covered in reads as these likely represent spurious or erroneous mapping, and are typically removed from such analyses [87]. Because we had sufficient sequencing depth due to enrichment, we also removed BAM files that had a median depth <5 (unique reads) to improve the quality of our consensus sequence calls and account for cytosine deamination (Supplementary Note 3). Table S10 lists the total number of genes enriched (325), and the number of genes that passed quality filtering criteria. Overall, an average of 127 (± 75) genes per sample passed these filtering criteria.

Consensus calling, concatenation, and Phylogenetic inference

For calling SNPs and sequence consensuses, we required a minimum depth of 5 and 75% consensus identity (Supplementary Note 3 and Figure S25). This approach differed to our genome analysis because enrichment allowed for greater depth of sequencing. Consensus sequences of each locus for a given species was concatenated in Geneious, adding a (250 bp stretch of N's between loci). For example, in sample A, *RpoB, DnaN, etc.* from species 1 were concatenated into a single sequence (Figure S35A). A species was deemed not present in a sample and removed from phylogenetic analysis if it did not have all of its enriched loci passing the filtering steps mentioned above (90% of the gene covered, minimum median depth 5). Concatenated consensus sequences from each species were aligned using MUSCLE [99] with default settings (Figure S35B), and phylogenetic reconstruction was done using RAxML [100].

Results

The resulting phylogenies (Figure S36-48) do not recapitulate the established relationships and past movement patterns between geographical regions. This could be due to various reasons, such as insufficient resolution or cross-mapping between species. The total number of nucleotide sites used for this method is substantially less than the whole-genome approach (thousands of sites compared to millions), which could lower the phylogenetic resolution. Cross-mapping of DNA reads from closely related microorganisms could have also confounded

these phylogenetic reconstructions. While we employed competitive mapping for the loci enriched in this study, we could not account for other genetic material currently not in reference databases, which could have interfered with phylogenetic reconstruction. Overall, it was unexpected that these putatively phylogenetically informative genes—which are thought to be recalcitrant to horizontal gene transfer—did not yield phylogenies corroborating the available evidence for the peopling of the Pacific.

References

1. Bedford S, Spriggs M, Regenvanu R. The Teouma Lapita site and the early human settlement of the Pacific Islands. Antiquity. 2006;80:812–828.

2. Sheppard PJ. Lapita Colonization across the Near/Remote Oceania Boundary. Curr Anthropol. 2011;52:799–840.

3. Skoglund P, Posth C, Sirak K, Spriggs M, Valentin F, Bedford S, et al. Genomic insights into the peopling of the Southwest Pacific. Nature. 2016;538:510–3.

4. Rieth TM, Hunt TL. A radiocarbon chronology for Sāmoan prehistory. J Archaeol Sci. 2008;35:1901–27.

5. Wilmshurst JM, Hunt TL, Lipo CP, Anderson AJ. High-precision radiocarbon dating shows recent and rapid initial human colonization of East Polynesia. Proc Natl Acad Sci. 2011;108:1815–20.

6. Athens JS, Rieth T, Dye T. A paleoenvironmental and archaeological model-based age estimate for the colonization of Hawai'i. Am Antiq. 2014;79:144–155.

7. Kirch PV, Green RC, Bellwood PS, Dunnell RC, Dye T, Gosden C, et al. History, Phylogeny, and Evolution in Polynesia [and Comments and Reply]. Curr Anthropol. 1987;28:431–56.

8. Kirch PV, Green RC. Hawaiki, Ancestral Polynesia: An Essay in Historical Anthropology. Cambridge University Press; 2001.

9. Wilson WH. Evidence for an Outlier Source for the Proto Eastern Polynesian Pronominal System. Ocean Linguist. 1985;24:85–133.

10. Wilson WH. Whence the East Polynesians?: Further Linguistic Evidence for a Northern Outlier Source. Ocean Linguist. 2012;51:289–359.

11. Hudjashov G, Endicott P, Post H, Nagle N, Ho SYW, Lawson DJ, et al. Investigating the origins of eastern Polynesians using genome-wide data from the Leeward Society Isles. Sci Rep. 2018;8:1823.

12. Eisenhofer R, Anderson A, Dobney K, Cooper A, Weyrich LS. Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events. J Isl Coast Archaeol. 2017;0:1–14.
13. Thomson VA, Lebrasseur O, Austin JJ, Hunt TL, Burney DA, Denham T, et al. Using ancient DNA to study the origins and dispersal of ancestral Polynesian chickens across the Pacific. Proc Natl Acad Sci. 2014;111:4826–4831.
14. Matisoo-Smith E, Robins JH. Origins and dispersals of Pacific peoples: Evidence from mtDNA phylogenies of the Pacific rat. Proc Natl Acad Sci U S A. 2004;101:9167–72.
15. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of Human Migrations in Helicobacter pylori Populations. Science. 2003;299:1582–5.
16. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336–44.
17. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature. 2017;544:357–61.
18. Li Y, Ismail AI, Ge Y, Tellez M, Sohn W. Similarity of Bacterial Populations in Saliva from African-American Mother-Child Dyads. J Clin Microbiol. 2007;45:3082–5.
19. Corby PM, Bretz WA, Hart TC, Schork NJ, Wessel J, Lyons-Weiler J, et al. Heritability of Oral Microbial Species in Caries-Active and Caries-Free Twins. Twin Res Hum Genet. 2007;10:821–828.
20. Stahringer SS, Clemente JC, Corley RP, Hewitt J, Knights D, Walters WA, et al. Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. Genome Res. 2012;22:2146–52.
21. Ebersole JL, Holt SC, Delaney JE. Acquisition of Oral Microbes and Associated Systemic Responses of Newborn Nonhuman Primates. Clin Vaccine Immunol CVI. 2014;21:21–8.
22. Shaw L, Ribeiro ALR, Levine AP, Pontikos N, Balloux F, Segal AW, et al. The human oral microbiome is shaped by shared environment rather than genetics: evidence from a large family of closely-related individuals. bioRxiv. 2017;131086.
23. Petchey F, Spriggs M, Bedford S, Valentin F, Buckley H. Radiocarbon dating of burials from the Teouma Lapita cemetery, Efate, Vanuatu. J Archaeol Sci. 2014;50:227–42.
24. Nelson GC, Fitzpatrick SM. Preliminary investigations of the Chelechol ra Orrak Cemetery, Republic of Palau: I, skeletal biology and paleopathology. Anthropol Sci. 2006;114:1–12.
25. Consortium THMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

26. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv. 2016;050559.
27. Eisenhofer R, Weyrich LS. Assessing alignment-based taxonomic classification of ancient microbial DNA. (Chapter III) in preparation. 2018;
28. Welch JLM, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci. 2016;113:E791–800.
29. Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, et al. Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. Npj Biofilms Microbiomes. 2017;3:2.
30. Proctor DM, Fukuyama JA, Loomer PM, Armitage GC, Lee SA, Davis NM, et al. A spatial gradient of bacterial diversity in the human oral cavity shaped by salivary flow. Nat Commun. 2018;9:681.
31. Farrer AG, Bekvalac J, Redfern R, Gully N, Dobney K, Cooper A, et al. Biological and cultural drivers of microbiota in Medieval and Post-Medieval London, UK. PhD Thesis: Ancient DNA studies of dental calculus. 2017;
32. Langendijk PS, Kulik EM, Sandmeier H, Meyer J, van der Hoeven JS. Isolation of Desulfomicrobium orale sp. nov. and Desulfovibrio strain NY682, oral sulfate-reducing bacteria involved in human periodontal disease. Int J Syst Evol Microbiol. 2001;51:1035–44.
33. Abusleme L, Dupuy AK, Dutzan N, Silva N, Burleson JA, Strausbaugh LD, et al. The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. ISME J. 2013;7:1016–25.
34. Nguyen- Hieu Tung, Khelaifia Saber, Aboudharam Gerard, Drancourt Michel. Methanogenic archaea in subgingival sites: a review. APMIS. 2012;121:467–77.
35. Siqueira JF, Rôças IN. Pseudoramibacter alactolyticus in Primary Endodontic Infections. J Endod. 2003;29:735–8.
36. Booth V., Downes J., Van den Berg J., Wade W. G. Gram- positive anaerobic bacilli in human periodontal disease. J Periodontal Res. 2004;39:213–20.
37. Lipson M, Skoglund P, Spriggs M, Valentin F, Bedford S, Shing R, et al. Population Turnover in Remote Oceania Shortly after Initial Settlement. Curr Biol [Internet]. 2018 [cited 2018 Mar 1];0. Available from: http://www.cell.com/current-biology/abstract/S0960-9822(18)30236-7
38. Posth C, Nägele K, Colleran H, Valentin F, Bedford S, Kami KW, et al. Language continuity despite population replacement in Remote Oceania. Nat Ecol Evol. 2018;1.

39. Fitzpatrick SM. Early human burials in the western Pacific: evidence for c.3000 year old occupation on Palau. Antiquity. 2003;77:719–31.
40. Stone JH, Fitzpatrick SM, Napolitano MF. Disproving claims for small-bodied humans in the Palauan archipelago. Antiquity. 2017;91:1546–60.
41. Fitzpatrick SM. The Archaeology of Western Micronesia. Oxf Handb Prehist Ocean [Internet]. 2018 [cited 2018 Apr 20]; Available from: http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199925070.001.0001/oxfordhb-9780199925070-e-012
42. Key FM, Posth C, Krause J, Herbig A, Bos KI. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. Trends Genet [Internet]. 2017 [cited 2017 Jul 9];0. Available from: http://www.cell.com/trends/genetics/abstract/S0168-9525(17)30086-0
43. Didelot X, Maiden MCJ. Impact of recombination on bacterial evolution. Trends Microbiol. 2010;18:315–22.
44. Posada D, Crandall KA. The Effect of Recombination on the Accuracy of Phylogeny Estimation. J Mol Evol. 2002;54:396–402.
45. Rannala B, Yang Z. Phylogenetic Inference Using Whole Genomes. Annu Rev Genomics Hum Genet. 2008;9:217–31.
46. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, Jane Adler C, et al. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nat Commun. 2013;4:1764.
47. Gamba C, Hanghøj K, Gaunitz C, Alfarhan AH, Alquraishi SA, Al-Rasheid KAS, et al. Comparing the performance of three ancient DNA extraction methods for high-throughput sequencing. Mol Ecol Resour. 2016;16:459–69.
48. Meyer M, Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc. 2010;2010:pdb.prot5448.
49. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes [Internet]. 2016 [cited 2018 Feb 26];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751634/
50. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Comput Biol. 2016;12:e1004957.
51. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinforma Oxf Engl. 2013;29:1682–4.

52. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.
53. Dabdoub SM, Fellows ML, Paropkari AD, Mason MR, Huja SS, Tsigarida AA, et al. PhyloToAST: Bioinformatics tools for species-level analysis and visualization of complex microbial datasets. Sci Rep. 2016;6:29123.
54. Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat Protoc. 2014;9:1056.
55. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.
56. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinforma Oxf Engl. 2012;28:1647–9.
57. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 2004;14:1394–403.
58. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.
59. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33:1870–4.
60. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci. 2004;101:11030–5.
61. Tamura K, Kumar S. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol Biol Evol. 2002;19:1727–36.
62. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.
63. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. Gut Pathog. 2016;8:24.
64. Lauder AP, Roche AM, Sherrill-Mix S, Bailey A, Laughlin AL, Bittinger K, et al. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. Microbiome. 2016;4:29.

65. Warinner C, Speller C, Collins MJ. A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. Philos Trans R Soc B Biol Sci. 2014;370:20130376–20130376.

66. Eisenhofer R, Cooper A, Weyrich LS. Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential. FEMS Microbiol Ecol [Internet]. 2017 [cited 2017 Jun 27];93. Available from: https://academic.oup.com/femsec/article/93/5/fix042/3089752/Reply-to-Santiago-Rodriguez-et-al-proper

67. Eisenhofer R, Weyrich LS. Proper Authentication of Ancient DNA Is Still Essential. Genes. 2018;9:122.

68. Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. Cold Spring Harb Perspect Biol. 2013;a012567.

69. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLOS ONE. 2012;7:e34131.

70. Weiß CL, Schuenemann VJ, Devos J, Shirsekar G, Reiter E, Gould BA, et al. Temporal patterns of damage and decay kinetics of DNA retrieved from plant herbarium specimens. R Soc Open Sci. 2016;3:160239.

71. Hyde ER, Haarmann DP, Lynne AM, Bucheli SR, Petrosino JF. The Living Dead: Bacterial Community Structure of a Cadaver at the Onset and End of the Bloat Stage of Decomposition. PLoS ONE [Internet]. 2013 [cited 2018 Mar 12];8. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3813760/

72. Warinner C, Herbig A, Mann A, Yates JAF, Weiß CL, Burbano HA, et al. A Robust Framework for Microbial Archaeology. Annu Rev Genomics Hum Genet. 2017;18:null.

73. Simón-Soro Á, Tomás I, Cabrera-Rubio R, Catalan MD, Nyvad B, Mira A. Microbial Geography of the Oral Cavity. J Dent Res. 2013;92:616–21.

74. Gerritsen J, Fuentes S, Grievink W, van Niftrik L, Tindall BJ, Timmerman HM, et al. Characterization of Romboutsia ilealis gen. nov., sp. nov., isolated from the gastro-intestinal tract of a rat, and proposal for the reclassification of five closely related members of the genus Clostridium into the genera Romboutsia gen. nov., Intestinibacter gen. nov., Terrisporobacter gen. nov. and Asaccharospora gen. nov. Int J Syst Evol Microbiol. 2014;64:1600–16.

75. Richards VP, Alvarez AJ, Luce AR, Bedenbaugh M, Mitchell ML, Burne RA, et al. Microbiomes of Site-Specific Dental Plaques from Children with Different Caries Status. Infect Immun. 2017;85.

76. Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, Sommer DD, et al. Deep Sequencing of the Oral Microbiome Reveals Signatures of Periodontal Disease. PLoS ONE. 2012;7:e37919.
77. Boutin S, Hagenfeld D, Zimmermann H, El Sayed N, Höpker T, Greiser HK, et al. Clustering of Subgingival Microbiota Reveals Microbial Disease Ecotypes Associated with Clinical Stages of Periodontitis in a Cross-Sectional Study. Front Microbiol [Internet]. 2017 [cited 2018 Apr 18];8. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2017.00340/full#h13
78. Hajishengallis G, Lamont RJ. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. Mol Oral Microbiol. 2012;27:409–19.
79. Simón-Soro A, Mira A. Solving the etiology of dental caries. Trends Microbiol. 2015;23:76–82.
80. Clarke NG, Hirsch RS. Two critical confounding factors in periodontal epidemiology. Community Dent Health. 1992;9:133–41.
81. Raitapuro-Murray T, Molleson TI, Hughes FJ. The prevalence of periodontal disease in a Romano-British population c. 200-400 AD. Br Dent J. 2014;217:459–66.
82. Chang A, Davis C, Bo C, Yang F, Zhao H, Xu J, et al. Predictive modeling of gingivitis severity and susceptibility via oral microbiota. ISME J. 2014;8:1768.
83. Huang S, Li Z, He T, Bo C, Chang J, Li L, et al. Microbiota-based Signature of Gingivitis Treatments: A Randomized Study. Sci Rep. 2016;6:24705.
84. Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci Rep. 2015;5:16498.
85. Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, Relman DA. Methanogenic Archaea and human periodontal disease. Proc Natl Acad Sci U S A. 2004;101:6176–81.
86. Dige I, Raarup MK, Nyengaard JR, Kilian M, Nyvad B. Actinomyces naeslundii in initial dental biofilm formation. Microbiology. 2009;155:2116–26.
87. Truong DT, Tett A, Pasolli E, Huttenhower C, Segata N. Microbial strain-level population structure and genetic diversity from metagenomes. Genome Res. 2017;27:626–38.
88. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.
89. Segata N, Waldron L, Ballarini A, Narasimhan V, Jousson O, Huttenhower C. Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 2012;9:811–4.

90. Truong DT, Franzosa EA, Tickle TL, Scholz M, Weingart G, Pasolli E, et al. MetaPhlAn2 for enhanced metagenomic taxonomic profiling. Nat Methods. 2015;12:902–3.
91. Jain R, Rivera MC, Lake JA. Horizontal gene transfer among genomes: The complexity hypothesis. Proc Natl Acad Sci. 1999;96:3801–6.
92. Hao W, Golding GB. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. BMC Genomics. 2008;9:235.
93. Wu M, Scott AJ. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics. 2012;28:1033–4.
94. Darling AE, Jospin G, Lowe E, Iv FAM, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243.
95. Ankenbrand MJ, Keller A. bcgTree: automatized phylogenetic tree building from bacterial core genomes. Genome. 2016;1–9.
96. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, et al. Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci. 1998;95:3140–5.
97. Falush D. Helicobacter pylori Traces of Human Migrations in. science. 2003;1080857:299.
98. Moodley Y, Linz B, Yamaoka Y, Windsor HM, Breurec S, Wu J-Y, et al. The Peopling of the Pacific from a Bacterial Perspective. Science. 2009;323:527–30.
99. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
100. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

Supplementary figures and tables

Table S1. Sample details

#SampleID	Methanobrevibacter-abundance	ToothType	ToothSurface	Perio	Caries
15384_China_2	5to1	Molar	Interproximal	N	N
15387_China_3	20+to10	Molar	Lingual	N	N
15388_China_4	0	Incisor	Lingual	Y	N
15389_China	20+to10	Molar	Buccal	N	N
15390_China_5_Hong_Kong	5to1	Molar	Buccal	N	N
15391_China_6	10to5	Molar	Lingual	Y	N
15392_China_7_Tian_Shan	5to1	Molar	Lingual	N	N
15393_China_8	10to5	Molar	Lingual	N	Y
15394_Nicobar_Islands_1	10to5	Molar	Buccal	N	N
15395_Nicobar_Islands_2	10to5	Molar	Lingual	N	N
15396_Andaman_4	0	Molar	Lingual	N	N
15397_Andaman_5	10to5	Premolar	Buccal	N	N
15398_Andaman_6	5to1	Molar	Buccal	Y	Y
15400_Andaman_7	0	Molar	Lingual	N	N
15401_Andaman_8	0	Molar	Buccal	N	N
15402_Andaman_9	20+to10	Molar	Distal	N	Y
15403_Andaman_10	5to1	Molar	Lingual	N	N
15404_Andaman_11	5to1	Premolar	Buccal	Y	Y
15405_Andaman_12	5to1	Premolar	Lingual	Y	N
15406_Andaman_1	5to1	Molar	Buccal	Y	Y
15407_Andaman_13	5to1	Molar	Lingual	Y	N
15408_Andaman_14	5to1	Incisor	Lingual	Y	N
15414_Sumatra_1	5to1	Molar	Buccal	N	N
15415_Sumatra_2	20+to10	Premolar	Lingual	Y	N
15416_Sumatra_3	5to1	Molar	Buccal	N	N
15417_Indonesia_1	20+to10	Premolar	Buccal	N	N
15418_Indonesia_2	0	Premolar	Lingual	N	N
15419_Indonesia_3	20+to10	Molar	Lingual	N	N
15420_Indonesia_4	20+to10	Canine	Buccal	N	Y
15421_Indonesia_5	5to1	Premolar	Lingual	N	N
15422_Indonesia_6	0	Canine	Lingual	Y	N
15423_Indonesia_7	5to1	Molar	Buccal	N	N
15424_Indonesia_8	5to1	Premolar	Buccal	N	N
15425_Indonesia_9	0	Molar	Buccal	N	N
15426_Indonesia_10	0	Molar	Buccal	Y	Y
15427_Indonesia_11	0	Molar	Lingual	N	N

15429_Fiji_1_Vanua_Balavu_Lomaloma	20+to10	Premolar	Interproximal	N	N
15430_Fiji_2_Suva	10to5	Molar	Lingual	Y	N
15432_Marquesas	5to1	Molar	Lingual	N	N
15433_Marquesas_1_Noukahivan	5to1	Molar	Lingual	N	N
15434_Marquesas_2_Worn_Trophy	5to1	Molar	Buccal	N	N
15435_Marquesas_3_Ohivaoan_Tafati	5to1	Molar	Buccal	N	N
15437_Marquesas_5_Ohivaoan_Tafati	20+to10	Molar	Lingual	N	N
15438_Marquesas_6	5to1	Molar	Buccal	N	N
15440_Marquesas_8_Fatuhivan	20+to10	Molar	Buccal	N	N
15441_Marquesas_9_Ohivaoan_Tafati	5to1	Molar	Buccal	N	N
15442_Marquesas_10_Uahugan	5to1	Molar	Lingual	N	N
15443_Marquesas_11_Noukahivan	5to1	Molar	Lingual	N	N
15444_Marquesas_12_Ohivaoan_Tafati	10to5	Premolar	Buccal	N	N
15447_New_Britain_5_Kokopo_NE_part	0	Molar	Buccal	N	N
15448_PNG_1_South_Cape	5to1	Molar	Lingual	Y	N
15450_PNG_4_AdmiraltyIslands_Baluan	5to1	Premolar	Buccal	N	N
15451_New_Britain_9_Kokopo_NE_part	0	Molar	Buccal	N	N
15452_New_Britain_11	5to1	Molar	Lingual	N	N
15453_New_Britain_8_Kokopo_NE_part	0	Molar	Lingual	N	N
15454_New_Britain_10_Kokopo_NE_part	0	Canine	Buccal	N	N
15455_Easter_Island_1	0	Canine	Buccal	Y	Y
15457_Easter_Island	20+to10	Premolar	Buccal	N	N
15458_Easter_Island_3	0	Premolar	Buccal	N	N
15459_Easter_Island_4	20+to10	Molar	Buccal	N	N
15460_Easter_Island_5	0	Molar	Buccal	N	N
15461_Easter_Island_6	0	Molar	Distal	N	N
15462_Vanuatu_1_ErromangoIsland	0	Molar	Buccal	N	N
15463_Vanuatu_13_NewHebrides_Mallicolo_Island	20+to10	Molar	Lingual	N	N
15464_Vanuatu_14	5to1	Molar	Lingual	N	N
15465_Vanuatu_ErromangoIsland	0	Canine	Buccal	N	N
15467_Society	10to5	Canine	Buccal	N	N
15470_Malaysia_2	5to1	Premolar	Buccal	N	N
15471_Tonga	10to5	Incisor	Lingual	N	N
15472_Solomon_Islands_8	5to1	Molar	Buccal	N	N
15477_Phillipines_1	20+to10	Canine	Buccal	N	N
15481_Manilla_1	5to1	Canine	Buccal	N	N
15732_Fiji_4	10to5	Molar	Lingual	Y	Y
15736_Fiji_8	20+to10	Molar	Buccal	Y	N
15737_Fiji_9	20+to10	Molar	Buccal	Y	N
15748_Marquesas_13	0	Molar	Buccal	N	N
15751_Marquesas_18	10to5	Molar	Buccal	N	N
15752_Marquesas_16	5to1	Molar	Buccal	N	Y
15753_Marquesas_17	10to5	Premolar	Buccal	Y	N

15754_Marquesas_19	5to1	Molar	Buccal	N	N
15757_New_Britain_2	5to1	Molar	Distal	Y	N
15758_New_Britain_3	0	Molar	Lingual	Y	N
15759_New_Britain_4	0	Premolar	Lingual	Y	N
15761_New_Britain_6	0	Molar	Lingual	Y	N
15763_New_Britain_7	5to1	Molar	Lingual	Y	N
15764_NewCaledonia	0	Molar	Lingual	N	N
15765_New_Caledonia_1	5to1	Molar	Buccal	Y	N
15766_New_Caledonia_2	5to1	Molar	Buccal	Y	N
15777_Solomon_Islands_1	5to1	Incisor	Buccal	N	N
15778_Solomon_Islands_2	0	Premolar	Lingual	Y	N
15781_Solomon_Islands_3	10to5	Premolar	Lingual	N	N
15782_Solomon_Islands_4	5to1	Canine	Buccal	N	N
15783_Solomon_Islands_5	0	Molar	Lingual	N	N
15784_Solomon_Islands_6	5to1	Premolar	Lingual	N	N
15785_Solomon_Islands_7	10to5	Incisor	Unknown	Y	N
15786_Samoa	5to1	Molar	Lingual	Y	N
15791_Tonga_1	10to5	Premolar	Lingual	Y	Y
15793_Vanuatu_3	20+to10	Molar	Buccal	Y	N
15794_Vanuatu_4	5to1	Molar	Buccal	Y	N
15795_Vanuatu_5	20+to10	Molar	Buccal	Y	N
15796_Vanuatu_6	0	Molar	Lingual	N	N
15797_Vanuatu_7	0	Molar	Buccal	Y	N
15799_Vanuatu_9	5to1	Molar	Buccal	Y	N
15800_Vanuatu_15	0	Molar	Buccal	N	N
15802_Vanuatu_11	5to1	Molar	Buccal	Y	N
15803_Vanuatu_12	5to1	Molar	Distal	Y	N
15815_Palau_1	5to1	Incisor	Unknown	N	N
15816_Palau_2	5to1	Molar	Unknown	N	N
15817_Palau_3	5to1	Molar	Unknown	N	N
15818_Palau_4	5to1	NA	Unknown	N	N
15819_Palau_5	5to1	Premolar	Unknown	N	N
15820_Palau_6	5to1	Canine	Unknown	N	N
15821_Palau_7	5to1	Incisor	Unknown	N	N
15822_Palau_8	5to1	Canine	Unknown	N	N
15856_Teouma_T46_B10B	5to1	Molar	Unknown	N	N
15858_Teouma_T48_B10C	5to1	Premolar	Unknown	N	N

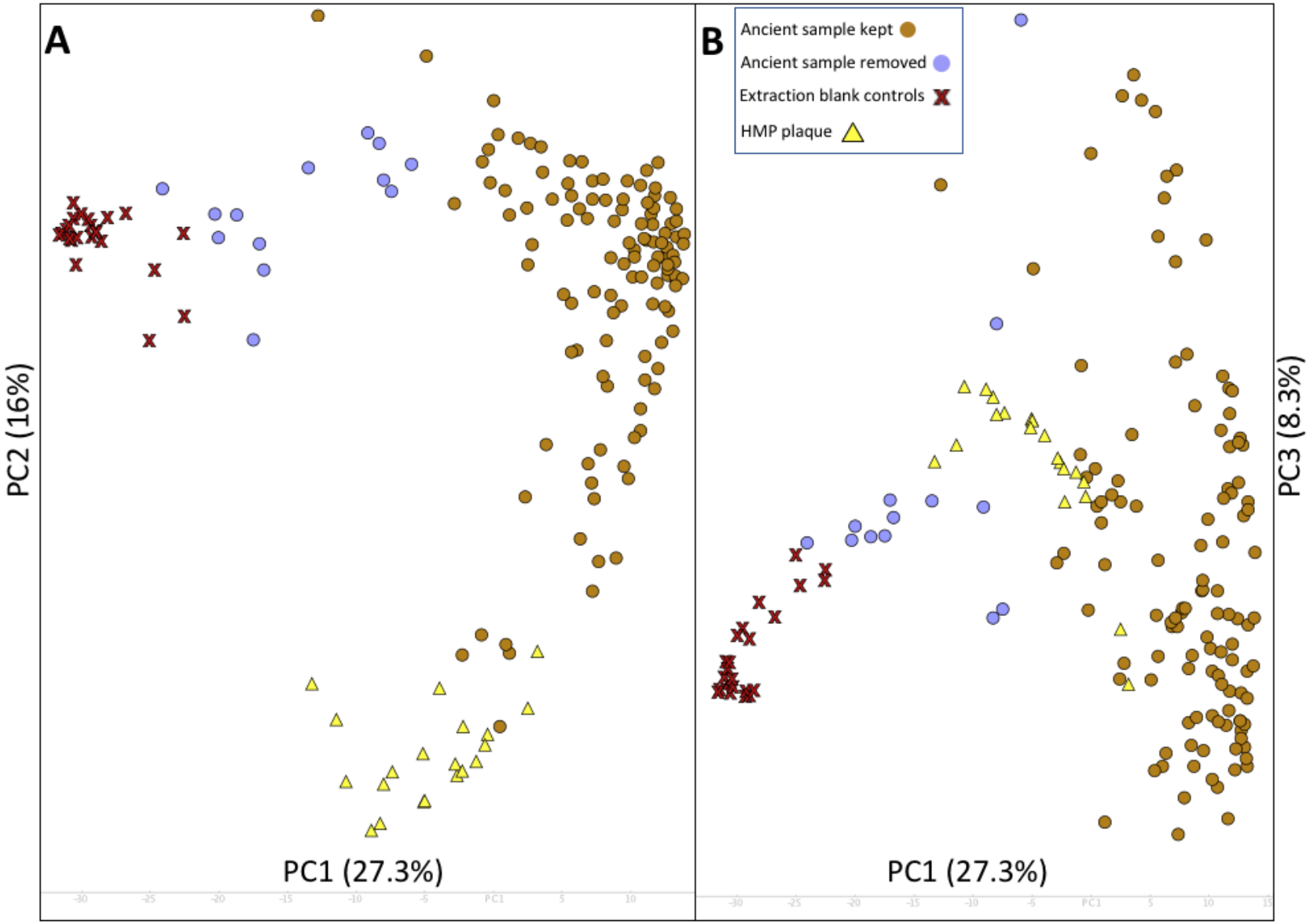


Figure S1. PCoA of Bray-Curtis distance metric showing the relation of samples to extraction controls

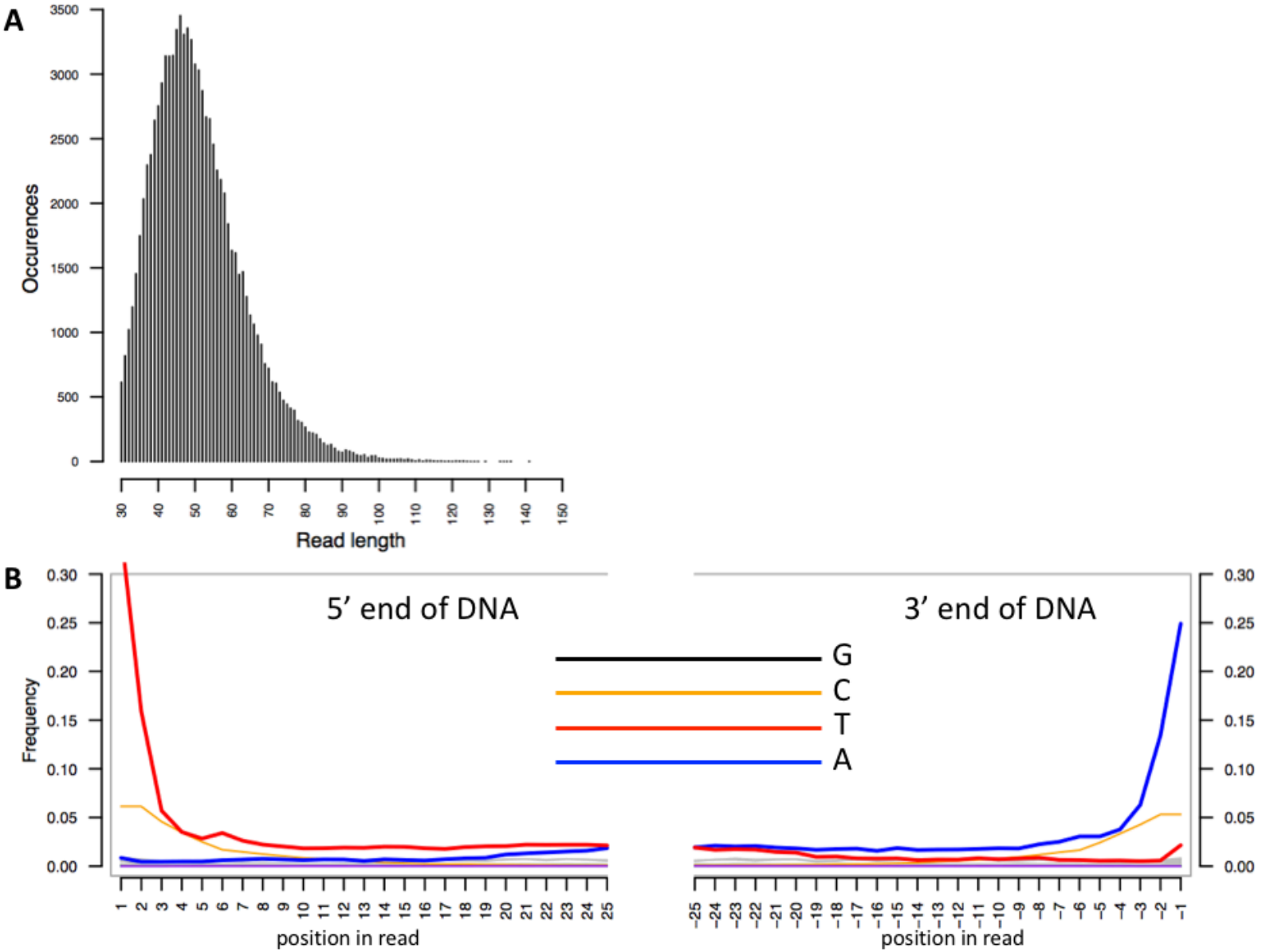


Figure S2. Example of ancient DNA damage patterns. (A) Log-normal DNA fragment length distribution. (B) Deamination-induced substitution pattern observed at ends of molecules.

Figure S3. Damage profile for Anaerolineaceae sp. oral taxon 439 at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

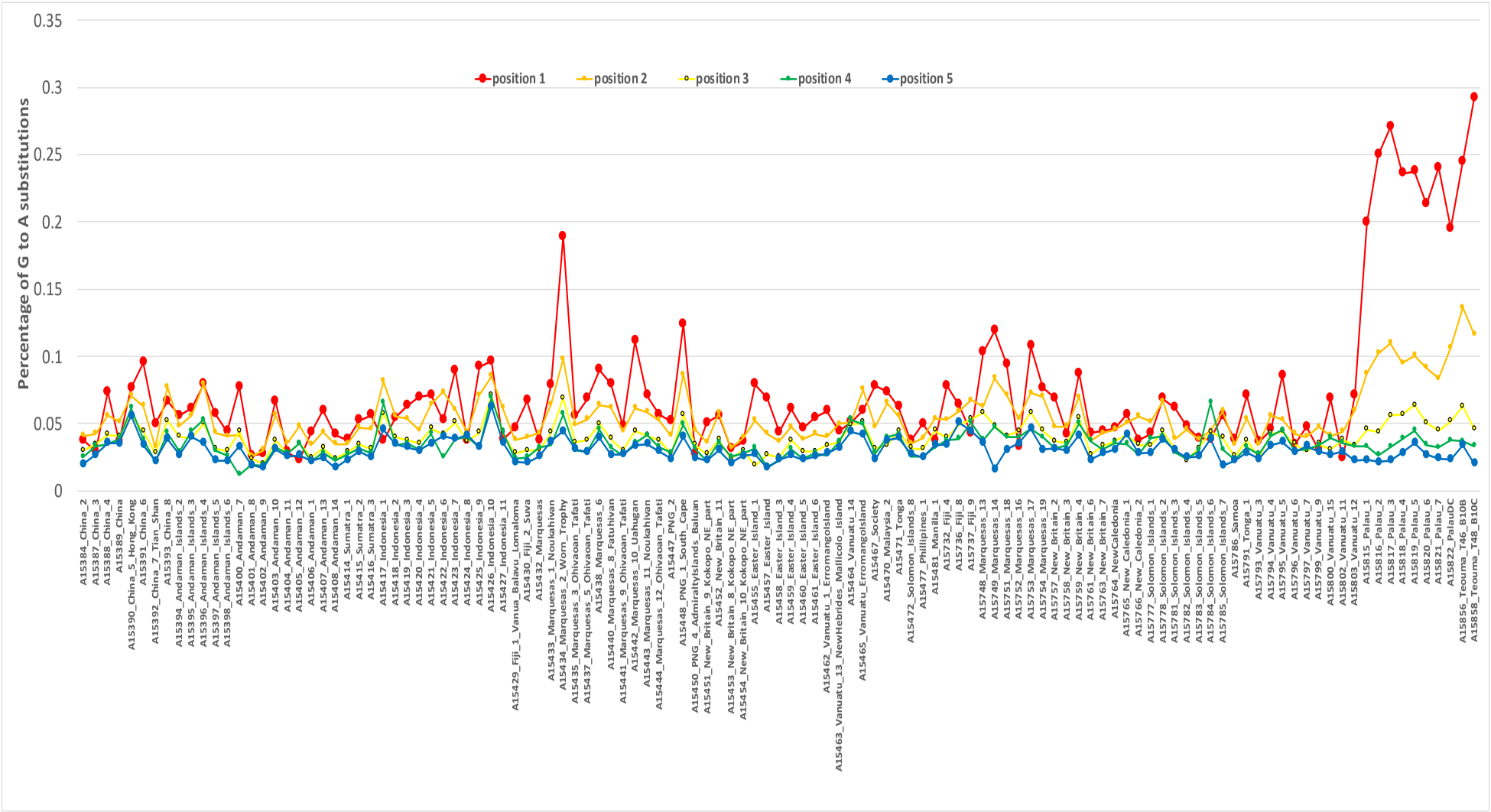


Figure S4. Damage profile for Anaerolineaceae sp. oral taxon 439 at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

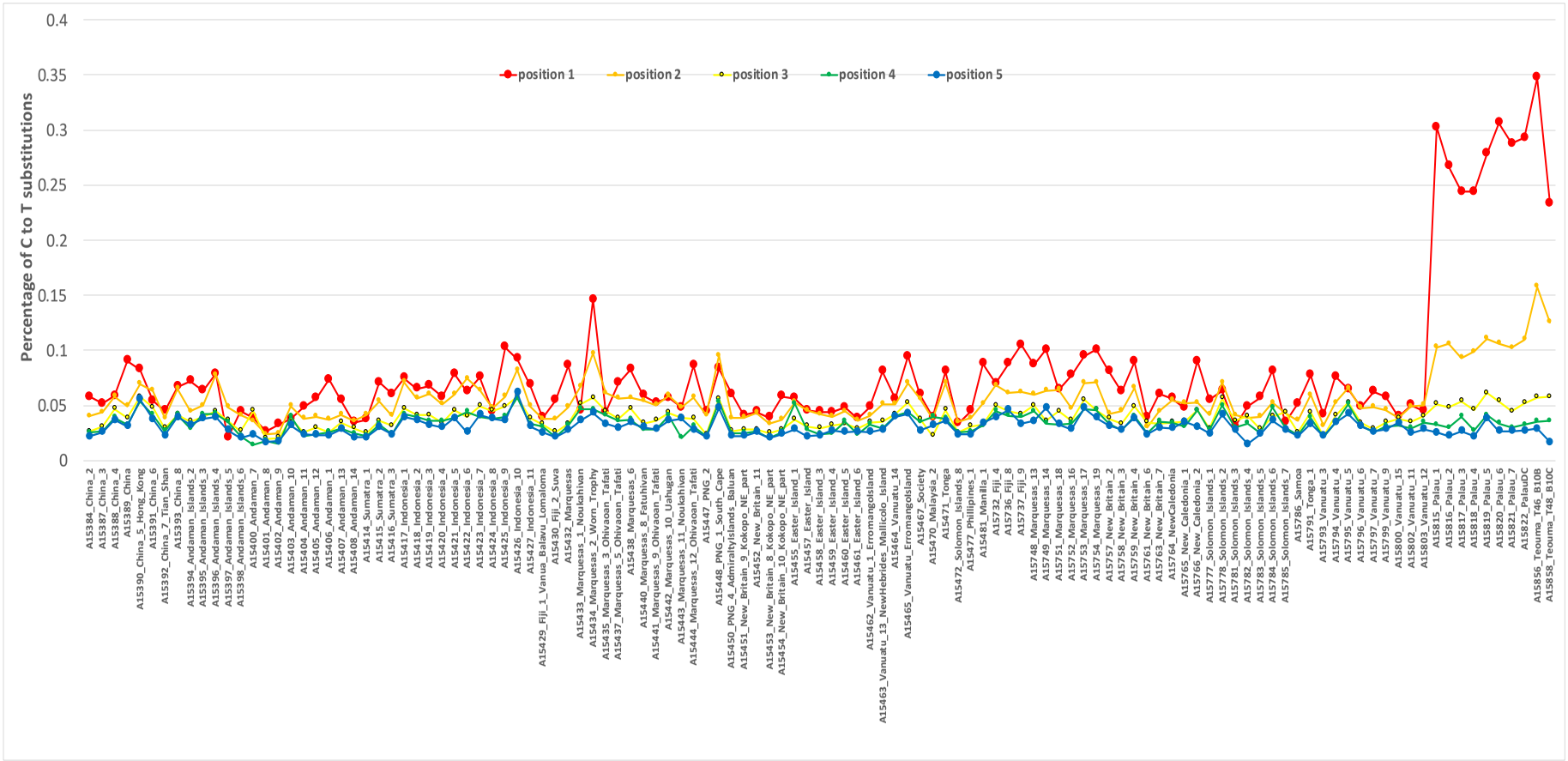


Figure S5. Damage profile for Actinomyces sp. oral taxon 414 at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

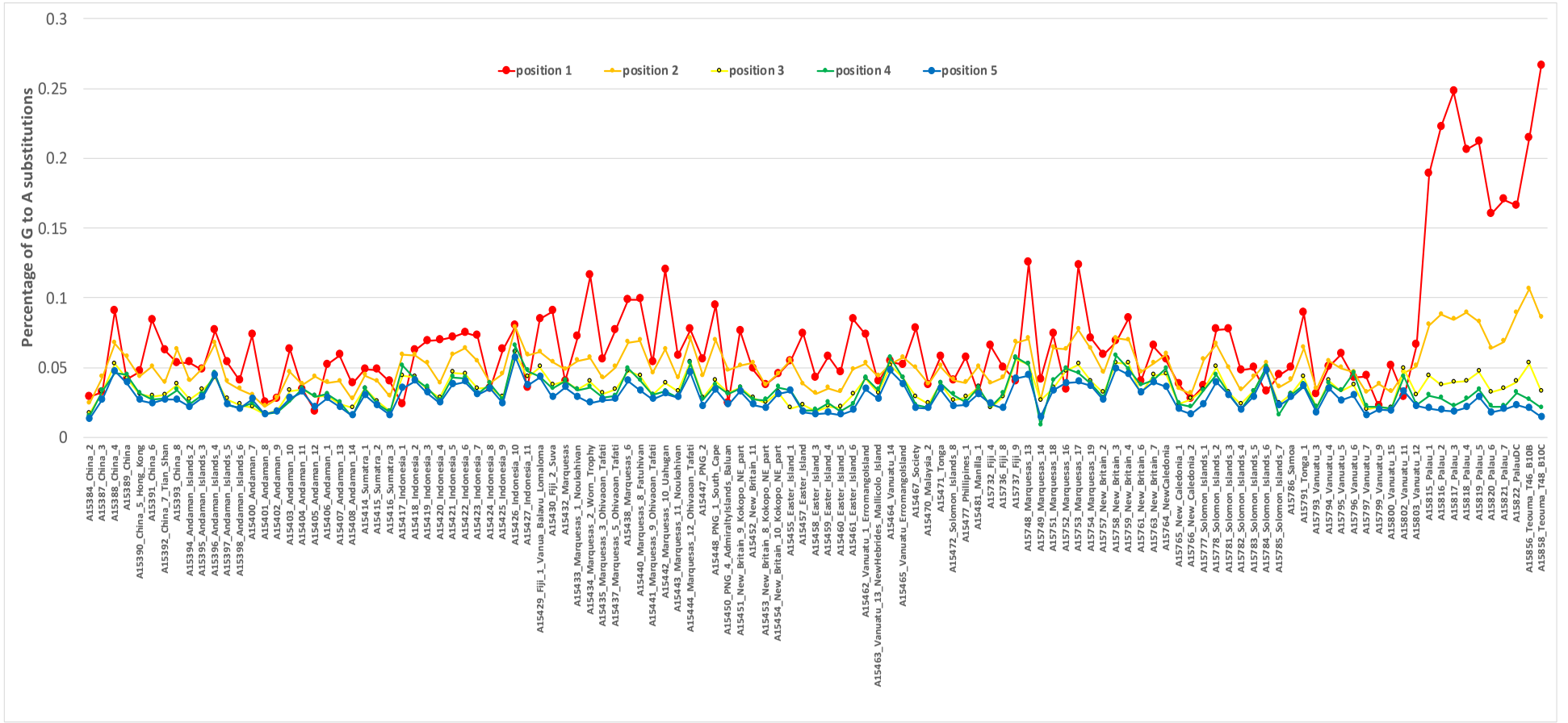


Figure S6. Damage profile for Actinomyces sp. oral taxon 414 at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

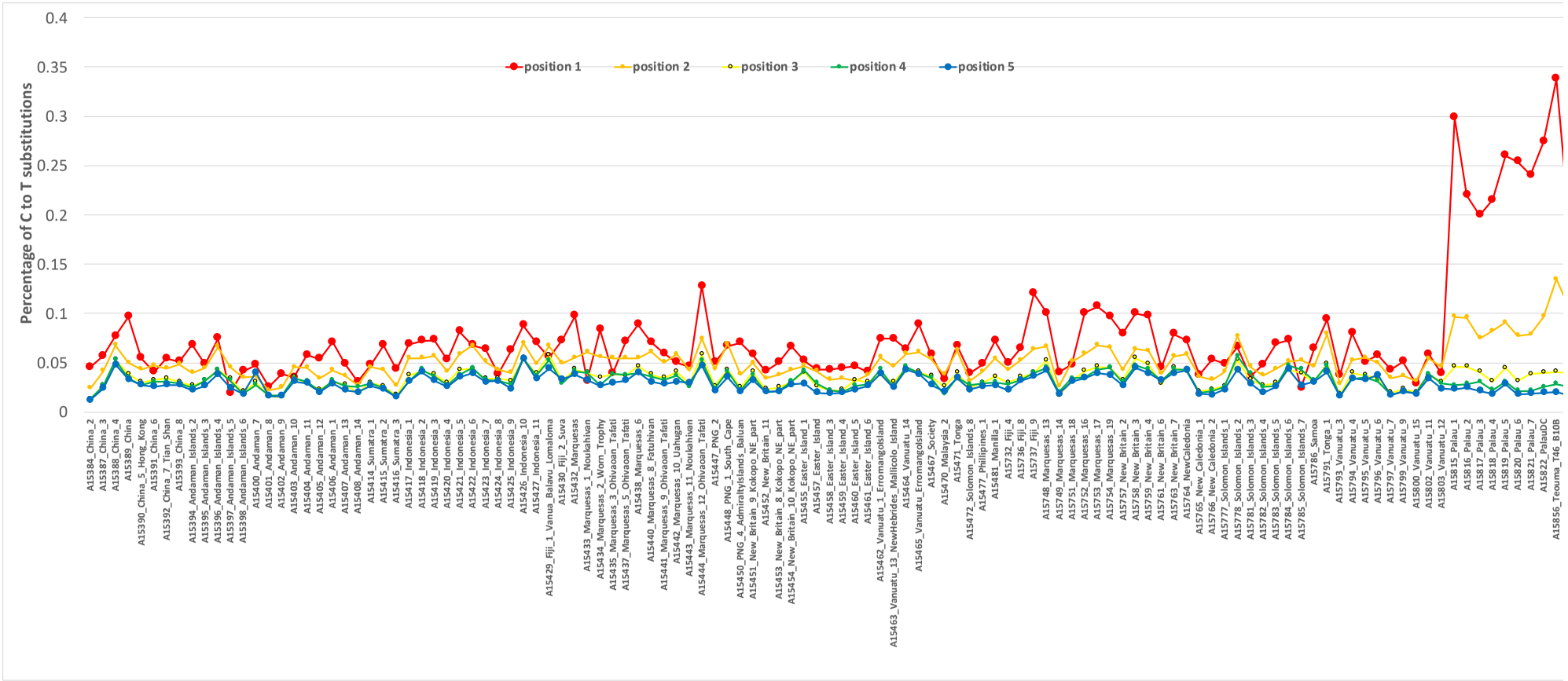


Figure S7. Damage profile for Fretibacterium fastidiosum at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

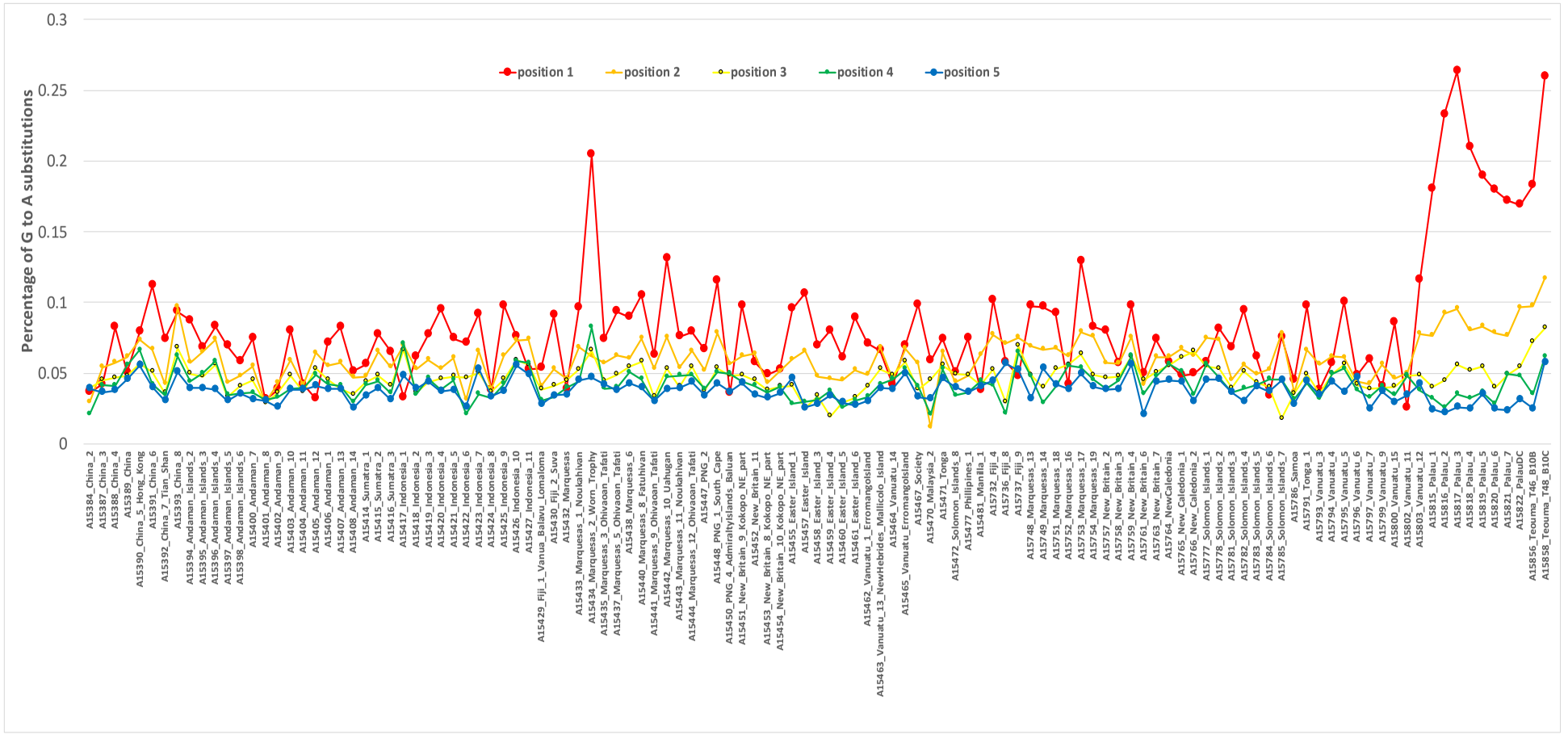


Figure S8. Damage profile for Fretibacterium fastidiosum at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

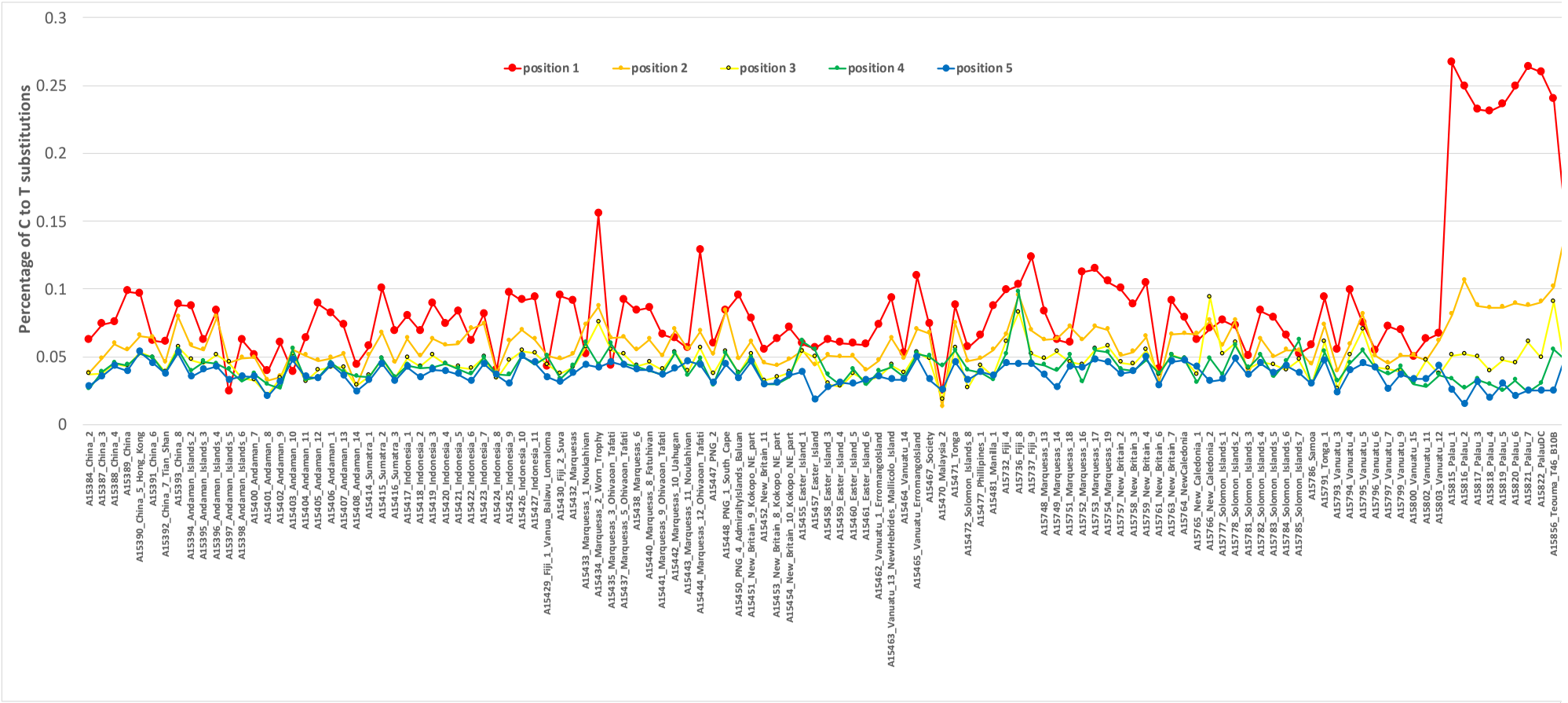


Figure S9. Damage profile for Methanobrevibacter oralis at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

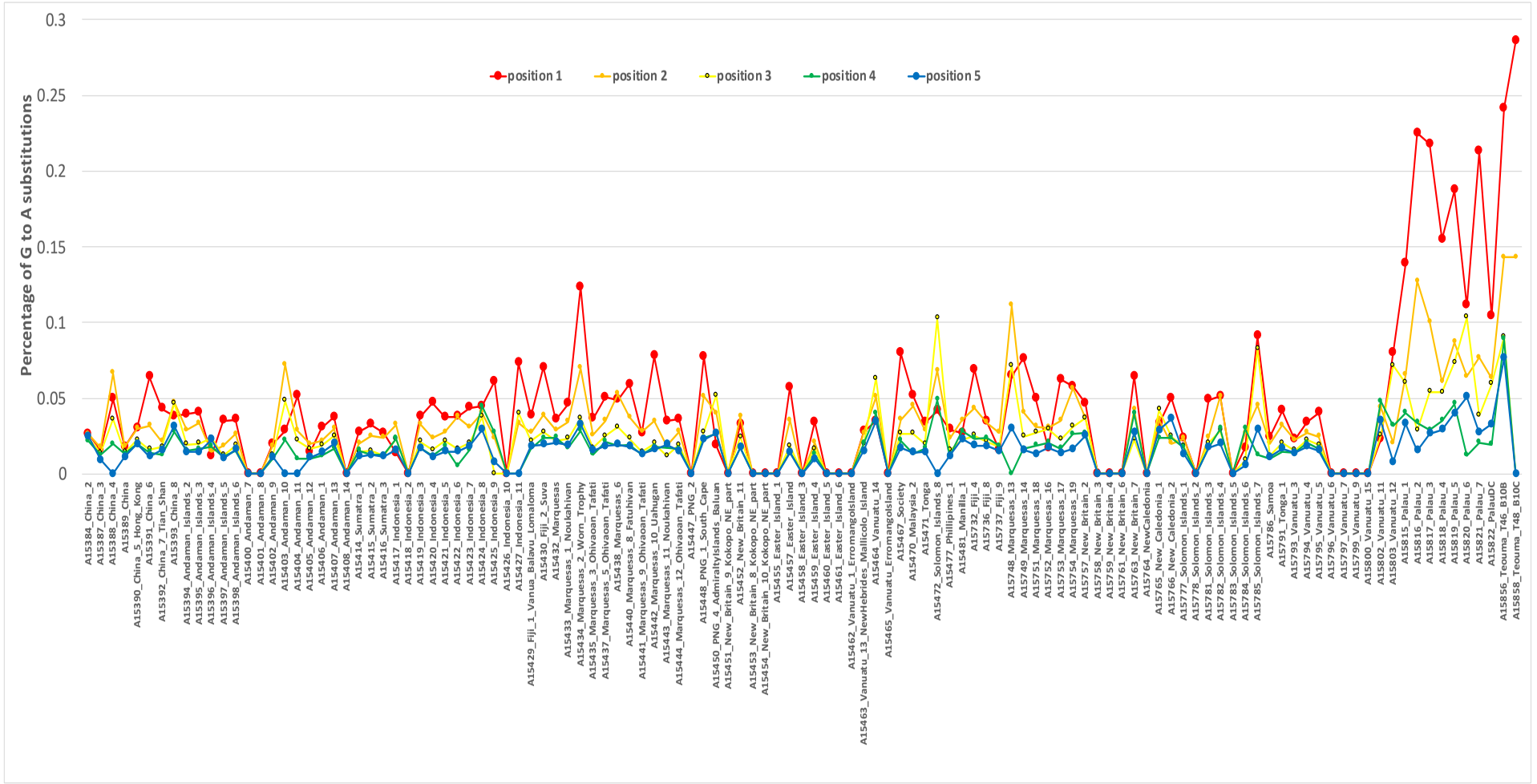
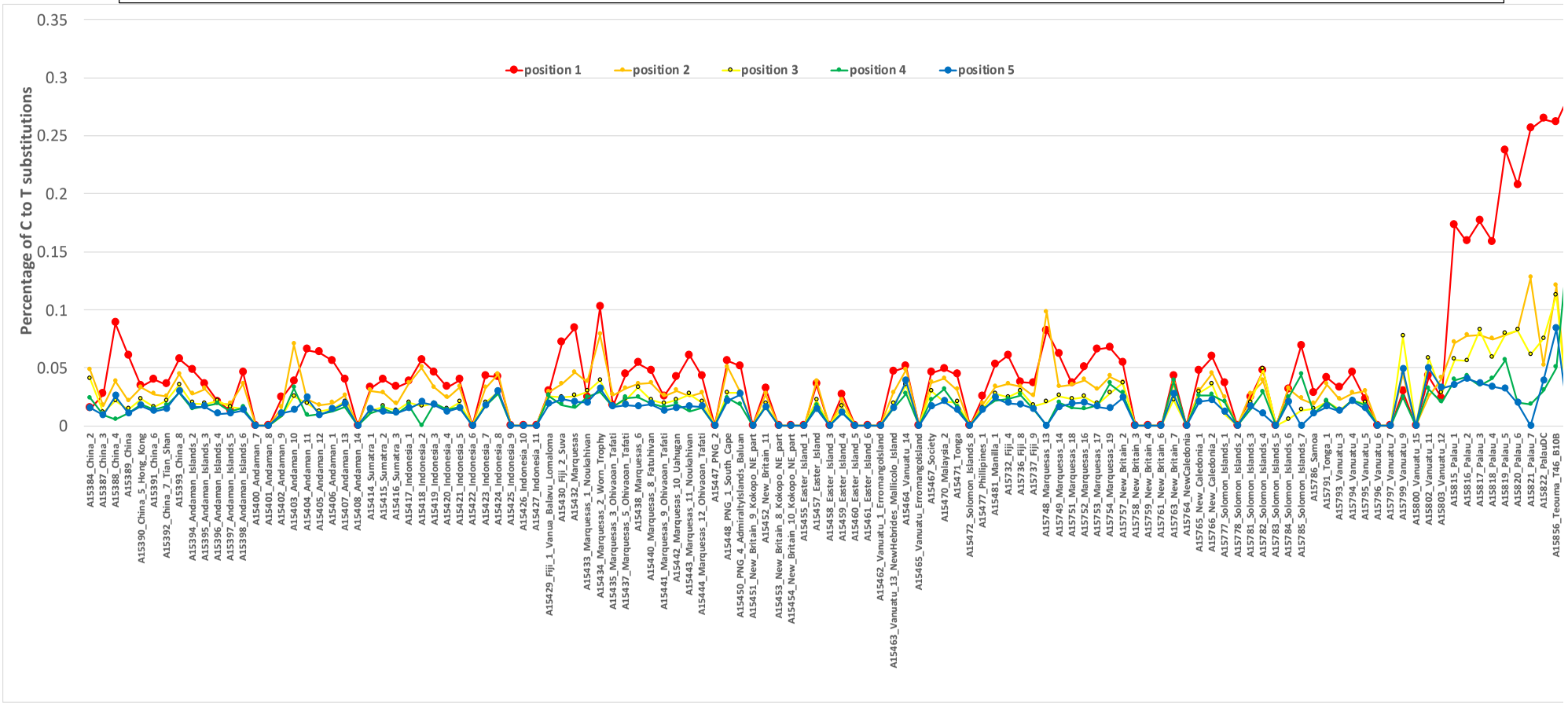


Figure S10. Damage profile for Methanobrevibacter oralis at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).



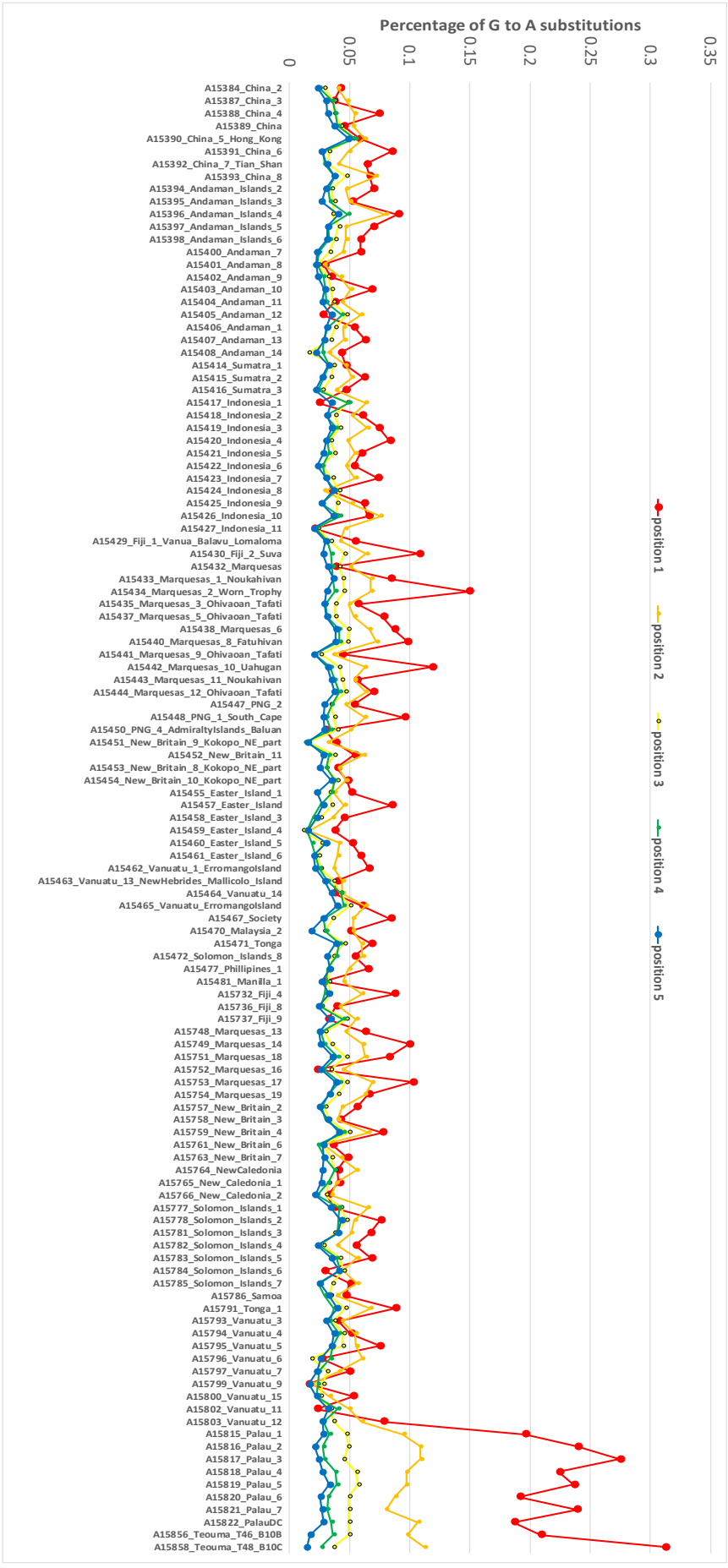
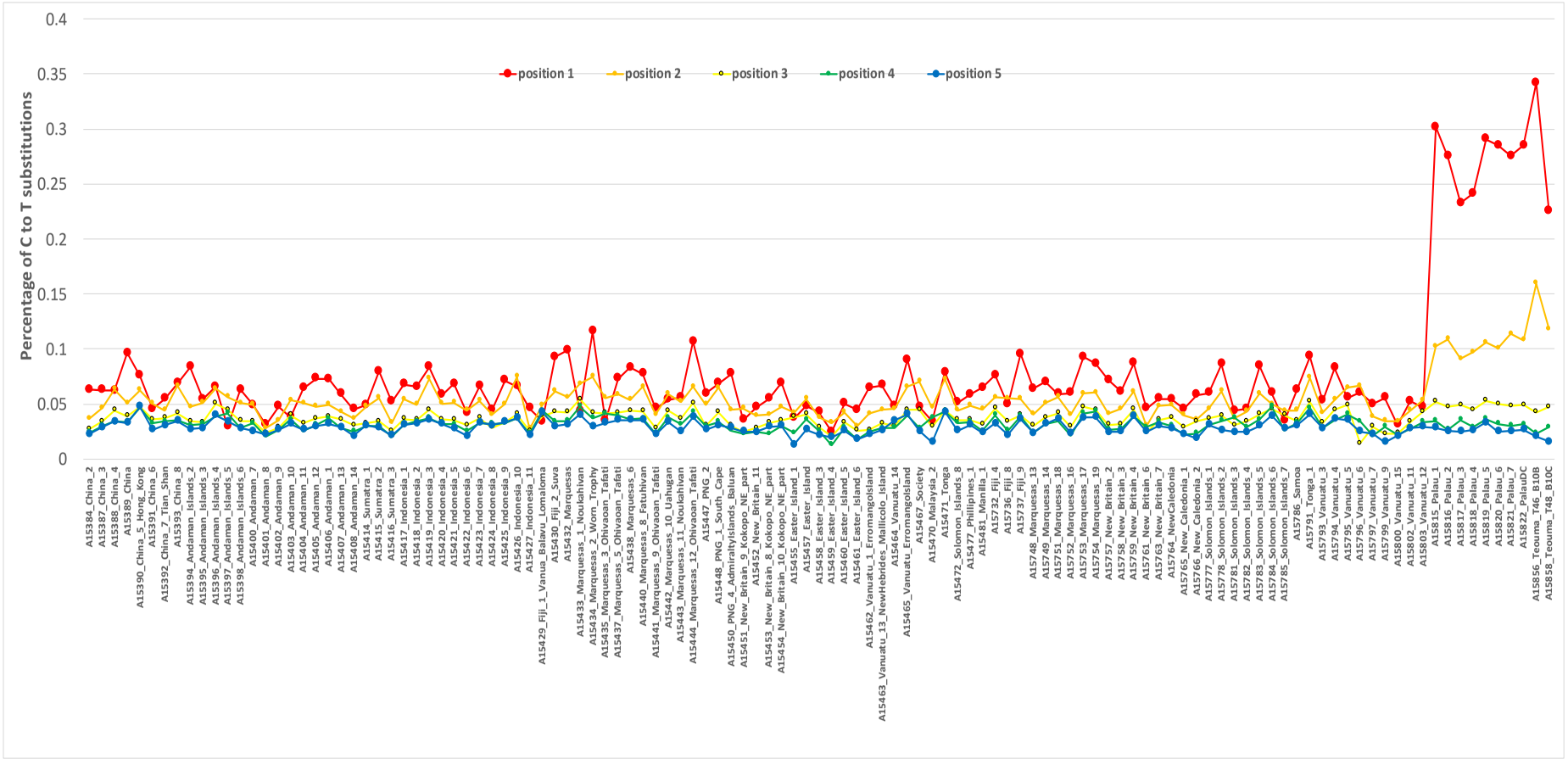


Figure S11. Damage profile for Olsenella sp. oral taxon 807 at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

Figure S12. Damage profile for Olsenella sp. oral taxon 807 at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).



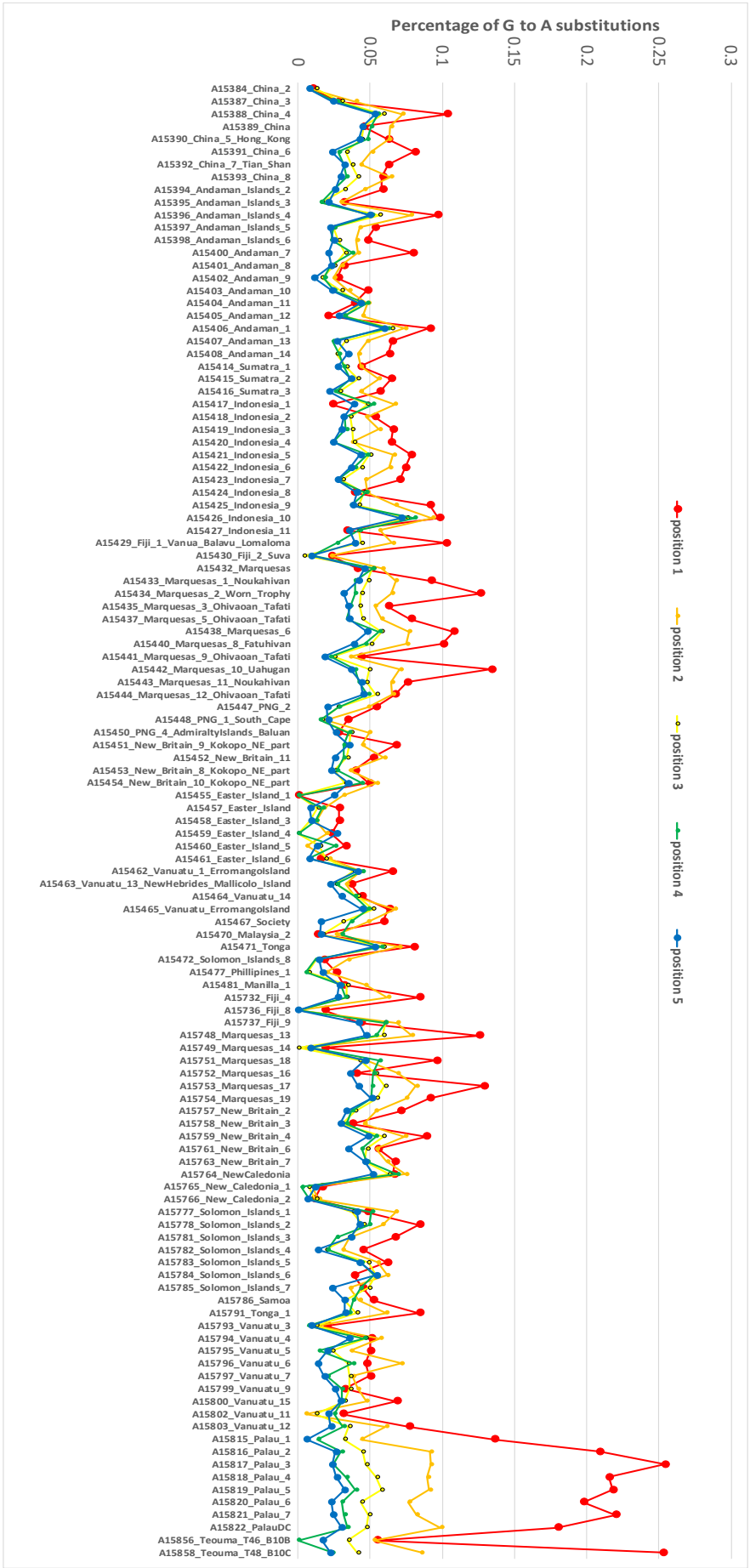


Figure S13. Damage profile for Pseudopropionibacterium propionicum at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

Figure S14. Damage profile for Pseudopropionibacterium propionicum at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

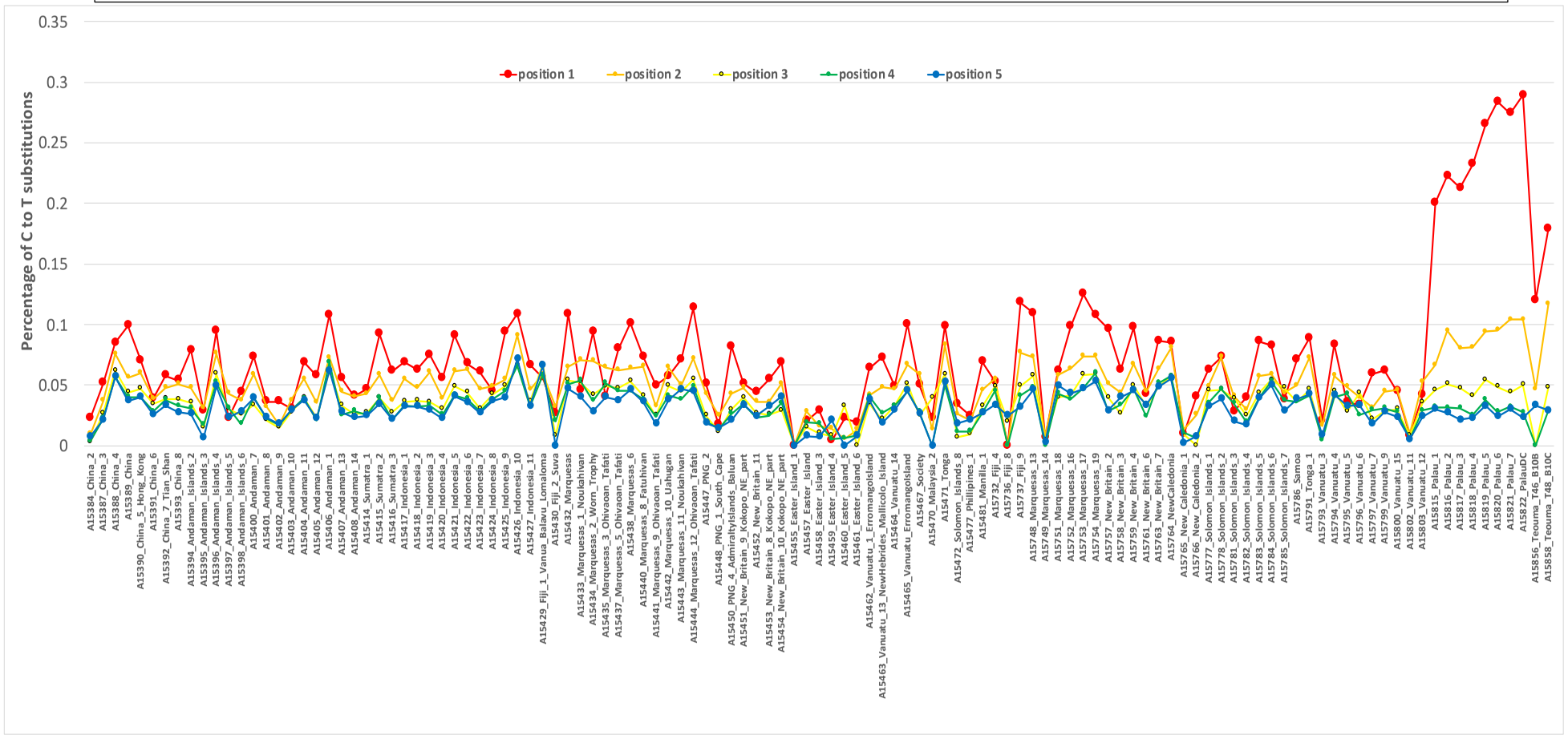


Figure S15. Damage profile for Tannerella forsythia at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

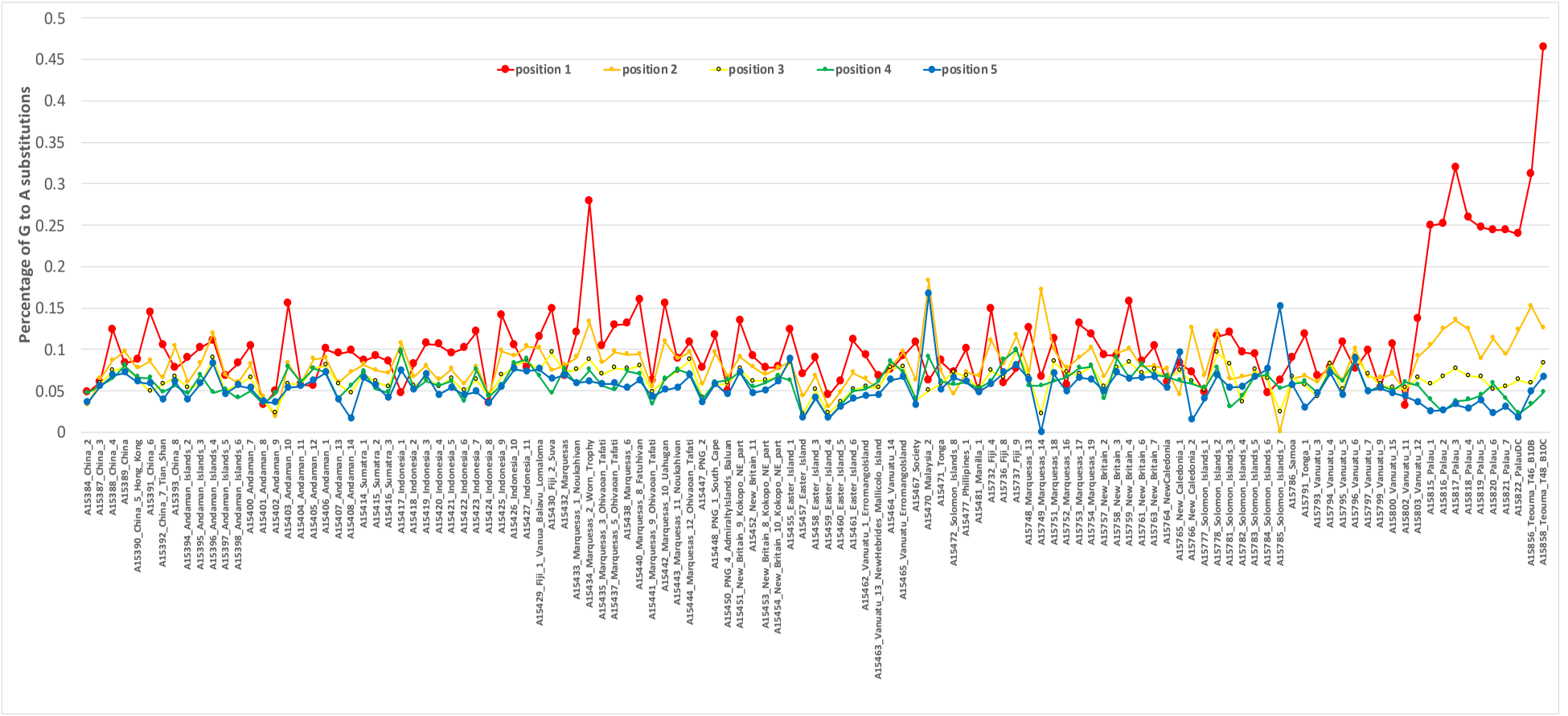


Figure S16. Damage profile for Tannerella forsythia at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

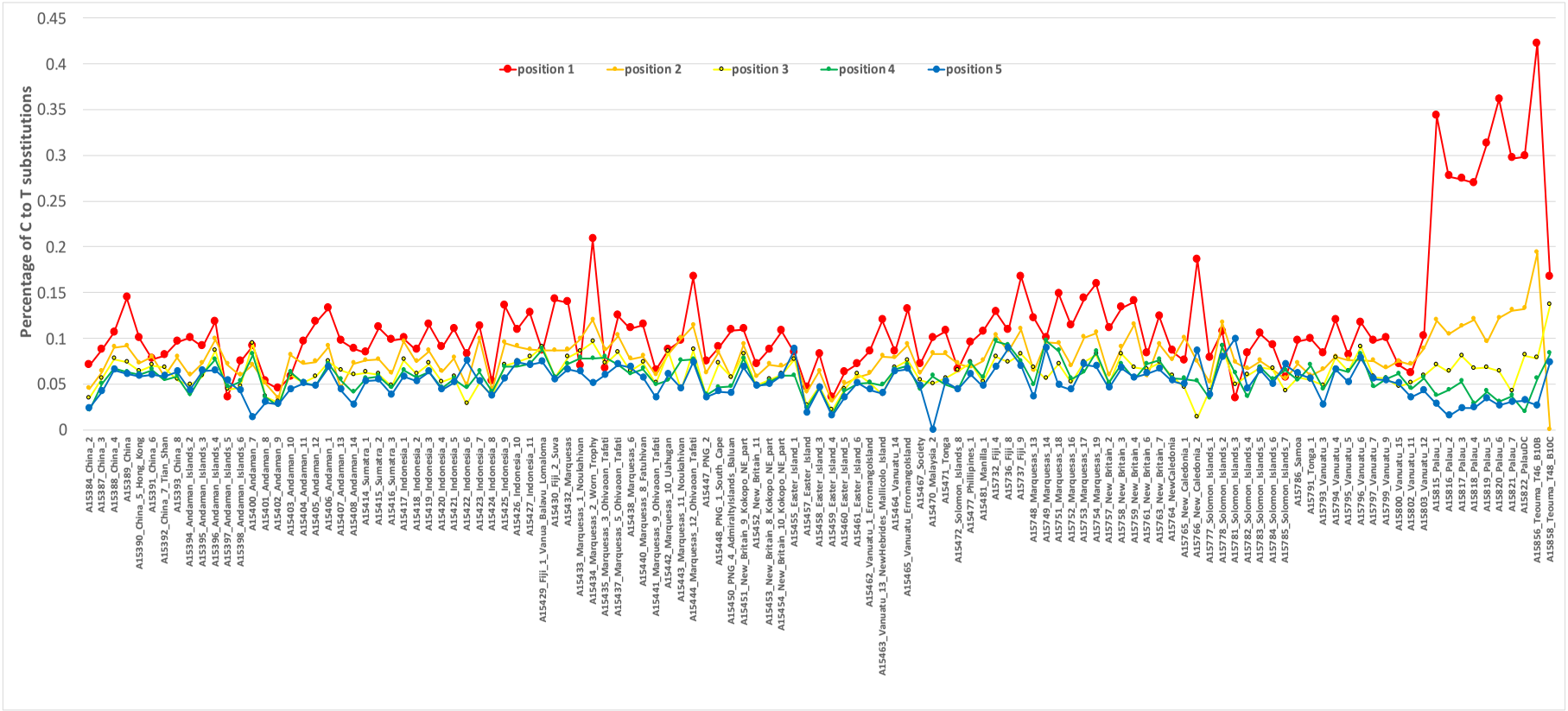


Figure S17. Damage profile for Actinomyces sp. oral taxon 414 at 3' terminus for modern HMP samples. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).

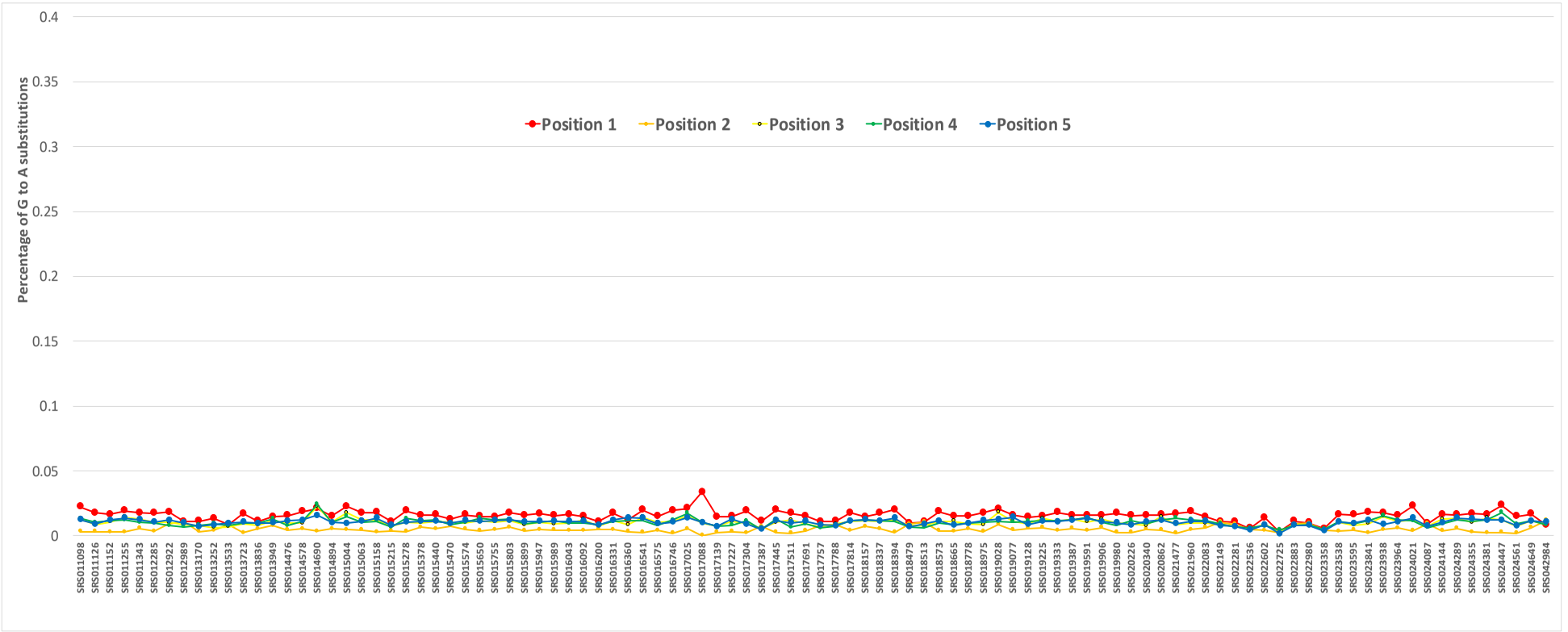


Figure S17 continued

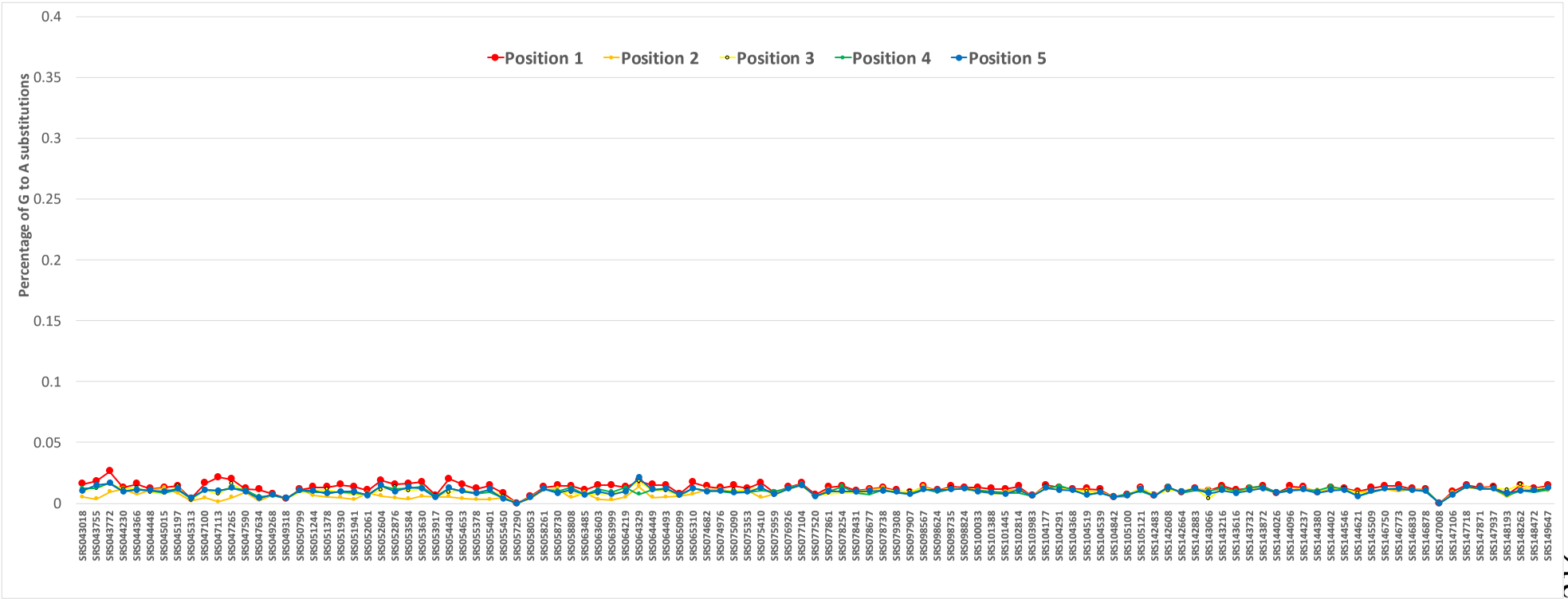
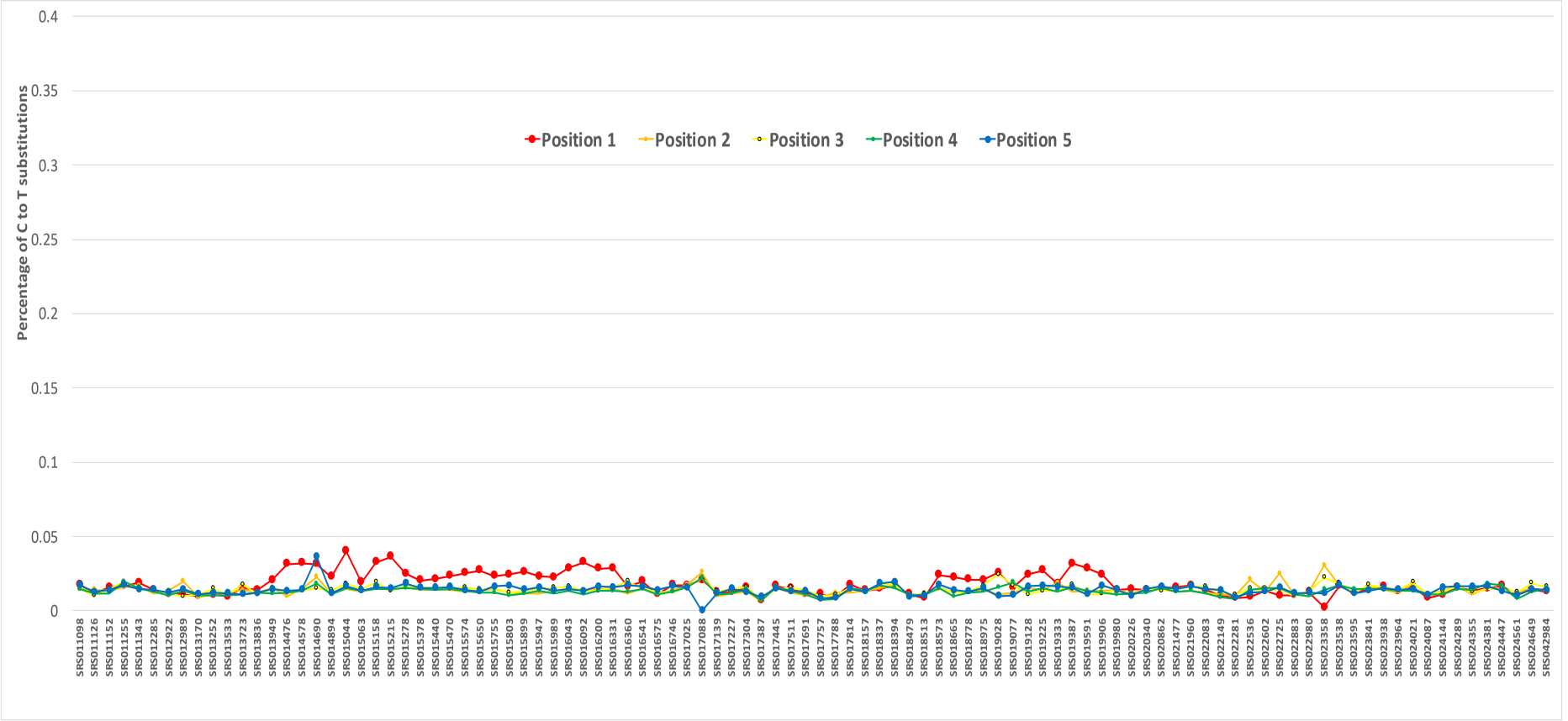


Figure S18. Damage profile for Actinomyces sp. oral taxon 414 for modern HMP samples at the 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules for all 117 samples. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red) and position 5 being five bases adjacent (blue).



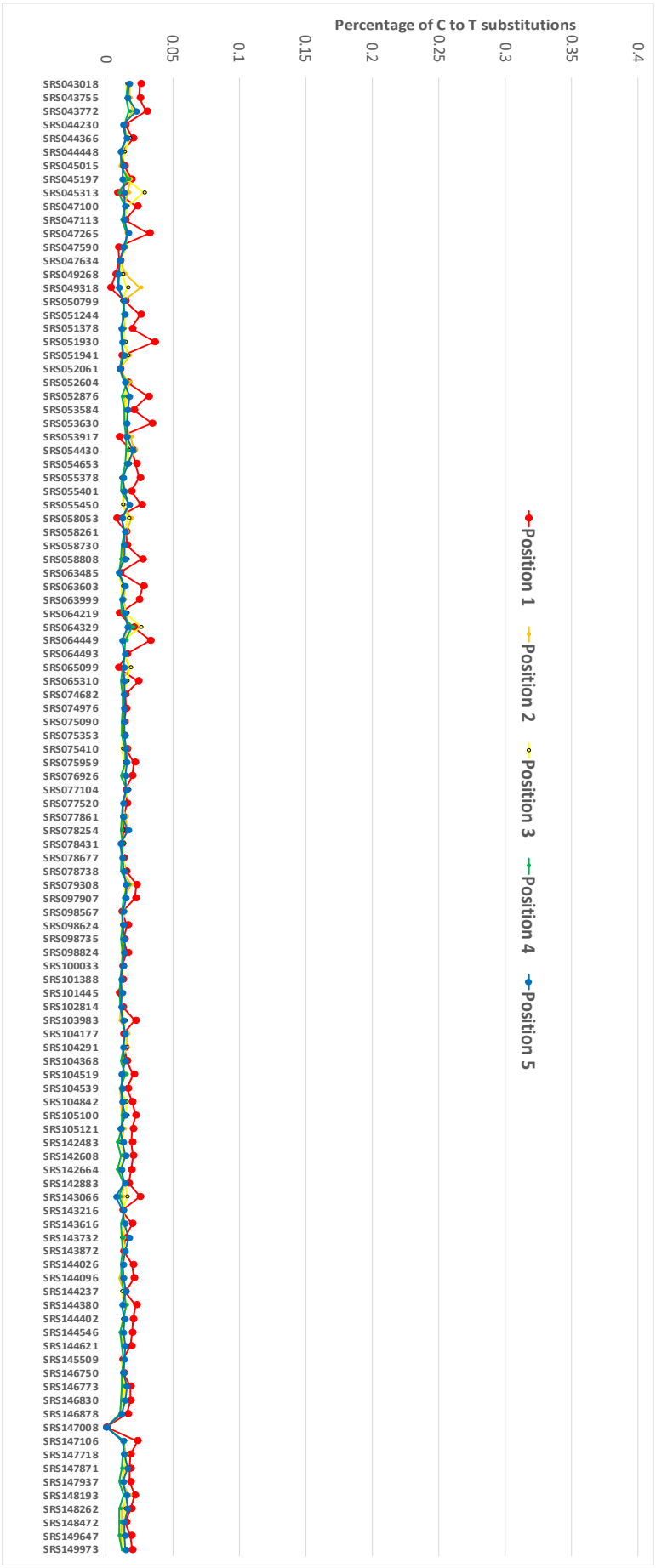
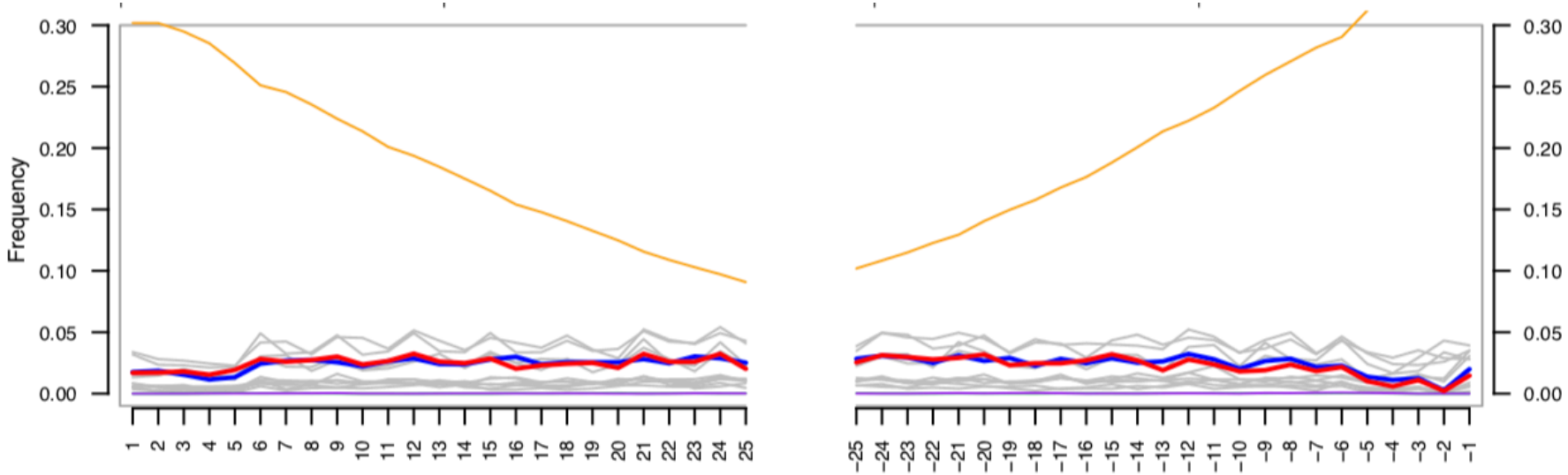


Figure S18 continued

Figure S19. Example mapDamage plot of modern HMP sample. C-to-T or G-to-A substitutions do not reach above background levels.



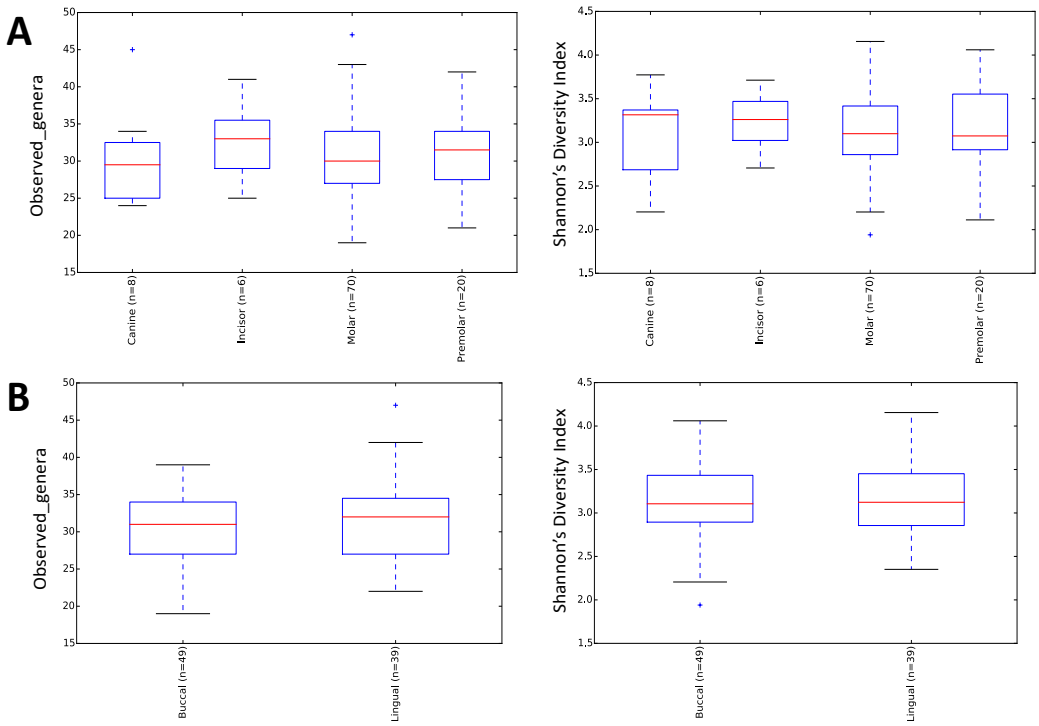


Figure S20. Alpha diversities for tooth type and tooth surface

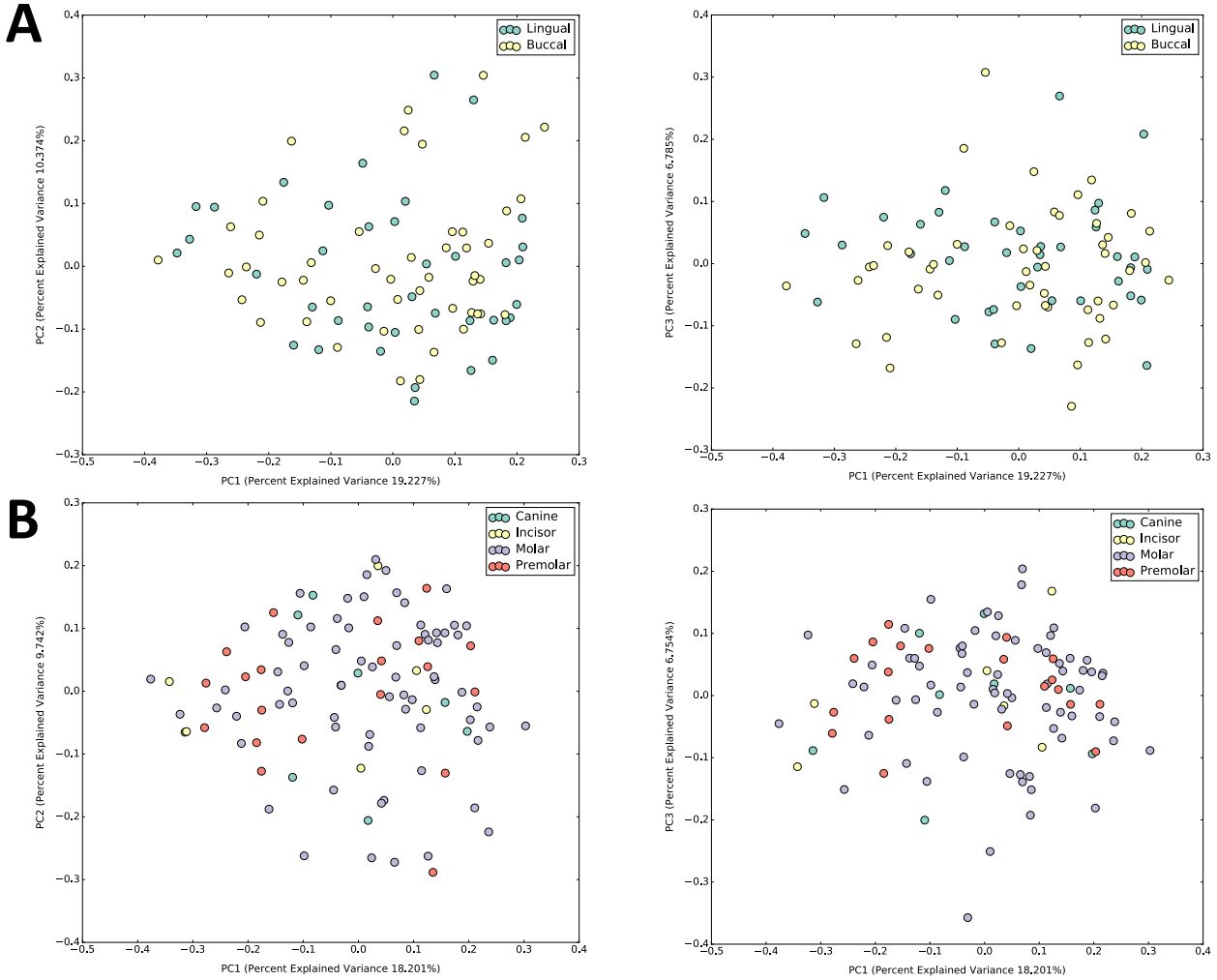


Figure S21. PCoA of Binary Jaccard distances between samples labelled by tooth surface (A) or tooth type (B)

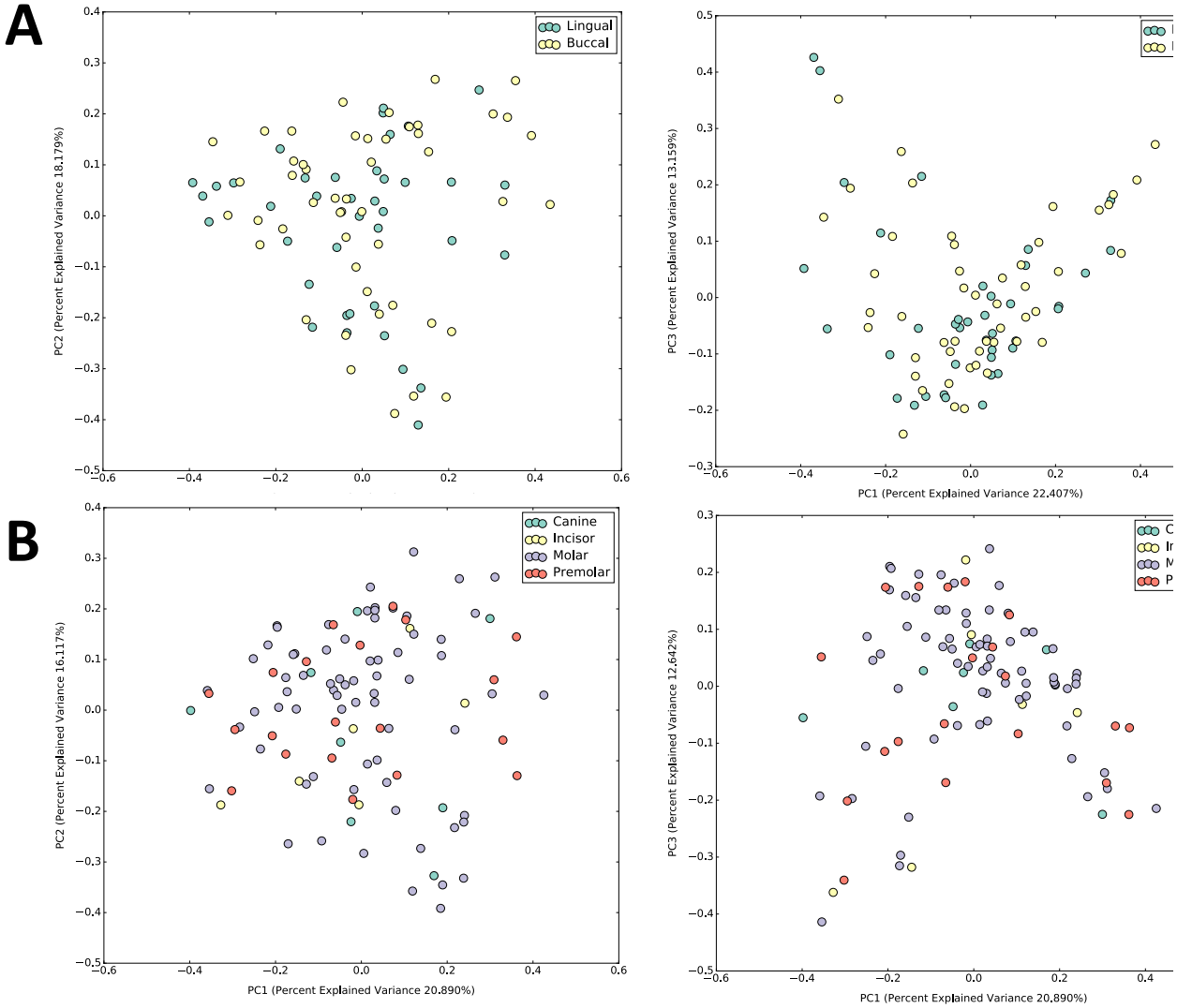


Figure S22. PCoA of Bray-Curtis distances between samples labelled by tooth surface (A) or tooth t

Table S2. Statistical analysis of beta diversity metrics by tooth type and su

Distance metric	Metadata category	Statistical test	R^2	Test statistic	p
Binary Jaccard	Tooth Surface	PERMANOVA		0.926	
Bray-Curtis	Tooth Surface	PERMANOVA		1.338	
Binary Jaccard	Tooth Type	PERMANOVA		1.066	
Bray-Curtis	Tooth Type	PERMANOVA		0.781	
Binary Jaccard	Tooth Surface	ANOSIM	0.010		
Bray-Curtis	Tooth Surface	ANOSIM	-0.004		
Binary Jaccard	Tooth Type	ANOSIM	0.003		
Bray-Curtis	Tooth Type	ANOSIM	0.078		

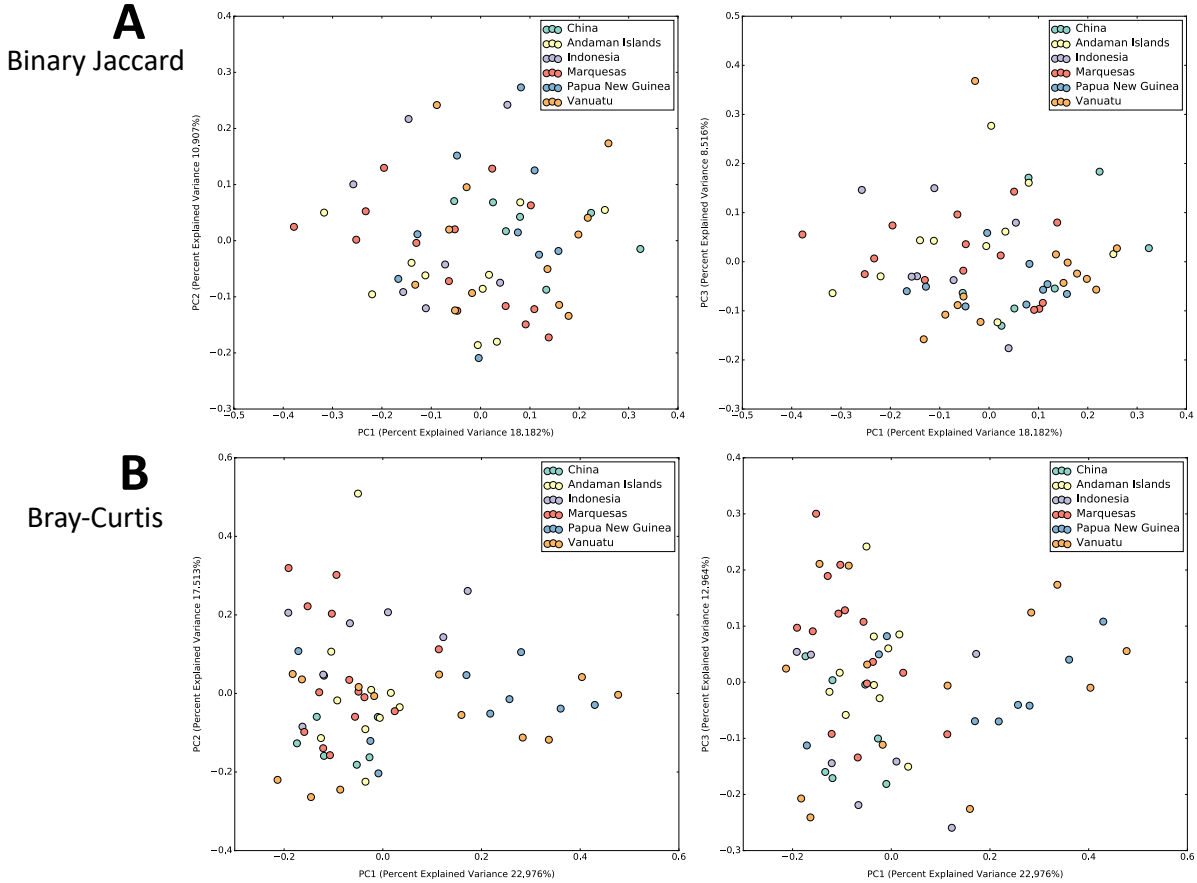


Figure S23. PCoA ordination of beta diversity labelled by Island

Table S3. Statistical analysis of beta diversity metrics by tooth type and surface

Distance metric	Metadata category	Statistical test	R^2	Test statistic	p-value
Binary Jaccard	Tooth Surface	PERMANOVA		1.695	0.003
Bray-Curtis	Tooth Surface	PERMANOVA		2.463	0.001
Binary Jaccard	Tooth Surface	ANOSIM	0.122		0.004
Bray-Curtis	Tooth Surface	ANOSIM	0.130		0.004

Table S4. Statistically significant genera detected between islands/regions

Test-Statistic	P	FDR_P	Bonferroni_P	Indonesia_mean	Papua New Guinea_mean	Marquesas_mean	China_mean	Andaman Islands_mean	Vanuatu_mean	taxonomy
23.771	0.000	0.013	**0.027**	1250	1505	1215	1449	1244	82	Desulfomicrobium
23.493	0.000	0.013	**0.030**	0	0	0	48	0	0	Asaccharospora
22.873	0.000	0.013	**0.040**	6580	1181	3148	1638	2168	794	Pseudopropionibacterium

Table S5. Statistical analysis of beta diversity in relation to disease state

Distance metric	Metadata category	Statistical test	R^2	Test statistic	p-value
Binary Jaccard	Caries	PERMANOVA		0.927	0.532
Bray-Curtis	Caries	PERMANOVA		0.483	0.869
Binary Jaccard	Periodontal disease	PERMANOVA		1.504	0.109
Bray-Curtis	Periodontal disease	PERMANOVA		1.799	0.070
Binary Jaccard	Caries	ANOSIM	-0.010		0.770
Bray-Curtis	Caries	ANOSIM	-0.135		0.877
Binary Jaccard	Periodontal disease	ANOSIM	-0.010		0.531
Bray-Curtis	Periodontal disease	ANOSIM	-0.008		0.518

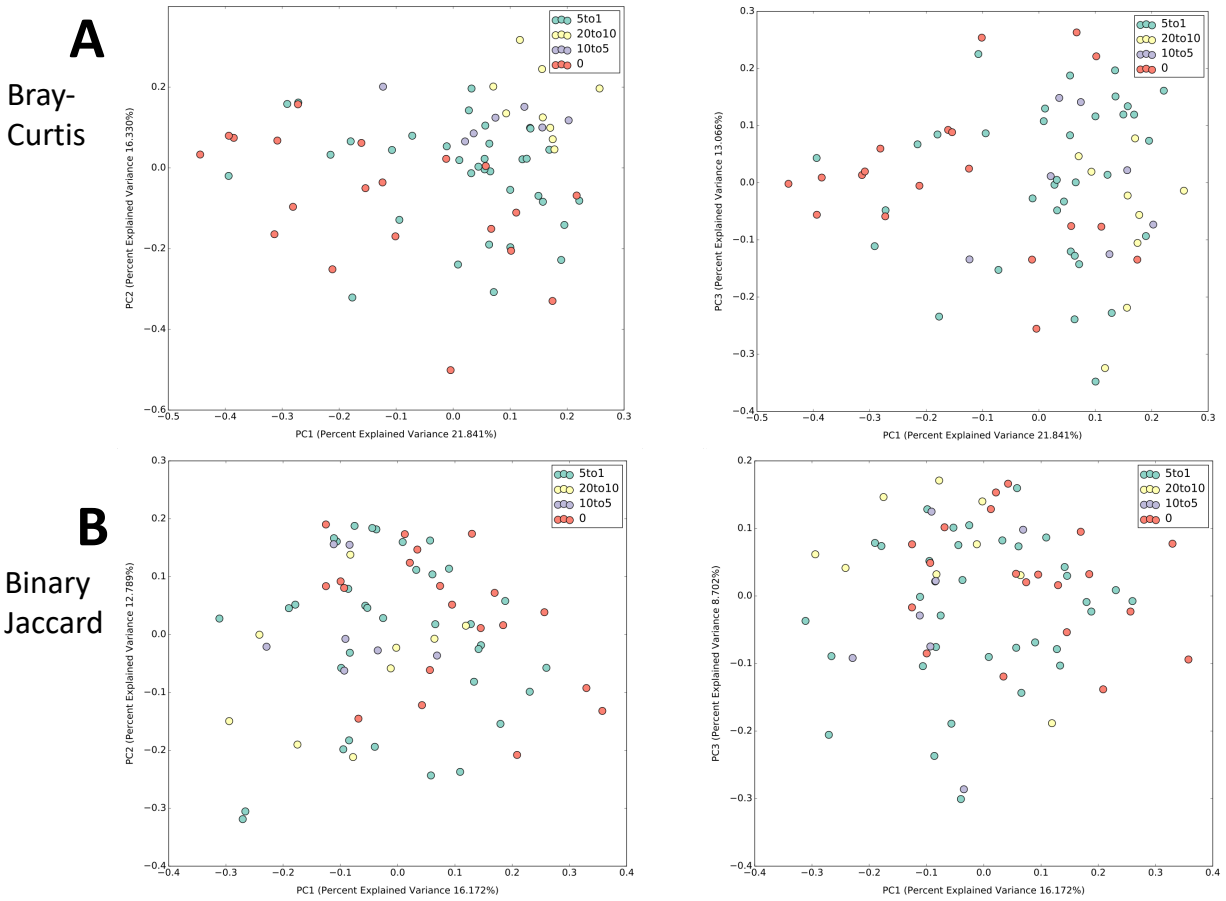


Figure S24. PCoA ordination of beta-diversity metrics labelled by abundance of *Methanobrevibacter*

Table S6. Genomes used for mapping and phylogenetic reconstruction

Genus	Species	Genome Assembly
Actinomyces	*cardiffensis*	F0333
Actinomyces	*dentalis*	DSM 19115
Actinomyces	*gerencseriae*	DSM 6844
Actinomyces	*israelii*	DSM 43320
Actinomyces	*massiliensis*	4401292
Actinomyces	*oris*	ASM155393v1
Actinomyces	*sp. oral taxon 414*	F0588
Anaerolineaceae	*sp. oral taxon 439*	ASM27771v1
Fretibacterium	*fastidiosum*	ASM21071v1
Methanobrevibacter	*oralis*	JMR01
Olsenella	*sp. oral taxon 807*	F0089
Pseudopropionibacterium	*propionicum*	F0230a
Tannerella	*forsythia*	ASM23821v1

Table S7. Mean abundance of *Actinomyces* species identified in MALT analysis

Genome	Mean abundance (%)
Actinomyces sp. oral taxon 414	**48.174**
Actinomyces israelii	**21.846**
Actinomyces dentalis	**18.571**
Actinomyces cardiffensis	**3.577**
Actinomyces oris	**2.307**
Actinomyces gerencseriae	**1.497**
Actinomyces massiliensis	**1.381**
Actinomyces naeslundii	0.655
Actinomyces odontolyticus	0.448
Actinomyces meyeri	0.272
Actinomyces glycerinitolerans	0.187
Actinomyces georgiae	0.177
Actinomyces sp. oral taxon 170	0.132
Actinomyces provencensis	0.131
Actinomyces slackii	0.125
Actinomyces radicidentis	0.125
Actinomyces turicensis	0.100
Actinomyces sp. oral taxon 849	0.068
Actinomyces sp. oral taxon 180	0.053
Actinomyces sp. oral taxon 848	0.047
Actinomyces sp. HPA0247	0.040
Actinomyces sp. oral taxon 448	0.033
Actinomyces timonensis	0.028
Actinomyces johnsonii	0.008
Actinomyces sp. oral taxon 877	0.006
Actinomyces sp. oral taxon 178	0.005
Actinomyces urogenitalis	0.004
Actinomyces sp. pika_114	0.003

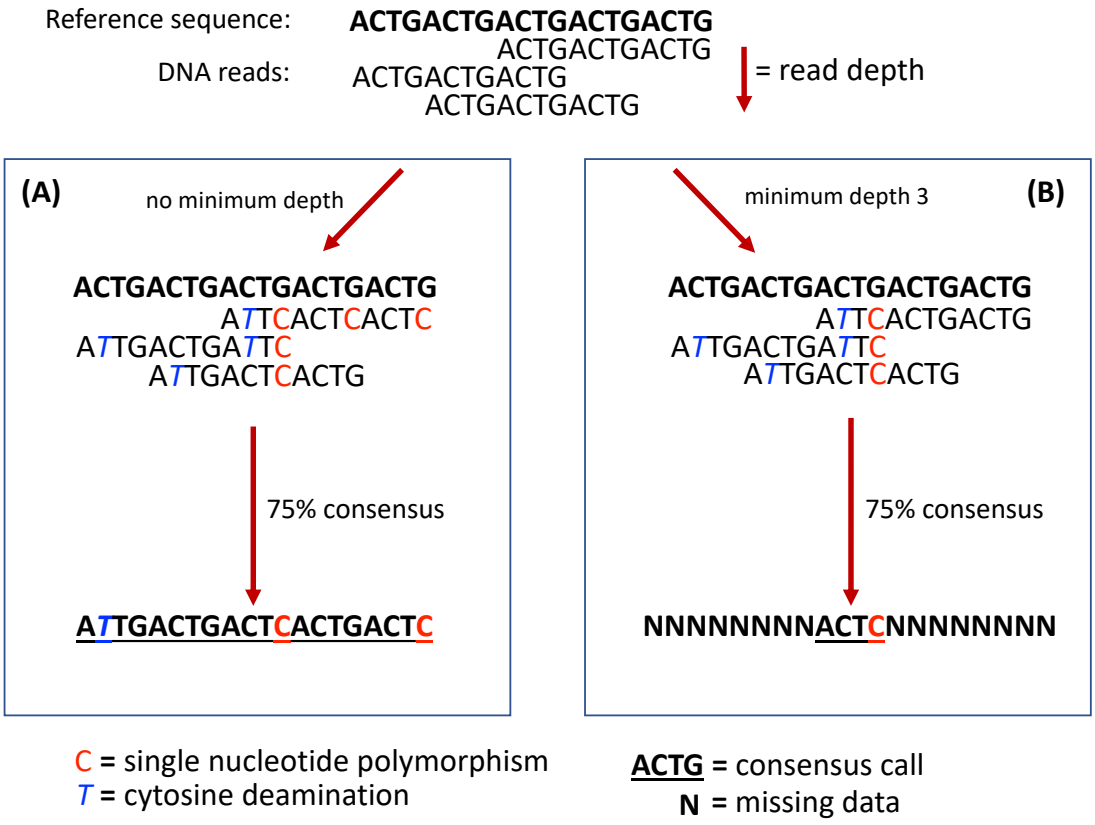


Figure S25. Illustration of consensus sequence call procedure

Figure S26. *Actinomyces sp. oral taxon 414*, no competitive mapping

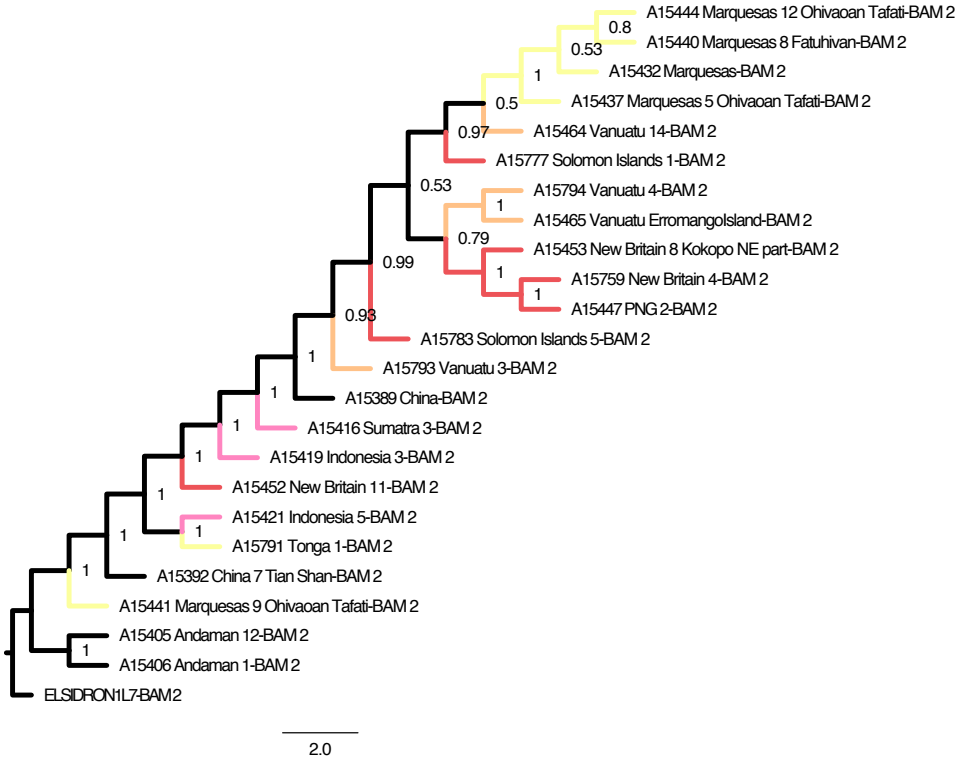


Figure S27. *Actinomyces sp. oral taxon 414*, competitive mapping

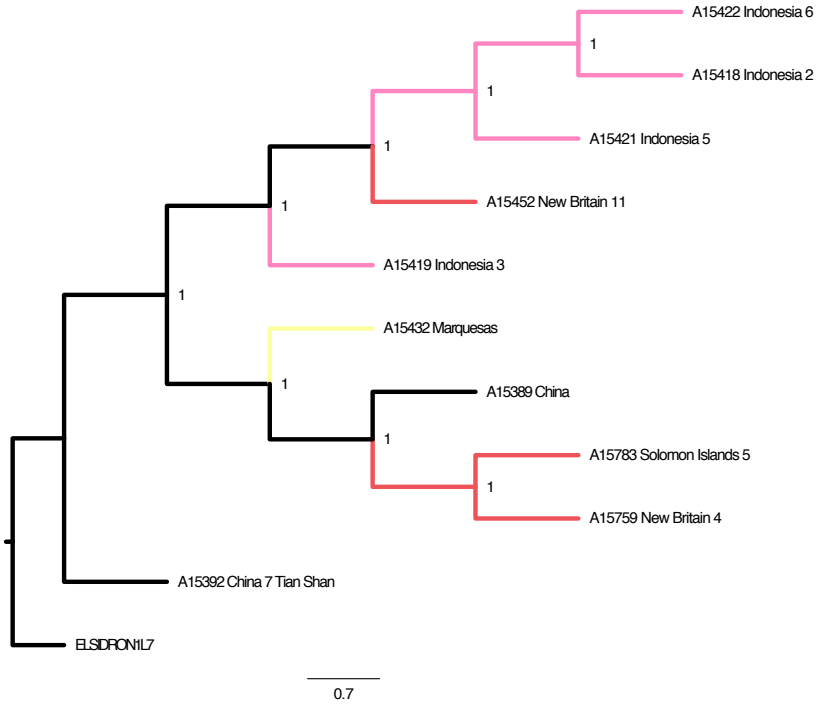


Figure S28. *Actinomyces dentalis*, competitive mapping

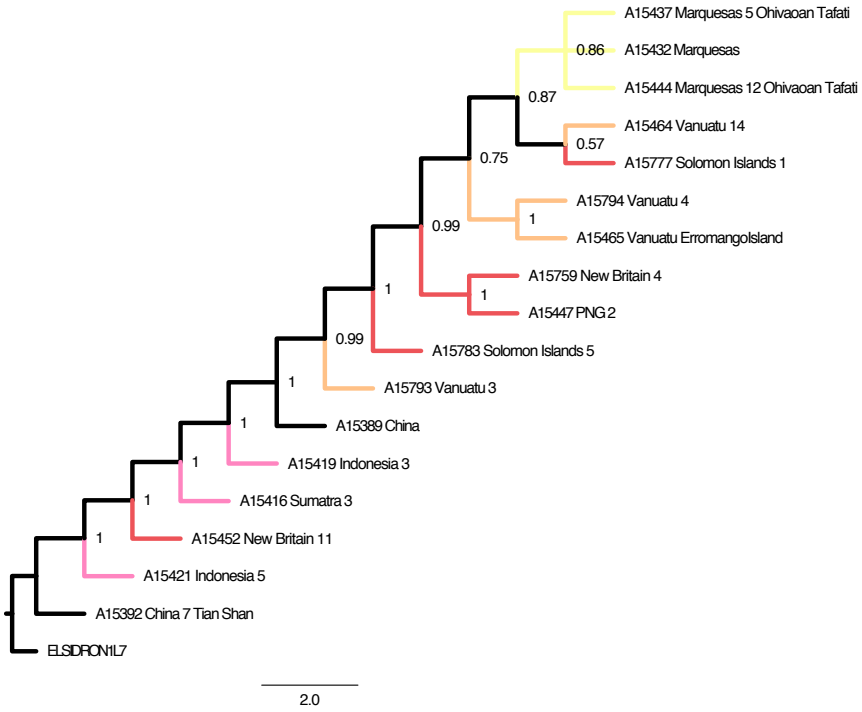


Figure S29. *Actinomyces israelii*, competitive mapping

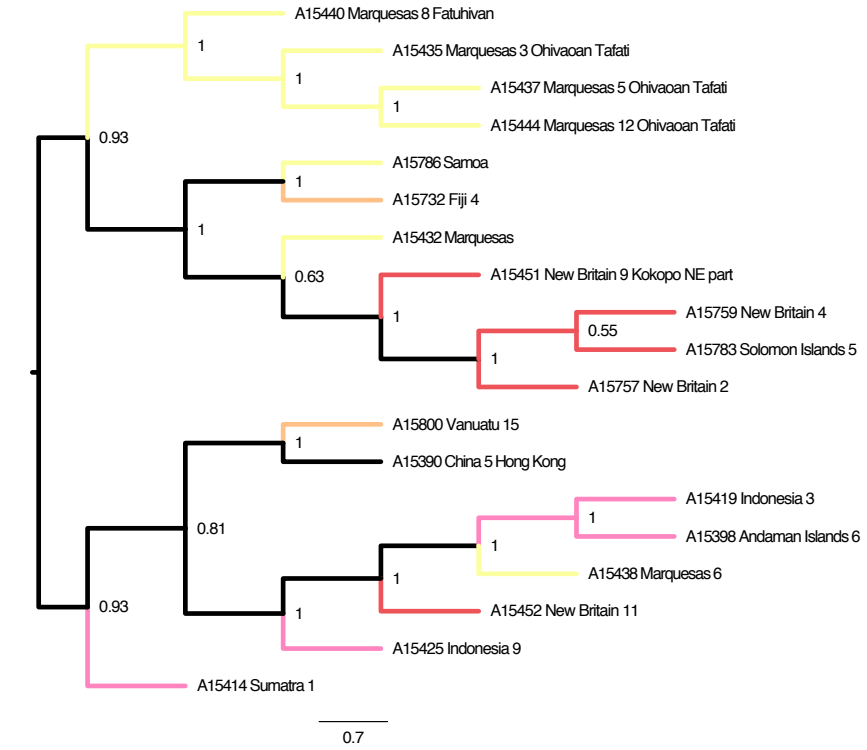


Figure S30. *Fretibacterium fastidiosum*

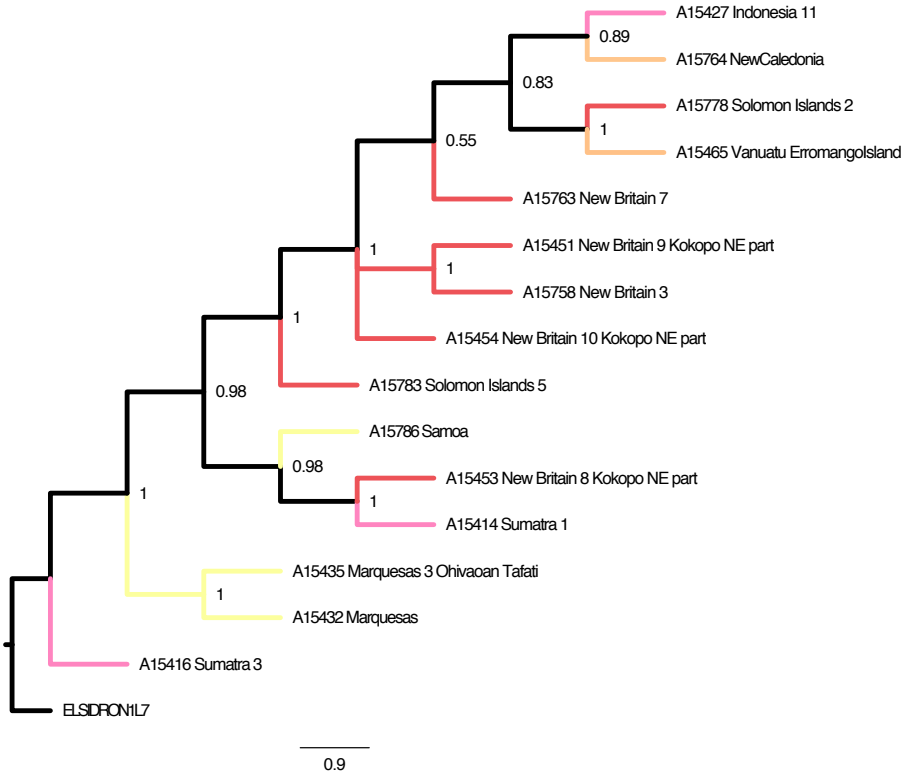


Figure S31. *Methanobrevibacter oralis*

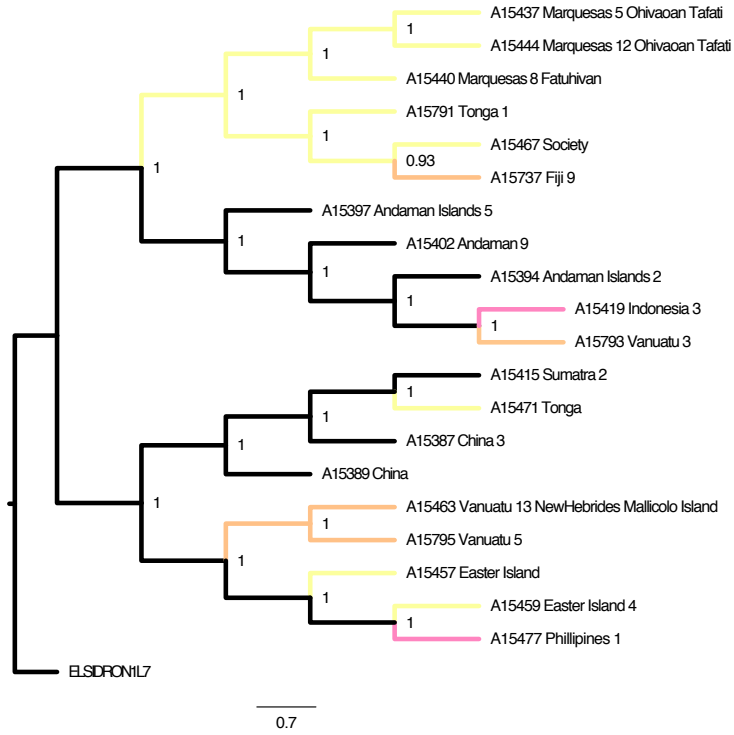


Figure S32. *Olsenella sp. oral taxon 807*

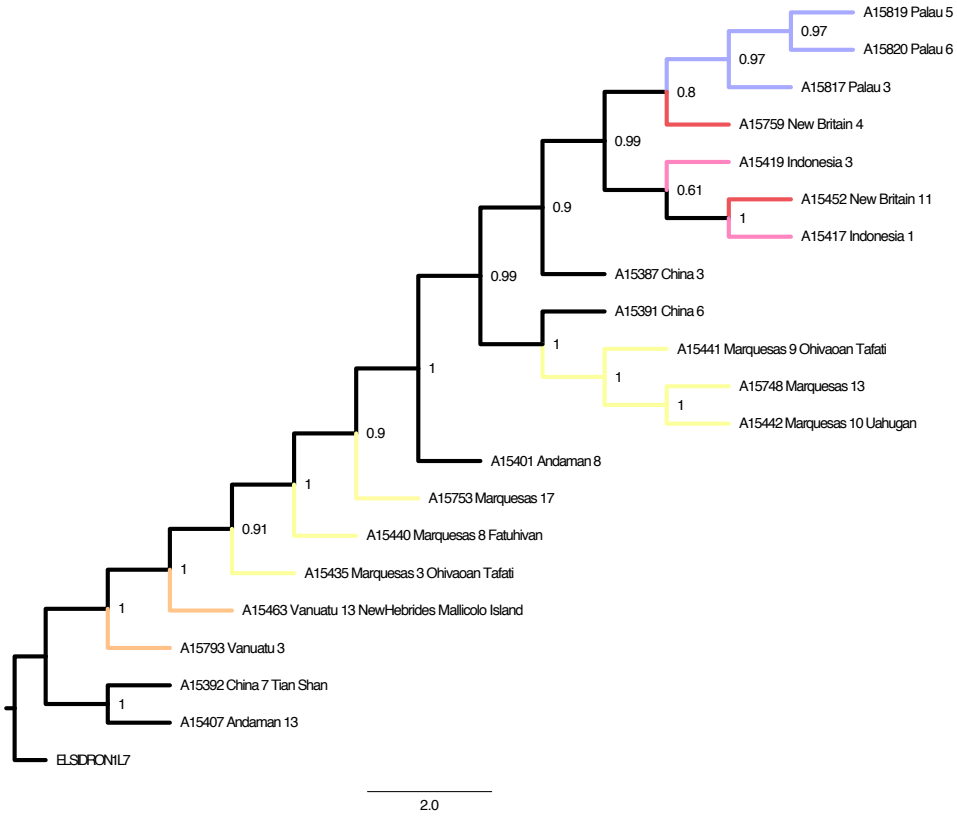


Figure S33. *Pseudopropionibacterium propionicum*

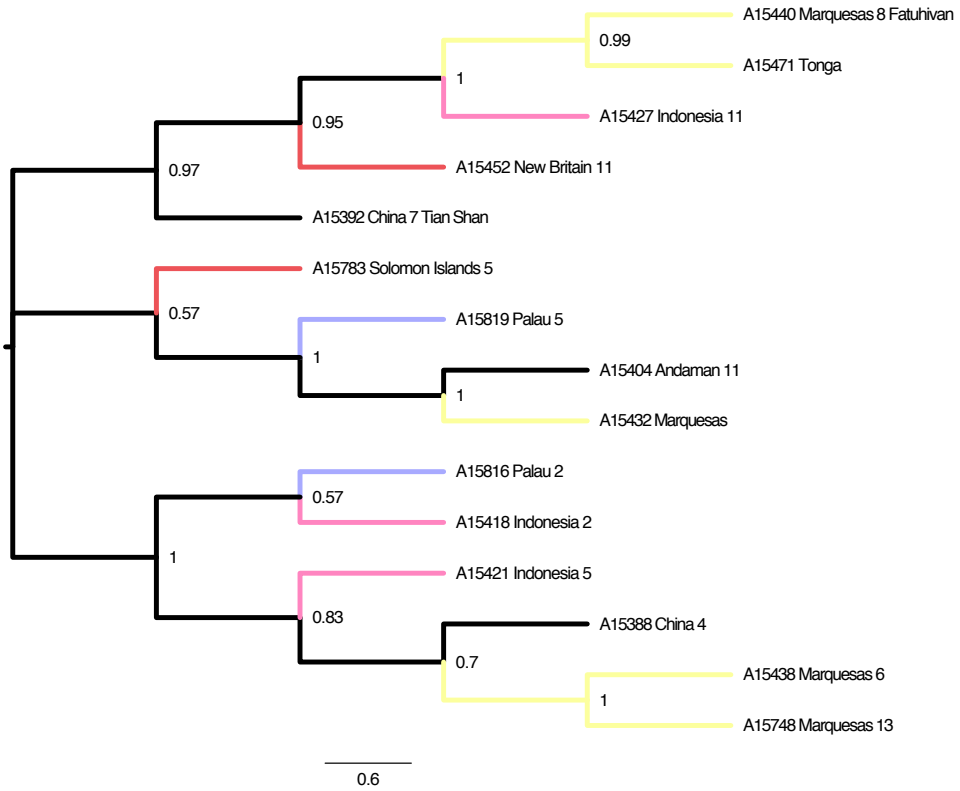


Figure S34. *Tannerella forsythia*

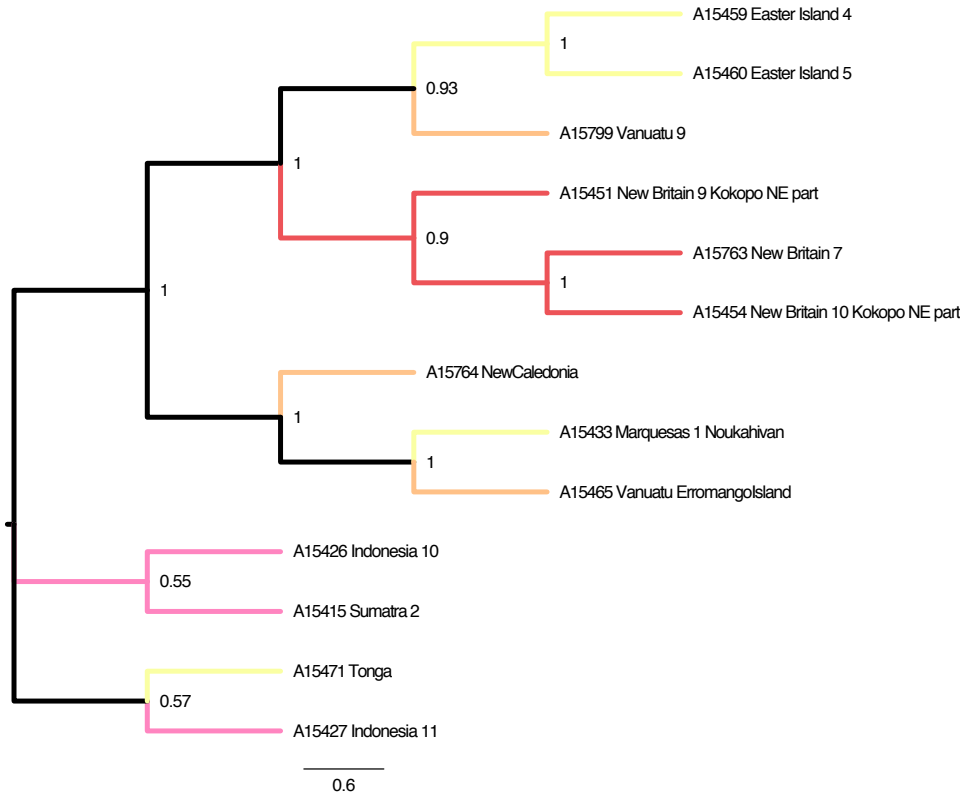


Table S8. Number of RNA probes per species

Species	Clade-specific markers	Single-copy core genes	MLST loci
Actinomyces massiliensis		8	
Bacteriodetes oral taxon 274		8	
Corynebacterium matchurotii		8	
Eubacterium saphenum		8	
Fretibacterium fastidiosum	62	8	
Pseudoramibacter alactolyticus	72	8	
Porphyromonas gingivalis		8	7
Pseudopropionibacterium propionicum	51	8	
Treponema denticola		8	
Tannerella forsythia	53	8	

Table S9. Enrichment statistics

Sample	hits_raw	hits_unique	hits_clonality	fold-enrichment
A15389_China_1_UnEnriched	3,591	3,504	2.4%	
A15389_China_1-Enriched	174,454	56,227	67.8%	16.0
A15402_Andaman_1_UnEnriched	2,332	2,052	12.0%	
A15402_Andaman_1-Enriched	621,877	19,606	96.8%	9.6
A15406_Andaman_2_UnEnriched	3,484	3,379	3.0%	
A15406_Andaman_2-Enriched	389,548	80,586	79.3%	23.8
A15435_Marquesas_1_UnEnriched	3,908	3,722	4.8%	
A15435_Marquesas_1-Enriched	355,873	68,949	80.6%	18.5
A15450_New_Guinea_1_UnEnriched	2,422	2,313	4.5%	
A15450_New_Guinea_1-Enriched	337,581	36,376	89.2%	15.7
A15451_New_Guinea_2_UnEnriched	9,151	8,973	1.9%	
A15451_New_Guinea_2-Enriched	830,465	647,932	22.0%	72.2
A15454_New_Guinea_3_UnEnriched	9,853	9,671	1.8%	
A15454_New_Guinea_3-Enriched	427,667	363,153	15.1%	37.6
A15458_Easter_Island_1_UnEnriched	2,956	2,278	22.9%	
A15458_Easter_Island_1-Enriched	895,407	10,958	98.8%	4.8
A15460_Easter_Island_2_UnEnriched	6,086	5,876	3.5%	
A15460_Easter_Island_2-Enriched	739,296	202,156	72.7%	34.4
A15462_Vanuatu_1_UnEnriched	2,415	1,955	19.0%	
A15462_Vanuatu_1-Enriched	874,348	15,055	98.3%	7.7
A15465_Vanuatu_3_UnEnriched	9,703	9,557	1.5%	
A15465_Vanuatu_3-Enriched	317,235	264,139	16.7%	27.6
A15467_Society_Islands_1_UnEnriched	6,944	6,438	7.3%	
A15467_Society_Islands_1-Enriched	751,524	54,765	92.7%	8.5
A15471_Tonga_2_UnEnriched	6,266	6,173	1.5%	
A15471_Tonga_2-Enriched	652,352	572,205	12.3%	92.7
A15732_Fiji_1_UnEnriched	1,083	797	26.4%	
A15732_Fiji_1-Enriched	62,881	3,168	95.0%	4.0
A15737_Fiji_2_UnEnriched	1,821	1,614	11.4%	
A15737_Fiji_2-Enriched	630,211	19,356	96.9%	12.0
A15753_Marquesas_2_UnEnriched	1,371	1,327	3.2%	

Table S9 continued. Enrichment statistics

Sample	hits_raw	hits_unique	hits_clonality	fold-enrichment
A15753_Marquesas_2-Enriched	179,263	38,583	78.5%	29.1
A15758_New_Britain_1_UnEnriched	6,380	5,711	10.5%	
A15758_New_Britain_1-Enriched	686,017	517,448	24.6%	90.6
A15761_New_Britain_2_UnEnriched	2,580	512	80.2%	
A15761_New_Britain_2-Enriched	532,066	1,274	99.8%	2.5
A15764_New_Caledonia_2_UnEnriched	9,420	8,989	4.6%	
A15764_New_Caledonia_2-Enriched	717,238	103,096	85.6%	11.5
A15765_New_Caledonia_1_UnEnriched	3,625	2,034	43.9%	
A15765_New_Caledonia_1-Enriched	936,969	6,867	99.3%	3.4
A15778_Solomon_Islands_1_UnEnriched	2,124	1,987	6.5%	
A15778_Solomon_Islands_1-Enriched	327,157	21,826	93.3%	11.0
A15783_Solomon_Islands_2_UnEnriched	5,401	5,298	1.9%	
A15783_Solomon_Islands_2-Enriched	379,184	220,159	41.9%	41.6
A15786_Samoa_1_UnEnriched	3,537	3,450	2.5%	
A15786_Samoa_1-Enriched	350,833	71,529	79.6%	20.7
A15791_Tonga_1_UnEnriched	2,369	2,328	1.7%	
A15791_Tonga_1-Enriched	205,312	161,396	21.4%	69.3
A15793_Vanuatu_2_UnEnriched	3,381	3,232	4.4%	
A15793_Vanuatu_2-Enriched	557,238	59,674	89.3%	18.5
A15811_English_Oakington_1_UnEnriched	5,103	4,311	15.5%	
A15811_English_Oakington_1-Enriched	274,447	30,108	89.0%	7.0
A15817_Palau_1_UnEnriched	1,142	1,120	1.9%	
A15817_Palau_1-Enriched	125,789	85,206	32.3%	76.1
A15818_Palau_2_UnEnriched	1,381	1,351	2.2%	
A15818_Palau_2-Enriched	220,627	88,097	60.1%	65.2
A15856_Teouma_1_UnEnriched	267	208	22.1%	
A15856_Teouma_1-Enriched	442,697	1,691	99.6%	8.1
A15858_Teouma_2_UnEnriched	495	245	50.5%	
A15858_Teouma_2-Enriched	175,503	567	99.7%	2.3
A17839_EBC_1_10_UnEnriched	0	0	0.0%	
A17839_EBC_1_10-Enriched	1	1	0.0%	0.0
A8338_English_Raunds_2_UnEnriched	7,941	7,832	1.4%	
A8338_English_Raunds_2-Enriched	559,264	157,267	71.9%	20.1
Average UnEnriched	4,146	3,814	12.2%	
Standard Deviation	2,807	2,842	17.3%	
Average Enriched	475,172	128,368	71.0%	27.8
Standard Deviation	245,718	170,674	30.2%	26.8

Table S10. Enriched loci passing criteria

Sample	Number of loci enriched	Number of loci passing filtering
A15389_China	325	161
A15402_Andaman	325	76
A15406_Andaman	325	155
A15435_Marquesas	325	188
A15450_New_Guinea	325	170
A15451_New_Guinea	325	153
A15454_New_Guinea	325	193
A15458_Easter_Island	325	76
A15460_Easter_Island	325	194
A15462_Vanuatu	325	71
A15465_Vanuatu	325	143
A15467_Society_Islands	325	87
A15471_Tonga	325	246
A15732_Fiji	325	0
A15737_Fiji	325	50
A15753_Marquesas	325	109
A15758_New_Britain	325	233
A15761_New_Britain	325	0
A15764_New_Caledonia	325	154
A15765_New_Caledonia	325	65
A15778_Solomon_Islands	325	55
A15783_Solomon_Islands	325	191
A15786_Samoa	325	133
A15791_Tonga	325	247
A15793_Vanuatu	325	80
A15811_English_Oakington	325	101
A15817_Palau	325	148
A15818_Palau	325	220
A15856_Teouma	325	0
A15858_Teouma	325	0
A17839_EBC_1_10	325	0
A8338_English_Raunds	325	224
AVERAGE	325	127
S.D.		75

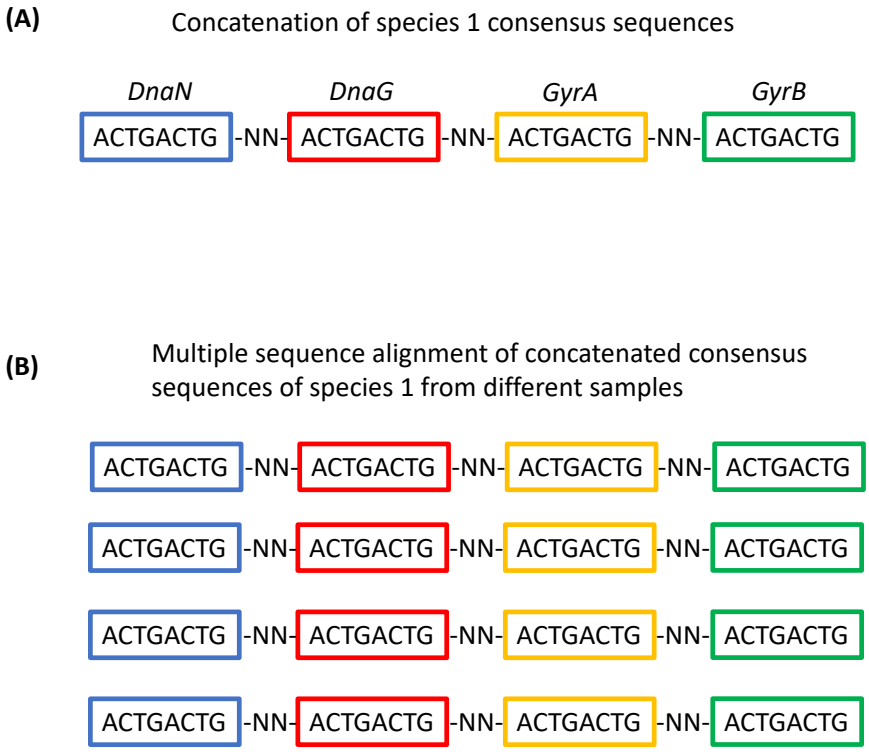


Figure S35. Illustration of consensus sequence concatenation (A) and multiple sequence alignment (B)

Figure S36. *Fretibacterium fastidiosum* clade-specific markers

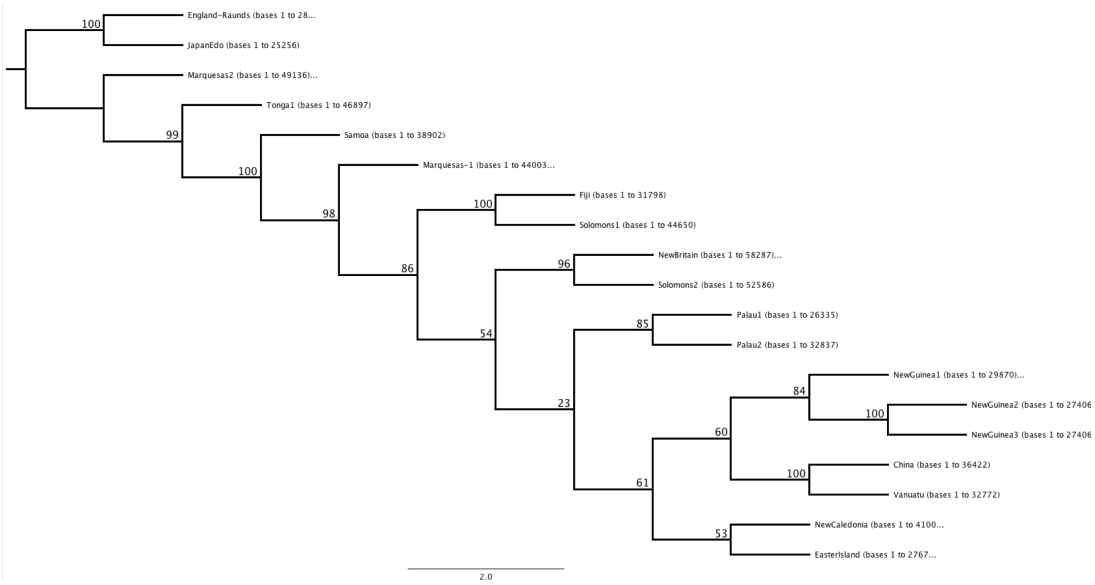


Figure S37. *Pseudoramibacter alactolyticus* clade-specific markers

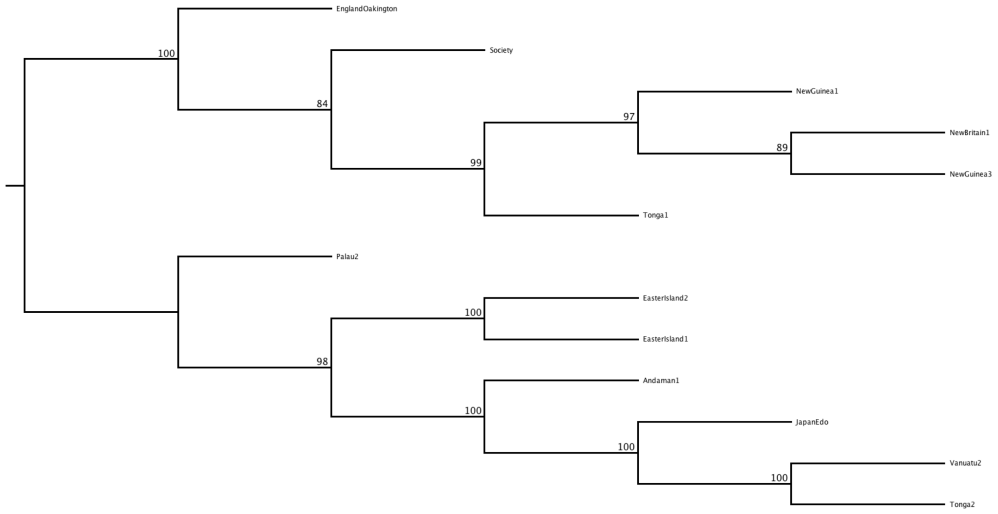


Figure S38. *Pseudopropionibacterium propionicum* clade-specific markers

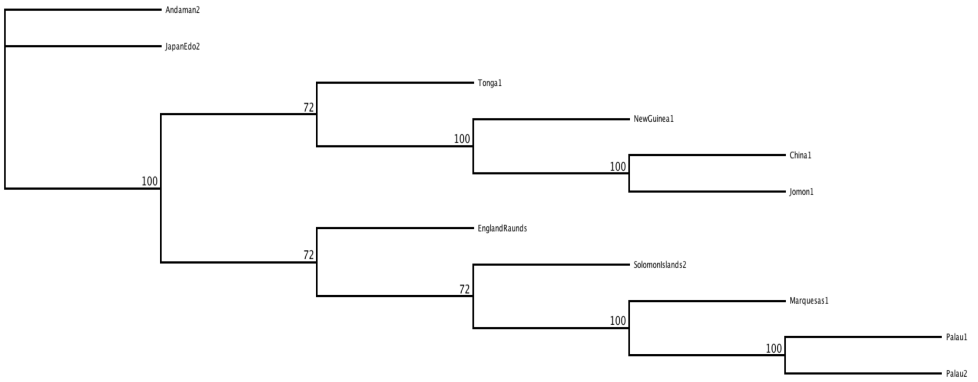


Figure S39. *Tannerella forsythia* clade-specific markers

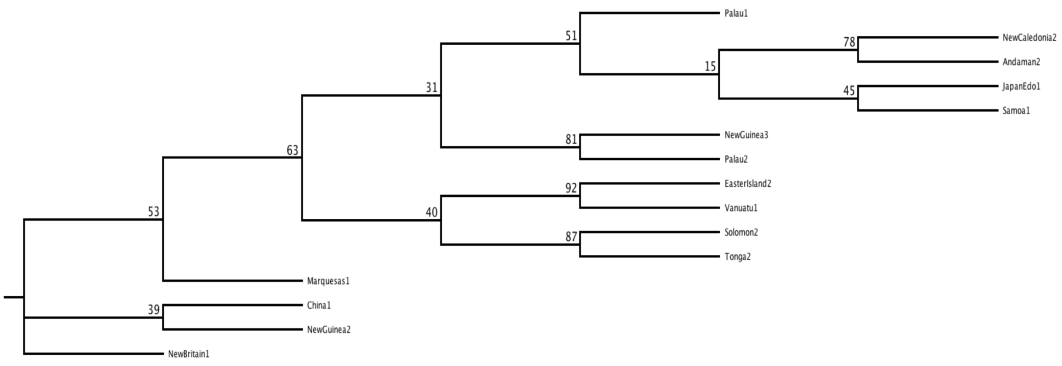


Figure S40. *Porphyromonas gingivalis* MLST loci

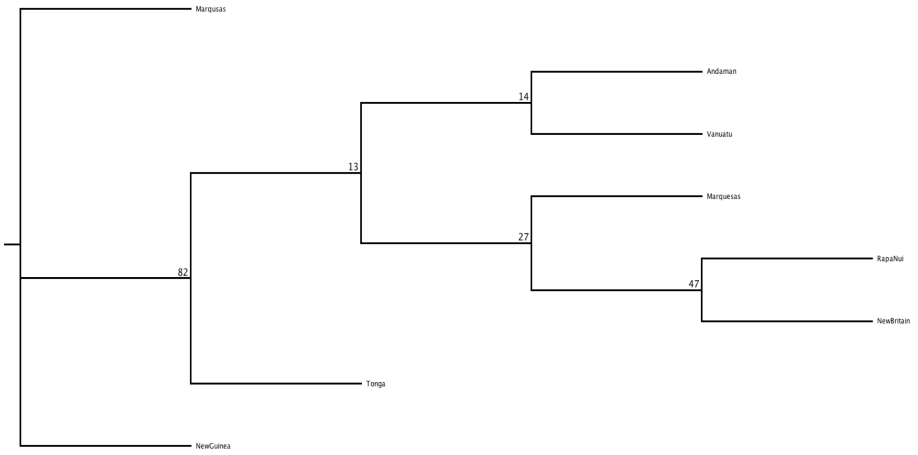


Figure S41. *Bacteriodetes sp oral taxon 274* single-copy core genes

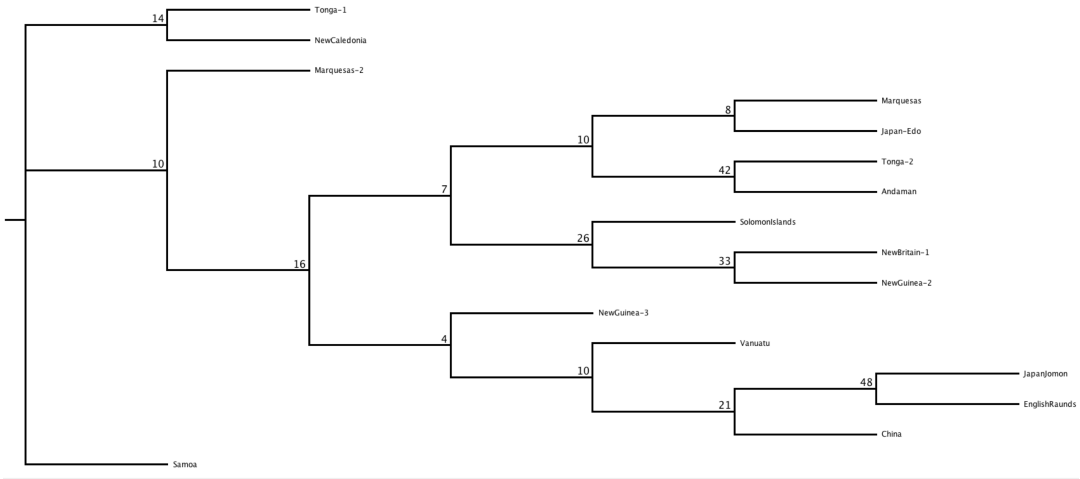


Figure S42. *Eubacterium saphenum* single-copy core genes

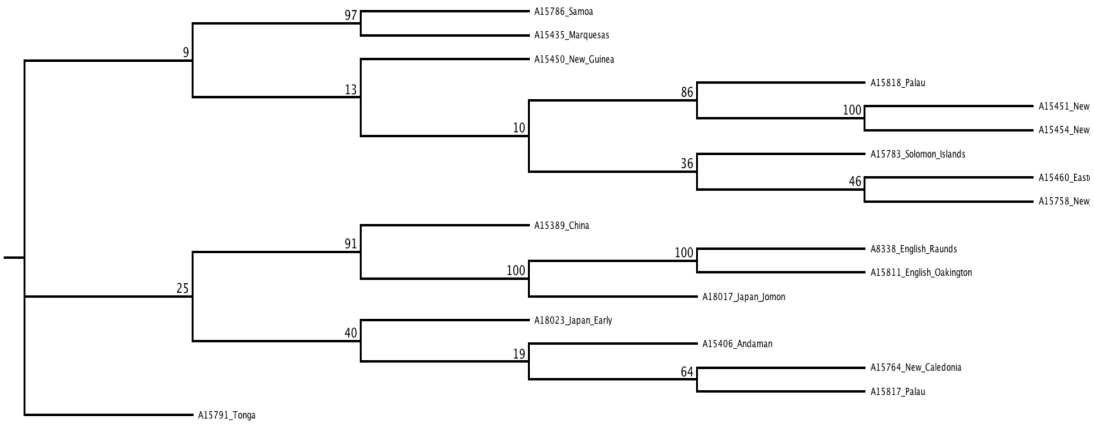


Figure S43. *Fretibacterium fastidiosum* single-copy core genes

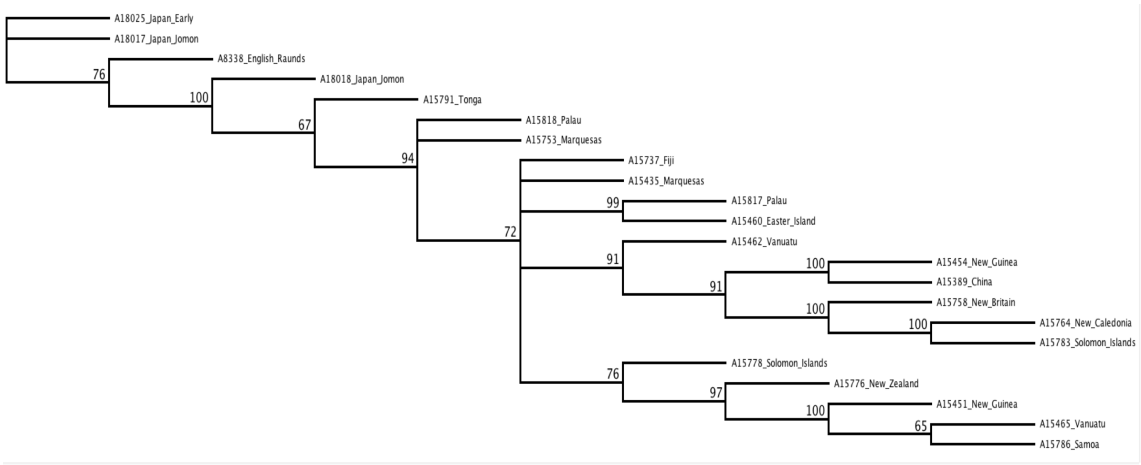


Figure S44. *Psuedoramibacter alactolyticus* single-copy core genes

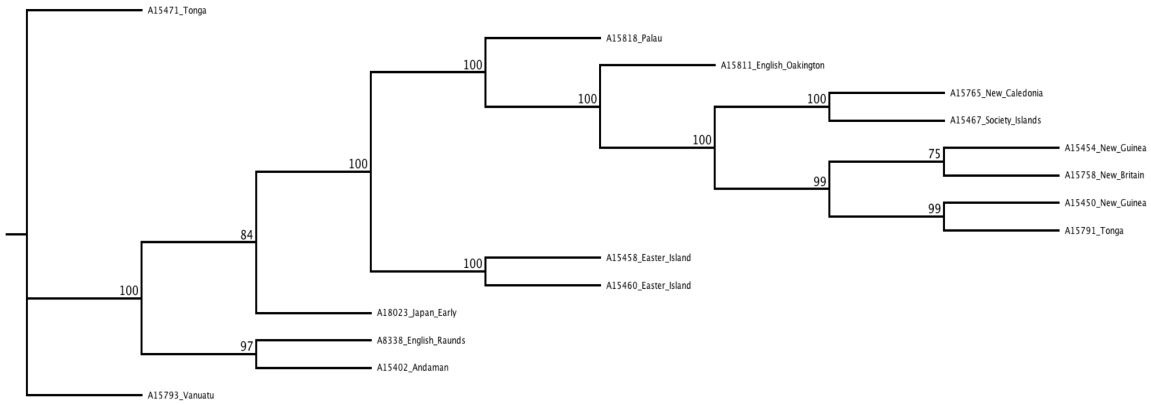


Figure S45. *Porphyromonas gingivalis* single-copy core genes

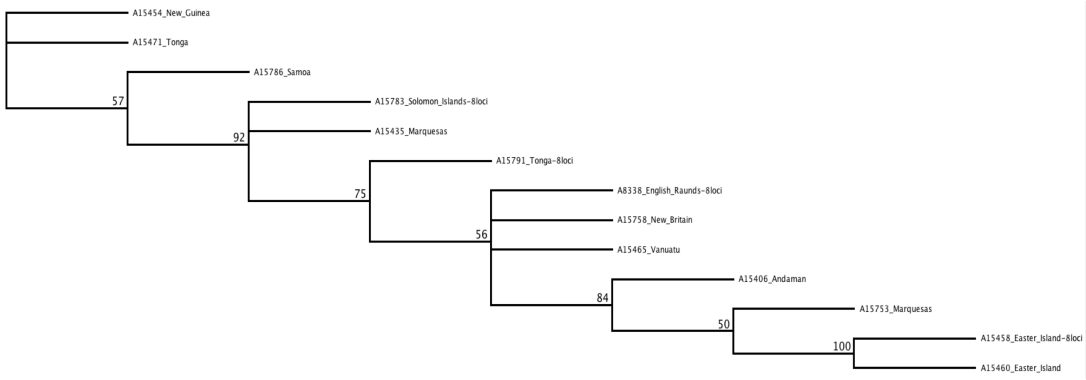


Figure S46. *Pseudopropionibacterium propionicum* single-copy core genes

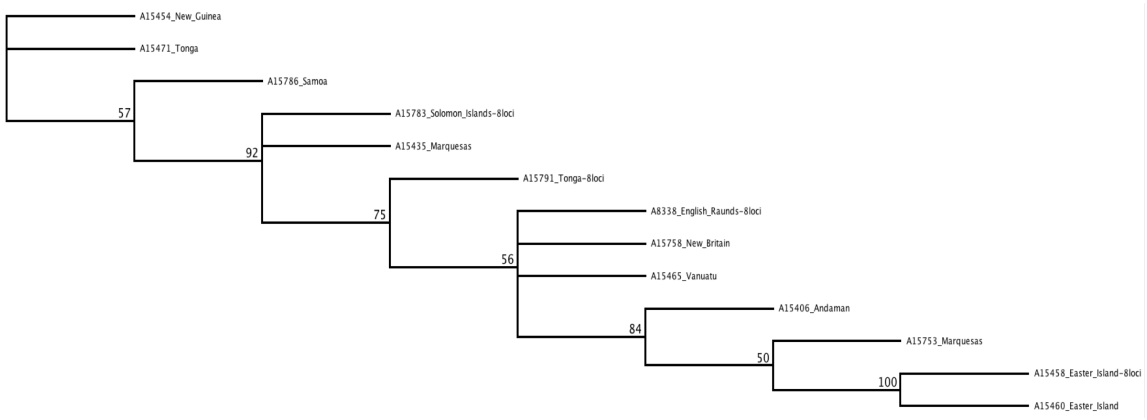


Figure S47. *Treponema denticola* single-copy core genes

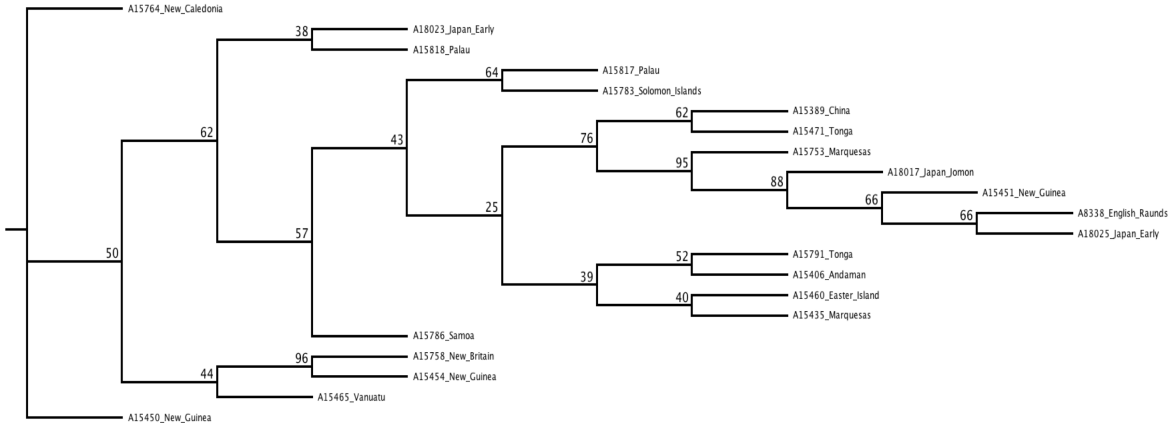
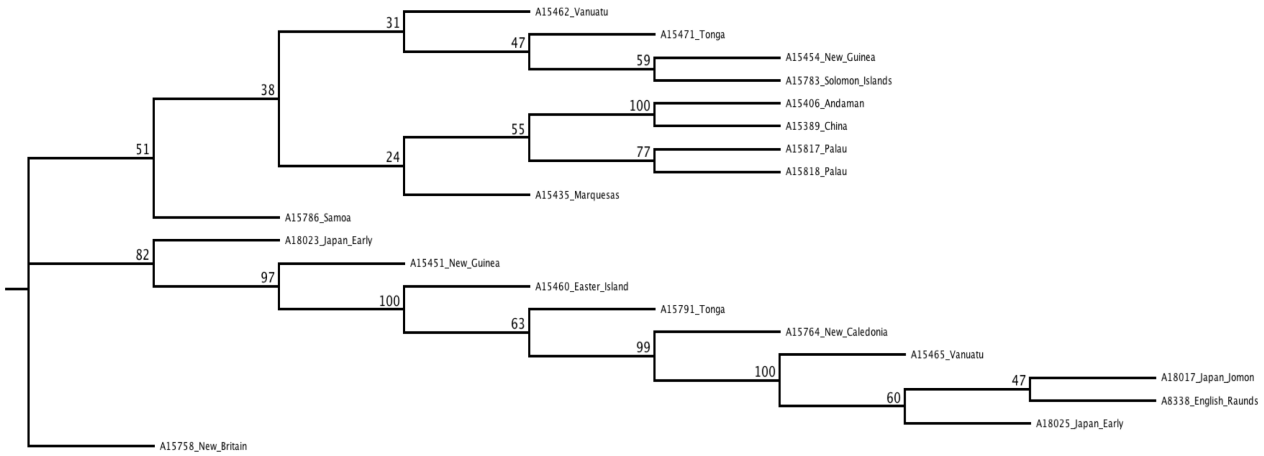


Figure S48. *Tannerella forsythia* single-copy core genes



Chapter VI

—

Insights into the demographic history of Japan using ancient oral microbiota

Statement of Authorship

Title of Paper	Insights into the demographic history of Japan using ancient oral microbiota
Publication Status	Unpublished work written in manuscript style.
Publication Details	Unpublished work written in manuscript style.

Principal Author

Name of Principal Author (Candidate)	Raphael Eisenhofer
Contribution to the Paper	Conceived the experiments. Performed all laboratory work and processed the data. Analysed the data and interpreted the results. 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 10/5/18

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	Laura S. Weyrich
Contribution to the Paper	collected funding, samples, metadata, and designed study, editted manuscript
Signature	Date 9/5/18

Name of Co-Author	Ken-ichi Shinoda		
Contribution to the Paper	I provided dental calculus of archaeological samples for this research, and input about the archaeological and anthropological context to this paper.		
Signature		Date	27/4/2018

Name of Co-Author	Hideaki Kanzawa-Kiriyama		
Contribution to the Paper	I provided dental calculus of archaeological samples for this research, and input about the archaeological and anthropological context to this paper.		
Signature		Date	27/4/2018

Insights into the demographic history of Japan using ancient oral microbiota

Authors: Raphael Eisenhofer[1], Hideaki Kanzawa-Kiriyama[2,3], Ken-ichi Shinoda[4], and Laura S. Weyrich[1]

Affiliations:

1: Australian Centre for Ancient DNA, University of Adelaide, Australia

2: Department of Genetics, School of Life Science, SOKENDAI (Graduate University for Advanced Studies), Mishima, Japan

3: Division of Population Genetics, National Institute of Genetics, Mishima, Japan

4: Department of Anthropology, National Museum of Nature and Science, Tsukuba, Japan

Abstract

Communities of microorganisms have been coevolving with humans throughout history. Studying how these communities of microorganisms that inhabit humans (human microbiota) have changed through time could lead to new medically-relevant insights, and expand our understanding of history. Recently, research has demonstrated that it is possible to study past human microbiota through the analysis of ancient microbial DNA preserved in calcified dental plaque (calculus). Here, using ancient dental calculus samples from two major cultural time periods in Japan: the Jomon period hunter-gatherers from ~3,000 BP (years before present), and the Edo period agriculturalists 400-150 BP, we investigate how human oral microbiota have changed in Japan through time. We also explore the presence and of oral diseases (periodontal disease, dental caries) in ancient Japan, and examine how these diseases influence community-level microbiota analyses. Finally, we perform phylogenetic analyses of ancient bacterial genomes and find support for bacterial lineage replacement potentially due to past human population replacement by migrants from mainland Asian populations. This research represents the first study of ancient oral microbiota from Japan and illustrates that the analysis of ancient bacterial genomes preserved in dental calculus can be used to learn about past human demographic events.

Introduction

Communities of microorganisms (microbiota) inhabit the human body [1] and encode functions that influence the development, physiology, behaviour, and the health of their hosts, collectively referred to as the human microbiome [2–9]. Disruptions to the human microbiome (dysbiosis) can compromise the health of the host [10,11]. Dysbiosis can occur due to a range of factors, such as the use of antibiotics [12,13], changes in diet [14], infection by pathogens [15,16], and the adoption of lifestyles associated with Industrialization [17]. Evidence suggests that many members of these microbial communities can be vertically inherited [18–21] and some have been co-evolving with humans over deep evolutionary time [22,23].

Different human populations have been observed to possess distinct microbiomes resulting from specific diets, exposure to unique environments, pathogens, and specific lifestyle traits [17,24–26]. However, little is known about how the human microbiome adapts and evolves when two cultures with distinct microbiomes meet and mix (*e.g.* when Europeans first met the peoples of the Americas). Such cultural admixtures could disrupt long-term evolutionary relationships between microbiota and host and potentially contribute to dysbiosis, and the resulting changes in lifestyle and diet could select for unique microbial communities [27]. Additionally, microbial replacement due to cultural admixture could also shape the microbiome in unique ways; for example, 'signatures' of past human interaction and population replacement (*e.g.* loss of particular species or strains) could be used to learn more about the demographic history of past human populations [28].

Recently, ancient human calcified dental plaque (calculus) has been identified as a robust source of ancient human-associated microbial DNA [29–31]. Dental calculus is a microbial biofilm that grows on teeth and undergoes periodic mineralisation events that locks oral microorganisms in place within a robust calcium phosphate matrix [32]. The direct association of dental calculus on human teeth, coupled with its robust nature, provide an unprecedented opportunity to examine the bioarchaeological record of past human oral microbiome, allowing researchers to identify past factors that have altered the oral microbiome through time [29–31]. For example, dental calculus research revealed that large shifts in the European microbiome were concordant with large-scale dietary and lifestyle changes (from hunting-gathering to an agricultural lifestyle) [29]. Dental calculus is, therefore, a tool that can be used to sample the oral microbiome of past human populations and explore how the microbiome adapts and evolves following major cultural and demographic shifts.

Ancient Japan is one such area where large-scale demographic changes occurred in the recent past. The Japanese Archipelago was largely inhabited by the Jomon culture from ~16,000 to 2,500 years before present (BP) [33,34]. Archaeological evidence suggests that Jomon

hunter-gatherers relied on both terrestrial and marine resources, including nuts, deer, boar, marine fish, and shellfish [35]. Carbon isotope ratios of human teeth also suggest that C3 plants and terrestrial mammals were major dietary resources for the Jomon people [36]. Agriculture-bearing migrants from continental Asia came to the Japanese Archipelago and admixed with Jomon during the early Yayoi period around 2,500 BP [37–39]. Both modern and ancient DNA studies suggest that the admixture was weighted towards migrants, with modern estimates of Jomon contribution to mainland Japanese populations being less than 20% [39]. Prior to this admixture, the mitochondrial divergence estimates suggesting that over 22,000 years of separation existed between the Jomon and continental Asian populations [40], which, coupled with their putatively disparate lifestyles (hunter-gatherer *vs.* agriculturalist), may have resulted in divergent co-evolution of their microbiome. This past demographic scenario provides an ideal testbed to measure the potential impacts of population admixtures on the human oral microbiome using ancient dental calculus. Here, we examine bacterial DNA preserved within ancient dental calculus from the Jomon (~3,000 BP) and Edo periods (400-150 BP) in Japan to examine the evolutionary history and the impacts of cultural admixture on the oral microbiome.

Methods

Ancient dental calculus samples

Ethics approval for this study was obtained from the University of Adelaide Human Research Ethics Committee (H-2012-108). Ancient dental calculus samples (5=Jomon, ~3,000 BP [41]), (10=Edo, 400-150 BP [42]) were collected from the Natural Museum of Nature and Science in Tsukuba, Ibaraki, Japan. Dental calculus was removed from specimens as previously described [43]. Briefly, a sterile dental pick was used to carefully remove dental calculus from one side of one tooth, and the specimen was placed in a sterile plastic bag for transport at room temperature to the Australian Centre for Ancient DNA at the University of Adelaide. Accompanying metadata was also collected at this time (Table S1).

DNA extraction and library preparation

All sample processing and molecular biology procedures prior to PCR amplification were carried out at the Australian Centre for Ancient DNA facility at the University of Adelaide. Experiments were performed within a specialised ancient DNA laboratory, which includes, positive air pressure, UV-treatment, regular 3% bleach cleanings, and still-air hoods located in isolated, still-air rooms to limit the introduction of modern contaminant DNA. All technicians entered the facility using a dedicated entry room and wore full-body clean suits, gloves, and

face-masks. Dental calculus samples were decontaminated to minimize environmental contamination by UV-irradiation for 15 minutes on each side, following by soaking in 2 ml of 5% sodium hypochlorite for 3 minutes, rinsing in 90% ethanol for 1 minute, and drying at room temperature for 2 minutes. Immediately post-decontamination, dental calculus samples were crushed on the side of plastic tubes with sterile tweezers, and DNA was extracted using an in-house silica-based method described previously [44].

Shotgun metagenomic libraries were constructed as previously described [2], using unique combinations of 7-bp forward and reverse barcodes. Thirteen cycles of PCR were used for the first amplification with P5/P7 barcoded adapters, followed by an additional 13 cycles for the addition of GAII-index and sequencing primers. Metagenomic shotgun libraries were cleaned using Ampure XP, quantified using an Agilent TapeStation, and pooled at equimolar concentrations prior to sequencing on the Illumina NextSeq platform (2 x 150 bp).

Data used from other previously published studied

Seventeen randomly selected modern dental plaque samples from the Human Microbiome Project [1] were downloaded (SRS076926, SRS078431, SRS058730, SRS077104, SRS075353, SRS023964, SRS024087, SRS063485, SRS024021, SRS077861, SRS078677, SRS074682, SRS058261, SRS075090, SRS075959, SRS077520, SRS078738). The paired reads (R1, R2) were concatenated into a single FASTQ file, then randomly subsampled to a depth of 1,500,000 sequences using SeqTK https://github.com/lh3/seqtk. Raw DNA sequences from ancient Chinese dental calculus data (Eisenhofer *et al.* Chapter VI) and the modern and ancient dental calculus sample data (https://www.oagr.org.au/experiment/view/65/) were obtained from previous studies [31].

Data processing and taxonomic composition analyses

The resulting data converted into FASTQ format using Illumina's bcl2fastq software, before being trimmed and demultiplexed using AdapterRemoval 2 based on unique P5/P7 barcodes [46]. Taxonomic composition was determined using MEGAN Alignment Tool (MALT) v 0.3.8 [47], whereby DNA reads from samples were aligned against a database created in-house that contains 47,696 archaeal and bacterial genome assemblies from the NCBI Assembly database [48]. The resulting blast-text files were converted into RMA files via the blast2rma script included in the program MEGAN v 6.11.1 [49], with the following Last Common Ancestor (LCA) parameters: Weighted-LCA=80%, minimum bitscore=42, minimum E-value=0.01, minimum support percent=0.1.

Samples were assessed for ancient DNA authenticity by comparison to extraction blank controls and by estimation of cytosine deamination using MapDamage [50]. Subtractive filtering was used to remove species found in the extraction blank controls from ancient dental calculus samples. The filtered samples were then normalised to equal sequencing depth in MEGAN before further analyses. Filtered, species-level taxonomic composition was exported from MEGAN into STAMP [51] for statistical analyses. For analysis in QIIME [52], filtered, species-level taxonomic composition was exported from MEGAN into BIOM format, and imported into QIIME 1.9.1. ANOSIM was used to test for statistical significance in composition between groups using the compare_categories.py script with 999 permutations. Differential abundance of species between groups was tested using the Kruskal-Wallis test with Bonferroni-correction in the group_significance.py script.

Whole-genome phylogenetic analysis

Genomic sequences were assembled by mapping to reference genomes using PALEOMIX [53] with the BWA-MEM aligner [54]. The resulting BAM files were imported into Geneious v. 10.2.3 [55], and consensus sequences called with the following parameters: call N if minimum depth <3, and 75% consensus threshold. Consensus sequences were then aligned using the Mauve Genome aligner with default settings [56]. Gblocks [57] with default settings was used to clean alignments. Phylogenetic reconstruction of transition and transversion substitutions was performed using RAxML [58], with the GTR-GAMMA substitution model and 1,000 bootstrap replicates. Phylogenetic reconstruction of only transversion substitutions was performed in MEGA v7.0 [59] using the Neighbour Joining method with evolutionary distances computed using Maximum Composite Likelihood [60] and 1,000 bootstrap replicates. The rate variation among sites was modelled with a gamma distribution (shape parameter = 1), and the differences in the composition bias among sequences were considered in evolutionary comparisons [61]. All positions with less than 75% site coverage were eliminated. That is, fewer than 25% alignment gaps, missing data, and ambiguous bases were allowed at any position.

Results

Authentic ancient microbial DNA was isolated from dental calculus

We applied metagenomic shotgun sequencing to 15 ancient Japanese dental calculus samples: 5 male Jomon period (~3000 BP) and 10 (5 male; 5 female) from the Edo period 400-150 BP). An average of 1,552,410 reads per sample was obtained (±716,139), with the fragment length

distributions being as expected for ancient DNA (average size 82 bp; Table S2). We used MALT (MEGAN Alignment Tool) to align DNA reads to a reference database containing 47,696 archaeal and bacterial genome assemblies, and as expected for ancient dental calculus studies [48], an average of 49.8% (±10.1%) of DNA reads in each sample could be assigned taxonomy. The ancient Japanese calculus samples looked similar to previously published ancient calculus samples (Figure 1) and are clearly distinct from extraction blank controls (EBCs) that were processed at the same time as the samples (Figure 1). Phyla present in the ancient calculus samples but not modern plaque samples from the Human Microbiome Project (HMP) include Synergistetes, Chloroflexi, Candidatus Saccharibacteria, and Euryarchaeota (Figure 1). These phyla contain species that are associated with periodontal disease and were identified in the ancient calculus samples: (Synergistetes; *Fretibacterium fastidiosum* [62], Chloroflexi; *Anaerolineaceae sp. oral taxon 439* [63], Candidatus Saccharibacteria; *TM7x* [64], and Euryarchaeota; *Methanobrevibacter oralis* [65]). Therefore, the absence of these phyla from the modern plaque samples could be explained by disease-state, as all HMP samples were taken from healthy individuals [1].

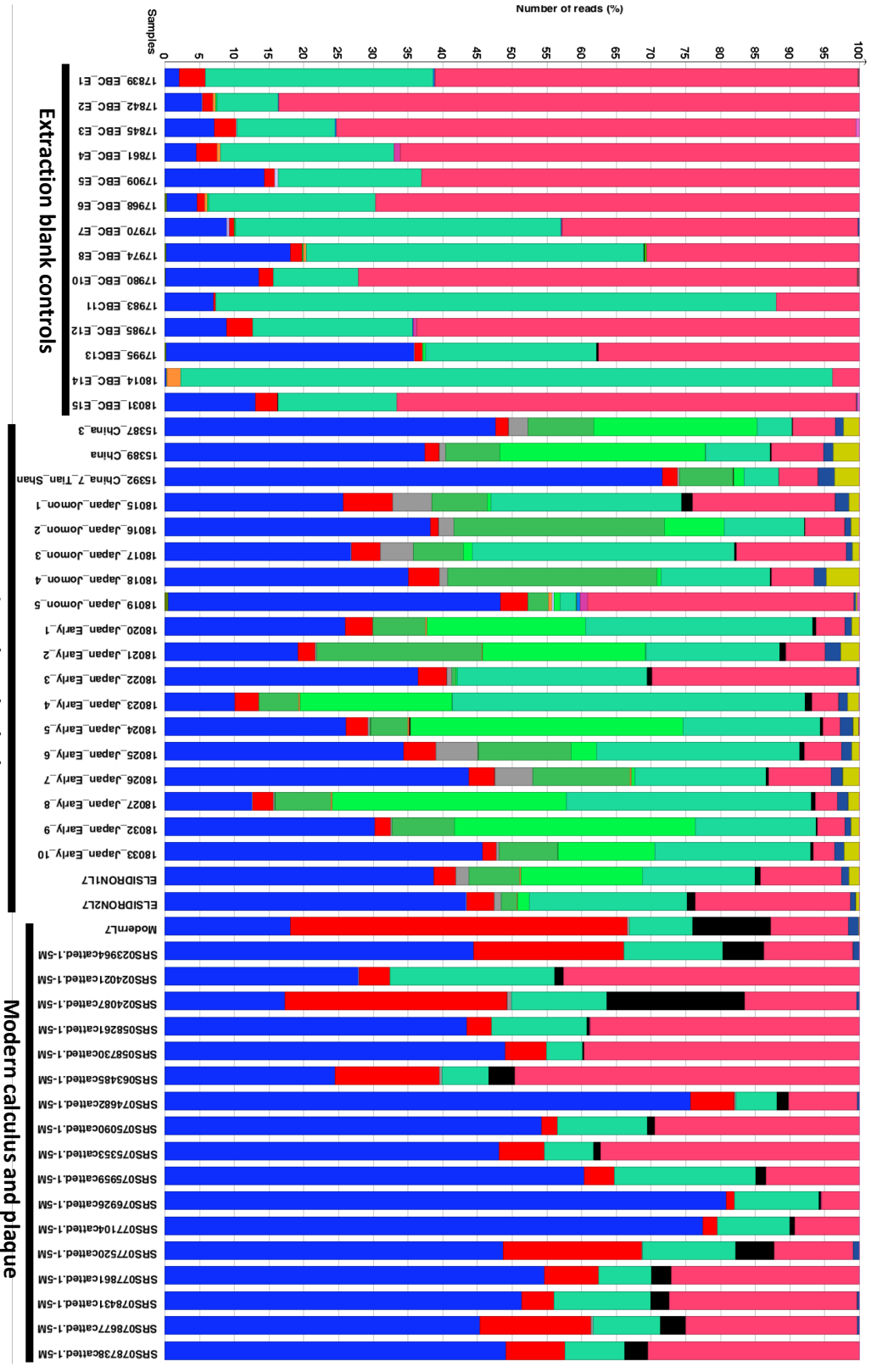


Figure 1. Phylum-level taxonomic composition of extraction blank controls and unfiltered samples

As background DNA contamination can influence ancient microbiome studies [66,67], we next assessed oral and contaminant DNA levels in the samples by ordinating Bray Curtis dissimilarity in a Principal Coordinates Analysis (PCoA) (Figure 2), which included EBCs, ancient Japanese and Chinese calculus samples, previously published ancient calculus specimens [31], and modern healthy plaque samples from the HMP [1]. Ancient Japanese calculus specimens clustered with published ancient calculus specimens and were dissimilar to EBCs, as expected (Figure 2). Except for one Edo calculus specimen, ancient Japanese samples were different from modern plaque samples from the HMP (Figure 2). Lastly, we took a conservative approach and removed any species found in the EBCs from the Japanese calculus samples to help eliminate the contributions of contaminant DNA [68]; an average of 94.9% (±7.1%) reads remained at the species level after filtering (Table S2), highlighting the great preservation of the specimens. To our knowledge, this is the oldest human-associated oral microbial DNA obtained from Asia to date (3,000 years old for the Jomon period calculus).

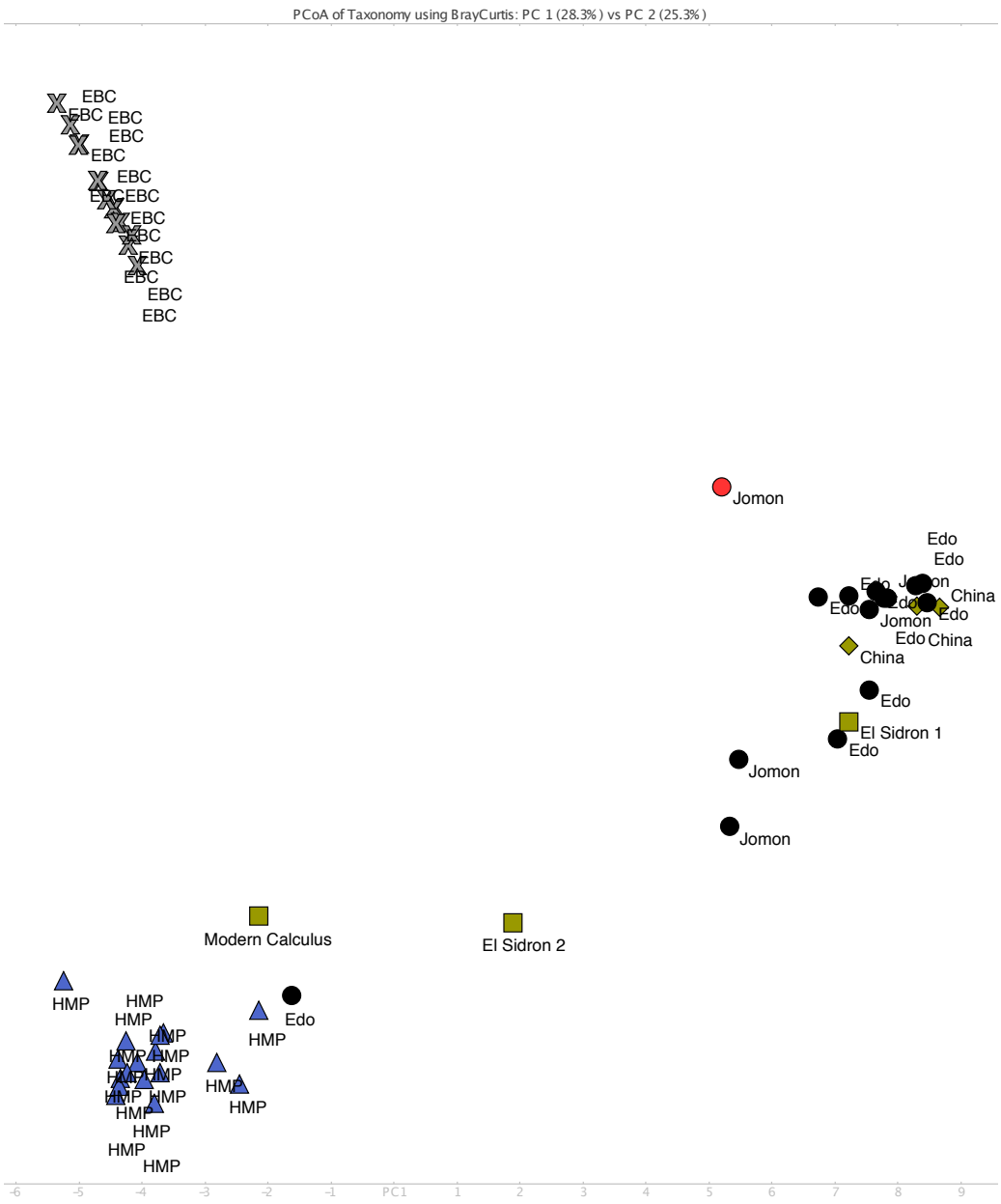


Figure 2. PCoA ordination of species-level Bray-Curtis distances. Extraction blank controls (EBCs, grey crosses), cluster separately from the rest of the ancient and modern oral samples. Ancient Japanese dental calculus samples (black circles) cluster with ancient dental calculus samples from China (orange diamonds) and the El Sidron 1 Neanderthal (orange square). Modern dental plaque samples from the Human Microbiome Project (blue triangles), and a modern dental calculus sample (orange square) cluster separately from ancient dental calculus samples, with the exception of one Edo period Japanese sample.

Comparing the oral microbiomes of Jomon and Edo period Japan

Species present in ancient Japanese samples after filtering by EBCs were predominantly oral (Figure 3), being previously identified in other oral microbiome studies [1,69] and having

entries in the Human Oral Microbiome Database (HOMD) [70]. We observed inter-individual variation in microbiome composition within groups (Jomon, Edo males, Edo females), especially for abundances of taxa (Figure 3). Such inter-individual variation has been observed previously in modern plaque microbiome studies [71,72].

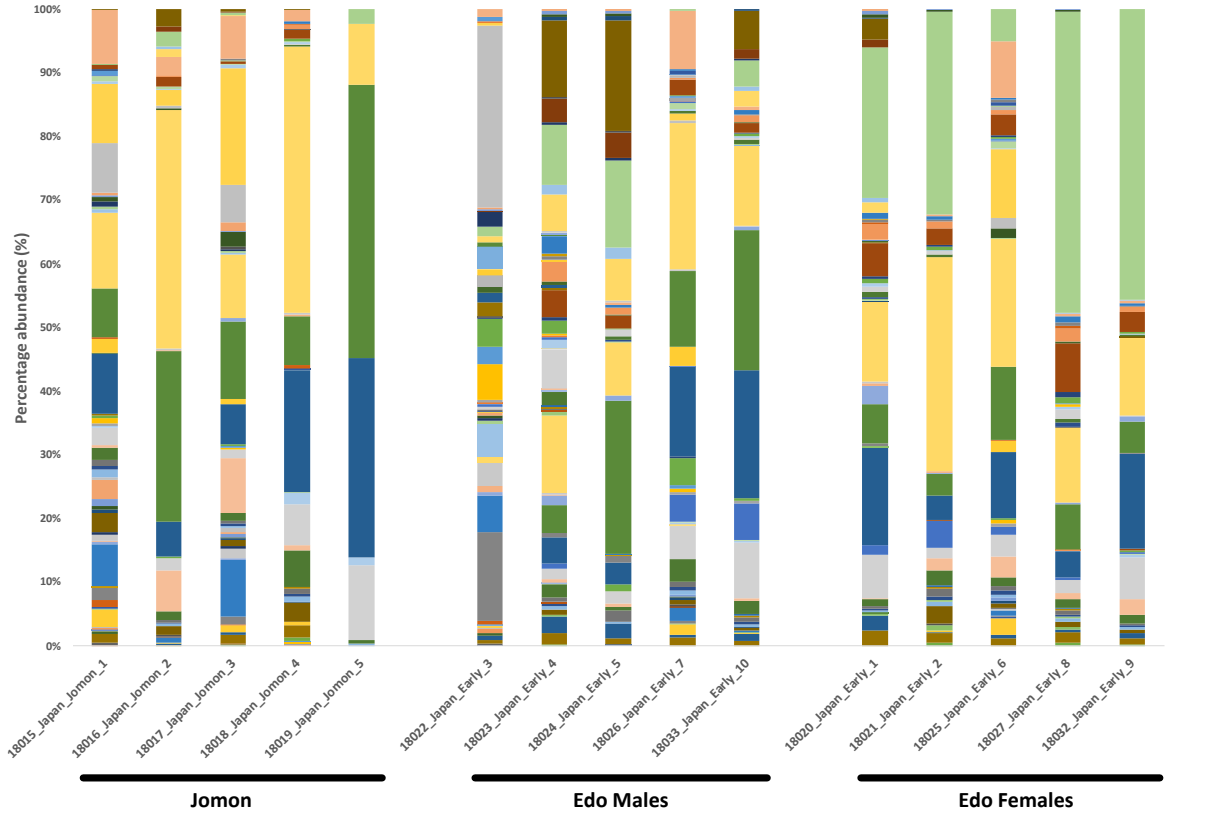


Figure 3. Species-level taxonomic composition of EBC-filtered ancient Japanese calculus samples

Endomicrobium proavitum
Proteiniphilum saccharofermentans
Porphyromonas sp. KLE 1280
Tannerella forsythia
Capnocytophaga ochracea
Bacteroidetes oral taxon 274
Leptotrichia buccalis
Brachymonas chironomi
Eikenella corrodens
Neisseria meningitidis
Neisseria sp. oral taxon 014
Desulfomicrobium orale
Cardiobacterium valvarum
Haemophilus parainfluenzae
Treponema maltophilum
Aminomonas paucivorans
Jonquetella anthropi
Actinomyces georgiae
Actinomyces israelii
Actinomyces naeslundii
Actinomyces radicidentis
Actinomyces sp. oral taxon 170
Actinomyces sp. oral taxon 877
Corynebacterium matruchotii
Propionibacterium freudenreichii
Olsenella sp. kh2p3
Olsenella uli
Anaerolineaceae bacterium oral taxon 439
Abiotrophia sp. HMSC24B09
Streptococcus constellatus
Streptococcus gordonii
Streptococcus sp. DD04
Clostridium sp. ATCC BAA-442
Geosporobacter ferrireducens
Anaerofustis stercorihominis
Blautia producta
Lachnoanaerobaculum saburreum
Shuttleworthia satelles
Peptoanaerobacter stomatis
Ruminiclostridium thermocellum
Ruminococcaceae bacterium CPB6
Selenomonas ruminantium
Selenomonas sputigena
Peptoniphilus indolicus
Tissierellia bacterium KA00581
Methanobrevibacter millerae
Methanobrevibacter ruminantium
Methanobrevibacter sp. YE315
Thermoplasmatales archaeon BRNA1
Bacteroides pyogenes
Petrimonas mucosa
Prevotella saccharolytica
Phocaeicola abscessus
Capnocytophaga sp. ChDC OS43
Fusobacterium hwasookii
Leptotrichia sp. oral taxon 212
Ottowia sp. oral taxon 894
Kingella denitrificans
Neisseria sicca
Desulfobulbus elongatus
Campylobacter gracilis
Aggregatibacter aphrophilus
Treponema denticola
Treponema pedis
Cloacibacillus porcorum
Actinomyces cardiffensis
Actinomyces gerencseriae
Actinomyces johnsonii
Actinomyces odontolyticus
Actinomyces slackii
Actinomyces sp. oral taxon 414
Gardnerella vaginalis
Rothia aeria
Propionibacterium sp. oral taxon 192
Olsenella sp. Marseille-P2300
Coriobacteriaceae bacterium 68-1-3
Gemella morbillorum
Granulicatella adiacens
Streptococcus intermedius
Streptococcus oralis
Streptococcus sp. HMSC061D01
Clostridium sp. BNL1100
Inediibacterium massiliense
Eubacterium limosum
Catonella morbi
Lachnoanaerobaculum sp. OBRC5-5
Lachnospiraceae bacterium oral taxon 500
Peptostreptococcaceae bacterium oral taxon 113
Ruminococcus albus
Eggerthia catenaformis
Selenomonas sp. oral taxon 138
Murdochiella vaginalis
Peptoniphilus sp. oral taxon 386
Candidatus Saccharibacteria oral taxon TM7x
Methanobrevibacter olleyae
Methanobrevibacter smithii
Methanobrevibacter wolinii
archaeon
Proteiniphilum acetatigenes
Porphyromonas gingivalis
Prevotella sp. oral taxon 472
Capnocytophaga granulosa
Capnocytophaga sp. oral taxon 329
Fusobacterium nucleatum
Lautropia mirabilis
Herminiimonas arsenicoxydans
Neisseria elongata
Neisseria sp. HMSC072F04
Desulfobulbus propionicus
Cardiobacterium hominis
Aggregatibacter segnis
Treponema lecithinolyticum
Treponema socranskii
Fretibacterium fastidiosum
Actinomyces dentalis
Actinomyces glycerinitolerans
Actinomyces massiliensis
Actinomyces oris
Actinomyces sp. HPA0247
Actinomyces sp. oral taxon 849
Corynebacterium durum
Propionibacterium acidifaciens
Atopobium sp. oral taxon 810
Olsenella sp. oral taxon 807
Slackia exigua
Abiotrophia defectiva
Streptococcus anginosus
Streptococcus cristatus
Streptococcus sanguinis
Clostridium botulinum
Clostridium sp. SY8519
Anaerovorax odorimutans
Pseudoramibacter alactolyticus
Johnsonella ignava
Oribacterium sp. oral taxon 078
Filifactor alocis
Angelakisella massiliensis
Ruminococcus flavefaciens
Selenomonas noxia
Selenomonas sp. oral taxon 892
Parvimonas micra
Dethiosulfatibacter aminovorans
Methanobrevibacter arboriphilus
Methanobrevibacter oralis
Methanobrevibacter sp. AbM4
Candidatus Methanomethylophilus alvus

As the Jomon and Edo cultures evolved in different locations and were associated with distinct diets, we wanted to explore the similarities and differences between the microbiomes found in both cultures. Jomon individuals tended to cluster separately from Edo period specimens based on microbiome composition (Clade 1; Figure 4); however, this difference was not statistically significant (ANOSIM R^2 = 0.226, p-value = 0.07), as several Edo individuals fell within the Jomon cluster (Clade 1; Figure 2). We found no species that were differentially abundant between Edo period or Jomon samples (Kruskal-Wallis test with Bonferroni-correction p-values >0.05) and no species that were specific to either culture.

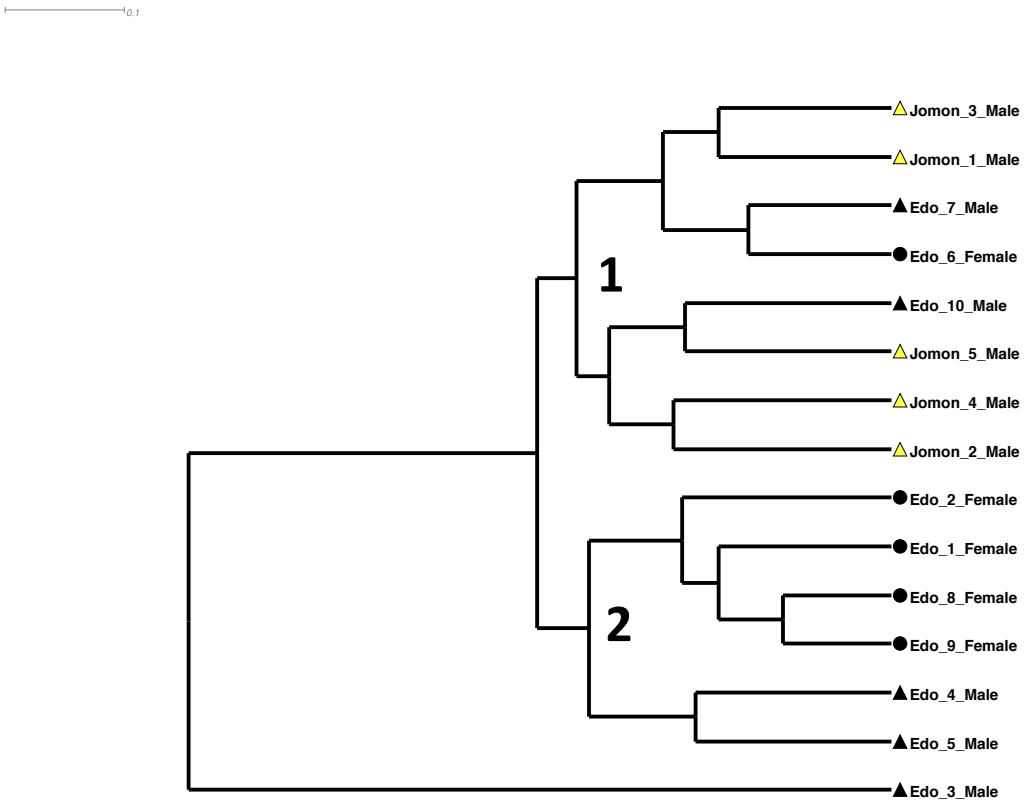


Figure 4. UPGMA tree of Bray-Curtis distances between samples. All Jomon period individuals fall in Clade 1, and Clade 2 contains strictly Edo period individuals, with further separation of male and females. Edo_3_Male is different from all other samples and is the same individual that clustered with Modern Human Microbiome Project plaque samples in Figure 1.

Within the Edo period clade (Clade 2; Figure 4) we found that the female individuals tended to cluster together, and this difference was statistically significant (ANOSIM R^2 = 0.28, *p*-value = 0.043). All females had evidence of periodontal disease, and all had their teeth dyed black which was a common cultural practice of the Edo period. We found no species that were differentially abundant between Edo period males and females (Kruskal-Wallis test with Bonferroni-correction *p*-values >0.05), and no species that were specific to either gender.

As all Edo females had signs of periodontal disease, which has been previously shown to influence microbiome composition in modern studies [73–75], we next tested for signatures of periodontal disease in our dataset. Within Edo individuals, no species were significantly associated with caries prevalence or periodontitis (Kruskal-Wallis test with Bonferroni-correction *p*-values >0.05), including members of the periodontitis-associated "red-complex" (*Treponema denticola, Tannerella forsythia, Porphyromonas gingivalis*) [76]. However, the abundance of the periodontitis-associated archaeon, *Methanobrevibacter oralis* [65], was substantially higher in the females (mean abundance in females=32%, mean abundance in males=5%) (Figure 5), though this difference was not statistically significant when controlling for multiple comparisons (Kruskal-Wallis uncorrected *p*-value = 0.028, Bonferroni-corrected *p*-value 3.795), potentially due to the small sample size and large degree of inter-individual variation observed (Figure 3).

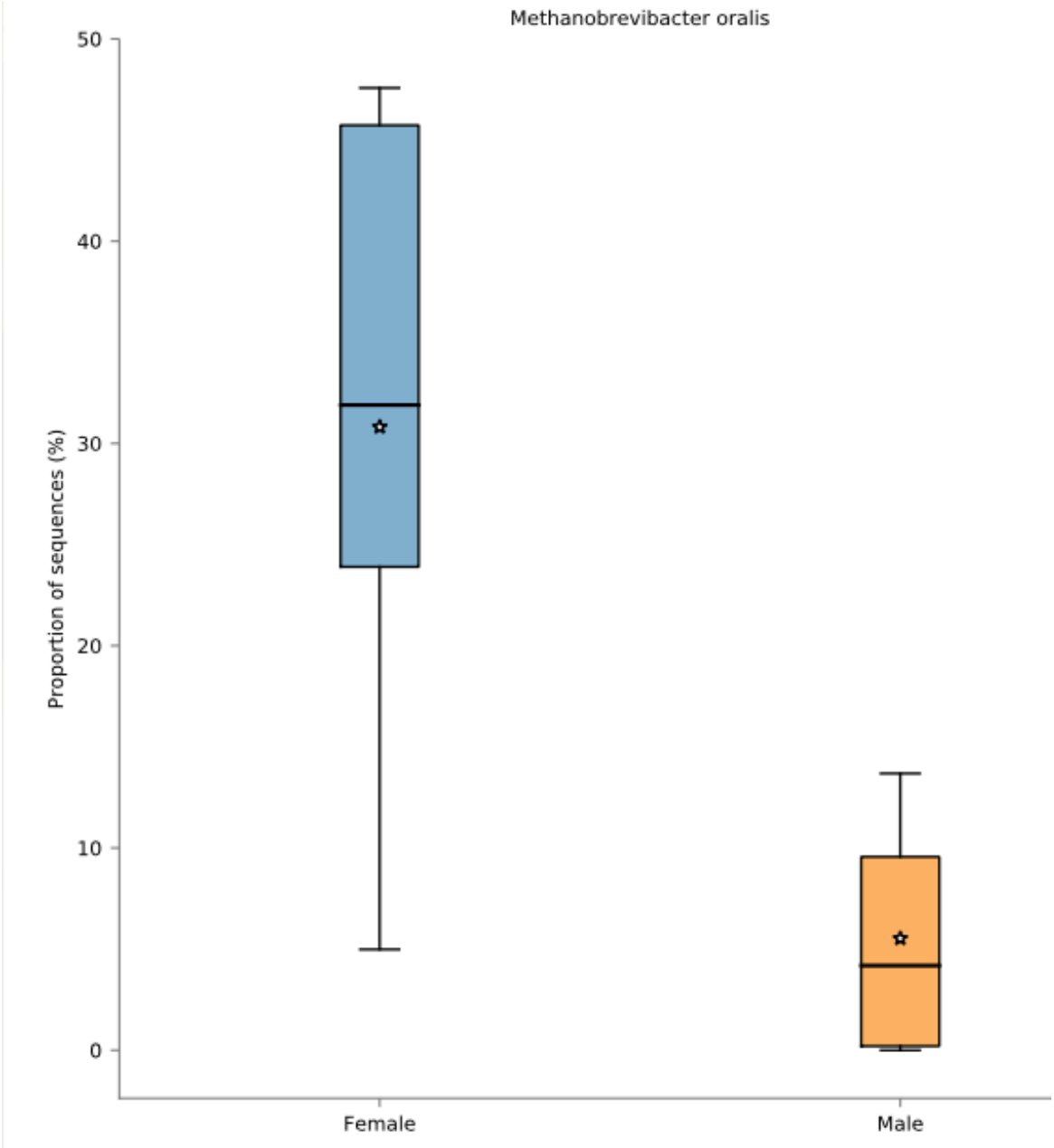


Figure 5. Box-plot of Methanobrevibacter oralis abundance in Edo period individuals. Females n=5 (blue) versus males n=5 (orange). Kruskal-Wallis test uncorrected *p*-value = 0.028, Bonferroni-corrected *p*-value 3.795.

Potential lineage replacement of bacteria during Jomon-Yayoi interaction

To test if the transition of Jomon to Yayoi culture in Japan resulted in a loss of Jomon microbial lineages, we focused on the specific genomes of microorganisms found in dental calculus. Given that the putative source of Yayoi culture into ancient Japan was mainland Asia, we added three Chinese dental calculus samples (circa 1800's) to test if mainland Asia was the potential source of any lineage replacement. These samples also clustered separately from EBCs (Figure 2) and had fragment length distributions and terminal deamination rates consistent with ancient DNA (Figures S1 & S2). To find suitable candidates for phylogenetic analysis, we determined the core oral microbiome in ancient Japan (*i.e.* species present in every sample). We found *Actinomyces sp. oral taxon 414, Actinomyces dentalis, Anaerolineaceae sp. oral taxon 439, and Olsenella sp. oral taxon 807* to be present in all samples. The oral bacterium *Anaerolineaceae sp. oral taxon 439* was chosen for phylogenetic analysis due to its high mean abundance within calculus samples (17.25%) which yielded a greater depth of coverage and higher quality variant calls. This bacterium is present at low abundance in healthy human plaque and higher abundance in individuals with periodontal disease [63]. Reads mapped against the *Anaerolineaceae sp. oral taxon 439* genome had terminal cytosine deamination typical of ancient DNA (Figures S1 & S2), with the Jomon and El Sidron Neanderthal samples having higher levels of cytosine deamination at terminal ends (13.9%) compared to the more recent (400-150-year-old) Chinese and Edo samples (6%), as expected with increasing age of sample [77]. Phylogenetic reconstruction found strong support for a distinct Jomon clade (Figure 6; yellow), while the *Anaerolineaceae* sequences in Edo calculus samples clustered with those from mainland China (Figure 6; orange). These findings were reproduced when restricting the analysis to transversions only to account for the potential influence of cytosine deamination on phylogenetic reconstruction (Figure S3).

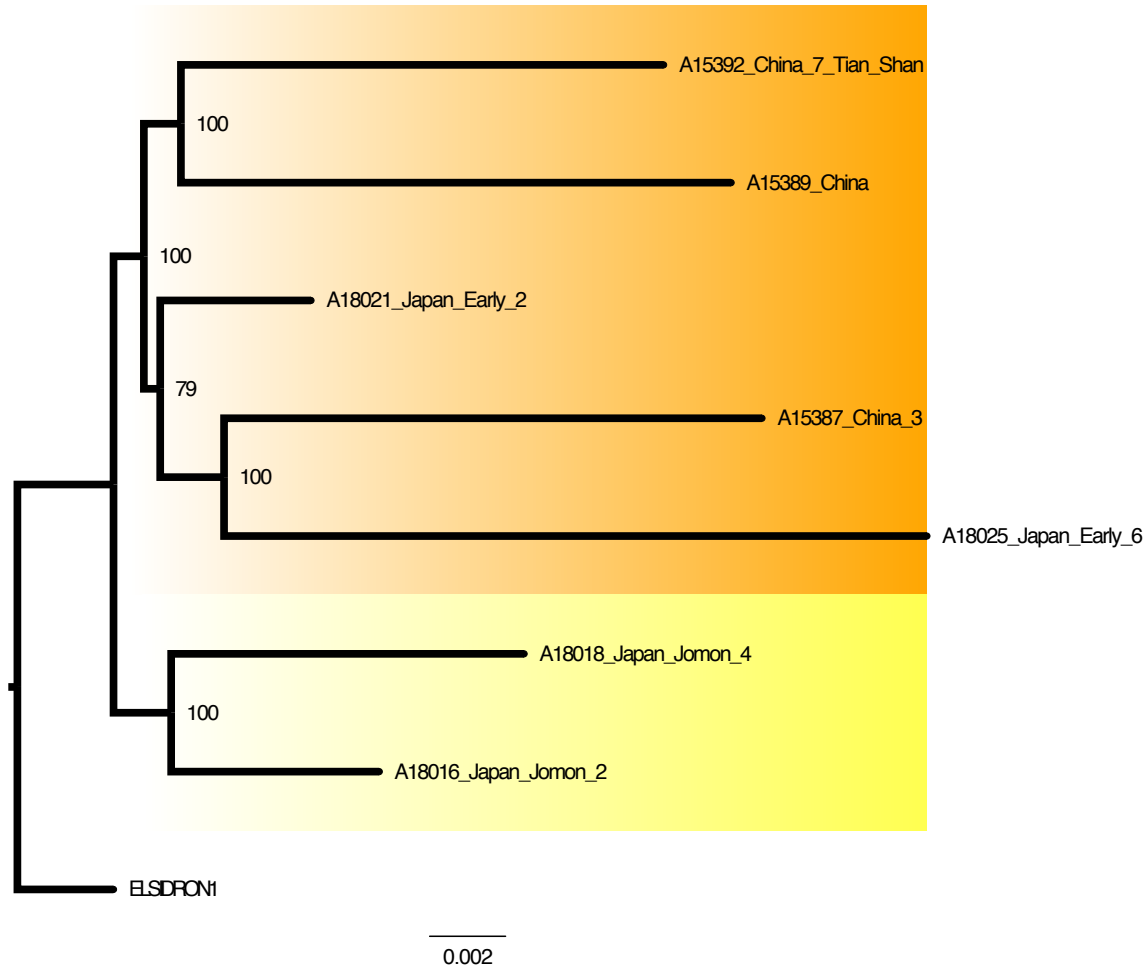


Figure 6. Maximum likelihood phylogenetic tree of reconstructed *Anaerolineaceae sp. oral taxon 439* genomes. Node labels represent percentage support of 1,000 bootstrap replicates. Elsidron 1 Neanderthal falls basal as an outgroup, with two separate clades containing Jomon (yellow), or a mixture of Chinese/Edo period Japanese samples (orange).

Discussion

This study is the first to explore the oral microbiome from ancient Japan. We found the DNA to be well-preserved and with minimal modern/ancient DNA contamination. The DNA also possessed characteristics of authentic ancient DNA, such as short fragment length distributions and terminal cytosine deamination [78]. We did not observe major differences between Jomon and Edo period microbiome compositions but did find differences between male and female Edo period Japanese which could be due to periodontal disease. Finally, using whole bacterial genomes we identified a possible bacterial lineage replacement, with Edo period Japanese possessing lineages more closely related to mainland Asian populations.

It has previously been reported that diet (hunter-gatherer vs. agricultural) may influence oral microbiome composition [31]. Here we reported a minor, but non-statistically significant difference in microbiome composition between Jomon (hunter-gatherer) and Edo (agricultural) period Japanese. No microbial species were found to be unique or differentially abundant between cultures, which could suggest that the classifiable oral microbiome composition has not changed drastically in Japan from Jomon to Edo period. This supports findings that oral microbiota are highly stable through time [71,79] and may be minimally influenced by certain dietary changes [80,81]. An alternate explanation for this finding is that the limited sample size of our study (5 Jomon and 10 Edo period) prevented the detection of such differences, as the large degree of inter-individual variation we observed (and which has been previously observed [71,72]) could have masked cultural/dietary signatures. Additionally, given that most modern oral reference genomes are generated from European and American isolates, the species that we classified may be biased towards core oral taxa that are stable through time [71,72,79,82]. Furthermore, we were unable to classify ~50% of reads from the ancient Japanese samples, and might, therefore, have missed microbial diversity present in these ancient samples which were unique to each culture or labile to dietary changes. Future improvements of analytical tools and further sampling of oral microbial genomes from broader human populations could allow for classification of the unclassified portion of our data, potentially providing enhanced bio-archaeological information from ancient dental calculus.

We found a significant difference between the microbiome composition of female and male Edo period Japanese—greater than the difference between Jomon and Edo cultures. This difference is not likely due to diet, as previously reported isotope data from the skeletons found no significant differences in dietary intake between male and female samples [42]. One potential driver of this difference is oral disease status, as all female samples had evidence of periodontal disease, which has been demonstrated in modern populations to impact microbiome composition [73,75,83]. In particular, we found the periodontal disease-associated archaeon *Methanobrevibacter oralis* [65,84] to be substantially more abundant in females versus males Edo period Japanese. Members of the periodontitis-associated "red-complex" were not found to be differentially abundant in females versus males [76]. However, this is unsurprising given recent recognition that periodontal disease is of complex aetiology, not the result of a handful of periopathogens [85]. Future studies with larger sample sizes including both periodontal-positive and negative individuals are needed to determine the influence of periodontal disease on the male/female split we observed in Edo period Japanese. Further studies controlling for periodontal disease could also test for influences of other cultural practices, such as teeth painting, which was common in Japan prior to the 20th century, and observed in the samples we

analysed. Overall, our findings suggest that periodontal disease is an important factor to control for when comparing microbial composition in ancient dental calculus studies, and future studies should aim to control for periodontal disease when making cultural comparisons.

We also explored the demographic history of Japan using whole bacterial genomes. The species we used for this analysis, *Anaerolineaceae sp. oral taxon 439,* is present in modern periodontally-healthy human plaque at low abundance, with increasing abundance observed in individuals with periodontal disease [63]. This bacterium was not identified in the extraction blank controls, and reads mapped against the genome were both short and possessed age-associated patterns of cytosine deamination typical of ancient DNA.

It is widely accepted that the modern Japanese population is the result of admixture between indigenous Jomon and later migrants from continental Asia during and after the Yayoi period [39]. We found evidence for a bacterial lineage replacement of Jomon associated *Anaerolineaceae* lineages. The Edo-associated *Anaerolineaceae* lineages appear were more closely related to historic Chinese strains than ancient Jomon. This could be interpreted as a microbial replacement event, whereby the continental Asian *Anaerolineaceae* lineage/s were brought by migrants to Japan and replaced the Jomon lineage. This is a plausible scenario if the continental Asian contribution to modern Japanese was larger than the Jomon, resulting in the loss of the lineage in a fashion analogous to genetic drift. Current estimates of Jomon genetic contribution to modern Japanese is <20%, supporting this scenario [39]. Another possibility is that the Jomon lineage has survived to this day, but that we did not detect it due to the relatively small sample size of our study. Future studies investigating modern individuals from across Japan could test for the survival of the Jomon *Anaerolineaceae* lineage. Spatially diverse sampling will be important, as it has been shown that genetic contribution from Jomon varied among populations across the Japanese Archipelago [37–39,86]. Further studies using ancient dental calculus could also assist in learning more about the source/s of Yayoi admixture, which remains undetermined. Future sequencing efforts will allow for the phylogenetic reconstruction of other human-associated microorganisms and permit investigations into how these genomes have changed through time—potentially yielding insights into their co-evolutionary history with humans.

In conclusion, we have reported the first ancient oral microbiome data from Asia. We also identified periodontal disease as being an important factor to control for when comparing microbial composition in ancient dental calculus studies. Finally, this study was the first to use ancient oral microbial genomes to investigate relationships between prehistoric populations, revealing insights into the demographic history of Japan.

References

1. Consortium THMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.

2. Agüero MG de, Ganal-Vonarburg SC, Fuhrer T, Rupp S, Uchimura Y, Li H, et al. The maternal microbiota drives early postnatal innate immune development. Science. 2016;351:1296–302.

3. Gensollen T, Iyer SS, Kasper DL, Blumberg RS. How colonization by microbiota in early life shapes the immune system. Science. 2016;352:539–44.

4. Belkaid Y, Hand TW. Role of the Microbiota in Immunity and Inflammation. Cell. 2014;157:121–41.

5. Rowland I, Gibson G, Heinken A, Scott K, Swann J, Thiele I, et al. Gut microbiota functions: metabolism of nutrients and other food components. Eur J Nutr. 2018;57:1–24.

6. Yano JM, Yu K, Donaldson GP, Shastri GG, Ann P, Ma L, et al. Indigenous Bacteria from the Gut Microbiota Regulate Host Serotonin Biosynthesis. Cell. 2015;161:264–76.

7. Sampson TR, Mazmanian SK. Control of Brain Development, Function, and Behavior by the Microbiome. Cell Host Microbe. 2015;17:565–76.

8. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. An obesity-associated gut microbiome with increased capacity for energy harvest. Nature. 2006;444:1027–131.

9. Ravussin Y, Koren O, Spor A, LeDuc C, Gutman R, Stombaugh J, et al. Responses of Gut Microbiota to Diet Composition and Weight Loss in Lean and Obese Mice. Obesity. 2012;20:738–47.

10. Tamboli CP, Neut C, Desreumaux P, Colombel JF. Dysbiosis in inflammatory bowel disease. Gut. 2004;53:1–4.

11. Carding S, Verbeke K, Vipond DT, Corfe BM, Owen LJ. Dysbiosis of the gut microbiota in disease. Microb Ecol Health Dis [Internet]. 2015 [cited 2018 Mar 18];26. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4315779/

12. Croswell A, Amir E, Teggatz P, Barman M, Salzman NH. Prolonged Impact of Antibiotics on Intestinal Microbial Ecology and Susceptibility to Enteric Salmonella Infection. Infect Immun. 2009;77:2741–53.

13. Blaser MJ. Antibiotic use and its consequences for the normal microbiome. Science. 2016;352:544–5.

14. Brown K, DeCoffe D, Molcan E, Gibson DL. Diet-Induced Dysbiosis of the Intestinal Microbiota and the Effects on Immunity and Disease. Nutrients. 2012;4:1095–119.

15. Pham TAN, Lawley TD. Emerging insights on intestinal dysbiosis during bacterial infections. Curr Opin Microbiol. 2014;17:67–74.

16. Beatty JK, Akierman SV, Motta J-P, Muise S, Workentine ML, Harrison JJ, et al. Giardia duodenalis induces pathogenic dysbiosis of human intestinal microbiota biofilms. Int J Parasitol. 2017;47:311–26.

17. Clemente JC, Pehrsson EC, Blaser MJ, Sandhu K, Gao Z, Wang B, et al. The microbiome of uncontacted Amerindians. Sci Adv. 2015;1:e1500183–e1500183.

18. Dominguez-Bello MG, Costello EK, Contreras M, Magris M, Hidalgo G, Fierer N, et al. Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc Natl Acad Sci U S A. 2010;107:11971–5.

19. Stahringer SS, Clemente JC, Corley RP, Hewitt J, Knights D, Walters WA, et al. Nurture trumps nature in a longitudinal survey of salivary bacterial communities in twins from early adolescence to early adulthood. Genome Res. 2012;22:2146–52.

20. Shaw L, Ribeiro ALR, Levine AP, Pontikos N, Balloux F, Segal AW, et al. The Human Salivary Microbiome Is Shaped by Shared Environment Rather than Genetics: Evidence from a Large Family of Closely Related Individuals. mBio. 2017;8:e01237-17.

21. Korpela K, Costea P, Coelho LP, Kandels-Lewis S, Willemsen G, Boomsma DI, et al. Selective maternal seeding and environment shape the human gut microbiome. Genome Res [Internet]. 2018 [cited 2018 Mar 26]; Available from: http://genome.cshlp.org/content/early/2018/03/14/gr.233940.117

22. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of Human Migrations in Helicobacter pylori Populations. Science. 2003;299:1582–5.

23. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, et al. Cospeciation of gut microbiota with hominids. Science. 2016;353:380–2.

24. Yatsunenko T, Rey FE, Manary MJ, Trehan I, Dominguez-Bello MG, Contreras M, et al. Human gut microbiome viewed across age and geography. Nature. 2012;486:222–7.

25. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, et al. Gut microbiome of the Hadza hunter-gatherers. Nat Commun. 2014;5:3654.

26. Rampelli S, Schnorr SL, Consolandi C, Turroni S, Severgnini M, Peano C, et al. Metagenome Sequencing of the Hadza Hunter-Gatherer Gut Microbiota. Curr Biol. 2015;25:1682–93.

27. Blaser MJ. Who are we? Indigenous microbes and the ecology of human diseases. EMBO Rep. 2006;7:956–60.

28. Eisenhofer R, Anderson A, Dobney K, Cooper A, Weyrich LS. Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events. J Isl Coast Archaeol. 2017;0:1–14.

29. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45:450–5.

30. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336–44.

31. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature. 2017;544:357–61.

32. Schroeder HE, Shanley D. Formation and Inhibition of Dental Calculus. J Periodontol. 1969;40:643–6.

33. Habu J. Ancient Jomon of Japan (Vol. 4). Cambridge University Press; 2004.

34. Imamura K. Prehistoric Japan: new perspectives on insular East Asia. Routledge; 2016.

35. Tomioka N. Animal resources and subsistence range during the Jomon period. Jomon archaeology, Vol. 4, Relationship between humans and animals [In Japanese]. 2010.

36. Kusaka S, Uno KT, Nakano T, Nakatsukasa M, Cerling TE. Carbon isotope ratios of human tooth enamel record the evidence of terrestrial resource consumption during the Jomon period, Japan. Am J Phys Anthropol. 2015;158:300–11.

37. Japanese Archipelago Human Population Genetics Consortium, Jinam T, Nishida N, Hirai M, Kawamura S, Oota H, et al. The history of human populations in the Japanese Archipelago inferred from genome-wide SNP data with a special reference to the Ainu and the Ryukyuan populations. J Hum Genet. 2012;57:787–95.

38. Jinam TA, Kanzawa-Kiriyama H, Inoue I, Tokunaga K, Omoto K, Saitou N. Unique characteristics of the Ainu population in Northern Japan. J Hum Genet. 2015;60:565–71.

39. Kanzawa-Kiriyama H, Kryukov K, Jinam TA, Hosomichi K, Saso A, Suwa G, et al. A partial nuclear genome of the Jomons who lived 3000 years ago in Fukushima, Japan. J Hum Genet. 2017;62:213.

40. Adachi Noboru, Shinoda Ken-ichi, Umetsu Kazuo, Kitano Takashi, Matsumura Hirofumi, Fujiyama Ryuzo, et al. Mitochondrial DNA analysis of Hokkaido Jomon skeletons: Remnants of archaic maternal lineages at the southwestern edge of former Beringia. Am J Phys Anthropol. 2011;146:346–60.

41. Hayashi K, Yamaguchi B, Dodo Y, Hiramoto Y. The middle of middle Jomon to the end of final Jomon: A Preliminary Report of the Survey of Localities B and C in the Miyano shell-mound: With Reference to the Human Skeletal Remains. Report of the research supported by the Grant-in-Aid for Scientific Research of the Ministry of Education, Science and Culture, Japan (In Japanese). 1981;

42. Tsutaya T, Nagaoka T, Kakinuma Y, Kondo O, Yoneda M. The diet of townspeople in the city of Edo: carbon and nitrogen stable isotope analyses of human skeletons from the Ikenohata-Shichikencho site. Anthropol Sci. 2016;124:17–27.

43. Weyrich LS, Dobney K, Cooper A. Ancient DNA analysis of dental calculus. J Hum Evol. 2015;79:119–24.

44. Brotherton P, Haak W, Templeton J, Brandt G, Soubrier J, Jane Adler C, et al. Neolithic mitochondrial haplogroup H genomes and the genetic origins of Europeans. Nat Commun. 2013;4:1764.

45. Meyer M, Kircher M. Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing. Cold Spring Harb Protoc. 2010;2010:pdb.prot5448.

46. Schubert M, Lindgreen S, Orlando L. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. BMC Res Notes [Internet]. 2016 [cited 2018 Feb 26];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4751634/

47. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv. 2016;050559.

48. Eisenhofer R, Weyrich LS. Assessing alignment-based taxonomic classification of ancient microbial DNA. (Chapter III) in preparation. 2018;

49. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, et al. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. PLOS Comput Biol. 2016;12:e1004957.

50. Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. Bioinforma Oxf Engl. 2013;29:1682–4.

51. Parks DH, Tyson GW, Hugenholtz P, Beiko RG. STAMP: statistical analysis of taxonomic and functional profiles. Bioinformatics. 2014;30:3123–4.

52. Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, et al. QIIME allows analysis of high-throughput community sequencing data. Nat Methods. 2010;7:335–6.

53. Schubert M, Ermini L, Sarkissian CD, Jónsson H, Ginolhac A, Schaefer R, et al. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. Nat Protoc. 2014;9:1056.

54. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25:1754–60.

55. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinforma Oxf Engl. 2012;28:1647–9.

56. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 2004;14:1394–403.

57. Castresana J. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol. 2000;17:540–52.

58. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.

59. Kumar S, Stecher G, Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. Mol Biol Evol. 2016;33:1870–4.

60. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci. 2004;101:11030–5.

61. Tamura K, Kumar S. Evolutionary distance estimation under heterogeneous substitution pattern among lineages. Mol Biol Evol. 2002;19:1727–36.

62. Marchesan JT, Morelli T, Moss K, Barros SP, Ward M, Jenkins W, et al. Association of Synergistetes and Cyclodipeptides with Periodontitis. J Dent Res. 2015;94:1425–31.

63. Campbell AG, Schwientek P, Vishnivetskaya T, Woyke T, Levy S, Beall CJ, et al. Diversity and genomic insights into the uncultured Chloroflexi from the human microbiota. Environ Microbiol. 2014;16:2635–43.

64. Brinig MM, Lepp PW, Ouverney CC, Armitage GC, Relman DA. Prevalence of bacteria of division TM7 in human subgingival plaque and their association with disease. Appl Environ Microbiol. 2003;69:1687–94.

65. Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, Relman DA. Methanogenic Archaea and human periodontal disease. Proc Natl Acad Sci U S A. 2004;101:6176–81.

66. Eisenhofer R, Cooper A, Weyrich LS. Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential. FEMS Microbiol Ecol [Internet]. 2017 [cited 2017 Jun 27];93. Available from: https://academic.oup.com/femsec/article/93/5/fix042/3089752/Reply-to-Santiago-Rodriguez-et-al-proper

67. Eisenhofer R, Weyrich LS. Proper Authentication of Ancient DNA Is Still Essential. Genes. 2018;9:122.

68. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. BMC Biol. 2014;12:87.

69. Dewhirst FE, Chen T, Izard J, Paster BJ, Tanner ACR, Yu W-H, et al. The Human Oral Microbiome. J Bacteriol. 2010;192:5002–17.

70. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic

and genomic information. Database [Internet]. 2010 [cited 2018 Feb 9];2010. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baq013/405450

71. Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, et al. Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. Npj Biofilms Microbiomes. 2017;3:2.

72. Proctor DM, Fukuyama JA, Loomer PM, Armitage GC, Lee SA, Davis NM, et al. A spatial gradient of bacterial diversity in the human oral cavity shaped by salivary flow. Nat Commun. 2018;9:681.

73. Li Y, He J, He Z, Zhou Y, Yuan M, Xu X, et al. Phylogenetic and functional gene structure shifts of the oral microbiomes in periodontitis patients. ISME J. 2014;8:1879–91.

74. Camelo-Castillo AJ, Mira A, Pico A, Nibali L, Henderson B, Donos N, et al. Subgingival microbiota in health compared to periodontitis and the influence of smoking. Front Microbiol [Internet]. 2015 [cited 2017 Nov 29];6. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2015.00119/full

75. Boutin S, Hagenfeld D, Zimmermann H, El Sayed N, Höpker T, Greiser HK, et al. Clustering of Subgingival Microbiota Reveals Microbial Disease Ecotypes Associated with Clinical Stages of Periodontitis in a Cross-Sectional Study. Front Microbiol [Internet]. 2017 [cited 2018 Apr 18];8. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2017.00340/full#h13

76. Socransky S s., Haffajee A d., Cugini M a., Smith C, Kent RL. Microbial complexes in subgingival plaque. J Clin Periodontol. 1998;25:134–44.

77. Sawyer S, Krause J, Guschanski K, Savolainen V, Pääbo S. Temporal Patterns of Nucleotide Misincorporations and DNA Fragmentation in Ancient DNA. PLOS ONE. 2012;7:e34131.

78. Dabney J, Meyer M, Pääbo S. Ancient DNA Damage. Cold Spring Harb Perspect Biol. 2013;a012567.

79. Utter DR, Mark Welch JL, Borisy GG. Individuality, Stability, and Variability of the Plaque Microbiome. Front Microbiol [Internet]. 2016 [cited 2016 Jul 8];7. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4840391/

80. Belstrøm D, Holmstrup P, Nielsen CH, Kirkby N, Twetman S, Heitmann BL, et al. Bacterial profiles of saliva in relation to diet, lifestyle factors, and socioeconomic status. J Oral Microbiol. 2014;6.

81. Keller MK, Kressirer CA, Belstrøm D, Twetman S, Tanner ACR. Oral microbial profiles of individuals with different levels of sugar intake. J Oral Microbiol [Internet]. 2017 [cited 2017 Dec 22];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5560414/

82. Cameron SJS, Huws S, Hegarty MJ, Smith DPM, Mur LAJ. The human salivary microbiome exhibits temporal stability in bacterial diversity. FEMS Microbiol Ecol. 2015;fiv091.

83. Kilian M, Chapple ILC, Hannig M, Marsh PD, Meuric V, Pedersen AML, et al. The oral microbiome – an update for oral healthcare professionals. Br Dent J. 2016;221:657.

84. Horz H-P, Conrads G. Methanogenic Archaea and oral infections — ways to unravel the black box. J Oral Microbiol [Internet]. 2011 [cited 2017 Sep 29];3. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3086593/

85. Hajishengallis G, Lamont RJ. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. Mol Oral Microbiol. 2012;27:409–19.

86. Nakagome S, Sato T, Ishida H, Hanihara T, Yamaguchi T, Kimura R, et al. Model-Based Verification of Hypotheses on the Origin of Modern Japanese Revisited by Bayesian Inference Based on Genome-Wide SNP Data. Mol Biol Evol. 2015;32:1533–43.

Supplementary figures and tables

Table S1. Japanese dental calculus samples with associated metadata

#SampleID	SpecificLocal	Culture	ToothType	ToothSurface	IndividualSex	Perio	Caries	MethanoProportion
18015_Japan_Jomon_1	Miyano	Jomon	Premolar	Buccal	Male	Y	N	5to1
18016_Japan_Jomon_2	Ikawazu	Jomon	Incisor	Buccal	Male	N	N	10to5
18017_Japan_Jomon_3	Miyano	Jomon	Unknown	Buccal	Male	N	Y	5to1
18018_Japan_Jomon_4	Miyano	Jomon	Molar	Buccal	Male	N	N	5to1
18019_Japan_Jomon_5	Ebishima	Jomon	Premolar	Lingual	Male	N	N	5to1
18020_Japan_Early_1	Tokyo	Edo	Premolar	NA	Female	Y	Y	40to30
18021_Japan_Early_2	Tokyo	Edo	Canine	Lingual	Female	Y	Y	40to30
18022_Japan_Early_3	Tokyo	Edo	Unknown	Lingual	Male	N	N	0
18023_Japan_Early_4	Tokyo	Edo	Canine	Buccal	Male	N	N	40to30
18024_Japan_Early_5	Tokyo	Edo	Molar	Lingual	Male	N	N	50to40
18025_Japan_Early_6	Tokyo	Edo	Premolar	Lingual	Female	Y	Y	5to1
18026_Japan_Early_7	Tokyo	Edo	Molar	Lingual	Male	N	N	5to1
18027_Japan_Early_8	Tokyo	Edo	Molar	Lingual	Female	Y	Y	50to40
18032_Japan_Early_9	Tokyo	Edo	Molar	Lingual	Female	Y	N	50to40
18033_Japan_Early_10	Tokyo	Edo	Molar	Buccal	Male	N	N	15to10

Table S2. Sequencing statistics for Japanese samples used in this study

Sample:	Number of DNA sequences	Mean fragment length	Reads without alignment (%)	Reads assigned to species (MEGAN normalised)	Reads assigned to species after EBC filtering	Reads remaining after EBC filtering (%)
18015_Japan_Jomon_1	595,632	97	44.7%	201,310	197,596	98.2%
18016_Japan_Jomon_2	1,201,049	68	47.7%	242,243	241,002	99.5%
18017_Japan_Jomon_3	2,925,395	62	46.3%	217,785	215,487	98.9%
18018_Japan_Jomon_4	1,084,967	69	55.0%	219,327	213,215	97.2%
18019_Japan_Jomon_5	1,079,836	103	46.9%	126,346	89,466	70.8%
18020_Japan_Early_1	1,510,459	70	62.1%	189,602	178,224	94.0%
18021_Japan_Early_2	1,838,395	76	52.7%	213,907	208,283	97.4%
18022_Japan_Early_3	1,081,657	82	20.3%	186,132	181,791	97.7%
18023_Japan_Early_4	3,233,252	77	66.9%	155,444	133,991	86.2%
18024_Japan_Early_5	1,099,036	137	44.2%	199,140	194,290	97.6%
18025_Japan_Early_6	2,125,863	87	47.2%	201,962	199,554	98.8%
18026_Japan_Early_7	1,515,743	91	51.4%	182,363	178,863	98.1%
18027_Japan_Early_8	746,478	70	59.1%	212,130	202,231	95.3%
18032_Japan_Early_9	1,600,982	71	49.2%	225,397	218,976	97.2%
18033_Japan_Early_10	1,647,411	79	52.6%	205,748	198,464	96.5%

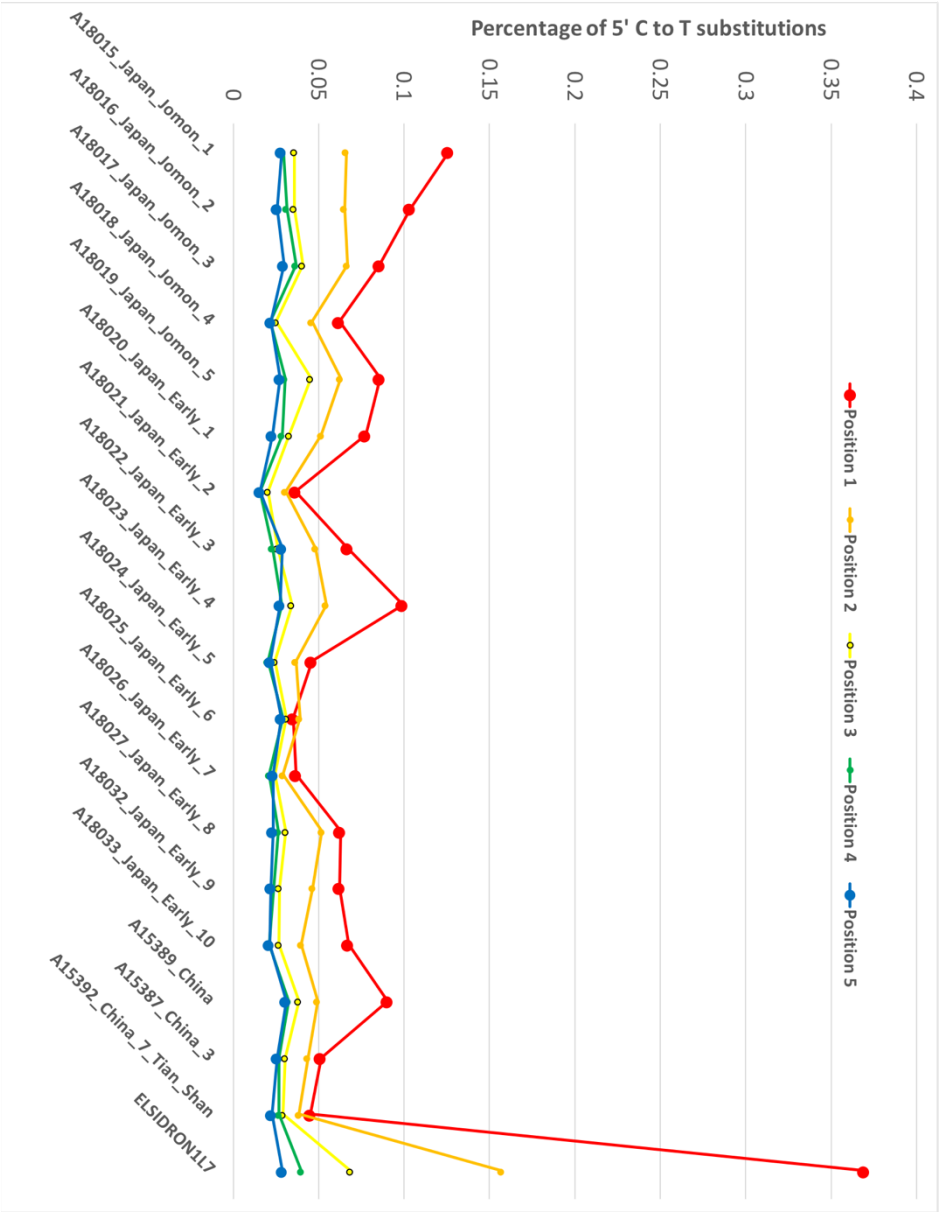
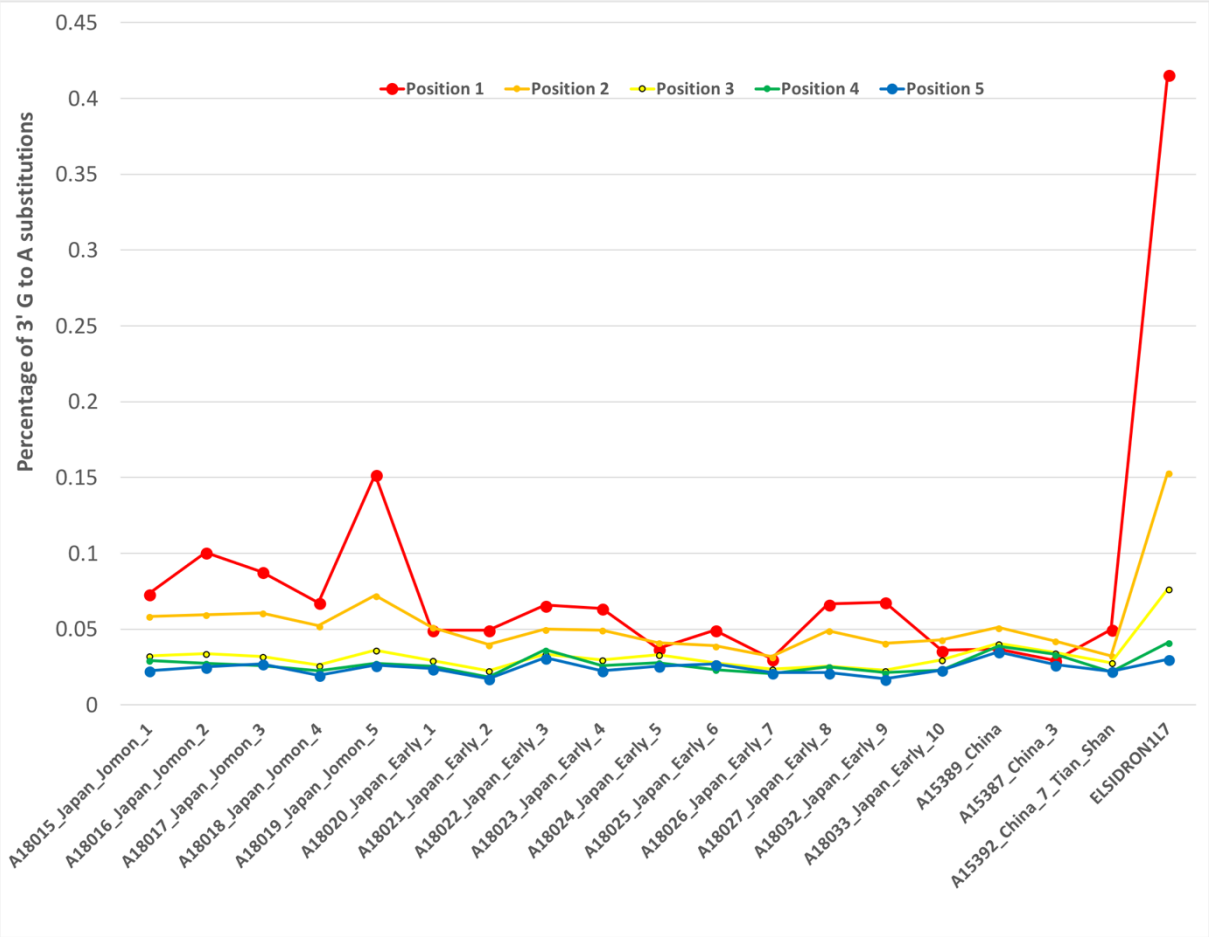


Figure S1. Damage profile for *Anaerolineaceae sp. oral taxon 439* at 5' terminus. Percentage of C-to-T substitutions at the five terminal bases of the 5' end of molecules. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red).

Figure S2. Damage profile for *Anaerolineaceae sp. oral taxon 439* at 3' terminus. Percentage of G-to-A substitutions at the five terminal bases of the 3' end of molecules. Positions 1-5 represent the bases adjacent to the fragmentation site from hot to cold, with position 1 being immediately adjacent (red).



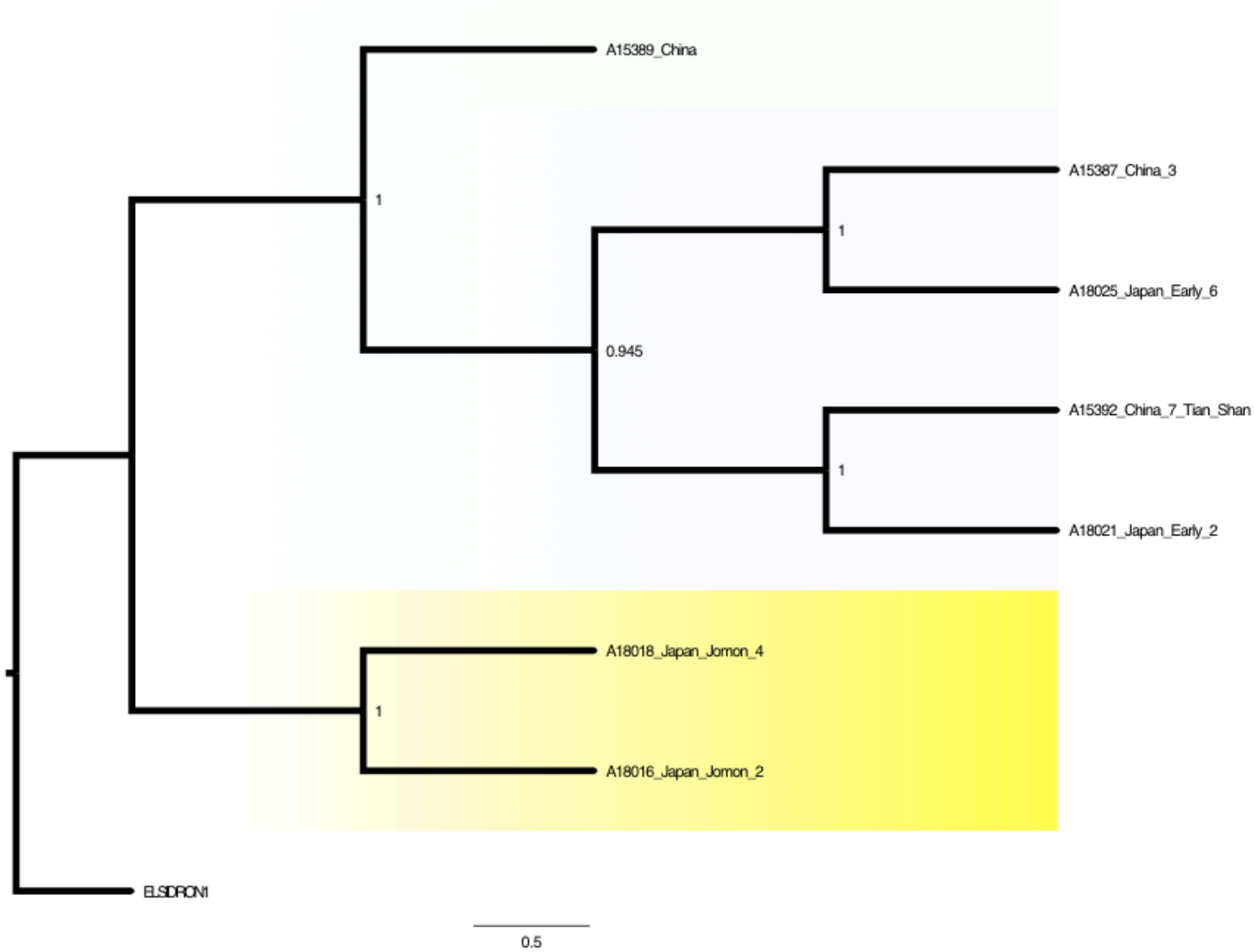


Figure S3. Transversions only Maximum Composite Likelihood Neighbour Joining tree. Node values indicate Percentage support for 1,000 bootstrap replicates.

Discussion

Palaeomicrobiological research can allow us to learn more about the lifeways, culture, and demographic history of our ancestors [1,2]. Such research can also help us understand how microbial communities have changed with humans throughout history, further adding to our understanding of the human microbiome in health and disease [3–5], and potentially informing and developing modern medical practices and therapies [5]. Critical to the realisation of these goals is the development and assessment of techniques for analysing ancient microbial communities, which is currently lacking due to the infancy of the field. This thesis develops new analytical techniques and refines existing ones to expand the scope, quality, and reproducibility of palaeomicrobiological research. In this discussion, I first highlight the major findings and contributions of this thesis and discuss their significance in a broader scientific and societal context. I then focus in on what this thesis means for the field of palaeomicrobiology, highlighting my key contributions, and identifying important issues moving forward while providing future directions for resolving them.

Broader significance of this thesis to science and society

Ancient dental calculus can be used to learn about past human migrations and demographic histories

Our ability to learn about and stitch together the rich tapestry of human history relies on studying the past. Such knowledge can help inform medical research [6–9], increase empathy and understanding for diverse human cultures [10,11], and help people (especially indigenous individuals) by connecting them with their culture/s [12,13]. This thesis demonstrated that ancient dental calculus can be used to learn about past human migrations and demographic histories, which could improve our understanding of human history.

In Chapter I [1], I investigated and proposed the use of ancient microbial DNA in dental calculus as a tool for inferring past human migration and demographic events. I also reviewed the archaeological, linguistic, and genetic data for human settlement in the Pacific, and highlighted reasons why ancient dental calculus may be the best (or sometimes only) approach

for understanding some past human migrations and demographic histories [1]. In Chapter V, I was the first to demonstrate that ancient dental calculus can be used to infer past human migrations and in doing so, added to our understanding of Pacific history. This chapter paves the way for future unprecedented discoveries about human history not just in the Pacific, but around the world. This thesis also demonstrated that it may be possible to use ancient dental calculus to learn about human demographic history. In chapter VI, I used ancient dental calculus from two cultural periods in Japan and found evidence for a potential bacterial lineage replacement event, supporting previous genetic and archaeological evidence for past population replacement in Japan.

Future developments to the work presented in this thesis could see ancient dental calculus used for familial or geographical and cultural identification. This could help efforts to repatriate museum skeletal remains, and the remains of individuals killed in conflicts or past wars, an issue important to people who have lost relatives in such situations [14]. While the analysis of human DNA is ideal for these purposes in most situations, the analysis of ancient dental calculus could be the best or only solution for repatriation of remains in geographic areas with poor DNA preservation and human populations with limited human genetic diversity (Chapter I). Because of the non-destructive nature of sampling dental calculus from human remains, its use can preserve skeletal remains, something that the analysis of human DNA cannot currently do. A further extension to the work presented in this thesis is to use microbial genomes in ancient dental calculus to add estimates of dates to archaeological and palaeontological contexts. Temporal estimates are essential for calibrating our understanding of history, and the analysis of biomolecular information using molecular clocks [15] has been used to add temporal context to diverse topics such as plant domestication [16], palaeontological estimates [17], human evolution [18], and archaeological sites [19]. Future analysis of ancient dental calculus could add temporal estimates to these contexts, with the advantages highlighted above and in Chapter I. In summary, this thesis unlocked ancient dental calculus as a tool for inferring past human migrations and demographic histories, and I look forward to future studies using ancient dental calculus to add new (and strengthen existing) threads of evidence to the complex tapestry of human history.

Refining methods for reconstructing ancient microbiomes

Recent research has demonstrated that the human microbiome is important for human health and disease [20–23]. Given the long association and coevolution of humans with microbial communities [24,25], understanding how the human microbiome has changed through time may be crucial to treating modern diseases and improving wellbeing [5,26–28]. Here, the

analysis of ancient microbial DNA in preserved dental calculus is poised to help [5,29,30]. However, critical to realising this potential is the accurate reconstruction of ancient microbiomes to better study them, compare them to modern microbiomes, and ultimately learn about how they have changed throughout history and contribute to modern health and disease. By developing new and refining current methods for reconstructing ancient microbiomes, this thesis lays the groundwork to aid in the reconstruction of ancient microbial communities.

In Chapter III, I performed an in-depth assessment of how the characteristics of ancient DNA influence alignment-based taxonomic classification of microbial communities. By using both simulated and real data, I demonstrated that the nucleotide-to-protein alignment method—which has been used previously in the field [2]—struggles to classify short DNA sequences that are typical of ancient DNA. Given that ancient samples vary in their DNA fragment length distributions, this represents an important source of bias for both taxonomic and functional comparisons in paleomicrobiological datasets. Therefore, I advocate against the use of nucleotide-to-protein alignments in favour of nucleotide-to-nucleotide alignments. Additionally, I highlighted the importance of reference database choice for alignment-based taxonomic classification, and stressed the importance for researchers (in modern and ancient microbiology) to carefully consider database choice when reconstructing microbiomes. In Chapter IV, I developed a new method for reconstructing ancient microbial communities using custom designed RNA probes to selectively enrich 16S rRNA gene fragments from ancient metagenomes. I demonstrated that this approach can be used to complement and add verification to other methods of reconstructing ancient microbiomes. While the use of 16S rRNA gene fragments from metagenomes has been previously used to reconstruct ancient microbial communities [2,31,32], there was yet to be a robust assessment of its applicability for doing so. This chapter also provided an in-depth assessment of the influence of ancient DNA characteristics on the ability to reconstruct ancient microbial communities using 16S rRNA gene fragments. Collectively, these chapters provide important findings and guidelines to improve future studies of ancient microbial communities.

These contributions will help future research using ancient microbiomes to inform modern medicine. Such research could identify health-associated microbial communities that could advise probiotic and prebiotic therapy, investigate important genomic changes in microbiomes relevant the human health and disease, and identify factors contributing to health and dysbiosis of human microbiomes to inform medical and societal interventions. While the reconstruction of ancient microbiomes is important for learning about past human health and disease, there are other fields of research where the work done in this thesis can contribute. For example, ancient microbial DNA locked in subglacial ice can be used to learn more about past

climatic changes [33,34]. Such research could assist in palaeoclimate reconstructions and explore microbial drivers of and responses to climate change. Another potential avenue of future research is the reconstruction of the ancient microbiomes from non-human animals, as other animals can form dental calculus [35–38]. Such research could shed light on diverse topics such as animal domestication, past interactions of animals with humans, the dietary habits of extinct animals, the taxonomic relationships between animals, and a greater understanding of host-microbiome coevolution throughout mammalian evolution. In summary, this thesis improves our ability to reconstruct ancient microbiomes, and in doing so, paves the way for future discoveries relevant to science and society.

Providing guidelines to improve the quality and reproducibility of future modern and ancient microbiome research

Scientific progress relies on findings that are well supported by robust evidence. Like building a house, knowledge built on faulty foundations can come crashing down, wasting valuable time and taxpayer money, and can potentially cause harm to society. For example, links between vaccinations and autism were found to be erroneous and subsequently retracted [39]. However, the impact this science had on the public's perception of vaccination led to a public health crisis resulting in many deaths that could have been prevented, and the echoes of this research are still being felt today [40]. Research suggests that science is undergoing a reproducibility crisis [41,42], and the rate of paper retraction has been found to correlate with journal impact factor [43]. Scientific misconduct accounts for the majority of article retractions [44], and the rate of fraud-induced retraction has increased ~10-fold since the 1970s [44]. This wastes valuable time and money, hinders the progress of medical research [45], and can erode public trust in science which could lead to the adoption of 'alternate facts' that are harmful to society [46].

The field of palaeomicrobiology is in its infancy, and there is still much to learn about confounding factors that can mislead scientific findings. As a result, there have been numerous claims made that have been challenged by other researchers [47–58]. Whether these controversial conclusions were made by honest mistakes or were the result of misconduct remains to be determined. However, to prevent future research from making mistakes, it is critical to identify and highlight pitfalls that could lead to poorly supported research and provide guidelines to improve future research. In Appendix Chapters I-III [47–49], I critically reviewed and responded to published paleomicrobiology studies that failed to include appropriate controls, used flawed methodology, and therefore produced conclusions that were unsubstantiated by sufficient evidence. This thesis, therefore, raises awareness of the pitfalls of

paleomicrobiological research and highlights key authentication criteria that must be met to ensure future reliable studies in the field.

On a similar note, I reviewed the state of modern low-biomass microbiome research in Chapter II, which shares similar challenges with paleomicrobiological research and also contains studies that have failed to meet sufficient authentication criteria [59–61]. Such research is typically done by individuals who have the best intentions, but who lack the required training and background knowledge to implement sufficient experimental controls and avoid confounding factors. By collaborating with leaders of low-biomass microbiome research, I used what I learned from the field of paleomicrobiology to help develop and put forth guidelines and a minimum set of criteria for future low-biomass microbiome research. Collectively, this thesis underscores the importance of a critical mindset when assessing ancient and low-biomass microbiome studies, and provides key examples and guidelines to help future researchers, reviewers, and editors safeguard the scientific rigour and replicability of coming research.

Contributions of this thesis to the field of palaeomicrobiology, limitations identified, and future directions to resolve them

Throughout my candidature, I identified numerous limitations inherent to the field of palaeomicrobiology. While I addressed some of these directly, some of them were beyond the scope of my given time and resources. By highlighting these limitations and providing potential solutions it is hoped that future work will be able to overcome them, and in doing so, offer new fundamental insights into microbiome research and human history.

1. *Classification of 'microbial dark matter'*

We are currently far from characterising the vast extent of microbial diversity on Earth [62,63]. The uncharacterised proportion of microbial communities often referred to as 'microbial dark matter', hinders the accurate reconstruction of microbial communities as alignment-based methods rely on prior genomic characterisation and representation in reference databases [64–67]. For example, Rinke *et al.* applied single-cell genomics to nine diverse microbial habitats and assembled 201 new microbial genomes [64]. These genomes allowed the authors to further classify up to 20% more of the metagenomic sequences from their samples, which allowed for further insights into community ecosystem structure and interactions. Given that alignment-based methods of taxonomic classification are currently the only tractable means of determining microbial composition in ancient samples (for reasons that I highlighted in Chapter III), the ability to study past microbial communities is limited by what is currently characterised in

modern reference databases. In Chapters III, V, and VI, I examined the current degree of this issue by analysing 157 ancient dental calculus samples from a broad range of geographic sites and time periods using to my knowledge, the largest MALT reference database built to date (47,696 microbial genome assemblies). Alarmingly, I found that an average of ~50% DNA sequences from ancient dental calculus samples could not be aligned and assigned taxonomy. This major limitation likely hindered my ability to detect and explore cultural differences between Jomon and Edo-period Japanese microbiomes in Chapter VI, and in the 16 different geographic regions across Asia-Pacific that I surveyed in Chapter V. In addition to hampering community-level analyses, this problem also limits genome-level analyses, as we are unable to detect species to study and are missing genomic regions that were lost from currently characterised species and are not present in modern reference genomes. Overall, this finding suggests that we are currently unable to characterise half of the DNA sequences within ancient dental calculus samples. Future paleomicrobiological research must recognise and report this limitation, and work towards addressing this critical issue.

Future directions for solving this problem include a larger effort to generate reference genomes from geographically diverse modern oral samples, and the *de novo* assembly of genomes from ancient samples. The latest Human Microbiome Project (HMP) study reported that for modern supragingival plaque samples from 265 individuals, an average of 75% reads could be aligned and assigned taxonomy ([68]; extended data figure 7). This number is close to the 80% I observed for the modern dental calculus sample analysed in Chapter III and suggests that further genomic characterisation is still required even for modern plaque microbiome studies. It should be noted that this number likely represents a best-case scenario for modern studies, as the HMP study used samples from healthy European/American individuals that have received the most attention in scientific research and from which the bulk of oral microbial genomes have been isolated and characterised from [69,70]. Therefore, it is likely that attempts to characterise modern oral microbiomes from other cultures (*e.g.* Asia, Africa) will yield lower percentages of classifiable DNA sequences. This same sampling bias towards healthy individuals of American and European descent may also hinder our ability to characterise the oral microbiomes of ancient cultures, such as the Pacific and Asian ones that I explored in Chapters V and VI. One approach to solving this problem is to further characterise the oral microbiomes of modern individuals on a global scale to improve representation in reference databases. This approach is becoming more feasible due to technological improvements in single-cell genomics [64,71] the culturing of fastidious microorganisms [72–74] (which allows for easier genome assembly), and bioinformatic techniques for *de novo* assembling of genomes from metagenomes [75,76].

While further characterisation of modern oral microbiome diversity is needed, this approach will be unable to characterise human-associated oral microorganisms that have become extinct throughout human history. Such microbial extinctions could arise due to the death or replacement of past human cultures *e.g.* the extensive mortality of historical South American populations due to disease brought by Europeans in the 15th century [77], or through population replacement of Japanese Jomon culture by Mainland Asian populations that I found evidence for in Chapter VI. Microbial extinctions could also have occurred in more recent times due to the adoption of 'Western' lifestyles/diets or the widespread use of antibiotics—both of which have been shown to change microbiome composition [78–81], and can lead to the loss of microbiome diversity [82]. To characterise 'extinct' microorganisms, *de novo* assembly of genomes from ancient metagenomes would be the only approach. Such analytical techniques already exist [75,76], and while a paleomicrobiological study found their performance on short and damaged DNA sequences was poor [23], further research is needed. Ideally, the first step would be to perform simulations (as done in Chapter III) to assess the tractability of *de novo* assembly in an ancient metagenome context. Such findings could also inform algorithmic improvements for *de novo* assembly tools to better cope with the characteristics of ancient DNA. Additionally, through the analysis of the 157 ancient dental calculus samples used in this thesis, I identified a large range of mean DNA fragment length distributions (*e.g.* 42-137 bp). Therefore, samples with longer mean fragment lengths could be preferentially targeted to make *de novo* assembly easier.

In summary, our current inability to characterise half of the DNA sequences from ancient dental calculus samples is a critical issue for the field. I believe that this problem is also directly related to the other limitations I describe below and must be addressed in order to expand the scope of future paleomicrobiological research.

2. *Computational constraints for searching ever-expanding reference databases*

While increasing the representation of microbial genomes in databases will help with characterising 'microbial dark matter', such expansion will also stress our ability to analyse them computationally. The development of MALT (MEGAN Alignment Tool) [83] made the alignment and assignment of millions of DNA reads against large reference databases feasible in a timely manner. MALT accomplishes this is by loading the hash table (database containing reference sequences) into RAM (Random Access Memory) to drastically improve runtime performance. Larger reference databases will, therefore, require larger quantities of RAM, which are only currently obtainable through expensive and sophisticated high-powered computer servers. For example, a standard desktop computer typically contains between 4-16

GB of RAM, whereas the computer server I used to load 47,696 genome assemblies had 1,500 GB of RAM. Such hardware constraints are beyond the reach of many researchers, and the addition of new reference genomes will push hardware requirements even further. I constructed a MALT database containing the sum of complete, chromosome, and scaffold-level genome assemblies (47,696) available as of July of 2017, and found that this number of references was the limit of what I could include in the within 1,500 GB of RAM. As of writing this discussion, the NCBI Assembly has expanded, and now contains 74,061 complete, chromosome, and scaffold-level bacterial genome assemblies. Clearly, computational and algorithmic advances must be developed to harness the growing number and diversity of reference genomes, especially if we are to reduce the proportion of 'microbial dark matter' mentioned in the previous section.

Due to computational constraints highlighted above, I could not include eukaryotic or viral genome assemblies, which, while only account for <1% of DNA sequences in ancient dental calculus [2,23] could potentially allow for the detection of additional health, dietary, and cultural information [2,23]. While I could have constructed a database containing only eukaryotic or viral genomes for such analysis, this is not recommended due to the potential for misclassification due to missing information and reference genome contamination. Misclassification of DNA sequences due to missing reference information is a problem that I identified in Chapter III. I found that by aligning simulated metagenomes that contained non-coding regions against a reference database that did not have non-coding sequences resulted in misclassification of these non-coding regions to taxa not present in my simulated community. Such an issue is likely to yield false-positive assignments if aligning data against a reference database containing only eukaryotic or viral assemblies (R Eisenhofer, unpublished data). This issue is also why I chose to use competitive mapping [84] for the phylogenetic analysis of microbial genes and genomes in Chapters V and VI. Another source of misclassification is contaminated genome assemblies [85–87]. A 2014 study found that a complete *Neisseria gonorrhoeae* genome assembly contained portions of DNA derived from cow and sheep genomes [85]. Another study in 2016 identified 154 genome assemblies that contained contamination from human DNA [86]. Recently, it was found that for 245 eukaryotic pathogen genome assemblies, an average of 11% of each genome contained contamination or low-complexity regions [87]. Therefore, a major issue moving forward is the identification and removal of contamination from genome assemblies prior to metagenomic classification, which could be accomplished using the tool developed in [87]. In summary, there are currently computational challenges that hinder the reconstruction of ancient microbial communities, and

future collaboration with computer scientists and bioinformaticists is needed to address these issues.

3. *Whole-genome analysis of ancient microorganisms*

In this thesis, I was the first to demonstrate that ancient microbial DNA in dental calculus could be used as a proxy for past human movements and interactions. Specifically, in Chapter V, I identified the bacterium *Anaerolineaceae sp. oral taxon 439* as a viable candidate for phylogenetic analysis, being both highly prevalent and abundant and recapitulating prior evidence for the peopling of the Pacific. While this finding expands the scope of ancient dental calculus research, there is more work to be done in further validating and extending this method to other microbial species. A major limitation I faced in this thesis was the relatively shallow depth of sequencing per sample (~2 million reads). This limited my phylogenetic analysis to the seven most abundant and prevalent taxa identified in Chapter V, and the reduced the quality of the whole-genome alignments. Current whole-genome alignment methods were designed to work on high-depth and coverage modern data [88,89], and it is possible that the issues I observed with other microbial taxa are due to poorly aligned genomes resulting from insufficient sequencing coverage. Future simulations (using the techniques applied in Chapter III) are needed to empirically test the minimum depth of sequencing required for ancient dental calculus data to ensure high-quality whole-genome alignments. These simulations would also benefit from testing for the influence of DNA fragment length, cytosine deamination, and competitive mapping on the quality of such alignments, which could help inform potential improvements to alignment algorithms for use with ancient DNA.

Once the issue of insufficient sequencing coverage is resolved, research is needed to determine if some oral microorganisms act as better proxies for past human movements than others. There are many reasons that could potentially explain why I found that *Anaerolineaceae sp. oral taxon 439* recapitulated past human movements in the Pacific, while the other species tested did not. Possible reasons for this include the biology of the microorganism tested, such as generation time, mutation rate, and/or degree of horizontal gene transfer. Different microbial species are known to have different generation times, ranging from minutes to weeks, and this variation corresponds to factors including metabolism, the availability of substrates, and temperature [90]. Faster generation times are expected to increase the frequency of mutations [91], and faster generation time has been empirically shown to increase the rate of neutral mutations [92]. Therefore, variation in the mutation rate between different species within dental plaque communities is likely, and future studies could identify microorganisms with different rates for use in different demographic scenarios. For example, a microorganism with a faster

mutation rate might be useful for inferring human movements that rapidly (*e.g.* settlement of East Polynesia), but may saturate and be ineffective at measuring deeper demographic events (*e.g.* out of Africa). Another factor that could violate clonal phylogenetic analysis of microbial genomes is horizontal gene transfer [93,94]. Some oral genera are known to have elevated rates of homologous recombination between related species (*e.g. Streptococcus* [95], *Neisseria* [96], and *Porphyromonas* [97]). Therefore, some oral taxa may be better suited for phylogenetic analysis than others, and future studies are needed to identify such microorganisms. One approach would be a time-series analysis of modern plaque microbiota from individuals, whereby individuals are sampled at multiple time points. Additionally, the degree of a microorganism's rate of recombination could be estimated in current ancient dental calculus datasets through the use of existing analytical techniques [98,99]. Ultimately, the discovery of other phylogenetically-informative oral species will further improve our confidence in and the resolution of inferring past human demographic events from ancient dental calculus. To summarise, the era of genomic exploration into past human microbiota is here, and future improvements in analytical techniques will undoubtedly expand our understanding of past human demographic history and microbial evolution.

4. *Identifying confounders of microbial community analyses*

A goal of the bulk of current ancient dental calculus research is to study the composition of oral microbiota through time, making comparisons between different cultures and periods throughout human existence. However, there are factors, both known and unknown, that can confound such community-level analyses. For example, recent research on both modern and ancient populations has shown that both tooth type (*e.g.* molar, incisor) and tooth surface (*e.g.* buccal and lingual) influence microbial community composition [100,101]. These findings forced me to control for tooth type in Chapter V and likely reduced my power to detect differences between cultures in both Chapters V and VI. Future studies with a focus on characterising and comparing microbial composition need to account for these findings by preferentially sampling a specific tooth/surface. Ideally, the field should agree upon a specific tooth/surface to allow better comparisons between datasets as they become publicly available.

Another factor that could confound comparisons of microbiota between cultures is the presence of oral disease, especially periodontal disease. Modern oral microbiome research has found that host periodontal disease state is a major driver of microbial composition [102–107]. Research prior to next-generation sequencing technology supported the idea that complexes of specific microbial species (*e.g.* the red complex) are signatures of periodontal disease [108]. While modern high-throughput studies have corroborated a shift in oral microbiota linked to

periodontal disease (*e.g.* increases in the abundance and diversity of anaerobic microorganisms), a reproducible, characteristic set of microorganisms present in disease has not been found. As such, it is increasingly recognised that periodontitis is a complex polymicrobial disease that can result from different community assemblages [109–112]. Such changes in community structure due to periodontal disease will make comparisons between cultures and time points difficult, as the disease phenotype could confound intra- and inter cultural comparisons. Additionally, disease-associated communities between different cultures or time points that share common trends (such as the increase in anaerobic membership/abundance) could hinder our ability to detect culture-specific microbiota. While future sampling efforts towards periodontally healthy samples could alleviate this issue, the classification of periodontal disease in ancient skulls is currently difficult due to taphonomic issues, such as post-mortem tooth loss, natural erosion, and skeletal preservation. Additionally, passive tooth eruption due to tooth wear can also lead to the loss of periodontal attachment over time — which is a primary paleopathological assessment of periodontal disease [113–115]. Furthermore, paleopathological assessments can only measure hard-tissue, and current modern medical diagnoses require the presence of soft gingival tissue, making accurate diagnoses difficult [116]. In addition, gingivitis could result in a disease-associated community structure without leaving evidence of periodontal disease [117–119]. Therefore, greater characterisation of microbial community trends pertaining to periodontal disease are needed in future modern and ancient studies, and given that different cultures likely possess distinct microbial assemblages, such assessments may be needed on a culture-by-culture basis.

While there may not be microorganisms that are universal in periodontal disease, some species are highly correlated to the disease [120,121]. In Chapters V and VI, the abundance of the periodontal disease-associated archaeon *Methanobrevibacter oralis* was a major correlate with microbial community differences between samples. *Methanobrevibacter oralis* is a methane-producing anaerobe that is associated with periodontal disease [122,120]. A 2004 study that examined 205 subgingival plaque samples from healthy and periodontally diseased individuals from the United States detected *Methanobrevibacter oralis* in 36% of diseased individuals and not in healthy controls [123]. Additionally, the severity of periodontitis correlated with increasing abundance of *Methanobrevibacter oralis* [123]. Another study of individuals from Japan corroborated these findings by only finding *Methanobrevibacter oralis* in periodontally diseased individuals and not in healthy controls [124]. These findings suggest that *Methanobrevibacter oralis* is potentially an important indicator of periodontal disease in some individuals. The absence of *Methanobrevibacter oralis* from some periodontally diseased individuals also lends support to the polymicrobial nature of periodontal disease that I

highlighted above. Additionally, the detection of *Methanobrevibacter oralis* in people from the United States and Japan suggests that this archaeon is present in diverse human populations. The widespread detection of *Methanobrevibacter oralis* in my dataset (prevalence of 75% from the 132 samples analysed in Chapters V and VI) suggest that this archaeon been with humans for a long time. This idea is supported by the detection of *Methanobrevibacter oralis* in a ~50,000-year-old Neanderthal individuals [23], and it will be interesting to see how far back this association goes in future studies. Practically speaking, the presence and abundance of *Methanobrevibacter oralis* in my datasets may have hindered my ability to detect community differences both within and between cultures. For example, in Chapter VI I found that the proportion of *Methanobrevibacter oralis* was the major distinguishing factor between Edo-period male and female individuals, and this appeared to be associated with a difference is periodontal health. Additionally, the higher prevalence and abundance of *Methanobrevibacter oralis* in Edo-period individuals likely hindered the identification of cultural-specific species differences between individuals within the Jomon and Edo-periods. Overall, my findings suggest that periodontal disease is a confounding factor for ancient microbiota comparisons both within and between cultures and that the presence of *Methanobrevibacter oralis* could be used as an indicator for past periodontal disease state.

Another potentially confounding factor that deserves future research is taphonomic biases, or determining whether DNA from some microbial taxa preserves better through time. If this is the case, the differential ages between samples could be misinterpreted as increasing or decreasing the abundance of certain taxa. Potential causes for such a phenomenon include differences in microbial cell walls, metabolism, the presence of proteins protecting DNA (*e.g.* histones), and GC-content. It was originally hypothesised that differences in cell walls may influence DNA fragmentation and damage patterns observed in ancient samples [66]. Such an influence of cell wall was reported for ancient *Mycobacterium* [125], but not in subsequent studies [2,31], suggesting that factors other than cell walls may contribute to the differential preservation of microbial DNA through time. Future studies should test for microbial metabolism as a potential contributor to differential preservation, as differences in microbial metabolism (*e.g.* redox states, pH) could promote/inhibit hydrolytic depurination. In addition, the presence of proteins in direct contact with DNA could lead to the differential preservation of microbial DNA; for example, it has been demonstrated that the observed periodicity in ancient human DNA fragment lengths is due its association with histones [126]. While it is generally accepted that most bacteria lack histones, archaea are known to possess them [127–129]. Given the high prevalence of the archaeon *Methanobrevibacter oralis* I observed in Chapters V and VI, it will be interesting to test if the presence of histones in

Methanobrevibacter oralis offer enhanced DNA preservation for this archaeon. Finally, differences in base composition between taxa, such as GC-content (*e.g.* 28% for *Methanobrevibacter oralis* vs. 73% for *Actinomyces dentalis*), will be another factor to test when comparing differential DNA preservation between microbial species through time. Future studies leveraging the growing number of ancient dental calculus samples with statistical modelling could identify and correct for such a bias.

Key to identifying and robustly reporting the confounding factors highlighted above will be obtaining larger sample sizes for future studies. Modern microbiota research has found that inter-individual differences in plaque microbiota can be large [101,130]. Corroborating these studies, in chapter VI I found that inter-individual variability was high in Jomon and Edo period Japanese samples. Such inter-individual differences likely reduced my power to detect potentially culture-specific microbiota and may limit future attempts at identifying confounding factors of ancient microbiota compositions. Future studies leveraging larger sample sizes will be key in identifying confounding factors and allow robust classifications of culture-specific microbiota. To summarise, factors both known and unknown can confound the study of past microbial composition. Future research is needed to characterise and potentially correct for such factors so that false conclusions are not made. Larger sample sizes will also be needed to help identify these confounding factors, and improve our ability to identify culture-specific microbiota.

5. *Functional analysis of ancient microbial DNA*

Most modern microbiome studies to date have focused on classifying taxonomic composition. However, it is becoming increasingly recognised that the functions of these communities are critical for understanding their roles in human health and disease [131,132]. The currently available tools used to characterise functional information from metagenomic datasets were designed for use with modern DNA and involve the translation of nucleotide sequences into amino acid sequences for search against a protein database [133,134]. In Chapter III, I demonstrated that there are issues with using nucleotide-to-protein alignments for ancient DNA, as short (<60 bp) DNA sequences are unable to be aligned once converted into amino acids. Given the large range of mean DNA fragment lengths for samples identified in Chapters V and VI (42-137 bp), such functional analyses would be severely biased. For example, samples obtained from a site with poor DNA preservation could be interpreted lacking specific functions when compared to better-preserved samples. While one could control for this by using samples with longer, uniform mean fragment lengths, this would severely reduce the number of samples available for analysis and the power of the statistical tests used. Therefore, there is an urgent

need for the development of new tools or adaptation of current ones to allow for the characterisation of functional information in paleomicrobiological studies. A promising direction will be the customisation of MALT to allow for the mapping of functional information to nucleotide-to-nucleotide alignments, which I demonstrated were robust to the characteristics of ancient DNA.

Once we can analyse functional information using ancient DNA, consensus must be reached on how best to analyse such data. Estimates for the number of genes in the human microbiome are in the millions [135], and the ability to analyse and make sense of this data will be challenging. For example, statistical analysis of the differential abundance of genes/functions between samples may be difficult, as multiple test corrections such as the Bonferroni-correction become harder to reach significance as more comparisons are made. Furthermore, simulations are needed to determine the depth of sequencing required to assess the full functional diversity present in a sample. In addition to functions identified with ancient DNA, future improvements in palaeoproteomics [136,137] could be used alongside DNA functional analysis to corroborate and enhance findings.

Further challenges to the functional analysis of ancient microbial DNA include the large proportion of 'microbial dark matter' present in ancient microbial datasets highlighted above, which will preclude the assignment of functions. Additionally, there are still many microbial genomes that have hypothetical proteins with uncharacterised functions [138]. Computational limitations such as those discussed earlier will also further hamper future functional analyses. Furthermore, identifying additional factors that bias community composition such as those discussed above are needed to ensure functional analyses are unbiased. Future work towards resolving these issues will allow for robust exploration of functional information in past microbial communities, which will be useful in determining how potential changes in these functions through human history may have contributed to oral health and disease.

6. *The paucity of comparative modern experiments and data*

Critical to our interpretations of ancient dental calculus data is a greater characterisation of the factors that influence oral microbiota in modern populations. The study of ancient dental calculus precludes manipulatable experiments for teasing apart such factors. Therefore, further modern dental plaque and calculus microbiota studies are needed to provide context to findings in ancient samples. Firstly, it will be important to further examine if there are notable differences between dental plaque and dental calculus microbiota, as dental plaque is easier to sample in modern individuals. Modern research suggests that plaque biofilm stage influences microbial composition [139]; therefore, it could be expected that dental calculus (which is

thought to form from late-stage biofilms [140]) may contain different microbial assemblages to early-stage dental plaque. Adler *et al.* compared modern dental plaque and calculus from the same tooth in six individuals and found small to moderate differences in community composition [66]. However, this study did not control for tooth type, tooth surface, or plaque biofilm stage, factors that have been recently found to influence plaque microbiota composition [100,101,139,141]. Future studies will larger sample sizes will allow for determination of whether dental plaque is a suitable proxy for dental calculus microbiota, which will inform experimental design for studies seeking to identify factors influencing plaque and calculus microbiota.

Future modern research should also seek to determine the influence of host dietary intake on plaque microbiota composition. This is yet to be fully determined in a modern context and would enhance our ability to infer dietary signals in past human populations, as well as corroborate previous paleomicrobiological studies on the topic [2,66]. The fermentation of dietary sugars by bacterial species (especially *Streptococci*) is a known risk factor for dental caries [142]. These dietary sugars give microorganisms able to exploit them an advantage over other members of the community both nutritionally, and by changing the chemistry of the local environment (*i.e.* reducing pH) [143]. However, such changes in plaque biofilms typically occur on the occlusal surfaces of teeth [144] which typically do not form calculus due to mastication- or dental hygiene-induced abrasion, and are therefore not sampled in ancient specimens. To my knowledge, the only modern study using dental plaque to test if host dietary intake influences microbiota composition is a recent study by Keller *et al.*, which found that different levels of sugar intake had minimal influence on buccal and interdental plaque microbiota composition [145]. Future experiments testing for the influence of diet on plaque microbiota should also take tooth type and surface into account, as differential salivary flow rates for different teeth/surfaces could impact plaque microbiota response to dietary intake [101,146,147].

While there is currently little support for host diet influencing plaque microbiota in modern studies, palaeomicrobiological studies have found evidence [2,66]. Adler *et al.* analysed 34 ancient dental calculus samples from different geographic sites and temporal periods in Europe. They found support for putatively dietary induced shifts in oral microbiota from pre-agriculture to agriculture times, and from medieval times to the Industrial Revolution. However, this study used amplification of the 16S rRNA gene to reconstruct ancient microbial communities, which was later demonstrated to be severely biased by the characteristics of ancient DNA [2,31] and could confound this conclusion. Weyrich *et al.* used the less biased shotgun metagenomic sequencing to taxonomically classify 19 ancient dental calculus samples

from different geographies and time periods (2 African gatherers, 2 African pastoralists, 2 European hunter-gatherers, 7 European farmers, 2 Europeans from the Industrial Revolution, 3 Neanderthals, and 1 wild chimpanzee). They identified three distinct groupings of samples by microbial composition relating to host dietary intake. However, they used nucleotide-to-protein alignments for taxonomic reconstruction which cannot align DNA fragments shorter than 60 bp (Chapter III), which, coupled with the range of different mean fragment lengths in the study, could potentially bias these groupings. Further reanalysis of this dataset with nucleotide-to-nucleotide alignments will be needed to confirm these host diet-induced groupings of ancient microbial communities.

Recent studies have found that host genetics can influence oral microbiota composition [148,149] (though see [150]). Future modern research characterising host genetic influences—especially for plaque microbiota—will be important to consider for interpreting changes in ancient dental calculus datasets. Furthermore, future ancient studies pairing host/microbiota samples from the same individuals could be used to directly test these potential influences and how they have changed throughout human history. In addition to teasing apart factors that influence oral microbiota composition, modern studies are desperately needed to act as points of reference for how these microbial communities have changed through time. Currently, the bulk of oral microbiome research is undertaken in the United States and Europe [69,70], and comparisons between these data and the data I generated in Chapters V and VI are not directly comparable for observing such changes. To address this issue, further modern sampling focusing on greater geographical and cultural representation is needed. This will potentially have the added benefit of improving the characterisation of 'microbial dark matter' highlighted above. Further extensive modern sampling could also tease apart global trends in plaque microbiota such as the identification of keystone species [151], or similarities/dissimilarities in dysbiotic microbial consortia. The history and inception of such global trends could then be tested using the temporal reach of ancient dental calculus.

Overall, the current paucity of comparative modern plaque microbiota studies limits our ability to interpret ancient dental calculus data. Future modern experiments and greater global sampling will enhance our ability to use ancient dental calculus data to its potential.

Conclusion

Ancient DNA analysis of past human microbiota in dental calculus has the potential to shape our understanding of how hosts and oral microbiota have evolved through time. Such findings may lead to improved medical insights and therapies for treating modern oral diseases and could

enhance our knowledge of human history and culture. Through the development, assessment, and improvement of analytical techniques, this thesis expands the scope of what we can learn from ancient microbial DNA. Additionally, this thesis provides 132 new and authenticated ancient dental calculus samples that have added to our understanding of global oral microbiota diversity and will act as a resource for further studies. Furthermore, by exploring and characterising pitfalls inherent in low-biomass and ancient DNA research, this thesis provides guidelines for researchers to improve the quality and reproducibility of future modern and ancient microbiome research.

References

1. Eisenhofer R, Anderson A, Dobney K, Cooper A, Weyrich LS. Ancient Microbial DNA in Dental Calculus: A New method for Studying Rapid Human Migration Events. J Isl Coast Archaeol. 2017;0:1–14.

2. Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. Nature. 2017;544:357–61.

3. Adler CJ, Dobney K, Weyrich LS, Kaidonis J, Walker AW, Haak W, et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45:450–5.

4. Warinner C, Rodrigues JFM, Vyas R, Trachsel C, Shved N, Grossmann J, et al. Pathogens and host immunity in the ancient human oral cavity. Nat Genet. 2014;46:336–44.

5. Weyrich LS. Evolution of the Human Microbiome and Impacts on Human Health, Infectious Disease, and Hominid Evolution. Reticul Evol [Internet]. Springer, Cham; 2015 [cited 2018 Mar 26]. p. 231–53. Available from: https://link.springer.com/chapter/10.1007/978-3-319-16345-1_9

6. Diversity NRC (US) C on HG. Scientific and Medical Value of Research on Human Genetic Variation [Internet]. National Academies Press (US); 1997 [cited 2018 May 8]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK100423/

7. Cardon LR, Bell JI. Association study designs for complex diseases. Nat Rev Genet. 2001;2:91–9.

8. Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. Nat Genet. 2004;36:512–7.

9. Leslie S, Winney B, Hellenthal G, Davison D, Boumertit A, Day T, et al. The fine scale genetic structure of the British population. Nature. 2015;519:309–14.
10. Chung RC-Y, Bemak F. The Relationship of Culture and Empathy in Cross-Cultural Counseling. J Couns Dev. 2002;80:154–9.
11. Durey A. Reducing racism in Aboriginal health care in Australia: where does cultural education fit? Aust N Z J Public Health. 2010;34 Suppl 1:S87-92.
12. Les Whitbeck B, Chen X, Hoyt DR, Adams GW. Discrimination, historical loss and enculturation: culturally specific risk and resiliency factors for alcohol abuse among American Indians. J Stud Alcohol. 2004;65:409–18.
13. Wexler L. The Importance of Identity, History, and Culture in the Wellbeing of Indigenous Youth. J Hist Child Youth. 2009;2:267–76.
14. Palo JU, Hedman M, Söderholm N, Sajantila A. Repatriation and identification of the Finnish World War II soldiers. Croat Med J. 2007;48:528–35.
15. Soares P, Ermini L, Thomson N, Mormina M, Rito T, Röhl A, et al. Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock. Am J Hum Genet. 2009;84:740–59.
16. Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, et al. Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci. 2011;108:8351–6.
17. Xiang Q-Y, Soltis DE, Soltis PS, Manchester SR, Crawford DJ. Timing the Eastern Asian–Eastern North American Floristic Disjunction: Molecular Clock Corroborates Paleontological Estimates. Mol Phylogenet Evol. 2000;15:462–72.
18. Meyer M, Arsuaga J-L, de Filippo C, Nagel S, Aximu-Petri A, Nickel B, et al. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. Nature. 2016;531:504–7.
19. Thomas KD. Molecular biology and archaeology: A prospectus for inter- disciplinary research. World Archaeol. 1993;25:1–17.
20. Kinross JM, Darzi AW, Nicholson JK. Gut microbiome-host interactions in health and disease. Genome Med. 2011;3:14.
21. Cho I, Blaser MJ. The human microbiome: at the interface of health and disease. Nat Rev Genet. 2012;13:260–70.
22. Zarco M, Vess T, Ginsburg G. The oral microbiome in health and disease and the potential impact on personalized dental medicine. Oral Dis. 2012;18:109–20.
23. Gilbert JA, Blaser MJ, Caporaso JG, Jansson JK, Lynch SV, Knight R. Current understanding of the human microbiome. Nat Med. 2018;24:392–400.

24. Falush D, Wirth T, Linz B, Pritchard JK, Stephens M, Kidd M, et al. Traces of Human Migrations in Helicobacter pylori Populations. Science. 2003;299:1582–5.
25. Moeller AH, Caro-Quintero A, Mjungu D, Georgiev AV, Lonsdorf EV, Muller MN, et al. Cospeciation of gut microbiota with hominids. Science. 2016;353:380–2.
26. Blaser MJ. Who are we? Indigenous microbes and the ecology of human diseases. EMBO Rep. 2006;7:956–60.
27. Levy S. Ancient Gut Microbiomes Shed Light on Modern Disease. Environ Health Perspect. 2013;121:a118.
28. Blaser MJ. Missing microbes: how the overuse of antibiotics is fueling our modern plagues. Macmillan; 2014.
29. Warinner C, Speller C, Collins MJ, Lewis Jr. CM. Ancient human microbiomes. J Hum Evol. 2015;79:125–36.
30. Schnorr SL, Sankaranarayanan K, Lewis CM, Warinner C. Insights into human evolution from ancient and contemporary microbiome studies. Curr Opin Genet Dev. 2016;41:14–26.
31. Ziesemer KA, Mann AE, Sankaranarayanan K, Schroeder H, Ozga AT, Brandt BW, et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. Sci Rep. 2015;5:16498.
32. Velsko IM, Overmyer KA, Speller C, Klaus L, Collins MJ, Loe L, et al. The dental calculus metabolome in modern and historic samples. Metabolomics [Internet]. 2017 [cited 2018 Apr 26];13. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5626792/
33. Hou W, Dong H, Li G, Yang J, Coolen MJL, Liu X, et al. Identification of Photosynthetic Plankton Communities Using Sedimentary Ancient DNA and Their Response to late-Holocene Climate Change on the Tibetan Plateau. Sci Rep. 2014;4:6648.
34. Frisia S, Weyrich LS, Hellstrom J, Borsato A, Golledge NR, Anesio AM, et al. The influence of Antarctic subglacial volcanism on the global iron cycle during the Last Glacial Maximum. Nat Commun. 2017;8:15425.
35. Baer PN, Keyes PH, White CL. Studies on experimental calculus formation in the rat. XII. On the transmissibility of factors affecting dental calculus. J Periodontol. 1968;39:86–8.
36. Dobney K, Brothwell D. A method for evaluating the amount of dental calculus on teeth from archaeological sites. J Archaeol Sci. 1987;14:343–51.
37. Wallis CV, Marshall-Jones ZV, Deusch O, Hughes KR. Canine and Feline Microbiomes. Underst Host-Microbiome Interact - Omics Approach [Internet]. Springer, Singapore; 2017 [cited 2017 Sep 7]. p. 279–325. Available from: https://link.springer.com/chapter/10.1007/978-981-10-5050-3_17

38. Miller SA, McFarlane G, Allen MS. Dental analysis demonstrates variability in diet and health of prehistoric Polynesian pigs. J Archaeol Sci Rep. 2017;15:203–12.
39. Wakefield A, Murch S, Anthony A, Linnell J, Casson D, Malik M, et al. RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. The Lancet. 1998;351:637–41.
40. Flaherty DK. The Vaccine-Autism Connection: A Public Health Crisis Caused by Unethical Medical Practices and Fraudulent Science. Ann Pharmacother. 2011;45:1302–4.
41. Peng Roger. The reproducibility crisis in science: A statistical counterattack. Significance. 2015;12:30–2.
42. Baker M. 1,500 scientists lift the lid on reproducibility. Nat News. 2016;533:452.
43. Fang FC, Casadevall A. Retracted Science and the Retraction Index. Infect Immun. 2011;79:3855–9.
44. Fang FC, Steen RG, Casadevall A. Misconduct accounts for the majority of retracted scientific publications. Proc Natl Acad Sci. 2012;109:17028–33.
45. Sarwar U, Nicolaou M. Fraud and deceit in medical research. J Res Med Sci Off J Isfahan Univ Med Sci. 2012;17:1077–81.
46. Spiegelhalter David. Trust in numbers. J R Stat Soc Ser A Stat Soc. 2017;180:948–65.
47. Eisenhofer R, Cooper A, Weyrich LS. Isolating Viable Ancient Bacteria: What You Put In Is What You Get Out. Genome Announc [Internet]. 2016 [cited 2017 Feb 8];4. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5000818/
48. Eisenhofer R, Cooper A, Weyrich LS. Reply to Santiago-Rodriguez et al.: proper authentication of ancient DNA is essential. FEMS Microbiol Ecol [Internet]. 2017 [cited 2017 Jun 27];93. Available from: https://academic.oup.com/femsec/article/93/5/fix042/3089752/Reply-to-Santiago-Rodriguez-et-al-proper
49. Eisenhofer R, Weyrich LS. Proper Authentication of Ancient DNA Is Still Essential. Genes. 2018;9:122.
50. Austin JJ, Ross AJ, Smith AB, Fortey RA, Thomas RH. Problems of reproducibility – does geologically ancient DNA survive in amber–preserved insects? Proc R Soc Lond B Biol Sci. 1997;264:467–74.
51. Gutiérrez G, Marín A. The most ancient DNA recovered from an amber-preserved specimen may not be as ancient as it seems. Mol Biol Evol. 1998;15:926–9.
52. Hazen RM, Roedder E. Biogeology. How old are bacteria from the Permian age? Nature. 2001;411:155–6.

53. Nickle DC, Learn GH, Rain MW, Mullins JI, Mittler JE. Curiously Modern DNA for a ``250 Million-Year-Old'' Bacterium. J Mol Evol. 2002;54:134–7.
54. Gilbert MTP, Cuccui J, White W, Lynnerup N, Titball RW, Cooper A, et al. Absence of Yersinia pestis-specific DNA in human teeth from five European excavations of putative plague victims. Microbiology. 2004;150:341–54.
55. Gilbert MTP, Cuccui J, White W, Lynnerup N, Titball RW, Cooper A, et al. Response to Drancourt and Raoult. Microbiology. 2004;150:264–5.
56. Vergnaud G. Yersinia pestis Genotyping. Emerg Infect Dis. 2005;11:1317–9.
57. Shapiro B, Rambaut A, Gilbert MTP. No proof that typhoid caused the Plague of Athens (a reply to Papagrigorakis et al.). Int J Infect Dis. 2006;10:334–5.
58. Weyrich LS, Llamas B, Cooper A. Reply to Santiago-Rodriguez et al.: Was luxS really isolated from 25- to 40-million-year-old bacteria? FEMS Microbiol Lett. 2014;353:85–6.
59. Aagaard K, Ma J, Antony KM, Ganu R, Petrosino J, Versalovic J. The placenta harbors a unique microbiome. Sci Transl Med. 2014;6:237ra65.
60. Amarasekara R, Jayasekara RW, Senanayake H, Dissanayake VHW. Microbiome of the placenta in pre-eclampsia supports the role of bacteria in the multifactorial cause of pre-eclampsia. J Obstet Gynaecol Res. 2015;41:662–9.
61. Bassols J, Serino M, Carreras-Badosa G, Burcelin R, Blasco-Baque V, Lopez-Bermejo A, et al. Gestational diabetes is associated with changes in placental microbiota and microbiome. Pediatr Res. 2016;80:777–84.
62. Schloss PD, Handelsman J. Status of the Microbial Census. Microbiol Mol Biol Rev. 2004;68:686–91.
63. Curtis TP, Head IM, Lunn M, Woodcock S, Schloss PD, Sloan WT. What is the extent of prokaryotic diversity? Philos Trans R Soc B Biol Sci. 2006;361:2023–37.
64. Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng J-F, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499:431–7.
65. Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523:208–11.
66. Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial Dark Matter Investigations: How Microbial Studies Transform Biological Knowledge and Empirically Sketch a Logic of Scientific Discovery. Genome Biol Evol. 2018;10:707–15.
67. Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci. 2007;104:11889–94.

68. Lloyd-Price J, Mahurkar A, Rahnavard G, Crabtree J, Orvis J, Hall AB, et al. Strains, functions and dynamics in the expanded Human Microbiome Project. Nature. 2017;550:61–6.
69. Chen T, Yu W-H, Izard J, Baranova OV, Lakshmanan A, Dewhirst FE. The Human Oral Microbiome Database: a web accessible resource for investigating oral microbe taxonomic and genomic information. Database [Internet]. 2010 [cited 2018 Feb 9];2010. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baq013/405450
70. Consortium THMP. Structure, function and diversity of the healthy human microbiome. Nature. 2012;486:207–14.
71. Campbell AG, Campbell JH, Schwientek P, Woyke T, Sczyrba A, Allman S, et al. Multiple Single-Cell Genomes Provide Insight into Functions of Uncultured Deltaproteobacteria in the Human Oral Cavity. PLOS ONE. 2013;8:e59361.
72. Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, et al. Use of Ichip for High-Throughput In Situ Cultivation of "Uncultivable" Microbial Species. Appl Environ Microbiol. 2010;76:2445–50.
73. Ma L, Kim J, Hatzenpichler R, Karymov MA, Hubert N, Hanan IM, et al. Gene-targeted microfluidic cultivation validated by isolation of a gut bacterium listed in Human Microbiome Project's Most Wanted taxa. Proc Natl Acad Sci. 2014;111:9768–73.
74. Kim HJ, Li H, Collins JJ, Ingber DE. Contributions of microbiome and mechanical deformation to intestinal bacterial overgrowth and inflammation in a human gut-on-a-chip. Proc Natl Acad Sci U S A. 2016;113:E7-15.
75. Imelfort M, Parks D, Woodcroft BJ, Dennis P, Hugenholtz P, Tyson GW. GroopM: an automated tool for the recovery of population genomes from related metagenomes. PeerJ [Internet]. 2014 [cited 2017 Oct 13];2. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4183954/
76. Parks DH, Rinke C, Chuvochina M, Chaumeil P-A, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2:1533.
77. O'Fallon BD, Fehren-Schmitz L. Native Americans experienced a strong population bottleneck coincident with European contact. Proc Natl Acad Sci U S A. 2011;108:20444–8.
78. Segata N. Gut Microbiome: Westernization and the Disappearance of Intestinal Diversity. Curr Biol. 2015;25:R611–3.
79. Broussard JL, Devkota S. The changing microbial landscape of Western society: Diet, dwellings and discordance. Mol Metab. 2016;5:737–42.
80. Blaser MJ. Antibiotic use and its consequences for the normal microbiome. Science. 2016;352:544–5.

81. Korpela K, Salonen A, Virta LJ, Kekkonen RA, Forslund K, Bork P, et al. Intestinal microbiome is related to lifetime antibiotic use in Finnish pre-school children. Nat Commun. 2016;7:10410.

82. Sonnenburg ED, Smits SA, Tikhonov M, Higginbottom SK, Wingreen NS, Sonnenburg JL. Diet-induced extinctions in the gut microbiota compound over generations. Nature. 2016;529:212–5.

83. Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. bioRxiv. 2016;050559.

84. Key FM, Posth C, Krause J, Herbig A, Bos KI. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. Trends Genet [Internet]. 2017 [cited 2017 Jul 9];0. Available from: http://www.cell.com/trends/genetics/abstract/S0168-9525(17)30086-0

85. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. PeerJ. 2014;2:e675.

86. Kryukov K, Imanishi T. Human Contamination in Public Genome Assemblies. PLOS ONE. 2016;11:e0162424.

87. Lu J, Salzberg S. Removing Contaminants from Metagenomic Databases. bioRxiv. 2018;261859.

88. Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. Genome Res. 2004;14:1394–403.

89. Darling AE, Mau B, Perna NT. progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. PLOS ONE. 2010;5:e11147.

90. Monod J. The Growth of Bacterial Cultures. Annu Rev Microbiol. 1949;3:371–94.

91. Lehtonen J, Lanfear R. Generation time, life history and the substitution rate of neutral mutations. Biol Lett [Internet]. 2014 [cited 2018 Apr 23];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4261869/

92. Weller C, Wu M. A generation-time effect on the rate of molecular evolution in bacteria. Evolution. 2015;69:643–52.

93. Schierup MH, Hein J. Consequences of recombination on traditional phylogenetic analysis. Genetics. 2000;156:879–91.

94. Ochman H, Lawrence JG, Groisman EA. Lateral gene transfer and the nature of bacterial innovation. Nature. 2000;405:299–304.

95. Håvarstein LS, Hakenbeck R, Gaustad P. Natural competence in the genus Streptococcus: evidence that streptococci can change pherotype by interspecies recombinational exchanges. J Bacteriol. 1997;179:6589–94.
96. Treangen TJ, Ambur OH, Tonjum T, Rocha EP. The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biol. 2008;9:R60.
97. Koehler A, Karch H, Beikler T, Flemmig TF, Suerbaum S, Schmidt H. Multilocus sequence analysis of Porphyromonas gingivalis indicates frequent recombination. Microbiol Read Engl. 2003;149:2407–15.
98. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;43:e15–e15.
99. Martin DP, Murrell B, Golden M, Khoosal A, Muhire B. RDP4: Detection and analysis of recombination patterns in virus genomes. Virus Evol [Internet]. 2015 [cited 2018 Apr 23];1. Available from: https://academic.oup.com/ve/article/1/1/vev003/2568683
100. Farrer AG, Bekvalac J, Redfern R, Gully N, Dobney K, Cooper A, et al. Biological and cultural drivers of microbiota in Medieval and Post-Medieval London, UK. PhD Thesis: Ancient DNA studies of dental calculus. 2017;
101. Proctor DM, Fukuyama JA, Loomer PM, Armitage GC, Lee SA, Davis NM, et al. A spatial gradient of bacterial diversity in the human oral cavity shaped by salivary flow. Nat Commun. 2018;9:681.
102. Liu B, Faller LL, Klitgord N, Mazumdar V, Ghodsi M, Sommer DD, et al. Deep Sequencing of the Oral Microbiome Reveals Signatures of Periodontal Disease. PLoS ONE. 2012;7:e37919.
103. Abusleme L, Dupuy AK, Dutzan N, Silva N, Burleson JA, Strausbaugh LD, et al. The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. ISME J. 2013;7:1016–25.
104. Camelo-Castillo AJ, Mira A, Pico A, Nibali L, Henderson B, Donos N, et al. Subgingival microbiota in health compared to periodontitis and the influence of smoking. Front Microbiol [Internet]. 2015 [cited 2017 Nov 29];6. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2015.00119/full
105. Califf KJ, Schwarzberg-Lipson K, Garg N, Gibbons SM, Caporaso JG, Slots J, et al. Multi-omics Analysis of Periodontal Pocket Microbial Communities Pre- and Posttreatment. mSystems. 2017;2:e00016-17.
106. Boutin S, Hagenfeld D, Zimmermann H, El Sayed N, Höpker T, Greiser HK, et al. Clustering of Subgingival Microbiota Reveals Microbial Disease Ecotypes Associated with

Clinical Stages of Periodontitis in a Cross-Sectional Study. Front Microbiol [Internet]. 2017 [cited 2018 Apr 18];8. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2017.00340/full#h13

107. Chen C, Hemme C, Beleno J, Shi ZJ, Ning D, Qin Y, et al. Oral microbiota of periodontal health and disease and their changes after nonsurgical periodontal therapy. ISME J. 2018;1.

108. Socransky S s., Haffajee A d., Cugini M a., Smith C, Kent RL. Microbial complexes in subgingival plaque. J Clin Periodontol. 1998;25:134–44.

109. Darveau RP. Periodontitis: a polymicrobial disruption of host homeostasis. Nat Rev Microbiol. 2010;8:481.

110. Hajishengallis G, Lamont RJ. Beyond the red complex and into more complexity: the polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. Mol Oral Microbiol. 2012;27:409–19.

111. Lamont RJ, Hajishengallis G. Polymicrobial synergy and dysbiosis in inflammatory disease. Trends Mol Med. 2015;21:172–83.

112. Marsh PD, Zaura E. Dental biofilm: ecological interactions in health and disease. J Clin Periodontol. 2017;44:S12–22.

113. Whittaker DK, Griffiths S, Robson A, Roger-Davies P, Thomas G, Molleson T. Continuing tooth eruption and alveolar crest height in an eighteenth-century population from Spitalfields, east London. Arch Oral Biol. 1990;35:81–5.

114. Clarke NG, Hirsch RS. Two critical confounding factors in periodontal epidemiology. Community Dent Health. 1992;9:133–41.

115. Raitapuro-Murray T, Molleson TI, Hughes FJ. The prevalence of periodontal disease in a Romano-British population c. 200-400 AD. Br Dent J. 2014;217:459–66.

116. Highfield J. Diagnosis and classification of periodontal disease. Aust Dent J. 2009;54 Suppl 1:S11-26.

117. Chang A, Davis C, Bo C, Yang F, Zhao H, Xu J, et al. Predictive modeling of gingivitis severity and susceptibility via oral microbiota. ISME J. 2014;8:1768.

118. Shaw L, Harjunmaa U, Doyle R, Mulewa S, Charlie D, Maleta K, et al. Distinguishing the Signals of Gingivitis and Periodontitis in Supragingival Plaque: a Cross-Sectional Cohort Study in Malawi. Appl Environ Microbiol. 2016;82:6057–67.

119. Huang S, Li Z, He T, Bo C, Chang J, Li L, et al. Microbiota-based Signature of Gingivitis Treatments: A Randomized Study. Sci Rep. 2016;6:24705.

120. Nguyen- Hieu Tung, Khelaifia Saber, Aboudharam Gerard, Drancourt Michel. Methanogenic archaea in subgingival sites: a review. APMIS. 2012;121:467–77.

121. Pérez-Chaparro PJ, Gonçalves C, Figueiredo LC, Faveri M, Lobão E, Tamashiro N, et al. Newly Identified Pathogens Associated with Periodontitis. J Dent Res. 2014;93:846–58.
122. Horz H-P, Conrads G. Methanogenic Archaea and oral infections — ways to unravel the black box. J Oral Microbiol [Internet]. 2011 [cited 2017 Sep 29];3. Available from: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3086593/
123. Lepp PW, Brinig MM, Ouverney CC, Palm K, Armitage GC, Relman DA. Methanogenic Archaea and human periodontal disease. Proc Natl Acad Sci U S A. 2004;101:6176–81.
124. Yamabe K, Maeda H, Kokeguchi S, Tanimoto I, Sonoi N, Asakawa S, et al. Distribution of Archaea in Japanese patients with periodontitis and humoral immune response to the components. FEMS Microbiol Lett. 2008;287:69–75.
125. Schuenemann VJ, Singh P, Mendum TA, Krause-Kyora B, Jäger G, Bos KI, et al. Genome-Wide Comparison of Medieval and Modern Mycobacterium leprae. Science. 2013;341:179–83.
126. Pedersen JS, Valen E, Velazquez AMV, Parker BJ, Rasmussen M, Lindgreen S, et al. Genome-wide nucleosome map and cytosine methylation levels of an ancient human genome. Genome Res. 2014;24:454–66.
127. Pereira SL, Grayling RA, Lurz R, Reeve JN. Archaeal nucleosomes. Proc Natl Acad Sci U S A. 1997;94:12633–7.
128. Reeve JN, Bailey KA, Li W-T, Marc F, Sandman K, Soares DJ. Archaeal histones: structures, stability and DNA binding. Biochem Soc Trans. 2004;32:227–30.
129. Henneman B, Dame RT. Archaeal histones: dynamic and versatile genome architects. Microbiol 2015 Vol 1 Pages 72-81 [Internet]. 2015 [cited 2017 Jul 11]; Available from: http://www.aimspress.com/article/10.3934/microbiol.2015.1.72/fulltext.html
130. Hall MW, Singh N, Ng KF, Lam DK, Goldberg MB, Tenenbaum HC, et al. Inter-personal diversity and temporal dynamics of dental, tongue, and salivary microbiota in the healthy oral cavity. Npj Biofilms Microbiomes. 2017;3:2.
131. Gosalbes MJ, Abellan JJ, Durban A, Perez-Cobas AE, Latorre A, Moya A. Metagenomics of human microbiome: beyond 16s rDNA. Clin Microbiol Infect. 2012;18:47–9.
132. Duran-Pinedo AE, Frias-Lopez J. Beyond microbial community composition: functional activities of the oral microbiome in health and disease. Microbes Infect Inst Pasteur. 2015;17:505–16.
133. Abubucker S, Segata N, Goll J, Schubert AM, Izard J, Cantarel BL, et al. Metabolic Reconstruction for Metagenomic Data and Its Application to the Human Microbiome. PLOS Comput Biol. 2012;8:e1002358.

134. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
135. Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI. The Human Microbiome Project [Internet]. Nature. 2007 [cited 2018 Apr 23]. Available from: https://www.nature.com/articles/nature06244
136. Warinner C, Hendy J, Speller C, Cappellini E, Fischer R, Trachsel C, et al. Direct evidence of milk consumption from ancient human dental calculus. Sci Rep. 2014;4:7104.
137. Hendy J, Welker F, Demarchi B, Speller C, Warinner C, Collins MJ. A guide to ancient protein studies. Nat Ecol Evol. 2018;1.
138. Ellrott K, Jaroszewski L, Li W, Wooley JC, Godzik A. Expansion of the Protein Repertoire in Newly Explored Environments: Human Gut Microbiome Specific Protein Families. PLoS Comput Biol [Internet]. 2010 [cited 2018 May 8];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2880560/
139. Yasui M, Furuta M, Eshima N, Takeshita T, Saeki Y, Yamashita Y, et al. Dental plaque development on a hydroxyapatite disk in young adults observed by using a barcoded pyrosequencing approach. Sci Rep. 2015;5:srep08136.
140. White DJ. Dental calculus: recent insights into occurrence, formation, prevention, removal and oral health effects of supragingival and subgingival deposits. Eur J Oral Sci. 1997;105:508–22.
141. Simón-Soro Á, Tomás I, Cabrera-Rubio R, Catalan MD, Nyvad B, Mira A. Microbial Geography of the Oral Cavity. J Dent Res. 2013;92:616–21.
142. Moynihan PJ, Kelly S a. M. Effect on caries of restricting sugars intake: systematic review to inform WHO guidelines. J Dent Res. 2014;93:8–18.
143. Baker JL, Faustoferri RC, Quivey RG. Acid-adaptive mechanisms of Streptococcus mutans–the more we know, the more we don't. Mol Oral Microbiol. 2017;32:107–17.
144. Carvalho JC. Caries Process on Occlusal Surfaces: Evolving Evidence and Understanding. Caries Res. 2014;48:339–46.
145. Keller MK, Kressirer CA, Belstrøm D, Twetman S, Tanner ACR. Oral microbial profiles of individuals with different levels of sugar intake. J Oral Microbiol [Internet]. 2017 [cited 2017 Dec 22];9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5560414/
146. Dawes C, Macpherson LMD. The Distribution of Saliva and Sucrose Around the Mouth During the Use of Chewing Gum and the Implications for the Site-specificity of Caries and Calculus Deposition. J Dent Res. 1993;72:852–7.
147. Dawes C. Why does supragingival calculus form preferentially on the lingual surface of the 6 lower anterior teeth? J Can Dent Assoc. 2006;72:923–6.

148. Gomez A, Espinoza JL, Harkins DM, Leong P, Saffery R, Bockmann M, et al. Host Genetic Control of the Oral Microbiome in Health and Disease. Cell Host Microbe. 2017;22:269-278.e3.
149. Demmitt BA, Corley RP, Huibregtse BM, Keller MC, Hewitt JK, McQueen MB, et al. Genetic influences on the human oral microbiome. BMC Genomics. 2017;18:659.
150. Shaw L, Ribeiro ALR, Levine AP, Pontikos N, Balloux F, Segal AW, et al. The Human Salivary Microbiome Is Shaped by Shared Environment Rather than Genetics: Evidence from a Large Family of Closely Related Individuals. mBio. 2017;8:e01237-17.
151. Welch JLM, Rossetti BJ, Rieken CW, Dewhirst FE, Borisy GG. Biogeography of a human oral microbiome at the micron scale. Proc Natl Acad Sci. 2016;113:E791–800.

Appendix I

Isolating Viable Ancient Bacteria: What You Put In Is What You Get Out

AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

LETTER TO THE EDITOR

Isolating Viable Ancient Bacteria: What You Put In Is What You Get Out

Raphael Eisenhofer, Alan Cooper, Laura S. Weyrich

Australian Centre for Ancient DNA, University of Adelaide, Adelaide, Australia

In the recent publication "Draft Genome Sequence of *Enterococcus faecium* Strain 58m, Isolated from Intestinal Tract Content of a Woolly Mammoth, *Mammuthus primigenius*" in *Genome Announcements* (1), Goncharov et al. claim to have isolated and grown in pure culture a 28,000-year-old *Enterococcus faecium* strain. However, the authors ignored a breadth of literature about the authentication of ancient DNA, failed to adhere to recommended guidelines (2), and did not provide the appropriate experimental controls and analyses required to substantiate such a claim. Here, we present a subsequent reanalysis of the Goncharov et al. isolate and demonstrate by multilocus sequence typing (MLST) that this strain likely represents a modern contaminant.

Previous efforts aimed at isolating viable ancient bacteria have been consistently controversial (3). Viable bacteria have been reported from a 250 million-year-old salt crystal (4) and 25- to 40 million-year-old amber (5). These unlikely findings have not been independently replicated, and failed molecular phylogenetic tests (6–8). In light of such dubious claims, a set of rigorous authentication criteria have been proposed (2). These include evolutionary rates tests, whereby phylogenetic comparisons of the ancient organism with its modern counterparts are expected to show substantial genetic differences, accumulated through time.

In the Goncharov et al. study, the authors admit that *E. faecium* is a common member of the human gut community and can be found from numerous environmental sources, yet strangely they did nothing to prevent or control for modern contamination at various stages of their experiment. Modern contaminants can enter during the sampling procedure (2) or during laboratory analysis (i.e., culturing or DNA sequencing). Contamination during laboratory analysis is especially probable when the isolate is cultured using broad-spectrum media (2), as used by the authors. Clearly, the authors should have considered these factors and demonstrated or minimally investigated to determine that their isolate did not represent a modern human or environmental contaminant, something they failed to do.

To test the authenticity of the authors' claims, we queried the genome assembly of the "ancient" *E. faecium* isolate against published sequences in the *E. faecium* MLST database (http://pubmlst.org/efaecium/), which contains >2,800 modern *E. faecium* isolates. The MLST sequence from the putatively ancient *E. faecium* isolate matches the previously identified sequence type 32 (ST32) with 100% sequence homology; this is unexpected if the genome is ancient. Modern isolates of ST32 are known from the Russian Federation, where this study took place. If the bacterium was an ancient resident of the mammoth gut, it should not be identical to a modern human isolate, given that many gut microorganisms coevolved with their hosts and that humans and mammoths diverged over 100 million years ago (9). The lack of even a single nucleotide difference within seven genetic loci, coupled with the fact that this bacterium is commonly found in the modern human gut community and other environmental sources, is damning evidence that the authors' isolate represents a modern contaminant.

The authors' "ancient" *E. faecium* isolate is highly similar to modern human isolates and is therefore almost certainly not an ancient mammoth strain.

REFERENCES

1. **Goncharov A, Grigorjev S, Karaseva A, Kolodzhieva V, Azarov D, Akhremenko Y, Tarasova L, Tikhonov A, Masharskiy A, Zueva L, Suvorov A.** 2016. Draft genome sequence of *Enterococcus faecium* strain 58m, isolated from intestinal tract content of a woolly mammoth, *Mammuthus primigenius*. Genome Announc **4**(1):e01706-15. http://dx.doi.org/10.1128/genomeA.01706-15.
2. **Hebsgaard MB, Phillips MJ, Willerslev E.** 2005. Geologically ancient DNA: fact or artefact? Trends Microbiol **13**:212–220. http://dx.doi.org/10.1016/j.tim.2005.03.010.
3. **Fischman J.** 1995. Have 25-million-year-old bacteria returned to life? Science **268**:977. http://dx.doi.org/10.1126/science.7754393.
4. **Vreeland RH, Rosenzweig WD, Powers DW.** 2000. Isolation of a 250 million-year-old halotolerant bacterium from a primary salt crystal. Nature **407**:897–900. http://dx.doi.org/10.1038/35038060.
5. **Cano RJ, Borucki MK.** 1995. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science **268**:1060–1064. http://dx.doi.org/10.1126/science.7538699.
6. **Nickle DC, Learn GH, Rain MW, Mullins JI, Mittler JE.** 2002. Curiously modern DNA for a "250 million-year-old" bacterium. J Mol Evol **54**:134–137. http://dx.doi.org/10.1007/s00239-001-0025-x.
7. **Yousten AA, Rippere KE.** 1997. DNA similarity analysis of a putative ancient bacterial isolate obtained from amber. FEMS Microbiol Lett **152**:345–347. http://dx.doi.org/10.1111/j.1574-6968.1997.tb10450.x.
8. **Weyrich LS, Llamas B, Cooper A.** 2014. Reply to Santiago-Rodriguez *et al.*: was *luxS* really isolated from 25- to 40-million-year-old bacteria? FEMS Microbiol Lett **353**:85–86. http://dx.doi.org/10.1111/1574-6968.12415.
9. **Meredith RW, Janečka JE, Gatesy J, Ryder OA, Fisher CA, Teeling EC, Goodbla A, Eizirik E, Simão TL, Stadler T, Rabosky DL, Honeycutt RL, Flynn JJ, Ingram CM, Steiner C, Williams TL, Robinson TJ, Burk-Herrick A, Westerman M, Ayoub NA, Springer MS, Murphy WJ.** 2011. Impacts of the Cretaceous terrestrial revolution and KPg extinction on mammal diversification. Science **334**:521–524. http://dx.doi.org/10.1126/science.1211028.

Published 25 August 2016

Citation Eisenhofer R, Cooper A, Weyrich LS. 2016. Isolating viable ancient bacteria: what you put in is what you get out. Genome Announc 4(4):e00712-16. doi:10.1128/genomeA.00712-16.

Address correspondence to Raphael Eisenhofer, raphael.eisenhoferphilipona@adelaide.edu.au.

For the author reply, see doi:10.1128/genomeA.00734-16.

Appendix II

—

Reply to Santiago-Rodriguez *et al.*: proper authentication of ancient DNA is essential

FEMS Microbiology Ecology, 93, 2017, fix042

doi: 10.1093/femsec/fix042
Advance Access Publication Date: 24 March 2017
Letter to the Editor

LETTER TO THE EDITOR

Reply to Santiago-Rodriguez *et al.*: proper authentication of ancient DNA is essential

Raphael Eisenhofer*†, Alan Cooper and Laura S Weyrich

Australian Centre for Ancient DNA, University of Adelaide, Adelaide 5005, Australia

*Corresponding author: Department of Genetics & Evolution, Darling Building, The University of Adelaide, North Terrace, Adelaide, South Australia, 5005, Australia. Tel: +61 8 8313 3952; E-mail: raph.eisenhofer@gmail.com

One sentence summary: Santiago-Rodriguez et al. (2016) failed to properly control and authenticate their ancient DNA study of South American mummies.

Editor: Marcus Horn

†Raphael Eisenhofer, http://orcid.org/0000-0002-3843-0749

Santiago-Rodriguez *et al.* (2016) attempt to characterize the gut microbiome from pre-Columbian Andean mummies. However, they fail to properly use basic standards required for the authentication of ancient bacterial DNA, compromising the authenticity of their results and setting an unacceptable standard for future work.

Authentic ancient DNA research is extremely difficult. This is especially true when studying ancient microorganisms, as their damaged and fragmented DNA is in low abundance relative to modern microorganisms which coat virtually every surface. Reagent and laboratory DNA contamination has been demonstrated to routinely impact microbiome analyses (Salter *et al.* 2014; Glassing *et al.* 2016; Lauder *et al.* 2016), and is especially problematic for samples with low biomass or endogenous DNA—such as ancient microbial samples. To identify and control for such contamination, multiple extraction blank controls and PCR-negatives need to be performed and sequenced, and any detected taxa should be subtracted from the sample results. Additionally, the latter must be screened against the growing list of common laboratory and reagent contaminants (Salter *et al.* 2014; Glassing *et al.* 2016; Lauder *et al.* 2016). Santiago-Rodriguez *et al.* (2016) fail to follow such precautions, improperly use a method of ancient DNA authentication (MapDamage, discussed below) and apply flawed methodologies, invalidating their results and potentially encouraging further problematic analyses.

Issues with 16S rRNA methodology

The authors state that their extraction controls did not show bands on agarose gels (with no supporting images in SI); however, this ignores the fact that bacterial contaminants are often present at levels not visible on relatively insensitive agarose gels. The sensitivity of 16S rRNA PCR (especially after 30 cycles of amplification) will nearly always result in the detection of contaminants, even if below visible levels. Such controls must be sequenced in the same way as the biological samples. The absence of sequenced extraction blank and negative PCR controls is a critical oversight given that the reported results include common contaminants of laboratory environments and reagents (Salter *et al.* 2014; Glassing *et al.* 2016; Lauder *et al.* 2016). These include: *Bacillaceae Bradyrhizobiacea, Clostridiaceae, Sphingobacteriaceae, Streptococcus*. To authenticate their results, the authors applied SourceTracker analysis to determine whether portions of the microbial community originate from the gut or from other sources, and conclude: 'The majority of the 16S rRNA gene sequences in mummy FI3 matched modern gut microbiomes, and those of mummies FI9 and FI12 did not match any of the sources included in the analysis, suggesting that no modern sources of contamination contributed to the findings presented in our study'. This is clearly an unjustified conclusion, as only four sources were tested (human skin, gut, oral and soil samples), and therefore cannot account for other common sources of modern contamination (e.g. laboratory, reagents, air). Another critical issue is the use of 16S ribosomal RNA sequencing to describe these samples, despite the recent recommendation against using this method to reconstruct ancient microbiomes due to known taphonomic biases (Ziesemer *et al.* 2015). The authors state that the V4 region of the 16S rRNA gene is fine for ancient DNA as it is 'within the recommended length for ancient DNA analyses'. This is clearly incorrect given that the V4 region is ~290 bp, and that the mean length of authentic ancient DNA typically ranges between 50 and 160 bp (Knapp and

Received: 22 February 2017; **Accepted:** 20 March 2017

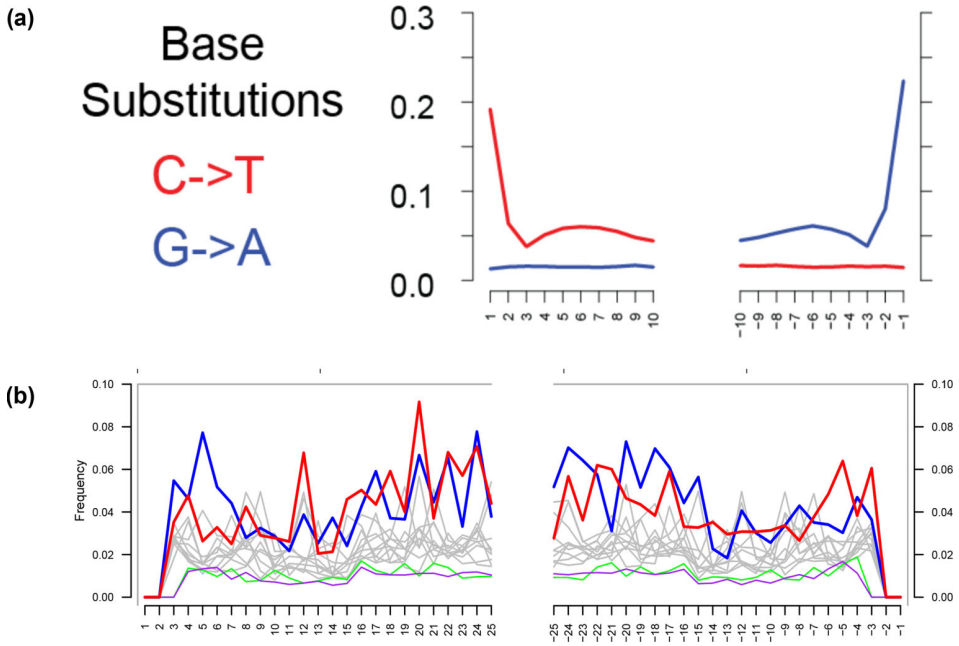


Figure 1. (**a**) Characteristic base substitution pattern of ancient DNA: increase of C→T and G→A substitutions at the 5′ and 3′ regions of the sequenced DNA, respectively (figure taken from Ziesemer *et al.* 2015). (**b**) Santiago-Rodriguez *et al.*'s MapDamage plot, showing no signs of substitutions at the DNA termini.

Hofreiter 2010). Targeting 16S rRNA markers that are longer than the expected fragment size can preferentially amplify modern contaminant sequences, increasing the representation of contaminant taxa over endogenous ones.

Issues with shotgun sequencing methodology

Again, the authors failed to control for modern DNA contamination by not making libraries of and sequencing their extraction blank controls. The authors then attempt to validate some of their 'ancient' microorganisms by performing MapDamage analyses (Ginolhac *et al.* 2011)—a standard in the field of ancient DNA. The authors state 'MapDamage analyses were performed with the contigs as described previously' citing the original MapDamage paper but providing no details of how the analyses were performed, such as what reference sequences were used. MapDamage requires DNA reads, a reference genome, and adequate coverage in order to quantify the C→T and G→A substitutions typical of ancient DNA damage. When MapDamage is used correctly on ancient DNA, an increase of C→T and G→A substitutions is observed at the 5′ and 3′ regions, respectively, of the sequenced DNA fragment (Fig. 1a). The authors do not observe such pattern, likely due to improper use of the software, or the fact that the species of interest are modern contaminants (Fig. 1b). The authors admit that no such damage patterns were detected in their metagenomes, but justify this by asserting that MapDamage is not useful in ancient microbiome studies. They incorrectly claim that this was previously identified by Ziesemer *et al.* (2015), when the latter do not state this and actually used MapDamage to authenticate ancient microbial DNA (Ziesemer *et al.* 2015). It appears that neither the authors nor reviewers understand the basis of using DNA damage to authenticate ancient DNA, and the lack of such damage patterns severely undermines the credibility of the results.

Conclusions

This is not the first time that the authors have failed to properly control an ancient DNA study. A contentious claim of the isolation of *luxS* from 25- to 40-million-year-old bacteria (Santiago-Rodriguez *et al.* 2014) also failed to provide sufficient controls to substantiate such a claim (Weyrich *et al.* 2014). Proper ancient DNA authentication is essential to the integrity of this field. Within the broader field of ancient DNA, a series of high-profile publications from the 1990's (of which a co-author of this study was part of) failed to provide adequate controls or authentication, and are now widely discredited. These studies damaged the credibility of the field, and wasted valuable time and money. We hope that history does not continue to repeat itself, and that editors, reviewers and researchers learn from this example to prevent this from happening again.

REFERENCES

Ginolhac A, Rasmussen M, Gilbert MT *et al.* mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* 2011;**27**:2153–5.

Glassing A, Scot ED, Galandiuk S *et al.* Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog* 2016;**8**:24.

Knapp M, Hofreiter M. Next generation sequencing of ancient DNA: requirements, strategies and perspectives. *Genes* 2010;**1**:227–43.

Lauder AP, Roche AM, Sherrill-Mix S *et al.* Comparison of placenta samples with contamination controls does not

provide evidence for a distinct placenta microbiota. *Microbiome* 2016;**4**:29.

Salter SJ, Cox, MJ, Turek, EM *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol* 2014;**12**:87.

Santiago-Rodriguez TM, Fornaciari G, Luciani S *et al.* Taxonomic and predicted metabolic profiles of the human gut microbiome in pre-columbian mummies. *FEMS Microbiol Ecol* 2016;**92**. doi:10.1093/femsec/fiw182.

Santiago-Rodriguez TM, Patrício AR, Rivera JI *et al.* luxS in bacteria isolated from 25- to 40-million-year-old amber. *FEMS Microbiol Lett* 2014;**350**:117–24.

Weyrich LS, Llamas, B, Cooper, A. Reply to Santiago-Rodriguez et al.: Was luxS really isolated from 25- to 40-million-year-old bacteria? *FEMS Microbiol Lett* 2014;**353**:85–6.

Ziesemer KA, Mann AE, Sankaranarayanan K *et al.* Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci Rep* 2015;**5**:16498.

Appendix III

—

Proper Authentication of Ancient DNA Is Still Essential

genes

MDPI

Comment

Proper Authentication of Ancient DNA Is Still Essential

Raphael Eisenhofer * and Laura S. Weyrich

Department of Genetics & Evolution, Darling Building, The University of Adelaide, North Terrace, Adelaide, SA 5005, Australia; laura.weyrich@gmail.com

* Correspondence: raphael.eisenhoferphilipona@adelaide.edu.au; Tel.: +61-4-3582-4262

Received: 31 January 2018; Accepted: 19 February 2018; Published: 26 February 2018

Keywords: ancient DNA; microbiome; paleomicrobiology; microbial ecology

Santiago-Rodriguez et al. [1] report on the putative gut microbiome and resistome of Inca and Italian mummies, and find that Italian mummies exhibit higher bacterial diversity compared to the Inca mummies. However, contaminant taxa in their negative control account for most of the biological signal observed. In addition, they fail to properly apply field-standard ancient DNA authentication techniques to their data and self-plagiarize a previously published figure. Poor standards in paleomicrobiological research are currently plaguing the field, despite numerous warnings [2–4] and reviews [5–8] on best practice.

DNA contamination from museums, curators, scientists, soil, and even the laboratory can drive signals present in modern and ancient metagenomics data sets [3,7,9–16]. Therefore, explicit rules and standards to avoid falsely reporting contaminants in metagenomics datasets have been put forth [3,4,6–8,17]. These standards typically include sampling and extraction blank controls (e.g., tubes processed without the addition of biological samples) to monitor contaminant DNA and correctly attribute its contribution in subsequent analyses. In the study by Santiago-Rodriquez et al. [1], a non-template or blank control was included in their 16S analysis. However, the authors failed to explore the contaminant species within this control during their analysis of differences between Incan and Italian mummies. We explored the taxa present within their blank control (Supplementary_Dataset_2.txt from their publication) and compared it to those identified within the mummies. We found that laboratory contaminants present within their blank control are driving the differences between Incan and Italian mummies (Figure 1). For example, the five most abundant taxa identified in the Italian mummies (*Sphingomonadales*, *Pseudomonadales*, *Rhizobiales*, *Bacillales*, and *Clostridiales* species) are all found in the blank control. It is also worth noting that these taxa have previously been identified as common laboratory or reagent contaminants in numerous studies [3,14,15]. This strongly suggests that the cultural differences reported by the authors are likely the result of laboratory contamination and calls into question the validity of their subsequent analyses.

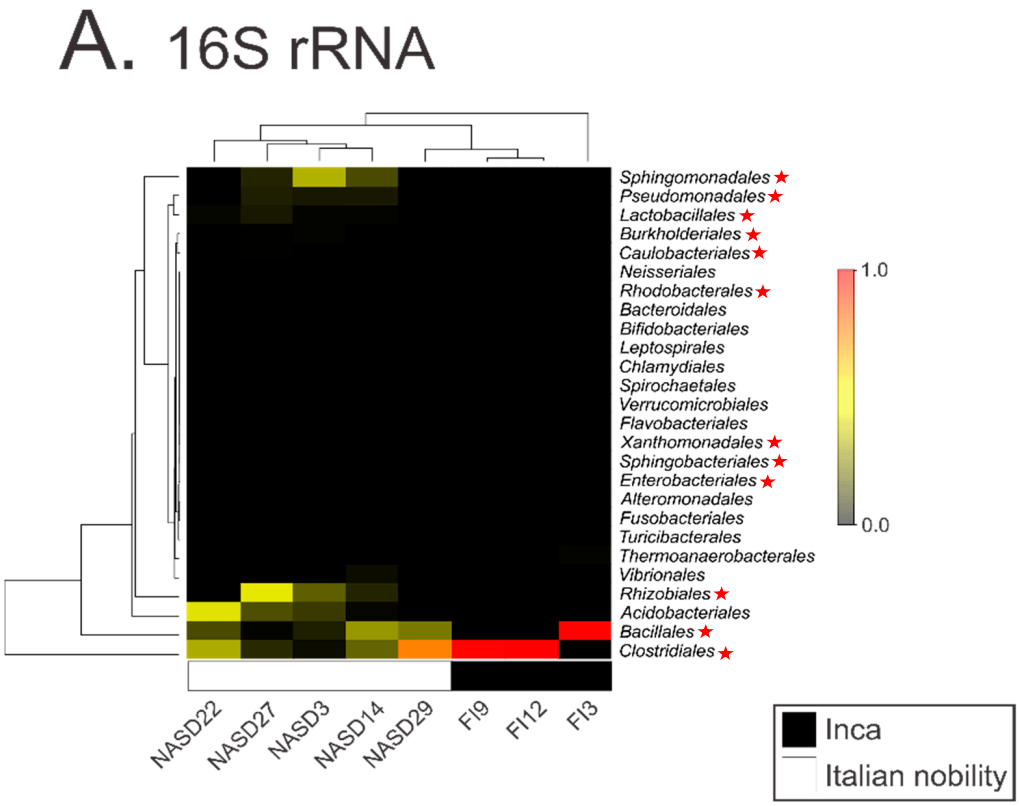


Figure 1. An altered reproduction of Figure 1A from Santiago-Rodriguez et al. [1] where taxa identified in the 16S rRNA blank control are identified in the mummy samples by red stars. The highest-abundance taxa identified in the 16S rRNA data are also present in the 16S rRNA blank control.

The authors then attempt to use MapDamage to assess the authenticity of their shotgun metagenomic ancient DNA; this tool is widely used within the paleomicrobiological field for detecting patterns of cytosine deamination that are characteristic of authentic ancient DNA [18]. Critically, the authors did not provide details as to how they ran the analysis; MapDamage calculates the deamination rate by comparing a reference genome to the mapped target sequences present in a given biological sample (i.e., the reference and target species are typically reported for the analysis). Despite this lack of information, the MapDamage plot provided by the authors in their supplementary information (Figure 2A) is identical to one in a previous publication by the team [4] (Figure 2B), suggesting that the authors self-plagiarized this figure and did not in fact run the analysis.

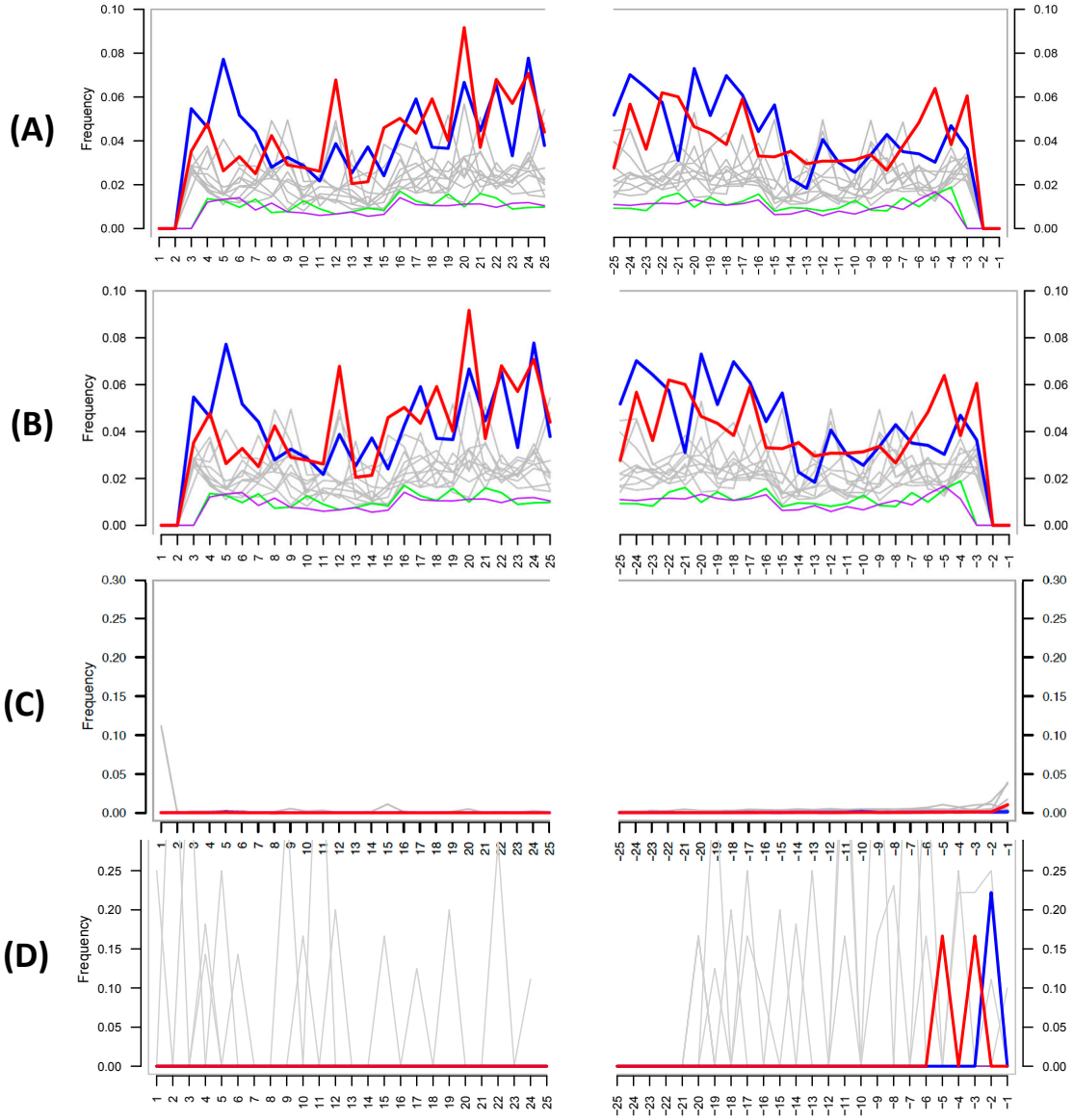


Figure 2. (**A**) MapDamage plot provided by the authors in their latest paper [1]. (**B**) MapDamage plot provided by authors in their previous publication [19]. Both plots are identical, and both show the absence of damage characteristic of authentic ancient DNA. (**C**) MapDamage plot obtained by using reads aligned from Italian mummy NASD14 from Santiago-Rodriquez et al. [1] against the *Sphingomonas sp. DC-6* genome (ASM71517v2). (**D**) Same as (**C**), except using *Vibrio parahaemolyticus* (ASM19609v1), a taxon not found in the authors' negative control. The lack of nucleotide misincorporation is indicative of modern DNA.

Despite this, the figure provided also does not support the authenticity of ancient DNA, as the expected C to T at the 5′ and G to A substitutions at 3′ ends of the DNA fragments are not present. The authors defend their lack of authentic ancient DNA signal by stating: "Damage-based ancient DNA authentication tools, such as mapDamage, may be incompatible with ancient microbiome studies unless a high sequencing coverage is reached". However, simulations and empirical data show that only a few thousand sequences from the genome of interest are required to assess the presence of cytosine deamination [6], and MapDamage has been successfully applied in several

310

paleomicrobiological studies [20–23]. To investigate if MapDamage could be appropriately applied to the Santiago-Rodriquez et al. data set [1], we downloaded the metagenomic reads from a mummy present within their study (NASD14) and identified species present in the sample using MALT and MEGAN [24,25] against a reference database containing >50-thousand bacterial and archaeal genomes obtained from NCBI Assembly. Similar to the authors' shotgun results, we identified ≈1.2 million reads assigned to *Sphingomonas* sp. DC-6 (Sphingomonadales), and 33,730 reads assigned to *Vibrio parahaemolyticus* (Vibrionales). We then mapped the metagenomic reads against these reference genomes with the BWA-backtrack (ALN) aligner [26]. The outputs were converted into SAM files then used as input for MapDamage, comparing the "ancient" *Sphingomonas* and *Vibrio* species to their respective reference genome. The resulting plots (Figure 2C,D) clearly illustrate no characteristic ancient DNA damage and are as expected for modern, likely contaminant, DNA. Given that *Sphingomonas* is one of the most abundant taxa in their data and is a known contaminant species, our reanalysis further strengthens the likelihood that contaminant DNA is driving their findings.

To conclude, a reanalysis of Santiago-Rodriquez et al.'s [1] findings strongly suggest that the observed signal is due to laboratory contamination of modern bacterial species. The authors also failed to compare their data to their extraction blanks controls, did not include shotgun metagenomic extraction blanks, and did not authenticate their ancient DNA using MapDamage. Paleomicrobiology is a new and rapidly growing field of research, with little room for plagiarized figures and blatant disregard for best-practice methods. In light of these findings, we suggest either heavy corrections or retraction of the article to prevent further erosion of the scientific integrity of paleomicrobiological research.

Author Contributions: R.E. analyzed the data, created the figures, and wrote the manuscript. L.S.W. provided feedback on the manuscript.

Conflicts of Interest: The authors work in the field of paleomicrobiology and want science published in this space to be of sufficient scientific rigor.

References

1. Santiago-Rodriguez, T.M.; Fornaciari, G.; Luciani, S.; Toranzos, G.A.; Marota, I.; Giuffra, V.; Cano, R.J. Gut Microbiome and Putative Resistome of Inca and Italian Nobility Mummies. *Genes* **2017**, *8*, 310. [CrossRef] [PubMed]
2. Weyrich, L.S.; Llamas, B.; Cooper, A. Reply to Santiago-Rodriguez et al.: Was luxS really isolated from 25- to 40-million-year-old bacteria? *FEMS Microbiol. Lett.* **2014**, *353*, 85–86. [CrossRef] [PubMed]
3. Salter, S.J.; Cox, M.J.; Turek, E.M.; Calus, S.T.; Cookson, W.O.; Moffatt, M.F.; Turner, P.; Parkhill, J.; Loman, N.J.; Walker, A.W. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **2014**, *12*, 87. [CrossRef] [PubMed]
4. Eisenhofer, R.; Cooper, A.; Weyrich, L.S. Reply to Santiago-Rodriguez et al.: Proper authentication of ancient DNA is essential. *FEMS Microbiol. Ecol.* **2017**, *93*. [CrossRef] [PubMed]
5. Cooper, A.; Poinar, H.N. Ancient DNA: Do It Right or Not at All. *Science* **2000**, *289*, 1139. [CrossRef] [PubMed]
6. Warinner, C.; Herbig, A.; Mann, A.; Yates, J.A.F.; Weiß, C.L.; Burbano, H.A.; Orlando, L.; Krause, J. A Robust Framework for Microbial Archaeology. *Annu. Rev. Genom. Hum. Genet.* **2017**, *18*, 321–356. [CrossRef] [PubMed]
7. Llamas, B.; Valverde, G.; Fehren-Schmitz, L.; Weyrich, L.S.; Cooper, A.; Haak, W. From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR Sci. Technol. Archaeol. Res.* **2017**, *3*, 1–14. [CrossRef]
8. Kim, D.; Hofstaedter, C.E.; Zhao, C.; Mattei, L.; Tanes, C.; Clarke, E.; Lauder, A.; Sherrill-Mix, S.; Chehoud, C.; Kelsen, J.; et al. Optimizing methods and dodging pitfalls in microbiome research. *Microbiome* **2017**, *5*, 52. [CrossRef] [PubMed]
9. Shen, H.; Rogelj, S.; Kieft, T.L. Sensitive, real-time PCR detects low-levels of contamination by Legionella pneumophila in commercial reagents. *Mol. Cell. Probes* **2006**, *20*, 147–153. [CrossRef] [PubMed]

10. Witt, N.; Rodger, G.; Vandesompele, J.; Benes, V.; Zumla, A.; Rook, G.A.; Huggett, J.F. An Assessment of Air As a Source of DNA Contamination Encountered When Performing PCR. *J. Biomol. Tech.* **2009**, *20*, 236–240. [PubMed]
11. Naccache, S.N.; Greninger, A.L.; Lee, D.; Coffey, L.L.; Phan, T.; Rein-Weston, A.; Aronsohn, A.; Hackett, J.; Delwart, E.L.; Chiu, C.Y. The Perils of Pathogen Discovery: Origin of a Novel Parvovirus-Like Hybrid Genome Traced to Nucleic Acid Extraction Spin Columns. *J. Virol.* **2013**, *87*, 11966–11977. [CrossRef] [PubMed]
12. Lusk, R.W. Diverse and Widespread Contamination Evident in the Unmapped Depths of High Throughput Sequencing Data. *PLoS ONE* **2014**, *9*, e110808. [CrossRef] [PubMed]
13. Adams, R.I.; Bateman, A.C.; Bik, H.M.; Meadow, J.F. Microbiota of the indoor environment: a meta-analysis. *Microbiome* **2015**, *3*, 49. [CrossRef] [PubMed]
14. Lauder, A.P.; Roche, A.M.; Sherrill-Mix, S.; Bailey, A.; Laughlin, A.L.; Bittinger, K.; Leite, R.; Elovitz, M.A.; Parry, S.; Bushman, F.D. Comparison of placenta samples with contamination controls does not provide evidence for a distinct placenta microbiota. *Microbiome* **2016**, *4*, 29. [CrossRef] [PubMed]
15. Glassing, A.; Dowd, S.E.; Galandiuk, S.; Davis, B.; Chiodini, R.J. Inherent bacterial DNA contamination of extraction and sequencing reagents may affect interpretation of microbiota in low bacterial biomass samples. *Gut Pathog.* **2016**, *8*, 24. [CrossRef] [PubMed]
16. Perez-Muñoz, M.E.; Arrieta, M.-C.; Ramer-Tait, A.E.; Walter, J. A critical assessment of the "sterile womb" and "in utero colonization" hypotheses: implications for research on the pioneer infant microbiome. *Microbiome* **2017**, *5*, 48. [CrossRef] [PubMed]
17. Key, F.M.; Posth, C.; Krause, J.; Herbig, A.; Bos, K.I. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet.* **2017**, *33*, 508–520. [CrossRef] [PubMed]
18. Jónsson, H.; Ginolhac, A.; Schubert, M.; Johnson, P.L.F.; Orlando, L. mapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinforma. Oxf. Engl.* **2013**, *29*, 1682–1684. [CrossRef]
19. Santiago-Rodriguez, T.M.; Fornaciari, G.; Luciani, S.; Dowd, S.E.; Toranzos, G.A.; Marota, I.; Cano, R.J. Taxonomic and predicted metabolic profiles of the human gut microbiome in pre-Columbian mummies. *FEMS Microbiol. Ecol.* **2016**, *92*, fiw182. [CrossRef] [PubMed]
20. Ziesemer, K.A.; Mann, A.E.; Sankaranarayanan, K.; Schroeder, H.; Ozga, A.T.; Brandt, B.W.; Zaura, E.; Waters-Rist, A.; Hoogland, M.; Salazar-García, D.C.; et al. Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Sci. Rep.* **2015**, *5*, 16498. [CrossRef] [PubMed]
21. Weyrich, L.S.; Duchene, S.; Soubrier, J.; Arriola, L.; Llamas, B.; Breen, J.; Morris, A.G.; Alt, K.W.; Caramelli, D.; Dresely, V.; et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* **2017**, *544*, 357–361. [CrossRef] [PubMed]
22. Valtueña, A.A.; Mittnik, A.; Key, F.M.; Haak, W.; Allmäe, R.; Belinskij, A.; Daubaras, M.; Feldman, M.; Jankauskas, R.; Janković, I.; et al. The Stone Age Plague and Its Persistence in Eurasia. *Curr. Biol.* **2017**, *27*, 3683–3691.e8. [CrossRef]
23. Vågene, Å.J.; Herbig, A.; Campana, M.G.; García, N.M.R.; Warinner, C.; Sabin, S.; Spyrou, M.A.; Valtueña, A.A.; Huson, D.; Tuross, N.; et al. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* **2018**, *2*, 520–528. [CrossRef] [PubMed]
24. Herbig, A.; Maixner, F.; Bos, K.I.; Zink, A.; Krause, J.; Huson, D.H. MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv* **2016**, 050559. [CrossRef]
25. Huson, D.H.; Beier, S.; Flade, I.; Górska, A.; El-Hadidi, M.; Mitra, S.; Ruscheweyh, H.-J.; Tappu, R. MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Comput. Biol.* **2016**, *12*, e1004957. [CrossRef] [PubMed]
26. Li, H.; Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **2009**, *25*, 1754–1760. [CrossRef] [PubMed]