Evolution of the Human Oral Microbiome and Resource Development for Ancient Metagenomics

Dissertation

in Partial Fulfilment of the Requirements for the Degree of
"doctor rerum naturalium" (Dr. rer. nat.)

Submitted to the Council of the Faculty of Biological Sciences of the Friedrich Schiller University Jena

by B.Sc. (Hons), M.Sc, James Alexander Fellows Yates

born on 1992-02-04 **in** Ascot, United Kingdom

Gutachter:

1. Prof. Dr. Christina Warinner (Max Planck Institute for Evolutionary Anthropology, Leipzipg, DE / Harvard University, Boston, USA / Friedrich-Schiller-Universität Jena, Jena, DE)

2. Prof. Dr. Johannes Krause (Max Planck Institute for Evolutionary Anthropology, Leipzig, DE / Friedrich-Schiller-Universität Jena, Jena, DE)

3. Assoc. Prof. Simon Rasmussen, Ph.D. (University of Copenhagen, Copenhagen, DK)

Beginn der Promotion: 2018-11-20
Dissertation eingereicht am: 2021-09-23
Tag der öffentlichen Verteidigung: 2022-05-17

Contents

1 Introduction

1.1 Background

The field of metagenomics is a fitting example of the complexity and challenges that 'big data' is bringing to biological sciences in the 21st century. Rather than analysing the genome of just a single organism, metagenomics endeavours to understand the genomic content of entire *communities* of organisms simultaneously (Handelsman et al., 1998). Metagenomics has become highly relevant in a wide-range of contexts, one such being medicine, where the human microbiome - most famously defined by Joshua Lederberg as the genetic material of the collection of microorganisms that resides within and on the human body (Lederberg and McCray, 2001) - has been shown to play a crucial role in the maintenance of health and disease (Cho and Blaser, 2012). For example, rapid fluctuation in the types and abundance of the gut microbiota has been correlated with incidents of Inflammatory Bowel Disease (Halfvarson et al., 2017), and synergistic interactions of anaerobic 'pathobionts' (opportunistic pathogens) in the oral cavity shown to accelerate instances of periodontitis (Tan et al., 2014).

This need to understand the many ways in which these microorganisms co-exist, interact, and compete, and the subsequent effects on the host organism or environment, has subsequently led to a series of large consortium-level projects such as Metagenomics of the Human Intestinal Tract (MetaHIT, Ehrlich, 2011) and the Human Microbiome Project (HMP, The NIH HMP Working Group et al., 2009). The questions of 'who, why, and how' of the human microbiome generally require large-scale (and expensive) consortium projects, as the microbial communities of each body site consist of a distinct, but diverse set of organisms, the interactions between which must be disentangled (Human Microbiome Project Consortium, 2012; Integrative HMP (iHMP) Research Network Consortium, 2014). The make-up and functioning of these communities are furthermore sensitive to many different environmental variables and conditions, much like in 'macro'-ecology (Relman, 2012; Costello et al., 2012; Barberán et al., 2014), and requires large numbers of individuals, samples, and metadata to control for these factors (Kelly et al., 2015; Casals-Pascual et al., 2020). This becomes particularly problematic when trying to produce representative datasets of different populations and societies. This was exemplified by the HMP project, which only included university students from two locations in the US as a representative population of humans (Human Microbiome Project Consortium, 2012; Integrative HMP (iHMP) Research Network Consortium, 2014). Thus, this data likely only represents a tiny fraction of true microbiome diversity across the world (Blaser and Falkow, 2009; Yatsunenko et al., 2012; Schnorr et al., 2014; Pasolli et al., 2019).

A crucial development that has enabled intensive investigation of metagenomes is 'culture-independent' high-throughput DNA sequencing (Chen and Pachter, 2005; Shendure and Ji, 2008). Circumventing the need for laborious optimisation of culturing conditions (to generate enough genetic material of only a *single* organism), these technologies instead allow for the direct sequencing of the genetic content of entire communities of taxa at once. This results in fewer biases related to the varying particular and sometimes incompatible (mono-) culturing conditions many organisms can have, allowing researchers to instead focus on exploring the relationships between different taxa, and their subsequent effects on their environment. However, it can be argued that initially, technological development outstripped the

pace in which computational analysis methods could deal with this industrialised-level of data production (Muir et al., 2016). This rate had been famously exemplified by the reduction of costs of sequencing falling faster than expected from 'Moore's law', at least until recently (https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, accessed Nov. 2020). Therefore, being able to efficiently process these sizes of data, as well as the sharing of these resources, remains of high interest.

A fortuitous by-product of the methods that 'second-generation' sequencing (SGS, colloquially known as next-generation sequencing or NGS) afforded was that the methods were ideal for the sequencing of *ancient* DNA (aDNA). DNA from long-dead and even extinct organisms can survive for over 1 million years (van der Valk et al., 2021). However, these molecules are often very fragmented, and samples contain large amounts of 'contaminating' nucleotide sequences from modern sources - including researchers themselves. Fortunately, these short fragments are perfect for SGS machines, which generally take the approach of sequencing millions of short fragments of DNA in parallel at high-accuracy. Retrieval of entire genomic sequences can then be subsequently computationally reconstructed through the large amount of data produced (van Dijk et al., 2014). Therefore aDNA fragments are highly compatible with the sequencing mechanisms of the machines, and enough DNA can be sequenced to pick up the trace amount of ancient fragments that exist amongst the often large amounts of contaminating biomass.

Despite these technological developments being simultaneously suited for both microbiome metagenomics and aDNA studies, until relatively recently most palaeogenetic research focused on amplicon sequencing (targeted amplification and sequencing of a single common gene present in many organisms, or of a single species-specific sequence), or the analysis of genomes or genome-wide data of single taxa (Slatkin and Racimo, 2016; Marciniak and Perry, 2017; Spyrou et al., 2019; Arning and Wilson, 2020). It is only in the last few years that metagenomic reconstruction of whole microbiome communities have started to become a research focus in the field (e.g. Tito et al., 2008; Warinner et al., 2014b; Weyrich et al., 2017; Lugli et al., 2017; Philips et al., 2017). In particular, the recent discovery of dental calculus - mineralised dental plaque - as a prevalent but surprisingly well-preserved reservoir of ancient biomolecules (de La Fuente et al., 2013; Adler et al., 2013; Warinner et al., 2014a,b; Weyrich et al., 2017) has opened up the opportunity for sufficiently large-scale studies to address questions about the microbiomes of ancient human societies, animals, and even food sources. However, the nature and potential of this new type of data remains largely unexplored, with open questions regarding the authentication of reconstructed ancient metagenomes, as well as how to efficiently process and interpret the data remaining unaddressed.

This thesis represents an intersection between aDNA, metagenomics, and microbiome research. Manuscript A represents a formative step towards reconstructing the early evolutionary states of the hominid oral microbiome that may be used to guide holistic approaches to modern oral healthcare. In addition, Manuscript A develops novel tools and approaches for improving palaeogenomic authentication of ancient microbiomes, as well as presenting a framework to address anthropological questions regarding the co-evolution of humans and their microorganisms. Manuscripts B and C address technical challenges that are currently encountered in large-scale ancient microbiome studies. Manuscript B describes the creation and curation of a novel and long-term resource of published metagenomic samples and their

essential metadata - *AncientMetagenomeDir*. This resource allows for more efficient exploration and gathering of comparative datasets as well as meta-analyses. Manuscript C presents *nf-core/eager*, an open-source bioinformatics pipeline designed specifically for the efficient processing of large palaeogenomic datasets using the latest bioinformatic practices. It also includes a new suite of functionalities dedicated for metagenomics, in a framework that can scale to the levels required for robust ancient microbiome analysis.

1.2 The human microbiome

1.2.1 Characterising the human microbiome

Microbes inhabit almost every environment on earth and are an essential part of the functioning of life (Stolp, 1988). The number of microbial cells living in and on the human body at any one time is currently estimated to match the number of human host cells themselves (Sender et al., 2016). These microbial communities are considered to be so important that theories such as the 'hologenome' are being developed (Rosenberg et al., 2007). In this concept the evolution and functioning of a given organism cannot be considered as the sole entity, but rather must always be considered as a 'super-organism' (Bordenstein and Theis, 2015). Processes of selection pressure and adaptation applied to the microbial organisms coexisting with the host are therefore as important to the survival of the host as those applied to the genes of the host - in addition to the pressures and adaptations made by host themselves on their microbiome, and vice versa (Rosenberg and Zilber-Rosenberg, 2018).

As in the macroecology of certain environments, microbes living in and on humans have adapted to different niches. This has led to a large amount of taxonomic diversity that inhabits different parts of the human body. In a foundational study, the HMP aimed to generate an initial comprehensive survey of this diversity by sampling from different parts of the body. They selected sites that had been shown to have relatively rich microbial communities, such as the oral cavity, gut, skin, and respiratory tract, across many individuals (Human Microbiome Project Consortium, 2012). The results of the study confirmed that while at higher taxonomic levels the microbes that were present were similar, most body sites harbored an overall distinct site-specific taxonomic signature at lower taxonomic levels (Human Microbiome Project Consortium, 2012). Furthermore, across this particular sample of healthy Americans adults, a remarkable stability in the functional interactions of the different microbiomes existed despite a certain level of inter-individual taxonomic variability. This observation has subsequently led to a shift in recent years in modern microbiome research to include functional reconstructions of different microbial communities, using other techniques such as metatranscriptomics, metaproteomics and high-throughput metabolomics (Integrative HMP (iHMP) Research Network Consortium, 2014). However, surveying taxonomic diversity in microbes inhabiting human populations can still be considered important. As in macroecology, maintaining taxonomic diversity is important to provide functional 'redundancy' when environmental conditions change (Relman, 2012; Dorrestein et al., 2014; Jacobson et al., 2020). For example 'backup' taxa or pathways are needed to ensure that the production or breakdown of certain useful or toxic metabolites still occurs, even if a particular taxon disappears (Tremaroli and Bäckhed, 2012).

Given that the changes of bacterial diversity in the human microbiome have been clinically shown to be associated with many chronic diseases in industrialised societies (e.g. Bisgaard et al., 2011; Blaser, 2016; Claesson et al., 2012; Yang et al., 2012), understanding the wider diversity - and how to maintain this - has become an important area of research (Davenport et al., 2017). Multiple studies of different body sites have shown that populations from non-industralised societies display a wider range of inter-individual and inter-population diversity than those in industrialised societies (e.g. Nasidze et al., 2009, 2011; Lassalle et al., 2018; Yatsunenko et al., 2012; Obregon-Tito et al., 2015). Cataloging this diversity and understanding the role of more rare taxa can therefore help further understand how to maintain more robust functional states and thus reduce the incidence of chronic disease (e.g. Keohane et al., 2020).

1.2.2 The human oral microbiome

The oral cavity is one of the most diverse body sites in terms of microbial colonisation (Human Microbiome Project Consortium, 2012) with nearly 700 different species characterised (Chen et al., 2010; Escapa et al., 2018), and between 100-300 different known taxa being present in a single individual at any one point (Lazarevic et al., 2010; Bik et al., 2010). The oral cavity consists of multiple unique niches with distinct microbial profiles, including the tongue, teeth, gingiva, palate, lips, cheek, and saliva (Dewhirst et al., 2010; Eren et al., 2014; Proctor et al., 2020). These different sites provide different environments, from stable mineralised surfaces to epithelial tissues to fluids, all of which that allow for the colonisation of different types of taxa (Dewhirst et al., 2010; Mark Welch et al., 2016; Human Microbiome Project Consortium, 2012; Mark Welch et al., 2019). Therefore the oral cavity is a dynamic environment, with various fluids, like saliva, contributing to fluctuating factors such as pH, movement of sugars, amino acids, and other potential nutrients, as well as the dispersal of cells (Proctor et al., 2020; Mark Welch et al., 2020). Given the high taxonomic diversity of the oral cavity, understanding and characterising the breadth of this diversity is likely important for developing improved oral healthcare across the world.

1.2.3 The human oral microbiome in health and disease

Oral disease is a major global healthcare challenge, with an estimated 3.5 billion people suffering from some form of oral disorder in 2017 (GBD 2017 Disease and Injury Incidence and Prevalence Collaborators, 2018), and with prevalence and incidence of periodontal disease remaining at consistent levels over the last few decades (Peres et al., 2019). Correspondingly, the global economic impact is significant, with estimates at $544.41 billion for 2015 (Righolt et al., 2018). Due to its association with periodontal disease, but also chronic diseases such as cardiovascular inflammation (Peres et al., 2019), one of the most studied microbiomes of the oral cavity is dental plaque. This microbial biofilm is attached to the surface of teeth, the only non-shedding surface in the mouth, and resides both above and below the gingival margin (i.e., the gumline). This type of microbial biofilm is also present in a animals (Dent, 1979). Originally, it was believed that some types of diseased states in the oral cavity were caused just by the presence of or amount of plaque (known as the 'Nonspecific Plaque Hypothesis'; Rosier et al., 2014; Colombo and Tanner, 2019). This changed in the 1970s when improvements in culturing and bacterial

taxonomic identification meant that researchers started associating caries or periodontal disease with the presence of specific pathogens, such as *Streptococcus mutans*, that directly isolated from caries (originally proposed as the 'Specific Plaque Hypothesis' by Loesche, 1976). By the 1980s a series of alternative concepts were introduced that considered disease being caused by a range of taxa (Rosier et al., 2014). Multi-taxa hypotheses, such as the 'Ecological Plaque Hypothesis' by Marsh (1994) and later extended with the 'Keystone Pathogen Hypothesis' by Hajishengallis et al. (2012), have since become widely accepted. These hypotheses suggest that *dysbiosis*, i.e., changes of presence/absence, interactions, and abundance of both taxa or environmental factors (Yost et al., 2015) away from a 'healthy' state, is the primary driver of oral disease (Kilian et al., 2016).

In the case of oral biofilms, 'sub-gingival' plaque is more difficult to remove during daily healthcare practices and thus can act as a reservoir of this biofilm. The lack of removal then leads to a concentration, and increasing diversity, of typically anaerobic taxa (Lamont et al., 2018), which has been correlated with inflammation of the gingiva and attachment loss between a tooth and the surrounding gingiva (White, 1997; Lamont et al., 2018). Historically, and as with the Keystone Pathogen Hypothesis, certain taxa have been more commonly associated with the development and progression of disease. This was summarised in a seminal paper by Socransky et al. (1998). In this paper, the authors defined different 'complexes' of taxa that were commonly associated with each other in different states of health and disease using DNA-DNA checkerboard hybridisation. Most famously, the 'red' complex consisted of taxa that were more prevalent in individuals with clinical measurements indicating severe periodontal disease - the current names of these taxa being *Tannerella forsythia*, *Porphyromonas gingivalis*, and *Treponema denticola*.

1.2.4 Formation of the plaque biofilm

The concepts of taxonomic stratification and succession of predictably associated taxa in the form of 'complexes' have been widely used in oral microbiology. These were originally summarised by Moore and Moore (1994), and more formally defined by Socransky et al. (1998), in the context of statistically testing anecdotal observations of associations of groups of species with periodontal disease. However, they have since acted as useful guides for describing the formation of the dental biofilm (particularly by ancient microbiome researchers). The original definition of the complexes as per Socransky et al. (1998) consisted of three main layers: 1) blue, purple, green, and yellow complexes of typically aerobic and saccharolytic taxa representing 'early colonisers', 2) co-aggregating anaerobic and more proteolytic orange complex taxa in an intermediate layer as 'bridging taxa', and 3) in a final 'mature' layer, disease-associated red complex species and other (facultative) anaerobic 'late colonisers'. While the general concept of this model still stands, more recent technological advances have been providing an increasingly higher resolution to the understanding of the processes in the formation and structure of the oral biofilm.

The idea of 'strict stratification' of these layers has since become more fuzzy, and a more dynamic formation and functioning of the biofilm is being observed. In particular, high-resolution imaging techniques such as fluorescence *in situ* hybridization (FISH) have been developed and repeatedly applied to oral biofilms (Zijnge et al., 2010; Mark Welch et al., 2016; Palmer et al.,

2017). Zijnge et al. (2010) developed and applied fluorescent probes of a wide range of well-known oral microbiota to sectioned biofilms of teeth from patients. Importantly, compared to previous FISH work, these were based on biofilms directly taken from teeth rather than artificial enamel plates placed in the mouth (e.g., Al-Ahmad et al., 2007). They identified successive layers within the subgingival plaque, defining these as: basal, intermediate, top, and 'outside' layers; whereas supragingival plaque only appeared to show two layers: basal and second (Zijnge et al., 2010). This structuring is thought to be due to the way in which dental plaque forms and develops.

The formation of dental plaque begins with the adsorption of saliva-derived small glycoproteins, proteins, lipids, and other biomolecules that have an affinity to the hydroxyapatite mineral of the surface of tooth enamel (see Siqueira et al., 2012; Jakubovics et al., 2021, for reviews). The thin film of rich organic molecules (termed the 'acquired enamel pellicle' or 'AEP'; Hannig and Hannig, 2009) acts an ideal place for the adhesion of a set of (often saccharolytic) early-colonising microbial taxa, via receptor-adhesin interactions. In particular, *Streptococcus*, *Actinomyces*, *Veillonella*, and *Haemophilus* are routinely observed (Palmer et al., 2003; Diaz et al., 2006; Mark Welch et al., 2016). When colonising a 'fresh' enamel surface, these early colonisers are typically derived from the shedding of other surfaces in the mouth and move to new locations via saliva flow (Mark Welch et al., 2020)

Once the basal layer has formed, cell-to-cell co-aggregation interaction between different taxa is a critical mechanism for the continued development of the biofilm (Kolenbrander et al., 2006; Valm et al., 2011; Palmer et al., 2017). Such aggregations and subsequent growth result in rounded structures termed 'hedgehogs', consisting of long filaments of taxa (such as *Corynebacterium*) that fan out from the basal layer, and span throughout the depth of the biofilm to the outer layer (Mark Welch et al., 2016). This intermediate layer typically consists of an anoxic environment, and therefore is colonised with (facultative) anaerobic taxa, such as *Capnocytophaga*, *Leptotrichia*, and *Fusobacterium* (Zijnge et al., 2010; Mark Welch et al., 2016). However there are many different coaggregation clusters occurring involving many different taxa throughout this layer and the entire biofilm, with many of the involved taxa being currently underappreciated (Palmer et al., 2017). These clusters, as well as environmental micro-gradients (von Ohle et al., 2010; Kim et al., 2020; Mark Welch et al., 2020), suggests a more dynamic biofilm than simple stratification, as is often depicted in the literature (e.g., Kolenbrander et al., 2006). Indeed, recently it was observed that the motile *Capnocytophaga gingivalis* can act as a 'transporter' of non-motile taxa (Shrivastava et al., 2018). This suggests that this common inhabitant of the oral biofilm may play an important role in biofilm formation by facilitating the dynamic distribution of taxa. Furthermore, *Fusobacterium*, which historically was considered to play a crucial role as a 'promiscuous' co-aggregator with multiple species (Kolenbrander et al., 2006), has only recently been shown to possibly play a lesser role compared to other more possibly important but less-studied taxa (such as *Cornyebacterium* Valm et al., 2011; Mark Welch et al., 2016; Palmer et al., 2017). That said, this remains debated (Diaz and Valm, 2020), and highlights that many unexplored avenues in oral biofilm research still exist for which high-resolution techniques such as FISH will be influential.

The highly anoxic 'top' layer consists of *Porphyromonas*, *Prevotella*, and *Tannerella* (Zijnge et al., 2010), many of which are considered disease-associated taxa (as originally reported by Socransky et al., 1998). Finally, the 'outer layer', 'perimeter', or 'outer shell' of the biofilm con-

sists of the tips of filaments of structural 'pylon' taxa, such as *Corynebacterium* (Zijnge et al., 2010; Mark Welch et al., 2016). On these tips, co-aggregation with aerobic and microaerophilic taxa, such as with *Streptococcus* and *Haemophilus*, are observed, but is also interspersed with anaerobic taxa such as *Porphyromonas* and spirochetes such as treponemes (Zijnge et al., 2010). Such structures of filaments with cocci at the ends have been termed 'corncob' structures (Jones, 1972; Mark Welch et al., 2016).

1.2.5 Plaque biofilms, disease and diversity

Periodontal disease typically begins to manifest at locations where a biofilm is at full maturity. The strata-like formation described above is typical of the progression of *supragingival* plaque - biofilms that exist above the gum line. Supragingival locations on the enamel surface are ideal for primarily aerobic, early-colonisers - given the ready access to oxygen and salivary carbohydrate and proteins used for binding (Marquis, 1995; Hojo et al., 2009). These taxa types are correspondingly reflected in Zijnge et al. (2010)'s two layer description. In contrast, *subgingival* plaque exists in the gingival crevice, which is an low-oxygen niche containing amino acid-rich gingival crevicular fluid, which is a suitable energy source rather for primarily proteolytic anaerobic taxa (Kolenbrander et al., 2010). Establishment and proliferation of these generally anaerobic and gram-negative taxa in this gap between the tooth root and the gingiva leads to inflammation, and therefore has lead to the designation of these taxa as 'periopathogens' (Rupf et al., 2000; Rylev and Kilian, 2008). Despite the more specific niche conditions, subgingival plaque has actually been shown to be *more* diverse than supragingival plaque (Darveau et al., 1998; Abusleme et al., 2013) and this is an important factor in hypotheses that *dybiosis* is what leads to disease. Imbalances in the biofilm environment can lead to 'enrichment' of certain taxa, which, without correction from the remaining community, can lead to issues such as greater chemical toxicity that affect host cells (Curtis et al., 2020).

For example, co-culturing of *Fusobacterium nucleatum* (Lee and Baek, 2013) and/or *T. denticola* (Zhu et al., 2013; Tan et al., 2014) with *P. gingivalis* leads to rapidly improved growth of the latter. The overgrowth of this species then inhibits accumulation of host immune-system Interleukin-8 chemokines (IL-8). The suppression of IL-8 by *P. gingivalis* allows other taxa to proliferate, resulting in a more aggressive host immune-system response, and ultimately leads to bone-loss and subsequently periodontal disease (Darveau et al., 1998). Another example of the multi-microbial nature of periodontitis is the relationship between streptococci species and *Aggregatibacter actinomycetemcomitans*. *A. actinomycetemcomitans* has adapted to up-regulate certain genes that protect it from the human immune responses, but only when in the presence of lactic acid and hydrogen peroxide rich environment produced by various streptococci species. At the same time, the acidic environment inhibits the growth of other taxa (Ramsey and Whiteley, 2009). This mechanism then leads to the increased growth of *A. actinomycetemcomitans*, which has the ability to degrade white blood cells and again further stimulate an overaggressive host response - resulting in inflammation-related disease (Åberg et al., 2015).

As the manifestation of some diseases can occur through disruption of the complex relationships between different taxa in the plaque biofilm, clinical solutions should also take into consideration the maintenance of a healthy microbiome in any prescribed intervention. In contrast to microbiota at other human body sites that can be more taxonomically dynamic

(Lozupone et al., 2012; Greenbaum et al., 2019), the taxonomic stability of plaque biofilm suggests it should be more straightforward to define what constitutes a 'healthy' oral microbiome state (Utter et al., 2016).

1.3 Evolution of the oral microbiome

1.3.1 Why reconstruct the evolutionary history of the oral microbiome?

Despite the importance of taxonomic redundancy towards the maintenance of the equilibrium of a healthy biofilm, our knowledge of the global diversity of the oral microbiome remains limited. The vast majority of oral microbiome research has been performed on populations from industrialised societies, particularly within the US (such as in the HMP), and has focused on more socio-economically rich groups within these societies. This thesis will use the term of 'industrialised' societies following their definition in sociocultural and economic terms. These groups have populations that overall typically rely on mechanised, mass production of foodstuffs, and have established access to modern healthcare, including antibiotics and hygiene practices (but access of which can vary within the society). These societies currently suffer the least oral disease (Peres et al., 2019), likely due to the ability for greater financing of oral health care and education. These sets of published microbiome data are thus likely poor representations of the global diversity of the oral microbiome.

Using such a limited subset of populations for analysing the oral microbiome further limits our understanding of the likely highly complex relationship between taxonomic diversity and the maintenance of health and disease. In clinical contexts, using currently characterised oral biofilm communities of heavily industrialised populations as reference material to develop broader healthcare treatments is not ideal. Resulting treatments may not be suitable for the microbial diversity and interactions of populations in other sociocultural and economic contexts (e.g. as noted for caries; Philip et al., 2018). For example, culturing studies have indicated possible differences in the species present in patients suffering from periodontal disease from different countries (Sanz et al., 2000; Herrera et al., 2008), although quantitative culturing studies have limitations due to qualitative morphological identification of different species. Different populations may require different taxonomic and functional profiles to maintain healthy oral microbiomes, due to the different environmental and cultural contexts in which each population lives. Without identifying all possible taxonomic or taxonomic-relationship contexts that result in a given illness, a treatment may not be sufficiently effective for everyone. This is because the data they are based on may only identify peripheral differences specific to the cultural context of that population, rather than the actual underlying causes or mechanisms that result in disease. Characterising this diversity will therefore also help improve our identification of the critical functional pathways and taxonomic relationships that cause diseases afflicting humans at global scales.

A range of studies have shown that in the gut and oral microbiome of diverse societies, there are many widespread taxonomic and functional characteristics that are not found in industrialised societies. Additionally, even within these industrialised populations, such differences in taxonomic presence and absence also exist in lower socio-economic classes, something which has not been captured in keystone projects such as the HMP (e.g. Schnorr et al., 2014; Rampelli

et al., 2015; Sankaranarayanan et al., 2015; Ozga et al., 2016; Li et al., 2014). Indeed it has been argued that for the gut microbiome, industrialisation of a society results in the development of a fundamentally dysbiotic microbiome across the population (e.g. Dominguez Bello et al., 2018; Sonnenburg and Sonnenburg, 2019). However, some researchers have questioned whether industrialised microbiomes are indeed 'fundamentally' dysbiotic, or whether the microbiome is rather just adapting to the environmental context of the industrialised culture (Broussard and Devkota, 2016; Carmody et al., 2021). Rather than assuming industrialisation 'automatically' should be equated with health or disease, characterising the diversity of different populations within their anthropological contexts will allow us to test this hypothesis. By doing so, we will be able to identify a wider of range environmental factors that can influence the oral microbiome and improve the contextualisation of disease in different societies.

However, expanding characterisation of the oral microbiome in living populations and of different lifestyles only provides a contemporary perspective on the total diversity. Living populations reflect only a singular snapshot at a limited point of time in human history. If we wish to understand how microbiomes adapt to different contexts, we require a *temporal* axis to understand how exactly microbiomes are adapting to different environmental contexts. By looking at the long-term evolution of microbiomes and comparing to long-term evolution of human behavioural or cultural change, we will be able to assess the which, when, and how much, and how adaptation occurred (Warinner et al., 2015). An important aspect of tracing adaption through time is that researchers require a starting point to compare against. A logical starting point would be the ancestors of all populations under study. Following evolutionary principles, inferring the ancestral state of an organism or microbiome can be made by finding common microbial taxa of related host species, such as, in this case, chimpanzees, gorillas and other hominids. Correspondingly, the microbiota of these host species also require characterisation before ancestral states can be reconstructed. Expanding diversity both in terms of present-day but also from ancient societies and species, could allow us to further identify whether diseases considered 'modern' are actually just certain mechanisms or responses of the microbiome to specific contexts - possibly having also occurred in the past. Identifying what factors resulted in certain changes could then potentially help allow predictions of 'reactions to certain changes, helping us more holistically manage oral health. Ultimately, by sampling more populations - both living and ancient - we will be able to assess how well inferences on the role of the microbiome in health and disease made from the current limited range of populations hold up both over short-, and long-term, evolution. Accessing such transects, however, is not trivial.

1.3.2 Accessing the global diversity of the human oral microbiome

Assessing the wider global diversity of the human oral microbiome has typically been previously performed by sampling living humans today, typically by focusing on saliva, which is an easily and non-invasively sampled substrate. Early work focused on analysis of terminal restriction fragment length polymorphism (T-RFLP), or amplicon sequencing of ribosomal RNA (16S rRNA, a gene present in all microbial organisms), to efficiently get broad surveys of the taxonomic content of samples, when SGS was still in its infancy and still relatively expensive. Studies such as Nasidze et al. (2009), Nasidze et al. (2011) and Li et al. (2013), sampled the saliva

of individuals from every major continent and identified a high diversity of taxa around the world, with many potentially uncharacterised microbial genera. This initial work appeared to suggest that between human populations there was no geographical partitioning of taxonomic composition (Nasidze et al., 2009). However a later study argued that the salivary microbiome profiles could be used, to a limited extent, to differentiate populations (Li et al., 2014; Takeshita et al., 2014). These studies found that less-industrialised hunter-gatherer societies in Western Africa displayed a higher level of diversity than more industrialised individuals in the same region (Nasidze et al., 2011; Li et al., 2014). Furthermore, when comparing the saliva microbiome between zoo and wild apes as a more extreme example of differences occurring from the result of industrialisation-like processes, Li et al. (2013) showed that zoo counterparts also had much less diversity.

However, saliva represents a transient and variable microbiome derived from multiple oral sites due to constant fluid flow in the oral cavity, and may not act as a good proxy for long-term oral microbiome composition. This transience is much more likely to reflect the environmental context of the individuals of different populations. While other 16S rRNA work on plaque samples reported possible differences between different ethnicities within a single population (African American versus Caucasian American; Mason et al., 2013), this has been strongly questioned due to a lack of control of these environmental differences (e.g., comments on the PLoS One article by Mason et al., 2013, Accessed March 2021), and overly-simplistic definitions of race (Benezra, 2020). Indeed, other studies of the plaque microbiome across industrialised societies in different geographic regions has been shown to be relatively similar, which emphasises that behaviour or environmental contexts, such oral health routines, is more likely to influence microbiome profiles than the genetics of the host themselves (Utter et al., 2016).

1.3.3 Reconstructing the ancestral oral microbiome

Identifying differences between living-day populations only gives a shallow view of the diversity of the human microbiome, however. Microbiomes will have been dynamic throughout our past, but constructing such a transect requires a starting point to start measuring changes from - such as from the microbiomes of our ancestors. Evolutionary concepts for reconstructing ancestral states of host genomes can also be applied to microbiomes through finding shared sequences between the (microbial) genomes of descendants of a common ancestor (Dunn et al., 2020). These concepts have been previously applied to the comparison of gut microbiomes of humans, chimpanzees, and other apes (e.g., Moeller et al., 2012, 2014). These studies broadly found that the diversity of gut taxa was indeed higher in primates and has since become less diverse in humans. Importantly, however, all primates analysed displayed a 'core' set of taxa. This showed that these different host species still share a set of common, and likely ancestrally-derived, collection of microorganisms, and that earlier forms of *Homo* may have also held this diversity. In addition, studies such as by Ochman et al. (2010) and Moeller et al. (2016) have showed the continued presence of these microbes in the host species, despite strain-level co-evolution during the diversification of the hosts themselves. However, evolutionary change (e.g., via inheritance, and microbial co-evolution with the host) is not the only factor that results in long-term changes, but ecological change (e.g., microbial species abundance fluctuation due to different environments) can also play a major role. Later studies including a wider

range of humans and Old World monkeys have shown that human gut microbiota appear to be taxonomically and functionally more similar to baboons and macaques, than to more closely related chimpanzees and gorillas (Amato et al., 2019b), and post-host divergence transfer of strains still occurred (Moeller et al., 2016). This reflects the complexities of the human microbiome in ways that depart from typical phylogenetic evolutionary relationships, with Amato et al. (2019b) arguing that dietary (Muegge et al., 2011), physiological, and behavioural (Gomez et al., 2015) traits play an equally (or even more so) influential role in the on-going co-evolution of host and their microbiota (Amato et al., 2019a). In both cases, humans were found to have an unusual rate and pattern of divergence (Amato et al., 2019b; Moeller et al., 2014), and that our relatively unique cultural and behavioural adaptability has influenced the evolution of our own microbiota. While this approach has provided some insights, biases remain. Firstly, many studies looking at nonhuman primates have used captive animals that have been shown, in a variety of contexts, to have modified microbiomes versus their wild counterparts (e.g. Clayton et al., 2016; McKenzie et al., 2017; Youngblut et al., 2019).

Much like in humans, recording and maintaining the diversity of the microbiota in animal species has been proposed to be an important extension of current host-genome focused conservation strategies (Redford et al., 2012; West et al., 2019; Trevelline et al., 2019; Ross et al., 2019). In this context, the recording of the diversity of microbiomes in threatened species must not only apply to simply microbial species, but also to the diversity of *body sites* that researchers should study. As can be observed from the previously cited studies, the vast majority of evolutionary research on host-associated microbiomes in animals (and also to a large extent, in humans) has been primarily focused on the gut. This bias can be partly explained by the fact that non-invasive sampling of the gut microbiota is relatively simple - through the sampling of naturally deposited faecal material. Furthermore, the gut is a relatively 'closed off' system compared to more 'exposed' body sites like the skin, and oral cavity. Having to control for the many more environmental variables that the latter sites encounter, makes studying these types of microbiomes more difficult. While physically sampling skin and saliva from humans is relatively straightforward, the same cannot be said of animals that may not be as cooperative. Therefore, the majority of these studies have relied heavily on zoo or domesticated animals (Council et al., 2016; Ross et al., 2018), who may not be suitable for the reconstruction of the potential wider biodiversity of the human ancestral microbiome, due to anthropological modification.

1.3.4 The oral microbiome outside clinical contexts

Microbiome research on the oral microbiome is primarily focused on health and disease. Given that one aspect of the human microbiome is that it is influenced by human behaviour (Herd et al., 2018), this concept could potentially be used to infer different aspects of human life such as diet, hygiene, and behaviour. For example, Shaw et al. (2017) demonstrated via 16S rRNA analysis of saliva microbiomes that the primary driver of taxonomic similarity of the saliva of related individuals was proximity. I.e., those who share long-term occupancy of the same households (Blekhman et al., 2015; Demmitt et al., 2017; Gomez et al., 2017), rather than host genetics. Interestingly, the particular taxa that drove the clustering of the different households only made up a small part of the overall saliva microbiome, and a consistent core microbiome

remained across all populations (Shaw et al., 2017). Despite this, these small differences showed that the saliva microbiome could in principle be used to test hypotheses of familial or household relationships between individuals, as an alternative to the genetics of the hosts themselves.

In another example, Lassalle et al. (2018) performed metagenomic analysis of saliva of individuals from the Philippines. The peoples in the study practise different subsistence strategies (hunter-gathering and 'traditional farming'), and the study found that some differences could be observed that distinguished the two populations. They also noted (as with Nasidze et al., 2009, 2011) that hunter-gatherer individuals from the Philippines showed higher variability and diversity in the taxonomic composition of their saliva microbiomes. At the same time, they also appeared to have a lower incidence of oral disease, an observation the authors hypothesised to possibly be related to the diet of the hunter-gatherer population containing less starch and sugar (Lassalle et al., 2018).

In both cases, while saliva is easier to access, the patterns observed in saliva may only reflect relatively short periods of human behaviour due to the transient nature of the fluid (e.g. Lazarevic et al., 2010; Jiang et al., 2015). A focus on saliva limits analysis of the oral microbiome to individuals living today and do not allow access to analyses of changes over long-term behavioural trends. As well as trying to use the oral microbiome to address more evolutionary questions, understanding the effects of *long-term* behavioural shifts can be considered important for designing improved prevention and treatment strategies of disease - rather than short-term 'reactionary' responses (e.g., an example of which could be the overuse of antibiotics; Ventola, 2015; Shallcross and Davies, 2014; Brealey et al., 2021). Instead, reconstruction of ancestral states of the oral microbiome from *past* individuals, could be used to identify past diversity and 'calibration points' at which certain changes occurred in recent and deep history. In this context, Manuscript A describes the use of ancient dental calculus from both close and distant primate relatives of humans to reconstruct deep-time ancestral states of the human oral microbiome.

1.4 Reconstruction of ancient oral microbiomes

1.4.1 Dental calculus: fossilised dental plaque

This concept, however, raises the question of *how* can 'deep-time' ancestral states of oral microbiomes be accessed? If not removed, dental plaque undergoes periodic 'fossilisation' events during the lifetime of an individual. For reasons that are still not fully understood, dental plaque spontaneously mineralises, during which the biofilm becomes supersaturated by minerals from saliva and from the microbes themselves (Jin and Yip, 2002; White, 1997). The minerals involved in this process primarily consists of calcium phosphates (Hayashizaki et al., 2008). During mineralisation, the plaque biofilm is 'petrified', killing a large proportion of the microbial cells (White, 1991) - the remains of which are preserved within the deposit alongside a range of other 'exogenous' material that may be present in the oral cavity at the time. This mineralisation process occurs unevenly throughout the biofilm, resulting in crystal aggregates of different ages and mineral composition (White, 1997). Once the mineralisation of the biofilm is complete, a new layer of dental plaque biofilm forms on the surface of the calculus, resulting in a layer-like structure of the deposit (Akcalı and Lang, 2017). Dental calculus is regularly found on skeletal

remains, as the hard mineral deposits protects against, and reduces the speed of postmortem degradation. It has therefore have been investigated by archaeological researchers since the 1970s (see Warinner et al., 2015, for a review) for a wide range of purposes - primarily for inferring diet from embedded plant microfossils (Klepinger et al., 1977; Middleton and Rovner, 1994; Fox et al., 1996; Cummings and Magennis, 1997; Henry and Piperno, 2008; Henry et al., 2011; Hardy et al., 2017; Henry et al., 2012, 2014; Power et al., 2015; Cristiani et al., 2016; Geber et al., 2019; Tromp et al., 2020) but also inferring self-treatment (Hardy et al., 2012, 2016), and tool working (Radini et al., 2016; D'Agostino et al., 2019; Radini et al., 2019).

This fossilisation process is critical for accessing the ancestral states of the oral microbiome. The maturation of minerals in dental calculus, after initial formation, results in the formation of a primary mineral type of whitlockite and hydroxyapatite (White, 1997). Hydroxyapatite is a calcium phosphate-based apatite that is one of the main inorganic substrates of bone, tooth, and enamel, and therefore preserves well in the archaeological record. Importantly, hydroxyapatite in bone and teeth has a physico-chemical affinity to biomolecules such as DNA, causing phosphate-based ionic adsorption of the biomolecules to the surface of the mineral and acts as a stabilising mechanism against DNA degradation (Grunenwald et al., 2014). Mineral growth can occur in a way that it replaces water present in the bone, something that improves preservation, as the presence of water is a key component to the hydrolysis-based breakdown of biomolecules, including DNA (Kendall et al., 2018). High density mineralisation further reduces microbial-based degradation as there are fewer 'channels' that allow microorganisms to penetrate into the interior of a skeletal element (Turner-Walker, 2007). As empirical evidence, more densely mineralised bone elements have been shown to more reliably yield higher levels of endogenous host aDNA, versus more porous or less hard elements (Gamba et al., 2014). Due to the similar mineral composition of bone and dental calculus, it is therefore not so surprising that dental calculus was recently shown to be not only a good reservoir for microfossils but also for ancient biomolecules including DNA, proteins, and other organic substances (Hardy et al., 2012; de La Fuente et al., 2013; Adler et al., 2013; Warinner et al., 2014a,b).

The presence of DNA in archaeological dental calculus was originally identified when researchers performed transmission electron microscopy (TEM) on sections of dental calculus (Preus et al., 2011), and found evidence of both preserved bacterial cells and reactive DNA in the centre of the cell impressions. Confirmation that these cells could be from oral taxa (rather than possible environmental contamination) later came from the successful targeted amplification of species-specific genes of known oral pathobionts such as *Streptococcus gordonii*, *F. nucleatum*, and *P. gingivalis* (de La Fuente et al., 2013). Since then, the possibility of extracting sufficient DNA deriving from the wider oral biofilm preserved in calculus has been repeatedly and independently observed (e.g. de La Fuente et al., 2013; Adler et al., 2013; Warinner et al., 2014b; Weyrich et al., 2017; Velsko et al., 2018; Brealey et al., 2020; Neukamm et al., 2020). Dental calculus has been observed across many archaeological populations and ancient societies (Lieverse, 1999; Novak, 2015; Austin et al., 2019), as well as being present on the skeletal remains of extinct hominid lineages as old as 12 million years (Fuss et al., 2018; Hershkovitz et al., 1997). Therefore, given the preservation potential, it is an ideal potential substrate for large-scale investigations into the historical diversity of the human microbiome.

1.4.2 Key stages in the development of ancient oral microbiome research

Adler et al. (2013) presented the first study attempting to take a broader taxonomic approach to analysing the oral microbiome of ancient individuals. They assessed the dental calculus of 34 archaeological individuals via 16S rRNA gene sequencing, dating in a time-transect from the Mesolithic era (7,000 years ago) to plaque from present-day individuals. The authors described shifts in the taxonomic profiles of bacteria in dental calculus during two major cultural shifts: adoption of farming and higher carbohydrate diets (around 8000-5000 years ago in Europe), and industrialisation during the Industrial Revolution (around 200 years ago in Northern Europe). In particular, they reported that the transition to agriculture resulted in an increased presence of taxa associated with periodontal disease such as *P. gingivalis*, *Tannerella* and *Treponema*. However, caries-associated taxa such as *S. mutans* only appeared with increased highly-refined sugar intake during the Industrial Revolution. While pioneering, this study suffered a range of issues that weakened some of the interpretations made about the response of the human oral microbiome to major archaeological transitions. Firstly, each temporal era was only represented by individuals from a single geographic region (Germany, Poland, United Kingdom, Australia), meaning site-specific characteristic biases were not controlled for. Secondly, while the study could confirm that known oral taxa were present in the dental calculus (de La Fuente et al., 2013), the use of amplicon sequencing, while cost effective, restricted the ability to authenticate that the amplified DNA was actually *ancient*. Amplification techniques using primers prevent the ability to evaluate ancient DNA degradation that exist primarily at DNA molecule termini (see section 1.5.3). A second factor in this was that the target regions of species-specific genes, or 16S rRNA genes, are much longer (150-300 base pairs, [bp]) than typical aDNA fragments (less than 100 bp) (Ziesemer et al., 2015), resulting in skews in taxonomic profiles. Thirdly, analytical processing reduced the number of sequences for assessing abundance variation of taxa associated with disease to just 34 DNA fragments - a level so low that is insufficient to describe biofilms consisting of hundreds of taxa - particularly when dealing with such a small sample size. Solutions were therefore needed to increase sample size and utilise techniques that allow less biased reconstruction and *authentication* of taxonomic profiles of the entire original endogenous microbial community in a calculus sample. Furthermore, these methods needed to be optimised to account for the short fragment lengths that occur with aDNA.

It was only when shotgun SGS sequencing was applied to dental calculus that researchers were able to demonstrate that the DNA associated with many different identifiable oral taxa were actually ancient. In two landmark papers, Warinner et al. (2014b) and Weyrich et al. (2017) were able to successfully confirm the presence of the large diversity of the oral microbiome preserved in ancient dental calculus - and in the case of Warinner et al. (2014b) - along with palaeoproteomic evidence. Both papers showed the long-term existence of the 'red complex' and more recently identified pathobionts, being possibly present as far back as 48 kya (Weyrich et al., 2017), and seemingly at a greater abundance than in modern plaque (Warinner et al., 2014b). Virulence factors, and more importantly some antibiotic resistance genes, were likely already present since the European Medieval era - showing analysis of ancient samples can have possible implications upon future strategies for combating antibiotic resistance (Warinner et al., 2014b). Both papers had sufficient preservation of the DNA in the dental cal-

culus of ancient individuals to yield complete and partial genomes of well-known oral taxa (the bacterium *T. forsythia* by Warinner et al. (2014b), and the archaeon *Methanobrevibacter oralis* by Weyrich et al. (2017)). Reconstruction of the partial *M. oralis* genome was sufficient enough to allow an attempt at molecular dating, which suggested that the taxon diverged from relatives *after* the divergence of Neanderthals and modern humans genomes. The authors subsequently proposed that oral strain sharing was still occurring after host divergence, emphasising evidence of inter-species hominin interaction that is currently being uncovered in host genomic studies (Hajdinjak et al., 2021; Prüfer et al., 2021).

Some aspects of these papers have since been criticised. For example, the molecular clock analysis by Weyrich et al. (2017) of a single only-partial ancient genome of *M. oralis*, with only a single modern relative and otherwise distantly related taxa, was noted to provide insufficient resolution to make interpretations of host interactions (Charlier et al., 2019). Indeed, sample sizes in shotgun ancient dental calculus papers aiming to characterise taxonomic trends or validating protocols, have either been small (e.g., less than 10 in Zhou et al., 2018c; Modi et al., 2020) or had heterogeneous sampling strategies (e.g. Mann et al., 2018; Weyrich et al., 2017) - although a handful of more recent studies have began to improve on this (e.g., ~50 in each of Velsko et al., 2018; Neukamm et al., 2020; Brealey et al., 2021). To tackle fundamental questions such as the development of the human microbiome and how this evolves in an entire species across millennia, large and comprehensive datasets are required. In this thesis, Manuscript A describes the largest shotgun-sequenced ancient dental calculus study to date (124 samples), with a sampling strategy designed to account for challenges in generating balanced sample sets in the context of palaeogenomics (see section 1.5.1). These issues in generating sufficient sample sizes in many ways can be attributed to the complications and challenges derived from the combination of metagenomics with aDNA.

1.5 Challenges in ancient metagenomics

1.5.1 Sample size

Generating sufficient and balanced sample sizes for robust metagenomic analysis often has added challenges when working with ancient DNA compared to modern genetic studies. Sample size is particularly important for microbiome studies as microbiome studies are generally focusing on the presence and abundance of particular taxa in dynamic ecosystem (when compared to the genomic analyses of single hosts, which are common in palaeogenomics). As such, the ecology of the microbiome can have many confounding and/or explanatory variables deriving from the different environments of different individuals. Such studies need sufficient sample size, and therefore power, to confirm any patterns or signatures that may be observed in the taxonomic and functional compositional analysis of microbiomes; the importance of which has been repeatedly emphasised (e.g. Casals-Pascual et al., 2020; Qian et al., 2020; Debelius et al., 2016; Silverman et al., 2018; Kelly et al., 2015).

In contrast to modern studies where samples are generally readily available from living individuals, or even grown in a laboratory, the majority of samples in the field of archaeogenetics are, as a whole, typically taken from archaeological and museum collections that are compiled from previous excavations and research. Sample availability is therefore reliant on the

types of archaeological sites that have been uncovered during excavation. Furthermore, in terms of geographical balance, regions such as Europe have had longer and more intensive archaeological investigations than other regions of the world, something reflected in the distribution of ancient human genomes (as demonstrated in Fig 1b., and Fig 5 of Marciniak and Perry, 2017; Orlando et al., 2021, respectively). Finally, for analysis of ancient oral microbiomes via dental calculus, museum curators historically have in some cases removed dental calculus deposits from teeth (Austin et al., 2019), under the assumption that they were 'dirt' and were not aesthetically pleasing and/or interfered with other dental morphology analyses . Generating sufficient sample sizes, while protecting the limited number of samples (representing unique cultural heritage) from over-zealous destructive analysis, is therefore a challenge that must be addressed in any microbiome study. Improving access to larger comparative datasets is addressed in this thesis in Manuscript B.

1.5.2 Data processing

Although collection of large sample sets of ancient samples is arguably more difficult than for modern metagenomics, increasing interest in ancient metagenomics is resulting in the creation of dedicated ancient microbiome research groups. Furthermore, established ancient population genetics groups are attempting to include these techniques alongside other research (typically added as 'disease' components to human population genetics studies). As this continues over time, sample and sequencing data sizes will naturally increase. However, these types of larger datasets needed for robust metagenomic analysis pose two primary challenges. Firstly, these analyses require heavy computational resources that are often performed on High-Performance-Computing (HPC) clusters and servers, and often consist of many complex linked steps. Secondly, given that archaeogenetics is a highly interdisciplinary research area, many different researchers come into the field with varying levels of computational backgrounds; something that is often incompatible with the scale of analysis required. This is particularly important when aDNA has a range of characteristics that require specialised analysis that 'off-the-shelf' modern metagenomic tools are not designed for (see section 1.5.3). Indeed, the only published metagenomics pipeline available at the beginning of this thesis that made reference to aDNA was metabit (Louvel et al., 2016). This pipeline however, performed no additional steps designed specifically for aDNA - such as characterisation of typical aDNA damage patterns or short fragments - and therefore no pipeline dedicated to ancient metagenomics existed, forcing researchers to design and execute their own workflows manually. In particular, the fundamental steps of taxonomic profiling (i.e., comparison of sequenced DNA molecules against thousands of genomes at once) and authentication (checking for aDNA damage patterns, etc., see section 1.5.3), was not automated. Pipelines such as HOPS (Hübler et al., 2019) have only recently started to be developed to automate some of these steps. However, pipelines like HOPS are often only designed for a particular step of a whole ancient metagenomic data processing workflow. Therefore, additional manual work linking with other steps of such a workflow (such as quality control and/or sequencing artefact removal) are required, something that is not optimal for scalability.

The most common solution for making computationally-heavy analysis more efficient is through the use of pipelines - tools that string together a series of required steps and automate

the execution of these commands. This reduces the amount of 'hands-on' work required of the researcher, as well as reducing the risk of user error. Furthermore, automation adds a level of standardisation in the analysis across the field. While aDNA pipelines to tackle this issue have been published in the past (PALEOMIX, EAGER, metabit, Schubert et al., 2014; Peltzer et al., 2016; Louvel et al., 2016), these do not follow current standards in software development, nor include analyses representing the latest workflows in ancient metagenomics. All three pipelines could be considered 'abandoned', or at least have not had any major updates in recent years. Such major updates would need to consist of either the addition of new tools or updating parameters to the latest recommendations. Furthermore, they were designed to only run on single machines (desktop, laptop, or single-node server), and therefore the level of parallelisation of analysis of multiple samples is limited. This is not ideal when more recent HPC clusters can have hundreds or thousands of nodes. Indeed, while the more recent HOPS ancient metagenomics pipeline has some level of integration with the SLURM HPC scheduling system, this only works for labs running this particular scheduler, and researchers on other systems are unable to efficiently run the pipeline.

While pipeline *development* is quite common in bioinformatics, this does not mean that they are accessible. Many researchers entering palaeogenomics, including ancient metagenomics, may often have little to no computational knowledge, such as for working on command-line interfaces, let alone on working on HPC cluster infrastructures. More specifically, for the young field of ancient metagenomics, routine practices are only just emerging, and therefore educational curricular or training courses in this area do not exist for the processing or interpretation of such sequencing data. To ensure that research groups are able to analyse large-scale datasets efficiently, but also to a sufficient quality required for aDNA studies, tools and pipelines need to be designed to be accessible to a wide range of labs in different contexts. Supporting this will correspondingly assist the increasing of sizes of datasets that can be used by other researchers. Accessibility can be considered to be primarily aided by two main factors: well-designed software for ease of use, and documentation. Looking at existing solutions, the metagenomics pipeline Metabit (Louvel et al., 2016) had little to no documentation on usage, nor output interpretation. While documentation was better with the ancient genomics pipeline PALEOMIX (Schubert et al., 2014), configuration of a pipeline run requires significant editing of a file format that is not well known outside of computer science, and again, available output documentation is limited. While PALEOMIX and Metabit offered command-line interfaces (CLI) - the preferred method of executing pipelines for experienced bioinformaticians - these are not particularly user-friendly for novices. The ancient genomics pipeline EAGER (Peltzer et al., 2016) took a different approach by developing a Graphical User Interface (GUI). GUIs are more familiar to users that are used to point-and-click applications such as word processors or spreadsheet software. However, as with the other pipelines, documentation is also relatively sparse, making it more difficult for most users to evaluate suitable parameters and expected results in the context of aDNA.

Finally, given the complexity of modern SGS data analyses, *reproducibility* has recently become a particular area of focus in the wider bioinformatics field (Sczyrba et al., 2017; Baker, 2016; Kim et al., 2018; Hothorn and Leisch, 2011; Gauthier et al., 2019). As well as for training purposes, being able to rerun analyses of other researchers and check that results are the same as reported in a publication, such as during peer-review, can be very useful. Previous

aDNA pipelines had not been designed to make this easy for users. The three pipelines required manual installation of all tools of the pipeline one by one, requiring system administrator maintenance to ensure versions are kept up to date. This can be problematic, as changes to a version of one tool can lead to incompatibilities with other tools in a pipeline. Latest bioinformatics practises nowadays promote the use of software containers, i.e., singular 'image' files that include an unmodifiable snapshot of all required tools and versions preconfigured for a pipeline (Gruening et al., 2018). Other users therefore can download the same image and run the same command as the original publication to successfully replicate the initial analysis. The use of this technology therefore allows *portability* across different computing infrastructures, and therefore supports reproducibility.

These issues are addressed in Manuscript C by the development of a completely re-written and extended ancient (meta-)genomic pipeline designed specifically for high-throughput processing. The development of this pipeline, named nf-core/eager, specifically focuses on reproducibility and accessibility.

1.5.3 DNA yield and quality

Next, even if sufficient ancient samples are collected and libraries are able to be generated and analysed, there is no guarantee that the sample will yield enough usable DNA for analysis in these pipelines. In the absence of repair mechanisms of living cells, all DNA molecules will eventually undergo forms of postmortem chemical change that slowly degrade the molecule. For aDNA, two main processes have been observed: fragmentation and nucleotide chemical modification (the latter often referred to as miscoding lesions or 'damage').

Fragmentation of DNA molecules comes most commonly following depurination. This consists of hydrolysis of purine bases (with guanine [G] undergoing hydrolysis at a higher rate than other bases) resulting in abasic sites, making the phosphate backbone also susceptible to hydrolysis. This susceptibility then increases the occurrence of breakage of the phosphate backbone at single sites on one of the strands, increasing the chance of the entire cleavage of the molecule around this location (Lindahl, 1993). If two of these breakages occur close to each other, the hydrogen bonds between the nucleotides are insufficient to hold the stands together, resulting in the cleavage of the molecule, and single-stranded overhangs at the end of the cleaved molecules. This is something that has been corroborated in observations of shotgun aDNA libraries, where it has been directly observed that when compared to a reference genome, there is an increase in the frequency of Gs in the reference prior to the beginning of sequenced molecules (Briggs et al., 2007). Over time this process thus results in shorter and shorter molecules, typically observed in aDNA to be less than 100 bp. This factor was a critical issue in early PCR-based aDNA studies that targeted single genes or sequences through amplification, as longer modern DNA molecules are more amenable for amplification than short sequences due to the presence of both priming sites on the same DNA strand (as demonstrated more recently in ancient microbiome studies by Ziesemer et al., 2015).

The second most well-characterised characteristic of aDNA is nucleotide misincorporations caused by deamination. This process is where hydrolysis of an amino-group of a nucleotide is modified or removed. In the context of aDNA, the most common deamination is of Cytosine (C) to Uracil (U) (Lindahl, 1993), which is subsequently read by non-proofreading polymerases

as a Thymine (T); something often known as C-T miscoding lesions. Deamination occurs more frequently on structurally less stable single-stranded molecules (e.g. Frederico et al., 1990). As above, after fragmentation via depurination, molecules will often have uneven strand ends where the backbone of each strand has been cleaved in different places (also known as single-stranded overhangs). This means that the exposed nucleotides of a longer-overhang strand are more susceptible to deamination than internal regions of molecules (Briggs et al., 2007). After the advent of SGS sequencing, it was observed that frequencies of T substitutions on the reference genome were higher than usual on the first base of the 5' end of sequenced molecules (known as 'reads') and gradually decreased further internally into the molecule (Briggs et al., 2007). C-T lesions on single-stranded overhangs are only typically observed on double-stranded SGS libraries at 5' ends. This is because library preparation protocols typically only repair via 'filling in' at this end, whereas 3' overhangs are enzymatically 'clipped off' (known as 'blunt ending'). Thus, complementary G-A misincorporations are reflected in the filled-in strand. Note that newer single-stranded library preparation retains C-T lesions at both ends (e.g., Gansauge et al., 2017).

These two factors result in a variety of technical challenges when carrying out ancient metagenomic studies. One of the initial steps of most metagenomic studies is to identify which taxa sequenced DNA reads are derived from, by comparing each sequence against a large database of reference sequences of known taxa (known as aligning or mapping; Quince et al., 2017). The accuracy and sensitivity of these alignments depend on the amount of informative sequences that each read contains - the shorter the read, the less informative it is. For example, low sequence complexity (i.e. a low diversity of nucleotides) in short reads means they are more likely to align to the reference genomes of many more taxa (that may either share stretches of a genome via relatedness, or by chance). The result of this is that aligners are often unable to know which is the correct genome the read is originally derived from (as demonstrated in Velsko et al., 2018). When dealing with strain-level analysis, short fragments also pose an obstacle when trying to separate the genomes of two highly similar strains. Typically, strain separation is performed by identifying multiple single-nucleotide polymorphisms (or SNPs) on a single read, which shows that these two genetic variations are derived from a single strain. When partially overlapping with another read containing the same sequence, this shows the genetic variation of the first read can be linked to another SNP on the new read. By repeatedly linking reads with overlapping SNPs together, strains can thus be separated (an approach called haplotyping). However, when fragments are short, this reduces the chance a read has more than one SNP and haplotyping cannot occur, as the genetic variation that *should* be associated together cannot be confidently linked (Nicholls et al., 2020). Miscoding lesions also complicates the confident calling of 'true' SNPs. In haploid organisms such as bacteria, calling a SNP typically occurs with the assumption that all reads aligned to the same position on a reference genome should have the same nucleotide. However, postmortem miscoding lesions adds uncertainty to SNP calling (e.g., as in Lindgreen et al., 2014), because in the presence of damage, not all nucleotides at a given position will be the same. Furthermore, reads from other very closely related species that are also present in a sample (see section 1.5.4) can *also* align to the same position but include a *different* SNP (Warinner et al., 2017), adding further complexity for reliable SNP calling for downstream analyses. In all cases, special analytical approaches are needed to account for the degraded nature of aDNA.

It is important to note that both of these natural chemical degradation processes occur at varying rates throughout existence of a sample. While the 'embedding' of DNA in mineral substrates of archaeological material stabilises different components of the DNA molecule 'in-place', and thus slows the rate of degradation, certain environmental conditions can further reduce (or accelerate) the rate of these processes (e.g., temperature, pH, and water content; see Allentoft et al., 2012; Kistler et al., 2017). Suitable preservation conditions (e.g., permafrost environments) can result in the preservation of DNA as old as 1.2 million years (van der Valk et al., 2021). In general, cold and dry environments are preferential, with cold environments reducing the kinetic energy available for chemical changes and desiccation reducing the amount of water available for hydrolysis (Bollongino et al., 2008). However, the relationship of these factors to aDNA preservation are still not well understood (Sawyer et al., 2012; Kistler et al., 2017), with age being shown to be a less informative factor in the prediction of preservation than previously thought (Smith et al., 2003; Sawyer et al., 2012). Increasing evidence of successful aDNA retrieval is being presented from warmer and wetter environments, such as in Africa or Polynesia (Gallego Llorente et al., 2015; Skoglund et al., 2016). This typically occurs when burial environments differ from the general climatic environment of the region (such as in caves). Therefore, in addition to potentially limited sample sizes, the amount of degradation and yield can vary highly within a sample set. This must therefore be accounted for when preparing sample- and analytical strategies for ancient microbiome studies.

1.5.4 Contamination

Even though DNA *can* be preserved (albeit in a fragmented and damaged state), this does not mean that the entire DNA present in an archaeological sample is from the original DNA of the organism (Der Sarkissian et al., 2014; Philips et al., 2017). A major challenge for ancient microbiome research is the separation of the already complex microbial community of the original microbiome from exogenous environmental biomass. Burial surroundings (e.g., sediment/soil) of archaeological samples are dynamic environments, and more recent, less-fragmented and undamaged 'contaminating' DNA often becomes more likely to be sequenced as the original DNA degrades. This contaminating DNA can come from a variety of sources, including microorganisms present in the soil or sediment (Jans et al., 2004), percolating water (Haile et al., 2007), and also in post-burial contexts such as handling by excavators (human DNA contamination) or storage in archaeological collections (microorganism) (Llamas et al., 2017). In addition, despite (in)famously defined ultra-strict setups and guidelines for working with aDNA (Cooper and Poinar, 2000; Llamas et al., 2017, e.g., dedicated laboratory for extraction, negative pressure flow, full body suits, face masks, regularly changed gloves, etc.), laboratory worker handling, and laboratory reagents (Leonard et al., 2007; Weyrich et al., 2019) remain sources of contamination in aDNA studies.

'Non-endogenous' taxa complicate downstream analysis in a variety of ways. Firstly, contaminating strains from modern sources can result in a false positive identification of a given taxa (i.e., that a taxon was not actually originally present in an ancient individual; Warinner et al., 2017; Harkins et al., 2015; Müller et al., 2016). During taxonomic profiling via the aligning of reads against large databases of reference genomes, the presence of uncharacterised environmental relatives of true (also present) endogenous taxa can result in the endogenous

taxa displaying a greater abundance than it actually had. In other words, when the reference genome of an environmental relative does not exist in a database, in lieu of the original genome, reads will generally align against the most similar genome that is present in the database. This artificial skewing of the abundance profiles subsequently makes the identification of important taxa that may have changed in response to variables of interest difficult to identify with confidence. Conversely, the presence of low-abundant 'rare' but informative taxa can become indistinguishable from 'noise' when equally low abundance environmental contamination is present (Davis et al., 2018). Additionally, as described above (section 1.5.3), cross-mapping from contaminating environmental taxa makes downstream analysis involving SNPs (e.g., for phylogenetics) more complicated, as the contamination of closely related modern taxa means that SNPs cannot be confidently made at true SNP sites in the genome of the original taxon.

It is also important to mention that the analysis of ancient microbes is not a brand new field of research (Spyrou et al., 2019). Ancient pathogen genomes were first reconstructed a decade ago (Bos et al., 2011), and accordingly, *some* dedicated for tools and techniques have been developed for such contexts (Bos et al., 2014; Zhou et al., 2018b; Hübler et al., 2019; Dimopoulos et al., 2020). While the challenges for ancient pathogen genome studies have a large overlap with the challenges in ancient microbiome research, microbiome work has the additional challenge of taxonomic scale. Ancient pathogen genomicists typically work on identifying and analysing a single pathogen species, whereas ancient microbiome research aims to characterise hundreds or even thousands of taxa at once. Pathogens that are typically studied are generally very well characterised in modern contexts - with extensive reference databases, high-quality and annotated genomes, and well researched diversity within the given genus. In contrast, many species detected in microbiome studies are under- or even uncharacterised, and only have partially assembled or draft genomes available. The high-quality reference genomes of pathogen species and their relatives are therefore amenable for the development of techniques such as DNA probe arrays. These allow for the 'capture' or 'enrichment' of the DNA of a given species of interest from the metagenomic 'soup' of environmental DNA (see Furtwängler et al., 2020, for a review). This improves the genomic coverage of ancient samples and increases confidence in SNP based strain separation. Yet for ancient microbiome work, the lack of diversity in databases or under-characterisation means that separating related species and even strains makes developing efficient captures more difficult. While a suite of techniques and tools have been developed for the assessment of preservation and separation of aDNA from contamination in ancient human genomics (see Peyrégne and Prüfer, 2020, for a review), equivalent *in silico* workflows and tools dedicated to ancient metagenomics remain heterogeneous and scarce. In this thesis, such a workflow and related tools are presented in Manuscripts A and C.

1.6 Aims

The scope of the present work can be considered to have two parts:

- Use large-scale shotgun sequenced ancient dental calculus datasets to trace deep-time evolutionary and anthropological histories of hominid oral microbiomes (Manuscript A)
- Address challenges currently present in the wider field of ancient metagenomics by producing pertinent and useful resources for the community (Manuscript A, B and C).

The research objectives of these two parts are as follows:

Part One

- Generate a large dataset of ancient dental calculus from ancient and extinct humans, and their close host-genomic relatives
- Develop a workflow for robust data analysis of oral microbiomes derived from ancient dental calculus
- Assess whether an ancestral state of the oral microbiome can be reconstructed
- Assess whether present-day understanding of healthy and diseased states of the oral microbiome still applies deep into human evolutionary history
- Assess whether inferences about human behavioural or cultural evolution can be made from large-scale shotgun ancient oral microbiome datasets

Part Two

- Develop new techniques and tools to assist in rapid assessment of ancient microbiome preservation in archaeological samples
- Build a useful and high-quality resource for accessing previously published ancient metagenomic public datasets
- Update existing palaeogenomic pipelines to bring in line with current high-throughput bioinformatic and software development practices
- Extend existing palaeogenomic pipelines to include tools and authentication specific to ancient metagenomics
- Build these ancient metagenomics resources in a sustainable and maintainable manner via community involvement

2 Overview of the manuscripts

2.1 Manuscript A

Status Published

Reference

Fellows Yates, J. A., Velsko, I. M., Aron, F., Posth, C., Hofman, C. A., Austin, R. M., Parker, C. E., Mann, A. E., Nägele, K., Weedman Arthur, K., Arthur, J. W., Bauer, C. C., Crevecoeur, I., Cupillard, C., Curtis, M. C., Dalén, L., Díaz-Zorita Bonilla, M., Carlos Díez Fernández-Lomana, J. C., Drucker, D. G., Escribano Escrivá, E., Francken, M., Gibbon, V. E., Gonzalez Morales, M., Grande Mateu, A., Harvati, K., Henry, A. G., Humphrey, L., Menéndez M., Mihailović, D., Peresani, M., Rodríguez Moroder, S., Roksandic, M., Rougier, H., Sázelová, S., Stock, J. T., Straus, L. G., Svoboda, J., Teßemann, B., Walker, M. J., Power, R. C., Lewis, C. M., Sankaranarayanan, K., Guschanski, K., Wrangham, R., Dewhirst, F. E., Salazar-García, D. C., Krause, J., Herbig, A., & Warinner, C. (2021). The evolution and changing ecology of the African hominid oral microbiome. *Proceedings of the National Academy of Sciences*. 118(20), e2021655118. https://doi.org/10.1073/pnas.2021655118

Summary

In this study, I demonstrate the use of large-scale aDNA analysis of dental calculus (n = 124) from howler monkeys, gorillas, chimpanzees, Neanderthals, and ancient and present-day human individuals to reconstruct ancestral states of the hominid oral microbiome. I develop a novel and extensive workflow dedicated for the assessment of preservation and authentication of aDNA of ancient oral microbial communities. This study demonstrates how the power of shotgun metagenomics coupled with large sample sets, 'minimally-invasively' collected from ancient individuals, can add new perspectives to research into the health and disease states of the human oral microbiome, as well as contribute evidence towards anthropological debates regarding of the development of modern humans.

2.2 Manuscript B

Status Published

Reference

Fellows Yates, J. A., Andrades Valtueña, A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-Lopez, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., Heintzman, P. D., Spyrou, M. A., Hübner, A., Gancz, A. S., Hider, J., Allshouse, A. F., Zaro, V., & Warinner, C. (2021). Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Scientific Data*, 8(1), 31. https://doi.org/10.1038/s41597-021-00816-y

Summary

This manuscript describes an open-access repository of curated metadata of more than 1000 published ancient metagenomic samples. By providing a centralised but accessible database of such samples, which includes dental calculus, the project will assist ancient metagenomics researchers improve the statistical power of their studies through larger and more robust sample sizes. In addition it aims to promote data reuse, therefore protecting the cultural heritage that is finite archaeological remains.

2.3 Manuscript C

Status Published

Reference

Fellows Yates, J. A., Lamnidis, T. C., Borry, M., Andrades Valtueña, A., Fagernäs, Z., Clayton, S., Garcia, M. U., Neukamm, J., & Peltzer, A. (2021). Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ*, 9, e10947. https://doi.org/10.7717/peerj.10947

Summary

This manuscript describes nf-core/eager, a cutting-edge bioinformatics pipeline for the processing of ancient genomic data. Within this, I have expanded the functionality of a popular but now undeveloped pipeline (EAGER) to include automated in-parallel taxonomic profiling and authentication of metagenomic samples (such as microbiomes). The complete reimplementation of this pipeline in the Nextflow programming framework allows for highly reproducible, scalable, and time/resource/cost- efficient processing and authentication of the large-scale datasets required for robust ancient microbiome analysis - in addition to standard host or pathogen genomic studies. Furthermore, through user-friendly usage and extensive documentation, nf-core/eager helps provide access to high-quality authentication and verification of aDNA datasets, allowing both new and established aDNA laboratories to more easily address the challenges that come from working with aDNA.

3 Manuscript A: The evolution and changing ecology of the African hominid oral microbiome

3.1 Overview and contribution

Manuscript Nr.: A

Title of Manuscript The evolution and changing ecology of the African hominid oral microbiome.

Authors Fellows Yates, J. A., Velsko, I. M., Aron, F., Posth, C., Hofman, C. A., Austin, R.M., Parker, C. E., Mann, A. E., Nägele, K., Weedman Arthur, K., Arthur, J. W., Bauer, C. C., Crevecoeur, I., Cupillard, C., Curtis, M. C., Dalén, L., Díaz-Zorita Bonilla, M., Díez Fernández-Lomana, J. C., Drucker, D. G., Escribano Escrivá, E., Francken, M., Gibbon, V. E., Gonzalez Morales, M., Grande Mateu, A., Harvati, K., Henry, A. G., Humphrey, L., Menéndez M., Mihailović, D., Peresani, M., Rodríguez Moroder, S., Roksandic, M., Rougier, H., Sázelová, S., Stock, J. T., Straus, L. G., Svoboda, J., Teßmann, B., Walker, M. J., Power, R. C., Lewis, C. M., Sankaranarayanan, K., Guschanski, K., Wrangham, R., Dewhirst, F. E., Salazar-García, D. C., Krause, J., Herbig, A., & Warinner, C.

Citation Fellows Yates, J. A., Velsko, I. M., Aron, F., Posth, C., Hofman, C. A., Austin, R. M., Parker, C. E., Mann, A. E., Nägele, K., Weedman Arthur, K., Arthur, J. W., Bauer, C. C., Crevecoeur, I., Cupillard, C., Curtis, M. C., Dalén, L., Díaz-Zorita Bonilla, M., Díez Fernández-Lomana, J. C., Drucker, D. G., Escribano Escrivá, E., Francken, M., Gibbon, V. E., Gonzalez Morales, M., Grande Mateu, A., Harvati, K., Henry, A. G., Humphrey, L., Menéndez M., Mihailović, D., Peresani, M., Rodríguez Moroder, S., Roksandic, M., Rougier, H., Sázelová, S., Stock, J. T., Straus, L. G., Svoboda, J., Teßmann, B., Walker, M. J., Power, R. C., Lewis, C. M., Sankaranarayanan, K., Guschanski, K., Wrangham, R., Dewhirst, F. E., Salazar-García, D. C., Krause, J., Herbig, A., & Warinner, C. (2021). The evolution and changing ecology of the African hominid oral microbiome. *Proceedings of the National Academy of Sciences.* 118(20), e2021655118. https://doi.org/10.1073/pnas.2021655118

The candidate is
☑ First author, ☐ Co-first author, ☑ Corresponding author, ☐ Co-author.

Status Published

Proportion (in %) of authors in the publication (indicated from 20%)

Author	Concept	Data Analysis	Experiment	Manuscript Composition	Material Provision
Fellows Yates, J. A.	35	70	0	50	0
Velsko, I. M.	0	20	0	15	0
Aron, F.	0	0	80	0	0
Posth, C.	0	0	0	0	0
Hofman, C. A.	0	0	0	0	0
Austin, R. M.	0	0	0	0	0
Parker, C. E.	0	0	0	0	0
Mann, A. E.	0	0	0	0	0
Nägele, K.	0	0	0	0	0
Weedman Arthur, K.	0	0	0	0	0
Arthur, J. W.	0	0	0	0	0
Bauer, C. C.	0	0	0	0	0
Crevecoeur, I.	0	0	0	0	0
Cupillard, C.	0	0	0	0	0
Curtis, M. C.	0	0	0	0	0
Dalén, L.	0	0	0	0	0
Díaz-Zorita Bonilla, M.	0	0	0	0	0
Díez Fernández-Lomana, J. C.	0	0	0	0	0
Drucker, D. G.	0	0	0	0	0
Escribano Escrivá, E.	0	0	0	0	0
Francken, M.	0	0	0	0	0
Gibbon, V. E.	0	0	0	0	0
Gonzalez Morales, M.	0	0	0	0	0
Grande Mateu, A.	0	0	0	0	0
Harvati, K.	0	0	0	0	0
Henry, A. G.	0	0	0	0	0
Humphrey, L.	0	0	0	0	0
Menéndez M.	0	0	0	0	0
Mihailović, D.	0	0	0	0	0
Peresani, M.	0	0	0	0	0
Rodríguez Moroder, S.	0	0	0	0	0
Roksandic, M.	0	0	0	0	0
Rougier, H.	0	0	0	0	0

Continued...

Author	Concept	Data Analysis	Experiment	Manuscript Composition	Material Provision
Sázelová, S.	0	0	0	0	0
Stock, J. T.	0	0	0	0	0
Straus, L. G.	0	0	0	0	0
Svoboda, J.	0	0	0	0	0
Teßmann, B.	0	0	0	0	0
Walker, M. J.	0	0	0	0	0
Power, R. C.	0	0	0	0	0
Lewis, C. M.	0	0	0	0	0
Sankaranarayanan, K.	0	0	0	0	0
Guschanski, K.	0	0	0	0	0
Wrangham, R.	0	0	0	0	0
Dewhirst, F. E.	0	0	0	0	0
Salazar-García, D. C.	0	0	0	0	0
Krause, J.,	0	0	0	0	0
Herbig, A.	0	0	0	0	0
Warinner, C.	40	0	0	30	30

3.2 Article

The evolution and changing ecology of the African hominid oral microbiome

James A. Fellows Yates[a,b,1], Irina M. Velsko[a], Franziska Aron[a], Cosimo Posth[a,c], Courtney A. Hofman[d,e], Rita M. Austin[d,e,f], Cody E. Parker[a,g], Allison E. Mann[h], Kathrin Nägele[a], Kathryn Weedman Arthur[i], John W. Arthur[i], Catherine C. Bauer[j], Isabelle Crevecoeur[k], Christophe Cupillard[l,m], Matthew C. Curtis[n], Love Dalén[o,p], Marta Díaz-Zorita Bonilla[q,r], J. Carlos Díez Fernández-Lomana[s], Dorothée G. Drucker[t], Elena Escribano Escrivá[u], Michael Francken[v], Victoria E. Gibbon[w], Manuel R. González Morales[x], Ana Grande Mateu[y], Katerina Harvati[t,z,aa], Amanda G. Henry[ab], Louise Humphrey[ac], Mario Menéndez[ad], Dušan Mihailović[ae], Marco Peresani[af,ag], Sofía Rodríguez Moroder[ah], Mirjana Roksandic[ai], Hélène Rougier[aj], Sandra Sázelová[ak], Jay T. Stock[al,am,an], Lawrence Guy Straus[ao], Jiří Svoboda[ak,ap], Barbara Teßmann[aq,ar], Michael J. Walker[as], Robert C. Power[b,at], Cecil M. Lewis[d], Krithivasan Sankaranarayanan[au], Katerina Guschanski[av,aw,ba], Richard W. Wrangham[ax], Floyd E. Dewhirst[ay,az], Domingo C. Salazar-García[at,bb,bc,bd], Johannes Krause[a,be], Alexander Herbig[a], and Christina Warinner[a,d,bf,1]

Edited by Robert R. Dunn, North Carolina State University, Raleigh, NC, and accepted by Editorial Board Member James F. O'Connell March 22, 2021 (received for review October 16, 2020)

The oral microbiome plays key roles in human biology, health, and disease, but little is known about the global diversity, variation, or evolution of this microbial community. To better understand the evolution and changing ecology of the human oral microbiome, we analyzed 124 dental biofilm metagenomes from humans, including Neanderthals and Late Pleistocene to present-day modern humans, chimpanzees, and gorillas, as well as New World howler monkeys for comparison. We find that a core microbiome of primarily biofilm structural taxa has been maintained throughout African hominid evolution, and these microbial groups are also shared with howler monkeys, suggesting that they have been important oral members since before the catarrhine–platyrrhine split ca. 40 Mya. However, community structure and individual microbial phylogenies do not closely reflect host relationships, and the dental biofilms of *Homo* and chimpanzees are distinguished by major taxonomic and functional differences. Reconstructing oral metagenomes from up to 100 thousand years ago, we show that the microbial profiles of both Neanderthals and modern humans are highly similar, sharing functional adaptations in nutrient metabolism. These include an apparent *Homo*-specific acquisition of salivary amylase-binding capability by oral streptococci, suggesting microbial coadaptation with host diet. We additionally find evidence of shared genetic diversity in the oral bacteria of Neanderthal and Upper Paleolithic modern humans that is not observed in later modern human populations. Differences in the oral microbiomes of African hominids provide insights into human evolution, the ancestral state of the human microbiome, and a temporal framework for understanding microbial health and disease.

dental calculus | microbiome | Neanderthal | primate | salivary amylase

The oral cavity is colonized by one of the most diverse sets of microbial communities of the human body, currently estimated at over 600 prevalent taxa (1). Dental diseases, such as caries and periodontitis, remain health burdens in all human populations despite hygiene interventions (2, 3), and oral microbes are often implicated in extraoral inflammatory diseases (4, 5). To date, most oral microbiome research has focused on clinical samples obtained from industrialized populations that have daily oral hygiene routines and access to antibiotics (1, 6), but far less is known about the global diversity of the oral microbiome, especially from diverse past and present nonindustrialized societies (7). The oral cavity contains at least six distinct habitats, but dental biofilms, including both supra- and subgingival dental plaque, are among the most diverse and clinically important (1, 6, 8). During life, these dental biofilms naturally and repeatedly calcify, forming dental calculus (tooth tartar) (9), a robust, long-term record of the oral microbiome (10). Archaeological dental calculus has been shown to preserve authentic oral bacterial metagenomes in a wide range of historic and prehistoric populations and up to 50 thousand years ago (ka) (10–13). As such, dental calculus presents an opportunity to directly investigate the evolution of the hominid microbiome and to reconstruct ancestral states of the modern human oral microbiome. In addition, because research has shown that evolutionary traits, diet, and cultural behaviors shape modern human microbiome structure and function at other body sites, such as the gut and skin microbiomes (14–18), investigating ancient oral metagenomes has the potential to reveal valuable information about major events in modern human evolution and prehistory, such as predicted dietary changes during the speciation of *Homo* (19–21) and the direct

Significance

The microbiome plays key roles in human health, but little is known about its evolution. We investigate the evolutionary history of the African hominid oral microbiome by analyzing dental biofilms of humans and Neanderthals spanning the past 100,000 years and comparing them with those of chimpanzees, gorillas, and howler monkeys. We identify 10 core bacterial genera that have been maintained within the human lineage and play key biofilm structural roles. However, many remain understudied and unnamed. We find major taxonomic and functional differences between the oral microbiomes of *Homo* and chimpanzees but a high degree of similarity between Neanderthals and modern humans, including an apparent *Homo*-specific acquisition of starch digestion capability in oral streptococci, suggesting microbial coadaptation with host diet.

Author contributions: J.A.F.Y., I.M.V., C.A.H., C.M.L., K.S., J.K., A.H., and C.W. designed research; J.K., A.H., and C.W. codirected research; J.A.F.Y., I.M.V., F.A., C.P., C.A.H., R.M.A., C.E.P., A.E.M., K.N., K.W.A., J.W.A., C.C.B., I.C., C.C., M.C.C., L.D., M.D.-Z.B., J.C.D.F.L., D.G.D., E.E.E., M.F., V.E.G., M.R.G.M., A.G.M., K.H., A.G.H., L.H., M.M., D.M., M.P., S.R.M., M.R., H.R., S.S., J.T.S., L.G.S., J.S., B.T., M.J.W., R.C.P., C.M.L., K.S., K.G., R.W.W., D.C.S.-G., and C.W. performed research; J.A.F.Y. contributed new reagents/analytic tools; J.A.F.Y., I.M.V., F.E.D., A.H., and C.W. analyzed data; and J.A.F.Y., I.M.V., and C.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. R.R.D. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: fellows@shh.mpg.de or warinner@shh.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2021655118/-/DCSupplemental.

Published May 10, 2021.

ANTHROPOLOGY

interaction of Neanderthals and modern humans during the Late Pleistocene (22).

To better understand the evolutionary ecology of the African hominid microbiome, we generated and analyzed 109 dental calculus metagenomes from present-day modern humans ($n = 8$), gorillas (*Gorilla*, $n = 29$), chimpanzees (*Pan*, $n = 20$), Neanderthals ($n = 13$), and two groups of archaeological modern humans associated with major lifestyle transitions (preagricultural, $n = 20$; preantibiotic, $n = 14$), as well as New World howler monkeys ($n = 5$) for comparison (*SI Appendix*, Fig. S1). To account for potential sampling biases, we analyzed multiple subspecies and populations of each African great ape genus, which were obtained from C20th or C21st-collected museum collections, and for modern humans we sampled multiple populations from both Africa and Europe. To this, we added previously published microbiome data from chimpanzees ($n = 1$) (13), Neanderthals ($n = 4$) (13), and present-day modern humans ($n = 10$) (23), for a total dataset of 124 individuals (Fig. 1*A*, *SI Appendix*, Table S1, and Dataset S1). We also generated eight new radiocarbon dates for archaeological individuals, for a total of 44 directly or indirectly dated ancient individuals in this study (Dataset S1).

Here, we investigate the structure, function, and core microbial members of the human oral microbiome within an evolutionary framework, seeking to determine whether a core microbiome can be defined for each African hominid group, whether the core is phylogenetically coherent, and whether some members of the core are specific to certain host groups. We test whether the oral microbiome of hominids reflects host phylogeny, finding that African hominid oral microbiota are distinguished by major taxonomic and functional differences that only weakly reflect host relationships and are likely influenced by other physiological, dietary, or behavioral factors. We compare the microbial profiles of Neanderthals and modern humans and, contrary to expectations (12, 13), find a high consistency of oral microbiome structure within *Homo*, regardless of geography, time period, or diet/lifestyle. We detect the persistence of shared genetic diversity in core taxa between Neanderthals and Upper Paleolithic humans prior to 14 ka, supporting a growing body of evidence for earlier admixture and interaction in Ice Age Europe (24, 25). Finally, we explore possible implications of our findings on *Homo*-associated encephalization (19, 26) and the role of dietary starch in human evolution (20, 21) by investigating the evolutionary history of amylase-binding capability by oral streptococci. We find that amylase binding is an apparent *Homo*-specific trait, suggestive of microbial coadaptation to starch-rich diets early in human evolution.

Results

Preservation of Oral Microbiota in Dental Calculus. Authenticating ancient DNA (aDNA) preservation is a necessary and essential step for all paleogenomic studies. However, these methods have been underdeveloped for ancient microbiomes. Here, we apply a multistep procedure of both conventional and new methods to evaluate and validate oral microbiome preservation in our dataset (*SI Appendix*, Fig. S2). First, we applied a reference-based metagenomic binning of reads to the National Center for Biotechnology Information (NCBI) nucleotide (nt) database (27) (Dataset S2) and then developed and applied a method to assess the decay of the cumulative percentage of known oral taxa in samples compared to a panel of oral and nonoral reference metagenomes (*SI Appendix*, Fig. S3*A* and section S3.4.1). This allowed us to remove samples that did not exhibit a taxonomic composition consistent with an oral origin. We then cross validated these results using SourceTracker (28) (*SI Appendix*, Fig. S3*B*) and inspection by principal coordinate analysis (PCoA, *SI Appendix*, Fig. S5 *A–D*). To samples exhibiting good oral microbiome preservation (Fig. 1*B*), we then applied the R package decontam (29) to detect and remove putative laboratory and environmental contaminants prior to downstream analysis (*SI Appendix*, section S3.6). Next, we examined each dataset and confirmed the presence of DNA damage characteristics of ancient samples, including short fragment lengths and elevated levels of

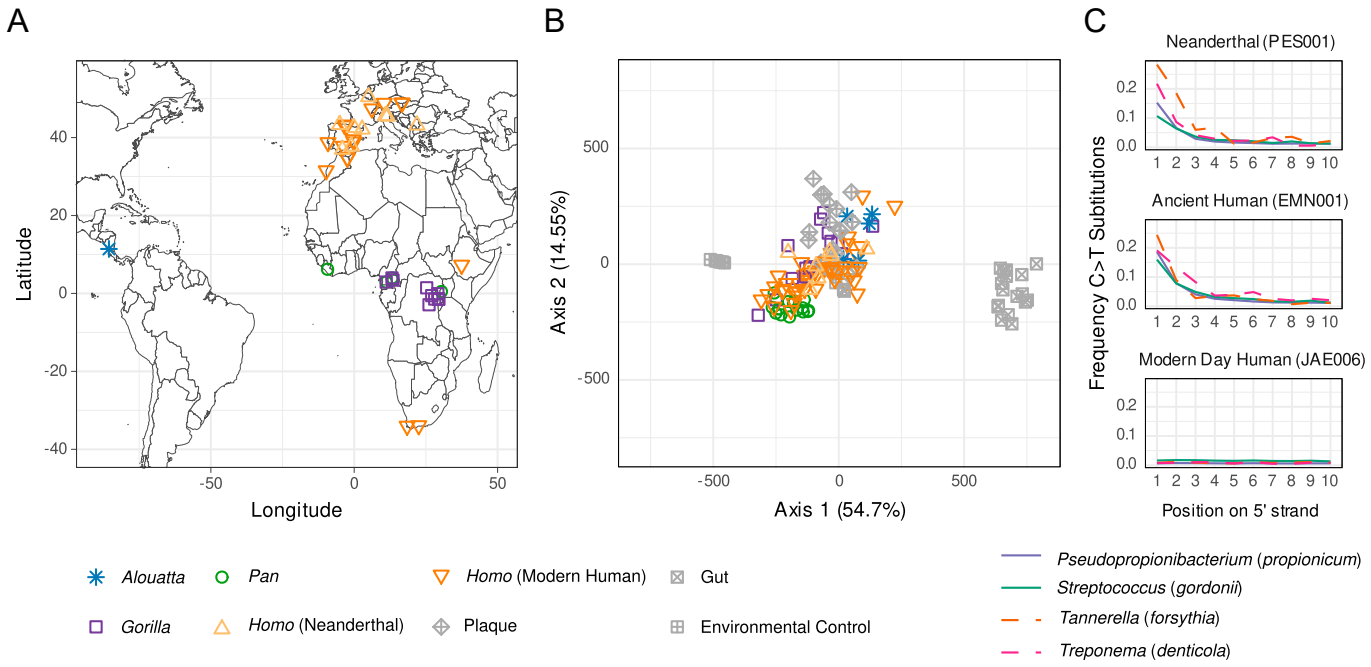

Fig. 1. Sample locations and oral microbiome authentication of ancient dental calculus. (*A*) Sample locations. (*B*) PCoA comparing euclidean distances of microbial genera of well-preserved ancient and present-day dental calculus to environmental proxy controls (degraded archaeological bone) and present-day dental plaque and feces. Ancient dental calculus is distinct from gut and archaeological bone but overlaps with present-day dental plaque. (*C*) Representative DNA damage patterns for Neanderthals and ancient and present-day modern humans for four oral-specific bacterial species. The Neanderthal and upper Paleolithic modern human individuals show expected damage patterns consistent with authentic aDNA, whereas the present-day individual does not. See also *SI Appendix*, Fig. S4.

cytosine to thymine deamination (Fig. 1*C* and *SI Appendix*, Fig. S4) (30). Finally, to reduce potentially spurious assignments for compositional analysis, we removed low-abundance taxa using thresholds optimized at different taxonomic levels (*SI Appendix*, Figs. S7 and S8 and sections S3.6 and S5.2). The resulting 89 well-preserved dental calculus datasets consist of samples ranging from the present day up to 100 ka.

The Core African Hominid Oral Microbiome. We performed PCoA on our dataset of well-preserved samples and found considerable overlap in the microbial composition of African hominid dental calculus, as well as howler monkeys (Fig. 1*B*), suggesting the existence of a core microbiome that has been maintained for more than 8 My, based on fossil and molecular evidence of host divergence among African hominids (31, 32), and possibly since before the catarrhine–platyrrhine split ca. 40 Mya (33, 34). At the same time, small but significant differences were indicated by Permutational Multivariate Analysis of Variance (PERMANOVA) (35) at both the microbial genus and species levels between each hominid genus (100 bootstrap replicates, $\alpha = 0.05$; genus: $F = 5.22 \pm 1.42$, $df = 3$, $R^2 = 0.27 \pm 0.05$, $P = 0.001$; species: $F = 6.67 \pm 2.52$, $df = 3$, $R^2 = 0.32 \pm 0.07$, $P = 0.001$; *SI Appendix*, Fig. S5), and this pattern remained robust after controlling for unequal sample sizes (*SI Appendix*, section S4.2).

Dental plaque biofilms in humans form by the microbial succession of early, bridging, and late colonizers (36), and in contrast to the gut, which has high interindividual variability at the microbial phylum level (37) and is sensitive to subsistence changes over short and long timescales (15, 38, 39), oral microbial communities have been found to be more stable and consistent, particularly at the genus level (40–42), and even when challenged by antibiotics (43). Because of this, we sought to begin to define the African hominid core oral microbiome as a group and for each genus separately. For a microbial taxon to be considered "core" (44, 45), we required it to be present in at least two-thirds of the populations making up a given host genus, counting as present only those populations in which it is found in at least half of individuals to account for variation in preservation (*SI Appendix*, Fig. S9*A* and section S5.2). We then calculated the intersection of each core microbial genus (Fig. 2*A*) and species (Fig. 2*B*) across all host taxa (Dataset S3). Most "core" taxa are shared across all three African hominid genera (*Gorilla*, *Pan*, and *Homo*) and howler monkeys, whereas fewer are "core" only to African hominids (*Gorilla*, *Pan*, and *Homo*), *Pan* and *Homo*, or *Homo* (Fig. 2 and *SI Appendix*, Fig. S1). Despite smaller sample sizes than studies of present-day microbiomes, bootstrapping analysis to assess consistency of calculations supported most core microbiome assignments, with lower values possibly indicating taxa influenced by factors such as biofilm maturity (*SI Appendix*, section S5.3). This suggests a high degree of genus-level microbial taxonomic conservation during African hominid, and possibly broader primate, host evolution and speciation.

Core taxa at both the genus and species levels include well-known members of each stage of plaque biofilm formation (8, 36), including the early colonizers *Streptococcus* and *Actinomyces*, the bridging taxa *Fusobacterium* and *Corynebacterium*, and the late colonizers *Porphyromonas* and *Treponema*, although the latter two are "core" to only chimpanzees and *Homo* (Fig. 2*C*). Major periopathogens, bacteria associated with periodontal disease, are found among the different host core combinations, and, focusing on *Porphyromonas* and *Tannerella* specifically because of their clinical significance today, we find that their major virulence factors are shared across multiple primates and thus are not specific to modern humans (*SI Appendix*, Fig. S9 *B* and *C*). The presence of periopathogens within the core microbiome supports the hypothesis that they are not pathogens in a conventional sense but rather that their pathogenic character in present-day humans may be related to an imbalance between the biofilm and the host, as has been suggested by recent ecological studies (46). Although some of the African hominid "core" taxa are periopathogens or their close relatives, most core members are known today to play important structural and functional roles in the formation and maturation of plaque, implying deep coevolutionary relationships between these taxa and their hosts.

African Hominid Oral Microbiome Structure Shows a Weak Relationship with Host Phylogeny. Hierarchical clustering shows that calculus metagenomes tend to cluster by host genus, confirming intragroup similarity, but these relationships exhibit differences from host phylogeny (Fig. 3). We find, for example, that howler monkeys and gorillas fall together in a single clade and a subset of *Homo* clusters with chimpanzees. With respect to the latter, available metadata do not provide any clear associations with factors such as geography, time period, or disease to explain this pattern (*SI Appendix*, section S4.3). Overall, gorillas and howler monkeys are characterized by a wide diversity of aerobic and facultatively anaerobic taxa, while chimpanzees have higher levels of obligately anaerobic taxa, including many putative periopathogens (e.g., *Porphyromonas gingivalis*, *Treponema denticola*, *Tannerella forsythia*, *Filifactor alocis*, and *Fretibacterium fastidosum*). Neanderthals consistently fall within the diversity of modern humans. *Homo* is notable for its high abundance of *Streptococcus* spp., while this genus is found at substantially lower levels in *Pan*.

Many of the taxa identified in human and nonhuman primate dental calculus are poorly characterized, making further exploration difficult. Indeed, several species within the human core genera remain unnamed (*Ottowia* sp. oral taxon 894, *Olsenella* sp. oral taxon 807) or understudied (*Pseudopropionibacterium propionicum*, *F. fastidiosum*) and some even lack genus designations ([*Eubacterium*] *minutum*, TM7x, Anaerolinaceae bacterium oral taxon 439). Their absence from most discussions of the modern human oral microbiome points to a major gap in current oral microbiology research, and targeted investigation of these species is needed to identify their functional and structural roles within plaque biofilms (47–49). Host genus patterns in community structure may be influenced by differences in salivary flow or composition (50), as well as differences in diet texture, quality, and nutrient content (51) (*SI Appendix*, sections S1, S5.1, and S5.6). We also investigated microbial community structure within modern humans, but in contrast to previous studies (13), we found no difference among broad dietary patterns or time periods (*SI Appendix*, Fig. S6 and section S4.5). These findings accord with the results of modern oral microbiome studies, which also show minimal, if any, broad and sustained compositional changes in response to diet (e.g., refs. 41, 52, and 53). The relative stability of the oral microbiome may be due in part to the extensive community interdependencies (54, 55) that have developed within these biofilms to metabolize complex host salivary glycoproteins, which are the major nutrient source for most members of the oral microbiota (56). This is in contrast to studies demonstrating strong associations between diet and taxonomic/functional composition in modern gut microbiomes (57, 58).

Evolutionary Histories of Oral Microbial Species Reflect *Homo* Interactions. We next examined the phylogenies of individual microbial taxa to determine if host evolutionary relationships are reflected at the microbial genome level. To improve genome coverage and reduce potential noise from DNA damage, we selected a representative subset of well-preserved calculus samples across all host genera ($n = 19$; *SI Appendix*, Table S1 and Dataset S1) and constructed uracil-DNA glycosylase-treated (UDG) libraries to remove deaminated cytosines (59), which we then deeply sequenced and analyzed together with a subset of four of the present-day modern humans. Genome-level sequence reconstruction from diversity-rich ancient microbiomes is challenging due to both the highly fragmented nature of aDNA and the low relative abundance of each species, which makes strain separation difficult (*SI Appendix*, section S6). Furthermore, a

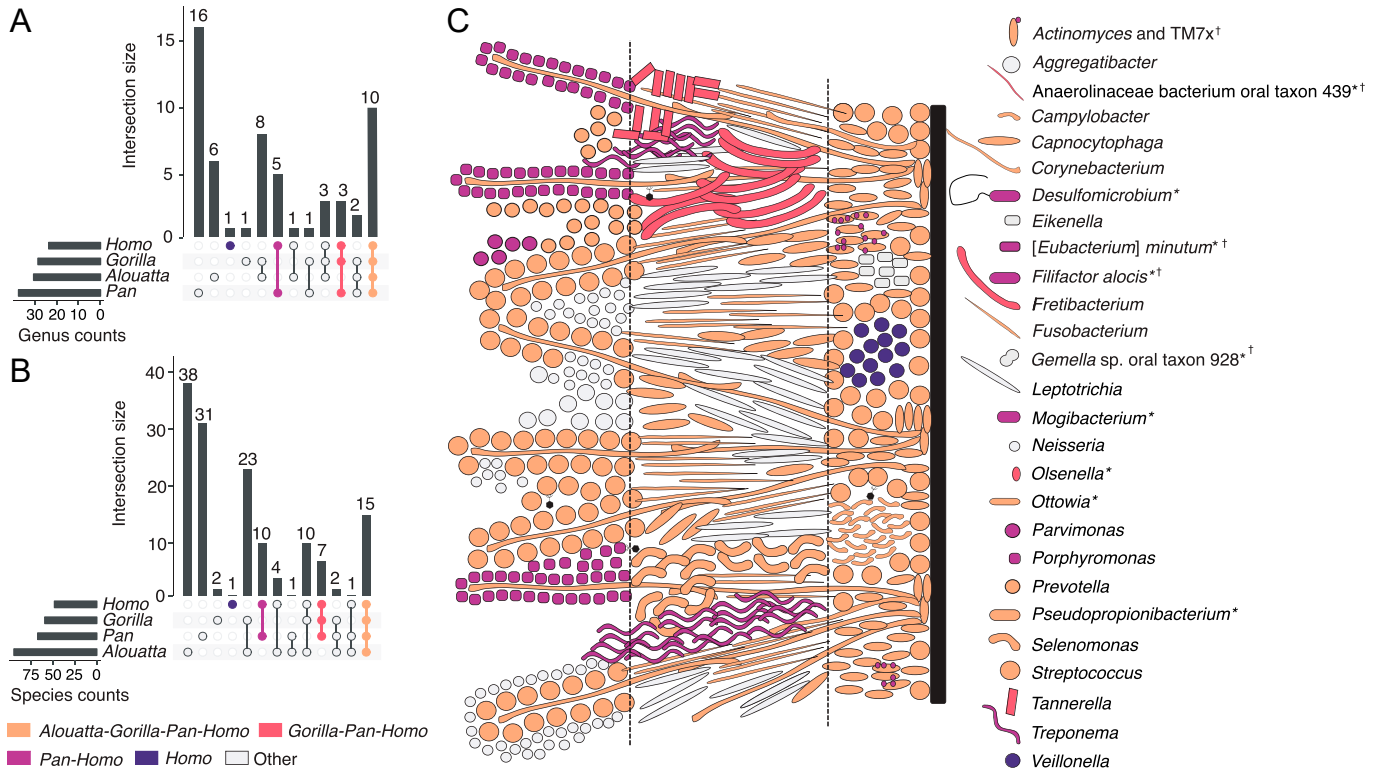

Fig. 2. Core oral microbiome of African hominids shows a deep evolutionary conservation of biofilm structure. UpSet plots showing the number of microbial genera (*A*) and species (*B*) core to host groups and group combinations. (*C*) Core taxa of the human oral microbiome (inclusive of all African hominid and howler monkey ranks). Human biofilm spatial organization based on refs. 8 and 100. Taxa are colored by the broadest host group for which they are core. "Other" taxa are those that fall into paraphyletic host groupings (e.g., *Alouatta*:*Homo*). Dashed lines separate the biofilm into basal, intermediate, and peripheral regions (100). Taxa with unknown spatial location are marked with an asterisk (*); taxa core to *Homo* with any combination of other host genera at the species level but not at the genus level are marked with a dagger (†). Reference Dataset S3 for additional information.

lack of sufficient reference genomes for many commensal oral microbes increases mismapping and noise artifacts when identifying single-nucleotide polymorphisms (SNPs) (*SI Appendix*, Fig. S10). Nevertheless, despite these challenges and using representative genomes from core taxa, we were able to reconstruct phylogenetic trees with high bootstrap support on internal nodes for eight oral bacteria (Fig. 4 and *SI Appendix*, Fig. S11).

As with compositional analysis, reconstructed genome-level sequences tend to cluster with those from the same host genus but do not closely reflect host phylogeny (Fig. 4 and *SI Appendix*, Figs. S11 and S12). Overall, genome-level sequences reconstructed from gorillas and chimpanzees fall closer to each other than do those of chimpanzees and *Homo*. Biases from the use of modern human-derived microbial reference genomes may in part contribute to this pattern, but microbial exchange due to overlapping territorial ranges of gorillas and chimpanzees throughout their evolution may also be a contributing factor. Within *Homo*, Neanderthals consistently group together, indicating shared within-species microbial diversity. However, we also note that the Upper Paleolithic individual from El Mirón in Iberia (18.6 ka) clusters in all trees with Neanderthals, rather than with other Pleistocene hunter-gatherers of the African Later Stone Age or more recent Holocene-era European or African populations. Recently published human genomic data including this individual has revealed that its associated genetic ancestry component was largely displaced across Europe after 14 ka (24, 60) during postglacial warming. Turning to our low-coverage metagenomic datasets, we assessed additional European Upper Paleolithic and Mesolithic groups (*SI Appendix*, section S6.6) and found that they show a similar pattern (albeit at lower resolution), with the oral taxa of individuals dated to before 14 ka mostly falling with Neanderthals and those after 14 ka mostly clustering with present-day modern humans (24, 60). This pattern suggests that the reconstructed oral bacterial genomes from El Mirón reflect a standing microbial diversity in *Homo* that was present in Europe during the Middle and Upper Paleolithic, but which was later replaced following subsequent migrations of modern human populations from elsewhere. Because oral microbiota are primarily inherited through caregivers (61, 62), additional sampling and ultradeep sequencing of Paleolithic European and Asian dental calculus may prove informative about the poorly understood interaction dynamics between archaic and modern humans.

***Homo*-Specific Shifts in Oral Biofilm Are Linked to Dietary Starch Availability.** The metabolic potential of a microbial community, which is inferred from its total gene content, can offer insights into biofilm ecology and function that cannot be understood from taxonomy alone. To better characterize the metabolic and functional differences among hominid oral biofilms, we compared the gene content of dental calculus metagenomes from well-preserved samples of the larger sequencing dataset using two different methods of functional classification, HUMAnN2 (63) and AADDER (64), and found moderate concordance in overall results. Principal Components Analysis (PCA) of the protein-level functional assignments cluster host genera distinctly with a high degree of separation between hosts and functional content (Fig. 5*A* and *SI Appendix*, Figs. S13 and S14), whereas we observe only a moderate degree of separation in the taxonomic PCoA (Fig. 1*B*), suggesting that gene content of the taxa shared by hominids is more host-specific than taxonomic assignments, a pattern that has also been seen for other microbial systems (65). The genes that drive separation of *Homo* from nonhuman primates consistently

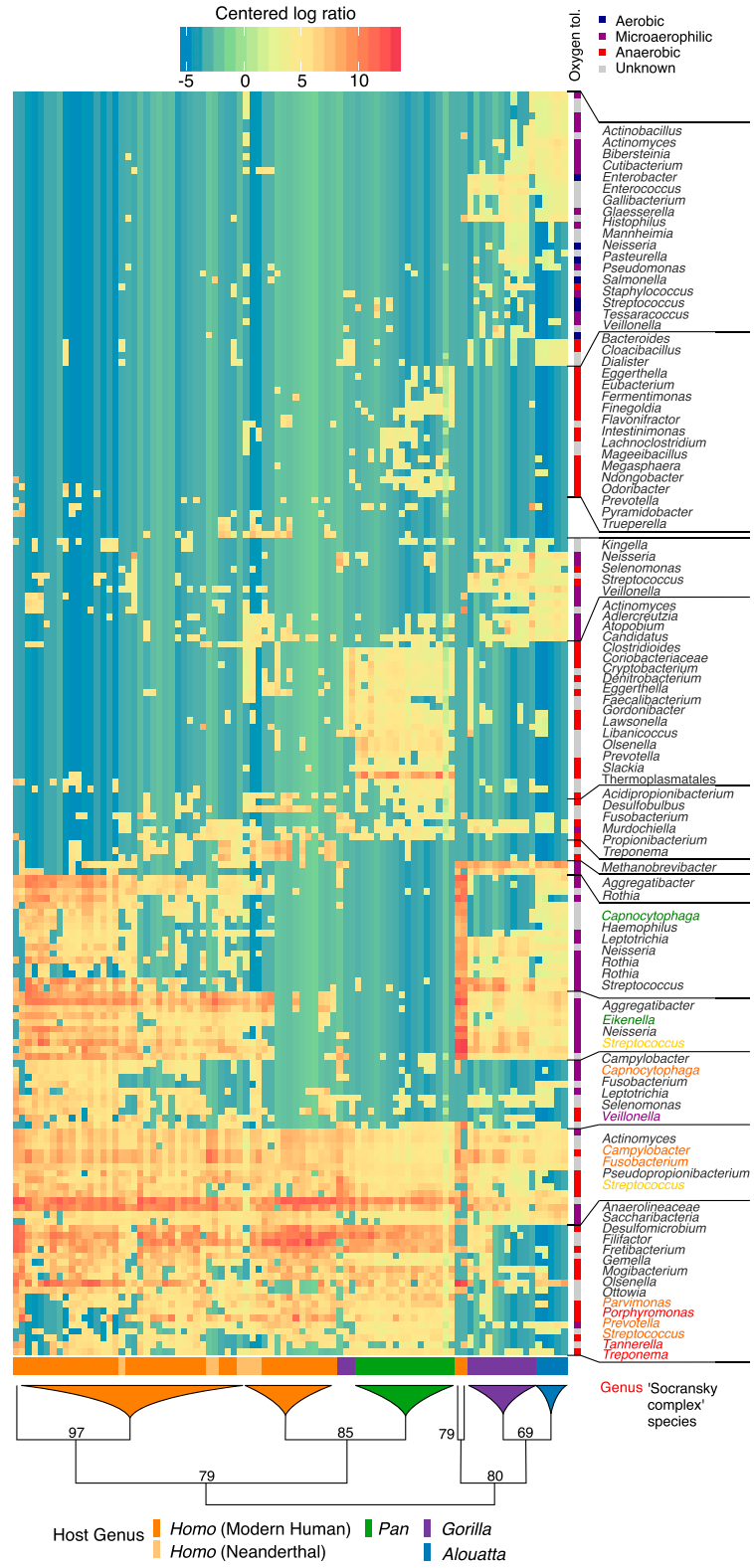

Fig. 3. African hominid dental calculus microbiomes cluster by host genus and other factors. Hierarchical clustering of howler monkeys, chimpanzees, gorillas, Neanderthals, and ancient and present-day modern humans based on species-level prokaryotic taxonomic assignments. Bacterial oxygen tolerance is associated with biofilm maturation stage in modern humans, and colored names indicate species corresponding to Socransky complexes (111) (reference *SI Appendix* section S5.1.1 for a summary). Microaerophilic is defined based on the BacDive database and is roughly synonymous to facultative anaerobe. The tree is schematic, and bifurcations are shown until all host genera are represented. Microbial species names are collapsed to genus level. Species and sample names can be located in *SI Appendix*, section S4.3.

relate to carbohydrate processing (*SI Appendix*, Fig. S14), are much more abundant in *Homo*, and largely derive from *Streptococcus* (*SI Appendix*, Fig. S13), something also observed in primate gut microbiomes (66). We therefore investigated the distribution of *Streptococcus* across our samples (Fig. 5*B*) using a classification system based on biochemical characteristics and genetic relatedness

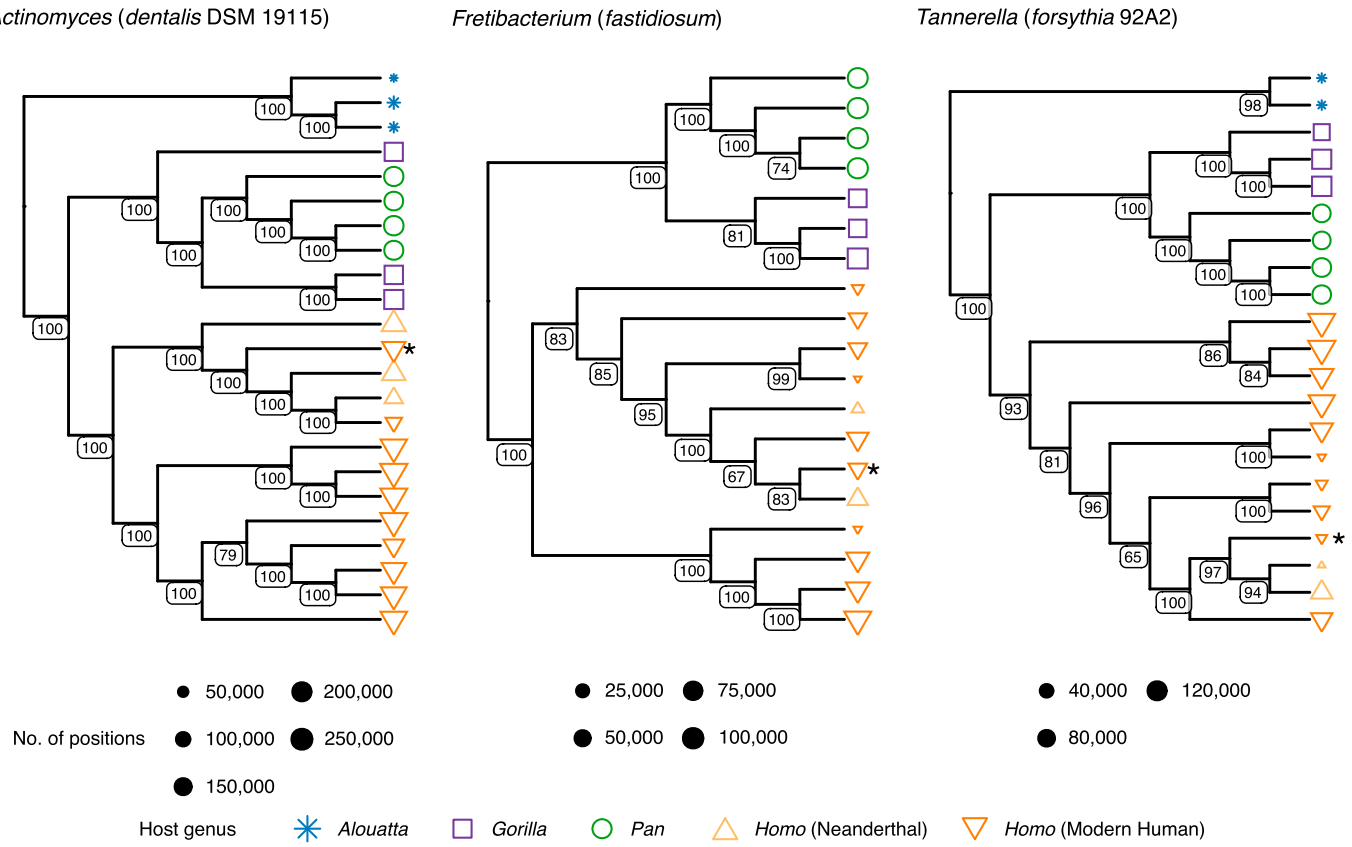

Fig. 4. African hominid oral taxa cluster phylogenetically by host genus. Selected neighbor-joining SNP-based phylogenetic cladograms of representative core oral microbiome genomes from deep-sequenced calculus metagenomes (*SI Appendix*, section S6.6). *Actinomyces* and *Tannerella* trees are rooted on the branch leading to howler monkeys (*Alouatta*, blue), *Fretibacterium* tree is midpoint rooted. Positions refer to non-N nucleotide calls in the alignment. Node values represent node support out of 100 bootstrap replicates. Asterisk (*) represents the Upper Paleolithic individual from El Mirón (EMN001), which consistently falls near Neanderthal individuals. The remaining eight trees, with tip labels, are provided in *SI Appendix*, Fig. S11.

(67). We find that *Streptococcus* species belonging to the Mitis, Sanguinis, and Salivarius groups are dominant in *Homo*, while these same groups are effectively absent in chimpanzees, and nonhuman primates in general are characterized by much higher proportions of *Streptococcus* species in the Anginosus, Mutans, and Pyogenic groups (Fig. 5*B*).

The Mitis, Sanguinis, and Salivarius groups are notable for their ability to express amylase-binding proteins to capture salivary α-amylase (68, 69), which they use for their own nutrient acquisition from dietary starch, as well as dental adhesion (70, 71). Amylase-binding protein genes (e.g., *abpA* and *abpB*) share no homology but rather confer a similar phenotype through convergent evolution, and they are found almost exclusively in oral *Streptococcus* species (68). Alpha-amylase is the most abundant enzyme in modern human saliva and modern humans express it at higher levels than any other hominid (50, 72). In contrast to most other nonhuman primates, modern humans exhibit high salivary α-amylase (*AMY1*) copy number variation, with a reported range of up to 30 diploid copies (16, 73, 74). This copy number expansion is estimated to have occurred along the modern human lineage after the divergence from Neanderthals in the Middle Pleistocene (75, 76). It has been argued this increase relates to dietary shifts during the evolutionary history of modern humans and specifically to an increased reliance on starch-rich foods (20, 21).

We next calculated the ratio of reads aligning to *abpA* and *abpB* sequences compared to all *Streptococcus* reads in the deep-sequenced dataset. We find that *abpA* and *abpB* reads are nearly absent in the nonhuman groups but are prevalent and significantly more abundant in *Homo* (Mann–Whitney *U* test *Homo* versus non-*Homo*: *abpB*, $\alpha = 0.05$, $U = 128$, $P = < 0.001$, 95% CI = 0.686 to 0.851; *abpA*, $\alpha = 0.05$, $U = 112$, $P = < 0.001$, 95% CI = 0.398 to 0.861). In particular, *abpB* is present in all deeply sequenced *Homo* individuals, and *abpA* is especially prevalent in modern humans (Fig. 5*C*). This suggests that oral streptococci evolved in association with changes in host diet and supports an early importance of starch-rich foods in *Homo* evolution.

Discussion

Commensal microbes of the oral microbiome represent an underutilized and independent source of information about host evolutionary and ecological differences (15, 77). With generation times orders-of-magnitude shorter than their hosts and the ability to acquire new functions through horizontal gene transfer across distantly related groups, microbes are a particularly dynamic and temporally resolved system for understanding human evolution. After applying a rigorous strategy to identify, decontaminate, and authenticate well-preserved dental calculus specimens up to 100 ka, we identify a core group of 10 bacterial genera within the African hominid primate oral microbiome that are also shared with howler monkeys, suggesting that these microbial groups have played a key role in oral biofilms since before the catarrhine–platyrrhine split ca. 40 Mya (33, 34). Today, these core taxa are primarily involved in providing structural support within the dental plaque biofilm, and their study holds promise for understanding biofilm growth and maturation in the ancestral human microbiome (78, 79). Identifying the role of such taxa is critical for the successful long-term treatment, prevention, and control of dysbiotic biofilms, such as those found in dental and periodontal diseases (80).

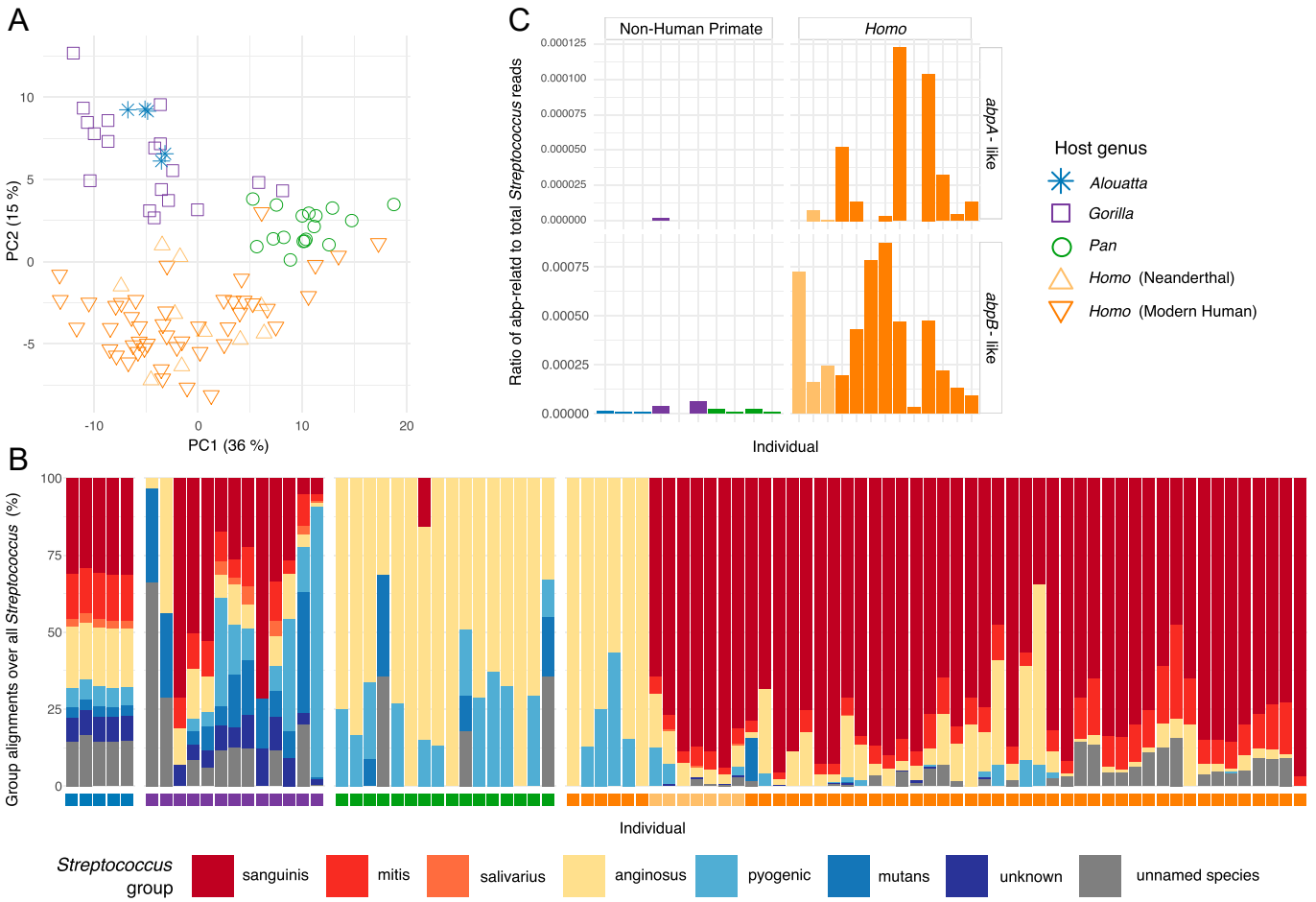

Fig. 5. Metabolic function and *Streptococcus* amylase-binding gene content is distinct between African hominid oral microbiomes. (*A*) PCA of microbial gene functions (SEED classification) clusters well-preserved samples by host genus (PERMANOVA $R^2 = 0.345$). *Homo* is functionally distinct from nonhuman African hominids and howler monkeys, particularly with respect to carbohydrate metabolism (*SI Appendix*, Fig. S14). (*B*) Bar plot of proportion of alignments to different *Streptococcus* groups show differences between host genera. Color of squares below bars corresponds to legend in *C*. Amylase-binding activity has been observed among members of the Sanguinis, Mitis, and Salivarius groups (68). (*C*) Ratios of reads aligning to amylase-binding-protein annotated sequences versus a genus-wide *Streptococcus* "superreference" show higher values in *Homo* than nonhuman primates, based on a deep-sequenced subset of samples and four present-day modern humans. Note the ratio on the *y*-axes of *abpA* and *abpB* are scaled differently.

Further, we identify 27 genus-level members of the *Homo* core oral microbiome, and these include many well-known and clinically relevant taxa, such as *Streptococcus* and the periopathogens *P. gingivalis*, *T. forsythia*, and *T. denticola*; however, nearly all of these are also core microbiome members of other African hominids. Only *Veillonella parvula*, a commensal species known to have a synergistic relationship with the cariopathogen *Streptococcus mutans* (81), is primarily found in humans. Surprisingly, not all members of the core *Homo* oral microbiome are well-known—three have no genus designation and several lack species names, revealing a major gap in oral microbiology research that in part relates to the difficulties in growing and propagating these microbes.

Focusing on oral microbiome evolution within *Homo*, we reconstruct authentic oral metagenomes of Neanderthals dating up to 100 ka and modern humans dating up to 30 ka, finding a high degree of similarity in microbial community structure, while also documenting indications of strain-level differences within core taxa. Interestingly, we find that Neanderthal-associated strain-level sequence variants are consistently present in Upper Paleolithic Europeans but not afterward, which accords with a described modern human genomic turnover around 14 ka (24, 60). Comparing human and nonhuman primates, we show that within *Streptococcus*, amylase-binding groups play a central role in the oral biofilms of *Homo*, likely aided by both their enhanced ability to colonize the dentition and their exclusive access to dietary starches. These *Streptococcus* groups and *abpB* are a general feature of *Homo*, suggesting that starch-rich foods, possibly modified by cooking (20) (*SI Appendix*, section S5.8), first became important early in *Homo* evolution prior to the split between Neanderthal and modern human lineages more than 600 ka (82, 83), a finding with potential implications for the energetics of *Homo*-associated encephalization (19–21, 26). Subsequent copy number expansion of *AMY1* in the modern human genome and the rise of *abpA* in oral streptococci may signal an even greater reliance on starch-rich foods by modern humans.

Further research on the evolution of abpA, abpB, and other amylase-binding proteins, including phylogenetic reconstruction and demographic modeling, promises to refine questions regarding biofilm formation and the nature and timing of dietary change in *Homo*. In addition, future research on non-African hominids (orangutans) and additional catarrhines, in particular cercopithecenes with high or unusual salivary amylase expression, such as gelada and hamadryas baboons (73, 84, 85), may yield further insights into the diverse evolutionary trajectories of primate oral microbiomes in response to habitat and dietary change. In addition, it is clear that more research on core genera is urgently needed, as many of the highly conserved and potentially key

structural taxa in hominid oral biofilms are understudied and even lack formal names. Furthermore, future sequencing projects focusing on within-species genomic diversity will be critical to understanding microbiome evolution and coadaptation within the human lineage. This study demonstrates that integrating evolutionary studies of the modern human microbiome with wild primate and ancient *Homo* metagenomic data provides valuable insights into the ancestral states of the human oral microbiome, the nature of microbial–host relationships, and major events in the evolution of modern humans and Neanderthals.

Materials and Methods

Materials. Our sampling strategy aimed to collect dental calculus from a minimum of two independent populations, each consisting of at least five individuals, for each host genus and modern human lifestyle group (excepting *Alouatta*) (*SI Appendix*, Table S1 and Dataset S1). Dental calculus was sampled from twentieth-century skeletal remains of wild *Alouatta* (*A. palliata*), *Gorilla* (*G. berengei beringei*; *G. berengei graueri*; *G. gorilla gorilla*), and *Pan* (*P. troglodytes schweinfurthii*; *P. troglodytes ellioti*; *P. troglodytes verus*) and from archaeological Neanderthals and modern humans using established protocols (DOIs: 10.17504/protocols.io.7vrhn56 and 10.17504/protocols.io.7hphj5n). Although many present-day human dental plaque datasets are publicly available, they have been shown to not be directly comparable to dental calculus (23), and consequently we generated dental calculus data for present-day humans. The study of deidentified present-day dental calculus was approved by the Institutional Review Board for Human Research Participant Protection at the University of Oklahoma (IRB no. 4543). All samples were collected under informed consent during routine dental cleaning procedures by practicing dental odontologists. For additional sample context descriptions and additional ethical approval information, reference *SI Appendix*, section S2.1.

Laboratory Methods. For all museum and field station samples, we performed DNA extraction in dedicated cleanroom facilities using a protocol optimized for the recovery of degraded and fragmentary DNA (86). Present-day calculus was extracted as previously described (23). For all samples, DNA was built into dual-indexed Illumina libraries (87) and shotgun sequenced. In addition, a subset of samples were separately subjected to UDG treatment (88), followed by deep sequencing. Negative controls were included in all extraction and library construction batches. Sequencing was performed on either Illumina NextSeq. 500 or HiSeq. 4000 platforms. For details, reference *SI Appendix*, section S2.2–S2.4 and protocols.io under DOI: 10.17504/protocols.io.bq7wmzpe.

Data Processing and Quality Filtering. For detailed descriptions of preprocessing and analysis procedures, including code, reference *SI Appendix* and external data repository (GitHub repository: https://github.com/jfy133/Hominid_Calculus_Microbiome_Evolution; Archive DOI: 10.5281/zenodo.3740493). Additional ancient (13) and present-day dental calculus (23) data from previous studies were downloaded from the Online Ancient Genome Repository (OAGR) (https://www.oagr.org.au/) and the European Bioinformatics Institute (EBI) European Nucleotide Archive (ENA) (https://www.ebi.ac.uk/ena/) databases, respectively. Comparative metagenomes from present-day modern human microbiome and environmental sources were additionally downloaded from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) database (https://www.ncbi.nlm.nih.gov/sra/). Accession numbers and download instructions for all FASTQ files are provided in *SI Appendix*, section S3.1. The EAGER pipeline (89) was used to perform initial preprocessing of sequencing data to remove possible modern human DNA sequences that can interfere with taxonomic profiling (due to present-day modern human DNA contamination in microbial reference genomes). We used relaxed bwa aln (90) mapping parameters for aDNA (–n 0.01), and nonhuman reads from replicate samples and libraries were then concatenated per individual. Human-mapped sequences were then poly-G clipped prior to reporting of mapping statistics. Processing statistics are provided in *SI Appendix*, section S3.2.

Taxonomic Binning and Preservation Assessment. For taxonomic binning, we used the aDNA-optimized high-throughput aligner MALT (27, 91) together with the NCBI nt database (October 2017; uploaded to Zenodo under DOI: 10.5281/zenodo.4382154) and a custom NCBI RefSeq database (containing bacteria, archaea, and *Homo sapiens*, October 2018, *SI Appendix*, section S3.3) and employed a relaxed percent identity parameter of 85% and a base tail cut off ("minimum support") of 0.01%. Resulting RMA6 files were loaded into MEGAN6 CE (64) and prokaryotic Operational Taxonomic Unit (OTU) tables were exported (Dataset S2). A comparison of the two databases is provided in *SI Appendix*, section S3.3. Given the challenges of low preservation and contamination in ancient microbiome studies, we performed a multistep procedure to screen for and remove poorly preserved samples and contaminant OTUs from the non-UDG-treated dataset (*SI Appendix*, Fig. S2). We developed a visualization for the identification of calculus samples with weak oral microbiome signatures (*SI Appendix*, Fig. S3). This procedure involves comparing identified taxa to their previously reported isolation source(s), ranking these taxa from most to least abundant, and tracking the cumulative percentage of oral taxa along this rank (termed here as "decay"). Samples with a low percentage of oral taxa after an initial "burn-in" based on stabilization of curve fluctuation were removed from downstream analysis. Reference *SI Appendix*, section S3.4 for details. We compared this method to results obtained using SourceTracker (28)—which was performed on 16S-mapped reads filtered from shotgun data using EAGER (with comparative present-day modern human and environmental metagenomes as sources), followed by closed-reference clustering using QIIME (92)—and found concordance between the two methods (*SI Appendix*, Fig. S3). We next used the R package decontam (29) to statistically detect putative laboratory and environmental contaminants (as present in negative controls and a set of archaeological bone samples—*SI Appendix*, section S3.6), which were then removed prior to downstream analysis. To authenticate the remaining OTUs, we utilized the output of MaltExtract (93) in the MaltExtract-Interactive Plotting App (MEx-IPA) tool (DOI: 10.5281/zenodo.3380011), which we developed for rapid visualization of characteristic aDNA patterns, such as cytosine to thymine deamination, short fragment lengths, and edit distance from reference (*SI Appendix*, Fig. S4 and section S3.5). After mapping with EAGER to well-known oral taxa, we also validated DNA damage patterns using DamageProfiler (Fig. 1*C*) (94).

Microbial Compositional Analysis. To remove low-abundance environmental contaminants or spurious hits, we selected a minimum abundance cutoff of 0.07% of alignments for genus-level and 0.04% of alignments for species-level identifications (*SI Appendix*, Figs. S7 and S8 and section 5.2). We normalized profiles through phylogenetic isometric-log-ratio transformation (95) of the abundance-filtered OTU tables and then performed PCoA on the resulting euclidean distances (*SI Appendix* Fig. S5 *C* and *D* and *SI Appendix*, section S4.1). To statistically verify host genus clusters, we used the adonis function from the R package vegan to perform PERMANOVA (35) analysis after controlling from unequal sample sizes (*SI Appendix*, section S4.2). After removal of poorly preserved samples, oral communities show distinct centroids for each host genus (bootstrapped PERMANOVA, $\alpha = 0.05$, $P = 0.001$, pseudo-$F = 5.23$, $R^2 = 0.28$); *Alouatta* was excluded due to small sample size. We performed bootstrapped hierarchical clustering (96) on the euclidean distances of centered log ratio–transformed OTU tables and visualized the results in the form of a heatmap (Fig. 3 and *SI Appendix* section S4.3). Sample and taxon clustering was performed with the McQuitty hierarchical clustering algorithm, and taxon blocks within the heatmap were selected by visual inspection. Bootstrap values of sample clusters were estimated through the R package pvclust (96). Species oxygen-tolerance metadata was obtained from the BacDive database (97) via the BacDiveR R package (DOI: 10.5281/zenodo.1308060). For validation of the observations made on the heatmaps, we also performed grouped indicator analysis (98) (*SI Appendix*, section S4.4). Clustering of human oral microbiomes by variables such as time, geography, and dietary subsistence was assessed using PCoA, PERMANOVA, and hierarchical clustering (*SI Appendix*, section S4).

Core Microbiome Analysis. Using the contaminant-filtered OTU tables of well-preserved samples, we converted all taxa above the minimum support threshold to a presence/absence profile. Taxa were required to be present in at least half (50%) of the members of a population for it to be considered core to the population and to be present in at least two-thirds (66%) of populations to be considered core to a host group (*SI Appendix*, Fig. S9; reference *SI Appendix*, section S5.2 for parameter experimentation details). We then generated UpSet plots (99) to visualize the microbial intersections of each host group at both the species and genus levels (Fig. 2 *A* and *B*), and we also compared the results between both databases. Further discussion on the exclusion of the common soil genus *Mycobacterium* from core genera is provided in *SI Appendix*, section S5.2. Validation of results through smaller sample sizes was carried out by bootstrapping analysis, which was performed by randomly subsampling (with replacement) individuals from each host genus and rerunning the core calculation procedure to 1,000 replicates (*SI Appendix*, section S5.3). We created a diagram of the core human oral microbiome (Fig. 2*C*) based on published fluorescence in situ hybridization (FISH) images of human dental plaque (8, 100). For species/genera that were not analyzed in these publications, literature searches were performed to find evidence of their localization within plaque based on immunohistochemistry, immunofluorescence, or FISH (*SI Appendix*, sections S5.1 and S5.4). All members of the human core microbiome are shown, including those also shared with other African hominids and howler monkeys. For further details, reference *SI Appendix*, section S5.3.

Genomic Analysis. We used EAGER to map (see below for more details) the deep-sequenced UDG-treated dataset and four samples from present-day individuals (*Alouatta*, 3; *Gorilla*, 3; *Pan*, 4; Neanderthal, 3; ancient modern human, 6; present-day modern human, 4; total: 23) against the reference genomes of *Tannerella forsythia* and *Porphyromonas gingivalis* (*SI Appendix*, section S5.5). We used bedtools (101) to calculate the breadth and depth coverage of a set of known virulence factors for these two taxa. To reduce the risk of spurious alignments (e.g., from cross mapping of conserved sequences), we filtered out genes that had a breadth of coverage less than 70% and/or that appeared to have strongly different coverage depths compared to the rest of the genome (reference *SI Appendix*, section S5.5 for more details). The resulting genes were visualized as a heatmap for comparison (*SI Appendix*, Fig. S9). We selected all species-level *Streptococcus* alignments from the shallow sequenced dataset minimum support filtered NCBI nt–MALT OTU tables and assigned them to one of eight species groups based on the literature (67) (reference *SI Appendix*, section S5.6 for group definitions). We then calculated the fraction of alignments for each species group over all taxonomic alignments for each sample (Fig. 5*B*). To further validate the results, we calculated a similar ratio but based on the mapping of the deep-sequenced dataset against a superreference of 166 *Streptococcus* genomes (see below). We identified *abpA*- and *abpB*-like gene coordinates from the superreference using panX (102), then extracted the number of reads mapping to these annotations and calculated the fraction of these reads over all *Streptococcus* superreference mapped reads. We then applied a Mann–Whitney *U* test to test the null hypothesis of no difference between the distributions of ratios of *Homo* and nonhuman primates, as well as compared these results to a distribution of *P* values of 100 randomly shuffled group assignments (reference *SI Appendix*, section S5.7 for more details). Reference sequences of *abpA* and *abpB* were extracted from *Streptococcus* genomes in RefSeq and indexed for mapping. All shallow sequencing dataset samples were mapped against all reference strains. For samples with a gene coverage of at least 40% at 1×, a consensus sequence was exported from the Integrative Genome Viewer (IGV) (103). An input file of the consensus sequences and references was generated in BEAUTi and used to run BEAST2 (104) for Bayesian skyline plot analysis. For details, reference *SI Appendix*, section S5.9.

Microbial Phylogenetics. We first attempted a competitive-mapping strategy against genus-wide superreferences of identified core taxa (reference *SI Appendix*, sections S6.1 and S6.2), but this approach yielded only limited results (*SI Appendix* Fig. S10 and section S6.3). We then instead performed phylogenetic reconstruction by mapping the same dataset to a single representative genome for each genus, considered as representing a population of related taxa. To account for challenges with low-coverage ancient data, we called SNPs using MultiVCFAnalyzer and required each SNP call to have a minimum of 2× coverage and a support of ≥70% of reads (*SI Appendix*, section S6.5). The resulting FASTA alignments were loaded into R. Samples with fewer than 1,000 SNPs were removed, and pairwise distances were calculated based on the JC69 model (105). A bootstrapped neighbor-joining algorithm from the R package ape (106) was applied to the distance matrices with 100 replicates (*SI Appendix*, section S6.6). Trees were visualized with ggtree (107). Finally, we retained trees where the basal internal nodes had bootstrap supports of ≥70% (*SI Appendix*, Fig. S11). The same procedure was then applied to the shallow sequencing dataset with the additional samples described above in the main text (*SI Appendix*, Fig. S12). To test whether pre-14 ka individuals clustered with Neanderthals due to reference bias, we calculated the median number of positions that were shared between EMN001 and Neanderthals to a histogram of median pairwise comparisons between all modern human individuals (*SI Appendix*, section S6.6).

Functional and Metabolic Pathway Analysis. We took two approaches to characterizing the functional profiles of the calculus metagenomes. First, we used HUMANn2 (63) [with MetaPhlAn2 (108) generated taxonomic profiles] to generate functional profiles based on the UniRef90 (109) and ChocoPhlAn (July 2018) (63) databases. Preservation was independently assessed for pathway abundance and Kyoto Encyclopedia of Genes and Genomes (KEGG) ortholog functional profiles, *SI Appendix*, section S7.1. We compared the functional profiles of well-preserved calculus between host groups using pathway abundance ($n = 94$) and gene families converted to KEGG orthologs ($n = 109$) using PCA (*SI Appendix*, Fig. S13). Orthologs with the strongest loadings were visualized with biplots (*SI Appendix*, Fig. S13 *A–C*), and the species from which these orthologs were derived were determined (*SI Appendix*, Fig. S13 *B–D*). The clustering of host genera in PCAs using only orthologs in specific pathways (carbohydrates, amino acids, lipids) was also explored (*SI Appendix*, Fig. S10 *A–C*). For details, reference *SI Appendix*, section S7.1.4. Second, we used AADDER (included within MEGAN6 CE) (64) to profile the number of alignments to annotations present in the custom RefSeq database as aligned by MALT (see above). We then used MEGAN6 to export SEED category (110) profiles. Preservation was independently assessed for SEED protein functional profiles, reference *SI Appendix*, section S7.2. We compared the functional profiles of well-preserved calculus ($n = 95$) between host groups using proteins but not higher-level pathways (*SI Appendix*, Fig. S13). The proteins with the strongest loadings were visualized using biplots (*SI Appendix*, Fig. S13 *E–G*), and the species from which these proteins were derived were determined (*SI Appendix*, Fig. S13 *F–H*). The clustering of host genera in PCAs using only proteins in specific pathways (carbohydrates, amino acids, lipids) was also explored (*SI Appendix*, Fig. S14 *D–F*). For details, reference *SI Appendix*, section S7.2.3.

Data Availability. All newly generated sequencing data have been deposited in the ENA repository (https://www.ebi.ac.uk/ena/browser/home) under project accession ID PRJEB34569. R notebooks, bioinformatic scripts, additional supporting figures, and intermediate analysis files are provided in an external data repository hosted on GitHub (http:github.com/jfy133/Hominid_Calculus_Microbiome_Evolution) and archived with Zenodo under DOI: 10.5281/zenodo.3740493.

ACKNOWLEDGMENTS. We thank Ethiopia's Authority for Research and Conservation of Cultural Heritage (ARCCH), the Museu de Prehistòria de València, Marta Negra at the Burgos Museum, Ottmar Kullmer at the Naturmuseum Senckenberg, Lyman Jellema at the Cleveland Museum of Natural History, Reneé Molina of the Maderas Rainforest Conservancy at the Ometepe Biological Field Station, Daniela C. Kalthoff at the Swedish Museum of Natural History, Emmanuel Gilissen at the Royal Museum for Central Africa, Tom Geerinckx and Patrick Semal at the Royal Belgian Institute of Natural Sciences, and Abdeljalil Bouzouggar at the Institut National des Sciences de l'Archéologie et du Patrimoine (INSAP) for archaeological and museum collection assistance. We thank the Uganda Wildlife Authority for giving permission to carry out research on chimpanzees from Kibale National Park. We thank Miriam Carbo Toran, Alicia Hernández Fuster, Juan Bautista Rodriguez Martinez, and Eros Chaves for present-day calculus sampling assistance. We thank Dr. Dominique Henry-Gambier (University of Bordeaux I) for her initial examination of the mandible of Rigney 1. We thank Dr. Dawn Mulhern and Jaelle Brealey for consultation on nonhuman primate samples. We thank Rachel Carmody for suggestions on early drafts. We thank Alex Hübner for advice on data analysis. We thank Zandra Fagernäs, Richard Hagan, Maria Spyrou, and Antje Wissgott for their additional laboratory assistance. Research at the De Nadale Cave is coordinated by the University of Ferrara within the framework of a project supported by the Ministry of Culture–Western Veneto Archaeological Superintendence, the Soprintendenza Archeologia, belle Arti e Paesaggio per le Provincie di Verona, Rovigo e Vicenza (SABAP), and the Zovencedo Municipality, financed by the H. Obermaier Society, local private companies (RAASM and Saf), and local sponsors. The Calleva Foundation supported the excavation and research during which sampling of the Taforalt calculus was carried out. This project was funded by grants from the US National Science Foundation (NSF) (BSC-1516633 to C.W. and C.M.L.; BSC-1027607 to K.W.A, M.C.C., and J.W.A.; SBR-0416125 to R.W.W.), the US National Institutes of Health (NIH) (2R01 GM089886 to C.M.L., C.W., and K.S.; R37DE016937 and R01DE024468 to F.E.D.), the European Research Council (ERC) (ERC-STG 677576 "HARVEST" to A.G.H.; ERC-CG 617627 "ADaPt" to J.T.S.), the Deutsche Forschungsgemeinschaft (DFG FOR 2237 to K.H.; EXC 2051-390713860 to C.W.), the National Research Foundation of South Africa (NRF 115257 and 12081 to V.E.G.), the Natural Sciences and Engineering Research Council of Canada (RGPIN-2017-04702 and RGPIN-2019-04113 to M.R.), Czech National Institutional Support (RVO 68081758 to S.S.), the Ministry of Culture and Information and the Ministry of Education, Science and Technological Development of the Republic of Serbia (177023 to D.M.), Junta de Castilla y León (BU028A09 to J.C.D.F-L.), the Swedish Research Council Formas (2016-00835 and 2019-00275 to K.G.), the University of South Florida, the University of Oklahoma, the Werner Siemens Foundation (Paleobiotechnology to C.W.), and the Max Planck Society. Any opinions, findings, and conclusions expressed in this study are those of the authors and do not necessarily reflect the views of the granting agencies.

[a]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, 07745 Jena, Germany; [b]Institute for Pre- and Protohistoric Archaeology and Archaeology of the Roman Provinces, Ludwig-Maximilians-University Münich, 80539 Münich, Germany; [c]Institute for Archaeological Sciences, Eberhard Karls University of Tübingen, 72070 Tübingen, Germany; [d]Department of Anthropology, University of Oklahoma, Norman, OK 73019; [e]Laboratories of Molecular Anthropology and Microbiome Research, University of Oklahoma, Norman, OK 73019; [f]Natural History Museum, University of Oslo, 0562 Oslo, Norway; [g]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287; [h]Department of Microbiology,

Immunology and Genetics, University of North Texas Health Science Center, Fort Worth, TX 76107; [i]Department of Anthropology, University of South Florida, St. Petersburg, FL 33701; [j]Palaeobiology, Biogeology, Department of Geosciences, Eberhard Karls University of Tübingen, 72074 Tübingen, Germany; [k]De la Préhistoire à l'Actuel: Culture, Environnement et Anthropologie (PACEA), CNRS UMR 5199, Université de Bordeaux, 33615 Pessac, France; [l]Laboratoire Chronoenvironnement, CNRS UMR 6249, 25030 Besançon, France; [m]Service Régional d'Archéologie de Bourgogne-Franche-Comté, Direction Régionale des Affaires Culturelles (DRAC) Bourgogne-Franche-Comté, 25043 Besançon, France; [n]Anthropology Program, California State University Channel Islands, Camarillo, CA 93012; [o]Centre for Palaeogenetics, 10691 Stockholm, Sweden; [p]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, 10405 Stockholm, Sweden; [q]Institut für Ur- und Frühgeschichte und Archäologie des Mittelalters, Eberhard Karls University of Tübingen, 72074 Tübingen, Germany; [r]Sonderforschungsbereiche 1070 Ressourcen Kulturen, Eberhard Karls University of Tübingen, 72074 Tübingen, Germany; [s]Prehistoria, Departamento de Historia, Geografía y Comunicación, Universidad de Burgos, 09001 Burgos, Spain; [t]Senckenberg Centre for Human Evolution and Palaeoenviroment, Eberhard Karls University of Tübingen, 72074 Tübingen, Germany; [u]Escribano Escrivá Clínica Dental, 38003 Santa Cruz de Tenerife, Spain; [v]Landesamt für Denkmalpflege im Regierungspräsidium Stuttgart, 78467 Konstanz, Germany; [w]Division of Clinical Anatomy and Biological Anthropology, Department of Human Biology, University of Cape Town, Cape Town 7925, South Africa; [x]Instituto Internacional de Investigaciones Prehistóricas de Cantabria, Universidad de Cantabria–Gobierno de Cantabria–Banco, 39071 Santander, Spain; [y]Clínica Dental Grande Mateu, 46004 València, Spain; [z]Paleoanthropology, Institute of Archaeological Sciences, Eberhard Karls University of Tübingen, 72070 Tübingen, Germany; [aa]Deutsche Forschungsgemeinschaft Centre for Advanced Studies "Words, Bones, Genes, Tools," Eberhard Karls University of Tübingen, 72070 Tübingen, Germany; [ab]Faculty of Archaeology, Leiden University, 2333CC Leiden, The Netherlands; [ac]Centre for Human Evolution Research, The Natural History Museum, London SW7 5BD, United Kingdom; [ad]Departamento de Prehistòria y Arqueología, Universidad Nacional de Educación, 28040 Madrid, Spain; [ae]Department of Archaeology, Faculty of Philosophy, University of Belgrade, 11000 Belgrade, Serbia; [af]Department of Humanities, University of Ferrara, 44121 Ferrara, Italy; [ag]Institute of Environmental Geology and Geoengineering, National Research Council, Milano, Lombardia, 20126, Italy; [ah]Clínica Alboraya 10, 46010 València, Spain; [ai]Department of Anthropology, University of Winnipeg, Winnipeg, MB R3T 3C7, Canada; [aj]Department of Anthropology, California State University, Northridge, CA 91330; [ak]Institute of Archaeology at Brno, Czech Academy of Sciences, 60200 Brno, Czech Republic; [al]Department of Anthropology, Western University, London, ON N6A 5C2, Canada; [am]Department of Archaeology, Max Planck Institute for the Science of Human History, 07745 Jena, Germany; [an]McDonald Institute for Archaeological Research, University of Cambridge, Cambridge CB2 3ER, United Kingdom; [ao]Department of Anthropology, University of New Mexico, Albuquerque, NM 87131; [ap]Department of Anthropology, Masaryk University, 61137 Brno, Czech Republic; [aq]Museum für Vor- und Frühgeschichte Berlin, Stiftung Preussischer Kulturbesitz, 10117 Berlin, Germany; [ar]Berliner Gesellschaft für Anthropologie, Ethnologie und Urgeschichte, 10117 Berlin, Germany; [as]Departamento de Zoología y Antropología Física, Universidad de Murcia, 30100 Murcia, Spain; [at]Department of Human Evolution, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany; [au]Department of Microbiology and Plant Biology, University of Oklahoma, Norman, OK 73019; [av]Animal Ecology, Department of Ecology and Genetics, Uppsala University, 75236 Uppsala, Sweden; [aw]Science for Life Laboratory, 75237 Uppsala, Sweden; [ax]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA 02138; [ay]Department of Microbiology, The Forsyth Institute, Cambridge, MA 02142; [az]Oral Medicine, Infection, and Immunity, Harvard School of Dental Medicine, Boston, MA 02115; [ba]Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, United Kingdom; [bb]Grupo de Investigación en Prehistòria IT-1223-19 (Universidad del País Vasco-Euskal Herriko Unibertsitatea), Ikerbasque, Basque Foundation for Science, 01006 Vitoria-Gasteiz, Spain; [bc]Departament de Prehistòria, Historia i Arqueología, Universitat de València, 46010 València, Spain; [bd]Department of Geological Sciences, University of Cape Town, Rondebosch 7701, South Africa; [be]Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, 04103 Leipzig, Germany; and [bf]Department of Anthropology, Harvard University, Cambridge, MA 02138

1. F. E. Dewhirst *et al*., The human oral microbiome. *J. Bacteriol.* **192**, 5002–5017 (2010).
2. N. J. Kassebaum *et al*., Global burden of untreated caries: A systematic review and metaregression. *J. Dent. Res.* **94**, 650–658 (2015).
3. P. I. Eke *et al*., Update on prevalence of periodontitis in adults in the United States: NHANES 2009 to 2012. *J. Periodontol.* **86**, 611–622 (2015).
4. F. A. Scannapieco, R. B. Bush, S. Paju, Associations between periodontal disease and risk for nosocomial bacterial pneumonia and chronic obstructive pulmonary disease. A systematic review. *Ann. Periodontol.* **8**, 54–69 (2003).
5. P. B. Lockhart *et al*.; American Heart Association Rheumatic Fever, Endocarditis, and Kawasaki Disease Committee of the Council on Cardiovascular Disease in the Young, Council on Epidemiology and Prevention, Council on Peripheral Vascular Disease, and Council on Clinical Cardiology, Periodontal disease and atherosclerotic vascular disease: Does the evidence support an independent association?: A scientific statement from the American heart association. *Circulation* **125**, 2520–2544 (2012).
6. The Human Microbiome Project Consortium, Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
7. J. C. Clemente *et al*., The microbiome of uncontacted Amerindians. *Sci. Adv.* **1**, e1500183 (2015).
8. J. L. Mark Welch, B. J. Rossetti, C. W. Rieken, F. E. Dewhirst, G. G. Borisy, Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad. Sci. U.S.A.* **113**, E791–E800 (2016).
9. A. Akcalı, N. P. Lang, Dental calculus: The calcified biofilm and its role in disease development. *Periodontol. 2000* **76**, 109–115 (2018).
10. C. Warinner *et al*., Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46**, 336–344 (2014).
11. C. de La Fuente, S. Flores, M. Moraga, DNA from human ancient bacteria: A novel source of genetic evidence from archaeological dental calculus. *Archaeometry* **55**, 767–778 (2013).
12. C. J. Adler *et al*., Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and industrial revolutions. *Nat. Genet.* **45**, 450–455 (2013).
13. L. S. Weyrich *et al*., Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* **544**, 357–361 (2017).
14. R. N. Carmody *et al*., Cooking shapes the structure and function of the gut microbiome. *Nat. Microbiol.* **4**, 2052–2063 (2019).
15. S. L. Schnorr, K. Sankaranarayanan, C. M. Lewis Jr, C. Warinner, Insights into human evolution from ancient and contemporary microbiome studies. *Curr. Opin. Genet. Dev.* **41**, 14–26 (2016).
16. A. C. Poole *et al*., Human salivary amylase gene copy number impacts oral and gut microbiomes. *Cell Host Microbe* **25**, 553–564.e7 (2019).
17. S. E. Council *et al*., Diversity and evolution of the primate skin microbiome. *Proc. Biol. Sci.* **283**, 20152586 (2016).
18. A. A. Ross, A. Rodrigues Hoffmann, J. D. Neufeld, The skin microbiome of vertebrates. *Microbiome* **7**, 79 (2019).
19. L. C. Aiello, P. Wheeler, The expensive-tissue hypothesis: The brain and the digestive system in human and primate evolution. *Curr. Anthropol.* **36**, 199–221 (1995).
20. R. N. Carmody, R. W. Wrangham, The energetic significance of cooking. *J. Hum. Evol.* **57**, 379–391 (2009).
21. K. Hardy, J. Brand-Miller, K. D. Brown, M. G. Thomas, L. Copeland, The importance of dietary carbohydrate in human evolution. *Q. Rev. Biol.* **90**, 251–268 (2015).
22. F. A. Villanea, J. G. Schraiber, Multiple episodes of interbreeding between Neanderthal and modern humans. *Nat. Ecol. Evol.* **3**, 39–44 (2019).
23. I. M. Velsko *et al*., Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome* **7**, 102 (2019).
24. Q. Fu *et al*., The genetic history of Ice Age Europe. *Nature* **534**, 200–205 (2016).
25. Q. Fu *et al*., An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
26. A. Navarrete, C. P. van Schaik, K. Isler, Energetics and the evolution of human brain size. *Nature* **480**, 91–93 (2011).
27. Å. J. Vågene *et al*., Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nat. Ecol. Evol.* **2**, 520–528 (2018).
28. D. Knights *et al*., Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 761–763 (2011).
29. N. M. Davis, D. M. Proctor, S. P. Holmes, D. A. Relman, B. J. Callahan, Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome* **6**, 226 (2018).
30. A. W. Briggs *et al*., Patterns of damage in genomic DNA sequences from a Neandertal. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14616–14621 (2007).
31. B. Wood, M. Grabowski, "Macroevolution in and around the hominin clade" in *Macroevolution: Explanation, Interpretation and Evidence*, E. Serrelli, N. Gontier, Eds. (Springer International Publishing, 2015), pp. 345–376.
32. J. Prado-Martinez *et al*., Great ape genetic diversity and population history. *Nature* **499**, 471–475 (2013).
33. M. Bond *et al*., Corrigendum: Eocene primates of South America and the African origins of New World monkeys. *Nature* **525**, 552 (2015).
34. C. G. Schrago, On the time scale of New World primate diversification. *Am. J. Phys. Anthropol.* **132**, 344–354 (2007).
35. M. J. Anderson, A new method for non-parametric multivariate analysis of variance: Non-parametric manova for ecology. *Austral Ecol.* **26**, 32–46 (2001).
36. P. E. Kolenbrander, R. J. Palmer Jr, S. Periasamy, N. S. Jakubovics, Oral multispecies biofilm development and the key role of cell-cell distance. *Nat. Rev. Microbiol.* **8**, 471–480 (2010).
37. P. J. Turnbaugh *et al*., A core gut microbiome in obese and lean twins. *Nature* **457**, 480–484 (2009).
38. L. A. David *et al*., Diet rapidly and reproducibly alters the human gut microbiome. *Nature* **505**, 559–563 (2014).
39. E. D. Sonnenburg *et al*., Diet-induced extinctions in the gut microbiota compound over generations. *Nature* **529**, 212–215 (2016).
40. D. R. Utter, J. L. Mark Welch, G. G. Borisy, Individuality, stability, and variability of the plaque microbiome. *Front. Microbiol.* **7**, 564 (2016).
41. I. Kato *et al*., Nutritional correlates of human oral microbiome. *J. Am. Coll. Nutr.* **36**, 88–98 (2017).
42. Y. Zhou *et al*., Biogeography of the ecosystems of the healthy human body. *Genome Biol.* **14**, R1 (2013).

43. E. Zaura *et al*., Same exposure but two radically different responses to antibiotics: Resilience of the salivary microbiome versus long-term microbial shifts in feces. *mBio* **6**, e01693-15 (2015).
44. A. Shade, J. Handelsman, Beyond the venn diagram: The hunt for a core microbiome. *Environ. Microbiol.* **14**, 4–12 (2012).
45. A. Risely, Applying the core microbiome to understand host-microbe systems. *J. Anim. Ecol.* **89**, 1549–1558 (2020).
46. G. Hajishengallis, R. J. Lamont, Beyond the red complex and into more complexity: The polymicrobial synergy and dysbiosis (PSD) model of periodontal disease etiology. *Mol. Oral Microbiol.* **27**, 409–419 (2012).
47. E. Pasolli *et al*., Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662.e20 (2019).
48. G. Aleti *et al*., Identification of the bacterial biosynthetic gene clusters of the oral microbiome illuminates the unexplored social language of bacteria during health and disease. *mBio* **10**, e00321-19 (2019).
49. A. Edlund *et al*., Metabolic fingerprints from the human oral microbiome reveal a vast knowledge gap of secreted small peptidic molecules. *mSystems* **2**, e00058-17 (2017).
50. S. Thamadilok *et al*., Human and nonhuman primate lineage-specific footprints in the salivary proteome. *Mol. Biol. Evol.* **37**, 395–405 (2019).
51. R. S. Scott, M. F. Teaford, P. S. Ungar, Dental microwear texture and anthropoid diets. *Am. J. Phys. Anthropol.* **147**, 551–579 (2012).
52. A. C. Anderson *et al*., Long-term fluctuation of oral biofilm microbiota following different dietary phases. *Appl. Environ. Microbiol.* **86**, e01421-20 (2020).
53. F. De Filippis *et al*., The same microbiota and a potentially discriminant metabolome in the saliva of omnivore, ovo-lacto-vegetarian and vegan individuals. *PLoS One* **9**, e112373 (2014).
54. J. L. Mark Welch, F. E. Dewhirst, G. G. Borisy, Biogeography of the oral microbiome: The site-specialist hypothesis. *Annu. Rev. Microbiol.* **73**, 335–358 (2019).
55. J. L. Mark Welch, S. T. Ramírez-Puebla, G. G. Borisy, Oral microbiome geography: Micron-scale habitat and niche. *Cell Host Microbe* **28**, 160–168 (2020).
56. P. D. Marsh, T. Do, D. Beighton, D. A. Devine, Influence of saliva on the oral microbiota. *Periodontol. 2000* **70**, 80–92 (2016).
57. S. L. Schnorr *et al*., Gut microbiome of the Hadza hunter-gatherers. *Nat. Commun.* **5**, 3654 (2014).
58. M. De Angelis *et al*., Diet influences the functions of the human intestinal microbiome. *Sci. Rep.* **10**, 4247 (2020).
59. A. W. Briggs *et al*., Removal of deaminated cytosines and detection of in vivo methylation in ancient DNA. *Nucleic Acids Res.* **38**, e87 (2010).
60. C. Posth *et al*., Pleistocene mitochondrial genomes suggest a single major dispersal of non-Africans and a late glacial population turnover in Europe. *Curr. Biol.* **26**, 827–833 (2016).
61. A. C. R. Tanner *et al*., Similarity of the oral microbiota of pre-school children with that of their caregivers in a population-based study. *Oral Microbiol. Immunol.* **17**, 379–387 (2002).
62. L. Shaw *et al*., The human salivary microbiome is shaped by shared environment rather than genetics: Evidence from a large family of closely related individuals. *mBio* **8**, e01237-17 (2017).
63. E. A. Franzosa *et al*., Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15**, 962–968 (2018).
64. D. H. Huson *et al*., MEGAN community edition–Interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* **12**, e1004957 (2016).
65. S. Louca *et al*., Function and functional redundancy in microbial systems. *Nat. Ecol. Evol.* **2**, 936–943 (2018).
66. K. R. Amato *et al*., Convergence of human and Old World monkey gut microbiomes demonstrates the importance of human ecology over phylogeny. *Genome Biol.* **20**, 201 (2019).
67. V. P. Richards *et al*., Phylogenomics and the dynamic genome evolution of the genus Streptococcus. *Genome Biol. Evol.* **6**, 741–753 (2014).
68. E. M. Haase *et al*., Comparative genomics and evolution of the amylase-binding proteins of oral streptococci. *BMC Microbiol.* **17**, 94 (2017).
69. A. E. Nikitkova, E. M. Haase, F. A. Scannapieco, Taking the starch out of oral biofilm formation: Molecular basis and functional significance of salivary α-amylase binding to oral streptococci. *Appl. Environ. Microbiol.* **79**, 416–423 (2013).
70. D. Deimling *et al*., Electron microscopic detection of salivary α-amylase in the pellicle formed in situ. *Eur. J. Oral Sci.* **112**, 503–509 (2004).
71. J. D. Rogers, R. J. Palmer Jr, P. E. Kolenbrander, F. A. Scannapieco, Role of Streptococcus gordonii amylase-binding protein A in adhesion to hydroxyapatite, starch metabolism, and biofilm formation. *Infect. Immun.* **69**, 7046–7056 (2001).
72. V. Behringer *et al*., Measurements of salivary alpha amylase and salivary cortisol in hominoid primates reveal within-species consistency and between-species differences. *PLoS One* **8**, e60773 (2013).
73. P. Pajic *et al*., Independent amylase gene copy number bursts correlate with dietary preferences in mammals. *eLife* **8**, e44628 (2019).
74. C. I. Fernández, A. S. Wiley, Rethinking the starch digestion hypothesis for AMY1 copy number variation in humans. *Am. J. Phys. Anthropol.* **163**, 645–657 (2017).
75. I. Lazaridis *et al*., Ancient human genomes suggest three ancestral populations for present-day Europeans. *Nature* **513**, 409–413 (2014).
76. C. E. Inchley *et al*., Selective sweep on human amylase genes postdates the split with Neanderthals. *Sci. Rep.* **6**, 37198 (2016).
77. C. Warinner, C. Speller, M. J. Collins, C. M. Lewis Jr, Ancient human microbiomes. *J. Hum. Evol.* **79**, 125–136 (2015).
78. C. Warinner, Dental calculus and the evolution of the human oral microbiome. *J. Calif. Dent. Assoc.* **44**, 411–420 (2016).
79. C. Warinner, C. M. Lewis, Microbiome and health in past and present human populations. *Am. Anthropol.* **117**, 740–741 (2015).
80. M. Kilian *et al*., The oral microbiome–An update for oral healthcare professionals. *Br. Dent. J.* **221**, 657–666 (2016).
81. S. Liu *et al*., Effect of Veillonella parvula on the physiological activity of Streptococcus mutans. *Arch. Oral Biol.* **109**, 104578 (2020).
82. K. Prüfer *et al*., A high-coverage Veandertal genome from Vindija Cave in Croatia. *Science* **358**, 655–658 (2017).
83. A. Gómez-Robles, Dental evolutionary rates and its implications for the Neanderthal-modern human divergence. *Sci. Adv.* **5**, eaaw1268 (2019).
84. M. Mau, K.-H. Südekum, A. Johann, A. Sliwa, T. M. Kaiser, Indication of higher salivary alpha-amylase expression in hamadryas baboons and geladas compared to chimpanzees and humans. *J. Med. Primatol.* **39**, 187–190 (2010).
85. M. Mau, K.-H. Südekum, A. Johann, A. Sliwa, T. M. Kaiser, Saliva of the graminivorous Theropithecus gelada lacks proline-rich proteins and tannin-binding capacity. *Am. J. Primatol.* **71**, 663–669 (2009).
86. J. Dabney *et al*., Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 15758–15763 (2013).
87. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010**, pdb.prot5448 (2010).
88. A. W. Briggs, P. Heyn, Preparation of next-generation sequencing libraries from damaged DNA. *Methods Mol. Biol.* **840**, 143–154 (2012).
89. A. Peltzer *et al*., EAGER: Efficient ancient genome reconstruction. *Genome Biol.* **17**, 60 (2016).
90. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
91. A. Herbig *et al*., MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman https://doi.org/10.1101/050559 (Accessed 28 April 2016).
92. J. G. Caporaso *et al*., QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7**, 335–336 (2010).
93. R. Hübler *et al*., HOPS: Automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biol.* **20**, 280 (2019).
94. J. Neukamm, A. Peltzer, K. Nieselt, DamageProfiler: Fast damage pattern calculation for ancient DNA. *Bioinformatics*, btab190, 10.1093/bioinformatics/btab190 (2021).
95. J. D. Silverman, A. D. Washburne, S. Mukherjee, L. A. David, A phylogenetic transform enhances analysis of compositional microbiota data. *eLife* **6**, e21887 (2017).
96. R. Suzuki, H. Shimodaira, Pvclust: An R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* **22**, 1540–1542 (2006).
97. L. C. Reimer *et al*., BacDive in 2019: bacterial phenotypic data for High-throughput biodiversity analysis. *Nucleic Acids Res.* **47**, D631–D636 (2019).
98. M. De Cáceres, P. Legendre, M. Moretti, Improving indicator species analysis by combining groups of sites. *Oikos* **119**, 1674–1684 (2010).
99. J. R. Conway, A. Lex, N. Gehlenborg, UpSetR: An R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
100. V. Zijnge *et al*., Oral biofilm architecture on natural teeth. *PLoS One* **5**, e9321 (2010).
101. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
102. W. Ding, F. Baumdicker, R. A. Neher, panX: Pan-genome analysis and exploration. *Nucleic Acids Res.* **46**, e5 (2018).
103. H. Thorvaldsdóttir *et al*., Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. *Brief. Bioinform.* **14**, 178–192 (2013).
104. R. Bouckaert *et al*., BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS Comput. Biol.* **15**, e100650 (2019).
105. T. H. Jukes, C. R. Cantor, "Evolution of protein molecules" in *Mammalian Protein Metabolism*, H. N. Munro, Ed. (Academic Pres, New York, USA, 1969), III, pp. 21–135.
106. E. Paradis, K. Schliep, Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
107. G. Yu, D. K. Smith, H. Zhu, Y. Guan, T. T.-Y. Lam, ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).
108. D. T. Truong *et al*., MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12**, 902–903 (2015).
109. B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu; UniProt Consortium, UniRef clusters: A comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics* **31**, 926–932 (2015).
110. R. Overbeek *et al*., The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.* **33**, 5691–5702 (2005).
111. S. S. Socransky, A. D. Haffajee, M. A. Cugini, C. Smith, R. L. Kent Jr, Microbial complexes in subgingival plaque. *J. Clin. Periodontol.* **25**, 134–144 (1998).

4 Manuscript B: Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir

4.1 Overview and contribution

Manuscript Nr.: B

Title of Manuscript Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir.

Authors Fellows Yates, J. A., Andrades Valtueña, A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-Lopez, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., Heintzman, P. D., Spyrou, M. A., Hübner, A., Gancz, A. S., Hider, J., Allshouse, A. F., Zaro, V., & Warinner, C.

Citation Fellows Yates, J. A., Andrades Valtueña, A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-Lopez, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., Heintzman, P. D., Spyrou, M. A., Hübner, A., Gancz, A. S., Hider, J., Allshouse, A. F., Zaro, V., & Warinner, C. (2021). Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Scientific Data*, 8(1), 31. https://doi.org/10.1038/s41597-021-00816-y

The candidate is
☑ First author, ☐ Co-first author, ☑ Corresponding author, ☐ Co-author.

Status Published

Proportion (in %) of authors in the publication (indicated from 20%)

Author	Concept	Data Analysis	Experiment	Manuscript Composition	Material Provision
Fellows Yates, J. A.	80	80	55	80	NA
Andrades Valtueña, A.	0	0	0	0	NA
Vågene, Å. J.	0	0	0	0	NA
Cribdon, B.	0	0	0	0	NA
Velsko, I. M.	0	0	0	0	NA
Borry, M.	0	0	0	0	NA
Bravo-Lopez, M. J.	0	0	0	0	NA
Fernandez-Guerra, A.	0	0	0	0	NA
Green, E. J.	0	0	0	0	NA
Ramachandran, S. L.	0	0	0	0	NA
Heintzman, P. D.	0	0	0	0	NA
Spyrou, M. A.	0	0	0	0	NA
Hübner, A.	0	0	0	0	NA
Gancz, A. S.	0	0	0	0	NA
Hider, J.	0	0	0	0	NA
Allshouse, A. F.	0	0	0	0	NA
Zaro, V.	0	0	0	0	NA
Warinner, C.	0	0	0	0	NA

4.2 Article

SCIENTIFIC DATA

OPEN

DATA DESCRIPTOR

Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir

James A. Fellows Yates[1,2]✉, Aida Andrades Valtueña[1], Åshild J. Vågene[3], Becky Cribdon[4], Irina M. Velsko[1], Maxime Borry[1], Miriam J. Bravo-Lopez[5], Antonio Fernandez-Guerra[6,7], Eleanor J. Green[8,9], Shreya L. Ramachandran[10], Peter D. Heintzman[11], Maria A. Spyrou[1], Alexander Hübner[1,12], Abigail S. Gancz[13], Jessica Hider[14,15], Aurora F. Allshouse[16,17], Valentina Zaro[18] & Christina Warinner[1,16]✉

Ancient DNA and RNA are valuable data sources for a wide range of disciplines. Within the field of ancient metagenomics, the number of published genetic datasets has risen dramatically in recent years, and tracking this data for reuse is particularly important for large-scale ecological and evolutionary studies of individual taxa and communities of both microbes and eukaryotes. AncientMetagenomeDir (archived at https://doi.org/10.5281/zenodo.3980833) is a collection of annotated metagenomic sample lists derived from published studies that provide basic, standardised metadata and accession numbers to allow rapid data retrieval from online repositories. These tables are community-curated and span multiple sub-disciplines to ensure adequate breadth and consensus in metadata definitions, as well as longevity of the database. Internal guidelines and automated checks facilitate compatibility with established sequence-read archives and term-ontologies, and ensure consistency and interoperability for future meta-analyses. This collection will also assist in standardising metadata reporting for future ancient metagenomic studies.

[1]Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, 07745, Jena, Germany. [2]Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, Ludwig-Maximilians-Universität München, München, 80539, Germany. [3]Section for Evolutionary Genomics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 1350, Denmark. [4]School of Life Sciences, University of Warwick, Coventry, CV4 7AL, United Kingdom. [5]International Laboratory for Human Genome Research, National Autonomous University of Mexico, Queretaro, 76230, Mexico. [6]Section for GeoGenetics, GLOBE Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, 1350, Denmark. [7]Microbial Genomics and Bioinformatics Research Group, Max Planck Institute for Marine Microbiology, Bremen, 28359, Germany. [8]BioArCh, Department of Archaeology, University of York, York, YO10 5DD, United Kingdom. [9]Department of Earth Sciences, Natural History Museum, London, SW7 5BD, United Kingdom. [10]Human Genetics, University of Chicago, Chicago, IL, 60637, USA. [11]The Arctic University Museum of Norway, UiT The Arctic University of Norway, Tromsø, 9037, Norway. [12]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, 04103, Germany. [13]Department of Anthropology, Pennsylvania State University, Pennsylvania, PA, 16802, USA. [14]Department of Anthropology, McMaster University, Hamilton, L8S4L9, Canada. [15]McMaster Ancient DNA Centre, McMaster University, Hamilton, L8S4L10, Canada. [16]Department of Anthropology, Harvard University, Cambridge, MA, 02138, USA. [17]Max Planck-Harvard Research Center for the Archaeoscience of the Ancient Mediterranean, Cambridge, MA, 02138, USA. [18]Department of Biology, Università degli Studi di Firenze, Florence, 50122, Italy. ✉e-mail: fellows@shh.mpg.de; warinner@shh.mpg.de

Background & Summary

A crucial, but sometimes overlooked, component of scientific reproducibility is the efficient retrieval of sample metadata. While the field of ancient DNA (aDNA) has been celebrated for its commitment to making sequencing data available through public archives[1], this data is not necessarily 'findable' (as defined in the FAIR principles[2]) - making the retrieval of relevant metadata time-consuming and complex. Metagenomic studies typically require large sample sizes, which are integrated with previously published datasets for comparative analyses. However, the current absence of standards in basic metadata reporting within ancient metagenomics can make data retrieval tedious and laborious, leading to analysis bottlenecks.

Ancient metagenomics can be broadly defined as the study of the *total* genetic content of samples that have degraded over time[3]. Areas of study that fall under ancient metagenomics include studies of host-associated microbial communities (e.g., ancient microbiomes[4]), genome reconstruction and analysis of specific microbial taxa (e.g., ancient pathogens[5]), and environmental reconstructions using sedimentary aDNA (sedaDNA)0[6]. Endogenous genetic material obtained from ancient samples has undergone a variety of degradation processes that can cause the original genetic signal to be overwhelmed by modern contamination. Therefore, to detect, quantify, and authenticate the remaining 'true' aDNA large DNA sequencing efforts are required[7,8]. These studies have only become feasible since the development of massively parallel 'next-generation sequencing', which enables the generation of large amounts of genetic data that are mostly uploaded to and stored on large generalised archives such as the European Bioinformatic Institute's (EBI) European Nucleotide Archive (ENA, https://www.ebi.ac.uk/ena/) or the US National Center for Biotechnology Information (NCBI)'s Sequence Read Archive (SRA, https://www.ncbi.nlm.nih.gov/sra). However, as these are generalised databases used for many kinds of genetic studies, searching for and identifying ancient metagenomic samples can be difficult and time consuming, partly because of the absence of standardised metadata reporting for ancient metagenomic data. Consequently, researchers must resort to repeated extensive literature searches of heterogeneously reported and inconsistently formatted publications to locate ancient metagenomic datasets. Overcoming the difficulty of finding previously published samples is particularly pertinent to studies of aDNA, as palaeontological and archaeological samples are by their nature limited, and avoiding repeated or redundant sampling is of high priority[9–11].

To address these issues, we established AncientMetagenomeDir, a CC-BY 4.0 licensed community-curated collection of annotated sample lists that aims to guide researchers to all published ancient metagenomics-related samples with publicly available sequence data. AncientMetagenomeDir was conceived by members of a recently established international and open community of researchers working in ancient metagenomics (Standards, Precautions and Advances in Ancient Metagenomics, or 'SPAAM' - https://spaam-community.github.io), whose aim is to foster research collaboration and define standards in analysis and reporting within the field. The collection aims to be comprehensive but lightweight, consisting of tab-separated value (TSV) tables for different major sub-disciplines of ancient metagenomics. These tables contain essential, sample-specific information for aDNA studies, including: geographic coordinates, temporal data, sub-discipline specific critical information, and public archive accession codes that guide researchers to associated sequence data (see Methods). This simple format, together with comprehensive guides and documentation, encourages continuous contributions from the community and facilitates usage of the resource by researchers coming from non-computational backgrounds, something common in interdisciplinary fields such as archaeo- and palaeogenetics.

AncientMetagenomeDir is designed to track the development of ancient metagenomics through regular releases. As of release v20.09, this includes 87 studies published since 2011, representing 443 ancient host-associated metagenome samples, 269 ancient microbial genome level sequences, and 312 sediment samples (Fig. 1) spanning 49 countries (Fig. 2). We expect AncientMetagenomeDir to deliver three key benefits. First, it will contribute to the longevity of important cultural heritage by guiding future sampling strategies, thereby reducing the risk of repeated or over-sampling of the same samples or regions. Second, it can serve as a starting point for the development of software to allow rapid aggregation of actual data files and field-specific data processing. Third, it will assist in expanding meta-analyses (such as[12,13]) to a wider range of sample types and DNA sources in order to tackle broader palaeogenetic, ecological, and evolutionary questions. Finally, as a community-curated resource designed specifically for widespread participation, AncientMetagenomeDir will help the field to define common standards of metadata reporting (such as with MIxS checklists[14]), facilitating the creation of future databases that are consistent, and richer, in useful metadata.

Methods

Repository Structure. AncientMetagenomeDir[15] is a community-curated set of tables maintained on GitHub containing metadata from published ancient metagenomic studies (https://github.com/SPAAM-community/AncientMetagenomeDir). While most submissions are made by SPAAM members, anyone with a GitHub account is welcome to propose (termed here 'proposer') and/or add publications for inclusion (termed 'contributor'). Proposers and contributors can be (but do not have to be) authors of the original publication(s) proposed for inclusion. Submitted studies must be published in a peer-reviewed journal because the purpose of AncientMetagenomeDir is not to act as a quality filter and we do not currently make assessments based on data quality. The tables are formatted as tab-separated value (TSV) files in order to maximize accessibility for all researchers and to allow portability between different data analysis software.

Valid samples for inclusion currently fall under three sub-fields: (1) host-associated metagenomes (i.e., host-associated or skeletal material microbiomes), (2) host-associated single genomes (i.e., pathogen or commensal microbial genomes), and (3) environmental metagenomes (e.g., sedaDNA). In addition, a fourth category is currently planned: (4) anthropogenic metagenomes (e.g., dietary and microbial DNA within pottery crusts, or microbial DNA and handling debris on parchment). The definitions under which a sample is considered 'ancient' is adapted on a per sub-field basis. Generally, samples are required to have had reported evidence of hydrolytic damage at molecule termini, short fragment lengths, and contain fraction of non-endogenous content

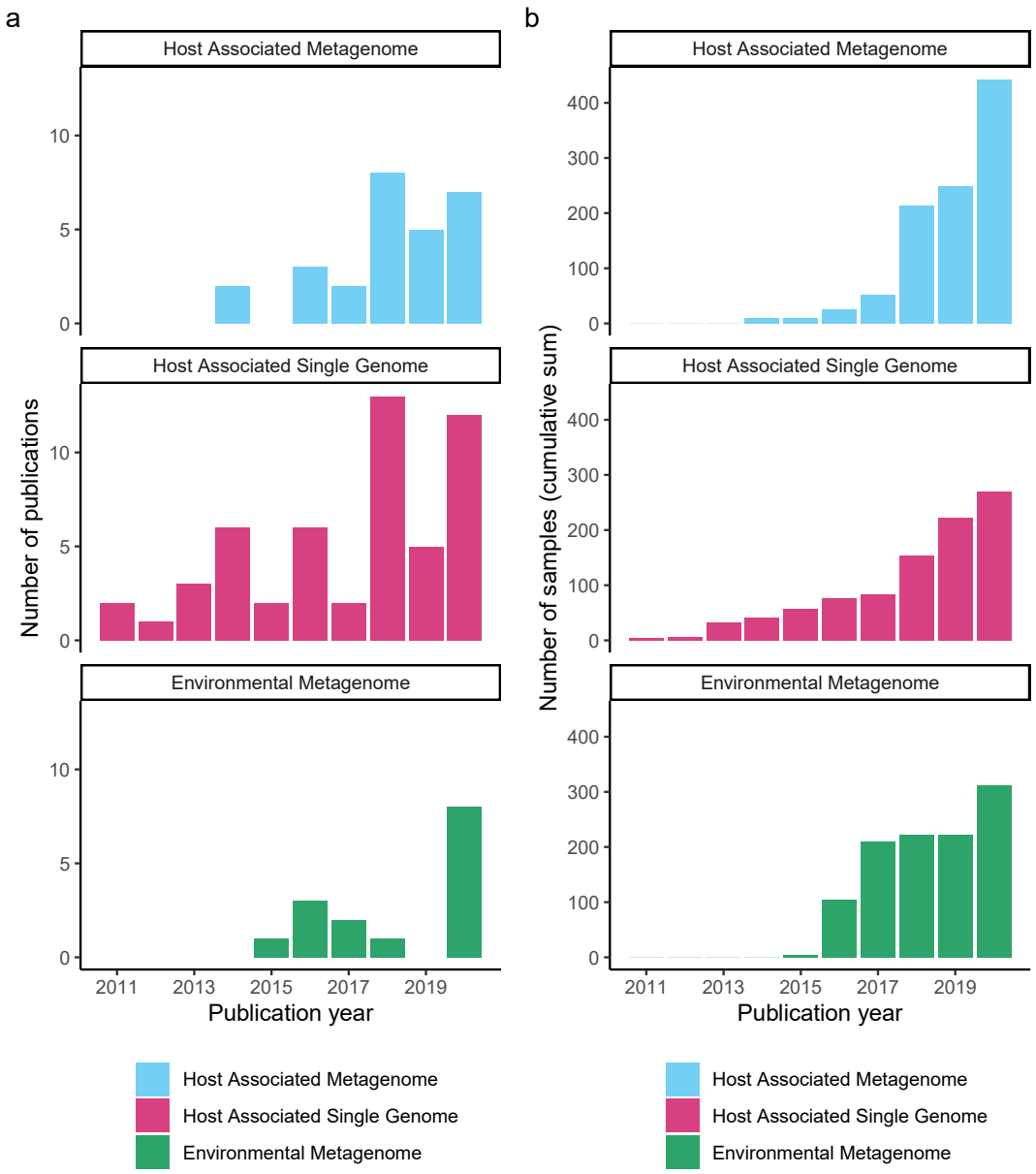

Fig. 1 Timelines depicting the development of the sub-disciplines of ancient metagenomics as recorded in AncientMetagenomeDir as per release v20.09. (**a**) Number of ancient metagenomic publications per year. (**b**) Cumulative sum of published samples with genetic sequencing data or sequences in publicly accessible archives.

(e.g. as summarised in[3]). However, for example, due to regular use in ancient pathogenomics studies, samples preserved in long-term medical collections from the last century that have limited degradation may also be included. In the first release of AncientMetagenomeDir, we have specified a minimum age of older than 1950 CE. Samples must have been sequenced using a shotgun metagenomic approach, or alternatively a whole organelle- or chromosome-level enrichment approach, and sequence data must be publicly available on an established or stable archive. INDSC-associated repositories such as the EBI's ENA or NCBI's SRA and Genbank databases are preferred, as they are the most accepted and commonly used archives for raw sequencing data. However, DOI-issuing long-term archives (such as Zenodo or Figshare), institutional repositories (such as institutional data services), or field-specific established repositories (e.g., TreeBASE) can also be accepted. Data on personal or lab websites are not accepted due to uncertain storage longevity. We currently do not include laboratory negative controls, as we consider these to be 'artefacts' of lab procedures and better addressed with experiment-level metadata. If required by a researcher, controls can be identified via sample-associated project accession codes.

Publications included in the current release of AncientMetagenomeDir were selected for inclusion based on direct contributions by authors of publications and also from literature reviews of each sub-field made by the SPAAM community. In this process, a proposer initially suggests a publication to be included via a GitHub 'Issue'. Publications may belong to multiple categories, and the corresponding issue is tagged with relevant category 'labels' to assist with faster evaluation and task distribution.

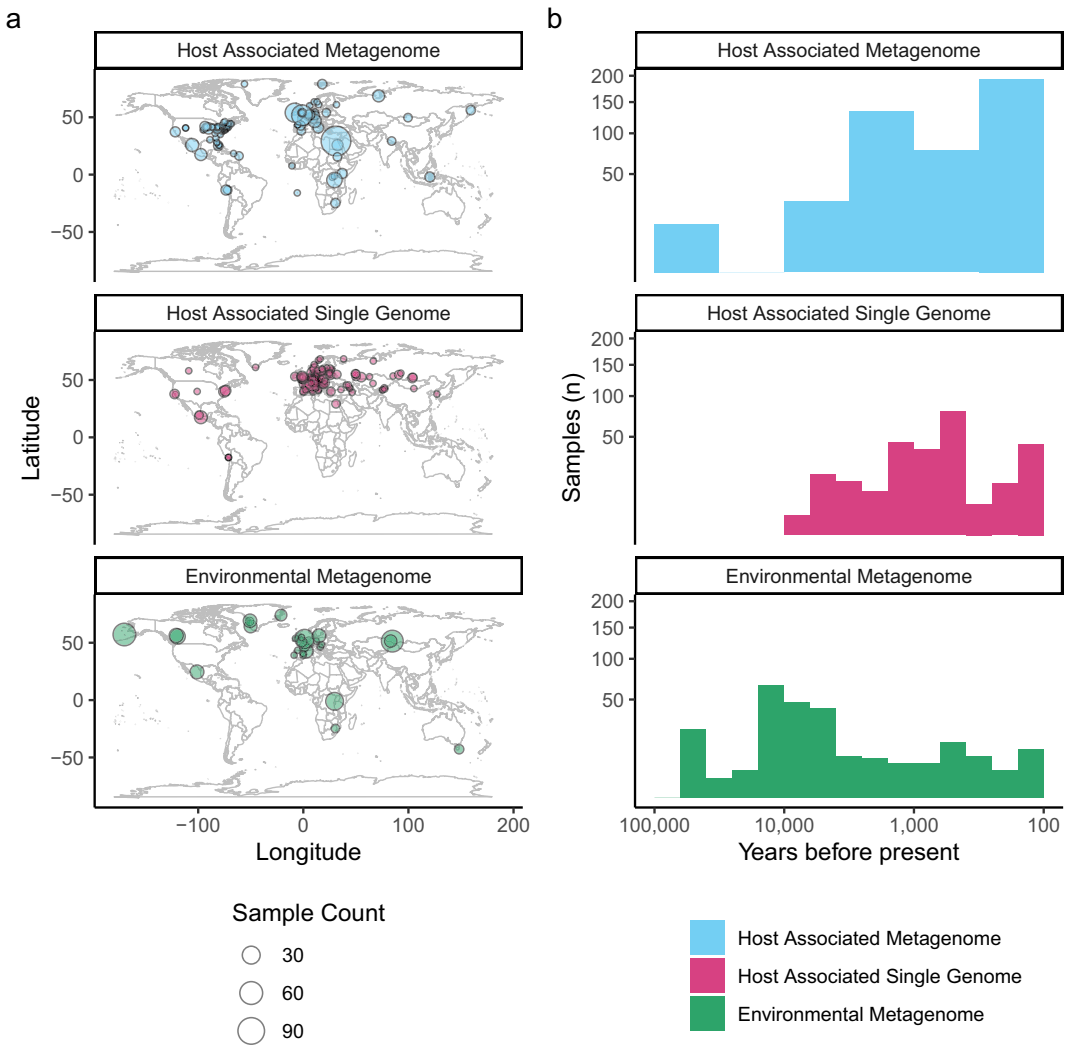

Fig. 2 Summary of temporal and spatial information of ancient metagenomic samples as recorded in AncientMetagenomeDir v20.09. (**a**) Maps depicting the geographic distribution of samples for each sub-discipline. (**b**) Histogram of sample ages for each sub-discipline. For visualisation purposes, plot axes are log-scaled, bins calculated using the 'Freedman–Diaconis' rule, and only samples dated to younger than 50,000 years are displayed.

Data acquisition. Members of the SPAAM community (termed 'curators') evaluate proposed publications for applicability under the criteria described above. Once approved, any member of the open SPAAM community can assign themselves to the corresponding Issue and will henceforth act as the contributor. A proposer from outside SPAAM who wishes to also be a contributor can be added to the SPAAM community by contacting a current member if desired. The contributor then creates a git branch from the main repository, manually extracts the relevant metadata from the given publication, and adds it to the assigned table (e.g., host associated metagenome, or environmental metagenome). Extensive documentation on submissions, including instructions on using GitHub, are available via tutorial documents and the associated repository wiki. Both are accessible via the main repository README under the 'Contributing' section. Furthermore, detailed documentation is also available to assist contributors and ensure correct entry of metadata, with one README file per table that contains column definitions and guidelines on how to interpret and record metadata.

The metadata in each table covers four main categories: publication metadata (project name, year, and publication DOI), geographic metadata (site name, coordinates, and country), sample metadata (sample name, sample age, material type, and (meta)genome type) and sequencing archive information (archive, sample archive accession ID). Due to inconsistency in the ways metadata are reported in publications and archives, and to maintain concise records, we have specified (standardised) approximations for the reporting of sample ages, geographic locations, and archive accessions, following MIxS[14] categories where possible. This approach allows researchers using the dataset to access sufficiently approximate information during search queries to identify samples of interest (e.g. all samples from Italy dating from between 4500-2500 Before Present (BP), i.e., from 1950), which they can subsequently manually check in the original publication to obtain the exact dating information (e.g., Late Bronze Age, 3725+/−15 BP). Due to inconsistency in dating and reporting methods, dates are reported

Field	Description	Field Type	Field Format
project_name	Unique AncientMetagenomeDir key for study	String	FirstAuthorYYYY
publication_year	Publication year of study	Integer	YYYY
publication_doi	Publication DOI (or library permalink)	String	Regex
site_name	Specific locality name where sample taken from	String	Free text
latitude	Latitude in decimal coordinate (WGS84 projection)	Number	Max. 3 decimals
longitude	Longitude in decimal coordinate (WGS84 projection)	Number	Max. 3 decimals
geo_loc_name	Present-day country name (INSDC) that locality resides in	String	Restricted enum
sample_name	Name of sample as reported in publication or archive	String	Free text
sample_age	Approximate date (before 1950, rounded to last 100 years)	Integer	YYYY
sample_age_doi	DOI of source of date. Can be more recent publication.	String	Regex
collection_date	Date of sampling of material for genetic analysis	Integer	YYYY
archive	Name of established data repository	String	Restricted enum
archive_accession	Sample-level accession code in data repository	String	Free text

Table 1. Core fields that are required for all AncientMetagenomeDir sub-discipline tables, including field type and standardised formatting description. Field formats are defined in a JSON schema, against which each new study submission is cross-checked by automated continuous integration (CI) checks and community peer-review. Further sub-discipline specific fields are included in the corresponding table, as required by the community.

(where relevant) as uncalibrated years BP, and rounded to the nearest 100 years, due to the range of calculation and reporting methods (radiocarbon dating vs. historical records, calibrated vs. uncalibrated radiocarbon dates, etc.). We hope that future extensions of AncientMetagenomeDir will include more exact dating information, such as raw dates and radiocarbon lab codes, to allow for consistent calibration of whole datasets for more precise dating information. Geographic coordinates are restricted to a maximum of three decimals, with fewer decimals indicating location uncertainty (e.g., if a publication only reports a region rather than a specific site). For sequence accession codes, we opted for using *sample* accession codes rather than direct sequencing data IDs. This is due to the myriad ways in which data are generated and uploaded to repositories (e.g., one sample accession per sample vs. one sample accession per library; or uploading raw sequencing reads vs. only consensus sequences). We found that in most cases sample accession codes are the most straightforward starting points for data retrieval. However, we did observe errors in some data accessions uploaded to public repositories, such as multiple sample codes assigned to different libraries of the same sample, and insufficient metadata to link accessions to specific samples reported in a study. Overall, we found that heterogeneity in sample (meta)data uploading was a common problem, which highlights the need for improvements in both training and community-agreed standards for data sharing and metadata reporting in public repositories (such as an ancient metagenomic MIxS extension). In addition to metadata recorded across all sample types, we have added table-specific metadata fields to individual categories as required (e.g., species for single genomes and community type for microbiomes). Such fields can be further extended or modified with the agreement of the community.

Data validation. After all metadata has been added, a contributor makes a Pull Request (PR) into the master branch. Every PR undergoes an automated 'continuous-integration' validation check via the open-source companion tool AncientMetagenomeDirCheck[16] (https://github.com/SPAAM-community/AncientMetagenomeDirCheck, License: GNU GPLv3). This tool automatically checks each submission for conformity against a specification schema of minimum required information and formatting consistency (see Technical Validation). Usage of controlled vocabularies, alongside stable linking (via DOIs), within the specifications ensures reliable querying of the dataset, and allows future expansion to include richer metadata by linking to other databases. Descriptions for the minimum required fields for an AncientMetagenomeDir table are provided in Table 1.

Once automated checks are cleared, a contributor then requests a minimum of one peer-review performed by another member of the SPAAM community (termed 'reviewer'). This reviewer checks the entered data for consistency against the table's README file and also for accuracy against the original publication. Once the automated and peer-review checks are both satisfied, the publication's metadata are then added to the master branch and the corresponding Issue is closed. For each added publication, a CHANGELOG is maintained to track the papers included in each release and to record any corrections that may have been made (e.g., if new radiocarbon dates are published for previously entered samples). The CHANGELOG or Issues pages on GitHub can be consulted to check whether a given publication has already been added (or excluded) from a table. Proposals and submissions can be made at any time, and contributed data is available on the main GitHub repository immediately after integration into the master branch. However, citable versions of the database are only made on each new (non-modifiable) release (see section Data Records). New submissions or corrections received after a release are included in subsequent versions.

Data Records

AncientMetagenomeDir (https://github.com/SPAAM-community/AncientMetagenomeDir) and AncientMetagenomeDirCheck (https://github.com/SPAAM-community/AncientMetagenomeDirCheck) are both maintained on GitHub. AncientMetagenomeDir has regular quarterly releases, each of which has a release-specific DOI assigned via the Zenodo long-term data repository. Both the collection and tools are

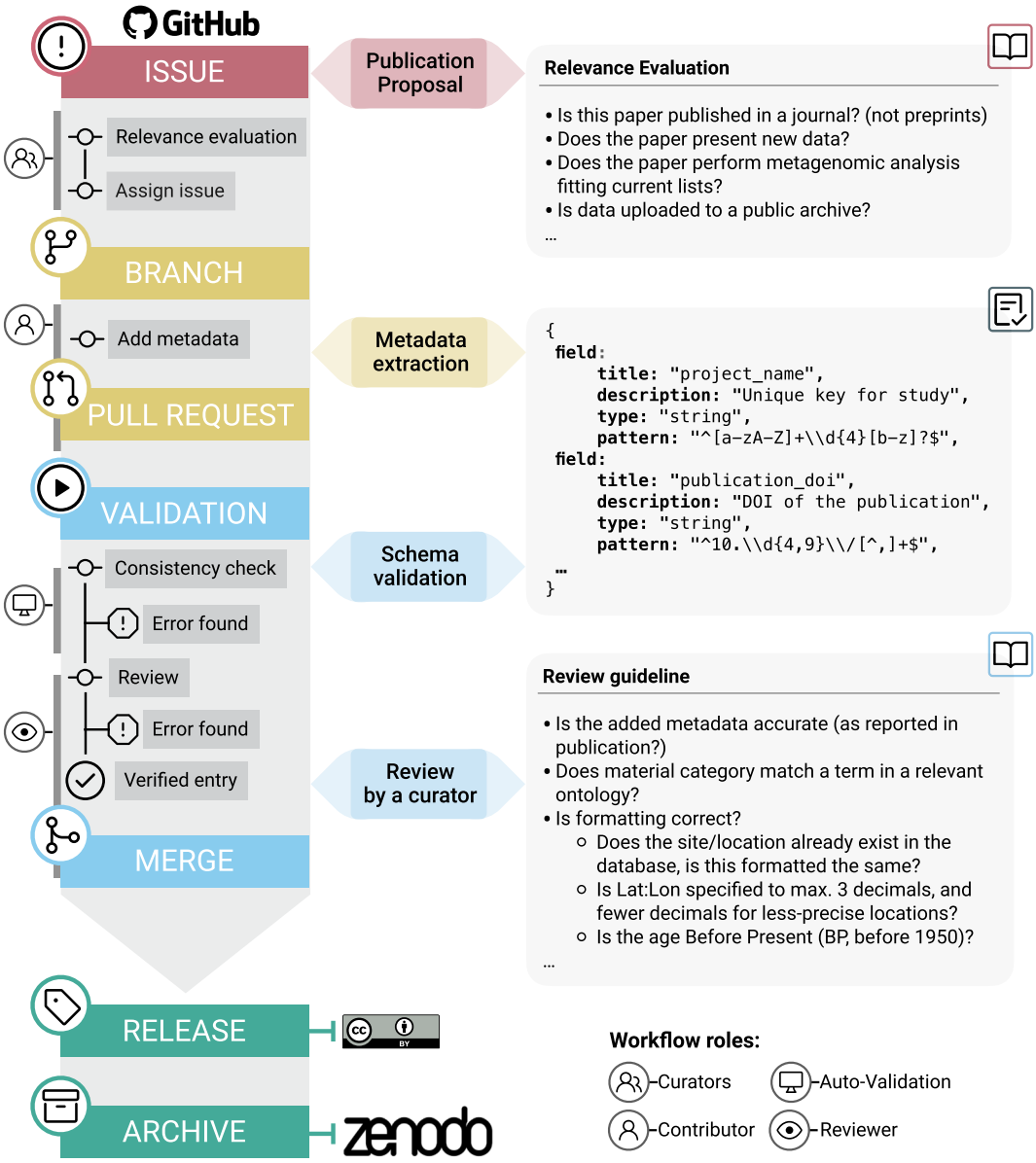

Fig. 3 AncientMetagenomeDir submission and update workflow. The submission workflow is carried out on GitHub, and final releases are archived at Zenodo. Submissions go through both automated computational validation and also peer-review for consistency and accuracy.

archived in the Zenodo repository with generalised DOIs[15] and[16], respectively. The full workflow can be seen in Fig. 3. Releases are made under a CC-BY 4.0 license (https://creativecommons.org/licenses/by/4.0/).

Technical Validation

All data entries to AncientMetagenomeDir undergo automated continuous-integration validation prior to submission into the protected main branch. These tests must pass before being additionally peer-reviewed by other member(s) of the community (see section Data Validation). Automated continuous-integration (CI) validation tests consist of regex patterns to control formatting of specified fields (e.g. DOIs, project IDs, date formats), and cross-checking of entries against controlled vocabularies defined in centralised JSON schema, often derived from established term-ontologies. For example, valid country codes are guided by the International Nucleotide Sequence Database Collaboration (INSDC) controlled vocabulary (http://www.insdc.org/country.html), host and microbial species names are defined by the NCBI's Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy), and material types are defined by the ontologies listed on the EBI's Ontology Look Up service (https://www.ebi.ac.uk/ols/index) - particularly the Uberon[17] and Envo ontologies[18,19]. Entries must also have valid sample accession IDs corresponding to shotgun metagenomic, genome-enriched sequence data, or - when only available - consensus sequences, uploaded to established and stable public archives.

Usage Notes

Usage of the resource typically consists of loading the TSV file of interest in software such as Microsoft Excel, LibreOffice Calc, or R. The data table can be subsequently sorted or queried to identify datasets of interest. It should be noted that certain metadata fields (e.g., sample_age, latitude, and longitude) are approximate and do not provide *exact* values; rather, if exact values for these fields are required, they must be retrieved from the original publication or requested from the publications' authors. All selected data retrieved using AncientMetagenomeDir and used in subsequent studies should be cited using the original publication citation as well as AncientMetagenomeDir.

Retrieval of sequencing data using sample accession codes can be achieved manually via a given archive's website, or via archive-supplied tools (e.g., Entrez Programming Utilities for NCBI's SRA (https://www.ncbi.nlm.nih.gov/books/NBK179288/), or enaBrowserTools for EBI's ENA (https://github.com/enasequence/enaBrowserTools).

Contributions to the tables are also facilitated by extensive step-by-step documentation on how to use GitHub and AncientMetagenomeDir, the locations of which are listed on the main README of the repository.

Code availability

An R notebook used for generating images with package versions can be found in the AncientMetagenomeDir repository at https://github.com/SPAAM-community/AncientMetagenomeDir/tree/master/assets/analysis (commit 4308bb7). Code for validation of the dataset (with version 1 used for the first release of AncientMetagenomeDir) can be found at https://github.com/SPAAM-community/AncientMetagenomeDirCheck and https://doi.org/10.5281/zenodo.4003826.

Received: 18 September 2020; Accepted: 13 December 2020;
Published online: 26 January 2021

References

1. Anagnostou, P. *et al.* When data sharing gets close to 100%: what human paleogenetics can teach the open science movement. *PloS one* **10**, e0121409, https://doi.org/10.1371/journal.pone.0121409 (2015).
2. Wilkinson, M. D. *et al.* The FAIR guiding principles for scientific data management and stewardship. *Scientific data* **3**, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
3. Warinner, C. *et al.* A robust framework for microbial archaeology. *Annual review of genomics and human genetics* **18**, 321–356, https://doi.org/10.1146/annurev-genom-091416-035526 (2017).
4. Warinner, C., Speller, C., Collins, M. J. & Lewis, C. M. Jr. Ancient human microbiomes. *Journal of human evolution* **79**, 125–136, https://doi.org/10.1016/j.jhevol.2014.10.016 (2015).
5. Spyrou, M. A., Bos, K. I., Herbig, A. & Krause, J. Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature reviews. Genetics* **20**, 323–340, https://doi.org/10.1038/s41576-019-0119-1 (2019).
6. Edwards, M. E. The maturing relationship between quaternary paleoecology and ancient sedimentary DNA. *Quaternary Research* **96**, 39–47, https://doi.org/10.1017/qua.2020.52 (2020).
7. Dabney, J., Meyer, M. & Pääbo, S. Ancient DNA damage. *Cold Spring Harbor perspectives in biology* **5**, https://doi.org/10.1101/cshperspect.a012567 (2013).
8. Peyrégne, S. & Prüfer, K. Present-Day DNA contamination in ancient DNA datasets. *BioEssays: news and reviews in molecular, cellular and developmental biology* e2000081, https://doi.org/10.1002/bies.202000081 (2020).
9. Prendergast, M. E. & Sawchuk, E. Boots on the ground in africa's ancient DNA 'revolution': archaeological perspectives on ethics and best practices. *Antiquity* **92**, 803–815, https://doi.org/10.15184/aqy.2018.70 (2018).
10. Pálsdóttir, A. H., Bläuer, A., Rannamäe, E., Boessenkool, S. & Hallsson, J. H. Not a limitless resource: ethics and guidelines for destructive sampling of archaeofaunal remains. *Royal Society open science* **6**, 191059, https://doi.org/10.1098/rsos.191059 (2019).
11. Wagner, J. K. *et al.* Fostering responsible research on ancient DNA. *American journal of human genetics* **107**, 183–195, https://doi.org/10.1016/j.ajhg.2020.06.017 (2020).
12. Allentoft, M. E. *et al.* The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences* **279**, 4724–4733, https://doi.org/10.1098/rspb.2012.1745 (2012).
13. Kistler, L., Ware, R., Smith, O., Collins, M. & Allaby, R. G. A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic acids research* **45**, 6310–6320, https://doi.org/10.1093/nar/gkx361 (2017).
14. Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature biotechnology* **29**, 415–420, https://doi.org/10.1038/nbt.1823 (2011).
15. Fellows Yates, J. A. *et al.* Spaam-community/ancientmetagenomedir: v20.09.1: Ancient ksour of ouadane. *Zenodo* https://doi.org/10.5281/zenodo.4011751 (2020).
16. Borry, M. & Fellows Yates, J. A. Spaam-commmunity/ancientmetagenomedircheck: Ancientmetagenomedircheck v1.0. *Zenodo* https://doi.org/10.5281/zenodo.4003826 (2020).
17. Mungall, C. J., Torniai, C., Gkoutos, G. V., Lewis, S. E. & Haendel, M. A. Uberon, an integrative multi-species anatomy ontology. *Genome biology* **13**, R5, https://doi.org/10.1186/gb-2012-13-1-r5 (2012).
18. Buttigieg, P. L. *et al.* The environment ontology: contextualising biological and biomedical entities. *Journal of biomedical semantics* **4**, 43, https://doi.org/10.1186/2041-1480-4-43 (2013).
19. Buttigieg, P. L. *et al.* The environment ontology in 2016: bridging domains with increased scope, semantic density, and interoperation. *Journal of biomedical semantics* **7**, 57, https://doi.org/10.1186/s13326-016-0097-6 (2016).

Acknowledgements

We would like to thank the wider SPAAM community (https://spaamcommunity.github.io) for their input in developing the project. J.A.F.Y., A.A.V., I.V., M.B., M.A.S, A.H. and C.W. acknowledge the Max Planck Society for financial support. J.A.F.Y was partly funded by the European Research Council (ERC) under the European Union's Horizon 2020 research innovation programme (ERC-2015-StG 678901-FoodTransforms to Philipp W. Stockhammer, Ludwig Maximilian University Munich, Germany). B.C. is supported by grant ERC-2014-ADG 670518 (to V. Gaffney, University of Bradford, United Kingdom). Å.J.V. is supported by Carlsbergfondet Semper Ardens grant CF18-1109 (to M. Thomas P. Gilbert, University of Copenhagen, Denmark). A.H. is partly supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's

Excellence Strategy–EXC 2051–Project-ID 390713860 (to C. Warinner, Friedrich Schiller University, Germany). E.J.G is supported by Arts & Humanities Research Council (grant number AH/N005015/1) and Natural History Museum (London, United Kingdom). M.J.B.-L. is supported by grant Wellcome Trust Seed Award in Science 208934/Z/17/Z, and by project IA201219 PAPIIT-DGAPA- UNAM (to María C. Ávila Arcos, LIIGH, Mexico). M.A.S. is supported by grant ERC-CoG 771234 PALEoRIDER (to Wolfgang Haak, Max-Planck-Institute for the Science of Human History, Germany). A.S.G is supported by NSF GRFP Grant No. DGE1255832 (any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation). S.L.R. is supported by NIH Genetics and Regulation Training Grant 5T32GM007197-46. I.V., M.B., and C.W. are supported by Werner Siemens Stiftung (Paleobiochemistry) (to C. Warinner, Leibnitz Institute for Natural Product Research and Infection Biology, Germany). Open Access funding enabled and organized by Projekt DEAL.

Author contributions

J.A.F.Y. and C.W. conceptualised the project. J.A.F.Y. designed the project and infrastructure with input from all co-authors. M.B. developed software. J.A.F.Y., A.A.V., Å.J.V., B.C., I.M.V., M.J.B.-L., A.F.-G., E.J.G., S.L.R., P.D.H., M.A.S., A.H., A.S.G., J.H., A.F.A., V.Z. and C.W. acquired data. J.A.F.Y. drafted the manuscript with input from all co-authors.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to J.A.F.Y. or C.W.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

5 Manuscript C: Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager

5.1 Overview and contribution

Manuscript Nr.: C

Title of Manuscript Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager

Authors Fellows Yates, J. A., Lamnidis, T. C., Borry, M., Andrades Valtueña, A., Fagernäs, Z., Clayton, S., Garcia, M. U., Neukamm, J., & Peltzer, A.

Citation Fellows Yates, J. A., Lamnidis, T. C., Borry, M., Andrades Valtueña, A., Fagernäs, Z., Clayton, S., Garcia, M. U., Neukamm, J., & Peltzer, A. (2021). Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ*, 9, e10947. `https://doi.org/10.7717/peerj.10947`

The candidate is
☑ First author, ☐ Co-first author, ☑ Corresponding author, ☐ Co-author.

Status Published

Proportion (in %) of authors in the publication (indicated from 20%)

Author	Concept	Data Analysis	Experiment	Manuscript Composition	Material Provision
Fellows Yates, J. A.	45	100	50	75	NA
Lamnidis, T. C.	0	0	0	0	NA
Borry, M.	0	0	0	0	NA
Andrades Valtueña, A.	0	0	0	0	NA
Fagernäs, Z.	0	0	0	0	NA
Clayton, S.	0	0	0	0	NA
Garcia, M.	0	0	0	0	NA
Neukamm, J.	0	0	0	0	NA
Peltzer, A.	45	0	30	0	NA

5.2 Article

Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager

James A. Fellows Yates[1,2], Thiseas C. Lamnidis[1], Maxime Borry[1], Aida Andrades Valtueña[1], Zandra Fagernäs[1], Stephen Clayton[1], Maxime U. Garcia[3,4], Judith Neukamm[5,6] and Alexander Peltzer[1,7]

[1] Department of Archaeogenetics, Max Planck Institute for the Science of Human History, Jena, Germany
[2] Institut für Vor- und Frühgeschichtliche Archäologie und Provinzialrömische Archäologie, Ludwig-Maximilians-Universität München, Münich, Germany
[3] National Genomics Infrastructure, Science for Life Laboratory, Stockholm, Sweden
[4] Barntumörbanken, Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden
[5] Institute of Evolutionary Medicine, University of Zurich, Zurich, Switzerland
[6] Institute for Bioinformatics and Medical Informatics, Eberhard-Karls University Tübingen, Tübingen, Germany
[7] Quantitative Biology Center, Eberhard-Karls University Tübingen, Tübingen, Germany

ABSTRACT

The broadening utilisation of ancient DNA to address archaeological, palaeontological, and biological questions is resulting in a rising diversity in the size of laboratories and scale of analyses being performed. In the context of this heterogeneous landscape, we present an advanced, and entirely redesigned and extended version of the EAGER pipeline for the analysis of ancient genomic data. This Nextflow pipeline aims to address three main themes: accessibility and adaptability to different computing configurations, reproducibility to ensure robust analytical standards, and updating the pipeline to the latest routine ancient genomic practices. The new version of EAGER has been developed within the nf-core initiative to ensure high-quality software development and maintenance support; contributing to a long-term life-cycle for the pipeline. nf-core/eager will assist in ensuring that a wider range of ancient DNA analyses can be applied by a diverse range of research groups and fields.

Subjects Anthropology, Bioinformatics, Evolutionary Studies, Genomics
Keywords Bioinformatics, Palaeogenomics, Ancient DNA, Pipeline, Nextflow, Reproducibility, Genomics, Metagenomics

Submitted 29 October 2020
Accepted 25 January 2021
Published 16 March 2021

Corresponding authors
James A. Fellows Yates, fellows@shh.mpg.de
Alexander Peltzer, peltzer@shh.mpg.de

Academic editor
Alexander Schliep

Additional Information and Declarations can be found on page 16

DOI **10.7717/peerj.10947**

INTRODUCTION

Ancient DNA (aDNA) has become a widely accepted source of biological data, helping to provide new perspectives for a range of fields including archaeology, cultural heritage, evolutionary biology, ecology, and palaeontology. The utilisation of short-read high-throughput sequencing has allowed the recovery of whole genomes and genome-wide data from a wide variety of sources, including (but not limited to), the skeletal remains of animals (*Palkopoulou et al., 2015*; *Orlando et al., 2013*; *Frantz et al., 2019*; *Star et al., 2017*), modern and archaic humans (*Damgaard et al., 2018*; *Green et al, 2010*; *Meyer et al, 2012*; *Slon et al, 2018*)-rv, bacteria (*Bos et al., 2014*; *Namouchi et al., 2018*;

How to cite this article Fellows Yates JA, Lamnidis TC, Borry M, Andrades Valtueña A, Fagernäs Z, Clayton S, Garcia MU, Neukamm J, Peltzer A. 2021. Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ* **9**:e10947
http://doi.org/10.7717/peerj.10947

Schuenemann et al., 2018), viruses (Mühlemann et al., 2018; Krause-Kyora et al., 2018), plants (Wales et al., 2019; Gutaker et al., 2019), palaeofaeces (Tett et al., 2019; Borry et al., 2020), dental calculus (Warinner et al., 2014; Weyrich et al., 2017), sediments (Willerslev et al., 2014; Slon et al., 2017), medical slides (Van Dorp et al., 2019), parchment (Teasdale et al., 2015), and recently, ancient 'chewing gum' (Jensen et al., 2019; Kashuba et al., 2019). Improvement in laboratory protocols to increase yields of otherwise trace amounts of DNA has at the same time led to studies that can total hundreds of ancient individuals (Olalde et al., 2018; Mathieson et al., 2018), spanning single (Bos et al., 2011) to thousands of organisms (Warinner et al., 2014). These differences of disciplines have led to a heterogeneous landscape in terms of the types of analyses undertaken, and their computational resource requirements (Tastan Bishop et al., 2015; Bah et al, 2018). Taking into consideration the unequal distribution of resources (and infrastructure such as internet connection), easy-to-deploy, streamlined and efficient pipelines can help increase accessibility to high-quality analyses.

The degraded nature of aDNA poses an extra layer of complexity to standard modern genomic analysis. Through a variety of processes (Lindahl, 1993) DNA molecules fragment over time, resulting in ultra-short molecules (Meyer et al., 2016). These sequences have low nucleotide complexity making it difficult to identify with precision which part of the genome a read (a sequenced DNA molecule) is derived from. Fragmentation without a 'clean break' leads to uneven ends, consisting of single-stranded 'overhangs' at ends of molecules that are susceptible to chemical processes such as deamination of nucleotides. These damaged nucleotides then lead to misincorporation of complementary bases during library construction for high-throughput DNA sequencing (Briggs et al., 2007). On top of this, taphonomic processes such as heat, moisture, and microbial- and burial-environment processes lead to varying rates of degradation (Kistler et al., 2017; Warinner et al., 2017). The original DNA content of a sample is therefore increasingly lost over time and supplanted by younger 'environmental' DNA. Later handling by archaeologists, museum curators, and other researchers can also contribute 'modern' contamination. While these characteristics can help provide evidence towards the 'authenticity' of true aDNA sequences (e.g., the aDNA cytosine to thymine or C to T 'damage' deamination profiles as by Ginolhac et al., 2011), they also pose specific challenges for genome reconstruction, such as unspecific DNA alignment and/or low coverage and miscoding lesions that can result in low-confidence genotyping. These factors often lead to prohibitive sequencing costs when retrieving enough data for modern high-throughput short-read sequencing data pipelines (such as more than 1 billion reads for a 1X depth coverage *Yersinia pestis* genome, as in Rasmussen et al., 2015), and thus aDNA-tailored methods and techniques are required to overcome these challenges.

Two previously published and commonly used pipelines in the field are PALEOMIX (Schubert et al., 2014) and EAGER (Peltzer et al., 2016). These two pipelines take a similar approach to link together standard tools used for Illumina high-throughput short-read data processing (sequencing quality control, sequencing adapter removal and/or paired-end read merging, mapping of reads to a reference genome, genotyping, etc.). However, they have a specific focus on tools that are designed for, or well-suited

for aDNA (such as the bwa aln algorithm for ultra-short molecules (*Li & Durbin, 2009*) and mapDamage (*Jónsson et al., 2013*) for evaluation of aDNA characteristics). Yet, neither of these genome reconstruction pipelines have had major updates to bring them in-line with current routine bioinformatic practices (such as continuous integration tests and software containers) and aDNA analyses. In particular, *Meta*genomic screening of off-target genomic reads for pathogens or microbiomes (*Warinner et al., 2014*; *Weyrich et al., 2017*) has become common in palaeo- and archaeogenetics, given its role in revealing widespread infectious disease and possible epidemics that have sometimes been previously undetected in the archaeological record (*Mühlemann et al., 2018*; *Krause-Kyora et al., 2018*; *Rasmussen et al., 2015*; *Andrades Valtueña et al., 2017*). Without easy access to the latest field-established analytical routines, ancient genomic studies risk being published without the necessary quality control checks that ensure aDNA authenticity, as well as limiting the full range of possibilities from their data. Given that material from samples is limited, there are both ethical as well as economical interests to maximise analytical yield (*Green & Speller, 2017*).

To address these shortcomings, we have completely re-implemented the latest version of the EAGER pipeline in Nextflow (*Di Tommaso et al., 2017*) (a domain-specific-language or 'DSL', specifically designed for the construction of omics analysis pipelines), and introduced new features and more flexible pipeline configuration. In addition, the renamed pipeline—nf-core/eager—has been developed in the context of the nf-core community framework (*Ewels et al., 2020*), which enforces strict guidelines for best-practices in software development.

MATERIALS AND METHODS

Updated Workflow

The new pipeline follows a similar structural foundation to the original version of EAGER (Fig. 1) and partially to PALEOMIX. Given Illumina short-read FASTQ and/or BAM files and a reference FASTA file, the core functionality of nf-core/eager can be split into five main stages:

1. Pre-processing:
   - Sequencing quality control: FastQC (*Andrews, 2010*)
   - Sequencing artefact clean-up (merging, adapter clipping): AdapterRemoval2 (*Schubert, Lindgreen & Orlando, 2016*), fastp (*Chen et al., 2018*)
   - Pre-processing statistics generation: FastQC
2. Mapping and post-processing:
   - Alignment against reference genome: BWA aln and mem (*Li & Durbin, 2009*; *Li, 2013*), CircularMapper (*Peltzer et al., 2016*), Bowtie2 (*Langmead & Salzberg, 2012*)
   - Mapping quality filtering: SAMtools (*Li et al., 2009*)
   - PCR duplicate removal: Picard MarkDuplicates (http://broadinstitute.github.io/picard/), DeDup (*Peltzer et al., 2016*)

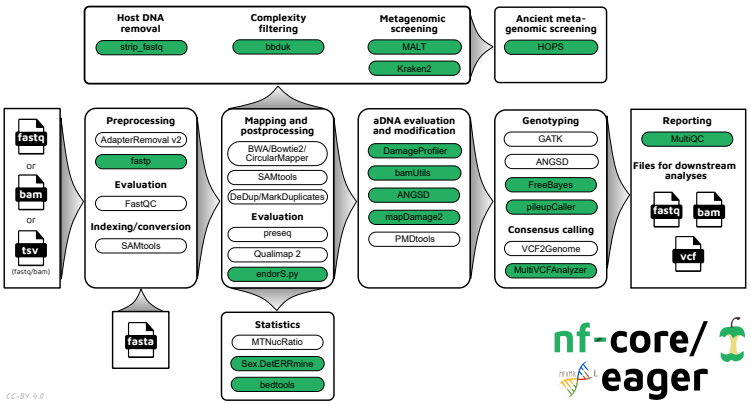

Figure 1 Simplified schematic of the nf-core/eager workflow pipeline. Green filled bubbles indicate new functionality added over the original EAGER pipeline.

Full-size DOI: 10.7717/peerj.10947/fig-1

- Mapping statistics generation: SAMtools, PreSeq (*Daley & Smith, 2013*), Qualimap2 (*Okonechnikov, Conesa & García-Alcalde, 2016*), bedtools (*Quinlan & Hall, 2010*), Sex.DetERRmine (*Lamnidis et al., 2018*)

3. aDNA evaluation and modification:
   - Damage profiling: DamageProfiler (*Neukamm, Peltzer & Nieselt, 2020*)
   - aDNA reads selection: PMDtools (*Skoglund et al., 2014*)
   - Damage removal/Base trimming: mapDamage2 (*Jónsson et al., 2013*), Bamutils (*Jun et al., 2015*)
   - Human nuclear contamination estimation: ANGSD (*Korneliussen, Albrechtsen & Nielsen, 2014*)
4. Variant calling and consensus sequence generation: GATK UnifiedGenotyper and HaplotypeCaller (*McKenna et al., 2010*), sequenceTools pileupCaller (https://github.com/stschiff/sequenceTools), VCF2Genome (*Peltzer et al., 2016*), MultiVCFAnalyzer (*Bos et al., 2014*)
5. Report generation: MultiQC (*Ewels et al., 2016*)

In nf-core/eager, all tools originally used in EAGER have been updated to their latest versions, as available on Bioconda (*Grüning et al., 2018*) and conda-forge (https://github.com/conda-forge), to ensure widespread accessibility and stability of utilised tools. The mapDamage2 (for damage profile generation) (*Jónsson et al., 2013*) and Schmutzi (for mitochondrial contamination estimation) (*Renaud et al., 2015*) methods have not been carried over to nf-core/eager, the first because a faster successor method is now available (DamageProfiler, *Neukamm, Peltzer & Nieselt, 2020*), and the latter because a stable release of the method could not be migrated to Bioconda at time of writing. We anticipate that there will be an updated version of Schmutzi in the near future that will allow us to integrate the method again into nf-core/eager. As an alternative, estimation of human *nuclear* contamination is now offered through ANGSD. mapDamage2 is however retained

to offer probabilistic *in silico* damage removal from BAM files. Support for the Bowtie2 aligner has been updated to have default settings optimised for aDNA (*Poullet & Orlando, 2020*).

New tools to the basic workflow include fastp for the removal of 'poly-G' sequencing artefacts that are common in 2-colour Illumina sequencing machines (such as the increasingly popular NextSeq and NovaSeq platforms). For variant calling, we have now included FreeBayes (*Garrison & Marth, 2012*) as an alternative to the human-focused GATK tools, and have also added pileupCaller for generation of genotyping formats commonly utilised in ancient human population analysis. We have also maintained the possibility of using the now officially unsupported GATK UnifiedGenotyper, as the supported replacement, GATK HaplotypeCaller, performs *de novo* assembly around possible variants; something that may not be suitable for low-coverage aDNA data.

Additional functionality tailored for ancient bacterial genomics includes integration of a SNP alignment generation tool, MultiVCFAnalyzer, which includes the ability to make an assessment of levels of cross-mapping from different related taxa to a reference genome - a common challenge in ancient bacterial genome reconstruction (as discussed in *Warinner et al., 2017*). The output SNP consensus alignment FASTA file can then be used for downstream analyses such as phylogenetic tree construction. Simple coverage statistics of particular annotations (e.g., genes) of an input reference is offered by bedtools, which can be used in cases such as for providing initial indications of functional differences between ancient bacterial strains (as in *Andrades Valtueña et al., 2017*). For analysis of human genomes, nf-core/eager can also give estimates of the relative coverage on the X and Y chromosomes with Sex.DetERRmine, which can be used to infer the biological sex of a given human individual. A dedicated 'endogenous DNA' calculator (endorS.py) is also included, to provide a percentage estimate of the sequenced reads matching the reference ('on-target') from the total number of reads sequenced per library.

Given the large amount of sequencing often required to yield sufficient genome coverage from aDNA data, palaeogenomicists tend to use multiple (differently treated) libraries, and/or merge data from multiple sequencing runs of each library or even samples. The original EAGER pipeline could only run a single library at a time, and in these contexts required significant manual user input in merging different FASTQ or BAM files of related libraries. A major upgrade in nf-core/eager is that the new pipeline supports automated processing of complex sequencing strategies for many samples, similar to PALEOMIX. This is facilitated by the optional use of a simple table (in TSV format, a format more commonly used in wet-lab stages of data generation, compared to PALEOMIX's YAML format) that includes file paths and additional metadata such as sample name, library name, sequencing lane, colour chemistry, and UDG treatment. This allows automated and simultaneous processing and appropriate merging and treatment of heterogeneous data from multiple sequencing runs and/or library types (Fig. 2).

The original EAGER and PALEOMIX pipelines required users to look through many independent output directories and files to make full assessment of their sequencing data. This has now been replaced in nf-core/eager with a much more extensive MultiQC report. This tool aggregates the log files of every supported tool into a single interactive report, and

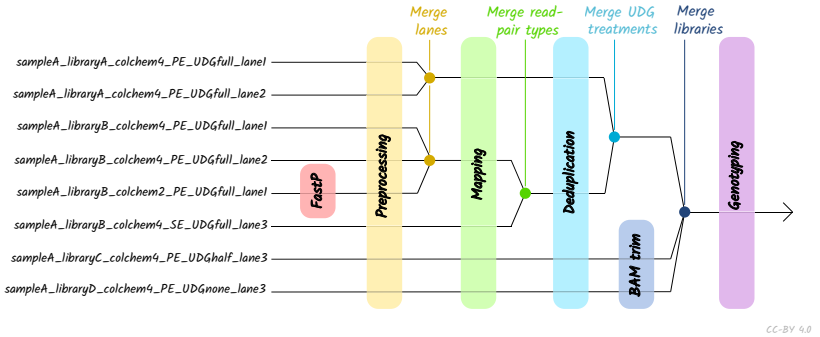

Figure 2 Diagram of different processing and library-merging points based on the nature of different libraries. Merge points represent merging of related BAM files as defined by metadata fields in an input TSV file. 'colchem' refers to the colour chemistry system of Illumina sequencers (2 for e.g., NextSeq or 4 for HiSeq machines). 'PE/SE' refers to paired-end and single-end sequencing chemistries. 'UDG' refers to uracil DNA glycosylase treatment, a laboratory procedure to completely or partially remove C to T mis-coding lesions. Lane refers to sequencing lane.

Full-size DOI: 10.7717/peerj.10947/fig-2

assists users in making a fuller assessment of their sequencing and analysis runs. We have developed a corresponding MultiQC module for every tool used by nf-core/eager, where possible, to enable comprehensive evaluation of all stages of the pipeline.

We have further extended the functionality of the original EAGER pipeline by adding ancient metagenomic analysis (Fig. 3); allowing reconstruction of the wider taxonomic content of a sample. We have added the possibility to screen all off-target reads (not mapped to the reference genome) with two metagenomic profilers: MALT (*Herbig et al., 2016*; *Vågene et al., 2018*) and Kraken2 (*Wood, Lu & Langmead, 2019*), in parallel to the mapping to a given reference genome (typically of the host individual, assuming the sample is a host organism). Pre-profiling removal of low-sequence-complexity reads that can slow down profiling and result in false-positive taxonomic identifications is offered through BBduk (Brian Bushnell: http://sourceforge.net/projects/bbmap/). Post-profiling characterisation of properties of authentic aDNA from metagenomic MALT alignments is carried out with MaltExtract of the HOPS pipeline (*Hübler et al., 2019*). This functionality can be used either for microbiome screening or putative pathogen detection. Ancient metagenomic studies sometimes include comparative samples from living individuals (*Velsko et al., 2019*). To support open data, whilst respecting personal data privacy, nf-core/eager includes a 'FASTQ host removal' script that creates raw FASTQ files, but with all reads successfully mapped to the reference genome removed. This allows for safe upload of metagenomic non-host sequencing data to public repositories after removal of identifiable (human) data, for example for microbiome studies.

An overview of the entire pipeline is shown in Fig. 1, and a tabular comparison of functionality between EAGER, PALEOMIX and nf-core/eager is in Table 1.

Usage

nf-core/eager can be run on POSIX-family operating systems (e.g., Linux and macOS) and has at minimum three dependencies, Java (>=8), Nextflow (*Di Tommaso et al., 2017*),

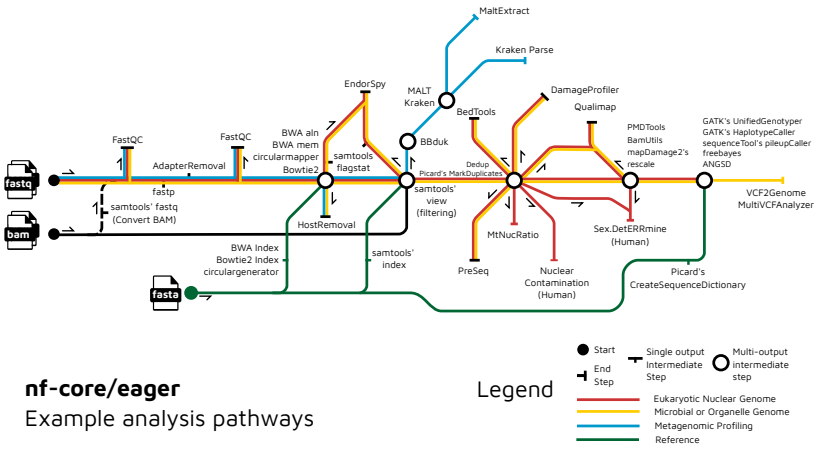

Figure 3 **Example routes of analyses offered by nf-core/eager.** nf-core/eager includes functionality that is applicable to different common aDNA research contexts, such as for population genetics, microbial genome reconstruction, and metagenomic screening of ancient microbiomes.
Full-size DOI: 10.7717/peerj.10947/fig-3

and a software container or environment system, e.g., Conda (https://conda.io), Docker (https://www.docker.com) or Singularity (https://sylabs.io), and can be run on both local machines as well as high performance computing (HPC) clusters. Once installed, Nextflow handles all subsequent pipeline-related software dependencies.

The pipeline is primarily a command-line interface (CLI), and therefore users execute the pipeline with a single command, and can specify additional options via parameters and flags. Alternatively, the nf-core initiative offers a graphical user interface (GUI) for this less familiar with CLIs (via https://nf-co.re/launch). Routinely used parameters or institutional configurations can be specified using Nextflow's in-built profile system, an example of which is provided in the basic test data, which can also be used for training and evaluation by new users. This allows specification of many pipeline parameters but with a single option. Specific versions of the pipeline or even specific git hash commits can be supplied as a parameter to ensure analytical reproducibility by other researchers.

The pipeline accepts as input raw short-read FASTQ files, BAM files, or a sample sheet in TSV format that contains extra metadata, and a reference genome in FASTA format. Usage of a TSV file allows processing of more complex sequencing strategies and heterogeneous data types - such as per-run processing of libraries sequenced across different models of sequencing machines, or customised per-library clipping depending on laboratory UDG-treatment (*Rohland et al., 2015*, Fig. 2). Additional files can be supplied for efficiency, such as pre-built reference indices, or annotation files, depending on the modules that wish to be run. Monitoring of pipeline progress is primarily performed via on-console logging.

By default, nf-core/eager performs sequencing QC, adapter clipping (and assuming paired-end data, merging), aligning of the processed reads to the supplied reference genome, PCR duplicate removal and finally aDNA damage and mapping-quality evaluation. If running on a HPC cluster utilising a scheduling system, Nextflow will generate and carry

Table 1 Comparison of pipeline functionality of common ancient DNA processing pipelines.

Category	Functionality	EAGER	PALEOMIX	nf-core/eager
Infrastructure	Software environments	Yes	No	Yes
	HPC scheduler integration	No	No	Yes
	Cloud computing integration	No	No	Yes
	Per-process resource optimisation	No	Partial	Yes
	Pipeline-step parallelisation	No	Yes	Yes
	Command line set up	No	Yes	Yes
	GUI set up	Yes	No	Yes
Preprocessing	Sequencing lane merging	Yes	Yes	Yes
	Sequencing quality control	Yes	No	Yes
	Sequencing artefact removal	No	No	Yes
	Adapter clipping/read merging	Yes	Yes	Yes
	Post-processing sequencing QC	No	No	Yes
Alignment	Reference mapping	Yes	Yes	Yes
	Reference mapping statistics	Yes	Yes	Yes
	Multi-reference mapping	No	Yes	No
Postprocessing	Mapped reads filtering	Yes	Yes	Yes
	Metagenomic complexity filtering	No	No	Yes
	Metagenomic profiling	No	No	Yes
	Metagenomic authentication	No	No	Yes
	Library complexity estimation	Yes	No	Yes
	Duplicate removal	Yes	No	Yes
	BAM merging	No	Yes	Yes
Authentication	Damage read filtering	Yes	No	Yes
	Human contamination estimation	Yes	No	Yes
	Human biological sex determination	No	No	Yes
	Genome coverage estimation	Yes	Yes	Yes
	Damage calculation	Yes	Yes	Yes
	Damage rescaling	No	Yes	Yes
Downstream	SNP calling/genotyping	Yes	Partial	Yes
	Consensus sequence generation	Yes	Partial	Yes
	Regions of interest statistics	Partial	Yes	Yes

out efficient submission strategies of jobs for the user. The pipeline produces a multitude of output files in various file formats, with a more detailed listing available in the user documentation. These include metrics, statistical analysis data, and standardised output files (BAM, VCF) for close inspection and further downstream analysis, as well as a MultiQC report. If an emailing daemon is set up on the server, the latter can be emailed to users automatically.

Benchmarking

Functionality demonstration

To demonstrate the simultaneous genomic analysis of human DNA and metagenomic screening for putative pathogens, as well as improved results reporting, we re-analysed data from *Barquera et al. (2020)* who performed a multi-discipline study of three 16th century individuals excavated from a mass burial site in Mexico City. The authors reported genetic results showing sufficient on-target human DNA (>1%) with typical aDNA damage (>20% C to T reference mismatches in the first base of the 5′ ends of reads) for downstream population-genetic analysis and Y-chromosome coverage indicative that the three individuals were genetically male. In addition, one individual (Lab ID: SJN003) contained DNA suggesting a possible infection by *Treponema pallidum*, a species with a variety of strains that can cause diseases such as syphilis, bejel and yaws, and a second individual (Lab ID: SJN001) displayed reads similar to the Hepatitis B virus. Both results were confirmed by the authors via in-solution enrichment approaches.

Full step-by-step instructions on the setup of the human and pathogen screening demonstration (including input TSV file and final command) can be seen in Data S1. In brief, we replicated the results of *Barquera et al. (2020)* using nf-core/eager v2.2.0 (commit: e7471a7 and Nextflow version: 20.04.1) by simultaneously aligning publicly available shotgun-sequencing reads against the human reference genome (hs37d5) using bwa aln and the off-target reads against the NCBI Nucleotide (nt) database (October 2017 - uploaded here to Zenodo under DOI: 10.5281/zenodo.4382153) with MALT. Alignment parameters were as close to as reported in the original publication, otherwise kept as default. The modified parameter values for pathogen detection were used, rather than nf-core/eager defaults, as these parameters can be highly target-species dependent and must be modified on a per-context basis. Additional modules turned on were mitochondrial-to-nuclear ratio calculation, nuclear contamination estimation, and biological sex determination.

To include the HOPS results from metagenomic screening in the report, we also re-ran MultiQC with the upcoming version v1.10 (to be integrated into nf-core/eager on release), which has an integrated HOPS module. After installing the development version of MultiQC (commit: 7584e64), as described in the MultiQC documentation (https://multiqc.info/), we re-ran the MultiQC command used with the pipeline.

Run-time comparison

We also compared pipeline run-times of two functionally equivalent and previously published pipelines to show that the new implementation of nf-core/eager is equivalent or more efficient than EAGER or PALEOMIX. We ran each pipeline on a subset of Viking-age genomic data of cod (*Gadus morhua*) from *Star et al. (2017)*. This data was originally run using PALEOMIX, and was re-run here as described, but with the latest version of PALEOMIX (v1.2.14), and with equivalent settings for the other two pipelines as close as possible to the original paper (EAGER with v1.92.33, and nf-core/EAGER with v2.2.0, commit 830c22d).

The respective benchmarking environment and exact pipeline run settings can be seen in the Data S1. Two samples each with three Illumina paired-end sequencing runs

were analysed, with adapter clipping and merging (AdapterRemoval), mapping (BWA aln), duplicate removal (Picard's MarkDuplicates) and damage profiling (PALEOMIX: mapDamage2, EAGER and nf-core/EAGER: DamageProfiler) steps being performed. Run-times comparisons were performed on a 32 CPU (AMD Opteron 23xx) and 256 GB memory Red Hat QEMU Virtual Machine running the Ubuntu 18.04 operating system (Linux Kernel 4.15.0-112). Resource parameters of each tool were only modified to specify the maximum available on the server and otherwise left as default. We ran the commands for each tool sequentially, but repeated these batches of commands 10 times - to account for variability in the cloud service's IO connection. Run times were measured using the GNU time tool (v1.7).

RESULTS

Functionality demonstration

We were able to successfully replicate the human and pathogen screening results in a single run of nf-core/eager. Mapping to the human reference genome (hs37d5) with BWA aln and binning of off-target reads with MALT to the NCBI Nucleotide database (2017-10-26), yielded the same results of all individuals having a biological sex of male, as well as the same frequency of C to T miscoding lesions and short mean fragment lengths (both characteristic of true aDNA). Metagenomic hits to both pathogens from the corresponding individuals that yielded complete genomes in the original publication were also detected. Both results and other processing statistics were identified via a single interactive MultiQC report, excerpts of which can be seen in Fig. 4. The full interactive report can be seen in the Data S1.

Run-time comparison

A summary of run-times of the benchmarking tests can be seen in Table 2. nf-core/eager showed fastest run-times across all three time metrics when running on default parameters. This highlights the improved efficiency of nf-core/eager's asynchronous processing system and per-process resource customisation (here represented by nf-core/eager defaults designed for typical HPC cluster setup).

As a more realistic demonstration of modern computing multi-threading setup, we also re-ran PALEOMIX with the flag –max-bwa-threads set to 4 (listed in Table 2 as 'optimised'), which is equivalent to a single BWA aln process of nf-core/eager. This resulted in a much faster run-time than that of default nf-core/eager, due to the approach of PALEOMIX of mapping each lane of a library separately, whereas nf-core/eager will map all lanes of a single library merged together. Therefore, given that each library was split across three lanes, increasing the threads of BWA aln to 4 resulted in 12 per library, whereas nf-core/eager only gave 4 (by default) for a single BWA aln process of one library. While the PALEOMIX approach is valid, we opted to retain the per-library mapping as it is often the longest running step of high-throughput sequencing genome-mapping pipelines, and it prevents flooding of HPC scheduling systems with many long-running jobs. Secondly, if users regularly use multi-lane data, due to nf-core/eager's fine-granularity control, they can simply modify nf-core/eager's BWA aln process resources via config files to account

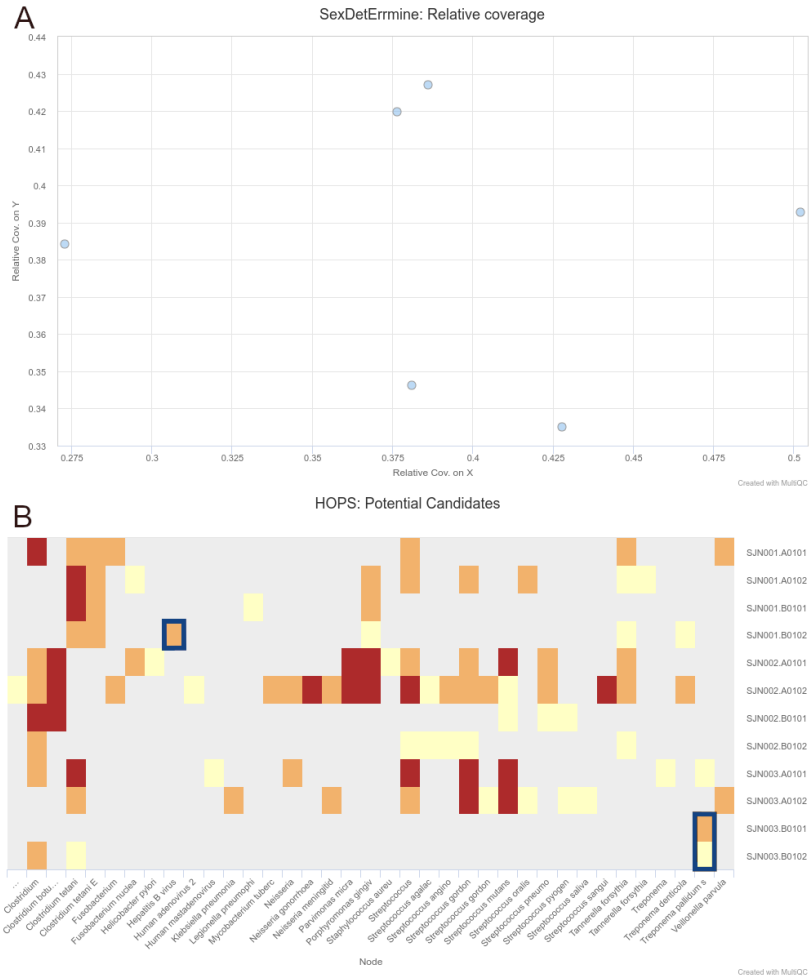

Figure 4 Sections of a MultiQC report (v1.10dev) with the outcome of simultaneous human DNA and microbial pathogen screening with nf-core/eager, including (A) Sex.DetERRmine output of biological sex assignment with coverages on X and Y being half of that of autosomes, indicative of male individuals, and (B) HOPS output with positive detection of both *Treponema pallidum* and Hepatitis B virus reads (indicated with blue boxes). Other taxa in HOPS output represent typical environmental contamination and oral commensal microbiota found in archaeological teeth. Data was Illumina shotgun sequencing data from *Barquera et al. (2020)*, and replicated results here were originally verified in the publication via enrichment methods. The full interactive reports for both MultiQC v1.9 and v1.10 can be seen in Data S1.

Full-size DOI: 10.7717/peerj.10947/fig-4

for this. When we optimised parameters that were used for BWA aln's multi-threading, and the number of multiple lanes to the same number of BWA aln threads as the optimised PALEOMIX run, nf-core/eager again displayed faster run-times (Table 2).

All metrics including mapped reads, percentage on-target, mean depth coverage and mean read lengths across all pipeline and replicates were extremely similar (Table 3).

Table 2 Comparison of run-times in minutes between three ancient DNA pipelines. PALEOMIX and nf-core/eager have additional runs with 'optimised' parameters with fairer computational resources matching modern multi-threading strategies. Values represent mean and standard deviation of run-times in minutes, calculated from the output of the GNU time tool. Real: real time, System: cumulative CPU system-task times, User: cumulative CPU time of all tasks.

Pipeline	Version	Environment	real	sys	user
nf-core-eager (optimised)	2.2.0dev	singularity	105.6 ± 4.6	13.6 ± 0.7	1593 ± 79.7
PALEOMIX (optimised)	1.2.14	conda	130.6 ± 8.7	12 ± 0.7	1820.2 ± 36.9
nf-core-eager	2.2.0dev	singularity	209.2 ± 4.4	11 ± 0.9	1407.7 ± 30.2
EAGER	1.92.37	singularity	224.2 ± 4.9	22.9 ± 0.3	1736.3 ± 70.2
PALEOMIX	1.2.14	conda	314.6 ± 2.9	10.7 ± 1	1506.7 ± 14

Table 3 Comparison of output statistics between three ancient DNA pipelines. Comparison of common result values of key high-throughput short-read data processing and mapping steps across the three pipelines, as reported by the equivalent value of the report from each tool. 'QF' stands for mapping-quality filtered reads. All values represent mean and standard deviation across 10 replicates of each pipeline.

Sample	Category	EAGER	nf-core/eager	PALEOMIX
COD076	Processed Reads	71,388,991 ± 0	71,388,991 ± 0	72,100,142 ± 0
COD092	Processed Reads	69,615,709 ± 0	6,9615,709 ± 0	70,249,181 ± 0
COD076	Mapped QF Reads	16,786,467.7 ± 106.5	16,786,491.1 ± 89.9	16,686,607.2 ± 91.3
COD092	Mapped QF Reads	16,283,216.3 ± 71.3	16,283,194.7 ± 37.4	16,207,986.2 ± 44.4
COD076	Percent QF On-target	23.5 ± 0	23.5 ± 0	23.1 ± 0
COD092	Percent QF On-target	23.4 ± 0	23.4 ± 0	23.1 ± 0
COD076	Deduplicated Reads	12,107,264.4 ± 87.8	12,107,293.7 ± 69.7	12,193,415.8 ± 86.7
COD092	Deduplicated Reads	13,669,323.7 ± 87.6	13,669,328 ± 32.4	13,795,703.3 ± 47.9
COD076	Mean Depth Coverage	0.9 ± 0	0.9 ± 0	0.9 ± 0
COD092	Mean Depth Coverage	1 ± 0	1 ± 0	1 ± 0
COD076	Mean Read Length	49.4 ± 0	49.4 ± 0	49.4 ± 0
COD092	Mean Read Length	48.8 ± 0	48.8 ± 0	48.7 ± 0

DISCUSSION

The re-implementation of EAGER into Nextflow offers a range of benefits over the original custom pipeline framework.

The new framework provides immediate integration of nf-core/eager into various job schedulers in POSIX HPC environments, cloud computing resources, as well as local workstations. This portability allows users to set up nf-core/eager regardless of the type of computing infrastructure or cluster size (if applicable), with minimal effort or configuration. This facilitates reproducibility and therefore maintenance of standards within the field. Portability is further assisted by the in-built compatibility with software environments and containers such as Conda, Docker and Singularity. These are isolated software 'sandbox' environments that include all software (with exact versions) required by the pipeline, in a form that is installable and runnable by users regardless of the setup of their local software environment. Another major change with nf-core/eager is that the primary user interaction mode of a pipeline run setup is now with a CLI, replacing the GUI of the original EAGER pipeline. This is more portable and compatible with most

HPC clusters (that may not offer display of a window system), and is in line with the vast majority of bioinformatics tools. We therefore believe this will not be a hindrance to new researchers from outside computational biology. However, a GUI-based pipeline set up is still available via the nf-core website's Launch page (https://nf-co.re/launch), which provides a common GUI format across multiple pipelines, as well as additional robustness checks of input parameters for those less familiar with CLIs. Typically the output of the launch functionality is a JSON file that can be used with a nf-core/tools launch command as a single parameter (similar to the original EAGER), however integration with Nextflow's companion monitoring tool tower.nf (https://tower.nf) also allows direct submission of pipelines without any command line usage.

Reproducibility is made easier through the use of 'profiles' that can define configuration parameters. These profiles can be managed at different hierarchical levels. *HPC cluster-level profiles* can specify parameters for the computing environment (job schedulers, cache locations for containers, maximum memory and CPU resources etc.), which can be centrally managed to ensure all users of a group use the same settings. *Pipeline-level profiles*, specifying parameters for nf-core/eager itself, allow fast access to routinely-run pipeline parameters via a single flag in the nf-core/eager run command, without having to configure each new run from scratch. Compared to the original EAGER, which utilised per-FASTQ XML files with hardcoded filepaths for a specific user's server, nf-core/eager allows researchers to publish the specific profile used in their runs alongside their publications, which can also be used by other groups to generate the same results. Usage of profiles can also reduce mistakes caused by insufficient 'prose' based reporting of program settings that can be regularly found in the literature. The default nf-core/eager profile uses parameters evaluated in different aDNA-specific contexts (e.g., in *Poullet & Orlando, 2020*), and will be updated in each new release as new studies are published.

nf-core/eager provides improved efficiency over the original EAGER pipeline by replacing sample-by-sample sequential processing with Nextflow's asynchronous job parallelisation, whereby multiple pipeline steps and samples are run in parallel (in addition to natively parallelised pipeline steps). This is similar to the approach taken by PALEOMIX, however nf-core/eager expands this by utilising Nextflow's ability to customise the resource parameters for every job in the pipeline; reducing unnecessary resource allocation that can occur with unfamiliar users to each step of a high-throughput short-read data processing pipeline. This is particularly pertinent given the increasing use of centralised HPC clusters or cloud computing that often use per-hour cost calculations.

Alongside the interactive MultiQC report, we have written extensive documentation on all parts of running and interpreting the output of the pipeline. Given that a large fraction of aDNA researchers come from fields outside computational biology, and thus may have limited computational training, we have written documentation and tutorials (https://nf-co.re/eager/) that also give guidance on how to run the pipeline and interpret each section of the report in the context of high-throughput sequencing data, but with a special focus on aDNA. This includes best practice or expected output schematic images that are published under CC-BY licenses to allow for use in other training material (an example can be seen in Fig. 5). We hope this open-access resource will make the study of

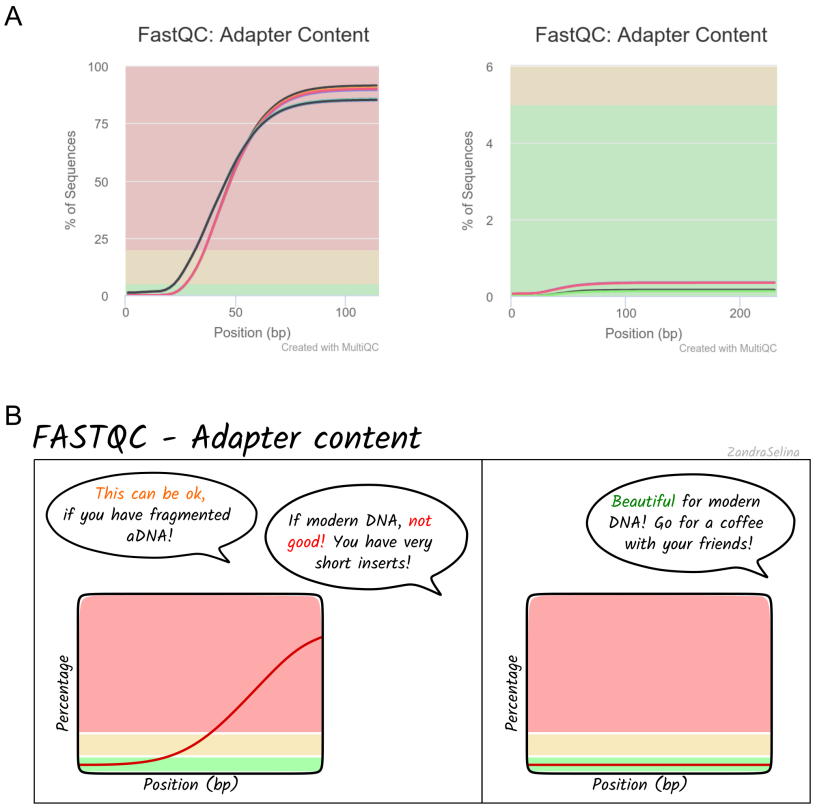

Figure 5 Example schematic image of pipeline output documentation with contextual guidance for aDNA. For each section of the nf-core/eager MultiQC pipeline run report, we provide schematic images that can assist new users in the interpretation of high-throughput sequencing of aDNA. For example, (A) represents MultiQC images of the FastQC report for the amount of sequencing reads with library adapters and (B) is the schematic version counterpart in the nf-core/eager documentation with notes specific for aDNA libraries.

Full-size DOI: 10.7717/peerj.10947/fig-5

aDNA more accessible to researchers new to the field, by providing practical guidelines on how to evaluate characteristics and effects of aDNA on downstream analyses.

The development of nf-core/eager in Nextflow and the nf-core initiative will also improve open-source development, while ensuring the high quality of community contributions to the pipeline. While Nextflow is written primarily in Groovy, the Nextflow DSL simplifies a number of concepts to an intermediate level that bioinformaticians without Java/Groovy experience can easily access (regardless of own programming language experience). Furthermore, Nextflow places ubiquitous and more widely known command-line interfaces, such as bash, in a prominent position within the code, rather than custom Java code and classes (as in EAGER). We hope this will motivate further bug fixes and feature contributions from the community, to keep the pipeline state-of-the-art and ensure a longer life-cycle. This will also be supported by the open and active nf-core community who provide general guidance and advice on developing Nextflow and nf-core pipelines.

It should be noted that the scope of nf-core/eager is as a generic, initial data processing and screening tool, and not to act as a tool for performing more experimental analyses that requires extensive parameter testing such as modelling. As such, while similar pipelines designed for aDNA have also been released, for example ATLAS (*Link et al., 2017*), these generally have been designed with specific contexts in mind (e.g., human population genetics). We therefore have opted to not include common downstream analysis such as Principal Component Analysis for population genetics, or phylogenetic analysis for microbial genomics, but rather focus on ensuring nf-core/eager produces useful files that can be easily used as input for common but more experimental and specialised downstream analysis. Secondly, given nf-core/eager's broad scope of allowing analysis of different target organisms, default parameters of the pipeline are selected as general 'sensible' defaults to account for typical ancient DNA characteristics - and are not necessarily optimised for every use-case. However, the extensive documentation should help researchers decide which parameter values are most suitable for their research.

CONCLUSION

nf-core/eager is an efficient, portable, and accessible pipeline for processing and screening ancient (meta)genomic data. This re-implementation of EAGER into Nextflow and nf-core will improve reproducibility and scalability of rapidly increasing aDNA datasets, for both large and small laboratories. Extensive documentation also enables newcomers to the field to get a practical understanding on how to interpret aDNA in the context of NGS data processing. Ultimately, nf-core/eager provides easier access to the latest tools and routine screening analyses commonly used in the field, and sets up the pipeline for remaining at the forefront of palaeogenetic analysis.

ACKNOWLEDGEMENTS

We thank the nf-core community for general support and suggestions during the writing of the pipeline. We also thank Arielle Munters, Hester van Schalkwyk, Irina Velsko, Katherine Eaton, Luc Venturini, Marcel Keller, Pierre Lindenbaum, Pontus Skoglund, Raphael Eisenhofer, Torsten Günter, Kevin Lord, and Åshild Vågene for bug reports and feature suggestions. We are grateful to the members of the Department of Archaeogenetics at the Max Planck Institute for the Science of Human History who performed beta testing of the pipeline. We thank the aDNA Twitter community for responding to polls regarding design decisions during development. The Gesellschaft für wissenschaftliche Datenverarbeitung (GWDG, Göttingen) kindly provided computational infrastructure for benchmarking. We also want to thank Selina Carlhoff, Maria Spyrou, Elizabeth Nelson, Alexander Herbig and Wolfgang Haak for providing comments and suggestions on this manuscript.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

James A. Fellows Yates, Thiseas C. Lamnidis., Maxime Borry, Aida Andrades Valtueña, Zandra Fagernäs, and Stephen Clayton were supported by the Max Planck Society. James A. Fellows Yates was supported by the ERC Starting Grant project FoodTransforms (ERC-2015-StG 678901-Food-Transforms) funded by the European Research Council awarded to Philipp W. Stockhammer (Ludwig Maximilian University, Munich). Thiseas C. Lamnidis was supported by the European Union's Horizon 2020 research and innovation programme (grant agreement No 851511) with the ERC Starting Grant project MICROSCOPE funded by the European Research Council awarded to Stephan Schiffels (Max Planck Institute for the Science of Human History). Zandra Fagernäs was supported by the Werner Siemens Stiftung funded project 'Paleobiotechnology' awarded to Christina Warinner (Max Planck Institute for the Science of Human History). Maxime U. Garcia is supported by the BarncancerFonden. There was no additional external funding received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Max Planck Society.
ERC Starting Grant project FoodTransforms: ERC-2015-StG 678901-Food-Transforms.
Werner Siemens Stiftung project Paleobiotechnology.
Ludwig Maximilian University, Munich.
European Union's Horizon 2020 Research and Innovation Programme: 851511.
Max Planck Institute for the Science of Human History.
BarncancerFonden.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- James A. Fellows Yates conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Thiseas C. Lamnidis, Maxime Borry, Aida Andrades Valtueña, Maxime U. Garcia and Judith Neukamm performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Zandra Fagernäs performed the experiments, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Stephen Clayton and Alexander Peltzer conceived and designed the experiments, performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

65

Data Availability

The following information was supplied regarding data availability:

All code is available at GitHub: https://github.com/nf-core/eager.

Code is also archived in Zenodo:

James A. Fellows Yates, Alexander Peltzer, Thiseas C. Lamnidis, Maxime Borry, ZandraFagernas, Aida Andrades Valtueña, ...Alex Hübner. (2021, January 14). nf-core/eager: [2.3.1] - Aalen (Patch) - 2021-01-14 (Version 2.3.1). Zenodo. http://doi.org/10.5281/zenodo.4438904.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.10947#supplemental-information.

REFERENCES

Andrades Valtueña A, Mittnik A, Key FM, Haak W, Allmäe R, Belinskij A, Daubaras M, Feldman M, Jankauskas R, Janković I, Massy K, Novak M, Pfrengle S, Reinhold S, Šlaus M, Spyrou MA, Szécsényi-Nagy A, Tõrv M, Hansen S, Bos KI, Stockhammer PW, Herbig A, Krause J. 2017. The stone age plague and its persistence in Eurasia. *Current Biology* **27(23)**:3683–3691.8 DOI 10.1016/j.cub.2017.10.025.

Andrews S. 2010. FastQC: a quality control tool for high throughput sequence data. *Available at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Bah SY, Morangá CM, Kengne-Ouafo JA, Amenga-Etego L, Awandare GA. 2018. Highlights on the application of genomics and bioinformatics in the fight against infectious diseases: Challenges and opportunities in Africa. *Frontiers in Genetics* **9**:575.

Barquera R, Lamnidis TC, Lankapalli AK, Kocher A, Hernández-Zaragoza DI, Nelson EA, Zamora-Herrera AC, Ramallo P, Bernal-Felipe N, Immel A, Bos K, Acuña-Alonzo V, Barbieri C, Roberts P, Herbig A, Kühnert D, Márquez-Morfín L, Krause J. 2020. Origin and health status of first-generation africans from early colonial Mexico. *Current Biology* **30**:2078–2091 DOI 10.1016/j.cub.2020.04.002.

Borry M, Cordova B, Perri A, Wibowo M, Honap TP, Ko J, Yu J, Britton K, Girdland-Flink L, Power RC, Stuijts I, Salazar-García DC, Hofman C, Hagan R, Kagoné TS, Meda N, Carabin H, Jacobson D, Reinhard K, Lewis C, Kostic A, Jeong C, Herbig A, Hübner A, Warinner C. 2020. CoproID predicts the source of coprolites and paleofeces using microbiome composition and host DNA content. *PeerJ* **8**:e9001 DOI 10.7717/peerj.9001.

Bos KI, Harkins KM, Herbig A, Coscolla M, Weber N, Comas I, Forrest SA, Bryant JM, Harris SR, Schuenemann VJ, Campbell TJ, Majander K, Wilbur AK, Guichon RA, Wolfe Steadman DL, Cook DC, Niemann S, Behr MA, Zumarraga M, Bastida R, Huson D, Nieselt K, Young D, Parkhill J, Buikstra JE, Gagneux S, Stone AC, Krause J. 2014. Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514(7523)**:494–497 DOI 10.1038/nature13591.

Bos KI, Schuenemann VJ, Golding GB, Burbano HA, Waglechner N, Coombes BK, McPhee JB, DeWitte SN, Meyer M, Schmedes S, Wood J, Earn DJD, Herring DA, Bauer P, Poinar HN, Krause J. 2011. A draft genome of Yersinia pestis from victims of the Black Death. *Nature* **478(7370)**:506–510 DOI 10.1038/nature10549.

Briggs AW, Stenzel U, Johnson PLF, Green RE, Kelso J, Prüfer K, Meyer M, Krause J, Ronan MT, Lachmann M, Pääbo S. 2007. Patterns of damage in genomic DNA sequences from a Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **104(37)**:14616–14621 DOI 10.1073/pnas.0704665104.

Chen S, Zhou Y, Chen Y, Gu J. 2018. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34(17)**:i884–i890 DOI 10.1093/bioinformatics/bty560.

Daley T, Smith AD. 2013. Predicting the molecular complexity of sequencing libraries. *Nature Methods* **10(4)**:325–327 DOI 10.1038/nmeth.2375.

Damgaard PdB, Marchi N, Rasmussen S, Peyrot M, Renaud G, Korneliussen T, Moreno-Mayar JV, Pedersen MW, Goldberg A, Usmanova E, Baimukhanov N, Loman V, Hedeager L, Pedersen AG, Nielsen K, Afanasiev G, Akmatov K, Aldashev A, Alpaslan A, Baimbetov G, Bazaliiskii VI, Beisenov A, Boldbaatar B, Boldgiv B, Dorzhu C, Ellingvag S, Erdenebaatar D, Dajani R, Dmitriev E, Evdokimov V, Frei KM, Gromov A, Goryachev A, Hakonarson H, Hegay T, Khachatryan Z, Khaskhanov R, Kitov E, Kolbina A, Kubatbek T, Kukushkin A, Kukushkin I, Lau N, Margaryan A, Merkyte I, Mertz IV, Mertz VK, Mijiddorj E, Moiyesev V, Mukhtarova G, Nurmukhanbetov B, Orozbekova Z, Panyushkina I, Pieta K, Smrčka V, Shevnina I, Logvin A, Sjögren K-G, Štolcová T, Tashbaeva K, Tkachev A, Tulegenov T, Voyakin D, Yepiskoposyan L, Undrakhbold S, Varfolomeev V, Weber A, Kradin N, Allentoft ME, Orlando L, Nielsen R, Sikora M, Heyer E, Kristiansen K, Willerslev E. 2018. 137 ancient human genomes from across the Eurasian steppes. *Nature* **557(7705)**:369–374 DOI 10.1038/s41586-018-0094-2.

Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. 2017. Nextflow enables reproducible computational workflows. *Nature Biotechnology* **35(4)**:316–319 DOI 10.1038/nbt.3820.

Ewels P, Magnusson M, Lundin S, Käller M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32(19)**:3047–3048 DOI 10.1093/bioinformatics/btw354.

Ewels PA, Peltzer A, Fillinger S, Patel H, Alneberg J, Wilm A, Garcia MU, Di Tommaso P, Nahnsen S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology* **38**:276–278 DOI 10.1038/s41587-020-0439-x.

Frantz LAF, Haile J, Lin AT, Scheu A, Geörg C, Benecke N, Alexander M, Linderholm A, Mullin VE, Daly KG, Battista VM, Price M, Gron KJ, Alexandri P, Arbogast R-M, Arbuckle B, Bălăşescu A, Barnett R, Bartosiewicz L, Baryshnikov G, Bonsall C, Borić D, Boroneanţ A, Bulatović J, Çakirlar C, Carretero J-M, Chapman J, Church M, Crooijmans R, De Cupere B, Detry C, Dimitrijevic V, Dumitraşcu V, Du Plessis L, Edwards CJ, Erek CM, Erim-Özdoğan A, Ervynck A, Fulgione D, Gligor M, Götherström A, Gourichon L, Groenen MAM, Helmer D, Hongo H, Horwitz LK, Irving-Pease EK, Lebrasseur O, Lesur J, Malone C, Manaseryan N,

Marciniak A, Martlew H, Mashkour M, Matthews R, Matuzeviciute GM, Maziar S, Meijaard E, McGovern T, Megens H-J, Miller R, Mohaseb AF, Orschiedt J, Orton D, Papathanasiou A, Pearson MP, Pinhasi R, Radmanović D, Ricaut F-X, Richards M, Sabin R, Sarti L, Schier W, Sheikhi S, Stephan E, Stewart JR, Stoddart S, Tagliacozzo A, Tasić N, Trantalidou K, Tresset A, Valdiosera C, van den Hurk Y, Van Poucke S, Vigne J-D, Yanevich A, Zeeb-Lanz A, Triantafyllidis A, Gilbert MTP, Schibler J, Rowley-Conwy P, Zeder M, Peters J, Cucchi T, Bradley DG, Dobney K, Burger J, Evin A, Girdland-Flink L, Larson G. 2019. Ancient pigs reveal a near-complete genomic turnover following their introduction to Europe. *Proceedings of the National Academy of Sciences of the United States of America* **116**:17231–17238 DOI 10.1073/pnas.1901169116.

Garrison E, Marth G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv*.

Ginolhac A, Rasmussen M, Gilbert M. TP, Willerslev E, Orlando L. 2011. mapDamage: testing for damage patterns in ancient DNA sequences. *Bioinformatics* **27(15)**:2153–2155 DOI 10.1093/bioinformatics/btr347.

Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, Patterson N, Li H, Zhai W, Fritz MH-Y, Hansen NF, Durand EY, Malaspinas A-S, Jensen JD, Marques-Bonet T, Alkan C, Prufer K, Meyer M, Burbano HA, Good JM, Schultz R, Aximu-Petri A, Butthof A, Hober BH, Öffner B, Siegemund M, Weihmann A, Nusbaum C, Lander ES, Russ C, Novod N, Affourtit J, Egholm M, Verna C, Rudan P, Brajkovic D, Kucan Z, Gušic I, Doronichev VB, Golovanova LV, Lalueza-Fox C, de la Rasilla M, Fortea J, Rosas A, Schmitz RW, Johnson PLF, Eichler EE, Falush D, Birney E, Mullikin JC, Slatkin M, Nielsen R, Kelso J, Lachmann M, Reich D, Paäbo S. 2010. A draft sequence of the neandertal genome. *Science* **328(5979)**:710–722.

Green EJ, Speller CF. 2017. Novel substrates as sources of ancient DNA: prospects and hurdles. *Genes* **8(7)**:180 DOI 10.3390/genes8070180.

Grüning B, Dale R, Sjödin A, Chapman BA, Rowe J, Tomkins-Tinch CH, Valieris R, Köster J, Bioconda Team. 2018. Bioconda: sustainable and comprehensive software distribution for the life sciences. *Nature Methods* **15(7)**:475–476 DOI 10.1038/s41592-018-0046-7.

Gutaker RM, Weiß CL, Ellis D, Anglin NL, Knapp S, Luis Fernández-Alonso J, Prat S, Burbano HA. 2019. The origins and adaptation of European potatoes reconstructed from historical genomes. *Nature Ecology & Evolution* **3(7)**:1093–1101 DOI 10.1038/s41559-019-0921-3.

Herbig A, Maixner F, Bos KI, Zink A, Krause J, Huson DH. 2016. MALT: fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. Preprint DOI 10.1101/050559.

Hübler R, Key FM, Warinner C, Bos KI, Krause J, Herbig A. 2019. HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biology* **20(1)**:280 DOI 10.1186/s13059-019-1903-0.

Jensen TZT, Niemann J, Iversen KH, Fotakis AK, Gopalakrishnan S, Vågene ÅJ, Pedersen MW, Sinding M-HS, Ellegaard MR, Allentoft ME, Lanigan LT, Taurozzi AJ,

Nielsen SH, Dee MW, Mortensen MN, Christensen MC, Sørensen SA, Collins MJ, Gilbert MTP, Sikora M, Rasmussen S, Schroeder H. 2019. A 5700 year-old human genome and oral microbiome from chewed birch pitch. *Nature Communications* **10(1)**:5520 DOI 10.1038/s41467-019-13549-9.

Jónsson H, Ginolhac A, Schubert M, Johnson PLF, Orlando L. 2013. mapDamage2.0: fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29(13)**:1682–1684 DOI 10.1093/bioinformatics/btt193.

Jun G, Wing MK, Abecasis GR, Kang HM. 2015. An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research* **25(6)**:918–925 DOI 10.1101/gr.176552.114.

Kashuba N, Kırdök E, Damlien H, Manninen MA, Nordqvist B, Persson P, Götherström A. 2019. Ancient DNA from mastics solidifies connection between material culture and genetics of mesolithic hunter–gatherers in Scandinavia. *Communications Biology* **2(1)**:185 DOI 10.1038/s42003-019-0399-1.

Kistler L, Ware R, Smith O, Collins M, Allaby RG. 2017. A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic Acids Research* **45(11)**:6310–6320 DOI 10.1093/nar/gkx361.

Korneliussen TS, Albrechtsen A, Nielsen R. 2014. ANGSD: analysis of next generation sequencing data. *BMC bioinformatics* **15**:356 DOI 10.1186/s12859-014-0356-4.

Krause-Kyora B, Susat J, Key FM, Kühnert D, Bosse E, Immel A, Rinne C, Kornell S-C, Yepes D, Franzenburg S, Heyne HO, Meier T, Lösch S, Meller H, Friederich S, Nicklisch N, Alt KW, Schreiber S, Tholey A, Herbig A, Nebel A, Krause J. 2018. Neolithic and Medieval virus genomes reveal complex evolution of Hepatitis B. *eLife* **7**:e36666 DOI 10.7554/eLife.36666.

Lamnidis TC, Majander K, Jeong C, Salmela E, Wessman A, Moiseyev V, Khartanovich V, Balanovsky O, Ongyerth M, Weihmann A, Sajantila A, Kelso J, Pääbo S, Onkamo P, Haak W, Krause J, Schiffels S. 2018. Ancient Fennoscandian genomes reveal origin and spread of Siberian ancestry in Europe. *Nature Communications* **9(1)**:5018 DOI 10.1038/s41467-018-07483-5.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9(4)**:357–359 DOI 10.1038/nmeth.1923.

Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv*.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25(14)**:1754–1760 DOI 10.1093/bioinformatics/btp324.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics* **25(16)**:2078–2079 DOI 10.1093/bioinformatics/btp352.

Lindahl T. 1993. Instability and decay of the primary structure of DNA. *Nature* **362(6422)**:709–715 DOI 10.1038/362709a0.

Link V, Kousathanas A, Veeramah K, Sell C, Scheu A, Wegmann D. 2017. ATLAS: analysis tools for low-depth and ancient samples. *Cold Spring Harbor Laboratory*. preprint DOI 10.1101/105346.

Mathieson I, Alpaslan-Roodenberg S, Posth C, Szécsényi-Nagy A, Rohland N, Mallick S, Olalde I, Broomandkhoshbacht N, Candilio F, Cheronet O, Fernandes D, Ferry M, Gamarra B, Fortes GG, Haak W, Harney E, Jones E, Keating D, Krause-Kyora B, Kucukkalipci I, Michel M, Mittnik A, Nägele K, Novak M, Oppenheimer J, Patterson N, Pfrengle S, Sirak K, Stewardson K, Vai S, Alexandrov S, Alt KW, Andreescu R, Antonović D, Ash A, Atanassova N, Bacvarov K, Gusztáv MB, Bocherens H, Bolus M, Boroneanț A, Boyadzhiev Y, Budnik A, Burmaz J, Chohadzhiev S, Conard NJ, Cottiaux R, Čuka M, Cupillard C, Drucker DG, Elenski N, Francken M, Galabova B, Ganetsovski G, Gély B, Hajdu T, Handzhyiska V, Harvati K, Higham T, Iliev S, Janković I, Karavanić I, Kennett DJ, Komšo D, Kozak A, Labuda D, Lari M, Lazar C, Leppek M, Leshtakov K, Vetro DL, Los D, Lozanov I, Malina M, Martini F, McSweeney K, Meller H, Mendušić M, Mirea P, Moiseyev V, Petrova V, Price TD, Simalcsik A, Sineo L, Šlaus M, Slavchev V, Stanev P, Starović A, Szeniczey T, Talamo S, Teschler-Nicola M, Thevenet C, Valchev I, Valentin F, Vasilyev S, Veljanovska F, Venelinova S, Veselovskaya E, Viola B, Virag C, Zaninović J, Zäuner S, Stockhammer PW, Catalano G, Krauß R, Caramelli D, Zariņa G, Gaydarska B, Lillie M, Nikitin AG, Potekhina I, Papathanasiou A, Borić D, Bonsall C, Krause J, Pinhasi R, Reich D. 2018. The genomic history of southeastern Europe. *Nature* **555(7695)**:197–203 DOI 10.1038/nature25778.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* **20(9)**:1297–1303 DOI 10.1101/gr.107524.110.

Meyer M, Kircher M, Gansauge M-T, Li H, Racimo F, Mallick S, Schraiber JG, Jay F, Prufer K, de Filippo C, Sudmant PH, Alkan C, Fu Q, Do R, Rohland N, Tandon A, Siebauer M, Green RE, Bryc K, Briggs AW, Stenzel U, Dabney J, Shendure J, Kitzman J, Hammer MF, Shunkov MV, Derevianko AP, Patterson N, Andres AM, Eichler EE, Slatkin M, Reich D, Kelso J, Paäbo S. 2012. A high-coverage genome sequence from an archaic denisovan individual. *Science* **338(6104)**:222–226.

Meyer M, Arsuaga J-L, De Filippo C, Nagel S, Aximu-Petri A, Nickel B, Martínez I, Gracia A, Bermúdez de Castro JM, Carbonell E, Viola B, Kelso J, Prüfer K, Pääbo S. 2016. Nuclear DNA sequences from the Middle Pleistocene Sima de los Huesos hominins. *Nature* **531(7595)**:504–507 DOI 10.1038/nature17405.

Mühlemann B, Jones TC, Damgaard PdB, Allentoft ME, Shevnina I, Logvin A, Usmanova E, Panyushkina IP, Boldgiv B, Bazartseren T, Tashbaeva K, Merz V, Lau N, Smrčka V, Voyakin D, Kitov E, Epimakhov A, Pokutta D, Vicze M, Price TD, Moiseyev V, Hansen AJ, Orlando L, Rasmussen S, Sikora M, Vinner L, Osterhaus ADME, Smith DJ, Glebe D, Fouchier RAM, Drosten C, Sjögren K-G, Kristiansen K, Willerslev E. 2018. Ancient hepatitis B viruses from the bronze age to the medieval period. *Nature* **557(7705)**:418–423 DOI 10.1038/s41586-018-0097-z.

Namouchi A, Guellil M, Kersten O, Hänsch S, Ottoni C, Schmid BV, Pacciani E, Quaglia L, Vermunt M, Bauer EL, Derrick M, Jensen AØ, Kacki S, Cohn Jr SK, Stenseth NC, Bramanti B. 2018. Integrative approach using Yersinia pestis genomes to revisit the historical landscape of plague during the Medieval Period. *Proceedings of the National Academy of Sciences of the United States of America* **115(50)**:E11790–E11797 DOI 10.1073/pnas.1812865115.

Neukamm J, Peltzer A, Nieselt K. 2020. DamageProfiler: fast damage pattern calculation for ancient DNA. preprint DOI 10.1101/2020.10.01.322206.

Okonechnikov K, Conesa A, García-Alcalde F. 2016. Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data. *Bioinformatics* **32(2)**:292–294 DOI 10.1093/bioinformatics/btv566.

Olalde I, Brace S, Allentoft ME, Armit I, Kristiansen K, Booth T, Rohland N, Mallick S, Szécsényi-Nagy A, Mittnik A, Altena E, Lipson M, Lazaridis I, Harper TK, Patterson N, Broomandkhoshbacht N, Diekmann Y, Faltyskova Z, Fernandes D, Ferry M, Harney E, De Knijff P, Michel M, Oppenheimer J, Stewardson K, Barclay A, Alt KW, Liesau C, Ríos P, Blasco C, Miguel JV, García RM, Fernández AA, Bánffy E, Bernabò-Brea M, Billoin D, Bonsall C, Bonsall L, Allen T, Büster L, Carver S, Navarro LC, Craig OE, Cook GT, Cunliffe B, Denaire A, Dinwiddy KE, Dodwell N, Ernée M, Evans C, Kuchařík M, Farré JF, Fowler C, Gazenbeek M, Pena RG, Haber-Uriarte M, Haduch E, Hey G, Jowett N, Knowles T, Massy K, Pfrengle S, Lefranc P, Lemercier O, Lefebvre A, Martínez CH, Olmo VG, Ramírez AB, Maurandi JL, Majó T, McKinley JI, McSweeney K, Mende BG, Modi A, Kulcsár G, Kiss V, Czene A, Patay R, Endrődi A, Köhler K, Hajdu T, Szeniczey T, Dani J, Bernert Z, Hoole M, Cheronet O, Keating D, Velemínský P, Dobeš M, Candilio F, Brown F, Fernández RF, Herrero-Corral A.-M, Tusa S, Carnieri E, Lentini L, Valenti A, Zanini A, Waddington C, Delibes G, Guerra-Doce E, Neil B, Brittain M, Luke M, Mortimer R, Desideri J, Besse M, Brücken G, Furmanek M, Hałuszko A, Mackiewicz M, Rapiński A, Leach S, Soriano I, Lillios KT, Cardoso JL, Pearson MP, Włodarczak P, Price TD, Prieto P, Rey P-J, Risch R, Rojo Guerra MA, Schmitt A, Serralongue J, Silva AM, Smrčka V, Vergnaud L, Zilhão J, Caramelli D, Higham T, Thomas MG, Kennett DJ, Fokkens H, Heyd V, Sheridan A, Sjögren K.-G, Stockhammer PW, Krause J, Pinhasi R, Haak W, Barnes I, Lalueza-Fox C, Reich D. 2018. The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature* **555(7695)**:190–196 DOI 10.1038/nature25738.

Orlando L, Ginolhac A, Zhang G, Froese D, Albrechtsen A, Stiller M, Schubert M, Cappellini E, Petersen B, Moltke I, Johnson PLF, Fumagalli M, Vilstrup JT, Raghavan M, Korneliussen T, Malaspinas A-S, Vogt J, Szklarczyk D, Kelstrup CD, Vinther J, Dolocan A, Stenderup J, Velazquez AMV, Cahill J, Rasmussen M, Wang X, Min J, Zazula GD, Seguin-Orlando A, Mortensen C, Magnussen K, Thompson JF, Weinstock J, Gregersen K, Røed KH, Eisenmann V, Rubin CJ, Miller DC, Antczak DF, Bertelsen MF, Brunak S, Al-Rasheid KAS, Ryder O, Andersson L, Mundy J, Krogh A, Gilbert M. TP, Kjær K, Sicheritz-Ponten T, Jensen LJ, Olsen JV, Hofreiter M, Nielsen R[71], Shapiro B, Wang J, Willerslev E. 2013. Recalibrating Equus

evolution using the genome sequence of an early Middle Pleistocene horse. *Nature* **499**(**7456**):74–78 DOI 10.1038/nature12323.

Palkopoulou E, Mallick S, Skoglund P, Enk J, Rohland N, Li H, Omrak A, Vartanyan S, Poinar H, Götherström A, Reich D, Dalén L. 2015. Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology* **25**(**10**):1395–1400 DOI 10.1016/j.cub.2015.04.007.

Peltzer A, Jäger G, Herbig A, Seitz A, Kniep C, Krause J, Nieselt K. 2016. EAGER: efficient ancient genome reconstruction. *Genome Biology* **17**(**1**):1–14 DOI 10.1186/s13059-016-0918-z.

Poullet M, Orlando L. 2020. Assessing DNA sequence alignment methods for characterizing ancient genomes and methylomes. *Frontiers in Ecology and Evolution* **8**:105 DOI 10.3389/fevo.2020.00105.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**(**6**):841–842 DOI 10.1093/bioinformatics/btq033.

Rasmussen S, Allentoft ME, Nielsen K, Orlando L, Sikora M, Sjögren K.-G, Pedersen AG, Schubert M, Van Dam A, Kapel C. MO, Nielsen HB, Brunak S, Avetisyan P, Epimakhov A, Khalyapin MV, Gnuni A, Kriiska A, Lasak I, Metspalu M, Moiseyev V, Gromov A, Pokutta D, Saag L, Varul L, Yepiskoposyan L, Sicheritz-Pontén T, Foley RA, Lahr MM, Nielsen R, Kristiansen K, Willerslev E. 2015. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. *Cell* **163**(**3**):571–582 DOI 10.1016/j.cell.2015.10.009.

Renaud G, Slon V, Duggan AT, Kelso J. 2015. Schmutzi: estimation of contamination and endogenous mitochondrial consensus calling for ancient DNA. *Genome Biology* **16**(**1**):224 DOI 10.1186/s13059-015-0776-0.

Rohland N, Harney E, Mallick S, Nordenfelt S, Reich D. 2015. Partial uracil-DNA-glycosylase treatment for screening of ancient DNA. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370**(**1660**):20130624 DOI 10.1098/rstb.2013.0624.

Schubert M, Ermini L, Der Sarkissian C, Jónsson H, Ginolhac A, Schaefer R, Martin MD, Fernández R, Kircher M, McCue M, Willerslev E, Orlando L. 2014. Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature Protocols* **9**(**5**):1056–1082 DOI 10.1038/nprot.2014.063.

Schubert M, Lindgreen S, Orlando L. 2016. AdapterRemoval v2: rapid adapter trimming, identification, and read merging. *BMC Research Notes* **9**:88 DOI 10.1186/s13104-016-1900-2.

Schuenemann VJ, Avanzi C, Krause-Kyora B, Seitz A, Herbig A, Inskip S, Bonazzi M, Reiter E, Urban C, Dangvard Pedersen D, Taylor GM, Singh P, Stewart GR, Velemínský P, Likovsky J, Marcsik A, Molnár E, Pálfi G, Mariotti V, Riga A, Belcastro MG, Boldsen JL, Nebel A, Mays S, Donoghue HD, Zakrzewski S, Benjak A, Nieselt K, Cole ST, Krause J. 2018. Ancient genomes reveal a high diversity of Mycobacterium leprae in medieval Europe. *PLOS Pathogens* **14**(**5**):e1006997 DOI 10.1371/journal.ppat.1006997.

Skoglund P, Northoff BH, Shunkov MV, Derevianko AP, Pääbo S, Krause J, Jakobsson M. 2014. Separating endogenous ancient DNA from modern day contamination in a Siberian Neandertal. *Proceedings of the National Academy of Sciences of the United States of America* **111(6)**:2229–2234 DOI 10.1073/pnas.1318934111.

Slon V, Hopfe C, Weiß CL, Mafessoni F, De la Rasilla M, Lalueza-Fox C, Rosas A, Soressi M, Knul MV, Miller R, Stewart JR, Derevianko AP, Jacobs Z, Li B, Roberts RG, Shunkov MV, de Lumley H, Perrenoud C, Gušić I, Kućan Ž, Rudan P, Aximu-Petri A, Essel E, Nagel S, Nickel B, Schmidt A, Prüfer K, Kelso J, Burbano HA, Pääbo S, Meyer M. 2017. Neandertal and Denisovan DNA from Pleistocene sediments. *Science* **356(6338)**:605–608 DOI 10.1126/science.aam9695.

Slon V, Mafessoni F, Vernot B, de Filippo C, Grote S, Viola B, Hajdinjak M, Peyrégne S, Nagel S, Brown S, Douka K, Higham T, Kozlikin MB, Shunkov MV, Derevianko AP, Kelso J, Meyer M, Prüfer K, Pääbo S. 2018. The genome of the offspring of a neanderthal mother and a denisovan father. *Nature* **561(7721)**:113–116 DOI 10.1038/s41586-018-0455-x.

Star B, Boessenkool S, Gondek AT, Nikulina EA, Hufthammer AK, Pampoulie C, Knutsen H, André C, Nistelberger HM, Dierking J, Petereit C, Heinrich D, Jakobsen KS, Stenseth NC, Jentoft S, Barrett JH. 2017. Ancient DNA reveals the Arctic origin of Viking Age cod from Haithabu, Germany. *Proceedings of the National Academy of Sciences of the United States of America* **114(34)**:9152–9157 DOI 10.1073/pnas.1710186114.

Tastan Bishop Ö, Adebiyi EF, Alzohairy AM, Everett D, Ghedira K, Ghouila A, Kumuthini J, Mulder NJ, Panji S, Patterton H-G, H3ABioNet Consortium, H3Africa Consortium. 2015. Bioinformatics education–perspectives and challenges out of Africa. *Briefings in Bioinformatics* **16(2)**:355–364 DOI 10.1093/bib/bbu022.

Teasdale MD, Van Doorn NL, Fiddyment S, Webb CC, O'Connor T, Hofreiter M, Collins MJ, Bradley DG. 2015. Paging through history: parchment as a reservoir of ancient DNA for next generation sequencing. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **370(1660)**:20130379 DOI 10.1098/rstb.2013.0379.

Tett A, Huang KD, Asnicar F, Fehlner-Peach H, Pasolli E, Karcher N, Armanini F, Manghi P, Bonham K, Zolfo M, De Filippis F, Magnabosco C, Bonneau R, Lusingu J, Amuasi J, Reinhard K, Rattei T, Boulund F, Engstrand L, Zink A, Collado MC, Littman DR, Eibach D, Ercolini D, Rota-Stabelli O, Huttenhower C, Maixner F, Segata N. 2019. The prevotella copri complex comprises four distinct clades underrepresented in westernized populations. *Cell Host & Microbe* **26(5)**:666–679.e7 DOI 10.1016/j.chom.2019.08.018.

Vågene ÅJ, Herbig A, Campana MG, Robles García NM, Warinner C, Sabin S, Spyrou MA, Andrades Valtueña A, Huson D, Tuross N, Bos KI, Krause J. 2018. Salmonella enterica genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution* **2(3)**:520–528 DOI 10.1038/s41559-017-0446-6.

Van Dorp L, Gelabert P, Rieux A, de Manuel M, de Dios T, Gopalakrishnan S, Carøe C, Sandoval-Velasco M, Fregel R, Olalde I, Escosa R, Aranda C, Huijben S, Mueller I,

Marquès-Bonet T, Balloux F, Gilbert MTP, Lalueza-Fox C. 2019. Plasmodium vivax Malaria viewed through the lens of an eradicated European strain. *Molecular Biology and Evolution* **37**:773–785 DOI 10.1093/molbev/msz264.

Velsko IM, Fellows Yates JA, Aron F, Hagan RW, Frantz LAF, Loe L, Martinez JBR, Chaves E, Gosden C, Larson G, Warinner C. 2019. Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome* **7**(**1**):102 DOI 10.1186/s40168-019-0717-3.

Wales N, Akman M, Watson R. HB, Sánchez Barreiro F, Smith BD, Gremillion KJ, Gilbert M. TP, Blackman BK. 2019. Ancient DNA reveals the timing and persistence of organellar genetic bottlenecks over 3,000 years of sunflower domestication and improvement. *Evolutionary Applications* **12**(**1**):38–53 DOI 10.1111/eva.12594.

Warinner C, Herbig A, Mann A, Fellows Yates JA, Weiß CL, Burbano HA, Orlando L, Krause J. 2017. A robust framework for microbial archaeology. *Annual Review of Genomics and Human Genetics* **18**:321–356 DOI 10.1146/annurev-genom-091416-035526.

Warinner C, Rodrigues J. FM, Vyas R, Trachsel C, Shved N, Grossmann J, Radini A, Hancock Y, Tito RY, Fiddyment S, Speller C, Hendy J, Charlton S, Luder HU, Salazar-García DC, Eppler E, Seiler R, Hansen LH, Castruita JAS, Barkow-Oesterreicher S, Teoh KY, Kelstrup CD, Olsen JV, Nanni P, Kawai T, Willerslev E, von Mering C, LewisJr CM, Collins MJ, Gilbert MTP, Rühli F, Cappellini E. 2014. Pathogens and host immunity in the ancient human oral cavity. *Nature Genetics* **46**(**4**):336–344 DOI 10.1038/ng.2906.

Weyrich LS, Duchene S, Soubrier J, Arriola L, Llamas B, Breen J, Morris AG, Alt KW, Caramelli D, Dresely V, Farrell M, Farrer AG, Francken M, Gully N, Haak W, Hardy K, Harvati K, Held P, Holmes EC, Kaidonis J, Lalueza-Fox C, de la Rasilla M, Rosas A, Semal P, Soltysiak A, Townsend G, Usai D, Wahl J, Huson DH, Dobney K, Cooper A. 2017. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* **544**(**7650**):357–361 DOI 10.1038/nature21674.

Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, Lorenzen ED, Vestergård M, Gussarova G, Haile J, Craine J, Gielly L, Boessenkool S, Epp LS, Pearman PB, Cheddadi R, Murray D, Bråthen KA, Yoccoz N, Binney H, Cruaud C, Wincker P, Goslar T, Alsos IG, Bellemain E, Brysting AK, Elven R, Sønstebø JH, Murton J, Sher A, Rasmussen M, Rønn R, Mourier T, Cooper A, Austin J, Möller P, Froese D, Zazula G, Pompanon F, Rioux D, Niderkorn V, Tikhonov A, Savvinov G, Roberts RG, MacPhee RDE, Gilbert MTP, Kjær KH, Orlando L, Brochmann C, Taberlet P. 2014. Fifty thousand years of Arctic vegetation and megafaunal diet. *Nature* **506**(**7486**):47–51 DOI 10.1038/nature12921.

Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. *Genome Biology* **20**(**1**):257 DOI 10.1186/s13059-019-1891-0.

6 Discussion

In this thesis I have demonstrated that ancient microbiome data can be used to increase our understanding of the taxonomic and functional evolution of the human oral microbiome, as well as add new perspectives on the behavioural history of our species (Manuscript A; Fellows Yates et al., 2021c). In response to the issues encountered during Manuscript A, I have also described the formation of a sustainable directory of all known openly available ancient metagenomic data, as a solution for the efficient compilation of comparative datasets and meta-analyses via public data reuse (Manuscript B; Fellows Yates et al., 2021a). In addition to the novel authentication and post-alignment analysis workflow for ancient microbiomes developed in Manuscript A, I describe the complete reimplementation and metagenomic extension of a dedicated aDNA pipeline that follows latest bioinformatics and palaeogenomics best practises (Manuscript C; Fellows Yates et al., 2021b). Together, Manuscript B and C will ultimately assist in the scaling of ancient metagenomics data processing to sample sizes needed for future studies.

The following discussion will describe how the three manuscripts achieve the overall objectives of this thesis, and the contributions of these manuscripts to the field of ancient metagenomics. The discussion is structured in a manner whereby the major phases of a standard ancient metagenomics study are covered. I will suggest how the outcomes of Manuscript A can be used to define future research directions in both the modern and ancient oral microbiome fields, and additionally how the resources from Manuscripts B and C will facilitate this at different stages of such a project, as well as for the wider ancient metagenomics community.

6.1 Research directions for future ancient dental calculus microbiome research

6.1.1 Oral biofilm conservation, diversity, and disease

Manuscript A describes the first attempt to understand deep-timescale ecological and genomic evolution of the human oral microbiome. Utilising aDNA samples from both living and archaeological modern humans, Neanderthals, chimpanzees, gorillas, and howler monkeys, we made the first observations into taxonomic and functional similarities and differences in the dental calculus of these related host taxa. While the focus of this thesis is primarily on modern humans, we decided to use these divergent host groups with the hypothesis that if the oral microbiome is highly sensitive to both host genomic and behavioural or cultural changes, these differences would be very clear when comparing evolutionary diverged species. This would therefore allow us to build on previous smaller-scale palaeogenomic research that has reported differences in modern human oral microbiomes in relation to major dietary shifts (Adler et al., 2013; Weyrich et al., 2017).

Core microbiome analysis comparing the microbial taxa that are found to be shared across all the different host genera, showed a surprising level of conservation of oral microbiota genera. This is true even across the estimated 40 mya years of evolution between the hosts, i.e., since the most recent common ancestor of primates and howler monkeys. While the similarity of the major 'players' of oral taxa across geographically diverse present-day populations has already been noted (Mark Welch et al., 2019), the core microbiome analysis in Manuscript A shows that this consistency also extends deep along a temporal axis. This conservation suggests that

there is sufficient stability in the oral microbiome to make reconstruction of ancestral states of the human oral microbiota possible, and also that the structure of the oral biofilm has been maintained with a high degree of continuity, despite significant host genomic and behavioural evolution. Importantly, many of the identified core genera have very few published reference genomes, or represent unnamed taxa (such as members of *Eubacterium*, Anaerolinaceae, or *Gemella*). Hierarchical clustering of the host genera based on their calculus microbial profiles showed that, for example, the species making up the *Streptococcus* genus-level 'core' assignment for gorilla and howler monkey was highly diverse, with more than 19 unique *Streptococcus* species identified to that group, compared to chimpanzees and humans. The results from the core analysis provide a potential guidance for future dental microbiology research as to which genera should be targeted for isolation, genome sequencing, and further characterisation of their relationships and roles in healthy oral biofilms. The identified taxa also represent good targets for the exploration of *de novo* assembly of aDNA (for further discussion, see section 6.4). These methods allow for reference-free genome reconstruction from ancient or nonhuman individuals of unrepresented, or now extinct, taxa that are not present in current databases, and therefore opens the possibility of the reconstruction of extinct diversity that would otherwise not be discoverable from living populations. As there was surprising consistency across individuals, it suggests that the human oral biofilm microbiome may have a possible 'ancestral' beginning state and stability that could act as a useful 'model biofilm' to understand how long-term influences of environmental factors may have influenced microbiome evolution. This is compared to microbiomes such as the gut, that would be more difficult to reconstruct due to the greater sensitivity or 'malleability' of the gut to short-term environmental factors.

A common misconception in studies analysing ancient dental calculus is that the presence of taxa associated with Socransky *et al.*'s 'red complex' bacteria (see section 1.2.4 for introduction), or even the presence of dental calculus in the first place, is indicative that the individual was suffering from some form of oral disease. This misconception is likely due to early and current palaeogenomic microbial research focusing on, and being successful with, the identification of well established pathogens such as *Yersinia pestis* (Perry and Fetherston, 1997; Spyrou et al., 2019). Furthermore, early oral biofilm literature such as the Non-Specific Plaque Hypothesis suggested just the presence and abundance of an oral biofilm (plaque) resulted in disease (see section 1.2.3). Indeed, more specific comparisons between plaque and dental calculus have demonstrated that ancient dental calculus appears to have these taxa almost ubiquitously, even in cases of individuals without (major) oral disease (Warinner et al., 2014a; Velsko et al., 2019). Expanding on this, we observed that these three main genera: *Tannerella*, *Treponema*, and *Porphyromonas*, can be considered core even across African hominids - and therefore should not necessarily be considered disease-causing 'pathogens' as commonly assumed. While gorilla and howler monkey individuals had much weaker signals (fewer reads) of the presence of these taxa, they were sufficiently prevalent enough to suggest that purely identifying the presence of these taxa alone is likely not a sufficient indication of a disease state. Furthermore, this confirms studies asserting plaque and dental calculus should not be directly compared, as they represent different biofilm states (Warinner et al., 2014a; Velsko et al., 2019). Importantly - it should be noted that the assignment of these pathobiont taxa as core was at genus level. While the specific *T. forsythia*, *T. denticola*, and *P. gingivalis* red complex taxa were highly prevalent and abundant in chimpanzees and humans, in gorillas and howler monkeys, reads to these genera were again

more highly spread over different species, and also with different species than for humans and chimpanzees. This again suggests the possible existence of a much higher diversity of these genera than assumed by clinicians and ancient microbiome researchers. For example, the number of genomes of both named and uncharacterised *Tannerella* strains archived on the NCBI Genome repository (`https://www.ncbi.nlm.nih.gov/genome/`) has doubled throughout this PhD (since 2015). More targeted efforts to separate strains in ancient samples is required (an outstanding challenge, see section 6.4) to improve the characterisation of the diversity of these important genera, which may allow for a more nuanced understanding of the role of these taxa in periodontal disease.

Indeed - when checking whether known virulence factors in *T. forsythia* and *P. gingivalis* were present within each of the host genera in Manuscript A, the genomic depth coverage of these genes were often found to be equal to the rest of the genome in at least one individual of each host genus. This suggests that these virulence factors are present or 'core' to all host genera, and are much more widespread than being limited to human periodontal disease. One limitation of Manuscript A, however, was that only minimal metadata, if any at all, regarding periodontal health status was available for most individuals. Therefore no direct correlation analysis could be performed either between disease state and presence or abundance of pathobionts, and nor of virulence factors. Retroactive assessment and improved sampling procedures for future projects may improve and allow for this type of analysis in the future.

Nonetheless, given the prevalence of both taxa and genes, these results support more recent theoretical concepts of oral biofilm-associated diseases in terms of the 'Ecological Plaque Hypothesis' (section 1.2.5). This hypothesis posits that it is a dysbiosis or disequilibrium in the *relationships* between taxa that results in disease rather than purely presence. This suggests that these pathobionts and known virulence factors were present in the oral biofilms of hominids deep in their evolution, and are therefore not necessarily human specific. Further research to understand what roles these genes play in the functioning of the wider biofilm will be necessary, however. Additionally, capturing the wider diversity of taxa that modulates the initiation of dysbiotic disease is still required. The analysis in Manuscript A consisted entirely of reference-based methods, which only allow insight into already known taxonomic and functional diversity. This is partly due to early attempts at true *de novo* reconstruction of bacterial genomes from intrinisically metagenomic ancient DNA data resulted in comparatively low-quality assemblies (e.g. Schuenemann et al., 2013). Only recently, with the development of dedicated shotgun metagenomic assemblers, have these approaches begun to be revisited by palaeogenomicists (Wibowo et al., 2021; Borry et al., 2021). Efforts should therefore be made by ancient metagenomicists to use reference-free methods to identify not only undercharacterised taxa, as shown in the core analysis, but also to identify completely unknown taxa that may now be extinct (see section 6.5), and that may harbour information about past microbial mechanisms for maintaining and changing states of biofilm ecology.

6.1.2 Learning about human history from oral microbiomes

In terms of using ancient dental calculus to infer behavioural and cultural change, the observation in Manuscript A of functional rather than taxonomic differences in the oral microbiota represents a major difference in regards to previous assumptions and reports from this area of

research. First, after analysis with a larger sample and data size, and a more balanced sampling strategy, I was unable to replicate the results of Adler et al. (2013) and Weyrich et al. (2017), both of which suggested that the taxonomic profiles of calculus biofilms can differ depending on general subsistence strategies (e.g., hunter-gathering cultures versus farming cultures). Our results rather reflect modern oral microbiome studies that appear to show that any dietary influence on both saliva and biofilm taxonomic profiles is generally short-lived, with the majority of taxa reverting to 'normal' diet profiles after the period of enforced dietary change (De Filippis et al., 2014; Anderson et al., 2020). While the results from Anderson et al. (2020) possibly suggest that with major and sustained dietary change, taxonomic profiles may be modified, these are generally quite subtle, and much larger ancient sample sizes would be required than currently available for both pre- and post dietary transition periods to detect this (such as for the Neolithic revolution and increase in carbohydrate rich diets).

This was demonstrated in Manuscript A with the identification of *Streptococcus* amylase-binding protein-related genes in Neanderthals and Humans to the exclusion of nonhuman primates. This observation appears to correspond with the hypothesised increase of carbohydrate consumption occurring along the hominin lineage, facilitated by behaviour changes (e.g., cooking), that resulted in the development of larger brains (Perry et al., 2007; Carmody and Wrangham, 2009; Carmody et al., 2011). Functional change in response to behavioural changes can also be seen in modern microbiome studies, such as by Nearing et al. (2020), who observed a range of carbohydrate-related degradation pathways in the saliva microbiota being reduced with increased refined grain intake. This reflects adaptation of the relationships between biofilm taxa to different environmental conditions to ensure the survival of the biofilm as a whole, rather than requiring entire taxonomic change of the biofilm itself. Palaeogenomicists should instead consider routinely analysing ancient dental calculus from a functional perspective alongside taxonomic analysis. For example, expanding ancient dental calculus sample sets to a more diverse set of ancient human populations, and possibly to other extinct human species, may allow molecular dating of the rise of amylase-binding ability in *Streptococcus*. In turn, this would possibly allow inferences of when such behavioural changes occurred in the past. Ultimately, while certain species or strains can act as 'markers' for certain behaviours, such as if they uniquely hold specific genes related to specific functions, given the remaining palaeogenomic issues of cross-mapping and strain separation of taxonomically rich but low coverage data (see section 6.4), direct characterisation of functional profiles as a primary analysis type might be a more useful endeavour. However, one limiting factor here is that while genome sequencing is now comparatively cheap and simple - resulting in thousands of genome sequences - the resulting assemblies are often released at various level of completion, with many not being published with associated genomic annotations (Lobb et al., 2020). Without further investment in finalising assemblies, such metagenomic functional analysis derived from annotated gene contents will be restricted.

Finally, Manuscript A has also demonstrated that phylogenetic tree reconstruction of multiple different oral microbiota species could offer another avenue into the understanding of the history and interactions between the human hosts themselves. I observed that across microbial genome-level phylogenies of 8 different taxa, that the Upper Palaeolithic individual from El Mirón (and also from Pavlov and a Mesolithic individual from Rigney - when analysing the shallow sequencing dataset) consistently fell with Neanderthal individuals, rather than other

modern humans. Curiously, this appears to match human genomic results indicative of a population bottleneck and ancestry turnover in Europe around 14 kya BP (Fu et al., 2015; Hajdinjak et al., 2021; Prüfer et al., 2021). Improving genomic reconstruction methods, as well developing more targeted studies with expanded datasets will allow us to evaluate the possibility of using ancient oral microbiomes as a proxy for host relationships, when host DNA is not available. Furthermore, due to the general inheritance of oral microbiota being derived from *all* caregivers to offspring and not just from direct parental relationships, strain-resolved ancient metagenomes may also allow fine-resolution observations into non direct-familial relationships - possibly even between different hominin populations (as observed in the host genomics of an offspring of Denisovan and Neanderthal parents by Slon et al., 2018).

During the review process of Manuscript A, two other studies were published that also attempted to use phylogenetics of oral microbiota as a proxy for host population history (Eisenhofer et al., 2020; Bravo-Lopez et al., 2020). An issue with these studies is that these analyses did not check for cross-mapping during phylogenetic reconstruction. These studies utilised methods that assume the use of high-quality variant calls that are typically generated from modern data, something that is an issue when applying to the comparatively 'low-quality' degraded nature of aDNA. This is particularly important in the case of Eisenhofer et al. (2020) who compared Jomon and Edo-era Japanese individuals and used an unnamed species as their target organism (*Anaerolineaceae* oral taxon 439). The use of this species is problematic as the diversity of this genus within the oral cavity is not yet characterised. Indeed, as of August 2021, the reference genome used by Eisenhofer et al. is flagged as problematic by the NCBI Genome database, possibly being derived from an environmental relative. In the same vein, Bravo-Lopez et al. (2020) mapped against the genome of *T. forsythia*, a genus that that, as described above (section 6.1.1), is increasingly being shown to have higher diversity than previously known.

Analysis of cross-mapping in Manuscript A identified that many of the human *T. forsythia* mappings had high levels of multi-allelic positions, suggesting that the phylogenies generated from these mappings represent bacterial 'populations' rather than isolates. Accordingly, we took a more conservative approach in analysing and reporting the results by calling variants with lower majority call thresholds, We then used simpler phylogenetic clustering algorithms to allow for the studying of such microbial 'populations', rather than analysing specifically as species-level phylogenies, and reporting as such. Furthermore, to add more support to our phylogenetic results, rather than analysing a single phylogeny from a single species, we generated trees using multiple different species reference genomes. By doing so, we were able to be more confidently report a possible reflection of host population history across independent microbial species 'populations', rather than the results being possibly derived from artefacts from cross-mapping. An alternative explanation of the reported functional differences between the reported taxa in (Bravo-Lopez et al., 2020), where pre-Hispanic strains were missing many antiobiotic resistance genes compared to later colonial ones, could be that the reconstructed genomes were of close-relatives of *T. forsythia*, rather than *T. forsythia* itself. That said, an absence of specifically *T. forsythia* in pre-Hispanic individuals would in itself be an interesting finding to further explore, given the perceived widespread presence across global present-day populations. Ultimately, methodological improvements are required to improve the genomic depth coverage of oral microbiota to levels sufficient for reliable variant calling. For example, very deep sequencing of a subset of samples was required for Manuscript A to reach a sufficient

number of variants for phylogenetic reconstruction, something that is currently often infeasible for many palaeogenomic studies. Equally, developing enrichment captures, as currently often performed in ancient pathogen studies, is currently difficult due to the uncharacterised nature of many of the oral microbiota (as described above). However, these initial results add support to the use of species present in the oral microbiota in ancient dental calculus for being potentially used as a proxy for studying host relationships. Indeed, once a more efficient reduction of cross-mapping for strain separation is possible (discussed further below; see section 6.4), this may open up possibilities for molecular dating of strains, and given that bacteria evolve at higher rates than their hosts, may give a finer temporal resolution to analyses of human evolution.

Overall, the exploratory nature of the large sample size study in Manuscript A has opened up a range of new potential research paths for future modern and ancient microbiome research. In particular, Manuscript A provides useful target lists for potentially important but under characterised genera of the human oral microbiome. Further research into these will help improve understanding of what effects may occur when modifying the relationships of these taxa with others in the oral biofilm - such as in medical treatment. Manuscript A also demonstrated that oral microbiomes can provide insight into the behavioural and possible cultural evolution of human populations. This is currently more technically challenging due to a lack of studies on the applicability of aDNA to modern functional profilers as well as less complete databases, functional and/or genome feature-level analysis. However functional analyses, rather than taxonomic analyses, may provide more productive avenues of research.

6.2 Sample sizes in ancient microbiome research

The complexity and nature of microbiomes, and the inherent degraded nature of ancient samples, means that sample size could be considered one of the most important factors in the analysis of ancient microbiomes. Degradation processes in burial environments mean that many samples often contain trace amounts or none of the original DNA that was present in the sample. Because of this, aDNA studies often have a high 'failure' rate, and many more samples are required to reach minimal numbers required for analysis, but also to ensure that any possible analytical patterns are not resulting from preservational artefacts. Previous ancient oral microbiome studies often had highly imbalanced sample sizes per comparative group - such as only one archaeological site per time period when attempting to infer dietary changes (Adler et al., 2013; Weyrich et al., 2017). This reduces the reliability of the results, as the differences between time periods are possibly the result of insufficient preservation or site-specific characteristics. While sample sizes have increased in more recent studies (Mann et al., 2018; Neukamm et al., 2020), these still have sparse strategies (few samples per site), or many samples from a single site. Both of which limits the inference of broader biological or cultural change.

The large dataset generated in Manuscript A provides two major advantages for the ancient microbiome field. Firstly, the dataset consists of a large number of individuals from multiple, balanced, geographic regions, time periods, and preservational qualities, with consistently generated sequencing libraries. It will therefore provide researchers a useful benchmarking dataset for assessing preservation compared to their own samples. Secondly, the nonhuman hominids in the dataset of Manuscript A will act as useful outgroups for future research focus-

ing on within-human variation - particularly for rooting phylogenetic trees of oral microbial genomes. Providing these outgroups could improve the molecular dating of such genomes in the approaches mentioned above, by offering deep calibration points of certain evolutionary events of the microbes. Using these outgroups in such a way would then subsequently improve the inference of when certain events in human cultural evolution occurred. One such example could be dating the adaptation of *Streptococcus* to bind human salivary amylase, possibly in response to increased starch consumption deep in human evolution, as put forward in Manuscript A. Indeed, bacteria generally have faster generation rates than host eukaryotes (e.g. Ochman et al., 1999; Lynch et al., 2016), and may provide higher resolution information on the evolution of behaviour traits than changes in the host genomes themselves.

One important consideration of the sampling strategy in Manuscript A was that each host genus included in the study should have multiple host species spanning different environmental and geographical locations - factors that can influence similarity and variation in the taxonomic profiles. We also required a minimum of 5 individuals per group and 10 per genus during sample collection. This accounts for natural intra-individual and population variation, as well as possible preservational differences derived from burial or museum collection conditions. The latter was reflected in the failure rate of samples yielding sufficient oral microbiome signatures for downstream analysis, where approximately only 70% of samples could be retained due to preservation. The number of discarded samples per host-group also varied, with some chimpanzee groups seeing no samples discarded, but with some gorilla groups seeing all but one sample having to be removed for downstream analysis. We also considered this issue of high failure rates and preservational skew when estimating which taxa could be considered 'core' to African Hominid groups. We required taxa to pass prevalence thresholds within each population, as well as passing this threshold in multiple populations to be considered core to hominids as a whole. This approach was therefore an important improvement over previous studies, showing that even if a given population or site of interest has few individuals, a broader sampling strategy can account for this by ensuring multiple populations or sites of a different group are also sampled. One aspect of this approach that could be questioned as implemented in Manuscript A, is having the rather low criterion of only 50% of individuals being required to have a taxon be assigned to the core of a particular group. This is because such a loose criterion could potentially allow taxa that are only transient in a particular population to be considered core. Indeed, experimentation in Manuscript A showed some variation in core species assignment when exploring different prevalence parameters. However, the results in Manuscript A should only be considered as a guide, and future ancient microbiome studies should utilise more considered, rather than opportunistic, sampling strategies, to account for both high failure rates and preservational skewing observed in Manuscript A. Furthermore, one limitation of the analytical design was that batch effects and other possible factors were not included in some of the statistical tests used (e.g., PERMANOVA; as performed in more recent ancient oral microbiome studies such as Brealey et al., 2021), due to the small sub-host genus population/group sizes. Larger sample sizes combined with improved metadata aggregation will help to further explore the results reported here.

Despite Manuscript A being the largest published ancient dental calculus dataset to date, some reviewers of the manuscript (presumed to be researchers working on modern microbiomes) remained critical of the sample size. In particular these reviewers commented that

results, particularly regarding the core microbiome analysis, will likely change with larger datasets. While this was an exploratory study that successfully demonstrated the potential information that could be gained from larger ancient oral microbiome studies, this remains a valid point. However, as stated in the introduction, ancient remains are often not as readily available as modern samples. In addition to skeletal remains generally being less abundant than modern individuals, palaeogenomicists must take into consideration the ethics of the destructive sampling of finite cultural heritage, as well as possible high failure rates in terms of DNA preservation. Obtaining new larger ancient datasets for improving statistical power in analysis is therefore not trivial, and this could be a limiting factor for future researchers.

Manuscript B offers one solution to this issue by exploiting the fortunate trend of the vast majority of palaeogenomicists uploading all their raw data to public repositories (Anagnostou et al., 2015). The community-curated 'AncientMetagenomeDir' repository (Fellows Yates et al., 2021a) is a list of all published and publicly available ancient metagenomic samples with basic but necessary metadata for ancient metagenomic studies. This resource allows researchers to rapidly search for and filter down to samples that are relevant for their own research, for example, either to geographic region, time-period, or sample type. This project also has longer term implications, such as for simplifying the tracking of the development of the field over time (such as in Orlando et al., 2021). It could also possibly allow the development of meta-analysis projects, such as improving predictions of the likelihood of microbial DNA preservation across different environmental conditions, as has been previously performed on ancient eukaryotic genomes (Allentoft et al., 2012; Kistler et al., 2017). Importantly, compared to other similar resources for ancient DNA (e.g., `https://www.oagr.org/`, or the Allen Ancient DNA Resource (AADR) available at `https://reich.hms.harvard.edu/`), this project has been built in a cross-lab collaborative and sustainable manner. This means that it can be easily maintained and continue to be developed in the future. This has been due to the important contribution of many members of the ancient metagenomics community, rather than just one student or lab, and the use of online collaborative tools, such as storing of the data on the intrinsically-collaborative and widely used GitHub platform (`https://github.com`; with long-term data archiving on Zenodo `https://zenodo.org/`). This longevity is important for emerging technologies such as machine learning, that require large data and associated metadata to produce models that can predict various characteristics of a dataset. Importantly, however, these models are designed to be improved over time through the inclusion of more data. Finally, this resource may assist in improving sampling strategies by allowing researchers or cultural heritage managers to check whether samples from a given site have already been sampled. By doing this researchers will reduce unnecessary repeated sampling of such finite resources, and allow projects to be more cost-effective; allowing funds to be invested in other areas.

Continued development of AncientMetagenomeDir is required to maximise the potential of the resource. Currently, the repository only includes metadata at sample level, and while accession numbers of each sample are stored, these are not always informative about the multiple types of sequencing data that may exist for each sample. Expanding the repository to library- and sequencing-run metadata will be an important step for promoting the adoption of the resource in analysis workflows. Providing direct links to the data would be a major benefit for researchers using the resource. For example, currently, researchers will still have to themselves find the exact sequencing read files for downloading. A major field-

wide problem that was identified during the compilation of the information was that sample metadata was reported in a very heterogeneous manner across publications, making it difficult for this information to be retrieved and standardised. For example, although radiocarbon date reporting standards already exist - such as standardised laboratory codes - (`http://www.radiocarbon.org/Info/labcodes.html`, Stuiver and Polach, 1977), these were not always followed in ancient metagenomics publications, thus making extracting date information difficult. This forced me to use practical but suboptimal workarounds such as the rounding of dates, or preferring less accurate uncalibrated radiocarbon dates. Such workarounds somewhat reduce the utility of this information in downstream analysis, outside of simple filtering to approximate time range of samples of interest or geographic regions (such as for precise tip dating in phylogenetic analysis). Furthermore, in many cases, the information reported in the publication itself did not correspond to the information sent along with the uploaded data on public data repositories, such as with inconsistencies between sample and library IDs. While the AncientMetagenomeDir project therefore acts as a first step towards improving this situation by reducing the need of individual researchers to go through each publication themselves, this heterogeneity needs to be corrected already at publication. Metadata reporting standards should be developed by the community to ensure wide-spread adoption and consistent reporting of necessary metadata for palaeogenomic studies. Such standards already exist in the modern genomics sphere, such as with 'Minimum Information about any (x) Sequence' (MIxS) checklists from the Genomic Standards Consortium (GSC), and are already used by public data repositories such as the European Bioinformatics Institute's (EBI) European Nucleotide Archive (ENA). Adoption of such a checklist dedicated to the palaeogenomics community would ultimately reduce the need for external projects such as AncientMetagenomeDir, as the public archives of data would already include this information, further speeding up access to data. Adoption of such common checklists would also have wider implications outside of purely academic purposes. Enforcing which, and how to record, such metadata of samples used for genetic analysis would mean other stakeholders, such as collection curators or cultural heritage managers, can monitor and track which samples are being used in the palaeogenomics field as a whole. For example, requiring the reporting of collection specimen IDs (not just internal lab codes) would both help reduce the risk of repeated sampling of (finite) samples, and allow cultural heritage offices to summarise how their samples are being used, as well as when results are available for further public dissemination.

Finally, increasing sample sizes for ancient microbiome studies should not be limited to purely genetic studies. As shown in Manuscript A, due to what appears to be a general taxonomic stability of the oral microbiome over deep evolutionary history, functional analyses may be more informative. However, more integrated biomolecular analyses to include proteomic (Mackie et al., 2017; Jeong et al., 2018; Hendy Jessica et al., 2018; Charlton et al., 2019; Geber et al., 2019; Scott et al., 2021; Bleasdale et al., 2021) or metabolomic (Velsko et al., 2017) analysis of dental calculus will be important, by showing what was actually being actively expressed by the biofilm, rather than what was potentially expressible (something that genomic data is limited to). Therefore, studies should start to be designed to not only increase sample sizes, but also increase the number of interdisciplinary analyses performed. While progress on this has begun to be made (e.g. Fagernäs et al., 2020; Fotakis et al., 2020), development should be invested in allowing these co-analyses to be scaled together to larger studies and reported in a

cohesive manner. Ideally, any push towards increasing interdisciplinary co-analyses of ancient dental calculus (and other ancient samples) should similarly see the necessity of making all data openly available. However, these must still follow FAIR and CARE practises to ensure the data is both reusable, but also *findable* in a responsible manner (Wilkinson et al., 2016; Carroll et al., 2020, 2021). This should occur for each of the analysis/data-types separately, but also with the data of all analyses being findable together.

Given all these archaeological science analyses would often derive from the same specimens (skeletal elements), one could envision a unified archaeological science metadata schema that could be used by a diverse range of archaeological science practitioners. Such a schema could have basic common information about a given specimen, and then submodules for each different type of analysis that is applied to it. Importantly, by having common specimen-level metadata, this would allow all other analyses to be found by researchers looking to integrate multiple lines of evidence. Indeed, such a biological metadata scheme already exists for a range of genetics-based techniques at transnational organisations such as the EBI's ENA archive (as described above). By developing resources in this direction, the information yield for both biological and archaeological studies will be maximised, producing more holistic evidence of changes to the human microbiome throughout human evolutionary and cultural history.

6.3 Improving throughput in ancient metagenomics analysis

Once future researchers have collected samples and generated sequencing data for large-scale studies, processing these large numbers of big sequencing files becomes a non-trivial task. Manuscript C provides a modern solution to this issue in the form nf-core/eager, a open-source and dedicated bioinformatics pipeline for ancient genomics, and now ancient metagenomics. Standard genomics pipelines and tools are often designed for 'pure' DNA extracts, high depth coverage, and long reads, and therefore do not include functionality to either validate or account for the degraded characteristics of aDNA (see 1.5 for summary). In contrast, aDNA data requires specialist tools and settings to produce accurate reconstructions of ancient (meta)genomes, to overcome the challenges related to environmental DNA contamination of samples, low sequencing coverage, and highly degraded 'endogenous' aDNA molecules. As an example, one of the latest 'best-practice' variant callers for human genomes is GATK HaplotypeCaller (Poplin et al., 2018), which superseded the previous GATK UnifiedGenotyper (DePristo et al., 2011). However, HaplotypeCaller performs *de novo* assembly of reads around potential variant sites. This is not possible with low-coverage and short-read aDNA, as these often do not have sufficient overlap to reconstruct a sequence without a reference. Therefore, when using HaplotypeCaller on a typical low-coverage ancient genome, many possible SNP sites are lost. In contrast, the original UnifiedGenotyper program only considered the alleles of a given position in the reads that mapped to a given site (compared to the reference). Therefore, the (now deprecated) UnifiedGenotyper algorithm is actually more suitable for ancient DNA contexts, as it does not require overlapping of reads to reconstruct longer sequences to account for insertions and deletions during variant calling.

While dedicated aDNA processing pipelines already existed at the start of this thesis, such as EAGER (Peltzer et al., 2016) and PALEOMIX (Schubert et al., 2014), these pipelines were designed in an era when only single or tens of genomes were being analysed in each study, and

with more limited computing resources being available. In contrast, palaeogenomic studies are now regularly generating or analysing hundreds of samples (Olalde et al., 2018; Mathieson et al., 2018; Margaryan et al., 2020). As palaeogenomics studies have increased in size, the complexity of these studies has correspondingly increased. Researchers are increasingly using ever more complex library construction and sequencing strategies, such as multiple library types and the use of different sequencing machines. nf-core/eager therefore addressed both challenges by reimplementing the original EAGER workflow in Nextflow (Di Tommaso et al., 2017). This domain-specific workflow manager performs automated, efficient, and sophisticated parallelisation of complex bioinformatic pipeline steps through integration with HPC schedulers. Furthermore, I coded a logic for automating the merging of libraries generated by heterogeneous library and sequencing setups at appropriate points. By using supplied metadata information to automatically calculate how each library should be processed, and when it can be merged with sister libraries, reduces a lot of complexity in early analysis 'housekeeping' steps of palaeogenomics studies. However, it is expected that as both the palaeogenomics and bioinformatics fields evolve, this complexity will continue to increase. By developing nf-core/eager within the nf-core initiative (Ewels et al., 2020), the pipeline will be easily adaptable to changes to bioinformatic software, infrastructure and best practises. This pipeline is already enjoying widespread interest and is currently being invested into by and supported by many different groups and disciplines, which, with the community-based development within the nf-core initiative, further reinforces the longevity of the pipeline.

Another aspect of further 'future-proofing' of the pipeline for bioinformatic and palaeogenomic evolution is the development of Nextflow's 'DSL2', a rewriting of the language that allows modulisation of the code. This means that different software and subroutines can be easily and rapidly integrated into pipelines as required. The adoption of these modules in a standardised manner by the nf-core community means that as new practices or analyses becomes routine in genomics or metagenomics, nf-core/eager can easily import an already written module (with tweaks to account for aDNA) that was developed originally for other nf-core pipelines. Therefore, the next phase of development of nf-core/eager will be refactoring the codebase to DSL2 to allow for this to happen. Facilitating rapid adoption of new analyses is important for palaeogenomics, as displayed in the recent history of the field, where laboratory improvements are allowing researchers to sometimes reach genomic coverage akin to modern DNA samples, allowing the application of more complex genomic analyses.

While earlier aDNA pipelines were originally designed purely for host genomics studies, the intrinsically metagenomic nature of aDNA samples and the (unexpected) success in finding possibly lethal pathogens from off-target reads (e.g. Rasmussen et al., 2015; Andrades Valtueña et al., 2017) has meant that researchers are now regularly metagenomically screening aDNA libraries for microbes alongside the host DNA. Allowing researchers to do both host-genomic and metagenomic screening - and critically, validation - of aDNA reads in a single pipeline is therefore important to improve the interdisciplinarity of studies. nf-core/eager added this functionality through the integration of in-parallel execution of the HOPS (Hübler et al., 2019) or Kraken2 (Wood et al., 2019) taxonomic profiling workflows - both allowing species identification of off target reads from host-genomic mapping, and in the case of HOPS, with aDNA authentication analysis. As the fields continue to evolve, being able to efficiently integrate newer analyses will allow nf-core/eager to offer further new lines of evidence. For example,

tools related to the metagenomic strand of analysis in nf-core/eager still remain somewhat limited. In particular, taxon or Operational Taxonomic Unit (OTU) tables (i.e., tables that list taxa found in each sample, and the number of reads of each taxon that were found), is a starting point for many microbiome analyses. However, currently, generation of OTU tables is only possible for the Kraken2 based metagenomic screening. Development of scripts or tools to generate OTU tables for MALT output of the HOPS pipeline (see below) would be advantageous as this something not natively supported by HOPS. This is due to the preferred use of this tool for aDNA, as MALT offers the ability to output direct alignments of reads against reference genomes, which are subsequently used for damage pattern analysis. Furthermore, the addition of dedicated quality control tools such as Nonpareil (Rodriguez-R et al., 2018), a metagenomic equivalent of PreSeq (currently included in nf-core/eager for host genome mapping) and is used for estimation of further sequencing effort, could be useful. The adoption of DSL2 in nf-core/eager would therefore facilitate the inclusion of these tools in the future.

The development of a new version of nf-core/eager within the nf-core initiative also allows improved access to interdisciplinary analysis to researchers. All nf-core pipelines (spanning genomics, metagenomics, metabolomics, and proteomics) share the same code base and template, therefore offering the same user interface. By removing the necessity to learn the peculiarities of different pipelines by having a common interaction interface, palaeogenomicists can very easily switch to analysing their data with a different nf-core pipeline to investigate other aspects of their data, and spend more time understanding the data itself rather than how to run the tool in the first place. The open-source and community characteristics of nf-core development also mean that the palaeogenomics community can easily offer their own customisation and software to allow each pipeline to adapt to the particularities of ancient DNA. While nf-core/eager offers many advantages to the palaeogenomics and ancient metagenomics communities for performing a wide range of different analyses, one outstanding issue is that the background of researchers is equally as diverse. This means that many researchers who begin to work with ancient DNA often have varying levels of knowledge of computational work and of the biological nature of ancient DNA, particularly in the case of researchers are coming from social sciences such as archaeology or anthropology. Within nf-core/eager, I therefore lead the writing of extensive documentation outside the typical scopes of describing basic usage instructions of pipelines, as often occurs in bioinformatics. We developed a documentation scheme that describes all outputs in a manner that provides researchers from introductory to expert-level backgrounds the rationale behind certain interpretations of report outputs, all in the context of aDNA. For example, many quality control reports from SGS processing tools such as FastQC, will consider aDNA as 'failing' various metrics and can lead users unfamiliar with aDNA to interpret these false positives as failures. Conversely, downstream analytical problems can occur when upstream SGS data processing was not carried out correctly, or artefacts remain. Therefore, we have written the documentation in an accessible manner, including graphical depictions of various report cases. Importantly, we designed all the documentation so that they can be adopted as training material by the wider palaeogenomics field (through generic descriptions, and creative commons open licensing). By making the documentation detailed, but also abstracted to palaeogenomics in general, this will improve general literacy into the fundamental components of SGS data processing. This open-source documentation will be a general boon to ancient metagenomics, which is a new but rapidly growing field,

and currently lacks formal training schemes enjoyed by other areas of palaeogenomics. This is important as many of the challenges of ancient DNA remain underappreciated resulting in problematic results, as described above (section 6.1.2), and despite attempts to broaden the awareness of these in microbial archaeology (Warinner et al., 2017). Therefore, expanding training material for the field in this open manner will also contribute to increasing accessibility of the field to a wider number of groups, particularly for those outside the established palaeogenomics community.

Finally, the nf-core initiative and the nf-core/eager pipeline highly recommends the use of software containers. These are singular image files that include all necessary software with required versions and configurations for the pipeline, meaning users do not need to install or maintain these themselves. This has important implications towards facilitating and improving reproducibility within the field. While public raw data sharing is widely performed within palaeogenomics (Anagnostou et al., 2015), description of analyses and workflows still remains limited to prose-based descriptions in manuscripts. This limits the ability for researchers to not only just reproduce the analysis but also reuse data from other publications in their own analyses. This is particularly important due to the complexity of current studies that generate multiple libraries and library treatments, the processing and merging of data of which are often not trivial to perform. By making nf-core/eager highly reproducible with the use of containers and portable configuration files, and assuming widespread adoption, I believe this will provide the benefit of standardising a lot of aDNA sequencing data preprocessing, something that should already be relatively routine given the low variation in data types. By doing so, it will be easier for future researchers to take an nf-core/eager input metadata table from previous publications (something that specifies exactly how various libraries were originally merged and treated), merge it with their own, and analyse large datasets together. Indeed, Nextflow already offers the ability to define URLs of data as input files, and automatically downloads these for users - further reducing the amount of hands-on work required compile such datasets. Specifically for ancient metagenomics, nf-core/eager could already be improved further, for example, with direct integration with AncientMetagenomeDir (Manuscript B) via automated conversion of future AncientMetagenomeDir library-level metadata tables to nf-core/eager input tables. This would be advantageous for making extremely automated and high-throughput analysis of comparative data, in addition to newly sequenced ancient metagenomic samples, possible.

6.4 Improving authentication in ancient metagenomics analysis

Ancient DNA analysis of microbes is not new, with a variety of different microbial pathogen genomes being recovered from skeletal remains over the last 10 years (Spyrou et al., 2019). However, the vast majority of these studies have focused on the recovery and authentication of a single or a small handful of taxa. In contrast, ancient microbiome work must contend not only with analyses at the scale of many samples, but also many taxa at once. A large component of this thesis was therefore invested into developing new workflows and methods for improving the throughput of preservation assessment of ancient microbial and metagenomic analysis. I will summarise a few key developments towards this challenge made during this thesis.

The new (ancient) metagenomic extension of the nf-core/eager pipeline (Manuscript C) in-

cludes the addition of MALT (Herbig et al., 2016; Vågene et al., 2018), an ultra-fast replacement for the widely used BLAST tool (Altschul et al., 1990), but designed for the taxonomic identification of SGS reads with optimisations to account for possible aDNA damage. This tool is applied to the unmapped reads from mapping to a host genome, and can be used both for screening of ancient pathogens, but also generation of taxonomic profiles for ancient microbiome samples (such as dental calculus) when combined with the GUI-based tool MEGAN6 (Huson et al., 2016). We also included Kraken2 (Wood et al., 2019) as a taxonomic profiling alternative, due to the large computing resources required by MALT. This is to ensure metagenomic screening remains accessible to as many researchers as possible, particularly those who do not have access to large-resource HPC clusters. More specifically for ancient microbiome analysis, screening for microbial genomes must also have some form of aDNA authentication. For those using the MALT integration in nf-core/eager, this is carried out by MaltExtract (a part of the HOPS pipeline; Hübler et al., 2019). This tool produces statistics for a list of user-defined taxonomic targets such as fragment lengths, damage profiles, and edit distance. In particular, edit distance is used for estimating how divergent the aDNA reads are to the reference genome. Currently there is no aDNA authentication step for the Kraken2 results, as this tool does not perform sequence alignment that allow users to identify nucleotide differences, and therefore users must manually validate their hits.

One issue with the default visualisation outputs from the HOPS pipeline (used for aDNA validation of metagenomic hits) are that they consist of singular PDF files for each sample and species - something not feasible for large scale microbiome studies such as in Manuscript A. I therefore developed an open-source interactive viewer, published with Manuscript A (MEx-IPA, `https://github.com/jfy133/MEx-IPA`), that allows users to rapidly load the output from HOPS and quickly switch between samples and species. This therefore results in a more efficient assessment of potential hits with authentic aDNA damage profiles and fragment lengths of many different original endogenous microbiome species, by integrating all results into a single window rather than having to open sometimes even hundreds of PDF files.

Another aspect of the metagenomic extension of nf-core/eager was the adding of specific tools used in the context of microbial genomic reconstruction (as also applied during Manuscript A for phylogenetic tree analysis). A common issue during ancient microbial genome reconstruction is the recruitment of reads from close environmental relatives to the reference genome of the target of interest; something that edit distance information can give indications of (Warinner et al., 2017). This 'cross-mapping' can complicate downstream phylogenetic analysis by causing incorrect clustering of genomes due to chimerically associated variants, and falsely increasing or decreasing mutation rates in molecular dating analysis. I therefore added MultiVCFAnalyzer (Bos et al., 2014) as a consensus calling tool, i.e., a tool that generates a single or multi-FASTA file with each sample's variants and used as input for downstream phylogenetic tools. MultiVCFAnalyzer has additional functionality that allow users to assess the level of this cross-mapping from environmental relatives, so as to allow users to further estimate the reliability of their genome reconstruction and phylogenetic analysis. Knowledge of these issues are somewhat lacking, particularly in the ancient microbiome field (see section 6.1.2). Therefore including this in an automated pipeline will help improve the awareness of the issue by making such tools that can assess and reduce cross-mapping more accessible.

Outside of phylogenetic reconstruction, (ancient) pathogen genomics and microbiome re-

search is starting to shift focus towards functional analysis of these genomes, such as the presence- and absence of genes (as also performed in Manuscript A, see section 6.1.2). To assist ancient metagenomics researchers to more readily access this information, I integrated into nf-core/eager the ability to automatically generate basic coverage statistics of all genome annotation features, such as genes, across a given reference genome via bedtools (Quinlan and Hall, 2010). This can thus be used to detect insertions or deletions within the population of ancient bacterial genomes, but also act as an additional species identification validation. For example, the species-specific marker gene *pla* on the pPCP1 plasmid (Parkhill et al., 2001) is often used in palaeogenomic studies to distinguish the pathogen *Y. pestis* from its environmental relatives (Schuenemann et al., 2011). Such analyses may become more important in future ancient microbiome analysis given the results of Manuscript A, that highlighted the need for further research into the currently many undercharacterised taxa identified as being possibly important taxa for the human oral biofilm. By identifying marker genes that can by used to indicate a particular microbiota species or strain being present in a sample, even if there is a wide diversity of closely related taxa also present, could be highly useful. Therefore, automation of such analyses in a single pipeline, will allow more for routine use across subdisciplines, and possibly help shift ancient microbial fields away from addressing purely 'who is there, when, where' questions to also 'how' and 'what are they doing' (Integrative HMP (iHMP) Research Network Consortium, 2014; Berg et al., 2020).

In general, the automated execution of these established (meta)genomic screening and reconstruction tools, but also with the results of these tools being included within a *single* interactive run report, will be useful for interdisciplinary fields such as palaeogenomics. Such data quality control and authentication information being aggregated into a single place will provide a smoother experience for users who are less bioinformatically experienced to improve evaluation of their own sequencing data (see section 6.3). This thus helps democratise and make the field more accessible for a wider range of research specialisms and topics. During Manuscript A, however, it became clear there were many areas of ancient microbiome analysis that are under-developed. In particular, many of the established validation tools included in nf-core/eager (Manuscript C) were developed with the primary aim of detecting or analysing only a single or a handful of (microbial) species. While nf-core/eager helps to scale across multiple samples, the available approaches for microbiome sciences were not scalable to the additional dimension of having multiple species to validate for each sample. I therefore developed additional tools and approaches towards this goal.

A critical component of ancient microbiome studies is the removal of samples displaying large taxonomic skews due to degradation and high amounts of environmental contamination. In previous and current ancient microbiome research (e.g., Ziesemer et al., 2015; Weyrich et al., 2017; Mann et al., 2018; Ottoni et al., 2019), the Bayesian source-estimation tool SourceTracker (Knights et al., 2011) has been regularly used to estimate the likely source of a given taxon when comparing a sample to different 'source environment' samples. However, this tool does not provide a means to specify cut-offs that allow 'include/exclude' selection of microbiome samples for downstream analyses. It also only gives a 'one-dimensional' view of the preservation of samples. For example, a sample dominated by a single taxon could be estimated to be most similar to the target environment (e.g., dental calculus) even if the remaining, very low abundance taxa, could be all derived from an environmental source. I developed a method termed

'cumulative percent decay curves' in Manuscript A, that allows for both finer evaluation of the level of the preservation of expected microbiome samples across all taxa in a sample, as well as methods to allow 'include/exclude' cut-off thresholds. This method, validated against both manual inspection of SourceTracker results, and qualitative assessment of comparative PCAs of dental calculus samples versus other environmental sources (such as soil, skin, and bone samples), showed that it was a useful heuristic for assessing whether samples have sufficient preservation for downstream analyses (Manuscript A). The method has also since been formalised in an R package, 'cuperdec', (`https://github.com/jfy133/cuperdec`) that has been published on the most widely-used R package repository CRAN under an open-source license, to allow other researchers to easily adopt the same method for their own ancient microbiome research. Importantly, the visualisation of such curves can be included even on a single plot. Therefore, in principle, this allows scalable visual inspection of the preservation of even hundreds of samples. This factor will be important as the sample sizes of ancient microbiome studies continues to grow (see section 6.2), but also as databases and sample types evolve, as well as new threshold algorithms are developed. All of these developments would then require regular visual inspection of the curves to ensure any identified thresholds are working as expected.

Even once samples with little original microbiome preservation have been removed, all other ancient samples will have some form of environmental contamination, either from the burial environment or introduced in lab contexts. Previous research in ancient microbiome research has used 'brute force' methods (e.g. Eisenhofer et al., 2020; Weyrich et al., 2017) to remove these taxa during analysis. This would typically consist of removing any taxon that is present in any control from all samples. This highly conservative approach may result in the removal of false positives, based on a spurious hits deriving from a single read hitting a taxon by chance. In contrast, I took a more nuanced approach via the recently published R package 'decontam', which uses the observation of an inverse correlation in the abundance or prevalence of contaminants in negative controls compared to samples. This ensured that I did not over-zealously remove oral taxa from downstream analysis. However, this only addresses possible laboratory contamination. To account for taxonomically diverse burial environmental contamination, we also further modified the typical use case of the package by using unrelated bone samples as a 'control source', in addition to negative controls, to include removal of common environmental contaminants. Using bone samples rather than soil is important, as environmental microbial communities of skeletal elements have a different profile to that of the surrounding soil, and the exact profile varies depending on anatomical element (e.g. Kazarina et al., 2019; Emmons et al., 2020; Pinzari et al., 2020). Therefore, the use of bone controls represent more similar (although not perfect) environments to identify environmental microbial communities in dental calculus over soil samples (which are also rarely stored in museum collections) as bones have more similar mineral and nutrient environments than soil. Comparing against bone samples therefore will improve the identification and separation of common environmental taxa from endogenous oral microbiome taxa. However, a more systematic investigation of this approach should be made. In Manuscript A, due to a lack of published raw FASTQ level shotgun data of multiple skeletal remains from the sites included in this study, we had to use bone samples from an unrelated site and location. Future research is required to address the feasibility of including bone samples from the same site as dental calculus samples, given that each burial

site may have different soil- and sediment taxonomic profiles.

Ultimately, after developing a comprehensive novel workflow to minimise biases derived from the characteristics of ancient metagenomic DNA as described in Manuscript A, we found that signatures of authentic ancient oral microbiomes could be preserved twice as long as previously thought. Calculus from Neanderthal individuals as old as 70 and 100 kya from De Nadale (Italy) and Pešturina (Serbia) caves respectively, both yielded DNA from well-characterised oral taxa and displayed both high levels of aDNA-associated deamination and high fragmentation. This is particularly important as dental calculus was noted to have unusually high DNA yield for ancient samples (Warinner et al., 2014b). Indeed, with the recent discovery that DNA can be preserved in permafrost samples to over one million years ago (van der Valk et al. (2021); exceeding theoretical expectations by Smith et al. (2003)), this highlights that more basic research should be directed to further understand the mechanisms of DNA preservation. Understanding particular characteristics of the organic and inorganic make-up of dental calculus may help revise DNA survival estimates, but also guide methods of non-destructive pre-screening of biomolecular content in the sample type, as is being developed for bone samples (e.g. Kontopoulos et al., 2020). Given the large number of scientific methods that can be performed on dental calculus deposits (see section 1.4.1), the development of such screening tools will help guide sampling strategies and which methods to apply. It will also help researchers estimate from how far back in time ancient oral microbiomes could potentially be reconstructed.

Finally, in Manuscript A, I made an attempt to address the problem of cross-mapping in single genome reconstruction of closely related species derived from complex microbial communities. As the plaque biofilm is taxonomically rich, there are many closely related or multiple strains of species that have similar genomes. During mapping-based genome reconstruction with a reference sequence, reads from these strain-specific SNPs can also map to the reference genome of the target of interest and result in 'multi-allelic' SNPs - despite bacteria generally having haploid genomes. While in Manuscript C, I described the inclusion of MultiVCFAnalyzer, which allows for the assessment of the level of such cross-mapping (see above), it currently does not provide sophisticated methods of reducing this effect, other than allowing relaxation of SNP calling parameters to make a 'best guess' of the original SNP at a given multi-allelic site. These multi-allelic SNPs result in uncertainty during downstream analysis. As described in the introduction (section 1.5.4), modern genomic solutions typically rely on high-coverage and long-read sequencing libraries to reliably 'haplotype' and separate strains of different taxa through co-association of different SNPs on single reads - something not possible with short read length and low coverage aDNA libraries. In Manuscript A, I attempted an alternative approach to reduce multi-allelic SNPs via 'competitive mapping' (i.e., mapping to multiple genomes at once) to pull away closely-related but off-target reads to their original sources (or strains more closely related to the off-target species), leaving only the correct position with the more likely original allele. Unfortunately this approach was not successful across the 15 species analysed. While the level of multi-allelic SNPs was generally reduced, coverage was significantly lower, making the confidence in SNP calling also lower, and thus equally problematic for downstream analyses. This highlights the challenges that remain in regard to genome reconstruction from complex ancient microbial communities, and that new approaches are needed to increase coverage, but that also allow differentiation between strains from aDNA data.

6.5 Outstanding challenges for ancient metagenomics analysis

Overall, in this thesis I generated a range of approaches, resources, and considerations for the ancient metagenomics community. These have primarily focused on automating and making scalable routine processing procedures, authentication of sample preservation in microbiome samples, and developed approaches to reduce the level of environmental contamination. However, a range of challenges were not resolved by the three manuscripts, which I believe should be the next steps that the field must take to improve the technical aspects of ancient microbiome work.

An ongoing issue - also in modern metagenomics - is in regard to metagenomic database sizes. As microbiomes are complex communities of many taxa, we must align our sequencing reads across many genomes at once. The number of genomes in databases are increasing at rapid rates as genomic sequencing techniques improve and become cheaper, and as such this becomes increasingly difficult to fit in available computing resources (Ye et al., 2019; Nasko et al., 2018; Kim et al., 2016; Zhou et al., 2018a). In modern metagenomics, most solutions utilise 'k-mer decomposition' based approaches that use multiple short sub-strings of a given nucleotide sequence to make a distinct but concise representation of the sequence (resulting in smaller database sizes). These representations are then compared to k-mer profiles of reads, rather than a nucleotide-per-nucleotide alignment against the reference genome (a computationally expensive procedure; Wood and Salzberg, 2014; Breitwieser et al., 2019), or nucleotide to amino acid conversion and amino acid to amino acid sequence comparison (e.g. Buchfink et al., 2015; Menzel et al., 2016). An issue with these approaches for aDNA is that very short reads (such as found in the late Pleistocene/Upper Palaeolithic individuals in Manuscript A) result in unspecific k-mers and reduces accuracy in taxonomic assignment and/or abundance estimation (Manekar and Sathe, 2018; Menzel et al., 2016; Song et al., 2014; Ye et al., 2019). Indeed, previous simulation work generally found that current k-mer based taxonomic profilers performed worse than alignment-based profilers for abundance estimation (Velsko et al., 2018). Equally, protein alignment approaches have demonstrated to be less accurate than nucleotide alignment (Eisenhofer and Weyrich, 2019). In both cases of k-mers or amino acid alignments, these approaches limit the ability of researchers to perform established authentication procedures, as the nucleotide information of C to T transitions are lost (as this requires direct nucleotide sequence comparison against a nucleotide reference). While researchers could do this in a separate step, this can be computationally expensive when being performed across hundreds of genomes. Further dedicated simulation and benchmarking studies are required to assess the limits of k-mer approaches for ultra-short aDNA samples (something that has started to be performed recently, such as by Cárdenas et al., 2021), and other approaches to reduce the computational resources required for metagenomic alignment need to be developed. The issue of the use of various databases within the field has also meant that there are currently no standard taxonomic profiling databases for ancient and modern microbiome researchers to use, which limits straightforward comparison of results between studies. The establishment of such 'standard databases' would also help improve reproducibility. Furthermore, this would also help improve portability of user-friendly pipelines such as nf-core/eager from Manuscript C, as it would further reduce the time required for hands-on database construction, and conversely increase the time spent by researchers analysing the biological aspects of their data.

Selection of standard database(s) and tool(s) for initial taxonomic profiling for ancient metagenomics would also assist in improving downstream tools for preservational assessment and filtering. One can imagine an equivalent to the curatedMetagenomicData dataset (Pasolli et al., 2017) for ancient samples by using the data from AncientMetagenomeDir of Manuscript B, which could then be used to develop improved methods of classifying well- and less-well preserved samples through large-scale comparisons. Furthermore, improvements could be made to the cumulative percent decay curve method introduced in Manuscript A, by adding more sophisticated filtering algorithms based on the source-decay curves. Currently only simple hard-threshold cut-offs have been added to the associated cuperdec package, and one more sophisticated method based on the settling of fluctuations in the curves (akin to convergence in Bayesian MCMC-chains). However, more statistically-informed methods that additionally estimate the level of uncertainty could also be developed. Standardised and curated databases would also assist in the cumulative percent decay curve method. This method requires lists of reference genomes and their known isolation sources to identify the number of genomes in a sample that are derived from the expected microbiome type (such as microbial genomes from the oral cavity expected for dental calculus). Such development would also require standardisation and curation of not just sequencing data but the genome sequences themselves. These genome sequences would also require corresponding metadata, such as isolation sources, something which is currently only sparsely available in the commonly used NCBI Genome database (`https://www.ncbi.nlm.nih.gov/genome`) or inconsistently included in the otherwise metadata rich BacDive cultural collection database (Reimer et al., 2019).

In addition to improving preservational screening, ancient microbiome researchers will need to more routinely apply statistical methods to assess the robusticity and reliably of their results. This is due to the the small sample sizes in currently published ancient studies, compared to that expected of modern microbiome studies. Indeed, this was a common worry of reviewers of Manuscript A, that the size was too small to make any significant interpretations. One factor in this is that most training resources in this vein focus on clinical case-control studies (e.g. Casals-Pascual et al., 2020), something that isn't necessarily applicable to more heterogeneous and somewhat intrinsically opportunistic study designs in palaeogenomics. However, in Manuscript A, to simulate the effect of stochasticity derived from small samples sizes, I used 'bootstrapping' (Efron, 1979) to assess the variability in taxonomic profiles when randomly removing samples across many resampling replicates. An issue with this approach, however, is that it is reliant on the underlying distribution of samples themselves. If consisting of many 'outliers' (such as if one archaeological site or sample collection has low preservation), this may skew results. This method also does not allow estimation of the statistical power of the dataset. Therefore, in addition to expanding their datasets via resources such as AncientMetagenomeDir (Manuscript B), researchers should explore different methods of estimating and accounting for the statistical power in their studies, while study sample sizes continue to grow over time.

Another limitation of Manuscript A towards the broader aim of using ancient microbiomes to expand our knowledge of the diversity of the human microbiome was that the analytical tools used require reference sequences. In other words, taxonomic identification required comparison of sequencing reads to already sequenced reference genomes of already isolated taxa. This therefore limits the detection of species in ancient individuals to known or closely related taxa,

all of which are derived from modern contexts. In contrast, *de novo* assembly is a reference-free method for reconstructing genomes. Via repeated overlapping of reads with partially similar sequences (and subsequently larger already-partially assembled 'contigs'), assemblers identify the most efficient and parsimonious 'path' through the overlaps to derive a single contiguous sequence - considered to be the most likely original whole-genomic sequence (Compeau et al., 2011). Reference-free methods are the ideal target method for the identification of the unknown diversity of ancient oral microbiomes, as they are not reliant on known diversity to identify taxa, and could potentially reconstruct extinct species. Taxonomic assignment of these reconstructed 'MAGs' (metagenomically assembled genomes) can then be made via phylogenetic placement within known diversity. Secondly, these methods also allow *de novo* identification of new predicted functional features in (extinct) strains of known taxa, as again it does not require a modern reference. This would be an important development following the opinion stated above (see section 6.1.2), that ancient oral microbiome studies will likely need to focus on analysing changes of the functional capacity of oral microbiomes to identify cultural change rather than purely taxonomic inference methods. However, most modern assemblers use k-mer based approaches that prefer longer-reads to resolve gaps during contig-overlap, something that theoretically appears to be incompatible or at least sub-optimal with very short aDNA reads (see above). Studies exploring the limitations and possibility of new approaches (such as Seitz and Nieselt, 2017; Wibowo et al., 2021; Borry et al., 2021), and new solutions towards reference-free assembly should be carried out.

Understanding how the microbiota of ancient individuals adapted to changing environments, something 'mostly' defined by the host, has potential as a proxy source of evidence for human behavioural and cultural change. However, the functional pathways of ancient oral microbiomes reconstructed by aDNA remain relatively unexplored (Warinner et al., 2014b; Jacobson et al., 2020). To date, methods such as ancient metaproteomics and metabolomics (Warinner et al., 2014b; Jersie-Christensen et al., 2018; Velsko et al., 2017, e.g.,) have been more often used, although these fields are younger and less established than palaeogenomics. This is partly as functional analysis of microbiomes is highly complex, and again highly reliant on databases of annotated genes (Kuczynski et al., 2011; Quince et al., 2017), which may often not be available for undercharacterised taxa. Furthermore, while this approach was partly applied in Manuscript A by co-author Irina Velsko as an exploratory analyses through the use of AADDER (Huson et al., 2016) and HUMAnN2 (Franzosa et al., 2018), the effect of DNA damage and short read lengths was not evaluated. However, as annotation databases and genome reconstruction methods improve in both modern and ancient metagenomics, this will become more feasible over time. Indeed, coupling *de novo* assembly of unknown species with homology-based profiling from functional databases has the potential to provide transformative insights into ancient human microbiomes in both contexts of health and disease and human cultural change.

7 Conclusion

This thesis has explored how using aDNA from dental calculus in large-scale studies can help us explore how the human microbiome has evolved across deep evolutionary timescales. Manuscript A reported observations regarding both the surprising conservation of the taxonomic diversity of the core hominid oral microbiome, but equally that large parts of this diversity is currently undercharacterised (Fellows Yates et al., 2021c). Manuscript A also demonstrated how ancient microbiomes can potentially give insight into not just human biology, but also potentially human history. Manuscript A uncovered functional differences between nonhuman primate and human oral microbiomes, the results of which may have implications for the dating of the adaption to starch-based dietary habits in past hominids, as well as showing that phylogenies of oral microbes could provide an alternative route for reconstructing relationships between different human populations.

From the experience of Manuscript A, however, it became very clear that in ancient microbiome research, robust scalability of crucial analysis steps is currently not possible. Without development of tools and approaches that address the fundamental problems of ancient metagenomics - such as preservation and authentication - the field will struggle to advance. Therefore, Manuscript A describes novel and scalable approaches to improve assessment and account for microbiome aDNA preservation in dental calculus (MEx-IPa, cuperdec). Manuscript B described a community-curated repository of all samples used and publicly published in ancient metagenomic studies (AncientMetagenomeDir), with metadata and accession codes to raw data, improving the reuse of data to increase sample sizes of future (meta)studies (Fellows Yates et al., 2021a). Manuscript C described a complete reimplementation of a popular ancient DNA genomics pipeline in a new pipeline framework to allow adaptation to the latest computational, bioinformatic, and software development infrastructure and best practices (nf-core/eager; Fellows Yates et al., 2021b). The pipeline also includes a range of new extensions over the original pipeline, including extensive documentation to facilitate wider adoption of robust ancient metagenomic analyses and authentication in palaeogenomics.

The outcomes of this thesis offers a wide range of exciting new research avenues for the young field of ancient oral microbiome research, and equally will facilitate future studies in these areas through the development of novel, open-source, and community-developed approaches and tools.

8 Summary / Zusammenfassung

8.1 English

The microbes that live in and on our bodies play major roles in health and disease due to their symbiotic relationship with the host. Understanding how these communities adapt to changes in their environment - either by natural or anthropological forces - is currently a critical area of research for improving holistic healthcare. However, modern research heavily relies on populations from industrialised societies. Due to this, the full diversity of the human microbiome is not known. Microbiomes from *past* societies and our Hominid relatives therefore have the potential to provide insights into how long-term human behaviour change has modified the human microbiome. Yet, this area of research is in its infancy and many analytical aspects remain underdeveloped. Large sample sizes are required to analyse complex microbial communities, and correspondingly workflows and tools to process these are needed.

The aim of this thesis was to demonstrate the potential of large-scale shotgun-sequenced ancient dental calculus. By doing so, I aimed to improve our understanding of ancestral oral microbiomes in both biological and anthropological contexts. Furthermore, I aimed to build tools and resources to improve throughput in ancient metagenomic analyses.

In Manuscript A, I have shown that ancient dental calculus can be used to improve the understanding of past human oral microbiome diversity (Fellows Yates et al., 2021c). Using the largest ancient dental calculus dataset to date of 124 individuals, I have identified that DNA from oral biofilms can be preserved at least until 100 kya in Neanderthals (doubling the age of previous results); showing accessing ancestral and extinct diversity is possible. I have found evidence suggesting that oral biofilms appear to have a long evolutionary taxonomic stability within the hominid lineage. Genera known to play important oral biofilm structural roles in modern humans were also found to be present in Neanderthals, chimpanzees, gorillas, and also in some cases howler monkeys. Importantly, taxa typically designated as markers for periodontal disease were found to be highly prevalent across all host species. This highlights that these taxa are not human-specific, and likely that the presence of which are not solely sufficient to diagnose disease, in neither ancient nor living individuals. However, many of these taxa are under-characterised, and may have implications towards oral health and disease. Manuscript A also indicated that inferences about human behavioural or cultural change from ancient microbiomes may be possible, albeit seemingly not in the way that was previously assumed. I was unable to replicate previous work suggesting general taxonomic shifts occurred in response to major cultural changes throughout human history (e.g., farming, antibiotic usage). Instead, *functional* changes within the oral microbiota is a more intriguing alternative. Manuscript A identified and began to pinpoint the 'arrival' of specific genes associated with carbohydrate processing in *Streptococcus* species. This appears to correspond to a critical point in the development of humans in regards to brain development when energy-dense foods were likely necessary. This exemplifies how the ability of microbes to rapidly adapt to their local environment can be used as a proxy for identifying the timing of important behavioural and cultural changes in humans.

Continued advances from Manuscript A will only be possible by routinely maximising sample sizes. However, samples from archaeological and museum remains represent finite and

precious cultural heritage. Manuscript B describes the repository 'AncientMetagenomeDir', that I established and lead development of (Fellows Yates et al., 2021a). This community-level resource lists all public ancient metagenomic sequencing datasets. The resource will allow researchers to efficiently re-use public data to ensure the robusticity and improve the statistical power of future studies. Furthermore, it will act as a useful starting point for meta-analyses. For example, it could potentially be a useful dataset for making predictions of aDNA preservation in a variety of contexts. Requiring large ancient metagenomic datasets in future studies only makes sense if processing is made with tools designed for these sizes. In Manuscript C, I describe an entirely rewritten user-friendly palaeogenomics pipeline following latest software development and bioinformatics best practises. In addition I wrote extensive documentation designed for the interdisciplinary user-base that makes up the palaeogenomics field (Fellows Yates et al., 2021b). The pipeline, nf-core/eager, has been developed in a way that allows for easy integration with large scale computing infrastructure required for such analyses. Importantly, I have extended this genomics pipeline to have in-parallel metagenomic profiling and screening of ancient DNA characteristics. During Manuscript A, I also developed a sophisticated workflow designed specifically for ancient microbiome analysis. Each analytical stage was designed with strategies towards accounting for and authenticating the intrinsic degraded nature of ancient DNA (MEx-IPA), including novel tools for preservational screening (cuperdec).

These manuscripts have contributed new insights into the biology and evolution of oral biofilms, but also introduced new open-source and sustainable tools and resources that will allow further investigation of ancient microbiomes. These contributions also provide guidance as to where future research should look towards. These include intensifying the investigation of undercharacterised but prevalent taxa, more routinely performing functional analysis of ancient microbiomes, and showing that more formalised tools designed for the analysis of low-coverage microbial genomes, will all be required to allow deeper understandings of the long term evolution of the human microbiome.

8.2 Deutsch

Übersetzung von Raphaela Stahl und Clemens Schmid

Der menschliche Organismus ist von mikrobiellen Gemeinschaften besiedelt und wir leben in einer dauerhaften, symbiotischen Beziehung mit ihnen. Dieser Beziehung ist es geschuldet, dass Mikroorganismen großen Einfluss auf unsere Gesundheit nehmen: Oft sind sie entscheidend dafür, ob und wie wir erkranken. Wenn man also die menschliche Gesundheit holistisch betrachten möchte, dann ist es von großem Interesse zu verstehen, wie das humane Mikrobiom auf ökologische Veränderungen reagiert und sich gegebenenfalls anpasst. Bedauerlicherweise fußt ein Großteil der Forschung bisher ausschließlich auf Studien mit Proben von industrialisierten Gesellschaften, woraus ein Erkenntnismangel zu Biodiversität und Potenzial von Mikrobiomen resultiert. Die Erforschung der Mikrobiomkomposition *vergangener*, menschlicher Gesellschaften und verwandter, hominider Spezies kann hier Abhilfe schaffen. Sie erlaubt Einblicke dazu, wie der Wandel von Verhaltensweisen und Ernährungsgewohnheiten das menschliche Mikrobiom langfristig beeinflusst. Diese Forschungperspektive ist noch in

ihren Anfängen — die Entwicklung und Erprobung relevanter, analytischer Methoden und Werkzeuge ist ein Desiderat. Weiterhin kann die Komplexität mikrobieller Gemeinschaften nur mit einer großen Anzahl an Proben aus unterschiedlichen Kontexten erschlossen werden. Das erfordert die Etablierung stabiler Arbeitsabläufe, die die neuen Analysewerkzeuge sinnvoll integrieren.

Ziel dieser Doktorarbeit war vor diesem Hintergrund, das Potenzial großangelegter Analysen von shotgun-sequenzierten, "alten" (d.h. z.B. aus archäologischen Kontexten) Zahnsteinproben aufzuzeigen. Das bisherige Wissen zur Geschichte des oralen Mikrobioms sollte sowohl aus biologischer als auch anthropologischer Sicht erweitert werden, während gleichermaßen Methoden entwickelt werden sollten, die den erforderlichen Durchsatz metagenomischer Analysen überhaupt erst möglich machen.

Manuskript A demonstriert, dass alter Zahnstein unser Verständnis der Diversität vergangener humaner oraler Mikrobiome tatsächlich verbessern kann (Fellows Yates et al., 2021c). Hierfür wurde eine — die bis dato größte — Probensammlung mit Zahnstein von 124 Individuen verarbeitet. Zahnstein konserviert DNA aus oralen Biofilmen vergleichsweise gut, und so konnte unter anderem DNA aus 100 000 Jahre alten Proben von Neandertalern extrahiert werden. Das ist doppelt so alt, wie die bis dahin ältesten analysierten Proben. Eine diachrone Analyse des Datensatzes ergab, dass orale Biofilme eine lange taxonomische Stabilität innerhalb der Hominiden aufweisen. Mikrobielle Genera, die nachweislich eine wichtige Rolle für den oralen Biofilm moderner Menschen spielen, konnten ebenfalls bei Neandertalern, Schimpansen, Gorillas und teilweise bei Brüllaffen gefunden werden. Typische Marker-Taxa der Pariodontitis sind unter allen Wirtsspezies verbreitet, was darauf hindeutet, dass diese Taxa nicht notwendigerweise humanspezifisch sind. Außerdem ist die bloße Anwesenheit dieser Taxa nicht hinreichend, um auf Krankheiten zu schließen — weder in früheren noch in heute lebenden Individuen. Ein Zusammenhang mit oraler Gesundheit kann jedoch auf dieser Grundlage auch nicht ausgeschlossen werden, da die Charakteristika dieser Taxa nicht vollständig erforscht sind. Unabhängig davon präsentiert Manuskript A Indizien, dass es potenziell Wechselwirkungen zwischen der evolutionären Entwicklung des Mikrobioms und menschlichem Kulturverhalten gegeben haben könnte; wenn auch nicht in bisher angenommenem Umfang. Von früheren Studien postulierte, starke taxonomische Verschiebungen aufgrund kultureller Veränderungen in der Menschheitsgeschichte (bspw. die Einführung der Landwirtschaft oder Nutzung von Antibiotika) konnten nicht reproduziert werden. Stattdessen gibt es Hinweise darauf, dass diese Prozesse mit *funktionalen* Änderungen der oralen Mikrobiota einhergegangen sein könnten. Manuskript A dokumentiert das Aufkommen neuer Gene in *Streptococcus* Spezies, die mit der Prozessierung von Kohlenhydraten assoziiert sind. Hier ist ein Zusammenhang mit der menschlichen Gehirnentwicklung plausibel, für die eine Intensivierung des Konsums energiereicher Nahrungsmittel angenommen werden kann. Dieses Beispiel veranschaulicht das Potenzial von Mikroorganismen, sich schnell an Umweltveränderungen anzupassen. Das kann ggf. als abstrakter Proxy herangezogen werden, um den Zeitpunkt wesentlicher Kultur- und Verhaltungsänderungen in der Menschheitsgeschichte näher zu bestimmen.

Die methodischen Ansätze in Manuskript A versprechen weiterführende Erkenntnisse, sofern die Probenanzahl weiter erhöht werden kann. Jedoch repräsentieren Proben aus archäologischen Funden oder Museen ein endliches, kulturelles Erbe. Hier setzt Manuskript B an. Es

beschreibt ein digitales Archiv — *AncientMetagenomeDir* — das in Vorbereitung des Manuskripts implementiert und veröffentlicht wurde (Fellows Yates et al., 2021a). AncientMetagenomeDir ist als Community-Resource konzipiert und listet alle ordentlich publizierten und metagenomisch-sequenzierten Datensätze aus aDNA (ancient DNA - alte DNA) Kontexten. Das Archiv soll es der Forschungscommunity ermöglichen, publizierte Daten effizient wieder- und weiter zu verwenden, und damit die Robustheit und statistische Aussagekraft zukünftiger Studien zu stärken. Es eignet sich als solider Startpunkt für Meta-Analysen — etwa um den Zusammenhang zwischen DNA-Erhaltung und biogeographischem Kontext zu untersuchen.

Die Anzahl und Komplexität von aDNA Datensätzen wird auch in Zukunft wachsen. Das hat zur Konsequenz, dass die zu ihrer Verarbeitung notwendigen Softwarewerkzeuge gleichermaßen weiter entwickelt werden müssen, um mit ihnen zu skalieren. Manuskript C führt die palaeogenomische Bioinformatik-Pipeline *nf-core/eager* ein, die auf Grundlage moderner Best Practises für Softwareentwicklung und mit einem Fokus auf Nutzerfreundlichkeit aufgebaut wurde. Aus Rücksicht auf die interdisziplinäre Nutzergruppe, die das gesamte Feld der Palaeogenetik umfasst, ist *nf-core/eager* umfangreich dokumentiert (Fellows Yates et al., 2021b). Die Pipeline soll sich möglichst einfach in Hochleistungsrechensysteme integrieren lassen, die für bioinformatische Analysen üblicherweise zur Anwendung kommen. Das Featureprofil schließt unter anderem metagenomisches Profiling und ein Screening auf aDNA-Charakteristika ein. Während der Arbeit an Manuskript A, wurde *nf-core/eager* außerdem um eine Pipeline erweitert, die für die Arbeit mit Mikrobiom-aDNA optimiert ist. Die Arbeitsschritte sind darauf angelegt, die besonderen Eigenschaften alter DNA zu erkennen und zu berücksichtigen (MEx-IPA), was insbesondere neue Werkzeuge für "preservational screening" (cuperdec) umfasst.

Zusammen dokumentieren die drei hier eingeführten Manuskripte neue Einsichten in Biologie und Evolution des oralen Biofilmes und stellen spezifisch dafür entwickelte, nachhaltige Open-Source-Softwarewerkzeuge vor, die die Arbeit mit alten Mikrobiom-Proben erleichtern. Sie eröffnen die Perspektive für zukünftige Forschungsansätze; etwa die Suche nach dominanten, aber schlecht erforschten Taxa, oder routinemäßige funktionale Analysen an alten Mikrobiomen. Ein tieferes Verständnis der langfristigen Evolution des menschlichen Mikrobioms ist abhängig von Software, die dediziert für die Analyse schlecht erhaltener, mikrobieller aDNA entwickelt wurde.

9 References

Åberg, C. H., Kelk, P., and Johansson, A. (2015). *Aggregatibacter actinomycetemcomitans*: virulence of its leukotoxin and association with aggressive periodontitis. *Virulence*, 6(3):188–195.

Abusleme, L., Dupuy, A. K., Dutzan, N., Silva, N., Burleson, J. A., Strausbaugh, L. D., Gamonal, J., and Diaz, P. I. (2013). The subgingival microbiome in health and periodontitis and its relationship with community biomass and inflammation. *The ISME journal*, 7(5):1016–1025.

Adler, C. J., Dobney, K., Weyrich, L. S., Kaidonis, J., Walker, A. W., Haak, W., Bradshaw, C. J. A., Townsend, G., Sołtysiak, A., Alt, K. W., Parkhill, J., and Cooper, A. (2013). Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genetics*, 45(4):450–5, 455e1.

Akcalı, A. and Lang, N. P. (2017). Dental calculus: the calcified biofilm and its role in disease development. *Periodontology 2000*, 76(1):109–115.

Al-Ahmad, A., Wunder, A., Auschill, T. M., Follo, M., Braun, G., Hellwig, E., and Arweiler, N. B. (2007). The in vivo dynamics of *Streptococcus* spp., *Actinomyces naeslundii*, *Fusobacterium nucleatum* and *Veillonella* spp. in dental plaque biofilm as analysed by five-colour multiplex fluorescence in situ hybridization. *Journal of medical microbiology*, 56(Pt 5):681–687.

Allentoft, M. E., Collins, M., Harker, D., Haile, J., Oskam, C. L., Hale, M. L., Campos, P. F., Samaniego, J. A., Gilbert, M. T. P., Willerslev, E., Zhang, G., Scofield, R. P., Holdaway, R. N., and Bunce, M. (2012). The half-life of DNA in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748):4724–4733.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410.

Amato, K. R., G Sanders, J., Song, S. J., Nute, M., Metcalf, J. L., Thompson, L. R., Morton, J. T., Amir, A., J McKenzie, V., Humphrey, G., Gogul, G., Gaffney, J., L Baden, A., A O Britton, G., P Cuozzo, F., Di Fiore, A., J Dominy, N., L Goldberg, T., Gomez, A., Kowalewski, M. M., J Lewis, R., Link, A., L Sauther, M., Tecot, S., A White, B., E Nelson, K., M Stumpf, R., Knight, R., and R Leigh, S. (2019a). Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes. *The ISME journal*, 13(3):576–587.

Amato, K. R., Mallott, E. K., McDonald, D., Dominy, N. J., Goldberg, T., Lambert, J. E., Swedell, L., Metcalf, J. L., Gomez, A., Britton, G. A. O., Stumpf, R. M., Leigh, S. R., and Knight, R. (2019b). Convergence of human and old world monkey gut microbiomes demonstrates the importance of human ecology over phylogeny. *Genome biology*, 20(1):201.

Anagnostou, P., Capocasa, M., Milia, N., Sanna, E., Battaggia, C., Luzi, D., and Destro Bisol, G. (2015). When data sharing gets close to 100%: what human paleogenetics can teach the open science movement. *PLoS One*, 10(3):e0121409.

Anderson, A. C., Rothballer, M., Altenburger, M. J., Woelber, J. P., Karygianni, L., Vach, K., Hellwig, E., and Al-Ahmad, A. (2020). Long-Term fluctuation of oral biofilm microbiota following different dietary phases. *Applied and environmental microbiology*, 86(20).

Andrades Valtueña, A., Mittnik, A., Key, F. M., Haak, W., Allmäe, R., Belinskij, A., Daubaras, M., Feldman, M., Jankauskas, R., Janković, I., Massy, K., Novak, M., Pfrengle, S., Reinhold, S., Šlaus, M., Spyrou, M. A., Szécsényi-Nagy, A., Tõrv, M., Hansen, S., Bos, K. I., Stockhammer, P. W., Herbig, A., and Krause, J. (2017). The stone age plague and its persistence in Eurasia. *Current biology*, 27(23):3683–3691.e8.

Arning, N. and Wilson, D. J. (2020). The past, present and future of ancient bacterial DNA. *Microbial genomics*, 6(7).

Austin, R. M., Sholts, S. B., Williams, L., Kistler, L., and Hofman, C. A. (2019). Opinion: To curate the molecular past, museums need a carefully considered set of best practices. *Proceedings of the National Academy of Sciences of the United States of America*, 116(5):1471–1474.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, 533(7604):452–454.

Barberán, A., Casamayor, E. O., and Fierer, N. (2014). The microbial contribution to macroecology. *Frontiers in microbiology*, 5:203.

Benezra, A. (2020). Race in the microbiome. *Science, technology & human values*, 45(5):877–902.

Berg, G., Rybakova, D., Fischer, D., Cernava, T., Vergès, M.-C. C., Charles, T., Chen, X., Cocolin, L., Eversole, K., Corral, G. H., Kazou, M., Kinkel, L., Lange, L., Lima, N., Loy, A., Macklin, J. A., Maguin, E., Mauchline, T., McClure, R., Mitter, B., Ryan, M., Sarand, I., Smidt, H., Schelkle, B., Roume, H., Kiran, G. S., Selvin, J., Souza, R. S. C. d., van Overbeek, L., Singh, B. K., Wagner, M., Walsh, A., Sessitsch, A., and Schloter, M. (2020). Microbiome definition re-visited: old concepts and new challenges. *Microbiome*, 8(1):103.

Bik, E. M., Long, C. D., Armitage, G. C., Loomer, P., Emerson, J., Mongodin, E. F., Nelson, K. E., Gill, S. R., Fraser-Liggett, C. M., and Relman, D. A. (2010). Bacterial diversity in the oral cavity of 10 healthy individuals. *The ISME journal*, 4(8):962–974.

Bisgaard, H., Li, N., Bonnelykke, K., Chawes, B. L. K., Skov, T., Paludan-Müller, G., Stokholm, J., Smith, B., and Krogfelt, K. A. (2011). Reduced diversity of the intestinal microbiota during infancy is associated with increased risk of allergic disease at school age. *The Journal of allergy and clinical immunology*, 128(3):646–52.e1–5.

Blaser, M. J. (2016). Antibiotic use and its consequences for the normal microbiome. *Science*, 352(6285):544–545.

Blaser, M. J. and Falkow, S. (2009). What are the consequences of the disappearing human microbiota? *Nature reviews. Microbiology*, 7(12):887–894.

Bleasdale, M., Richter, K. K., Janzen, A., Brown, S., Scott, A., Zech, J., Wilkin, S., Wang, K., Schiffels, S., Desideri, J., Besse, M., Reinold, J., Saad, M., Babiker, H., Power, R. C., Ndiema, E., Ogola, C., Manthi, F. K., Zahir, M., Petraglia, M., Trachsel, C., Nanni, P., Grossmann, J., Hendy, J., Crowther, A., Roberts, P., Goldstein, S. T., and Boivin, N. (2021). Ancient proteins provide evidence of dairy consumption in eastern Africa. *Nature communications*, 12(1):632.

Blekhman, R., Goodrich, J. K., Huang, K., Sun, Q., Bukowski, R., Bell, J. T., Spector, T. D., Keinan, A., Ley, R. E., Gevers, D., and Clark, A. G. (2015). Host genetic variation impacts microbiome composition across human body sites. *Genome biology*, 16:191.

Bollongino, R., Tresset, A., and Vigne, J.-D. (2008). Environment and excavation: Pre-lab impacts on ancient DNA analyses. *Comptes rendus. Palevol*, 7(2–3):91–98.

Bordenstein, S. R. and Theis, K. R. (2015). Host biology in light of the microbiome: Ten principles of holobionts and hologenomes. *PLoS biology*, 13(8):e1002226.

Borry, M., Hübner, A., Rohrlach, A. B., and Warinner, C. (2021). PyDamage: automated ancient damage identification and estimation for contigs in ancient DNA de novo assembly. *PeerJ*, 9:e11845.

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., Forrest, S. A., Bryant, J. M., Harris, S. R., Schuenemann, V. J., Campbell, T. J., Majander, K., Wilbur, A. K., Guichon, R. A., Wolfe Steadman, D. L., Cook, D. C., Niemann, S., Behr, M. A., Zumarraga, M., Bastida, R., Huson, D., Nieselt, K., Young, D., Parkhill, J., Buikstra, J. E., Gagneux, S., Stone, A. C., and Krause, J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, 514(7523):494–497.

Bos, K. I., Schuenemann, V. J., Golding, G. B., Burbano, H. A., Waglechner, N., Coombes, B. K., McPhee, J. B., DeWitte, S. N., Meyer, M., Schmedes, S., Wood, J., Earn, D. J. D., Herring, D. A., Bauer, P., Poinar, H. N., and Krause, J. (2011). A draft genome of *Yersinia pestis* from victims of the black death. *Nature*, 478(7370):506–510.

Bravo-Lopez, M., Villa-Islas, V., Rocha Arriaga, C., Villaseñor-Altamirano, A. B., Guzmán-Solís, A., Sandoval-Velasco, M., Wesp, J. K., Alcantara, K., López-Corral, A., Gómez-Valdés, J., Mejía, E., Herrera, A., Meraz-Moreno, A., Moreno-Cabrera, M. d. l. L., Moreno-Estrada, A., Nieves-Colón, M. A., Olvera, J., Pérez-Pérez, J., Iversen, K. H., Rasmussen, S., Sandoval, K., Zepeda, G., and Ávila-Arcos, M. C. (2020). Paleogenomic insights into the red complex bacteria *Tannerella forsythia* in Pre-Hispanic and colonial individuals from Mexico. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 375(1812):20190580.

Brealey, J. C., Leitão, H. G., Hofstede, T., Kalthoff, D. C., and Guschanski, K. (2021). The oral microbiota of wild bears in sweden reflects the history of antibiotic use by humans. *Current biology*, In Press.

Brealey, J. C., Leitão, H. G., van der Valk, T., Xu, W., Bougiouri, K., Dalén, L., and Guschanski, K. (2020). Dental calculus as a tool to study the evolution of the mammalian oral microbiome. *Molecular biology and evolution*, 37(10):3003–3022.

Breitwieser, F. P., Lu, J., and Salzberg, S. L. (2019). A review of methods and databases for metagenomic classification and assembly. *Briefings in bioinformatics*, 20(4):1125–1136.

Briggs, A. W., Stenzel, U., Johnson, P. L. F., Green, R. E., Kelso, J., Prüfer, K., Meyer, M., Krause, J., Ronan, M. T., Lachmann, M., and Pääbo, S. (2007). Patterns of damage in genomic DNA sequences from a neandertal. *Proceedings of the National Academy of Sciences of the United States of America*, 104(37):14616–14621.

Broussard, J. L. and Devkota, S. (2016). The changing microbial landscape of western society: Diet, dwellings and discordance. *Molecular metabolism*, 5(9):737–742.

Buchfink, B., Xie, C., and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature methods*, 12(1):59–60.

Cárdenas, Y. O. A., Neuenschwander, S., and Malaspinas, A.-S. (2021). Benchmarking metagenomics classifiers on ancient viral DNA: a simulation study. *bioRxiv*, page 2021.04.30.442132. Preprint.

Carmody, R. N., Sarkar, A., and Reese, A. T. (2021). Gut microbiota through an evolutionary lens. *Science*, 372(6541):462–463.

Carmody, R. N., Weintraub, G. S., and Wrangham, R. W. (2011). Energetic consequences of thermal and nonthermal food processing. *Proceedings of the National Academy of Sciences of the United States of America*, 108(48):19199–19203.

Carmody, R. N. and Wrangham, R. W. (2009). The energetic significance of cooking. *Journal of human evolution*, 57(4):379–391.

Carroll, S. R., Garba, I., Figueroa-Rodríguez, O. L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J. D., Anderson, J., and Hudson, M. (2020). The CARE principles for indigenous data governance. *Data science journal*, 19.

Carroll, S. R., Herczog, E., Hudson, M., Russell, K., and Stall, S. (2021). Operationalizing the CARE and FAIR principles for indigenous data futures. *Scientific data*, 8(1):108.

Casals-Pascual, C., González, A., Vázquez-Baeza, Y., Song, S. J., Jiang, L., and Knight, R. (2020). Microbial diversity in clinical microbiome studies: Sample size and statistical power considerations. *Gastroenterology*, 158(6):1524–1528.

Charlier, P., Gaultier, F., and Héry-Arnaud, G. (2019). Interbreeding between Neanderthals and modern humans: Remarks and methodological dangers of a dental calculus microbiome analysis. *Journal of human evolution*, 126:124–126.

Charlton, S., Ramsøe, A., Collins, M., Craig, O. E., Fischer, R., Alexander, M., and Speller, C. F. (2019). New insights into Neolithic milk consumption through proteomic analysis of dental calculus. *Archaeological and anthropological sciences*, 11(11):6183–6196.

Chen, K. and Pachter, L. (2005). Bioinformatics for whole-genome shotgun sequencing of microbial communities. *PLoS computational biology*, 1(2):106–112.

Chen, T., Yu, W.-H., Izard, J., Baranova, O. V., Lakshmanan, A., and Dewhirst, F. E. (2010). The human oral microbiome database: a web accessible resource for investigating oral microbe taxonomic and genomic information. *Database: the journal of biological databases and curation*, 2010:baq013.

Cho, I. and Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature reviews. Genetics*, 13(4):260–270.

Claesson, M. J., Jeffery, I. B., Conde, S., Power, S. E., O'Connor, E. M., Cusack, S., Harris, H. M. B., Coakley, M., Lakshminarayanan, B., O'Sullivan, O., Fitzgerald, G. F., Deane, J., O'Connor, M., Harnedy, N., O'Connor, K., O'Mahony, D., van Sinderen, D., Wallace, M., Brennan, L., Stanton, C., Marchesi, J. R., Fitzgerald, A. P., Shanahan, F., Hill, C., Ross, R. P., and O'Toole, P. W. (2012). Gut microbiota composition correlates with diet and health in the elderly. *Nature*, 488(7410):178–184.

Clayton, J. B., Vangay, P., Huang, H., Ward, T., Hillmann, B. M., Al-Ghalith, G. A., Travis, D. A., Long, H. T., Tuan, B. V., Minh, V. V., Cabana, F., Nadler, T., Toddes, B., Murphy, T., Glander, K. E., Johnson, T. J., and Knights, D. (2016). Captivity humanizes the primate microbiome.

Proceedings of the National Academy of Sciences of the United States of America, 113(37):10376–10381.

Colombo, A. P. V. and Tanner, A. C. R. (2019). The role of bacterial biofilms in dental caries and periodontal and peri-implant diseases: A historical perspective. *Journal of dental research*, 98(4):373–385.

Compeau, P. E. C., Pevzner, P. A., and Tesler, G. (2011). How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991.

Cooper, A. and Poinar, H. N. (2000). Ancient DNA: Do it right or not at all. *Science*, 289(5482):1139–1139.

Costello, E. K., Stagaman, K., Dethlefsen, L., Bohannan, B. J. M., and Relman, D. A. (2012). The application of ecological theory toward an understanding of the human microbiome. *Science*, 336(6086):1255–1262.

Council, S. E., Savage, A. M., Urban, J. M., Ehlers, M. E., Skene, J. H. P., Platt, M. L., Dunn, R. R., and Horvath, J. E. (2016). Diversity and evolution of the primate skin microbiome. *Proceedings of the Royal Society B: Biological Sciences*, 283(1822).

Cristiani, E., Radini, A., Edinborough, M., and Borić, D. (2016). Dental calculus reveals Mesolithic foragers in the Balkans consumed domesticated plant foods. *Proceedings of the National Academy of Sciences of the United States of America*, 113(37):10298–10303.

Cummings, L. S. and Magennis, A. (1997). A phytolith and starch record of food and grit in mayan human tooth tartar. *Estado actual de los estudios de*, pages 211–218.

Curtis, M. A., Diaz, P. I., and Van Dyke, T. E. (2020). The role of the microbiota in periodontal disease. *Periodontology 2000*, 83(1):14–25.

D'Agostino, A., Gismondi, A., Di Marco, G., Lo Castro, M., Olevano, R., Cinti, T., Leonardi, D., and Canini, A. (2019). Lifestyle of a roman imperial community: ethnobotanical evidence from dental calculus of the ager curensis inhabitants. *Journal of ethnobiology and ethnomedicine*, 15(1):62.

Darveau, R. P., Belton, C. M., Reife, R. A., and Lamont, R. J. (1998). Local chemokine paralysis, a novel pathogenic mechanism for *Porphyromonas gingivalis*. *Infection and immunity*, 66(4):1660–1665.

Davenport, E. R., Sanders, J. G., Song, S. J., Amato, K. R., Clark, A. G., and Knight, R. (2017). The human microbiome in evolution. *BMC biology*, 15(1):127.

Davis, N. M., Proctor, D. M., Holmes, S. P., Relman, D. A., and Callahan, B. J. (2018). Simple statistical identification and removal of contaminant sequences in marker-gene and metagenomics data. *Microbiome*, 6(1):226.

De Filippis, F., Vannini, L., La Storia, A., Laghi, L., Piombino, P., Stellato, G., Serrazanetti, D. I., Gozzi, G., Turroni, S., Ferrocino, I., Lazzi, C., Di Cagno, R., Gobbetti, M., and Ercolini, D. (2014). The same microbiota and a potentially discriminant metabolome in the saliva of omnivore, ovo-lacto-vegetarian and vegan individuals. *PLoS One*, 9(11):e112373.

de La Fuente, C., Flores, S., and Moraga, M. (2013). DNA from human ancient bacteria: a novel source of genetic evidence from archaeological dental calculus. *Archaeometry*, 55(4):767–778.

Debelius, J., Song, S. J., Vazquez-Baeza, Y., Xu, Z. Z., Gonzalez, A., and Knight, R. (2016). Tiny microbes, enormous impacts: what matters in gut microbiome studies? *Genome biology*, 17(1):217.

Demmitt, B. A., Corley, R. P., Huibregtse, B. M., Keller, M. C., Hewitt, J. K., McQueen, M. B., Knight, R., McDermott, I., and Krauter, K. S. (2017). Genetic influences on the human oral microbiome. *BMC genomics*, 18(1):659.

Dent, V. E. (1979). The bacteriology of dental plaque from a variety of zoo-maintained mammalian species. *Archives of oral biology*, 24(4):277–282.

DePristo, M. A., Banks, E., Poplin, R., Garimella, K. V., Maguire, J. R., Hartl, C., Philippakis, A. A., del Angel, G., Rivas, M. A., Hanna, M., McKenna, A., Fennell, T. J., Kernytsky, A. M., Sivachenko, A. Y., Cibulskis, K., Gabriel, S. B., Altshuler, D., and Daly, M. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics*, 43(5):491–498.

Der Sarkissian, C., Ermini, L., Jónsson, H., Alekseev, A. N., Crubezy, E., Shapiro, B., and Orlando, L. (2014). Shotgun microbial profiling of fossil remains. *Molecular ecology*, 23(7):1780–1798.

Dewhirst, F. E., Chen, T., Izard, J., Paster, B. J., Tanner, A. C. R., Yu, W.-H., Lakshmanan, A., and Wade, W. G. (2010). The human oral microbiome. *Journal of bacteriology*, 192(19):5002–5017.

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature biotechnology*, 35(4):316–319.

Diaz, P. I., Chalmers, N. I., Rickard, A. H., Kong, C., Milburn, C. L., Palmer, Jr, R. J., and Kolenbrander, P. E. (2006). Molecular characterization of subject-specific oral microflora during initial colonization of enamel. *Applied and environmental microbiology*, 72(4):2837–2848.

Diaz, P. I. and Valm, A. M. (2020). Microbial interactions in oral communities mediate emergent biofilm properties. *Journal of dental research*, 99(1):18–25.

Dimopoulos, E. A., Carmagnini, A., Velsko, I. M., Warinner, C., Larson, G., Frantz, L. A. F., and Irving-Pease, E. K. (2020). HAYSTAC: A bayesian framework for robust and rapid species identification in high-throughput sequencing data. *bioRxiv*, page 2020.12.16.419085. Preprint.

Dominguez Bello, M. G., Knight, R., Gilbert, J. A., and Blaser, M. J. (2018). Preserving microbial diversity. *Science*, 362(6410):33–34.

Dorrestein, P. C., Mazmanian, S. K., and Knight, R. (2014). Finding the missing links among metabolites, microbes, and the host. *Immunity*, 40(6):824–832.

Dunn, R. R., Amato, K. R., Archie, E. A., Arandjelovic, M., Crittenden, A. N., and Nichols, L. M. (2020). The internal, external and extended microbiomes of hominins. *Frontiers in Ecology and Evolution*, 8:25.

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Ehrlich, S. D. (2011). MetaHIT: The european union project on metagenomics of the human intestinal tract. In Nelson, K. E., editor, *Metagenomics of the Human Body*, pages 307–316. Springer New York, New York, NY.

Eisenhofer, R., Kanzawa-Kiriyama, H., Shinoda, K.-I., and Weyrich, L. S. (2020). Investigating the demographic history of Japan using ancient oral microbiota. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 375(1812):20190578.

Eisenhofer, R. and Weyrich, L. S. (2019). Assessing alignment-based taxonomic classification of ancient microbial DNA. *PeerJ*, 7:e6594.

Emmons, A. L., Mundorff, A. Z., Keenan, S. W., Davoren, J., Andronowski, J., Carter, D. O., and DeBruyn, J. M. (2020). Characterizing the postmortem human bone microbiome from surface-decomposed remains. *PLoS One*, 15(7):e0218636.

Eren, A. M., Borisy, G. G., Huse, S. M., and Mark Welch, J. L. (2014). Oligotyping analysis of the human oral microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 111(28):E2875–84.

Escapa, I. F., Chen, T., Huang, Y., Gajare, P., Dewhirst, F. E., and Lemon, K. P. (2018). New insights into human nostril microbiome from the expanded human oral microbiome database (eHOMD): a resource for the microbiome of the human aerodigestive tract. *mSystems*, 3(6).

Ewels, P. A., Peltzer, A., Fillinger, S., Patel, H., Alneberg, J., Wilm, A., Garcia, M. U., Di Tommaso, P., and Nahnsen, S. (2020). The nf-core framework for community-curated bioinformatics pipelines. *Nature biotechnology*, 38(3):276–278.

Fagernäs, Z., García-Collado, M. I., Hendy, J., Hofman, C. A., Speller, C., Velsko, I., and Warinner, C. (2020). A unified protocol for simultaneous extraction of DNA and proteins from archaeological dental calculus. *Journal of archaeological science*, 118:105135.

Fellows Yates, J. A., Andrades Valtueña, A., Vågene, Å. J., Cribdon, B., Velsko, I. M., Borry, M., Bravo-Lopez, M. J., Fernandez-Guerra, A., Green, E. J., Ramachandran, S. L., Heintzman, P. D., Spyrou, M. A., Hübner, A., Gancz, A. S., Hider, J., Allshouse, A. F., Zaro, V., and Warinner, C. (2021a). Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Scientific data*, 8(1):31.

Fellows Yates, J. A., Lamnidis, T. C., Borry, M., Andrades Valtueña, A., Fagernäs, Z., Clayton, S., Garcia, M. U., Neukamm, J., and Peltzer, A. (2021b). Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ*, 9:e10947.

Fellows Yates, J. A., Velsko, I. M., Aron, F., Posth, C., Hofman, C. A., Austin, R. M., Parker, C. E., Mann, A. E., Nägele, K., Arthur, K. W., Arthur, J. W., Bauer, C. C., Crevecoeur, I., Cupillard, C., Curtis, M. C., Dalén, L., Bonilla, M. D.-Z., Carlos Díez Fernández-Lomana, J., Drucker, D. G., Escrivá, E. E., Francken, M., Gibbon, V. E., González Morales, M. R., Mateu, A. G., Harvati, K., Henry, A. G., Humphrey, L., Menéndez, M., Mihailović, D., Peresani, M., Moroder, S. R., Roksandic, M., Rougier, H., Sázelová, S., Stock, J. T., Straus, L. G., Svoboda, J., Teßmann, B., Walker, M. J., Power, R. C., Lewis, C. M., Sankaranarayanan, K., Guschanski, K., Wrangham, R. W., Dewhirst, F. E., Salazar-García, D. C., Krause, J., Herbig, A., and Warinner, C. (2021c). The evolution and changing ecology of the African hominid oral microbiome. *Proceedings of the National Academy of Sciences of the United States of America*, 118(20):e2021655118.

Fotakis, A. K., Denham, S. D., Mackie, M., Orbegozo, M. I., Mylopotamitaki, D., Gopalakrishnan, S., Sicheritz-Pontén, T., Olsen, J. V., Cappellini, E., Zhang, G., Christophersen, A., Gilbert, M. T. P., and Vågene, Å. J. (2020). Multi-omic detection of *Mycobacterium leprae* in archaeological human dental calculus. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 375(1812):20190584.

Fox, C. L., Juan, J., and Albert, R. M. (1996). Phytolith analysis on dental calculus, enamel surface, and burial soil: information about diet and paleoenvironment. *American journal of physical anthropology*, 101(1):101–113.

Franzosa, E. A., McIver, L. J., Rahnavard, G., Thompson, L. R., Schirmer, M., Weingart, G., Lipson, K. S., Knight, R., Caporaso, J. G., Segata, N., and Huttenhower, C. (2018). Species-level functional profiling of metagenomes and metatranscriptomes. *Nature methods*, 15(11):962–968.

Frederico, L. A., Kunkel, T. A., and Shaw, B. R. (1990). A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, 29(10):2532–2537.

Fu, Q., Hajdinjak, M., Moldovan, O. T., Constantin, S., Mallick, S., Skoglund, P., Patterson, N., Rohland, N., Lazaridis, I., Nickel, B., Viola, B., Prüfer, K., Meyer, M., Kelso, J., Reich, D., and Pääbo, S. (2015). An early modern human from Romania with a recent Neanderthal ancestor. *Nature*, 524(7564):216–219.

Furtwängler, A., Neukamm, J., Böhme, L., Reiter, E., Vollstedt, M., Arora, N., Singh, P., Cole, S. T., Knauf, S., Calvignac-Spencer, S., Krause-Kyora, B., Krause, J., Schuenemann, V. J., and Herbig, A. (2020). Comparison of target enrichment strategies for ancient pathogen DNA. *BioTechniques*, 69(6):455–459.

Fuss, J., Uhlig, G., and Böhme, M. (2018). Earliest evidence of caries lesion in hominids reveal sugar-rich diet for a Middle Miocene dryopithecine from Europe. *PLoS One*, 13(8):e0203307.

Gallego Llorente, M., Jones, E. R., Eriksson, A., Siska, V., Arthur, K. W., Arthur, J. W., Curtis, M. C., Stock, J. T., Coltorti, M., Pieruccini, P., Stretton, S., Brock, F., Higham, T., Park, Y., Hofreiter, M., Bradley, D. G., Bhak, J., Pinhasi, R., and Manica, A. (2015). Ancient Ethiopian genome reveals extensive Eurasian admixture throughout the African continent. *Science*, 350(6262):820–822.

Gamba, C., Jones, E. R., Teasdale, M. D., McLaughlin, R. L., Gonzalez-Fortes, G., Mattiangeli, V., Domboróczki, L., Kővári, I., Pap, I., Anders, A., Whittle, A., Dani, J., Raczky, P., Higham, T. F. G., Hofreiter, M., Bradley, D. G., and Pinhasi, R. (2014). Genome flux and stasis in a five millennium transect of European prehistory. *Nature communications*, 5:5257.

Gansauge, M.-T., Gerber, T., Glocke, I., Korlevic, P., Lippik, L., Nagel, S., Riehl, L. M., Schmidt, A., and Meyer, M. (2017). Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic acids research*, 45(10):e79.

Gauthier, J., Vincent, A. T., Charette, S. J., and Derome, N. (2019). A brief history of bioinformatics. *Briefings in bioinformatics*, 20(6):1981–1996.

GBD 2017 Disease and Injury Incidence and Prevalence Collaborators (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 392(10159):1789–1858.

Geber, J., Tromp, M., Scott, A., Bouwman, A., Nanni, P., Grossmann, J., Hendy, J., and Warinner, C. (2019). Relief food subsistence revealed by microparticle and proteomic analyses of dental calculus from victims of the Great Irish Famine. *Proceedings of the National Academy of Sciences of the United States of America*, 116(39):19380–19385.

Gomez, A., Espinoza, J. L., Harkins, D. M., Leong, P., Saffery, R., Bockmann, M., Torralba, M., Kuelbs, C., Kodukula, R., Inman, J., Hughes, T., Craig, J. M., Highlander, S. K., Jones, M. B., Dupont, C. L., and Nelson, K. E. (2017). Host genetic control of the oral microbiome in health and disease. *Cell host & microbe*, 22(3):269–278.e3.

Gomez, A., Petrzelkova, K., Yeoman, C. J., Vlckova, K., Mrázek, J., Koppova, I., Carbonero, F., Ulanov, A., Modry, D., Todd, A., Torralba, M., Nelson, K. E., Gaskins, H. R., Wilson, B., Stumpf, R. M., White, B. A., and Leigh, S. R. (2015). Gut microbiome composition and metabolomic profiles of wild western lowland gorillas (*Gorilla gorilla gorilla*) reflect host ecology. *Molecular ecology*, 24(10):2551–2565.

Greenbaum, S., Greenbaum, G., Moran-Gilad, J., and Weintraub, A. Y. (2019). Ecological dynamics of the vaginal microbiome in relation to health and disease. *American journal of obstetrics and gynecology*, 220(4):324–335.

Gruening, B., Sallou, O., Moreno, P., da Veiga Leprevost, F., Ménager, H., Søndergaard, D., Röst, H., Sachsenberg, T., O'Connor, B., Madeira, F., Dominguez Del Angel, V., Crusoe, M. R., Varma, S., Blankenberg, D., Jimenez, R. C., BioContainers Community, and Perez-Riverol, Y. (2018). Recommendations for the packaging and containerizing of bioinformatics software. *F1000Research*, 7.

Grunenwald, A., Keyser, C., Sautereau, A. M., Crubézy, E., Ludes, B., and Drouet, C. (2014). Adsorption of DNA on biomimetic apatites: Toward the understanding of the role of bone and tooth mineral on the preservation of ancient DNA. *Applied surface science*, 292:867–875.

Haile, J., Holdaway, R., Oliver, K., Bunce, M., Gilbert, M. T. P., Nielsen, R., Munch, K., Ho, S. Y. W., Shapiro, B., and Willerslev, E. (2007). Ancient DNA chronology within sediment deposits: are paleobiological reconstructions possible and is DNA leaching a factor? *Molecular biology and evolution*, 24(4):982–989.

Hajdinjak, M., Mafessoni, F., Skov, L., Vernot, B., Hübner, A., Fu, Q., Essel, E., Nagel, S., Nickel, B., Richter, J., Moldovan, O. T., Constantin, S., Endarova, E., Zahariev, N., Spasov, R., Welker, F., Smith, G. M., Sinet-Mathiot, V., Paskulin, L., Fewlass, H., Talamo, S., Rezek, Z., Sirakova, S., Sirakov, N., McPherron, S. P., Tsanova, T., Hublin, J.-J., Peter, B. M., Meyer, M., Skoglund, P., Kelso, J., and Pääbo, S. (2021). Initial Upper Palaeolithic humans in Europe had recent Neanderthal ancestry. *Nature*, 592(7853):253–257.

Hajishengallis, G., Darveau, R. P., and Curtis, M. A. (2012). The keystone-pathogen hypothesis. *Nature reviews. Microbiology*, 10(10):717–725.

Halfvarson, J., Brislawn, C. J., Lamendella, R., Vázquez-Baeza, Y., Walters, W. A., Bramer, L. M., D'Amato, M., Bonfiglio, F., McDonald, D., Gonzalez, A., McClure, E. E., Dunklebarger, M. F., Knight, R., and Jansson, J. K. (2017). Dynamics of the human gut microbiome in inflammatory bowel disease. *Nature microbiology*, 2:17004.

Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, 5(10):R245–9.

Hannig, C. and Hannig, M. (2009). The oral cavity—a key system to understand substratum-dependent bioadhesion on solid surfaces in man. *Clinical oral investigations*, 13(2):123–139.

Hardy, K., Buckley, S., Collins, M. J., Estalrrich, A., Brothwell, D., Copeland, L., García-Tabernero, A., García-Vargas, S., de la Rasilla, M., Lalueza-Fox, C., Huguet, R., Bastir, M., Santamaría, D., Madella, M., Wilson, J., Cortés, A. F., and Rosas, A. (2012). Neanderthal medics? Evidence for food, cooking, and medicinal plants entrapped in dental calculus. *Die Naturwissenschaften*, 99(8):617–626.

Hardy, K., Radini, A., Buckley, S., Blasco, R., Copeland, L., Burjachs, F., Girbal, J., Yll, R., Carbonell, E., and de Castro, J. M. B. (2017). Diet and environment 1.2 million years ago revealed through analysis of dental calculus from Europe's oldest hominin at Sima del Elefante, Spain. *The Science of Nature*, 104(1-2):2.

Hardy, K., Radini, A., Buckley, S., Sarig, R., Copeland, L., Gopher, A., and Barkai, R. (2016). Dental calculus reveals potential respiratory irritants and ingestion of essential plant-based nutrients at Lower Palaeolithic Qesem Cave Israel. *Quaternary international*, 398(0):129–135.

Harkins, K. M., Buikstra, J. E., Campbell, T., Bos, K. I., Johnson, E. D., Krause, J., and Stone, A. C. (2015). Screening ancient tuberculosis with qPCR: challenges and opportunities. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 370(1660):20130622.

Hayashizaki, J., Ban, S., Nakagaki, H., Okumura, A., Yoshii, S., and Robinson, C. (2008). Site specific mineral composition and microstructure of human supra-gingival dental calculus. *Archives of oral biology*, 53(2):168–174.

Hendy Jessica, Warinner Christina, Bouwman Abigail, Collins Matthew J., Fiddyment Sarah, Fischer Roman, Hagan Richard, Hofman Courtney A., Holst Malin, Chaves Eros, Klaus Lauren, Larson Greger, Mackie Meaghan, McGrath Krista, Mundorff Amy Z., Radini Anita, Rao Huiyun, Trachsel Christian, Velsko Irina M., and Speller Camilla F. (2018). Proteomic evidence of dietary sources in ancient dental calculus. *Proceedings of the Royal Society B: Biological Sciences*, 285(1883):20180977.

Henry, A. G., Brooks, A. S., and Piperno, D. R. (2011). Microfossils in calculus demonstrate consumption of plants and cooked foods in Neanderthal diets (Shanidar III, Iraq; Spy I and II, Belgium). *Proceedings of the National Academy of Sciences of the United States of America*, 108(2):486–491.

Henry, A. G., Brooks, A. S., and Piperno, D. R. (2014). Plant foods and the dietary ecology of Neanderthals and early modern humans. *Journal of human evolution*, 69:44–54.

Henry, A. G. and Piperno, D. R. (2008). Using plant microfossils from dental calculus to recover human diet: a case study from Tell al-Raqā'i, Syria. *Journal of archaeological science*, 35(7):1943–1950.

Henry, A. G., Ungar, P. S., Passey, B. H., Sponheimer, M., Rossouw, L., Bamford, M., Sandberg, P., de Ruiter, D. J., and Berger, L. (2012). The diet of *Australopithecus sediba*. *Nature*, 487(7405):90–93.

Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., and Huson, D. H. (2016). MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman. *bioRxiv*, page 050559. Preprint.

Herd, P., Palloni, A., Rey, F., and Dowd, J. B. (2018). Social and population health science approaches to understand the human microbiome. *Nature human behaviour*, 2(11):808–815.

Herrera, D., Contreras, A., Gamonal, J., Oteo, A., Jaramillo, A., Silva, N., Sanz, M., Botero, J. E., and León, R. (2008). Subgingival microbial profiles in chronic periodontitis patients from Chile, Colombia and Spain. *Journal of clinical periodontology*, 35(2):106–113.

Hershkovitz, I., Kelly, J., Latimer, B., Rothschild, B. M., Simpson, S., Polak, J., and Rosenberg, M. (1997). Oral bacteria in Miocene *Sivapithecus*. *Journal of human evolution*, 33(4):507–512.

Hojo, K., Nagaoka, S., Ohshima, T., and Maeda, N. (2009). Bacterial interactions in dental biofilm development. *Journal of dental research*, 88(11):982–990.

Hothorn, T. and Leisch, F. (2011). Case studies in reproducibility. *Briefings in bioinformatics*, 12(3):288–300.

Hübler, R., Key, F. M., Warinner, C., Bos, K. I., Krause, J., and Herbig, A. (2019). HOPS: automated detection and authentication of pathogen DNA in archaeological remains. *Genome Biology*, 20(1):280.

Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, 486(7402):207–214.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H.-J., and Tappu, R. (2016). MEGAN community edition - interactive exploration and analysis of Large-Scale microbiome sequencing data. *PLoS computational biology*, 12(6):e1004957.

Integrative HMP (iHMP) Research Network Consortium (2014). The integrative human microbiome project: dynamic analysis of microbiome-host omics profiles during periods of human health and disease. *Cell host & microbe*, 16(3):276–289.

Jacobson, D. K., Honap, T. P., Monroe, C., Lund, J., Houk, B. A., Novotny, A. C., Robin, C., Marini, E., and Lewis, Jr, C. M. (2020). Functional diversity of microbial ecologies estimated from ancient human coprolites and dental calculus. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, 375(1812):20190586.

Jakubovics, N. S., Goodman, S. D., Mashburn-Warren, L., Stafford, G. P., and Cieplik, F. (2021). The dental plaque biofilm matrix. *Periodontology 2000*, 86(1):32–56.

Jans, M. M. E., Nielsen-Marsh, C. M., Smith, C. I., Collins, M. J., and Kars, H. (2004). Characterisation of microbial attack on archaeological bone. *Journal of archaeological science*, 31(1):87–95.

Jeong, C., Wilkin, S., Amgalantugs, T., Bouwman, A. S., Taylor, W. T. T., Hagan, R. W., Bromage, S., Tsolmon, S., Trachsel, C., Grossmann, J., Littleton, J., Makarewicz, C. A., Krigbaum, J., Burri, M., Scott, A., Davaasambuu, G., Wright, J., Irmer, F., Myagmar, E., Boivin, N., Robbeets, M., Rühli, F. J., Krause, J., Frohlich, B., Hendy, J., and Warinner, C. (2018). Bronze age population dynamics and the rise of dairy pastoralism on the eastern eurasian steppe. *Proceedings of the National Academy of Sciences of the United States of America*, 115(48):E11248–E11255.

Jersie-Christensen, R. R., Lanigan, L. T., Lyon, D., Mackie, M., Belstrøm, D., Kelstrup, C. D., Fotakis, A. K., Willerslev, E., Lynnerup, N., Jensen, L. J., Cappellini, E., and Olsen, J. V. (2018). Quantitative metaproteomics of medieval dental calculus reveals individual oral health status. *Nature communications*, 9(1):4744.

Jiang, W.-X., Hu, Y.-J., Gao, L., He, Z.-Y., Zhu, C.-L., Ma, R., and Huang, Z.-W. (2015). The impact of various time intervals on the supragingival plaque dynamic core microbiome. *PLoS One*, 10(5):e0124631.

Jin, Y. and Yip, H.-K. (2002). Supragingival calculus: formation and control. *Critical reviews in oral biology and medicine*, 13(5):426–441.

Jones, S. J. (1972). A special relationship between spherical and filamentous microorganisms in mature human dental plaque. *Archives of oral biology*, 17(3):613–616.

Kazarina, A., Gerhards, G., Petersone-Gordina, E., Kimsis, J., Pole, I., Zole, E., Leonova, V., and Ranka, R. (2019). Analysis of the bacterial communities in ancient human bones and burial soil samples: Tracing the impact of environmental bacteria. *Journal of archaeological science*, 109:104989.

Kelly, B. J., Gross, R., Bittinger, K., Sherrill-Mix, S., Lewis, J. D., Collman, R. G., Bushman, F. D., and Li, H. (2015). Power and sample-size estimation for microbiome studies using pairwise distances and PERMANOVA. *Bioinformatics*, 31(15):2461–2468.

Kendall, C., Eriksen, A. M. H., Kontopoulos, I., Collins, M. J., and Turner-Walker, G. (2018). Diagenesis of archaeological bone and tooth. *Palaeogeography, palaeoclimatology, palaeoecology*, 491(Supplement C):21–37.

Keohane, D. M., Ghosh, T. S., Jeffery, I. B., Molloy, M. G., O'Toole, P. W., and Shanahan, F. (2020). Microbiome and health implications for ethnic minorities after enforced lifestyle changes. *Nature medicine*, 26(7):1089–1095.

Kilian, M., Chapple, I. L. C., Hannig, M., Marsh, P. D., Meuric, V., Pedersen, A. M. L., Tonetti, M. S., Wade, W. G., and Zaura, E. (2016). The oral microbiome - an update for oral healthcare professionals. *British dental journal*, 221(10):657–666.

Kim, D., Barraza, J. P., Arthur, R. A., Hara, A., Lewis, K., Liu, Y., Scisci, E. L., Hajishengallis, E., Whiteley, M., and Koo, H. (2020). Spatial mapping of polymicrobial communities reveals a precise biogeography associated with human dental caries. *Proceedings of the National Academy of Sciences of the United States of America*, 117(22):12375–12386.

Kim, D., Song, L., Breitwieser, F. P., and Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome research*, 26(12):1721–1729.

Kim, Y.-M., Poline, J.-B., and Dumas, G. (2018). Experimenting with reproducibility: a case study of robustness in bioinformatics. *GigaScience*, 7(7).

Kistler, L., Ware, R., Smith, O., Collins, M., and Allaby, R. G. (2017). A new model for ancient DNA decay based on paleogenomic meta-analysis. *Nucleic acids research*, 45(11):6310–6320.

Klepinger, L. L., Kuhn, J. J., and Thomas, Jr, J. (1977). Prehistoric dental calculus gives evidence for coca in early coastal Ecuador. *Nature*, 269(5628):506–507.

Knights, D., Kuczynski, J., Charlson, E. S., Zaneveld, J., Mozer, M. C., Collman, R. G., Bushman, F. D., Knight, R., and Kelley, S. T. (2011). Bayesian community-wide culture-independent microbial source tracking. *Nature methods*, 8(9):761–763.

Kolenbrander, P. E., Palmer, Jr, R. J., Periasamy, S., and Jakubovics, N. S. (2010). Oral multispecies biofilm development and the key role of cell-cell distance. *Nature reviews. Microbiology*, 8(7):471–480.

Kolenbrander, P. E., Palmer, Jr, R. J., Rickard, A. H., Jakubovics, N. S., Chalmers, N. I., and Diaz, P. I. (2006). Bacterial interactions and successions during plaque development. *Periodontology 2000*, 42:47–79.

Kontopoulos, I., Penkman, K., Mullin, V. E., Winkelbach, L., Unterländer, M., Scheu, A., Kreutzer, S., Hansen, H. B., Margaryan, A., Teasdale, M. D., Gehlen, B., Street, M., Lynnerup, N., Liritzis, I., Sampson, A., Papageorgopoulou, C., Allentoft, M. E., Burger, J., Bradley, D. G., and Collins, M. J. (2020). Screening archaeological bone for palaeogenetic and palaeoproteomic studies. *PLoS One*, 15(6):e0235146.

Kuczynski, J., Lauber, C. L., Walters, W. A., Parfrey, L. W., Clemente, J. C., Gevers, D., and Knight, R. (2011). Experimental and analytical tools for studying the human microbiome. *Nature reviews. Genetics*, 13(1):47–58.

Lamont, R. J., Koo, H., and Hajishengallis, G. (2018). The oral microbiota: dynamic communities and host interactions. *Nature reviews. Microbiology*, 16(12):745–759.

Lassalle, F., Spagnoletti, M., Fumagalli, M., Shaw, L., Dyble, M., Walker, C., Thomas, M. G., Bamberg Migliano, A., and Balloux, F. (2018). Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Molecular ecology*, 27(1):182–195.

Lazarevic, V., Whiteson, K., Hernandez, D., François, P., and Schrenzel, J. (2010). Study of inter- and intra-individual variations in the salivary microbiota. *BMC genomics*, 11:523.

Lederberg, J. and McCray, A. T. (2001). 'ome sweet 'omics– a genealogical treasury of words. *The Scientist*, 15:8.

Lee, S. H. and Baek, D. H. (2013). Characteristics of *Porphyromonas gingivalis* lipopolysaccharide in co-culture with *Fusobacterium nucleatum*. *Molecular oral microbiology*, 28(3):230–238.

Leonard, J. A., Shanks, O., Hofreiter, M., Kreuz, E., Hodges, L., Ream, W., Wayne, R. K., and Fleischer, R. C. (2007). Animal DNA in PCR reagents plagues ancient DNA research. *Journal of archaeological science*, 34(9):1361–1366.

Li, J., Nasidze, I., Quinque, D., Li, M., Horz, H.-P., André, C., Garriga, R. M., Halbwax, M., Fischer, A., and Stoneking, M. (2013). The saliva microbiome of *Pan* and *Homo*. *BMC microbiology*, 13:204.

Li, J., Quinque, D., Horz, H.-P., Li, M., Rzhetskaya, M., Raff, J., Hayes, M., and Stoneking, M. (2014). Comparative analysis of the human saliva microbiome from different climate zones: Alaska, Germany, and Africa. *BMC microbiology*, 14(1):316.

Lieverse, A. R. (1999). Diet and the aetiology of dental calculus. *International Journal of Osteoarchaeology*, 9(4):219–232.

Lindahl, T. (1993). Instability and decay of the primary structure of DNA. *Nature*, 362(6422):709–715.

Lindgreen, S., Krogh, A., and Pedersen, J. S. (2014). SNPest: a probabilistic graphical model for estimating genotypes. *BMC research notes*, 7:698.

Llamas, B., Valverde, G., Fehren-Schmitz, L., Weyrich, L. S., Cooper, A., and Haak, W. (2017). From the field to the laboratory: Controlling DNA contamination in human ancient DNA research in the high-throughput sequencing era. *STAR: Science & Technology of Archaeological Research*, 3(1):1–14.

Lobb, B., Tremblay, B. J.-M., Moreno-Hagelsieb, G., and Doxey, A. C. (2020). An assessment of genome annotation coverage across the bacterial tree of life. *Microbial genomics*, 6(3).

Loesche, W. J. (1976). Chemotherapy of dental plaque infections. *Oral sciences reviews*, 9:65–107.

Louvel, G., Der Sarkissian, C., Hanghøj, K., and Orlando, L. (2016). metaBIT, an integrative and automated metagenomic pipeline for analysing microbial profiles from high-throughput sequencing shotgun data. *Molecular ecology resources*, 16(6):1415–1427.

Lozupone, C. A., Stombaugh, J. I., Gordon, J. I., Jansson, J. K., and Knight, R. (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, 489(7415):220–230.

Lugli, G. A., Milani, C., Mancabelli, L., Turroni, F., Ferrario, C., Duranti, S., van Sinderen, D., and Ventura, M. (2017). Ancient bacteria of the Ötzi's microbiome: a genomic tale from the Copper Age. *Microbiome*, 5(1):5.

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., and Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature reviews. Genetics*, 17(11):704–714.

Mackie, M., Hendy, J., Lowe, A. D., Sperduti, A., Holst, M., Collins, M. J., and Speller, C. F. (2017). Preservation of the metaproteome: variability of protein preservation in ancient dental calculus. *Science and technology of archaeological research*, 3(1):74–86.

Manekar, S. C. and Sathe, S. R. (2018). A benchmark study of k-mer counting methods for high-throughput sequencing. *GigaScience*, 7(12).

Mann, A. E., Sabin, S., Ziesemer, K., Vågene, Å. J., Schroeder, H., Ozga, A. T., Sankaranarayanan, K., Hofman, C. A., Fellows Yates, J. A., Salazar-García, D. C., Frohlich, B., Aldenderfer, M., Hoogland, M., Read, C., Milner, G. R., Stone, A. C., Lewis, Jr, C. M., Krause, J., Hofman, C., Bos, K. I., and Warinner, C. (2018). Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Scientific reports*, 8(1):9822.

Marciniak, S. and Perry, G. H. (2017). Harnessing ancient genomes to study the history of human adaptation. *Nature reviews. Genetics*, 18(11):659–674.

Margaryan, A., Lawson, D. J., Sikora, M., Racimo, F., Rasmussen, S., Moltke, I., Cassidy, L. M., Jørsboe, E., Ingason, A., Pedersen, M. W., Korneliussen, T., Wilhelmson, H., Buś, M. M., de Barros Damgaard, P., Martiniano, R., Renaud, G., Bhérer, C., Moreno-Mayar, J. V., Fotakis, A. K., Allen, M., Allmäe, R., Molak, M., Cappellini, E., Scorrano, G., McColl, H., Buzhilova,

A., Fox, A., Albrechtsen, A., Schütz, B., Skar, B., Arcini, C., Falys, C., Jonson, C. H., Błaszczyk, D., Pezhemsky, D., Turner-Walker, G., Gestsdóttir, H., Lundstrøm, I., Gustin, I., Mainland, I., Potekhina, I., Muntoni, I. M., Cheng, J., Stenderup, J., Ma, J., Gibson, J., Peets, J., Gustafsson, J., Iversen, K. H., Simpson, L., Strand, L., Loe, L., Sikora, M., Florek, M., Vretemark, M., Redknap, M., Bajka, M., Pushkina, T., Søvsø, M., Grigoreva, N., Christensen, T., Kastholm, O., Uldum, O., Favia, P., Holck, P., Sten, S., Arge, S. V., Ellingvåg, S., Moiseyev, V., Bogdanowicz, W., Magnusson, Y., Orlando, L., Pentz, P., Jessen, M. D., Pedersen, A., Collard, M., Bradley, D. G., Jørkov, M. L., Arneborg, J., Lynnerup, N., Price, N., Gilbert, M. T. P., Allentoft, M. E., Bill, J., Sindbæk, S. M., Hedeager, L., Kristiansen, K., Nielsen, R., Werge, T., and Willerslev, E. (2020). Population genomics of the viking world. *Nature*, 585(7825):390–396.

Mark Welch, J. L., Dewhirst, F. E., and Borisy, G. G. (2019). Biogeography of the oral microbiome: The Site-Specialist hypothesis. *Annual review of microbiology*, 73:335–358.

Mark Welch, J. L., Ramírez-Puebla, S. T., and Borisy, G. G. (2020). Oral microbiome geography: Micron-Scale habitat and niche. *Cell host & microbe*, 28(2):160–168.

Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E., and Borisy, G. G. (2016). Biogeography of a human oral microbiome at the micron scale. *Proceedings of the National Academy of Sciences of the United States of America*, 113(6):E791–800.

Marquis, R. E. (1995). Oxygen metabolism, oxidative stress and acid-base physiology of dental plaque biofilms. *Journal of industrial microbiology*, 15(3):198–207.

Marsh, P. D. (1994). Microbial ecology of dental plaque and its significance in health and disease. *Advances in dental research*, 8(2):263–271.

Mason, M. R., Nagaraja, H. N., Camerlengo, T., Joshi, V., and Kumar, P. S. (2013). Deep sequencing identifies ethnicity-specific bacterial signatures in the oral microbiome. *PLoS One*, 8(10):e77287.

Mathieson, I., Alpaslan-Roodenberg, S., Posth, C., Szécsényi-Nagy, A., Rohland, N., Mallick, S., Olalde, I., Broomandkhoshbacht, N., Candilio, F., Cheronet, O., Fernandes, D., Ferry, M., Gamarra, B., Fortes, G. G., Haak, W., Harney, E., Jones, E., Keating, D., Krause-Kyora, B., Kucukkalipci, I., Michel, M., Mittnik, A., Nägele, K., Novak, M., Oppenheimer, J., Patterson, N., Pfrengle, S., Sirak, K., Stewardson, K., Vai, S., Alexandrov, S., Alt, K. W., Andreescu, R., Antonović, D., Ash, A., Atanassova, N., Bacvarov, K., Gusztáv, M. B., Bocherens, H., Bolus, M., Boroneanţ, A., Boyadzhiev, Y., Budnik, A., Burmaz, J., Chohadzhiev, S., Conard, N. J., Cottiaux, R., Čuka, M., Cupillard, C., Drucker, D. G., Elenski, N., Francken, M., Galabova, B., Ganetsovski, G., Gély, B., Hajdu, T., Handzhyiska, V., Harvati, K., Higham, T., Iliev, S., Janković, I., Karavanić, I., Kennett, D. J., Komšo, D., Kozak, A., Labuda, D., Lari, M., Lazar, C., Leppek, M., Leshtakov, K., Vetro, D. L., Los, D., Lozanov, I., Malina, M., Martini, F., McSweeney, K., Meller, H., Menđušić, M., Mirea, P., Moiseyev, V., Petrova, V., Price, T. D., Simalcsik, A., Sineo, L., Šlaus, M., Slavchev, V., Stanev, P., Starović, A., Szeniczey, T., Talamo, S., Teschler-Nicola, M., Thevenet, C., Valchev, I., Valentin, F., Vasilyev, S., Veljanovska, F., Venelinova, S., Veselovskaya, E., Viola, B., Virag, C., Zaninović, J., Zäuner, S., Stockhammer,

P. W., Catalano, G., Krauß, R., Caramelli, D., Zariņa, G., Gaydarska, B., Lillie, M., Nikitin, A. G., Potekhina, I., Papathanasiou, A., Borić, D., Bonsall, C., Krause, J., Pinhasi, R., and Reich, D. (2018). The genomic history of southeastern Europe. *Nature*, 555(7695):197–203.

McKenzie, V. J., Song, S. J., Delsuc, F., Prest, T. L., Oliverio, A. M., Korpita, T. M., Alexiev, A., Amato, K. R., Metcalf, J. L., Kowalewski, M., Avenant, N. L., Link, A., Di Fiore, A., Seguin-Orlando, A., Feh, C., Orlando, L., Mendelson, J. R., Sanders, J., and Knight, R. (2017). The effects of captivity on the mammalian gut microbiome. *Integrative and comparative biology*, 57(4):690–704.

Menzel, P., Ng, K. L., and Krogh, A. (2016). Fast and sensitive taxonomic classification for metagenomics with Kaiju. *Nature communications*, 7:11257.

Middleton, W. D. and Rovner, I. (1994). Extraction of opal phytoliths from herbivore dental calculus. *Journal of archaeological science*, 21(4):469–473.

Modi, A., Pisaneschi, L., Zaro, V., Vai, S., Vergata, C., Casalone, E., Caramelli, D., Moggi-Cecchi, J., Mariotti Lippi, M., and Lari, M. (2020). Combined methodologies for gaining much information from ancient dental calculus: testing experimental strategies for simultaneously analysing DNA and food residues. *Archaeological and anthropological sciences*, 12(1):10.

Moeller, A. H., Caro-Quintero, A., Mjungu, D., Georgiev, A. V., Lonsdorf, E. V., Muller, M. N., Pusey, A. E., Peeters, M., Hahn, B. H., and Ochman, H. (2016). Cospeciation of gut microbiota with hominids. *Science*, 353(6297):380–382.

Moeller, A. H., Degnan, P. H., Pusey, A. E., Wilson, M. L., Hahn, B. H., and Ochman, H. (2012). Chimpanzees and humans harbour compositionally similar gut enterotypes. *Nature communications*, 3:1179.

Moeller, A. H., Li, Y., Mpoudi Ngole, E., Ahuka-Mundeke, S., Lonsdorf, E. V., Pusey, A. E., Peeters, M., Hahn, B. H., and Ochman, H. (2014). Rapid changes in the gut microbiome during human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 111(46):16431–16435.

Moore, W. E. and Moore, L. V. (1994). The bacteria of periodontal diseases. *Periodontology 2000*, 5:66–77.

Muegge, B. D., Kuczynski, J., Knights, D., Clemente, J. C., González, A., Fontana, L., Henrissat, B., Knight, R., and Gordon, J. I. (2011). Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science*, 332(6032):970–974.

Muir, P., Li, S., Lou, S., Wang, D., Spakowicz, D. J., Salichos, L., Zhang, J., Weinstock, G. M., Isaacs, F., Rozowsky, J., and Gerstein, M. (2016). The real cost of sequencing: scaling computation to keep pace with data generation. *Genome biology*, 17:53.

Müller, R., Roberts, C. A., and Brown, T. A. (2016). Complications in the study of ancient tuberculosis: Presence of environmental bacteria in human archaeological remains. *Journal of archaeological science*, 68:5–11.

Nasidze, I., Li, J., Quinque, D., Tang, K., and Stoneking, M. (2009). Global diversity in the human salivary microbiome. *Genome research*, 19(4):636–643.

Nasidze, I., Li, J., Schroeder, R., Creasey, J. L., Li, M., and Stoneking, M. (2011). High diversity of the saliva microbiome in Batwa Pygmies. *PLoS One*, 6(8):e23352.

Nasko, D. J., Koren, S., Phillippy, A. M., and Treangen, T. J. (2018). RefSeq database growth influences the accuracy of k-mer-based lowest common ancestor species identification. *Genome biology*, 19(1):165.

Nearing, J. T., DeClercq, V., Van Limbergen, J., and Langille, M. G. I. (2020). Assessing the variation within the oral microbiome of healthy adults. *mSphere*, 5(5).

Neukamm, J., Pfrengle, S., Molak, M., Seitz, A., Francken, M., Eppenberger, P., Avanzi, C., Reiter, E., Urban, C., Welte, B., Stockhammer, P. W., Teßmann, B., Herbig, A., Harvati, K., Nieselt, K., Krause, J., and Schuenemann, V. J. (2020). 2000-year-old pathogen genomes reconstructed from metagenomic analysis of Egyptian mummified individuals. *BMC biology*, 18(1):108.

Nicholls, S. M., Aubrey, W., De Grave, K., Schietgat, L., Creevey, C. J., and Clare, A. (2020). On the complexity of haplotyping a microbial community. *Bioinformatics*, 37(10):1360–1366.

Novak, M. (2015). Dental health and diet in early medieval Ireland. *Archives of oral biology*, 60(9):1299–1309.

Obregon-Tito, A. J., Tito, R. Y., Metcalf, J., Sankaranarayanan, K., Clemente, J. C., Ursell, L. K., Zech Xu, Z., Van Treuren, W., Knight, R., Gaffney, P. M., Spicer, P., Lawson, P., Marin-Reyes, L., Trujillo-Villarroel, O., Foster, M., Guija-Poma, E., Troncoso-Corzo, L., Warinner, C., Ozga, A. T., and Lewis, C. M. (2015). Subsistence strategies in traditional societies distinguish gut microbiomes. *Nature communications*, 6:6505.

Ochman, H., Elwyn, S., and Moran, N. A. (1999). Calibrating bacterial evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 96(22):12638–12643.

Ochman, H., Worobey, M., Kuo, C.-H., Ndjango, J.-B. N., Peeters, M., Hahn, B. H., and Hugenholtz, P. (2010). Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS biology*, 8(11):e1000546.

Olalde, I., Brace, S., Allentoft, M. E., Armit, I., Kristiansen, K., Booth, T., Rohland, N., Mallick, S., Szécsényi-Nagy, A., Mittnik, A., Altena, E., Lipson, M., Lazaridis, I., Harper, T. K., Patterson, N., Broomandkhoshbacht, N., Diekmann, Y., Faltyskova, Z., Fernandes, D., Ferry, M., Harney, E., de Knijff, P., Michel, M., Oppenheimer, J., Stewardson, K., Barclay, A., Alt, K. W., Liesau, C., Ríos, P., Blasco, C., Miguel, J. V., García, R. M., Fernández, A. A., Bánffy, E., Bernabò-Brea, M., Billoin, D., Bonsall, C., Bonsall, L., Allen, T., Büster, L., Carver, S., Navarro, L. C., Craig, O. E., Cook, G. T., Cunliffe, B., Denaire, A., Dinwiddy, K. E., Dodwell, N., Ernée, M., Evans, C., Kuchařík, M., Farré, J. F., Fowler, C., Gazenbeek, M., Pena, R. G., Haber-Uriarte, M., Haduch, E., Hey, G., Jowett, N., Knowles, T., Massy, K., Pfrengle, S., Lefranc, P., Lemercier, O., Lefebvre, A., Martínez, C. H., Olmo, V. G., Ramírez, A. B., Maurandi, J. L., Majó, T., McKinley,

J. I., McSweeney, K., Mende, B. G., Modi, A., Kulcsár, G., Kiss, V., Czene, A., Patay, R., Endrődi, A., Köhler, K., Hajdu, T., Szeniczey, T., Dani, J., Bernert, Z., Hoole, M., Cheronet, O., Keating, D., Velemínský, P., Dobeš, M., Candilio, F., Brown, F., Fernández, R. F., Herrero-Corral, A.-M., Tusa, S., Carnieri, E., Lentini, L., Valenti, A., Zanini, A., Waddington, C., Delibes, G., Guerra-Doce, E., Neil, B., Brittain, M., Luke, M., Mortimer, R., Desideri, J., Besse, M., Brücken, G., Furmanek, M., Hałuszko, A., Mackiewicz, M., Rapiński, A., Leach, S., Soriano, I., Lillios, K. T., Cardoso, J. L., Pearson, M. P., Włodarczak, P., Price, T. D., Prieto, P., Rey, P.-J., Risch, R., Rojo Guerra, M. A., Schmitt, A., Serralongue, J., Silva, A. M., Smrčka, V., Vergnaud, L., Zilhão, J., Caramelli, D., Higham, T., Thomas, M. G., Kennett, D. J., Fokkens, H., Heyd, V., Sheridan, A., Sjögren, K.-G., Stockhammer, P. W., Krause, J., Pinhasi, R., Haak, W., Barnes, I., Lalueza-Fox, C., and Reich, D. (2018). The Beaker phenomenon and the genomic transformation of northwest Europe. *Nature*, 555(7695):190–196.

Orlando, L., Allaby, R., Skoglund, P., Sarkissian, C. D., Stockhammer, P. W., Ávila-Arcos, M. C., Fu, Q., Krause, J., Willerslev, E., Stone, A. C., and Warinner, C. (2021). Ancient DNA analysis. *Nature Reviews Methods Primers*, 1(1):1–26.

Ottoni, C., Guellil, M., Ozga, A. T., Stone, A. C., Kersten, O., Bramanti, B., Porcier, S., and Van Neer, W. (2019). Metagenomic analysis of dental calculus in ancient Egyptian baboons. *Scientific reports*, 9(1):19637.

Ozga, A. T., Nieves-Colón, M. A., Honap, T. P., Sankaranarayanan, K., Hofman, C. A., Milner, G. R., Lewis, Jr, C. M., Stone, A. C., and Warinner, C. (2016). Successful enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus. *American journal of physical anthropology*, 160(2):220–228.

Palmer, Jr, R. J., Gordon, S. M., Cisar, J. O., and Kolenbrander, P. E. (2003). Coaggregation-mediated interactions of streptococci and actinomyces detected in initial human dental plaque. *Journal of bacteriology*, 185(11):3400–3409.

Palmer, Jr, R. J., Shah, N., Valm, A., Paster, B., Dewhirst, F., Inui, T., and Cisar, J. O. (2017). Interbacterial adhesion networks within early oral biofilms of single human hosts. *Applied and environmental microbiology*, 83(11).

Parkhill, J., Wren, B. W., Thomson, N. R., Titball, R. W., Holden, M. T., Prentice, M. B., Sebaihia, M., James, K. D., Churcher, C., Mungall, K. L., Baker, S., Basham, D., Bentley, S. D., Brooks, K., Cerdeño-Tárraga, A. M., Chillingworth, T., Cronin, A., Davies, R. M., Davis, P., Dougan, G., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Karlyshev, A. V., Leather, S., Moule, S., Oyston, P. C., Quail, M., Rutherford, K., Simmonds, M., Skelton, J., Stevens, K., Whitehead, S., and Barrell, B. G. (2001). Genome sequence of *Yersinia pestis*, the causative agent of plague. *Nature*, 413(6855):523–527.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M. C., Rice, B. L., DuLong, C., Morgan, X. C., Golden, C. D., Quince, C., Huttenhower, C., and Segata, N. (2019). Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*, 0(0):649–662.E20.

Pasolli, E., Schiffer, L., Manghi, P., Renson, A., Obenchain, V., Truong, D. T., Beghini, F., Malik, F., Ramos, M., Dowd, J. B., Huttenhower, C., Morgan, M., Segata, N., and Waldron, L. (2017). Accessible, curated metagenomic data through ExperimentHub. *Nature methods*, 14(11):1023–1024.

Peltzer, A., Jäger, G., Herbig, A., Seitz, A., Kniep, C., Krause, J., and Nieselt, K. (2016). EAGER: efficient ancient genome reconstruction. *Genome biology*, 17(1):1–14.

Peres, M. A., Macpherson, L. M. D., Weyant, R. J., Daly, B., Venturelli, R., Mathur, M. R., Listl, S., Celeste, R. K., Guarnizo-Herreño, C. C., Kearns, C., Benzian, H., Allison, P., and Watt, R. G. (2019). Oral diseases: a global public health challenge. *The Lancet*, 394(10194):249–260.

Perry, G. H., Dominy, N. J., Claw, K. G., Lee, A. S., Fiegler, H., Redon, R., Werner, J., Villanea, F. A., Mountain, J. L., Misra, R., Carter, N. P., Lee, C., and Stone, A. C. (2007). Diet and the evolution of human amylase gene copy number variation. *Nature genetics*, 39(10):1256–1260.

Perry, R. D. and Fetherston, J. D. (1997). *Yersinia pestis*–etiologic agent of plague. *Clinical microbiology reviews*, 10(1):35–66.

Peyrégne, S. and Prüfer, K. (2020). Present-Day DNA contamination in ancient DNA datasets. *BioEssays*, 42(9):e2000081.

Philip, N., Suneja, B., and Walsh, L. (2018). Beyond *Streptococcus mutans*: clinical implications of the evolving dental caries aetiological paradigms and its associated microbiome. *British dental journal*, 224(4):219–225.

Philips, A., Stolarek, I., Kuczkowska, B., Juras, A., Handschuh, L., Piontek, J., Kozlowski, P., and Figlerowicz, M. (2017). Comprehensive analysis of microorganisms accompanying human archaeological remains. *GigaScience*, 6(7):1–13.

Pinzari, F., Cornish, L., and Jungblut, A. D. (2020). Skeleton bones in museum indoor environments offer niches for fungi and are affected by weathering and deposition of secondary minerals. *Environmental microbiology*, 22(1):59–75.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., and Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, page 201178. Preprint.

Power, R. C., Salazar-García, D. C., Straus, L. G., González Morales, M. R., and Henry, A. G. (2015). Microremains from El Mirón Cave human dental calculus suggest a mixed plant–animal subsistence economy during the Magdalenian in Northern Iberia. *Journal of archaeological science*, 60(Supplement C):39–46.

Preus, H. R., Marvik, O. J., Selvig, K. A., and Bennike, P. (2011). Ancient bacterial DNA (aDNA) in dental calculus from archaeological human remains. *Journal of archaeological science*, 38(8):1827–1831.

Proctor, D. M., Shelef, K. M., Gonzalez, A., Davis, C. L., Dethlefsen, L., Burns, A. R., Loomer, P. M., Armitage, G. C., Ryder, M. I., Millman, M. E., Knight, R., Holmes, S. P., and Relman, D. A. (2020). Microbial biogeography and ecology of the mouth and implications for periodontal diseases. *Periodontology 2000*, 82(1):26–41.

Prüfer, K., Posth, C., Yu, H., Stoessel, A., Spyrou, M. A., Deviese, T., Mattonai, M., Ribechini, E., Higham, T., Velemínský, P., Brůžek, J., and Krause, J. (2021). A genome sequence from a modern human skull over 45,000 years old from Zlatý kůň in Czechia. *Nature ecology & evolution*, 5(6):820–825.

Qian, X.-B., Chen, T., Xu, Y.-P., Chen, L., Sun, F.-X., Lu, M.-P., and Liu, Y.-X. (2020). A guide to human microbiome research: study design, sample collection, and bioinformatics analysis. *Chinese medical journal*, 133(15):1844–1855.

Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J., and Segata, N. (2017). Shotgun metagenomics, from sampling to analysis. *Nature biotechnology*, 35(9):833–844.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842.

Radini, A., Buckley, S., Rosas, A., Estalrrich, A., de la Rasilla, M., and Hardy, K. (2016). Neanderthals, trees and dental calculus: new evidence from El Sidrón. *Antiquity*, 90(350):290–301.

Radini, A., Tromp, M., Beach, A., Tong, E., Speller, C., McCormick, M., Dudgeon, J. V., Collins, M. J., Rühli, F., Kröger, R., and Warinner, C. (2019). Medieval women's early involvement in manuscript production suggested by lapis lazuli identification in dental calculus. *Science Advances*, 5(1):eaau7126.

Rampelli, S., Schnorr, S. L., Consolandi, C., Turroni, S., Severgnini, M., Peano, C., Brigidi, P., Crittenden, A. N., Henry, A. G., and Candela, M. (2015). Metagenome sequencing of the Hadza Hunter-Gatherer gut microbiota. *Current biology*, 25(13):1682–1693.

Ramsey, M. M. and Whiteley, M. (2009). Polymicrobial interactions stimulate resistance to host innate immunity through metabolite perception. *Proceedings of the National Academy of Sciences of the United States of America*, 106(5):1578–1583.

Rasmussen, S., Allentoft, M. E., Nielsen, K., Orlando, L., Sikora, M., Sjögren, K.-G., Pedersen, A. G., Schubert, M., Van Dam, A., Kapel, C. M. O., Nielsen, H. B., Brunak, S., Avetisyan, P., Epimakhov, A., Khalyapin, M. V., Gnuni, A., Kriiska, A., Lasak, I., Metspalu, M., Moiseyev, V., Gromov, A., Pokutta, D., Saag, L., Varul, L., Yepiskoposyan, L., Sicheritz-Pontén, T., Foley, R. A., Lahr, M. M., Nielsen, R., Kristiansen, K., and Willerslev, E. (2015). Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell*, 163(3):571–582.

Redford, K. H., Segre, J. A., Salafsky, N., Martinez del Rio, C., and McAloose, D. (2012). Conservation and the microbiome. *Conservation biology*, 26(2):195–197.

Reimer, L. C., Vetcininova, A., Carbasse, J. S., Söhngen, C., Gleim, D., Ebeling, C., and Overmann, J. (2019). BacDive in 2019: bacterial phenotypic data for high-throughput biodiversity analysis. *Nucleic acids research*, 47(D1):D631–D636.

Relman, D. A. (2012). The human microbiome: ecosystem resilience and health. *Nutrition reviews*, 70 Suppl 1:S2–9.

Righolt, A. J., Jevdjevic, M., Marcenes, W., and Listl, S. (2018). Global-, regional-, and Country-Level economic impacts of dental diseases in 2015. *Journal of dental research*, 97(5):501–507.

Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R., and Konstantinidis, K. T. (2018). Nonpareil 3: Fast estimation of metagenomic coverage and sequence diversity. *mSystems*, 3(3).

Rosenberg, E., Koren, O., Reshef, L., Efrony, R., and Zilber-Rosenberg, I. (2007). The role of microorganisms in coral health, disease and evolution. *Nature reviews. Microbiology*, 5(5):355–362.

Rosenberg, E. and Zilber-Rosenberg, I. (2018). The hologenome concept of evolution after 10 years. *Microbiome*, 6(1):78.

Rosier, B. T., De Jager, M., Zaura, E., and Krom, B. P. (2014). Historical and contemporary hypotheses on the development of oral diseases: are we there yet? *Frontiers in cellular and infection microbiology*, 4:92.

Ross, A. A., Müller, K. M., Weese, J. S., and Neufeld, J. D. (2018). Comprehensive skin microbiome analysis reveals the uniqueness of human skin and evidence for phylosymbiosis within the class mammalia. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25):E5786–E5795.

Ross, A. A., Rodrigues Hoffmann, A., and Neufeld, J. D. (2019). The skin microbiome of vertebrates. *Microbiome*, 7(1):79.

Rupf, S., Kannengiesser, S., Merte, K., Pfister, W., Sigusch, B., and Eschrich, K. (2000). Comparison of profiles of key periodontal pathogens in periodontium and endodontium. *Endodontics & dental traumatology*, 16(6):269–275.

Rylev, M. and Kilian, M. (2008). Prevalence and distribution of principal periodontal pathogens worldwide. *Journal of clinical periodontology*, 35(8 Suppl):346–361.

Sankaranarayanan, K., Ozga, A. T., Warinner, C., Tito, R. Y., Obregon-Tito, A. J., Xu, J., Gaffney, P. M., Jervis, L. L., Cox, D., Stephens, L., Foster, M., Tallbull, G., Spicer, P., and Lewis, C. M. (2015). Gut microbiome diversity among Cheyenne and Arapaho individuals from Western Oklahoma. *Current biology*, 25(24):3161–3169.

Sanz, M., van Winkelhoff, A. J., Herrera, D., Dellemijn-Kippuw, N., Simón, R., and Winkel, E. (2000). Differences in the composition of the subgingival microbiota of two periodontitis populations of different geographical origin. a comparison between Spain and The Netherlands. *European journal of oral sciences*, 108(5):383–392.

Sawyer, S., Krause, J., Guschanski, K., Savolainen, V., and Pääbo, S. (2012). Temporal patterns of nucleotide misincorporations and DNA fragmentation in ancient DNA. *PLoS One*, 7(3):e34131.

Schnorr, S. L., Candela, M., Rampelli, S., Centanni, M., Consolandi, C., Basaglia, G., Turroni, S., Biagi, E., Peano, C., Severgnini, M., Fiori, J., Gotti, R., De Bellis, G., Luiselli, D., Brigidi, P., Mabulla, A., Marlowe, F., Henry, A. G., and Crittenden, A. N. (2014). Gut microbiome of the Hadza hunter-gatherers. *Nature communications*, 5:3654.

Schubert, M., Ermini, L., Der Sarkissian, C., Jónsson, H., Ginolhac, A., Schaefer, R., Martin, M. D., Fernández, R., Kircher, M., McCue, M., Willerslev, E., and Orlando, L. (2014). Characterization of ancient and modern genomes by SNP detection and phylogenomic and metagenomic analysis using PALEOMIX. *Nature protocols*, 9(5):1056–1082.

Schuenemann, V. J., Bos, K., DeWitte, S., Schmedes, S., Jamieson, J., Mittnik, A., Forrest, S., Coombes, B. K., Wood, J. W., Earn, D. J. D., White, W., Krause, J., and Poinar, H. N. (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 of the United States of America*, 108(38):E746–52.

Schuenemann, V. J., Singh, P., Mendum, T. A., Krause-Kyora, B., Jäger, G., Bos, K. I., Herbig, A., Economou, C., Benjak, A., Busso, P., Nebel, A., Boldsen, J. L., Kjellström, A., Wu, H., Stewart, G. R., Taylor, G. M., Bauer, P., Lee, O. Y.-C., Wu, H. H. T., Minnikin, D. E., Besra, G. S., Tucker, K., Roffey, S., Sow, S. O., Cole, S. T., Nieselt, K., and Krause, J. (2013). Genome-wide comparison of medieval and modern *Mycobacterium leprae*. *Science*, 341(6142):179–183.

Scott, A., Power, R. C., Altmann-Wendling, V., Artzy, M., Martin, M. A. S., Eisenmann, S., Hagan, R., Salazar-García, D. C., Salmon, Y., Yegorov, D., Milevski, I., Finkelstein, I., Stockhammer, P. W., and Warinner, C. (2021). Exotic foods reveal contact between South Asia and the Near East during the second millennium BCE. *Proceedings of the National Academy of Sciences of the United States of America*, 118(2).

Sczyrba, A., Hofmann, P., Belmann, P., Koslicki, D., Janssen, S., Dröge, J., Gregor, I., Majda, S., Fiedler, J., Dahms, E., Bremges, A., Fritz, A., Garrido-Oter, R., Jørgensen, T. S., Shapiro, N., Blood, P. D., Gurevich, A., Bai, Y., Turaev, D., DeMaere, M. Z., Chikhi, R., Nagarajan, N., Quince, C., Meyer, F., Balvočiūtė, M., Hansen, L. H., Sørensen, S. J., Chia, B. K. H., Denis, B., Froula, J. L., Wang, Z., Egan, R., Don Kang, D., Cook, J. J., Deltel, C., Beckstette, M., Lemaitre, C., Peterlongo, P., Rizk, G., Lavenier, D., Wu, Y.-W., Singer, S. W., Jain, C., Strous, M., Klingenberg, H., Meinicke, P., Barton, M. D., Lingner, T., Lin, H.-H., Liao, Y.-C., Silva, G. G. Z., Cuevas, D. A., Edwards, R. A., Saha, S., Piro, V. C., Renard, B. Y., Pop, M., Klenk, H.-P., Göker, M., Kyrpides, N. C., Woyke, T., Vorholt, J. A., Schulze-Lefert, P., Rubin, E. M., Darling, A. E., Rattei, T., and McHardy, A. C. (2017). Critical assessment of metagenome interpretation-a benchmark of metagenomics software. *Nature methods*, 14(11):1063–1071.

Seitz, A. and Nieselt, K. (2017). Improving ancient DNA genome assembly. *PeerJ*, 5:e3126.

Sender, R., Fuchs, S., and Milo, R. (2016). Revised estimates for the number of human and bacteria cells in the body. *PLoS biology*, 14(8):e1002533.

Shallcross, L. J. and Davies, D. S. C. (2014). Antibiotic overuse: a key driver of antimicrobial resistance. *The British journal of general practice*, 64(629):604–605.

Shaw, L., Ribeiro, A. L. R., Levine, A. P., Pontikos, N., Balloux, F., Segal, A. W., Roberts, A. P., and Smith, A. M. (2017). The human salivary microbiome is shaped by shared environment rather than genetics: Evidence from a large family of closely related individuals. *mBio*, 8(5).

Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature biotechnology*, 26(10):1135–1145.

Shrivastava, A., Patel, V. K., Tang, Y., Yost, S. C., Dewhirst, F. E., and Berg, H. C. (2018). Cargo transport shapes the spatial organization of a microbial community. *Proceedings of the National Academy of Sciences of the United States of America*, 115(34):8633–8638.

Silverman, J. D., Shenhav, L., Halperin, E., Mukherjee, S., and David, L. A. (2018). Statistical considerations in the design and analysis of longitudinal microbiome studies. *bioRxiv*, page 448332. Preprint.

Siqueira, W. L., Custodio, W., and McDonald, E. E. (2012). New insights into the composition and functions of the acquired enamel pellicle. *Journal of dental research*, 91(12):1110–1118.

Skoglund, P., Posth, C., Sirak, K., Spriggs, M., Valentin, F., Bedford, S., Clark, G. R., Reepmeyer, C., Petchey, F., Fernandes, D., Fu, Q., Harney, E., Lipson, M., Mallick, S., Novak, M., Rohland, N., Stewardson, K., Abdullah, S., Cox, M. P., Friedlaender, F. R., Friedlaender, J. S., Kivisild, T., Koki, G., Kusuma, P., Merriwether, D. A., Ricaut, F.-X., Wee, J. T. S., Patterson, N., Krause, J., Pinhasi, R., and Reich, D. (2016). Genomic insights into the peopling of the southwest Pacific. *Nature*, 538(7626):510–513.

Slatkin, M. and Racimo, F. (2016). Ancient DNA and human history. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23):6380–6387.

Slon, V., Mafessoni, F., Vernot, B., de Filippo, C., Grote, S., Viola, B., Hajdinjak, M., Peyrégne, S., Nagel, S., Brown, S., Douka, K., Higham, T., Kozlikin, M. B., Shunkov, M. V., Derevianko, A. P., Kelso, J., Meyer, M., Prüfer, K., and Pääbo, S. (2018). The genome of the offspring of a Neanderthal mother and a Denisovan father. *Nature*, 561(7721):113–116.

Smith, C. I., Chamberlain, A. T., Riley, M. S., Stringer, C., and Collins, M. J. (2003). The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of human evolution*, 45(3):203–217.

Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C., and Kent, Jr, R. L. (1998). Microbial complexes in subgingival plaque. *Journal of clinical periodontology*, 25(2):134–144.

Song, L., Florea, L., and Langmead, B. (2014). Lighter: fast and memory-efficient sequencing error correction without counting. *Genome biology*, 15(11):509.

Sonnenburg, J. L. and Sonnenburg, E. D. (2019). Vulnerability of the industrialized microbiota. *Science*, 366(6464).

Spyrou, M. A., Bos, K. I., Herbig, A., and Krause, J. (2019). Ancient pathogen genomics as an emerging tool for infectious disease research. *Nature reviews. Genetics*, 20(6):323–340.

Stolp, H. (1988). *Microbial Ecology: Organisms, Habitats, Activities*. Cambridge University Press, Cambridge.

Stuiver, M. and Polach, H. A. (1977). Discussion reporting of 14C data. *Radiocarbon*, 19(3):355–363.

Takeshita, T., Matsuo, K., Furuta, M., Shibata, Y., Fukami, K., Shimazaki, Y., Akifusa, S., Han, D.-H., Kim, H.-D., Yokoyama, T., Ninomiya, T., Kiyohara, Y., and Yamashita, Y. (2014). Distinct composition of the oral indigenous microbiota in South Korean and Japanese adults. *Scientific reports*, 4:6990.

Tan, K. H., Seers, C. A., Dashper, S. G., Mitchell, H. L., Pyke, J. S., Meuric, V., Slakeski, N., Cleal, S. M., Chambers, J. L., McConville, M. J., and Reynolds, E. C. (2014). *Porphyromonas gingivalis* and *Treponema denticola* exhibit metabolic symbioses. *PLoS pathogens*, 10(3):e1003955.

The NIH HMP Working Group, Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., Baker, C. C., Di Francesco, V., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., Giblin, C., David, H., Mills, M., Salomon, R., Mullins, C., Akolkar, B., Begg, L., Davis, C., Grandison, L., Humble, M., Khalsa, J., Little, A. R., Peavy, H., Pontzer, C., Portnoy, M., Sayre, M. H., Starke-Reed, P., Zakhari, S., Read, J., Watson, B., and Guyer, M. (2009). The NIH human microbiome project. *Genome research*, 19(12):2317–2323.

Tito, R. Y., Macmil, S., Wiley, G., Najar, F., Cleeland, L., Qu, C., Wang, P., Romagne, F., Leonard, S., Ruiz, A. J., Reinhard, K., Roe, B. A., and Lewis, Jr, C. M. (2008). Phylotyping and functional analysis of two ancient human microbiomes. *PLoS One*, 3(11):e3703.

Tremaroli, V. and Bäckhed, F. (2012). Functional interactions between the gut microbiota and host metabolism. *Nature*, 489(7415):242–249.

Trevelline, B. K., Fontaine, S. S., Hartup, B. K., and Kohl, K. D. (2019). Conservation biology needs a microbial renaissance: a call for the consideration of host-associated microbiota in wildlife management practices. *Proceedings of the Royal Society B: Biological Sciences*, 286(1895):20182448.

Tromp, M., Matisoo-Smith, E., Kinaston, R., Bedford, S., Spriggs, M., and Buckley, H. (2020). Exploitation and utilization of tropical rainforests indicated in dental calculus of ancient Oceanic Lapita culture colonists. *Nature human behaviour*, 4(5):489–495.

Turner-Walker, G. (2007). The chemical and microbial degradation of bones and teeth. In *Advances in Human Palaeopathology*, pages 3–29. John Wiley & Sons, Ltd, Chichester, UK.

Utter, D. R., Mark Welch, J. L., and Borisy, G. G. (2016). Individuality, stability, and variability of the plaque microbiome. *Frontiers in microbiology*, 7:564.

Vågene, Å. J., Herbig, A., Campana, M. G., Robles García, N. M., Warinner, C., Sabin, S., Spyrou, M. A., Andrades Valtueña, A., Huson, D., Tuross, N., Bos, K. I., and Krause, J. (2018). *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature ecology & evolution*, 2(3):520–528.

Valm, A. M., Mark Welch, J. L., Rieken, C. W., Hasegawa, Y., Sogin, M. L., Oldenbourg, R., Dewhirst, F. E., and Borisy, G. G. (2011). Systems-level analysis of microbial community organization through combinatorial labeling and spectral imaging. *Proceedings of the National Academy of Sciences of the United States of America*, 108(10):4152–4157.

van der Valk, T., Pečnerová, P., Díez-Del-Molino, D., Bergström, A., Oppenheimer, J., Hartmann, S., Xenikoudakis, G., Thomas, J. A., Dehasque, M., Sağlıcan, E., Fidan, F. R., Barnes, I., Liu, S., Somel, M., Heintzman, P. D., Nikolskiy, P., Shapiro, B., Skoglund, P., Hofreiter, M., Lister, A. M., Götherström, A., and Dalén, L. (2021). Million-year-old DNA sheds light on the genomic history of mammoths. *Nature*, 591(7849):265–269.

van Dijk, E. L., Auger, H., Jaszczyszyn, Y., and Thermes, C. (2014). Ten years of next-generation sequencing technology. *Trends in genetics*, 30(9):418–426.

Velsko, I. M., Fellows Yates, J. A., Aron, F., Hagan, R. W., Frantz, L. A. F., Loe, L., Martinez, J. B. R., Chaves, E., Gosden, C., Larson, G., and Warinner, C. (2019). Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome*, 7(1):102.

Velsko, I. M., Frantz, L. A. F., Herbig, A., Larson, G., and Warinner, C. (2018). Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. *mSystems*, 3(4).

Velsko, I. M., Overmyer, K. A., Speller, C., Klaus, L., Collins, M. J., Loe, L., Frantz, L. A. F., Sankaranarayanan, K., Lewis, Jr, C. M., Martinez, J. B. R., Chaves, E., Coon, J. J., Larson, G., and Warinner, C. (2017). The dental calculus metabolome in modern and historic samples. *Metabolomics*, 13(11):134.

Ventola, C. L. (2015). The antibiotic resistance crisis: part 1: causes and threats. *Pharmacy and Therapeutics*, 40(4):277–283.

von Ohle, C., Gieseke, A., Nistico, L., Decker, E. M., DeBeer, D., and Stoodley, P. (2010). Real-time microsensor measurement of local metabolic activities in ex vivo dental biofilms exposed to sucrose and treated with chlorhexidine. *Applied and environmental microbiology*, 76(7):2326–2334.

Warinner, C., Hendy, J., Speller, C., Cappellini, E., Fischer, R., Trachsel, C., Arneborg, J., Lynnerup, N., Craig, O. E., Swallow, D. M., Fotakis, A., Christensen, R. J., Olsen, J. V., Liebert, A., Montalva, N., Fiddyment, S., Charlton, S., Mackie, M., Canci, A., Bouwman, A., Rühli, F., Gilbert, M. T. P., and Collins, M. J. (2014a). Direct evidence of milk consumption from ancient human dental calculus. *Scientific reports*, 4:7104.

Warinner, C., Herbig, A., Mann, A., Fellows Yates, J. A., Weiß, C. L., Burbano, H. A., Orlando, L., and Krause, J. (2017). A robust framework for microbial archaeology. *Annual review of genomics and human genetics*, 18:321–356.

Warinner, C., Rodrigues, J. F. M., Vyas, R., Trachsel, C., Shved, N., Grossmann, J., Radini, A., Hancock, Y., Tito, R. Y., Fiddyment, S., Speller, C., Hendy, J., Charlton, S., Luder, H. U., Salazar-García, D. C., Eppler, E., Seiler, R., Hansen, L. H., Castruita, J. A. S., Barkow-Oesterreicher, S., Teoh, K. Y., Kelstrup, C. D., Olsen, J. V., Nanni, P., Kawai, T., Willerslev, E., von Mering, C., Lewis, Jr, C. M., Collins, M. J., Gilbert, M. T. P., Rühli, F., and Cappellini, E. (2014b). Pathogens and host immunity in the ancient human oral cavity. *Nature genetics*, 46(4):336–344.

Warinner, C., Speller, C., and Collins, M. J. (2015). A new era in palaeomicrobiology: prospects for ancient dental calculus as a long-term record of the human oral microbiome. *Philosophical Transactions Of The Royal Society of London. Series B, Biological sciences*, 370(1660):20130376.

West, A. G., Waite, D. W., Deines, P., Bourne, D. G., Digby, A., McKenzie, V. J., and Taylor, M. W. (2019). The microbiome in threatened species conservation. *Biological conservation*, 229:85–98.

Weyrich, L. S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., Morris, A. G., Alt, K. W., Caramelli, D., Dresely, V., Farrell, M., Farrer, A. G., Francken, M., Gully, N., Haak, W., Hardy, K., Harvati, K., Held, P., Holmes, E. C., Kaidonis, J., Lalueza-Fox, C., de la Rasilla, M., Rosas, A., Semal, P., Soltysiak, A., Townsend, G., Usai, D., Wahl, J., Huson, D. H., Dobney, K., and Cooper, A. (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature*, 544(7650):357–361.

Weyrich, L. S., Farrer, A. G., Eisenhofer, R., Arriola, L. A., Young, J., Selway, C. A., Handsley-Davis, M., Adler, C. J., Breen, J., and Cooper, A. (2019). Laboratory contamination over time during low-biomass sample analysis. *Molecular ecology resources*, 19(4):982–996.

White, D. J. (1991). Processes contributing to the formation of dental calculus. *Biofouling*, 4(1-3):209–218.

White, D. J. (1997). Dental calculus: recent insights into occurrence, formation, prevention, removal and oral health effects of supragingival and subgingival deposits. *European journal of oral sciences*, 105(5 Pt 2):508–522.

Wibowo, M. C., Yang, Z., Borry, M., Hübner, A., Huang, K. D., Tierney, B. T., Zimmerman, S., Barajas-Olmos, F., Contreras-Cubas, C., García-Ortiz, H., Martínez-Hernández, A., Luber, J. M., Kirstahler, P., Blohm, T., Smiley, F. E., Arnold, R., Ballal, S. A., Pamp, S. J., Russ, J., Maixner, F., Rota-Stabelli, O., Segata, N., Reinhard, K., Orozco, L., Warinner, C., Snow, M., LeBlanc, S., and Kostic, A. D. (2021). Reconstruction of ancient microbial genomes from the human gut. *Nature*, 594(7862):234–239.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag,

T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific data*, 3:160018.

Wood, D. E., Lu, J., and Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome biology*, 20(1):257.

Wood, D. E. and Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome biology*, 15(3):R46.

Yang, F., Zeng, X., Ning, K., Liu, K.-L., Lo, C.-C., Wang, W., Chen, J., Wang, D., Huang, R., Chang, X., Chain, P. S., Xie, G., Ling, J., and Xu, J. (2012). Saliva microbiomes distinguish caries-active from healthy human populations. *The ISME journal*, 6(1):1–10.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., Magris, M., Hidalgo, G., Baldassano, R. N., Anokhin, A. P., Heath, A. C., Warner, B., Reeder, J., Kuczynski, J., Caporaso, J. G., Lozupone, C. A., Lauber, C., Clemente, J. C., Knights, D., Knight, R., and Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, 486(7402):222–227.

Ye, S. H., Siddle, K. J., Park, D. J., and Sabeti, P. C. (2019). Benchmarking metagenomics tools for taxonomic classification. *Cell*, 178(4):779–794.

Yost, S., Duran-Pinedo, A. E., Teles, R., Krishnan, K., and Frias-Lopez, J. (2015). Functional signatures of oral dysbiosis during periodontitis progression revealed by microbial metatranscriptome analysis. *Genome medicine*, 7(1):27.

Youngblut, N. D., Reischer, G. H., Walters, W., Schuster, N., Walzer, C., Stalder, G., Ley, R. E., and Farnleitner, A. H. (2019). Host diet and evolutionary history explain different aspects of gut microbiome diversity among vertebrate clades. *Nature communications*, 10(1):2200.

Zhou, W., Gay, N., and Oh, J. (2018a). ReprDB and panDB: minimalist databases with maximal microbial representation. *Microbiome*, 6(1):15.

Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C., and Achtman, M. (2018b). Accurate reconstruction of microbial strains from metagenomic sequencing using representative reference genomes. In *Research in Computational Molecular Biology*, pages 225–240. Springer International Publishing.

Zhou, Z., Lundstrøm, I., Tran-Dien, A., Duchêne, S., Alikhan, N.-F., Sergeant, M. J., Langridge, G., Fotakis, A. K., Nair, S., Stenøien, H. K., Hamre, S. S., Casjens, S., Christophersen, A., Quince, C., Thomson, N. R., Weill, F.-X., Ho, S. Y. W., Gilbert, M. T. P., and Achtman, M. (2018c). Pan-genome analysis of ancient and modern *Salmonella enterica* demonstrates genomic stability of the invasive para C lineage for millennia. *Current biology*, 28(15):2420–2428.e10.

Zhu, Y., Dashper, S. G., Chen, Y.-Y., Crawford, S., Slakeski, N., and Reynolds, E. C. (2013). *Porphyromonas gingivalis* and *Treponema denticola* synergistic polymicrobial biofilm development. *PLoS One*, 8(8):e71727.

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., Aldenderfer, M., Speller, C., Hendy, J., Weston, D. A., MacDonald, S. J., Thomas, G. H., Collins, M. J., Lewis, C. M., Hofman, C., and Warinner, C. (2015). Intrinsic challenges in ancient microbiome reconstruction using 16S rRNA gene amplification. *Scientific reports*, 5:16498.

Zijnge, V., van Leeuwen, M. B. M., Degener, J. E., Abbas, F., Thurnheer, T., Gmür, R., and Harmsen, H. J. M. (2010). Oral biofilm architecture on natural teeth. *PLoS One*, 5(2):e9321.

10 Ehrenwörtliche Erklärung

James Alexander Fellows Yates

Max Planck Institute for the Science of Human History and,
Friedrich-Schiller-Universität Jena

Hiermit erkläre ich,

(a) dass mir die geltende Promotionsordnung bekannt ist,

(b) dass ich die Dissertation selbst angefertigt habe, keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in der Arbeit angegeben habe,

(c) dass ich alle Personen habe, die mir bei der Auswahl und Auswertung sowie bei der Herstellung des Manuskriptes unterstützt haben, in der Autorenliste der Manuskripte und den entsprechenden Danksagungen namentlich erwähnt,

(d) dass ich die Hilfe einer kommerziellen Promotionsvermittlung nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,

(e) dass ich die Dissertation noch nicht als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe,

Jena, den

James A. Fellows Yates

11 Annexes

- Description of candidate-specific contributions to publications in this thesis
- List of locations of supplementary information for the publications contained in this thesis.

11.1 Candidate-specific contributions to publications

11.1.1 Manuscript Nr. A

Short Citation
Fellows Yates et al. (2021). *Proc Natl Acad Sci U S A*, *118*(20), e2021655118.
DOI: 10.1073/pnas.2021655118

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data:

Figure(s) #1-4	☑	100% (the data reproduced in this figure originate entirely from experimental work carried out by the candidate)
	☐	0% (the data reproduced in this figure are based exclusively on the work of other co-authors)
	☐	Any contribution of the doctoral candidate to the figure: __% Brief description of the contribution:
Figure(s) #5	☐	100% (the data reproduced in this figure originate entirely from experimental work carried out by the candidate)
	☐	0% (the data reproduced in this figure are based exclusively on the work of other co-authors)
	☑	Any contribution of the doctoral candidate to the figure: 66% Brief description of the contribution: All data analysis and figure generation for figure panels B and C.

Experimental data is used here to describe novel scientific analysis and interpretation, as sequencing data generation was performed via a core facility.

Long description

J.A.F.Y., I.M.V., A.H., and C.W. designed the study with input from C.A.H., C.M.L, K.S., and J.K.. I helped design the sampling strategy, contacting museums and collaborators for relevant samples, and selection of analysis pathways. K.W.A., J.W.A., C.C.B., I.C., C.C., M.C.C., L.D., J.C.F.M-L., M.D-Z.B., D.G.D., E.E.F., M.F., V.E.G., M.R.G.M., A.G.M, K.H., A.G.H., L.H., M.M., D.M., M.P., S.R.M., M.R., H.R., S.S., J.T.S., L.G.S., J.S., B.T., M.J.W., and J.K., provided materials and resources. **J.A.F.Y**, F.A., C.P., R.M.A, C.E.P., L.D., A.G.H., R.C.P., K.G., R.W., D.C.S-G., J.K., and C.W., performed sampling and collection coordination. I assisted in sample selection from sample providers, shipping, metadata logging prior to and on arrival at the ancient DNA laboratories, as well as going to museums in the US, France, and Germany to perform sampling from skeletons and teeth myself. F.A., C.P., C.A.H., R.M.A., C.E.P., A.E.M, and K.N. performed laboratory analysis. I coordinated the labwork for the project with batch planning and updating the laboratory database. **J.A.F.Y.** and I.M.V. analysed the data. I developed a novel end-to-end analysis workflows for the processing, authentication, and analysis of ancient microbiome. Within this I developed software multiple software (MEx-IPA, cumulative

percent decay curves) and scripts for steps in the workflow without existing software. I performed initial sequencing quality control, data processing and clean-up, database construction for taxonomic and functional profiling as well continued curation during the project. I also performed aDNA preservational and authentication screening (MaltExtract), taxonomic compositional analysis (QIIME, MALT) and statistical testing (PERMANOVA, bootstrapping, hierarchical clustering evaluation), phylogenetic analysis, and prepared data for functional analysis (AADDER, HUMANn2). I.M.V performed functional analysis. **J.A.F.Y.** and I.M.V. and prepared analytical figures. **J.A.F.Y.**, I.M.V, A.H., and C.W. interpreted data with assistance from K.G., A.H., R.W., and F.E.D.. **J.A.F.Y.**, I.M.V., and C.W. wrote the paper and supplementary information, with contributions from the other co-authors. I co-wrote the main text and wrote the majority of the supplementary information. Overall, **J.A.F.Y** contributed 65% of the study.

11.1.2 Manuscript Nr. B

Short Citation
Fellows Yates et al. (2020), *Scientific Data*, *8*(1), 31. DOI: 10.1038/s41597-021-00816-y

Contribution of the doctoral candidate
Contribution of the doctoral candidate to figures reflecting experimental data:

Figure(s) #1-2	☐	100% (the data reproduced in this figure originate entirely from experimental work carried out by the candidate)
	☐	0% (the data reproduced in this figure are based exclusively on the work of other co-authors)
	☑	Any contribution of the doctoral candidate to the figure: 50% Brief description of the contribution: Designed project and infrastructure, defined the metadata fields represented, contributed some of the metadata submissions, performed the analysis, and generated figure.
Figure(s) #3	☐	100% (the data reproduced in this figure originate entirely from experimental work carried out by the candidate)
	☐	0% (the data reproduced in this figure are based exclusively on the work of other co-authors)
	☑	Any contribution of the doctoral candidate to the figure: 95% Brief description of the contribution: Designed project, infrastructure, and workflow. Drafted initial figure.

Note: Experimental data is used here to describe infrastructure design, novel metadata collection from literature review, and interpretation, as per bioinformatics studies.

Long description

J.A.F.Y. and C.W. designed the study with input from A.A.V., Å.J.V, I.M.V, M.B., A.F-.G and P.D.H.. I designed the project aims and scope, set up repository, designed submission workflow, designed tables, defined metadata columns (with input from co-authors), set up automated CI-testing, developed training material, and recruited co-authors and contributors to the project. M.B. developed software. I provided design input and testing for the software **J.A.F.Y.** and E.G. wrote documentation with input from the other co-authors. **J.A.F.Y.** performed analysis. I summarised the content of the repository across multiple factors including bibliographic trends and geographic and temporal trends of the samples. **J.A.F.Y** and A.F-.G. generated figures. **J.A.F.Y.**, A.A.V., Å.J.V., B.C., I.M.V., M.J.B.-L., A.F.-G., E.J.G., S.L.R., P.D.H., M.A.S., A.H., A.S.G., J.H., A.F.A., V.Z. and C.W. acquired data. I performed the main literature to gather publications for inclusion in the repository with additional contributions from co-authors. I also screened publications and gathered metadata, in some cases also contacting and coordinating with authors of these publications for post-publication upload of data to public archives. **J.A.F.Y.** wrote the paper with input from the co-authors. I planned and wrote the majority of the publication with comments from all co-authors. Overall, **J.A.F.Y** contributed 70% to the

study.

11.1.3 Manuscript Nr. C

Short Citation
Fellows Yates et al. (2021) *PeerJ*, *9*, e10947 DOI: 10.7717/peerj.10947

Contribution of the doctoral candidate

Contribution of the doctoral candidate to figures reflecting experimental data:

Figure(s) #1,3,5	☐	100% (the data reproduced in this figure originate entirely from experimental work carried out by the candidate)
	☐	0% (the data reproduced in this figure are based exclusively on the work of other co-authors)
	☑	Any contribution of the doctoral candidate to the figure: 55% Brief description of the contribution: Shared project design and carried out the majority of project coordination, contributed most of the code and testing, designed all figures, and generated figure 3.
Figure(s) #2	☐	100% (the data reproduced in this figure originate entirely from experimental work carried out by the candidate)
	☐	0% (the data reproduced in this figure are based exclusively on the work of other co-authors)
	☑	Any contribution of the doctoral candidate to the figure: 95% Brief description of the contribution: Designed data processing logic, tested data processing, designed figure.
Figure(s) #4	☑	100% (the data reproduced in this figure originate entirely from experimental work carried out by the candidate)
	☐	0% (the data reproduced in this figure are based exclusively on the work of other co-authors)
	☐	Any contribution of the doctoral candidate to the figure: —%

Experimental data is used here to describe project design, novel code contribution, data analysis and interpretation, as per bioinformatic studies.

Long description

J.A.F.Y., S.C. and A.P. conceived and designed the project. I spearheaded the expansion of functionality of the original pipeline to include newer routine analyses, and designed a more complex but flexible data-path to address the latest sequencing strategies used by aDNA laboratories. **J.A.F.Y.** coordinated the project. I recruited contributors to the project, monitored and distributed bug-fixing tasks and also ran community engagement to get broader input into design decisions in the pipeline. **J.A.F.Y.**, and A.P. predominantly wrote the code, with input from T.C.L., M.B., A.A.V., S.C., M.U.G., and J.N., who also developed additional tools. I added new modules, in particular the majority of the metagenomic and pathogen tools, and also carried out the majority of bug-fixes. I also ensured the pipeline code was up-to-date

against the generic nf-core pipeline template, monitored tool versioning and updating code when necessary, and performed both automated and manual testing of code. **J.A.F.Y** wrote documentation. I wrote the majority of usage, parameter, and output documentation, as well as writing one tutorial (metagenomic analysis) and designed the outline of the other tutorials (human and pathogen genomics). Z.F. generated documentation images. **J.A.F.Y.** generated test data, prepared testing environment, performed benchmarking tests and generated figures. I setup the server, installed all software, researched and downloaded suitable test data, and ran all commands and run time logging. **J.A.F.Y** and A.P. wrote the paper with contributions from the other co-authors. I wrote the majority of the main text with A.P., all supplementary information text, and as generated all publication-specific figures. Overall, **J.A.F.Y** contributed 80% of the study.

11.2 Supplementary information

The entire supplementary information for all three manuscripts are either very long or include files that are too large to be included in the printed version and in some cases in the electronic (CD) version of this thesis. Please see the supplementary information directories in the electronic version of this thesis for files published alongside Manuscript A and C by the corresponding journal.

Alternatively, for the entire supplementary information of each, please consult the following locations for long-term archived versions alongside the original manuscripts. All long-term archived versions have been deposited on the Zenodo platform that has a retention policy of 20 years (see `https://about.zenodo.org/policies/`).

11.2.1 Manuscript A

- SI Appendix [URL]
  - `https://www.pnas.org/content/118/20/e2021655118/tab-figures-data`
- SI Appendix [DOI]
  - `https://doi.org/10.1073/pnas.2021655118`
- GitHub repository [URL]
  - `https://github.com/jfy133/Hominid_Calculus_Microbiome_Evolution`
- Zenodo Archive of GitHub repository [DOI]
  - `https://doi.org/10.5281/zenodo.3740492`
- GitHub repository (MEx-IPA) [URL]
  - `https://github.com/jfy133/MEx-IPA`
- Zenodo archive of repository (MEx-IPA) [DOI]
  - `https://doi.org/10.5281/zenodo.3380011`
- GitHub repository (cuperdec) [URL]
  - `https://github.com/jfy133/cuperdec`
- Zenodo archive of repository (cuperdec) [DOI]
  - `https://doi.org/10.5281/zenodo.4561901`

11.2.2 Manuscript B

- GitHub repository (Data) [URL]
  - `https://github.com/SPAAM-community/AncientMetagenomeDir`
- Zenodo archive of GitHub repository (Data) [DOI]
  - `https://doi.org/10.5281/zenodo.3980833`

11.2.3 Manuscript C

- Supplementary Information [URL]
  - `https://peerj.com/articles/10947/#supplemental-information`
- Supplementary Information [DOI]
  - `https://doi.org/10.7717/peerj.10947/supp-1`
- GitHub repository (Supplement) [URL]
  - https://github.com/apeltzer/eager2-paper/tree/master/content/supplement
- GitHub repository (Code) [URL]
  - `https://github.com/nf-core/eager`
- Zenodo archive of GitHub repository (Code) [URL]
  - `https://doi.org/10.5281/zenodo.1465061`