UNIVERSITY
OF TRENTO - Italy
Department of Cellular, Computational
and Integrative Biology - CIBIO

International PhD Program in Biomolecular Sciences

Cycle 33rd

Evolutionary analysis of the human microbiome using ancient metagenomic samples

Supervisor: Prof. Nicola Segata, CIBIO, University of Trento
Co-supervisor: Prof. Omar Rota-Stabelli, C3A, University of Trento

Ph.D. Thesis of

Kun Huang

Department of Cellular, Computational and Integrative Biology

University of Trento

Academic year: 2019-2020

Declaration

I Kun Huang confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Table of Contents

Abstract

Studying the evolutionary history of human microbiome members is indispensable to understanding how our microbiota complexity has been shaped over time. While recent decades have witnessed remarkable progress in investigation of the human microbiome within different evolutionary contexts, reconstructing a time-resolved divergence history for human microbiome members is still extremely challenging. Longitudinal metagenomics studies have shown the potential of addressing such challenges, but it is limited in tracking the evolutionary change in a short term. By contrast, leveraging the microbial genomic information preserved in the ancient metagenomic samples from the long past is emerging as a powerful strategy to study the long-established evolutionary history of the human microbiome. In this thesis, I aimed to devise a novel methodology that allows for efficiently reconstructing a time-resolved evolutionary history of human microbiome species using ancient metagenomic data. To this end, I firstly started from comparing four newly excavated paleofeces samples to a large body of contemporary metagenomic datasets. I observed that our human gut microbiota has diverged from its ancestral state in both microbial composition and metabolic pathways. This could be related to the change of lifestyle during human history. To better understand the divergence time of microbiome members, I secondly developed a novel computational pipeline which can precisely reconstruct and date strain-level phylogenies for microbiome species using carbon dated ancient metagenomic samples as calibration under Bayesian molecular clocking framework. The application of this tool has uncovered the unprecedented evolutionary diversity, in the context of geography and time period, of three *Methanobrevibacter* species from the human oral microbiome, and reconstructed a delicate time-resolved evolutionary history for common gut microbial species, such as *Prevotella copri*, *Eubacterium rectale*, *Methanobrevibacter smithii* and others mainly populating in the human gut microbiota. This approach promises that more underlying evolutionary history about the human microbiome will be unveiled in the foreseen future.

Chapter 1 | Introduction to the dissertation

1.1 Introduction to the dissertation

Whipps and colleagues proposed in 1988 for the first time the definition of a microbiome: a “characteristic microbial community” living in a “reasonably well-defined habitat” with distinct properties, functions, and interactions with the external environment, which together shape specific ecological niches (Berg et al. 2020). While microorganisms are not visible to human eyes, their presence might be anywhere. From agricultural soil to different sites of the human body, these invisible entities are thriving in every possible corner on earth.

Human microbiome is certainly an important ecosystem to explore. The number of microbial genes hosted by human body is more than 100 times the number of our own genes (Qin et al. 2010; Turnbaugh et al. 2007; Consortium and The Human Microbiome Project Consortium 2012) and human cells only accounts for around 47% of the total number of cells carried in and on our body (Sender, Fuchs, and Milo 2016). Being likened to our “last organ”, the human microbiome plays a pivotal role as an inseparable functional unit in human health. For example, it actively interacts with host tissues (Bron, van Baarlen, and Kleerebezem 2011), balances the immune system (Needham et al. 2013), and regulates drug metabolism (Spanogiannopoulos et al. 2016). In the last decade, we have witnessed an impressive progress in exploring the unknown of the human microbiome from a few large investigations (Consortium and The Human Microbiome Project Consortium 2012; Qin et al. 2010; Turnbaugh et al. 2009; Yatsunenko et al. 2012) to numerous smaller studies focused on specific scientific questions. Most studies have nonetheless been focused on exploring the overall microbiome diversity, and neglected its complex evolutionary history which is the key to understanding how such diversity has been shaped through time.

Unquestionably, untying the knot of the evolution in the human microbiome can address questions like how the stable composition has been achieved (Gálvez et al. 2020), how the variability and individuality have been formed in the short- and long-term (Zhao et al. 2019), and from where and when certain members of microbiome have emerged (Lebreton et al. 2017; Tett et al. 2019). Most studies exploring evolution in the human microbiome have been focused on reconstructing phylogenetic relationships between microbial members (Truong et al. 2017; Langille et al. 2013; Lozupone, Hamady, and Kelley 2007; Asnicar et al. 2020), but our understanding of the evolutionary timeline remains limited in the case of microbes directly from metagenomic samples. The success of obtaining ancient microbial DNA

from archaeological samples of human remains manifests the potential for revealing the deep history of microbiome evolution in the past. Weyrich et al., combined metagenomic advances and molecular clock dating to reconstruct the earliest divergence time for *Methanobrevibacter oralis* strains using the oral microbiome from a ~48,000 years old Neanderthal individual (Weyrich et al. 2017), which illuminated the possible microbiome transferring between Neanderthals and modern humans during subsequent interactions in the pre-history. Other studies focused on more recent evolutionary histories include uncovering the spreading route of *Yersinia pestis* basal strain during Neolithic decline (Rascovan et al. 2019), discovering the emergence of human-adapted *Salmonella enterica* in the Neolithization process (Key et al. 2020), finding a high diversity of *Mycobacterium leprae* in medieval Europe (Schuenemann et al. 2018) and many others. These studies unleashed an avalanche of surprises about the evolutionary history of microbes which we can hardly unearth solely using contemporary metagenomic data.

Mining DNA from the microbiome of the individuals from the past has shown remarkable success in extending our understanding about microbial evolutionary history, particularly coupling ancient DNA technology with next-generation sequencing. Current relative studies mostly concentrate on pathogens which are the culprits of historical epidemics and pandemics (Rascovan et al. 2019; Vågene et al. 2018; Sabin et al. 2020a; Schuenemann et al. 2013). For instance, *Yersinia pestis*, the etiological agent of several plagues in the history, has been described at length regarding its evolutionary origin (Rascovan et al. 2019), the diversification timeline of its lineages from Bronze Age to the present day (Spyrou et al. 2018), and its dispersal pattern in Europe (Morozova et al. 2020). However, it just scratched the surface of evolutionary history in the tremendous microbial world.

The major challenges of using ancient metagenomic samples to study evolutionary history of microbes directly from the microbiome include several aspects. Firstly, though next-generation sequencing permits an immense volume of genomic information from samples, ancient molecules which can be precisely captured in archaeological material remain an extremely low proposition (Key et al. 2017). As a result, many studies turned to screen abundant members using laboratory techniques, such as PCR, and enrich DNA molecules for the targeted ancient microbes (Maixner et al. 2016; Spyrou et al. 2018, 2019). Secondly, ancient DNA from the long dead has shown clear post-mortem damage on two ends of the sequencing reads (Briggs et al. 2007) and the impact on the subsequent analysis has been reported by few studies

(Axelsson et al. 2008; Andrew Rambaut et al. 2009). For instance, it can potentially generate a considerable amount of artificial single-nucleotide variants (SNVs) which are not from the genuine evolution. Thirdly, many current ancient microbiome studies heavily depend on mapping reads against a single reference genome and extracting only variable sites for evolutionary analysis (Peltzer et al. 2016; Spyrou et al. 2016, 2018). The phylogeny reconstructed in this fashion has been reported to be biased and unreliable (Bertels et al. 2014).

With the continuing drop of the cost of next-generation sequencing and the growing number of ancient microbiome samples being unearthed, metagenomic data collected from both contemporary and ancient materials have become increasingly available, which promises a greater expansion of our understanding about microbial evolutionary history from the deep past, recent past, to the present day. Therefore, a methodology which can efficiently reconstruct and date the evolutionary history of microbes directly from ancient and contemporary metagenomic samples is needed.

1.2 Aims of the dissertation

In this dissertation, I aimed to develop a methodological approach able to exploit ancient microbiological samples to unravel the evolutionary history of human-associated bacteria. The second main aim was to apply the methodology on some representative human-associated species and validate the hypothesis that precise time-resolved phylogeny can be built for them.

Specifically, I strove to accommodate four major methodological challenges:

1. design a methodology which can efficiently confront a continuingly growing metagenomic data either as sequencing reads or assembled genomes.

2. tailor the methodology in a way of extracting maximum genomic information from ancient materials which are known to contain low biomass and meanwhile guarantee the reliability of the phylogenies built on the extracted information.

3. extend the reference-guided approach from the conventional fashion based on single-reference to a manner including multiple references in order to minimize bias when inferring genome phylogeny.

4. devise rich flexible utilities for results assessment, visualization, post-processing, and results interpretation to achieve a fine-grained phylogenetic and molecular clocking analysis.

For biological relevance, I aimed to answer four main questions:

1. How our gut microbiota have altered in microbial composition and metabolic functions in the past. Can such divergence be explained by concepts of molecular evolution?

2. Once sufficient understanding of the divergence of the human gut microbiome has been gained, how can we measure the divergence time for human microbiome species? How to apply this concept to microbes which are commonly observed in human microbiomes, such as *Prevotella copri*, *Eubacterium rectale* and *Methanobrevibcater smithii*?

3. What evolutionary history can be explored in the other human microbial ecosystem (e.g. oral cavity) using the same approach? Can oral metagenomic data from ancient populations help gain better understanding of the microbial evolutionary history where using solely contemporary data cannot fully explain? Is evolutionary diversity related to external factors such as geography and time period?

4. Is there any link between the evolutionary history of our microbiota and that of humans, for example, coevolution?

1.3 Structure of the dissertation

This dissertation is organized into four main chapters, each of **Chapter 2, 3 and 4** discussing one first-authored (or joint first-authored) submitted manuscript (except the study reported in **Chapter 3** will be submitted in coming weeks) and **Chapter 5** reporting three published works which are related collaborative studies I carried out in my doctoral training. A brief introduction to each work is given at the beginning of each chapter, followed by a detailed statement about my contribution.

- **Chapter 2** reports the manuscript "Paleofeces analyses indicate blue cheese and beer consumption by Iron Age Hallstatt salt miners and a non-Westernized gut microbiome structure in Europe until the Baroque period". In this work, as a joint first author, I computationally analyzed the microbial composition and functional features of four human paleofeces from Hallstatt, Austria and compared them with a large body of contemporary metagenomic datasets.

- **Chapter 3** reports the manuscript "MetaClock: An integrated framework to infer evolutionary history of microbes from ancient and contemporary metagenomic data". In this work, based on the foundation of **Chapter 2**, I devised a novel computational pipeline to analyze evolutionary relationships and divergence timeline of microbiome members from ancient and contemporary populations.

- **Chapter 4** reports the manuscript "Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal *Methanobrevibacter*". In this work, as a joint first author, I combined metagenomic assembly advances and the approach established in **Chapter 3** to reconstruct the geographical and temporal diversity of three most enriched *Methanobrevibacter* species in ancient calculus samples.

- **Chapter 5** reports the abstracts, main sections, and brief introductions to three collaborative studies I contributed to during my doctoral training. These works are representative of the applications in metagenomic contexts using the tool I reported in **Chapter 3.**

Chapter 2 | Paleofeces analysis indicates blue cheese and beer consumption by Iron Age Hallstatt salt miners and a non-Westernized gut microbiome structure in Europe until the Baroque period

2.1 Introduction to the chapter

This chapter reports the initial exploration of paleofeces samples from Iron Age Hallstatt salt miners, with particular interest in reconstructing our ancestors' lifestyle and ancestral human gut microbiome structure. I started from exploring ancient gut contents because the preserved microbial DNA can provide valuable information for reconstructing the evolutionary history of human-associated microbes. In the study presented in this chapter, I first employed metagenomic analysis on four human paleofeces samples from a mining community dated to the Bronze Age, Iron Age, and early Modern times and compared their microbial composition and metabolic functions to those of vast amounts of contemporary individuals. I next analyzed the prevalence of species enriched in paleofeces samples in a large body of present-day gut metagenomic data from 8,986 healthy adults. As a result, I demonstrate that the gut microbiome structure of miners resembles that of non-Westernized individuals of contemporary populations and our gut microbiota has diverged from its ancestral state due to recent dietary changes. This finding is further supported by considerable traces of bran, glumes of cereals, and other highly-fibrous plant fragments identified by other co-authors using in-depth microscopic and molecular analysis. The divergence of the human gut microbiota represents an important proxy of our long established co-evolved microbiome.

Contribution: In this study, I mainly contributed to the reconstruction of the microbiome structure of four newly excavated paleofeces samples and to compare them with a large body of modern metagenomic samples (N = 823) from different ecosystems. I next sought to assess the prevalence of most enriched microbes in paleofeces in 8,968 modern human gut metagenomes representing a global population, and quantified the abundance of *Prevotella copri* complex in ancient samples as well. Besides large-scale metagenomic analysis, I also interacted closely with other joint first authors in preprocessing sequencing data, provided bioinformatics assistance through the whole work, contributed to data interpretation and manuscript writing.

Paleofeces analysis indicates blue cheese and beer consumption by Iron Age Hallstatt salt miners and a non-Westernized gut microbiome structure in Europe until the Baroque period

Frank Maixner*, Mohamed S. Sarhan*, Kun D. Huang*, Adrian Tett, Alexander Schoenafinger, Stefania Zingale, Aitor Blanco-Míguez, Paolo Manghi, Jan Cemper-Kiesslich, Omar Rota-Stabelli, Thomas Rattei, Robert L. Moritz, Klaus Oeggl, Nicola Segata*, Albert Zink*, Hans Reschreiter*, Kerstin Kowarik*

* these authors contributed equally

2.2 Abstract

The UNESCO World Heritage region Hallstatt-Dachstein/Salzkammergut represents one of Europe's oldest cultural and industrial landscapes with salt mines in the Hallstatt mountain dating back at least to the 14th century BC. The site gave name to the Hallstatt Period (800 to 400 BC) of the Early Iron Age in Europe. The high salt concentrations and the constant annual temperature at around 8°C inside the isolated mine workings perfectly preserved organic archaeological artefacts (e.g. clothing, mining tools, human excrements) that provide unique insights into the daily life of a progressive community in Hallstatt. Here we subjected the human paleofeces from the mining system dated to the Bronze Age- Iron Age, and early Modern times to in-depth microscopic, metagenomic and proteomic analysis. This allowed us to reconstruct the diet of the former population and gain insights into their ancient gut microbiome composition. Our dietary survey identified bran and glumes of different cereals as one of the most prevalent plant fragments. This highly fibrous, carbohydrate-rich diet was supplemented with proteins from broad beans and occasionally with fruits, nuts, or animal food. Linked to these traditional dietary habits all ancient miners up to the early Modern times have gut microbiome structures akin to modern non-Westernized individuals which may indicate a shift in our gut community composition due to quite recent dietary changes. When we extended our microbial survey to fungi present in the paleofeces, we observed in one of the Iron Age samples a high abundance of *Penicillium roqueforti* and *Saccharomyces cerevisiae* DNA. Genome-wide analysis indicates that both fungi were involved in food fermentation and provide the first molecular evidence for blue cheese and beer consumption during Iron Age Europe.

Keywords

Hallstatt, salt mine, paleofeces, microbiome, diet, fermented food, cheese, beer, DNA, proteins

2.3 Introduction

Paleofeces is naturally desiccated ancient feces found in dry caves or desert areas. It can also be preserved in mummies, in ancient latrines, in bogs or in soils when the environmental conditions prevent its deterioration (R. I. Gilbert and Mielke 1985). Previous studies have shown that paleofecal material still contains plant macro- and microfossils, parasite eggs, and even ancient biomolecules (DNA, proteins, metabolites) (Shillito et al. 2020). Ancient paleofeces have therefore recently been used as a source of information to study prehistoric nutrition patterns (M. T. P. Gilbert et al. 2008; Poinar et al. 2001; Maixner et al. 2018), health (Mitchell 2017; Reinhard et al. 2013), and to analyze single representatives (Maixner et al. 2016; Tett et al. 2019) or the overall composition of the intestinal microbiome of our ancestors (Tito et al. 2012; Borry et al. 2020; Wibowo et al. 2021).

Well preserved paleofeces are rare as they require specific preservation conditions. One of the few archaeological sites where paleofeces can be found are the prehistoric salt mines of the Austrian UNESCO World Heritage area Hallstatt-Dachstein/Salzkammergut. Prehistoric salt mines offer ideal preservation conditions for organic materials. The high salt concentrations and the constant annual temperature, at around 8°C inside the isolated mine workings, perfectly preserve organic artefacts. The Hallstatt salt mines located in the Eastern Alps (Figure 1), offers the world's oldest and most continuous record of underground salt mining (Stöllner 2004; Harding 2013; Reschreiter and Kowarik 2019). Large scale underground salt mining in the Hallstatt salt mountain dates back at least to the 14th century BC (late Bronze Age). Several prehistoric mining (Bronze, Iron Age) and historic (14th AD to present) phases are well documented. The site also gave name to the Hallstatt Period (800 to 400 BC) of the Early Iron Age in Europe.

In the prehistoric Bronze Age and Iron Age mine workings of Hallstatt, dense layers of production waste reaching several meters of thickness were excavated uncovering thousands of wooden tools and construction elements, implements made from fur, rawhide, hundreds of woolen textile fragments, and grass and bast ropes and human excrements (M. Grabner et al. 2019). These objects provide unique insights into the daily life of a Bronze Age and Iron Age mining community ranging from mining technology, organization of production and resource management to human health, dietary habits, social organization of production processes and social status within a mining system. These aspects have been studied extensively in Hallstatt based on a

unique combination of data sources encompassing the prehistoric mine working, Bronze Age meat curing facilities and the large Iron Age cemetery (Festi et al. 2021; M. Grabner et al. 2019; Reschreiter and Kowarik 2019).

Here we focus on the question of the structure and evolution of dietary habits as well as the human gut microbiome in one Europe's most important early production community. We used microscopic and metagenomic analysis to characterize nutrition patterns of the prehistoric (Bronze Age, early Iron Age) and historic (early Modern Period) miners and metagenomic analysis to determine the structure and evolution of the gut microbiome. Our findings will enhance the understanding of early European dietary habits (especially the production and consumption of processed foodstuffs) and provide further evidence of the recent change in gut microbiome structure as a result of industrialization/Westernization.

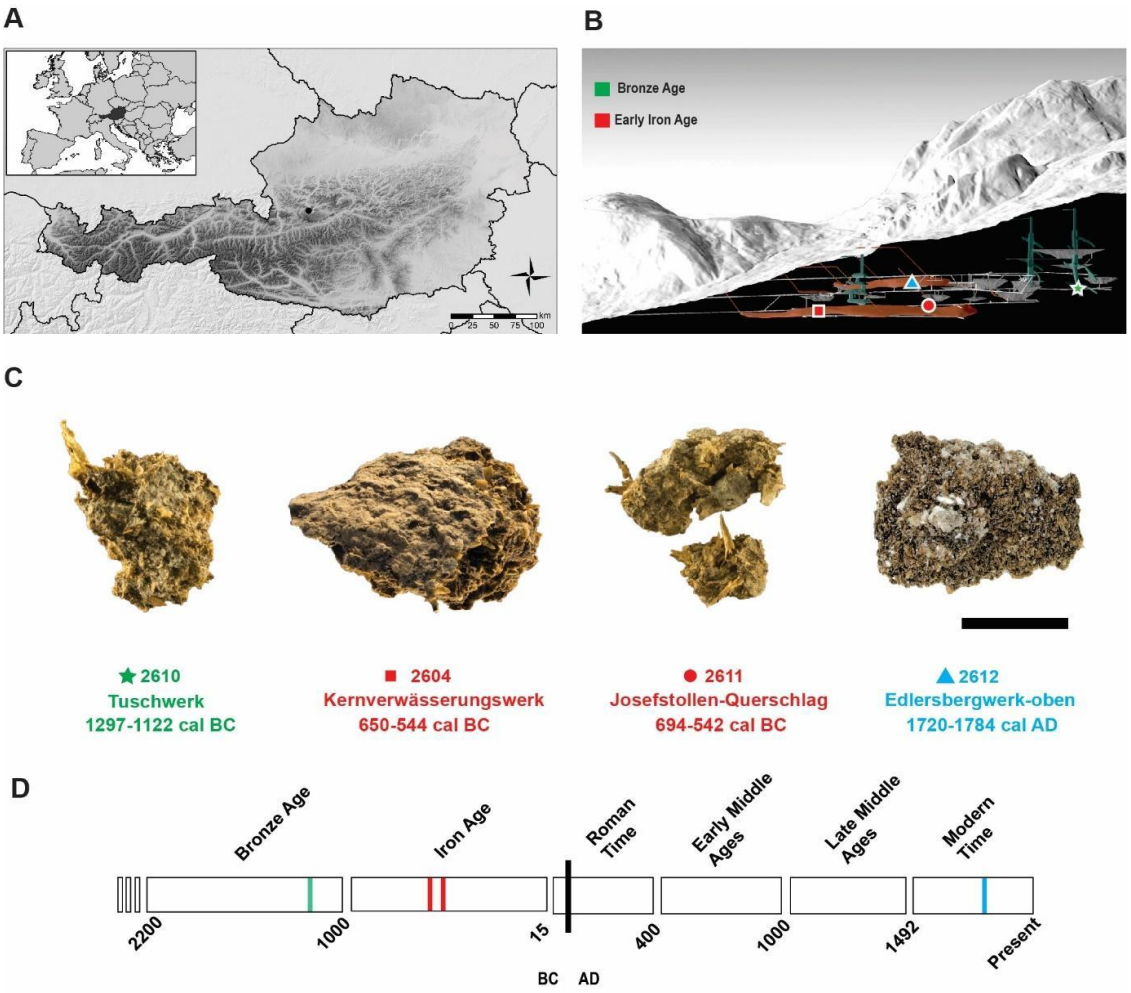


Figure 1. The Hallstatt salt mine and radiocarbon dated paleofeces samples used in this study. (A) The salt mines are located in Upper Austria **(B)** Finding sites of the four paleofeces samples in the Bronze Age and Iron Age mine shafts. The symbol color corresponds to the radiocarbon date of the paleofeces. **(C)** Macroscopic appearance of the four paleofeces samples. The scale bar corresponds to 1cm of length. The sample description includes the sample ID, the mine workings name, and the radiocarbon date. The provided radiocarbon date range corresponds to the Cal 2-sigma values with the highest probability. **(D)** Temporal assignment of the radiocarbon dated paleofeces to the major European time periods from the Bronze Age onwards.

2.4 Results

2.4.1 Paleofeces from Bronze Age to Baroque Period contains ancient endogenous DNA

In this study, we initially subjected four paleofeces samples, collected from Bronze Age and Iron Age Hallstatt mine workings to radiocarbon dating, then to in-depth microscopic and molecular analysis (Figure 1A, B, C, Table S1, and Table S2). The four paleofeces samples can be macroscopically differentiated in samples containing a high amount of fibrous plant material (2610, 2604, 2611) and in one more homogeneous sample (2612) that does not contain any visible larger plant fragments (Figure 1). Radiocarbon analyses date the roughly structured samples to the Late Bronze Age (2610) and Iron Age (2604, 2611), which is in perfect accordance with the proposed period of usage of the mine workings where the paleofeces has been found (Michael Grabner et al. 2021). In contrast, the fine textured paleofeces 2612 sampled in an Iron Age mine working dates to the Early Modern times, the Baroque period (Figure 1C and Table S2). For this part of the salt mine however, it is historically documented that the mine workings had started to be re-used from the beginning of 18th century onwards (Wochenberichte 1723). Independently of the paleofeces age, their storage time since excavation (some samples were recovered in the year 1983), or the mode of excavation (wet-sieving vs. direct sampling) (Figure S1), we could retrieve from all samples biomolecules (DNA and protein) for the subsequent molecular analysis (Table S1 and Data S1). Proteomics analysis provided first evidence for the presence of endogenous biomolecules in the paleofeces material. The most abundant peptides were assigned to human intestinal tract proteins that are involved in food digestive processes (Data S1). The DNA of the paleofeces material was further subjected to a deep shotgun-sequencing approach resulting in 57,130,584 to 221,314,691 quality-filtered reads (Table S3). A first taxonomic overview using

DIAMOND against the NCBI non-redundant database revealed that the majority of reads in the samples are assigned to the *Bacteria* (93.9 to 78.9% of all assigned reads) with *Firmicutes* and *Bacteroidetes* being the most abundant phyla of this kingdom (Figure S2). Less than 7.5% of the reads were eukaryotic with up to 6.7% fungal reads in sample 2604. The *Metazoa* and *Viridiplantae* reads, important for the molecular reconstruction of the diet, comprised 0.5 to 0.01% of all assigned reads. Further analysis of the human DNA in the paleofeces revealed an endogenous DNA content between 0.26 and 0.06%, sufficient for molecular sex and mitochondrial haplogroup assignment (Table S4). The highly fragmented human reads display low deamination patterns at the 5′ ends, typical of aDNA damage (Figure S3). Except for the most recent sample 2612 where the reads appear less fragmented and display almost no DNA damage. Our analyses show that the four paleofeces come from male individuals which carry distinct mitogenomes with low contamination estimates (1 to 2%), indicating that each sample represents unique unmixed ancient feces.

2.4.2 Ancient paleofeces display a gut microbiome structure similar to modern non-Westernized individuals

We compared the microbiome structure of the paleofeces to a large number of contemporary metagenomes (N = 823). Ordination analysis (PCoA) performed on species-level taxonomic composition shows that the paleofeces from the Bronze Age to the Baroque period cluster with stool samples from contemporary non-Westernized individuals (Figure 2A), i.e. those whose diet is mainly composed of unprocessed food and fresh- fruits and vegetables (Brewster et al. 2019). This clustering is similarly observed for encoded metabolic pathways (Figure 2B). All the paleofecal samples were distinct from the oral and, more importantly, from the soil samples, suggesting little evidence of soil contamination which is sometimes observed in ancient metagenomics studies (Key et al. 2017).

To further assess the paleofeces samples, we analyzed the prevalence of the top 15 most enriched species in the paleofeces compared to 8,968 gut microbiomes of healthy Westernized and non-Westernized adults. As a result, 13 out of 15 most abundant species were identified to be associated with the human gut environment, of which 11 species were found to be more prevalent in modern non-Westernized populations compared to Westernized populations. Five of these species, *Bifidobacterium angulatum*, *Lactobacillus ruminis*, *Catenibacterium mitsuokai*, *Prevotella copri*, and *Clostridium ventriculi*, were over twice as prevalent in

non-Westernized populations (Figure 2C and Table S5). One of the two species not associated with the human gut is the halophilic archaeon *Halococcus morrhuae* which survives on high concentration of salt (Grant 2015). It was observed in low abundance in the paleofeces sample 2612, the only sample that was not subjected to wet-sieving. Therefore, we assume that the archaeon was introduced from the environment via the salt crystals. In the paleofeces samples 2604 and 2611 we identified in addition *Clostridium perfringens,* a known intestinal foodborne pathogen (García et al. 2019), but also occurs free living in the soil and other environments (Voidarou et al. 2011). It is therefore difficult to assert whether it is an environmental contamination or a remnant of food spoilage in the miners' gut.

The *Prevotella copri* complex which is highly prevalent in non-Westernized populations and prevalent in previously investigated ancient samples (Tett et al. 2019), was identified in all paleofecal samples representing, on average, 7.3% (1.6% ~ 14.7%) of the relative abundance (Figure 2C). Consistent with previous findings (Tett et al. 2019), we found multiple clades of the complex to be present in each of the paleofeces samples (Figure 2D) with the exception of Clade D which was barely detectable in sample 2604 and 2612. All other clades were detected with relative abundances > 0.01% in all samples (Figure 2D and Table S6). Of note, the sample 2604 displayed higher abundance, in contrast to other samples, of bacterial species, such as *Lactobacillus brevis*, *Bifidobacterium merycicum*, *Bifidobacterium angulatum*, and *Lactobacillus plantarum* (Table S7), that are known to be of probiotic activities or involved in processing of dairy products (Pasolli et al. 2020).

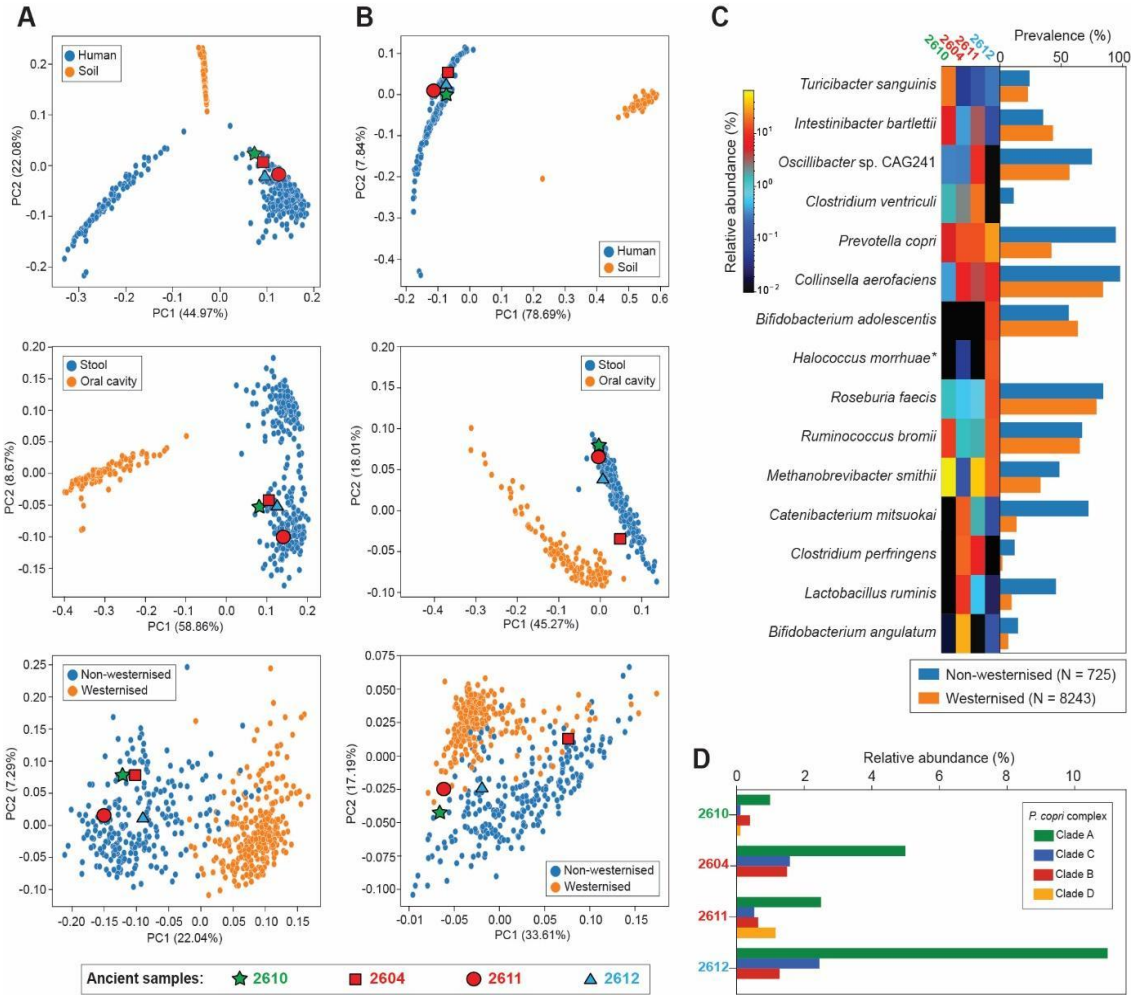


Figure 2. Overview of microbial composition and metabolic pathways of paleofeces samples in comparison to a large collection of contemporary metagenomes. (A) Ordination analysis (PCoA) based on microbial abundance profiled using MetaPhlAn 3.0 (Beghini et al. 2021) between four paleofeces samples and 823 contemporary samples characterized by sampling environment, body site, and non-Westernized lifestyle. **(B)** Ordination analysis (PCoA) based on metabolic pathway abundance profiled using HUMAnN 3.0 (Franzosa et al. 2018) between four paleofeces samples and the same contemporary samples used in (A). **(C)** Prevalence of the top 15 most enriched species of four paleofeces samples in non-Westernized and Westernized datasets comprising 8,968 stool samples from healthy adult individuals. Asterisk indicates species which is likely from external contamination. **(D)** Relative abundance of *P. copri* four clades estimated using MetaPhlAn 3.0 (Beghini et al. 2021) in each paleofecal sample.

2.4.3 Microscopic and molecular reconstruction of the Hallstatt miner's diet

Next, we aimed to reconstruct the dietary components in the paleofeces using both a microscopic and a molecular survey. The above-mentioned structural differences

between the paleofeces became even more evident in the microscopic analyses. The Modern Times sample 2612 was much finer textured than all other samples from prehistory (Figure 1C). This was reflected also in the macro remain composition of the paleofeces, showing that samples 2610, 2604 and 2611 contained a lot of seeds contrary to 2612, which consisted of frequent tissues of fruit husks and seed coats (Table S8). Generally, all samples displayed a predominance of cereal remains.

Microscopic analysis revealed that the Bronze Age sample (2610) consisted more or less exclusively of cereal remains, which originated from barley (*Hordeum vulgare*), spelt (*Triticum spelta*), some emmer (*Triticum dicoccum*), proso millet (*Panicum miliaceum*) and a few weeds, e.g. corn cockle (*Agrostemma githago*), poison parsley (*Aethusa cynapium*). The Iron Age sample (2604, 2611) were characterized by a predominance of cereal remains from barley (*Hordeum vulgare*), spelt (*Triticum spelta*), millets (*P. miliaceum, Setaria italica*) and a little emmer (*T. dicoccum*). Furthermore, in sample 2611 testa remains of broad bean (*Vicia faba*) and seeds of opium poppy (*Papaver somniferum*) were observed. Crab apple (*Malus sylvestris*) and bilberries/cranberries (*Vaccinium myrtillus*/*vitis-idea*) in sample 2604 document the consumption of gathered wild fruits. Striking in the Iron Age sample 2604 was the contamination with weeds, in particular with corn cockle (*Agrostemma githago*). In the sample 2612 from the Baroque time, the microscopic pattern was notably different to the other samples. The plant material was finely ground, and entire fruits were missing apart from a digested mericarp of anise (*Pimpinella anisum*). The plant tissue belonged to bran (fragments of cereal testa, pericarp, hairs, hilum, endosperm) of wheat (*Triticum* sp.). The precise species was unidentifiable, but according to the rare occurrence of tube cells in the pericarp fragments a member of the tetra- or hexaploid wheat group is suggested. In minor quantities also bran of barley (*H. vulgare*) was observed. Furthermore, the consumption of legumes is documented by testa remains of garden bean (*Phaseolus vulgare*) in this sample.

In addition to the microscopic analysis, we subjected paleofeces biomolecules (DNA and proteins) to molecular dietary analyses. Both metagenomic and proteomic analyses included a homology search against different databases, followed by strict filtering steps of the obtained hits, and a subsequent in-depth analysis of selected identified taxa (See methods, Figure S4 and Tables S9). For the plant diet, we could confirm the presence of the most abundant domesticated plant macro-remains including broomcorn millet (*P. milliaceum*), barley (*H. vulgare*), and wheat (*Triticum* spp.) (Figure 3B). In addition, we found DNA-based evidence for the presence of

walnut (*Juglans regia*) in the sample 2604 and protein-based support for the occurrence of opium poppy seeds (*P. somniferum*) in the sample 2611. All the low abundant wild plants unveiled by the microscopic investigation, in addition to the foxtail millet (*S. italica*) which appeared with high grains number, were not identified in our molecular survey, which has undergone strict filtering to minimize the false positives (Figure S4). Further phylogenetic analysis assigned the *Triticum* spp. chloroplast genomes of the samples 2604 and 2610 closest to the chloroplasts of tetraploid (emmer, durum) and hexaploid (spelt wheat, bread wheat) wheat varieties, respectively (Figure 3C, Table S10, Table S11). Additional comparison with the bread wheat genome revealed an equal subgenome (A, B, and D) representation in the 2604 and 2610 metagenomic reads, which suggest in combination with the microscopic identification of numerous characteristic grains, glumes and spikelets the presence of hexaploid spelt wheat (*T. spelta*) in these paleofeces (Figure 3D, Figure S5). Beside the plant diet, we obtained molecular evidence for the consumption of cattle- (*Bos taurus*) and swine meats (*Sus scrofa*) throughout all time periods (Figure 3B, Figure S6). Interestingly, the most abundant cattle proteins in sample 2611 (hemoglobin and coagulation proteins) indicate that the plant diet was supplemented by cattle blood or blood-rich tissues (e.g., liver) (Data S1). The molecular analyses revealed in addition, that individuals from both the Iron Age (2611) and the Baroque (2612) suffered from intestinal infections with whipworms (*Trichuris trichuria*) and roundworm (*Ascaris* spp.) (Figure 3B, Figure S7). Finally, all samples showed a continuous low background with fungal DNA mainly coming from different *Ascomycota*.

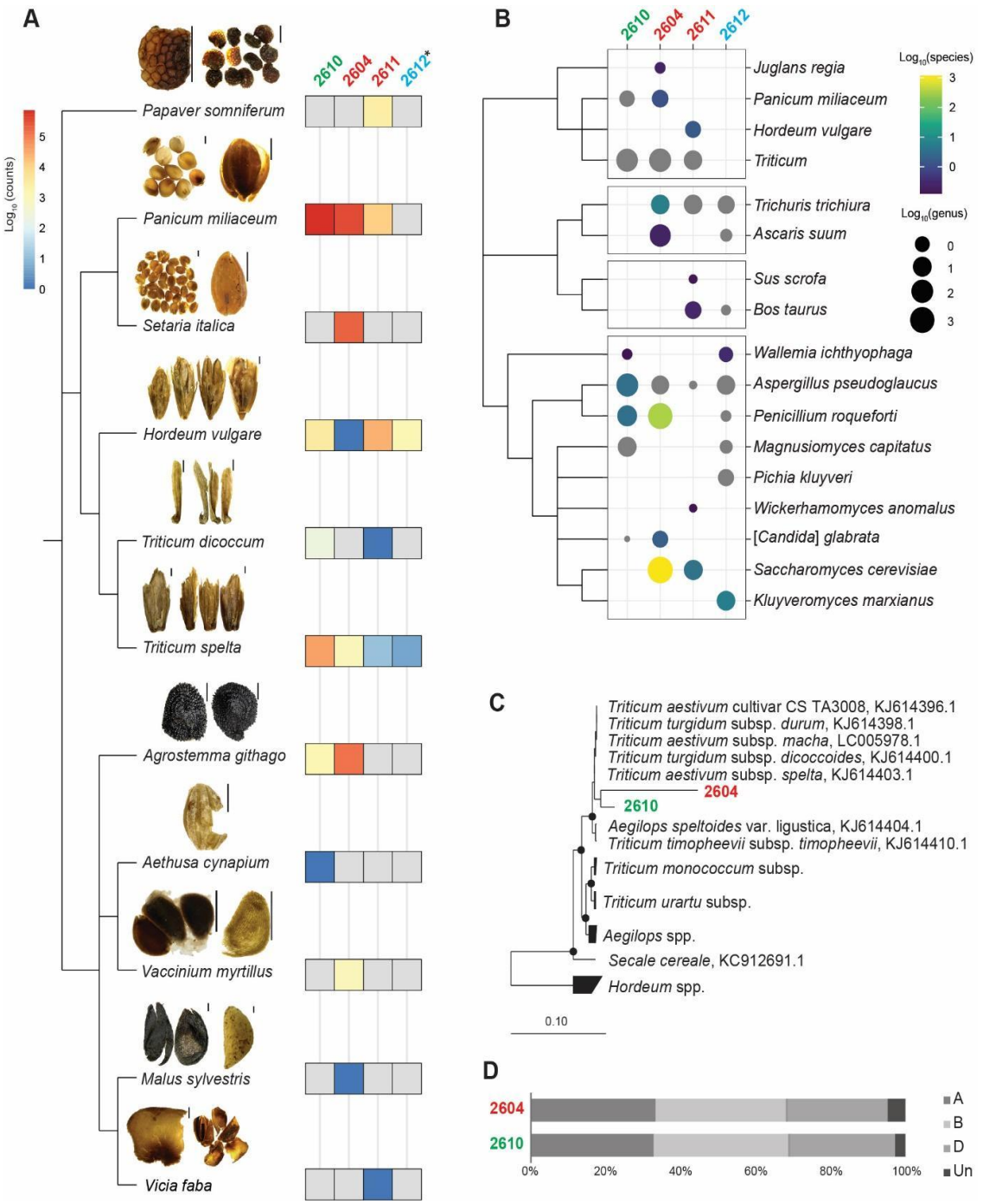


Figure 3. Microscopic and molecular dietary analysis of the Hallstatt paleofeces. (A) Plant macro-remains microscopically detected in the four paleofeces samples. The scale bar indicates 1 mm of length. The heatmap shows the log-scale macro-remain counts normalized to 3.7g sample. The sample with asterisk was assessed in a semi-quantitative manner; For further details, please refer to Table S8. **(B)** Most abundant taxa (plants, nematodes, animals, fungi) detected in the four paleofeces metagenomes and proteomes. The circle size and circle color correspond to log10 “normalized” number of reads per million at genus-

and species levels, respectively. The asterisks in the proteome heatmap means the peptides were assigned only to genus-level. **(C)** Phylogenetic assignment of two partial *Triticum* chloroplast genomes in the 2604 and 2610 metagenomes. The comparative dataset included complete chloroplast genomes of selected members of the *Triticeae* tribe (NCBI Accession Numbers are provided in the figure). The tree was calculated using the maximum-likelihood algorithm (PhyML) based on 136,160 informative positions. Black circles symbolize parsimony and neighbor joining bootstrap support (>90%) based on 100 and 1000 iterations, respectively. The scale bar indicates 10% estimated sequence divergence. **(D)** Wheat subgenome (A, B, and D) representation in the 2604 and 2610 metagenomes (Please refer to Table S11), aligned to the modern hexaploid bread wheat reference genome (Acc. No. GCA_900519105). Both the wheat chloroplast and nuclear reads were highly fragmented and display aDNA specific damage patterns (Figure S5).

2.4.4 Molecular evidence for blue cheese and beer consumption during the Iron Age

The Iron Age sample 2604 displayed in contrast to all other samples exceptionally high prevalence of *Penicillium roqueforti* and *Saccharomyces cerevisiae* proteins (Data S1) and DNA (Table S9), making up to 7-22% of total eukaryotic reads. This was very characteristic to this sample, compared with the other samples that did not show such prevalence, even the sample 2611 which was taken from the same sampling site and dated back to the same time point. To authenticate the data and to gain further insights into their potential ecological significance we mapped the high-quality reads of sample 2604 against the reference genomes of these two fungi (Table S10). With 11-13X coverage, we were able to reconstruct > 92% of both genomes. To confirm whether these two fungi are of ancient origin and not modern contaminants, we initially checked the ancient DNA damage pattern of the mapped reads. Both fungi displayed typical ancient DNA damage patterns, with levels comparable to the human endogenous DNA (Figure S8). Hence, and considering their extraordinarily high abundance and exclusive incidence in this sample, we assumed their endogenous originality to the coprolite microbial community. Additionally, both fungi are commonly used nowadays in food processing; *P. roqueforti* is used for cheese fermentation and *S. cerevisiae* is used for fermenting bread and alcoholic beverages including beer, mead and wine. Therefore, we assume that they could have been involved in food processing at that time. To test this assumption, we used the reconstructed genomes for further comparative phylogeny and population genetic analyses to infer whether they had been truly involved in food processing or were just transient environmental microbes.

First, we compared our putative *P. roqueforti* strain, to other 33 sequenced modern *P. roqueforti* strains coming from different functional niches (Dumas et al. 2020). The comparative dataset included 18 cheese-fermenting and 15 non-cheese fermenting strains (Table S12), in addition to *Penicillium psychrosexualis* and *Penicillium carneum* as an outgroup. After mapping the raw reads of all strains to the reference *P. roqueforti* genome FM164 and data filtering, we resolved 120,337 SNPs, which were used for inferring maximum likelihood (ML) phylogenetic relationships among the tested strains (Methods). In consistency with the original publication of Dumas and colleagues, the resulting phylogeny revealed four distinct clades, a non-Roquefort cheese clade, a Roquefort cheese clade, a silage/food spoilage clade, and a wood/food spoilage clade (Figure 4A). Initially, the phylogenetic analysis separated the non-Roquefort cheese clade from the other, then the Roquefort-cheese clade has been diverged from the other food spoilage clades. The ancient *P. roqueforti* strain showed highest similarity to the non-Roquefort cheese strains, being clustered together with their corresponding clade as an earlier divergent. The reason behind such early divergence might be attributed to the recent acquisition of some genomic regions, most importantly *CheesyTer* and *Wallaby*, by the non-Roquefort-cheese strains. This gene acquisition happened most likely via repeated multiplication of selected spores of the best cheeses on bread used as a growth medium in the late 19th century and early 20th century before the advent of microbiological in vitro culturing techniques (Ropars et al. 2015). Thereby, modern non-Roquefort strains were exposed to extensive selection coupled to horizontal gene transfer events from other cheese producing *Penicillum* spp. or even other genera (Dumas et al. 2020; Cheeseman et al. 2014; Ropars et al. 2020, 2015). Importantly, our Iron Age strain did not contain any of those recently acquired fragments, which comes in congruence with the hypothesis that such domestication events occurred during the last two centuries.

We further used the tool ADMIXTURE to infer the degree of admixture among the strains. By assuming the presence of 3 ancestries (K=3), we could clearly distinguish the non-Roquefort-, Roquefort-, and food spoilage strains (Figure 4B). Our putative strain displayed ~ 70% cheese producing ancestry (60% of the non-Roquefort- and 10% of Roquefort cheese) and ~ 30% food spoiler ancestry. Both the phylogenetic placement and ADMIXTURE profile indicate that the ancient *P. roqueforti* has been already under positive selection towards the non-Roquefort cheese-cluster, a selection process that most likely occurred during the process of cheese production. Some archeological findings excavated from the mines might have been used for that purpose (Figure 4C), showing some traces of fatty food products.

Next, we compared the ancient *S. cerevisiae* genome to 157 recent strains coming from different ecological niches, i.e. food, alcoholic beverages (e.g. beer, wine, sake, and spirit), biofuels, and laboratory, as well as wild strains (Table S13). ML phylogenetic analysis, based on 375,629 SNP positions, distinguished 2 main clades. The first main clade splits into two subclades, with one containing most of the beer strains (beer 1 clade), which show a successive sub-clustering based on the origin of the strains (Figure 4D). The other sub-clade (henceforth referred to as “mixed” clade) included a mixture of bread, wine, beer, and spirit strains. The second main clade is composed of two sub-clades, a wine clade and another beer clade (beer 2). All other wild-, laboratory-, and sake strains fall to the base of the whole phylogeny. The ancient *S. cerevisiae* strain clustered basal to the second main clade which includes the wine and beer 2 strains. Further population structure analysis displayed high admixture in our putative strain, resembling primarily the wine ancestral population with 47%, followed by 29.2% beer ancestries (Figure S9). and only 19% wild strain ancestry. Therefore, and considering the ML phylogenetic assignment, we assume that the possibility of our strain to be of wild origin is unlikely. The results rather indicate higher similarity to wine and beer strains. Principal component analysis (PCA) provided further indication for the domestication of our strain in alcoholic beverage fermentation. Along the PC1 that explains 25.42% of the variation, our strain clustered closer to the strains of beer 2 then to the strains of the wine clade (Figure 4E). This was further supported with proteomic analysis (Data S1), that unveiled that most of the peptides assigned to the genus *Saccharomyces* derived from proteins involved in alcohol fermentation pathways (e.g. glycolysis). To further narrow down the possible routes of domestication we decided to differentiate the strains based on functional marker genes. According to recent literature (Pontes et al. 2020; Gonçalves et al. 2016; Marsit et al. 2017), the genes *RTM1*, *BIO1*/*BIO6*, and the chromosomal regions A/B/C can be used to differentiate yeast strains based on their functional niches. The gene *RTM1* is a strong domestication marker responsible for conferring resistance against the toxicity of molasses and other rich-sugar substrates and assumed to be positively selected in beer yeast strains (Pontes et al. 2020; Ness and Aigle 1995). While the genes *BIO1* and *BIO6*, which are involved in *de novo* biosynthesis of biotin, are highly selected in sake fermenting yeasts, due to lack of biotin in the fermentation substrates, such as rice (Wu, Ito, and Shimoi 2005). The regions A, B, and C are horizontally transferred genomic regions from other yeast genera, e.g. *Kluyveromyces*, *Pichia*, and *Zygosaccharomyces* (Novo et al. 2009). These regions contain 39 genes distributed over 3 different chromosomes (Table S14) and are assumed to play a role in wine fermentation.

Therefore, we searched for the presence of these marker genes in our comparative dataset, including our ancient strain. In accordance with the literature, almost all beer strains, either of clade 1 or clade 2, were positive for *RTM1*, while the wine clade was mainly positive for the genomic regions A/B/C. The mixed clade contained both *RTM1* and the regions A/B/C. The sake clade exclusively contained the *BIO1*/*BIO6* genes and partially the *RTM1*. Remarkably, the wild strains isolated from cocoa in Africa, clustered in the basal clade and did not contain any of these marker genes (Figure 4D, Table S15), contrary to our strain that contained the *RTM1* and lacked to the *BIO1*/*BIO6* genes and the regions A/B/C.

Based on the previous givens, i.e. i) the ancient DNA damage profile; ii) the high prevalence of *S. cerevisiae* reads; iii) the presence of fermentable cereal substrates such as wheat and barley; iv) the phylogenetic assignment of the ancient yeast strain; v) the population admixture profile; as well as vi) the distribution of marker genes, we assume that this yeast is of ancient origin and has been involved in beer fermentation, albeit the mode of fermentation still might be arguable, i.e. being bottom-, top-, or spontaneous fermentation.

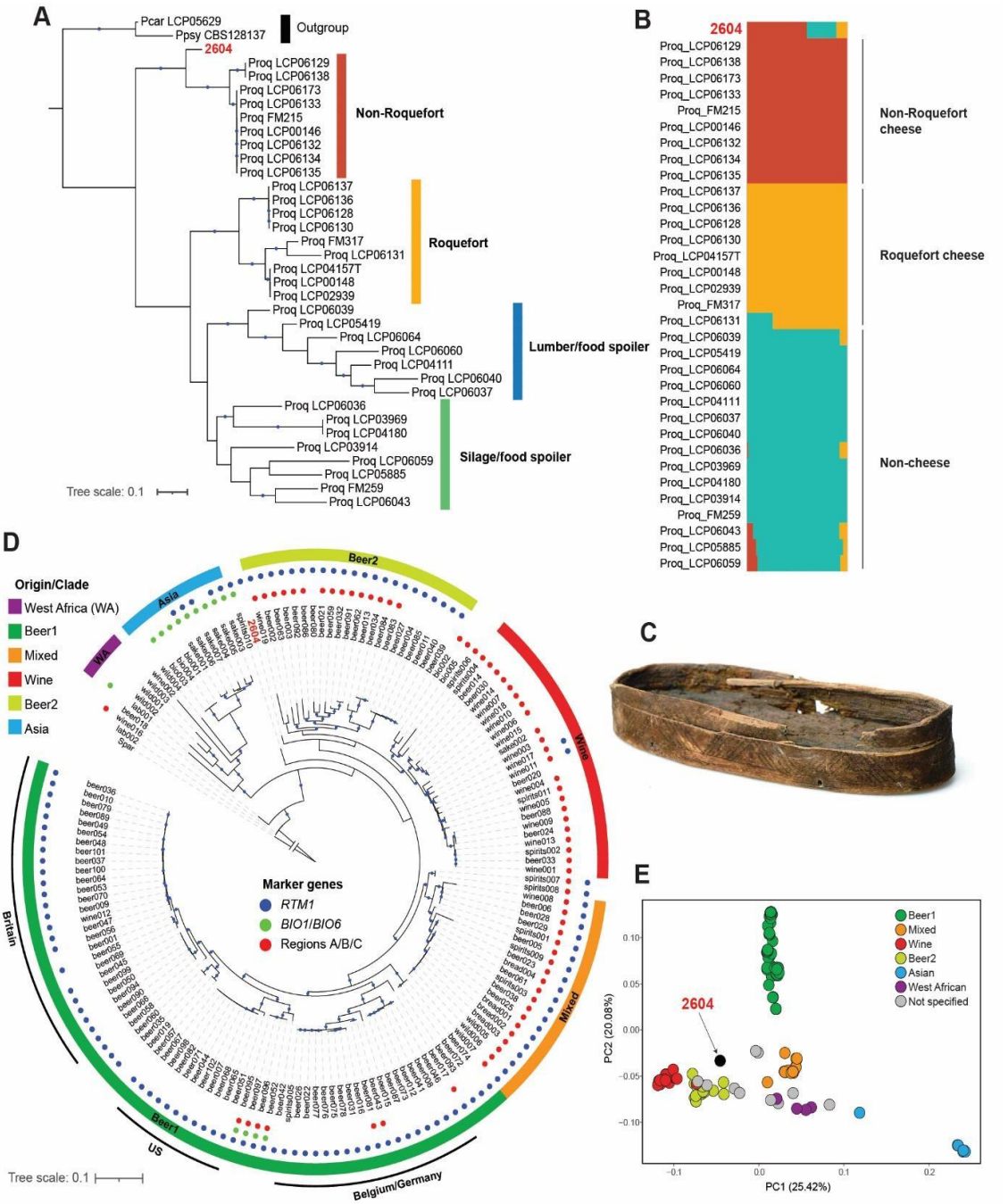


Figure 4. Genome-wide SNP analysis of ancient fungal "strains" vs. modern industrial and wild/environmental strains. (**A**) Maximum likelihood (ML) phylogenetic analysis of the *Penicillium roqueforti* genome assembled from the sample 2604 in addition to other previously published *P. roqueforti* genomes (Dumas et al. 2020). A total number of 120,359 SNP positions were used for the analysis. *P. roqueforti* FM164 was used as a reference, while *P. carneum* and *P. psychrosexualis* were used as outgroups. Colored strips indicate the *P. roqueforti* population as previously inferred (Dumas et al. 2020). For further information on the comparison dataset, please refer to Table S12. (**B**) Population structure analysis of *P. roqueforti*

2604 with the same previous dataset, considering 3 ancestries (K=3 with lowest cross-validation error), based on 120,337 SNPs. The order of labels in panel **B** corresponds to the clustering in panel **A**. (**C**) Wooden containers that has been found among other archeological findings in the mines and assumed to be used as cheese strainers (**D**) ML phylogenetic analysis of *Saccharomyces cerevisiae* genome assembled from the sample 2604 compared with other published *S. cerevisiae* genomes (Gallone et al. 2016). The dataset for the analysis included 375,629 SNPs. The *Saccharomyces paradoxus* CBS432 was used as an outgroup. The colored strips indicate the clade/origin as reported previously (Gallone et al. 2016). The colored dots at the tree edges refer to the presence/absence of functional marker genes (Gonçalves et al. 2016). Blue dots in **A** and **C** indicate bootstrap support > 80% based on 1000 bootstrap replicates. (**E**) Principal component analysis based on 136712 SNPs, of the *S. cerevisiae* strains. For further information, please refer to Table S14 and Table S15.

2.5 Discussion

Our interdisciplinary analyses of the samples have given a unique insight into the microbiome evolution, dietary habits and food processing techniques of the Hallstatt miners over the past three millennia. Molecular and microscopic investigations revealed that the miner´s diet was mainly composed of cereals, such as domesticated wheats (emmer and spelt), barely, common millets, and foxtail millets. This carbohydrate-rich diet was supplemented with proteins from broad beans and occasionally with fruits, nuts, or animal diet. The food remains in the prehistoric sample with a lot of entire fruits and seeds were less processed than those of the Modern times, which consisted of finely ground wheat. This suggests that the prehistoric miners consumed the cereals and legumes in a sort of gruel or porridge, whereas Modern Times miners ate their cereals as bread or biscuit (Heiss et al. 2021).

In general, such carbohydrate-rich fibrous dietary components as observed in the Bronze Age and Iron Age samples are typical for traditional communities and are considered to be the main drivers of the non-Westernized microbiome structure (Makki et al. 2018; Crittenden and Schnorr 2017). Consistent with this observation, our analysis showed that the Hallstatt paleofecal samples contain microbial features similar to gut microbiomes of modern non-Westernized populations (Figure 4A and 4B). Species identified in the samples, such as *Lactobacillus ruminis*, *Catenibacterium mitsuokai*, and *Prevotella copri*, were also found to be highly prevalent in individuals with a more traditional lifestyle (Figure 4C). Furthermore, our paleofecal samples were rich in the *P. copri* complex (Figure 4D), including the four clades that are nearly ubiquitous and co-present in non-Westernized populations (Tett et al. 2019). Of

particular interest, *P. copri* members have been shown to be associated with the degeneration of complex carbohydrates (De Filippis et al. 2019; Gálvez et al. 2020; Tett et al. 2019; Fehlner-Peach et al. 2019), which are the major component of unprocessed fibrous plant diet. Finding paleofeces highly resembling, in terms of microbiome structure, that of non-Westernized individuals supports previous observations (Tito et al. 2012; Wibowo et al. 2021). It also adds weight to the hypothesis that the modern human gut microbiome has diverged from an ancestral state, probably due to modern lifestyle, diet, or medical advances. Interestingly, this non-Westernized microbiome structure has been observed in all four paleofeces dating from the Bronze Age to the Baroque period, which would indicate quite a recent change in the gut community. However, to spot the critical time points when this shift in the human gut microbiome began requires more ancient samples spanning a wider time range; of particular interest would be samples from the past two or three centuries, where major dietary and medical changes occurred. Overall, our results support the theory that the shift from traditional- to an industrial Westernized lifestyle might be the driving force for changing the human gut microbiome from its ancestral state (Blaser 2017; Pasolli et al. 2019; Tett et al. 2019; Sonnenburg and Sonnenburg 2019).

In one of the Iron Age samples (i.e. 2604), the molecular analyses indicated consumption of fermented food and beverages. The fungal analysis revealed a high prevalence of *Penicillium roqueforti* and *Saccharomyces cerevisiae*, with clear signs of domestication, which are nowadays involved in fermenting blue cheeses and alcoholic beverages, respectively. Furthermore, the consumption of walnut (*Juglans regia*) in an Eastern Alpine valley demonstrates that exotic foodstuffs were also part of long-distance trade networks during prehistory (Scott et al. 2021).

Following rapidly in the wake of ruminant animal domestication (mainly cattle, sheep, and goat), cheese production represents one of the oldest and widespread food preservation techniques developed by humans (Fox et al. 2017). The oldest reported chemical evidence for processing of milk into fermented products (i.e. cheese making) is dated back to 5400 BCE in northern Europe (Salque et al. 2013). Other indications, including actual preserved pieces of cheese, whey strainers, and recipes for cheese production, were found in the middle east and the near east and in the Mediterranean basin (Bottéro 1985; Greco et al. 2018; McClure et al. 2018). Here, we report strong evidence for the domestication of the fungus *Penicillium roqueforti* in the course of food processing, most likely to produce a cheese which closely resembles modern blue

cheeses (non-Roquefort cheese clade, in Figure 4A). To our knowledge, this represents the earliest known evidence for directed cheese ripening and affinage in Europe, adding a crucial aspect to an emerging picture of highly sophisticated culinary traditions in European prehistory (Heiss et al. 2021). Importantly, the production of blue cheese involves a surface application of dry salt and therefore it is characterized by a high salt content of up to 7.5% (w/w) (Cakmakci et al. 2012). The cheese curd could have been collected, desiccated, and inoculated with the fungi in wooden cheese containers like the ones excavated in the Hallstatt mines (Figure 4C). The presence of *P. roqueforti* indicates a major step in ruminant milk processing from fresh to ripened cheese, which offered in addition to new flavors several advantages to the Hallstatt miners including longer storage, i.e. months, and less lactose content in the fermented dairy product (Monti et al. 2019). The latter reduced lactose content may have helped the ancient minors to better digest milk products, living in a time where lactose persistence frequencies only started to rise in Europe (Burger et al. 2020). The presence of salt as well as the constant temperature (8℃) and humidity inside the Hallstatt mine workings represent ideal conditions for blue cheese production, following the current cheese production standards (Cantor et al. 2017). It is noteworthy that the early discovery of the Roquefort cheese was linked to Roquefort-sur-Soulzon caves, in France, which maintain a temperature of 10 ℃ and ~ 90% humidity over the year. With such conditions protecting the cheese from desiccation, these caves have been used exclusively for centuries for ripening and ageing of the “Roquefort” cheese (Capozzi and Spano 2011; Desmasures 2014).

Indications for the production of fermented alcoholic beverages in prehistory are abundant, albeit frequently ambiguous (Heiss et al. 2020), and can be found in the Near East, Middle East, Far East, and Europe (Schöttke and Rögener 2021; Liu et al. 2018; J. Wang et al. 2016; Guerra-Doce 2015). Evidence for the production of grape wine in Europe and viniculture in the Near East date back to the 6th and 7th millennium BCE (McGovern et al. 2017). Such evidence was mainly based on chemical residue analysis, archaeobotanical analysis or indicated in ancient inscriptions. Recently scientists claimed that they were able to revive an ancient yeast strain from Egyptian potteries and used it to ferment beer (Aouizerat et al. 2019). Here we were able to reconstruct > 90% of the *S. cerevisiae* genome from an Iron Age-dated paleofecal sample. We used different molecular analysis at the genome level to infer the possible routes of domestication for this yeast. Our results strongly suggested its being used in beer fermentation. Together with the results of the dietary analysis that showed

presence of different fermentable cereals, e.g. wheat, barley, and millets, we can envisage how the fermentation was carried out.

It might be assumed that the fermentation was carried out in a spontaneous manner, i.e. by adding water to wort, and allowing the fermentation process to take place by the wild air-borne yeasts or the constitutional microbiota of the used cereals (Basso, Alcarde, and Portugal 2016). We do not see, however, indications for other yeasts species, such as *Brettanomyces bruxellensis*, that co-occur in spontaneously fermented beers (Spitaels et al. 2014). In addition, we see clear indications of domestication and continuous supply of new admixture components to this yeast, which might suggest that fermentation vessels were repeatedly used for this purpose or the inoculation of the fermentation batches has been done by back-slopping, i.e. inoculation of new fermentation batches with portions of previous batches (Spitaels et al. 2017). Albeit varied evidence for beer production in prehistoric Europe exists (Heiss et al. 2020; Guerra-Doce 2015), these beers could not be preserved for longer time periods and would have had to be consumed rapidly after production, which also presupposes that the beer would have had to produced either in Hallstatt itself or in the very near surroundings.

Considering the constant temperature of 8 °C inside the Hallstatt mines, it might be expected that this yeast was used for production of lager-like beer, where fermentation is carried out at low temperatures (also known as bottom-fermentation), and results in a beer that can be stored for longer time periods (E. P. Baker et al. 2019). Historically, however, the bottom-fermentation was most likely developed after the year 1553, when the Duke of Bavaria Albrecht V forbade brewing during summer months (Dornbusch 1998). Additionally, Gonçalves and colleagues demonstrated that *Saccharomyces pastorianus* strains, which are hybrids of *S. cerevisiae* and another *Saccharomyces* species and are used for production of lager beers, belong to the main beer clade (Figure 4D) (Gonçalves et al. 2016). Therefore, we postulate that the beer produced at that time is similar to nowadays pale beer, produced mainly by top-fermenting *S. cerevisiae* strains.

Conclusions

Paleofeces material displays a unique archaeological information source that provides precious insights into the diet and gut microbiome composition of our ancestors. Here we had access to four paleofeces samples from the Hallstatt salt mine dated from the

Bronze Age to the Baroque period. The constant annual temperature and high salt concentrations inside the mine perfectly preserved both plant macroremains and biomolecules (DNA and protein) in the paleofeces. We demonstrate the indispensable complementarity of using microscopic and molecular approaches in resolving the paleofecal dietary residual components and to reconstruct the ancient gut microbiome. Furthermore, we extended our paleofeces microbiome analysis to focusing on key microbes that are involved in food processing which opens new avenues in understanding fermentation history. Additional samples from different time points will provide in the future a more fine-scaled diachronic picture which may help to understand the role of dietary changes in shaping our gut microbiome and how much this was further influenced by modern lifestyles or medical advances recently introduced with the industrialization/Westernization.

2.6 Materials and Methods

2.6.1 Sampling site, paleofeces samples, radiocarbon dating

In this study, we subjected four paleofeces samples to in-depth microscopic and molecular analysis. The paleofeces material stems from Bronze Age and Iron Age mine workings in the Hallstatt salt mines in Upper Austria (Figure 1A, B, C and Table S1). Three paleofeces (2610, 2604, 2611) were recovered in 1983, 1989 (2604) and 2003 (2610) through wet-sieving of larger blocks of prehistoric production debris excavated in the mine workings. One additional paleofeces (2612) was sampled in 2019 in situ at the site Edlersbergwerk-oben with sterile sampling tools. This sample was not subjected to wet-sieving. All four samples (using approx. 500 mg each) were subjected to radiocarbon dating at the Curt-Engelholm-Centre for Archaeometry in Mannheim, Germany (Figure 1C and Table S2). The remaining material has been used for microscopic and molecular analyses.

2.6.2 Microscopic analysis of the paleofeces

Before chemical treatment of the paleofeces the surface of each sample was stripped off to avoid contamination. Then rehydration in a 0.5% solution of trisodium phosphate for 72 hours (Pearsall 2015). After olfactory and visual testing, the liquid was screened through 500, 250, 125, and 63µm steel meshes, and the outwash was kept for further microfossil studies. Plant macro remains were picked out of the fractionated residues in the steel meshes and then identified and quantified under a stereo microscope with magnifications up x63. The identification of the plant remains was conducted with

identification keys (Cappers, Bekker, and Jans 2006; Neef, Cappers, and Bekker 2012) and the reference collection of the Department of Botany, Innsbruck University, was consulted for comparative purposes.

2.6.3 DNA extraction, library preparation and sequencing

The molecular analysis of the paleofeces samples was conducted at the ancient DNA laboratory of the EURAC Institute for Mummy Studies in Bolzano, Italy. Sample documentation, sample preparation and DNA extraction were performed in a dedicated pre-PCR area following the strict procedures required for studies of ancient DNA: use of protective clothing, UV-light exposure of the equipment and bleach sterilization of surfaces, use of PCR workstations and filtered pipette tips. The DNA was extracted from the paleofeces samples (200 mg) using a chloroform-based DNA extraction method according to the protocol of Tang and colleagues with minor modifications (Tang et al. 2008). To test whether the wet-sieving step during the classical archaeological excavation results in an “wash-out” effect of DNA from the paleofeces, we re-hydrated additional 200 mg material of two samples (2611, 2612) in 1ml of 0.5N Na_3PO_4 (tri-sodium phosphate)-buffer for 18h at RT (R. I. Gilbert and Mielke 1985). After re-hydration, the samples were centrifuged (5000g, 5 min) and both the supernatant (2611RW, 2612RW) and the pelleted paleofeces (2611R, 2612R) were subjected to DNA extraction as described above. From all DNA extracts double-indexed libraries were generated for the sequencing runs with a modified protocol for Illumina multiplex sequencing (Kircher, Sawyer, and Meyer 2012; Matthias Meyer et al. 2012). Libraries were first shallow sequenced on an Illumina MiSeq platform and then deep sequenced on Illumina HiSeq2500 and HiSeqX10 platforms using 101–base pair (HiSeq2500) and 151–base pair (MiSeq, HiSeqX) paired-end sequencing kits. For details to the metagenomic datasets please refer to Table S3. Data are available from the European Nucleotide Archive under accession no. PRJEB44507.

2.6.4 Pre-processing of the sequence data, general taxonomic overview, human DNA analysis

The paired Illumina reads were firstly quality-checked and processed (adapter removal and read merging) as previously described in (Maixner et al. 2018). Initially we tested for the above-described possible DNA wash-out effect by assigning taxonomically the microbial reads using MetaPhlAn 3.0 (Beghini et al. 2021) in the shallow sequenced MiSeq datasets of the untreated samples (2611, 2612), the washed and pelleted

paleofeces (2611R, 2612R), and the supernatant (2611RW, 2612RW). For all subsequent analyses, we used the combined sequencing data available for each sample (Table S3). First, we assessed a general taxonomic profile of the sequencing reads using DIAMOND (v2.0.7) blastx search (Buchfink, Reuter, and Drost 2021) against the RefSeq non-redundant protein database (nr). The DIAMOND tables were converted to rma6 (blast2rma tool) format (--minPercentIdentity 97), imported into MEGAN6 software (Huson et al. 2016), and subsequently visualized using the Krona tool (Ondov, Bergman, and Phillippy 2011). Next, we assessed the endogenous human DNA content in the paleofeces samples by aligning the sequence reads against the human genome (build Hg19) (Rosenbloom et al. 2015) and the human mtDNA reference genome (rCRS) (Andrews et al. 1999) using BWA (H. Li and Durbin 2009) with a seed length of 1,000. The minimum mapping and base quality were both 30. To deduplicate the mapped reads, we used the DeDup tool (https://github.com/apeltzer/DeDup). For details to the mapping results please refer to Table S10. The resulting bam files were used to check for characteristic aDNA nucleotide misincorporation frequency patterns and for the fragment length distribution using the DamageProfiler tool (Neukamm, Peltzer, and Nieselt, n.d.). Mitochondrial human contamination rates were assessed using Schmutzi (Renaud et al. 2015). The sex of the mapped human reads was assigned using a Maximum likelihood method, based on the karyotype frequency of X and Y chromosomal reads (Skoglund et al. 2013). Variants in the mitochondrial genome were called using SAMtools mpileup and bcftoools (H. Li and Durbin 2009) with stringent filtering options (quality>30). The haplogroup was identified by submitting the variant calling file to the HaploGrep website (Weissensteiner et al. 2016).

2.6.5 Comparing paleofeces microbiome structure to contemporary metagenomic datasets

To compare the paleofeces microbiome structure with contemporary individuals', we downloaded 9,368 publicly available shotgun metagenomes representing modern-day human populations (N = 9,207), and, as a control, environmental soil sample (N = 161). Each metagenome was characterized by source (human or soil), and in the case of human samples, by body-site, and whether from a Westernized or non-Westernized population. The term "non-Westernized" describes a population that follows a traditional non-urbanized lifestyle encompassing factors such as diet, hygiene, and with no or limited access to medical healthcare and pharmaceuticals (e.g. antibiotics) as previously described (Brewster et al. 2019; Pasolli et al. 2019; Tett et al. 2019).

Afterwards, we performed profiling of the microbial composition of each metagenomic sample using MetaPhlAn 3.0 (Beghini et al. 2021) using default settings.

We first analyzed the prevalence of top 15 ancient-sample enriched species in the context of global populations using a subset (N = 8,968) of shotgun metagenomes that are stool samples from healthy adult individuals characterized by non-Westernization (Table S16). A species was counted as presence in a sample if the relative abundance was above 0.01%. Next, we randomly selected 662 metagenomes (132 oral cavity and 530 stool samples) from all publicly available human metagenomes, along with 161 soil metagenomes, for profiling of the metabolic pathways using HUMAnN 3.0 (Franzosa et al. 2018) with default settings (Table S17). The Python script humann_renorm_table.py as part of the pipeline was used to normalize the metabolic pathways to relative abundance, followed by converting individual profiles into a merged abundance table using the utility script humann_join_tables.py. For the same selected metagenomic samples, the individual profiles of microbial composition generated by MetaPhlAn 3.0 as described above were similarly merged into a single abundance table using a python script merge_metaphlan_tables.py. Subsequently, the merged abundance tables were used for calculating Bray-Curtis's distance with logarithmic transformation based on which a Principal Coordinates Analysis was performed using python package scikit-bio (version 0.5.6; http://scikit-bio.org/).

2.6.6 Molecular characterization of ancient diet

The merged reads were searched for homology, using BLAST search, against different specific DNA barcoding databases (Figure S4). Specifically, the databases of plant Internal transcribed spacers (Banchi et al. 2020), RuBisCo-large subunit and Maturase K genes (rbcL/matK, v3.boldsystems.org), and chloroplast genomes (ncbi.nlm.nih.gov/genome/organelle) were used to identify the plant dietary components. The BOLD database of the cytochrome c oxidase subunit I gene (COI, v3.boldsystems.org) was used to specifically target the animal diet and intestinal parasites. The UNITE database of the fungal ITS gene database (unite.ut.ee) was used to identify potential food processing fungi. Finally, to target all the eukaryotic dietary and diet-related components, the reads were aligned against the currently available full mitochondrial and chloroplast genomes of the NCBI database (Coordinators and NCBI Resource Coordinators 2012) using BWA (H. Li and Durbin 2009) with default parameters. Subsequently, to obtain a taxonomic overview of the aligned reads we performed a sequence similarity search using BLASTn (Altschul et al.

1990) with default parameters against the complete NCBI-nt database (Coordinators and NCBI Resource Coordinators 2012).

Both the resulting BLAST and the previously created DIAMOND tables were converted to rma6 (blast2rma tool) format and imported into MEGAN6 software (Huson et al. 2016). The read counts at genus- and species-level were exported as csv files and imported into R-Studio software for further analysis.

To distinguish the true hits from the false hits, we applied two filters as follows: i) For the BLAST and DIAMOND tables, we only considered hits of ≥ 90 % identity and ≥ 90 % coverage of each query read; then ii) We considered the sample positive for a certain component only if it shows incidence in majority of databases of that component, e.g. the sample 2604 was considered positive for *Bos taurus*, because it contained true hits for mitochondrial-, COI-, and nr-databases; and finally iii) for the plant dietary components, we used the plantiTS database as a primary filter, since it is highly curated and recently updated. So, we initially filtered out all the plantiTS negative taxa then considered the same majority rule.

Finally, we used transformed sums of counts of the genera and species that passes the aforementioned filters into normalized log base 2 ($\log_{10}$ reads million^{-1}) and plotted them as dot-plots proportional to their abundances. The dendrogram to the left of the dot-plot represent the phylogeny based on NCBI taxonomy (ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi).

Selected identified taxa were further subjected to in-depth analysis by mapping the quality-filtered sequence reads against organellar (mitochondrial and chloroplast) and nuclear genomes using bowtie2 (v1.2.1.1) and the parameter "*end-to-end*" (B. Langmead and Salzberg 2013) (for details to the references used please refer to the Table S10). In two cases where the previously described molecular dietary analysis resulted in an assignment down to the genus level (*Triticum* spp., *Ascaris* spp.) we assessed an appropriate reference sequence by mapping the sequence reads against selected chloroplast and mitochondrial genomes using BWA (H. Li and Durbin 2009) (with default parameters) implemented in the program FastQ Screen (Wingett and Andrews 2018), and selected as reference the species that belongs to one of these two genera and has the most specific hits. After the bowtie2 alignment against the reference sequences the mapped reads were deduplicated and checked for damage patterns as described above for the human DNA. For details to the mapping results

please refer to Table S10. Organellar sequences that showed higher than 70% read coverage were further subjected to phylogenetic analysis as previously detailed (Maixner et al. 2018). In brief, a consensus FASTA sequence was generated using the ANGSD tool (Korneliussen, Albrechtsen, and Nielsen 2014) and together with other comparative datasets subjected to a multiple alignment using the MAFFT multiple sequence alignment program (Katoh et al. 2002). Phylogenetic analysis were performed by applying distance-matrix, maximum-parsimony, and maximum-likelihood methods implemented in the ARB software package (Ludwig et al. 2004): neighbor-joining (using the Jukes-Cantor algorithm for nucleic acid correction with 1000 bootstrap iterations), DNA parsimony (PHYLIP version 3.66 with 100 bootstrap iterations), and DNA maximum-likelihood [PhyML (Guindon and Gascuel 2003) with the HKY substitution model]. The number of informative nucleotide positions used for the phylogenetic analysis and the bootstrap support is indicated in the respective figure captions.

2.6.7 Genome-level analysis of ancient fungi - Variant calling

The sample 2604 showed high prevalence of *Penicillium roqueforti* and *Saccharomyces cerevisiae* DNA allowing further genome-level comparative analysis with modern datasets (Table S12, Table S13). As comparative datasets, we included the recently published genomic data of 34 *P. roqueforti* (Dumas et al. 2020) as well as 157 *S. cerevisiae* (Gallone et al. 2016).

We mapped the quality-filtered reads against the reference genomes of *Penicillium roqueforti* FM164 and *Saccharomyces cerevisiae* S288c, independently, using bowtie2 (v1.2.1.1) and the parameter "*end-to-end*" (B. Langmead and Salzberg 2013). Using samtools (H. Li and Durbin 2009), the mapping qualities of < 30 were filtered out and the bam files were sorted and indexed. The read duplicates were marked and removed using the DeDup tool (github.com/apeltzer/DeDup/) and new read-groups were assigned using the "*AddOrReplaceReadGroups*" of Picard tools (broadinstitute.github.io/picard/). Then, we used the Genome Analysis Toolkit 4 (GATK4) (gatk.broadinstitute.org) to call genome variants, following the best practices workflow as follows: i) Variants within each of the single bam files were called using the tool "*HaplotypeCaller*"; ii) The resulting VCF files were combined into a single file using the tool "*CombineGVCFs*"; iii) The combined VCF file was then genotyped using the tool "*GenotypeGVCFs*"; iv) InDels were filtered out and only single nucleotide polymorphisms (SNPs) were kept using the tool "*SelectVariants*"; and finally v) the VCF

file was filtered to include SNPs with minor allele frequency (MAF) of ≥ 0.05-0.1 and ≤ 10% missing data. A total of 120359 and 375629 SNPs were resolved for *P. roqueforti* and *S. cerevisiae*, respectively.

2.6.8 Phylogenetic- and population structure analysis

To analyze the phylogenetic relationship of the stains, we used the vcf-kit to convert the SNP datasets (vcf files) to fastA alignment format (Cook and Andersen 2017). Then we used the Fasta2phylip.pl script to create PHYLIP alignment files (mullinslab.microbiol.washington.edu), which were then subjected to maximum likelihood (ML) phylogenetic analysis using RAxML (v8.2.12), following generalized time reversible substitution model (GTR) and GAMMA model of rate heterogeneity (Stamatakis 2014). Best-scoring ML trees were searched after 1000 bootstrap replicates and visualized and annotated using the interactive tree of life, iTOL tool (itol.embl.de).

To infer the degree of admixture and the number of populations in the analyzed datasets (120337 SNPs for *P. roqueforti* and 136712 *S. cerevisiae*), we carried out unsupervised population structure analysis using ADMIXTURE (v1.3.0), testing K in the range of 3 to 10 (Alexander and Lange 2011). The best K values were determined based on the lowest cross-validation error (cv) and maximum likelihood.

2.6.9 Functional marker genes

To infer the potential functions of the modern and ancient *Saccharomyces cerevisiae* yeast strains based on their genome sequence, we searched for the presence of functional marker genes (*RTM1*, *BIO1*/*BIO6*, and Regions A/B/C) that were recently described by (Gonçalves et al. 2016). The *RTM1* gene is responsible for conferring resistance to the toxicity of molasses, and therefore it is assumed to be positively selected in beer yeast strains (Ness and Aigle 1995). While the genes *BIO1* and *BIO6*, which are involved in *de novo* synthesis of biotin, are highly selected in sake fermenting yeasts (Wu, Ito, and Shimoi 2005). The regions A, B, and C are horizontally transferred genomic regions from other yeast genera, e.g. *Kluyveromyces*, *Pichia*, and *Zygosaccharomyces* (Novo et al. 2009). These regions contain 39 genes distributed over 3 different chromosomes (Table S14) and are assumed to play a role in wine fermentation. To search for these marker genes, we used bowtie2 to map the quality filtered reads of our sample as well as the recently published *S. cerevisiae* comparative genomic data (Gallone et al. 2016), against the individual genes. Then, we considered

a gene as present if it is covered ≥ 90 % with 3x depth (Table S10). For phylogenetic tree annotation (Figure 3D), the regions A/B/C were considered present if any of the genes was positive, while *BIO1*/*BIO6* were considered positive if both genes were positive.

2.6.10 Proteomic analysis

Paleofeces samples (ID 2610, 2604, 2611, 2612) and a blank control sample to highlight any contamination were resuspended in 800 µL 5% sodium dodecyl sulfate, 50 mM triethylammonium bicarbonate (TEAB, pH 8.5) and disrupted at 4°C using three 2.8 mm ceramic beads (Qiagen, USA) and a Precellys 24 homogenizer (Bertin Corp, USA) at 6500 rpm for 30 s. Samples were transferred to Eppendorf tubes and centrifuged at 4,000 rpm for 7 min, the supernatant removed and centrifuged twice at 13,000 g for 10 min. The supernatant was subjected to the S-Trap™ mini spin column digestion protocol (ProtiFi, USA). Briefly, samples were acidified with 12% phosphoric acid in water (pH ≤ 1). Binding/wash buffer (100 mM TEAB (final) in 90% methanol) was added to each sample at 6.4 x the volume of each sample. Samples were vortexed and applied to the S-Trap column in ≤ 500 µL aliquots until the entire sample was loaded. After each sample load the S-Trap column was centrifuged at 4,000 g for 30 s. Samples were cleaned by adding 3x 400 µL binding/wash buffer and centrifugation at 4,000 g for 30 s after each wash, and 1 min after the last wash. Proteins were digested on the S-Trap at 37 °C for 13 h by adding 125 µL digestion buffer containing trypsin gold (Promega, USA) in 125 mM NH_4HCO_3. The S-Trap was rehydrated with 100 µL 125 mM NH_4HCO_3 and peptides eluted with 80 µL 125 mM NH_4HCO_3, 80 µL 0.2% formic acid in water, and 80 µL 50% acetonitrile in water. Peptide eluates were pooled and dried under centrifugal evaporation (Savant, Thermo-Fisher Scientific, USA).

Samples were analyzed by high-resolution nano LC-MS/MS on an Orbitrap Eclipse Tribrid mass spectrometer (Thermo-Fisher Scientific, USA) equipped with an Easy-nLC 1000 (Thermo-Fisher Scientific, USA). Peptides were resolubilized in 50 µL 0.1% formic acid in water. 5 µL of each sample were loaded onto a 2 cm Acclaim PepMap 100 trap column (75 um ID, C18 3 µm, Thermo-Fisher Scientific, USA) and separated using a 50 cm C18 2 µm Easy Spray column (ES903, Thermo-Fisher Scientific, USA) using a 120-minute gradient. Mobile phase A consisted of 0.1% formic acid in water, and mobile phase B consisted of 0.1% formic acid in acetonitrile. The gradient used 5% to 22% mobile phase B over 105 minutes, followed by separation from 22% to 32%

mobile phase B over the remaining 15 minutes. The column was washed and equilibrated at 95% and 5% mobile phase B, respectively, over 40 minutes following each run. The flow rate was set at 300 nL/min and the column was heated at 45°C. Mass spectra were acquired on the Eclipse mass spectrometer using data dependent acquisition (DDA) with dynamic exclusion. The precursor scan range was from 375-1500 *m/z* at 120,000 resolution and 100% normalized target AGC (4e5) with a 50 ms maximum injection time. The duty cycle was set to acquire as many MS/MS spectra as possible over a three second period following each precursor scan. A 1.2 *m/z* selection window was used to acquire MS/MS spectra at 30,000 resolution, a normalized AGC target of 100% (5e4), 54 ms maximum injection time, and fragmented using HCD with a normalized collision energy of 30. Dynamic exclusion was set to 60 seconds, with peptide match set to preferred and isotope exclusion turned on. Charge exclusion was set to 1 and greater than 7.

The raw mass spectra files were converted to mzML (Martens et al. 2011) format using msConvert from ProteoWizard (Kessner et al. 2008) and analyzed using the Trans-Proteomic Pipeline (TPP v6.0.0-rc16 Noctilucent) (Deutsch et al. 2015). The analysis pipeline consisted of database searching with Comet (version 2018.01 rev. 4) (Eng, Jahan, and Hoopmann 2013) against a 17-species UniProt FASTA database (Table S18) and shuffled decoy sequences. Comet parameters included variable modifications of +15.994915 Da (Met, Cys, His, Tyr, Trp, and Phe), and +0.984016 (Asn and Gln). A precursor tolerance of 20 ppm was set, a fragment bin tolerance of 0.2 and fragment bin offset of 0. Semi-enzymatic cleavage with up to 3 missed cleavages was allowed, and an isotope error tolerance of 3. Peptide-spectrum matches (PSMs) were validated using PeptideProphet (Andrew Keller et al. 2002) and iProphet (Shteynberg et al. 2011). Protein inference was performed using ProteinProphet (Nesvizhskii et al. 2003). Relative quantitation of protein groups was performed using StPeter (Hoopmann et al. 2018). Observed protein groups for each sample were derived from the set of protein groups validated with ProteinProphet at an estimated 1% false discovery rate (FDR). Those protein groups were then curated by removing any group that intersected with the set of protein groups identified in the blank control analysis. A list of observed peptides was created from the peptide sequences that were used to infer the final set of protein groups. The peptide list was then filtered to contain only those peptides with a probability below a 1% FDR threshold as determined from the iProphet analysis.

Further, the peptide sequences of each sample were imported into Unipept 4.0 Desktop version (Singh et al. 2019), and subjected to LCA assignment, and filtered out

the following: i) peptides assigned to bacteria; ii) peptides assigned to multiple species; iii) single peptide incidences that uniquely infer a species presence (Data S1). Finally, functional annotation of the samples was carried out on different taxonomic levels to infer the molecular function of the obtained peptides.

2.7 Supplementary materials

2.7.1 Supplemental figures

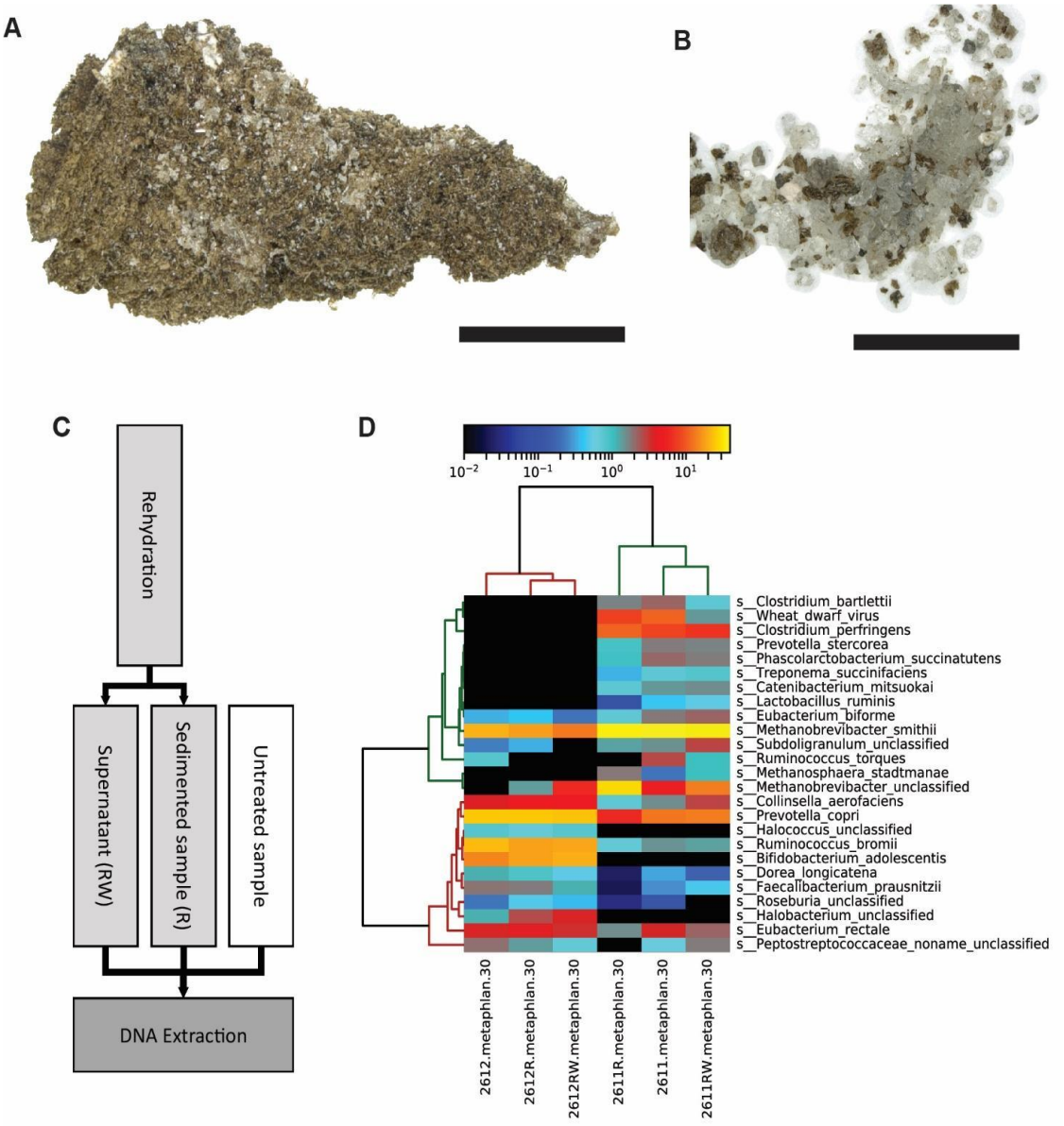
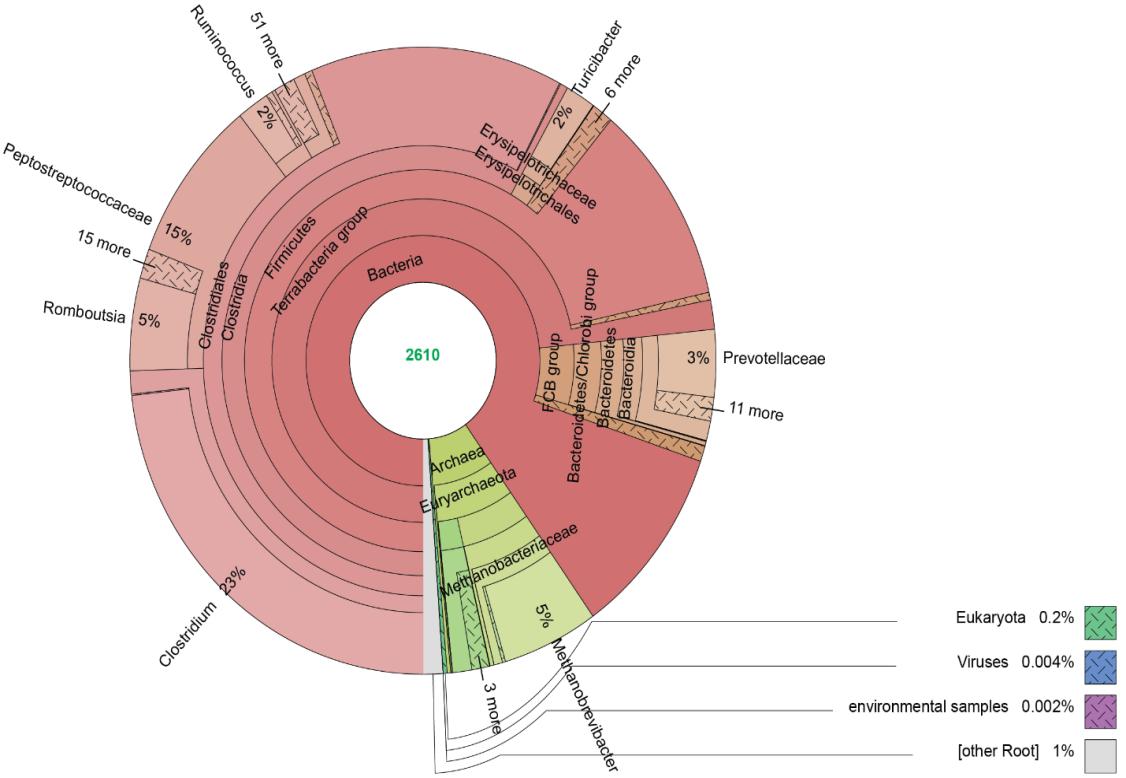


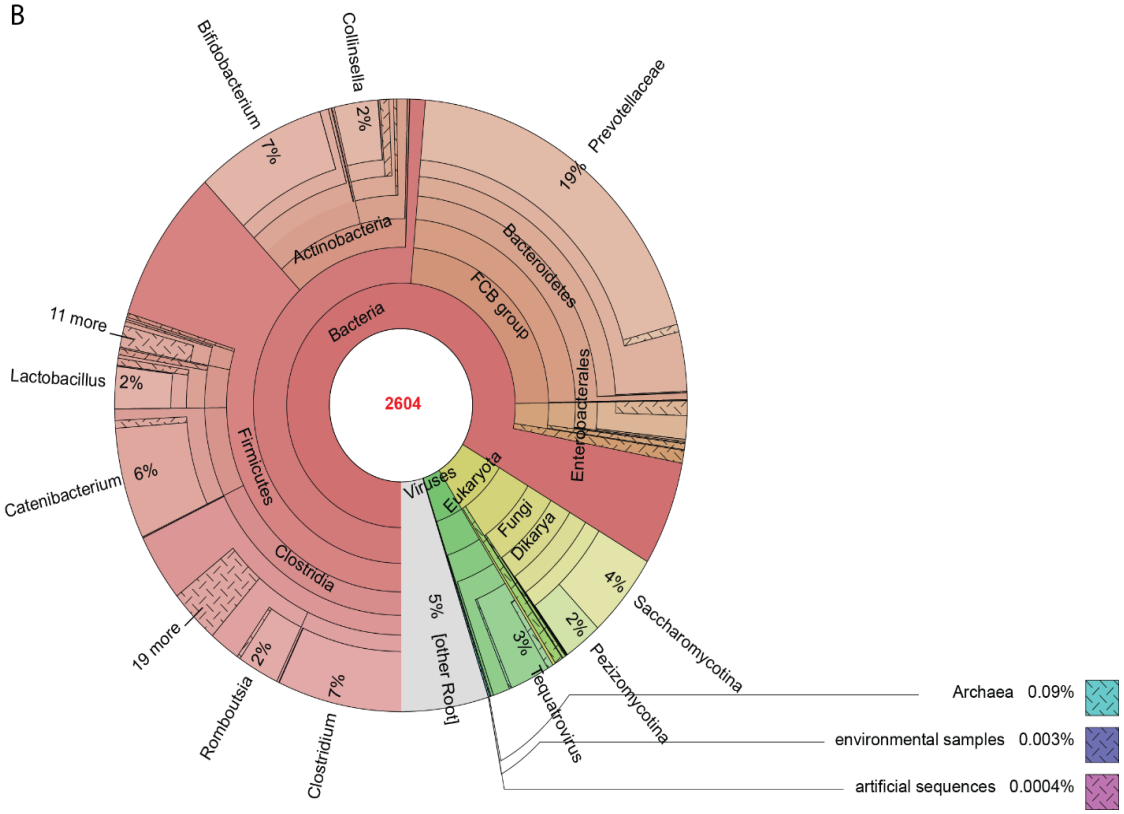
Figure S1. Wet-sieving step during the classical archaeological excavation results in an "wash-out" effect of DNA from the paleofeces. Before subjecting

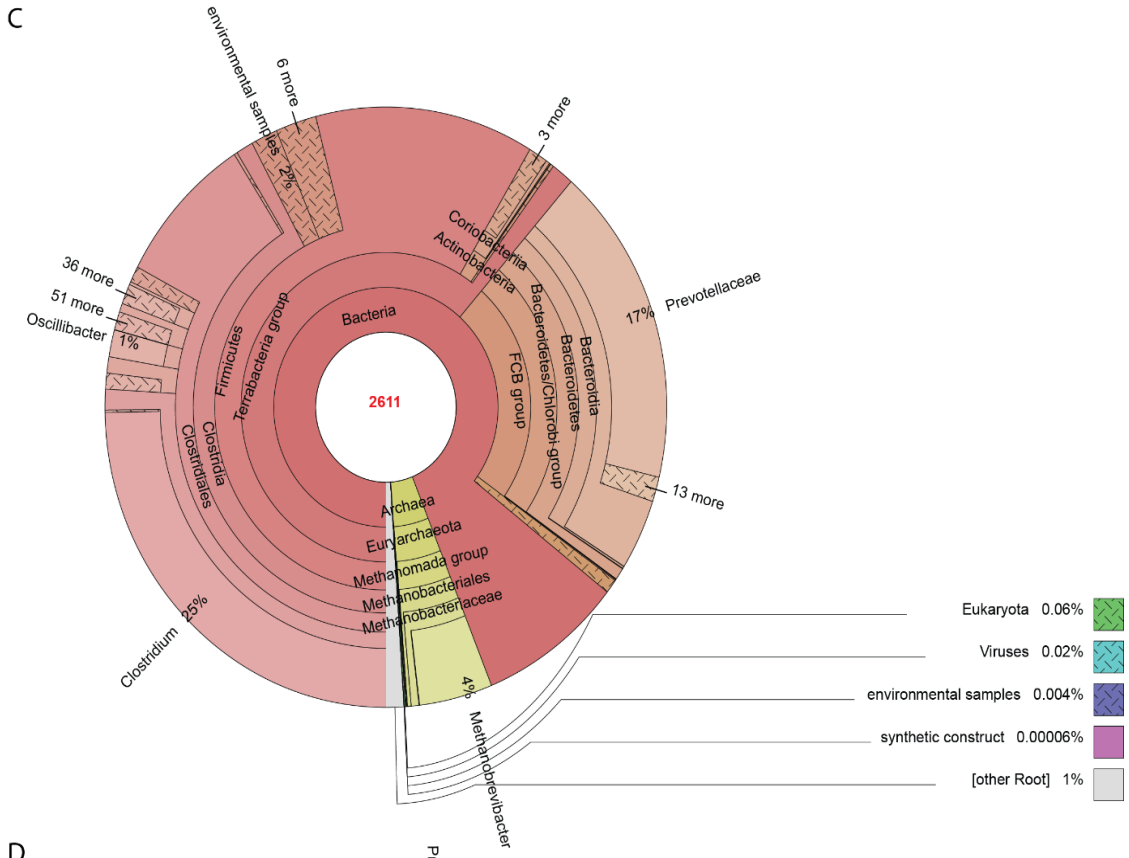
the material to further in-depth molecular analysis, we tested whether the currently used wet-sieving step during the archaeological excavation and recovery of the paleofeces from the salt mine may have a negative “wash-out” effect on the DNA that is initially present in the material. We tested this effect on the paleofeces sample 2611, that has been recovered in the year 1983 via wet-sieving, and sample 2612 **(A)** that has not been subjected to wet-sieving is still covered by salt crystals **(B). (C)** Workflow to test the DNA “wash out effect. Both the “fresh” sample 2612 and the wet-sieved sample 2611 were subjected to a re-hydration treatment and DNA has been extracted from the untreated sample, the sedimented sample (R) and the supernatant (RW). **(D)** Taxonomic analysis of the microbial DNA of the untreated, rehydrated (R) paleofeces samples and the supernatant (RW). MetaPhlAn 3.0 heatmap of the top 25 species found in the different MiSeq datasets. Our “wash-out” test analyzing the microbial DNA of shallow sequenced data of two re-hydrated paleofeces (2611, 2612) revealed that “free” endogenous DNA becomes lost into the aqueous phase when the material gets into contact with water (Figure S2). Importantly, however, even after 18h of re-hydration, the paleofeces material still retains sufficient DNA that resembles the bacterial taxonomic profile of the original untreated sample. Therefore, we also consider the wet-sieved samples (2604, 2610, 2611) which were considerably shorter in contact with water as an important information source for our molecular analysis.

A


Ruminococcus
51 more
2%
Turicibacter
6 more
2%
Erysipelotrichaceae
Erysipelotrichales
Peptostreptococcaceae
15%
15 more
Romboutsia
5%
Clostridiales
Clostridia
Firmicutes
Terrabacteria group
Bacteria
2610
FCB group
Bacteroidetes/Chlorobi group
Bacteroidetes
Bacteroidia
3% Prevotellaceae
11 more
Archaea
Euryarchaeota
Methanobacteriaceae
5%
Clostridium
23%
3 more
Methanobrevibacter
Eukaryota 0.2%
Viruses 0.004%
environmental samples 0.002%
[other Root] 1%

B


Bifidobacterium
7%
Collinsella
2%
Actinobacteria
Prevotellaceae
19%
Bacteroidetes
FCB group
Enterobacterales
11 more
Lactobacillus
2%
Catenibacterium
6%
Firmicutes
Clostridia
Bacteria
2604
Viruses
Eukaryota
Fungi
Dikarya
4% Saccharomycotina
2% Pezizomycotina
3%
Sequatrovirus
5% [other Root]
19 more
2%
Romboutsia
7%
Clostridium
Archaea 0.09%
environmental samples 0.003%
artificial sequences 0.0004%


C
environmental samples
6 more
2%
3 more
Coriobacteriia
Actinobacteria
Bacteroidetes/Chlorobi group
Bacteroidetes
Bacteroidia
17%
Prevotellaceae
13 more
36 more
51 more
Oscillibacter
1%
Bacteria
2611
Firmicutes
Terrabacteria group
Clostridia
Clostridiales
FCB group
Archaea
Euryarchaeota
Methanomada group
Methanobacteriales
Methanobacteriaceae
4%
Clostridium
25%
Methanobrevibacter
Eukaryota 0.06%
Viruses 0.02%
environmental samples 0.004%
synthetic construct 0.00006%
[other Root] 1%

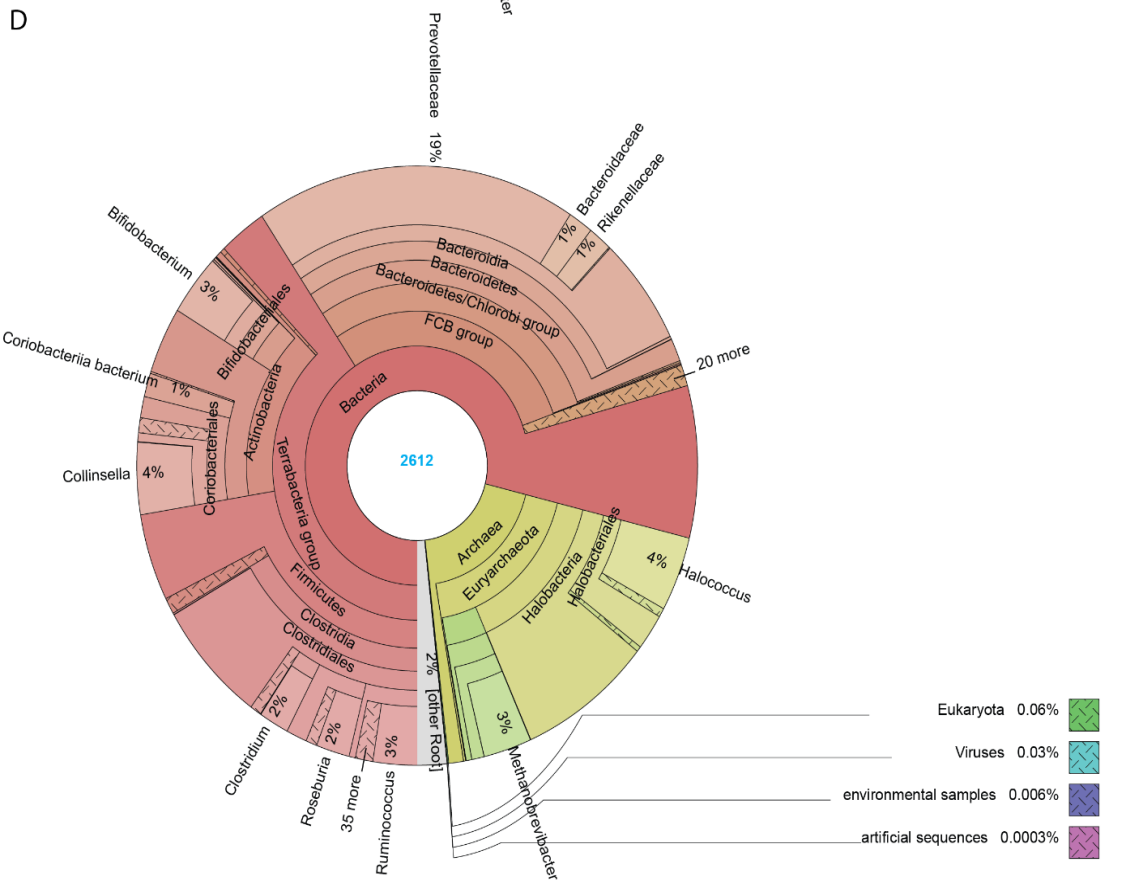

D
Prevotellaceae
19%
Bacteroidaceae
Rikenellaceae
1%
1%
Bacteroidia
Bacteroidetes
Bacteroidetes/Chlorobi group
FCB group
20 more
Bifidobacterium
3%
Bifidobacteriales
Coriobacteriia bacterium
1%
Coriobacteriales
Actinobacteria
Collinsella
4%
Bacteria
2612
Terrabacteria group
Firmicutes
Clostridia
Clostridiales
Archaea
Euryarchaeota
Halobacteria
Halobacteriales
4%
Halococcus
2%
[other Root]
3%
Methanobrevibacter
Clostridium
2%
Roseburia
2%
35 more
Ruminococcus
3%
Eukaryota 0.06%
Viruses 0.03%
environmental samples 0.006%
artificial sequences 0.0003%

Figure S2. Taxonomic overview of the sequence reads in the merged shotgun datasets of the paleofeces samples 2610 (A), 2604 (B), 2611 (C), and 2612 (D). The metagenomic reads were taxonomically assigned using the Diamond tool (Buchfink, Xie, and Huson 2015) against the NCBI non-redundant protein database.

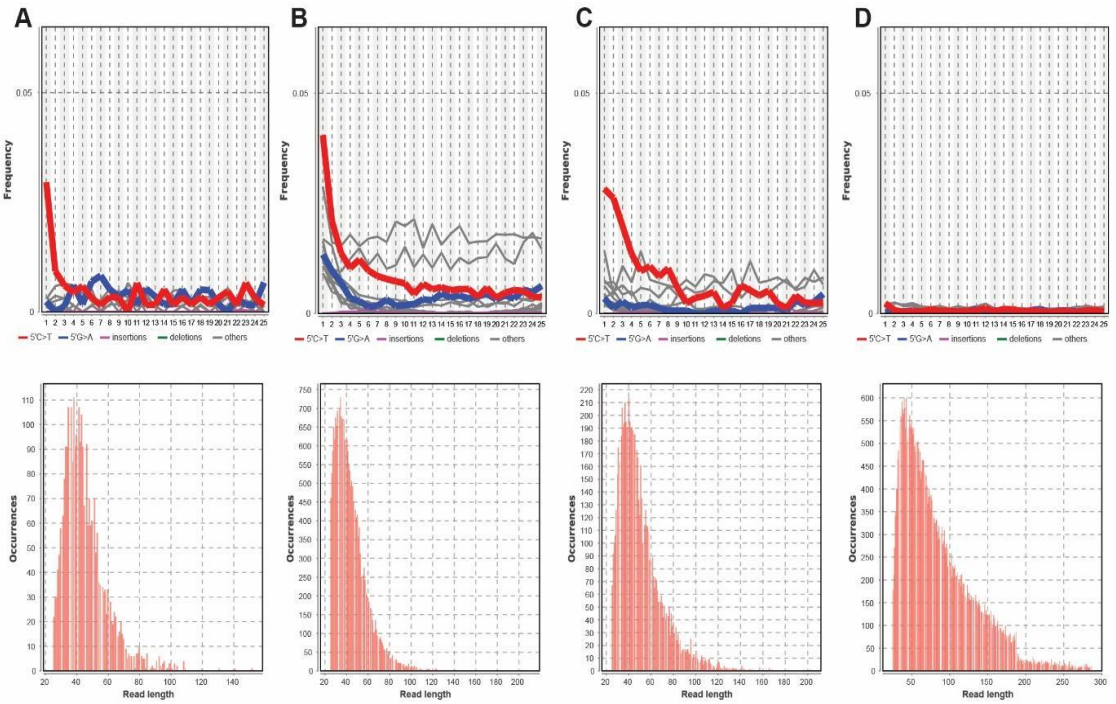


Figure S3. DNA damage patterns (left) and read length distribution (right) of the mitochondrial human reads in the paleofeces samples 2610 (A), 2604 (B), 2611 (C), and 2612 (D). The frequency of C to T base misincorporations at the 5´end of the reads and the length distribution of the aligned reads has been assessed using the DamageProfiler tool. For the details to the reference sequence used please refer to the Supplementary Table S4.

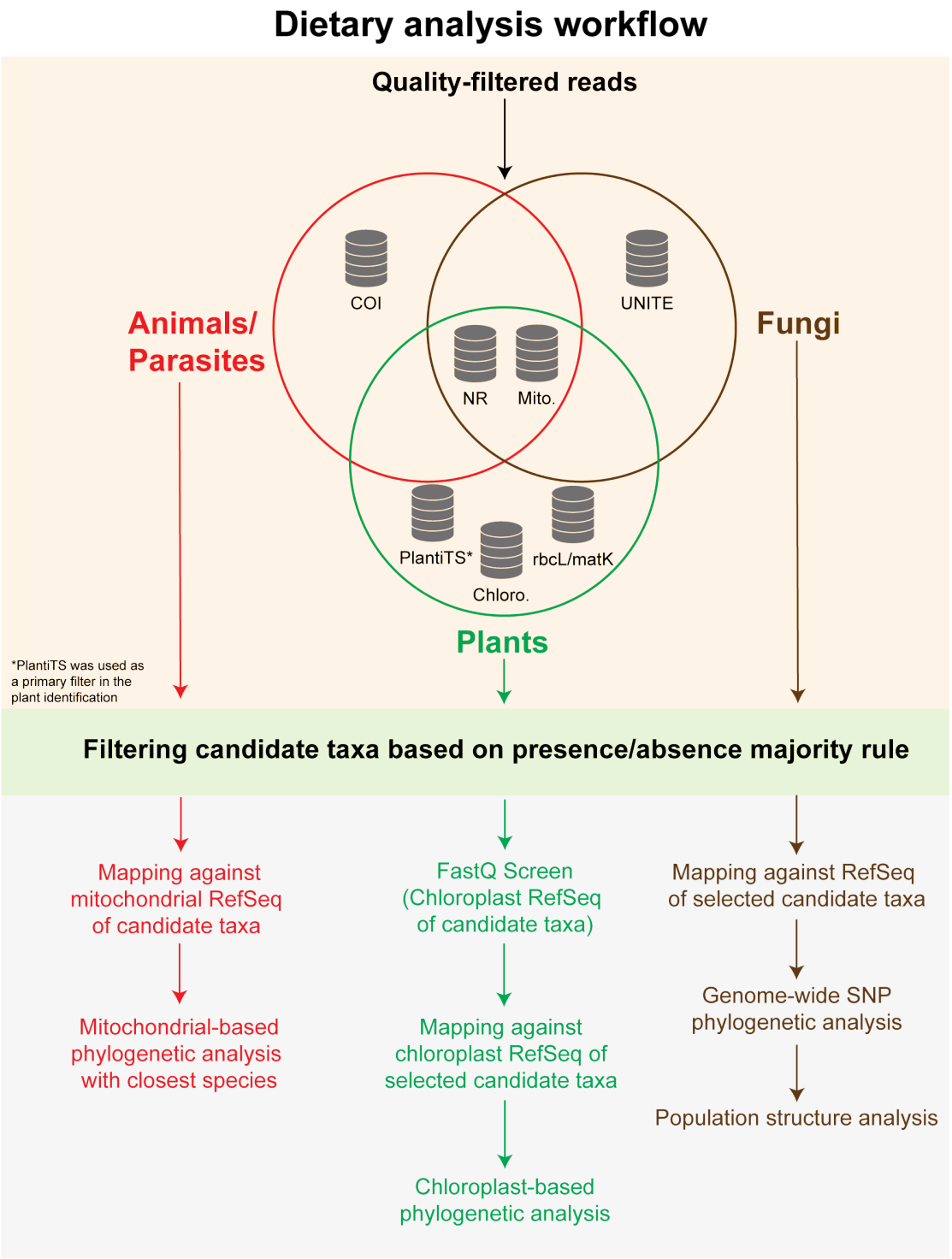


Figure S4. Workflow of the metagenome-based dietary analysis. To identify animal, parasite, plant, and fungal components in the four paleofeces samples we subjected the quality-filtered reads to a homology search (BLAST) against different databases. After filtering candidate taxa using a presence/absence majority rule we subjected selected taxa to further in-depth comparative sequence analysis.

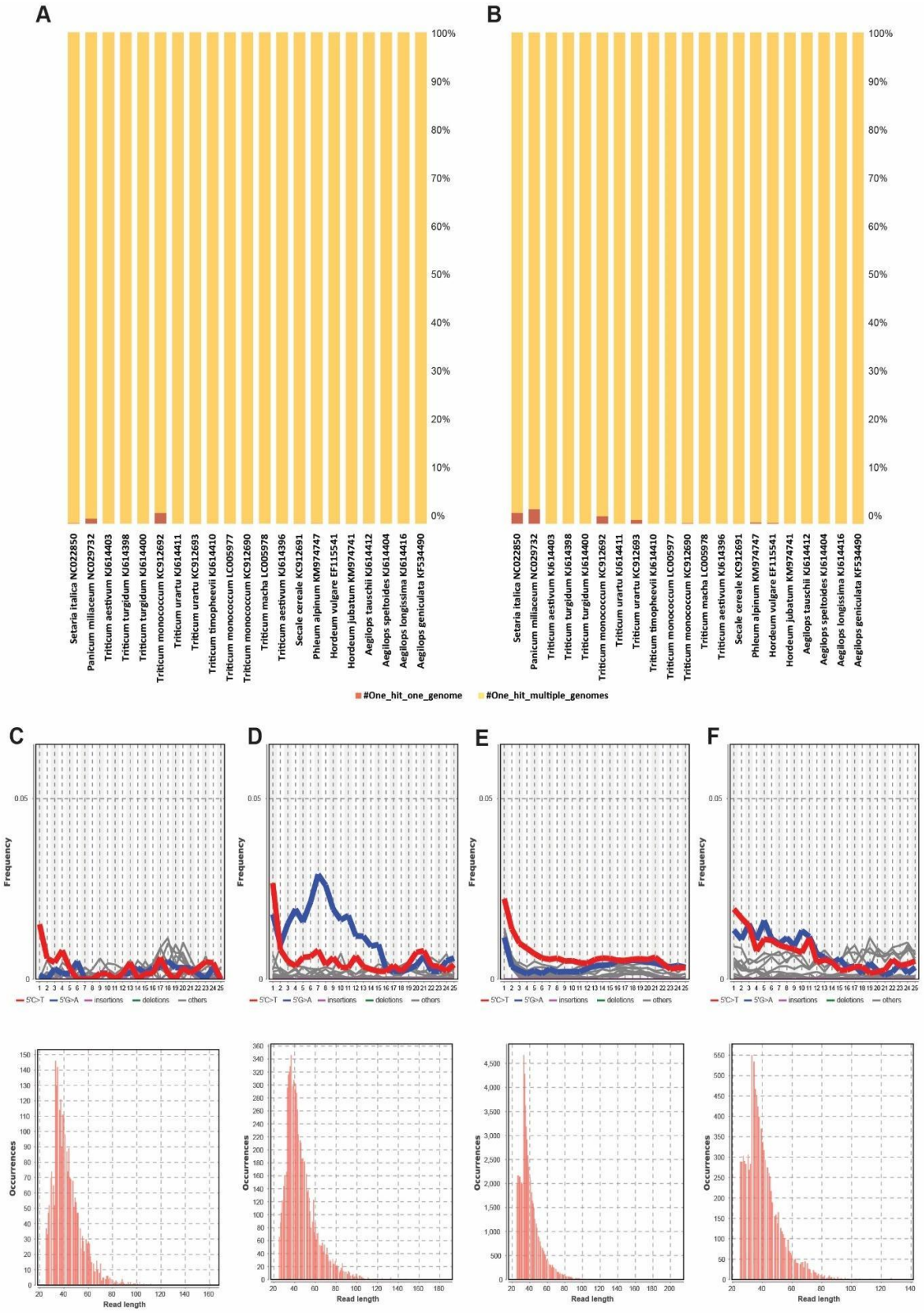


Figure S5. FastQ-Screen- and ancient DNA damage analyses of the ancient plant dietary components in the samples 2610 and 2604. FastQ Screen analysis of the quality-filtered sequence reads of the paleofeces samples 2610 (A) and 2604 (B) against the chloroplast genomes of selected members of the family *Poaceae*.

FastQ Screen (Andrew 2011) was used with default parameters using BWA (Li and Durbin 2010) as the alignment tool. The frequency of C to T base misincorporations at the 5´end of the reads and the length distribution of the aligned reads has been assessed using the DamageProfiler tool. DNA damage patterns and read length distribution of the wheat (*Triticum* spp.) autosomal and chloroplast (C,E) reads and chloroplast reads only (D,F) in the paleofeces sample 2610 (C,D) and 2604 (E,F). For the details to the reference sequence used please refer to the Supplementary Table S10.

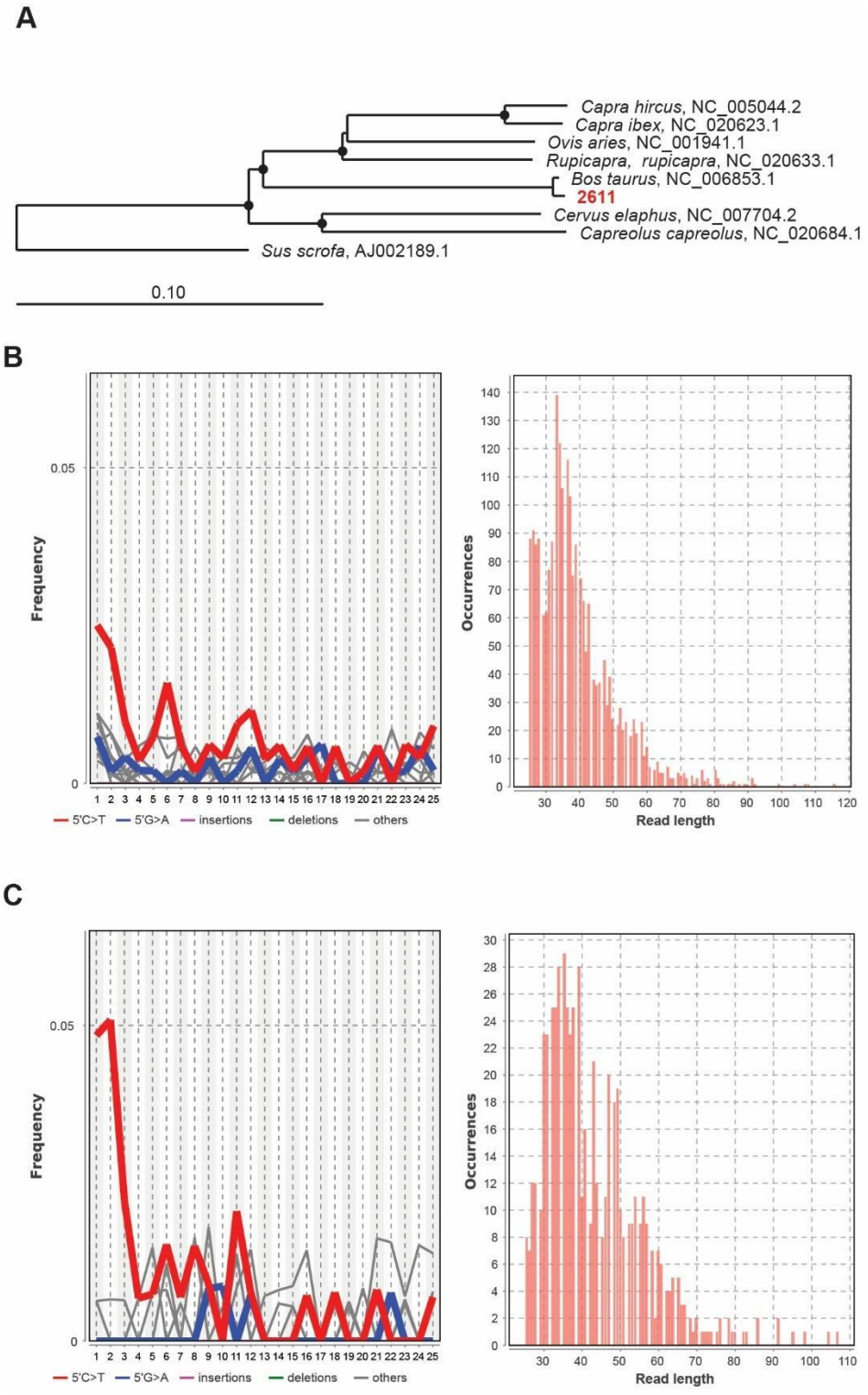


Figure S6. Phylogenetic analysis and ancient DNA damage analyses of the ancient animal dietary components in the samples 2611. A, Phylogenetic assignment of the partial cattle (*Bos taurus*) mitochondrial genome recovered from the paleofeces sample 2611. The displayed tree was calculated using the maximum-likelihood algorithm (PhyML) based on 16,412 informative positions. Black circles symbolize parsimony and neighbor joining bootstrap support (>90%) based on 100 and 1000 iterations, respectively. The scale bar depicts 0.1 substitutions per

residue. The frequency of C to T base misincorporations at the 5´end of the reads and the length distribution of the aligned reads has been assessed using the DamageProfiler tool. DNA damage patterns (left) and read length distribution (right) of the cattle (*Bos taurus*) autosomal and mitochondrial (B) reads and mitochondrial reads only (C) in the paleofeces sample 2611. For the details to the reference sequence used please refer to the Supplementary Table S10.

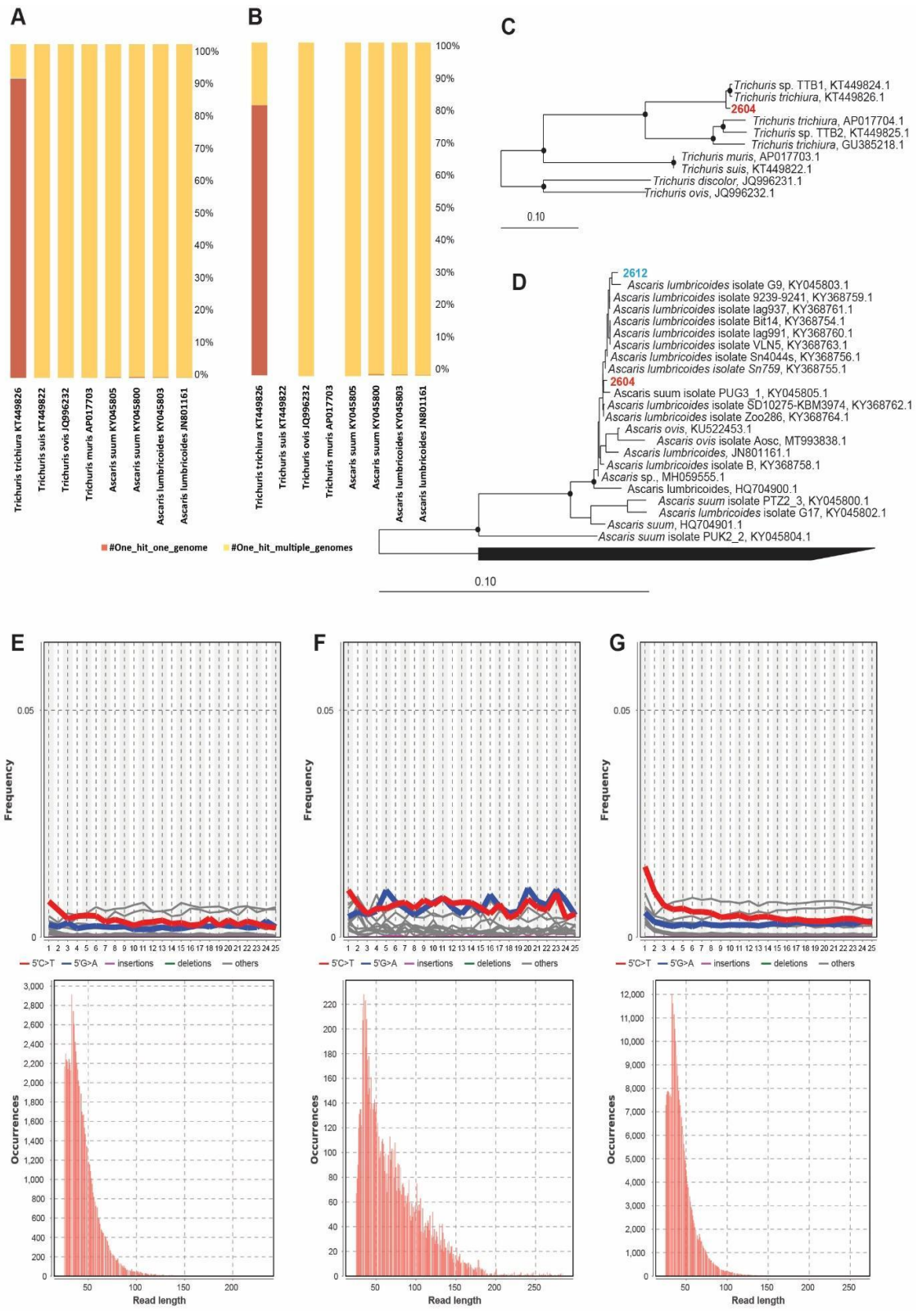


Figure S7. FastQ Screen-, phylogenetic-, and ancient DNA damage analyses of the ancient parasites of the samples 2604 and 2612. FastQ Screen analysis of the quality-filtered sequence reads of the paleofeces samples 2604 (A) and 2612 (B) against the mitochondrial genomes of selected members of the genus *Trichuris* spp. and *Ascaris* spp. FastQ Screen (Andrew 2011) was used with default parameters

using BWA (Li and Durbin 2010) as the alignment tool. C and D, Phylogenetic assignment of the partial mitochondrial genomes of *Trichuris* spp. (A) and *Ascaris* spp. (B) recovered from the paleofeces samples 2604 and 2612. The displayed trees were calculated using the maximum-likelihood algorithm (PhyML) based on 13,189 (*Trichuris* spp.) and 14,265 (*Ascaris* spp.) informative positions. Black circles symbolize parsimony and neighbor joining bootstrap support (>90%) based on 100 and 1000 iterations, respectively. The scale bars depict 0.1 substitutions per residue. The frequency of C to T base misincorporations at the 5´end of the reads and the length distribution of the aligned reads has been assessed using the DamageProfiler tool. DNA damage patterns (top) and read length distribution (bottom) of the roundworm (*Ascaris* spp.) autosomal and mitochondrial reads in the paleofeces sample 2604 (E) and 2612 (F) and of the whipworm (*Trichuris trichuria*) autosomal and mitochondrial reads in the paleofeces sample 2604 (G). For the details to the reference sequence used please refer to the Supplementary Table S10.

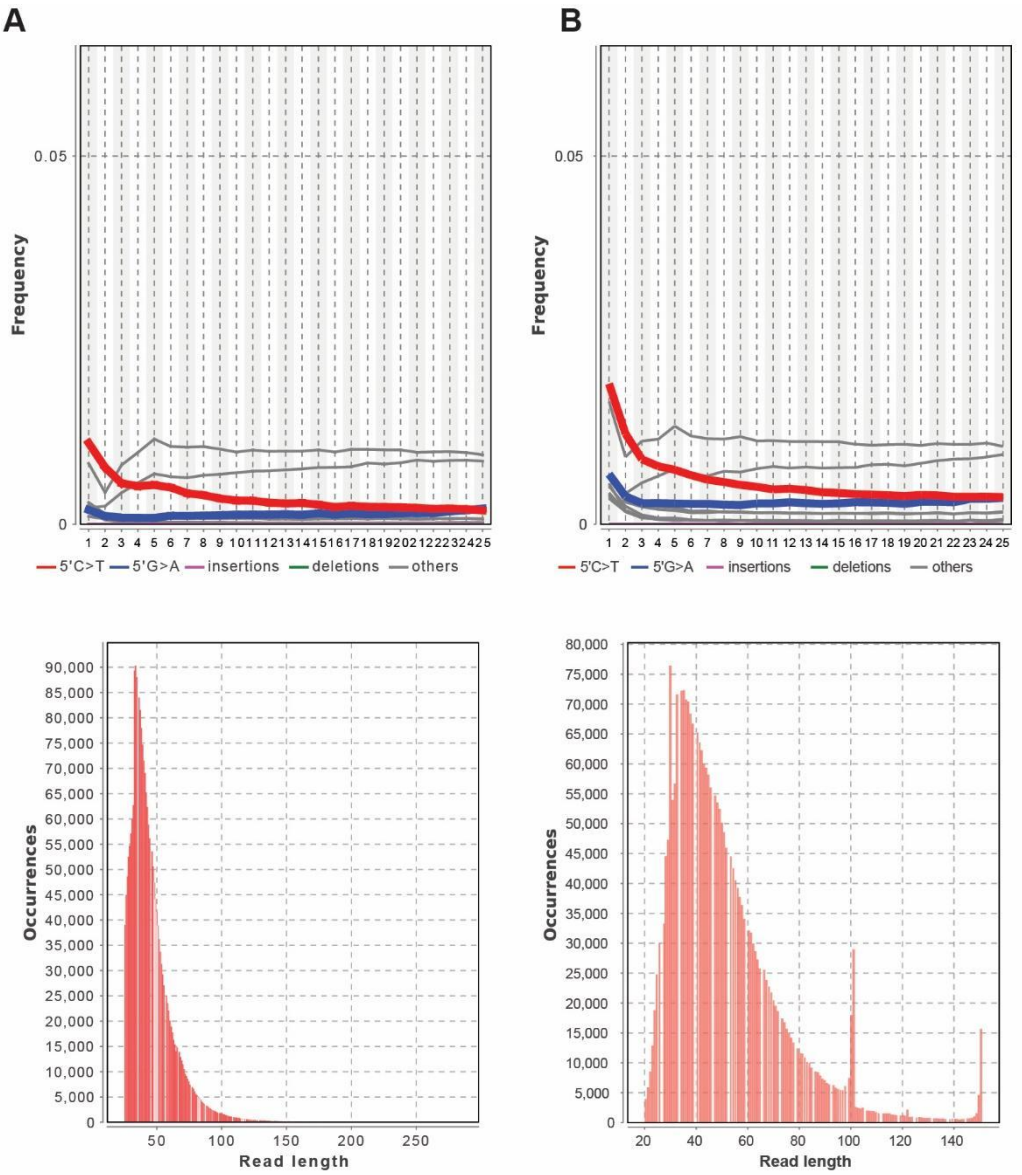


Figure S8. DNA damage patterns (top) and read length distribution (bottom) of the autosomal reads of *Penicillium roqueforti* (A) and *Saccharomyces cerevisiae* (B) in the paleofeces sample 2604. The frequency of C to T base misincorporations at the 5´end of the reads and the length distribution of the aligned reads has been assessed using the DamageProfiler tool. For the details to the reference sequence used please refer to the Supplementary Table S10.

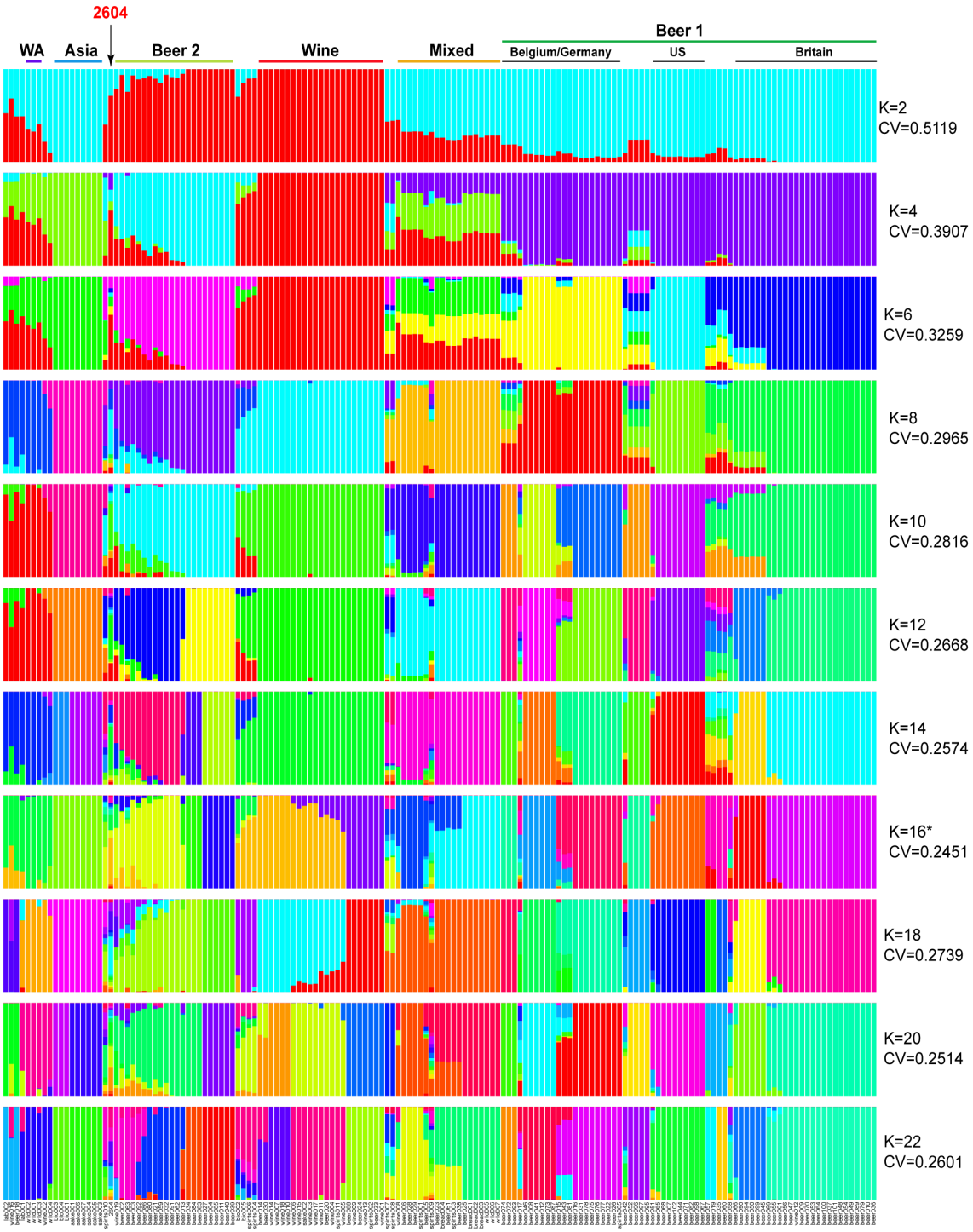


Figure S9. Population admixture profiles as inferred by ADMIXTURE tool. The analysis was carried out assuming K=2 to 22. Asterisk refer to the K values of lowest cross validation error (cv). A total of 136712 SNPs were included in the final analyzed dataset.

2.7.2 Supplemental tables

Tables and the respective captions are available at:
https://drive.google.com/drive/folders/1SJw250VIC13evZWL6p8O6RMAmMZ039nT?usp=sharing

Chapter 3 | MetaClock: An integrated framework to infer evolutionary history of microbes from ancient and contemporary metagenomic data

3.1 Introduction to the chapter

In **Chapter 3** I introduce a novel computational framework, MetaClock, which facilitates strain-level phylogenetics and molecular clock analysis for microbiome members from ancient and present-day populations. **Chapter 2** showed how the general structure of our gut microbiome has diverged from the ancestral state leveraging invaluable information from paleofeces samples. In this Chapter, I will elaborate an automatic computational pipeline which allows for effortlessly studying how and when strains of a species have diverged in history directly using ancient and contemporary metagenomic data. MetaClock can efficiently integrate microbial genomic information from ancient and modern metagenomic data into a whole-genome alignment based on a reference genome. It performs (i) strict quality control via authenticating microbial aDNA, (ii) correction of misincorporated sites, and (iii) mapping of microbial aDNA sequences against multiple modern references. Thus biases can be significantly minimised when inferring downstream phylogenetic analyses. Whole genome divergence times can then be estimated by using, as calibration in the molecular clock analysis, the carbon dated ancient strain. By applying this approach to four paleofeces samples reported in **Chapter 2** and publicly available modern metagenomic datasets, we largely extended the discovery in **Chapter 2** uncovering detailed time-resolved diversification trajectories for six common human microbiome members. This shows a great example of how and when strains of microbiome members have diverged in the course of evolution using ancient and contemporary metagenomic samples.

Contribution: In this study, I formulated the analysis concept with the support from my supervisors Prof. Nicola Segata and Prof. Omar Rota-Stabelli, developed the methodology, and implemented the software pipeline (in Python) and wrote supporting software documentation. I carried out methodology evaluation using (semi)synthetic data, performed evolutionary analysis applying this tool on six human gut microbes, interpreted data and wrote the manuscript.

MetaClock: an integrated framework to infer evolutionary history of microbes from ancient and large-scale metagenomic datasets

Kun D. Huang, Frank Maixner, Adrian Tett, Aitor Blanco-Míguez, Fabio Cumbo, Moreno Zolfo, Lena Granehäll, Mohamed S. Sarhan, Kerstin Kowarik, Hans Reschreiter, Andrea Silverj, Francesco Asnicar, Albert Zink, Omar Rota-Stabelli, Nicola Segata

3.2 Abstract

Microbial Ancient DNA (maDNA) can be extracted from dental calculus, coprolites and mummified specimens so as to help reconstruct the evolutionary history of human-associated microbes. In the last decade, high-throughput sequencing of maDNA coupled with genome-based phylogenetic approaches have shed light on various human pathogenic bacteria. A comprehensive framework to extract, assemble, and analyse maDNA from ancient metagenomes is however still missing, particularly in the context of vast contemporary metagenomic data available. Here, we introduce a novel approach to accurately reconstruct phylogenies from human microbiome species and to estimate their divergence times. The method exploits ancient metagenomes and large-scale modern metagenomic datasets. It is based on the extraction of strain-level genomic information from ancient metagenomes by performing (i) strict quality control via authenticating maDNA, (ii) correction of misincorporated sites, and (iii) mapping of maDNA sequences against multiple modern references. This pipeline assures to minimise biases when inferring downstream phylogenetic analyses. Whole genome divergence times can then be estimated within our computational framework, by using the carbon dated ancient strain to calibrate phylogenies based on contemporary metagenomic data. In this study, we applied this tool to reveal time-resolved evolutionary trajectories of six common human gut bacteria represented by 11,378 draft genomes from >10,000 metagenomes. With the increasing availability of ancient microbial samples, our framework enables strain-resolved molecular-clocked phylogenetic analyses to interrogate the evolutionary history of the human microbiota.

3.3 Introduction

Estimation of divergence times on a phylogenetic framework (molecular clock dating) plays an imperative role in understanding microbial evolutionary history. Divergence estimates allow for example to trace the recent emergence of novel disease-associated strains (Manara et al. 2018), or the appearance of pathogens back to the Paleozoic Era (Lebreton et al. 2017). Divergences are typically estimated from sequence alignments using a molecular clock approach on a bayesian framework (dos Reis, Donoghue, and Yang 2016). Like all other clock studies, those on bacteria require one or more *a priori* time (or rate) information to calibrate the number of mutations observed in the alignments and scale them to time. In the absence of fossils, it is possible for microbial clock studies to use empirically pre-estimated evolutionary rates: this approach is however complicated by the generation time, a piece of information unknown for unculturable bacteria. A more efficient way of calibrating the clock is to use the tip-dating approach based on the age of samples. The presence of at least one ancient strain in an alignment of present time bacteria is now an essential prerequisite because the ancient strain serves as a calibration point for the whole dataset. Bacteria however evolve relatively slowly compared to viruses for which tip-dating is being routinely implemented (Kupczok et al. 2018; Sebastián Duchêne and Holmes 2018; Sun, Wu, and Ke 2017). To accomodate for less number of mutations per site per unit of time, the tip dating approach on bacteria requires therefore a large scale dataset based on the whole genomes.

Ancient microbial genomes for molecular clock studies can be reconstructed from ancient metagenomes, an approach still in its infancy and that has been successfully used in the last few years. For example, the genome of an ancient *Yersinia pestis,* the causal agent of plague, reconstructed from the metagenome of a tooth dated ~5,000 BP was successfully used to calibrate the *Yersinia* phylogeny and to prove that emergence of plague in Eurasia much earlier than initially thought (Rascovan et al. 2019). We recently showed how the genome of an ancient *Prevotella copri* reconstructed from pre-Columbian paleofeces could be used as a calibration to uncover the evolutionary trajectory of the whole *P. copri* complex prior to human migratory waves out of Africa (Tett et al. 2019). Similarly, we have shown the potential of ancient genomes from paleofeces excavated in southwestern U.S.A by using metagenome assembled genomes (MAGs) to infer the divergence times of a key human symbiont *Methanobrevibacter smithii*, which provides a fresh clue of

interrogating the evolutionary history of the human gut microbiota (Wilbowo et al. 2021)

It is clear that maDNA preserved in ancient specimens can provide valuable information for investigating the divergences of the human microbiome. How to precisely reconstruct ancient genomes from ancient metagenomic samples is however a challenging task. The development of computational metagenomics (Nurk et al. 2017; Asnicar et al. 2020; Truong et al. 2017) over the past decade has provided many strategies to study microbial evolution in the human microbiome, but the difficulty becomes apparent when exploiting metagenomic samples drawn from ancient remains. *De novo* metagenomic assembly has emerged as a powerful tool to recover microbial genomic information with particular advantages of reconstructing genomes for uncharacterized species (Pasolli et al. 2019; Karcher et al. 2020; Scholz et al. 2020; Nagarajan and Pop 2013). Nonetheless, metagenomic assembly is complicated in constructing fragmentary short sequences typically characteristic of aDNA into extended contigs, and computationally unbearable in the case of deeply sequenced samples which are common in ancient metagenomics studies (Weyrich et al. 2017; Maixner et al. 2016; Jensen et al. 2019). Obtaining accurate assemblies from ancient metagenomes requires laborious manual curation to mitigate chimeric contamination, strain heterogeneity, and artificial mutations possibly stemming from aDNA damage (Seitz and Nieselt 2017; Luhmann, Doerr, and Chauve 2017). Hence, it is difficult to generalize existing approaches to ancient metagenomic datasets and low abundance microbes.

For pathogenic members such as *Yersinia pestis* and *Mycobacterium tuberculosis* which are highly clonal (Luhmann, Doerr, and Chauve 2017; Dos Vultos et al. 2008) and with access to sufficient isolated reference sequences, assembly-free approaches based on mapping metagenomic reads against reference genomes and extracting single-nucleotide variants (SNVs) is a valid approach to capture evolutionary signals of a species from ancient samples (Key et al. 2020; Schuenemann et al. 2018; Spyrou et al. 2018; Peltzer et al. 2016). Similar strategies which also include non-polymorphic sites have been employed in previous studies (Tett et al. 2019; Weyrich et al. 2017). However, not only do they require tedious separate analysis relying on many computational tools, but also introduce considerable biases in inferring phylogeny when mapping reads against only a single reference (Bertels et al. 2014). As such, it has remained challenging to extract microbial genomic information, reconstruct a reliable strain-level phylogeny, and estimate the evolutionary timeline for a given

species efficiently from ancient metagenomes in the context of enormous contemporary metagenomic data.

Here we introduce MetaClock, an automated and integrated framework for studying microbial evolutionary history exploiting ancient and present-day metagenomic datasets. MetaClock can precisely extract genomic information directly from ancient metagenomic samples and contemporary metagenomic datasets (metagenomes or MAGs), authenticate the endogenous origin of aDNA and accommodate artificial effects applying detailed quality controls. Performing it with multiple reference genomes can significantly enhance the accuracy and reliability for downstream phylogenetic and molecular clocking analyses. In this study, MetaClock allowed us to explore the time-revolved evolutionary history of six gut non-pathogenic bacteria by leveraging a large number of contemporary metagenomic datasets and newly discovered paleofeces metagenomes.

3.4 Results

3.4.1 Automated reconstruction of microbial genomic information from large contemporary metagenomics data and ancient shotgun metagenomes for molecular clocking

We developed a novel computational framework to study the strain-level evolutionary history of microbes, using both contemporary and ancient (meta)genomic data with molecular clock techniques (**Fig. 1a**). The method can take as input both ancient and contemporary (meta)genomic datasets to reconstruct the microbial genomic information for genome-based phylogenetic and molecular clock analyses. It firstly extracts genomic information directly from metagenomic datasets based on a representative reference genome of the species of interest and reconstructs a genome multiple-sequence alignment (MSA). If (meta)genomic datasets are provided without assembly, base quality controls implemented as tunable parameters would be performed on short-sequence reads in order to avoid the noise from strain heterogeneity, sequencing and alignment errors in building consensus sequences with bowtie2 (B. Langmead and Salzberg 2013) (**Fig. 1a** and see **Methods)**. If assembled genomes (single-amplified genomes (SAGs), isolate genomes or MAGs) are provided, a homology-guided strategy is acquired to extract the genomic region of assembled genomes having substantial sequence similarity with the reference, by tunably restricting alignment length and percentage identity of BLASTn hits. Secondly, the genomic information from both ancient and present-day samples are precisely

integrated into one genome-wide MSA based on the same reference genome. Thus, the pipeline captures genomic information from the whole genome and retains sufficiently initial data to select the valuable signals for carrying out strain-level evolutionary analysis in the subsequent steps. In addition, utilities are available for efficient post-processing, such as cleaning of missing information, selection of targeted genome region, results visualization, fast estimation for clocking signals and many others, in order to assist users to achieve their study-dependent aims (see **Methods** and **Software availability**).

Additionally, merging MSAs from mappings to multiple reference sequences allows for unbiased phylogeny reconstruction (Bertels et al. 2014), hence within this framework we exploited this concept and devised a novel strategy of combining multiple MSAs from different references (**Fig. 1b**; **Supplemental Fig. 1**; see **Methods**). This new feature complements existing approaches - they can result in unreliable phylogeny reconstruction due to SNP extraction and mapping to a single reference (Bertels et al. 2014) - which are commonly used in phylogeny reconstruction in aDNA studies (Peltzer et al. 2016; Spyrou et al. 2018; Schuenemann et al. 2018). Taken together, MetaClock provides a complete computational environment to perform fine-grained reliable investigation of microbial evolutionary history at the strain level from both ancient and modern datasets with flexible parameters and without restriction on forms of (meta)genomics data.

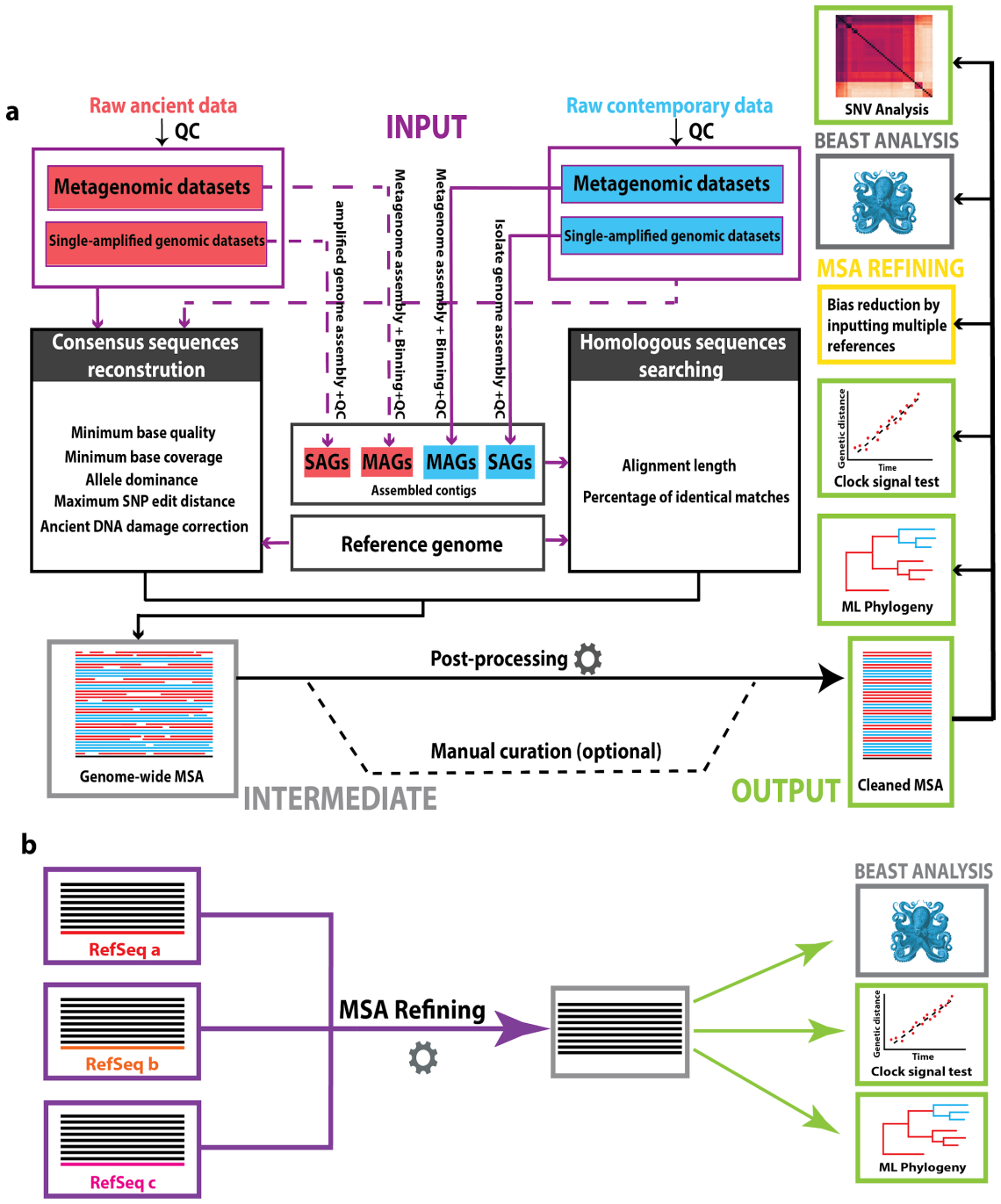


Figure 1. The MetaClock pipeline. **a**, Metagenomic data as either short-sequence reads or assembled contigs (SAGs or MAGs) are taken as input, with tunable parameters for quality control, to reconstruct the genome multiple-sequence alignment (MSA) which is guided by a reference sequence (prepared by users) representative of one microbial species. A rich suite of post-processing and curation features are provided in MetaClock, to generate a cleaned MSA based on which a maximum likelihood (ML) phylogenetic tree is reconstructed. Single-nucleotide variants (SNVs) analysis and temporal signal estimation can be optionally performed to interpret strain-level evolutionary relations. The MSA can be further refined using the multiple-reference mode of MetaClock (**Fig. 1b**), or be used in molecular clocking

analysis by BEAST. Dashed lines represent optional procedures. **b**, MSAs from different references can be produced through the pipeline described in (**Fig. 1a**) and then combined using our newly devised algorithm (**Supplemental Fig. 1**) in order to ameliorate the reliability of phylogenetic and clocking analyses.

3.4.2 Precise strain-resolved phylogeny reconstruction from metagenomes and improved phylogeny accuracy when using multiple reference genomes

We first compared the method performance with a commonly used approach in reconstructing a strain level phylogeny, i.e. generating assembled contigs using metagenomic assembly followed by strict quality control, and subsequently building a core gene alignment to reconstruct a phylogenetic tree. Focusing on *Ruminococcus bromii*, we used semisynthetic metagenomic datasets (**Fig. 2a**; see **Methods** and **Data availability**) and obtained 89 high-quality (completeness ≥90% and contamination ≤5%) *R. bromii* MAGs out of a total of 946 medium-quality MAGs (completeness ≥50% and contamination ≤5%), reconstructed from semisynthetic metagenomic samples containing synthetic *R. bromii* reads with coverage greater than 5X. We then compared the phylogenies reconstructed with MetaClock using the recovered *R. bromii* MAGs or the raw semisynthetic metagenomes directly against phylogenies reconstructed with the aforementioned classical approach, and obtained similar topologies (**Supplemental Fig. 2**). The congruence was further confirmed by an overall correlation of 0.9783 (0.9716 when using as input contigs; Pearson's correlation) between pairwise branch length distances of the MetaClock-phylogenies and the ones obtained with the classic approach (**Fig. 2b**). Moreover, our framework also shows a credible consistency when using different types of DNA sequences (short-sequence reads and assembled contigs), displaying an overall correlation of 0.9621 between pairwise branch length distances from the inferred phylogenies (**Fig. 2b**). MetaClock allows for extracting microbial genomic information from a complex metagenomic environment for genetically close strains of a given species and to reconstruct a genome phylogeny comparable with a classical more laborious method.

We further validated the method's improvement when combining multiple MSAs from different references using synthetic datasets generated from a known phylogeny which is regarded as the reference tree (see **Methods** and **Data availability**). As a result, an increasing number of references showed a clear decrease in the Robinson-Foulds distance weighted on branch lengths between trees inferred by MetaClock and the given phylogeny (**Fig. 2c** and **2d**). The inferred phylogeny from combining nine reference MSAs was, on average, 17.54% closer to the given tree compared to the one

relying on only a single reference when validating on the synthetic dataset containing only short-sequence reads (**Fig. 2c**). The performance even improved when using the dataset containing both short-sequence reads and contigs-like sequences, showing a 38.85% drop (on average) in the deviation of the inferred trees (on nine references) to the given tree (**Fig. 2e**). Likewise, the tree variability followed the same trend with an increasing number of reference MSAs (**Fig. 2e** and **2f**). The newly-devised algorithm of combining multiple MSAs in MetaClock allows for an improved accuracy and a decreased variability in strain-resolved phylogeny reconstruction, and it can ameliorate the reliability of phylogenetic analysis based on only one single reference genome which is to date widely adopted in ancient metagenomic studies (Peltzer et al. 2016; Sabin et al. 2020b; Schuenemann et al. 2018; Spyrou et al. 2018; Key et al. 2020).

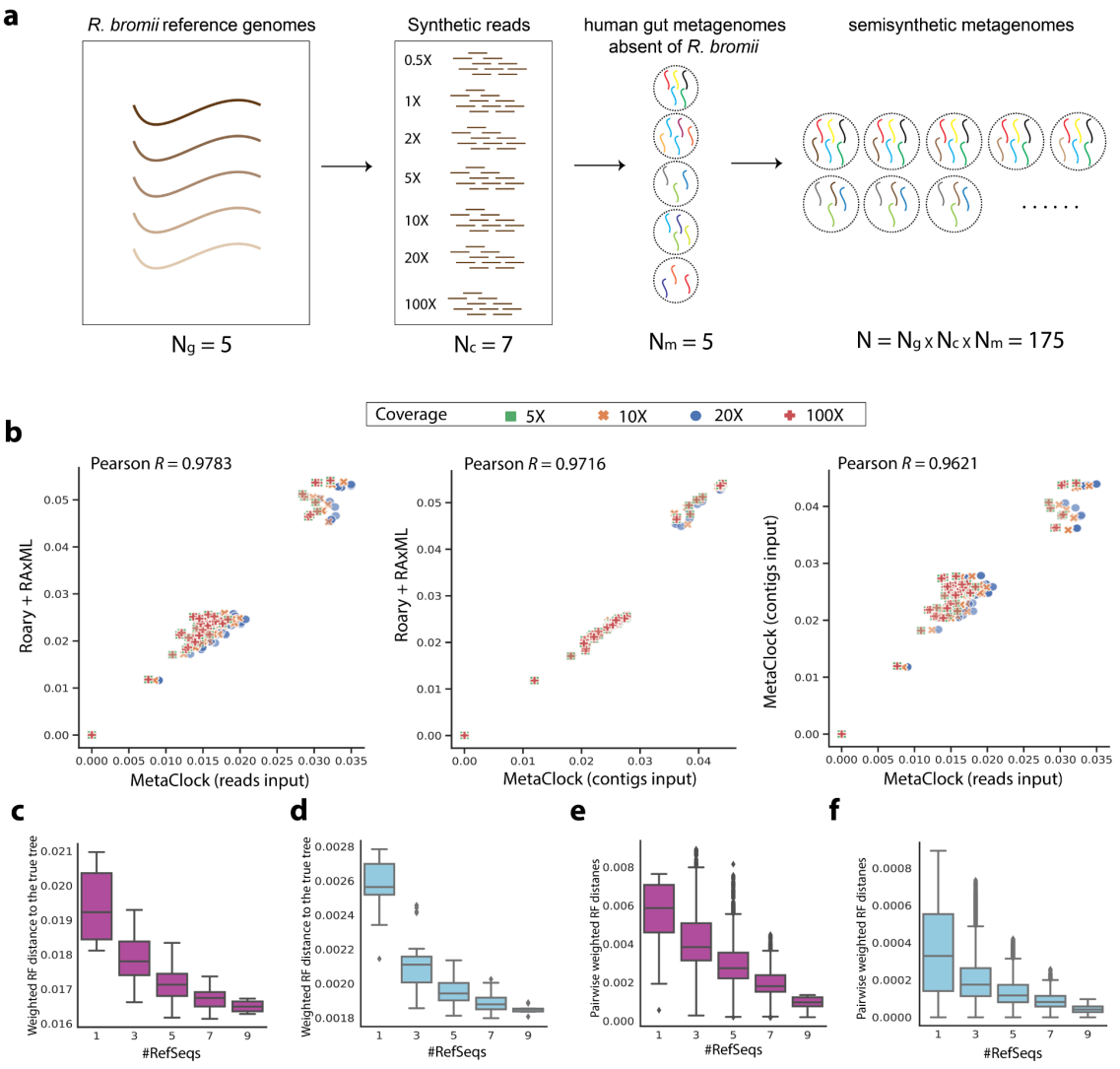


Figure 2. Method validation on (semi)synthetic data. a, The procedure of generating semisynthetic metagenomes with empirical sequencing noise and

confounding non-target species (see **Methods**). N_g indicates the number of reference genomes used in generating synthetic reads; N_c gives the number of increasing coverage depths chosen in generating synthetic reads; N_m is the number of real metagenomes lacking *Ruminococcus bromii* and inhabited by other species; N is the total number of semisynthetic metagenomes produced in the end of the procedure. **b**, Correlation analysis of phylogenetic distances in the resulting trees reconstructed manually using Roary (Page et al. 2015) followed by RAxML (Stamatakis 2014) and automatically by MetaClock. **c** and **d**, Deviation of inferred trees on multiple references to the given phylogeny, measured with Robinson-Foulds distance weighted on branch lengths. The distance was measured between the given phylogeny and each inferred tree from MetaClock using different numbers of reference genomes in combining MSAs. **e** and **f**, Tree variability when using different numbers of reference genomes, measured with pairwise Robinson-Foulds distances weighted on branch length. The pairwise distances were measured between trees built on the same number of references. Magenta-colored plot is based on the dataset comprising only short-sequence reads, and blue-colored plot is based on including both short-sequence reads and contigs-like sequences.

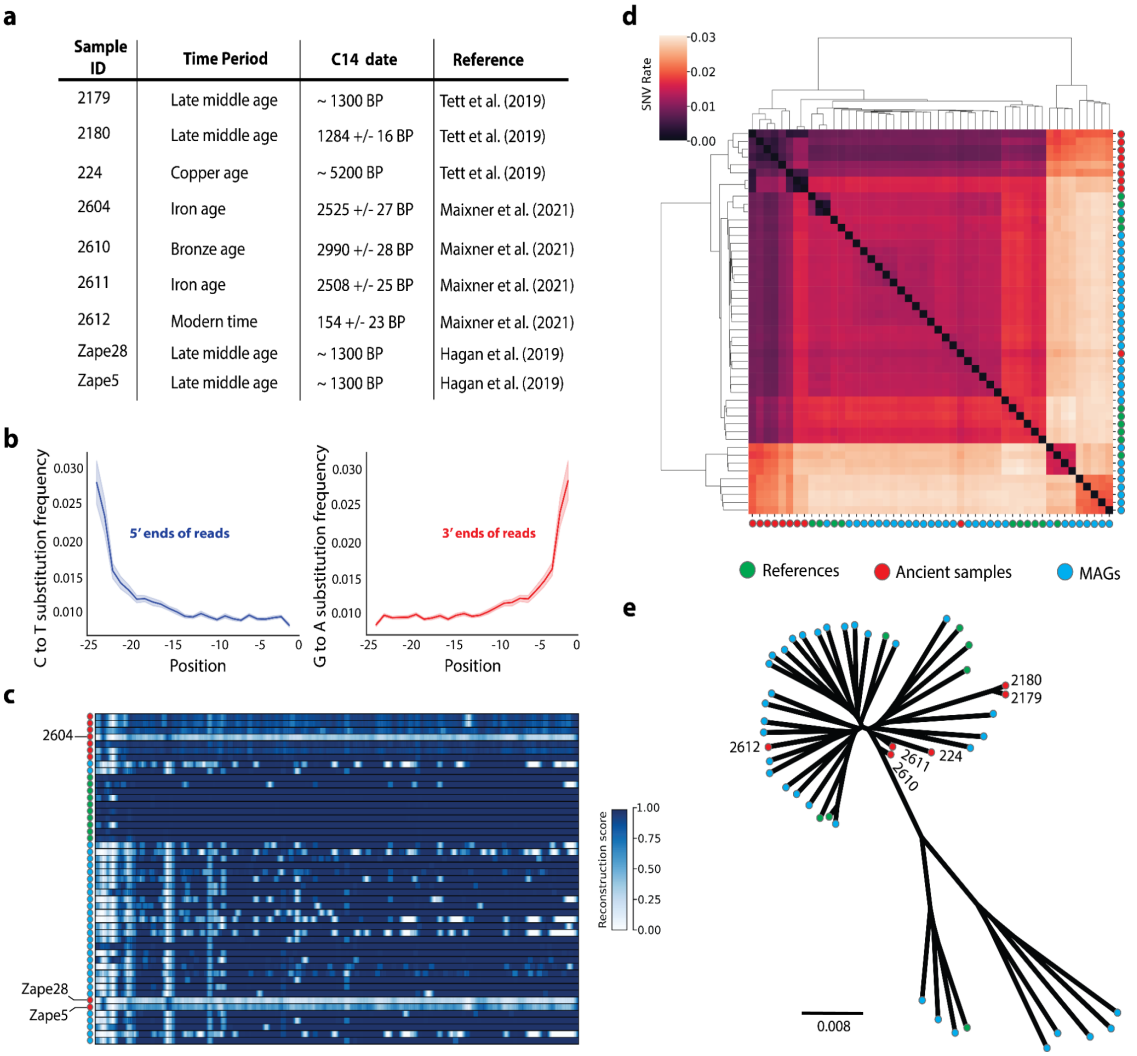
3.4.3 MetaClock reconstructed a high-resolution evolutionary relationship of ancient and modern *Ruminococcus bromii* strains

R. bromii is prevalent in the human gut in present-day populations (Truong et al. 2017) and ancient microbial samples (**Supplemental Fig. 4**). It is a keystone species and known for degrading complex carbohydrates in the human colon (Ze et al. 2012), and has been reported to be an important factor in human health (Petrov et al. 2017; Lv et al. 2016; Ze et al. 2015). However, little is known about the evolutionary history of this microbe, a fact that makes it an ideal candidate to demonstrate MetaClock's utility in reconstructing its evolution at the strain level. By applying MetaClock on *R. bromii* from ancient metagenomic samples of gut contents and a large amount of contemporary MAGs, we provide an automated comparative analysis and the first highly-resolved evolutionary relationship of ancient and modern *R. bromii* strains.

We screened a total of 16 ancient gut metagenomes for the presence of *R. bromii* using MetaPhlAn2 (Truong et al. 2015) and selected nine metagenomes with positive signals for subsequent MetaClock analysis (**Supplemental Fig. 4** and **Fig. 3a**, see **Data availability**). We reconstructed the most complete phylogeny for *R. bromii* including 3,976 MAGs (**Supplemental Fig. 6** and see **Methods**) from which we selected 40 representative genomes (30 high-quality MAGs and 10 reference genomes, see **Data availability**) that were used with MetaClock for building the phylogeny. As a result, MetaClock provides a merged MSA from multiple reference MSAs and the respective visualization for reconstruction information (**Fig. 3c**).

Optionally, endogenous origin of ancient metagenomic reads used in building consensus genome sequences can be authenticated using aDNA damage patterns (**Fig. 3b**). Pairwise SNV rates between strains can be calculated and visualized with a clustered heatmap to guide the interpretation of the genetic closeness information (**Fig. 3d**). Finally, a strain-level phylogeny can be inferred based on the merged MSA which can be manually curated with utility scripts provided in MetaClock (**Fig. 3e** and see **Methods**). All these optional analyses can also be performed automatically by selecting their respective parameters. Despite the small number of ancient metagenomic samples, with MetaClock we were able to study the evolutionary relationship of *R. bromii* strains from both ancient and present-day populations. From these results, it appears that ancient strains are genetically distinct with respect to their modern counterparts based on SNVs, except for sample 2612, which is the most recent ancient specimen dating back to ~154 years (**Fig. 3d**). The resulting strain-level phylogeny provides a finer structure for the phylogenetic placement of ancient and contemporary strains (**Fig. 3e**).

Importantly, the results were generated based on verified aDNA evidenced by aDNA damage patterns (**Fig. 3b**) and noises from for example aDNA damage effect and single-reference mapping bias can be mitigated via our functions for damaged site trimming and MSAs combing (see **Methods**). This analysis highlights the use of MetaClock for efficiently reconstructing strain-level evolutionary relationships for a given species leveraging ancient metagenomic samples and contemporary MAGs.



Sample ID	Time Period	C14 date	Reference
2179	Late middle age	~ 1300 BP	Tett et al. (2019)
2180	Late middle age	1284 +/- 16 BP	Tett et al. (2019)
224	Copper age	~ 5200 BP	Tett et al. (2019)
2604	Iron age	2525 +/- 27 BP	Maixner et al. (2021)
2610	Bronze age	2990 +/- 28 BP	Maixner et al. (2021)
2611	Iron age	2508 +/- 25 BP	Maixner et al. (2021)
2612	Modern time	154 +/- 23 BP	Maixner et al. (2021)
Zape28	Late middle age	~ 1300 BP	Hagan et al. (2019)
Zape5	Late middle age	~ 1300 BP	Hagan et al. (2019)

Figure 3. Reconstructing strain-level evolutionary relationships for *Ruminococcus bromii* using rich contemporary MAGs and ancient gut microbiome samples in MetaClock. a, Ancient microbiome samples with their historical time period, age determined by C14 dating (those dated by archaeological methods rather than carbon dating are prefixed with tilde signs), and reference. **b**, Damage patterns automatically estimated with mapDamage2.0 (Jónsson et al. 2013) (as part of the framework) of metagenomic reads used in building consensus sequences from ancient samples. The blue line indicates the average frequency of C to T substitutions at the 5' ends of reads from mapping against multiple references (N=10), and the red line indicates the average frequency of G to A substitutions at the 3' ends of reads from mapping against multiple references (N=10). The shaded areas illustrate standard deviation. **c**, Visualization of the reconstructed genome MSA. Reconstruction score indicates non-missing information in genome regions. Samples showing a poor genome-wide reconstruction score are highlighted with sample names. **d**, Pairwise SNV rates of reconstructed *R. bromii* genomes. **e**, Automated phylogenetic placement by ML phylogenetic approach (Stamatakis 2014) (implemented as part of the framework) of ancient samples, reference genomes and MAGs after excluding poorly-reconstructed samples (2604, Zape28, and Zape5,

highlighted in **Fig. 3c**).

3.4.4 MetaClock reconstructs evolutionary timelines for typical gut microbial members using molecular dating analyses

Measuring the evolution of human-associated microbes exploiting carbon dated ancient metagenomic samples has gathered interest because it provides evolutionary insights into our ancestors' microbiome (Rascovan et al. 2019; Spyrou et al. 2018; Weyrich et al. 2017; Key et al. 2020) and how our microbiome has diverged in the past (Tett et al. 2019; Wibowo et al. 2021). Here, we selected six common human gut species (*Ruminococcus bromii*, *Prevotella copri* Clade A, *Clostridium ventriculi*, *Catenibacterium mitsuokai*, *Eubacterium rectale*, and *Intestinibacter bartlettii*) because of the abundant species-level marker genes detected in ancient samples and rich contemporary reconstructed genomes (N = 11,378) (**Supplemental Fig. 4** and **Supplemental Table 2**). We next calibrated their genome phylogenies reconstructed using MetaClock with intensive molecular clock models using carbon dated ancient samples (see **Methods** and **Data availability**). The evolutionary modelling provided fresh insights into gut microbiome evolution from three complementary aspects. Firstly, we observed different demographic patterns of the target species (**Supplemental Table 3**). Four species (*R. bromii*, *P. copri* Clade A, *C. ventriculi*, and *C. mitsuokai*) were estimated to have undergone an exponential growth population, while the *E. rectale* population has been showing a constant growth rate in the past. The demographic pattern of *I. bartlettii* was instead inferred to fit a coalescent Bayesian skyline model characterized by a higher number of parameters, suggesting a piecewise-constant change in population size through time (A. J. Drummond et al. 2005) (**Supplemental Table 3** and **Supplemental Fig. 7**). Secondly, the mutation rates vary between species, exemplified by the heterogeneous ranges of median rate estimates for strains of each species (**Fig. 4** and **Supplemental Table 3**). Likewise, except for *E. rectale* whose strains were estimated to evolve at the same rate, at least one order of magnitude difference in evolutionary rate was observed separately in strains of other five species (**Fig. 4** and **Supplemental Table 3**). Of note, our rate estimates of target species are in agreement with genome-scale evolutionary rates reported for other bacteria, and are slower than those of pathogenic members such as *Acinetobacter baumannii* which was reported to evolve at $>10^{-5}$ mutation per site per year (Sebastian Duchêne et al. 2016). Generally evolutionary rates have been estimated for pathogenic microbes (Rascovan et al. 2019; Spyrou et al. 2018; Key et al. 2020; Sabin et al. 2020b). The analysis of these six microbes demonstrates

MetaClock offers the ability to easily study the evolutionary rates of human microbes of broader interest.

The evolutionary timeline of human microbiome species can represent an important proxy for the long-established association with the host's evolutionary history. Here, the mean age estimates of whole *R. bromii* species can trace back as early as around 750 ka (**Fig. 4a**) when hominins started developing sophisticated human behaviors such as stone knapping, tool use and floral and faunal consumption (Alperson-Afil et al. 2009). The lineages of *P. copri* subspecies Clade A were estimated to begin radiating ~2 million years ago (**Fig. 4b**), with a 95% highest posterior density (HPD) interval 70.42 ka ~ 4.88 mya (**Supplemental Table 3**) which is in line with a previous observation in terms of HPD overlap (Tett et al. 2019). Intriguingly, the radiation of *C. ventriculi*, *C. mitsuokai*, and *I. bartlettii* were estimated to have occurred in the time window when *Homo sapiens* began to establish facial, mandibular and dental features which aligns to modern humans at around 315 thousand years before present (Hublin et al. 2018), with median estimates ranging from ~250 to ~450 thousand years ago (**Fig. 4c**, **4d**, and **4f**). The evolutionary timeline of *E. rectale* based on 44 modern genomes representing a phylogenetic diversity of ~8,000 contemporary strains and three carbon dated ancient genomes as calibration indicates that the initial diversification occurred ~120 ka (95% highest posterior density interval: 44-217ka) (**Fig. 4e** and **Supplemental Table 3**; see **Methods**), coincides with the period when modern humans migrated from Africa 90-194 ka before present (Hershkovitz et al. 2018). The evolutionary modelling analysis within MetaClock provided the first time-resolved evolutionary investigation based on multiple species into the human gut microbiota, which can serve the best proxy for studying further the long-term relationship of hosted members and our human ancestors.

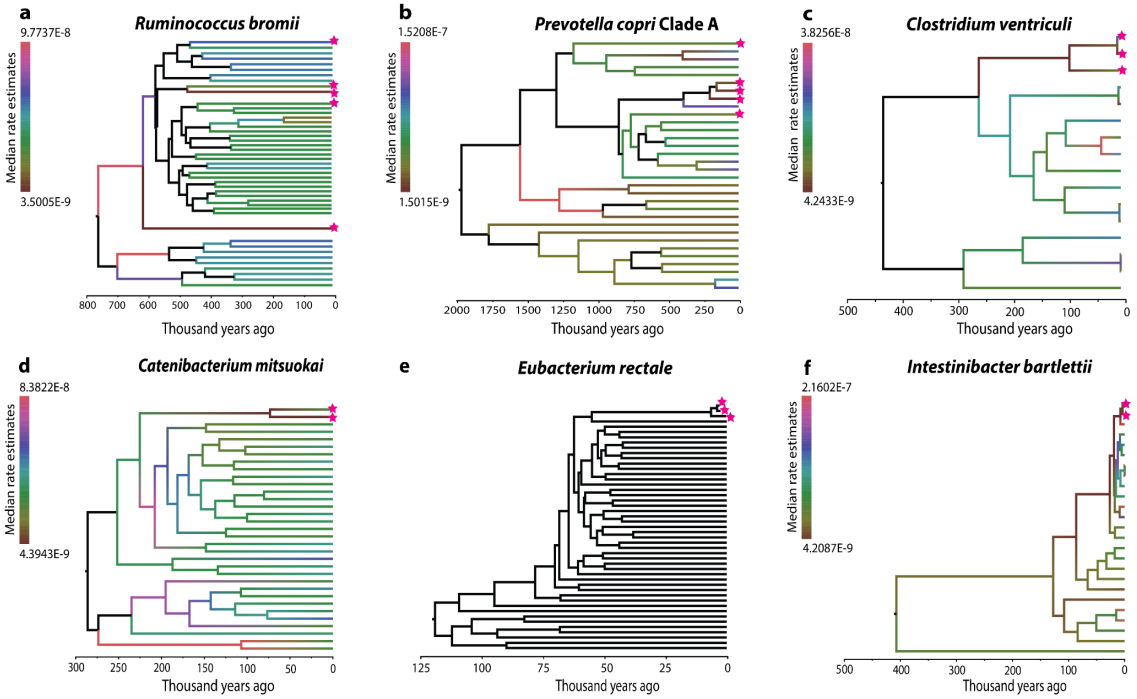


Figure 4. Time-resolved phylogenetic trees of six typical intestinal species inferred from MetaClock using molecular clocking techniques. Ancient samples used as tip calibrations were highlighted with magenta stars. Based on an intensive comparison of clocking models, a relaxed clock was used to best explain the evolutionary history for **a**, *Ruminococcus Bromii*, **b,** *Prevotella copri* Clade A, **c**, *Clostridium ventriculi*, **d,** *Catenibacterium mitsuokai*, and **f,** *Intestinibacter bartlettii*. A strict clock model was used for **e,** *Eubacterium rectale.* The inferred rate variants among strains are indicated by a gradient color scale for branches in each tree, except for *E. rectale* whose strains were estimated to evolve at a constant rate).

3.5 Methods

MetaClock provides an automated computational environment for reconstructing the evolutionary history of human microbiome species from contemporary and ancient populations. This method involves constructing whole-genome MSA for strains of a given species directly from the metagenomic reads, MAGs, SAGs or isolated genomes. Simultaneously, it performs quality controls, strain-level SNV analysis, automated aDNA authentication, sophisticated MSA curation, genome phylogeny inference (based on maximum-likelihood principle), estimation of temporal signal (**Fig. 1a**). Moverover, having acknowledged the possible aDNA damage effect and biases from mapping reads against only one single reference genome, we developed two novel features to mitigate these effects: (i) *a priori* trimming on the end of reads where post-mortem damage was frequently observed, which can be tuned with a specific

parameter, and (ii) combining multiple MSAs from different reference genomes (**Fig. 1b** and **Supplemental Fig. 1**). Evolutionary timelines and long-term evolutionary rates for microbial species can then be estimated within our framework by leveraging molecular clocking on both contemporary metagenomic datasets and carbon dated ancient metagenomic samples (see **Software availability**).

3.5.1 The MetaClock algorithm

A MetaClock analysis starts with the python script *metaclock_mac.py* (Metagenome Alignment Constructor) to align metagenomic reads (no assembly provided as input) in each sample against a single reference genome using Bowtie2 (B. Langmead and Salzberg 2013) (**Fig. 1a**). A consensus genome sequence is then reconstructed for each sample that minimizes all the following criteria: (i) alignment quality, (ii) length of aligned reads, (iii) site coverage, (iv) identity of aligned reads, and (v) SNP edit distance. All the criteria for quality control can be tuned with specific parameters. Sites are filled with gaps if any of the criteria above are not met. For sequencing reads from ancient metagenomic data, in order to rule out potential effect of aDNA damage on downstream phylogenetic and clocking analysis, the read ends where post-mortem cytosine deamination frequently occurs can be trimmed to a user defined trimming distance. Optionally, if assembled contigs are provided they can be aligned against the same reference using blastn (Altschul et al. 1990) (**Fig. 1a**). Contig-to-reference alignment hits are selected based on alignment length and percentage identity which are tunable parameters, with lowest e-value and highest bit-score set as default for selecting best unique hits. Afterwards, either consensus sequences from reads or genomic regions where substantial sequence similarity is shared by contigs and the reference are integrated into a whole-genome MSA based on the same reference genome as backbone. Thus, the resulting MSA and reference genome have the same length, with missing information being filled with gaps. Additionally, aDNA authentication can be automatically performed prior to trimming on read ends by estimating damage patterns by the integration of mapDamage2 (Jónsson et al. 2013), and pairwise SNV rates (the number of SNVs divided by the number of sites, when only considering sites without gaps) can be estimated between strains as well.

A handful of post-processing steps are available for assessing and tailoring raw MSA in order to maintain sufficient phylogenetic signals for subsequent fine-grained analyses. Specifically, the overall completeness of MSA is first assessed, using *metaclock_visualizer.py* which estimates genome-wide gap ratio in a sliding-window

manner with the user-defined window size and step size. The resulting visualization serves as an intuitive guide for further tailoring steps. Once confidence is obtained for MSA completeness, a tailoring procedure can be executed using *metaclock_tailor.py* to precisely remove missing information in the MSA. The tailoring procedure can be performed by two modes: *metaclock_tailoring.py basic* and *metaclock_tailoring.py landscape*. The *basic* mode allows users to remove missing information by specifying target samples to be retained and a maximum threshold for missing information in a column, or by just simply performing automated trimming with TrimAl (Capella-Gutiérrez, Silla-Martínez, and Gabaldón 2009) embedded as part of the script. On the other hand, the *landscape* mode provides a more detailed investigation into MSA based on genomic landscape. It partitions the whole MSA into many divided MSAs based on the reference genome annotation profile provided by the user. If no annotations are provided, MetaClock will automatically annotate the reference genome using Prokka (Seemann 2014). For each partitioned MSA, the length, missing information, SNV density, pairwise genetic distances, and biological features (tRNA, coding and noncoding sequences) can be estimated (see **Software availability**). Users can then select and concatenate the interesting partitioned MSAs into a single alignment concatenation based on the assessment of genomic landscape.

Next, the processed MSA is then used by MetaClock to reconstruct a ML genome phylogeny using RAxML (Stamatakis 2014) and to estimate temporal signal (Andrew Rambaut et al. 2016) if sample ages are provided. Afterwards, a parameterized clocking analysis can be performed on the MSA using external softwares like BEAST2(Bouckaert et al. 2014) or a further MSA refining can be achieved using multiple references before clocking analysis as detailed below.

3.5.2 Refining MSA combining multiple references for reliable phylogenetic analysis

Bertels et al. have shown that biases can be considerably reduced by combining MSAs from mappings to multiple reference sequences (Bertels et al. 2014). We were inspired to exploit this concept and implemented it within our framework. Alignment columns from all MSAs built on different references are merged into a final set of alignment columns by performing the following steps (**Fig. 1b** and **Supplemental Fig. 1**). Firstly, substantial similar regions between references are searched using blastn (Altschul et al. 1990) and resulting hits are selected, restricting a flexible parameter of alignment length and percentage identity by users. Secondly, similar sites (the sites within similar

regions) are all indexed for extracting corresponding columns from reference MSAs. To have a flexible control over the length of final combined alignment, we thirdly acquired the concept of ‘coreness’ to select similar sites for selecting columns from reference MSAs. Users can decide the number of reference genomes a site must be present in order to extract alignment columns from reference MSAs. Lastly, selected columns from different reference MSAs are estimated to generate a consensus column based on a majority ratio given by the user, and all consensus columns are then concatenated into a final merged MSA.

3.5.3 Method validation and evaluation

To evaluate the method's performance, MetaClock was applied to semisynthetic and synthetic datasets. The validation was first performed on 175 semisynthetic metagenomic datasets (**Supplemental Table 5** and see **Data Availability**) which were made by mixing synthetic reads of *R. bromii* strains with real metagenomic samples (in which *R. bromii* was screened for absence using MetaPhlAn2 (Truong et al. 2015) **Fig. 2a**). Synthetic reads were generated using GemSIM (McElroy, Luciani, and Thomas 2012), with the provided error model for Illumina paired-end reads (101bp), from five reference genomes at coverages from 0.5X to 100X (**Fig. 2a**). The semisynthetic metagenomes were subjected to *de novo* metagenome assembly using metaSPAdes (Nurk et al. 2017), contigs binning using MetaBAT2 (Kang et al. 2019), quality assessment using CheckM (Parks et al. 2015), and taxonomic assignment using PhyloPhlAn 3.0 (the *phylophlan_metagenomic* script with “-d SGB.Sep20”) (Asnicar et al. 2020). *R. bromii* MAGs above high quality (completeness ≥90%; contamination ≤5%) and the corresponding semisynthetic metagenomes, along with other eight *R. bromii* genomes (which were not used for producing synthetic reads; accession numbers: GCF_002834225, GCF_002834235, GCF_003464295, GCF_003466165, GCF_003466205, GCF_003466225, GCF_003466725, GCF_003469265), were then used for comparing MetaClock with a classic method based on core gene alignment.

Firstly, we tested MetaClock on semisynthetic metagenomes containing *R. bromii* reads at coverage greater than 5X (because assembly failed for samples with *R. bromii* coverage lower than 5X) and the eight reference genomes. Secondly, we tested MetaClock on the corresponding *R. bromii* MAGs and the same eight reference genomes. Default settings were used. Thirdly, coding sequences of same MAGs and reference genomes were predicted using Prokka (Seemann 2014) first and the core gene alignment was then reconstructed using PRANK (Löytynoja 2014) as part of the

Roary pipeline (Page et al. 2015) with default parameters. The core gene alignment was then used to infer a ML phylogeny using RAxML (Stamatakis 2014) under a GTR model of substitution with 4 gamma categories and 100 bootstrap pseudoreplicates. We visually inspected the consistency of phylogenetic structures generated by those three manners using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/), and a further quantification was provided by a Pearson correlation analysis of pairwise branch length distances from phylogenies built by different approaches.

Next, the method's accuracy and variability was evaluated on synthetic datasets generated by a given phylogeny when using multiple references. A known phylogeny (see **Data availability**) was first given to simulate 10 nucleotide sequences at length of 2Mbp using Seq-Gen (v1.3.4) (A. Rambaut and Grassly 1997) under GTR model with gamma distribution of four categories. Synthetic data were generated by sampling reads from each simulated nucleotide sequence with coverage 5X using GemSIM (McElroy, Luciani, and Thomas 2012) with an error model for paired-end Illumina reads (101bp). We then used *metaclock_mac.py* with default configuration settings to reconstruct 10 MSAs using each of the simulated sequences as a reconstruction reference and the synthetic reads as input from the remaining nine genomes, separately. (**Supplemental Fig. 3a**). MSAs from different references were combined from two to nine references using *metaclock_combiner.py* with default parameters (**Supplemental Table 1**), and *metaclock_tailor.py basic* was used to clean the combined MSA on which a ML genome phylogeny was inferred simultaneously. The accuracy was determined by the deviation of inferred phylogeny of combined MSAs to the known phylogeny measuring branch length weighted Robinson-Foulds Matrix (Robinson and Foulds 1979). Likewise, we estimated the pairwise tree distances, using the same matrix, between trees inferred combining the same number of references in order to evaluate phylogeny variability with the change of number of reference genomes used.

In addition, to evaluate the performance in a more realistic scenario - the input comprises short-sequence reads which can hardly be assembled as is common in ancient metagenomic studies and assembled contigs that are relatively easy to obtain from contemporary samples - we repeated the evaluation described above on a more complex synthetic datasets. We first simulated 20 sequences with length of 2Mbp based on a known phylogeny (see **Data availability**) comprising 20 taxa. Ten simulated sequences were used to generate synthetic reads with 5X coverage and the remaining 10 sequences were used directly as contigs-like sequences. MetaClock was

applied as described above using this new dataset composed of both synthetic reads and continuous sequences (**Supplemental Fig. 3b**), and the accuracy and variability were validated in the same manner as described above.

3.5.4 Ancient samples from gut content

In this study, we considered 16 shotgun sequencing metagenomes from ancient human gut content to demonstrate the main usages of MetaClock computational framework. The analyzed metagenomes include the gut content of the Iceman and 11 ancient coprolite samples reported previously (Tett et al. 2019; Borry et al. 2020; Hagan et al. 2020), as well as four newly discovered ancient coprolite specimens from Hallstatt, Austria (see **Data Availability**). The details about these four paleofeces samples are reported in another tandem study (Maxiner et al. 2021).

3.5.5 Preprocessing ancient metagenomic samples

Sequencing reads of all ancient metagenomics samples used in this study were subjected to stringent quality checks using an integrated preprocessing pipeline (https://github.com/SegataLab/preprocessing). The preprocessing procedure includes adaptor removal, trimming position with sequencing quality <15, discarding reads with mean quality <15 and with length <30nt, and filtering out human DNA sequences by mapping reads to the *Homo sapiens* reference (build hg19) already implemented in the pipeline.

3.5.6 Species-level taxonomic abundance estimation for ancient metagenomic samples

To decide target species which should be enriched in ancient samples, all preprocessed ancient metagenome samples used in this study were estimated for taxonomic abundance at the species level using MetaPhlAn2 (Truong et al. 2015) with default parameters. Afterwards, the estimated species abundance was visualized using *hclust2.py* with parameters “--ftop 25 --f_dist_f braycurtis --s_dist_f braycurtis”.

3.5.7 Exploring phylogenetic diversity of selected species using a large body of contemporary MAGs

We first retrieved 33 reference genomes for six species enriched in ancient samples from the NCBI genome database available as of (**Supplemental Table 6**). Due to the underrepresentation of genetic diversity from reference genomes we also considered

MAGs available for the five species in public repositories. To this end, we downloaded all the MAGs considered in six previous metagenomics assembly investigations (Pasolli et al. 2019; Almeida et al. 2019; Forster et al. 2019; Asnicar et al. 2021; Nayfach et al. 2019) and further applied the whole metagenomic assembly pipeline on additional metagenomes from other eight studies (Tett et al. 2019; Shao et al. 2019; Rosa et al. 2018; Hansen et al. 2018; Mehta et al. 2018; Visconti et al. 2019; Wampach et al. 2018; Jie et al. 2017; Thomas et al. 2019). The metagenomic assembly pipeline is based on the procedure described in a previous study (Pasolli et al. 2019). Specifically, metagenomic reads were assembled into contigs using MEGAHIT (D. Li et al. 2015), and then were aligned back against the assembled contigs (>1000nt) using Bowtie2 (version 2.2.3; "--very-sensitive-local") (B. Langmead and Salzberg 2013). The resulting BAM files were sorted and used for contig binning using MetaBAT2 (Kang et al. 2019) with default settings to generate putative genomes. Quality assessment was performed on the putative genomes using CheckM with lineage-specific workflow using default parameters (Parks et al. 2015). We only retained MAGs of medium- and high-quality. Finally, we assigned a taxonomic label to MAGs downloaded or reconstructed in this study using PhyloPhlAn 3.0 (the *phylophlan_metagenomic* script with parameter "-d SGB.Sep20") (Asnicar et al. 2020), and selected those with taxonomy assigned to our target species for subsequent analysis.

Finally, we reconstructed a large-scale strain-resolved phylogeny for each target species using corresponding reference genomes and MAGs using PhyloPhlAn3.0 (Asnicar et al. 2020) with parameters "--diversity low --fast" in order to capture the maximum genetic diversity of each selected species in the contemporary datasets.

3.5.8 Reconstructing a high-resolution evolutionary relationship of *R. bromii* genomes using MetaClock

We selected a common human gut bacterial species, *R. bromii*, as an example to reconstruct its high-resolution evolutionary relationships using ancient paleofeces samples and large-scale present-day metagenomic datasets. We firstly selected as ancient input data nine ancient metagenomic samples containing abundant genomic information from *R. bromii* (**Supplemental Fig. 4**), and selected as contemporary input data 30 high quality MAGs, along with 10 reference genomes (accession numbers: GCF_002834165, GCF_002834225, GCF_002834235, GCF_003458325, GCF_003459405, GCF_003463575, GCF_003464295, GCF_003466165, GCF_003466725, GCF_003469265), representative of a large *R. bromii* phylogeny

comprising 3,669 genomes (**Supplemental Fig. 6** and see **Data availability**). Secondly, *metaclock_mac.py* script was used to reconstruct 10 genome-wide MSAs, one for each of the reference genomes listed above, with default configuration settings and parameters: "--authentication --SNV_rate". Here, *metaclock_mac.py* reconstructed genomic information directly from ancient metagenome samples and contemporary assembled genomes guided by 10 different reference genomes. In the meantime, ancient DNA authentication was executed automatically on aligned reads using mapDamage2 (Jónsson et al. 2013) embedded as part of the pipeline, and pairwise SNV rates between genomes were estimated and visualized simultaneously. Thirdly, individual MSAs were combined into one refined MSA using the *metaclock_combiner.py* script with default parameters, in order to minimize bias from mapping against a single reference genome.

The refined MSA was visualized using the *metaclock_visualizer.py* script, with parameters: "--window-size 10000 --step-size 2000", for assessing the genome-wide reconstruction quality. Using *metaclock_tailor.py basic,* poorly-reconstructed genomes were excluded first and then columns in the alignment containing >5% missing information were removed, and a ML phylogeny based on the curated alignment was simultaneously inferred with parameter: "--raxml_tree". To confirm the consistency we repeated the same procedure for the alignment in which the poorly-reconstructed genomes were retained (**Supplemental Fig. 5**).

3.5.9 Estimating time-resolved evolutionary history for multiple common human gut species with molecular clocking in this framework

We applied MetaClock on real metagenomic datasets, combining with molecular clock models, in order to estimate evolutionary timescales for six species. These six species were selected for their representativeness as typical human symbionts and presence in at least nine ancient samples (**Supplemental Fig. 4**). For the species with >100 modern genomes, we selected from 20 to 40 high-quality modern genomes including MAGs and reference genomes, in order to represent the diversity of the species (**Supplemental Fig. 6** and **Supplemental Table 2**). For *P. copri* Clade A, we directly used 27 high-quality representative MAGs from a previous study (Tett et al. 2019), and for *C. ventriculi*, we used all available modern publicly available genomes (N=13).

Having prepared the modern genomes to represent the genetic diversity for each target species, we applied *metaclock_mac.py* on the ancient samples and the selected

contemporary genome datasets for each target species using a default configuration setting in order to generate individual MSAs from different references (**Supplemental Table 2**). Afterwards, individual MSAs from each species were combined into one refined alignment using *metaclock_combiner.py* with default parameters, followed by a simple post-processing step using *metaclock_tailor.py basic* to remove poorly-reconstructed genomes (missing information ≥50%), reconstructed ancient genomes without direct carbon dated ages, and then trim gappy columns (missing information ≥5%). The combined MSA for each species was used to explore the evolutionary history with intensive molecular clocking modelling in the subsequent steps.

Here, BEAST2 (version 2.5.1) (Bouckaert et al. 2014) was used to infer the evolutionary timescale for each target species. Well-reconstructed ancient genomes with the corresponding origin samples being accurately radiocarbon dated were used as tip calibrations (**Supplemental Table 4**). For each target species, we analysed eight alternative evolutionary models (**Supplemental Table 3**) and selected the best-fit clock and demographic models to interpret its evolutionary history. Model comparison was performed calculating Log Bayes factors based on marginal likelihood estimation by path sampling (Baele et al. 2012), only considering simulations with adequate effective sample size (ESS) for majority parameters (**Supplemental Table 3**) (Andrew Rambaut et al. 2018). A consensus tree was summarized for each species using TreeAnnotator (Alexei J. Drummond and Rambaut 2007) with 10% burnin, 95% posterior probability limit, maximum clade credibility to determine target tree type. Each consensus tree estimated under the model suggested by the highest Log Bayes factor was visualized using FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/).

3.6 Discussion

Here, we have devised a new computational framework that allows strain-resolved phylogeny and evolutionary timeline to be estimated directly from metagenomic reads, MAGs, SAGs or isolated genomes from ancient and contemporary populations, and we employed it to investigate the time-measured evolutionary history for six common human gut species using a large number of present-day metagenomic samples and ancient metagenomes. This framework enables the reconstruction of evolutionary history using both contemporary and ancient metagenomic data for many common microbes without the laborious efforts such as screening presence of target species relying on PCR techniques (Schuenemann et al. 2018; Maixner et al. 2016), in-solution

capture to enrich DNA for a target species (Maixner et al. 2016; Sabin et al. 2020b; Spyrou et al. 2018), and can be utilised without *de novo* metagenome assembly which can be particularly challenging for ancient metagenomes (Quince et al. 2017; Seitz and Nieselt 2017). The approach utilizes reference genomes as representatives of species to efficiently extract full genomic information from metagenomes, and searches for sufficiently similar regions. It precisely integrates genomic information from different types of sources (short-sequence reads and assembled contigs) into a whole-genome MSA. By combining multiple MSAs from different references, MetaClock significantly improves the accuracy of phylogenetic inference compared to using only one single reference. Moreover, MetaClock provides other features (including post-processing utilities, visualization, aDNA damage analysis, SNV analysis, genome phylogeny inference and temporal signal estimation for further molecular clock analysis) to efficiently study ancient metagenomes in the context of vast amounts of contemporary metagenomic data available. Combining molecular clock techniques within MetaClock allows us to infer time-resolved evolutionary history of microbes from large-scale metagenomic datasets at an unprecedented scale. In this study, we reconstructed a detailed evolutionary history of common human gut members which serves as a valuable proxy to study the association of common microbes with our human ancestors.

In this work we detailed the evolutionary trajectories of non-pathogenic human gut microbiome species, which helps explain the diversity of commensals observed in contemporary datasets (Tett et al. 2019; Karcher et al. 2020). Strains from the same species (with the exception for *E. rectale*) were evolving at various rates and the rate of one strain varies through its evolutionary timeline, with lower evolutionary rates than those of pathogenic members (Sebastian Duchêne et al. 2016). Non-pathogenic members clearly could have appeared very early, even with the most recent one, *E. rectale* which started radiating before the first wave of human migration out of Africa and with the earliest one, *P. copri* Clade A whose diversification occurred way before the *Homo sapiens* began to develop physiological features which aligns to modern humans. These observations about the evolution of non-pathogenic species contrasts to the swifty and new emergence frequently observed in pathogenic agents, for example, the very recent radiation of *Yersinia pestis (Rascovan et al. 2019)* and *Mycobacterium tuberculosis (Sabin et al. 2020b)*.

More essentially, the capability to reconstruct the highly-resolved evolutionary history for microbes directly from contemporary and ancient metagenomics datasets is a

critical move to investigate, to a large scale, how symbionts have been associated with their human hosts in the course of long-term evolution. Ancient DNA advances, with particular strength in pathogens, have unearthed the buried historical past between agents and hosts (Key et al. 2020; Spyrou et al. 2018; Rascovan et al. 2019; Schuenemann et al. 2018). It will be equally important to explore the evolutionary past of microbial members common in contemporary populations even without evidence of relating to diseases. Similar approaches can also start to interrogate the association of the evolutionary history of the human gut microbiota and that of human, e.g. the relationship of *P. copri* complex with Westernisation (Tett et al. 2019) and the diversification timeline of *Methanobrevibacter smithii*, a key human symbiont, reconstructed from paleofeces (Marsha et al., 2021). This is of particular interest in the era when a large effort has been put into studying human evolution and migration using ancient DNA advances, with which a detailed microbial evolutionary history can complement understanding what happened between our ancestors and their symbionts in the deep past.

Reconstructing evolutionary relationships between strains will also increasingly scale up shotgun metagenomics analysis in ancient microbial communities. We have only demonstrated how to establish the linkage between ancient gut microbiota and rich present-day metagenomic data focusing on six typical intestinal species, and a greater picture of the microbial evolutionary past can be drawn applying our framework to other host-associated environments like the oral cavity, for instance. Such studies relying on ancient and contemporary metagenomes have been carried out within scope of specific species (Weyrich et al. 2017), but the capability to integrate vast amounts of modern metagenomics data with fast-growing ancient shotgun metagenomes for many species makes MetaClock framework an advanced approach for effortlessly interrogating the microbial evolutionary history.

Limitation of methodology

Our methodology has limitations. First, our method depends on reference genomes provided by users. Studies have shown that unknown microbes are common in ancient samples (Wibowo et al. 2021). Thus, our tool is still inadequate in exploring novel species from ancient samples. Second, while MetaClock is able to estimate temporal signals for the input dataset, a proper divergence estimate should be preceded by a model comparison in order to define the most fitting set of evolutionary models/priors (replacement models, clock models, demographic model) in BEAST2 (Bouckaert et al.

2014). We will continue developing the pipeline in the near future by automatizing the part of extensive clocking model comparison.

3.7 Author Contributions

K.D.H., N.S. and O.R.S. conceived of the computational framework. K.D.H with assistance of F.A. and M.Z. diviced core algorithms. K.D.H. with assistance of Z.W.L.L. built the software package. K.K., H.R. and F.M. collected archaeological samples. F.M., L.G. and M.S.S. processed archaeological samples and constructed sequencing libraries. K.D.H. with assistance of A.B.M and F.C. constructed the contemporary metagenomic datasets. K.D.H. with assistance of A.S. performed evolutionary analysis. K.D.H. wrote the manuscript with the assistance of N.S., A.T., O.R.S., and A.Z., and critical comments from the remaining co-authors.

3.8 Acknowledgements

We thank Nicola Zadra for the critical input of the evolutionary analysis. We also thank all the members of the Segata laboratory for the thorough discussions and support. We thank the team of the NGS Core Facility at CIBIO and Macrogen for metagenomic sequencing and the high-performance computing team at the University of Trento.

3.9 Declaration of interests

Authors declare that they have no competing interests.

3.10 Data availability

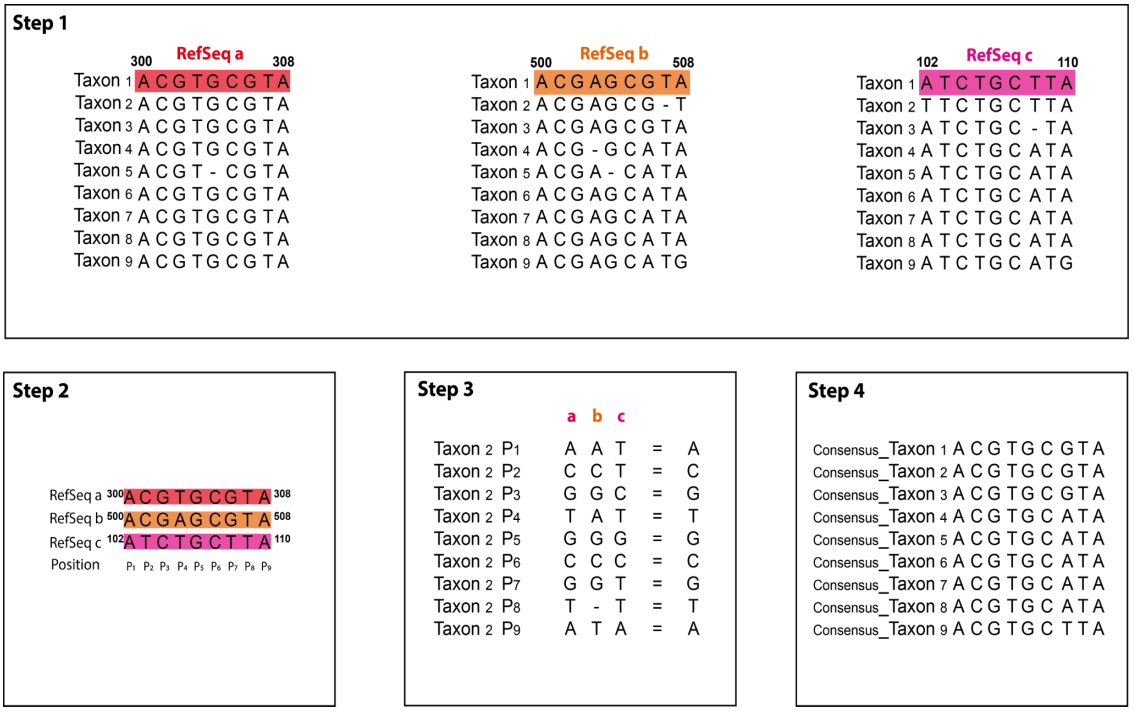
The semi-synthetic metagenomes, netwick files of two given phylogenies used for method validation, and all metagenomically assembled genomes (MAGs) are available at http://cmprod1.cibio.unitn.it/MetaClock. The downloaded ancient metagenomic samples are available in the European Nucleotide Archive (ERA) under the accession PRJEB33577, PRJEB35362, and in the NCBI database under accession: PRJEB31971. The newly sequenced ancient metagenomic samples are available in the European Nucleotide Archive (ERA) under accession PRJEB44507.

3.11 Software availability

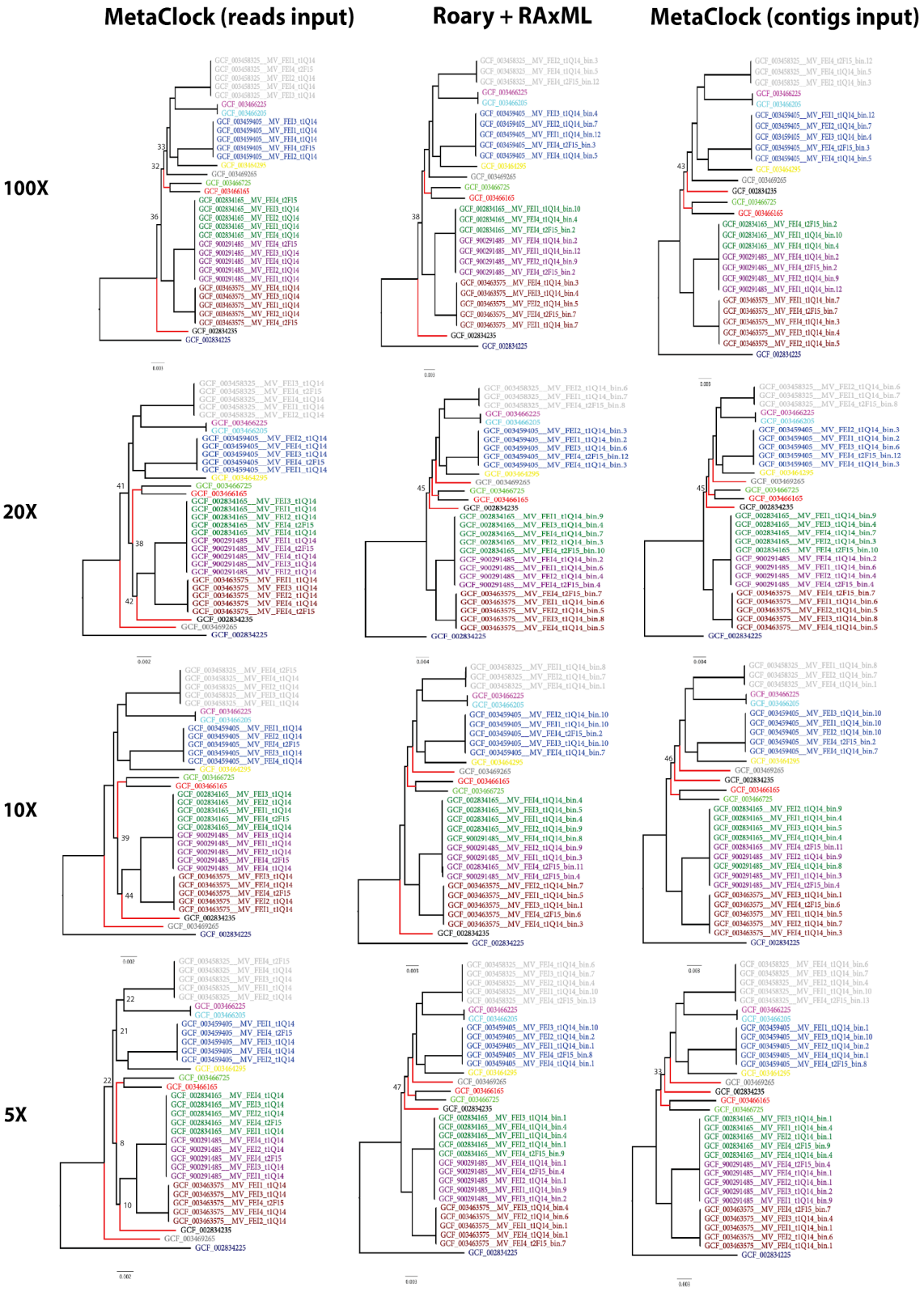
MetaClock (version 1.0) is available with source code, manual, tutorials, and a support user group at https://github.com/SegataLab/metaclock.

3.12 Supplementary materials

3.12.1 Supplemental figures

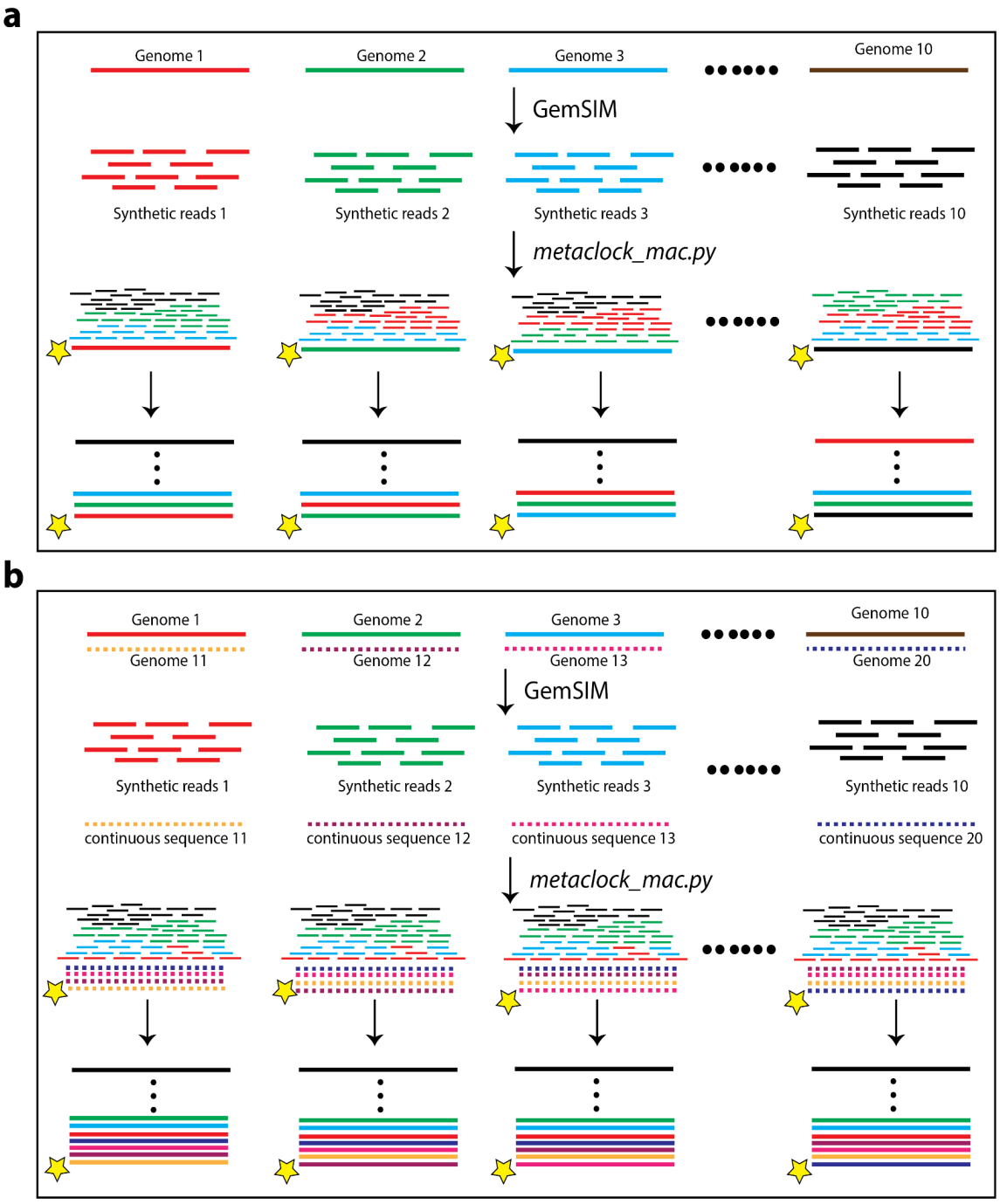


Supplemental Figure 1. Combining multiple MSAs from different references. (1) Generate genome MSAs using different reference sequences (Fig. 1a), separately. (2) Search for highly similar sequences between references using blastn(Altschul et al. 1990) with the hit length and percentage identity being set as tunable parameters, and document positions within similar regions for site selection in the sequences in the subsequent step. (3) Taking Taxon2 as an example taxon, select the site in the sequences from different reference alignments based on the documented position coordinates and calculate the consensus by applying nucleotide dominance ratio (majority rule by default). (4) Concatenate all taxon sequences calculated by merging sites from multiple references.



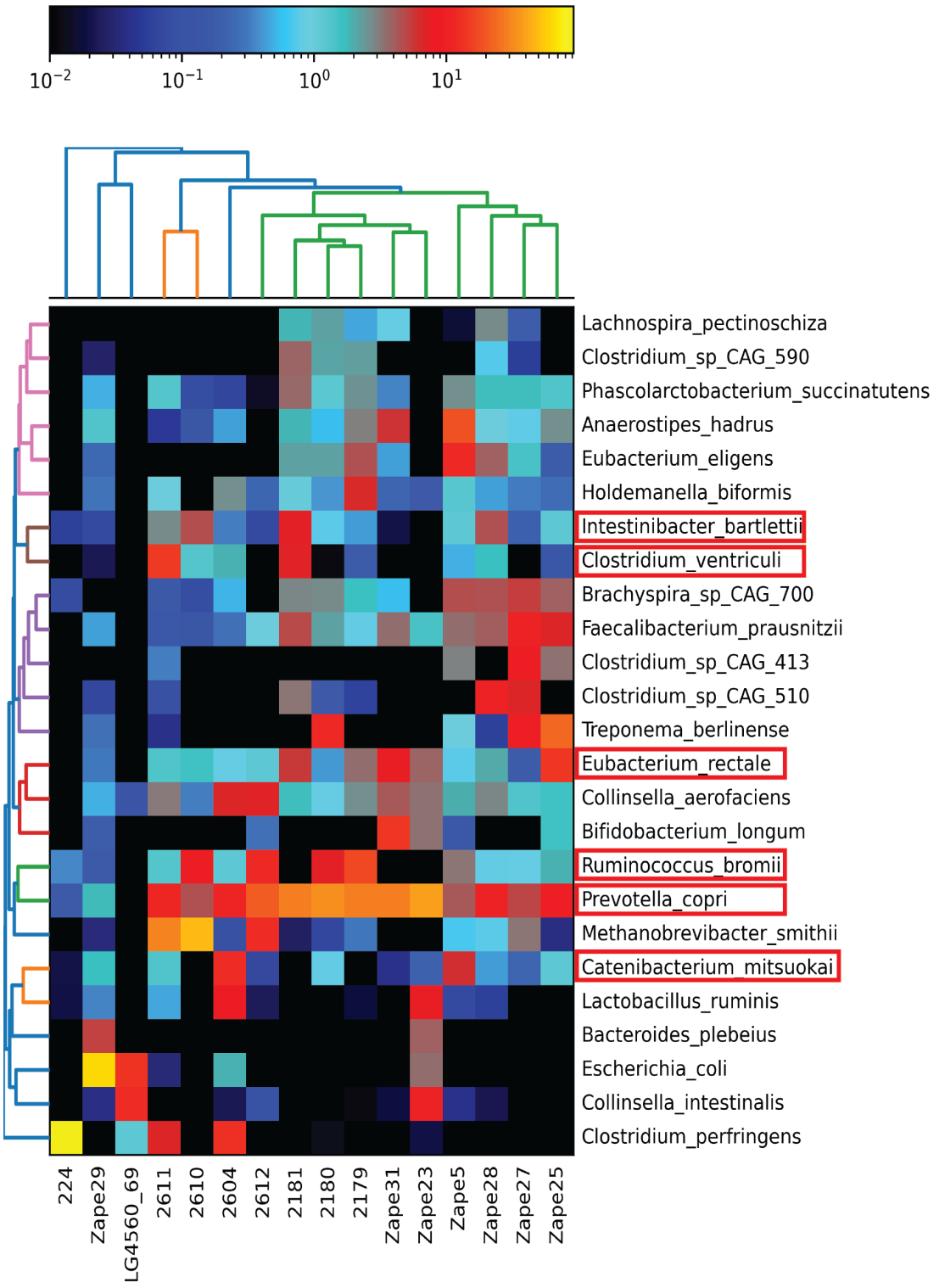
Supplemental Figure 2. The visual inspection with FigTree v1.4.4 (http://tree.bio.ed.ac.uk/software/figtree/) of phylogenetic trees reconstructed on semisynthetic metagenomic datasets (coverage of 5X, 10X, 20X, and 100X) and the corresponding *R. bromii* MAGs using MetaClock (reads input and contigs input) and those based on core gene alignment generated with Roary (Page et al. 2015)

followed by RAxML (Stamatakis 2014). Nodes with bootstrap values lower than 75 were highlighted and inconsistent branches were marked as red.

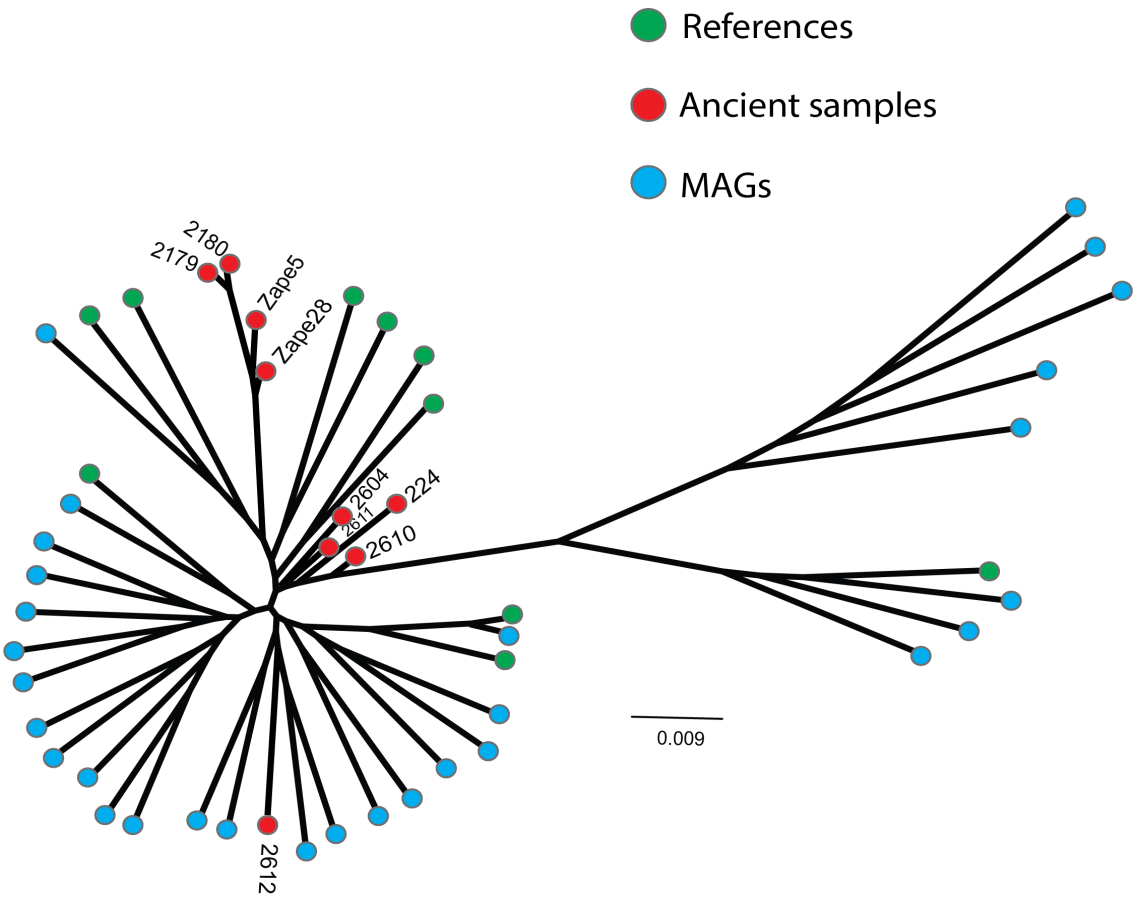


Supplemental Figure 3. Schematic illustration of reconstructing individual genome alignment in the MetaClock framework using synthetic datasets. **a**, Genome sequences were generated by a given phylogeny containing 10 taxa using Seq-Gen(v1.3.4) (A. Rambaut and Grassly 1997) and the corresponding synthetic reads were generated using GemSIM (McElroy, Luciani, and Thomas 2012). *metaclock_mac.py* was applied on the synthetic data to produce 10 MSAs, each

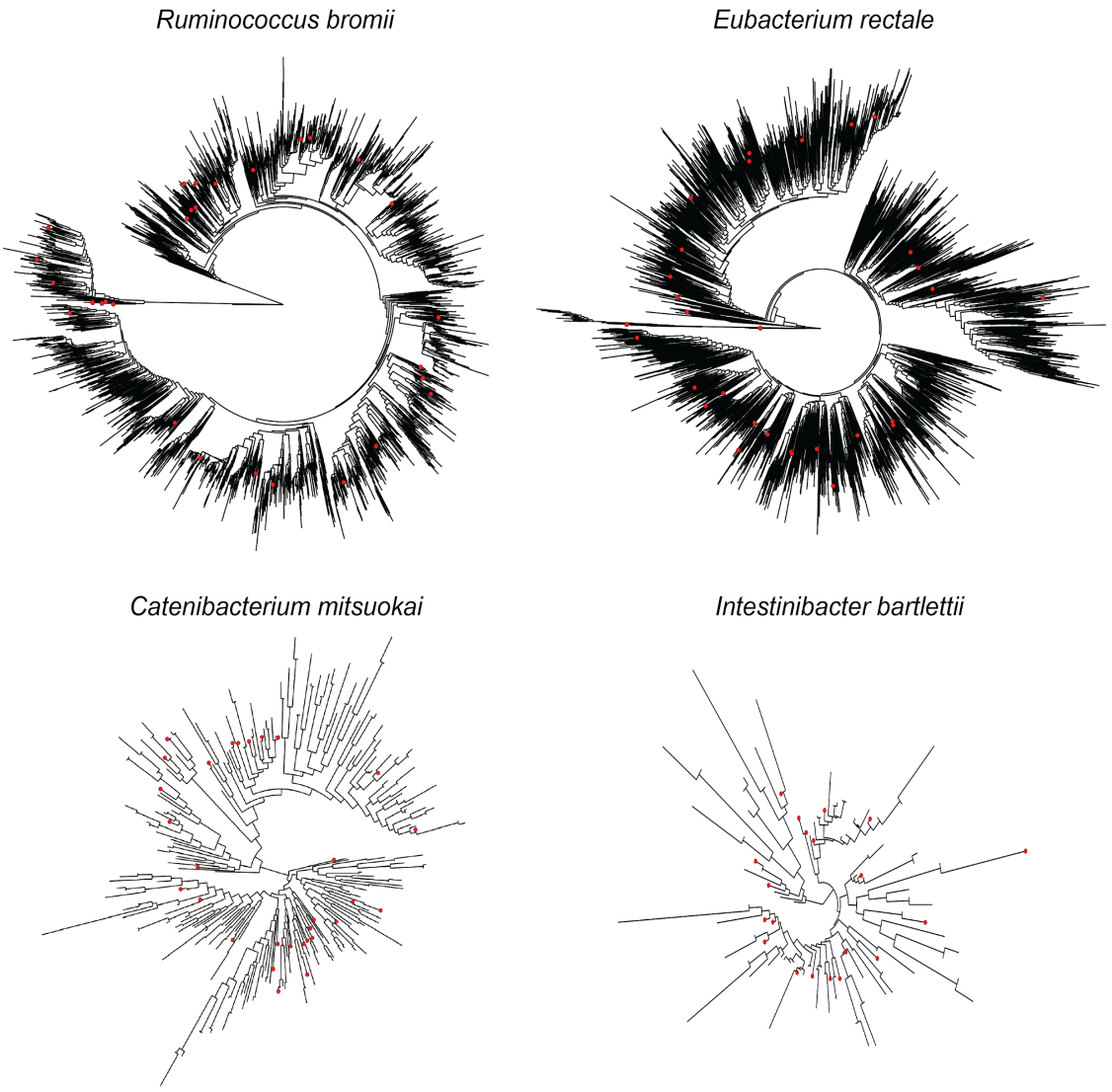
from one simulated sequence as a reference. **b**, Genome sequences were generated by a given phylogeny comprising 20 taxa using the same way as (a.)**,** but genome 1-10 (illustrated as solid lines) were used to generate synthetic reads using GemSIM (McElroy, Luciani, and Thomas 2012) and genome 11-20 (illustrated as dashed lines) were used as input of continuous sequences. Afterwards, *metaclock_mac.py* was employed on this synthetic data with reference being available only from continuous sequences. Yellow stars indicate reference genomes used in alignment reconstruction with *metaclock_mac.py*.



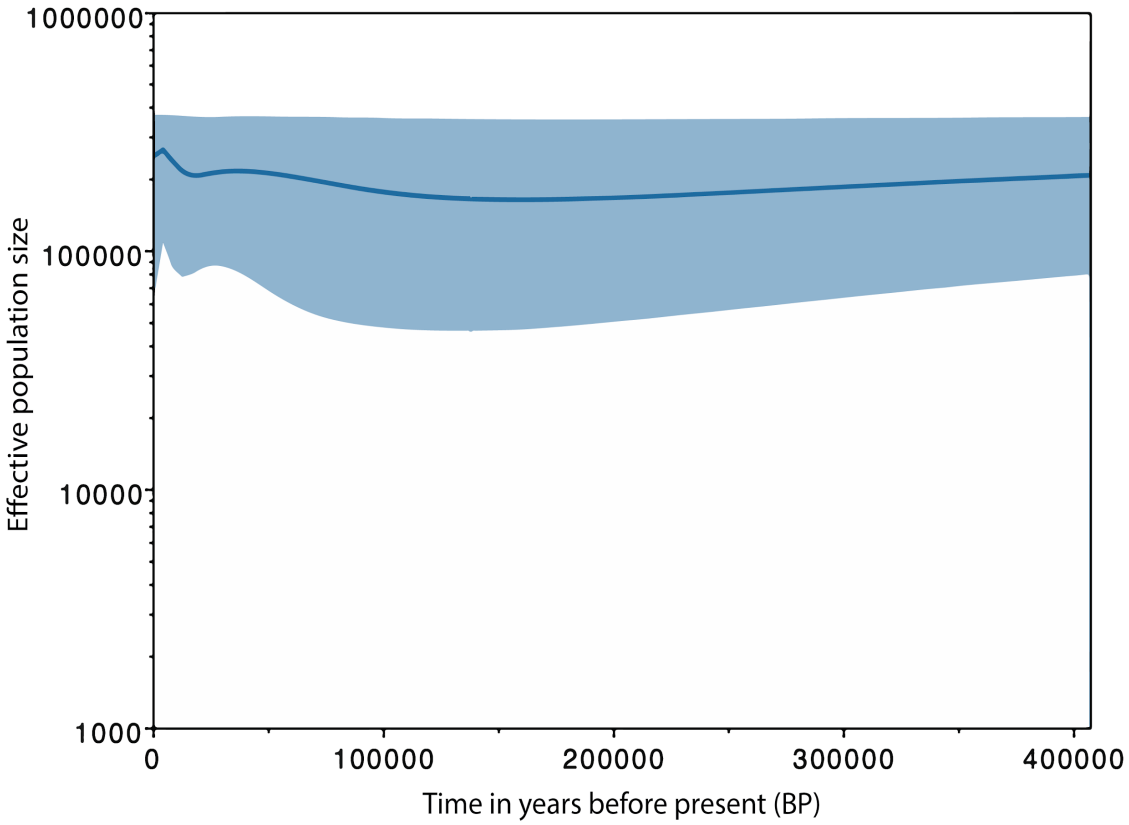
Supplemental Figure 4. Specie-level taxonomic abundance of top 25 species found in all ancient samples used in this study. Species highlighted with red box are common gut species and found to be abundant in ancient samples as well. Hence, they are selected as example species to be used in MetaClock in order to study their time-revolved evolutionary histories.



Supplemental Figure 5. Phylogenetic placement based on ML approach (Stamatakis 2014) implemented in MetaClock of ancient samples, reference genomes and MAGs when automated missing information trimming was executed.



Supplemental Figure 6. Phylogenetic diversity of four target species which have >100 MAGs available from contemporary populations. Each phylogenetic tree was reconstructed using PhyloPhlAn3.0 (Asnicar et al. 2020) and red dots indicate those selected in MetaClock analysis.



Supplemental Figure 7. Coalescent Bayesian skyline plot. The effective population size of *I. bartlettii*, estimated by the coalescent Bayesian skyline model using BEAST2 (Bouckaert et al. 2014), changes in the course of evolution. The plot was generated using Tracer1.7 (Andrew Rambaut et al. 2018), and the solid line indicates mean value and the blue shade suggests 95% highest density posterior interval.

3.12.2 Supplemental tables

Captions of supplemental tables are reported below. Tables are available at: https://drive.google.com/drive/folders/1UvSGp06COxxGeJKqBiiXGMzWx9kNTgtt?usp=sharing

Supplemental Table 1. The number of possible alignments resulting from combining various numbers of alignments from different references when using 10 reference genomes.

Supplemental Table 2. Basic statistics of genome alignment reconstruction using MetaClock for each species. #Positive samples indicates the number of ancient samples in which the target species were detected using MetaPhlAn 2.0 (Truong et al. 2015)**;** #Modern genomes available gives the total number of available

reference genomes and MAGs from contemporary metagenomic datasets; #Modern MAGs selected for MetaClock indicates the number of modern MAGs selected from #Modern genomes available based on the large-scale phylogenetic structure built using PhyloPhlAn3.0 (Asnicar et al. 2020) (**Supplemental Figure 6**); #Modern reference genomes selected for MetaClock indicate the number of reference genomes retrieved directly from the NCBI RefSeq database. #Taxa (in total) in the combined alignment for clocking is the total number of taxa (comprising those from both ancient and contemporary datasets) in the combined alignment refined by MetaClock and subsequently used for molecular clocking analysis.

Supplemental Table 3. BEAST2 model comparison for each species. The best clock and demographic models (highlighted with *) were selected, comparing the null model (highlighted with ^) with alternative models by Log Bayes factor based on marginal likelihood estimation using path sampling (Baele et al. 2012) for each target species. Model comparison was performed only on simulations with sufficient MCMC mixing which was visually checked using Tracer1.7 (Andrew Rambaut et al. 2018) for adequate effective sample size (ESS) of majority parameters.

Supplemental Table 4. The archaeological periods, C14 dating ages and geographical locations of ancient samples used as tip calibrations.

Supplemental Table 5. The description of semi-synthetic metagenomes used for MetaClock validation. *real_metagenome_sample_name* indicates five real metagenomic samples drawn from a previous study (Asnicar et al. 2017) and screened for absence of target species *R. bromii*. *MetaClock_target_genome* gives seven strains of *R. bromii* used for generating synthetic reads and then added in the real metagenomic samples; *MetaClock_target_genome_coverage* is the coverage used for synthesizing reads from each genome strain.

Supplemental Table 6. The accession number of reference genomes from six selected species under NCBI database.

Chapter 4 | Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal *Methanobrevibacter*

4.1 Introduction to the chapter

In this chapter, I extend the application of the tool introduced in **Chapter 3** to interrogate the evolutionary history of microbes from the other human microbial ecosystem, the oral cavity. This is an important extension because the success will promise a wider user community considering the existing and continuingly growing number of ancient oral metagenomes (Weyrich et al. 2017; Mann et al. 2018; Velsko et al. 2019; Warinner et al. 2014). In the study presented in this chapter, we firstly described the microbial composition in the 20 newly sequenced human calculus metagenomes from archaeological samples and then compared with publicly available oral samples, with an intriguing discovery of rich archeal *Methanobrevibacter* traces. Afterwards, We next sought to reconstruct ancient *Methanobrevibacter* genomes employing *de novo* metagenomic assembly. We uncovered two uncharacterized *Methanobrevibacter* species by placing the newly reconstructed genomes in the whole phylogeny of *Methanobrevibacter* genus. We then established their evolutionary diversity regarding time span and geographical locations using the approach I elaborated in **Chapter 3**. This work sets as an important example of the potential of the tool described in **Chapter 3** in establishing a fine-grained evolutionary diversity for microbiome members which are present in both ancient and contemporary populations in the other human microbial ecosystem such as the oral cavity.

Contribution: In this study, I mainly contributed to *de novo* metagenomic assembly on our unpublished ancient calculus samples, to identify the novel species in the *Methanobrevibacter* genus, to reconstruct the evolutionary diversity of three most abundant *Methanobrevibacter* species in the ancient samples. I interpreted the data and wrote the manuscript with the other joint first author.

Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal *Methanobrevibacter*

Lena Granehäll*, Kun D. Huang*, Adrian Tett, Paolo Manghi, Alice Paladin, Niall O'Sullivan, Omar Rota-Stabelli, Nicola Segata, Albert Zink, Frank Maixner

* these authors contributed equally

4.2 Abstract

Background: Dental calculus (mineralized dental plaque) preserves many types of microfossils and biomolecules, including microbial and host DNA, and ancient calculus are thus an important source of information regarding our ancestral human oral microbiome. In this study, we taxonomically characterized the dental calculus microbiome from 20 ancient human skeletal remains originating from Trentino-South Tyrol, Italy, dating from the Neolithic (6000-3500 BCE) to the Early Middle Ages (400-1000 CE).

Results: We found a high abundance of the archaeal genus *Methanobrevibacter* in the calculus. However, only a fraction of the sequences showed high similarity to *Methanobrevibacter oralis*, the only described *Methanobrevibacter* species in the human oral microbiome so far. To further investigate the diversity of this genus, we used *de novo* metagenome assembly to reconstruct 11 *Methanobrevibacter* genomes from the ancient calculus samples. Besides the presence of *M. oralis* in one of the samples, our phylogenetic analysis revealed two hitherto uncharacterised and unnamed oral *Methanobrevibacter* species, that are prevalent in ancient calculus samples sampled from a broad range of geographical locations and time periods.

Conclusions: We have shown the potential of using *de novo* metagenomic assembly on ancient samples to explore microbial diversity and evolution. Our study suggests that some members of the pre-industrial human oral microbiome such as the newly discovered archaeal species are now rare in the modern human oral microbiome.

Keywords: ancient DNA, ancient dental calculus, oral microbiome, metagenomics, *de novo* assembly, *Methanobrevibacter*

4.3 Background

Dental calculus develops via the mineralization of plaque, which can remain for millennia on ancient skeletal remains. It displays a specific microniche in the oral cavity and has been shown to perfectly preserve ancient biomolecules (DNA, proteins, metabolites) (Mackie et al. 2017; Velsko et al. 2017; Warinner, Speller, and Collins 2015) and dietary microfossils (pollen, starch) (Hardy, Buckley, and Copeland 2018). Ancient dental calculus have therefore been used as a source of information to study the composition and functional properties of oral microbial communities, diets and health in the past (Warinner et al. 2014; Weyrich et al. 2017). For example, changes in the oral microbial composition from the Neolithic to the Industrial Revolution has indicated that dietary shifts have played a key role in altering our oral microbial ecosystems (Adler et al. 2013). Further studies of ancient calculus have identified the presence of various keystone oral pathogens (Warinner et al. 2014), in particular members of the so called “red-complex”, a group of bacteria highly associated with periodontal disease, as well as the presence of inflammatory host response proteins (Warinner et al. 2014; Bravo-Lopez et al. 2020; Jersie-Christensen et al. 2018; Mann et al. 2018; Neukamm et al. 2020; Velsko et al. 2019; Ziesemer et al. 2015). Despite this, little is known about the role and diversity of non-bacterial microbes in ancient dental calculus. In comparison to modern plaque and calculus, ancient calculus have reportedly a higher abundance of archaea, dominated by the genus *Methanobrevibacter* (Velsko et al. 2019). Currently, *Methanobrevibacter oralis* is the only isolated and characterized *Methanobrevibacter* species in the human oral microbiome (Ferrari et al. 1994), but undetermined *Methanobrevibacter* species have been continuously found in ancient dental calculus (Weyrich et al. 2017; Mann et al. 2018; Ziesemer et al. 2015; Huynh et al. 2016) in as high abundances as >60 % (Ziesemer et al. 2015). This frequent occurrence has allowed the reconstruction of a draft genome of a member of *Methanobrevibacter* from a Neanderthal calculus which was denominated *Methanobrevibacter oralis* neanderthalensis, the only reconstructed ancient *Methanobrevibacter* genome to date (Weyrich et al. 2017). Despite the high abundance of *Methanobrevibacter spp.* in ancient dental calculus and its possible involvement in periodontal disease (Griffen et al. 2012; Lepp et al. 2004), the current knowledge on the diversity and role of these methanogens in the human oral microbiome is limited. To better understand the complex diversity and evolution of the oral microbiome, here we analysed dental calculus from ancient human remains from Trentino-South Tyrol region in Northern Italy, spanning from the Neolithic to the Early Middle Ages (Figure 1). Apart from studying the overall microbial community of the

samples, we focused on the less studied archaeal component of the oral microbiome and *de novo* metagenomic assembled ancient genomes of 76 members of the genus *Methanobrevibacter*. We discovered two hitherto unknown archaeal *Methanobrevibacter* species and show that the actual *Methanobrevibacter* genetic diversity is much larger than previously thought. This has important implications for understanding the evolution of the dental calculus microbiome in humans in the last few millennia.

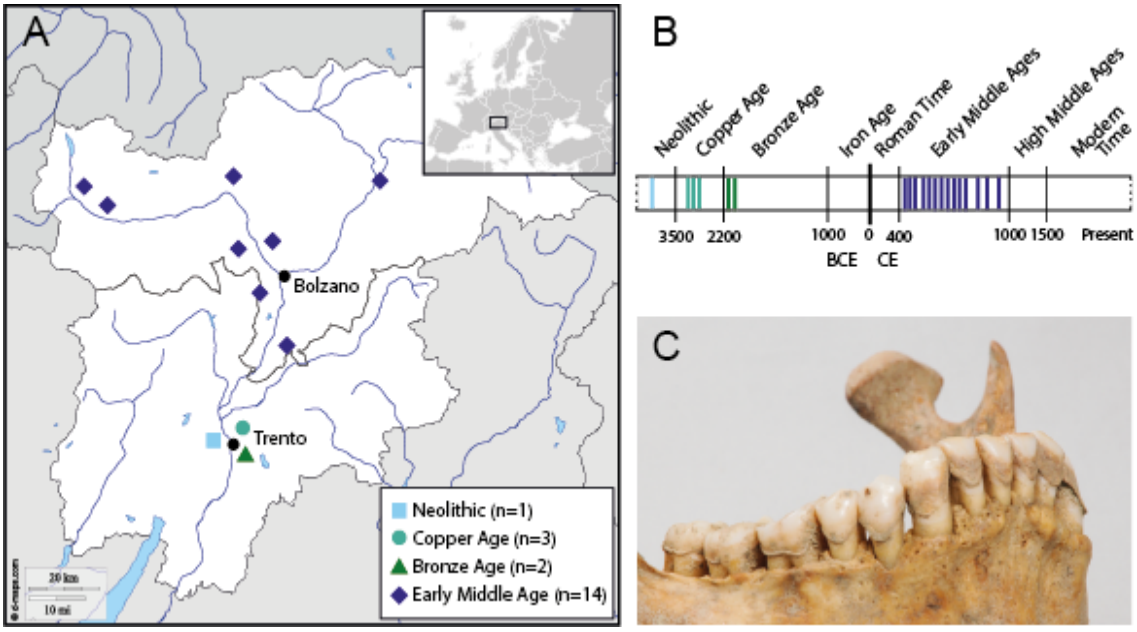


Figure 1. A) Spatial and B) temporal location of the 20 dental calculus samples in this study. Archaeological time periods based on Festi et al. (2014). C) Dental calculus on the mandibular teeth of the early medieval individual 2100, found in the burial site of Burgusio Santo Stefano (St. Stephan ob Burgeis) in South Tyrol, Italy.

4.4 Methods

4.4.1 Individuals

Ancient dental calculus from 20 individuals found in various archaeological sites in Trentino-South Tyrol, Italy, spanning from the Neolithic to the Early Middle Ages (Figure 1, Additional file 2: Table S1), were subjected to in-depth metagenomic analysis to characterise the oral microbiome and further describe the diversity of the genus *Methanobrevibacter.* Anthropological analysis (sex and age at death estimations) and radiocarbon dating of the human remains from the Early Middle Age were previously published in (Paladin et al. 2020), and the prehistoric individuals were analysed according to the same methods. Periodontitis and caries assessments were performed by qualitative measurements (Additional file 2: Table S1). Periodontitis was determined to be present or absent. Severity of caries was determined using a scoring system of 0-4, where 0 is “absent” and 4 indicates a “severe destruction of the dental crown due to caries”.

4.4.2 Sampling, DNA extraction and sequencing

All genetic laboratory work was conducted in the dedicated ancient DNA lab of the Institute for Mummy Studies at Eurac Research, Bolzano, Italy. Dental calculus (10-66mg) was removed from the surface of each tooth using sterile tweezers and probes. Each sample was sprayed with 3% H_2O_2, and then exposed to UV light for 10 minutes to sterilize the calculus surface. Overnight DNA extraction was done using a solution of 0.5M EDTA, 20 mg/ml proteinase K and 0.1M N-Laurylsarcosine, modified from (Rohland, Siedel, and Hofreiter 2010) and (Damgaard et al. 2015). Extracts were concentrated using 10K Amicon® filter devices and purified using Qiagen® MinElute PCR Purification Kit. Extraction blanks were included for every five samples. Single-indexed adapted libraries were prepared according to Meyer and Kircher 2010 (M. Meyer and Kircher 2010) with unique 7 bp indexes. The libraries were pooled and paired-end sequenced on an Illumina Hiseq4000, including one of the extraction blanks. Data is available from the European Nucleotide Archive under accession no. PRJEB43389.

4.4.3 Pre-processing sequencing data

Paired Illumina sequences were merged and adapters were trimmed using PEAR (Zhang et al. 2014) with a minimum merging overlap and minimum sequence length of

25 bp. Per base quality was set to 25, and trimmed using QualityFilterFastQ (Kircher 2012). After quality control, all samples were filtered against the reads in the extraction blank to reduce cross contamination effects (https://sourceforge.net/projects/bbmap/) (Bowers et al. 2017).

4.4.4 Pre-processing comparative datasets

For comparative analysis, previously published ancient and modern dental calculus metagenomic sequences from humans, Neanderthals and baboons (Warinner et al. 2014; Weyrich et al. 2017; Mann et al. 2018; Velsko et al. 2019; Ottoni et al. 2019) were downloaded and pre-processed as described above. Cutadapt (Martin 2011) was used to remove adapters from single end read dataset, with otherwise the same quality cut-offs. Modern plaque, tongue dorsum, stool and skin samples were collected from the human microbiome project (HMP) (Consortium and The Human Microbiome Project Consortium 2012) (Additional file 2: Table 122 S2). Soil samples were downloaded from a study of Johnston and colleagues (Johnston et al. 2016). Modern tongue dorsum, skin, stool, plaque and soil datasets were trimmed and adapters removed with Cutadapt, with a minimum sequence length of 30 bp, and the per base quality cut off of 20. Only samples reported with low contamination and with >6,000,000 reads after pre-processing were used for comparative and statistical analysis to match the amount of reads in the Trentino-South Tyrolean calculus dataset.

4.4.5 Authentication

After pre-processing, taxonomic assignment was performed using MetaPhlAn2 (version 2.7.7) (Truong et al. 2015). For ancient samples, the non-default minimum read length threshold was set to 30 bp to adjust for short aDNA fragments (--read_min_len 30).

To authenticate the metagenomic sequences as coming from an ancient oral source we first used Sourcetracker2 (Knights et al. 2011) on species-level data to identify the source of possible contaminants. Other comparative datasets were used as sources (soil, tongue dorsum, skin, and modern calculus).

A species-level PCoA was also performed to confirm the oral origin of the ancient dental calculus sequences in comparison to microbial communities from plaque, tongue dorsum, skin, modern and ancient calculus. Bray-Curtis distances were

calculated based on the normalized taxonomic assignment (https://ggplot2.tidyverse.org/) (Wickham 2016).

To assess typical ancient DNA damage of the metagenomic reads, MapDamage2 (Jónsson et al. 2013) was run on sequences aligned to the genomes of microorganisms detected in the calculus samples. Pre-processed reads were aligned to the highest abundant identified taxa (> 2.3% average abundance) in the Trentino-South Tyrolean calculus dataset, as well as the members of the red complex: Bacteroidetes oral taxon 274 (NZ_GG774889.1), Desulfobulbus oral taxon 041 (GCA_000349345.1), *Eubacterium saphenum* (NZ_GG688422.1), *Fretibacterium fastidiosum* (GCA_000210715.1), *Porphyromonas* (NC_018142.1), *Streptococcus sanguinis* (NC_009009.1), *Tannerella forsythia* (NC_016610.1) and *Treponema denticola* (NC_002967.9). Alignment to microbial genomes was performed using Bowtie2 (Ben Langmead and Salzberg 2012), with the 'very sensitive local setting' (-D 20 -R 3 -N 1 -L 20 -I S,1,0.50) and deduplicated using DeDup v0.11.3 (https://github.com/apeltzer/DeDup). Alignment quality was set to 30.

4.4.6 Human DNA analysis

Pre-processed sequences from the ancient dental calculus were aligned to the human reference genome (build hg19) using BWA (H. Li and Durbin 2009) with seed disabled and then deduplicated using DeDup. Minimum mapping quality was set to 30. MapDamage2 was used to assess ancient DNA damage on human reads. Genetic sexing was performed as described in (Skoglund et al. 2013). Genetically determined sex was used for subsequent analysis, except when the genetic sex could not be determined in which case anthropologically determined sex was used.

4.4.7 Calculus microbiome taxonomic characterisation

Prior to the diversity and statistical analysis all samples were normalized to 6,000,000 reads (using seqtk version 1.3-r106, https://github.com/ lh3/seqtk) to match the lowest number of sequences in the Trentino-South Tyrolean dataset. Bray-Curtis distances between oral microbiomes (ancient calculus, modern calculus, plaque, tongue dorsum) were calculated based on the normalized taxonomic assignment from MetaPhlan2 using Qiime2 (Bolyen et al. 2019) and visualized with ggplot2.

4.4.8 Analysis of *Methanobrevibacter* abundance

To determine differences in microbial composition between ancient and modern calculus, especially differences in *Methanobrevibacter* abundance, a linear discriminant analysis was performed on phylum, genus and species abundances using LEfSe (Segata et al. 2011). The threshold on the logarithmic LDA score for discriminative features was set to 4. Differences in phylum abundance between ancient and modern was visualized using pheatmap in R (Kolde and Kolde 2015), only showing phyla >2% abundance. A t-test from the R package ggpubr (Alboukadel 2018) was used to determine differences in each phyla that were statistically significant between the two groups.

We tested whether there were any correlation between the abundance of *Methanobrevibacter* and various categories of metadata (level of periodontitis, age, sex, time period, etc.) using Kruskal-Wallis rank sum test and pairwise Wilcoxon tests in R. Correlations between species were calculated with Spearman's correlation with the Hmisc package in R (Harrell and Harrell 2019).

4.4.9 Microbial genome reconstruction from ancient calculus samples

Each of the 20 pre-processed ancient calculus samples in this study were subjected to *de novo* metagenome assembly using metaSpades (version 3.10.1; default parameters) which was evaluated to outperform among other metagenome assemblers (Forouzan et al. 2018; Nurk et al. 2017). We obtained 92,485 contigs (> 1,000 nt) which were kept for further processing. Reads were aligned against contigs using Bowtie2 (version 2.2.9; '--very-sensitive-local') and the output was used for contigs binning using MetaBAT2 (version 2.2.9; '-m 1500'), resulting in 117 bins (i.e., putative genomes) (Ben Langmead and Salzberg 2012; Kang et al. 2015). The putative genomes were assessed for completeness, contamination and strain heterogeneity using CheckM (version 1.0.7; lineage specific workflow), to select the set of final draft genomes considered (Parks et al. 2015). Based on recent guidelines, we selected medium-quality (MQ) genomes that had completeness > 50% and contamination < 5%, resulting in 76 metagenome assembled genomes (MAGs) (Bowers et al. 2017). Each of the microbial genomes was assessed for genome size (bp), number of contigs, contig N50 values, mean contig length and the longest contig (Table S7) using QUAST (Gurevich et al. 2013). To assert the endogenous origin of reconstructed microbial genomes, we checked each genome for nucleotide misincorporation rate patterns

which instrument to authenticate ancient sequences. The BAM files of alignment between reads and contigs were processed using mapDamage2 (default parameters) (Jónsson et al. 2013). Finally, we assigned a taxonomic label to each reconstructed microbial genome using PhyloPhlAn3.0 (phylophlan_metagenomic, '-d SGB.Aug19') (Asnicar et al. 2020).

4.4.10 Phylogenetic analysis of *Methanobrevibacter* genomes

To place the newly reconstructed ancient *Methanobrevibacter* genomes in the phylogenetic context of *Methanobrevibacter* genus, we used a total of 64 modern assembled genomes (52 reference genomes (Additional file 2: Table S13) and 12 MAGs from a previous large-scale metagenomic investigation (Pasolli et al. 2019)) representative of 17 known *Methanobrevibacter* species. Core genes were searched within ancient MAGs (n = 11) and their contemporary counterparts (n = 64), and were then concatenated into a core gene alignment of 42,225 bp length using PRANK (Löytynoja 2014) which is implemented in Roary pipeline (version 3.13.0; '-i 80 -cd 90 -e -mafft') (Page et al. 2015). To reconstruct the phylogenetic tree, we used RAxML (8.1.15) (Stamatakis 2014) under a GTR model of substitution with 4 gamma categories and 100 bootstrap pseudo replicates.

We then sought to reconstruct a more precise phylogeny for three subtrees partitioned from the *Methanobrevibacter* genus tree built as described above, using two similar but complementary methods. Firstly, we performed the phylogenetic analysis using RAxML (as described above) on the core gene alignment using genomes from three subtrees, respectively. Core gene alignments were produced using PRANK with parameters of 85% identity (90% for genomes in the subtree 3 due to the fact that subtree 3 contains fewer genomes) for gene clustering and of gene presence in >90% across genomes for defining core genes for each subtree. Secondly, we used the same RAxML phylogenetic method on the multiple sequence alignment (MSA) reconstructed based on the whole genome region. The whole-genome MSA for three subtrees were reconstructed, by firstly aligning whole genome sequences (as query genomes) from each subtree against one selected reference genome (GCA_003111605 for subtree 1; GCA_003111625 for subtree 2; GCA_001639275 for subtree 3) with BLASTn ('-word_size 9') (Altschul et al. 1990). Afterwards, genomic regions where query genomes and the reference genome share a substantial sequence similarity (BLASTn hits with length >500 bp and identity percentage >95%) were selected to generate the whole-genome MSA, followed by excluding columns having >10% missing.

To further support the phylogenetic distances with a method based on the genome similarity, we also estimated the average nucleotide identity (ANI) pairwise distances using pyani (version 0.2.6; option '-m ANIm') (Pritchard et al. 2016) for two subtrees (subtree 1 and subtree 2) which comprise the 10 newly reconstructed and uncharacterized *Methanobrevibacter* genomes from our ancient calculus samples. The measurement was performed on the whole genome sequences and on the core gene alignments generated above as well.

Next, we explored the strain-level phylogenetic diversity of the three *Methanobrevibacter* species with particular interest (the two newly discovered *Methanobrevibacter oralis*) including 82 publicly available oral microbiome samples (from 10 contemporary and 72 ancient individuals) (Additional file 2: Table S2). An alignment-based approach was used to reconstruct the whole genome alignment. GCA_001639275 was selected as a reference for a known species, *M. oralis*, and reconstructed MAGs with highest completeness and lowest contamination were selected as representatives of ancient lineages from subtree 1 and those from subtree 2 (here calc_2086.bin.1 for those from subtree 1 and calc_2094.bin.7 for those from subtree 2) due to a lack of reported reference in the public database (Additional file 2: Table S7). For previously published metagenomic samples, draft genomes were generated using a python script consensus.py from package cmseq (https://github.com/SegataLab/cmseq). It aligned metagenomic reads of each sample against single references and extracting consensus sites of aligned reads. Sites covered by reads were filled with gaps if: (1) mapping quality < 30, 243 (2) coverage < 3 folds, (3) minimum identity of reads < 97%, (4) aligned read length < 30nt, (5) minimum dominant allele frequency < 80%. Newly sequenced samples of this study from which *Methanobrevibacter* genomes could not be obtained by *de novo* assembly were also subjected to the same approach. For other newly sequenced samples whose MAGs are available, we aligned contigs against the same selected references using BLASTn ('-word_size 9'). A whole-genome MSA was compiled based on the same single reference, integrating reconstructed draft genomes and highly-similar sequences (>95% identity percentage and >500 bp length) of aligned MAGs. We cleaned each alignment, excluding sequences with >50% gaps and then removing columns containing >10% missing data. The cleaned alignments were used in reconstructing the phylogenetic trees with RAxML as above. Recombination events for these two subtrees were analysed using ClonalFrameML (Didelot and Wilson 2015) (v1.25, default parameters) under a ML phylogenetic context. Alignments were masked from recombination by replacing genomic regions affected by recombination with gaps using

python script maskrc-svg (https://github.com/kwongj/maskrc-svg). The ML phylogeny was estimated for masked alignments. The process was iteratively repeated until no recombination events were detected. To confirm the phylogeny built on the MAGs from this study, we also reconstructed draft genomes from short-sequencing reads of the same metagenomic samples using the alignment-based approach as described above and then repeated the same phylogenetic analysis.

4.4.11 Functional annotation and pangenome analysis of the ancient *Methanobrevibacter* bins

In order to analyse the potential difference in function between the three *Methanobrevibacter* species (*M. oralis*, TS-1 and TS-2) the assembled ancient *Methanobrevibacter* bins were annotated with prokka (Seemann 2014) and uniref90. The pangenome for each of the three species was determined using Roary (Page et al. 2015). The *M. oralis* pangenome was created using the four *M. oralis* assembled genomes (GCA_001639275, GCF_000529525.1, GCF_900289035.1 and GCF_902384065.1) present in Genbank, as well as the ancient *M. oralis* bin from sample 2102. We determined genes that were unique to the two ancient *Methanobrevibacter* pangenomes by excluding genes found in the *M. oralis* pangenome. To find common genes missed by the automatic selection we then manually inspected the gene tables, removed M. oralis orthologs found in OrthoDB (vs 10.1) and finally conducted a blastn search to find any similar sequences in *M. oralis* and other *Methanobrevibacter* species.

4.4.12 Methyl coenzyme M reductase *(mcrA) gene* analysis

The methyl coenzyme M reductase complex (Mcr) is a key enzyme in methanogenesis, and one of its subunits, *mcrA*, is a common marker gene for methanogenic archaea (Evans et al. 2019). To determine if this pathway was present in the ancient genomes, we extracted the *mcrA* gene sequences from the *Methanobrevibacter*-assembled genomes after annotation with prokka. For samples that had not generated any *Methanobrevibacter* bins the metagenomic sequences were aligned to the *mcrA* gene from their phylogenetically closest genomes (*M. oralis* DSM 7256, TS-1 or TS-2). These additional *mcrA* sequences were extracted from the aligned bam-files with ANGSD (version 0.918) and samtools faidx (version 1.10) (Korneliussen, Albrechtsen, and Nielsen 2014). Only bases covered at least 3 times and with a minimum alignment score of 30 were used for the consensus sequence. (angsd -dofasta 2 -doCounts 1 -setMinDepth 3 -minQ 30). The *mcrA* dataset was complemented with *mcrA*

sequences from currently available genomic datasets of species in the genus *Methanobrevibacter* (n=13). First, all DNA sequences were translated into amino acids by using the Perform Translation tool in the ARB software package (Ludwig et al. 2004). The *mcrA* sequence alignment was automatically inferred with the ClustalW protein alignment program (Thompson, Higgins, and Gibson 1994), implemented in the ARB software package and then manually refined by using the ARB sequence editor. The amino acid alignment of selected samples were further examined for the conservation of important positions in the catalytic site (Ermler et al. 1997; Borrel et al. 2019). For the phylogenetic analysis of the *mcrA* gene, the nucleic acid sequences were re-aligned according to the corresponding amino acid sequences with the respective tool in the ARB software. Phylogenetic analysis was performed by applying distance-matrix, maximum-parsimony, and maximum- likelihood methods implemented in the ARB software package: neighbour-joining (using the Jukes-Cantor algorithm for nucleic acid correction with 1000 bootstrap iterations), DNA parsimony (PHYLIP version 3.66 with 100 bootstrap iterations), and DNA maximum-likelihood [PhyML (Guindon and Gascuel 2003) with the HKY substitution model]. In total 1541 alignment columns were used for phylogenetic analysis.

4.5 Results

4.5.1 Taxonomic characterisation and authentication of the ancient calculus microbiome

We sampled the dental calculus from 20 individuals found in 11 burial sites located in Trentino-South Tyrol (Figure 1, Additional file 2: Table S1), Italy, and subjected them to shotgun metagenomic sequencing. Most of the individuals showed signs of oral diseases, with 80% (16/20) of them affected by periodontitis, which were severe (level 2) in 70% (14/20) of the cases. 75% (15/20) of the individuals showed carious lesions of very high level (destructive decays) (Additional file 2: Table 2). Only two individuals (10%) did not show any signs of oral disease.

To authenticate the sequences as coming from an oral source and to determine the composition of the microbiome, they were first taxonomically profiled with MetaPhlAn 2 (Figure 2A). We found a total of 222 microbial species in the Trentino-South Tyrolean dataset (Additional file 2: Table 3). The most abundant phyla were *Firmicutes*, *Proteobacteria*, *Actinobacteria*, and *Bacteroidetes*, consistent with the reported major phyla in the expanded Human Oral Microbiome Database eHOMD and other ancient

calculus microbiomes (Warinner et al. 2014; Weyrich et al. 2017; Escapa et al. 2018). Of the 25 highest abundant species (Figure 2A), 21 were categorised as oral in eHOMD, and 13 of them have been associated with periodontitis. These include the “red complex” members *Treponema denticola* and *Tannerella forsythia* (Socransky et al. 1998), but also species from the genera *Methanobrevibacter* and *Desulfobulbus*, anaerobes found in periodontal deep pockets and regarded as "late colonisers" of oral plaque (Griffen et al. 2012; Lepp et al. 2004). Further species-level PCoA analysis based on Bray-Curtis distances showed that our dental calculus samples cluster together with previously published ancient calculus samples (Figure 2B), and showed similarities to modern calculus (mineralized plaque sampled from living humans, as opposed to the living biofilm which constitutes plaque), but were distinctly different from the other modern human datasets used for comparison (plaque, tongue dorsum, skin and stool) (Additional file 1: Figure S1). We did not find any further clustering of the calculus samples based on the time period of the sample or the health status, origin, sex, and age at death of the individuals (Additional file 1: Figure S2). Additional analysis using Sourcetracker2 also showed that of the reads stemming from a known source, a majority were from oral sources (predominantly modern calculus and plaque) (Additional file 1: Figure S3). All samples displayed post-mortem damage typical for ancient DNA for the selected highest abundant microbial species (Additional file 1: Figure S4, Additional file 2: Table S4), indicating that the reads are stemming from the ancient calculus and not from recent microorganisms. In addition to the microbial DNA, the calculus metagenomic sequences contained on average 0.12% human DNA (Additional file 2: Table S5). Genetic sex of the human host could be determinated for 18 of the individuals. Post-mortem damage in the human reads was detected in all samples (Additional file 2: Table S5). Altogether, all samples were considered authentic and used for downstream analysis, except sample 2100 which had substantially fewer reads and was therefore not used for further comparative analysis.

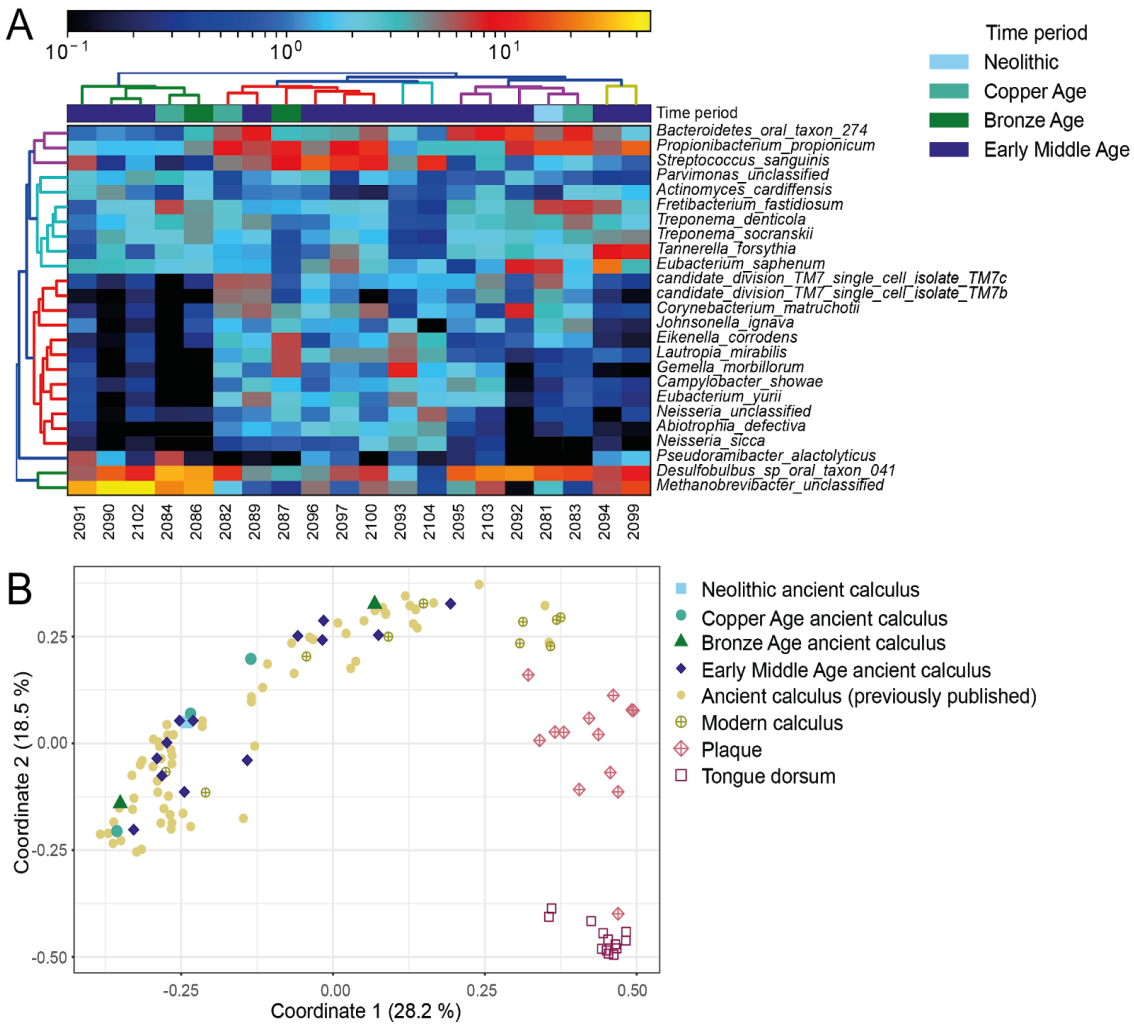


Figure 2. Taxonomic analysis of the microbial composition of the ancient dental calculus samples collected. A) MetaPhlAn heatmap of the top 25 species found in the ancient calculus. B) Species-level Principal Coordinate Analysis (PCoA) of β-diversity (Bray Curtis distances) considering the microbiome from ancient calculus (this study, (Warinner et al. 2014; Weyrich et al. 2017; Mann et al. 2018; Velsko et al. 2019; Ottoni et al. 2019) modern calculus (Velsko et al. 2019), plaque and tongue dorsum (HMP) (Consortium and The Human Microbiome Project Consortium 2012).

4.5.2 *Methanobrevibacter* is highly abundant in ancient calculus

We sought to compare the taxonomic profiles of the Trentino-South Tyrolean ancient calculus to a total number of 76 published ancient and modern calculus samples (Additional file 2: Table S2) to determine if there was a difference in microbial community composition between the different tissues. We observed in both ancient and modern calculus microbiomes high abundances of members of *Firmicutes*, *Proteobacteria*, *Actinobacteria*, and *Bacteroidetes*, whereas members of

Euryarchaeota, *Fusobacteria*, *Synergistetes*, *Spricohaetes* and *Candidatus Saccharibacteria* were underrepresented, indicating a similar pattern between the ancient calculus and its contemporary counterpart regarding the taxonomic composition at the phylum level (Figure 3A). However, ancient calculus had on average 8 times higher amounts of *Euryarchaeota* than modern calculus (p=2.4e-07) (Figure 3A). This was represented by the genus *Methanobrevibacter*, the only archaeal genus present in the calculus samples which was found to be one of the major features to explain the differences between ancient and modern calculus microbiomes (LDA score = 4.62) (Figure 3B). Other species significantly represented in the ancient calculus include *T. denticola*, *E. saphenum* and *Desulfobulbus* oral sp. 041 (Additional file 1: Figure S5, Additional file 2: Table S6).

Methanobrevibacter sequences were present in all Trentino-South Tyrolean samples (0.11-47.3% relative abundance), and in five samples it constituted the dominant taxa of the whole microbiome (22.5-47.3% relative abundance) (Additional file 2: Table S3). The abundance of *Methanobrevibacter* was one of the driving components of the clustering of calculus microbiomes, displaying a gradient of abundance across the first coordinate of the PCoA plot which attributed to 34.4% of the variance in the dataset (Figure 3C). Statistical analysis revealed no correlations between the abundance of *Methanobrevibacter* and different groups of metadata in all calculus samples, including the recorded oral diseases, sex, and age at death. Looking at correlations between abundances between *Methanobrevibacter* and other species, we found several positive and negative correlations (Additional file 2: Table S6). Two out of four observed positive correlations of *Methanobrevibacter* was to other bacterial species that have been associated with periodontitis, e.g. *Filifactor alocis*, (corr=0.3328, p=0.000982), and *Streptococcus anginosus* (corr= 0.3377, p= 0.000815).

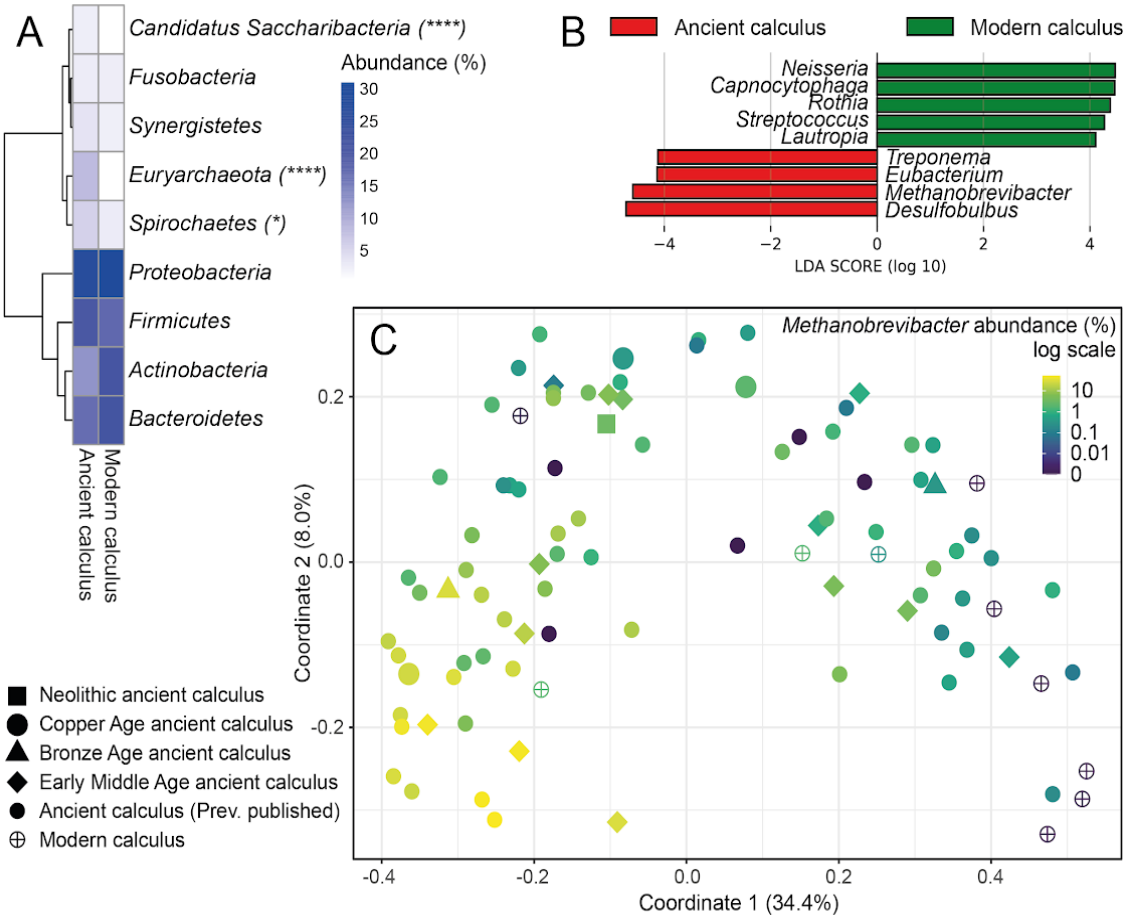


Figure 3. Differences in microbial taxonomic profiles between ancient and modern calculus and *Methanobrevibacter* abundance in dental calculus A) Phylum abundance (%) in ancient and modern calculus. Phyla that are statistically different between the two tissues are marked with asterisks (* $p \leq 0.05$, **** $p \leq 0.0001$). B) Differences in genera between ancient and modern calculus (LEfSe). C) Species-level Principal Coordinate Analysis of the β-diversity (Bray Curtis distances) of all ancient and modern calculus samples showing the abundance of *Methanobrevibacter*. Previously published ancient calculus from Mann et al. 2018, Warinner et al. 2014, Weyrich et al. 2017, Velsko et al. 2019, Ottoni et al. 2019, and modern calculus from Velsko et al. 2019.

4.5.3 Metagenome assembly of calculus samples reveals two novel *Methanobrevibacter* species, TS-1 and TS-2

We sought to reconstruct bacterial and archaeal genomes from ancient calculus specimens from the Italian Trentino-South Tyrolean region, performing *de novo* metagenome assembly on each calculus metagenomic sample. From the 20 calculus, we reconstructed a total of metagenome assembled genomes (MAGs). All reconstructed MAGs were processed by strict quality measurement including

completeness, contamination, and strain heterogeneity, resulting in 76 of them to be considered of medium quality (MQ) or above as proposed by recent guidelines (Bowers et al. 2017) (completeness > 50% and contamination < 5%). The endogenous origin of these 76 genomes were authenticated with each displaying clear damage patterns at the both ends of reads (Additional file 1: Figure S6). At the species level only 16 of the genomes are close enough (Mash distance (Ondov et al. 2016) < 5%) to the genome with a known species-level taxonomic label as defined by the Species-level Genome Bin (SGB) analysis (Asnicar et al. 2020; Pasolli et al. 2019) (Additional file 2: Table S8). The remaining genomes (n=60) were assigned as unknown.

Additionally, a total of 11 ancient genomes were identified to be within the *Methanobrevibacter* genus (Additional file 2: Table S8) and fell in three separated subtrees when placed in phylogenetic context with all publicly available representatives of the genus *Methanobrevibacter* (Figure 4A). One genome fell within the *M. oralis* clade with a Mash distance < 2% from the *M. oralis* reference genome, the other 10 ancient genomes formed two independent clades which were clearly distinct to other *Methanobrevibacter* species. The phylogeny of each subtree can be repeated and with an improved resolution at branch length when reconstructing the maximum likelihood tree separately using genomes from each subtree (Additional file 1: Figure S7). The ANI pairwise distances between the newly reconstructed *Methanobrevibacter* genomes and the closest modern relatives showed a limited distance among ancient genomes based on both the core genome and whole genome (Additional file 1: Figure S8). Conversely, ancient genomes were strikingly distant from the modern closest relatives, with mean distance >15% in subtree 1 and subtree 2, in terms of both core and whole genome (Additional file 1: Figure S8). The high ANI distance between ancient genomes and modern representatives, in both subtree 1 and subtree 2, suggests that the ancient genomes could represent two newly discovered species which have not yet been previously reported based on current operational ANI-based consensus on assigning strains to species (Pasolli et al. 2019; Tett et al. 2019; Goris et al. 2007; Jain et al. 2018; Konstantinidis and Tiedje 2005).

The distinction of ancient genome clades in subtree 1 and subtree 2 is further supported by the clade separation with a clearly long branch length in the subtree phylogeny based on core genome as well as based on whole genome (Additional file 1: Figure S7A, S7B, S7C and S7D). These two candidate species (TS-1 for ancient genomes from subtree 1 and TS-2 for those from subtree 2) thus represent novel

lineages that are however not present anymore in the contemporary oral microbiome, or at least not prevalent enough to be detected by comparing with *Methanobrevibacter* genomes from modern metagenomics datasets.

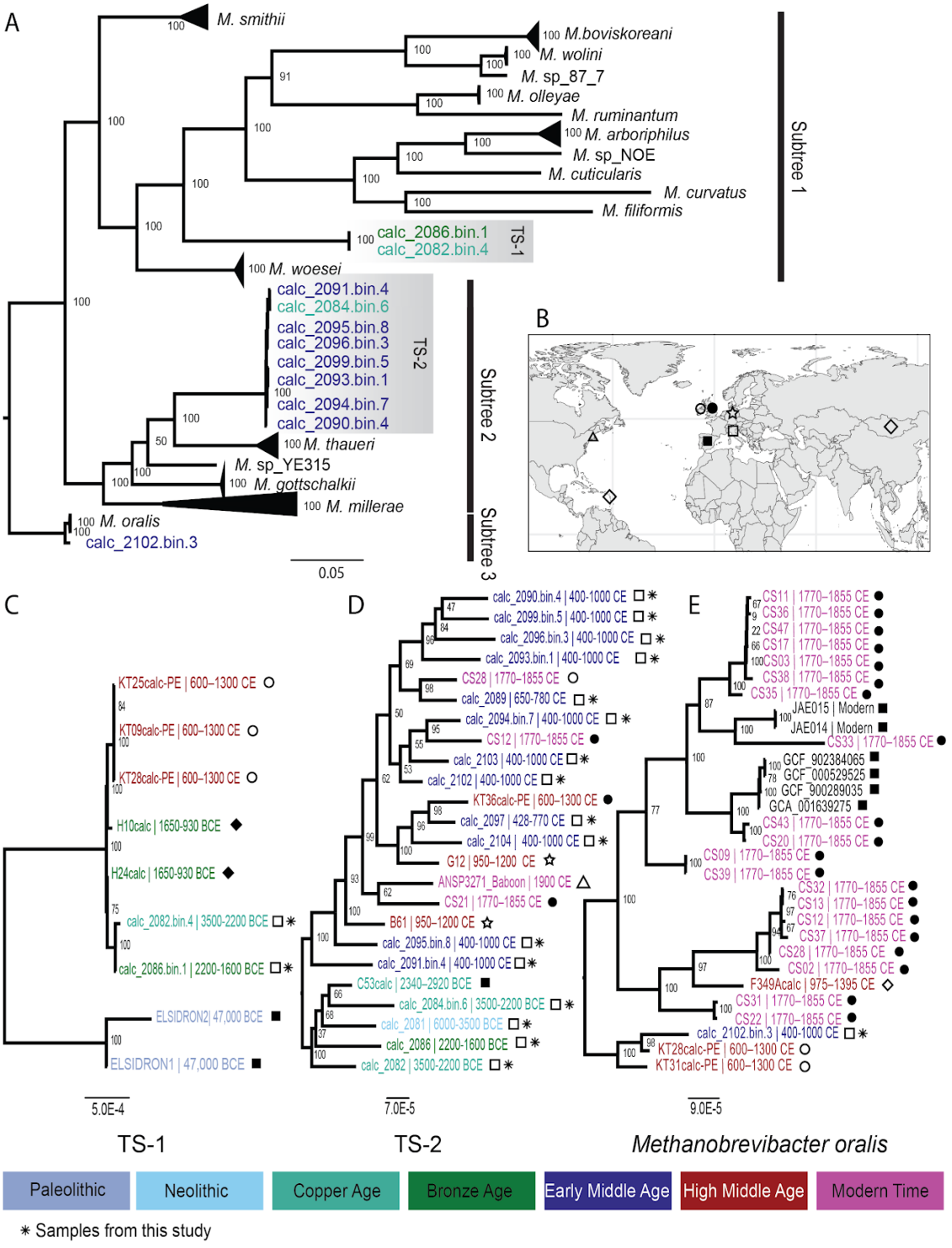


Figure 4. Phylogenetic analysis of ancient and modern *Methanobrevibacter* diversity.

The colors in the trees corresponds to the age of the samples: Paleolithic: >12000 BCE, Neolithic: 6000-3500 BCE, Copper Age: 3500-2200 BCE, Bronze Age: 2200-1000 BCE, Early Middle Age: 400-1000 CE, High Middle Age: 1000-1300 CE, Modern time: 1492 CE - present. A) Phylogenetic tree showing the *Methanobrevibacter* MAGs with modern *Methanobrevibacter* genomes B) Geographical location of ancient calculus samples placed in the phylogenetic context in (C, D and E). C) Phylogenetic tree of ancient calculus samples aligned to the highest quality TS-1 MAG (>50% covered at least 3 folds). D) Phylogenetic tree of ancient calculus samples aligned to the highest quality TS-2 MAG (>50% covered at least 3 folds). E) Phylogenetic tree of samples aligned to M. oralis (50% covered at least 3 folds).

4.5.4 Phylogenomic analysis provides historical insights into temporal diversity of three ancient-sample enriched *Methanobrevibacter* species

Studies focusing on modern oral microbiomes have consensus on the pivotal role methanogenic archaea play in human oral health (Lepp et al. 2004; Horz and Conrads 2011; Grine et al. 2018). The good-quality *Methanobrevibacter* genomes we recovered from ancient calculus samples of this study allowed us to study their evolutionary diversity, during a long-term period, at an unprecedented resolution. We firstly explored the phylogenetic position of newly reconstructed *Methanobrevibacter* MAGs in the whole diversity of genus *Methanobrevibacter* (Figure 4A). This was further supported by the consistent phylogenetic structure reconstructed independently using genomes from each subtree (Additional file 1: Figure S7).

Secondly, to study the biodiversity of three species represented by the 11 *Methanobrevibacter* MAGs obtained in this study, we performed comparative analysis exploiting 82 publicly available calculus metagenomes (Additional file 2: Table S2). Reconstructed *Methanobrevibacter* MAGs were divided into three groups based on their phylogenetic placement: group 1 included genomes from subtree 1; group 2 comprised genomes placed in the subtree 2; the single genome clustering with *M. oralis* lineages in the subtree 3 was assigned to group 3 (Figure 4A). We detected genomic signals of those three clades in the publicly available oral metagenomic samples by aligning metagenomic reads against the selected representative genomes. At length, seven samples showed at least 50% of the TS-1 representative genome covered at >3X, 12 samples for TS-2 and 25 samples for *M. oralis* (Additional file 1: Figure S9). These high-coverage samples were further included in phylogenetic analysis along with *Methanobrevibacter* MAGs reconstructed in this study.

Leveraging a total of 102 calculus samples (20 of this study and 82 from previous studies) which spanned ~50,000 years and were collected from eight countries across three continents shed light on the temporal and geographical diversity, from prehistory to modern era, of three *Methanobrevibacter* species enriched in ancient samples. We used a maximum likelihood approach to build a strain-resolved phylogeny based on the MSA reconstructed aligning metagenomic reads (or MAGs if available in the samples) against genome sequences calc_2084.bin.1, calc_2094.bin.7, and GCA_001639275, respectively representative of TS-1, TS-2, and *M. oralis*. To maximize the number of samples possibly being included without losing phylogenetic resolution, we used only samples which have >50% reference genome covered at >3X and retained only positions covered across >90% of genomes. Additionally, to ameliorate the phylogeny reliability, we iteratively removed the regions which were affected by homologous recombination until no recombination events were detected. We observed that the novel candidate species TS-1 have continuously presented from Late Pleistocene Age, Copper Age, Bronze Age, Iron Age, to Late Middle Ages (Figure 4C). The earliest evidence of its presence could be found in two 48,000 year old Neanderthal samples in one of which a *Methanobrevibacter oralis* like genome was previously reported to be present as well (Weyrich et al. 2017). The two Neanderthal lineages of TS-1 seem to be the ancestral to the more recent lineages from Italy, Mongolia and Ireland (Figure 4C). We did not observe any signature of this species in the modern calculus samples, and the latest record could be only traced in medieval samples from Ireland (Figure 4C, Additional file 1: Figure 9). Likewise, TS-2 has dispersed over multiple countries across continents during at least the last 5,000 years (Figure 4D). Compared to TS-1 and TS-2, *M. oralis*-like strains appear to be more prevalent in relatively recent samples from the 6th century CE to the modern era with a wide genetic diversity across multiple geographical locations (Figure 4E). Taken together, our phylogenetic analysis indicates that the three *Methanobrevibacter* (*M. oralis*, TS-1, TS-2) found in ancient dental calculus from a broad geographic range show first signs of a temporal diversification. To further verify these findings, it is however necessary to increase the statistical power by including more calculus samples in the future.

To verify the phylogeny which was reconstructed using MAGs we also performed the same phylogenetic analysis using draft genomes reconstructed by mapping metagenomic reads against a single reference genome. We obtained identical phylogenies for TS-1 and M. oralis (Additional file 1: Figure S10) when metagenomic reads were used directly. This added analysis further supports that the MAGs

generated in this study possess good sequence quality to perform high-resolution phylogenetic analysis. While a certain extent of phylogeny variability was observed in the TS-2 between using MAGs and using draft genomes reconstructed by the alignment-approach, most of these tips were arisen from nodes supported by low bootstrapping values (Additional file 1: Figure S10). The alternated phylogeny observed in TS-2 could also be because of effects from homologous recombination. Our recombination analysis using ClonalFrameML (Didelot and Wilson 2015) showed 3 times the amount of recombination events in TS-2 than in *M. oralis*, and 30 times than in TS-1 (Additional file 1: Figure S11).

4.5.5 Functional analysis of the calculus *Methanobrevibacter* species reveal similarities to other *Methanobrevibacter* species

To determine if there were any functional differences between the two newly discovered *Methanobrevibacter* species present in the ancient calculus samples compared to *M. oralis*, we characterised and annotated their respective pangenomes. 1628 genes were found in the TS-1 pangenome, 2676 genes in the TS-2 pangenome, compared to the 2649 genes in the *M. oralis* pangenome (Additional file 2: Tables S9-S10). Due to the characteristics of ancient DNA (e.g. damage and loss of DNA over time), we only compared genes that were present in TS-1 and TS-2, but not present in *M. oralis*. Uniref90 annotations provided 5 genes that were only present in TS-1. These genes were either involved in metabolic or genetic information processing pathways (Additional file 2: Tables S11-S12). Further comparative sequence analysis show that all 5 genes are also present in *Methanobrevibacter* species others than *M. oralis*, eg. *M. smithii*, *M. millerae* and *M. olleyae*. 21 genes were present in TS-2 but not *M. oralis*, the majority being involved in metabolic pathways. 68.6% (16/21) of the genes showed a high sequence similarity to other *Methanobrevibacter* species, of which more than half (12/21) were most similar to *Methanobrevibacter* sp. YE315. Three genes involved in metal uptake and glycolysis indicate a functional differentiation of TS-2 to *M. oralis*. These include the Molybdate ABC transporter permease protein ModB present in 7 of the 8 TS-2 MAGs, the Nicotianamine synthase-like protein present in all TS-2 MAGs, and Phosphoenolpyruvate synthase that was found in two TS-2 MAGs.

4.5.6 Methyl Coenzyme M Reductase *mcrA* gene in oral *Methanobrevibacter* species

To characterize further functionality in the three calculus *Methanobrevibacter* species, we extended our analysis to one component of the Methyl Coenzyme M Reductase, the key enzyme in methanogenesis which catalyses the final and rate-limiting step in methane biogenesis (Evans et al. 2019). We focused our analysis on the *mcrA* subunit encoding gene, a widely-used marker gene for methanogen classification. Importantly, all ancient *Methanobrevibacter* MAGs in the newly described clades harbour the *mcrA* gene. The *mcrA* gene phylogeny of all calculus samples and selected modern *Methanobrevibacter* isolates shows a highly similar tree topology to the whole genome phylogeny and also separates the two new clades from *M. oralis* and other known *Methanobrevibacter* species. (Additional file 1: Figure S12). Further comparison of the *mcrA* amino acid alignment revealed in all three oral *Methanobrevibacter* clades a high conservation of most amino acid residues that were previously described to interact with CoM, CoB, F430 cofactors and that are part of the substrate cavity wall or have post-translational modifications (Ermler et al. 1997; Borrel et al. 2019). Only in one catalytic site, which is part of the substrate cavity wall, TS-1 encodes a tyrosine and not a phenylalanine, which both *M. oralis* and TS-2 share at that position (Additional file 1: Figure S13).

4.6 Discussion

In this study we analysed ancient calculus samples utilizing *de novo* metagenome assembly techniques to increase the knowledge of the oral microbiome and its diversity over time, with a particular focus on the periodontitis-associated archaeal genus *Methanobrevibacter*. All analysed samples, independently of the site and the time period, contained both ancient endogenous microbial DNA and traces of the human host DNA. Taxonomic characterisation and authentication of the metagenomic data showed that the studied samples mainly consist of oral microbial members and that there are only few signs of external contamination from other environments (e.g. soil). Overall, the calculus samples of all analysed individuals (independently on the sex, age at death or diseases) had a taxonomic composition similar to what has been previously found in other ancient calculus studies (Weyrich et al. 2017; Mann et al. 2018; Velsko et al. 2019) and a major proportion of the discovered microorganisms have been associated with periodontitis (Socransky et al. 1998; Pérez-Chaparro et al. 2014; Elabdeen et al. 2015; Takeuchi et al. 2001; Haririan et al. 2014). These include the red complex members *Treponema denticola* and *Tannerella forsythia* (Socransky et al. 1998) but also species from the genera *Methanobrevibacter* and *Desulfobulbus*, late colonisers of oral plaque and found in deep periodontal pockets (Griffen et al. 2012; Lepp et al. 2004). Many of these taxa were also defining for the ancient calculus microbiome, compared to modern calculus samples which were instead characterised by taxa associated with healthy plaque, eg. *Streptococcus sanguinis*, *Lautropia mirabilis*, *Rothia dentocariosa*, *Neisseria sicca* and *Neisseria elongate* (Donati et al. 2016; Colombo et al. 2009). This may indicate a difference in the level of maturation, and, possibly, disease state, in the calculus samples, between ancient and modern calculus as previously reported (Velsko et al. 2019).

Consistent with previous observations (Mann et al. 2018; Velsko et al. 2019; Ziesemer et al. 2015), we also found that *Methanobrevibacter* members are profoundly abundant in our samples (up to 47.3%), which could be related to the association of *Methanobrevibacter* enrichment and oral disease. In modern patients the abundance of *Methanobrevibacter* spp. have been correlated to the severity of periodontitis (up to 18.5% in severe periodontitis), and *M. oralis* and *Candidatus Methanobrevibacter* sp. has been found in plaque from periodontitis patients as well (Lepp et al. 2004; Huynh et al. 2015; Belay et al. 1988). However, we could not find any such statistically significant correlation between *Methanobrevibacter* abundance and periodontitis in our

dataset, although 75% of the samples showed signs of periodontitis, possibly making a skewed sampling bias. While the correlation was still unclear even when we expanded our dataset by including more previously published ancient and modern calculus samples with reported health status, there were indications of the presence of a non-healthy microbiome in the samples and *Methanobrevibacter* was positively correlated to other species associated with periodontitis (Additional file 2: Table S6. It is thus clear that further studies are needed to shed light on the complexity of the non-healthy oral microbiome and its implications for human health. The role of *Methanobrevibacter* members in human health will be characterized further in the future with an increasing availability of metagenomic samples from ancient and contemporary oral environments.

Of note, the high abundance of *Methanobrevibacter* observed in ancient calculus could also be related with taphonomic processes (Ziesemer et al. 2015). Post-mortem decay could skew the abundance of microorganisms in ancient calculus through overgrowth of certain taxa after the individual has died. Similarly, DNA preservation in ancient samples could be biased between microbial species due to their genomic GC-content (Mann et al. 2018) or differences in cell wall composition (Warinner et al. 2014, 2017) which could both confer stability to DNA after death. Furthermore, as calculus has been created over the lifetime of an individual, the abundance of taxa in ancient calculus most likely does not represent a single snapshot of a microbial community. The plaque microbiome can also change over time in one individual (Utter, Mark Welch, and Borisy 2016), which makes quantitative comparisons between plaque, modern and ancient calculus difficult.

The classical alignment-based methods depending on reference genomes are common in ancient DNA studies (Weyrich et al. 2017; Spyrou et al. 2019; Key et al. 2020), but they hamper the detection of new species, genome rearrangement, and large indels or regions which are no longer present in extant microbial genomes. Abundant archaeal reads in the calculus samples allowed us to perform *de novo* metagenomic assembly, which resulted in the reconstruction of 11 novel ancient *Methanobrevibacter* genomes. Unexpectedly, besides only one genome clustering with a known species, *M. oralis*, the other 10 ancient genomes formed two independent clades (TS-1 and TS-2.) which are clearly distinct to other *Methanobrevibacter* species (Figure 4A). The ancient genomes within each new clade were all closely related to each other, and highly dissimilar from modern genomes within their respective subtree (>15% ANI) (Additional file 1: Figure S8). This suggests the presence of a higher

diversity of human oral *Methanobrevibacter* than has been previously known and confirms that *de novo* assembly is a valuable tool for identifying new microbial species (J. L. Baker et al. 2021; McLean et al. 2015) studying microbial evolution and ecology, even in ancient metagenomics investigations (Brealey et al. 2020; Luhmann, Doerr, and Chauve 2017; Krause-Kyora et al. 2018). Many challenges still come with the assembly of ancient DNA due to its fragmented and damaged nature, one limitation being the risk of incorporation of ancient DNA damage into the assemblies. However, we found no evidence of misincorporation in either the damage patterns retrieved after realigning the reads from each sample to their respective MAG (Additional file 1: Figure S6), or the comparison of genomes reconstructed by alignment-based methods or *de novo* assembly (Additional file 1: Figure S10).

After screening our samples, and the 82 previously published ancient and modern calculus samples, we found sequences from at least one of the three oral *Methanobrevibacter* species (*M. oralis*, TS-1 and TS-2) in 61 samples (59.8%) (Fig 4C-E). We also found a trend in the prevalence of *Methanobrevibacter* pertaining to sample age. *M. oralis* was more prevalent in modern and younger ancient calculus, and the oldest calculus sample containing >50% coverage of *M. oralis* was the sample 2102 dating back to the 7th century. In contrast, TS-1 was more prevalent in older samples and was completely absent in calculus younger than 700 years. Of the 8 samples containing the highest coverage of TS-1, six were over 3000 years old, including the 48,000-year-old Neanderthal sample. TS-2 was found in both prehistoric and younger samples, spanning at least 6000 years. The phylogenetic analysis of TS-2 also showed that the prehistoric genomes within this species fall together in the tree. We did not observe a clear geographical pattern associating with the phylogeny of the three *Methanobrevibacter* species, but they were found in a broad range of geographical regions in Eurasia and Americas (Figure 4B-4E).

The discovery of two new oral *Methanobrevibacter* clades and the coexistence of different clade members in the same ancient calculus raises the question whether there exists functional differentiation that could explain the sharing of the same ecological niche. The annotation of the newly assembled *Methanobrevibacter* genomes revealed a similar gene content between *M. oralis* and TS-1 and TS-2. The two TS-1 MAGs had a lower completeness than the TS-2 MAGs, as well as only two genomes instead of 8, which could account for the fewer found genes in their pangenomes (1628 genes in the TS-1 pangenome, 2676 genes in the TS-2 pangenome). All MAGs belonging to TS-1 and TS-2 contained the *mcrA* gene, one of

the key genes in methanogenesis, indicating that members of the two new clades use the same anaerobic respiratory pathway like *M. oralis*. Besides this major functional similarity, we detected three genes in TS-2 that are absent in *M. oralis* and that could potentially confer additional nutritional advantages and niche adaptation within the oral cavity. The first one, *modB*, encodes the Molybdate ABC transporter permease protein ModB, a part of the molybdopterin biosynthesis pathway and methanogenesis (Lurie-Weinberger, Peeri, and Gophna 2012). Genes in this pathway have been found to be the result of lateral gene transfer in *M. smithii*, potentially as an adaptation to the human gut and increasing metal uptake. Secondly, one of the genes unique to TS-2, and present in all 8 TS-2 MAGs, is coding for a Nicotianamine synthase protein. Archaeal homologues of this gene have been found in other *Methanobrevibacter* species, e.g. *M. ruminatum* (Laffont and Arnoux 2020). Nicotianamine likely also facilitates in metal uptake. Several metals, like iron, are a limited resource in the oral cavity (R. Wang et al. 2012) and these genes could therefore potentially confer a nutritional advantage to TS-2. Thirdly, a putative gene coding for Phosphoenolpyruvate synthase (Pyruvate, water dikinase) was found in two TS-2 MAGs. The gene is present in other archaeal microbes, including *M. smithii* and is involved in carbohydrate metabolism and the Embden–Meyerhof–Parnas (EMP) pathway (Samuel et al. 2007; Tjaden et al. 2006), thus playing a role in autotrophic growth (Eyzaguirre, Jansen, and Fuchs 1982).

4.7 Conclusions

In this study we have shown the potential of using *de novo* metagenomic assembly on ancient DNA sequences to explore the diversity and evolution of oral microbial members. Our analysis unearthed two newly discovered *Methanobrevibacter* species prevalent in calculus from individuals living several thousands of years ago and indicated a possible decline of *Methanobrevibacter* diversity in the human oral microbiome over time. Previous studies have similarly suggested a change in the oral microbiome over time, displaying less diversity observed in modern samples (Adler et al. 2013). This is in line with a 16S rRNA gene analysis based on calculus samples dating back to the 14th to 19th century, showing a decreasing *Methanobrevibacter* diversity but with an increase of *M. oralis* in the modern population (Huynh et al. 2016). However, such decline in the microbiome diversity over time seems not to be specific to the oral environment. Much of the human microbiome diversity today is found in non-westernised populations (Pasolli et al. 2019; New and Brito 2020). For example, the diversity of a common gut microbe, *Prevotella copri*, has been shown to be higher in modern non-westernised populations, as well as in ancient individuals, compared to westernised populations, indicating a shift in diet is a likely factor of the decline in human microbiome diversity (Tett et al. 2019). Conversely, some human pathogens have been shown to become more specialised and diverse in the beginning of the Neolithisation process (Key et al. 2020). Changing lifestyles and diets over the centuries and the modern-day medical usages (e.g. antibiotics) might play a considerable role in depleting human microbiome diversity (Adler et al. 2013; Ferrer et al. 2017), and hence it may help explain the observed loss of *Methanobrevibacter* diversity in the recent past. Further studies with fast-growing metagenomic data from both ancient and contemporary populations will certainly enhance our understanding of *Methanobrevibacter* members whose diversity has not been fully unravelled yet.

4.8 Declarations

Ethics approval and consent to participate

Non-applicable

Consent for publication

Non-applicable

Availability of data and material

The datasets generated and analysed during the current study are available in the European Nucleotide Archive repository under accession no. PRJEB43389

Competing interests

The authors declare that they have no competing interests.

Funding

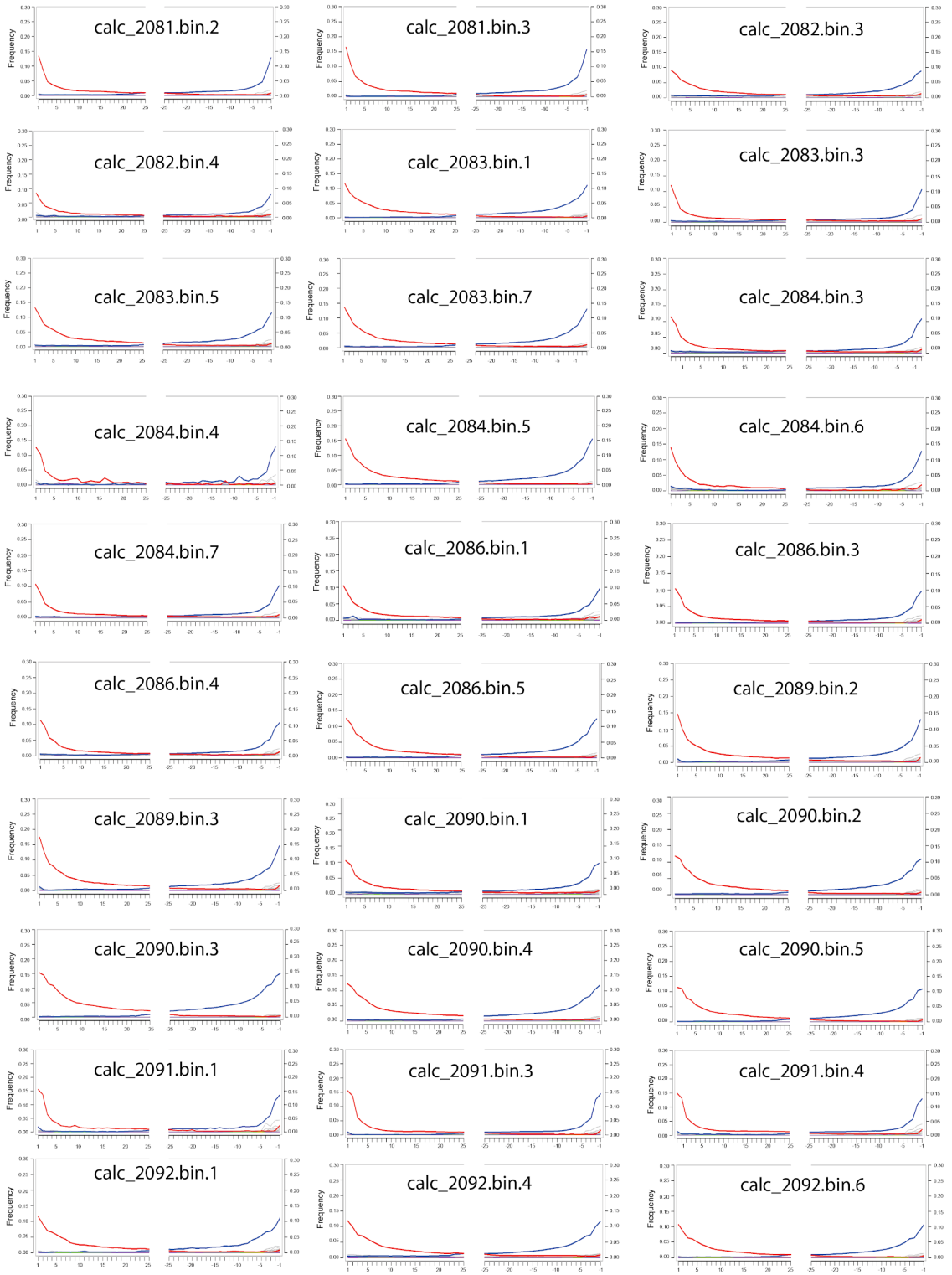
This study was supported by the Programma Ricerca Budget prestazioni Eurac 2017 of the Province of Bolzano, Italy.

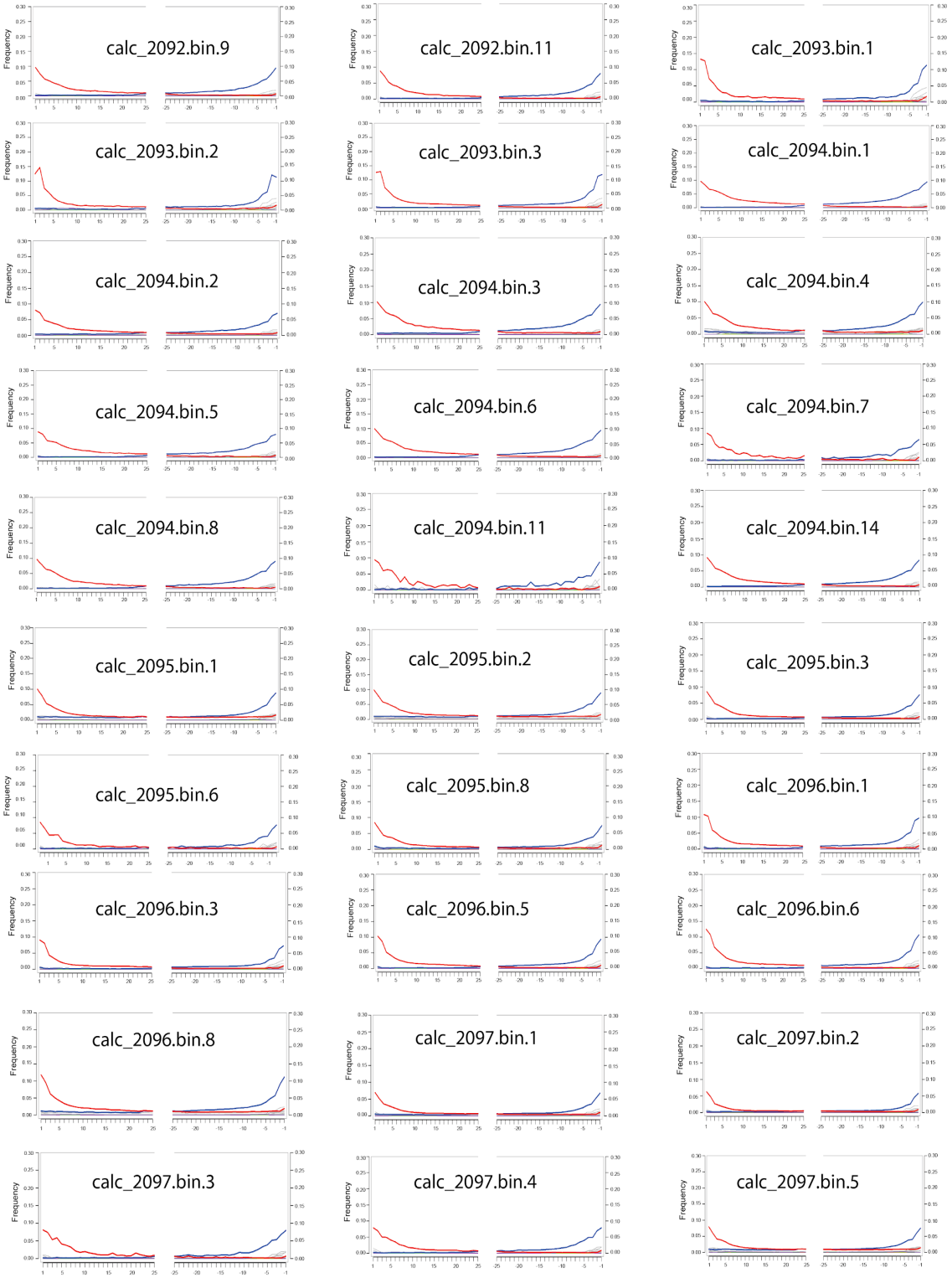
Authors' contributions

LG and KDH participated in the design of the study, data analysis and wrote the manuscript. AT participated in data analysis and contributed to manuscript preparation. AP conducted data collection and contributed to manuscript preparation. NO'S conducted data collection. PM and ORS contributed to manuscript preparation. NS and AZ participated in the design of the study and contributed to manuscript preparation. FM participated in the design and coordination of the study, data analysis and contributed to manuscript preparation. All authors read and approved the final manuscript.

4.9 Additional files

4.9.1 Supplemental figures




calc_2092.bin.9
calc_2092.bin.11
calc_2093.bin.1
calc_2093.bin.2
calc_2093.bin.3
calc_2094.bin.1
calc_2094.bin.2
calc_2094.bin.3
calc_2094.bin.4
calc_2094.bin.5
calc_2094.bin.6
calc_2094.bin.7
calc_2094.bin.8
calc_2094.bin.11
calc_2094.bin.14
calc_2095.bin.1
calc_2095.bin.2
calc_2095.bin.3
calc_2095.bin.6
calc_2095.bin.8
calc_2096.bin.1
calc_2096.bin.3
calc_2096.bin.5
calc_2096.bin.6
calc_2096.bin.8
calc_2097.bin.1
calc_2097.bin.2
calc_2097.bin.3
calc_2097.bin.4
calc_2097.bin.5

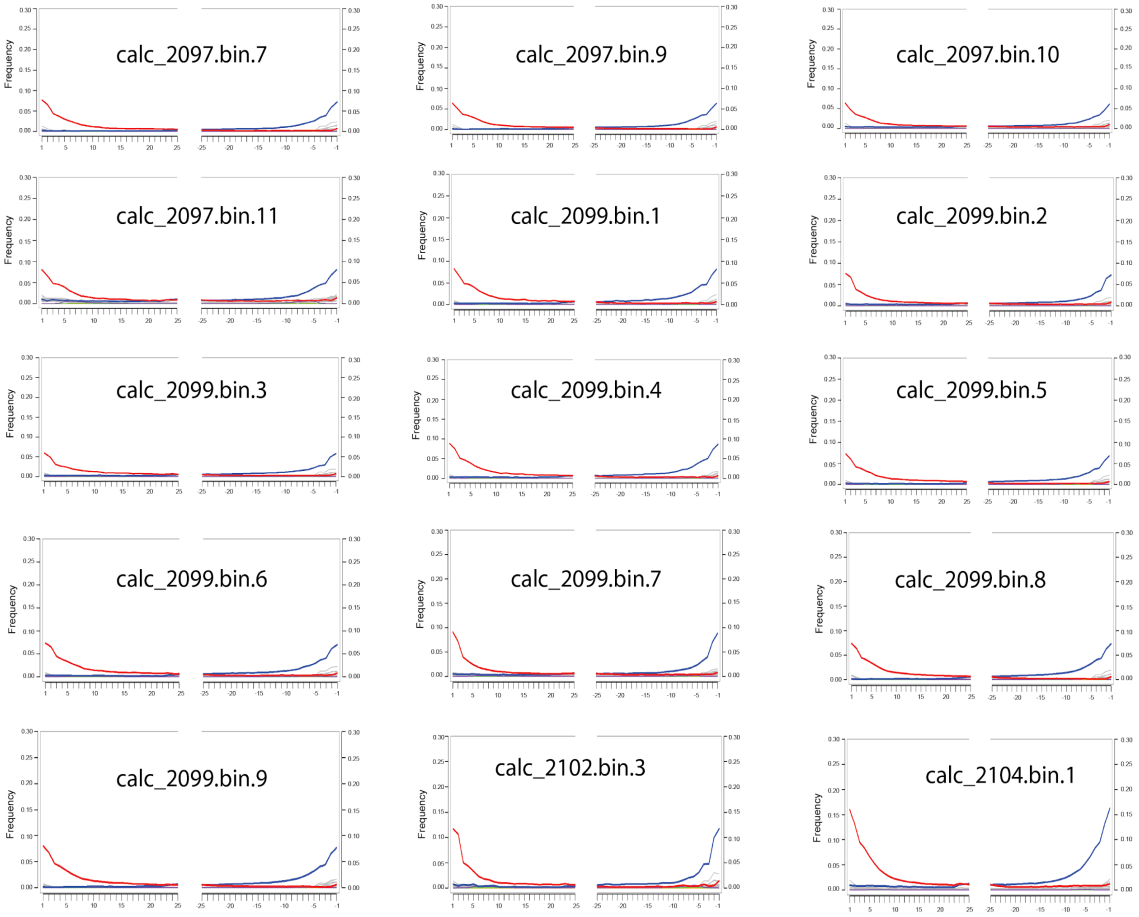


Figure S1. Authentication of endogenous origin for 76 newly reconstructed microbial genomes.

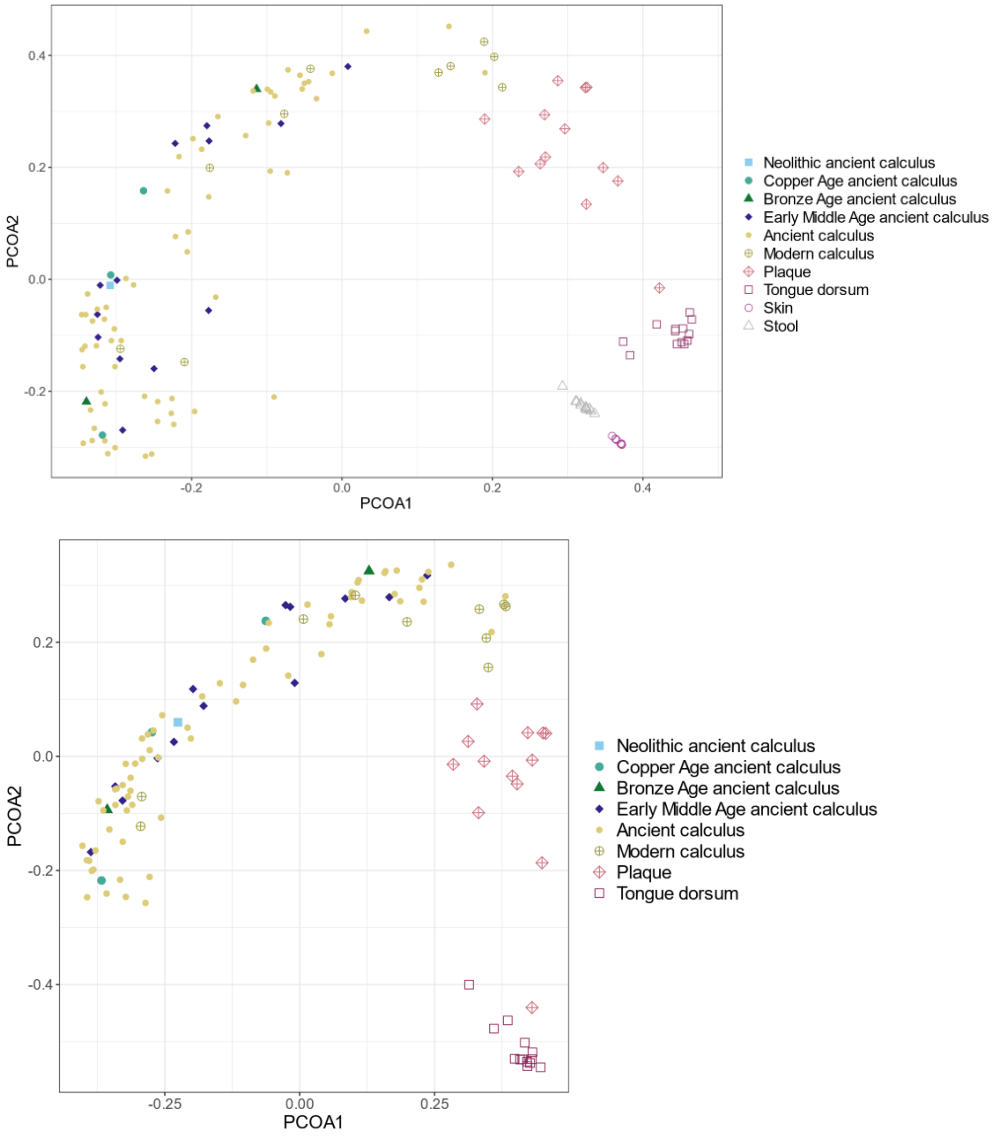


Figure S2. A) Species-level principal coordinate analysis of all comparative datasets used in this study (ancient calculus, modern calculus, plaque, tongue dorsum, skin, stool and soil). B) Species-level principal coordinate analysis with oral datasets and filtered to only oral species. The list of 235 oral taxa used for this analysis were determined by identification of taxa present in all oral datasets from the HMP (buccal mucosa, hard palate, keratinized gingiva, palatine tonsils, saliva, subgingival plaque, supragingival plaque, throat, tongue dorsum), where the taxa needed to be present at > 0.1% in at least one oral sample as well as present at any abundance in at least 10% of the total number of oral samples to be included.

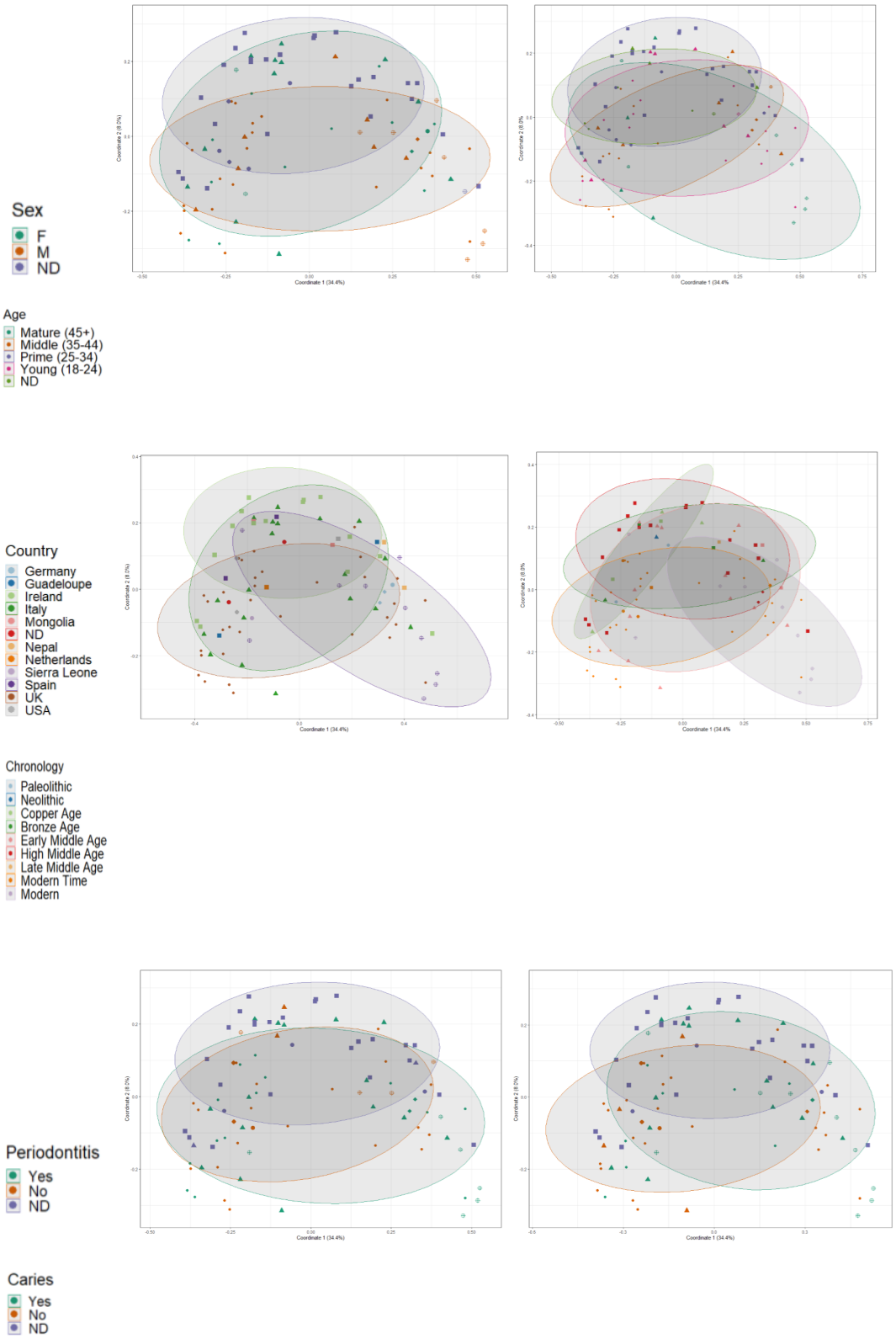


Figure S3. Species-level Principal coordinate analysis plots based on Bray-curtis distances, visualising normal confidence ellipses (levels=0.75) for each group of metadata: A) Sex B) Age at death C) Country D) Time period of individual (based on archaeological periods from Festi et al. 2014). E) Occurrence of periodontitis. F)

Occurrence of caries.

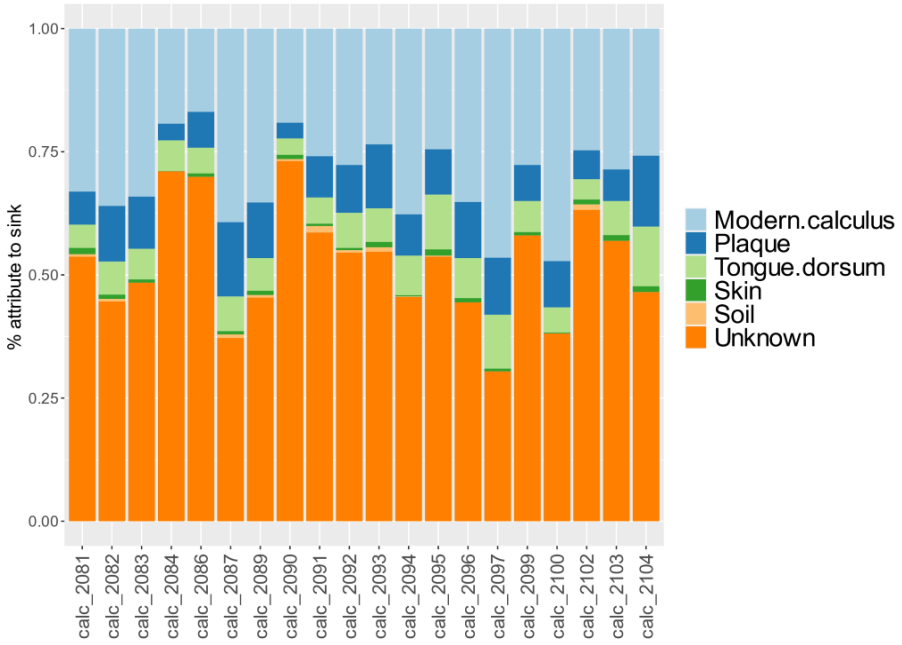


Figure S4. Sourcetracker analysis for all Trentino-South Tyrolean calculus samples indicating the proportions of each source dataset based on species-level taxonomic abundances.

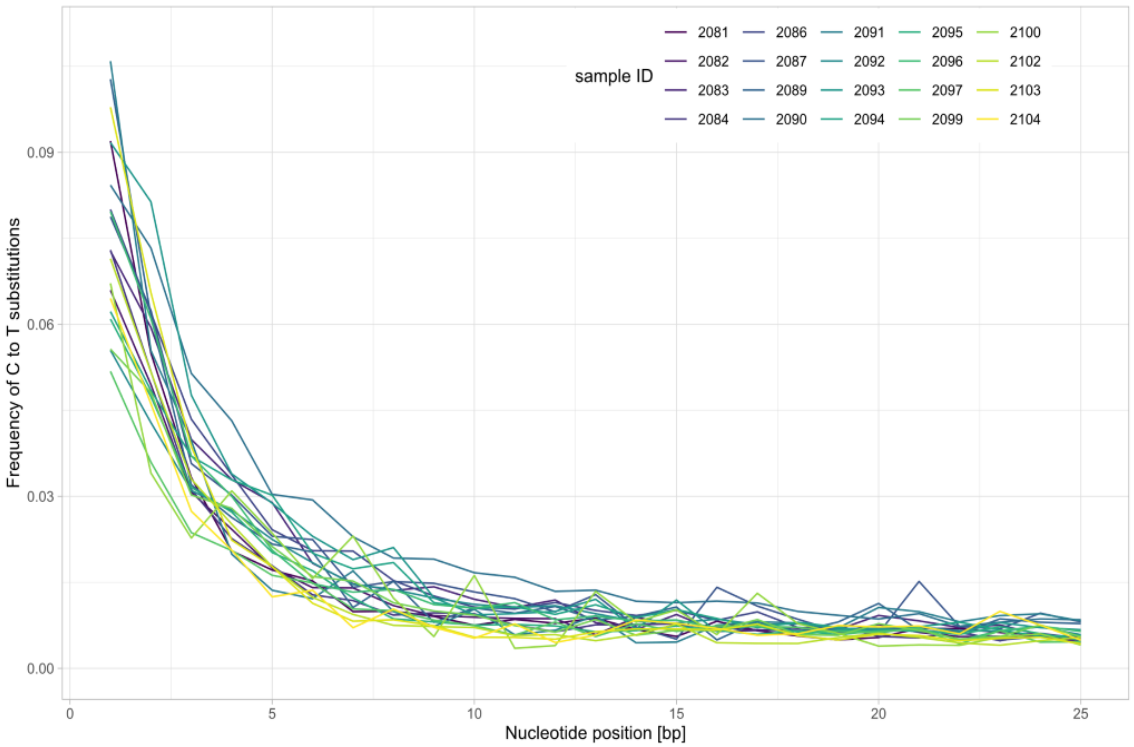


Figure S5. Frequency of C to T transitions in reads mapped to the highest abundant taxa present in the Trentino-South Tyrolean calculus dataset, *Desulfobulbus* oral taxon 041 (GCA_000349345.1), estimated by MapDamage.

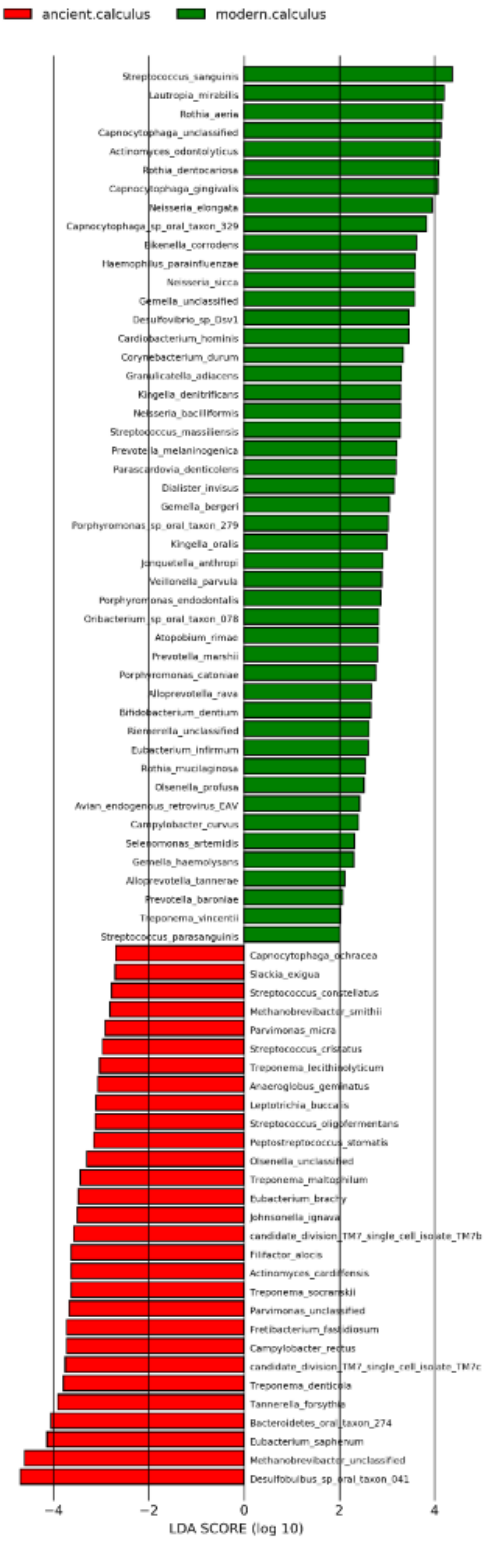


Figure S6. Species level linear discriminant analysis (LDA) effect size (LEfSe) between ancient and modern calculus.

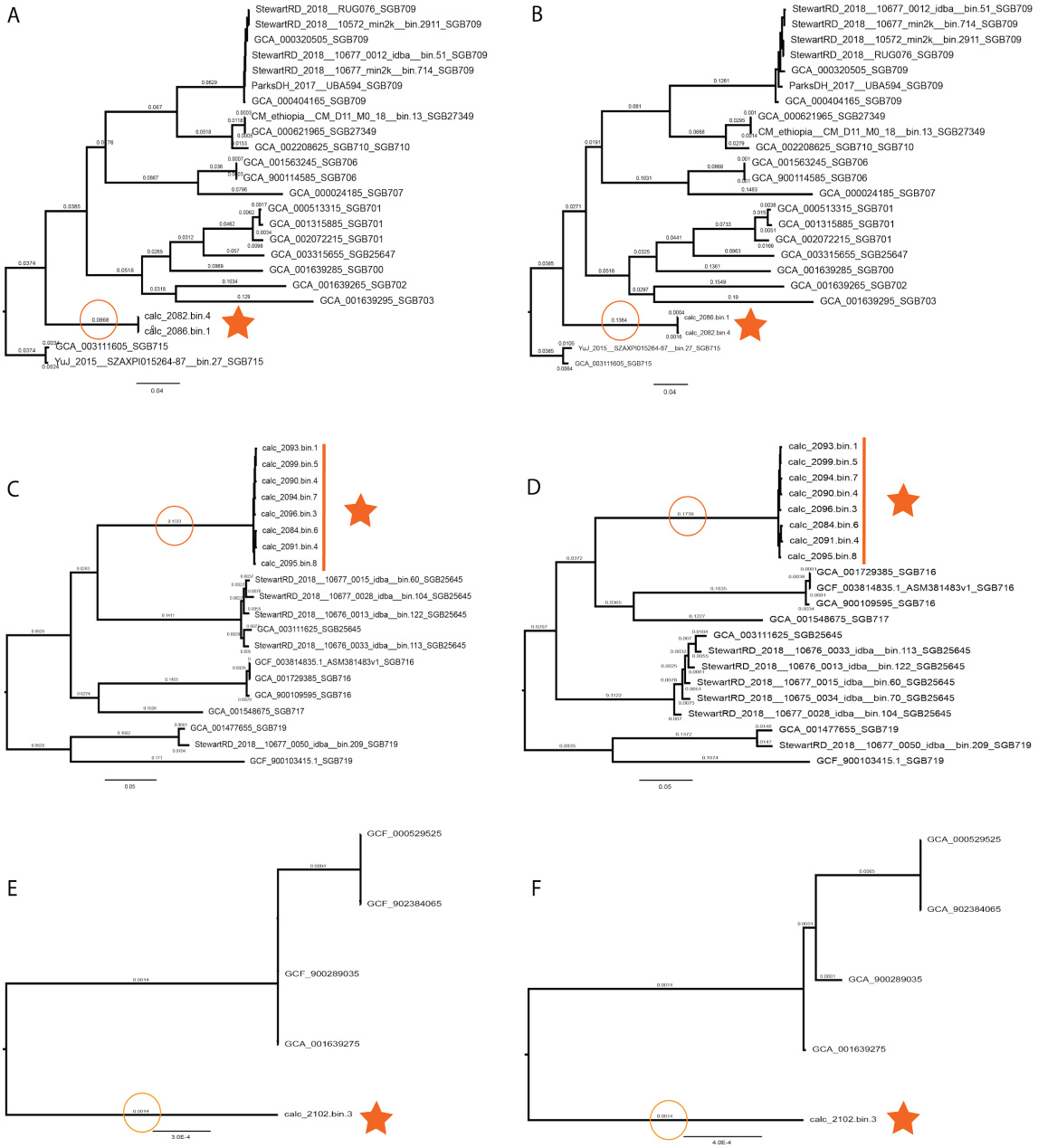


Figure S7. Resolution-improved phylogeny of subtrees based on the core and whole genome. (A), (C), and (E) indicate subtree1, subtree2, and subtree3, based on core gene alignment generated by MAFFT(Katoh and Standley 2013), respectively. (B), (D), and (F) indicate subtree1, subtree2, and subtree3, based on the whole genome alignment, respectively (see Method). The orange star highlights ancient genomes and the red circle emphasizes the long branch length.

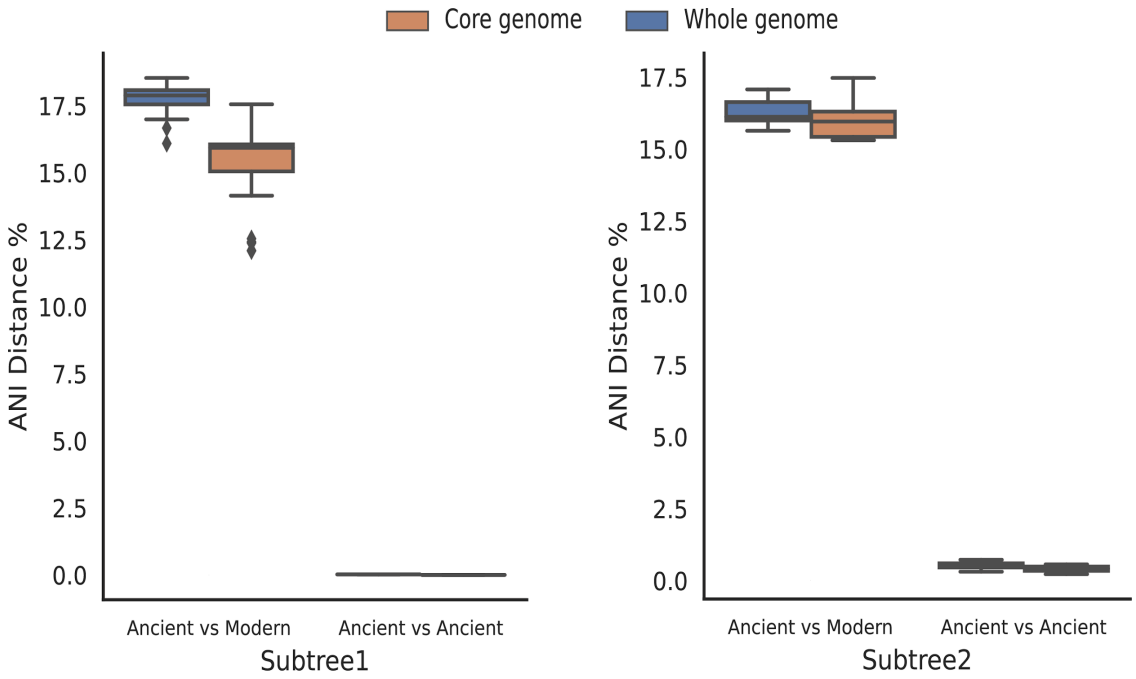


Figure S8. Genetic distance pairwise distances within ancient genomes and between ancient and modern genomes in subtree 1 (TS-1) and subtree 2 (TS-2) respectively, measured by ANI.

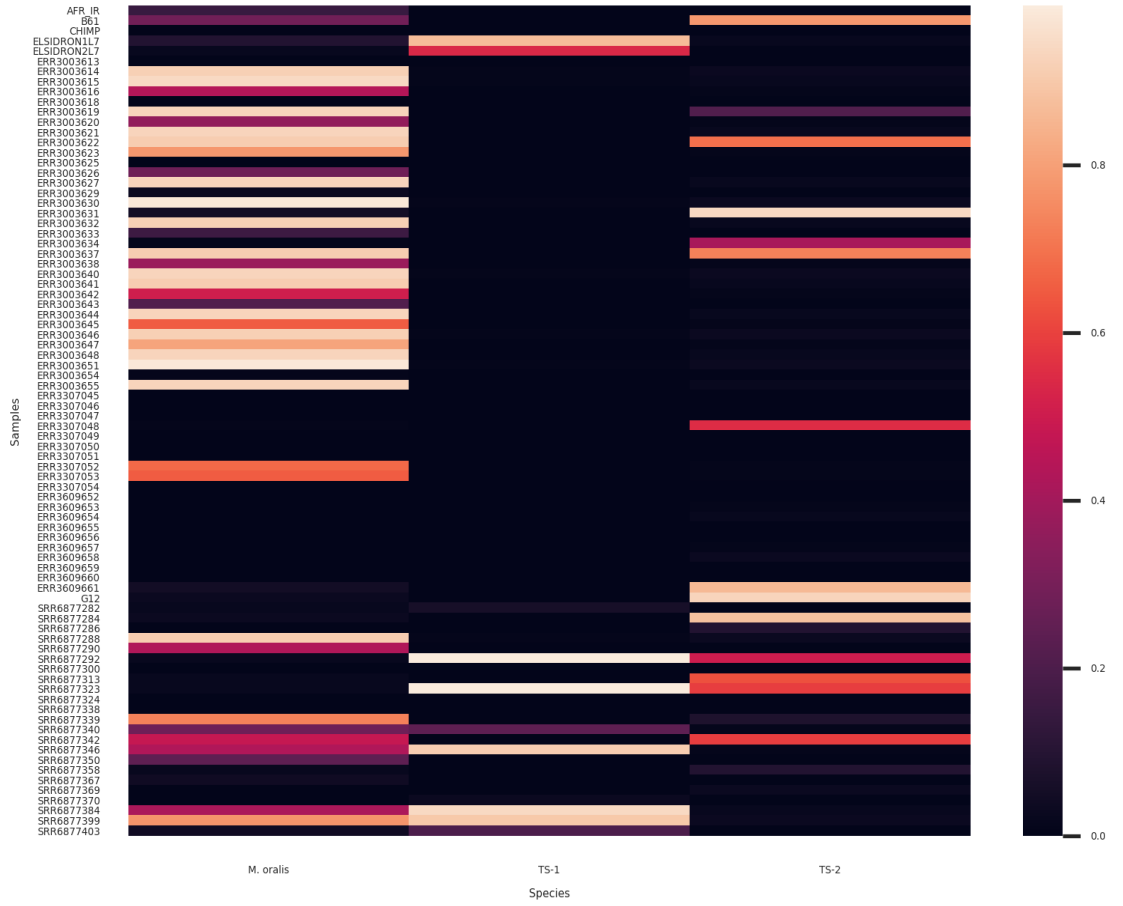


Figure S9. Alignment-based assessment for species coverage across publicly available metagenomic samples. Reference genome GCA_001639275 represents *M. oralis*, newly reconstructed genome calc_2084.bin.1 represents *TS-1*, and calc_2094.bin.7 was used for another species *TS-2*.

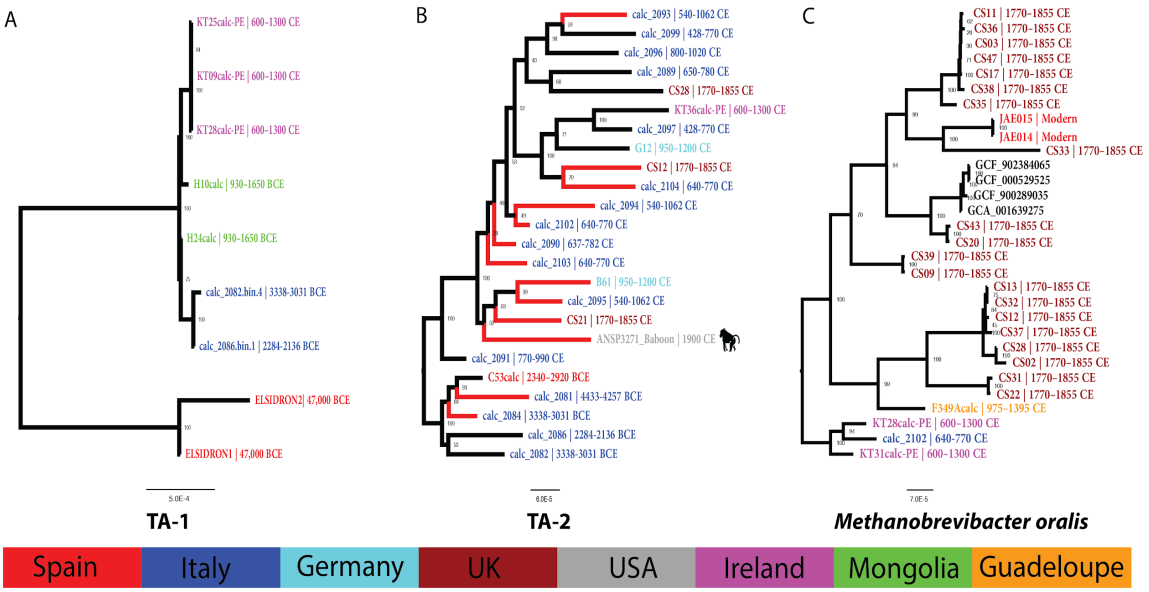


Figure S10. Strain-resolved phylogeny of newly discovered *Methanobrevibacter* species TS-1 (A), TS-2 (B) and *Methanobrevibacter oralis* (C), reconstructed using draft ancient genomes from alignment-based approach. Branches variable in phylogenetic position from those in Figure 4D are marked as red.

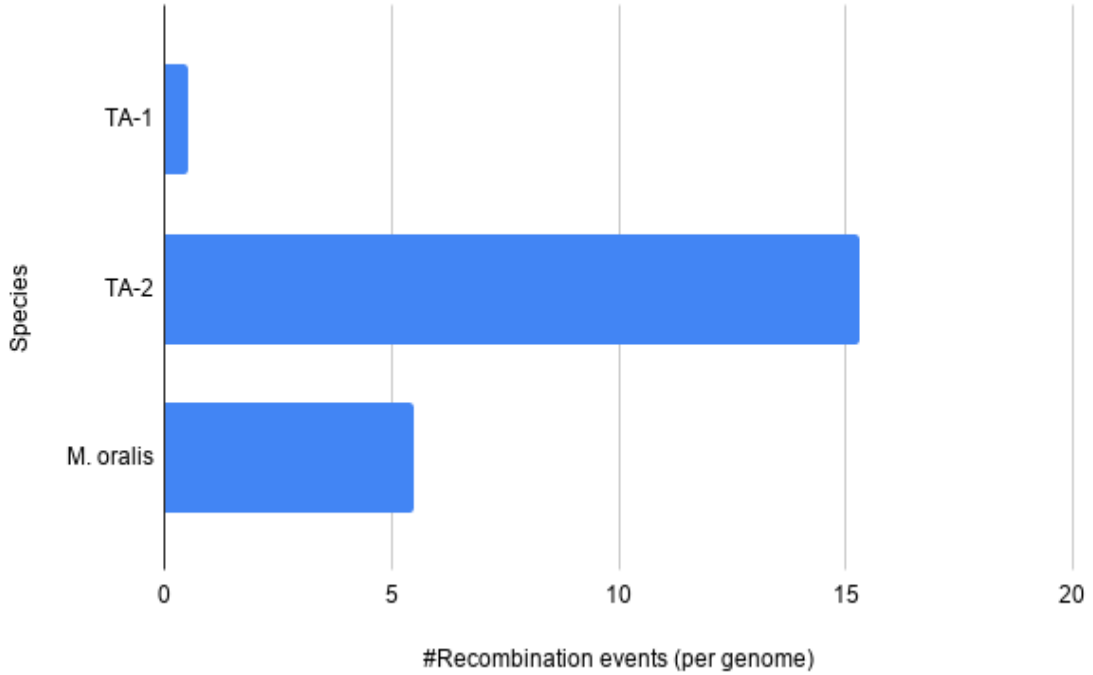
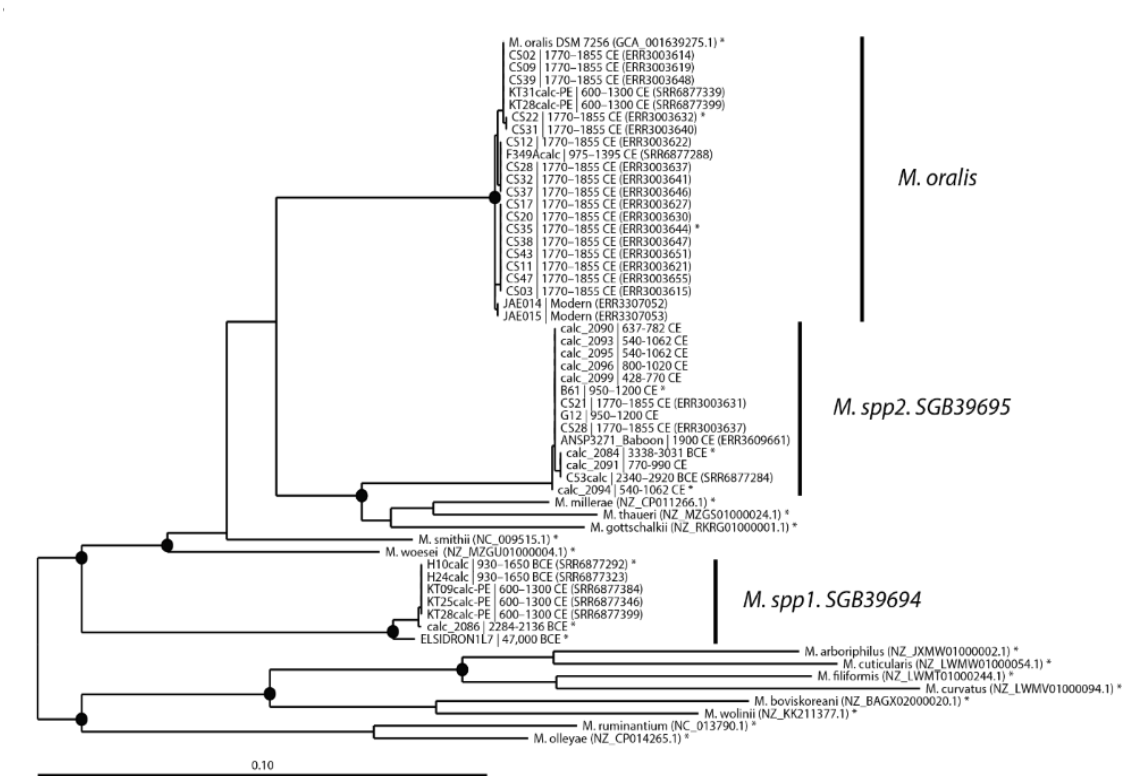


Figure S11. The number of recombination events (normalized by the total number of genomes in the alignment) detected in TS-1, TS-2 and *M. oralis*, using ClonalFrameML (Didelot and Wilson 2015).



Supplemental Figure S12. Maximum likelihood tree based on a *mcrA* nucleic acid sequence alignment. Sequences marked with an asterisk indicate *mcrA* sequences used for amino acid analysis. Black circles symbolize parsimony and neighbour joining bootstrap support (>60%) based on 100 and 1000 iterations, respectively. The scale bar depicts 0.1 substitutions per residue.

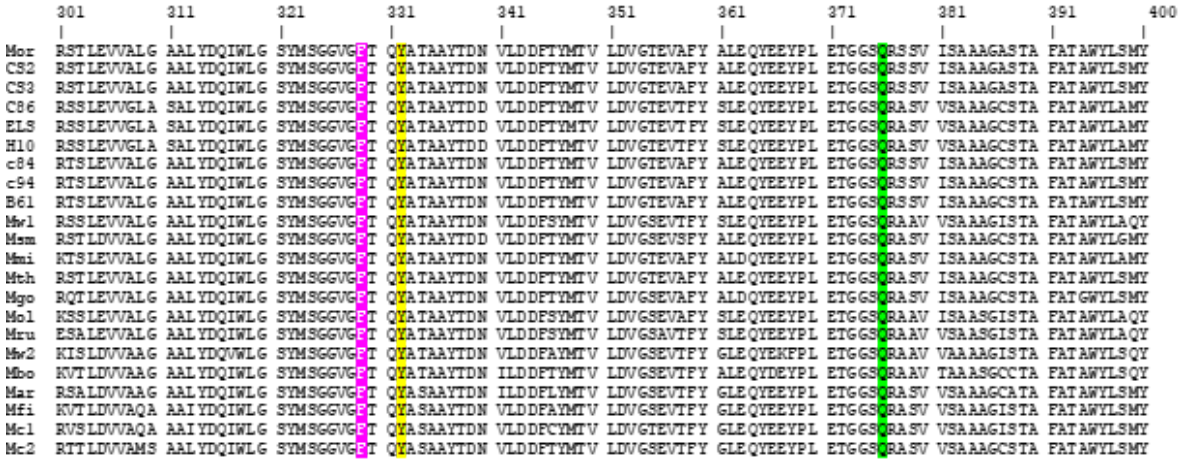
Methyl-coenzyme M reductase I subunit alpha

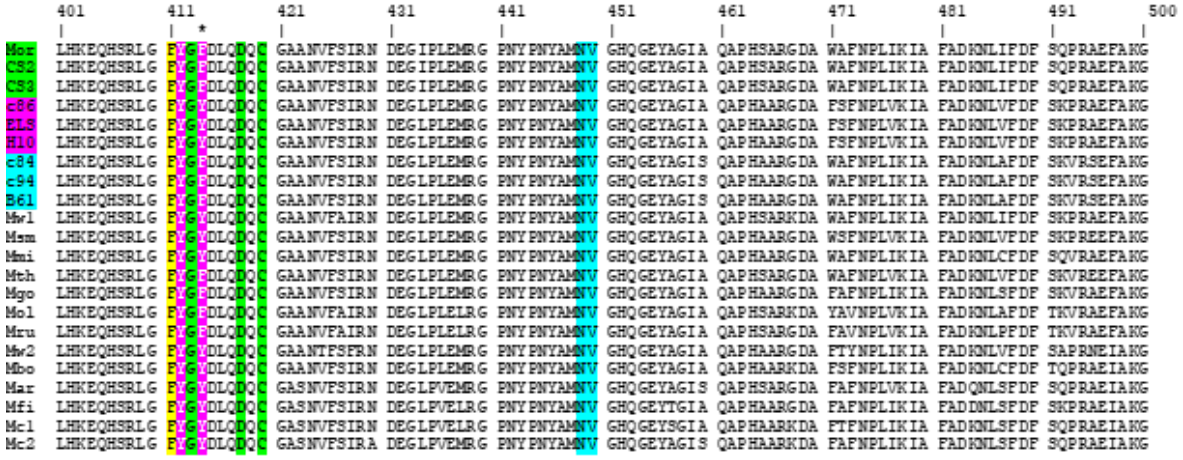
```
Name: Mor          Len:   GCA_001639275.1, M. oralis DSM 7256
Name: CS2          Len:   CS22 ERR3003632 Moralis2
Name: CS3          Len:   CS35 ERR3003644 Moralis2
Name: c86          Len:   SGB39694_calc_2086
Name: ELS          Len:   ELSIDRONIL7 SGB39694
Name: H10          Len:   H10calc SRR6877292 SGB39694
Name: c84          Len:   SGB39695_calc_2084
Name: c94          Len:   SGB39695_calc_2094
Name: B61          Len:   B61 SGB39695
Name: Mwl          Len:   NZ_MZGU01000004.1 M. woesei strain DSM 11979
Name: Msm          Len:   NC_009515.1 M. smithii ATCC 35061
Name: Mmi          Len:   NZ_CP011266.1 M. millerae strain SM9
Name: Mth          Len:   NZ_MZGS01000024.1 M. thaueri strain DSM 11995
Name: Mgo          Len:   NZ_RKRG01000001.1 M. gottschalkii DSM 11977
Name: Mol          Len:   NZ_CP014265.1 M. olleyae strain YLM1
Name: Mru          Len:   NC_013790.1 M. ruminantium M1
Name: Mw2          Len:   NZ_KK211377.1 M. wolinii SH
Name: Mbo          Len:   NZ_BAGX02000020.1 M. boviskoreani JH1
Name: Mar          Len:   NZ_JXMW01000002.1 M. arboriphilus JCM 13429
Name: Mfi          Len:   NZ_LWMT01000244.1 M. filiformis strain DSM 11501
Name: Mc1          Len:   NZ_LWMV01000094.1 M. curvatus strain DSM 11111
Name: Mc2          Len:   NZ_LWMW01000054.1 M. cuticularis strain DSM 11139
```

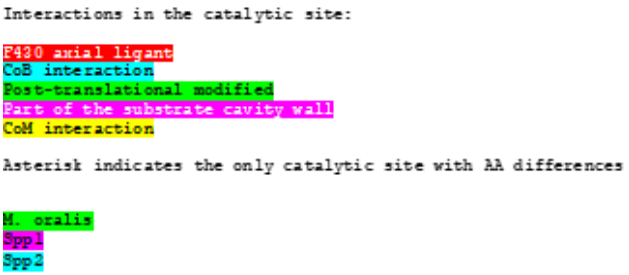
```
     1          11         21         31         41         51         61         71         81         91        100
     |          |          |          |          |          |          |          |          |          |          |
Mor  MADKKFLDAM KQKFSEDPTE KRTQFYNMGG WKQSERKTAF VNEGKKIAEE RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NSAIQQAWDD
CS2  MADKKFLDAM KQKFSEDPTE KRTQFYNMGG WKQSERKTAF VNEGKKIAEE RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NSAIQQAWDD
CS3  MADKKFLDAM KQKFSEDPTE KRTQFYNMGG WKQSERKTAF VNEGKKIAEE RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NSAIQQAWDD
SGB  MADKRFLEAM KKKFKEDPTD KRTSFYNMDG WKQSERKCAF VKEGEEIAKE RGIPMYNPDI GNPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
ELS  MADKRFLEAM KKKFKEDPTD KRTSFYNMDG WKQSERKCAF VKEGEEIAKE RGIPMYNPDI GNPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
H10  MADKRFLEAM KKKFKEDPTD KRTSFYNMDG WKQSERKCAF VKEGEEIAKE RGIPMYNPDI GNPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
c84  MADKKFLDAM KQKFSEDPTE KTTQFYNMGG WTQSERKTAF VNEGKEIAEK RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
c94  MADKKFLDAM KQKFSEDPTE KTTQFYNMGG WTQSERKTAF VNEGKEIAEK RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
B61  MADKKFLDAM KQKFSEDPTE KTTQFYNMGG WTQSERKTAF VNEGKEIAEK RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
Mwl  MADKKFLDAM KQKFSEDPTD DRTTFYNMGG WKQSERKSAF VKEGQEIAEK RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NSAIQQAWDD
Msm  MADKKFLDAM KNKFSEGFTD KRTTFYNMGG WKQSERKSAF VKEGKEIAEK RGIPMYNPDI GTPLGQRALM SYQLSTTDTY VEGDDLHFIN NAAIQQAWDD
Mmi  MADKKFLDAM KQKFSEDPTE KTTQFYNMGG WTQSERKTAF VNEGKEIAEK RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
Mth  MADKKFLDAM KKKFTEDPTE KRTQFYNMGG WTQSERKTAF VNEGKEIAAK RGIPMYNPDI GSPLGQRALM SYQLSTTDTF VEGDDLHYIN NAAIQQAWDD
Mgo  MADKKFLNAM KQKFSEDPTE KSTQFYNMGG WTQSERKTAF VNEGKEIAEK RGIPMYNPDI GTPLGQRALM SYQLSTTDTF VEGDDLHFIN NAAIQQAWDD
Mol  MADKKFLDAM TKKFKEAPDE RTTTFYNMGG WTQSERKTEF VNEGKSIAEK RGIPMYNPDI GNPLGQRALM SYQLSGTDTF VEGDDLHFIN NAAMQQAWDD
Mru  MADKKFLDAM TKKFKEAPEE KTTTFYNMGG WTQSERKTEF VNEGKAIAEA RGIPMYNPDI GNPLGQRALM SYQLSGTDTF VEGDDLHFIN NAAMQQAWDD
Mw2  MADKKFIDAL NKKFKESPED KTTTFYNLGG WRQSERKTEF YEAGKEIAEK RGIPQFNPDV GSPLGQRTLM PYQLSGTDTF VEGDDLHFIN NAAIQQMWDD
Mbo  MADKKFYEAL NKKFKESPED KTTTFYNLGG WKQSERKTQF YEAGKEIAEE RGIPMYNPDI GSPLGQRTLM PYQLSTTDTF VEGDDLHFIN NAAIQQLWDD
Mar  MADKKFIEAL NKKFKEDPEE KTTTFYNLGG WKQSERKTEF ANAGKEIAEK RGIPQYNPDV GSPLGQRALM PYQVSTTDTY VEGDDFHFVN NAAIQQMWDD
Mfi  MADKKFIQAL DKKFKEAPED KTTTFYNLGG WKQSERKTEF YNAGKEIAEK RGIPQYNPDV GSPLGQRTLM PYQVSTTDTY VEGDDFHFVN NAAIQQMWDD
Mc1  MADKKFIDAL NKKFKEDPTE KSTTFYNLGG WTQSERKTEF ANAGKEIAEK RGIPQYNPDV GSPLGQRALM PYQVSTTDTY VEGDDFHFVN NAAIQQMWDD
Mc2  MANKKFIEAL NKKFKESPEE KTTTFYNLGG WKQSERKSEF AKEGKEIATK RGIPQYNPDV GSPLGQRALM PYQVSTTETF VEGDDFHFVN NAAIQQMWDD
```

```
     101        111        121        131        141        151        161        171        181        191        200
     |          |          |          |          |          |          |          |          |          |          |
Mor  IRKTVIVGLN TAHNVLEKRL GIEVTPETIT HYLETVNHAM PGAAVVQPHM VETDPLVVSD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
CS2  IRKTVIVGLN TAHNVLEKRL GIEVTPETIT HYLETVNHAM PGAAVVQPHM VETDPLVVSD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
CS3  IRKTVIVGLN TAHNVLEKRL GIEVTPETIT HYLETVNHAM PGAAVVQPHM VETDPLVVSD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
C86  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT EYLEIVNHAM PGAAVVQPHM VETDPLVVSD SYVKVFTGDD ELADEIDSAF VLDINKEFND EQANALKNEV
ELS  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT EYLEIVNHAM PGAAVVQPHM VETDPLVVSD SYVKVFTGDD ELADEIDSAF VLDINKEFND EQANALKNEV
H10  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT EYLEIVNHAM PGAAVVQPHM VETDPLVVSD SYVKVFTGDD ELADEIDSAF VLDINKEFND EQANALKNEV
c84  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT NYLEIVNHAM PGAAVVQPHM VETDPLVVQD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
c94  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT NYLEIVNHAM PGAAVVQPHM VETDPLVVQD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
B61  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT NYLEIVNHAM PGAAVVQPHM VETDPLVVQD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
Mwl  IRKTVIVGLN TAHNVLEKRL GIEVTPETIT EYLETVNHAM PGAAVVQPHM VETDPLVVSD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
Msm  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT EYLETVNHAM PGAAVVQPHM VETDPLVVQD SYVKVFTGDD ELADEIDSAF VLDINKEFNE EQAAALKEEV
Mmi  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT NYLETVNHAM PGAAVVQPHM VETDPLVVQD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
Mth  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT NYLETVNHAM PGAAVVQPHM VETDPLVVQD SYVKVFTGDD ELADEIDSAF VLDINKEFPE EQAEALKAEV
Mgo  IRRTVIVGLN TAHNVLEKRL GIEVTPETIT NYLETVNHAM PGAAVVQPHM VETDPLLVAD SYVKVFTGDD ELADEIDSAF VLDINKEFNE EQAEALKAEV
Mol  IRRTVIVGLN TAHNVLEKRL GMEVTPETIT NYLEVVNHAM PGAAVIQPHM VETNPLLVED SYVKIFTGDD DLADEIDAAF VLDINKEFPE EQAEVLKSEV
Mru  IRKTVIVGLN TAHNVLEKRL GMEVTPETIT NYLEVVNHAM PGAAAVQPHM VETNPLLVDD SYVKVFTGDD DLAAEIDPAF VLDINKEFPE EQAEALKAEV
Mw2  IRRTVIVGLN TAHNVLERRL GMEVTPETIT RYLETLNHAM PGGAVVQPHM VEINPYLAAD SYVKVFTGDD ELADEIDKCY VLDINKEFPE EQAEALKAEV
Mbo  IRRTVLVGLN TAHNVLQRRL GMEVTPETIT NYLETVNHAM PGAAVVQPHM VETNPYLVSD SYVKIFTGDD ELADEIDQCY VLDINKEFPD DQAAQLKAEV
Mar  IRRTVIVGLN TAHNVLEKRL GMEVTPETIT EYLETVNHAM PGAAVVQPHM VETNPSLVAD SYVKVFTGDD ELADEIDSAF VLDINKEFPE DQAEALKAEV
Mfi  IRRTVVVGLN TAHTVLEKRL GIEVTPETIT NYLETVNHAM PGAAVVQPHM VETNPLLVAD SYVKVFTGDD ELADEIDSAF VLDINKEFPE DQAAVLKEEV
Mc1  IRRTVIVGLN TAHNVIEKRL GMEVTPETIT HYLETVNHAM PGAAVVQPHM VETNPLLVAD SYVKIFTGDD ELADEIDKAY VLDINKEFNE EQAAALKAEV
Mc2  IRRTVIVGLN TAHTVLEKRL GIEVTPETIT EYLETVNHAM PGAAVVQPHM VETNPALVAD SYVKIFTGDD EVADEIDSAF VLDINKEFSK EQAEALKAEV
```


201 211 221 231 241 251 261 271 281 291 300
Mor GGSVWQAVRI PAIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENELGG VPFGFMADIC QGTRVHDDPV
CS2 GGSVWQAVRI PAIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENELGG VPFGFMADIC QGTRVHDDPV
CS3 GGSVWQAVRI PAIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENELGG VPFGFMADIC QGTRVHDDPV
C86 GGSVWQVVRI PSVVGRICDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VHMGTYLPVR RARAENELGG VPFGFMADIC QASRVNDDPV
ELS GGSVWQVVRI PSVVGRICDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VHMGTYLPVR RARAENELGG VPFGFMADIC QASRVNDDPV
H10 GGSVWQVVRI PSVVGRICDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VHMGTYLPVR RARAENELGG VPFGFMADIC QASRVNDDPV
c84 GGAVWQAVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VHMGTYLPVR RARAENELGG VPFGFMADIC QASRVNDDPV
c94 GGAVWQAVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VHMGTYLPVR RARAENELGG VPFGFMADIC QGSRVYDDPV
B61 GGAVWQAVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VHMGTYLPVR RARAENELGG VPFGFMADIC QGSRVYDDPV
Mwl GGSIWQAVRI PGIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPIR RARAENELGG VPFGFMADIC QATRLYDDPV
Msm GGSVWQAVRI PAIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENELGG VPFGFMADIC QSSRVNDDPV
Mmi GGAVWQAVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VHMGTYLPVR RARAENELGG VPFGFMADIC QASRVSDDPV
Mth GGAIWQAVRI PSIVGRVCDG GNTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENELGG VPFGFMADIC QGSRAYDDPV
Mgo GSAVWQAVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV VQMGTYLPVR RARAANELGG VPFGFMADIC QGSRVYDDPV
Mol GGAIWQIVRI PSVVGRVCDG GTTSRWSAMQ IGMSMISAYG QCAGEGATGD FAYASKHAEV IGMGTYLPIR RARADNELGG VPFGFMADIC QATRVSDDPV
Mru GGAIWQIVRV PSVVGRVCDG GTTSRWSAMQ IGMSMISAYG QCAGEGATGD FAYASKHAEV IGMGTYLPIR RARAGNELGG VPFGFMADIC QATRVTDDPV
Mw2 GDKIWQAVRI PSVVGRVCDG GVTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IQMGTPLPQR RARAANEPGG IPFGYMADIV QSSRVNDDPV
Mbo GDKIWQVVRV PSLVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAFATKHAEV IQMGAPLPQR RARAANEPGG VPFGYIADIV QSSRVNDDPV
Mar GDKVWQAVRI PTIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENEPGG IAFGFLADMV QSTRVNEDPV
Mfi GDKIWQVVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISSYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENEPGG IPFGVMADIN QSSRVNDDPV
Mc1 GDNIWQVVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENEPGG VAFGYMADIC QSSRINDDPV
Mc2 GDKIWQVVRI PSIVGRVCDG GTTSRWSAMQ IGMSMISAYN QCAGEGATGD FAYASKHAEV IHMGTYLPVR RARAENEPGG IPFGYMADIC QSSRVNDDPV


301 311 321 331 341 351 361 371 381 391 400
Mor RSTLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALEQYEEYPL ETGGSQRSSV ISAAAGASTA FATAWYLSMY
CS2 RSTLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALEQYEEYPL ETGGSQRSSV ISAAAGASTA FATAWYLSMY
CS3 RSTLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALEQYEEYPL ETGGSQRSSV ISAAAGASTA FATAWYLSMY
C86 RSSLEVVGLA SALYDQIWLG SYMSGGVGFT QYATAAYTDD VLDDFTYMTV LDVGTEVTFY SLEQYEEYPL ETGGSQRASV VSAAAGCSTA FATAWYLAMY
ELS RSSLEVVGLA SALYDQIWLG SYMSGGVGFT QYATAAYTDD VLDDFTYMTV LDVGTEVTFY SLEQYEEYPL ETGGSQRASV VSAAAGCSTA FATAWYLAMY
H10 RSSLEVVGLA SALYDQIWLG SYMSGGVGFT QYATAAYTDD VLDDFTYMTV LDVGTEVTFY SLEQYEEYPL ETGGSQRASV VSAAAGCSTA FATAWYLAMY
c84 RTSLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALEQYEEYPL ETGGSQRSSV ISAAAGCSTA FATAWYLSMY
c94 RTSLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALEQYEEYPL ETGGSQRSSV ISAAAGCSTA FATAWYLSMY
B61 RTSLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALEQYEEYPL ETGGSQRSSV ISAAAGCSTA FATAWYLSMY
Mwl RSSLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFSYMTV LDVGSEVTFY SLEQYEEYPL ETGGSQRAAV VSAAAGISTA FATAWYLAQY
Msm RSTLDVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDD VLDDFTYMTV LDVGSEVSFY ALEQYEEYPL ETGGSQRASV ISAAAGCSTA FATAWYLGMY
Mmi KTSLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALDQYEEYPL ETGGSQRASV ISAAAGCSTA FATAWYLAMY
Mth RSTLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGTEVAFY ALEQYEEYPL ETGGSQRASV ISAAAGCSTA FATAWYLSMY
Mgo RQTLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFTYMTV LDVGSEVAFY ALDQYEEYPL ETGGSQRASV ISAAAGCSTA FATGWYLSMY
Mol KSSLEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFSYMTV LDVGSEVAFY SLEQYEEYPL ETGGSQRAAV ISAASGISTA FATAWYLAQY
Mru ESALEVVALG AALYDQIWLG SYMSGGVGFT QYATAAYTDN VLDDFSYMTV LDVGSAVTFY SLEQYEEYPL ETGGSQRAAV VSAASGISTA FATAWYLAQY
Mw2 KISLDVVAAG AALYDQVWLG SYMSGGVGFT QYATAAYTDN VLDDFAYMTV LDVGSEVTFY GLEQYEKFPL ETGGSQRAAV VAAAAGISTA FATAWYLSQY
Mbo KVTLDVVAAG AALYDQIWLG SYMSGGVGFT QYATAAYTDN ILDDFTYMTV LDVGSEVTFY ALEQYDEYPL ETGGSQRAAV TAAASGCCTA FATAWYLSQY
Mar RSALDVVAAG AALYDQIWLG SYMSGGVGFT QYASAAYTDN ILDDFLYMTV LDVGSEVTFY GLEQYEEYPL ETGGSQRASV VSAAAGCATA FATAWYLSMY
Mfi KVTLDVVAQA AAIYDQIWLG SYMSGGVGFT QYASAAYTDN VLDDFAYMTV LDVGSEVTFY GLEQYEEYPL ETGGSQRASV VSAAAGISTA FATAWYLSMY
Mc1 RVSLDVVAQA AAIYDQIWLG SYMSGGVGFT QYASAAYTDN VLDDFCYMTV LDVGTEVTFY GLEQYEEYPL ETGGSQRASV VSAAAGISTA FATAWYLSMY
Mc2 RTTLDVVAMS AALYDQIWLG SYMSGGVGFT QYASAAYTDN VLDDFTYMTV LDVGSEVTFY GLEQYEEYPL ETGGSQRASV VSAAAGCSTA FATAWYLSMY


401 411 421 431 441 451 461 471 481 491 500
Mor LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGIPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHSARGDA WAFNPLIKIA FADKNLIFDF SQPRAEFAKG
CS2 LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGIPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHSARGDA WAFNPLIKIA FADKNLIFDF SQPRAEFAKG
CS3 LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGIPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHSARGDA WAFNPLIKIA FADKNLIFDF SQPRAEFAKG
C86 LHKEQHSRLG FYGYDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARGDA FSFNPLVKIA FADKNLVFDF SKPRAEFAKG
ELS LHKEQHSRLG FYGYDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARGDA FSFNPLVKIA FADKNLVFDF SKPRAEFAKG
H10 LHKEQHSRLG FYGYDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARGDA FSFNPLVKIA FADKNLVFDF SKPRAEFAKG
c84 LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIS QAPHAARGDA WAFNPLIKIA FADKNLAFDF SKVRSEFAKG
c94 LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIS QAPHAARGDA WAFNPLIKIA FADKNLAFDF SKVRSEFAKG
B61 LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIS QAPHAARGDA WAFNPLIKIA FADKNLAFDF SKVRSEFAKG
Mwl LHKEQHSRLG FYGYDLQDQC GAANVFAIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHSARKDA WAFNPLIKIA FADKNLIFDF SKPRAEFAKG
Msm LHKEQHSRLG FYGYDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARGDA WSFNPLVKIA FADKNLVFDF SKPREEFAKG
Mmi LHKEQHSRLG FYGYDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARGDA WAFNPLIKIA FADKNLCFDF SQVRAEFAKG
Mth LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHSARGDA WAFNPLVKIA FADKNLVFDF SKVREEFAKG
Mgo LHKEQHSRLG FYGFDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARGDA FAFNPLIKIA FADKNLSFDF SKVRAEFAKG
Mol LHKEQHSRLG FYGFDLQDQC GAANVFAIRN DEGLPLELRG PNYPNYAMNV GHQGEYAGIA QAPHSARKDA YAVNPLVKIA FADKNLAFDF TKVRAEFAKG
Mru LHKEQHSRLG FYGFDLQDQC GAANVFAIRN DEGLPLELRG PNYPNYAMNV GHQGEYAGIA QAPHSARGDA FAVNPLVKIA FADKNLPFDF TKVRAEFAKG
Mw2 LHKEQHSRLG FYGYDLQDQC GAANTFSFRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARGDA FTYNPLIKIA FADKNLVFDF SAPRNEIAKG
Mbo LHKEQHSRLG FYGYDLQDQC GAANVFSIRN DEGLPLEMRG PNYPNYAMNV GHQGEYAGIA QAPHAARKDA FSFNPLIKIA FADKNLCFDF TQPRAEIAKG
Mar LHKEQHSRLG FYGYDLQDQC GASNVFSIRN DEGLPVEMRG PNYPNYAMNV GHQGEYAGIS QAPHSARGDA FAFNPLVKIA FADQNLSFDF SQPRAEIAKG
Mfi LHKEQHSRLG FYGYDLQDQC GASNVFSIRN DEGLPVELRG PNYPNYAMNV GHQGEYTGIA QAPHAARGDA FAFNPLIKIA FADDNLSFDF SKPRAEIAKG
Mc1 LHKEQHSRLG FYGYDLQDQC GASNVFSIRN DEGLPVELRG PNYPNYAMNV GHQGEYSGIA QAPHAARKDA FTFNPLIKIA FADKNLSFDF SQPRAEIAKG
Mc2 LHKEQHSRLG FYGYDLQDQC GASNVFSIRA DEGLPVEMRG PNYPNYAMNV GHQGEYAGIS QAPHAARKDA FAFNPLIKIA FADKNLSFDF SQPRAEIAKG



Supplemental Figure S13. Alignment of mcrA amino acid sequences from selected *Methanobrevibacter* genomes. Catalytic sites are indicated with color and an asterisk indicates the only site with an amino acid difference between the *Methanobrevibacter* spp.

4.9.2 Supplemental tables

Tables and the captions are available at:

https://drive.google.com/drive/folders/1B0TdPAbzW9R8a5EAaSjHYTmjmL21H2lu?usp=sharing

Chapter 5 | Applications of the methodology in other works

In this chapter, I discuss three published collaborative works I contributed to and was involved as a co-author during my doctoral training. All works reported in this chapter are closely related to the methodology development I introduced in **Chapter 3**; the article in **5.1** includes evolutionary history reconstruction for *Prevotella copri*, accomplished using analytical approach that was the prototype of the developed version of the methodology elaborated in **Chapter 3**; the article in **5.2** introduces another example of calibrating the evolutionary timeline of *Methanobrevibacter smithii* using ancient gut microbiome within the computational framework I detailed in **Chapter 3**; the work in **5.3** includes an example of applying the methodology described in **Chapter 3** to reconstructing the genome-based phylogeny of *Eubacterium rectale*. Each work is introduced by a brief statement of the project rationale, followed by the linkage with the methodology elaborated in **Chapter 3**, and closed by a short description of my role in the research. For each work, I demonstrate only the abstract, the section I contributed to and the access to the article published.

5.1 The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations

This article in which I am the second author was published in Cell Host & Microbe in 2019. The approach I designed for reconstructing the evolutionary history for *Prevotella copri* is the most original version of the methodology elaborated in **Chapter 3**. This is the first published work describing the evolutionary history of commensal bacteria directly using rich metagenomic data. In this study, we first reported that *P copri* is not a monotypic species but comprises four clades characterized by distinct phylogenetic placement, functions, and metabolic pathways. Secondly, we discovered the *P. copri* complex is more prevalent in populations with non-Westernized lifestyles and clades are frequently co-present within non-Westernized individuals. Thirdly, combining ancient gut content and molecular clocks we reconstructed a diversification timeline for the four clades of the whole complex, which postulates a trend that the presence of *P. copri* is decreasing in the Westernized world. In **Chapter 3**, I refined and generalized the approach in order to reconstruct the evolutionary history for more given microbial species which are present in human populations using ancient and contemporary metagenomic data.

Contribution:

In this study, my main contribution is reconstructing ancient strains of *P. copri*, genome-based phylogenetic analysis, and calibrating the phylogeny using reconstructed ancient strains within the Bayesian molecular clocking framework. I initially sought to reconstruct ancient genomes of *P. copri* from paleofeces samples for four clades. The resulting reconstructed genomes were inspected with stringent quality controls and ancient DNA authentication in order to ascertain the reliability of genomic information integrated for downstream analysis. Afterwards, qualified reconstructed ancient genomes were aligned against core genes shared among strains from four clades, generating core gene alignments containing both modern and ancient strains. Phylogeny congruence was inspected for all core genes before generating alignment concatenation in order to minimize the effect of horizontal gene transfer. The curated genome alignment concatenation was used for placing ancient strains in the whole *P. copri* genome-based phylogeny and reconstructing a time-revolved evolutionary history. In the part of reconstructing evolutionary history for the *P. copri* complex, I performed Bayesian molecular clocking analysis and carried out intensive model selection in order to describe the evolutionary history with sufficient reliability. As a result, we revolved a clear divergence timeline of *P. copri* four clades: the whole

complex began to diversify ~6.5 million years ago; the diversification of three main clades is around ~3.5 and ~2.5 million years ago; despite the recent date estimation of clade divergence, even at the lowest estimation (420 ka), this occurred well before the first human migration waves out of Africa which is around 90 - 194 ka years ago. Taken together, all our estimates indicate that the four clades of the *P. copri* complex were a feature of our pre-migratory human ancestors.

The abstract and the section related to this thesis is reported below, and the full-text article can be accessed here: https://www.sciencedirect.com/science/article/pii/S1931312819304275

The *Prevotella copri* Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations

Adrian Tett, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, Paolo Manghi, Kevin Bonham, Moreno Zolfo, Francesca De Filippis, Cara Magnabosco, Richard Bonneau, John Lusingu, John Amuasi, Karl Reinhard, Thomas Rattei, Fredrik Boulund, Lars Engstrand, Albert Zink, Maria Carmen Collado, Dan R. Littman, Daniel Eibach, Danilo Ercolini, Omar Rota-Stabelli, Curtis Huttenhower, Frank Maixner, Nicola Segata

Cell Host & Microbe, 2019

Abstract

Prevotella copri is a common human gut microbe that has been both positively and negatively associated with host health. In a cross-continent meta-analysis exploiting >6,500 metagenomes, we obtained >1,000 genomes and explored the genetic and population structure of *P. copri*. *P. copri* encompasses four distinct clades (>10% inter-clade genetic divergence) that we propose constitute the *P. copri* complex, and all clades were confirmed by isolate sequencing. These clades are nearly ubiquitous and co-present in non-Westernized populations. Genomic analysis showed substantial functional diversity in the complex with notable differences in carbohydrate metabolism, suggesting that multi-generational dietary modifications may be driving reduced prevalence in Westernized populations. Analysis of ancient metagenomes highlighted patterns of *P. copri* presence consistent with modern non-Westernized populations and a clade delineation time pre-dating human migratory waves out of Africa. These findings reveal that P. copri exhibits a high diversity that is underrepresented in Western-lifestyle populations.

5.1.1 *P. copri* Diversity in Ancient Human Gut Contents Resembles that of Non-Westernized Populations and Gives Insights into Its Evolutionary History

To ascertain if the high *P. copri* prevalence and co-presence of the four clades in non-Westernized populations reflects the composition in ancient human gut microbiomes, we analyzed the gut content of four archaeological samples. We studied material from the lower intestinal tract and lung tissue of the Iceman, a 5,300-year-old natural ice mummy (Spindler 2013). The Iceman genetically belongs to the Early European Farmers and originated and lived in Southern Europe, in the Eastern Italian Alps (Haak et al. 2015; Andreas Keller et al. 2012; Lazaridis et al. 2014; Müller et al. 2003) (Figure 5A). We also analyzed three coprolite samples (fossilized feces) (Figure 5A) recovered from the pre-Columbian (1,300 ± 100 BP) site “La Cueva de los Muertos Chiquitos” from Durango, a Northwestern state of Mexico (Brooks et al. 1962) (see STAR Methods).

We found the *P. copri* complex to be present in both the Mexican and the European ancient gut metagenomes (Figure 5B). All samples had at least two *P. copri* clades (clades A and C), and in two coprolites, all four clades could be detected. The higher prevalence of clade A and C in our ancient samples mirrors the tendency of modern-day populations (both Westernized and non-Westernized) where these two clades are more prevalent (Figure 3A). To discount the possibility of a non-ancient gut origin, we verified that the *P. copri* reads displayed damage patterns indicative of ancient DNA (Figures S5A and S5B) (Orlando, Gilbert, and Willerslev 2015), and in a control sample (Iceman lung tissue), no *P. copri* clades were detected (positive for only a single marker of the 2,448 *P. copri* complex specific markers) (Figure 5B). Two characteristics point toward a similarity between the ancient samples and modern non-Westernized populations. First, *P. copri* is common in the ancient samples like non-Westernized samples. Second, the ancient samples are characterized by a high clade co-presence (presence of 2 to 4 distinct clades) as observed in non-Westernized individuals. While we have only analyzed a small number of ancient metagenomes, we show that the likelihood of observing a high co-presence in the non-Westernized samples by chance is very low (Figure S5C). Together, the similarities between ancient and contemporary non-Westernized individuals suggest that the *P. copri* carriage pattern in non-Western populations is more akin to our ancestors.

To calibrate a *P. copri* phylogeny, we screened all ancient samples and found that one coprolite (sample 2180, radiocarbon dated AD 673 to 768, see STAR Methods) had

sufficiently high coverage of clade A to be used for tip calibration (see STAR Methods). Model selection indicated that this dataset is best modeled by a strict clock suggesting a constant rate of evolution through time and in different *P. copri* clades (see STAR Methods). All our divergence estimates converged satisfactorily on clear posterior means (Figure 5C) and the age estimates indicate that *P. copri* began to diversify (split of clade B) ~6.5 million years ago (Figure 5C). The diversification of clades A, D, and C is estimated to have occurred between ~3.5 and ~2.5 million years ago (Ungar and Sponheimer 2011; Wood and Collard 1999). The differentiation within each of the clades is instead relatively recent with the median estimates following that of the emergence of Homo sapiens circa 315 ka (Hublin et al. 2017). Despite the range in estimated clade divergence, even at the lowest estimation (420 ka), this occurred well before the first human migration waves out of Africa circa 90–194 ka years ago (Grün et al. 2005; Hershkovitz et al. 2018). This would indicate that the four clades of the *P. copri* complex were a feature of our pre-migratory human ancestors.

Further support of *P. copri* clade diversification prior to migration is that the high prevalence of all four clades and clade co-presence within an individual is a consistent feature of disparate non-Westernized populations in Africa, Oceania, South America, and Asia. This, together with the estimation of clade divergence, implies the *P. copri* complex has been a long-standing feature of the human microbiome. This analysis and the observed multi-generational decrease in the prevalence of *Prevotella* strains in non-Western migrants upon Westernization (Vangay et al. 2018) suggests that the underrepresentation of *P. copri* in Westernized populations could be due to its loss in response to Westernization. This loss has rapidly occurred in an almost infinitesimal time frame relative to host-microbe coevolution.

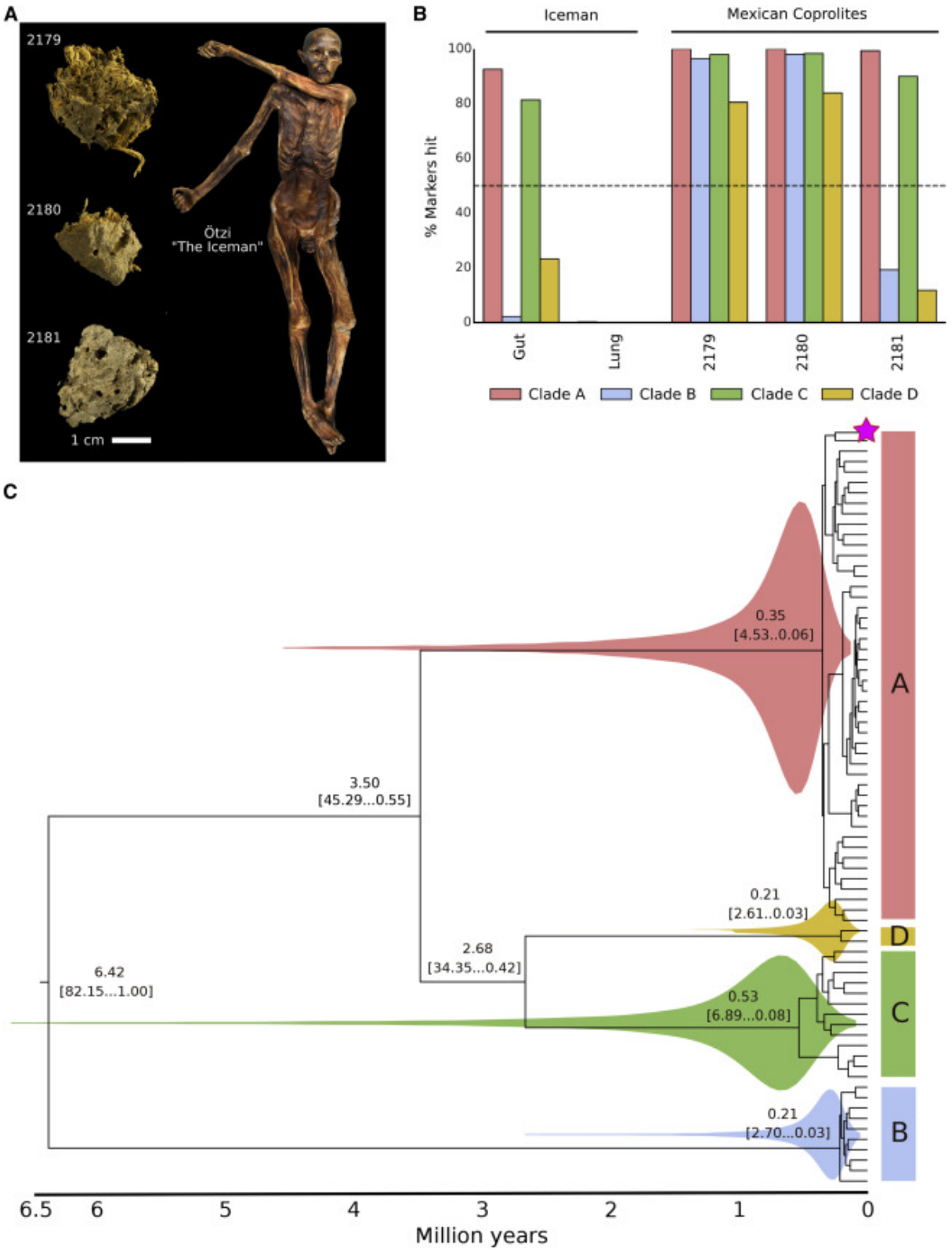


Figure 5. Ancient Microbiomes and the Evolutionary History of the *P. copri* Complex (A) Ancient Mexican coprolite samples and intestinal and lung tissue sampled from the Iceman, a natural ice mummy.

(B) Percentage of positive *P. copri* clade-specific markers identified in each ancient metagenomic sample.
(C) Time-resolved phylogenetic tree of the *P. copri* complex; magenta star indicates the ancient coprolite sample, 2180 (see STAR Methods).

5.2 Reconstruction of ancient microbial genomes from the human gut

This work I collaborated with a large panel of other researchers around the world was published in Nature in 2021. The part of exploring *Methanobrevibacter smithii* evolutionary history in this study can be taken as a great example for how to employ the approach elaborated in **Chapter 3** on real metagenomic data for investigating the evolutionary history of a given species. In this work, we firstly reconstructed a good number of high- and medium-quality ancient microbial genomes, and compared them with modern metagenomic data in the aspect of phylogenetic relationship, functional features. Secondly, we focused on reconstructing the evolutionary history for a pivotal human species, *Methanobrevibacter smithii*, by leveraging the ancient samples and rich contemporary metagenomic data. Finally, we uncovered that our ancestral gut microbiome resembles the present non-industrial human gut microbiome, which postulates an unignorable trend - our gut microbiome diversity is declining in the procedure of industrialization. Moreover, using the approach reported in **Chapter 3** can potentially open the black box of evolutionary history for many microbial species.

Contribution:

In this work, my contribution concentrated on exploring the evolutionary history of *M. smithii* performing the approach reported in **Chapter 3**. Firstly, I placed reconstructed ancient *M. smithii* genomes in the whole *M. smithii* phylogeny with available modern genomes in order to investigate the evolutionary relationship between ancient and modern strains. Secondly, I reconstructed a genome-based phylogeny using ancient strains and selected modern strains which represent the whole diversity of this species. Lastly, I built an evolutionary timeline of *M. smithii*, calibrating the phylogeny using the temporal signals from ancient strains within the Bayesian framework. The result demonstrates how *M. smithii* strains have radiated in evolution during the past 100,000 years ago.

The abstract is reported below, and the full-text article can be accessed here: https://doi.org/10.1038/s41586-021-03532-0

Reconstruction of ancient microbial genomes from the human gut

Marsha C. Wibowo, Zhen Yang, Maxime Borry, Alexander Hübner, Kun D. Huang, Braden T. Tierney, Samuel Zimmerman, Francisco Barajas-Olmos, Cecilia Contreras-Cubas, Humberto García-Ortiz, Angélica Martínez Hernández, Jacob M. Luber, Philipp Kirstahler, Tre Blohm, Francis E. Smiley, Sonia A. Ballal, Sünje Johanna Pamp, Julia Russ, Frank Maixner, Omar Rota-Stabelli, Nicola Segata, Karl Reinhard, Lorena Orozco, Christina Warinner, Meradeth Snow, Steven LeBlanc, Aleksander D. Kostic

Nature 2021

Abstract

Loss of gut microbial diversity (Smits et al. 2017; De Filippo et al. 2010; Yatsunenko et al. 2012; Obregon-Tito et al. 2015; Angelakis et al. 2016; Tett et al. 2019) in industrial populations is associated with chronic diseases (Blaser 2017), underscoring the importance of studying our ancestral gut microbiome. However, relatively little is known about the composition of pre-industrial gut microbiomes. Here we performed a large-scale *de novo* assembly of microbial genomes from palaeofaeces. From eight authenticated human palaeofaeces samples (1,000–2,000 years old) with well-preserved DNA from southwestern USA and Mexico, we reconstructed 498 medium- and high-quality microbial genomes. Among the 181 genomes with the strongest evidence of being ancient and of human gut origin, 39% represent previously undescribed species-level genome bins. Tip dating suggests an approximate diversification timeline for the key human symbiont Methanobrevibacter smithii. In comparison to 789 present-day human gut microbiome samples from eight countries, the palaeofaeces samples are more similar to non-industrialized than industrialized human gut microbiomes. Functional profiling of the palaeofaeces samples reveals a markedly lower abundance of antibiotic-resistance and mucin-degrading genes, as well as enrichment of mobile genetic elements relative to industrial gut microbiomes. This study facilitates the discovery and characterization of previously undescribed gut microorganisms from ancient microbiomes and the investigation of the evolutionary history of the human gut microbiota through genome reconstruction from palaeofaeces.

5.2.1 *Methanobrevibacter smithii* tip dating

Next, we estimated the divergence times of *M. smithii* using two filtered (contigs < 1% damage were removed) ancient M. smithii genomes from samples UT30.3 and UT43.2 for tip calibrations (Methods and Supplementary Fig. 3a). Bayesian inference under a strict clock and the most fitting demographic model (Supplementary Table 7) shows that the ancient *M. smithii* genomes fall within the known diversity of contemporary *M. smithii* genomes (Fig. 3 and Supplementary Fig. 3a) and that *M. smithii* began to diversify around 85,000 years ago with a 95% highest posterior density (HPD) interval of 51,000–128,000 years (Fig. 3). This timeline is moderately later than the timeline of its sister species *Methanobrevibacter oralis* (HPD = 112,000–143,000 years) (Weyrich et al. 2017). The two estimates are compatible in terms of HPD overlap, and both occurred within or slightly after the estimated first human migration waves out of Africa around 90,000–194,000 years ago (Grün et al. 2005; Hershkovitz et al. 2018). In addition, the origin of the lineage leading to the two ancient *M. smithii* genomes is between 40,000 and 16,000 years ago (mean = 27,000 years ago). These estimates predate (although there is overlap towards the earlier 95% posterior estimates) the accepted age of human entry into North America through the Beringia bridge (20,000–16,000 years ago). The results did not significantly change when potential aDNA damage sites were removed (Supplementary Fig. 3b and Supplementary Information section 8), suggesting that damage did not notably affect our MAGs. We also validated these divergence date estimates using raw sequence divergence calculations (Extended Data Fig. 10 and Supplementary Information section 8). Overall, we show that using ancient genomes for calibrating M. smithii phylogenies, we could evolutionarily match previous studies of *M. oralis* (Weyrich et al. 2017). This supports the potential of using ancient MAGs to study the evolutionary history of gut symbionts. However, whether species within the genus actually follow the indicated diversification timeline needs to be investigated with additional ancient Methanobrevibacter genomes that span different time periods.

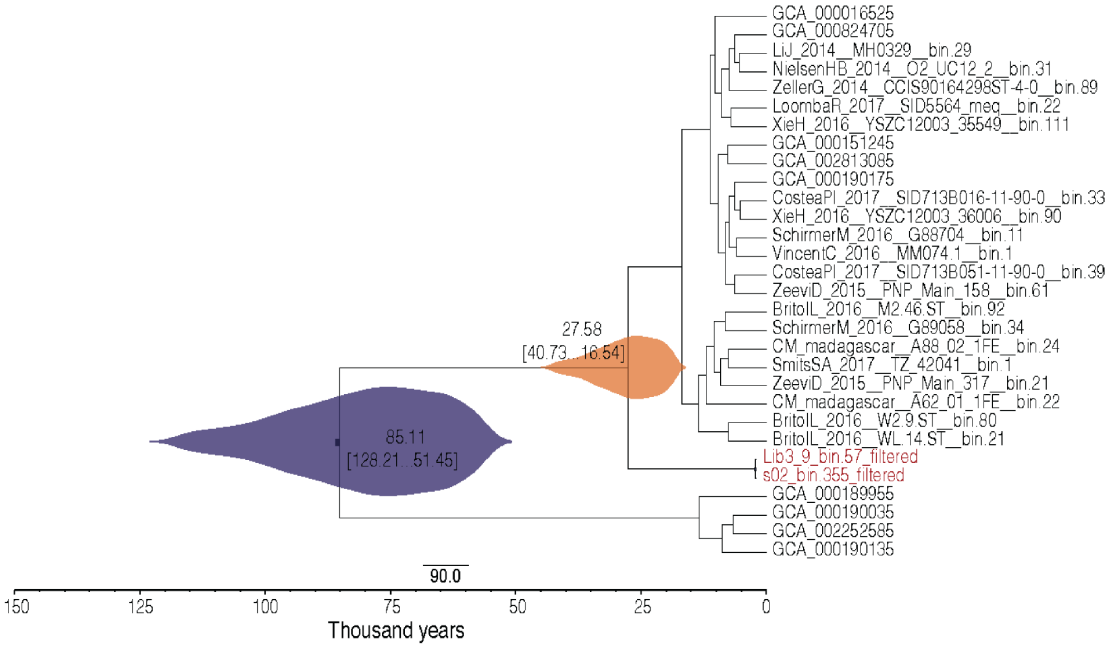
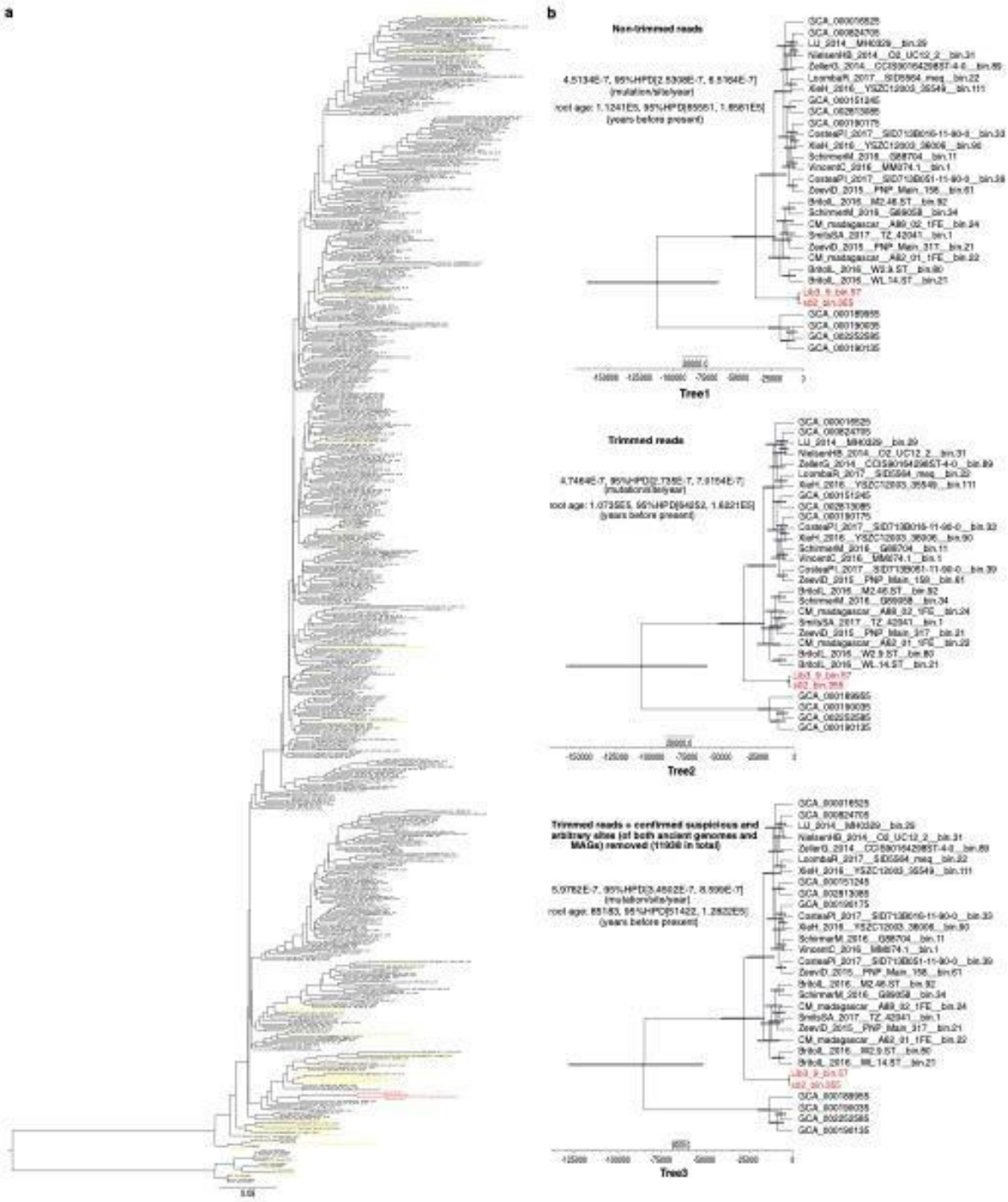
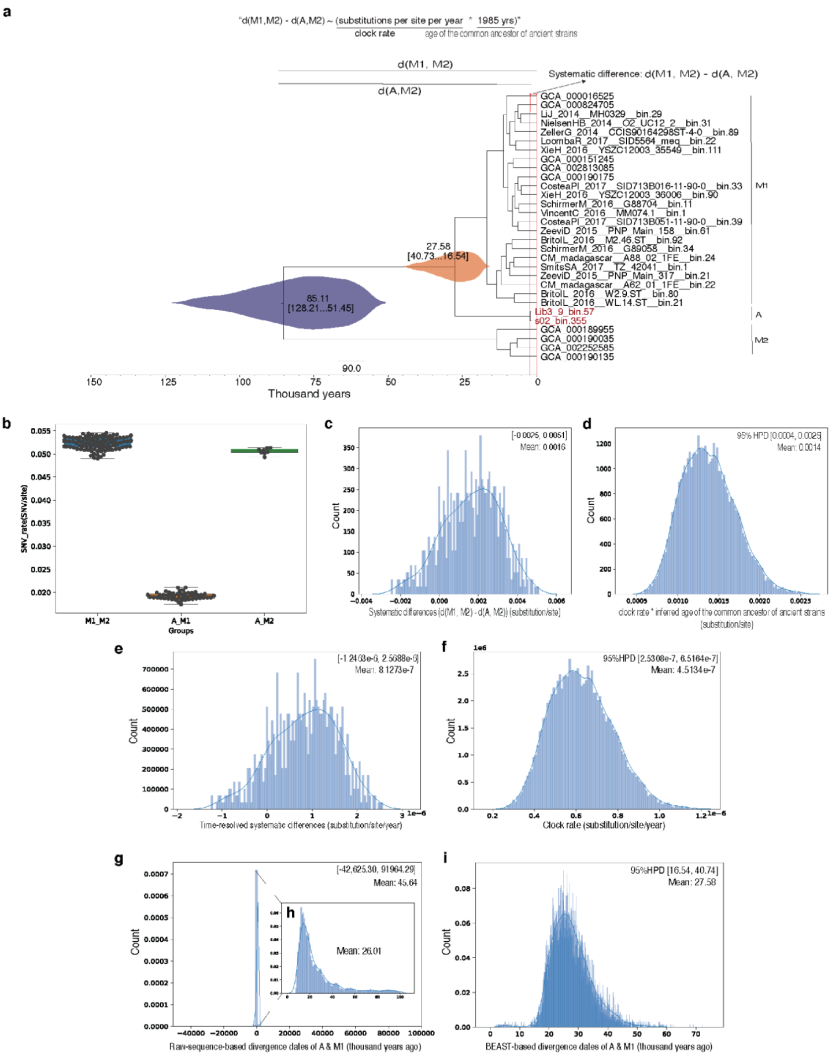


Fig. 3 | Evolutionary context of a key human gut symbiont. A time-measured phylogenetic tree of *M. smithii* reconstructed based on the core genome using a Bayesian approach under a strict clock model. Purple and orange violin plots illustrate the 95% HPD values (in parentheses) of estimated mean ages for the diversification of *M. smithii* and the split of the lineage leading to ancient *M. smithii* (highlighted in red), respectively.



Supplementary Fig. 3 | *Methanobrevibacter smithii* phylogeny and tip dating. **a,** A whole phylogeny of *M. smithii*. Ancient *M. smithii* MAGs are highlighted in red. Contemporary *M. smithii* genomes included in the Bayesian phylogenetic analysis are displayed in yellow. **b,** Comparison of tip dating results using three different versions of ancient *M. smithii* genomes: pre-filtered genomes (left panel), filtered genomes (middle panel; contigs with <1% damage were removed), and filtered genomes after trimming of the first and last 5bp of the reads (right panel). The results are consistent among the three versions.



Extended Data Fig. 10 | Comparison of *Methanobrevibacter smithii* divergence dates from BEAST2 analysis vs. raw genetic distance calculations (Supplementary Information section 8; related to Fig. 3). a, Diagram showing the different *M. smithii* groups and genetic distances calculated. **b,** Pairwise sequence divergences between M1 & M2 strains, A & M2 strains, and A & M1 strains (M1_M2 n = 96; A_M1 n = 48; A_M2 n = 8). Data are presented as boxplots as described in the Methods section. **c-d,** Comparison of distribution of systematic differences between M1 & M2 and A & M2 divergences and BEAST2 estimates. **c,** The systematic differences based on pairwise sequence divergences (measured by SNV rate) between M1 & M2 and A & M2 strains. **d,** The products of the clock rates (substitution/site/year) inferred using BEAST2(Bouckaert et al. 2014) (Supplementary Table 7) and the inferred age of the common ancestor of the ancient strains. **e-f,** Comparison of distribution of pairwise time-resolved systematic differences based on raw sequences divergence and the distribution of existing inferred clock rates. **e,** Time-resolved systematic differences calculated by dividing systematic differences (Extended Data Fig. 10c) with the average C14 date of the

paleofeces used in molecular clocking analysis. **f,** Clock rates inferred by BEAST2 analysis (Supplementary Table 7). **g,** Raw-sequence-based divergence dates between A & M1 strains, re-calibrated using time-resolved systematic differences. **h,** Distribution of raw-sequence-based divergence dates when low-frequency outliers are excluded. **i**, Distribution of estimated divergence dates between A & M1 strains based on BEAST2 analysis.

5.3 Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations

This work in which I was involved as a co-author was published in Genome Biology in 2020. In this work, I applied the approach reported in **Chapter 3** on reconstruction of genome phylogeny for *Eubacterium rectale* within the Bayesian framework. This study mainly focused on characterizing the geographical diversity and population structure of *Eubacterium rectale*, one of the most prevalent human gut bacteria, combining large-scale metagenomic analysis, population genetics, and phylogenetics. It revealed an unprecedented diversity of *Eubacterium rectale* regarding the geographical dispersal and host population structure.

Contribution:

In this study, I mainly focused on reconstructing a rooted genome phylogeny for exploring the divergence order of four subspecies in the course of *E. rectale* evolution. The strategies of accomplishing this analysis were based on the computational framework which has been discussed in detail in **Chapter 3.**

The abstract is reported below, and the full-text article can be accessed here:
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-020-02042-y

Analysis of 1321 *Eubacterium rectale* genomes from metagenomes uncovers complex phylogeographic population structure and subspecies functional adaptations

Nicolai Karcher, Edoardo Pasolli, Francesco Asnicar, Kun D. Huang, Adrian Tett, Serena Manara, Federica Armanini, Debbie Bain, Sylvia H. Duncan, Petra Louis, Moreno Zolfo, Paolo Manghi, Mireia Valles-Colomer, Roberta Raffaetà, Omar Rota-Stabelli, Maria Carmen Collado, Georg Zeller, Daniel Falush, Frank Maixner, Alan W. Walker, Curtis Huttenhower and Nicola Segata

Genome Biology 2020

Abstract

Background

Eubacterium rectale is one of the most prevalent human gut bacteria, but its diversity and population genetics are not well understood because large-scale whole-genome investigations of this microbe have not been carried out.

Results

Here, we leverage metagenomic assembly followed by a reference-based binning strategy to screen over 6500 gut metagenomes spanning geography and lifestyle and reconstruct over 1300 *E. rectale* high-quality genomes from metagenomes. We extend previous results of biogeographic stratification, identifying a new subspecies predominantly found in African individuals and showing that closely related non-human primates do not harbor *E. rectale*. Comparison of pairwise genetic and geographic distances between subspecies suggests that isolation by distance and co-dispersal with human populations might have contributed to shaping the contemporary population structure of *E. rectale*. We confirm that a relatively recently diverged *E. rectale* subspecies specific to Europe consistently lacks motility operons and that it is immotile in vitro, probably due to ancestral genetic loss. The same subspecies exhibits expansion of its carbohydrate metabolism gene repertoire including the acquisition of a genomic island strongly enriched in glycosyltransferase genes involved in exopolysaccharide synthesis.

Conclusions

Our study provides new insights into the population structure and ecology of *E. rectale* and shows that shotgun metagenomes can enable population genomics studies of microbiota members at a resolution and scale previously attainable only by extensive isolate sequencing.

Conclusions

Reconstructing and dating the divergence of microbiome species is of great interest for investigating the evolutionary history of the human microbiota. The state of the art is to recover microbial genomes from carbon dated ancient metagenome samples and use these genomes as calibration points for molecular clock dating analyses. It is a laborious task. An integrated computational framework is needed to scale up such investigations, particularly at the circumstance of continuously growing amounts of both ancient and contemporary metagenomic data available.

To address this challenge and facilitate metagenomics studies focused on interrogating the evolutionary history of the human microbiota, in this thesis I presented a novel computational framework and showcased its successful applications on real metagenomic datasets to uncover the underlying delicate evolutionary history of many human microbiome species.

In **Chapter 2** I showed how our gut microbiota has diverged from its ancestral state from the bird's-eye view of microbial composition and metabolic features, and this divergence could be associated with Westernized lifestyles including diet change and modern medical access. This work set the best proxy for investigating further the evolution of the human microbiome by comparing ancient metagenomic samples with present-day metagenomic datasets.

In **Chapter 3** I presented an integrated computational framework, MetaClock for reconstructing and dating the diversification timeline of microbiome species. I devised an automated pipeline which can precisely extract and integrate genomic information from ancient and contemporary metagenomic samples, and further build a time-resolved phylogeny for strains of microbiome species with molecular clock dating. In addition, I applied this approach to six common human gut members and uncovered underlying coevolution between microbiome and the host, for example the estimated diversification time of *Ruminococcus bromii* is approximately 750 ka when hominins started developing sophisticated human behaviors such as stone knapping(Alperson-Afil et al. 2009). This integrated pipeline has greatly contributed to the field by allowing for reconstructing the evolutionary history of other microbiome members (see **Chapter 4**).

In **Chapter 4** I demonstrated a detailed application of MetaClock in the circumstance of oral microbiome, recovering the evolutionary diversity of three *Methanobrevibacter* species two of which are novel species. In this work, the strain-level phylogenetic analysis performed using MetaClock revealed striking evolutionary diversity of three *Methanobrevibacter* species in the deep past but the diversity is clearly declining at present. This could be related to the recent change in the human lifestyle. This work promised a wider usage of MetaClock expanding applicable ecosystems from human gut to oral cavity, thus it can also contribute to the field focused on oral microbiome.

In **Chapter 5** I reported other studies in which MetaClock played an important role in investigating the fine-grained evolutionary history of the human microbiota.

Overall, the work discussed in this dissertation demonstrates a contribution to computational metagenomics and particularly to the direction where the interest of investigating the evolution of human microbiota using ancient metagenomic samples is gathering pace. The framework has been successfully applied in several circumstances to uncover the underlying evolutionary history of microbiota and the co-evolution with the host by focusing on certain members. To reconstruct a more complete map which can depict how one strain diverged into the other and the corresponding time for many species in the human microbiota, we need to discover more species that are prevalent and abundant in both ancient and present-day metagenomic samples. The continuously growing number of archaeological human microbiome samples excavated will promise more microbial species to be recovered from the long past. Analyzing them with their modern counterparts in this framework will give fresh insights into the unexplored evolutionary history of the human microbiota.

Future perspectives

The studies discussed in this dissertation promise several interesting aspects where combining metagenomics, phylogenetic concepts and molecular clock dating is emerging as a powerful tool in investigating the evolutionary history of the human microbiota. I would like to elaborate three potential research directions to follow up in the future.

1. **Dating the evolutionary history of oral microbiome members**. In **Chapter 2** and **Chapter 3** we explored a detailed divergence pattern of the human

microbiome structure before and after the industrialization, and further reconstructed time-resolved phylogenies for six common human gut bacteria. This largely expanded our understanding about how the human gut microbiome has diverged in the past from the level of general microbiome structure to that of specific species. While in **Chapter 4** we investigated the evolutionary diversity of oral microbiome using three *Methanobrevibacter* species as examples, unlike gut microbiome species studied in **Chapter 3** and **4,** an evolutionary timeline for these oral species is still missing. Thus, I would like to carry on MetaClock analysis on these three *Methanobrevibacter* species in order to reconstruct a time-measured evolutionary history of the human oral microbiome.

2. **Expanding MetaClock analysis coupling with metagenomic assembly advances.** Applications of MetaClock are largely focused on known microbial species in the studies reported in my dissertation. It could hinder a better understanding about our microbiome evolution because the human microbiome consists of large amounts of the unknown. In **Chapter 4**, we demonstrated how we recovered two novel *Methanobrevibacter* species and built their strain-level phylogenies by combining *de novo* metagenomic assembly and MetaClock analysis. In the future, we would like to employ this strategy to human gut microbiome members so as to complement the limitations in **Chapter 2** and **Chapter 3** where only known species were studied.

3. **Fully Automatizing the section of molecular clock dating in the MetaClock pipeline**. MetaClock is a highly integrated framework for studying the evolutionary history of the human microbiome directly using ancient and contemporary metagenomic samples. It is fully automated from extracting metagenomic reads and reconstructing a strain-level phylogeny to estimating temporal signals for further molecular dating. However, to perform intensive model selection for molecular clock dating analysis users still need to execute external softwares such as BEAST2 following our detailed instructions in the tutorial. Model selection is a very daunting task and requires expertise in Bayesian phylogenetic analysis. I wish to fully automate this part in the MetaClock pipeline, so that users will feel more comfortable with performing likewise analysis in the MetaClock framework without overloaded prior knowledge.

With a continuously growing number of ancient metagenomes being available and the increase of computational capability, this methodology will enable us to reconstruct a more complete evolutionary history of human microbiome members and their association with the host evolution. It can certainly increase our understanding about how our microbiota has been shaped in the past.

References

Adler, Christina J., Keith Dobney, Laura S. Weyrich, John Kaidonis, Alan W. Walker, Wolfgang Haak, Corey J. A. Bradshaw, et al. 2013. “Sequencing Ancient Calcified Dental Plaque Shows Changes in Oral Microbiota with Dietary Shifts of the Neolithic and Industrial Revolutions.” *Nature Genetics* 45 (4): 450–55, 455e1.

Alboukadel, Kassambara. 2018. “ggpubr:‘ggplot2’ Based Publication Ready Plots.” *R Package Version 0. 4. 0*.

Alexander, David H., and Kenneth Lange. 2011. “Enhancements to the ADMIXTURE Algorithm for Individual Ancestry Estimation.” *BMC Bioinformatics*. https://doi.org/10.1186/1471-2105-12-246.

Almeida, Alexandre, Alex L. Mitchell, Miguel Boland, Samuel C. Forster, Gregory B. Gloor, Aleksandra Tarkowska, Trevor D. Lawley, and Robert D. Finn. 2019. “A New Genomic Blueprint of the Human Gut Microbiota.” *Nature* 568 (7753): 499–504.

Alperson-Afil, Nira, Gonen Sharon, Mordechai Kislev, Yoel Melamed, Irit Zohar, Shosh Ashkenazi, Rivka Rabinovich, et al. 2009. “Spatial Organization of Hominin Activities at Gesher Benot Ya’aqov, Israel.” *Science* 326 (5960): 1677–80.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10.

Andrews, Richard M., Iwona Kubacka, Patrick F. Chinnery, Robert N. Lightowlers, Douglass M. Turnbull, and Neil Howell. 1999. “Reanalysis and Revision of the Cambridge Reference Sequence for Human Mitochondrial DNA.” *Nature Genetics*. https://doi.org/10.1038/13779.

Angelakis, Emmanouil, Muhammad Yasir, Dipankar Bachar, Esam I. Azhar, Jean-Christophe Lagier, Fehmida Bibi, Asif A. Jiman-Fatani, et al. 2016. “Gut Microbiome and Dietary Patterns in Different Saudi Populations and Monkeys.” *Scientific Reports* 6 (August): 32191.

Aouizerat, Tzemach, Itai Gutman, Yitzhak Paz, Aren M. Maeir, Yuval Gadot, Daniel Gelman, Amir Szitenberg, et al. 2019. “Isolation and Characterization of Live Yeast Cells from Ancient Vessels as a Tool in Bio-Archaeology.” *mBio* 10 (2). https://doi.org/10.1128/mBio.00388-19.

Asnicar, Francesco, Sarah E. Berry, Ana M. Valdes, Long H. Nguyen, Gianmarco Piccinno, David A. Drew, Emily Leeming, et al. 2021. “Microbiome Connections with Host Metabolism and Habitual Diet from 1,098 Deeply Phenotyped Individuals.” *Nature Medicine* 27 (2): 321–32.

Asnicar, Francesco, Serena Manara, Moreno Zolfo, Duy Tin Truong, Matthias Scholz, Federica Armanini, Pamela Ferretti, et al. 2017. “Studying Vertical Microbiome Transmission from Mothers to Infants by Strain-Level Metagenomic Profiling.” *mSystems* 2 (1). https://doi.org/10.1128/mSystems.00164-16.

Asnicar, Francesco, Andrew Maltez Thomas, Francesco Beghini, Claudia Mengoni, Serena Manara, Paolo Manghi, Qiyun Zhu, et al. 2020. “Precise Phylogenetic Analysis of Microbial Isolates and Genomes from Metagenomes Using PhyloPhlAn 3.0.” *Nature Communications* 11 (1): 2500.

Axelsson, Erik, Eske Willerslev, M. Thomas P. Gilbert, and Rasmus Nielsen. 2008.

"The Effect of Ancient DNA Damage on Inferences of Demographic Histories." *Molecular Biology and Evolution* 25 (10): 2181–87.
Baele, Guy, Philippe Lemey, Trevor Bedford, Andrew Rambaut, Marc A. Suchard, and Alexander V. Alekseyenko. 2012. "Improving the Accuracy of Demographic and Molecular Clock Model Comparison While Accommodating Phylogenetic Uncertainty." *Molecular Biology and Evolution* 29 (9): 2157–67.
Baker, Emilyclare P., David Peris, Ryan V. Moriarty, Xueying C. Li, Justin C. Fay, and Chris Todd Hittinger. 2019. "Mitochondrial DNA and Temperature Tolerance in Lager Yeasts." *Science Advances* 5 (1): eaav1869.
Baker, Jonathon L., James T. Morton, Márcia Dinis, Ruth Alvarez, Nini C. Tran, Rob Knight, and Anna Edlund. 2021. "Deep Metagenomics Examines the Oral Microbiome during Dental Caries, Revealing Novel Taxa and Co-Occurrences with Host Molecules." *Genome Research* 31 (1): 64–74.
Banchi, Elisa, Claudio G. Ametrano, Samuele Greco, David Stanković, Lucia Muggia, and Alberto Pallavicini. 2020. "PLANiTS: A Curated Sequence Reference Dataset for Plant ITS DNA Metabarcoding." *Database*. https://doi.org/10.1093/database/baz155.
Basso, Rafael Felipe, André Ricardo Alcarde, and Cauré Barbosa Portugal. 2016. "Could Non-Saccharomyces Yeasts Contribute on Innovative Brewing Fermentations?" *Food Research International*. https://doi.org/10.1016/j.foodres.2016.06.002.
Beghini, Francesco, Lauren J. McIver, Aitor Blanco-Míguez, Leonard Dubois, Francesco Asnicar, Sagun Maharjan, Ana Mailyan, et al. 2021. "Integrating Taxonomic, Functional, and Strain-Level Profiling of Diverse Microbial Communities with bioBakery 3." *eLife* 10 (May). https://doi.org/10.7554/eLife.65088.
Belay, N., R. Johnson, B. S. Rajagopal, E. Conway de Macario, and L. Daniels. 1988. "Methanogenic Bacteria from Human Dental Plaque." *Applied and Environmental Microbiology* 54 (2): 600–603.
Berg, Gabriele, Daria Rybakova, Doreen Fischer, Tomislav Cernava, Marie-Christine Champomier Vergès, Trevor Charles, Xiaoyulong Chen, et al. 2020. "Microbiome Definition Re-Visited: Old Concepts and New Challenges." *Microbiome* 8 (1): 103.
Bertels, Frederic, Olin K. Silander, Mikhail Pachkov, Paul B. Rainey, and Erik van Nimwegen. 2014. "Automated Reconstruction of Whole-Genome Phylogenies from Short-Sequence Reads." *Molecular Biology and Evolution* 31 (5): 1077–88.
Blaser, Martin J. 2017. "The Theory of Disappearing Microbiota and the Epidemics of Chronic Diseases." *Nature Reviews. Immunology* 17 (8): 461–63.
Bolyen, Evan, Jai Ram Rideout, Matthew R. Dillon, Nicholas A. Bokulich, Christian C. Abnet, Gabriel A. Al-Ghalith, Harriet Alexander, et al. 2019. "Reproducible, Interactive, Scalable and Extensible Microbiome Data Science Using QIIME 2." *Nature Biotechnology* 37 (8): 852–57.
Borrel, Guillaume, Panagiotis S. Adam, Luke J. McKay, Lin-Xing Chen, Isabel Natalia Sierra-García, Christian M. K. Sieber, Quentin Letourneur, et al. 2019. "Wide Diversity of Methane and Short-Chain Alkane Metabolisms in Uncultured Archaea." *Nature Microbiology* 4 (4): 603–13.
Borry, Maxime, Bryan Cordova, Angela Perri, Marsha Wibowo, Tanvi Prasad Honap, Jada Ko, Jie Yu, et al. 2020. "CoproID Predicts the Source of Coprolites and Paleofeces Using Microbiome Composition and Host DNA Content." *PeerJ* 8 (April): e9001.
Bottéro, Jean. 1985. "The Cuisine of Ancient Mesopotamia." *The Biblical Archaeologist*. https://doi.org/10.2307/3209946.
Bouckaert, Remco, Joseph Heled, Denise Kühnert, Tim Vaughan, Chieh-Hsi Wu, Dong Xie, Marc A. Suchard, Andrew Rambaut, and Alexei J. Drummond. 2014. "BEAST 2: A Software Platform for Bayesian Evolutionary Analysis." *PLoS Computational Biology* 10 (4): e1003537.

Bowers, Robert M., The Genome Standards Consortium, Nikos C. Kyrpides, Ramunas Stepanauskas, Miranda Harmon-Smith, Devin Doud, T. B. K. Reddy, et al. 2017. “Minimum Information about a Single Amplified Genome (MISAG) and a Metagenome-Assembled Genome (MIMAG) of Bacteria and Archaea.” *Nature Biotechnology*. https://doi.org/10.1038/nbt.3893.

Bravo-Lopez, Miriam, Viridiana Villa-Islas, Carolina Rocha Arriaga, Ana B. Villaseñor-Altamirano, Axel Guzmán-Solís, Marcela Sandoval-Velasco, Julie K. Wesp, et al. 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, Jaelle C., Henrique G. Leitão, Tom van der Valk, Wenbo Xu, Katia Bougiouri, Love Dalén, and Katerina Guschanski. 2020. “Dental Calculus as a Tool to Study the Evolution of the Mammalian Oral Microbiome.” *Molecular Biology and Evolution* 37 (10): 3003–22.

Brewster, Ryan, Fiona B. Tamburini, Edgar Asiimwe, Ovokeraye Oduaran, Scott Hazelhurst, and Ami S. Bhatt. 2019. “Surveying Gut Microbiome Research in Africans: Toward Improved Diversity and Representation.” *Trends in Microbiology* 27 (10): 824–35.

Briggs, Adrian W., Udo Stenzel, Philip L. F. Johnson, Richard E. Green, Janet Kelso, Kay Prüfer, Matthias Meyer, et al. 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): 14616–21.

Bron, Peter A., Peter van Baarlen, and Michiel Kleerebezem. 2011. “Emerging Molecular Insights into the Interaction between Probiotics and the Host Intestinal Mucosa.” *Nature Reviews. Microbiology* 10 (1): 66–78.

Brooks, Richard H., Lawrence Kaplan, Hugh C. Cutler, and Thomas W. Whitaker. 1962. “Plant Material from a Cave on the Rio Zape, Durango, Mexico.” *American Antiquity* 27 (3): 356–69.

Buchfink, Benjamin, Klaus Reuter, and Hajk-Georg Drost. 2021. “Sensitive Protein Alignments at Tree-of-Life Scale Using DIAMOND.” *Nature Methods*. https://doi.org/10.1038/s41592-021-01101-x.

Burger, Joachim, Vivian Link, Jens Blöcher, Anna Schulz, Christian Sell, Zoé Pochon, Yoan Diekmann, et al. 2020. “Low Prevalence of Lactase Persistence in Bronze Age Europe Indicates Ongoing Strong Selection over the Last 3,000 Years.” *Current Biology*. https://doi.org/10.1016/j.cub.2020.08.033.

Cakmakci, Songül, Engin Gundogdu, Ali A. Hayaloglu, Elif Dagdemir, Mustafa Gurses, Bulent Cetin, and Deren Tahmas-Kahyaoglu. 2012. “Chemical and Microbiological Status and Volatile Profiles of Mouldy Civil Cheese, a Turkish Mould-Ripened Variety.” *International Journal of Food Science & Technology*. https://doi.org/10.1111/j.1365-2621.2012.03116.x.

Cantor, Mette Dines, Tatjana van den Tempel, Tine Kronborg Hansen, and Ylva Ardö. 2017. “Blue Cheese.” *Cheese*. https://doi.org/10.1016/b978-0-12-417012-4.00037-5.

Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. 2009. “trimAl: A Tool for Automated Alignment Trimming in Large-Scale Phylogenetic Analyses.” *Bioinformatics* 25 (15): 1972–73.

Capozzi, Vittorio, and Giuseppe Spano. 2011. “Food Microbial Biodiversity and ?Microbes of Protected Origin?” *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2011.00237.

Cappers, René T. J., Renée M. Bekker, and Judith E. A. Jans. 2006. *Digital seed atlas of the Netherlands*. Barkhuis Publishing.

Cheeseman, Kevin, Jeanne Ropars, Pierre Renault, Joëlle Dupont, Jérôme Gouzy, Antoine Branca, Anne-Laure Abraham, et al. 2014. “Multiple Recent Horizontal Transfers of a Large Genomic Region in Cheese Making Fungi.” *Nature*

Communications 5: 2876.
Colombo, Ana Paula V., Susan K. Boches, Sean L. Cotton, J. Max Goodson, Ralph Kent, Anne D. Haffajee, Sigmund S. Socransky, et al. 2009. “Comparisons of Subgingival Microbial Profiles of Refractory Periodontitis, Severe Periodontitis, and Periodontal Health Using the Human Oral Microbe Identification Microarray.” *Journal of Periodontology* 80 (9): 1421–32.
Consortium, The Human Microbiome Project, and The Human Microbiome Project Consortium. 2012. “Structure, Function and Diversity of the Healthy Human Microbiome.” *Nature*. https://doi.org/10.1038/nature11234.
Coordinators, Ncbi Resource, and NCBI Resource Coordinators. 2012. “Database Resources of the National Center for Biotechnology Information.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gks1189.
Crittenden, Alyssa N., and Stephanie L. Schnorr. 2017. “Current Views on Hunter‐gatherer Nutrition and the Evolution of the Human Diet.” *American Journal of Physical Anthropology*. https://doi.org/10.1002/ajpa.23148.
Damgaard, Peter B., Ashot Margaryan, Hannes Schroeder, Ludovic Orlando, Eske Willerslev, and Morten E. Allentoft. 2015. “Improving Access to Endogenous DNA in Ancient Bones and Teeth.” *Scientific Reports* 5 (June): 11184.
De Filippis, Francesca, Edoardo Pasolli, Adrian Tett, Sonia Tarallo, Alessio Naccarati, Maria De Angelis, Erasmo Neviani, et al. 2019. “Distinct Genetic and Functional Traits of Human Intestinal Prevotella Copri Strains Are Associated with Different Habitual Diets.” *Cell Host & Microbe* 25 (3): 444–53.e3.
De Filippo, Carlotta, Duccio Cavalieri, Monica Di Paola, Matteo Ramazzotti, Jean Baptiste Poullet, Sebastien Massart, Silvia Collini, Giuseppe Pieraccini, and Paolo Lionetti. 2010. “Impact of Diet in Shaping Gut Microbiota Revealed by a Comparative Study in Children from Europe and Rural Africa.” *Proceedings of the National Academy of Sciences of the United States of America* 107 (33): 14691–96.
Desmasures, N. 2014. “CHEESE | Mold-Ripened Varieties.” *Encyclopedia of Food Microbiology*. https://doi.org/10.1016/b978-0-12-384730-0.00060-4.
Deutsch, Eric W., Luis Mendoza, David Shteynberg, Joseph Slagel, Zhi Sun, and Robert L. Moritz. 2015. “Trans-Proteomic Pipeline, a Standardized Data Processing Pipeline for Large-Scale Reproducible Proteomics Informatics.” *Proteomics. Clinical Applications* 9 (7-8): 745–54.
Didelot, Xavier, and Daniel J. Wilson. 2015. “ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes.” *PLoS Computational Biology* 11 (2): 1–18.
Donati, Claudio, Moreno Zolfo, Davide Albanese, Duy Tin Truong, Francesco Asnicar, Valerio Iebba, Duccio Cavalieri, et al. 2016. “Uncovering Oral Neisseria Tropism and Persistence Using Metagenomic Sequencing.” *Nature Microbiology* 1 (7): 16070.
Dornbusch, Horst D. 1998. *Prost!: The Story of German Beer*. Brewers Publications.
Dos Vultos, Tiago, Olga Mestre, Jean Rauzier, Marcin Golec, Nalin Rastogi, Voahangy Rasolofo, Tone Tonjum, Christophe Sola, Ivan Matic, and Brigitte Gicquel. 2008. “Evolution and Diversity of Clonal Bacteria: The Paradigm of Mycobacterium Tuberculosis.” *PloS One* 3 (2): e1538.
Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus. 2005. “Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences.” *Molecular Biology and Evolution* 22 (5): 1185–92.
Drummond, Alexei J., and Andrew Rambaut. 2007. “BEAST: Bayesian Evolutionary Analysis by Sampling Trees.” *BMC Evolutionary Biology* 7 (November): 214.
Duchêne, Sebastián, and Edward C. Holmes. 2018. “Estimating Evolutionary Rates in Giant Viruses Using Ancient Genomes.” *Virus Evolution* 4 (1): vey006.
Duchêne, Sebastian, Kathryn E. Holt, François-Xavier Weill, Simon Le Hello, Jane Hawkey, David J. Edwards, Mathieu Fourment, and Edward C. Holmes. 2016.

"Genome-Scale Rates of Evolutionary Change in Bacteria." *Microbial Genomics* 2 (11): e000094.
Dumas, Emilie, Alice Feurtey, Ricardo C. Rodríguez de la Vega, Stéphanie Le Prieur, Alodie Snirc, Monika Coton, Anne Thierry, et al. 2020. "Independent Domestication Events in the Blue-Cheese Fungus Penicillium Roqueforti." *Molecular Ecology* 29 (14): 2639–60.
Elabdeen, H. R. Z., M. Mustafa, H. Hasturk, V. Klepac-Ceraj, R. W. Ali, B. J. Paster, T. Van Dyke, and A. I. Bolstad. 2015. "Subgingival Microbial Profiles of Sudanese Patients with Aggressive Periodontitis." *Journal of Periodontal Research* 50 (5): 674–82.
Eng, Jimmy K., Tahmina A. Jahan, and Michael R. Hoopmann. 2013. "Comet: An Open-Source MS/MS Sequence Database Search Tool." *PROTEOMICS*. https://doi.org/10.1002/pmic.201200439.
Ermler, U., W. Grabarse, S. Shima, M. Goubeaud, and R. K. Thauer. 1997. "Crystal Structure of Methyl-Coenzyme M Reductase: The Key Enzyme of Biological Methane Formation." *Science* 278 (5342): 1457–62.
Escapa, Isabel F., Tsute Chen, Yanmei Huang, Prasad Gajare, Floyd E. Dewhirst, and Katherine P. Lemon. 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). https://doi.org/10.1128/mSystems.00187-18.
Evans, Paul N., Joel A. Boyd, Andy O. Leu, Ben J. Woodcroft, Donovan H. Parks, Philip Hugenholtz, and Gene W. Tyson. 2019. "An Evolving View of Methane Metabolism in the Archaea." *Nature Reviews. Microbiology* 17 (4): 219–32.
Eyzaguirre, Jaime, Kathrin Jansen, and Georg Fuchs. 1982. "Phosphoenolpyruvate Synthetase inMethanobacterium Thermoautotrophicum." *Archives of Microbiology* 132 (1): 67–74.
Fehlner-Peach, Hannah, Cara Magnabosco, Varsha Raghavan, Jose U. Scher, Adrian Tett, Laura M. Cox, Claire Gottsegen, et al. 2019. "Distinct Polysaccharide Utilization Profiles of Human Intestinal Prevotella Copri Isolates." *Cell Host & Microbe*. https://doi.org/10.1016/j.chom.2019.10.013.
Ferrari, Annamaria, Tullio Brusa, Anna Rutili, Enrica Canzi, and Bruno Biavati. 1994. "Isolation and Characterization ofMethanobrevibacter Oralis Sp. Nov." *Current Microbiology*. https://doi.org/10.1007/bf01570184.
Ferrer, Manuel, Celia Méndez-García, David Rojo, Coral Barbas, and Andrés Moya. 2017. "Antibiotic Use and Microbiome Function." *Biochemical Pharmacology* 134 (June): 114–26.
Festi, Daniela, Daniel Brandner, Michael Grabner, Wolfgang Knierzinger, Hans Reschreiter, and Kerstin Kowarik. 2021. "3500 Years of Environmental Sustainability in the Large-Scale Alpine Mining District of Hallstatt, Austria." *Journal of Archaeological Science: Reports* 35 (February): 102670.
Forouzan, Esmaeil, Parvin Shariati, Masoumeh Sadat Mousavi Maleki, Ali Asghar Karkhane, and Bagher Yakhchali. 2018. "Practical Evaluation of 11 de Novo Assemblers in Metagenome Assembly." *Journal of Microbiological Methods* 151 (August): 99–105.
Forster, Samuel C., Nitin Kumar, Blessing O. Anonye, Alexandre Almeida, Elisa Viciani, Mark D. Stares, Matthew Dunn, et al. 2019. "A Human Gut Bacterial Genome and Culture Collection for Improved Metagenomic Analyses." *Nature Biotechnology* 37 (2): 186–92.
Fox, Patrick F., Timothy P. Guinee, Timothy M. Cogan, and Paul L. H. McSweeney. 2017. "Cheese: Historical Aspects." *Fundamentals of Cheese Science*. https://doi.org/10.1007/978-1-4899-7681-9_1.
Franzosa, Eric A., Lauren J. McIver, Gholamali Rahnavard, Luke R. Thompson, Melanie Schirmer, George Weingart, Karen Schwarzberg Lipson, et al. 2018. "Species-Level Functional Profiling of Metagenomes and Metatranscriptomes."

Nature Methods 15 (11): 962–68.
Gallone, Brigida, Jan Steensels, Troels Prahl, Leah Soriaga, Veerle Saels, Beatriz Herrera-Malaver, Adriaan Merlevede, et al. 2016. “Domestication and Divergence of Saccharomyces Cerevisiae Beer Yeasts.” *Cell* 166 (6): 1397–1410.e16.
Gálvez, Eric J. C., Aida Iljazovic, Lena Amend, Till Robin Lesker, Thibaud Renault, Sophie Thiemann, Lianxu Hao, et al. 2020. “Distinct Polysaccharide Utilization Determines Interspecies Competition between Intestinal Prevotella Spp.” *Cell Host & Microbe*, October. https://doi.org/10.1016/j.chom.2020.09.012.
García, Santos, Jorge E. Vidal, Norma Heredia, and Vijay K. Juneja. 2019. “Clostridium Perfringens.” *Food Microbiology*. https://doi.org/10.1128/9781555819972.ch19.
Gilbert, M. Thomas P., Dennis L. Jenkins, Anders Götherstrom, Nuria Naveran, Juan J. Sanchez, Michael Hofreiter, Philip Francis Thomsen, et al. 2008. “DNA from Pre-Clovis Human Coprolites in Oregon, North America.” *Science* 320 (5877): 786–89.
Gilbert, Robert I., and James H. Mielke. 1985. *The Analysis of Prehistoric Diets*. Academic Press.
Gonçalves, Margarida, Ana Pontes, Pedro Almeida, Raquel Barbosa, Marta Serra, Diego Libkind, Mathias Hutzler, Paula Gonçalves, and José Paulo Sampaio. 2016. “Distinct Domestication Trajectories in Top-Fermenting Beer Yeasts and Wine Yeasts.” *Current Biology: CB* 26 (20): 2750–61.
Goris, Johan, Konstantinos T. Konstantinidis, Joel A. Klappenbach, Tom Coenye, Peter Vandamme, and James M. Tiedje. 2007. “DNA--DNA Hybridization Values and Their Relationship to Whole-Genome Sequence Similarities.” *International Journal of Systematic and Evolutionary Microbiology* 57 (1): 81–91.
Grabner, Michael, Elisabeth Wächter, Kurt Nicolussi, Monika Bolka, Trivun Sormaz, Peter Steier, Eva Maria Wild, et al. 2021. “Prehistoric Salt Mining in Hallstatt, Austria. New Chronologies out of Small Wooden Fragments.” *Dendrochronologia*. https://doi.org/10.1016/j.dendro.2021.125814.
Grabner, M., H. Reschreiter, E. Wächter, M. Konrad, G. Winner, and K. Kowarik. 2019. “Die Verwendeten Holzarten Im Prähistorischen Salzbergbau von Hallstatt.” *Kowarik, K. , Hallstätter Beziehungsgeschichten. Wirtschaftsstrukturen Und Umfeldbeziehungen Der Bronze-Und ältereisenzeitlichen Salzbergbaue von Hallstatt/OÖ. Studien Zur Kulturgeschichte von Oberösterreich* 50: 83–96.
Grant, William D. 2015. “Halococcus.” In *Bergey’s Manual of Systematics of Archaea and Bacteria*, 1–10. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118960608.gbm00484.
Greco, Enrico, Ola El-Aguizy, Mona Fouad Ali, Salvatore Foti, Vincenzo Cunsolo, Rosaria Saletti, and Enrico Ciliberto. 2018. “Proteomic Analyses on an Ancient Egyptian Cheese and Biomolecular Evidence of Brucellosis.” *Analytical Chemistry*. https://doi.org/10.1021/acs.analchem.8b02535.
Griffen, Ann L., Clifford J. Beall, James H. Campbell, Noah D. Firestone, Purnima S. Kumar, Zamin K. Yang, Mircea Podar, and Eugene J. Leys. 2012. “Distinct and Complex Bacterial Profiles in Human Periodontitis and Health Revealed by 16S Pyrosequencing.” *The ISME Journal*. https://doi.org/10.1038/ismej.2011.191.
Grine, Ghiles, Elodie Terrer, Mahmoud Abdelwadoud Boualam, Gérard Aboudharam, Hervé Chaudet, Raymond Ruimy, and Michel Drancourt. 2018. “Tobacco-Smoking-Related Prevalence of Methanogens in the Oral Fluid Microbiota.” *Scientific Reports* 8 (1): 9197.
Grün, Rainer, Chris Stringer, Frank McDermott, Roger Nathan, Naomi Porat, Steve Robertson, Lois Taylor, Graham Mortimer, Stephen Eggins, and Malcolm McCulloch. 2005. “U-Series and ESR Analyses of Bones and Teeth Relating to the Human Burials from Skhul.” *Journal of Human Evolution* 49 (3): 316–34.
Guerra-Doce, Elisa. 2015. “The Origins of Inebriation: Archaeological Evidence of the Consumption of Fermented Beverages and Drugs in Prehistoric Eurasia.” *Journal of Archaeological Method and Theory*. https://doi.org/10.1007/s10816-014-9205-z.

Guindon, Stéphane, and Olivier Gascuel. 2003. “A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood.” *Systematic Biology* 52 (5): 696–704.
Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75.
Haak, Wolfgang, Iosif Lazaridis, Nick Patterson, Nadin Rohland, Swapan Mallick, Bastien Llamas, Guido Brandt, et al. 2015. “Massive Migration from the Steppe Was a Source for Indo-European Languages in Europe.” *Nature* 522 (7555): 207–11.
Hagan, Richard W., Courtney A. Hofman, Alexander Hübner, Karl Reinhard, Stephanie Schnorr, Cecil M. Lewis Jr, Krithivasan Sankaranarayanan, and Christina G. Warinner. 2020. “Comparison of Extraction Methods for Recovering Ancient Microbial DNA from Paleofeces.” *American Journal of Physical Anthropology* 171 (2): 275–84.
Hansen, Lea B. S., Henrik M. Roager, Nadja B. Søndertoft, Rikke J. Gøbel, Mette Kristensen, Mireia Vallès-Colomer, Sara Vieira-Silva, et al. 2018. “A Low-Gluten Diet Induces Changes in the Intestinal Microbiome of Healthy Danish Adults.” *Nature Communications* 9 (1): 4630.
Harding, Anthony. 2013. *Salt in Prehistoric Europe*. Sidestone Press.
Hardy, Karen, Stephen Buckley, and Les Copeland. 2018. “Pleistocene Dental Calculus: Recovering Information on Paleolithic Food Items, Medicines, Paleoenvironment and Microbes.” *Evolutionary Anthropology: Issues, News, and Reviews*. https://doi.org/10.1002/evan.21718.
Haririan, Hady, Oleh Andrukhov, Kristina Bertl, Stefan Lettner, Sonja Kierstein, Andreas Moritz, and Xiaohui Rausch-Fan. 2014. “Microbial Analysis of Subgingival Plaque Samples Compared to that of Whole Saliva in Patients with Periodontitis.” *Journal of Periodontology* 85 (6): 819–28.
Harrell, Frank E., Jr, and Maintainer Frank E. Harrell Jr. 2019. “Package ‘hmisc.’” *CRAN2018* 2019: 235–36.
Heiss, Andreas G., Marian Berihuete Azorín, Ferran Antolín, Lucy Kubiak-Martens, Elena Marinova, Elke K. Arendt, Costas G. Biliaderis, et al. 2020. “Mashes to Mashes, Crust to Crust. Presenting a Novel Microstructural Marker for Malting in the Archaeological Record.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0231696.
Heiss, Andreas G., Thorsten Jakobitsch, Silvia Wiesinger, and Peter Trebsche. 2021. “Dig Out, Dig in! Plant-Based Diet at the Late Bronze Age Copper Production Site of Prigglitz-Gasteil (Lower Austria) and the Relevance of Processed Foodstuffs for the Supply of Alpine Bronze Age Miners.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0248287.
Hershkovitz, Israel, Gerhard W. Weber, Rolf Quam, Mathieu Duval, Rainer Grün, Leslie Kinsley, Avner Ayalon, et al. 2018. “The Earliest Modern Humans Outside Africa.” *Science* 359 (6374): 456–59.
Hoopmann, Michael R., Jason M. Winget, Luis Mendoza, and Robert L. Moritz. 2018. “StPeter: Seamless Label-Free Quantification with the Trans-Proteomic Pipeline.” *Journal of Proteome Research*. https://doi.org/10.1021/acs.jproteome.7b00786.
Horz, Hans-Peter, and Georg Conrads. 2011. “Methanogenic Archaea and Oral Infections - Ways to Unravel the Black Box.” *Journal of Oral Microbiology* 3 (February). https://doi.org/10.3402/jom.v3i0.5940.
Hublin, Jean-Jacques, Abdelouahed Ben-Ncer, Shara E. Bailey, Sarah E. Freidline, Simon Neubauer, Matthew M. Skinner, Inga Bergmann, et al. 2017. “New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of Homo Sapiens.” *Nature*. https://doi.org/10.1038/nature22336.
———. 2018. “Author Correction: New Fossils from Jebel Irhoud, Morocco and the Pan-African Origin of Homo Sapiens.” *Nature* 558 (7711): E6.

Huson, Daniel H., Sina Beier, Isabell Flade, Anna Górska, Mohamed El-Hadidi, Suparna Mitra, Hans-Joachim Ruscheweyh, and Rewati Tappu. 2016. “MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data.” *PLOS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1004957.

Huynh, Hong T. T., Vanessa D. Nkamga, Michel Signoli, Stéfan Tzortzis, Romuald Pinguet, Gilles Audoly, Gérard Aboudharam, and Michel Drancourt. 2016. “Restricted Diversity of Dental Calculus Methanogens over Five Centuries, France.” *Scientific Reports*. https://doi.org/10.1038/srep25775.

Huynh, Hong T. T., Marion Pignoly, Vanessa D. Nkamga, Michel Drancourt, and Gérard Aboudharam. 2015. “The Repertoire of Archaea Cultivated from Severe Periodontitis.” *PloS One* 10 (4): e0121565.

Jain, Chirag, Luis M. Rodriguez-R, Adam M. Phillippy, Konstantinos T. Konstantinidis, and Srinivas Aluru. 2018. “High Throughput ANI Analysis of 90K Prokaryotic Genomes Reveals Clear Species Boundaries.” *Nature Communications*. https://doi.org/10.1038/s41467-018-07641-9.

Jensen, Theis Z. T., Jonas Niemann, Katrine Højholt Iversen, Anna K. Fotakis, Shyam Gopalakrishnan, Åshild J. Vågene, Mikkel Winther Pedersen, et al. 2019. “A 5700 Year-Old Human Genome and Oral Microbiome from Chewed Birch Pitch.” *Nature Communications* 10 (1): 5520.

Jersie-Christensen, Rosa R., Liam T. Lanigan, David Lyon, Meaghan Mackie, Daniel Belstrøm, Christian D. Kelstrup, Anna K. Fotakis, et al. 2018. “Quantitative Metaproteomics of Medieval Dental Calculus Reveals Individual Oral Health Status.” *Nature Communications*. https://doi.org/10.1038/s41467-018-07148-3.

Jie, Zhuye, Huihua Xia, Shi-Long Zhong, Qiang Feng, Shenghui Li, Suisha Liang, Huanzi Zhong, et al. 2017. “The Gut Microbiome in Atherosclerotic Cardiovascular Disease.” *Nature Communications* 8 (1): 845.

Johnston, Eric R., Luis M. Rodriguez-R, Chengwei Luo, Mengting M. Yuan, Liyou Wu, Zhili He, Edward A. G. Schuur, et al. 2016. “Metagenomics Reveals Pervasive Bacterial Populations and Reduced Community Diversity across the Alaska Tundra Ecosystem.” *Frontiers in Microbiology*. https://doi.org/10.3389/fmicb.2016.00579.

Jónsson, Hákon, Aurélien Ginolhac, Mikkel Schubert, Philip L. F. Johnson, and Ludovic Orlando. 2013. “mapDamage2.0: Fast Approximate Bayesian Estimates of Ancient DNA Damage Parameters.” *Bioinformatics* 29 (13): 1682–84.

Kang, Dongwan D., Jeff Froula, Rob Egan, and Zhong Wang. 2015. “MetaBAT, an Efficient Tool for Accurately Reconstructing Single Genomes from Complex Microbial Communities.” *PeerJ*. https://doi.org/10.7717/peerj.1165.

Kang, Dongwan D., Feng Li, Edward Kirton, Ashleigh Thomas, Rob Egan, Hong An, and Zhong Wang. 2019. “MetaBAT 2: An Adaptive Binning Algorithm for Robust and Efficient Genome Reconstruction from Metagenome Assemblies.” *PeerJ* 7 (July): e7359.

Karcher, Nicolai, Edoardo Pasolli, Francesco Asnicar, Kun D. Huang, Adrian Tett, Serena Manara, Federica Armanini, et al. 2020. “Analysis of 1321 Eubacterium Rectale Genomes from Metagenomes Uncovers Complex Phylogeographic Population Structure and Subspecies Functional Adaptations.” *Genome Biology* 21 (1): 138.

Katoh, Kazutaka, Kazuharu Misawa, Kei-Ichi Kuma, and Takashi Miyata. 2002. “MAFFT: A Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” *Nucleic Acids Research* 30 (14): 3059–66.

Katoh, Kazutaka, and Daron M. Standley. 2013. “MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability.” *Molecular Biology and Evolution* 30 (4): 772–80.

Keller, Andreas, Angela Graefen, Markus Ball, Mark Matzas, Valesca Boisguerin, Frank Maixner, Petra Leidinger, et al. 2012. “New Insights into the Tyrolean Iceman’s

Origin and Phenotype as Inferred by Whole-Genome Sequencing." *Nature Communications* 3 (February): 698.
Keller, Andrew, Alexey I. Nesvizhskii, Eugene Kolker, and Ruedi Aebersold. 2002. "Empirical Statistical Model to Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search." *Analytical Chemistry* 74 (20): 5383–92.
Kessner, Darren, Matt Chambers, Robert Burke, David Agus, and Parag Mallick. 2008. "ProteoWizard: Open Source Software for Rapid Proteomics Tools Development." *Bioinformatics* 24 (21): 2534–36.
Key, Felix M., Cosimo Posth, Luis R. Esquivel-Gomez, Ron Hübler, Maria A. Spyrou, Gunnar U. Neumann, Anja Furtwängler, et al. 2020. "Emergence of Human-Adapted Salmonella Enterica Is Linked to the Neolithization Process." *Nature Ecology & Evolution* 4 (3): 324–33.
Key, Felix M., Cosimo Posth, Johannes Krause, Alexander Herbig, and Kirsten I. Bos. 2017. "Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication." *Trends in Genetics: TIG* 33 (8): 508–20.
Kircher, Martin, Susanna Sawyer, and Matthias Meyer. 2012. "Double Indexing Overcomes Inaccuracies in Multiplex Sequencing on the Illumina Platform." *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkr771.
Knights, Dan, Justin Kuczynski, Emily S. Charlson, Jesse Zaneveld, Michael C. Mozer, Ronald G. Collman, Frederic D. Bushman, Rob Knight, and Scott T. Kelley. 2011. "Bayesian Community-Wide Culture-Independent Microbial Source Tracking." *Nature Methods*. https://doi.org/10.1038/nmeth.1650.
Kolde, Raivo, and Maintainer Raivo Kolde. 2015. "Package 'pheatmap.'" *R Package* 1 (7): 790.
Konstantinidis, Konstantinos T., and James M. Tiedje. 2005. "Genomic Insights That Advance the Species Definition for Prokaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 102 (7): 2567–72.
Korneliussen, Thorfinn Sand, Anders Albrechtsen, and Rasmus Nielsen. 2014. "ANGSD: Analysis of Next Generation Sequencing Data." *BMC Bioinformatics* 15 (November): 356.
Krause-Kyora, Ben, Julian Susat, Felix M. Key, Denise Kühnert, Esther Bosse, Alexander Immel, Christoph Rinne, et al. 2018. "Neolithic and Medieval Virus Genomes Reveal Complex Evolution of Hepatitis B." *eLife* 7 (May). https://doi.org/10.7554/eLife.36666.
Kupczok, Anne, Horst Neve, Kun D. Huang, Marc P. Hoeppner, Knut J. Heller, Charles M. A. P. Franz, and Tal Dagan. 2018. "Rates of Mutation and Recombination in Siphoviridae Phage Genome Evolution over Three Decades." *Molecular Biology and Evolution* 35 (5): 1147–59.
Laffont, Clémentine, and Pascal Arnoux. 2020. "The Ancient Roots of Nicotianamine: Diversity, Role, Regulation and Evolution of Nicotianamine-like Metallophores." *Metallomics: Integrated Biometal Science* 12 (10): 1480–93.
Langille, Morgan G. I., Jesse Zaneveld, J. Gregory Caporaso, Daniel McDonald, Dan Knights, Joshua A. Reyes, Jose C. Clemente, et al. 2013. "Predictive Functional Profiling of Microbial Communities Using 16S rRNA Marker Gene Sequences." *Nature Biotechnology* 31 (9): 814–21.
Langmead, Ben, and Steven L. Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nature Methods* 9 (4): 357–59.
Langmead, B., and S. L. Salzberg. 2013. "Langmead. 2013. Bowtie2." *Nature Methods* 9: 357–59.
Lazaridis, Iosif, Nick Patterson, Alissa Mittnik, Gabriel Renaud, Swapan Mallick, Karola Kirsanow, Peter H. Sudmant, et al. 2014. "Ancient Human Genomes Suggest Three Ancestral Populations for Present-Day Europeans." *Nature* 513 (7518): 409–13.
Lebreton, François, Abigail L. Manson, Jose T. Saavedra, Timothy J. Straub, Ashlee M. Earl, and Michael S. Gilmore. 2017. "Tracing the Enterococci from Paleozoic

Origins to the Hospital." *Cell* 169 (5): 849–61.e13.
Lepp, Paul W., Mary M. Brinig, Cleber C. Ouverney, Katherine Palm, Gary C. Armitage, and David A. Relman. 2004. "Methanogenic Archaea and Human Periodontal Disease." *Proceedings of the National Academy of Sciences of the United States of America* 101 (16): 6176–81.
Li, Dinghua, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. 2015. "MEGAHIT: An Ultra-Fast Single-Node Solution for Large and Complex Metagenomics Assembly via Succinct de Bruijn Graph." *Bioinformatics* 31 (10): 1674–76.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
Liu, Li, Jiajing Wang, Danny Rosenberg, Hao Zhao, György Lengyel, and Dani Nadel. 2018. "Fermented Beverage and Food Storage in 13,000 Y-Old Stone Mortars at Raqefet Cave, Israel: Investigating Natufian Ritual Feasting." *Journal of Archaeological Science: Reports*. https://doi.org/10.1016/j.jasrep.2018.08.008.
Löytynoja, Ari. 2014. "Phylogeny-Aware Alignment with PRANK." *Methods in Molecular Biology* 1079: 155–70.
Lozupone, C. A., M. Hamady, and S. T. Kelley. 2007. "Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities." *Applied and*. https://aem.asm.org/content/73/5/1576.short.
Ludwig, Wolfgang, Oliver Strunk, Ralf Westram, Lothar Richter, Harald Meier, Yadhukumar, Arno Buchner, et al. 2004. "ARB: A Software Environment for Sequence Data." *Nucleic Acids Research* 32 (4): 1363–71.
Luhmann, Nina, Daniel Doerr, and Cedric Chauve. 2017. "Comparative Scaffolding and Gap Filling of Ancient Bacterial Genomes Applied to Two Ancient Yersinia Pestis Genomes." *Microbial Genomics* 3 (9): e000123.
Lurie-Weinberger, Mor N., Michael Peeri, and Uri Gophna. 2012. "Contribution of Lateral Gene Transfer to the Gene Repertoire of a Gut-Adapted Methanogen." *Genomics* 99 (1): 52–58.
Lv, Long-Xian, Dai-Qiong Fang, Ding Shi, De-Ying Chen, Ren Yan, Yi-Xin Zhu, Yan-Fei Chen, et al. 2016. "Alterations and Correlations of the Gut Microbiome, Metabolism and Immunity in Patients with Primary Biliary Cirrhosis." *Environmental Microbiology* 18 (7): 2272–86.
Mackie, Meaghan, Jessica Hendy, Abigail D. Lowe, Alessandra Sperduti, Malin Holst, Matthew J. Collins, and Camilla F. Speller. 2017. "Preservation of the Metaproteome: Variability of Protein Preservation in Ancient Dental Calculus." *Science and Technology of Archaeological Research* 3 (1): 74–86.
Maixner, Frank, Ben Krause-Kyora, Dmitrij Turaev, Alexander Herbig, Michael R. Hoopmann, Janice L. Hallows, Ulrike Kusebauch, et al. 2016. "The 5300-Year-Old Helicobacter Pylori Genome of the Iceman." *Science* 351 (6269): 162–65.
Maixner, Frank, Dmitrij Turaev, Amaury Cazenave-Gassiot, Marek Janko, Ben Krause-Kyora, Michael R. Hoopmann, Ulrike Kusebauch, et al. 2018. "The Iceman's Last Meal Consisted of Fat, Wild Meat, and Cereals." *Current Biology*. https://doi.org/10.1016/j.cub.2018.05.067.
Makki, Kassem, Edward C. Deehan, Jens Walter, and Fredrik Bäckhed. 2018. "The Impact of Dietary Fiber on Gut Microbiota in Host Health and Disease." *Cell Host & Microbe*. https://doi.org/10.1016/j.chom.2018.05.012.
Manara, Serena, Edoardo Pasolli, Daniela Dolce, Novella Ravenni, Silvia Campana, Federica Armanini, Francesco Asnicar, et al. 2018. "Whole-Genome Epidemiology, Characterisation, and Phylogenetic Reconstruction of Staphylococcus Aureus Strains in a Paediatric Hospital." *Genome Medicine* 10 (1): 82.
Mann, Allison E., Susanna Sabin, Kirsten Ziesemer, Åshild J. Vågene, Hannes Schroeder, Andrew T. Ozga, Krithivasan Sankaranarayanan, et al. 2018. "Differential Preservation of Endogenous Human and Microbial DNA in Dental Calculus and Dentin." *Scientific Reports* 8 (1): 9822.

Marsit, Souhir, Jean-Baptiste Leducq, Éléonore Durand, Axelle Marchant, Marie Filteau, and Christian R. Landry. 2017. “Evolutionary Biology through the Lens of Budding Yeast Comparative Genomics.” *Nature Reviews. Genetics* 18 (10): 581–98.
Martens, Lennart, Matthew Chambers, Marc Sturm, Darren Kessner, Fredrik Levander, Jim Shofstahl, Wilfred H. Tang, et al. 2011. “mzML—a Community Standard for Mass Spectrometry Data.” *Molecular & Cellular Proteomics*. https://doi.org/10.1074/mcp.r110.000133.
Martin, Marcel. 2011. “Cutadapt Removes Adapter Sequences from High-Throughput Sequencing Reads.” *EMBnet.journal*. https://doi.org/10.14806/ej.17.1.200.
McClure, Sarah B., Clayton Magill, Emil Podrug, Andrew M. T. Moore, Thomas K. Harper, Brendan J. Culleton, Douglas J. Kennett, and Katherine H. Freeman. 2018. “Fatty Acid Specific δ13C Values Reveal Earliest Mediterranean Cheese Production 7,200 Years Ago.” *PLOS ONE*. https://doi.org/10.1371/journal.pone.0202807.
McElroy, Kerensa E., Fabio Luciani, and Torsten Thomas. 2012. “GemSIM: General, Error-Model Based Simulator of next-Generation Sequencing Data.” *BMC Genomics* 13 (February): 74.
McGovern, Patrick, Mindia Jalabadze, Stephen Batiuk, Michael P. Callahan, Karen E. Smith, Gretchen R. Hall, Eliso Kvavadze, et al. 2017. “Early Neolithic Wine of Georgia in the South Caucasus.” *Proceedings of the National Academy of Sciences of the United States of America* 114 (48): E10309–18.
McLean, Jeffrey S., Quanhui Liu, John Thompson, Anna Edlund, and Scott Kelley. 2015. “Draft Genome Sequence of ‘ Candidatus Bacteroides Periocalifornicus,’ a New Member of the Bacteriodetes Phylum Found within the Oral Microbiome of Periodontitis Patients.” *Genome Announcements*. https://doi.org/10.1128/genomea.01485-15.
Mehta, Raaj S., Galeb S. Abu-Ali, David A. Drew, Jason Lloyd-Price, Ayshwarya Subramanian, Paul Lochhead, Amit D. Joshi, et al. 2018. “Stability of the Human Faecal Microbiome in a Cohort of Adult Men.” *Nature Microbiology* 3 (3): 347–55.
Meyer, Matthias, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G. Schraiber, et al. 2012. “A High-Coverage Genome Sequence from an Archaic Denisovan Individual.” *Science* 338 (6104): 222–26.
Meyer, M., and M. Kircher. 2010. “Illumina Sequencing Library Preparation for Highly Multiplexed Target Capture and Sequencing.” *Cold Spring Harbor Protocols*. https://doi.org/10.1101/pdb.prot5448.
Mitchell, Piers D. 2017. “Human Parasites in the Roman World: Health Consequences of Conquering an Empire - CORRIGENDUM.” *Parasitology*, May, 1.
Monti, Lucia, Valeria Pelizzola, Milena Povolo, Stefano Fontana, and Giovanna Contarini. 2019. “Study on the Sugar Content of Blue-Veined ‘Gorgonzola’ PDO Cheese.” *International Dairy Journal*. https://doi.org/10.1016/j.idairyj.2019.03.009.
Morozova, Irina, Artem Kasianov, Sergey Bruskin, Judith Neukamm, Martyna Molak, Elena Batieva, Aleksandra Pudło, Frank J. Rühli, and Verena J. Schuenemann. 2020. “New Ancient Eastern European Yersinia Pestis Genomes Illuminate the Dispersal of Plague in Europe.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 375 (1812): 20190569.
Müller, Wolfgang, Henry Fricke, Alex N. Halliday, Malcolm T. McCulloch, and Jo-Anne Wartho. 2003. “Origin and Migration of the Alpine Iceman.” *Science* 302 (5646): 862–66.
Nagarajan, Niranjan, and Mihai Pop. 2013. “Sequence Assembly Demystified.” *Nature Reviews. Genetics* 14 (3): 157–67.
Nayfach, Stephen, Zhou Jason Shi, Rekha Seshadri, Katherine S. Pollard, and Nikos C. Kyrpides. 2019. “New Insights from Uncultivated Genomes of the Global Human Gut Microbiome.” *Nature* 568 (7753): 505–10.
Needham, Brittany D., Sean M. Carroll, David K. Giles, George Georgiou, Marvin

Whiteley, and M. Stephen Trent. 2013. “Modulating the Innate Immune Response by Combinatorial Engineering of Endotoxin.” *Proceedings of the National Academy of Sciences of the United States of America* 110 (4): 1464–69.
Neef, Reinder, René T. J. Cappers, and Renée M. Bekker. 2012. *Digital Atlas of Economic Plants in Archaeology*. Barkhuis.
Ness, F., and M. Aigle. 1995. “RTM1: A Member of a New Family of Telomeric Repeated Genes in Yeast.” *Genetics* 140 (3): 945–56.
Nesvizhskii, Alexey I., Andrew Keller, Eugene Kolker, and Ruedi Aebersold. 2003. “A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry.” *Analytical Chemistry*. https://doi.org/10.1021/ac0341261.
Neukamm, Judith, Alexander Peltzer, and Kay Nieselt. n.d. “DamageProfiler: Fast Damage Pattern Calculation for Ancient DNA.” https://doi.org/10.1101/2020.10.01.322206.
Neukamm, Judith, Saskia Pfrengle, Martyna Molak, Alexander Seitz, Michael Francken, Partick Eppenberger, Charlotte Avanzi, et al. 2020. “2000-Year-Old Pathogen Genomes Reconstructed from Metagenomic Analysis of Egyptian Mummified Individuals.” *BMC Biology*. https://doi.org/10.1186/s12915-020-00839-8.
New, Felicia N., and Ilana L. Brito. 2020. “What Is Metagenomics Teaching Us, and What Is Missed?” *Annual Review of Microbiology* 74 (September): 117–35.
Novo, Maite, Frédéric Bigey, Emmanuelle Beyne, Virginie Galeote, Frédérick Gavory, Sandrine Mallet, Brigitte Cambon, et al. 2009. “Eukaryote-to-Eukaryote Gene Transfer Events Revealed by the Genome Sequence of the Wine Yeast Saccharomyces Cerevisiae EC1118.” *Proceedings of the National Academy of Sciences of the United States of America* 106 (38): 16333–38.
Nurk, Sergey, Dmitry Meleshko, Anton Korobeynikov, and Pavel A. Pevzner. 2017. “metaSPAdes: A New Versatile Metagenomic Assembler.” *Genome Research* 27 (5): 824–34.
Obregon-Tito, Alexandra J., Raul Y. Tito, Jessica Metcalf, Krithivasan Sankaranarayanan, Jose C. Clemente, Luke K. Ursell, Zhenjiang Zech Xu, et al. 2015. “Subsistence Strategies in Traditional Societies Distinguish Gut Microbiomes.” *Nature Communications* 6 (March): 6505.
Ondov, Brian D., Nicholas H. Bergman, and Adam M. Phillippy. 2011. “Interactive Metagenomic Visualization in a Web Browser.” *BMC Bioinformatics* 12 (1): 385.
Ondov, Brian D., Todd J. Treangen, Páll Melsted, Adam B. Mallonee, Nicholas H. Bergman, Sergey Koren, and Adam M. Phillippy. 2016. “Mash: Fast Genome and Metagenome Distance Estimation Using MinHash.” *Genome Biology* 17 (1): 132.
Orlando, Ludovic, M. Thomas P. Gilbert, and Eske Willerslev. 2015. “Reconstructing Ancient Genomes and Epigenomes.” *Nature Reviews. Genetics* 16 (7): 395–408.
Ottoni, Claudio, Meriam Guellil, Andrew T. Ozga, Anne C. Stone, Oliver Kersten, Barbara Bramanti, Stéphanie Porcier, and Wim Van Neer. 2019. “Metagenomic Analysis of Dental Calculus in Ancient Egyptian Baboons.” *Scientific Reports*. https://doi.org/10.1038/s41598-019-56074-x.
Page, Andrew J., Carla A. Cummins, Martin Hunt, Vanessa K. Wong, Sandra Reuter, Matthew T. G. Holden, Maria Fookes, Daniel Falush, Jacqueline A. Keane, and Julian Parkhill. 2015. “Roary: Rapid Large-Scale Prokaryote Pan Genome Analysis.” *Bioinformatics* 31 (22): 3691–93.
Paladin, Alice, Negahnaz Moghaddam, Agnieszka Elzbieta Stawinoga, Inga Siebke, Valentina Depellegrin, Umberto Tecchiati, Sandra Lösch, and Albert Zink. 2020. “Early Medieval Italian Alps: Reconstructing Diet and Mobility in the Valleys.” *Archaeological and Anthropological Sciences*. https://doi.org/10.1007/s12520-019-00982-6.
Parks, Donovan H., Michael Imelfort, Connor T. Skennerton, Philip Hugenholtz, and Gene W. Tyson. 2015. “CheckM: Assessing the Quality of Microbial Genomes Recovered from Isolates, Single Cells, and Metagenomes.” *Genome Research* 25

(7): 1043–55.
Pasolli, Edoardo, Francesco Asnicar, Serena Manara, Moreno Zolfo, Nicolai Karcher, Federica Armanini, Francesco Beghini, et al. 2019. “Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle.” *Cell* 176 (3): 649–62.e20.
Pasolli, Edoardo, Francesca De Filippis, Italia E. Mauriello, Fabio Cumbo, Aaron M. Walsh, John Leech, Paul D. Cotter, Nicola Segata, and Danilo Ercolini. 2020. “Large-Scale Genome-Wide Analysis Links Lactic Acid Bacteria from Food with the Gut Microbiome.” *Nature Communications* 11 (1): 2610.
Pearsall, Deborah M. 2015. *Paleoethnobotany, Third Edition: A Handbook of Procedures*. Left Coast Press.
Peltzer, Alexander, Günter Jäger, Alexander Herbig, Alexander Seitz, Christian Kniep, Johannes Krause, and Kay Nieselt. 2016. “EAGER: Efficient Ancient Genome Reconstruction.” *Genome Biology* 17 (March): 60.
Pérez-Chaparro, P. J., C. Gonçalves, L. C. Figueiredo, M. Faveri, E. Lobão, N. Tamashiro, P. Duarte, and M. Feres. 2014. “Newly Identified Pathogens Associated with Periodontitis: A Systematic Review.” *Journal of Dental Research* 93 (9): 846–58.
Petrov, V. A., I. V. Saltykova, I. A. Zhukova, V. M. Alifirova, N. G. Zhukova, Yu B. Dorofeeva, A. V. Tyakht, et al. 2017. “Analysis of Gut Microbiota in Patients with Parkinson’s Disease.” *Bulletin of Experimental Biology and Medicine* 162 (6): 734–37.
Poinar, H. N., M. Kuch, K. D. Sobolik, I. Barnes, A. B. Stankiewicz, T. Kuder, W. G. Spaulding, V. M. Bryant, A. Cooper, and S. Pääbo. 2001. “A Molecular Analysis of Dietary Diversity for Three Archaic Native Americans.” *Proceedings of the National Academy of Sciences of the United States of America* 98 (8): 4317–22.
Pontes, Ana, Mathias Hutzler, Patrícia H. Brito, and José Paulo Sampaio. 2020. “Revisiting the Taxonomic Synonyms and Populations of Saccharomyces cerevisiae—Phylogeny, Phenotypes, Ecology and Domestication.” *Microorganisms* 8 (6): 903.
Pritchard, Leighton, Rachel H. Glover, Sonia Humphris, John G. Elphinstone, and Ian K. Toth. 2016. “Genomics and Taxonomy in Diagnostics for Food Security: Soft-Rotting Enterobacterial Plant Pathogens.” *Analytical Methods* 8 (1): 12–24.
Qin, Junjie, MetaHIT Consortium, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, et al. 2010. “A Human Gut Microbial Gene Catalogue Established by Metagenomic Sequencing.” *Nature*. https://doi.org/10.1038/nature08821.
Quince, Christopher, Alan W. Walker, Jared T. Simpson, Nicholas J. Loman, and Nicola Segata. 2017. “Shotgun Metagenomics, from Sampling to Analysis.” *Nature Biotechnology* 35 (9): 833–44.
Rambaut, A., and N. C. Grassly. 1997. “Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution along Phylogenetic Trees.” *Computer Applications in the Biosciences: CABIOS* 13 (3): 235–38.
Rambaut, Andrew, Alexei J. Drummond, Dong Xie, Guy Baele, and Marc A. Suchard. 2018. “Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.” *Systematic Biology* 67 (5): 901–4.
Rambaut, Andrew, Simon Y. W. Ho, Alexei J. Drummond, and Beth Shapiro. 2009. “Accommodating the Effect of Ancient DNA Damage on Inferences of Demographic Histories.” *Molecular Biology and Evolution* 26 (2): 245–48.
Rambaut, Andrew, Tommy T. Lam, Luiz Max Carvalho, and Oliver G. Pybus. 2016. “Exploring the Temporal Structure of Heterochronous Sequences Using TempEst (formerly Path-O-Gen).” *Virus Evolution* 2 (1): vew007.
Rascovan, Nicolás, Karl-Göran Sjögren, Kristian Kristiansen, Rasmus Nielsen, Eske Willerslev, Christelle Desnues, and Simon Rasmussen. 2019. “Emergence and Spread of Basal Lineages of Yersinia Pestis during the Neolithic Decline.” *Cell* 176

(1-2): 295–305.e10.
Reinhard, K. J., L. F. Ferreira, F. Bouchet, L. Sianto, J. M. F. Dutra, A. Iniguez, D. Leles, et al. 2013. “Food, Parasites, and Epidemiological Transitions: A Broad Perspective.” *International Journal of Paleopathology* 3 (3): 150–57.
Reis, Mario dos, Philip C. J. Donoghue, and Ziheng Yang. 2016. “Bayesian Molecular Clock Dating of Species Divergences in the Genomics Era.” *Nature Reviews. Genetics* 17 (2): 71–80.
Renaud, Gabriel, Viviane Slon, Ana T. Duggan, and Janet Kelso. 2015. “Schmutzi: Estimation of Contamination and Endogenous Mitochondrial Consensus Calling for Ancient DNA.” *Genome Biology*. https://doi.org/10.1186/s13059-015-0776-0.
Reschreiter, Hans, and Kerstin Kowarik. 2019. “Bronze Age Mining in Hallstatt. A New Picture of Everyday Life in the Salt Mines and Beyond.” *Archaeologia Austriaca* 103: 99–136.
Robinson, D. F., and L. R. Foulds. 1979. “Comparison of Weighted Labelled Trees.” In *Combinatorial Mathematics VI*, 119–26. Springer Berlin Heidelberg.
Rohland, Nadin, Heike Siedel, and Michael Hofreiter. 2010. “A Rapid Column-based Ancient DNA Extraction Method for Increased Sample Throughput.” *Molecular Ecology Resources*. https://doi.org/10.1111/j.1755-0998.2009.02824.x.
Ropars, Jeanne, Estelle Didiot, Ricardo C. Rodríguez de la Vega, Bastien Bennetot, Monika Coton, Elisabeth Poirier, Emmanuel Coton, Alodie Snirc, Stéphanie Le Prieur, and Tatiana Giraud. 2020. “Domestication of the Emblematic White Cheese-Making Fungus Penicillium Camemberti and Its Diversification into Two Varieties.” *Current Biology: CB* 30 (22): 4441–53.e4.
Ropars, Jeanne, Ricardo C. Rodríguez de la Vega, Manuela López-Villavicencio, Jérôme Gouzy, Erika Sallet, Émilie Dumas, Sandrine Lacoste, et al. 2015. “Adaptive Horizontal Gene Transfers between Multiple Cheese-Associated Fungi.” *Current Biology: CB* 25 (19): 2562–69.
Rosa, Bruce A., Taniawati Supali, Lincoln Gankpala, Yenny Djuardi, Erliyani Sartono, Yanjiao Zhou, Kerstin Fischer, et al. 2018. “Differential Human Gut Microbiome Assemblages during Soil-Transmitted Helminth Infections in Indonesia and Liberia.” *Microbiome*. https://doi.org/10.1186/s40168-018-0416-5.
Rosenbloom, Kate R., Joel Armstrong, Galt P. Barber, Jonathan Casper, Hiram Clawson, Mark Diekhans, Timothy R. Dreszer, et al. 2015. “The UCSC Genome Browser Database: 2015 Update.” *Nucleic Acids Research* 43 (Database issue): D670–81.
Sabin, Susanna, Alexander Herbig, Åshild J. Vågene, Torbjörn Ahlström, Gracijela Bozovic, Caroline Arcini, Denise Kühnert, and Kirsten I. Bos. 2020a. “A Seventeenth-Century Mycobacterium Tuberculosis Genome Supports a Neolithic Emergence of the Mycobacterium Tuberculosis Complex.” *Genome Biology*. https://doi.org/10.1186/s13059-020-02112-1.
———. 2020b. “A Seventeenth-Century Mycobacterium Tuberculosis Genome Supports a Neolithic Emergence of the Mycobacterium Tuberculosis Complex.” *Genome Biology* 21 (1): 201.
Salque, Mélanie, Peter I. Bogucki, Joanna Pyzel, Iwona Sobkowiak-Tabaka, Ryszard Grygiel, Marzena Szmyt, and Richard P. Evershed. 2013. “Earliest Evidence for Cheese Making in the Sixth Millennium Bc in Northern Europe.” *Nature*. https://doi.org/10.1038/nature11698.
Samuel, Buck S., Elizabeth E. Hansen, Jill K. Manchester, Pedro M. Coutinho, Bernard Henrissat, Robert Fulton, Philippe Latreille, Kung Kim, Richard K. Wilson, and Jeffrey I. Gordon. 2007. “Genomic and Metabolic Adaptations of Methanobrevibacter Smithii to the Human Gut.” *Proceedings of the National Academy of Sciences of the United States of America* 104 (25): 10643–48.
Scholz, Matthias, Davide Albanese, Kieran Tuohy, Claudio Donati, Nicola Segata, and Omar Rota-Stabelli. 2020. “Large Scale Genome Reconstructions Illuminate Wolbachia Evolution.” *Nature Communications* 11 (1): 5235.

Schöttke, Niklas, and Frank Rögener. 2021. “Cold Mashing - Analysis and Optimization of Extraction Processes at Low Temperatures in the Brewing Process.” *E3S Web of Conferences*. https://doi.org/10.1051/e3sconf/202124701036.
Schuenemann, Verena J., Charlotte Avanzi, Ben Krause-Kyora, Alexander Seitz, Alexander Herbig, Sarah Inskip, Marion Bonazzi, et al. 2018. “Ancient Genomes Reveal a High Diversity of Mycobacterium Leprae in Medieval Europe.” *PLoS Pathogens* 14 (5): e1006997.
Schuenemann, Verena J., Pushpendra Singh, Thomas A. Mendum, Ben Krause-Kyora, Günter Jäger, Kirsten I. Bos, Alexander Herbig, et al. 2013. “Genome-Wide Comparison of Medieval and Modern Mycobacterium Leprae.” *Science* 341 (6142): 179–83.
Scott, Ashley, Robert C. Power, Victoria Altmann-Wendling, Michal Artzy, Mario A. S. Martin, Stefanie Eisenmann, Richard Hagan, et al. 2021. “Exotic Foods Reveal Contact between South Asia and the Near East during the Second Millennium BCE.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.2014956117.
Seemann, Torsten. 2014. “Prokka: Rapid Prokaryotic Genome Annotation.” *Bioinformatics* 30 (14): 2068–69.
Segata, Nicola, Jacques Izard, Levi Waldron, Dirk Gevers, Larisa Miropolsky, Wendy S. Garrett, and Curtis Huttenhower. 2011. “Metagenomic Biomarker Discovery and Explanation.” *Genome Biology* 12 (6): R60.
Seitz, Alexander, and Kay Nieselt. 2017. “Improving Ancient DNA Genome Assembly.” *PeerJ* 5 (April): e3126.
Sender, Ron, Shai Fuchs, and Ron Milo. 2016. “Are We Really Vastly Outnumbered? Revisiting the Ratio of Bacterial to Host Cells in Humans.” *Cell* 164 (3): 337–40.
Shao, Yan, Samuel C. Forster, Evdokia Tsaliki, Kevin Vervier, Angela Strang, Nandi Simpson, Nitin Kumar, et al. 2019. “Stunted Microbiota and Opportunistic Pathogen Colonization in Caesarean-Section Birth.” *Nature* 574 (7776): 117–21.
Shillito, Lisa-Marie, John C. Blong, Eleanor J. Green, and Eline N. van Asperen. 2020. “The What, How and Why of Archaeological Coprolite Analysis.” *Earth-Science Reviews* 207 (103196): 103196.
Shteynberg, David, Eric W. Deutsch, Henry Lam, Jimmy K. Eng, Zhi Sun, Natalie Tasman, Luis Mendoza, Robert L. Moritz, Ruedi Aebersold, and Alexey I. Nesvizhskii. 2011. “iProphet: Multi-Level Integrative Analysis of Shotgun Proteomic Data Improves Peptide and Protein Identification Rates and Error Estimates.” *Molecular & Cellular Proteomics*. https://doi.org/10.1074/mcp.m111.007690.
Singh, Robbert Gurdeep, Alessandro Tanca, Antonio Palomba, Felix Van der Jeugt, Pieter Verschaffelt, Sergio Uzzau, Lennart Martens, Peter Dawyndt, and Bart Mesuere. 2019. “Unipept 4.0: Functional Analysis of Metaproteome Data.” *Journal of Proteome Research*. https://doi.org/10.1021/acs.jproteome.8b00716.
Skoglund, Pontus, Jan Storå, Anders Götherström, and Mattias Jakobsson. 2013. “Accurate Sex Identification of Ancient Human Remains Using DNA Shotgun Sequencing.” *Journal of Archaeological Science*. https://doi.org/10.1016/j.jas.2013.07.004.
Smits, Samuel A., Jeff Leach, Erica D. Sonnenburg, Carlos G. Gonzalez, Joshua S. Lichtman, Gregor Reid, Rob Knight, et al. 2017. “Seasonal Cycling in the Gut Microbiome of the Hadza Hunter-Gatherers of Tanzania.” *Science* 357 (6353): 802–6.
Socransky, S. S., A. D. Haffajee, M. A. Cugini, C. Smith, and R. L. Kent Jr. 1998. “Microbial Complexes in Subgingival Plaque.” *Journal of Clinical Periodontology* 25 (2): 134–44.
Sonnenburg, Erica D., and Justin L. Sonnenburg. 2019. “The Ancestral and Industrialized Gut Microbiota and Implications for Human Health.” *Nature Reviews. Microbiology* 17 (6): 383–90.
Spanogiannopoulos, Peter, Elizabeth N. Bess, Rachel N. Carmody, and Peter J.

Turnbaugh. 2016. “The Microbial Pharmacists within Us: A Metagenomic View of Xenobiotic Metabolism.” *Nature Reviews. Microbiology* 14 (5): 273–87.
Spindler, Konrad. 2013. *The Man In The Ice*. Hachette UK.
Spitaels, Freek, Anneleen Diane Wieme, Isabel Snauwaert, Luc De Vuyst, and Peter Vandamme. 2017. “Microbial Ecology of Traditional Beer Fermentations.” *Brewing Microbiology: Current Research, Omics and Microbial Ecology*. https://doi.org/10.21775/9781910190616.07.
Spitaels, Freek, Anneleen D. Wieme, Maarten Janssens, Maarten Aerts, Heide-Marie Daniel, Anita Van Landschoot, Luc De Vuyst, and Peter Vandamme. 2014. “The Microbial Diversity of Traditional Spontaneously Fermented Lambic Beer.” *PLoS ONE*. https://doi.org/10.1371/journal.pone.0095384.
Spyrou, Maria A., Marcel Keller, Rezeda I. Tukhbatova, Christiana L. Scheib, Elizabeth A. Nelson, Aida Andrades Valtueña, Gunnar U. Neumann, et al. 2019. “Phylogeography of the Second Plague Pandemic Revealed through Analysis of Historical Yersinia Pestis Genomes.” *Nature Communications* 10 (1): 4470.
Spyrou, Maria A., Rezeda I. Tukhbatova, Michal Feldman, Joanna Drath, Sacha Kacki, Julia Beltrán de Heredia, Susanne Arnold, et al. 2016. “Historical Y. Pestis Genomes Reveal the European Black Death as the Source of Ancient and Modern Plague Pandemics.” *Cell Host & Microbe* 19 (6): 874–81.
Spyrou, Maria A., Rezeda I. Tukhbatova, Chuan Chao Wang, Aida Andrades Valtueña, Aditya K. Lankapalli, Vitaly V. Kondrashin, Victor A. Tsybin, et al. 2018. “Analysis of 3800-Year-Old Yersinia Pestis Genomes Suggests Bronze Age Origin for Bubonic Plague.” *Nature Communications* 9 (1): 1–10.
Stamatakis, Alexandros. 2014. “RAxML Version 8: A Tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies.” *Bioinformatics* 30 (9): 1312–13.
Stöllner, Th. 2004. “Salz, Salzgewinnung, Salzhandel, Archäologisch.” *Hoops, J. , Reallexikon* 26: 357–79.
Sun, Jiufeng, De Wu, and Changwen Ke. 2017. “Molecular Phylogenetic Analyses of Imported Zika Virus Genomes in China.” *The Lancet* 390 (December): S36.
Takeuchi, Y., M. Umeda, M. Sakamoto, Y. Benno, Y. Huang, and I. Ishikawa. 2001. “Treponema Socranskii, Treponema Denticola, and Porphyromonas Gingivalis Are Associated with Severity of Periodontal Tissue Destruction.” *Journal of Periodontology* 72 (10): 1354–63.
Tang, Jun-Ni, Zhi-Guang Zeng, Hong-Ning Wang, Tai Yang, Peng-Ju Zhang, Yu-Ling Li, An-Yun Zhang, et al. 2008. “An Effective Method for Isolation of DNA from Pig Faeces and Comparison of Five Different Methods.” *Journal of Microbiological Methods*. https://doi.org/10.1016/j.mimet.2008.07.014.
Tett, Adrian, Kun D. Huang, Francesco Asnicar, Hannah Fehlner-Peach, Edoardo Pasolli, Nicolai Karcher, Federica Armanini, et al. 2019. “The Prevotella Copri Complex Comprises Four Distinct Clades Underrepresented in Westernized Populations.” *Cell Host & Microbe* 26 (5): 666–79.e7.
Thomas, Andrew Maltez, Paolo Manghi, Francesco Asnicar, Edoardo Pasolli, Federica Armanini, Moreno Zolfo, Francesco Beghini, et al. 2019. “Metagenomic Analysis of Colorectal Cancer Datasets Identifies Cross-Cohort Microbial Diagnostic Signatures and a Link with Choline Degradation.” *Nature Medicine* 25 (4): 667–78.
Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. “CLUSTAL W: Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice.” *Nucleic Acids Research* 22 (22): 4673–80.
Tito, Raul Y., Dan Knights, Jessica Metcalf, Alexandra J. Obregon-Tito, Lauren Cleeland, Fares Najar, Bruce Roe, et al. 2012. “Insights from Characterizing Extinct Human Gut Microbiomes.” *PloS One* 7 (12): 1–8.
Tjaden, Britta, André Plagens, Christine Dörr, Bettina Siebers, and Reinhard Hensel. 2006. “Phosphoenolpyruvate Synthetase and Pyruvate, Phosphate Dikinase of Thermoproteus Tenax: Key Pieces in the Puzzle of Archaeal Carbohydrate

Metabolism." *Molecular Microbiology* 60 (2): 287–98.
Truong, Duy Tin, Eric A. Franzosa, Timothy L. Tickle, Matthias Scholz, George Weingart, Edoardo Pasolli, Adrian Tett, Curtis Huttenhower, and Nicola Segata. 2015. "MetaPhlAn2 for Enhanced Metagenomic Taxonomic Profiling." *Nature Methods* 12 (10): 902–3.
Truong, Duy Tin, Adrian Tett, Edoardo Pasolli, Curtis Huttenhower, and Nicola Segata. 2017. "Microbial Strain-Level Population Structure and Genetic Diversity from Metagenomes." *Genome Research* 27 (4): 626–38.
Turnbaugh, Peter J., Micah Hamady, Tanya Yatsunenko, Brandi L. Cantarel, Alexis Duncan, Ruth E. Ley, Mitchell L. Sogin, et al. 2009. "A Core Gut Microbiome in Obese and Lean Twins." *Nature* 457 (7228): 480–84.
Turnbaugh, Peter J., Ruth E. Ley, Micah Hamady, Claire M. Fraser-Liggett, Rob Knight, and Jeffrey I. Gordon. 2007. "The Human Microbiome Project." *Nature* 449 (7164): 804–10.
Ungar, Peter S., and Matt Sponheimer. 2011. "The Diets of Early Hominins." *Science* 334 (6053): 190–93.
Utter, Daniel R., Jessica L. Mark Welch, and Gary G. Borisy. 2016. "Individuality, Stability, and Variability of the Plaque Microbiome." *Frontiers in Microbiology* 7 (April): 564.
Vågene, Åshild J., Alexander Herbig, Michael G. Campana, Nelly M. Robles García, Christina Warinner, Susanna Sabin, Maria A. Spyrou, et al. 2018. "Salmonella Enterica Genomes from Victims of a Major Sixteenth-Century Epidemic in Mexico." *Nature Ecology and Evolution* 2 (3): 520–28.
Vangay, Pajau, Abigail J. Johnson, Tonya L. Ward, Gabriel A. Al-Ghalith, Robin R. Shields-Cutler, Benjamin M. Hillmann, Sarah K. Lucas, et al. 2018. "US Immigration Westernizes the Human Gut Microbiome." *Cell* 175 (4): 962–72.e10.
Velsko, Irina M., James A. Fellows Yates, Franziska Aron, Richard W. Hagan, Laurent A. F. Frantz, Louise Loe, Juan Bautista Rodriguez Martinez, et al. 2019. "Microbial Differences between Dental Plaque and Historic Dental Calculus Are Related to Oral Biofilm Maturation Stage." *Microbiome*. https://doi.org/10.1186/s40168-019-0717-3.
Velsko, Irina M., Katherine A. Overmyer, Camilla Speller, Lauren Klaus, Matthew J. Collins, Louise Loe, Laurent A. F. Frantz, et al. 2017. "The Dental Calculus Metabolome in Modern and Historic Samples." *Metabolomics: Official Journal of the Metabolomic Society* 13 (11): 134.
Visconti, Alessia, Caroline I. Le Roy, Fabio Rosa, Niccolò Rossi, Tiphaine C. Martin, Robert P. Mohney, Weizhong Li, et al. 2019. "Interplay between the Human Gut Microbiome and Host Metabolism." *Nature Communications* 10 (1): 4505.
Voidarou, C., E. Bezirtzoglou, A. Alexopoulos, S. Plessas, C. Stefanis, I. Papadopoulos, S. Vavias, et al. 2011. "Occurrence of Clostridium Perfringens from Different Cultivated Soils." *Anaerobe* 17 (6): 320–24.
Wampach, Linda, Anna Heintz-Buschart, Joëlle V. Fritz, Javier Ramiro-Garcia, Janine Habier, Malte Herold, Shaman Narayanasamy, et al. 2018. "Birth Mode Is Associated with Earliest Strain-Conferred Gut Microbiome Functions and Immunostimulatory Potential." *Nature Communications* 9 (1): 5091.
Wang, Jiajing, Li Liu, Terry Ball, Linjie Yu, Yuanqing Li, and Fulai Xing. 2016. "Revealing a 5,000-Y-Old Beer Recipe in China." *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.1601465113.
Wang, Renke, Aida Kaplan, Lihong Guo, Wenyuan Shi, Xuedong Zhou, Renate Lux, and Xuesong He. 2012. "The Influence of Iron Availability on Human Salivary Microbial Community Composition." *Microbial Ecology* 64 (1): 152–61.
Warinner, Christina, Alexander Herbig, Allison Mann, James A. Fellows Yates, Clemens L. Weiß, Hernán A. Burbano, Ludovic Orlando, and Johannes Krause. 2017. "A Robust Framework for Microbial Archaeology." *Annual Review of Genomics and Human Genetics* 18 (August): 321–56.

Warinner, Christina, João F. Matias Rodrigues, Rounak Vyas, Christian Trachsel, Natallia Shved, Jonas Grossmann, Anita Radini, et al. 2014. “Pathogens and Host Immunity in the Ancient Human Oral Cavity.” *Nature Genetics* 46 (4): 336–44.
Warinner, Christina, Camilla Speller, and Matthew J. Collins. 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.
Weissensteiner, Hansi, Dominic Pacher, Anita Kloss-Brandstätter, Lukas Forer, Günther Specht, Hans-Jürgen Bandelt, Florian Kronenberg, Antonio Salas, and Sebastian Schönherr. 2016. “HaploGrep 2: Mitochondrial Haplogroup Classification in the Era of High-Throughput Sequencing.” *Nucleic Acids Research* 44 (W1): W58–63.
Weyrich, Laura S., Sebastian Duchene, Julien Soubrier, Luis Arriola, Bastien Llamas, James Breen, Alan G. Morris, et al. 2017. “Neanderthal Behaviour, Diet, and Disease Inferred from Ancient DNA in Dental Calculus.” *Nature* 544 (7650): 357–61.
Wibowo, Marsha C., Zhen Yang, Maxime Borry, Alexander Hübner, Kun D. Huang, Braden T. Tierney, Samuel Zimmerman, et al. 2021. “Reconstruction of Ancient Microbial Genomes from the Human Gut.” *Nature*, May. https://doi.org/10.1038/s41586-021-03532-0.
Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer.
Wingett, Steven W., and Simon Andrews. 2018. “FastQ Screen: A Tool for Multi-Genome Mapping and Quality Control.” *F1000Research* 7 (August): 1338.
Wood, B., and M. Collard. 1999. “The Human Genus.” *Science* 284 (5411): 65–71.
Wu, Hong, Kiyoshi Ito, and Hitoshi Shimoi. 2005. “Identification and Characterization of a Novel Biotin Biosynthesis Gene in Saccharomyces Cerevisiae.” *Applied and Environmental Microbiology* 71 (11): 6845–55.
Yatsunenko, Tanya, Federico E. Rey, Mark J. Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, et al. 2012. “Human Gut Microbiome Viewed across Age and Geography.” *Nature* 486 (7402): 222–27.
Ze, Xiaolei, Yonit Ben David, Jenny A. Laverde-Gomez, Bareket Dassa, Paul O. Sheridan, Sylvia H. Duncan, Petra Louis, et al. 2015. “Unique Organization of Extracellular Amylases into Amylosomes in the Resistant Starch-Utilizing Human Colonic Firmicutes Bacterium Ruminococcus Bromii.” *mBio* 6 (5): e01058–15.
Ze, Xiaolei, Sylvia H. Duncan, Petra Louis, and Harry J. Flint. 2012. “Ruminococcus Bromii Is a Keystone Species for the Degradation of Resistant Starch in the Human Colon.” *The ISME Journal* 6 (8): 1535–43.
Zhao, Shijie, Tami D. Lieberman, Mathilde Poyet, Kathryn M. Kauffman, Sean M. Gibbons, Mathieu Groussin, Ramnik J. Xavier, and Eric J. Alm. 2019. “Adaptive Evolution within Gut Microbiomes of Healthy People.” *Cell Host & Microbe* 25 (5): 656–67.e8.
Ziesemer, Kirsten A., Allison E. Mann, Krithivasan Sankaranarayanan, Hannes Schroeder, Andrew T. Ozga, Bernd W. Brandt, Egija Zaura, et al. 2015. “Intrinsic Challenges in Ancient Microbiome Reconstruction Using 16S rRNA Gene Amplification.” *Scientific Reports* 5 (1): 16498.

Acknowledgments

No man is an island. Any of the work I discussed above would never have been accomplished without the expertise contributed by everyone involved. I would firstly thank my supervisor, Prof. Nicola Segata, from whom I have learnt so much how to conduct novel investigations, how to collaborate with colleagues from diverse backgrounds, and most importantly, how to be persistent and resilient as a devoted researcher. These qualities I acquired from him armed me with enough confidence to confront any challenges in the future. I also would like to thank my co-supervisor, Prof. Omar Rota-Stabelli, who taught me how to think out of box and how to enjoy research purely without letting strong purpose stop the seed of novelty sprouting. I am also grateful for his tolerance of my stubbornness at some moments in past years. Another important person in my doctoral training I would like to thank is Dr. Adrian Tett. I learnt so much from him how to present research work in public and write a beautiful story for a wider readership. I wish to express my gratitude to my key collaborators, Dr. Maixner Frank and Lena Granehäl, who generously shared the precious ancient metagenomic samples so that I could make many intriguing discoveries. Through a close collaboration with them in the last four years, their expertise in the ancient DNA field taught me so much the underlying significance of our biological findings. And more importantly, I learnt how to be a good collaborator.

Thanks to all my colleagues and friends at Segata Lab and Omar's group. Last four years would not have been the same without you. Any coffee break, beer time, and big/small parties we had together never stopped reminding me how privileged I have been to go through this journey with you all around.

In the end, I would like to thank my ultra liberal family for supporting me to practice whatever wacky ideas and to encourage me to be bold in the journey of my life, without any doubt and complaint. I would never be who I really am without being born and raised in such a family.