**Dental calculus as a tool to study the evolution of the oral microbiome in mammals**

Jaelle C. Brealey[1], Henrique G. Leitão[1], Wenbo Xu[1], Love Dalén[2], Katerina Guschanski[1]

[1]Department of Ecology and Genetics, Animal Ecology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden

[2]Department of Bioinformatics and Genetics, Swedish Museum of Natural History, 104 05, Stockholm, Sweden

**Correspondence:**

Jaelle Brealey, Department of Ecology and Genetics, Animal Ecology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden. Email: jaelle.brealey@ebc.uu.se

Katerina Guschanski, Department of Ecology and Genetics, Animal Ecology, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Uppsala, Sweden. Email: katerina.guschanski@ebc.uu.se

**Abstract**

The microbiome and its host have a long history of co-evolution, defined by a complex interplay between environmental and host genetic factors. Yet, we have a very limited understanding of how the host-associated microbiome is influenced by extrinsic and intrinsic long-term processes, such as climate change and host demography. To this end, we need to study microbiome changes through time, over known historical events that have affected the host. Dental calculus, the calcified form of the microbial biofilm that forms on vertebrate teeth, is one of few microbial substrates that preserves virtually unchanged through time. Metagenomics sequencing of ancient dental calculus has provided valuable insights in humans, but few studies have attempted to characterise the dental calculus microbiome of non-human animals. We establish dental calculus as a valuable research tool

for the study of host-associated microbiome evolution. We recover a recognisable oral microbiome signature from species as diverse as gorillas, bears and reindeer, and identify a subset of microorganisms that may be unique to each host species. We detect oral pathogens in specimens with evidence of oral disease, identify microbial metabolic pathways, characterise antibiotic resistance genes and recover host DNA using metagenomic analyses of dental calculus. In addition, we identify important strategies and current limitations for working with historical metagenomics data from non-human mammals. With this new tool in hand, we can now study changes in microbiome composition and function as result of extrinsic and intrinsic processes in host history over millennia.

**Introduction**

The microbiome and its host have a long history of co-evolution, influenced by a complex interplay between host genetics and environmental factors with direct relevance for host health [1–4]. Interspecific comparisons between modern host populations and experimental work in model organisms have provided important insights into a magnitude of factors shaping the host-associated microbial community. Such studies have uncovered phylogenetic patterns of co-diversification between the host and the microbiome, and the effect of diet, reproductive status, and host genotype on microbiome composition [5–9]. However, there is limited understanding of how the host-associated microbiome is influenced by extrinsic and intrinsic long-term processes that span multiple host generations, such as climate change and host population demography. The use of microbial fossils represents a powerful and so far little-explored approach to studying microbiome evolution, allowing us to quantify how these complex communities and individual members change through time. Unfortunately, host-associated microbial communities rarely fossilise, with the exception of coprolites and dental calculus [10]. Coprolites, petrified faecal matter which capture components of the gut microbiome, are rare and show poor DNA preservation, because faeces are rapidly colonised by environmental bacteria and decompose quickly after deposition [10]. More readily available and

well-preserved is dental calculus, the calcified form of the dental plaque microbial biofilm which forms on mammalian teeth. Dental plaque comprises an important component of the oral microbiome, as well as capturing invading pathogens, host cell molecules and biomolecules related to diet [11–14]. It undergoes periodic mineralisation during an individual's life, which effectively seals the biotic material from external contamination and preserves it virtually unchanged through time [10]. Furthermore, dental calculus is readily available from specimens in natural history collections and from archaeological samples. In humans, shotgun sequencing of dental calculus has revealed shifts in oral microbiome composition associated with cultural transitions [15] and tracked pathogen co-evolution through time [16].

While the focus to date has been on human-microbiome and human-pathogen co-evolution, many other mammals produce dental calculus. Collections in natural history museums provide the opportunity to study the effects of human-driven changes in animal populations over the last 200 years, including population declines and the introduction of pollutants like antibiotics [17–20]. Few studies to date have investigated microbiome changes in modern wild animal populations [21–23], and historical populations remain virtually unexplored. Working with historical microbial DNA in any host species is challenging, and non-model host species pose additional difficulties. Historical DNA is typically of low quantity, fragmented, chemically modified, and easily contaminated by modern DNA sources, such as the personnel processing the samples, trace environmental microorganisms and contaminants in reagents [24–26]. Characterisation of historical microbiomes has relied on public reference databases, which are heavily biased towards microbial species with medical and agricultural significance [27], restricting microbiome research of non-model host organisms and environments. The extent to which dental calculus can be studied in non-human mammals thus remains unclear.

To establish dental calculus as a standard research tool for the study of host-associated microbiome evolution in diverse mammalian species, we characterised the historical dental calculus microbiome of three evolutionarily distant mammalian species with distinct physiology and ecology: European reindeer (*Rangifer tarandus*), Scandinavian brown bear (*Ursus arctos*), and eastern gorilla (*Gorilla beringei*). Reindeer are group-living herbivores with a multigastric digestive system including the rumen, and specialised hypsodont molars adapted to an abrasive, fibrous diet. Brown bears are solitary omnivores with brachydont molars more adapted for a partially carnivorous diet. Gorillas are group-living folivores and specialised hindgut fermenters, and were included to aid characterisation of previously unexplored microbiomes from the other two host species. The close evolutionary relationship between gorillas and humans (the major source of the microbial reference databases) and the previous successful reconstruction of the chimpanzee oral microbiome from dental calculus [13] prompted us to use gorilla dental calculus to verify our laboratory and bioinformatics approaches. We successfully recover an oral microbiome signature from all three host species, detect potential oral pathogens and antibiotic resistance genes, and characterise microbial metabolic pathways and host DNA in dental calculus samples. Our study solidly establishes dental calculus as a valuable tool for studies of host-microbiome co-evolution and opens doors for future investigations of oral pathogen evolution and anthropogenic effects on wildlife and the environment.

**Results and Discussion**

*Oral microbiome signature can be successfully recovered from dental calculus of non-human mammals and contains host-specific taxa*

Total DNA extraction, Illumina shotgun sequencing and metagenomics analyses were carried out on dental calculus collected from five reindeer, six brown bear and two eastern gorilla specimens dating from 1861 to 1961 (Table S1). Three extractions were performed for each gorilla dental calculus sample: surface decontamination with UV exposure, surface decontamination with an EDTA wash,

and no surface decontamination. We also processed and sequenced two blank negative controls from the extraction and library preparation stages, to quantify background contamination.

Following data preprocessing, which included read merging and removal of adapter, duplicate and host reads (see Materials and Methods), the reads were passed on to Kraken2 [28] for microbial taxonomic assignment. On average, we assigned 2.6 million reads per sample (range 0.04-9.0 million reads). Putative contaminating taxa were identified by presence of longer DNA fragment lengths [24], presence in extraction and library preparation blank control samples [25], and those at higher absolute abundance in low biomass samples [29]. After removal of these contaminants, 931 microbial taxa remained (798 bacteria, 19 archaea and 96 viruses).

To determine the proportion of the microbial communities in our samples that could be assigned to either well-characterised oral microbiome taxa or putative contaminating sources, we used SourceTracker [30], a Bayesian classification tool. As potential sources, we used the human plaque microbiome [1,31], representing an oral microbiome signature, the human skin [32] and gut [1,31] microbiomes as the most likely contaminants in museum environments and during sample processing, an independently sequenced laboratory reagent microbiome [25], and a soil microbiome [33], representing environmental contamination. While there was substantial variation between samples, an oral microbiome signature was successfully recovered from all three host species (Figure 1). Generally, the oral microbiome signature was highest in the two gorilla individuals (Figure S1) as expected given their close evolutionary relationship with humans, the species from which the oral microbiome reference database was constructed. Human skin and laboratory reagents were the most common contamination sources, and were particularly high in specimens where surface decontamination of the calculus was not performed before DNA extraction (Figure S2). One such bear sample (Ua6) contained high levels of contaminants (>70%) and no detectable oral microbiome signature, and was therefore excluded from all further analyses. For both gorilla specimens, samples

decontaminated with an EDTA wash before DNA extraction yielded the highest proportion of oral taxa and were therefore the only gorilla samples retained for downstream analyses.

The oral microbiome signature differed between host species, particularly using presence/absence measures (Figure 2A; for abundance-based comparison, see Figure S3 and Table S2). Over 30% of the variation between samples could be explained by host species ($p$=0.002) and < 10% each by surface decontamination, sequencing depth or human DNA contamination (Table S2). Microbial taxa that showed the greatest host species specificity (i.e. were only present in a single host species) were generally associated with either the mammalian oral microbiome or other mammalian body sites (Table 1). To independently verify the host species-specific differences between reindeer and bear, we used a random forests classifier on the reindeer and bear presence/absence data (after the exclusion of one reindeer sample (Rt5), which was consistently identified as an outlier in all analyses (Figure S4)). The correct host species could be assigned in 88.9% of cases, and the most important taxa for determining the host species generally included oral taxa, such as *Streptococcus* species, which were present and highly abundant in the bears (Figure 2B-C). Reindeer also differed from bears and gorillas in measures of alpha diversity (Figure S5), with significantly lower Shannon indices in the majority of samples (median ± interquartile range: 1.762 ± 0.274 (reindeer) vs 3.776 ± 0.048 (gorilla) vs 3.796 ± 0.578 (bear), p=0.003).

The differences in microbiome composition between the three host species, and particularly the greater similarity between bears and gorillas compared to reindeer (Figure 2), may reflect differences in host-associated microbiome communities between monogastric mammals like gorillas and bears, and ruminants like reindeer. A few of the microbial taxa specific to the reindeer were identified as *Methanobrevibacter* species associated with the rumen of domesticated bovines and ovines [34]. Ruminants regurgitate large amounts of rumen material into the oral cavity when chewing cud, and a study in sheep has found that oral swabs contain a proportion of rumen-associated

microbes [35], suggesting that dental calculus of ruminants may also capture a subset of the rumen microbiome.

*Functional analyses provide evidence for oral pathogens in wild animals*

By performing shotgun sequencing, we were able to gain a first insight into the functional potential of the microbial communities captured in dental calculus. We used the HumanN2 pipeline [36] to assign KEGG orthologues [37] to our filtered microbial reads, and cluster these functions into MetaCyc metabolic pathways [38]. In contrast to the taxonomic results, we observed no clear functional differences between host species (Figure S6 and Table S2. Instead, we identified core functional pathways shared across individuals and host species (Figure 3). The core pathways identified in this study were generally involved in essential metabolic processes, such as energy production and biomolecule synthesis, functions common to most living organisms [31]. Conserved functional properties despite differing community compositions is a common phenomenon in ecosystems, including the human microbiome [1,31]. However, this inference could also be driven by limitations of the functional reference databases, which are mostly centred on global mechanisms [39] and do not allow for more fine-grained functional characterisation.

Among general functions that can be expected from an oral microbial community, we identified a number of metabolic pathways related to carbohydrate degradation and acid production, which are commonly performed by bacteria that colonise the oral cavity. These pathways are also associated with the emergence and progression of dental caries [40] and were particularly abundant in one bear specimen (Ua9), which was sampled from a caries lesion with evidence of periodontitis (Figure 4B). Correspondingly, the most abundant microbial taxa in Ua9 included members of the *Lactobacillus casei* group and *Bifidobacterium* genus, species that have been associated with active caries lesions in humans [41] (Figure 4A). Among the studied specimens, one additional bear individual had signs of caries and periodontitis (Ua7). While other cariogenic bacteria (e.g. *Streptococcus mitis* group

species) and those associated with the periodontal 'red complex' pathogens (*Porphyromonas gingivalis*, *Treponema denticola*, and *Tannerella forsythia*) were identified in Ua7, these bacterial species also occurred in bear specimens with no signs of dental diseases. One possible explanation for this result is the turn-over of bacterial communities related to the stage of the infection (active versus inactive). For example, both *Lactobacillus* and *Bifidobacterium* species found in the oral cavity are more frequently associated with active caries lesions [41]. Furthermore, Ua7 was sampled from a "healthy" tooth, rather than a caries lesion, and both taxonomically (Figure 4A) and functionally (Figure 3) appeared more similar to the other "healthy" bear samples. The biogeography of the oral microbiome in humans has been found to be site-specific [42], suggesting that even within a diseased individual, the oral microbiome could differ between healthy and carious sites. Increased sample numbers from healthy individuals and those with oral diseases (including paired samples from different sites within an individual) are needed to substantiate these preliminary findings. Nevertheless, our results demonstrate that dental calculus can be used to study oral diseases in wild animals and to characterise bacterial players involved in oral pathologies.

*Antimicrobial resistance genes are present in wild animal microbiomes*

The presence of bacteria carrying AMR factors has been documented in the human oral microbiome [15,43]. We therefore investigated whether oral microbial communities of wild animals contain AMR genes and whether abundance differs between host species. To this end, we investigated both global AMR potential in the entire microbial community and the diversity of AMR genes encoded by two oral bacterial pathogens with reported multi-drug resistance, *Acinetobacter baumannii* and *Klebsiella pneumoniae* [44]. To survey global AMR potential, we blasted the contamination-filtered reads against the Comprehensive Antibiotic Resistance Database (CARD) [45] and assigned the top match for each read to its respective gene family under the Antibiotic Resistance Ontology (ARO). To identify AMR genes encoded by *A. baumannii* and *K. pneumoniae*, which were both present in our studied host species (Figure S7), we extracted reads that mapped to these bacterial genomes with

bwa [46,47]. We confirmed that reads from both bacterial species exhibited characteristic deamination patterns consistent with historical DNA damage (Figure 5B-C) and subsequently processed these reads in CARD, as above. Both globally and in *A. baumannii* and *K. pneumoniae*, gorillas and bears tended to have greater AMR potential compared to reindeer (Figure 6). With our limited sample numbers, we observed no obvious increase in AMR potential in samples collected after the start of industrial-scale antibiotics production in the 1940s [48]. However, our analyses add support to previous studies on historical human and permafrost microbiomes by demonstrating that many of the underlying molecular mechanisms conferring resistance to modern antibiotics have existed in the environment long before widespread antibiotics production [15,49]. This investigation opens doors for the use of dental calculus as a tool for the study of AMR evolution in wild animals from diverse geographic locations through time, and for determining the potential of wildlife to serve as reservoirs for clinically-relevant AMR factors.

*Host DNA is preserved in dental calculus*

In addition to microbial remains, dental calculus also entraps host cells and thus can serve as a source for host DNA [14,50]. In our samples, a median of 3.1% ± 11.1% (range 0.06-90.6%) of post-processing reads mapped to their respective host genomes, with a median average coverage per site of 0.002X ± 0.02 (range 0.0001-0.4X) (Table S3). A previous study using shotgun sequencing of human dental calculus recovered lower levels of host DNA (< 1% of reads) from calculus [14]. It is unclear whether these disparities are due to intrinsic differences in calculus composition between human and non-human host species, differences in sampling and study methods, or the presence of more readily identifiable authentic host reads, as we are working with non-human mammals, which are more easily distinguishable from the most likely mammalian contamination source – the human. We were also able to recover complete mitochondrial genomes (>1X average coverage per site) for one gorilla, three reindeer and two bear samples. Mitochondrial genomes provide a rich source of information about the host evolutionary history, host population structure and changes in host

population sizes over time [18,50]. With deeper metagenomics sequencing and/or employment of target capture techniques, the relatively high proportion of host reads allows for investigations into the importance of host genetic diversity and demography in shaping microbiome composition from the same dental calculus sample.

*Dealing with current limitations: Contamination and database biases*

Contamination of samples with modern DNA is a major problem facing all historical genomic and metagenomic studies [24]. In this study, we used a combination of laboratory procedures and bioinformatics processing steps to reduce contamination and its effects on the results. By subjecting samples to two different decontamination approaches, we determined that a quick incubation of the dental calculus sample in a 0.5M EDTA solution reduced contaminants without loss of an oral microbiome signature (Figure 1) and thus offers an efficient decontamination strategy. We have also employed a multi-step bioinformatics approach to identify contaminant taxa, which consists of removal of taxa with long DNA fragment lengths [24], those detected in the extraction and library preparation blank control samples [25], and those showing an inverse relationship between absolute abundance and the total number of reads per sample [29]. Removal of these contaminants from our data systematically increased the proportion of the bacterial communities being attributed to the oral microbiome and decreased contaminant microbiome signals (Figure S8). Even after filtering, some proportion of human skin taxa were reported by SourceTracker in our samples (Figure 1, Figure S8). Several bacterial species colonise multiple niches within the host [1,31], which can obscure distinction of a genuine signal from a likely contaminant. For example, *Streptococcus mitis*, *Staphylococcus epidermidis* and *Corynebacterium matruchotii* are found in the human mouth, nostrils and skin [1]. This limitation is not specific to our dataset, but poses greater problems for studies based on historical samples that are expected to be subjected to contamination, like those sampling from museum specimens.

For select bacterial taxa, we confirmed the presence of DNA damage patterns consistent with historical DNA (Figure 5). We included host DNA damage as a control for expected damage patterns (Figure 5A) and a known environmental bacterium (*Pseudomonas psychrophila*) as an example of a likely modern contaminant (Figure 5D).

One limitation of our study is the reliance on microbial reference databases. These databases are heavily biased towards microbial species with medical or agricultural significance [27], restricting read-based analyses of metagenomics data from non-human hosts. A large proportion of microbial taxa identified in our samples remained unassigned to any source microbiome by SourceTracker (Figure 1). Although some of these taxa may be members of other contaminating sources not included in our analysis, it can be expected that by studying a novel environment (the non-human oral microbiome) we will encounter unique microbial taxa that do not share similarity with the human oral microbiome. In the absence of a dedicated reference database from our study species, an alternative approach to taxonomic assignment in metagenomics studies is *de novo* assembly of metagenome-assembled genomes (MAGs) [51]. Given the fragmented and damaged nature of ancient DNA, this technique poses great challenges for historical microbiome studies. Yet, one study of invasive *Salmonella enterica* strains responsible for paratyphoid fever in humans successfully recovered a historical *S. enterica* MAG from deeply sequenced bone and tooth samples from an 800-year-old skeleton, as well as another likely endogenous MAG from an novel *Eubacterium* species from dental calculus from the same specimen [16]. We attempted MAG assembly on our dental calculus data with MEGAHIT [52] and MetaBAT [53], and successfully recovered six high quality MAGs (>90% genome completeness and <5% strain contamination as estimated by CheckM [54]) from four samples (Table 2). The highest quality genome was identified as *Lactobacillus paracasei* from the carious bear sample Ua9. The presence of typical damage patterns suggests that the MAG is likely endogenous (Figure 4C). Furthermore, *L. paracasei* was detected in Ua9 at a relative abundance of almost 30% based on analyses with Kraken2 (Figure 4A, Table 2). Metabolic pathways performed by

*L. paracasei*, such as 'lactose and galactose degradation' and 'hexitol fermentation to lactate, formate, ethanol and acetate', were identified at high abundance in the functional analysis (Figure 3), and this bacterium is abundant in active caries lesions [41]. Thus, all of these lines of evidence support the accurate identification of the *de novo* MAG.

The other five recovered MAGs had less similarity to bacteria references in the Genome Taxonomy Database [55] and were only identified at the genus level. Two appear to be related to potentially oral bacteria (*Streptococcus* and *Haemophilus*), suggesting that with deeper sequencing MAG recovery may be able to complement read-based analyses of historical microbiome composition. This is particularly important in non-model species, where reference database bias is a problem. Ultimately, however, dedicated reference databases constructed from modern, high-quality samples from the host species and body sites of interest are needed to more completely characterise historical microbiomes.

**Conclusion**

With the development of high-throughput sequencing techniques and advancements in methods for metagenomic analyses of ancient samples, the time is ripe to venture into the study of the diversity of environmental and host-associated microbial communities. The unique temporal perspective provided by historical and ancient samples allows us to study many fundamental evolutionary processes, including those with direct relevance to human and ecosystem health. Questions of interest include how the host-associated microbiome evolves through multi-generational changes in the host environment, such as changes in host population demography and genetic diversity, and the invasion of new habitats or changes in competitive regimes. Temporal sampling also provides insights into disease emergence and the progression of AMR evolution in the host-associated microbiome, which is of interest to both evolutionary biologists and the public health sector, since wild animal populations can act as sources and reservoirs for emerging zoonotic pathogens [56,57] and

contribute to the spread of AMR [58,59]. Our study describes a rigorous roadmap for the analyses of historical microbiomes and identifies important current limitations. We believe that in the future, this integrative approach of studying the microbiome and its host will be applied to many different questions in ecology and evolution.

## Materials and methods

### *Sample collection*

Dental calculus was collected from two eastern gorilla (*Gorilla beringei*) specimens from the Royal Museum for Central Africa (Brussels, Belgium), and five reindeer (*Rangifer tarandus*) and six brown bear (*Ursus arctos*) specimens from the Swedish Natural History Museum (Stockholm, Sweden). Skulls were macroscopically examined for dental calculus deposits and evidence of oral diseases. Any calculus was removed from the surfaces of the teeth with disposable sterile scalpel blades into sterile siliconized tubes.

### *Sample processing and DNA extraction*

All laboratory protocols were performed in a dedicated ancient DNA laboratory following stringent procedures to minimise contamination [24]. Initially, calculus samples were processed without surface decontamination (Table S1). After testing surface decontamination protocols on the two gorilla samples (Table S1), each calculus sample was exposed to UV light (245 nm) for 10 min and subsequently washed with 500 µl of 0.5M ethylenediaminetetraacetate (EDTA) for 30 seconds. The pellet was taken forward for DNA extraction. Sample weight ranged from < 5 mg to 20 mg. DNA extractions were performed following a silica-based method [60]. Briefly, samples were incubated overnight at 37°C in extraction buffer (0.45M EDTA, 0.25 mg/ml Proteinase K). The DNA from the supernatant was combined with binding buffer (3M sodium acetate, 5M guanidine-hydrochloride, 40% (v/v) isopropanol, 0.05% (v/v) Tween-20) and processed through the spin columns from High Pure Viral Nucleic Acid Large Volume kits (Roche, Switzerland). Purified DNA was eluted in either 45

µl of EB buffer (10 mM tris-hydrochloride (pH 8.0) (Qiagen, The Netherlands) or 45 µl of TE buffer (10 mM tris-hydrochloride (pH 8.0), 1 mM EDTA), both supplemented with 0.05% (v/v) Tween-20 (Table S1). Ten blanks were included during five extraction batches, and one was randomly selected to be taken forward to sequencing.

*Library preparation and sequencing*

Double-stranded Illumina libraries were prepared following an established protocol [61] and included the double-barcoding double-indexing strategy to guard against index hopping and retain absolute certainty about sample of origin [62]. Briefly, blunt-end repair reactions were performed using 20 µl of each extract and purified using MinElute columns with elutions in 22 µl of EB buffer (Qiagen, The Netherlands). Adapters containing inline 7 bp barcodes (Table S4) were ligated to both ends of the blunt-ended DNA, which was subsequently purified with MinElute columns and eluted in 22 µl of EB buffer. After the adapter fill-in reaction, *Bst* 2.0 polymerase (New England BioLabs, USA) was inactivated with a 15 min incubation at 80°C. At this stage, some samples (Table S1) went through an additional cleaned-up with MinElute columns. Seven library blanks were included in four library preparation batches, and one was randomly selected to be taken forward to sequencing. The adapter-ligated libraries were quantified using a real-time PCR assay with preHyb primers [63] (Table S4) and the estimated fragment number was used to approximate the number of indexing PCR cycles needed for sequencing. Libraries were double-indexed with unique P5 and P7 indices so that each sample had a unique barcode-index combination (Table S1). Indexing PCR reactions were performed with 18 µl of adapter-ligated library in 50 µl reactions, with 1 µl PfuTurbo $C_x$ hotstart polymerase (2.5 U/µl, Agilent Technologies, USA), 5 µl 10X PfuTurbo $C_x$ reaction buffer, 0.5 µl dNTP mix (25 mM) and 1 µl of each indexing primer (10 µM). After an initial incubation for 2 min at 95°C, 12 cycles of 30 sec at 95°C, 30 sec at 59°C and 1 min at 72°C were performed, followed by a final step of 10 min at 72°C. Reactions were purified with MinElute columns and eluted in 10 µl of EBT buffer. The indexed libraries were quantified using a real-time PCR assay with i7 and i5 indexing

primers [63] (Table S4) and library DNA fragment length distribution was determined by the 2200 TapeStation system. The mean fragment length after library preparation and excluding the 148 bp adapter sequences was 70 bp, as would be expected for degraded historical samples [64]. Five microliters of each sample library were pooled along with a randomly selected extraction blank and library blank. Size selection was performed on the pooled library with AMPure XP beads (Beckman Coulter, USA) and the purified library eluted in 36 µl of EB buffer. The final pooled library was quantified using a Qubit High Sensitivity fluorometer and on the 2200 TapeStation system. The pooled library was first sequenced by SciLifeLab Uppsala on 2 lanes of the Illumina HiSeq 2500 using paired-end 125 bp read length v4 chemistry, followed by an additional 2 lanes on the Illumina HiSeq 2500 in rapid mode using paired-end 100 bp read length v2 chemistry.

*Data processing*

Sequencing data were demultiplexed and assigned to each sample with an in-house python script based on the unique combination of barcodes and indices. Paired-end reads were merged and adapters and low quality terminal bases (phred scores ≤ 30) were removed with AdapterRemoval v2.2.2 [65]. Barcode sequences were removed from the 5' and 3' ends of merged reads with an in-house python script. The set of forward reads from the unmerged read pairs (i.e. pairs that did not contain overlapping regions of at least 11 bp between the forward and reverse reads) were also retained for analyses. The 5' barcode was removed with an in-house python script and the 3' barcode with any remaining adapter sequence removed with AdapterRemoval. Reads from the two lanes within the same sequencing run were concatenated into a single file per sample. Merged reads from the two separate runs were also concatenated into a single file per sample. Reads with a length < 30 bp were filtered out with AdapterRemoval and reads with mean base quality < 30 were filtered out with PrinSeq-Lite v0.20.4 [66]. Duplicate reads were removed with an in-house python script, which takes into account fragment length and sequence identity. The Illumina sequencing control phage PhiX was spiked into our sequencing runs and has previously been reported to have been

erroneously integrated into many microbial genomes [23,67]. Reads were therefore mapped to PhiX (accession: GCA_000819615.1) with bwa mem v0.7.17 [46,47] and the unmapped reads retained with SAMTools v1.9 [68] and BEDTools v2.21.0 [69]. To remove reads originating from the host organism and human contamination, we mapped all reads in a sample to a combined reference consisting of the human genome [70] (RefSeq accession: GCF_000001405.38) and the respective host genome (GCF_000151905.2 (*Gorilla gorilla gorilla*) [71], GCF_003584765.1 (*U. arctos horribilis*) [72] and GCA_004026565.1 (*Rangifer tarandus*)) with bwa mem. The unmapped reads were retained with SAMTools for downstream microbial taxonomic analyses.

*Taxonomic assignment*

Merged and unmerged unmapped reads were assigned taxonomy using Kraken2 v2.0.7 [28] with the standard Kraken2 database (all archaea, bacteria, viruses and human taxa in RefSeq; built 2019-03-01). Kraken2 is a taxonomic classifier that uses *k*-mers to map reads to the lowest common ancestor in the reference database. This means that if an assignment is not possible at the strain or species level, the algorithm will move up the tree until the read can be assigned to a higher taxonomic rank. We used Kraken-biom (https://github.com/smdabdoub/kraken-biom) to extract the summarised number of reads assigned at the genus and species levels (i.e. strain level assignments were pushed up to species level). These assignments were taken into R (https://www.R-project.org/) for further processing and analysis.

*Identifying contamination*

We used several complementary approaches to identify and remove contaminating bacterial taxa.

*SourceTracker analysis*: Potential contribution of source microbiomes to samples was estimated with SourceTracker v1.0 [30] in R, using the Kraken2 genus- and species-level assignments. Source

sequencing reads were processed through the same pipeline as sample reads, and included soil [33], human skin [32], human gut [1,31], human oral [1,31] and laboratory reagent [25] microbiomes (Table S5).

*Fragment length*: Based on the distribution of fragment lengths in our libraries as determined by TapeStation, we expected the majority of our authentic historical microbial reads to be approximately 70 bp in length. The forward and reverse reads of a sequenced pair should thus overlap and be successfully merged by AdapterRemoval. We therefore expected that modern contaminating taxa with longer fragment length would be predominantly found among the unmerged reads. Thus, we juxtaposed the read counts of each taxon per sample obtained from the merged and unmerged reads (Figure S9), and calculated the difference between the two values. Species-level taxa with a difference in absolute abundance between unmerged and merged reads larger than 1.5 multiplied by the interquartile range of all differences were filtered out as putative contaminants. While it is possible that certain microbial structural features, e.g. the thick cell wall in Gram-positive bacteria, might bias preservation towards specific taxa and thereby affect our fragment length analysis [24], a recent study of microbial DNA perseveration in human dental calculus found no association between cellular structures and microbial DNA fragmentation [14]. Nonetheless, we confirmed that the majority of species-level taxa present in this list were non-host-associated environmental microorganisms through a literature search.

*Presence in blanks*: SourceTracker analysis demonstrated that the two blank samples contained taxa associated with soil and human skin microbiomes (Figure S8). We therefore removed all species-level taxa that were present in the blanks from the calculus data.

*Statistical identification of contaminants*: Contamination derived from environmental sources (including the laboratory) is expected to be present in samples at approximately similar absolute amounts and will therefore be disproportionally more abundant in samples with low numbers of

total sequenced reads [25,73]. This concept has been implemented in the R package decontam [29], which was used to identify and remove all species-level taxa that showed an inverse relationship between taxon abundance and estimated total number of sequences included in the sequencing pool per sample.

*Removal of contaminants from reads:* For all downstream analyses, reads mapping to all species-level taxa (bacteria, archaea and viruses) identified as putative contaminants were removed from the fastq files. To this end, one reference genome for each taxon was downloaded from GenBank (Table S6; accessed 2019-03-05), reads were mapped to the reference genomes with bwa mem v0.7.17, relaxing the mismatch parameter (-B) to 3 in order to map reads from closely related strains. Only unmapped reads were passed onto further analyses.

*Microbial analyses*

*Genome size normalisation:* Genome size can influence the number of reads assigned to a given taxon in metagenomic data and hence taxon abundance estimates [74]. To account for this bias, the number of reads per taxon within a sample was normalised by dividing by the estimated average genome size of each respective taxon. Estimated average prokaryotic and viral genome sizes were calculated using publicly available genome sizes from the RefSeq database (accessed 2019-02-15) [75]. In cases where no genome size data were available for a given species, the average genome size of taxa in that genus was used.

*Abundance filtering:* Taxa present at < 0.03% relative abundance (genome size normalised value divided by sum of genome size normalised values in a sample) were removed, as filtering of low-abundance species reduces false-positive taxonomic assignments [76]. The filtering threshold was selected by testing a series of thresholds commonly applied in metagenomics studies, ranging from 0.01 – 0.1% [77](Figure S10). From this analysis, we identified the threshold (0.03%) yielding a

microbial community with a complexity which was most similar to what has been observed in previous dental calculus [14] and oral microbiome studies [78–81] (i.e. approximately 100-300 taxa).

*Abundance normalization:* Lastly, in high-throughput sequencing microbiome datasets the total number of reads obtained is limited by the sequencing instrument. Thus, organisms cannot be treated as independent, as an increase in the proportion of one taxon must necessarily lead to a decrease in the proportion of another taxon [82]. To account for the compositional nature of the dataset, we applied the centred log-ratio (CLR) transformation, which calculates the logarithm of the ratio between the number of reads assigned to a taxon in a sample and the geometric mean of all reads assigned to all taxa in that sample. Because log transformation is only possible for positive values, we dealt with 0 count values by adding a pseudo-count to the genome size normalised value for every taxon in every sample. Due to the genome size normalisation, we set the pseudo-count as the inverse of the average genome size for all taxa.

*Statistical analyses*: We used the R package vegan [83] for diversity estimates. The Shannon index [84] was used to estimate alpha diversity, via the diversity function on the raw read count data (i.e. prior to genome size normalisation and CLR transformation), because the calculation of the metric required positive integers. Differences between host species were investigated with an AIC-based stepwise regression to determine the best-fit general linear model, with surface decontamination, sequencing depth (number of unmapped reads per sample) and proportion of human reads per sample included as covariates.

Beta diversity was investigated based on the presence/absence of microbial taxa. Jaccard distances, calculated using the vegdist function, were used for ordination with non-metrical multidimensional scaling (NMDS), and permutational multivariate analysis of variance (PERMANOVA) with the adonis function in vegan. Host species, surface decontamination, number of unmapped reads per sample,

and proportion of human reads per sample were included as covariates in the adonis model. To determine whether differences in within-group variation between host species was biasing inferences of a host-specific oral microbiome signature, a distance-based test for homogeneity of multivariate dispersions was performed with the vegan function betadisper. No such differences were detected, adding confidence to the PERMANOVA results. The same set of beta analyses were repeated using Euclidean distances calculated on the CLR normalised abundance data (Figure S3, Table S2).

To identify taxa which discriminated between host species, we carried out a random forest classification based on presence/absence data using the R package randomForest [85]. For this analysis, the gorilla samples were excluded due to low sample size. During initial testing, it became apparent that one reindeer sample (Rt5) was driving most of the variation in the dataset, due to its unique community composition. Thus, we removed this individual from this analysis, as the goal here was to identify taxa unique to each species, not individuals. The built-in out-of-bag error, in this case an estimation of how often an individual was correctly assigned to a species, was used to test and validate the model.

To classify taxa as oral, we investigated a subset of taxa unique to each host species and shared by > 50% of samples from this species. As the two gorilla samples did not share any taxa, we investigated all of the taxa unique to each of the two samples. We then compared these sets of taxa to the Human Oral Microbiome Database (HOMD) [86]. Taxa not present in the HOMD were manually classified as 'oral', 'host-associated' (in non-oral mammalian microbiomes) or 'not host-associated' through a literature search.

*Functional analysis*

*Classification*: The functional genic content of the microbial community in dental calculus was characterised by running the contamination-filtered reads through the HUMAnN2 pipeline [36]. HUMAnN2 first uses the taxonomic profiler MetaPhlAn2 [87] to assign taxonomy to reads based on single copy marker genes. The successfully identified species are used to build a sample-specific reference database of functionally annotated pangenomes containing the identified species, to which all sample reads are mapped. All reads which do not map to the pangenome (i.e. those with ambiguous taxonomy) are aligned with DIAMOND [88] against the UniRef90 [89] database. The mappings are weighted by quality and sequence length to estimate species-specific and total community gene family abundance. Metabolic pathways are also reconstructed based on genes annotated to metabolic enzymes, and the pathway abundance and coverage are reported. Because MetaPhlAn2 relies on microbial marker genes identified from well-characterised species, the majority of our reads were unable to be taxonomically assigned. We therefore focus on the total community results (i.e. not stratified by microbial species) and report pathway abundance/coverage based on MetaCyc [38] information.

*Analysis*: CLR normalisation, NMDS ordination and PERMANOVA were carried out for the pathway abundances as for the microbial analyses outlined above. Core pathways were defined as those containing > 50% of the required enzymes in a sample (i.e. per sample pathway coverage > 0.5).

*Antimicrobial resistance profiles*

*Global*: Contamination-filtered reads were aligned to the Comprehensive Antibiotic Resistance Database (CARD) v3.0.1 (modified 2019-02-19) [45] with blast v2.7.1+ [90,91] using default parameters. The Antibiotic Resistance Ontology (ARO) accession number associated with each CARD sequence was used to obtain the gene family, resistance mechanism and resistance drug classes of each sequence. Where reads matched multiple sequences in the CARD database, the best hit was identified based on highest bit score. Where multiple hits had the same bit score, we compared the

ARO terms and if all hits shared the same ARO information, we randomly chose one hit to carry forward. There were a few cases where the AMR gene family differed between hits but remained within the broader gene group (e.g. beta-lactamases). These cases were assigned as "unclassified" within that broader group and the drug class information was concatenated. Once every read was assigned to a single ARO accession, we calculated the number of reads per ARO accession per sample and normalised it by sample sequencing depth (number of reads post-contamination-filtering). The abundance of ARO gene families for each sample was calculated by summing across the ARO accession abundances associated with each gene family.

*Targeted*: All complete genomes for *K. pneumoniae* (n=332) and *A. baumannii* (n=133) available at NCBI Genome (accessed 2019-03-28) were downloaded and concatenated into one 'pangenome' for each species. Contamination-filtered reads were mapped to each pangenome with bwa mem. All mapped reads were extracted with SAMTools (samtools fastq -F4) and aligned to CARD as per the global AMR analysis outlined above. The abundance of *K. pneumoniae* or *A. baumannii* ARO gene families for each sample was calculated by summing across ARO accession abundances normalised by number of reads mapping to each bacteria.

*Metagenome-assembled genome recovery*

We were unable to taxonomically assign many reads in our samples using the read-based reference database methods. Yet, these reads are particularly interesting, as some of them are likely to represent uncharacterised microbial biodiversity within the new sampled environment – the wild mammal oral microbiome. We therefore attempted to recover MAGs from our samples following a similar strategy as described in Zhou *et al.* 2018 [16] and Parks *et al.* 2017 [51]. Reads were assembled into contigs with MEGAHIT [52], which uses succinct *de Bruijn* graphs to *de novo* assemble complex metagenomics data. Sample depth profiles were generated by mapping reads back to the contigs with BBMap v38.08 (Bushnell B., *BBMap*, https://sourceforge.net/projects/bbmap/). Contigs > 1500

bp in length were grouped into genome bins based on co-abundance and sequence composition probabilities using MetaBAT2 [53], where each bin theoretically represents the genome of a single strain. The quality of each bin was assessed using the lineage-specific workflow in CheckM [54]. CheckM identifies the lineage of each bin within the reference genome tree and uses lineage-specific marker genes to estimate the genome completeness and strain contamination of each bin. High quality bins were classified as those with > 90% completeness and < 5% contamination [51]. The taxonomic lineage of each high quality bin, as defined in the Genome Taxonomy Database (GTDB) [55], was determined with GTDB-Tk (https://github.com/ecogenomics/gtdbtk) using the classify workflow.

*Host genome recovery*

*Genome coverage*: The average host genome coverage of each sample was calculated with SAMTools (samtools depth -a -Q10) based on the reads that bwa mem aligned to the host reference genome (see *Data processing*).

*Mitogenome assembly*: Reads that mapped to the host genome were extracted (samtools fastq -F4) and mapped with bwa mem to the host mitochondrial genome (brown bear: NC_003427.1; western gorilla: NC_011120.1; white-tailed deer: NC_015247.1 [the reindeer genome assembly did not contain an annotated mitochondrial genome, hence a close relative was used instead]). Average coverage was calculated with SAMTools, as per *Genome coverage* outlined above.

*mapDamage profiles*

Deamination patterns are a characteristic feature of damaged DNA fragments. As deamination rates increase over time [92], these patterns may be used to confirm the authenticity of microbiome data from ancient and historical specimens. To this end, we investigated damage patterns in two oral bacterial pathogens, *Acinetobacter baumanii* and *Klebsiella pneumoniae*, and also *Pseudomonas*

*psychrophila*, an aquatic bacteria and likely modern contaminant [93]. We calculated deamination rates using mapDamage [94] via the EAGER pipeline [95], mapping non-host reads pooled across reindeer samples to the pangenomes of *A. baumanii* and *K. pneumoniae* and to the representative genome of *P. psychrophila* (GCA_000282975.1). For the reindeer samples, host reads were pooled and mapped against the reindeer reference genome, to verify the presence of deamination in the study samples and thus validate this approach for confirming the authenticity of an oral microbiome signature. For *Lactobacillus paracasei*, contamination-filtered reads from bear Ua9 were mapped to the reference genome of *L. paracasei* identififed by GTDB-Tk (GCF_000829035.1).

**Acknowledgements**

We thank Daniela Kalthoff and Emmanuel Gilissen for providing access to the mammalian museum specimens, and James Fellows Yates, Franziska Aron, Johanna von Seth and Tatiana Feuerborn for laboratory and analysis advice. Sequencing was performed by the SNP&SEQ Technology Platform in Uppsala. The facility is part of the National Genomics Infrastructure (NGI) Sweden and Science for Life Laboratory. The SNP&SEQ Platform is also supported by the Swedish Research Council and the Knut and Alice Wallenberg Foundation. We also acknowledge the National Bioinformatics Infrastructure (NBIS) for providing computational resources to this project. The project was supported by the Formas grant (2016-00835) and the National Projects grant (NP00039) to KG.

**Tables**

Table 1. Likely origin[1] of microbial taxa unique to each host species.

| Host species | No. of taxa | Unique taxa | Unique taxa shared by >50% of samples | Oral | Host-associated | Other |
|---|---|---|---|---|---|---|
| Gorilla[2] | 304 | 71 | 0 | 23.9% | 23.9% | 60.6% |
| Bear | 571 | 277 | 30 | 36.7% | 53.3% | 10.0% |
| Reindeer | 371 | 172 | 5 | 0.0% | 60.0% | 40.0% |

[1] Microbial taxa were classified as oral, non-oral but host-associated in mammals, and not associated with mammals, based on literature searches.

[2] Percentages for gorillas were calculated based on the number of unique taxa, whereas for reindeer and bear percentages were calculated based on the number of unique shared taxa.

Table 2. Metagenome-assembled genomes recovered from dental calculus.

| Sample | Taxonomy | Completeness (%) | Contamination (%) | Proportion of reads identified by Kraken[1] |
|---|---|---|---|---|
| Ua9 | *Lactobacillus C paracasei* | 95.79 | 0.15 | 0.272 |
| Rt13 | *Haemophilus B* sp. | 94.28 | 1.86 | 0.006 |
| Rt7 | *Streptococcus* sp. | 92.69 | 1.25 | 0.665 |
| Ua14 | *Petrimonas* sp. | 95.30 | 1.64 | 0.004 |
| Ua14 | 'UBA11524' sp. (Christensenellales) | 93.93 | 2.42 | NA |
| Ua14 | *Basfia* sp. | 90.37 | 3.30 | 0.012 |

[1] Summed across species at the genus level, except for *L. paracasei*, for which species-level assignment was used.

**Figures**

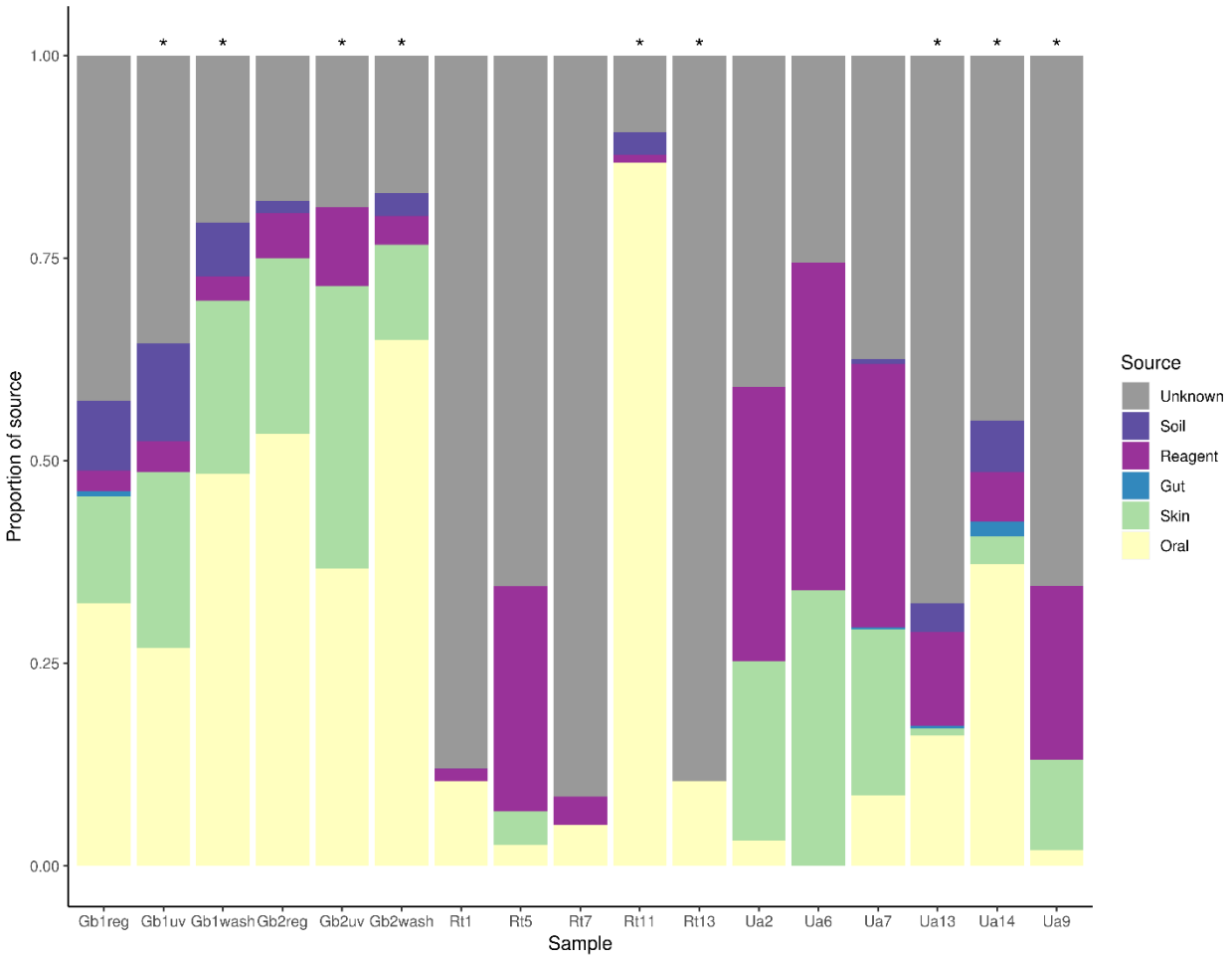


Figure 1. An oral microbiome signature can be recovered from the dental calculus of gorillas (Gb), reindeer (Rt), and brown bears (Ua). Proportions of source contributions to the microbial communities (identified taxonomically at the species level) contained in the dental calculus samples estimated by SourceTracker, using modern human oral, human skin, human gut, laboratory reagent, and soil microbiome datasets as sources. Asterisks (*) indicate samples in which surface decontamination (EDTA wash and/or UV exposure) was performed before DNA extraction.

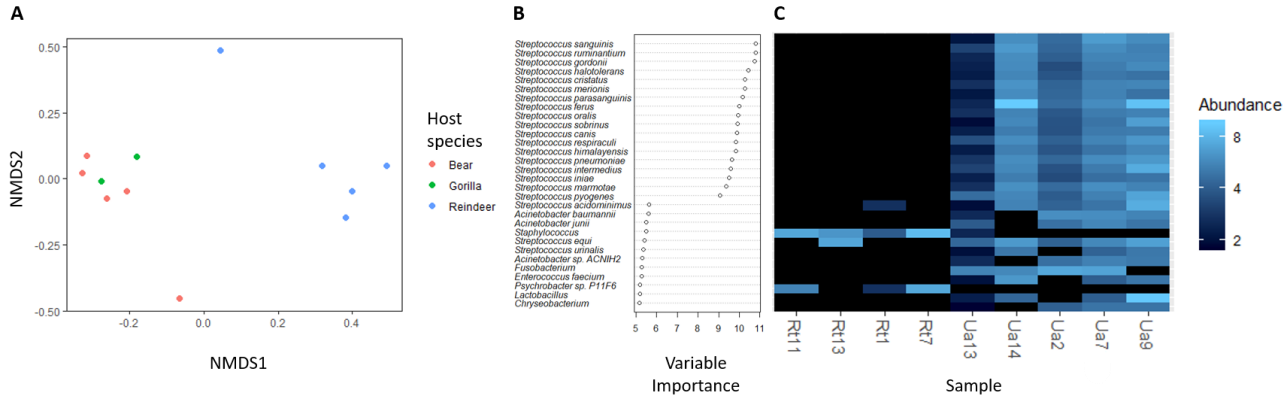


Figure 2. The oral microbiome signature differences between host species. (A) NMDS ordination on Jaccard distance matrix of presence/absence of microbial taxa per sample, coloured by host species. (B) Random forest variable importance plot of the 30 most discriminatory taxa comparing bear and reindeer samples (excluding Rt5, see below), based on presence/absence data. (C). CLR normalised abundance of the top 30 taxa in (B) in the bear (Ua) and reindeer (Rt) samples (excluding Rt5, which was identified as an outlier in these analyses).

Host
PWY-7013: L-1,2-propanediol degradation
PWY-5083: NAD/NADH phosphorylation and dephosphorylation
PWY-6545: pyrimidine deoxyribonucleotides de novo biosynthesis III
PWY-7279: aerobic respiration II (cytochrome c) (yeast)
PWY0-1415: superpathway of heme biosynthesis from uroporphyrinogen-III
PWY-6527: stachyose degradation
LACTOSECAT-PWY: lactose and galactose degradation I
P461-PWY: hexitol fermentation to lactate, formate, ethanol and acetate
PWY-7219: adenosine ribonucleotides de novo biosynthesis
PWY-7111: pyruvate fermentation to isobutanol (engineered)
ILEUSYN-PWY: L-isoleucine biosynthesis I (from threonine)
VALSYN-PWY: L-valine biosynthesis
BRANCHED-CHAIN-AA-SYN-PWY: superpathway of branched amino acid biosynthesis
PWY-5103: L-isoleucine biosynthesis III
PWY-7221: guanosine ribonucleotides de novo biosynthesis
PWY-6126: superpathway of adenosine nucleotides de novo biosynthesis II
PWY-7229: superpathway of adenosine nucleotides de novo biosynthesis I
PWY-6936: seleno-amino acid biosynthesis
PWY-6151: S-adenosyl-L-methionine cycle I
PWY-6609: adenine and adenosine salvage III
PWY-6700: queuosine biosynthesis
PWY-7220: adenosine deoxyribonucleotides de novo biosynthesis II
PWY-7222: guanosine deoxyribonucleotides de novo biosynthesis II
PWY-6125: superpathway of guanosine nucleotides de novo biosynthesis II
PWY-7228: superpathway of guanosine nucleotides de novo biosynthesis I
PEPTIDOGLYCANSYN-PWY: peptidoglycan biosynthesis I (meso-diaminopimelate containing)
PWY-6386: UDP-N-acetylmuramoyl-pentapeptide biosynthesis II (lysine-containing)
PWY-6387: UDP-N-acetylmuramoyl-pentapeptide biosynthesis I (meso-diaminopimelate containing)
PWY-6121: 5-aminoimidazole ribonucleotide biosynthesis I
PWY-6122: 5-aminoimidazole ribonucleotide biosynthesis II
PWY-6277: superpathway of 5-aminoimidazole ribonucleotide biosynthesis
PWY-3001: superpathway of L-isoleucine biosynthesis I
THRESYN-PWY: superpathway of L-threonine biosynthesis
ARGSYN-PWY: L-arginine biosynthesis I (via L-ornithine)
PWY-7400: L-arginine biosynthesis IV (archaebacteria)
PWY-5686: UMP biosynthesis
PWY-2942: L-lysine biosynthesis III
PWY-5097: L-lysine biosynthesis VI
PWY-841: superpathway of purine nucleotides de novo biosynthesis I
PWY-724: superpathway of L-lysine, L-threonine and L-methionine biosynthesis II
PWY-6163: chorismate biosynthesis from 3-dehydroquinate
ARO-PWY: chorismate biosynthesis I
COMPLETE-ARO-PWY: superpathway of aromatic amino acid biosynthesis
1CMET2-PWY: N10-formyl-tetrahydrofolate biosynthesis
PWY-7197: pyrimidine deoxyribonucleotide phosphorylation
PWY-3841: folate transformations II
PWY0-166: superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis (E. coli)
PWY-7184: pyrimidine deoxyribonucleotides de novo biosynthesis I
PWY-7208: superpathway of pyrimidine nucleobases salvage
ARGSYNBSUB-PWY: L-arginine biosynthesis II (acetyl cycle)
PWY-1042: glycolysis IV (plant cytosol)
PWY-6385: peptidoglycan biosynthesis III (mycobacteria)
PWY-5973: cis-vaccenate biosynthesis
DAPLYSINESYN-PWY: L-lysine biosynthesis I
PWY-5188: tetrapyrrole biosynthesis I (from glutamate)
NONOXIPENT-PWY: pentose phosphate pathway (non-oxidative branch)
PWY-7663: gondoate biosynthesis (anaerobic)
CALVIN-PWY: Calvin-Benson-Bassham cycle
TRNA-CHARGING-PWY: tRNA charging
PWY-5695: urate biosynthesis/inosine 5'-phosphate degradation
GLUTORN-PWY: L-ornithine biosynthesis
PWY0-162: superpathway of pyrimidine ribonucleotides de novo biosynthesis
PWY-6168: flavin biosynthesis III (fungi)
PWY0-1479: tRNA processing
PWY-1269: CMP-3-deoxy-D-manno-octulosonate biosynthesis I
FERMENTATION-PWY: mixed acid fermentation
PWY-5667: CDP-diacylglycerol biosynthesis I
PWY0-1319: CDP-diacylglycerol biosynthesis II
PWY-5913: TCA cycle VI (obligate autotrophs)
PWY-7117: C4 photosynthetic carbon assimilation cycle, PEPCK type
PWY-2941: L-lysine biosynthesis II
PWY-5100: pyruvate fermentation to acetate and lactate II
PWY-6147: 6-hydroxymethyl-dihydropterin diphosphate biosynthesis I
PWY-7539: 6-hydroxymethyl-dihydropterin diphosphate biosynthesis III (Chlamydia)
SER-GLYSYN-PWY: superpathway of L-serine and glycine biosynthesis I
DTDPRHAMSYN-PWY: dTDP-L-rhamnose biosynthesis I
PWY-7187: pyrimidine deoxyribonucleotides de novo biosynthesis II
PWY-6123: inosine-5'-phosphate biosynthesis I
PWY-6124: inosine-5'-phosphate biosynthesis II
PWY-7234: inosine-5'-phosphate biosynthesis III
HISTSYN-PWY: L-histidine biosynthesis
PWY-7199: pyrimidine deoxyribonucleosides salvage
HOMOSER-METSYN-PWY: L-methionine biosynthesis I
PWY-5347: superpathway of L-methionine biosynthesis (transsulfuration)
MET-SAM-PWY: superpathway of S-adenosyl-L-methionine biosynthesis
METSYN-PWY: L-homoserine and L-methionine biosynthesis
PWY66-409: superpathway of purine nucleotide salvage
PWY0-1296: purine ribonucleosides degradation
PWY0-1297: superpathway of purine deoxyribonucleosides degradation
HISDEG-PWY: L-histidine degradation I
FASYN-INITIAL-PWY: superpathway of fatty acid biosynthesis initiation (E. coli)
PWY0-862: (5Z)-dodec-5-enoate biosynthesis
PWY-7388: octanoyl-[acyl-carrier protein] biosynthesis (mitochondria, yeast)
PWY-6282: palmitoleate biosynthesis I (from (5Z)-dodec-5-enoate)
FASYN-ELONG-PWY: fatty acid elongation -- saturated
PWY-7664: oleate biosynthesis IV (anaerobic)
ASPASN-PWY: superpathway of L-aspartate and L-asparagine biosynthesis
PWY0-1061: superpathway of L-alanine biosynthesis
PWY0-1586: peptidoglycan maturation (meso-diaminopimelate containing)
UDPNAGSYN-PWY: UDP-N-acetyl-D-glucosamine biosynthesis I
PWY-5690: TCA cycle II (plants and fungi)
PWY-7254: TCA cycle VII (acetate-producers)
PWY-3781: aerobic respiration I (cytochrome c)
PHOSLIPSYN-PWY: superpathway of phospholipid biosynthesis I (bacteria)
PWY4FS-7: phosphatidylglycerol biosynthesis I (plastidic)
PWY4FS-8: phosphatidylglycerol biosynthesis II (non-plastidic)
TCA: TCA cycle I (prokaryotic)
NAGLIPASYN-PWY: lipid IVA biosynthesis
PYRIDOXSYN-PWY: pyridoxal 5'-phosphate biosynthesis I
COA-PWY-1: coenzyme A biosynthesis II (mammalian)
TRPSYN-PWY: L-tryptophan biosynthesis
PWY-4981: L-proline biosynthesis II (from arginine)
COA-PWY: coenzyme A biosynthesis I
PWY-4242: pantothenate and coenzyme A biosynthesis III
PWY-6703: preQ0 biosynthesis
NONMEVIPP-PWY: methylerythritol phosphate pathway I
HEME-BIOSYNTHESIS-II: heme biosynthesis I (aerobic)
HEMESYN2-PWY: heme biosynthesis II (anaerobic)
PANTO-PWY: phosphopantothenate biosynthesis I
PWY-7357: thiamin formation from pyrithiamine and oxythiamine (yeast)
PWY-7282: 4-amino-2-methyl-5-phosphomethylpyrimidine biosynthesis (yeast)
PWY0-845: superpathway of pyridoxal 5'-phosphate biosynthesis and salvage
Role
Ua9
Rt11
Ua13
Ua2
Rt13
Rt5
Ua7
Ua14
Rt7
Gb1wash
Gb2wash
Sample
Abundance
6
4
2
0
-2
Annotation
Host
Gorilla
Reindeer
Bear
Role
Biosynthesis
Central metabolism
Degradation
Fermentation
Photosynthesis
Respiration

Figure 3. Core functions are shared by the oral microbiome of gorillas (Gb), bears (Ut), and reindeer (Rt). CLR normalised abundance of MetaCyc metabolic pathways containing >50% of required enzymes, as identified in by HumanN2 in each study sample. Metabolic pathways are annotated with their main role in metabolism. Samples and pathways are hierarchically clustered by dissimilarity implemented by pheatmap.

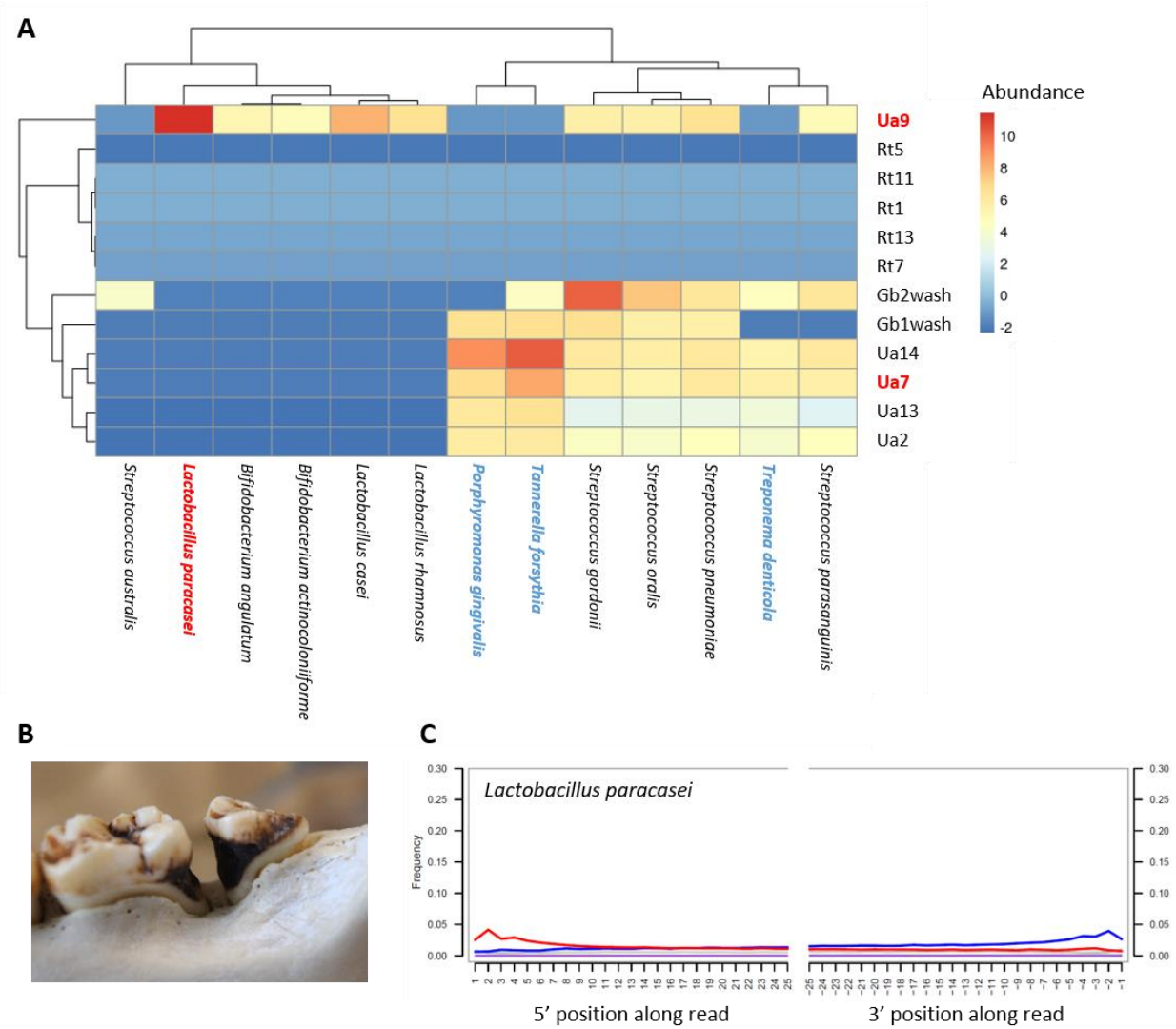


Figure 4. Cariogenic bacteria can be detected in dental calculus. (A) Kraken2 CLR normalised abundance of cariogenic bacteria (*Streptococcus*, *Lactobacillus* and *Bifidobacterium* species) and periodontal pathogens (*Porphyromonas gingivalis*, *Treponema denticola*, and *Tannerella forsythia*, in bold blue text) in the 12 dental calculus samples. Ua9 and Ua7 (in bold red text) both had evidence of caries and periodontitis based on the macroscopic examination of the skull; however, only Ua9 was sampled from a caries site. *L. paracasei* (bold red text) was highly abundant in Ua9 and investigated further in (C). (B) Sampling site from caries lesion with periodontitis from bear specimen Ua9. (C) Deamination plot of *L. paracasei* showing frequency of C to T (red) and G to A (blue) substitutions in Ua9 reads compared to the bacteria reference genome.

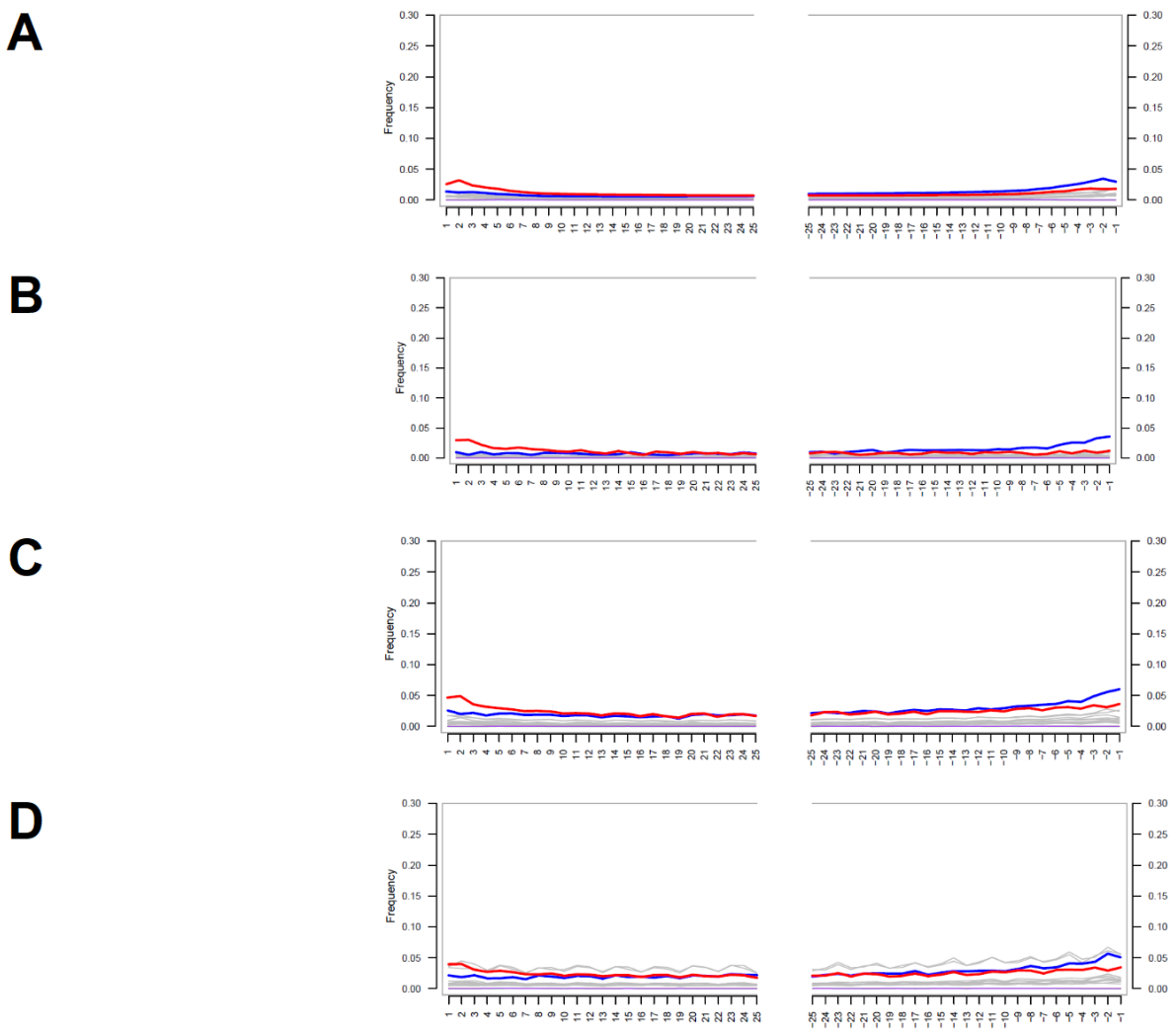


Figure 5. Deamination rates generated by mapDamage for (A) host DNA (*Rangifer tarandus*); (B) *Klebsiella pneumoniae* and (C) *Acinetobacter baumannii*, clinically relevant oral bacteria carrying multi-drug resistance in modern human populations; and (D) *Pseudomonas psychrophila*, an aquatic bacteria and likely modern contaminant. Reads were pooled across all reindeer samples. C to T (red), G to A (blue) and all other (grey) substitutions are shown. In (D), the terminal 5' C to T and 3' G to A substitutions appear similar to authentic DNA damage; however, background (grey) substitution rates are also high, suggesting that the frequency calculations are affected by low read numbers mapping to the reference.

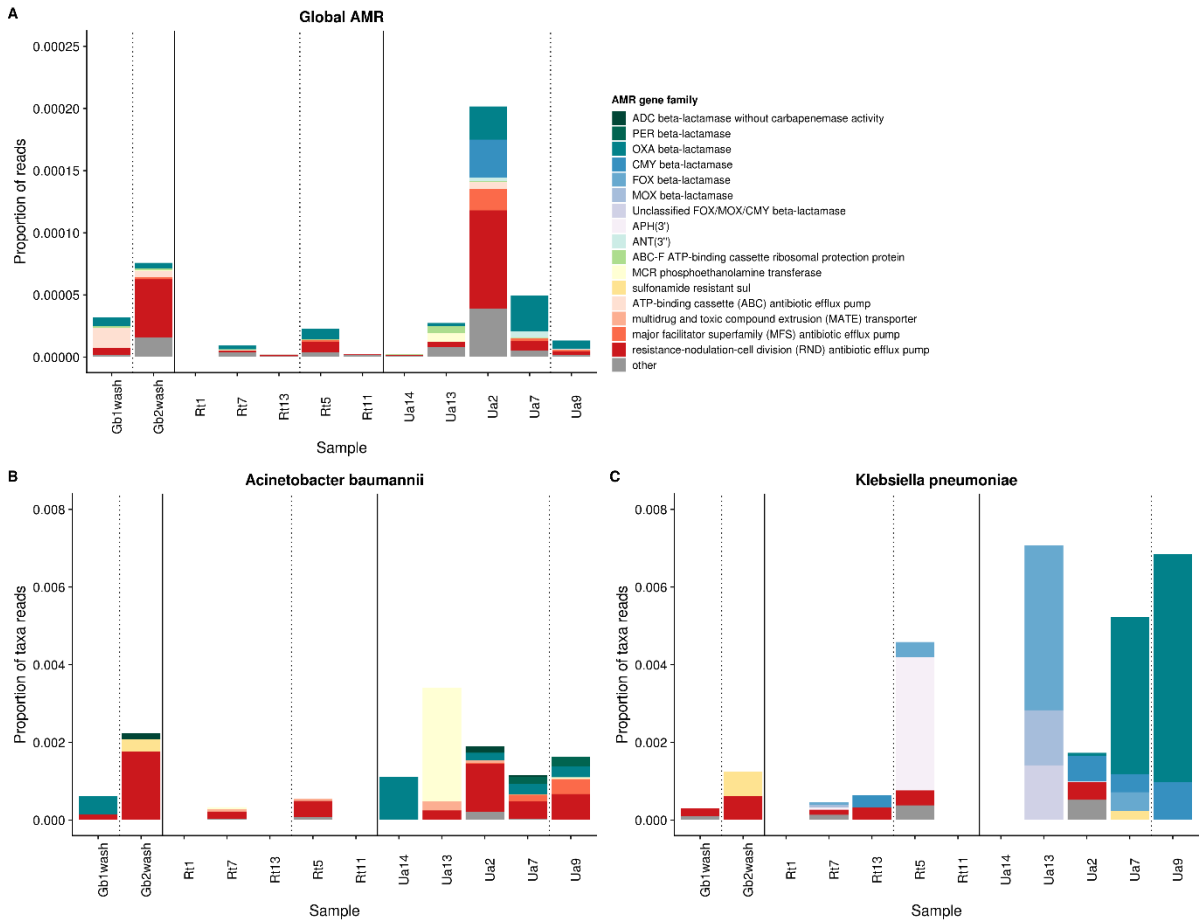


Figure 6. Antimicrobial resistance (AMR) genes can be recovered from historical dental calculus from specimens collected prior to industrial-scale antibiotics production in the 1940s. (A) Proportion of global AMR genes identified in all contamination-filtered reads in each sample. (B) Proportion of *Acinetobacter baumannii* and (C) *Klebsiella pneumoniae* specific AMR genes identified in each sample. Samples are grouped by host species (Gb: gorilla, Rt: reindeer, Ua: bear, separated by solid vertical lines) and ordered by year with pre-1940 samples separated from post-1940 samples by dashed vertical lines. The top eight most abundant AMR gene families are shown for each analysis, with the rest grouped into 'other'.

**References**


1. Huttenhower, C. *et al.* Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
2. Petersen, C. & Round, J. L. Defining dysbiosis and its influence on host immunity and disease. *Cell. Microbiol.* **16,** 1024–1033 (2014).
3. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16,** 191 (2015).
4. Gomez, A. *et al.* Host genetic control of the oral microbiome in health and disease. *Cell Host Microbe* **22,** 269–278 (2017).
5. Groussin, M. *et al.* Unraveling the processes shaping mammalian gut microbiomes over evolutionary time. *Nat. Commun.* **8,** 14319 (2017).
6. Nishida, A. H. & Ochman, H. Rates of gut microbiome divergence in mammals. *Mol. Ecol.* (2018). doi:10.1111/mec.14473
7. Chevalier, C. *et al.* Gut Microbiota Orchestrates Energy Homeostasis during Cold. *Cell* **163,** 1360–1374 (2015).
8. Carmody, R. N. *et al.* Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host Microbe* **17,** 72–84 (2015).
9. Koren, O. *et al.* Host remodeling of the gut microbiome and metabolic changes during pregnancy. *Cell* **150,** 470–480 (2013).
10. Warinner, C., Speller, C., Collins, M. J. & Lewis, C. M. Ancient human microbiomes. *J. Hum. Evol.* **79,** 125–136 (2015).
11. Adler, C. J. *et al.* Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nat. Genet.* **45,** 450–455 (2013).
12. Warinner, C. *et al.* Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46,** 336–344 (2014).
13. Weyrich, L. S. *et al.* Neanderthal behaviour, diet, and disease inferred from ancient DNA in

dental calculus. *Nature* **544,** 357–361 (2017).

14. Mann, A. E. *et al.* Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Sci. Rep.* **8,** 9822 (2018).

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

16. Zhou, Z. *et al.* Pan-genome Analysis of Ancient and Modern Salmonella enterica Demonstrates Genomic Stability of the Invasive Para C Lineage for Millennia. *Curr. Biol.* **28,** 2420–2428 (2018).

17. Holmes, M. W. *et al.* Natural history collections as windows on evolutionary processes. *Mol. Ecol.* **25,** 864–881 (2016).

18. Van Der Valk, T. *et al.* Significant loss of mitochondrial diversity within the last century due to extinction of peripheral populations in eastern gorillas. *Sci. Rep.* **8,** 1–10 (2018).

19. van der Valk, T., Díez-del-Molino, D., Marques-Bonet, T., Guschanski, K. & Dalén, L. Historical genomes reveal the genomic consequences of recent population decline in eastern gorillas. *Curr. Biol.* **29,** 165–170 (2019).

20. van't Hof, A. E., Edmonds, N., Dalíková, M. & Saccheri, I. J. Industrial Melanism in British Peppered Moths Has a Singular and. *Science (80-. ).* **332,** 958–60 (2011).

21. Hird, S. M. Evolutionary biology needs wild microbiomes. *Front. Microbiol.* **8,** 1–10 (2017).

22. Amato, K. R. *et al.* Habitat degradation impacts black howler monkey (Alouatta pigra) gastrointestinal microbiomes. *ISME J.* **7,** 1344–1353 (2013).

23. Hooper, R. *et al.* Host-derived population genomics data provides insights into bacterial and diatom composition of the killer whale skin. *Mol. Ecol.* **28,** 484–502 (2019).

24. Key, F. M., Posth, C., Krause, J., Herbig, A. & Bos, K. I. Mining Metagenomic Data Sets for Ancient DNA: Recommended Protocols for Authentication. *Trends Genet.* **33,** 508–520 (2017).

25. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).

26. Weyrich, L. *et al.* Laboratory contamination over time during low-biomass sample analysis. *Mol. Ecol. Resour.* (2019). doi:10.1111/1755-0998.13011

27. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499,** 431–437 (2013).

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

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

30. Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8,** 761–763 (2011).

31. Lloyd-price, J. *et al.* Strains, functions, and dynamics in the expanded Human Microbiome Project. *Nature* **550,** 61–66 (2018).

32. Oh, J. *et al.* Biogeography and individuality shape function in the human skin metagenome. *Nature* **514,** 59–64 (2014).

33. Johnston, E. R. *et al.* Metagenomics reveals pervasive bacterial populations and reduced community diversity across the Alaska tundra ecosystem. *Front. Microbiol.* **7,** 1–16 (2016).

34. Janssen, P. H. & Kirs, M. Structure of the archaeal community of the rumen. *Appl. Environ. Microbiol.* **74,** 3619–3625 (2008).

35. Kittelmann, S., Kirk, M. R., Jonker, A., McCulloch, A. & Janssen, P. H. Buccal swabbing as a noninvasive method to determine bacterial, archaeal, and eukaryotic microbial community structures in the rumen. *Appl. Environ. Microbiol.* **81,** 7470–7483 (2015).

36. Franzosa, E. A. *et al.* Species-level functional profiling of metagenomes and metatranscriptomes. *Nat. Methods* **15,** 962–968 (2018).

37. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopaedia of Genes and Genomes. *Nucl. Acids Res.* **28,** 27–30 (2000).

38. Caspi, R. *et al.* The MetaCyc database of metabolic pathways and enzymes. *Nucleic Acids Res.* **46,** D633–D639 (2018).

39. Prakash, T. & Taylor, T. D. Functional assignment of metagenomic data: Challenges and applications. *Brief. Bioinform.* **13,** 711–727 (2012).

40. Takahashi, N. & Nyvad, B. The role of bacteria in the caries process: Ecological perspectives. *J. Dent. Res.* **90,** 294–303 (2011).

41. Neves, B. G., Stipp, R. N., da Silva Bezerra, D., de Figueiredo Guedes, S. F. & Rodrigues, L. K. A. Molecular detection of bacteria associated to caries activity in dentinal lesions. *Clin. Oral Investig.* **21,** 2053–2061 (2017).

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

43. Xie, G. *et al.* Community and gene composition of a human dental plaque microbiota obtained by metagenomic sequencing. *Mol. Oral Microbiol.* **25,** 391–405 (2010).

44. Tommasi, R., Brown, D. G., Walkup, G. K., Manchester, J. I. & Miller, A. A. ESKAPEing the labyrinth of antibacterial discovery. *Nature Reviews Drug Discovery* **14,** 529–542 (2015).

45. Jia, B. *et al.* CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **45,** D566–D573 (2017).

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

47. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr.* arXiv:1303.3997 (2013). doi:10.1186/s13756-018-0352-y

48. Landecker, H. Antibiotic resistance and the biology of history. *Body Soc.* **22,** 19–52 (2016).

49. Dcosta, V. M. *et al.* Antibiotic resistance is ancient. *Nature* **477,** 457–461 (2011).

50. Ozga, A. T. *et al.* Successful enrichment and recovery of whole mitochondrial genomes from ancient human dental calculus. *Am. J. Phys. Anthropol.* **160,** 220–228 (2016).

51. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially

expands the tree of life. *Nat. Microbiol.* **2,** 1533–1542 (2017).

52. Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31,** 1674–1676 (2015).

53. Kang, D. D., Froula, J., Egan, R. & Wang, Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ* **3,** e1165 (2015).

54. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–55 (2015).

55. Parks, D. H. *et al.* A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nat. Biotechnol.* **36,** 996–1004 (2018).

56. Gong, S. & Bao, L. The battle against SARS and MERS coronaviruses: Reservoirs and Animal Models. *Anim. Model. Exp. Med.* **1,** 125–133 (2018).

57. Reperant, L. A., Moesker, F. M. & Osterhaus, A. D. M. E. Influenza: from zoonosis to pandemic. *ERJ Open Res.* **2,** 00013-2016 (2016).

58. Arnold, K. E., Williams, N. J. & Bennett, M. 'Disperse abroad in the land': The role of wildlife in the dissemination of antimicrobial resistance. *Biology Letters* **12,** 20160137 (2016).

59. Vittecoq, M. *et al.* REVIEW: Antimicrobial resistance in wildlife. *J. Appl. Ecol.* **53,** 519–529 (2016).

60. Dabney, J. *et al.* Complete mitochondrial genome sequence of a Middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. *Proc. Natl. Acad. Sci.* **110,** 15758–15763 (2013).

61. Meyer, M. & Kircher, M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* **2010,** pdb.prot5448 (2010).

62. Valk, T. van der, Vezzi, F., Ormestad, M., Dalen, L. & Guschanski, K. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. *Mol. Ecol. Resour.*

(2019). doi:10.1111/1755-0998.13009

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

64. van der Valk, T., Lona Durazo, F., Dalén, L. & Guschanski, K. Whole mitochondrial genome capture from faecal samples and museum-preserved specimens. *Mol. Ecol. Resour.* **17,** e111–e121 (2017).

65. Schubert, M., Lindgreen, S. & Orlando, L. AdapterRemoval v2: Rapid adapter trimming, identification, and read merging. *BMC Res. Notes* **9,** 88 (2016).

66. Schmieder, R. & Edwards, R. Quality control and preprocessing of metagenomic datasets. *Bioinformatics* **27,** 863–864 (2011).

67. Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N. C. & Pati, A. Large-scale contamination of microbial isolate genomes by illumina Phix control. *Stand. Genomic Sci.* **10,** 18 (2015).

68. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

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

70. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27,** 849–864 (2017).

71. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483,** 169–175 (2012).

72. Taylor, G. A. *et al.* The genome of the North American brown bear or grizzly: Ursus arctos ssp. horribilis. *Genes (Basel).* **9,** 598 (2018).

73. Lusk, R. W. Diverse and widespread contamination evident in the unmapped depths of high

throughput sequencing data. *PLoS One* **9,** e110808 (2014).

74. Frank, J. A. & Sørensen, S. J. Quantitative Metagenomic Analyses Based on Average Genome Size Normalization. *Appl. Environ. Microbiol.* **77,** 2513–2521 (2011).

75. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44,** D733–D745 (2016).

76. Velsko, I. M., Frantz, L. A. F., Herbig, A., Larson, G. & Warinner, C. Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. *mSystems* **3,** e00080-18 (2018).

77. Peabody, M. A., Van Rossum, T., Lo, R. & Brinkman, F. S. L. Evaluation of shotgun metagenomics sequence classification methods using in silico and in vitro simulated communities. *BMC Bioinformatics* **16,** 363 (2015).

78. Dewhirst, F. E. *et al.* The human oral microbiome. *J. Bacteriol.* **192,** 5002–5017 (2010).

79. Kilian, M. *et al.* The oral microbiome - an update for oral healthcare professionals. *Br. Dent. J.* **221,** 657–666 (2016).

80. Dewhirst, F. E. *et al.* The canine oral microbiome. *PLoS One* **7,** 1–12 (2012).

81. Dewhirst, F. E. *et al.* The feline oral microbiome: A provisional 16S rRNA gene based taxonomy with full-length reference sequences. *Vet. Microbiol.* **175,** 294–303 (2015).

82. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome datasets are compositional: And this is not optional. *Front. Microbiol.* **8,** 1–6 (2017).

83. Oksanen, A. J. *et al.* Vegan: Community ecology package. R package version 2.5-3. https://CRAN.R-project.org/package=vegan. (2018). doi:10.4135/9781412971874.n145

84. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* (1948). doi:10.1002/j.1538-7305.1948.tb00917.x

85. Liaw, A. & Wiener, M. Classification and Regression by randomForest. *R news* **2,** 18–22 (2002).

86. Chen, T. *et al.* The Human Oral Microbiome Database: a web accessible resource for

investigating oral microbe taxonomic and genomic information. *Database (Oxford).* (2010). doi:10.1093/database/baq013

87. Truong, D. T. *et al.* MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nat. Methods* **12,** 902–903 (2015).

88. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* (2015). doi:10.1038/nmeth.3176

89. Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R. & Wu, C. H. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics* (2007). doi:10.1093/bioinformatics/btm098

90. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).

91. Madden, T. L. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10,** 421 (2009).

92. Dabney, J., Meyer, M. & Paabo, S. Ancient DNA Damage. *Cold Spring Harb. Perspect. Biol.* **5,** a012567–a012567 (2013).

93. Abraham, W. P. & Thomas, S. Draft Genome Sequence of Pseudomonas psychrophila MTCC 12324, Isolated from the Arctic at 79°N . *Genome Announc.* **3,** 12324 (2015).

94. Jónsson, H., Ginolhac, A., Schubert, M., Johnson, P. L. F. & Orlando, L. MapDamage2.0: Fast approximate Bayesian estimates of ancient DNA damage parameters. *Bioinformatics* **29,** 1682–1684 (2013).

95. Peltzer, A. *et al.* EAGER: efficient ancient genome reconstruction. *Genome Biol.* **17,** 60 (2016).

**Supplementary information**

Includes Figures S1-10 and the following tables:

Table S1. Sample information and laboratory processing.

Table S2. Beta diversity PERMANOVA analyses for microbial presence/absence, CLR normalised abundance and metabolic pathway abundance data.

Table S3. Host DNA genome and mitochondrial genome recovery per sample.

Table S4. Illumina adapter and qPCR sequences.

Table S5. Sources and SRA accessions used for SourceTracker analysis.

Table S6. GenBank accessions of reference genomes used to filter identified microbial contaminants.