npj | biofilms and microbiomes

Article

Published in partnership with Nanyang Technological University

https://doi.org/10.1038/s41522-024-00642-1

Streptococcus abundance and oral site tropism in humans and non-human primates reflects host and lifestyle differences

Irina M. Velsko [1,2] ✉ & Christina Warinner[1,2,3,4,5] ✉

The genus *Streptococcus* is highly diverse and a core member of the primate oral microbiome. *Streptococcus* species are grouped into at least eight phylogenetically-supported clades, five of which are found almost exclusively in the oral cavity. We explored the dominant *Streptococcus* phylogenetic clades in samples from multiple oral sites and from ancient and modern-day humans and non-human primates and found that clade dominance is conserved across human oral sites, with most *Streptococcus* reads assigned to species falling in the Sanguinis or Mitis clades. However, minor differences in the presence and abundance of individual species within each clade differentiated human lifestyles, with loss of *S. sinensis* appearing to correlate with toothbrushing. Of the non-human primates, only baboons show clade abundance patterns similar to humans, suggesting that a habitat and diet similar to that of early humans may favor the growth of Sanguinis and Mitis clade species.

Streptococcus is a diverse and heavily-studied bacterial genus, with a wide range of hosts and habitats including humans and other mammals, amphibians, fish, and food fermentation cultures. While species of this genus include pathogenic host-generalists that infect multiple host species[1], as well as pathogenic host-specialists[2], many *Streptococcus* species are commensal host-associated microbiome members that do not inherently cause disease. Phylogenetic analysis of the genus revealed eight well-supported clades, five of which (Sanguinis, Mitis, Anginosus, Salivarius, and Mutans) make up the so-called viridans group[3]. The species within these clades are particularly prominent within the human oral microbiome, exhibiting a highly specific host-niche adaptation within this genus.

In healthy, industrialized US American populations, the resident oral *Streptococcus* species exhibit oral site tropism, where particular species preferentially reside on distinct oral surfaces such as the tongue, buccal mucosa, or the tooth surface in dental plaque biofilm[4]. The functional characteristics that distinguish *Streptococcus* species living in different oral niches have been explored in healthy North American populations[4], which advanced our understanding of the genetic and biochemical drivers of site tropism. Unexpectedly, gene content poorly differentiates closely related species with distinct site tropism, yet biofilm spatial structuring at the micron scale may be determined by species-specific interactions, including both direct contact and indirect metabolite sharing[5], highlighting the importance of community species composition to site tropism of any particular *Streptococcus* species. However, whether the species partitioning we observe in studied populations are characteristic of human populations globally, or whether they are affected by global market integration (i.e., access to globally sourced and processed items including foods, often used to indicate level of "industrialization") and urbanization, as well as the evolutionary origins of site-tropism, have not yet been investigated.

Recent work on dental plaque of non-industrial populations and ancient dental calculus demonstrated differences in the microbial communities of these sample types compared to dental plaque samples from healthy North American populations that could be attributed to dental hygiene practices[6,7]. Dental plaque biofilms grow and develop in a predictable succession of species[8], and frequent removal of the dental plaque biofilm through regular toothbrushing disrupts the natural biofilm growth and maturation process. The repeated removal and regrowth of early stage plaque, in which *Streptococcus* is abundant, delays biofilm maturation, such that late-colonizer taxa are infrequently detected and/or at low abundance compared to biofilms that mature relatively undisturbed. The apparent difference in the oral microbiome profile between industrial, non-industrial, and historic populations, therefore likely represent two ends of community

[1]Department of Archaeogenetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [2]Archaeogenetics Research Unit, Leibniz Institute for Natural Products Research and Infection Biology Hans Knöll Institute, Jena, Germany. [3]Faculty of Biological Sciences, Friedrich Schiller University, Jena, Germany. [4]Radcliffe Institute for Advanced Study, Cambridge, MA, USA. [5]Department of Anthropology, Harvard University, Cambridge, MA, USA.
✉e-mail: irina_marie_velsko@eva.mpg.de; warinner@fas.harvard.edu

succession rather than distinct communities that develop in response to unique environmental input such as diet.

Streptococcus species play a major role in colonizing tooth surfaces and initiating dental plaque biofilm formation in humans[9]. Recently, *Streptococcus* was shown to be a core genus of the primate dental biofilm by analyzing dental calculus[10], a mineralized version of dental plaque that forms in situ on tooth surfaces during life, and which preserves well in the archeological record[11]. Fellows Yates et al.[10] investigated the distribution of *Streptococcus* in ancient and modern human and non-human primate dental calculus by grouping the *Streptococcus* species by the phylogenetic clades in which they fall, and comparing the abundance of each clade across host species. The streptococcal profiles of ancient humans, including several from Neanderthals, were largely indistinguishable from those of modern humans, yet chimpanzees, gorillas, and howler monkeys each exhibited a distinctive *Streptococcus* clade profile. This opened the question of the extent to which oral *Streptococcus* diversity is shared or unique across primates, and whether oral *Streptococcus* species have co-evolved with their primate hosts from a deep-time shared common ancestor, or been acquired at unique points in primate evolution. Further, whether oral *Streptococcus* site tropism is observed across primates or a unique characteristic of humans is unknown.

In ancient and present-day humans, who have relatively high relative abundance of *Streptococcus*, the Sanguinis clade was the most abundant clade, while in chimpanzees, who have low relative abundance of *Streptococcus*, the Anginosus clade was the most abundant[10]. Curiously, however, in a small number of ancient human samples (~10%), the Sanguinis clade species were nearly absent, and these instead had predominantly Anginosus clade species, strongly resembling the chimpanzee *Streptococcus* clade profiles. Due to the high heterogeneity of samples in that study, no explanation for the chimpanzee-like profile in human samples could be proposed. Whether this taxonomic profile was a broad characteristic of human dental calculus microbiomes, and which possible biological variables contribute to this unique profile remained unresolved. Therefore, the relevance of this minority *Streptococcus* profile on understanding oral microbiome community composition and metabolic functional potential remains to be investigated using larger datasets with more detailed sample metadata.

Here we investigated the distribution of *Streptococcus* clades in a large dataset of living and ancient human and non-human primate oral samples (Fig. 1A, Supplementary Fig. 6), to better understand the extent of oral *Streptococcus* site tropism through time and at a global ecological scale. We find that the distribution of *Streptococcus* clades in ancient human calculus is consistent across time and space, with a majority of humans having a Sanguinis and Mitis dominated streptococcal profile, and a minority of humans (~10%) mostly lacking these clades. Streptococcal profiles are moreover largely consistent within individuals, with teeth across the dentition generally exhibiting similar streptococcal clade patterns. In living human populations, the distribution of clades in each oral site is largely consistent across varying levels of global market integration and urbanization, suggesting that present-day lifestyle factors are not major drivers of streptococcal clade colonization, although species-level differences are observed within clades. Among these is an apparent reduction of *S. sinensis* in the dental plaque of industrialized populations, which may be due to toothbrushing. Each non-human primate investigated has a distinctive *Streptococcus* clade profile, with baboons being most similar to humans. No distinct functional differences underlying site and host specialization were found, suggesting that further characterization of the genes in commensal *Streptococcus* may be necessary to understand niche specialization.

Results

Presence of Sanguinis clade species determines clade abundance in ancient humans

We first determined which *Streptococcus* species are found across a large ancient dental calculus dataset comprised of samples from across the globe, spanning 100,000 years, and processed and sequenced in various labs (Fig. 1B, Supplementary Figs. 1–5, 6B, C). This allowed us to investigate whether the trend reported by Fellows Yates et al.[10], wherein a majority of human ancient dental calculus samples have predominantly Sanguinis clade species and a minority have a chimpanzee-like Anginosus clade-dominated profile, is a universal feature of human ancient dental calculus, or whether this was a feature of the dataset originally used. Following taxonomic profiling with the Genome Taxonomy Database (GTDB), we assigned each *Streptococcus* species in the species table to one of the following previously described *Streptococcus* clades: Sanguinis, Mitis, Salivarius, Anginosus, Bovis, Pyogenic, Mutans, Downei, Other, or Unknown. We then calculated the proportion of reads assigned to species falling in each clade out of all *Streptococcus* read counts (Fig. 2A, Supplementary Table 3).

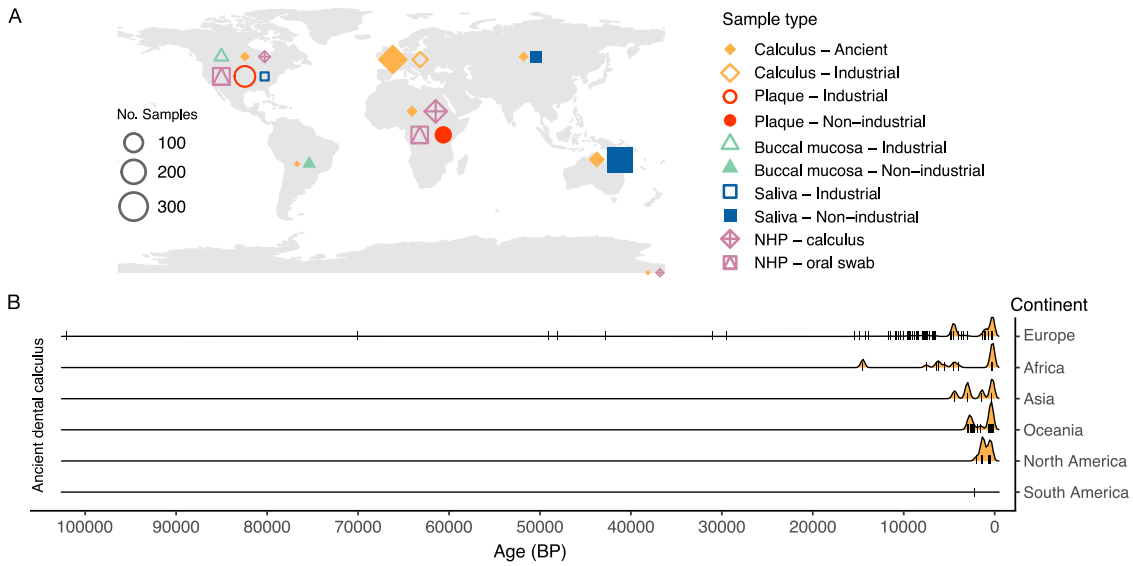

Fig. 1 | Geographic and temporal distribution of samples included in this study. **A** The continent of origin for all samples included in this study. Point shape and color indicate sample type, while point size indicates the number of samples. The two points in the lower right corner represent museum specimens for which the geographic origin is uncertain. **B** Temporal origin of ancient dental calculus samples separated by continent. Samples are indicated by a tick mark across the line for each continent, binned per 200 years. Histograms demonstrate the ages with the highest density of sample counts. The majority of samples from all continents are from ≤ 1000 BP (Before Present).

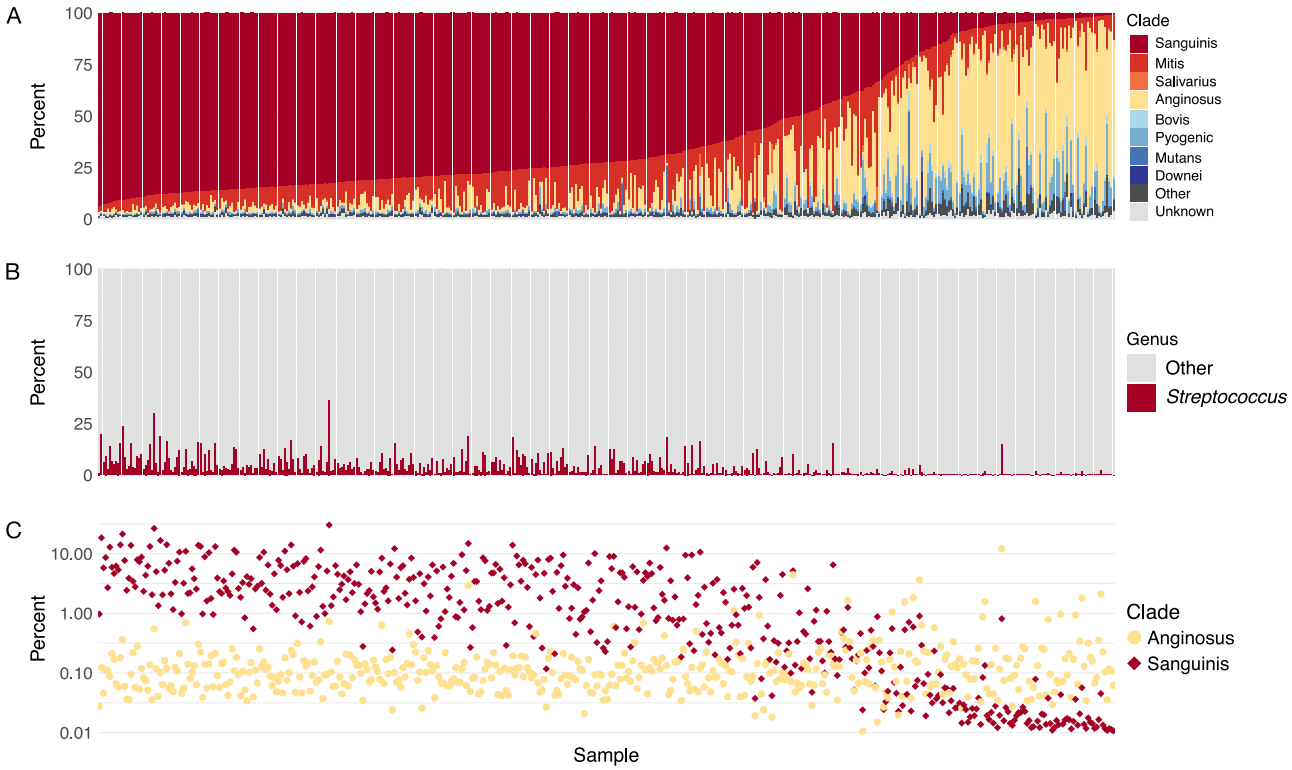

Fig. 2 | Distribution of *Streptococcus* clades in ancient dental calculus samples. **A** Percent of *Streptococcus* reads that were assigned to each clade out of all reads assigned to *Streptococcus*, ordered by decreasing relative abundance of Sanguinis clade and increasing relative abundance of Anginosus clade. **B** Percent of reads assigned to species in the genus *Streptococcus* and to all other genera. **C** Percent of reads assigned to species in the Sanguinis and Anginosus clades out of all species-level read assignments.

We found that dental calculus in our global, deep-time dataset, regardless of age (Supplementary Fig. 6), replicates the same pattern of *Streptococcus* clade relative abundance first described in ref.[10], with the majority having most reads assigned to species falling in the Sanguinis clade, while a small number of samples (71/483, 14.7%) have reads assigned primarily to species falling in the Anginosus clade. Both a Pearson's correlation test on CLR-transformed data and a CoDA correlation test confirmed a significant negative correlation between the relative abundance of the two clades in this dataset ($\rho = -0.68, p < 0.0001$ and $\rho = -0.67$, respectively). This pattern was further replicated in each dataset individually, demonstrating that this *Streptococcus* species profile is a characteristic of human ancient dental calculus generally and not the result of biases in laboratory processing, and is moreover not restricted to a particular geographic region or time period. An exception, however, was observed in a dental calculus dataset from Oceania[12] (Supplementary Fig. 7), in which the majority of streptococcal reads identified in more than half of the samples fell within species of the Anginosus clade. The microbial community profile of these samples was shown to fall within the known global species variation of ancient calculus, but was enriched in as-yet unidentified taxa.

We next tested whether there was a difference in the relative abundance of *Streptococcus* overall in samples at either end of the clade spectrum (i.e., between samples with highest Sanguinis relative abundance and samples with highest Anginosus relative abundance). We found that samples with high relative abundance of Sanguinis species typically had a high relative abundance of *Streptococcus* overall in the dental calculus, while samples containing predominantly Anginosus clade species had a very low relative abundance of *Streptococcus*, similar to the chimpanzee calculus samples in Fellows Yates et al.[10] (Fig. 2B). We then tested whether the samples with high relative abundance of Anginosus clade species show this pattern due to a loss of Sanguinis clade species, or due to an increase in relative abundance of Anginosus clade species, and found that it is due to a loss of Sanguinis clade species (Fig. 2C), which also explains the overall lower relative abundance of *Streptococcus* in these samples.

We additionally investigated the abundance of several other genera with associations to *Streptococcus* described in the literature[13]. As other species in addition to *Streptococcus* may act as early (e.g., *Actinomyces*, *Neisseria*, *Veillonella*, *Rothia*, *Gemella*, *Granulicatella*, *Eikenella*, *Haemophilus*, *Prevotella)* and intermediate (*Corynebacterium*, *Capnocytophaga*, *Fusobacterium*)[14,15] dental plaque colonizers, we investigated whether these taxa have higher relative abundance in samples with low *Streptococcus* relative abundance. None of the other early colonizing genera we investigated showed this pattern however (Supplementary Fig. 8, Supplementary Table 4), and most had abundance patterns with a moderately strong positive correlation to that of *Streptococcus* ($\rho = 0.3–0.82, p < 0.0001$), suggesting that the physiological conditions of these calculus biofilms with low relative abundance of *Streptococcus* did not support growth of the predominantly aerotolerant early colonizer taxa. *Actinomyces*, a predominantly facultative anaerobe, was the only genus that was uniformly abundant across all ancient dental calculus samples, suggesting that it may play a foundational role in biofilm formation that is minimally impacted by actions of *Streptococcus*. Further, the relative abundance of two late colonizer genera associated with *Streptococcus, Porphyromonas*[5,13,16,17] and *Methanobrevibacter*[18], showed only weak correlations with the relative abundance of *Streptococcus* ($\rho = 0.34, 0.075$, respectively, $p < 0.01$). *Methanobrevibacter* relative abundance in particular varied widely across all samples (Supplementary Fig. 9).

Streptococcus clade abundance is not associated with dental health, time period, geography, or processing laboratory

We next addressed whether the difference in relative abundance of *Streptococcus* and of Sanguinis and Anginosus clade species is correlated with dental pathology, laboratory processing, or sequencing outcomes. In living populations, dental health is the strongest factor associated with altered oral microbiome profiles[19], although this does not seem to be true of ancient

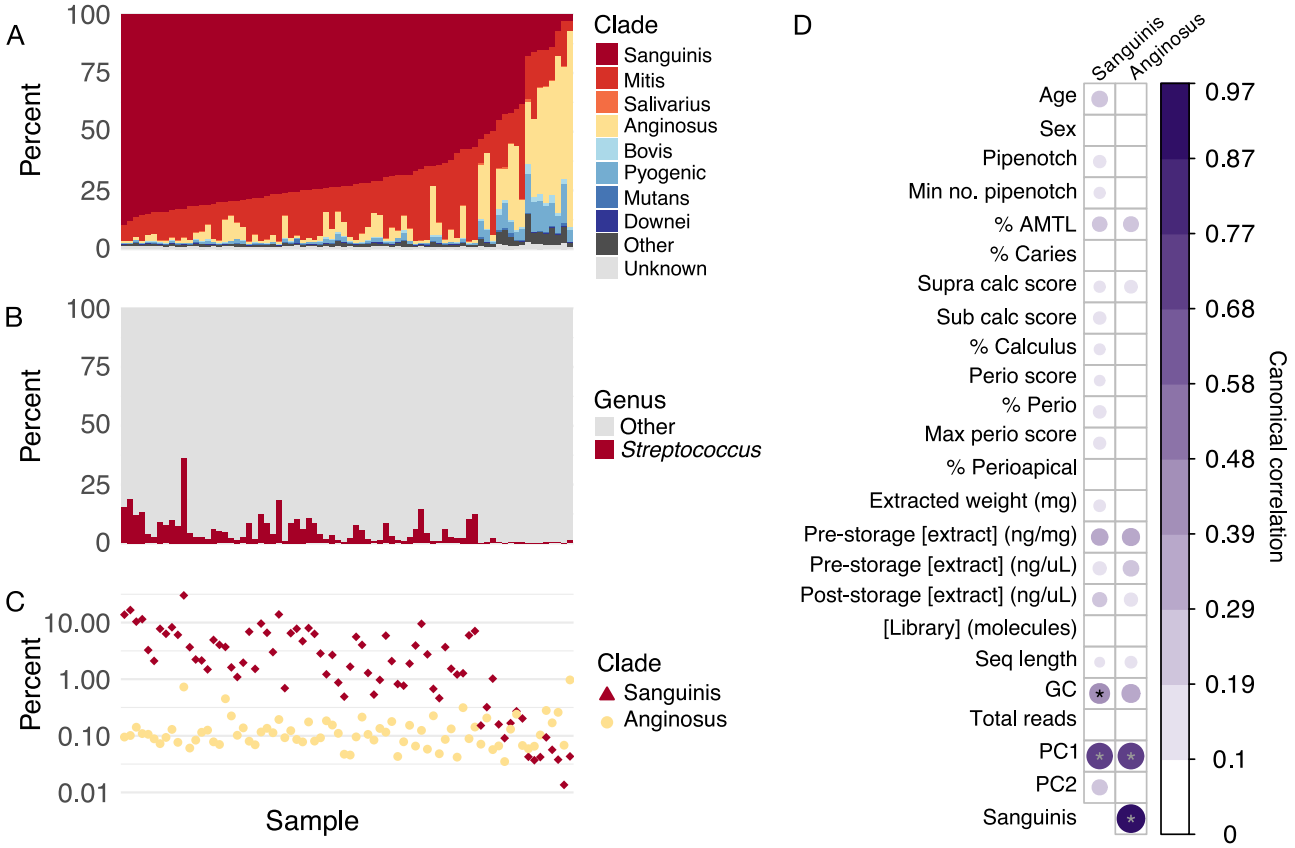

Fig. 3 | Distribution of *Streptococcus* clades in ancient dental calculus samples from Middenbeemster, the Netherlands and canonical correlations with sample parameters. A Percent of *Streptococcus* reads that were assigned to each clade out of all reads assigned to *Streptococcus*, ordered by decreasing relative abundance of Sanguinis clade and increasing relative abundance of Anginosus clade. **B** Percent of reads assigned to species in the genus *Streptococcus* and to all other genera. **C** Percent of reads assigned to species in the Sanguinis and Anginosus clades out of all species-level read assignments. **D** Canonical correlations between the percent of Sanguinis clade or Anginosus clade (from **A**) and archeological metadata, laboratory, and sequencing metrics.

dental calculus microbiome profiles[6,20]. We used a large metadata-rich dataset from a single cemetery in Middenbeemster, the Netherlands[20], which was used for a defined period of time, restricting variation due to geography and sample age.

The dental calculus in this dataset showed the same pattern of *Streptococcus* clade relative abundances that we observed in our global dataset (Fig. 3A–C), and we found no strong correlations (correlation ≥ 0.4, $p < 0.01$) between the relative abundance of Sanguinis or Anginosus clades with any of the laboratory extraction metrics, sequencing outcomes, dental records, or pathology (Fig. 3D). Instead, we found that the relative abundance of the Sanguinis clade was strongly correlated with the PC1 and PC2 loadings in a PCA, as well as with mean library GC content, and relative abundance of Anginosus clade (Fig. 3D). While the correlation between the relative abundance of Sanguinis and Anginosus clades is negative, the correlation between the Sanguinis clade relative abundance and mean GC content is weakly positive (Supplementary Fig. 12).

Individual factors influence the dominant *Streptococcus* clade in ancient dental calculus

In addition to dental health, there are numerous other factors that could potentially affect the oral microbiome species profile. Many of these, such as diet, oral hygiene, drug use, and genetics, are currently difficult or impossible to investigate in ancient dental calculus. However, we may be able to determine if there are factors specific to an individual that influence an individual's microbiome profile without knowing what those factors are. To this end, we assessed whether *Streptococcus* distribution patterns may be intrinsic to an individual by profiling the *Streptococcus* clades in multiple calculus samples collected from teeth across the dentition of four individuals. This dataset includes calculus samples representing nearly half of the dentition of each individual[21].

While three individuals had high relative abundance of *Streptococcus* out of all genera (Fig. 4A, Supplementary Fig. 13) and exhibited mostly Sanguinis-dominated *Streptococcus* clade profiles, one individual showed the alternative pattern (4.1% ± 3.7% vs. 1.3% ± 1.8%, respectively; $p < 0.001$, effect size = 0.44). For this individual, the overall relative abundance of *Streptococcus* was low, and approximately half of the calculus samples had low relative abundance of Sanguinis clade (<50%) and relatively high proportions of Anginosus clade (>5%, Fig. 4, Supplementary Fig. 13). These results suggest that the low relative abundance of *Streptococcus* due to lower abundance of Sanguinis clade species may be an individual-specific phenomenon, with a high probability of being observed in a single piece of calculus randomly selected for analysis.

Global market integration/urbanization minimally impacts modern-day oral microbiome *Streptococcus* clade profiles

While dental calculus is the only oral sample type to readily preserve in the archeological record, there are at least seven distinct surfaces in the mouth, each of which harbors a distinct microbial community[22,23]. Oral *Streptococcus* species are known site-tropists, with different species preferentially prevalent and abundant at selected oral sites such as tongue, saliva, or dental plaque[4]. Further, because recent studies on the impacts of urbanization and industrialization have demonstrated differences in microbiomes of people living in highly urbanized, industrialized conditions relative to those living in less urbanized and industrialized locations[7,23–26], we next assessed whether there are differences between *Streptococcus* clade distributions in oral samples of present-day individuals living across a spectrum of urbanization

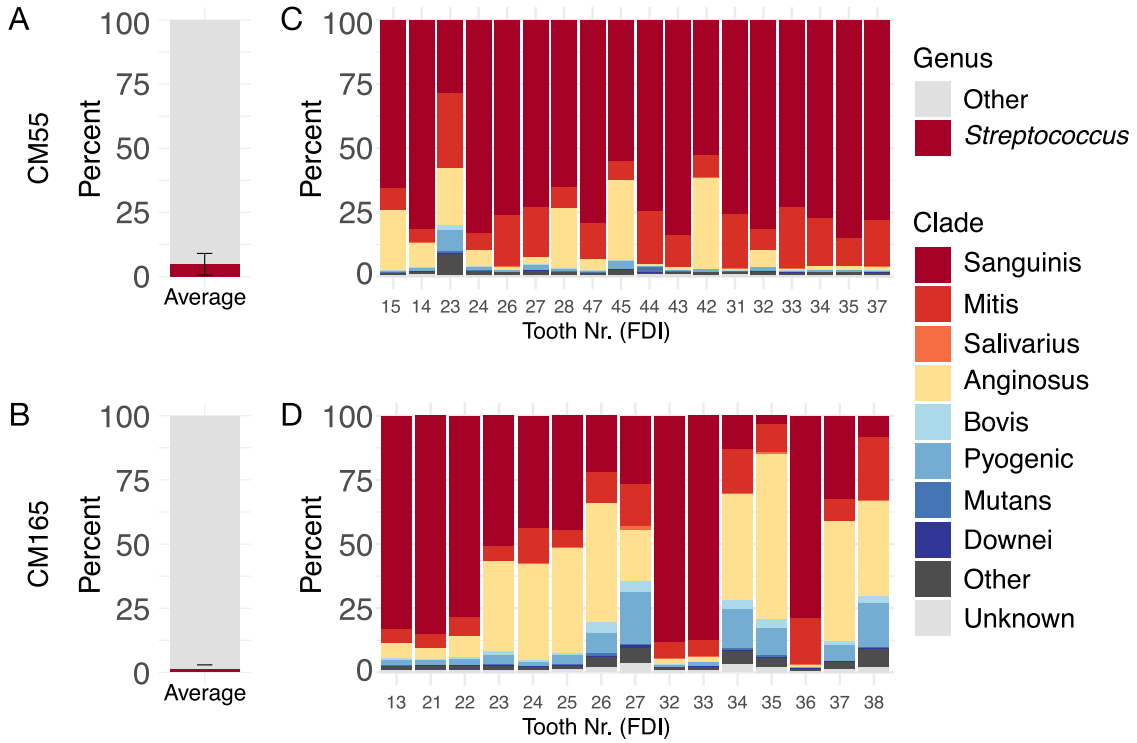

Fig. 4 | Distribution of *Streptococcus* groups in calculus of each tooth sampled from two individuals from the Chalcolithic site (ca. 4500-5000 BP) Camino del Molino, Spain. Average percent of reads assigned to the genus *Streptococcus* compared to all other genera in individual CM55 (**A**) and CM165 (**B**), averaged across all teeth sampled. Relative abundance of reads assigned to species within each *Streptococcus* clade out of all reads assigned to *Streptococcus*, by tooth, in individual CM55 (**C**) and CM165 (**D**). Tooth numbers are in FDI World Dental Federation notation.

and global market integration. We chose three oral sites to focus on: tooth surface (dental plaque/dental calculus), buccal mucosa (cheek swabs), and saliva, as there are publicly available microbiome datasets for these sites from groups with differing levels of urbanization/global market integration (Fig. 1A, Supplementary Table 1).

The tooth surface samples, calculus and plaque, contained predominantly species from the Sanguinis and Mitis clades, and they had notably higher mean relative abundance of Mitis clade (modern calculus 19% ± 8%, $p < 0.01$, effect size = 0.14; non-industrial plaque 27% ± 16%, $p < 0.001$, effect size = 0.35; industrial plaque 43% ± 15%, $p < 0.001$, effect size = 0.65) than ancient dental calculus (13% ± 9%) (Fig. 5A, Supplementary Table 3). Calculus samples have slightly higher relative abundance of Sanguinis than Mitis clades compared to plaque; however, none of the plaque or calculus samples lacked Sanguinis clade species, in contrast to what we observed in ancient dental calculus (Fig. 5C). A small number of plaque samples from individuals with low urbanization/industrialization lifestyles have high relative abundance of *S. mutans*, as previously noted[7], which is due to a rise in relative abundance of this species and not a drop-out of other clade species.

We observed a higher relative abundance of *Streptococcus* in both industrial and non-industrial buccal mucosa (39% ± 19% and 16% ± 8%, respectively) compared to dental calculus (5% ± 3%) or industrial or non-industrial dental plaque (7% ± 6% and 4% ± 3%, respectively) ($p < 0.01$, effect size ≥ 0.39 for each comparison, Supplementary Table 5) and for industrial and non-industrial saliva (14% ± 3% and 19% ± 7%, respectively) compared to dental calculus and industrial or non-industrial dental plaque ($p < 0.01$, 0.11 ≥ effect size ≤ 0.54, for each comparison except industrial saliva vs. non-industrial plaque where $p > 0.05$, Supplementary Table 5) (Fig. 5B), and they contain predominantly Mitis clade species. There were no substantial differences in the clade relative abundances between low and high levels of urbanization/market integration across sample types, although there is overall slightly lower relative abundance of *Streptococcus* in samples from low urbanization/market integration samples (Fig. 5B). Buccal mucosa samples from highly industrialized contexts have much higher relative abundance of *Streptococcus* than from low industrialization contexts (39% ± 18% vs 16% ± 8%, respectively, $p > 0.001$, effect size = 0.76), yet the clade proportions remain consistent. The relative abundance of other early colonizer taxa is distinct by oral site (Supplementary Fig. 10) and their relation with the relative abundance of *Streptococcus* varies (Supplementary Table 4). While both *Capnocytophaga* and *Fusobacterium* have a significant difference in relative abundance and large effect size ($p < 0.001$; effect size = 0.59, 0.65, respectively) between industrial plaque and non-industrial plaque (10% ± 5.2% vs 3.7% ± 2.9% and 3.0% ± 2.1% vs 1.1% ± 0.8%, respectively), *Capnocytophaga* has a higher relative abundance in both oral sites, suggesting it may play a more important role in structuring dental biofilms.

Species-level *Streptococcus* distributions show minor differences by global market economy integration

To investigate which species were dominant at each oral site within the Sanguinis, Mitis, and Anginosus clades, and whether dominant species differed between samples from high and low industrialized/urbanized contexts, we created heatmaps with the relative abundance of all species in these three clades that were detected in any of the samples (Supplementary Figs. 14, 15). In ancient dental calculus and modern non-industrial plaque samples where Sanguinis was the dominant clade, *S. sinensis* was the most abundant Sanguinis clade species (282/361—78%, Fig. 6A, and 32/67—48%, Fig. 6B, respectively), which was unexpected given that we had previously noted that *S. sanguinis* is the most abundant Sanguinis clade species in these samples[6,7,10,12,20]. In the ancient dental calculus samples in which the Anginosus clade had higher relative abundance, *S. constellatus* was the most abundant species in this clade (97/109, 89%).

In modern oral samples, the most abundant *Streptococcus* species were largely consistent between high and low industrialization/urbanization samples for each oral site, with notable exceptions in plaque. Many non-industrial plaque samples have higher relative abundance of the Sanguinis

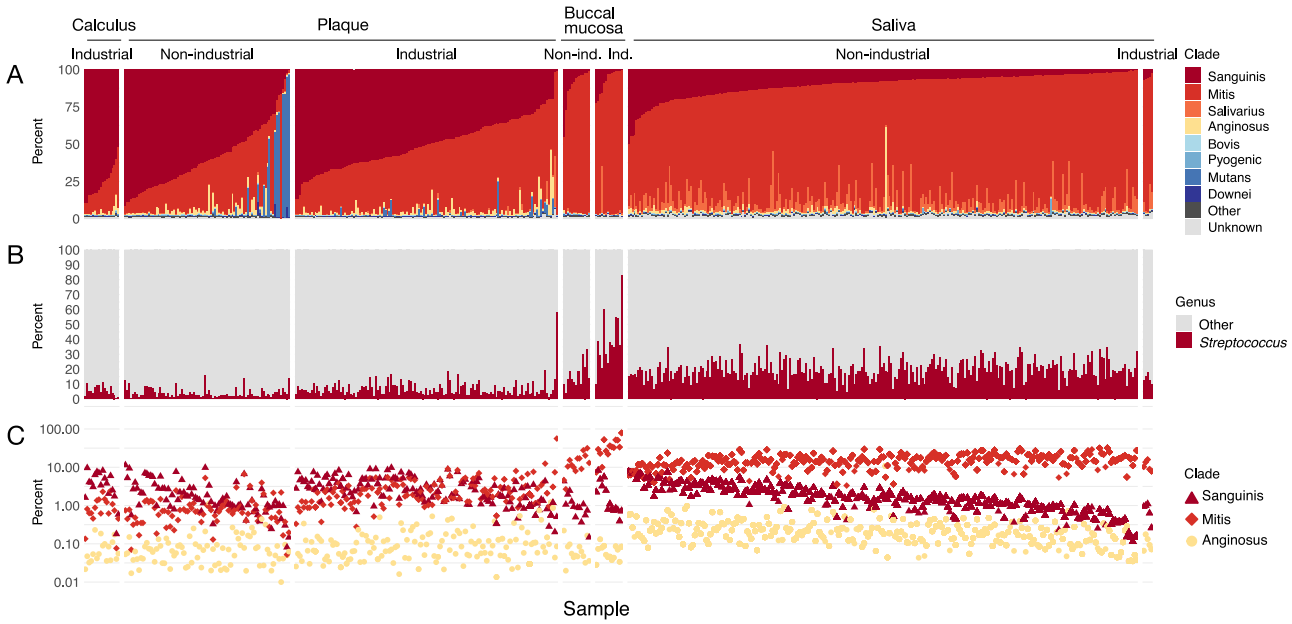

Fig. 5 | Distribution of *Streptococcus* clades in modern oral samples of calculus, dental plaque, buccal mucosa, and saliva. A Percent of *Streptococcus* reads that were assigned to each clade out of all reads assigned to *Streptococcus*, ordered by decreasing abundance of Sanguinis clade and increasing abundance of Mitis clade. **B** Percent of reads assigned to species in the genus *Streptococcus* and to all other genera. **C** Percent of reads assigned to species in the Sanguinis, Mitis, and Anginosus clades out of all species-level read assignments.

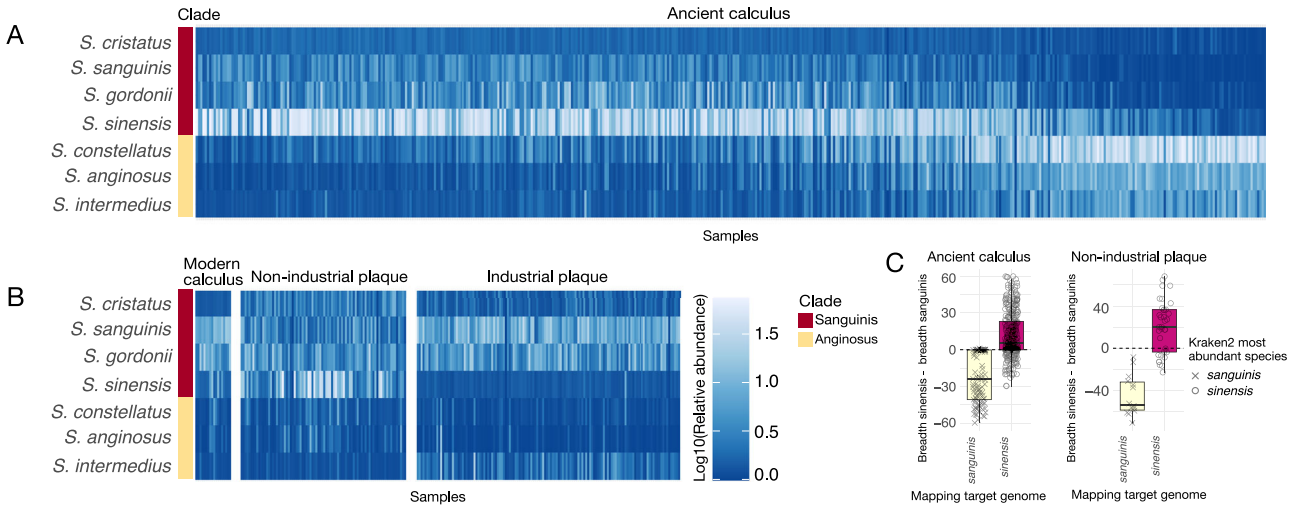

Fig. 6 | Relative abundance of Sanguinis and Anginosus clade species in tooth-adherent oral microbiome samples. Color scale is log10 of the percent relative abundance. **A** Ancient dental calculus. Sample order is identical to Fig. 2. **B** Modern dental calculus and modern dental plaque. Sample order is identical to Fig. 5. **C** Difference in the breadth of coverage (minimum 1X depth) of *S. sanguinis* and *S. sinensis* genomes in ancient dental calculus and modern non-industrial dental plaque. Shapes indicate the *Streptococcus* species that was most abundant in each sample based on profiling with Kraken2 using the GTDB r202 database.

clade species *S. sinensis* and *S. cristatus* than do any of the industrial plaque samples (*S. sinensis* 11% ± 15% *vs* 0.43% ± 0.49%, respectively, $p < 0.001$, effect size = 0.65; *S. cristatus* 2.3% ± 2.2% *vs* 0.88% ±0.70%, respectively, $p < 0.001$, effect size = 0.38). Conversely, the industrialized plaque samples have higher relative abundance of the Sanguinis clade species *S. sanguinis* and the Anginosus clade species *S. intermedius* than many non-industrial plaque samples (*S. sanguinis* 7.5% ± 5.6% *vs* 2.7% ± 2.9%, respectively, $p < 0.001$, effect size = 0.48; *S. intermedius* 1.5% ± 3.3% vs 0.31% ± 0.37%, respectively, $p < 0.001$, effect size = 0.35). In contrast, *S. oralis_S* has higher relative abundance in industrial plaque than non-industrial plaque (2.2% ± 2.6% *vs* 0.34% ± 0.35%, respectively, $p < 0.001$, effect size = 0.65) (Supplementary Fig. 15). Because of the notable difference in relative abundance of *S. sinensis* and *S. sanguinis* between different human sample types, these two species appear to fulfill distinct roles in biofilm establishment and growth, such that their relative abundance is linked to the biofilm developmental stage, with *S. sanguinis* abundant in early-stage biofilms, but being overtaken by *S. sinensis* as the biofilm grows and matures.

Streptococcus sinensis abundance and distribution in dental plaque and calculus

Streptococcus sinensis has been infrequently reported in dental plaque studies to date and was missed in our earlier ancient dental calculus taxonomic profiling[6,10,12,20] because it was not included in the database we used. We used two steps to confirm the presence of *S. sinensis* in the samples. First, we determined which *Streptococcus* species was the most abundant in the Kraken2 taxonomic profile for each sample, then we used genome mapping

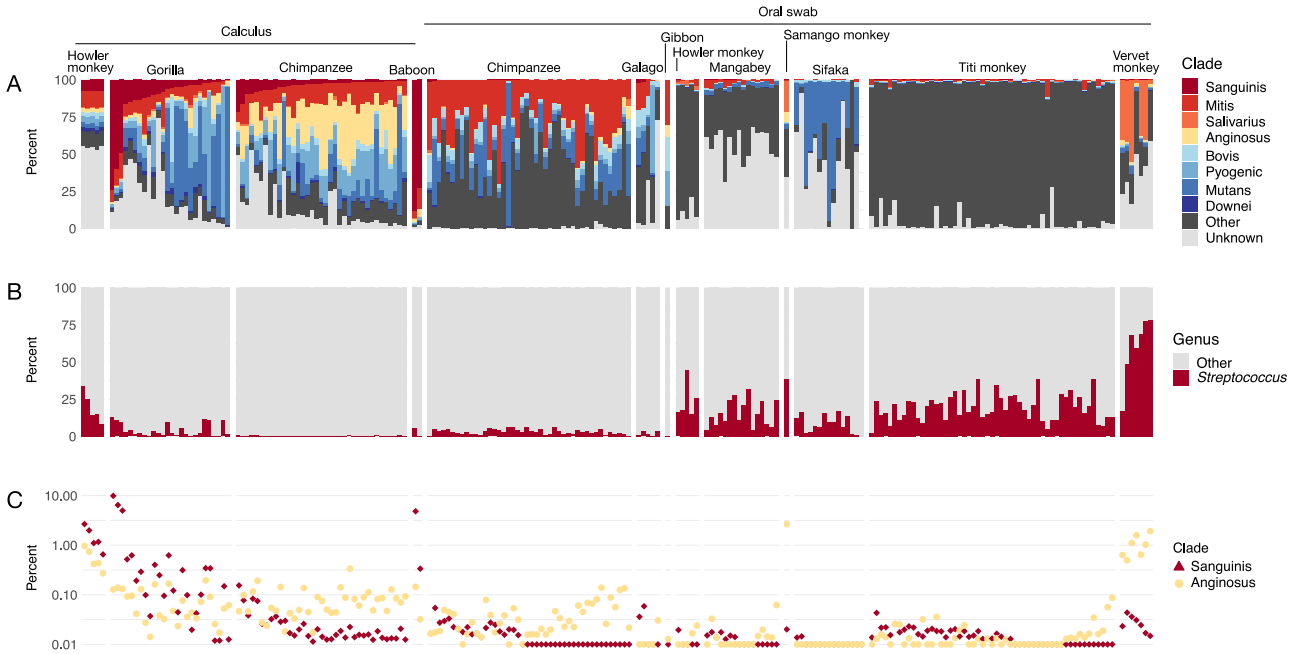

Fig. 7 | Distribution of *Streptococcus* clades in dental calculus and modern oral swabs of non-human primates. A Percent of *Streptococcus* reads that were assigned to each clade, ordered by decreasing relative abundance of Sanguinis clade and increasing relative abundance of Mitis clade. **B** Percent of reads assigned to species in the genus *Streptococcus* and to all other genera. **C** Percent of reads assigned to species in the Sanguinis, Mitis, and Anginosus clades out of all species-level read assignments.

to confirm the assignment of *S. sinensis* (Fig. 6C, Supplementary Figs. 16, 17). After mapping all human dental calculus datasets against genomes for *S. sanguinis* and *S. sinensis*, we found that Kraken2 read assignments correlated well with mapping breadth and depth of coverage for each species in a majority of samples; however, for a small number of samples, we observed greater mapping to the alternative reference genome, suggesting that the most abundant Sanguinis clade species in these samples is a closely-related, not-yet-described species. Future assembly of MAGs from these samples may help resolve the identity of the highly abundant Sanguinis clade species found in these samples; however, at present, MAG assembly and binning remain challenging for *Streptococcus*[27].

Sanguinis clade species are minimally represented in non-human primate oral microbiomes

We next assessed whether non-human primates have distinctive distributions of *Streptococcus* clades compared to humans by examining the species present in calculus and oral swabs. The overall relative abundance of *Streptococcus* varied substantially across non-human primates and oral sites, with *Streptococcus* species being generally lower abundance in dental calculus than in oral swabs (Fig. 7). Chimpanzees had among the lowest relative abundance of *Streptococcus* in both dental calculus (0.33% ± 0.23%) and oral swabs (3.1% ± 1.6%), while *Streptococcus* made up more than half of the genera identified in oral swabs from vervet monkeys (59% ± 20%). Each non-human primate species had a distinct distribution of *Streptococcus* clades, and many of the most abundant *Streptococcus* species did not fall into previously described clades (Fig. 7A, Supplementary Table 3), particularly in the oral swab samples. Intriguingly, we observe distinct *Streptococcus* clade profiles between dental calculus and oral swabs for both Chimpanzees and Howler monkeys, the only non-human primates for which we have both data types, hinting that oral *Streptococcus* site tropism is a characteristic of primates and not unique to humans. A PCA of beta-diversity differences based on the full species profile, not just *Streptococcus*, highlights how the relative abundance of particular clades of *Streptococcus* may contribute to overall diversity differences between hosts and oral sites (Fig. 8, Supplementary Fig. 18).

In contrast to humans, only 5 of 107 non-human primate dental calculus samples (~5%) had a predominantly Sanguinis clade profile: three gorilla samples and two baboon samples. While this represents a minority profile for gorillas (3 of 26), further study of baboon dental calculus is needed as only two dental calculus samples were sufficiently preserved for analysis. From these results, it appears that dental calculus dominated by Sanguinis and Mitis clades may be a particularly human feature, although further investigation of baboon samples will be necessary to confirm this. Other early colonizer genera were generally infrequently detected and low in relative abundance across non-human primate oral samples, with a few exceptions (Supplementary Fig. 11).

Streptococcus gene content differs between hosts and oral sites

Given the ecological diversity and distinct niches occupied by *Streptococcus* species in human and non-human oral microbiota, we tested whether there are particular genes in *Streptococcus* that are enriched or depleted in the primate hosts and oral sites. This includes genes that are significantly enriched in human shedding surfaces (buccal mucosa/saliva) compared to non-shedding surfaces (dental plaque and calculus), between human and non-human primates for either shedding and non-shedding surfaces, and between human ancient and modern dental calculus (Supplementary Table 7). A total of 2,615,774 genes attributed to *Streptococcus* were identified across all samples. The majority of genes that we found significantly enriched between groups were genes of unknown function (Supplementary Table 7), which limits the conclusions we can draw regarding functional specificities of *Streptococcus* in different hosts and at different oral sites. This highlights the necessity of continued laboratory functional characterization of host-associated microbes to improve our understanding of microbial metabolic functioning and the impacts it has on both the microbiome community and the host.

Discussion

The highly diverse genus *Streptococcus* shows distinct host and site-tropism within the oral cavities of primates. While these differences may reflect host differences in salivary composition[28] and dietary differences among primate species[10], human-associated *Streptococcus* appear to a small extent to also be

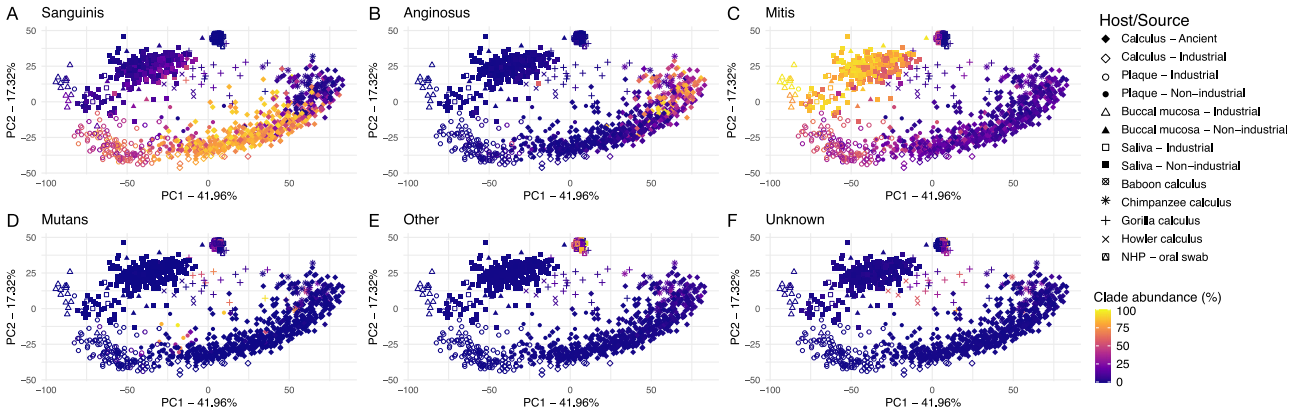

Fig. 8 | Principal components analysis (PCA) plot of ancient and modern human and non-human primate oral microbiomes based on all species detected, not just *Streptococcus*. Shapes indicate sample host, colored by proportion of reads in each clade out of all *Streptococcus* reads: **A** Sanguinis clade, **B** Anginosus clade, **C** Mitis clade, **D** Mutans clade, **E** Other clades, **F** Unknown clades.

affected by human hygiene practices and level of global market economy integration/urbanization. Humans appear to be uniquely enriched in species from the Sanguinis and Mitis clades, which are otherwise rare in non-human primates with the possible exception of baboons. As the savannah territory and tuber consumption of baboons is hypothesized to be similar to that of early humans, the ability of Sanguinis and Mitis clade species to utilize dietary starch may have provided an ecological advantage that lead to the dominance of these species in the mouths of starch-consuming primates[10]. This deep evolutionary dietary transition has been maintained by nearly all human populations across the globe today, where the primary source of starch, such as the grain or tuber variety most commonly consumed, may differ across populations but starch consumption, albeit in varying amounts, remains a nearly ubiquitous feature of human diets[29], with few exceptions.

Using a large and diverse dataset of more than 400 well-preserved dental calculus samples, we were able to confirm that human ancient dental calculus can be grouped into two categories based on *Streptococcus* relative abundance profiles, which are consistently found across time and geography. By using publicly available data that was produced in different labs using different DNA extraction and library build techniques, we also demonstrate that this observation is robust to laboratory processing methods, and is likely a true biological phenomenon. The factors driving these distinct ecological profiles are not yet known, and although we found no associations between streptococcal profiles and osteological measures of oral pathology, we cannot rule out that there may be associations with specific immune responses or soft tissue pathology, which cannot be measured by DNA analysis or osteological examination of archeological remains.

However, other host physiology-related explanations seem likely drivers of different *Streptococcus* clade profiles. For example, early studies of in situ biofilm formation on tooth surfaces reported two distinct groups of study participants that developed plaque at different rates[30–32]. Further work found differences in salivary properties and composition between "rapid" and "slow" plaque-forming participants that influenced *Streptococcus* species[33–35]. We speculate that the *Streptococcus* profiles of ancient dental calculus may reflect the "rapid" and "slow" plaque formation documented in these early plaque development studies; however, whether there is a relationship between the two would require further in situ or in vitro modeling to understand.

The near absence of early colonizer species other than *Actinomyces* in dental calculus with low overall *Streptococcus* relative abundance suggests that those biofilms may have a distinct pattern of species acquisition and turnover that differs from the well-described oral biofilm succession model[8]. In ancient dental calculus, we found moderate to strong correlations between the relative abundance of *Streptococcus* and of most early/intermediate colonizer species we examined, which was particularly striking for *Corynebacterium* and *Capnocytophaga*, two important species for spatial structuring of the dental plaque biofilm[13,36] and are abundant across non-industrial and industrial dental plaque samples.

Capnocytophaga was on average less abundant in non-industrial plaque samples than industrial plaque samples, which hints that overall plaque biofilm species and structural turnover to a more anaerobic, mature community may be related to a loss of this genus, given that the taxonomic profile of these samples appears intermediate between ancient dental calculus (which represents fully mature dental biofilms) and dental plaque from industrialized populations (which represents earlier stage dental biofilm formation due to toothbrushing). The low relative abundance of *Capnocytophaga* in gorilla and chimpanzee calculus suggests that it may be particularly important in structuring human plaque biofilms, despite being a core member of the primate oral microbiome[10], as at higher relative abundance it may affect a wider area of the biofilm, interact with a wider variety of species, and/or interact with a higher number of total microbial cells (of one or more species). As *Actinomyces* are uniformly abundant in the tooth-adherent biofilm samples we examined, including ancient and modern human dental calculus, human dental plaque, and non-human primate dental calculus, this genus may be the ancestral early colonizer for dental biofilms, with humans later adding *Streptococcus*, with concomitant changes in the taxa that successively colonize the biofilm and now differentiate human and non-human primate dental biofilms.

In previous work, we reported *S. sanguinis* to be the most abundant *Streptococcus* species in human dental calculus[6,7,10,20], but here we find that while *S. sanguinis* is present, *S. sinensis* is actually the most abundant Sanguinis clade species present in both ancient dental calculus and modern non-industrial plaque samples. This discrepancy is due to the fact that the custom RefSeq database used for taxonomic classification in prior studies did not include any *S. sinensis* genomes, and reads from this species were likely mis-assigned to the closely-related *S. sanguinis* instead, a known phenomenon in taxonomic assignments using incomplete databases[37]. Inconsistencies in NCBI taxonomy may also account for the absence of *S. sinensis* in taxonomic profiles, such as with the NCBI genome *Streptococcus* sp. DD04, which has been reclassified as *S. sinensis* by GTDB. Despite their close phylogenetic relationship, *S. sinensis* and *S. sanguinis* appear to thrive in different biofilm environments.

Streptococcus sinensis was originally isolated from a patient with infective endocarditis[38] and was later found to be part of the oral microbiome[39], but its physiology and biochemistry have not yet been extensively explored[40]. Although it was not included in the phylogeny presented in Richards et al.[3], our clustering of the type strain genome by average nucleotide identity placed it in the Sanguinis clade, with high similarity to *S. cristatus*, reflecting the phylogenetic placement of the species reported by others[41,42]. The high prevalence and abundance of *S. sinensis* in both ancient dental calculus and modern plaque from populations with low

global market integration, but not in plaque from highly industrialized populations suggests that *S. sinensis* may therefore represent the first described VANISH (volatile and/or associated negatively with industrialized societies of humans) taxon of the human oral microbiota[43,44].

The high relative abundance of *S. sinensis* in ancient and non-industrialized oral biofilms suggests that the species prefers more mature, anaerobic biofilm environments, although further work characterizing *S. sinensis* growth conditions is needed to confirm this. This pattern contrasts with the pattern observed in dental plaque from heavily industrialized populations, in which early biofilm colonizer streptococcal species preferring more aerobic environments, such as *S. sanguinis* and *S. gordonii*, predominate. Regular and/or frequent toothbrushing to remove dental plaque in heavily studied industrial populations may prevent the oral biofilm from maturing to an anaerobic, reduced state that can support growth of *S. sinensis*, potentially explaining why it is not commonly reported in dental plaque of heavily industrialized populations. Dental hygiene, in particular regular toothbrushing, has been proposed to account for differences in species prevalence and abundance between dental plaque and ancient dental calculus as well as between dental plaque from populations with high and low global market integration and urbanization[6,7,20], and may be one of the strongest factors shaping oral microbiome composition today.

Certain genetic differences between *S. sinensis* and. *S. gordonii/sanguinis* may explain the differences in relative abundance between these species in ancient dental calculus. Nearly all sequenced genomes of *S. gordonii* and *S. sanguinis* contain a gene encoding the protein amylase-binding protein A (AbpA) that allows them to bind human salivary amylase[10]. Expression of AbpA could offer a colonization advantage to *S. gordonii* and *S. sanguinis*, as salivary amylase is part of the acquired enamel pellicle that forms the base layer on which dental plaque biofilms grow[45]. This protein was suggested to play a role in shaping the human-specific dental plaque biofilm[10], and the near ubiquity of the gene in sequenced genomes of *S. gordonii* and *S. sanguinis* suggests it plays an important role in the physiology of these species.

In addition, many *S. gordonii* and *S. sanguinis* genomes contain *gspB* or a homolog (*hsa, srpA*) that may play a role in substrate binding or nutrient acquisition, as it allows binding to sialic acids[46], which are abundant on salivary mucins. In contrast, none of the four *S. sinensis* genomes in NCBI contain *abpA* or *gspB/hsa/srpA*, suggesting *S. sinensis* does not share the colonization advantage of *S. gordonii* and *S. sanguinis* for early biofilm formation. The high relative abundance of *S. sanguinis* in dental plaque from industrial populations that practice regular toothbrushing, which represents early-stage dental plaque biofilms, and the contrasting high relative abundance of *S. sinensis* in ancient dental calculus and dental plaque of groups with low global market economy integration, which represent more mature dental plaque biofilms, supports the preference of these *Streptococcus* species for different stages of biofilm development.

In support of a role for ApbA in oral microbiome composition, human salivary amylase (AMY1) gene copy number was shown to be positively correlated with salivary microbiome richness[47], indicating that AMY1 and dietary starch consumption may play a role in shaping the salivary oral microbiome. However, the individual taxa most strongly associated with high AMY1 copy numbers were the genera *Porphyromonas* and *Prevotella*, a curious finding given that many oral *Porphyromonas* species do not use carbohydrates for energy sources, raising questions about how a gene that processes dietary starch and controls free starch availability in the mouth might promote proliferation of taxa that largely do not use starch or its breakdown products. Further work on the interactions of AMY1 and the dental plaque microbiome are needed to understand the interactions of human genetics, salivary protein composition, diet, and oral microbiome species composition and abundance.

An inverse relationship between the relative abundance of *Streptococcus* and *Methanobrevibacter* has been reported in ancient dental calculus[20], which is argued to reflect the overall oxygen tolerance of other abundant species in these samples. Although we see this trend in our large ancient dental calculus dataset here, *Methanobrevibacter* relative abundance is highly variable between samples, ranging from 0% to nearly 40% even across samples in which *Streptococcus* is nearly absent, and the correlation between the genera is weak. *S. constellatus* has been reported to support *Methanobrevibacter* growth[18], and this is the Anginosus clade species that is most abundant in the dental calculus with low *Streptococcus* relative abundance, perhaps providing support for a metabolic interdependence between the two. However, the ecological role of *Methanobrevibacter* in shaping oral communities with low *Streptococcus* relative abundance may also be filled by other taxa in its absence, calling into question the importance of *Methanobrevibacter* itself in shaping biofilm community structure[48], and emphasizing instead a set of as yet undefined metabolic features that may be shared by numerous taxa.

Working with non-human primate oral samples creates particular challenges for accurate taxonomic profiling, as the microbiota in these hosts are not extensively characterized biochemically or genomically, and few sequenced genomes are publicly available. However, our taxonomic profiling approach appears to be sufficient to capture the diversity of a wide range of host oral streptococci, despite differences in the number of species per clade in the database we used. For example, of the 303 *Streptococcus* species detected across all samples, 162 (53%) belong to the Mitis clade, yet we do not observe a bias to higher relative abundance of Mitis clade species across or within sample types. The next two most prevalent clades in the database by species count, Pyogenic and Other (34 and 32 species, respectively) are nearly undetected in human samples, and their relative abundance varies substantially across non-human primates. The Anginosus clade includes the fewest species (five) yet is consistently detected in humans although not non-human primates, while the Other and Unknown (32 and 11 species, respectively) clades are rarely detected in humans but represent upwards of 50% of the reads in seven non-human primate host species. This indicates that the species found in human oral samples likely represent all of the clades with species found in the human oral cavity, but that the species present in non-human primate mouths probably fall into currently undescribed clades that need further phylogenetic work to define. Broader sampling, cultivation, sequencing, and genomic analyses of non-human primate oral streptococci are needed to clarify this.

Despite high genetic heterogeneity and horizontal gene transfer within the oral streptococci, species-specific preferences for distinct oral niches appear to be relatively consistent across time, geography, and cultural practices in humans, and may be characteristic of non-human primates as well. The role oral streptococci fill in dental plaque biofilm development appears to strongly affect the mature biofilm species profile. Specialization of the Sanguinis clade species within the human oral cavity concomitant with the increase in starch in human diets may have been critical step in shaping the human oral microbiome profile that exists today, while the recent adoption of regular toothbrushing may be driving population-wide loss of species like *S. sinensis* from the oral microbiome. Further investigation of ancient dental calculus and oral biofilm samples from under-studied living populations is necessary to provide a better understanding of the evolutionary history of oral streptococci and how changing cultural practices are impacting oral microbiome communities today. Employing an anthropological approach informed by human evolutionary biology and paleogenomics enables us to broaden our understanding of what makes a healthy, stable oral biofilm community.

Methods

Data selection and download

Ancient dental calculus metagenomic data from studies published prior to June 2022, which included more than 2 samples that were Illumina shotgun sequenced, and were not explicitly used for extraction or decontamination method testing were downloaded from the European Nucleotide Archive (ENA)[6,10–12,20,49–56]. All samples are listed on the Ancient Metagenome Directory[57]. Modern human[6,7,10,23–26] and non-human primate[10,58–62] Illumina shotgun sequenced data were likewise downloaded from the ENA. A list of all samples and accessions is in Supplementary Table 1. This resulted in a starting dataset of 541 ancient human calculus samples, 537 modern

human samples (18 calculus, 220 plaque, 28 buccal mucosa, 271 saliva), 107 dental calculus samples from non-human primates (chimpanzee, gorilla, baboon, howler monkey), and 197 oral swabs from modern non-human primate (chimpanzee, galago, gibbon, howler monkey, mangabey, samango monkey, sifaka, titi monkey, vervet monkey). Samples were plotted on a map using the R packages spData[63], sf[64,65], ggplot[66], and ggpointgrid[67], and sample age histograms were plotted using the R package ggridges[68].

Data processing

Raw fastq files for all samples were processed with the nf-core/eager pipeline[69]. Settings were left in default except for bwa on ancient samples, for which the following flags were used -l 32 -n 0.01. All samples regardless of host species were mapped against the human genome, and all reads that mapped were discarded from downstream analysis. All remaining reads were taxonomically classified using Kraken2[70] and the GTDB[71–73] r202 database provided on the Struo2 ftp server[74]. Metaphlan-formatted output tables were joined using the KrakenTools[75] script combine_mpa.py. Differences in the proportion of reads that were assigned taxonomy are not related to the sequencing depth (Supplementary Fig. 1). We confirmed that the *Streptococcus* species profile was similar to that published by Fellows Yates et al.[10] (Supplementary Fig. 5), and placed the *Streptococcus* genomes found in the GTDB r202 database but not the custom RefSeq database used by Fellows Yates et al. into clades by clustering based on ANI with dRep[76] (Supplementary Table 2 for details see the section Assessment of *Streptococcus* clade distributions).

Sample preservation assessment

Preservation of the oral microbiome community in all samples was assessed with the R package cuperdec[10] to ensure the oral microbiome had not been contaminated by other microbial sources such as skin or environmental taxa. Samples were grouped as non-human primate, ancient human calculus, or modern human oral, and preservation cut-offs were determined individually for each of the three groups (Supplementary Figs. 2–4). All samples that were determined to be poorly preserved were discarded from downstream analysis (Supplementary Table 1). This left 482 ancient human samples, 532 modern human samples (18 calculus, 220 plaque, 28 buccal mucosa, 267 saliva), 70 ancient primate calculus samples, and 147 modern primate oral swabs.

Comparison of MALT RefSeq and Kraken2 GTDB r202 Streptococcus profiles

Fellows Yates et al.[10] used MALT with a custom RefSeq database to profile the species in their ancient calculus samples. As MALT requires a substantial amount of memory to run and takes many hours per sample, and this database is now out-dated, it was not feasible to use MALT for profiling the samples in this study. We chose to use Kraken2 and the GTDB database r202 (the most recent release at the time this study was performed) because of Kraken's speed, and the comprehensive species representation in GTDB. Tables listing the taxa identified in each sample can be found on the project github site. To confirm that the *Streptococcus* species profiles we found with Kraken2/GTDB were similar to that seen with MALT/customRefSeq, we compared the *Streptococcus* clade profiles for two datasets for which we already had MALT/customRefSeq species tables: Fellows Yates et al.[10], and Velsko et al.[20]. We found the *Streptococcus* clade profiles were highly comparable (Supplementary Fig. 5A-C), although there were some notable differences with the non-human primate clade distributions. This was likely due to the differences in *Streptococcus* species in each database (Supplementary Fig. 5D).

Assessment of *Streptococcus* clade distributions

All *Streptococcus* genomes with hits in any sample were assigned to a phylogenetic clade from Richards et al.[3]. To be able to group *Streptococcus* genomes that had reads assigned by Kraken2 but for which the clade was unknown (because it is unnamed, or uploaded to NCBI after the publication of Richards et al.[3], we clustered all *Streptococcus* genomes with hits in any sample by average nucleotide identity (ANI) with the wrapper dRep[76] using the programs MASH[77] and fastANI[78]. Species clusters were defined as genomes with ≥95% ANI. *Streptococcus* genomes were then assigned to a clade, defined by Richards et al.[3], based on dRep primary clustering, referring to the named species found within each species cluster. If a genome fell outside of these clades with named species that were included in Richards et al.[3] but was not basal to all named clades in the dendrogram produced by dRep, it was assigned "Other", while genomes that fell outside of these named clades and were basal to all known/named clades were assigned "Unknown" (Supplementary Table 2).

We calculated the proportion of reads from *Streptococcus* species in each of the *Streptococcus* clades out of all *Streptococcus* species-assigned reads in each sample. Further, within each sample, we calculated the proportion of reads that were assigned to any species in the genus *Streptococcus* vs. all other genus assignments. We additionally calculated the proportion of reads assigned to all species per clade out of all species assignments. Lastly, we calculated the relative abundance of all species in each sample and selected out the *Streptococcus* species for plotting in a heat map. Percentages were log10-transformed after adding a value of +1 to all percents, to better visualize the different relative abundances across species and samples. The relative abundance of additional taxa was calculated in the same way.

Correlation coefficients between the relative abundance of *Streptococcus* and other genera or between *Streptococcus* clades were calculated with two compositionally-aware data analysis (CoDA) approaches: Pearson correlation coefficient ρ was calculated on a center log ratio (CLR)-transformed count matrix with the r package rstatix[79], an approach shown to perform comparably or identically to explicitly CoDA-designed software tools[80,81], and CoDA coefficient ρ was calculated on the same CLR-transformed count matrix with the R package propr[82,83]. Principal components analysis was performed on the CLR-transformed Kraken2 full species table with the R package mixOmics[84].

Principal components analysis was performed on the center-log ratio (CLR)-transformed Kraken2 table of all species with the R package mixOmics[84]. The table was first filtered to include only species-level assignments, and species with an abundance less than 0.001% were filtered out to remove spurious low-abundance hits. The proportion of each *Streptococcus* clade used to color the PCA plots are those proportions plotted in panel A of Figs. 1, 4, and 6.

Correlations with oral pathology

Correlations between the proportion of *Streptococcus* clades and oral pathology in the historic Middenbeemster dataset were assessed with canonical correlation analysis, as performed in Velsko et al.[20], following[85]. Input tables contained selected metadata categories (Supplementary Table 1 from source publication), as well as the proportion of Sanguinis and Anginosus clade *Streptococcus* species out of all *Streptococcus* species detected from the Kraken2 taxonomic table generated for this study.

PCA was performed for only well-preserved samples from Middenbeemster[20] on the center log-ratio-transformed species table produced from Kraken2 profiling with the GTDB r202 database. The function canCorPairs from the R package variancePartition[86,87] was used to perform canonical correlations, while the cor.mtest function in the R package corrplot[88] was used to perform statistical tests. Correlation matrix plots were generated with the function corrplot in the same package. To focus on the strongest correlations, we considered only correlations ≥0.4 with a significance of $p \leq 0.01$ to be significant.

Assessment of species distributions of additional taxa

We additionally investigated the abundance of eleven selected genera for their role in early biofilm colonization and development (*Actinomyces*, *Eikenella*, *Gemella*, *Granulicatella*, *Haemophilus*, *Neisseria*, *Prevotella*, *Veillonella*), or biofilm structuring (*Capnocytophaga*, *Corynebacterium*, *Fusobacterium*), as well as two late colonizer species that are known to interact with *Streptococcus* (*Porphyromonas gingivalis*, *Methanobrevibacter oralis*). To determine the abundance of these taxa, we used the same

approach as we used to calculate the proportion of *Streptococcus* vs all other genera. Within each sample, we calculated the proportion of reads that were assigned to any species in each of the above genera vs. all other genus assignments, or, for the two late colonizer species we calculated the proportion of reads assigned to each species vs. all other species assignments.

Correlation coefficients between the relative abundance of *Streptococcus* and other genera were calculated with two CoDA approaches: Pearson correlation coefficient ρ was calculated on a CLR-transformed count matrix with the R package rstatix[79], an approach shown to perform comparably or identically to explicitly CoDA-designed software tools[80,81], and CoDA coefficient ρ was calculated on the same CLR-transformed count matrix with the R package propr[82,83]. For the Pearson correlation $p < 0.05$ was considered significant. For the CoDA ρ the cut-off value for an FDR of 5% was calculated.

Streptococcus sanguinis and *S. sinensis* genome mapping

Representative genomes of 9 *S. sanguinis* species from GTDB classification and the 4 available *S. sinensis* genomes were downloaded from NCBI (Supplementary Table 6) and concatenated into a single fasta file. All non-industrial plaque samples and all human ancient calculus samples were mapped to this concatenated file with bwa aln using the flags -l 32 -n 0.01. Variants were called with bcftools mpileup, and the calls were filtered for those with a quality greater than 20 and a depth of ≥2. The *S. sanguinis* (GCF_003943655.1) and *S. sinensis* (GCF_000767835.1) genomes with the highest breadth and depth of coverage were selected and all samples mapped against these individual genomes with bwa aln using the same parameters as above, then variants were called with bcftools in the same way as above. The vcf files were converted to tsv using bcftools norm -m and bcftools query -f "%CHROM\t%POS\t%REF\t%DP\t%ALT\n" to separate multiallelic snps into one observed snp per line per site for data analysis.

In addition, we mapped all non-industrial plaque samples and all human ancient calculus samples to a file containing *S. sanguinis* (GCF_003943655.1) and *S. sinensis* (GCF_000767835.1) genomes as well as four MAGs of Sanguinis clade *Streptococcus* that were assembled from modern and ancient dental calculus in ref. 27, using bwa aln and the same parameters as above. This allowed us to determine if the dominant Sanguinis clade species in our samples may be one that was not represented by genomes in NCBI RefSeq or Genbank.

Gene content enrichment

To assess differences in the gene content between sample groups, we annotated the gene content of all samples using the Global Microbial Gene Catalog (GMGC)[89]. For gene coverage normalization across samples as reads per kilobase, we used RRAP[90]. Genes attributed to *Streptococcus* based on the GMGC-provided taxonomy list were subsetted from the normalized table and used to assess differences in presence/absence of genes between groups. In detail, collapsed reads from all samples were mapped against the GMGC database GMGC10.95nr with bowtie2 and the following flags: -D 20 -R 3 -N 1 -L 20 -i S,1,0.50 --no-unal. This allowed for mapping ancient damaged reads, but was applied to all samples, both ancient and modern.

The bowtie2-mapped bam files of each sample mapped against the GMGC10.95nr were used as input for RRAP[90] for normalization by reads per kilobase, with the parameters –skip-indexing and –skip-rr. Since reads used for mapping were collapsed read pairs, we used the collapsed read fastq file as both a read 1 and a read 2 fastq file for RRAP. Three groups of samples were run individually through RRAP: the modern human oral samples, the ancient human calculus, and the non-human primate samples. This produced three output files of gene content normalized by reads-per-kilobase, one for each sample group. Significant differences in gene presence/absence between groups was assessed with a Wilcox test with FDR correction and effect size using the R package rstatix[79]. Prior to testing for significant differences, tables were subsetted to include only genes present in at least 30% of all samples being compared. Genes were considered significantly enriched in one group if multiple test-corrected p values were less than 0.05 and the effect size was at least 0.4. Tables listing the enriched genes in each comparison can be found on the project github site.

Data visualization

All plots were generated in R with ggplot2[66] unless otherwise noted. Plots were assembled with the R package patchwork[91]. Statistics were calculated with the R package rstatix[79]. The viridis color package[92] was used for continuous colors.

Data availability

All data analyzed here was published in other studies and is publicly available. All accessions are listed in Supplementary Table 1.

Code availability

The underlying code for this study is available at https://github.com/ivelsko/oral_streptococcus_clades.

Received: 24 May 2024; Accepted: 19 December 2024;
Published online: 17 January 2025

References

1. Richards, V. P. et al. Population gene introgression and high genome plasticity for the zoonotic pathogen *Streptococcus agalactiae*. *Mol. Biol. Evol.* **36**, 2572–2590 (2019).
2. Brouwer, S. et al. Pathogenesis, epidemiology and control of Group A Streptococcus infection. *Nat. Rev. Microbiol.* **21**, 431–447 (2023).
3. Richards, V. P. et al. Phylogenomics and the dynamic genome evolution of the genus *Streptococcus*. *Genome Biol. Evol.* **6**, 741–753 (2014).
4. McLean, A. R., Torres-Morales, J., Dewhirst, F. E., Borisy, G. G. & Mark Welch, J. L. Site-tropism of streptococci in the oral microbiome. *Mol. Oral. Microbiol.* **37**, 229–243 (2022).
5. Morillo-Lopez, V., Sjaarda, A., Islam, I., Borisy, G. G. & Mark Welch, J. L. Corncob structures in dental plaque reveal microhabitat taxon specificity. *Microbiome* **10**, 145 (2022).
6. Velsko, I. M. et al. Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome* **7**, 102 (2019).
7. Velsko, I. M., Gallois, S., Stahl, R., Henry, A. G. & Warinner, C. High conservation of the dental plaque microbiome across populations with differing subsistence strategies and levels of market integration. *Mol. Ecol*. https://doi.org/10.1111/mec.16988 (2023).
8. Kolenbrander, P. E. et al. Bacterial interactions and successions during plaque development. *Periodontology 2000* **42**, 47–79 (2006).
9. Kolenbrander, P. E. et al. Communication among oral bacteria. *Microbiol. Mol. Biol. Rev.* **66**, 486–505 (2002).
10. Fellows Yates, J. A. et al. The evolution and changing ecology of the African hominid oral microbiome. *Proc. Natl. Acad. Sci. USA.* **118**, e2021655118 (2021).
11. Warinner, C. et al. Pathogens and host immunity in the ancient human oral cavity. *Nat. Genet.* **46**, 336–344 (2014).
12. Velsko, I. M. et al. Exploring the potential of dental calculus to shed light on past human migrations in Oceania. *Nat. Commun.* **15**, 10191 (2024).
13. Mark Welch, J. L., Rossetti, B. J., Rieken, C. W., Dewhirst, F. E. & Borisy, G. G. Biogeography of a human oral microbiome at the micron scale. *Proc. Natl. Acad. Sci. USA* **113**, E791–E800 (2016).
14. Ritz, H. L. Microbial population shifts in developing human dental plaque. *Arch. Oral. Biol.* **12**, 1561–1568 (1967).
15. Diaz, P. I. et al. Molecular characterization of subject-specific oral microflora during initial colonization of enamel. *Appl. Environ. Microbiol.* **72**, 2837–2848 (2006).
16. Wang, Q., Wang, B.-Y., Pratap, S. & Xie, H. Oral microbiome associated with differential ratios of *Porphyromonas gingivalis* and *Streptococcus cristatus*. *Microbiol. Spectr.* **12**, e0348223 (2024).
17. Park, Y. et al. Short fimbriae of *Porphyromonas gingivalis* and their role in coadhesion with *Streptococcus gordonii*. *Infect. Immun.* **73**, 3983–3989 (2005).

18. Djemai, K., Drancourt, M. & Tidjani Alou, M. Bacteria and methanogens in the human microbiome: a review of syntrophic interactions. *Microb. Ecol.* **83**, 536–554 (2022).
19. Socransky, S. S. & Haffajee, A. D. Periodontal microbial ecology. *Periodontol 2000* **38**, 135–187 (2005).
20. Velsko, I. M. et al. Ancient dental calculus preserves signatures of biofilm succession and interindividual variation independent of dental pathology. *PNAS Nexus* **1**, gac148 (2022).
21. Fagernäs, Z. et al. Understanding the microbial biogeography of ancient human dentitions to guide study design and interpretation. *FEMS Microbes.* https://doi.org/10.1093/femsmc/xtac006 (2022).
22. Aas, J. A., Paster, B. J., Stokes, L. N., Olsen, I. & Dewhirst, F. E. Defining the normal bacterial flora of the oral cavity. *J. Clin. Microbiol.* **43**, 5721–5732 (2005).
23. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
24. Clemente, J. C. et al. The microbiome of uncontacted Amerindians. *Sci. Adv.* **1**, e1500183 (2015).
25. Lassalle, F. et al. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Mol. Ecol.* **27**, 182–195 (2018).
26. Brito, I. L. et al. Mobile genes in the human microbiome are structured from global to individual scales. *Nature* **535**, 435–439 (2016).
27. Klapper, M. et al. Natural products from reconstructed bacterial genomes of the Middle and Upper Paleolithic. *Science* **380**, 619–624 (2023).
28. Thamadilok, S. et al. Human and nonhuman primate lineage-specific footprints in the salivary proteome. *Mol. Biol. Evol.* **37**, 395–405 (2020).
29. Perry, G. H. et al. Diet and the evolution of human amylase gene copy number variation. *Nat. Genet.* **39**, 1256–1260 (2007).
30. Zee, K. Y., Samaranayake, L. P. & Attström, R. Predominant cultivable supragingival plaque in Chinese 'rapid' and 'slow' plaque formers. *J. Clin. Periodontol.* **23**, 1025–1031 (1996).
31. Theilade, E., Wright, W. H., Jensen, S. B. & Löe, H. Experimental gingivitis in man. II. A longitudinal clinical and bacteriological investigation. *J. Periodontal Res.* **1**, 1–13 (1966).
32. Listgarten, M. A., Mayo, H. E. & Tremblay, R. Development of dental plaque on epoxy resin crowns in man. A light and electron microscopic study. *J. Periodontol.* **46**, 10–26 (1975).
33. Simonsson, T. Aspects of dental plaque formation with special reference to colloid-chemical phenomena. *Swed. Dent. J. Suppl.* **58**, 1–67 (1989).
34. Simonsson, T., Edwardsson, S. & Glantz, P.-O. Biophysical and microbiologic studies of 'heavy' and 'light' plaque formers. *Eur. J. Oral. Sci.* **95**, 43–48 (1987).
35. Simonsson, T., Rönström, A., Rundegren, J. & Birkheb, D. Rate of plaque formation ? some clinical and biochemical characteristics of 'heavy' and 'light' plaque formers. *Eur. J. Oral. Sci.* **95**, 97–103 (1987).
36. Shrivastava, A. et al. Cargo transport shapes the spatial organization of a microbial community. *Proc. Natl. Acad. Sci. USA* **115**, 8633–8638 (2018).
37. Warinner, C. et al. A robust framework for microbial archaeology. *Annu. Rev. Genom. Hum. Genet.* **18**, 321–356 (2017).
38. Woo, P. C. Y. et al. *Streptococcus sinensis* sp. nov., a novel species isolated from a patient with infective endocarditis. *J. Clin. Microbiol.* **40**, 805–810 (2002).
39. Woo, P. C. Y. et al. The oral cavity as a natural reservoir for *Streptococcus sinensis*. *Clin. Microbiol. Infect.* **14**, 1075–1079 (2008).
40. Brennan, A. A. et al. Investigating the *Streptococcus sinensis* competence regulon through a combination of transcriptome analysis and phenotypic evaluation. *Microbiology* **168**, 001256 (2022).
41. Jensen, A., Scholz, C. F. P. & Kilian, M. Re-evaluation of the taxonomy of the Mitis group of the genus *Streptococcus* based on whole genome phylogenetic analyses, and proposed reclassification of *Streptococcus dentisani* as *Streptococcus oralis* subsp. *dentisani* comb. nov., *Streptococcus tigurinus* as *Streptococcus oralis* subsp. *tigurinus* comb. nov., and *Streptococcus oligofermentans* as a later synonym of *Streptococcus cristatus*. *Int. J. Syst. Evol. Microbiol.* **66**, 4803–4820 (2016).
42. Teng, J. L. L. et al. Phylogenomic and MALDI-TOF MS analysis of *Streptococcus sinensis* HKU4T reveals a distinct phylogenetic clade in the genus *Streptococcus*. *Genome Biol. Evol.* **6**, 2930–2943 (2014).
43. Fragiadakis, G. K. et al. Links between environment, diet, and the hunter-gatherer microbiome. *Gut Microbes* **10**, 216–227 (2019).
44. Sonnenburg, E. D. & Sonnenburg, J. L. The ancestral and industrialized gut microbiota and implications for human health. *Nat. Rev. Microbiol.* **17**, 383–390 (2019).
45. Scannapieco, F. A., Torres, G. & Levine, M. J. Salivary α-amylase: role in dental plaque and caries formation. *Crit. Rev. Oral. Biol. Med.* **4**, 301–307 (1993).
46. Deng, L. et al. Oral streptococci utilize a Siglec-like domain of serine-rich repeat adhesins to preferentially target platelet sialoglycans in human blood. *PLoS Pathog.* **10**, e1004540 (2014).
47. Poole, A. C. et al. Human salivary amylase gene copy number impacts oral and gut microbiomes. *Cell Host Microbe* **25**, 553–564.e7 (2019).
48. Gancz, A. S. et al. Ancient dental calculus reveals oral microbiome shifts associated with lifestyle and disease in Great Britain. *Nat. Microbiol* **8**, 2315–2325 (2023).
49. Eerkens, J. W. et al. A probable prehistoric case of meningococcal disease from San Francisco Bay: next generation sequencing of Neisseria meningitidis from dental calculus and osteological evidence. *Int. J. Paleopathol.* **22**, 173–180 (2018).
50. Eisenhofer, R., Kanzawa-Kiriyama, H., Shinoda, K.-I. & Weyrich, L. S. Investigating the demographic history of Japan using ancient oral microbiota. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190578 (2020).
51. Granehäll, L. et al. Metagenomic analysis of ancient dental calculus reveals unexplored diversity of oral archaeal *Methanobrevibacter*. *Microbiome* **9**, 197 (2021).
52. Jacobson, D. K. et al. Functional diversity of microbial ecologies estimated from ancient human coprolites and dental calculus. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **375**, 20190586 (2020).
53. Mann, A. E. et al. Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Sci. Rep.* **8**, 9822 (2018).
54. Neukamm, J. et al. 2000-year-old pathogen genomes reconstructed from metagenomic analysis of Egyptian mummified individuals. *BMC Biol.* **18**, 108 (2020).
55. Ottoni, C. et al. Tracking the transition to agriculture in Southern Europe through ancient DNA analysis of dental calculus. *Proc. Natl. Acad. Sci. USA.* **118**, e2102116118 (2021).
56. Weyrich, L. S. et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* **544**, 357–361 (2017).
57. Fellows Yates, J. A. et al. Community-curated and standardised metadata of published ancient metagenomic samples with AncientMetagenomeDir. *Sci. Data* **8**, 31 (2021).
58. Asangba, A. E. et al. Large comparative analyses of primate body site microbiomes indicate that the oral microbiome is unique among all body sites and conserved among nonhuman primates. *Microbiol. Spectr.* **10**, e0164321 (2022).
59. Brealey, J. C. et al. Dental calculus as a tool to study the evolution of the mammalian oral microbiome. *Mol. Biol. Evol.* **37**, 3003–3022 (2020).
60. Moraitou, M. et al. Ecology, not host phylogeny, shapes the oral microbiome in closely related species. *Mol. Biol. Evol.* **39**, msac263 (2022).
61. Ottoni, C. et al. Metagenomic analysis of dental calculus in ancient Egyptian baboons. *Sci. Rep.* **9**, 19637 (2019).
62. Ozga, A. T. et al. Oral microbiome diversity in chimpanzees from Gombe National Park. *Sci. Rep.* **9**, 17354 (2019).

63. Bivand, R., Nowosad, J. & Lovelace, R. spData: Datasets for Spatial Analysis. https://jakubnowosad.com/spData/ (2024).
64. Pebesma, E. Simple features for R: standardized support for spatial vector data. *R. J.* **10**, 439 (2018).
65. Pebesma, E. & Bivand, R. *Spatial Data Science: With Applications in R*. https://doi.org/10.1201/9780429459016 (2023).
66. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer, 2016).
67. Schmid, C. ggpointgrid: Rearrange scatter plot points on a regular grid. https://github.com/nevrome/ggpointgrid (2022).
68. Wilke, C. O. ggridges: Ridgeline Plots in 'ggplot2'. https://wilkelab.org/ggridges/ (2024).
69. Fellows Yates, J. A. et al. Reproducible, portable, and efficient ancient genome reconstruction with nf-core/eager. *PeerJ* **9**, e10947 (2021).
70. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
71. 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).
72. Parks, D. H. et al. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res.* **50**, D785–D794 (2022).
73. Parks, D. H. et al. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.* **38**, 1079–1086 (2020).
74. Youngblut, N. D. & Ley, R. E. Struo2: efficient metagenome profiling database construction for ever-expanding microbial genome datasets. *PeerJ* **9**, e12198 (2021).
75. Lu, J. et al. Metagenome analysis using the Kraken software suite. *Nat. Protoc.* **17**, 2815–2839 (2022).
76. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).
77. Ondov, B. D. et al. Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.* **17**, 132 (2016).
78. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
79. Kassambara, A. rstatix: pipe-friendly framework for basic statistical tests. R package version 0.4. 0. https://cran.r-project.org/web/packages/rstatix/index.html (2020).
80. Jensen, I. T., Janss, L., Radutoiu, S. & Waagepetersen, R. Compositionally aware estimation of cross-correlations for microbiome data. *PLoS One* **19**, e0305032 (2024).
81. Fuschi, A. et al. Correlation measures in metagenomic data: the blessing of dimensionality. *bioRxiv* 2024.02.29.582875 https://doi.org/10.1101/2024.02.29.582875 (2024).
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**, 2224 (2017).
83. Quinn, T., Richardson, M. F., Lovell, D. & Crowley, T. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* **7**, 16252 (2017).
84. Rohart, F., Gautier, B., Singh, A. & Lê Cao, K.-A. mixOmics: an R package for 'omics feature selection and multiple data integration. *PLoS Comput. Biol.* **13**, e1005752 (2017).
85. Briscoe, L., Balliu, B., Sankararaman, S., Halperin, E. & Garud, N. R. Evaluating supervised and unsupervised background noise correction in human gut microbiome data. *PLoS Comput. Biol.* **18**, e1009838 (2022).
86. Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC Bioinform.* **17**, 483 (2016).
87. Hoffman, G. E. & Roussos, P. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics* **37**, 192–201 (2021).
88. Wei, T. et al. corrplot: Visualization of a correlation matrix. *R package version 0. 73 230* (2013).
89. Coelho, L. P. et al. Towards the biogeography of prokaryotic genes. *Nature* **601**, 252–256 (2022).
90. Kojima, C. Y., Getz, E. W. & Thrash, J. C. RRAP: RPKM recruitment analysis pipeline. *Microbiol. Resour. Announc.* **11**, e0064422 (2022).
91. Pedersen, T. L. patchwork: The Composer of ggplots. *R package version 0. 0* 1 (2017).
92. Garnier, S., Ross, N., Rudis, B. & Sciaini, M. Package 'viridis'. *Colorblind-Friendly Color Maps for R* (2018).

Acknowledgements

We thank Alexander Hübner for discussion on gene content analyses. This work was supported by the Werner Siemens Stiftung ("Paleobiotechnology" to C.W.), the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy (EXC 2051 Project-ID 390713860), the American School for Prehistoric Research (ASPR), the Max Planck Society, and Harvard University.

Author contributions

C.W. and I.M.V. conceived the project. I.M.V. performed the analyses and wrote the manuscript with input from C.W.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41522-024-00642-1.

Correspondence and requests for materials should be addressed to Irina M. Velsko or Christina Warinner.