Analysis of the human oral microbiome from modern and historical samples with SPARSE and EToKi

Mark Achtman and Zhemin Zhou

Warwick Medical School, University of Warwick, Coventry, UK

Orcid ID: MA, 0000-0001-6815-0070; ZZ, 0000-0001-9783-0366

Subject Areas:

methodology, metagenomics, population genomics, microbiology

Author for correspondence:

Mark Achtman

e-mail: m.achtman@warwick.ac.uk

Abstract

Elsewhere we have provided details on a program for the probabilistic classification of sequence reads from metagenomes to microbial species (SPARSE) [1]. Similarly, we have also provided details on a stand-alone set of pipelines (EToKi) that are used to perform backend calculations for EnteroBase [2]. We have also provided examples of analyses of genomes reconstructed from metagenomes from ancient skeletons together with genomes from their modern relatives [3], which can also be visualized within EnteroBase. Finally, we have also described GrapeTree [4], a graphic visualizer of genetic distances between large numbers of genomes. Here we combine all of these approaches, and examine the microbial diversity within the human oral microbiome from hundreds of metagenomes of saliva, plaque and dental calculus, from modern as well as from historical samples. 1591 microbial species were detected by these methods, some of which differed dramatically in their frequency by source of the metagenomes. We had anticipated that the oral complexes of Socransky *et al*. [5] would predominate among such taxa. However, although some of those species did discriminate between different sources, we were unable to confirm the very existence of the oral complexes. As a further example of the functionality of these pipelines, we reconstructed multiple genomes in high coverage of *Streptococcus mutans* and *Streptococcus sobrinus*, two species that are associated with dental caries. Both were very rare in historical dental calculus but they were quite common in modern plaque, and even more common in saliva. The reconstructed genomes were compared with modern genomes from RefSeq, providing a detailed overview of the core genomic diversity of these two species.

1.Introduction

Multiple research areas have undergone revolutionary changes in the last 10 years due to broad accessibility to high throughput DNA sequencing at reduced costs. These include the evolutionary biology of microbial pathogens based on metagenomic sequencing. Studies on *Mycobacterium tuberculosis* [6,7], *Mycobacterium leprae* [8,9], *Yersinia pestis* [10-14] and *Salmonella enterica* [3,15,16] have yielded important insights into the history of infectious diseases by combining modern and historical genomes. In most cases, interpretations of the newly deciphered historical genomes benefitted greatly because they could be slotted into an existing framework for the modern population genomic structure of the causative bacteria [17-19]. However, interesting questions about other microbes are more difficult to address where a framework based on extensive analyses of modern organisms is still lacking.

Exploring the genetic diversity of microbes directly from metagenomic sequences is usually performed by classifying the sequence reads into taxonomic units. Taxonomic assignments can be performed by the *de novo* assembly of the metagenomic reads into MAGs (metagenomic assembled genomes), or by assigning individual sequence reads to existing reference genomes. However most current metagenomic classifiers rely on the public genomes in NCBI, which are subject to an extreme sample bias and have a preponderance of genomes from pathogenic bacteria. Furthermore, shotgun metagenomes often include DNA from environmental sources, which include multiple micro-organisms that have never been cultivated, and may belong to unknown or poorly classified microbial taxa whose abundance is not reflected by existing databases. Recent evaluations have also demonstrated that current taxonomic classifiers either lack sufficient sensitivity for species-level assignments, or suffer from false positives, and that they overestimate the number of species in the metagenome [20-22]. Both tendencies are especially problematic for the identification of microbial species which are only present at low-abundance, e. g. detecting pathogens in ancient metagenomic samples.

We designed SPARSE to provide accurate taxonomic assignments of metagenomic reads and published a technical description in 2018 [1]. SPARSE accounts for the existing bias in reference databases by creating a subset that represents the entire genetic diversity according to 99% average nucleotide identity (ANI99%) but assigns taxonomic designations based on its ANI95% superset, which is roughly equivalent to individual bacterial species [23,24]. To this end, it groups genomes of Bacteria, Archaea, Viruses and Protozoa from RefSeq into sequence similarity-based hierarchical clusters, and the resulting dataset includes only one reference genome per ANI99% cluster for fine-grained taxonomic assignments. SPARSE assigns metagenomic sequence reads to these clusters by using Minimap2 [25]. Unreliable alignments are excluded if the successful alignments were widely dispersed because wide dispersion reflects either ultra-conserved elements of uncertain specificity or a high probability of homoplasies due to horizontal gene transfer. The remaining metagenomic reads are then assigned to unique clusters on the basis of a probabilistic model, and labelled according to the taxonomic labels and pathogenic potential of the genomes within those clusters. The

probabilistic model specifically penalizes non-specific mappings of reads from unknown sources, and hence reduces false-positive assignments. Our methodological comparisons demonstrated that SPARSE has greater precision and sensitivity with simulated metagenomic data than 10 other taxonomic classifiers, and, unlike five other methods, consistently yielded correct identifications of pathogen reads within metagenomes of ancient DNA [1].

After SPARSE has assigned reads to taxa, it can also be used to extract reads corresponding to an ANI95% taxon of interest from the metagenomic data, which corresponds to a species [24]. The next step for genomic analyses is provided by EToKi [2], a stand-alone package of pipelines that we developed for EnteroBase to support manipulations with 100,000s of modern microbial genomes. The filter function in EToKi merges overlapping pair-ended reads within each data set, removes low quality bases and trims adapter sequences. It then conducts a fine-grained evaluation of the reads for greater sequence similarities to an in-group of genomes from the taxon of interest than to other genomes from a related but distinct out-group. EToKi then masks all nucleotides in an appropriate reference genome, and creates a pseudo-MAG equivalent by unmasking nucleotides with sufficient coverage among the reads that have passed the in-group/out-group comparisons. Finally, EToKi creates a SNP matrix from pseudo-MAGs plus additional draft genomes, and generates a Maximum-Likelihood phylogeny (RAxML 8.2 [26]). The ML tree can be imported together with its metadata for visual interrogation by graphic interfaces in EnteroBase such as GrapeTree [4] or Dendrogram, as recently illustrated for ancient metagenomes of *Y. pestis* [2].The availability of these tools is not restricted to ancient DNA genomes, and they can be readily used metagenomic analyses of modern samples.

Here we illustrate how these tools can be used to investigate the taxonomic composition of metagenomes from the modern and ancient human oral flora, and also examine in greater detail the population genomic structures of *Streptococcus mutans* and *Streptococcus sobrinus*, which are associated with dental caries in some human populations [27-29].

2. Results

(a) SPARSE analysis of oral metagenomes

In its original incarnation in August 2017 [1], SPARSE used MASH [30] to assign 101,680 genomes from the NCBI RefSeq database to 28,732 ANI99% clusters of genomes [1]. By May 2018, 21,540 additional genomes had been added to NCBI RefSeq. These were merged into the existing database in the same manner as previously, either by creating a new cluster containing one genome if the ANI to all existing clusters was less than 99%, or else by merging that genome into an existing cluster. The resulting database contained 33,212 ANI99% clusters of microbial genomes from Eukaryotes, Prokaryotes and Viruses. One representative genome was chosen for each of the 32,378 ANI99% clusters containing Bacteria, Archaea or Viruses. The ANI99% representative database was supplemented by a human reference genome (Genome Reference Consortium Human Build 38) such that reads from human DNA could also be called. All the representative genomes were assigned to a superset of 20,054 ANI95% clusters, and this was used for species assignments and genomic extractions.

We identified 17 public archives containing 1,016 sets of metagenomic sequences (Table 1) from 791 oral samples which had been obtained from modern human saliva, modern human dental plaque or historical dental calculus from a variety of global sources (Table S1). Individual sequence reads from those metagenomes were assigned to taxa with SPARSE, and then cleaned with EToKi filter. Seven metagenomes lacked bacterial reads from the oral microbiome (ancient dental calculus: 5; modern saliva: 2). These seven metagenomes were ignored for further analyses, leaving metagenomes from a total of 784 samples (Table 2). The Table S2 reports the percentage assignment of the reads in each sample to each of 1,592 taxa, except that assignments at a frequency of <0.0005% of the sequence reads are reported as 0%. That spreadsheet includes a column identifying potential pathogens, including assignments to the oral microbial complexes defined by Socransky *et al*. [5]. SPARSE also identified 158 samples containing Archaea from six species and 314 samples containing human viruses or bacteriophages (Table 3).

(b) Comparisons of microbiomes from saliva, plaque and historical dental calculus

We tested whether the quantitative levels of microbial taxa differed by sample source with multiple approaches. UMAP (Uniform Manifold Approximation and Projection), a recently described [31] high performance algorithm for dimensional reduction of diversity within large amounts of data by non-linear multidimensional clustering, was applied to the abundances of each taxon in each sample. Visual examination of a UMAP plot (Fig. 1A) shows three discrete clusters. The discrete nature of these clusters was confirmed by an independent application of machine learning *via* optimal k-mean clustering based on the first three components from the UMAP analysis (Supplemental Fig. S1A). Most members of each cluster were from a common source, i.e. one cluster from modern saliva, one from modern dental calculus, and one from ancient dental calculus (Fig. 1A). However, the correlation was imperfect, and each cluster also contained some metagenomes from alternative sources. Similar results were obtained with a

classical principal component analysis (PCA), except that the clusters were not as clearly distinguished and each cluster contained a higher proportion of exceptions (Supplemental Fig. S1B). The assignments of source affiliations to cluster were largely consistent between UMAP and PCA, with occasional exceptions (Supplemental Fig. S1C).

We also compared the correlation between clusters and source by hierarchical clustering. To this end, we calculated the Euclidean p- distances between each pair of samples, and subjected them to hierarchical clustering by the neighbor-joining algorithm with the results shown in Fig. 1B. This approach also largely separated the samples by source, with only few exceptions. Samples from modern saliva formed one large cluster. Samples from modern dental plaque formed two related but discrete sub-clusters, one of which included a sub-sub cluster of samples from historical dental calculus. These clusters also largely corresponded to the clusters found by k-mean clustering of UMAP data (Supplemental Fig. S1A).

Thus, three independent methods each found three primary and distinct clusters of the quantitative numbers of reads in microbial taxa, and these clusters largely corresponded to modern saliva, modern plaque and historical dental calculus. This finding indicates that there are source-specific taxa in the metabiomes from these sources.

(c) Source-specific taxa

We used the SVM machine learning approach to identify the most important bacterial taxa for the observed clustering of microbial taxon composition by sample source. The results for the 40 most discriminating ANI95% taxa with the most optimal of 300 variants of the SVM model are presented in descending order of their SVM weights in Fig. 2.

Fig. 2 also includes mini-histograms for each of the 40 taxa that summarize their relative abundance of sequence reads by source. Multiple taxa were dramatically more prominent in samples from one source than from either of the two other sources. However, the most prominent sample source varied with the taxon. Eleven of the 40 discriminatory taxa belonged to the oral complexes that are associated with periodontitis according to Socransky *et al*. [5]. In general agreement with this observation, seven species from oral complexes (*Veillonella parvula*, *Fusobacterium nucleatum, Capnocytophaga gingivalis, Streptococcus gordonii, Actinomyces naeslundii, Actinomyces viscosus,* and *Capnocytophaga sputigena*) were most abundant in modern plaque and two other species (*Streptococcus sanguinis, Tannerella forsythia*) were most abundant in historical dental calculus. The yellow complex includes *Streptococcus mitis,* which SPARSE subdivided into two distinct ANI95% clusters designated *S. mitis* A and *S. mitis* B in accordance with a recent publication [32]. Each of the two taxa was most frequent in saliva than in dental plaque or dental calculus.

We were somewhat surprised that 17 other taxa that were assigned to an oral complex by Socransky *et al*. [5] were not included in the 40 most discriminatory taxa. We therefore examined the relative abundances of all 28 taxa from oral complexes in greater detail (Fig. 3). Three of the four taxa in the Blue and Purple Complexes are very abundant in oral

metagenomes, and all four are preferentially found in modern plaque. However, the other oral complexes are not uniform in their patterns of relative abundances. For example, within the Red complex, both *T. forsythia* and *Treponema denticola* were most frequently found in historical dental calculus but *Porphyromonas gingivalis* is is most frequent in modern plaque, and is generally much less abundant. Similar intra-complex discrepancies were found for the Orange, Yellow, and Green Complexes.

The inconsistent frequencies of members of an oral complex raises questions about whether the compositions of those complexes are consistent in individual samples. Careful reading of Socransky *et al.* [5] revealed that the oral complexes were considered as a hypothesis, whereas they have now attained the status of conventional wisdom, and even play a prominent role in routine laboratory investigations of periodontitis. The data in that publication were based on 28 cultivated bacterial species, whose presence or absence was determined by DNA hybridization against a small number of probes. This technology is now outdated; the number of oral taxa has increased dramatically; and the data presented here are for relative abundance rather than presence or absence. We have therefore examined the strengths of association into the oral complexes from the current data presented here according to similar criteria and similar methods as those used in Socransky *et al.* publication.

The original composition of the oral complexes depended strongly on results from hierarchical clustering based on a concordance between pairs of species for presence or absence in individual samples. The tree in Fig. 4 shows neighbour-joining clustering of the common microbial taxa detected by SPARSE, some of which have been cultivated whereas others have not. The taxa were clustered by the similarities of their abundances over all samples. This tree provides no support for the original composition of the oral complexes because the four areas of the tree where oral complex taxa are clustered each contain representatives from multiple complexes, and none of those clusters corresponds to the original compositions proposed by Socransky *et al*. [5].

It seemed possible that the discrepancies between Fig. 4 and the original compositions of the oral complexes might reflect the fact that this study identified many additional taxa. We therefore performed cluster analyses with our current data for the original set of 31 cultivatable bacterial species examined by Socransky *et al.* We compared the Neighbor-joining algorithm used here with the less powerful, agglomerative clustering method (UPGMA, Unweighted Pair Group Method with Arithmetic Mean) used by Socransky *et al*. We also compared the abundances across all samples with abundances in plaque, which was the primary source for bacteria tested by Socransky *et al.* The results (Fig. S3) show dramatic inconsistencies between independent trees in regard to the clustering of the oral complex bacteria. For example, *T. forsythia, T. dentocola* and *P. gingivalis* of the Red Complex cluster together in Fig. S3A,C,F,G. However, *T. denticola* and *T. forsythia* are separated from *P. gingivalis* in the four other parts of Fig. S3. And do not even cluster together with each other Fig. S3E. Similar or even greater discrepancies are visible for the other oral complexes in Fig. S3.

These inconsistencies in clustering patterns across minor differences in sampling and clustering algorithms raise severe doubts about the very existence of the oral complexes as defined by Socransky *et al.* [5].

(d) Numbers of taxa per source

The rarefaction curves in Fig. 5A provide a breakdown of taxa by sample source as additional samples are tested. SPARSE detected 1591 microbial taxa over all 784 metagenomic samples: 1,389 from modern saliva; 842 from modern plaque and 696 from historical calculus. These estimates will increase as additional samples are added, but at increasingly slower rates because the rarefaction curves seem to be reaching a plateau, except possibly for historical dental calculus where the fewest samples have been evaluated until now.

The median numbers of taxa per sample were much more moderate than the total numbers, ranging from 177 (historical dental calculus) to 288 (modern saliva). These median values reflect a bimodal distribution for numbers of taxa per sample (Fig. 5B), wherein a few samples had jackpots of large numbers of taxa but all other samples had only few.

The analyses described above focused on differences in taxon composition by source. However, the Venn diagram in Fig. 5C shows that 447 taxa were common to all three sources, even if their relative abundance varied. Modern plaque yielded only 34 taxa which were not found in either historical dental calculus or modern saliva. More source-specific taxa were found in historical dental calculus, which may possibly reflect some contamination with environmental material. Alternatively, some taxa may be absent in modern dental plaque because historical lineages have become extinct [3]. Saliva yielded 504 unique taxa, some of which might be transient, and do not persist long enough to be incorporated into plaque.

(e) Population genomics of organisms associated with dental caries

The causes of dental caries remain controversial [28,33-35], other than that the disease is usually preceded by dental plaque, and that modern dental plaque contains high concentrations of *Streptococcus mutans* and *Streptococcus sobrinus*. Our data confirms that reads belonging to these two taxa are abundant in modern dental plaque, and even more abundant in modern saliva (Fig. 6A,C). Our data also shows that the abundance of both *S. mutans* and *S. sobrinus* was extremely low in historical dental calculus, supporting prior conclusions that *S. mutans* first became common in the last 200 years after industrialization introduced high levels of sugar to human diets [36]. *S. sobrinus* was undetectable in the historical samples (<10 reads per metagenome) but up to 0.01% of all reads in some historical samples were assigned to *S. mutans*. These reads showed an increase of deamination in their 5'-ends (Fig. S4), confirming that they were truly from ancient DNA, and that *S. mutans* has been part of the human oral microbiome for millennia [36].

We exploited the high frequency of sequence reads from these two *Streptococcus* species in modern dental plaque and saliva to illustrate how SPARSE and EToKi can be used to extract MAGs from metagenomic sequence reads, and combine them with genomes sequenced from

cultivated bacteria. To this end, we extracted all sequence reads specific to the ANI95% clusters for *S. mutans* and *S. sobrinus* from metagenomes in which there was high coverage (Fig. 6E,F), and cleaned them with the EToKi prepare module. The reads were filtered with EToKi assemble by specifying a reference genome as well as an ingroup and an outgroup of additional genomes. Sequence reads were excluded which had higher alignment scores to the outgroup than to the ingroup genomes. EToKi creates a pseudo-genome from each reference genome (*S. mutans*: UA159; *S. sobrinus*: NCTC12279) in which all nucleotides are masked in order to ensure that only nucleotides supported by metagenomic reads will be used for phylogenetic analysis. Sites in the pseudo-genome were unmasked which were covered by ≥3 reads. These procedures resulted in a total of 31 MAGs for *S. mutans* and 15 MAGs for *S. sobrinus* in which over 70% of the reference genome had been unmasked, most of which were from Chinese samples [37]. The MAGs were combined with genomes from cultivated bacteria of the same species from Brazil, the U.S. and the UK as well as other countries (Table 2) and EToKi align was used to extract non-repetitive SNP matrices, which were subjected to calculation of Maximum Likelihood (ML) phylogenies (Fig. 7) using EToKi phylo.

The ML phylogenies of the two species showed interesting differences. Geographic clustering was apparent within the ML tree of *S. sobrinus*. All Chinese MAGs clustered together, separate from the bacterial genomes from Brazil. In contrast, *S. mutans* MAGs from Chinese isolates did not show obvious phylogeographic specificities, and were inter-dispersed among bacterial genomes from multiple geographic locations. Similar conclusions about a lack of phylogeographic specificity have been reached with a subset of 57 genomes in previous studies [38].

3. Discussion

Several years ago, we accidentally became interested in comparing genomes assembled from metagenomes from historical sources with draft genomes from cultivated bacteria. Our initial efforts involved the deployment of individual bioinformatic tools, comparisons of multiple publicly available algorithms, and compilation of draft genomes from publicly available sequence read archives of short read sequences [7]. In parallel we were also involved in developing EnteroBase, a compendium of 100,000s of draft genomes assembled from multiple genera that can cause enteric disease in humans. Including *Salmonella* [2,19]. These two directions overlapped when we had the opportunity to examine the evolutionary history of *Salmonella enterica* based on metagenomics sequences from 800 year old bones, teeth and dental calculus [3]. In that case, reads from *S. enterica* were found in teeth and bone, but nor in dental calculus. However, we were motivated to examine further samples of dental calculus, and quickly realized that life was too short for manual analyses. Optimised pipelines were needed, but none of the existing tools proved to be adequately reliable and sufficiently sensitive for assigning sequence reads from historical metagenomes to the tree of microbial life. We therefore took a step back, and developed SPARSE to satisfy our requirements. In parallel, we developed EToKi to be able to efficiently handle manipulations with sequence reads

and genomic assemblies for the back end operations in EnteroBase. Finally, we developed GrapeTree [4], which supports the graphic visualization and manipulation of phylogenetic trees representing large numbers of genomes. Here we have demonstrated how to combine these tools to obtain an overview of the microbial flora in samples from human oral saliva, modern dental plaque and historical dental calculus. We also demonstrate how these tools can be used to reconstruct genomes of taxa present at moderate concentrations within metagenomics sequences, and compare them with conventional draft genomes. The experimental procedures for processing 791 metagenomes consisted of running SPARSE in the background for 2 months (~100,000 CPU hours). The pipelines described here permitted all other procedures and evaluations described here to be completed in less than two weeks.

All data produced here are freely accessible, including genomes, taxon abundances per sample, and examples of the commands used https://github.com/zheminzhou/OralMicrobiome. The programs and source code are also publicly available.

As novices to the area of oral biology, we were prepared to believe in the existence of oral complexes of bacteria [5]. However, we were not able to reliably reconstruct the original patterns which led to their definition (Fig. 4, Fig. S3), and suggest that their existence may have depended on currently outdated technology and a focus on limited samples of microbial taxa. Given our inability to corroborate this concept, we suggest that the existence and composition of the oral complexes should be re-examined by others, with other datasets and other methods.

We were also curious about organisms associated with dental caries and late stages of dental plaque formation, *S. mutans* and *S. sobrinus,* especially because there were sufficient reads to assemble partial genomes (MAGs) from multiple samples as well as multiple draft genomes from cultivated bacteria. We were also intrigued by the claim that *S. mutans* was rare in historical plaque [36]. Our data support that claim, and we found only very low frequencies of reads of either organism in the historical calculus samples that we examined. Our data also support prior conclusions of a lack of phylogeographic differentiation within *S. mutans* [38]. However, although the data are still somewhat limited, there seems to be a clear distinction between *S. sobrinus* from China and Brazil, which might reflect phylogeographical signals. Unfortunately, the two sets of genomes are also different by source because the Chinese genomes were MAGs reconstructed from metagenomes and the Brazil genomes were from cultivated bacteria. Other scientists with an interest in this organism might want to obtain additional genomes of *S. sobrinus* from other geographical areas to determine whether the phylogeographical trends stand up to further investigation. Such efforts could also be facilitated by creating an EnteroBase for *Streptococcus*, which would be relatively easy to do if there were any interested curators and sufficient interest in the *Streptococcus* community.

In summary, we illustrate the use of a variety of reliable, high throughput tools for determining the microbial diversity and extracting microbial genomes from metagenomic data. We illustrate

these tools with metagenomes from modern and historical samples, and release all the data and methods for further use by others.

4. Methods

(a) Dimension reduction of frequencies of reads.

The SPARSE results were submitted to two forms of dimensional reduction of diversity. UMAP analysis was performed with its Python implementation [31], using the parameters min_neighbors=5 and min_dist=0.0. PCA was performed using the decomposition.PCA module of the scikit-learn Python library [39]. Optimal k-mean clusters of the first three components from the UMAP analysis were calculated with the sklearn.cluster module of the scikit-learn Python library.

(b) Ranking of microbial species by their SVM weights

A supervised Support Vector Machine [40] classification of samples was performed using the SVM module of the scikit-learn Python library on the raw SPARSE results (Dataset S3). The SVM classification was performed 300 times on a randomly chosen training set consisting of 60% of all samples with varying penalty hyper-parameter C, and scored using 5-fold cross-validation. The model was then tested with the optimal hyper-parameter on the remaining 40% of samples, and correctly inferred the oral source for >96% of the test samples. The optimal SVM coefficients for each individual species were estimated by training that model once again on all the oral samples. The order of the species in Fig. 2 consists of the SVM weights (squares of the coefficients; [41]) in descending order.

(c) Genome reconstructions for *Streptococcus mutans* and *Streptococcus sobrinus*

We set a minimum criterion for reconstruction of MAGs at least two million nucleotides covered by sequence reads in a metagenome. SPARSE identified 66 and 28 samples which met these criteria for *S. mutans* (ANI95% cluster s5) or *S. sobrinus* (s3465), respectively (Figs. 4B,D). The species specific reads from these samples were extracted from the metagenomes and subjected to reference-guided assemblies using 'EToKi assemble' as described elsewhere [2]. These assemblies were performed against reference genomes UA159 (*S. mutans*, GCF_000007465) and NCTC12279 (*S. sobrinus*, GCF_900475395). All other genomes deposited in RefSeq from the same species (*S. mutans*: 194, *S. sobrinus*: 49) were used as an ingroup and 262 genomes from other species in the *Streptococcus* Mutans group were used as an outgroup (Table S3). Individuals SNPs were unmasked which were supported by at least three reads, and whose consensus base was supported by ≥70% of the mapped reads. The successfully reconstructed genomes for *S. mutans* and *S. sobrinus* can be downloaded at https://github.com/zheminzhou/OralMicrobiome. An alignment (EToKi align) of the 31 *S. mutans* MAGs plus all 195 *S. mutans* genomes plus the only *S. troglodytae* genome in RefSeq (Table S3) included 1.73 MB that were shared by ≥ 95% of the genomes, and 181,321 core SNPs. Similarly, a *S. sobrinus* alignment of 15 MAGs as well as 50 draft or complete genomes of *S. sobrinus* plus 6 genomes of *Streptococcus downei* from RefSeq spanned 1.16 MB and contained 160,863 core SNPs. These alignments were subjected to Maximum Likelihood phylogeny reconstruction using EToKi phylo. Both ML trees were then visualised with GrapeTree [4].

Table 1. Sources of metagenomic reads.

Archive	Accession	Sets of short reads	Number of. samples	Source	Institute	Citation
1	PRJNA445215	62	48	calculus	Max Planck Institute for the Science of Human History	[42]
2	PRJEB30331, PRJNA454196	45	44	calculus	University of Oxford	[43]
3	PRJNA216965	9	2	calculus	University of Oklahoma	[44]
4	PRJNA383868	87	87	plaque	J. Craig Venter Institute	[45]
5	PRJNA255922	48	48	plaque	University of California, Los Angeles	[46]
6	PRJNA78025	7	4	plaque	University of Maryland	[47]
7	PRJNA289925	1	1	plaque	University of Washington	[48]
8	PRJEB6997	298	298	plaque & saliva	BGI	[37]
9	PRJNA230363	12	12	plaque & saliva	Chinese Academy of Sciences	[49]
10	PRJEB24090	61	61	saliva	University of California San Diego	[50]
11	PRJNA380727	56	55	saliva	Peking University School of Stomatology	
12	PRJNA396840	30	30	saliva	University of Copenhagen	[51]
13	PRJEB14383	28	28	saliva	University College London	[52]
14	PRJDB4115	26	26	saliva	University of Tokyo	[53]
15	PRJNA217052	217	18	saliva	Broad Institute	[54]
16	PRJNA188481	8	8	saliva	Broad Institute	[55]
17	http://dx.doi.org/10.4225/55/584775546a409	21	21	calculus	OAGR, University of Adelaide	[56]

Note: Calculus refers to ancient dental calculus from historical samples. Plaque and saliva refer to modern dental plaque and saliva.

All Sets were downloaded from GenBank except for Archive 17, which was downloaded from the Online Ancient Genome Repository.

Table 2. Sources and properties of single-colony isolated bacterial strains and metagenomic samples from which genomes were used for analyses.

Category	Sub-category	Number
Bacterial genomes		262
	S. mutans	195
	S. sobrinus	50
	others	17
Metagenome source		791
	Ancient dental calculus	114
	Modern plaque	288
	Modern saliva	389
Sample size	(nucleotides)	
	0-2GB	348
	2-4GB	130
	4-6GB	162
	6-8GB	93
	8-10GB	45
	>10GB	13
Countries		
	Asia	423
	China	365
	Japan	26
	Philippines	28
	Others	4
	North America	122
	U.S.A.	120
	Guadeloupe	2
	Europe	127
	Denmark	30
	Ireland	36
	Germany	44
	Others	17
	Oceania	110
	Australia	92
	Fiji	18
	Africa	**7**
	South Africa	4
	Sudan	2
	Sierra Leone	1

Table 3. Detailed summary of Archaea and Viruses in all 786 samples.

Taxonomy	No. ancient samples (109)	% reads (relative)	No. plaque (288)	% reads (relative)	No. saliva (387)	% reads (relative)
Host (Human)	**109**	**0.31**	**244**	**2E-4 (100)**	**335**	**7.05**
Archaea (6)#	**81**	**1.78 (100)**	**28**	2E-4 (100)	**49**	**1E-4 (100)**
Methanobrevibacter oralis	79	1.78 (100)	28		47	1E-4 (98.0)
Methanobrevibacter smithii	1	3E-5 (2E-3)			1	8E-7 (0.77)
Candidatus Nitrosoarchaeum koreensis	1	1E-5 (7E-4)			0	
**Thermoplasmatales archaeon BRNA1*	1	7E-6 (4E-4)			0	
Methanobrevibacter smithii	1	1E-6 (8E-5)		**4E-4 (100)**	1	1E-6 (1.32)
Virus (18)#	**4**	**1E-5 (100)**	**54**	9E-6 (2.49)	**256**	**4E-3 (100)**
Human betaherpesvirus 7			9	3E-4 (88.1)	150	6E-4 (16.3)
Human gammaherpesvirus 4			19	5E-6 (1.32)	86	3E-3 (75.3)
Human alphaherpesvirus 1			1		9	8E-5 (2.13)
Human betaherpesvirus 6B				3E-5 (8.06)	7	2E-5 (0.62)
Bacterophages (12)	4	1E-5 (100)	30		128	2E-4 (5.62)

Data accessibility. Python scripts and source material for mathematical modelling and dimension reduction of SPARSE results, as well as all 46 MAGs are deposited at https://github.com/zheminzhou/OralMicrobiome as Datasets S1 to S3.

Authors' contributions. Z.Z. analysed data and prepared the figures. M.A. and Z.Z. interpreted the results and wrote the manuscript.

Competing interests. We have no competing interests.

Funding. This project was supported by the Wellcome Trust (202792/Z/16/Z) and EnteroBase development was funded by the BBSRC (BB/L020319/1).

References

1. Zhou, Z., Luhmann, N., Alikhan, N.-F., Quince, C., and Achtman, M. 2018 Accurate reconstruction of microbial strains from metagenomic sequencing using representative reference genomes. In *RECOMB 2018*, pp. 225-240: Springer, Cham.

2. Zhou, Z., Alikhan, N.-F., Mohamed, K., The Agama Study Group, and Achtman, M. 2019 The user's guide to comparative genomics with EnteroBase. Three case studies: micro-clades within *Salmonella enterica,* serovar Agama, ancient and modern populations of *Yersinia pestis*, and core genomic diversity of all *Escherichia*. *BioRxiv*, 613554. (doi:doi:10.1101/613554v1)

3. Zhou, Z. *et al.*. 2018 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. (doi:https://doi.org/10.1016/j.cub.2018.05.058)

4. Zhou, Z., Alikhan, N.-F., Sergeant, M. J., Luhmann, N., Vaz, C., Francisco, A. P., Carrico, J. A., and Achtman, M. 2018 GrapeTree: Visualization of core genomic relationships among 100,000 bacterial pathogens. *Genome Res* **28**, 1395-1404. (doi:DOI: 10.1101/gr.232397.117)

5. Socransky, S. S., Haffajee, A. D., Cugini, M. A., Smith, C., and Kent, R. L., Jr. 1998 Microbial complexes in subgingival plaque. *J Clin Periodontol* **25**, 134-144.

6. Bos, K. I. *et al.*. 2014 Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature* **514**, 494-497. (doi:nature13591 [pii];10.1038/nature13591 [doi])

7. Kay, G. L. *et al.*. 2015 Eighteenth-century genomes show that mixed infections were common at time of peak tuberculosis in Europe. *Nat Commun* **6**, 6717. (doi:ncomms7717 [pii];10.1038/ncomms7717 [doi])

8. Schilling, A. K. *et al.*. 2019 British red squirrels remain the only known wild rodent host for leprosy bacilli. *Front Vet Sci* **6**, 8. (doi:10.3389/fvets.2019.00008 [doi])

9. Schuenemann, V. J. *et al.*. 2018 Ancient genomes reveal a high diversity of *Mycobacterium leprae* in medieval Europe. *PLoS Pathog* **14**, e1006997. (doi:10.1371/journal.ppat.1006997 [doi];PPATHOGENS-D-17-02430 [pii])

10. Bos, K. I. *et al.*. 2011 A draft genome of *Yersinia pestis* from victims of the Black Death. *Nature* **478**, 506-510.

11. Rasmussen, S. *et al.*. 2015 Early divergent strains of *Yersinia pestis* in Eurasia 5,000 years ago. *Cell* **163**, 571-582. (doi:S0092-8674(15)01322-7 [pii];10.1016/j.cell.2015.10.009 [doi])

12. Damgaard, P. B. *et al.*. 2018 137 ancient human genomes from across the Eurasian steppes. *Nature* **557**, 369-374. (doi:10.1038/s41586-018-0094-2 [doi];10.1038/s41586-018-0094-2 [pii])

13. Keller, M. *et al.*. 2019 Ancient *Yersinia pestis* genomes from across Western Europe reveal early diversification during the First Pandemic (541-750). *Proc Natl Acad Sci U S A* **116**, 12363-12372. (doi:1820447116 [pii];10.1073/pnas.1820447116 [doi])

14. Spyrou, M. A. *et al.*. 2019 Phylogeography of the second plague pandemic revealed through analysis of historical *Yersinia pestis* genomes. *Nat Commun* **10**, 4470. (doi:10.1038/s41467-019-12154-0 [doi];10.1038/s41467-019-12154-0 [pii])

15. Vågene, Å. J. *et al.*. 2018 *Salmonella enterica* genomes from victims of a major sixteenth-century epidemic in Mexico. *Nature Ecology & Evolution* **2**, 520-528.

16. Key, F. M. *et al.* 2020 Evolution towards human-specific *Salmonella enterica* accompanied the Western Eurasian Neolithization process. In (Anon.).

17. Coll, F., McNerney, R., Guerra-Assuncao, J. A., Glynn, J. R., Perdigao, J., Viveiros, M., Portugal, I., Pain, A., Martin, N., and Clark, T. G. 2014 A robust SNP barcode for typing *Mycobacterium tuberculosis* complex strains. *Nat Commun* **5**, 4812. (doi:ncomms5812 [pii];10.1038/ncomms5812 [doi])

18. Achtman, M. 2016 How old are bacterial pathogens? *Proc Biol Sci* **283**, 1836.

19. Alikhan, N.-F., Zhou, Z., Sergeant, M. J., and Achtman, M. 2018 A genomic overview of the population structure of *Salmonella*. *PLoS Genet* **14**, e1007261. (doi:10.1371/journal.pgen.1007261 [doi];PGENETICS-D-18-00122 [pii])

20. McIntyre, A. B. R. *et al.*. 2017 Comprehensive benchmarking and ensemble approaches for metagenomic classifiers. *Genome Biol* **18**, 182. (doi:10.1186/s13059-017-1299-7 [doi];10.1186/s13059-017-1299-7 [pii])

21. Sczyrba, A. *et al.*. 2017 Critical assessment of metagenome interpretation - a benchmark of computational metagenomics software. *BioRxiv*. (doi:10.1101/099127)

22. Velsko, I. M., Frantz, L. A. F., Herbig, A., Larson, G., and Warinner, C. 2018 Selection of appropriate metagenome taxonomic classifiers for ancient microbiome research. *mSystems* **3**. (doi:10.1128/mSystems.00080-18 [doi];mSystems00080-18 [pii])

23. Konstantinidis, K. T. and Tiedje, J. M. 2005 Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci USA* **102**, 2567-2572.

24. Jain, C., Rodriguez, R., Phillippy, A. M., Konstantinidis, K. T., and Aluru, S. 2018 High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114. (doi:10.1038/s41467-018-07641-9 [doi];10.1038/s41467-018-07641-9 [pii])

25. Li, H. 2018 Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100. (doi:4994778 [pii];10.1093/bioinformatics/bty191 [doi])

26. Stamatakis, A. 2014 RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-1313. (doi:btu033 [pii];10.1093/bioinformatics/btu033 [doi])

27. Abranches, J., Zeng, L., Kajfasz, J. K., Palmer, S. R., Chakraborty, B., Wen, Z. T., Richards, V. P., Brady, L. J., and Lemos, J. A. 2018 Biology of oral streptococci. *Microbiol Spectr* **6**. (doi:10.1128/microbiolspec.GPP3-0042-2018 [doi])

28. Johansson, I., Witkowska, E., Kaveh, B., Lif, H. P., and Tanner, A. C. 2016 The microbiome in populations with a low and high prevalence of caries. *J Dent Res* **95**, 80-86. (doi:0022034515609554 [pii];10.1177/0022034515609554 [doi])

29. Oda, Y., Hayashi, F., and Okada, M. 2015 Longitudinal study of dental caries incidence associated with *Streptococcus mutans* and *Streptococcus sobrinus* in patients with intellectual disabilities. *BMC Oral Health* **15**, 102. (doi:10.1186/s12903-015-0087-6 [doi];10.1186/s12903-015-0087-6 [pii])

30. Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., and Phillippy, A. M. 2016 Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132. (doi:10.1186/s13059-016-0997-x [doi];10.1186/s13059-016-0997-x [pii])

31. Becht, E., McInnes, L., Healy, J., Dutertre, C. A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. 2019 Dimensionality reduction for visualizing single-cell data using UMAP. *Nat Biotechnol* **37**, 38-44. (doi:nbt.4314 [pii];10.1038/nbt.4314 [doi])

32. Velsko, I. M., Perez, M. S., and Richards, V. P. 2019 Resolving phylogenetic relationships for *Streptococcus mitis* and *Streptococcus oralis* through core- and pan-genome analyses. *Genome Biol Evol* **11**, 1077-1087. (doi:5371073 [pii];10.1093/gbe/evz049 [doi])

33. Simón-Soro, A. and Mira, A. 2015 Solving the etiology of dental caries. *Trends Microbiol* **23**, 76-82. (doi:S0966-842X(14)00225-X [pii];10.1016/j.tim.2014.10.010 [doi])

34. Richards, V. P., Alvarez, A. J., Luce, A. R., Bedenbaugh, M., Mitchell, M. L., Burne, R. A., and Nascimento, M. M. 2017 Microbiomes of site-specific dental plaques from children with different caries status. *Infect Immun* **85**. (doi:IAI.00106-17 [pii];10.1128/IAI.00106-17 [doi])

35. Bowen, W. H., Burne, R. A., Wu, H., and Koo, H. 2018 Oral biofilms: Pathogens, matrix, and polymicrobial interactions in microenvironments. *Trends Microbiol* **26**, 229-242. (doi:S0966-842X(17)30213-5 [pii];10.1016/j.tim.2017.09.008 [doi])

36. Adler, C. J. *et al.*. 2013 Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. *Nature Genet*.

37. Zhang, X. *et al.*. 2015 The oral and gut microbiomes are perturbed in rheumatoid arthritis and partly normalized after treatment. *Nat Med* **21**, 895-905. (doi:nm.3914 [pii];10.1038/nm.3914 [doi])

38. Cornejo, O. E. *et al.*. 2013 Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. *Mol Biol Evol* **30**, 881-893.

39. Pedregosa, F. *et al.*. 2011 Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830.

40. Platt, J. C. 2019 Probabilistic outputs for Support Vector Machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*.

41. Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. 2002 Gene selection for cancer classification using Support Vector Machines. *Machine Learning* **46**, 389-422. (doi:10.1023/A:1012487302797)

42. Mann, A. E. *et al.*. 2018 Differential preservation of endogenous human and microbial DNA in dental calculus and dentin. *Sci Rep* **8**, 9822. (doi:10.1038/s41598-018-28091-9 [doi];10.1038/s41598-018-28091-9 [pii])

43. Velsko, I. M. *et al.*. 2019 Microbial differences between dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome* **7**, 102. (doi:10.1186/s40168-019-0717-3 [doi];10.1186/s40168-019-0717-3 [pii])

44. Warinner, C. *et al.*. 2014 Pathogens and host immunity in the ancient human oral cavity. *Nature Genet* **46**, 336-344. (doi:ng.2906 [pii];10.1038/ng.2906 [doi])

45. Espinoza, J. L. *et al.*. 2018 Supragingival plaque microbiome ecology and functional potential in the context of health and disease. *MBio* **9**. (doi:mBio.01631-18 [pii];10.1128/mBio.01631-18 [doi])

46. Shi, B., Chang, M., Martin, J., Mitreva, M., Lux, R., Klokkevold, P., Sodergren, E., Weinstock, G. M., Haake, S. K., and Li, H. 2015 Dynamic changes in the subgingival microbiome and their potential for diagnosis and prognosis of periodontitis. *MBio* **6**, e01926-14. (doi:mBio.01926-14 [pii];10.1128/mBio.01926-14 [doi])

47. Liu, B. *et al.*. 2012 Deep sequencing of the oral microbiome reveals signatures of periodontal disease. *PLoS ONE* **7**, e37919. (doi:10.1371/journal.pone.0037919 [doi];PONE-D-11-24763 [pii])

48. McLean, J. S., Liu, Q., Thompson, J., Edlund, A., and Kelley, S. 2015 Draft genome sequence of "Candidatus *Bacteroides periocalifornicus*," a new member of the *Bacteriodetes* phylum found within the oral microbiome of periodontitis patients. *Genome Announc* **3**. (doi:3/6/e01485-15 [pii];10.1128/genomeA.01485-15 [doi])

49. Wang, J., Jia, Z., Zhang, B., Peng, L., and Zhao, F. 2019 Tracing the accumulation of in vivo human oral microbiota elucidates microbial community dynamics at the gateway to the GI tract. *Gut*. (doi:gutjnl-2019-318977 [pii];10.1136/gutjnl-2019-318977 [doi])

50. Marotz, C. A., Sanders, J. G., Zuniga, C., Zaramela, L. S., Knight, R., and Zengler, K. 2018 Improving saliva shotgun metagenomics by chemical host DNA depletion. *Microbiome* **6**, 42. (doi:10.1186/s40168-018-0426-3 [doi];10.1186/s40168-018-0426-3 [pii])

51. Belstrom, D., Constancias, F., Liu, Y., Yang, L., Drautz-Moses, D. I., Schuster, S. C., Kohli, G. S., Jakobsen, T. H., Holmstrup, P., and Givskov, M. 2017 Metagenomic and metatranscriptomic analysis of saliva reveals disease-associated microbiota in patients with periodontitis and dental caries. *NPJ Biofilms Microbiomes* **3**, 23. (doi:10.1038/s41522-017-0031-4 [doi];31 [pii])

52. Lassalle, F., Spagnoletti, M., Fumagalli, M., Shaw, L., Dyble, M., Walker, C., Thomas, M. G., Bamberg, M. A., and Balloux, F. 2018 Oral microbiomes from hunter-gatherers and

traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Mol Ecol* **27**, 182-195. (doi:10.1111/mec.14435 [doi])

53. Takayasu, L. *et al.*. 2017 Circadian oscillations of microbial and functional composition in the human salivary microbiome. *DNA Res* **24**, 261-270. (doi:3052236 [pii];10.1093/dnares/dsx001 [doi])

54. Brito, I. L. *et al.*. 2019 Transmission of human-associated microbiota along family and social networks. *Nat Microbiol* **4**, 964-971. (doi:10.1038/s41564-019-0409-6 [doi];10.1038/s41564-019-0409-6 [pii])

55. Franzosa, E. A. *et al.*. 2014 Relating the metatranscriptome and metagenome of the human gut. *Proc Natl Acad Sci U S A* **111**, E2329-E2338. (doi:1319284111 [pii];10.1073/pnas.1319284111 [doi])

56. Weyrich, L. S. *et al.*. 2017 Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* **544**, 357-361. (doi:nature21674 [pii];10.1038/nature21674 [doi])

57. Lefort, V., Desper, R., and Gascuel, O. 2015 FastME 2.0: A Comprehensive, Accurate, and Fast Distance-Based Phylogeny Inference Program. *Mol Biol Evol* **32**, 2798-2800. (doi:msv150 [pii];10.1093/molbev/msv150 [doi])

Figure 1. Source specificity of the percentage of species composition in 784 oral microbiomes according to SPARSE. (A) X-Y plot of the first two components from a UMAP (Uniform Manifold Approximation and Projection) [31] dimensional reduction of taxon abundances. (B) Neighbour-joining (FastMe2; [57]) hierarchical clustering based on the Euclidean distances between pairs of microbiomes. Euclidean p- distances were calculated between each pair as the square root of the sum of the squared pairwise differences in the percentage of reads assigned by SPARSE to each microbial taxon. Nodes whose cluster location was inconsistent with the UMAP clustering in part A are highlighted with black perimeters. Tree visualization: GrapeTree [4].

Figure 2. Average percentage abundance (left axis) of bacterial species by source for the 40 most influential species according to Support Vector Machine analysis. The relative abundances per source are indicated by the three bars for each species in mini-histograms. Species are ordered from left to right according to decreasing SVM weight (right axis). SVM weight was calculated as the squared coefficients in the optimal SVM model that separated samples according to oral source. Species belonging to oral complexes are indicated by filled circles of the corresponding colours. Legend: Source colours and symbol for SVM weight

Figure 3. Average percentage abundances of 28 species from six oral complexes [5] in 784 metagenomes by oral source. The percentage abundances per source are indicated by the three bars for each species in mini-histograms. Species are ordered from left to right by oral complex (colours). Within each oral complex, the order is by decreasing total abundance.

Figure 4. Neighbour-joining (FastMe2; [57]) hierarchical clustering based on the Euclidean distances between pairs of 245 microbial species whose percentage abundance was >2% in at least one microbiome. Species names are indicated for members of the six oral complexes [5], and four apparent clusters of oral complexes are highlighted in gray. An expanded version of the same tree including all species labels is available in Fig. S2.

Figure 5. Numbers of microbial taxa by source. A). Rarefaction curves of numbers of species with 95% confidence estimates (shadow) by source. Inset data indicates median numbers of species per source as well as the total numbers for all samples. Rarefactions were performed with the program script called SPARSE_curve.py using 1000 randomized permutations of the order of samples. B). Binned histogram of number of species by percentage of samples containing those numbers. The data for this plot was also calculated with SPARSE_curve. C) Venn diagram of overlapping presence of taxa (>0.0005% abundance) for the three oral sources.

Figure 6. Reconstruction of *S. mutans* and *S. sobrinus* MAGs from oral metagenomes. (A, C) Numbers of oral samples by source internally binned by the percentage of reads specific to *S. mutans* (A) and *S. sobrinus* (C). (B, D) Numbers of oral samples by source internally binned by the predicted read coverage of a reference genome of *S. mutans* (B) and *S. sobrinus* (D). (E, F) Read coverage (left) and percentage of the reference genome that was unmasked (≥3 reads; ≥ 70% consistency) (right) in *S. mutans* (E) and *S. sobrinus* (F). Ordered by decreasing coverage.

Figure 7. ML phylogenies of *S. mutans* and *S. sobrinus* genomes. (A) A RaxML tree [26] of 195 genomes from RefSeq and 31 MAGs (metagenomic assembled genome) of *S. mutans* plus one genome of *S. troglodytae* as an outgroup. The tree was based on 181,321 non-repetitive SNPs in 1.73 Mb. (B) A RaxML tree of 50 genomes from RefSeq and 16 *S. sobrinus* MAGs plus 6 *S. downei* genomes as an outgroup. This tree was based on 160,863 non-repetitive SNPs in 1.13 Mb. MAGs are highlighted by thick black perimeters. Visualisation with GrapeTree [4]. Branches with a genetic distance of >0.1 were shortened for clarity, and shown as dashed lines. Legend: Numbers of strains by country of origin for both trees.

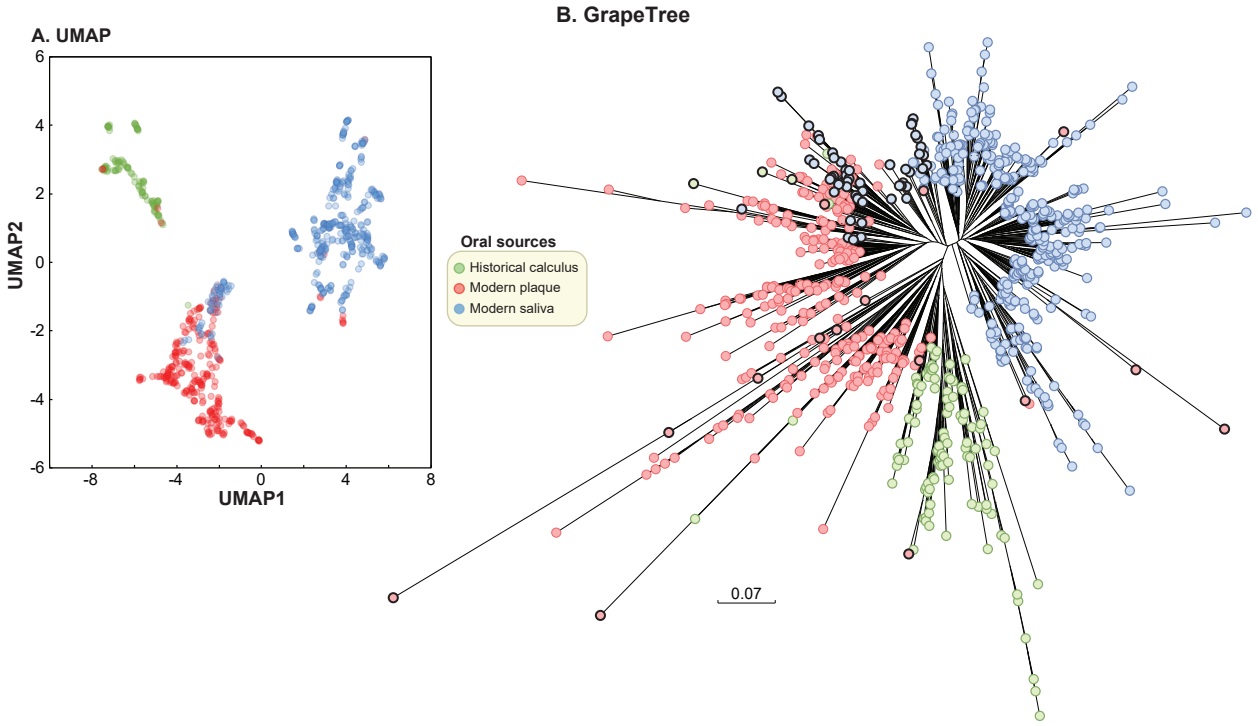


Figure 1. Source specificity of the percentage of species composition in 784 oral microbiomes according to SPARSE. (A) X-Y plot of the first two components from a UMAP (Uniform Manifold Approximation and Projection) [31] dimensional reduction of taxon abundances. (B) Neighbour-joining (FastMe2; [57]) hierarchical clustering based on the Euclidean distances between pairs of microbiomes. Euclidean p- distances were calculated between each pair as the square root of the sum of the squared pairwise differences in the percentage of reads assigned by SPARSE to each microbial taxon. Nodes whose cluster location was inconsistent with the UMAP clustering in part A are highlighted with black perimeters. Tree visualization: GrapeTree [25].

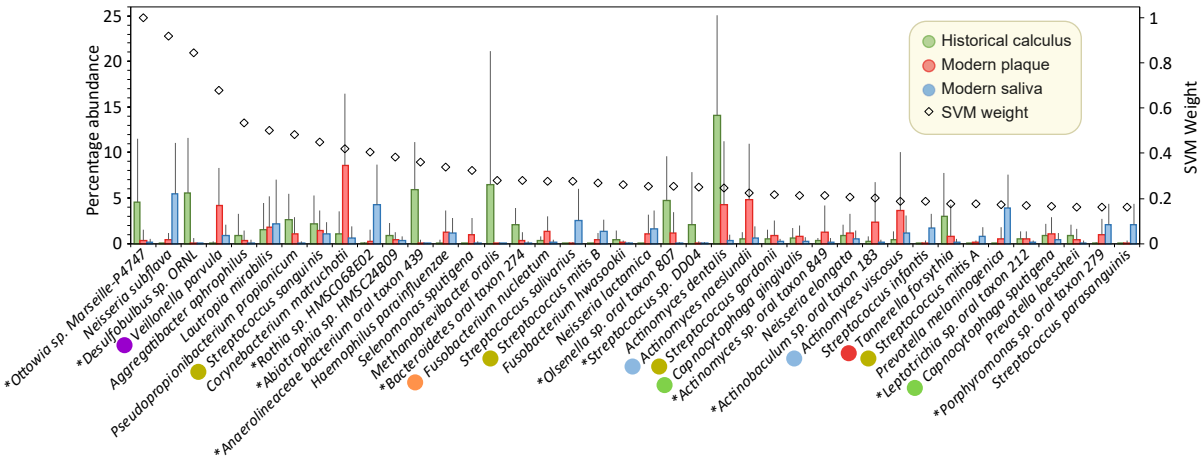


Figure 2. Percentage abundance (left axis) of bacterial species by source for the 40 most influential species according to Support Vector Machine analysis. The relative abundances per source are indicated by the three bars for each species in mini-histograms. Species are ordered from left to right according to decreasing SVM weight (right axis). SVM weight was calculated as the squared coefficients in the optimal SVM model that separated samples according to oral source. Species belonging to oral complexes are indicated by filled circles of the corresponding colours. Legend: Source colours and symbol for SVM weight.

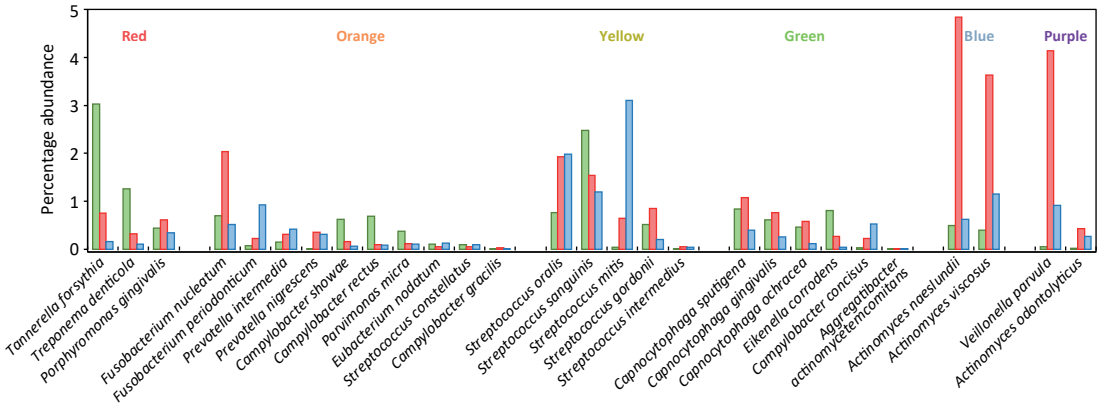


Figure 3. Average percentage abundances of 28 species from six oral complexes [30] in 784 metagenomes by oral source. The percentage abundances per source are indicated by the three bars for each species in mini-histograms. Species are ordered from left to right by oral complex (colours). Within each oral complex, the order is by decreasing total abundance.

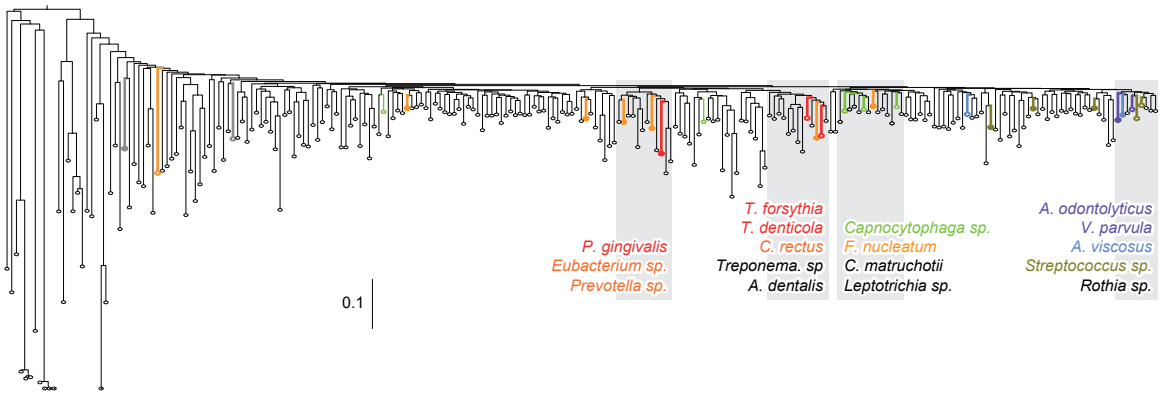
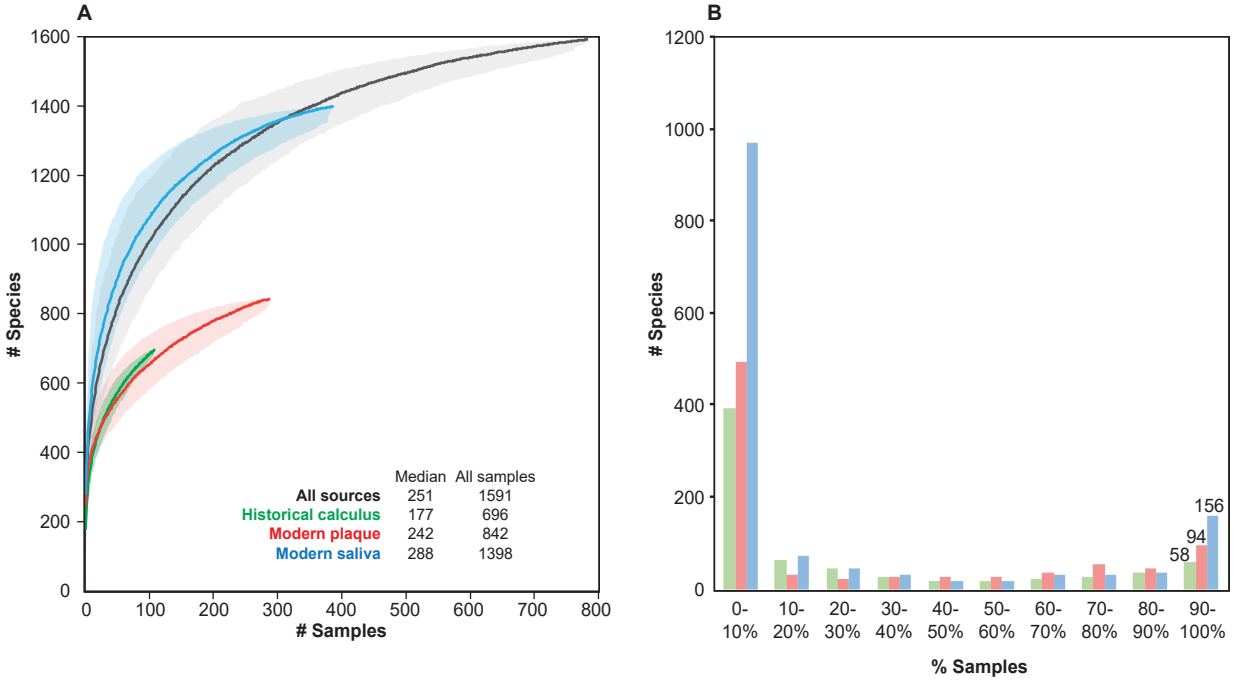


Figure 4. Neighbour-joining (FastMe2; [57]) hierarchical clustering based on the Euclidean distances between pairs of 245 microbial species whose percentage abundance was >2% in at least one microbiome. Species names are indicated for members of the six oral complexes [30], and four apparent clusters of oral complexes are highlighted in gray. An expanded version of the same tree including all species labels is available in Fig. S2.



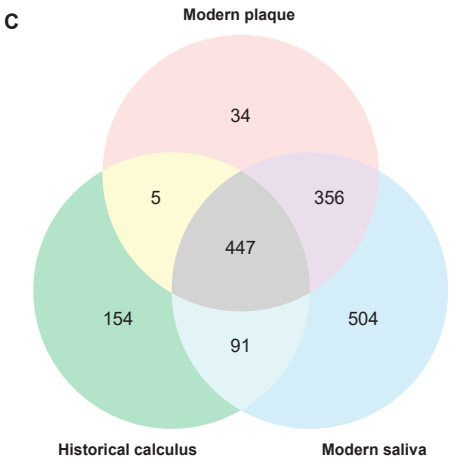


Figure 5. Numbers of microbial taxa by source. A). Rarefaction curves of numbers of species with 95% confidence estimates (shadow) by source. Inset data indicates median numbers of species per source as well as the total numbers for all samples. Rarefactions were performed with the program script called SPARSE_curve.py using 1000 randomized permutations of the order of samples. B). Binned histogram of number of species by percentage of samples containing those numbers. The data for this plot was also calculated with SPARSE_curve. C) Venn diagram of overlapping presence of taxa (>0.0005% abundance) for the three oral sources.

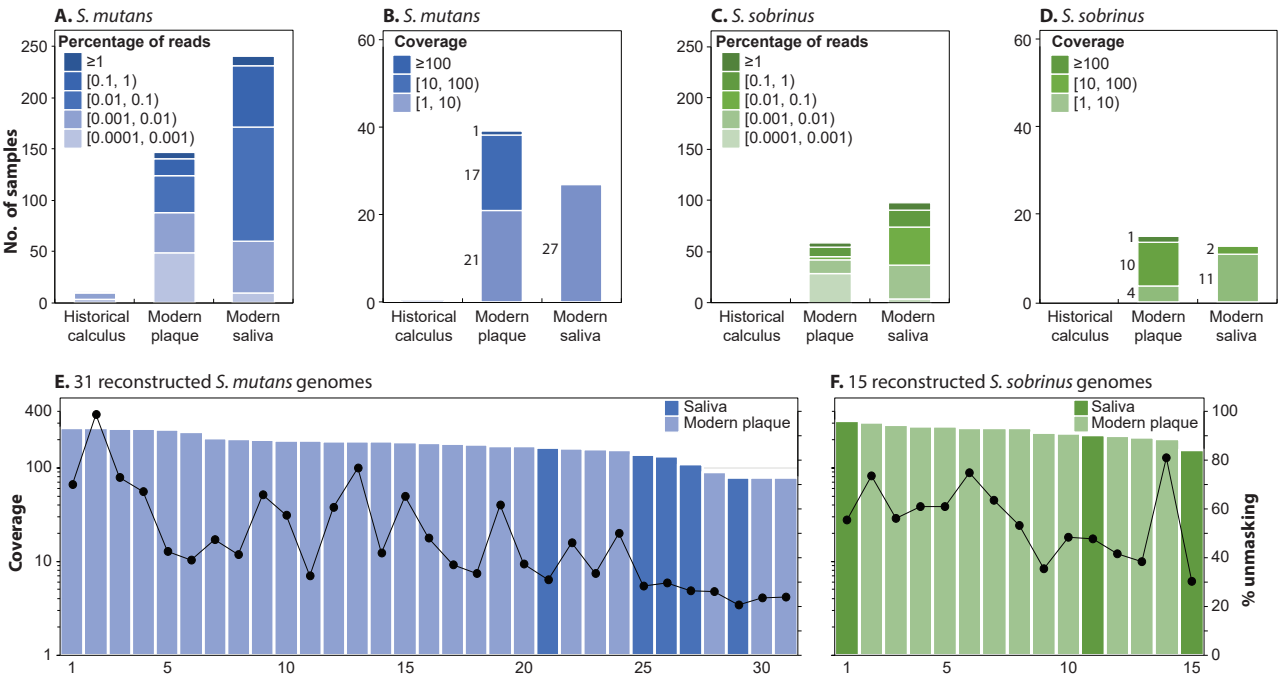


Figure 6. Reconstruction of *S. mutans* and *S. sobrinus* MAGs from oral metagenomes. (A, C) Numbers of oral samples by source internally binned by the percentage of reads specific to *S. mutans* (A) and *S. sobrinus* (C). (B, D) Numbers of oral samples by source internally binned by the predicted read coverage of a reference genome of *S. mutans* (B) and *S. sobrinus* (D). (E, F) Read coverage (left) and percentage of the reference genome that was unmasked (≥3 reads; ≥70% consistency) (right) in *S. mutans* (E) and *S. sobrinus* (F). Ordered by decreasing coverage.

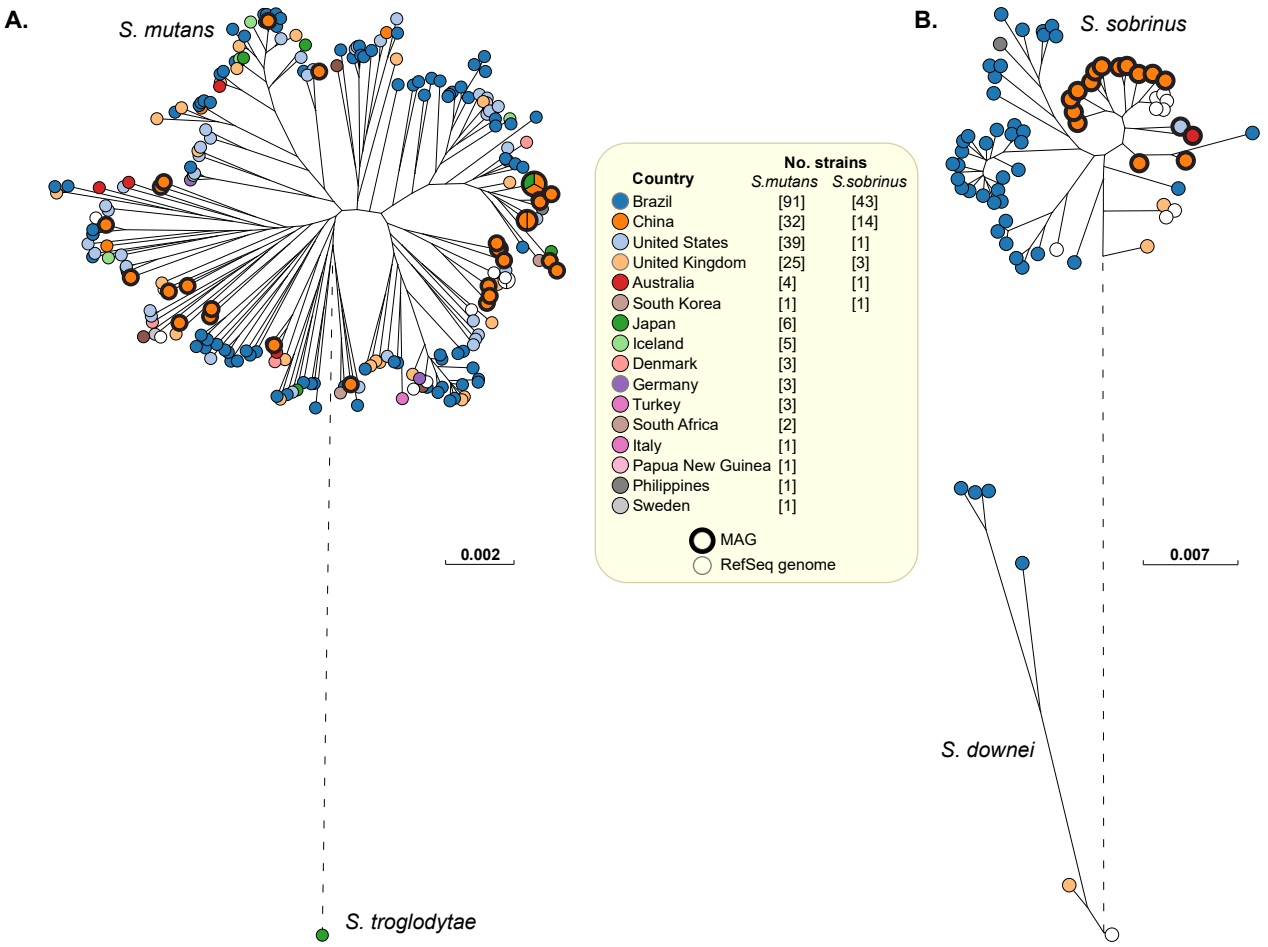
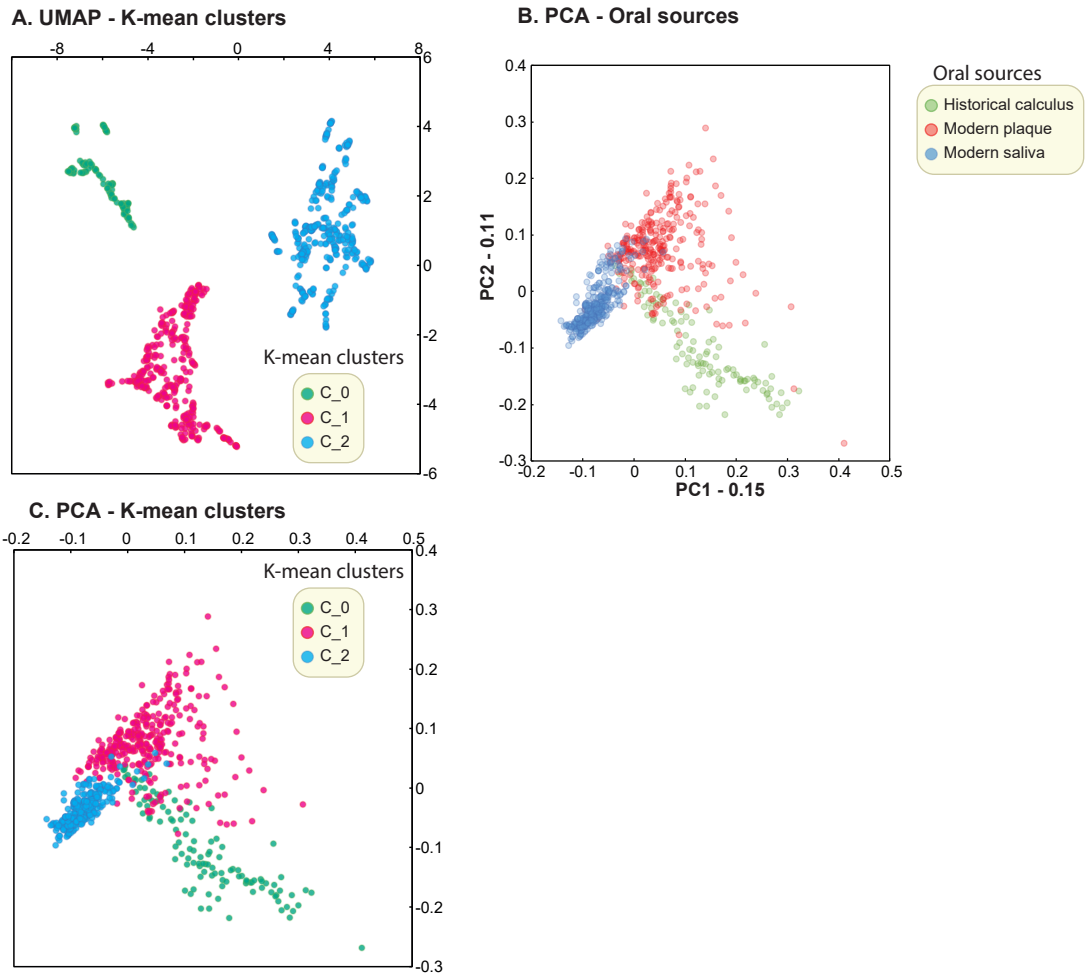


Figure 7. ML phylogenies of *S. mutans* and *S. sobrinus* genomes. (A) A RaxML tree [24] of 195 genomes from RefSeq and 31 MAGs (metagenomic assembled genome) of *S. mutans* plus one genome of *S. troglodytae* as an outgroup. The tree was based on 181,321 non-repetitive SNPs in 1.73 Mb. (B) A RaxML tree of 50 genomes from RefSeq and 16 *S. sobrinus* MAGs plus 6 *S. downei* genomes as an outgroup. This tree was based on 160,863 non-repetitive SNPs in 1.13 Mb. MAGs are highlighted by thick black perimeters. Visualisation with GrapeTree [25]. Branches with a genetic distance of >0.1 were shortened for clarity, and shown as dashed lines. Legend: Numbers of strains by country of origin for both trees.



Supplemental Figure S1. The K-mean clusters of the first three components of the UMAP analyses and their visualization in PCA plots. (A) The visualization of the three clusters in the plane of UMAP components. (B) Plot of the first two components of the PCA analysis. (C) The same plot as (B) with nodes color-coded by K-mean clusters in (A).

Supplemental FIgure S2. Neighbor-joining clustering of microbial species based on the similarities of their abundances in all samples as in Figure 4.

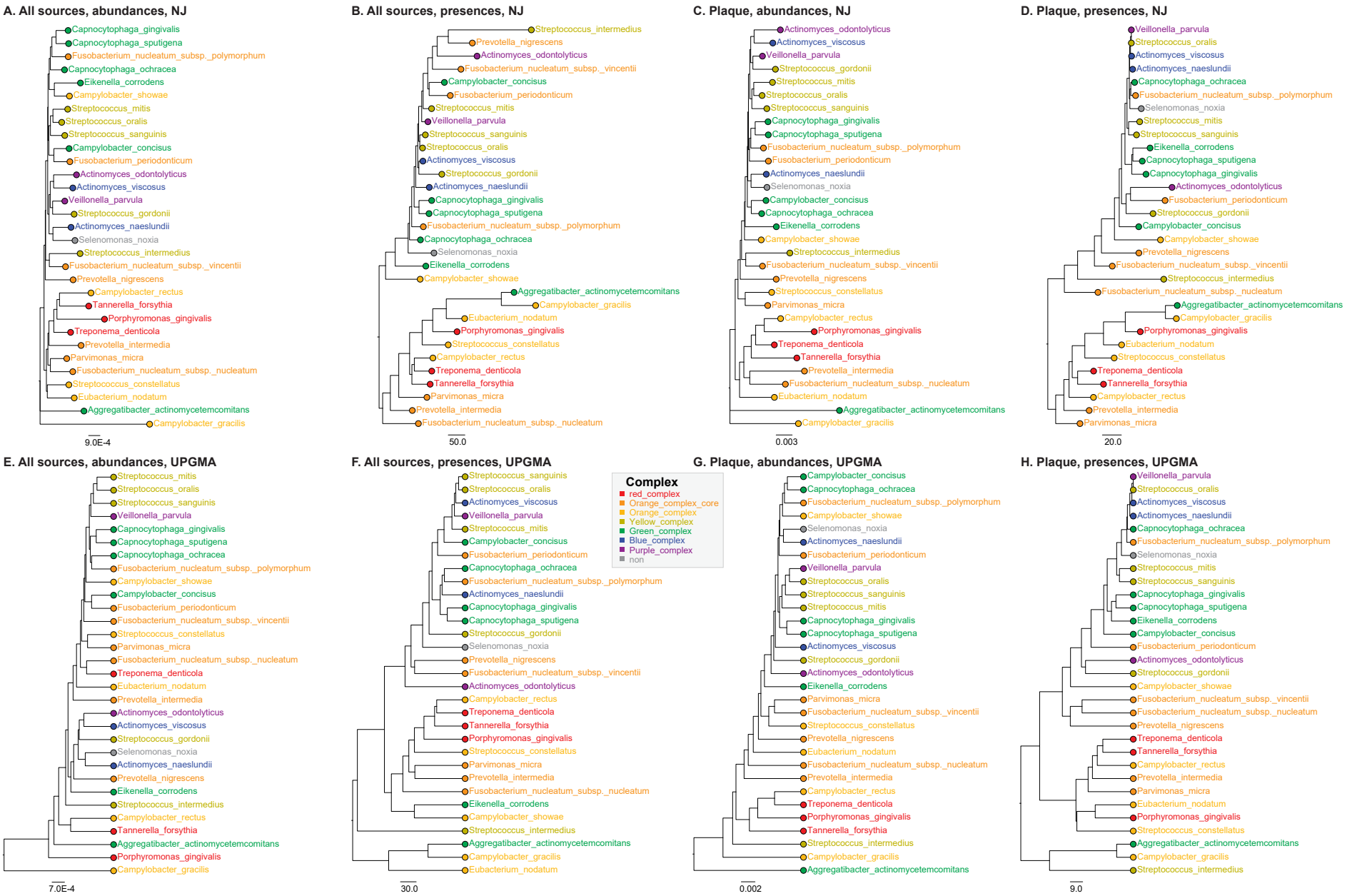


Figure S3. Neighbor-joining and UPGMA clustering of the 28 species described in Socransky et al. based on their abundances or presences in the oral samples.

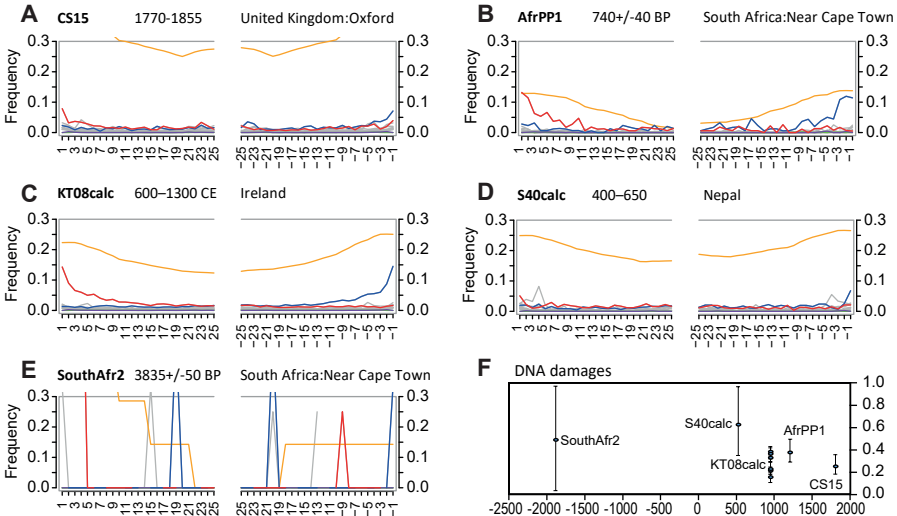


Figure S4. Increased deamination frequencies at the 5′-end of sequencing reads from ancient metagenomes. (A-E) The results of MapDamage2 on selected read sets from historical samples. (F) A summary of mapDamage estimates (Δ_S) of *S. mutans* reads within the historical metagenomes. MapDamage2 was run on a BAM alignment of species-specific reads against the reference genome of *S. mutans* UA159 using default parameters.