Genome Biology

https://doi.org/10.1186/s13059-026-04018-w

Article in Press

The genomic history of *Streptococcus mutans* from the Mesolithic until modern times

Received: 11 Sep 2024

Accepted: 24 Feb 2026

Published online: 12 March 2026

Cite this article as: Thygesen, V., Farahani, M., Nielsen, S. *et al.* The genomic history of *Streptococcus mutans* from the Mesolithic until modern times. *Genome Biol* (2026). https://doi.org/10.1186/s13059-026-04018-w

Vincent Thygesen, Motahare Farahani, Sofie Nielsen, Florentin Constancias, Michael Givskov, Jacqueline Abranches, Gabriele Scorrano, Marie Jørkov, Ghader Ebrahimi, Julio Bendezu-Sarmiento, Fabrice Demeter, Kristian Kristiansen, Daniel Belstrøm & Martin Sikora

The genomic history of *Streptococcus mutans* from the Mesolithic until modern times

Vincent F.D.F. Thygesen[1], Motahare Feizabadi Farahani[2], Sofie Holtsmark Nielsen[3], Florentin Constancias[1], Michael Givskov[4], Jacqueline Abranches[5], Gabriele Scorrano[6], Marie Louise S. Jørkov[7], Ghader Ebrahimi[8], Julio C. Bendezu-Sarmiento[9], Fabrice Demeter[7,9], Kristian Kristiansen[7,10], Daniel Belstrøm[1*#], Martin Sikora[7*#]

1: Department of Odontology, Section for Clinical Oral Microbiology, Faculty of Health and Medical Sciences, University of Copenhagen, Denmark.

2: SciLifeLab Ancient DNA unit, Dept. of Archaeology and Classical Studies, Stockholm University, Sweden

3: Department of Bacteria, Parasites and Fungi, Statens Serum Institut, Denmark

4: Costerton Biofilm Center, University of Copenhagen, Denmark

5: Department of Oral Biology, College of Dentistry, University of Florida, United States of America

6: Center for Molecular Anthropology for the study of ancient DNA, Department of Biology, University of Rome Tor Vergata, Italy

7: Section for GeoGenetics, Globe Institute, Faculty of Health Science, University of Copenhagen, Denmark

8: Department of Archaeology, University of Mohaghegh Ardabili, Iran

9: Eco-anthropologie (EA), Dpt ABBA, Muséum national d'Histoire naturelle, CNRS, Université Paris Cité, France

10: Department of Historical Studies, University of Gotenburg, Sweden

#: Corresponding author

*: These authors contributed equally.

Corresponding authors:

Daniel Belstrøm, Professor

Department of Odontology, Section for Clinical Oral Microbiology

Faculty of Health and Medical Sciences

University of Copenhagen

Nørre alle 20

2200 Copenhagen, Denmark

E-mail: dbel@sund.ku.dk

Phone: +45 21300580

Martin Sikora, Associate Professor

Globe institute, Section for Geogenetics

Faculty of Health and Medical Sciences

University of Copenhagen

Øster Voldgade 5-7

1350 Copenhagen, Denmark

E-mail: martin.sikora@sund.ku.dk

Phone: +45 93565403

Keywords: Ancient oral microbiota, Pathogen genomics, Microbial evolution, Pangenomics, Streptococcus mutans

Abstract

Background:

Streptococcus mutans is a member of the human oral microbiota and is considered one of the most important cariogenic organisms. Previous studies have suggested an expansion of *S. mutans* populations about 10,000 years ago with the onset of agriculture, yet direct molecular evidence of its presence from ancient DNA remains sparse.

Results:

Here, we present population genomic analyses of 25 ancient *S. mutans* genomes (average read depth 0.1X – 387X) recovered from archaeological remains across Eurasia spanning ~8,000 years of human evolution. Recombination-corrected phylogenomic analyses using Gubbins show a star-like phylogeny indicative of an early radiation, with the ancient genomes falling within the genomic diversity of modern isolates but restricted to one of the major clades of the phylogeny (D). Analyses of genes encoding present day virulence factors reveals that the presence of the mutanobactin operon involved in oxygen tolerance is restricted to specific subclades (A & B) and absent among the ancient samples. Using the MEGAHIT assembler followed by binning of contigs with CONCOCT, we recover metagenome-assembled genomes (MAG) of 7 high-coverage ancient S. mutans strains, including a 7,500-year-old sample from an early European Neolithic farmer. Pangenome analysis with modern isolates using the anvi'o's suite revealed the presence of specific functional genes in the ancient isolates, which were lost through time.

Conclusions:

Our study demonstrates that *Streptococcus mutans* DNA is well preserved in tooth samples from archaeological remains and show that it formed part of the human oral microbiota already before

the onset of agriculture, consistent with a radiation and population expansion well before 8,000 years ago.

Background

Dental caries is the most widespread microbially-driven disease affecting more than 2 billion people worldwide[1,2]. Dental caries is an oral biofilm-induced disease, where fermentable carbohydrates are metabolized, generating acidic byproducts like lactic acid[3,4]. This induces an ecological shift that creates an acidic anaerobic environment, in which specific oral bacterial species such as *Streptococcus mutans* thrives[5]. Consequently, *S. mutans* has for many decades been considered the most important cariogenic organism [6].

S. mutans is a gram positive facultative anaerobic coccus, residing in oral biofilms as a member of the resident oral microbiota[7]. Since its discovery in 1924[8], *S. mutans* has for many decades been considered as the model bacterium of dental caries[7]. The reason for the prominent position of *S. mutans* in caries research is that *S. mutans* can perform all required biochemical processes driving or triggering development of the disease. As such, *S. mutans* can degrade carbohydrates into organic acids, which is essential to establishing the acidic environment of the caries lesion. In addition, *S. mutans* is proficient in synthesizing the extracellular polymers needed to create a biofilm that can maintain a low pH and anaerobic conditions of a caries lesion, in which *S. mutans* unlike most other oral species is able to thrive[7,9]. Previous genomic studies have identified and described several virulence factors of *S. mutans*, primarily involved in carbohydrate metabolism and acid tolerance[10-11]. However, the evolution of the *S. mutans* genome remains to be uncovered.

Ancient samples from the oral cavity are an important source of information in the field of paleomicrobiology, as DNA is highly preserved in mineralized tissues of the oral cavity, such as teeth and dental calculus[12.] However, only a few studies have attempted to reconstruct ancient oral

microbiotas[13-15]. When used in combination with data from modern samples, ancient DNA sequencing provides a unique perspective, due to the ability to reconstruct ancient bacterial genomes and trace the composition and likely evolution of the microbiota and its adaptations through time.

Metagenomics based studies have successfully demonstrated the presence of *S. mutans* DNA in ancient oral samples[16-18], which suggests that *S. mutans* has been part of the resident oral microbiota for millennia. Previous studies have collectively reported a low frequency of caries lesions in ancient skulls as compared to modern data, albeit an increase in lesions seems to be associated with the transition from hunter-gatherers to agriculture[19]. Collectively, these findings raise the question if specific core genes of present day *S. mutans* are the consequence of adaptation to a modern lifestyle, characterized by increased consumption of simple, fermentable carbohydrates[20].

To answer these questions, the purpose of the present study was to characterize spatiotemporal variation of the genome of *S. mutans*. We employed state of the art methodologies[21-22] for reconstruction of *S. mutans* genomes from 1,313 ancient human samples, originating from across Eurasia, spanning the period from the Mesolithic and Neolithic Ages (10000 - 4500 BCE) up to the end of the Viking Age (~1000 CE)[23-24], and compared data with modern reference genomes of *S. mutans*. We tested the hypothesis that the genome of *S. mutans* has been subject to significant modification through time because of changes in human diet and lifestyles

Methods

Reference-based dataset generation

Ancient samples for analysis were selected based on having been identified as positive for *Streptococcus mutans* (N>5,000 reads mapped) in a screening study for ancient pathogens including

1,313 ancient samples[25]. We further included three high coverage samples (>6X) identified from screening shotgun sequencing data of additional >5,000 ancient samples available at Globe Institute. For each positive sample, adapter-trimmed shotgun sequencing reads were mapped against the *S. mutans* NCH105 reference assembly (accession GCF_009738105.1) using bowtie2[26] (https://github.com/BenLangmead/bowtie2 2.5.1) with the very sensitive bowtie2 preset: "-D 20 -R 3 -N 1 -L 20 -i S,1,0.50", modified by N set to 1 to further increase sensitivity.

The resulting alignment files files were sorted, indexed, and filtered using samtools[27] (1.15 https://github.com/samtools/ filters: "-q1 -F 0x400"), coverage statistics were calculated using bedtools[28] (2.30.0 https://github.com/arq5x/bedtools2) and duplicate reads were marked using picard[29] (3.0.0 https://broadinstitute.github.io/picard). Authentication of ancient DNA and estimation of damage patterns was carried out using metaDMG[30] (https://github.com/metaDMG-dev/metaDMG-core 1.37). Summary statistics for read mappings were calculated and plotted using the Rsamtools[31] (2.16 https://bioconductor.org/packages/release/bioc/html/Rsamtools.html) and tidyverse[32] (2.0.0 https://www.tidyverse.org/) packages in R. Coverage statistics for annotated genes were obtained using mosdepth[33] (0.3.4 https://github.com/brentp/mosdepth) and bedtools. For ancient samples, the gene coverage was normalized as the observed/expected breadth of coverage: $covPratio_{gene} = min(1, coverageP_{gene}/expcoverage_{genome})$ (Additional file 3: Table S1)

Where:

- $coverageP_{gene}$= the percentage of the gene covered (0,1) (Additional file 3: Table 1)
- $expcoverage_{genome} = 1 - e^{-(PercentageReferenceCovered_{genome})}$ (0,1)

To determine strain multiplicity for high coverage samples (>5X read depth), we carried out genotype calling using 'bcftools call' with '—ploidy 2' option, and calculated the rate of observing multiple alleles at all sites covered by at least five reads in each sample. Haploid genotype calls for samples without evidence for multiple infections were then generated using 'bcftools call' with '—

ploidy 1' option. For the remaining samples (high coverage multi-infection as well as low coverage), we generated consensus sequences by calling the majority allele across all aligned reads with mapping quality >20 at each reference genome position covered.

To construct a reference dataset of modern genomes, we downloaded 255 *S. mutans* isolates available May 2023 from NCBI (ncbi-genome-download --format fasta,assembly-report -P --genera "Streptococcus mutans" --metadata strep_metadata.tsv bacteria). Each modern sample was aligned with the *S. mutans* NCH105 reference assembly (accession GCF_009738105.1) using minimap2[34] (2.26-r1175 https://github.com/lh3/minimap2), using parameters (-cx asm5) recommended for intra-species asm-to-asm alignment. Variants against the reference were called and extracted in VCF format using the paftools.js extension, using a minimum alignment length of 10kb, and complemented with the reference genome against all aligned regions using bedtools, to generate the final reference aligned fasta files for each assembly. These resulting reference-aligned fasta files were concatenated and converted back to VCF format using snp-sites[35] (2.5.1 https://github.com/sanger-pathogens/snp-sites), and used to generate core genome alignments for *S. mutans*, requiring >= 99% of modern samples able to be aligned against the reference. For samples with identical core genomes, we removed all but one of the samples for downstream analyses.

The dataset for the PCA and clustering was obtained by extracting bi-allelic SNPs across modern and high coverage ancient samples across the core genome, with genotypes for lower coverage ancient samples set to missing if neither of the two alleles was observed.

Population genomics and phylogenetics

Genetic variation among modern and ancient *S. mutans* genomes was investigated using principal component analysis (PCA) using plink[36] (1.90b6.21 www.cog-genomics.org/plink/1.9/) and GCTA[37] (1.94.1). Haplotype sharing and clustering analyses were carried out using

ChromoPainterV2[38] (https://github.com/sahwa/ChromoPainterV2) and fineSTRUCTURE v2[38], with all samples designated as donors and recipients.

Maximum-likelihood phylogenetic analysis was carried out using RaxML-ng[39] (1.2.0 https://github.com/amkozlov/raxml-ng) using the GTR+GAMMA model with 200 nonparametric bootstrap replicates. Using this tree as a starting tree, a recombination corrected phylogenetic tree was then constructed using Gubbins run_gubbins.py (3.3.1 http://nickjcroucher.github.io/gubbins/) over 20 iterations of recombinant region masking on the MSA and subsequent RaxML-ng tree construction, again with a GTR+GAMMA substitution model using the previous RaXMl-NG final tree as a starting tree. A final tree was constructed afterwards with 200 nonparametric bootstrap replicates. From the workflow, a recombination masked alignment of the MSA was extracted using mask_gubbins_aln.py. Low coverage ancient samples were placed onto a RaxML-ng phylogenetic tree built from the masked MSA from Gubbins using EPA-ng[40] (0.3.8 https://github.com/pierrebarbera/epa-ng), and merged with the rest of the tree using gappa[41] (0.8.2 https://github.com/lczech/gappa). Phylogenetic trees were visualized and compared for bootstrap confidence, ChromoPainter cluster and sample metadata concordance in R using tidyverse (2.0.0 https://www.tidyverse.org/), together with ggtree[42] (3.8.0 https://bioconductor.org/packages/release/bioc/html/ggtree.html) ape[43] (5.7.1 https://github.com/emmanuelparadis/ape) and aplot[44] (0.22 https://cran.r-project.org/web/packages/aplot/index.html). These packages were also used to visualize the distribution of genes across the phylogeny, and for placement weight evaluation, we also used treeIO[45] (1.24.1 https://bioconductor.org/packages/release/bioc/html/treeio.html).

Construction of MAGs

To generate a genus-wide representative panel of modern *Streptococcus* reference genomes, a total of 656 *Streptococcus* genomes were downloaded from NCBI on September 2021 and clustered based on a 99% average nucleotide identity (ANI) similarity threshold computed through fastANI[46] (https://github.com/ParBLiSS/FastANI 1.33). From each cluster, a representative reference was picked based on its centrality in the cluster through TraMineR[47] (http://traminer.unige.ch/ 2.25) in R[48] (4.3.0 https://www.r-project.org/).

After removal of human reads using KrakenUniq[49] (v0.7 https://github.com/fbreitwieser/krakenuniq) with a custom database of 25272 human and human associated microbial genomes[50], those ancient samples with over 85 percent of the reference genomes covered for a *S. mutans* reference had their reads assembled using three different MEGAHIT[51] (v1.2.7 https://github.com/voutcn/megahit) settings - henceforth referred to as strategy "A", "B" and "C" (Additional file 3: Table 2). Each assembly strategy was explored using both a contig cutoff of 1kb and 2.5kb.

Contigs databases were created for each sample assembly using anvi'o[52] (v7.1 https://anvio.org/). The contigs databases were annotated with anvi'os default hidden Markov model (HMM) taxonomic profiles using HMMER (3.4 http://hmmer.org/)[53], NCBI-COGS[54] gene functions through DIAMOND (v2.1.12.166 http://www.diamondsearch.org)[55] and a Genome Taxonomy Database (GTDB) based Single-copy Core Gene (SCG) database for gene taxonomy[56]. The original short reads for each sample were mapped back to their respective contigs databases using bowtie2, which were used together to create sample contig profiles using anvi'o.

The contigs were binned using CONCOCT[57] (1.1.0 https://github.com/BinPro/CONCOCT) and METABAT2[58] (2.15 https://gensoft.pasteur.fr/docs/MetaBAT/2.15/) and anvi'o was used to summarize the binning results by linking their constituent contigs with the information in their corresponding sample contig and profile databases.

To evaluate which method would most consistently deliver high quality *S. mutans* bins, we compared different summary metrics such as N50, estimated completion, length and redundancy. Once the ideal assembly and binning strategy had been chosen, the *S. mutans* consensus sequence for each sample, corresponding to the contigs binned as *S. mutans*, were used to compute anvi'o contigs databases again, which were compared which each other in an anvi'o pangenomic analysis[59]. From this comparative analysis, unique gene clusters were identified, and the sequences of these were linked back to the contigs they belonged to in the original *S. mutans* bins. Read damage estimates were then made using metaDMG, by which the authenticity of each contig could be quantified by tallying the damage estimations for all reads that corresponded to that contig. These two pieces of contig information were padded unto the anvi'o contig and profile databases, which were used to inform the manual refinement of the *S. mutans* bins into their final MAGs[60]. The refined MAG's were checked for contamination using gunc[61] (1.0.6 https://github.com/grp-bork/gunc) which analysed the taxonomy of genes identified in contigs against GTDB. Contaminated MAGs were filtered by removing contigs which were assigned to a phylogenetic lineage different from *S. mutans* using mmseqs2[62] (15.6f452 https://github.com/soedinglab/MMseqs2) taxonomy with GTDB until they all passed the gunc analysis (See Additional file 1: Supplemental information A).

Our ancient *S. mutans* MAGs and the set of modern *S. mutans* assembly used for the phylogenetic analysis were annotated using the bakta[63] (1.9.3 https://github.com/oschwengers/bakta) database. The presence of genes annotated in the ancient and modern isolates through bakta which could not be found within the reference genome (GCF_009738105.1) were visualized as a heatmap against our phylogenetic tree including the 3 placements that were MAGs, as had been done previously for the phylogenetic analysis.

Pangenomics

Besides our ancient MAG's, 10 modern *S. mutans* reference genomes, and 15 representative species from across the genus *Streptococcus* were chosen for this analysis.

Contigs databases were created for each sample and annotated with hits from the KEGG+KOfam[64] databases through anvi'o. Anvi'o was then used on these databases to estimate metabolic pathway completion across the dataset according to the KEGG database[65], enabling a comparative analysis of the enrichment of metabolic modules between ancient/modern *S. mutans* and the other *Streptococcus* species.

The modules which differed in enrichment across these 3 groups were selected, and their presence in the dataset were compared in a heatmap in R through the gplots[66] package (3.1.3 https://www.rdocumentation.org/packages/gplots/versions/3.1.3/topics/heatmap.2).

Results

Overview of the dataset

To investigate the origins and past genetic diversity of *Streptococcus mutans*, we generated a dataset of ancient *S. mutans* genomes by mapping shotgun-sequenced ancient metagenomic DNA reads to a modern *S. mutans* reference assembly (GCF_009738105.1). A total of 25 ancient *S. mutans* genomes passed stringent ancient DNA authentication criteria, with average read depth ranging from 0.08X to 387.5X, including seven high coverage genomes (read depth >7X), three of which were deeply sequenced to >35X (Additional file 3: Table 1). Genomic similarity of the ancient genomes to the modern reference as measured through average nucleotide similarity (ANI) was high (98 % - 99 %; Additional file 3: Table 1, Additional file 2: Fig. S8). Among the genomes

we found evidence for infections by a clonal strain (multi-allele rate $<10^{-4}$) and infection by a mixture of strains (multi-allele rate $\geq 10^{-4}$).

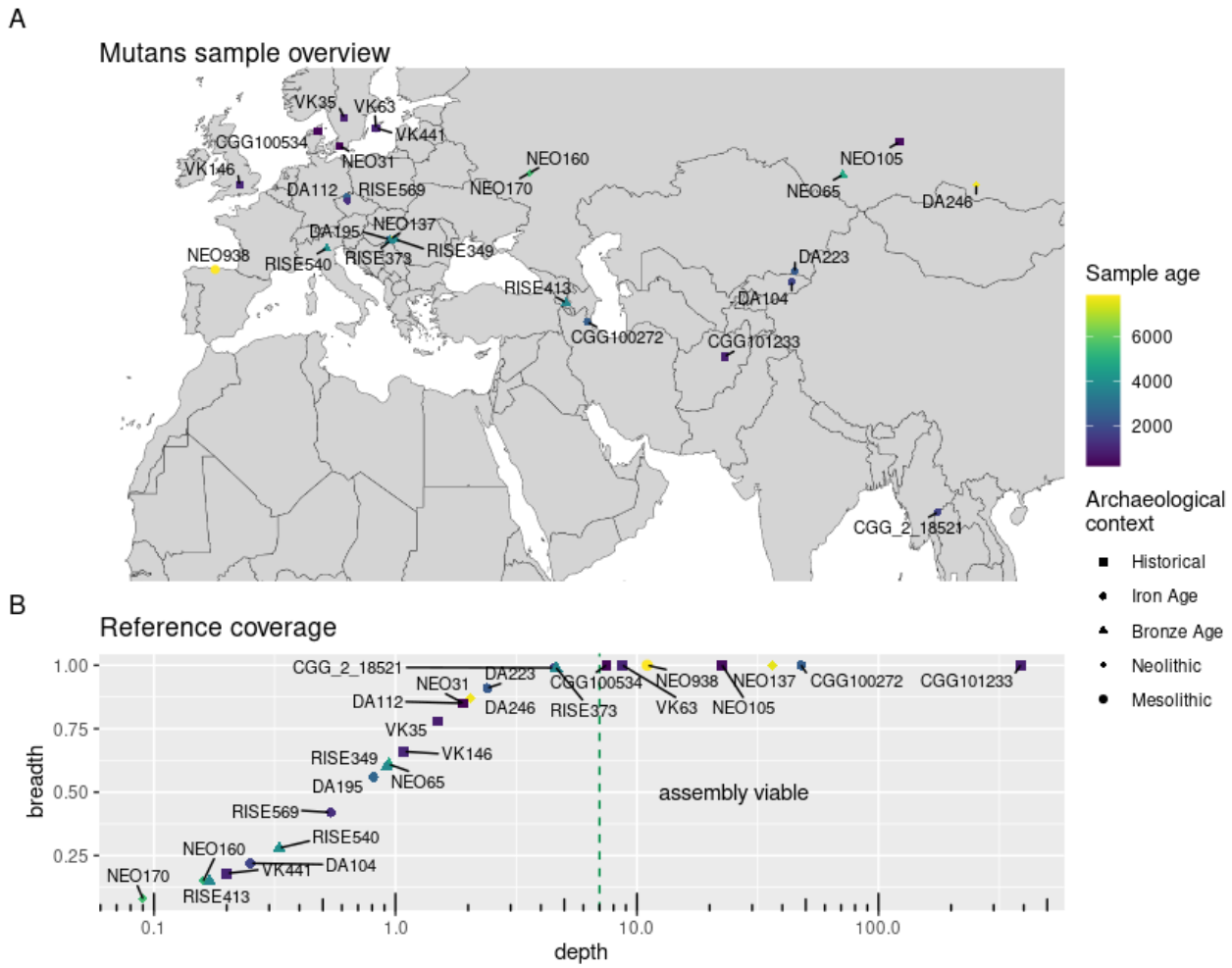

Figure 1. **Dataset overview**. **(A)** Spatiotemporal distribution of the 25 ancient *S. mutans* genomes reported in this study. Sample ages and archaeological period are indicated with symbol color and shape, respectively. **(B)** Mapping statistics showing average read depth (x axis) and breadth of genomic coverage (y axis) for each ancient genome. Minimum read depth cutoff for samples used in metagenomic assembly is indicated by dashed line.

The ancient genomes in our dataset spanned a broad geographic range from Europe to North and Southeast Asia, with sample ages from ~7,800 to 280 years before present (Figure 1; Additional file 3: Table 1). The earliest occurrences of *S mutans* were found in a Mesolithic individual from Spain, Southern Europe (NEO938; El Mazo; 7,966 – 7,789 years cal. BP) and a Neolithic individual from

Lake Baikal, North Asia (DA246; Shamanka; 7,827 – 7,626 years cal. BP), both associated with hunter-gatherer contexts.

Genetic structure of S. mutans

To investigate the population structure of our ancient *S. mutans* within the context of modern diversity, we merged the ancient genomes with publicly available genome assemblies of 255 modern *S. mutans* isolates (Additional file 2: Fig. S1). As *S. mutans* has previously shown to be a highly recombining species, we first investigated its population structure with a recombination-aware chromosome painting approach (Methods). We painted the core genomes of each modern assembly and the four clonal high coverage (>7X) ancient genomes with each of them as potential donors using ChromoPainter. The resulting matrix of haplotype chunk counts representing genetic similarity among genomes (Additional file 2: Fig. S2) was then used to perform a hierarchical clustering using fineSTRUCTURE. At the highest resolution, the 259 genomes formed 180 distinct clusters, which were aggregated into five higher level clusters to highlight the broad genetic structure among the modern and ancient genomes (Figure 2; Additional file 2: Fig. S3). Recombination between *S. mutans* isolates was evident in the chunk count matrix, which showed many instances of donor genomes donating high numbers of chromosome chunks to recipient genomes in distant clusters (Additional file 2: Fig. S2). The three older high coverage ancient genomes (NEO137, 7,672 – 7,512 cal. BP; CGG101233, 1042 - 1222 cal. BP; VK63, 975 years BP) were found in closely related subclusters within one of the high level clusters (cluster D; Figure 2; Additional file 2: Fig. S2), whereas the most recent sample (NEO105, 516 - 334 cal. BP) was found in a different cluster (cluster A; Figure 2; Additional file 2: Fig. S2).

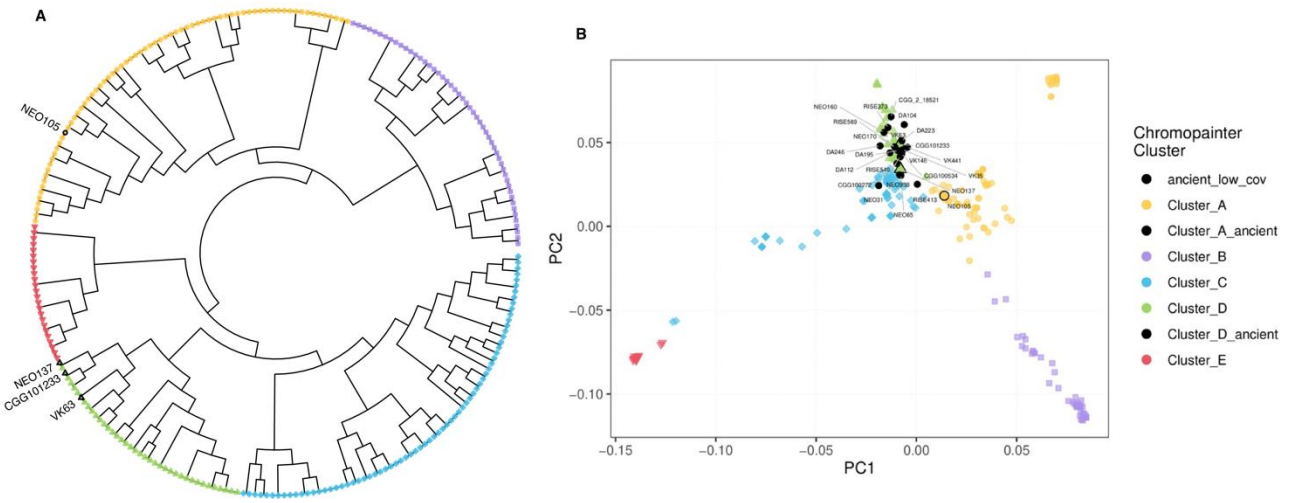

Figure 2. **Genetic structure of ancient and modern *S. mutans.* (A)** Dendrogram showing fineSTRUCTURE clustering hierarchy inferred from ChromoPainter haplotype chunk count sharing matrix for modern and high coverage ancient *S. mutans* genomes. **(B)** Principal component analysis of modern and ancient *S. mutans* based on SNP genotype covariance matrix. Cluster membership for modern and high coverage ancient genomes is indicated by symbol color and shape. The three ancient samples included in the tree are marked as triangles, whereas NEO105 is marked as a dot. The placement of the lower coverage ancient samples in the PC space are indicated simply as black dots.

We next carried out principal component analysis (PCA) based on the matrix of pairwise genotype covariances to place the lower coverage ancient genomes within the genetic diversity of the modern and high coverage ancient genomes. Interestingly, we found that most ancient genomes fell within a tight area in principal component space, close to the three older ancient high coverage ancient genomes (cluster D; Fig 2B; Additional file 2: Fig. S2).

Phylogenetics

To further investigate the evolutionary history of *S. mutans* lineages, we carried out maximum-likelihood phylogenetic reconstruction of the core genome alignments of the modern and high

coverage ancient genomes. The topology of the resulting phylogeny reflected the population structure obtained from ChromoPainter, with the high level clusters broadly corresponding to major clades (Additional file 2: Fig. S3). The branching among the deep clades was generally poorly resolved with low bootstrap support (Additional file 1: Supplemental information B), consistent with the effects of recombination obscuring deeper evolutionary relationships as seen in other recombining *Streptococcus* species such as *S. pyogenes*[67].

To account for the effects of recombination, we inferred patterns of recombination along the genome and reconstructed recombination-masked phylogenies using Gubbins. Consistent with the results from ChromoPainter, we found widespread evidence for recombination across the core genome (Additional file 2: Fig. S4), which was inferred to occur across 99.9% of the alignment (genome-wide r/m = 0.2502). While bootstrap support of the deep branches was slightly improved in the masked phylogeny, the topology was still characterized by an almost star-like phylogeny of early lineages radiating in poorly resolved order from the last common ancestor (Additional file 2: Fig. S5, Additional file 1: Supplemental information C).

Phylogenetic placement of the assembly of the sister species *S. troglodytae* was uncertain, but its Maximum Likelihood position was nonetheless deemed the most objective measure with which to root the final tree (Additional file 2: Fig. S6 A-U, Additional file 1: Supplemental information C). Lower coverage ancient genomes showed that most were placed within a clade of modern and high coverage ancient genomes corresponding to cluster A inferred from ChromoPainter (Figure 3; Additional file 2: Fig. S6 A-U). We observed no evident geographic structure among the isolates in the recombination-masked alignment, nor did source of isolation correlate with the topology (Figure 3; Additional file 2: Fig. S4).

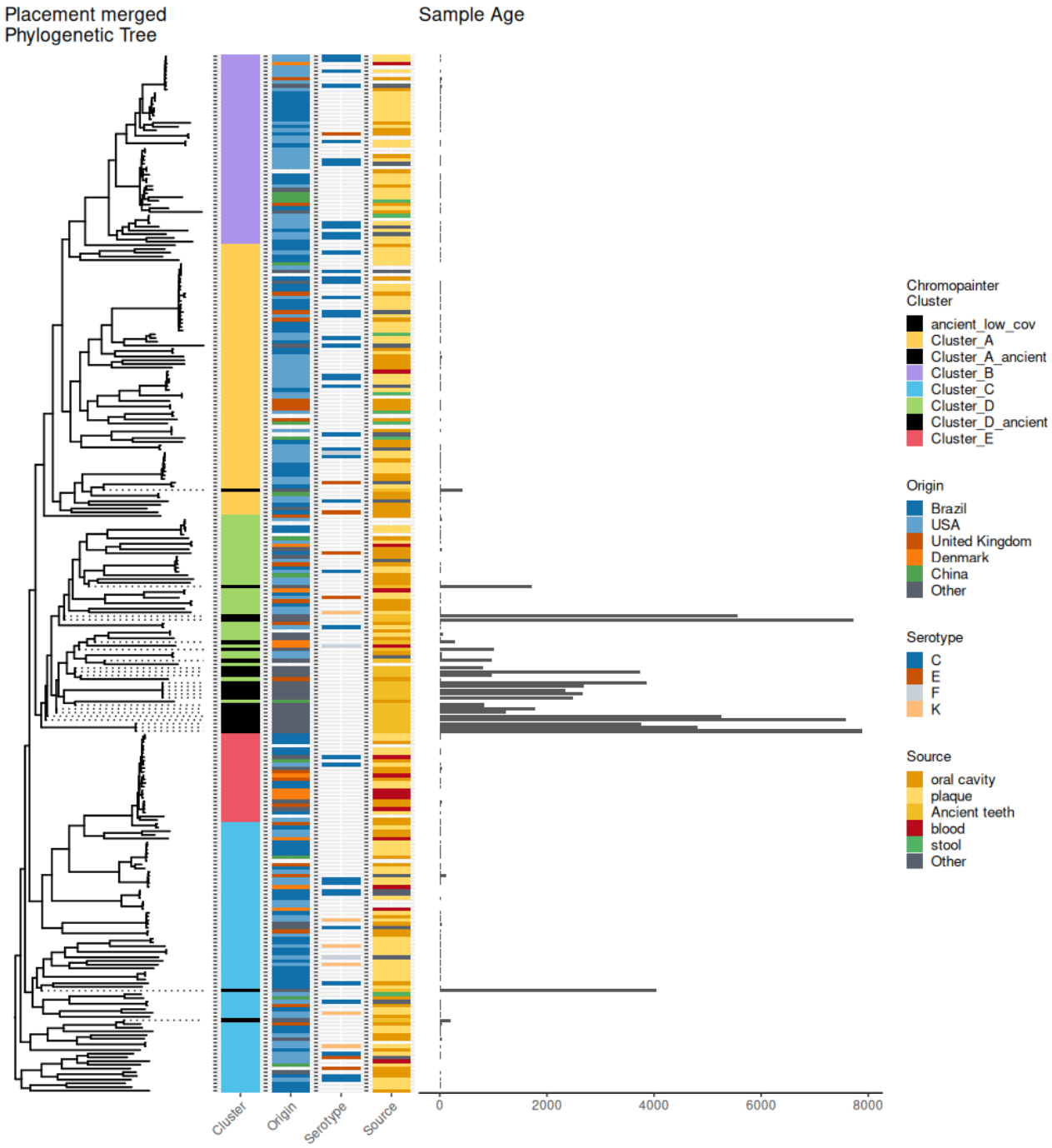

Figure 3. Annotated phylogeny of *S. mutans*. Maximum-likelihood phylogenetic tree inferred using RAxML-ng from recombination masked-alignment of modern and high coverage ancient genomes. Lower coverage ancient genomes as well as the *S. troglodytae* assembly (GCF_002355215.1), which was used to root the tree, were added to the tree using phylogenetic placement with EPA-ng (Additional file 2: Fig. S6 A-V). Annotation columns show different metadata of samples as indicated in legend. Ancient samples are represented as dotted lines with branch lengths artificially set for ease of comparison with these columns.

Genetic adaptations for oxidative stress resistance

We next aimed at identifying putative targets of recent positive selection in *S. mutans.* In a scenario of adaptation through newly acquired genes, those that confer a strong selective advantage for *S. mutans* populations are expected to spread rapidly, leading to a characteristic signature of high frequency combined with low genetic diversity among lineages. Having identified such genes, we can then use their occurrence among the ancient genomes to determine a lower bound for the time when they were acquired by *S. mutans* populations during their evolutionary history.

To scan for genes that were putative targets of recent positive selection, we quantified the average pairwise difference (i. e. nucleotide diversity π) and frequency of each gene annotated in the reference assembly (GCF_009738105.1). We identified several candidate genes that were outliers with low genetic diversity, the most striking example of which were genes which are part of the mutanobactin A (*mub)* operon (Figure 4) involved in oxidative stress resistance[68]. We found that a region containing 13 genes (*mubT – mubP*) with low genetic diversity ($\pi < 10^{-3}$) was observed at high frequency (~100 modern isolates) within a clade of genomes from ChromoPainter clusters A and B, but virtually absent from genomes in other clusters and ancient genomes (Figure 4). The notable exception was sample RISE540 (Arano, Italy; 4,148 – 3,927 cal. BP), one of only two ancient genomes placed within ChromoPainter cluster C outside the main ancient genome cluster D, which showed evidence for reads mapping across all genes of the operon (Figure 4). The genes *arcC aguA* and *ptcA* associated with acid tolerance[69] were likewise found in around 50% of modern isolates but were largely absent from the ancient *S. mutans* cluster (Figure 4).

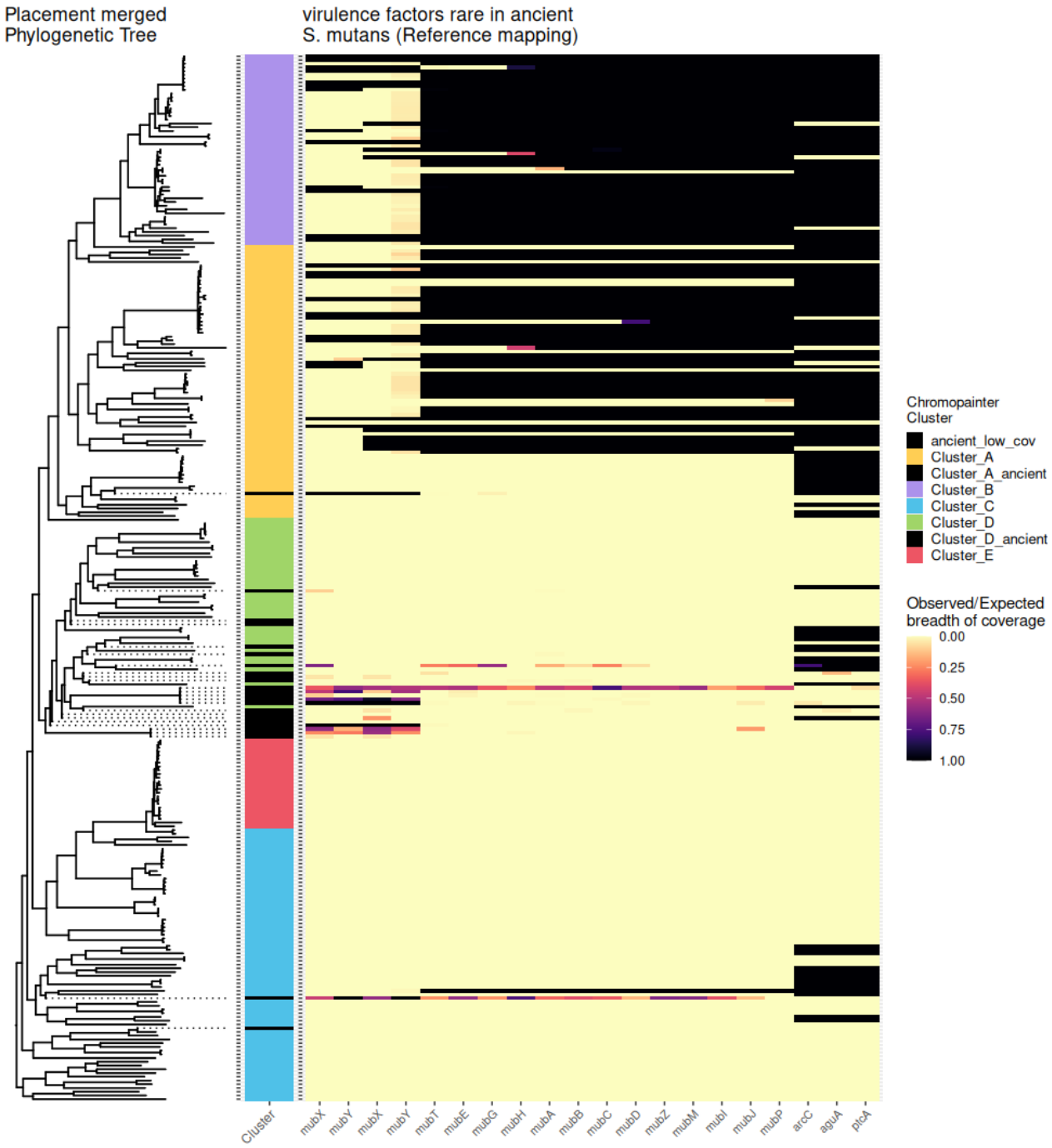

Figure 4. Gene presence in the mub operon. Heatmap showing gene presence (indicated using fraction of the gene covered) of *mub* operon genes and genes *arcC aguA* and *ptcA* across modern and ancient *S. mutans*. Rows of heatmap are ordered based on the recombination masked phylogeny, shown in the margin of the heatmap. Color bar indicates high level fineSTRUCTURE cluster membership, with ancient genomes shown in black. Ancient samples are represented as dotted lines with branch lengths artificially set for ease of comparison with these columns.

Ancient S. mutans pan-genomics

To characterize the full diversity of the pan-genome of our high coverage ancient *S. mutans* genomes outside of what could be found in our reference genome (GCF_009738105.1), we constructed MAGs and compared their pangenome with the modern isolates. We first determined whether metagenomic assembly was feasible for these samples by investigating how unique the metagenomic reads mapped to *S. mutans* compared to a panel of 115 reference genomes of other *Streptococcus* species (Additional file 2: Fig. S7). We found that read mapping rates were generally low for reference genomes of species other than *S. mutans* and its close relative *S. troglodytae*, indicating that few of the *Streptococcus* sequencing reads in the samples originate from other species (Additional file 2: Fig. S8).

Among the assembly methods tested, we achieved consistently the highest bin quality using the "A" settings for MEGAHIT and CONCOCT as a binning tool, as well as a 1kb contig length threshold (Additional file 3: Table 2, Additional file 2: Fig. S9 and S10).

After manual refinement of contigs in the *S. mutans* bins based on comparative gene uniqueness, local read damage estimations, and BLASTp identity with possible contaminants of suspect contigs (Additional file 2: Fig. S11), there were seven MAGs with estimated >80% completion and <5% redundancy according to anvi'o available for downstream pangenomic analysis. NEO105 only passed the gunc contamination check after 25 contigs were filtered out using mmseqs2 taxonomy (Additional file 3: Table 3, Additional file 1: Supplemental information A).

Three of our MAGs were of very high quality according to anvi'o (completion=100%, redundancy=0%, N50>20k, num_contigs<100) NEO105, NEO137 and CGG101233 (Additional file 3: Table 1, 4 and Additional file 2: Supplemental information D).

We compared our ancient MAGs with the 10 most phylogenetically distinct modern *S. mutans* isolates, and 15 representative *Streptococcus* species from across the genus in a pangenomics analysis focusing on metabolic pathways. Using KEGG metabolic modules detected through anvi'o among the gene clusters, we found six modules that differed in enrichment between the samples. For these six modules, the three ancient MAGs CGG101233, NEO105, NEO137 had the same enrichment profile as the modern samples, whereas the other ancient MAGs had lower abundances (Additional file 2: Fig. S12). This lower abundance was likely due to missing or partial genes rather than biological differences, since these MAGs were the ones of the lower quality (Additional file 3: Table 4).

We could not find any virulence factors from our reference mapping which were systematically absent from our 25 ancient genomes compared to the modern isolates beyond the *mub* operon, *arcC*, *aguA* and *ptcA* (Figure 4). The partial absence of the *comY* operon however in ancient *S. mutans*, could suggest a lower competency for internalizing DNA from other organisms (Additional file 2: Fig. S13)[70].

Our ancient *S. mutans* MAGs were then compared with the modern *S. mutans* isolates using gene annotations found through BAKTA, where we focused on the genes which could not be found through reference mapping (Additional file 2: Fig. S14). Among these genes, few were rare among the modern isolates, and those that were unique to the MAGs, were then usually only present in one of them. These rare cases included genes coding for bacteriocins *bacA*, *bhtE* and *pksJ*, each found among the oldest MAGs (Additional file 2: Fig. S15).

We sorted out genes which had previously been described as bacteriocins in *S. mutans*[7], which were not present in the reference genome GCF_009738105.1 (Additional file 3: Table S2). This gave a list of 17 virulence genes, all of which could be found in at least one of our ancient MAGs (Additional file 2: Fig. S16)[7]. This included the eponymous antimicrobial mutacin genes

specifically produced by some *S. mutans*[7] (NEO105) and the Ess/Type VII secretion system (CGG100272), previously described as a virulence factor among *Staphylococcus aureus*[71] (Additional file 2: Fig. S16).

Discussion

Ancient genomics provides a unique data source to directly survey the past genetic diversity of human pathogens[72]. In this study, we applied these tools to carry out the first large-scale genetic characterization of the human oral pathobiont *Streptococcus mutans* recovered from ancient human remains. Until now, only a single ancient *S. mutans* genome, recovered from ~4,000-year-old teeth from Ireland, has been reported[20]. Our 25 newly reported samples substantially expands the number of ancient *S. mutans* genomes, which contradicts the notion that it can only be rarely reported in ancient genomic datasets[73]. Furthermore, we observed exceptional DNA preservation and very high abundance of bacterial DNA in some of our samples. The most striking case was a ~1000 year old ancient tooth sample from Afghanistan, which yielded an ultra-deep sequenced genome with average read depth of 387X, using a metagenomic shotgun-sequencing approach without genomic enrichment. Together with three other genomes with read depth >20X, our results thus point to exceptionally high bacterial load of *S. mutans* in the oral microbiotas for some individuals in ancient human populations (See Additional file 1: Supplemental information D).

The ancient *S. mutans* genomes spanned a wide spatiotemporal range. The earliest genomes date to ~9,000 years ago and were recovered in ancient teeth from individuals associated with hunter-gatherer archaeological context. Moreover, the two earliest cases were found in distant geographic regions, ranging from northern Spain in Southern Europe to the Lake Baikal region in Siberia.

These results demonstrate that *S. mutans* already formed part of the oral microbial community of human populations before the onset of agriculture and was already widespread across Eurasia by ~9,000 years ago. The dates of our earliest samples are compatible with a previously inferred rapid expansion of *S. mutans* populations around ~10,000 years ago[10]. However, their wide geographic distribution and occurrence in archaeological contexts of hunter-gatherer populations question whether the transition to agriculture was indeed a major factor in this expansion, as previously suggested[10].

Our analyses of the genetic structure of *S. mutans* also support previous results of recombination being a major force in shaping the genetic diversity of the species (Additional file 2: Fig. S4, Additional file 1: Supplemental information B)[10]. These high rates of horizontal gene transfer likely contribute to the ability of *S. mutans* to adapt to fast-changing environments, as well as explain the lack of association of phylogeny with sample metadata such as geographic origin or isolation source, and the difficulty in finding a robust temporal signal (Additional file 1: Supplemental information B and C). On the other hand, our results also clearly highlight a non-random placement of most ancient *S. mutans* genomes within a restricted part of the phylogeny of present-day isolate genomes studied to date, despite spanning ~8,000 years of *S. mutans* evolution.

Species with high rates of recombination such as observed in *S. mutans* facilitate genetic adaptations through individual genes or functional elements proliferating in the population through selective sweeps. Consistent with this prediction, we find strong evidence for a recent selective sweep involving the *mub* operon in our dataset, as the low genetic diversity and high frequency we observed among the modern isolates suggests that it spread rapidly through a subset of populations (genomes from fineSTRUCTURE clusters A and B; Figure 4). The fact that it was absent in almost all our ancient genomes also supports a recent spread. As the *mub* operon provides *S. mutans*

tolerance to oxidative stress, we hypothesize that recent changes in dental hygiene such as tooth brushing, resulting in more frequent exposure to more aerobic environments could be a driving factor in this adaptation. It will also have made *S. mutans* more resistant to peroxide producing commensals.

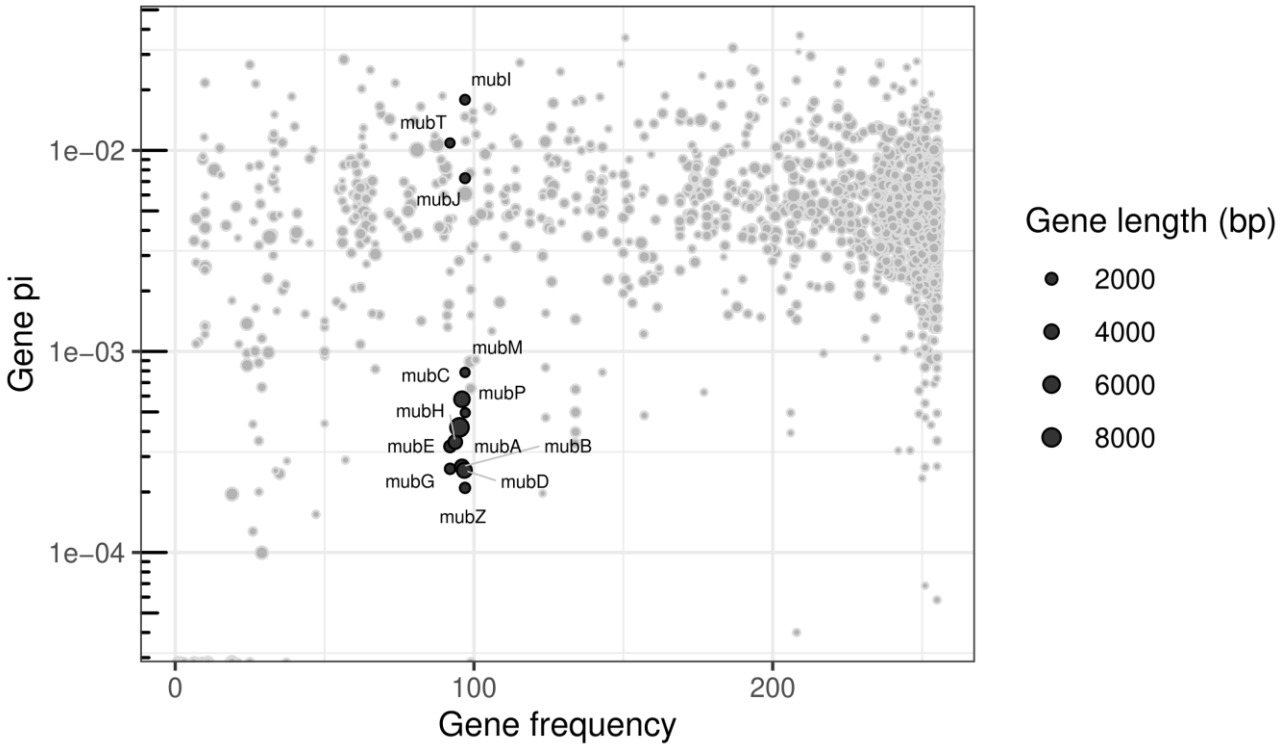

Figure 5. **Gene-level nucleotide diversity.** Plot showing average nucleotide diversity (π) for 1,913 genes annotated in the *S. mutans* reference assembly GCF_009738105.1 as a function of their frequency observed among the 255 modern isolates. Genes of the *mub* operon are indicated with black symbols and labelled. Plot symbol size indicates length in base pairs for a particular gene.

Nevertheless, a single ancient genome with evidence for its presence was identified, from a ~4,000 years old sample in Italy, thus demonstrating that it had already been acquired by *S. mutans* by that time in Europe. We thus do not contest that additional data might be able to nuance our hypothesis about aerobic adaptation in *S. mutans*. The fact that we only have few samples representing broad periods of human history also limits us to exploring the presence of, but not the frequency of S. mutans in general or its genes in particular.

Streptococcus mutans is a pathogen that, as far as we can elucidate, has retained the core genes that define its pathogenicity. This should not preclude necessarily, that there hasn't occurred a significant change in its phenotype as previously described in other oral bacterial species, such as *Aggregatibacter Actinomycetemcomitans,* where a single SNP change resulted in a 20-fold increase in pathogenicity, due to increased leukotoxin production[74]. Importantly, such subtle changes are difficult to trace in ancient pathogen genomes, as most do not have deep coverage, and for those that do, we lack clinical metadata. Nonetheless, we do not find structure in the phylogeny which would suggest such a development happening among the ancient samples. The ancient *S. mutans* samples do not cluster themselves in any way as to show such a development, which has been seen with historically virulent pathogens such as *Yersinia pestis*, which separate into distinct, age dependent clusters, compared with modern samples[75]. The lack of such a temporal or geographical structure in the *S. mutans* phylogeny, suggests that the story of this bacteria is different than the one we see for *A. Actinomycetemcomitans* or *Y. Pestis*, whose pathogenicity demonstrably has waxed and waned through periodic outbreaks, triggered by changes in genetic makeup or human conditions.

Overall, we did not find compelling genetic evidence to support that *S. mutans* has evolved to the changes in human diet in the last 8,000 years. This challenges the hypothesis that *S. mutans* has either emerged as, or evolved into, an agent directly responsible for the development of caries, because of the post-industrial diet of carbohydrates. Since we can find *S. mutans* across diets and lifestyles, and since archaeological and anthropological evidence suggests that caries was a rare disease among people, who did not actively mitigate the development of a rich biofilm, the overriding factors that determine the development of caries likely relate to a modern lifestyle with frequent carbohydrate intake. Our study therefore supports a discourse of caries being a multifactorial disease with many necessary etiological factors[4]. Primarily a carbohydrate rich diet

that can be capitalized by resident acidogenic bacteria such as *S. mutans*, which, if given enough time to anchor themselves in a mature biofilm on the tooth, creates a caries lesion[76].

Conclusions

Our results demonstrate the utility of ancient pathogen genomics to elucidate the evolutionary history of a major human pathobiont, and how we can use the perspective of the past to inform how to combat the rise of caries as an endemic disease in the future.

Ethics approval and consent to participate:

Not applicable

Consent for publication

Not applicable

Competing interests

The authors declare that they have no competing interests

Funding

The Lundbeck Foundation GeoGenetics Centre is supported by the Lundbeck Foundation (grant nos. R302-2018-2155, R155-2013-16338), the Novo Nordisk Foundation (grant no. NNF18SA0035006), the Wellcome Trust (grant no. UNS69906), Carlsberg Foundation (grant no. CF18-0024), the Danish National Research Foundation (grant nos. DNRF94, DNRF174) and Ferring Pharmaceuticals A/S. The study was funded by an internal PhD grant from the University of Copenhagen

Authors' contributions

M.S. and D.B. conceptualized the study. V.T., M.S., F.C., S.H.N. and J.A were involved in data analysis, method development and implementation. M.F.F., M.J, G.E. and J.B.S. were involved in data generation and M.F.F., G.S.. F.D. and K.K. curated bioarchaeological data. D.B. M.S. F.C. and M.G. supervised the research. V.T. D.B. M.S. F.C. and M.G. wrote the first draft of the paper and

were involved in reviewing drafts and editing. V.T., D.B., M.S. and M.F.F. wrote the supplemental information. All authors read and approved the final manuscript.

Availability of Data and Materials:

Data for the three samples for whom sequencing data as trimmed read files (FASTQ) are released in this study (Additional file 1: Supplemental information E) have been deposited at the European Nucleotide Archive under accession no. PRJEB105110. *S. mutans* read alignments (BAM) for all 25 samples and the 7 MAGs (FASTA) have also been released under the same accession no.. Sequencing read data for the 22 samples had been previously published in Sikora et al. 2025[25] or in Allentoft et al. 2024[77] and are available under accession no. PRJEB65256 and PRJEB64656 respectively. The rest of the data sets used in the study were obtained from public sources[50,54,56,63,64]. Custom scripts and workflows used in this manuscript are available in Zenodo[78] and is licensed under a Creative Commons Attribution 4.0 International License.

Acknowledgements

Thank you to Nicolas Delgado Clausen and Signe Klemm for taking high definition pictures of CGG101233 as well as Frederik Valeur Seersholm for help with data upload to ENA. We would also like to thank Eske Willerslev for funding and overall project support.

Peer review information

Andrew Cosgrove and Claudia Feng were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Additional file information:

Additional file 1: Supplemental information A-E on selected topics

Additional file 2: Supplemental figures 1-26

Addtional file 3: Tables 1-4 and supplemental tables 1-3

References

1. WHO. Global oral health status report. 2022.

2. Kassebaum NJ, Bernabé E, Dahiya M, Bhandari B, Murray CJL, Marcenes W. Global burden of untreated caries: a systematic review and metaregression. J Dent Res. 2015;94(5):650–8.

3. Nyvad B, Crielaard W, Mira A, Takahashi N, Beighton D. Dental caries from a molecular microbiological perspective. Caries Res. 2013;47(2):89–102.

4. Pitts NB, Zero DT, Marsh PD, Ekstrand K, Weintraub JA, Ramos-Gomez F. Dental caries. Nat Rev Dis Primers. 2017;3:17030.

5. Schwendicke F, Frencken JE, Bjørndal L, Maltz M, Manton DJ, Ricketts. Managing Carious Lesions: Consensus Recommendations on Carious Tissue Removal. Adv Dent Res. 2016;28(2):58–67.

6. Takahashi N, Nyvad B. The role of bacteria in the caries process: ecological perspectives. J Dent Res. 2011;90(3):294–303.

7. Lemos JA, Palmer SR, Zeng L, Wen ZT, Kajfasz JK, Freires IA, Abranches J, Brady LJ. The Biology of Streptococcus mutans. Microbiol Spectr. 2019 Jan;7(1):10.1128/microbiolspec.GPP3-0051-2018.

8. Loesche WJ. Role of Streptococcus mutans in human dental decay. Microbiol Rev. 1986 Dec;50(4):353-80.

9. Lemos JA, Burne RA. A model of efficiency: stress tolerance by Streptococcus mutans. Microbiology (Reading). 2008 Nov;154(Pt 11):3247-3255.

10. Cornejo OE, Lefébure T, Bitar PD, Lang P, Richards VP, Eilertson K, Do T, Beighton D, Zeng L, Ahn SJ, Burne RA, Siepel A, Bustamante CD, Stanhope MJ. Evolutionary and population genomics of the cavity causing bacteria Streptococcus mutans. Mol Biol Evol. 2013 Apr;30(4):881-93

11. Palmer SR, Miller JH, Abranches J, Zeng L, Lefebure T, Richards VP, Lemos JA, Stanhope MJ, Burne RA. Phenotypic heterogeneity of genomically-diverse isolates of Streptococcus mutans. PLoS One. 2013 Apr 16;8(4):e61358.

12. Metcalf JL, Ursell LK, Knight R. Ancient human oral plaque preserves a wealth of biological data. Nat Genet. 2014;46(4):321-3.

13. Adler CJ et al. Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions. Nat Genet. 2013;45(4):450-5.

14. Weyrich LS et al. Neanderthal behaviour, diet, and disease inferred from ancient DNA In dental calculus. Nature. 2017;544(7650):357-361.

15. Fellows Yates JA et al. The evolution and changing ecology of the African hominid oral microbiome. Proc Natl Acad Sci U S A. 2021;118(20):e2021655118.

16. Achtman M, Zhou Z. Metagenomics of the modern and historical human oral microbiome with phylogenetic studies on Streptococcus mutans and Streptococcus sobrinus. Philos Trans R Soc Lond B Biol Sci. 2020 Nov 23;375(1812):20190573.

17. Scorrano G, Nielsen SH, Vetro DL, Sawafuji R, Mackie M, Margaryan A, Fotakis AK, Martínez-Labarga C, Fabbri PF, Allentoft ME, Carra M, Martini F, Rickards O, Olsen JV, Pedersen MW, Cappellini E, Sikora M. Genomic ancestry, diet and microbiomes of Upper Palaeolithic hunter-gatherers from San Teodoro cave. Commun Biol. 2022 Nov 18;5(1):1262.

18. Jensen TZT, Niemann J, Iversen KH, Fotakis AK, Gopalakrishnan S, Vågene ÅJ, Pedersen MW, Sinding MS, Ellegaard MR, Allentoft ME, Lanigan LT, Taurozzi AJ, Nielsen SH, Dee MW, Mortensen MN, Christensen MC, Sørensen SA, Collins MJ, Gilbert MTP, Sikora M, Rasmussen S, Schroeder H. A 5700 year-old human genome and oral microbiome from chewed birch pitch. Nat Commun. 2019 Dec 17;10(1):5520.

19. Lanfranco LP, Eggers S. The usefulness of caries frequency, depth, and location in determining cariogenicity and past subsistence: a test on early and later agriculturalists from the Peruvian coast. Am J Phys Anthropol. 2010 Sep;143(1):75-91.

20. Jackson, I., Woodman, P., Dowd, M., Fibiger, L., & Cassidy, L. M. (2024). Ancient Genomes From Bronze Age Remains Reveal Deep Diversity and Recent Adaptive Episodes for Human Oral Pathobionts. *Molecular Biology and Evolution*, *41*(3). https://doi.org/10.1093/molbev/msae017

21. Rascovan N et al. Emergence and Spread of Basal Lineages of Yersinia pestis during the Neolithic Decline. Cell. 2019;176(1-2):295-305.e10.

22. Key FM et al. Emergence of human-adapted Salmonella enterica is linked to the Neolithization process. Nat Ecol Evol. 2020 Mar;4(3):324-333.

23. Allentoft ME et al. Population genomics of Bronze Age Eurasia. Nature. 2015;522(7555):167-72.

24. Damgaard PB et al. 137 ancient human genomes from across the Eurasian steppes. Nature. 2018;557(7705):369-374.

25. Sikora, M., Canteri, E., Fernandez-Guerra, A. *et al.* The spatiotemporal distribution of human pathogens in ancient Eurasia. *Nature* **643**, 1011–1019 (2025). https://doi.org/10.1038/s41586-025-09192-8

26. Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods*, *9*(4), 357–359. https://doi.org/10.1038/nmeth.1923

27. Danecek, P. et al. Twelve years of SAMtools and BCFtools. GigaScience 10. giab008. issn: 2047-217X. eprint: https://academic.oup.com/gigascience/article-pdf/10/2/giab008/36332246/giab008.pdf. https://doi.org/10.1093/gigascience/giab008 (Feb. 2021)

28. Quinlan AR and Hall IM, 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26, 6, pp. 841–842.

29. "Picard Toolkit." 2019. Broad Institute, GitHub Repository. https://broadinstitute.github.io/picard/; Broad Institute

30. Michelsen, C. et al. metaDMG – A Fast and Accurate Ancient DNA Damage Toolkit for Metagenomic Data. bioRxiv. https://www.biorxiv.org/content/early/2022/12/09/2022.12.06.519264 (2022)

31. Morgan M, Pagès H, Obenchain V, Hayden N (2024). *Rsamtools: Binary alignment (BAM), FASTA, variant call (BCF), and tabix file import*. R package version 2.20.0, https://bioconductor.org/packages/Rsamtools.

32 .McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. PLOS ONE8,e61217.issn:1932-6203.h

ttps://journals.plos.org/plosone/article?id=10.1371/journal.pone.0061217 (Apr. 2013)

Phylogenetics

33. Pedersen, B. S., & Quinlan, A. R. (2018). Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, *34*(5), 867–868. https://doi.org/10.1093/BIOINFORMATICS/BTX699

34. Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/BIOINFORMATICS/BTY191

35. "SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments", Andrew J. Page, Ben Taylor, Aidan J. Delaney, Jorge Soares, Torsten Seemann, Jacqueline A. Keane, Simon R. Harris, Microbial Genomics 2(4), (2016)

36. Chang CC, Chow CC, Tellier LCAM, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. GigaScience, 4.

37. Yang et al. (2011) GCTA: a tool for Genome-wide Complex Trait Analysis. Am J Hum Genet. 88(1): 76-82. [PubMed ID: 21167468]

38. Lawson, D. J., Hellenthal, G., Myers, S. & Falush, D. Inference of population structure using dense haplotype data. *PLoS Genet.* **8**, e1002453 (2012).

39. Stamatakis A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9): 1312-1313. doi:10.1093/bioinformatics/btu033

40. Pierre Barbera, Alexey M Kozlov, Lucas Czech, Benoit Morel, Diego Darriba, Tomáš Flouri, Alexandros Stamatakis; EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences, Systematic Biology, syy054, https://doi.org/10.1093/sysbio/syy054

41. Genesis and Gappa: processing, analyzing and visualizing phylogenetic (placement) data. Lucas Czech, Pierre Barbera, and Alexandros Stamatakis. Bioinformatics, 2020. https://doi.org/10.1093/bioinformatics/btaa070

42. Yu G (2022). *Data Integration, Manipulation and Visualization of Phylogenetic Treess*, 1st edition edition. Chapman and Hall/CRC. doi:10.1201/9781003279242, https://www.amazon.com/Integration-Manipulation-Visualization-Phylogenetic-Computational-ebook/dp/B0B5NLZR1Z/.

43. Paradis E, Schliep K (2019). “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.” _Bioinformatics_, *35*, 526-528. doi:10.1093/bioinformatics/bty633 <https://doi.org/10.1093/bioinformatics/bty633>.

44. Yu G (2023). _aplot: Decorate a 'ggplot' with Associated Information_. R package version 0.1.10, <https://CRAN.R-project.org/package=aplot>.

45. Yu G (2022). *Data Integration, Manipulation and Visualization of Phylogenetic Treess*, 1st edition edition. Chapman and Hall/CRC. https://www.amazon.com/Integration-Manipulation-Visualization-Phylogenetic-Computational-ebook/dp/B0B5NLZR1Z/.

46. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications 2018 9:1*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-07641-9

47. Gabadinho, A., Ritschard, G., Müller, N.S. & Studer, M. (2011), Analyzing and visualizing state sequences in R with TraMineR, *Journal of Statistical Software*. Vol. 40(4), pp. 1-37.

48. R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

49. KrakenUniq: confident and fast metagenomics classification using unique k-mer counts. Breitwieser FP, Baker DN, Salzberg SL. Genome Biology, Dec 2018. https://doi.org/10.1186/s13059-018-1568-0

50. Sikora, M., Iversen, AKN., et al. The spatiotemporal distribution of human pathogens in ancient Eurasia supplementary datasets. 2025. Data Sets. ERDA. https://doi.org/10.17894/ucph.f0f75211-7bc3-445d-90c0-b72a22ba0930. Analyzed in 2023 prior to publication

51. Li, D., Liu, C-M., Luo, R., Sadakane, K., and Lam, T-W., (2015) MEGAHIT: An ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, doi: 10.1093/bioinformatics/btv033 [PMID: 25609793].

52. Eren, A. M., Kiefl, E., Shaiber, A., Veseli, I., Miller, S. E., Schechter, M. S., Fink, I., Pan, J. N., Yousef, M., Fogarty, E. C., Trigodet, F., Watson, A. R., Esen, Ö. C., Moore, R. M., Clayssen, Q., Lee, M. D., Kivenson, V., Graham, E. D., Merrill, B. D., … Willis, A. D. (2020). Community-led, integrated, reproducible multi-omics with anvi'o. *Nature Microbiology 2020 6:1*, *6*(1), 3–6. https://doi.org/10.1038/s41564-020-00834-3

53. Simon C Potter, Aurélien Luciani, Sean R Eddy, Youngmi Park, Rodrigo Lopez, Robert D Finn, HMMER web server: 2018 update, *Nucleic Acids Research*, Volume 46, Issue W1, 2 July 2018, Pages W200–W204, https://doi.org/10.1093/nar/gky448

54. Galperin MY, Makarova KS, Wolf YI, Koonin EV. Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. 2015 Jan;43(Database issue):D261-9. doi: 10.1093/nar/gku1223. Epub 2014 Nov 26. PMID: 25428365; PMCID: PMC4383993.

55. Buchfink, B., Reuter, K. & Drost, HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods 18, 366–368 (2021). https://doi.org/10.1038/s41592-021-01101-x

56. Donovan H Parks, Pierre-Alain Chaumeil, Aaron J Mussig, Christian Rinke, Maria Chuvochina, Philip Hugenholtz, GTDB release 10: a complete and systematic taxonomy for 715 230 bacterial

and 17 245 archaeal genomes, *Nucleic Acids Research*, Volume 54, Issue D1, 6 January 2026, Pages D743–D754, https://doi.org/10.1093/nar/gkaf1040

57. Johannes Alneberg, Brynjar Smári Bjarnason, Ino de Bruijn, Melanie Schirmer, Joshua Quick, Umer Z Ijaz, Leo Lahti, Nicholas J Loman, Anders F Andersson & Christopher Quince. 2014. Binning metagenomic contigs by coverage and composition. *Nature Methods*, doi: 10.1038/nmeth.3103

58. Kang, D. D., Li, F., Kirton, E., Thomas, A., Egan, R., An, H., & Wang, Z. (2019). MetaBAT 2: An adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*, *2019*(7). https://doi.org/10.7717/PEERJ.7359/SUPP-3

59. Delmont TO, Eren AM. 2018. Linking pangenomes and metagenomes: the Prochlorococcus metapangenome. PeerJ 6:e4320 https://doi.org/10.7717/peerj.4320

60. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3:e1319 https://doi.org/10.7717/peerj.1319

61. Orakov, A., Fullam, A., Coelho, L. P., Khedkar, S., Szklarczyk, D., Mende, D. R., Schmidt, T. S. B., & Bork, P. (2021). GUNC: detection of chimerism and contamination in prokaryotic genomes. *Genome Biology*, *22*(1), 1–19. https://doi.org/10.1186/S13059-021-02393-0/FIGURES/3

62. Steinegger M and Soeding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology, doi: 10.1038/nbt.3988 (2017).

63. Schwengers O., Jelonek L., Dieckmann M. A., Beyvers S., Blom J., Goesmann A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microbial Genomics, 7(11). https://doi.org/10.1099/mgen.0.000685

64. Takuya Aramaki, Romain Blanc-Mathieu, Hisashi Endo, Koichi Ohkubo, Minoru Kanehisa, Susumu Goto, Hiroyuki Ogata, KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold, Bioinformatics, Volume 36, Issue 7, April 2020, Pages 2251–2252, https://doi.org/10.1093/bioinformatics/btz859

65. Shaiber, A., Willis, A.D., Delmont, T.O. *et al.* Functional and genetic markers of niche partitioning among enigmatic members of the human oral microbiome. *Genome Biol* **21**, 292 (2020). https://doi.org/10.1186/s13059-020-02195-w

66.Warnes G, Bolker B, Bonebakker L, Gentleman R, Huber W, Liaw A, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B (2022). _gplots: Various R Programming Tools for Plotting Data_. R package version 3.1.3, <https://CRAN.R-project.org/package=gplots>.

67. Davies, M. R., McIntyre, L., Mutreja, A., Lacey, J. A., Lees, J. A., Towers, R. J., Duchêne, S., Smeesters, P. R., Frost, H. R., Price, D. J., Holden, M. T. G., David, S., Giffard, P. M., Worthing, K. A., Seale, A. C., Berkley, J. A., Harris, S. R., Rivera-Hernandez, T., Berking, O., … Walker, M. J. (2019). Atlas of group A streptococcal vaccine candidates compiled using large-scale comparative genomics. *Nature Genetics 2019 51:6*, *51*(6), 1035–1043. https://doi.org/10.1038/s41588-019-0417-8

68. Wu, C., Cichewicz, R., Li, Y., Liu, J., Roe, B., Ferretti, J., Merritt, J., & Qi, F. (2010). Genomic Island TnSmu2 of Streptococcus mutans Harbors a Nonribosomal Peptide Synthetase-Polyketide Synthase Gene Cluster Responsible for the Biosynthesis of Pigments Involved in Oxygen and H2O2 Tolerance. *Applied and Environmental Microbiology*, *76*(17), 5815. https://doi.org/10.1128/AEM.03079-0

69. Griswold, A. R., Chen, Y. Y. M., & Burne, R. A. (2004). Analysis of an Agmatine Deiminase Gene Cluster in Streptococcus mutans UA159. *Journal of Bacteriology*, *186*(6), 1902. https://doi.org/10.1128/JB.186.6.1902-1904.2004

70. Merritt, J., Qi, F., & Shi, W. (2005). A unique nine-gene comY operon in Streptococcus mutans. *Microbiology*, *151*(1), 157–166. https://doi.org/10.1099/MIC.0.27554-0/CITE/REFWORKS

71. Warne, B., Harkins, C. P., Harris, S. R., Vatsiou, A., Stanley-Wall, N., Parkhill, J., Peacock, S. J., Palmer, T., & Holden, M. T. G. (2016). The Ess/Type VII secretion system of Staphylococcus aureus shows unexpected genetic diversity. *BMC Genomics*, *17*(1). https://doi.org/10.1186/S12864-016-2426-7

72. Spyrou, M.A., Bos, K.I., Herbig, A. *et al.* Ancient pathogen genomics as an emerging tool for infectious disease research. *Nat Rev Genet* **20**, 323–340 (2019). https://doi.org/10.1038/s41576-019-0119-1

73. Velsko, I. M., Fellows Yates, J. A., Aron, F., Hagan, R. W., Frantz, L. A. F., Loe, L., Martinez, J. B. R., Chaves, E., Gosden, C., Larson, G., & Warinner, C. (2019). Microbial differences between

dental plaque and historic dental calculus are related to oral biofilm maturation stage. *Microbiome*, *7*(1). https://doi.org/10.1186/S40168-019-0717-3

74. Haubek D. et al. Microevolution and Patterns of Dissemination of the JP2 Clone of Aggregatibacter (Actinobacillus) actinomycetemcomitans. Infect Immun. 2007 Jun; 75(6): 3080–3088.

75. Rasmussen S. et al. Early divergent strains of Yersinia pestis in Eurasia 5,000 years ago. Cell 2015 Oct 22;163(3):571-82. doi: 10.1016/j.cell.2015.10.009. Epub 2015 Oct 22.

76. Lara Jochum, Bärbel Stecher,Label or Concept – What Is a Pathobiont?,Trends in Microbiology,Volume 28, Issue 10,2020,Pages 789-792,ISSN 0966-842X,https://doi.org/10.1016/j.tim.2020.04.011.(https://www.sciencedirect.com/science/article/pii/S0966842X20301049)

77. Allentoft, ME., Sikora, M., Refoyo-Martínez, A. et al. Population genomics of post-glacial western Eurasia. *Nature* 625, 301–311 (2024). https://doi.org/10.1038/s41586-023-06865-0

78. Thygesen, VFDF., Belstrøm, D., Sikora, M. et al. Analysis of S. mutans in ancient tooth samples Zenodo. 2025. https://doi.org/10.5281/zenodo.17881924

79. Chen, L. X., Anantharaman, K., Shaiber, A., Murat Eren, A., & Banfield, J. F. (2020). Accurate and complete genomes from metagenomes. *Genome Research*, *30*(3), 315–333. https://doi.org/10.1101/GR.258640.119/-/DC1

80. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P., & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research, 25*(7), 1043–1055. https://doi.org/10.1101/GR.186072.114

81. Didelot et al (2018). Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Research, 46:e134.

82. Lee, I.P.A., Andam, C.P. Frequencies and characteristics of genome-wide recombination in *Streptococcus agalactiae*, *Streptococcus pyogenes*, and *Streptococcus suis*. *Sci Rep* **12**, 1515 (2022). https://doi.org/10.1038/s41598-022-04995-5

83. Paul Zaharias, Frédéric Lemoine, Olivier Gascuel, Robustness of Felsenstein's Versus Transfer Bootstrap Supports With Respect to Taxon Sampling, *Systematic Biology*, Volume 72, Issue 6, November 2023, Pages 1280–1295, https://doi.org/10.1093/sysbio/syad052

84. Robert Lanfear, Matthew W Hahn, The Meaning and Measure of Concordance Factors in Phylogenomics, *Molecular Biology and Evolution*, Volume 41, Issue 11, November 2024, msae214, https://doi.org/10.1093/molbev/msae214

85. Bui Quang Minh, Matthew W Hahn, Robert Lanfear, New Methods to Calculate Concordance Factors for Phylogenomic Datasets, *Molecular Biology and Evolution*, Volume 37, Issue 9, September 2020, Pages 2727–2733, https://doi.org/10.1093/molbev/msaa106

86. Yu K Mo, Robert Lanfear, Matthew W Hahn, Bui Quang Minh, Updated site concordance factors minimize effects of homoplasy and taxon sampling, *Bioinformatics*, Volume 39, Issue 1, January 2023, btac741, https://doi.org/10.1093/bioinformatics/btac741

87. Krzyściak W, Jurczak A, Kościelniak D, Bystrowska B, Skalniak A. The virulence of Streptococcus mutans and the ability to form biofilms. Eur J Clin Microbiol Infect Dis. 2014 Apr;33(4):499-515.

88. Aas JA, Griffen AL, Dardis SR, Lee AM, Olsen I, Dewhirst FE, Leys EJ, Paster BJ. Bacteria of dental caries in primary and permanent teeth in children and young adults. J Clin Microbiol. 2008 Apr;46(4):1407-17.

89. Kouchi Y, Ninomiya J, Yasuda H, Fukui K, Moriyama T, Okamoto H. Location of Streptococcus mutans in the dentinal tubules of open infected root canals. J Dent Res. 1980 Dec;59(12):2038-46.

90. Damgaard, P., Margaryan, A., Schroeder, H. *et al.* Improving access to endogenous DNA in ancient bones and teeth. *Sci Rep* 5, 11184 (2015). https://doi.org/10.1038/srep11184

91. M. Meyer, M. Kircher, Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harb. Protoc.* 10.1101/pdb.prot5448 (2010). doi:10.1101/pdb.prot5448 Medline

92. Gansauge MT, Gerber T, Glocke I, et al. Single-stranded DNA library preparation from highly degraded DNA using T4 DNA ligase. *Nucleic Acids Res*. 2017;45(10):e79. doi:10.1093/nar/gkx033

93. Bendezu-Sarmiento (in press)**,** Archaeological Survey and First Preliminary Results of the Site of Shahr-i Gholghola (Afghanistan) – the Bamiyan Valley as a Centre of Trade and Cultural Exchange, Second International Conference on Central Asian Archaeology held at the University of Bern, Switzerland, 13-15 February 2020.

94. Gardin J.-C., 1957, Poteries de Bamiyan, in *Ars Orientalis*, vol. 2, p. 227-245.

95. Ebrahimi, Gh., 2021. Kültepe Sareyn (Anahita) Strategraphic Excavation and Sounding Report. In Ruhollah Shirazi (Ed.). 18th Annual Symposium on the Iranian Archaeology. Tehran: Research Institute of Cultural Heritage & Tourism. Pp. 242-245. (In Persian)

96. Badirzadeh, A., Niyyati, M., Babaei, Z., Amini, H., Badirzadeh, H. and Rezaeian, M., 2011. Isolation of free-living amoebae from Sareyn hot springs in ardebil province, iran. Iranian journal of parasitology, 6(2), p.1-8.

97. Мансури-Фар, С., 2019. Geothermal regime and geochemistryof hot springs in Mount Sabalan (Northwest of Iran). Природные ресурсы, (2), pp.23-33.

98. Khani, B., Abedi, A., Eskandari, N., & Ebrahimi, G. (2024). Ritual Practices in the Kura-Araxes Culture: Hearths and Figurines as Markers of Religious Identity. *Journal of Archaeological Studies*, *16*(2), 229–267. https://doi.org/10.22059/jarcs.2025.387728.143323

99. Azarnosh, M. and Rezaloo, R., 2006. Reviewing the Chronology of Northwestern Iran in the Bronze Age, Case Study: Qalla Khosrow. The International Journal of Humanities, 13(3), pp.1-15.

100. Rezaloo, Reza 2007. The Emergence of Complex Society during Late Bronze Age in Southern Regions of Aras River: Case Study Qalla Khosrow. Ph.D. Thesis in Archaeology, Tehran: Tarbiat Modares University.

101. Pourfaraj, A., 2017. A Chronological Explanation of The First Millennium BC Cultures of Nir County, Ardebil. Journal of Archaeological Studies, 9(1), pp.37-54.

102. Hejebri Nobari, A., Khanali, H., Yilmaz, A. and Mosavi Kohpar, S.M., 2020. Archaeological analysis and investigations on Shaharyeri in Ardabil province. Journal of Archaeological Studies, 12(2), pp.267-283.

103. Dyson, R.H., 1999. The Achaemenid painted pottery of Hasanlu IIIA. Anatolian studies, 49, pp.101-110.