scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Genome resequencing and comparative analysis of *Streptococcus mutans* in adults with high and low caries risk

Merve Yildirim Ucuncu[1], Musa Kazim Ucuncu[2], Ilker Karacan[3,4] & Nursen Topcuoglu[5] ✉

***Streptococcus mutans*, is considered the main microbial etiological agent of dental caries, therefore it has been proposed as a useful predictor of caries risk as well as a target for caries prevention strategies. We aimed to compare the genomic characteristics of *S. mutans* strains isolated from individuals with high and low caries risk, in order to determine their genotypic features related to dental caries in adults. A total of 25 *S. mutans* isolates, obtained from the saliva of 13 volunteers with high dental caries activity and 12 caries-free individuals, were analysed using whole-genome sequencing techniques. A total of 2904 protein-coding gene sequences were detected as a result of the pan-genome analysis. The number of core genes detected in all genomes sequenced in the study was found to be 1563. A total of 50584 mutations were detected using ATCC 25175 strain as a reference. This is a large genome dataset of 25 *S. mutans* strains which can be further used for all *S. mutans* genome analysis.**

Background & Summary

Streptococcus mutans is known as one of the major contributors to dental caries due to its ability to synthesize extracellular polysaccharides, develop biofilm matrices in dental tissues, produce organic acids by metabolizing various carbohydrates, and survive in low pH conditions[1].

The primary virulence factors associated with the cariogenic potential of *S. mutans* can be described as adhesion, acid production (acidogenesis), and acid tolerance (aciduricity). These factors coordinate changes in the physicochemical properties of the biofilm, leading to the proliferation of *S. mutans* and other acidogenic and aciduric species in the oral cavity. *S. mutans* executes its pathogenesis by sensing both cellular and non-cellular environments further regulating gene transcripts in response to environmental changes. Consequently, this regulatory mechanism potentially allows for the expression of virulence factors[2].

S. mutans can also be isolated from individuals without a history of caries. We thought that comparing the differences in gene mutations and gene presence/absence status of *S. mutans* isolates, a highly competent species that lives in a large gene pool with other members of the oral microbiota, in individuals with and without widespread dental caries would be important in understanding its role in the basis of dental caries. In recent years, advancements in next-generation sequencing techniques have facilitated the exploration of structural variations, repetitive regions, DNA methylation information, and previously unsequenced regions through whole-genome sequencing analyses. Previous clinical studies of the *S. mutans* genome have focused predominantly on pediatric population, examining children exhibiting a clinical condition known as severe early childhood caries (S-ECC)[2–4]. It has been observed that control mechanisms and risk factors of the oral microbiota associated with dental caries differ between children and adults, and the progression of caries development also differs in clinical settings. Therefore, existing studies may not fully reflect the dental caries issues of the general population, highlighting the need for new studies focusing on adult individuals.

[1]Institute of Graduate Studies in Health Sciences, Istanbul University, Istanbul, Turkey. [2]Department of Restorative Dentistry, Faculty of Dentistry, Altinbas University, Istanbul, Turkey. [3]Department of Molecular Biology and Genetics, Faculty of Engineering and Natural Sciences, Istanbul Medeniyet University, Istanbul, Turkey. [4]Department of Neurogenetics, Institute of Neurological Sciences, Istanbul University-Cerrahpasa, Istanbul, Turkey. [5]Department of Basic Medical Sciences, Faculty of Dentistry, Istanbul University, Istanbul, Turkey. ✉e-mail: nurtopcu@istanbul.edu.tr

	Control Mean ± Sd (Min - Max) (n = 12)	Experimental (n = 13) Mean ± Sd (Min - Max) (n = 13)
Age	24.42 ± 3.63 (20–31)	23.54 ± 6.92 (18–40)
DMFT	0.33 ± 0.65 (0–2)	9.77 ± 3.46 (6–17)
Salivary flow rate (ml/min)	1.38 ± 0.68 (0.4–3)	1.17 ± 0.69 (0.4–3)
Salivary buffering capacity (pH)	4.62 ± 0.56 (4–5.5)	4.5 ± 0.64 (4.5–5.5)
Salivary *S. mutans* counts (cfu/ml)	$4.33 \times 10^4 \pm 8.01 \times 10^4$ (10^3–2.7×10^5)	$2.31 \times 10^6 \pm 3.74 \times 10^6$ (10^5–1.1×10^7)
Caries risk (%)	8.91 ± 6.25 (2–21)	83.15 ± 8.49 (61–98)

Table 1. Oral health findings of the individuals participating in the study. DMFT: the sum of the number of Decayed, Missing due to caries, and Filled Teeth. Sd: Standard deviation.

Group/sample	Number of samples	Number of read pairs	Number of bases (Gb)	Read length in bp	Percentage of bases with >Q30
Control	12	1,150,558 ± 138,381	345.17 ± 41.51	150 paired-end	95.78 ± 0.89
Experimental	13	1,371,355 ± 212,413	411.41 ± 63.72	150 paired-end	94.63 ± 3.01
Sample SM*	1	1,544,996	463.50	150 paired-end	91.16 ± 4.53

Table 2. Raw sequencing data amount and quality summary table. *Sample SM was the laboratory strain originated from ATCC 25175 and used/sequenced as internal laboratory control.

Type of mutation	Total number (%)
Insertion/deletion without frame shift	157 (0.31%)
Loss of function mutations (Insertion/deletion with frame shift consequences or start/stop codon mutations)	573 (1.13%)
Missense	15759 (31.15%)
Synonym	23568 (46.59%)
Intergenic	10527 (20.81%)

Table 3. Mutation counts according to their functional effects.

Based on this premise, our study aims to investigate and compare the genomic sequences of *S. mutans* strains isolated from the saliva of adults with and without dental caries, using whole-genome sequencing techniques, to elucidate the genotypic characteristics of *S. mutans* and to compare it in terms of its association with dental caries.

In this clinical study, a total of 25 strains were isolated from saliva samples using specific culturing methods. 25 isolates together with a cultured ATCC 25175 strain, were then subjected to species specific PCR testing to ensure all samples are belonging to the *S. mutans*. The experimental group with caries comprised 13 isolates, while the caries-free control group comprised 12 *S. mutans* isolates. The mean age of the control group was determined to be 24.42 ± 3.63, while that of the experimental group was 23.54 ± 6.92. Both the control and experimental groups' DMFT indices, salivary flow rates, buffering capacities, and *S. mutans* (cfu/ml) values and caries risk situations calculated according to the Cariogram model are presented in Table 1. Clinical information for all participants included in the study is provided in Supplementary Table 1.

The concentrations of 26 DNA samples averaged 36.8 ng/μl. A total of ~34 million paired reads were obtained, resulting in a cumulative ~10 Gbp of sequencing data. For each sample, an average of 1-276127 ± 212735 paired DNA reads, each 150 bp in length were obtained. Since the genome size of *S. mutans* is known to be around 2 Mbp, the genome sequencing depth was calculated roughly 188x per isolate. Detailed raw data amount and quality information was given in Table 2.

A total of 50584 genomic mutations were detected in all samples when using ATCC 25175 genome as a reference (obtained from https://www.atcc.org). Annotation of those mutations showed that 47 % were synonymous, 31% were missense mutations, and 21% were located in intergenic regions (Table 3). A statistically significant difference was detected in the presence/absence ratios of 67 mutations in the experimental and control groups ($p < 0.05$) (Supplementary Table 2).

A pan-genome analysis was conducted on a total of 26 samples, including 25 clinical isolates, the genome sequence of the ATCC 25175 strain available in the ATCC database. Totally 2904 coding gene sequences were detected, 1563 of which were identified as core genes, detected in all samples. Only one of the four different forms of the *Lantibiotic mutacin-1140* (*lanA*), gene showed a significant presence ratio difference between the experimental and control groups among all detected genes ($p = 0.03$). When using amino acid sequences of the genes identified in all samples, four different gene sequences showing at least 90 % similarity were detected for the *lanA* gene. One of these detected gene sequences was found in five isolates in the experimental group, while none of the isolates in the control group harboured it. No statistically significant differences were observed for other three *lanA* sequences (Table 4).

A phylogenetic tree was constructed by incorporating the whole genome sequences of *S. mutans* strains isolated from various geographic locations worldwide, obtained from the public databases (Supplementary Table 3), alongside the whole-genome sequences from our study (Fig. 1). Additionally, another tree was

Name	Annotation	Control (n = 12)	Experimental (n = 13)	Sensitivity	Specificity	Odds ratio	*p*
lanA	*Lantibiotic mutacin-1140*	0	5	38.46153846	100	inf	0.03*
		3	2	15.38461538	75	0.545454545	0.64
		3	5	38.46153846	75	1.875	0.67
		0	1	7.692307692	100	Inf	1

Table 4. Presence/absence status of '*lanA*' gene sequences in experimental and control groups. *p < 0.05, Fisher Exact Test (two tailed).

constructed based on a matrix created using the presence/absence status of all detected genes, revealing the genetic relationships among the analysed isolates (Fig. 2).

Methods

Study design. This study was conducted at Faculty of Dentistry, Istanbul University, involving patients presenting to the clinic. The study comprised both experimental and control groups, all without any systemic diseases. The experimental group included individuals with a high prevalence of dental caries, while the control group consisted of individuals selected from a similar age group without dental caries. Prior to the commencement of the study, ethical approval was obtained from the Istanbul University Faculty of Dentistry Clinical Research Ethics Committee (Date: 28.06.2021 - No: 268053). Voluntary informed consent was obtained from all participants prior to their participation in the study.

Clinical examination and sample collection. To assess the oral-dental health status of the volunteers, a comprehensive anamnesis form was created, capturing the individual's age, gender, DMFT (Decayed, Missing, Filled Teeth), oral hygiene index, high-risk indicators for dental caries, and protective factors. Oral hygiene was categorized according to the evaluation system in the Cariogram. Stimulated saliva samples were collected from the volunteers to determine dental caries risks, and caries activity tests were conducted.

Caries risk analysis. Individual caries risk for all volunteers was calculated using the Cariogram application based on laboratory tests and anamnestic data. Subsequently, individuals were divided into two groups according to their caries risk: those with high and very high caries risk were considered as the experimental group, while those with low and very low caries risk were categorized as the control group[5].

Mutans streptococci isolation. Mutans streptococci were isolated from the samples using quantitative cultural techniques: All the saliva samples were inoculated onto Mitis Salivarius Bacitracin (MSB) Agar (Acumedia Man Inc., Baltimore, Maryland, USA) after ten-fold dilution[6]. The inoculated plates were then incubated at 37 °C in a CO_2-enriched atmosphere (5–7%) for 48 hours. After the incubation period, the opaque, and rough colonies on the surface of MSB Agar were counted using a magnifying glass. The colony-forming units (cfu/ml) were calculated, and the level of mutans streptococci in saliva was determined. The isolates obtained from a single colony were subcultured on Brain Heart Infusion (BHI) Agar (Merck KGaA, Darmstadt, Germany). Subsequently, the subcultures were stored at −20 °C, prepared for PCR analysis.

PCR analysis, DNA isolation, and Whole Genome Sequencing. DNA isolation was performed initially for 40 isolates, which were grown on the selective media, obtained from clinical samples and one *Streptococcus mutans* ATCC 25175 isolate obtained from the American Type Culture Collection using the High Pure PCR Template Preparation Kits (Roche, Mannheim, Germany) following manufacturer's recommendations. The resulting DNA extracts were stored at −20 °C. Species-specific primers (F: 5′-GGTCAGGAAAGTCTGGAGTAAAAAGGCTA-3′ and R :5′-GCGGTAGCTCCGGCACTAAGCC-3′) was used for PCR confirmation[7]. Following amplification, 10 µl of PCR product (expected amplicon size was 282 bp) was analysed by running a 2% agarose gel. The agarose gels were stained and visualized under ultraviolet light.

The concentration values of the resulting DNA samples were recorded after quality control using the NanoDrop-2000 device. DNA samples passing the quality control parameters were prepared for next-generation sequencing using the "Illumina DNA Prep kit" according to manufacturer's protocols and were sequenced on the Illumina NextSeq. 2000 instrument in order to obtain 2 × 150 bp reads. Raw sequencing data were stored in FASTQ format after demultiplexing.

Sequencing data processing. Demultiplexing and removing of sequencing adapters from raw sequences were performed by CASAVA data analysis software (Illumina, USA). Quality control of the raw sequencing data was initially performed using FASTQC and MultiQC[8]. Read counts and sequence quality scores were examined and visualized using these programs. Sequencing data quality statistics are given in Table 2.

As the first step of bioinformatics analysis, de novo genome assembly was carried out for each sample using the default parameters of Unicycler v0.5.0[9], and the assembled genomes for each sample were used for further analyses. Subsequently, MLST (multilocus sequence typing) analysis was performed for each genome using the autoMLST program with default parameters[10] to confirm whether if assembled contigs were belong to *S. mutans*. Samples not identified as the *S. mutans* bacterium were excluded from the analysis at this step.

A comparison of the genome sequences generated by *de novo* genome assembly with other *S. mutans* genome sequences was conducted. For this purpose, other *S. mutans* genome sequences were selected by various

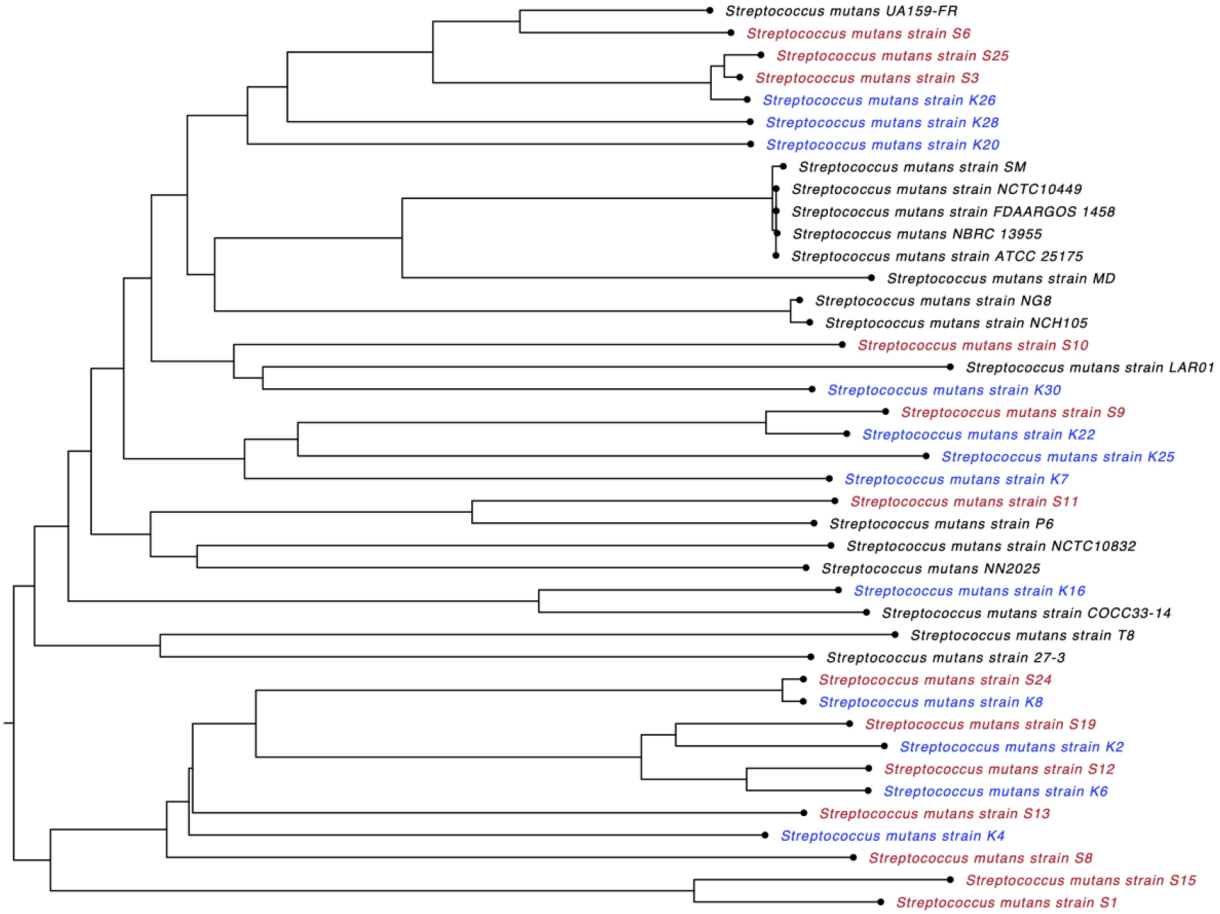

Fig. 1 Phylogenetic tree constructed for the 26 genome sequences in this study and the genome sequences of 16 other *S. mutans* strains isolated selected from different geographical regions (Red: Experimental/Blue: Control).

geographical regions from which they were isolated. Genome-wide distances were calculated using the Genome Blast Distance Phylogeny (GBDP) method implemented in the FastME v2.1.6.1 program[11], and phylogenetic analysis was performed to draw a phylogenetic tree along with all genome sequences included in the study.

Mutation analysis. For the investigation of mutations in all samples included in the analysis, raw FASTQ data was analysed using the Snippy v4.6.0 program (https://github.com/tseemann/snippy) with default parameters, using the ATCC 25175 reference isolate genome as the reference. The identified variants were written to VCF files and merged using the bcftools v1.18 program[12]. The effects of each mutation on genes/proteins were determined using Snpeff v4.3[13], and a table including occurrences and annotations of all mutations was created. Using this table, differences in the presence rates of each mutation between the experimental and control groups were tested using Fisher's exact test, and p-values were calculated. A threshold of $p < 0.05$ was used for statistical significance.

Pan-genome analysis. The genomes generated from the *de novo* assembly of the isolates were subjected to pan-genome analysis to investigating presence/absence of the detected genes. For this purpose, we first used the assembled genome sequences to determine the gene/protein sequences using the Prokka v1.14.6[14]. The obtained protein sequences and annotations were analysed using the "Roary: the Pan Genome Pipeline" v3.13.0 program[15], and the detected amino acid sequences were clustered based on a 90% similarity threshold to determine the presence/absence status in the samples. Core genes were identified by searching genes which were detected in all samples. A phylogenetic tree was prepared based on the presence/absence status of all detected genes. Using the outputs of the Roary program, statistical comparison of the presence ratios of genes between the experimental and control groups was performed using the Scoary v1.6.16 program[16], and p-values were calculated for each gene/protein clusters with 90% similarity. A threshold of $p < 0.05$ was used for statistical significance.

Data Records

The study data consists of three types of data including raw data in FASTQ format, assembled genome sequences in FASTA format, and the table of all detected mutations and their annotations. Raw sequencing data[17] and assembled genome sequences as contigs[18–42] have been deposited with links to BioProject accession number PRJNA1114739 in the NCBI BioProject database (https://www.ncbi.nlm.nih.gov/bioproject/) and detected mutations with their annotations for all *Streptococcus mutans* isolates in this study is available at Figshare[43].

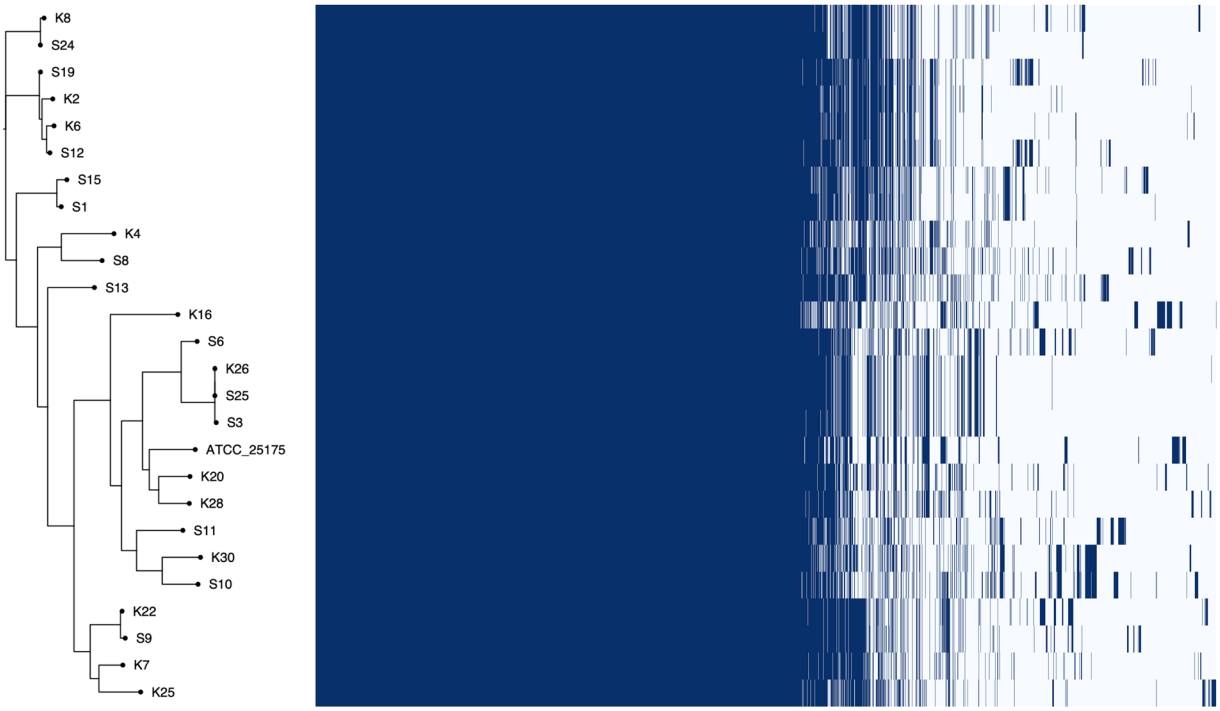

Fig. 2 Phylogenetic tree constructed based on the gene presence/absence matrix data from pangenome analysis. Presence/absence data was visualized as heatmap on the right side of the figure (Dark and white lines in the heatmap represent the presence and absence status, respectively) (S: Experimental/K: Control).

Technical Validation

All saliva samples and patient examinations were collected and conducted by a single researcher (M.Y.U) with more than five years of experience. The concentration and quality of the obtained genomic DNA samples were evaluated using NanoDrop 2000 (Thermo Fisher Scientific, USA) spectrophotometer. *S. mutans* isolates were grown on selective Mitis Salivarius Bacitracin Agar. *Streptococcus* colonies were collected and used for DNA extraction and then subjected for PCR confirmation using *S. mutans* specific primers in order to ensure isolates are belonging to *S. mutans*. To minimize PCR bias, all PCRs were performed in duplicates. The concentration of the sequencing library was assessed using a Qubit Fluorometer (Thermo Fisher Scientific, USA). Assembled contigs per isolate were subjected to ribosomal Multilocus Sequence Typing (rMLST) analysis[44] which is an approach that analyses the 53 bacterial ribosome protein subunit genes and calculates a support score for species level typing. All 25 assembled contigs were identified as *Streptococcus mutans* according to rMLST analysis. We used the ATCC 25175 strain, which was obtained in our laboratory in 2015 directly from ATCC, for only sequencing validation purposes. The resequencing of this sample resulted in only eight mutations, which was considered a natural result of expected *de novo* mutations.

Code availability

No custom scripts were used in this study. All software and pipelines were used with default parameters unless otherwise noted as described in methods section.

Received: 10 June 2024; Accepted: 2 January 2025;
Published online: 21 February 2025

References

1. Lemos, J. A. & Burne, R. A. A model of efficiency: stress tolerance by *Streptococcus mutans*. *Microbiology (Reading).* **154**, 3247–3255 (2008).
2. Bedoya-Correa, C. M., Rodríguez, R. J. R. & Parada-Sanchez, M. T. Genomic and phenotypic diversity of *Streptococcus mutans*. *J. Oral Biosci.* **61**, 22–31 (2019).
3. Argimón, S. *et al.* Comparative genomics of oral isolates of *Streptococcus mutans by in silico* genome subtraction does not reveal accessory DNA associated with severe early childhood caries. *Infection, Genetics and Evolution.* **21**, 269–278 (2014).
4. Gautam, S. S., Rajendra, K. C., Leong, K. W. C., Mac Aogáin, M. & O'Toole, R. F. A step-by-step beginner's protocol for whole genome sequencing of human bacterial pathogens. *J. Biol. Methods.* **6**, e110 (2019).
5. Bratthall, D., Petersson, G. H., Stjernswärd, J. R. & Cariogram M. Cariogram Internet version 2.01. https://mau.se/en/about-us/faculties-and-departments/faculty-of-odontology/cariogram/ (2004).
6. Gold, O. G., Jordan, H. V. & Van Houte, J. A selective medium for *Streptococcus mutans*. *Arch. Oral Biol.* **18**, 1357–1364 (1973).
7. Corpet, F. Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res.* **16**, 10881–10890 (1988).
8. Ewels, P., Magnusson, M., Lundin, S. & Käller, M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **32**, 3047–3048 (2016).
9. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).

10. Alanjary, M., Steinke, K. & Ziemert, N. AutoMLST: an automated web server for generating multi-locus species trees highlighting natural product potential. *Nucleic Acids Res.* **47**, W276–W282 (2019).
11. Lefort, V. & Desper, R. FastME 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Mol. Biol. Evol.* **32**, 2798–2800 (2015). Gascuel & O.
12. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
13. Cingolani, P. *et al*. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**, 80–92 (2012).
14. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
15. Page, A. J. *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
16. Brynildsrud, O., Bohlin, J., Scheffer, L. & Eldholm, V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol.* **17**, 1–9 (2016).
17. *NCBI Sequence Read Archive* https://identifiers.org/ncbi/insdc.sra:SRP509243 (2024).
18. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046037035.1 (2024).
19. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046036975.1 (2024).
20. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046036935.1 (2024).
21. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035415.1 (2024).
22. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046036895.1 (2024).
23. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046036915.1 (2024).
24. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035455.1 (2024).
25. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035315.1 (2024).
26. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035345.1 (2024).
27. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035335.1 (2024).
28. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035355.1 (2024).
29. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035395.1 (2024).
30. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035295.1 (2024).
31. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035255.1 (2024).
32. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035235.1 (2024).
33. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035215.1 (2024).
34. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035275.1 (2024).
35. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035175.1 (2024).
36. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035195.1 (2024).
37. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035115.1 (2024).
38. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035155.1 (2024).
39. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035135.1 (2024).
40. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035095.1 (2024).
41. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035075.1 (2024).
42. *NCBI GenBank* https://identifiers.org/ncbi/insdc.gca:GCA_046035035.1 (2024).
43. Ucuncu, M. Y., Ucuncu, M. K., Karacan, I. & Topcuoglu, N. Genome resequencing and comparative analysis of *Streptococcus mutans* isolates obtained from patients with high and low caries risk. Detected mutations among *Streptococcus mutans* isolates, *Figshare.*, https://doi.org/10.6084/m9.figshare.25866511.v1 (2024).
44. Jolley, K. A. *et al*. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. *Microbiology (Reading).* **158**, 1005–1015 (2012).

Acknowledgements

This study is supported by the Research Fund of Istanbul University (project no: 38340).

Author contributions

M.Y.U. and M.K.U. conducted data collection, performed dental examinations, as well as classifications into dental caries indexes. I.K. and N.T. conducted data analysis, interpreted the results, and reviewed the manuscript. All authors have contributed to writing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-025-04399-w.

Correspondence and requests for materials should be addressed to N.T.