A Core Genome Multilocus Sequence Typing Scheme for *Streptococcus mutans*

Shanshan Liu,[a,b] Xiaoliang Li,[a] Zhenfei Guo,[a] Hongsheng Liu,[b] Yu Sun,[c] Yudong Liu,[d] Qinglong Wang,[b] Shengkai Liao,[a]
Kai Zhang[a]

[a]Department of Stomatology, The First Affiliated Hospital of Bengbu Medical College, Bengbu, China
[b]Department of Stomatology, Bengbu Medical College, Bengbu, China
[c]Department of Biochemistry and Molecular Biology, Bengbu Medical College, Bengbu, China
[d]Department of histology and Embryology, Bengbu Medical College, Bengbu, China

ABSTRACT *Streptococcus mutans* is one of the primary pathogens responsible for the development of dental caries. Recent whole-genome sequencing (WGS)-based core genome multilocus sequence typing (cgMLST) approaches have been employed in epidemiological studies of specific human pathogens. However, this approach has not been reported in studies of *S. mutans*. Here, we therefore developed a cgMLST scheme for *S. mutans*. We surveyed 199 available *S. mutans* genomes as a means of identifying cgMLST targets, developing a scheme that incorporated 594 targets from the *S. mutans* UA159 reference genome. Sixty-eight sequence types (STs) were identified in this cgMLST scheme (cgSTs) in 80 *S. mutans* isolates from 40 children that were sequenced in this study, compared to 35 STs identified by multilocus sequence typing (MLST). Fifty-six cgSTs (82.35%) were associated with a single isolate based on our cgMLST scheme, which is significantly higher than in the MLST scheme (11.43%). In addition, 58.06% of all MLST profiles with ≥2 isolates were further differentiated by our cgMLST scheme. Topological analyses of the maximum likelihood phylogenetic trees revealed that our cgMLST scheme was more reliable than the MLST scheme. A minimum spanning tree of 145 *S. mutans* isolates from 10 countries developed based upon the cgMLST scheme highlighted the diverse population structure of *S. mutans*. This cgMLST scheme thus offers a new molecular typing method suitable for evaluating the epidemiological distribution of this pathogen and has the potential to serve as a benchmark for future global studies of the epidemiological nature of dental caries.

IMPORTANCE *Streptococcus mutans* is regarded as a major pathogen responsible for the onset of dental caries. *S. mutans* can transmit among people, especially within families. In this study, we established a new epidemiological approach to *S. mutans* classification. This approach can effectively differentiate among closely related isolates and offers superior reliability relative to that of the traditional MLST molecular typing method. As such, it has the potential to better support effective public health strategies centered around this bacterium that are aimed at preventing and treating dental caries.

KEYWORDS molecular biology, *Streptococcus mutans*, cgMLST, whole-genome sequence

Dental caries are highly prevalent bacterial infections in humans, and they can have a significant adverse impact on the mental and physical health of affected individuals (1–3). One of the primary drivers of these caries is the Gram-positive bacterium *Streptococcus mutans*, as its colonization of the oral cavity is closely linked to caries initiation and progression (4). *S. mutans* bacteria are able to readily adhere to the

surface of a tooth, forming a biofilm and releasing acidic compounds following carbohydrate metabolism, thereby corroding the tooth surface (5). There is strong epidemiological evidence that *S. mutans* is more commonly detected in the oral cavity of children affected by dental caries compared to that in children free of such caries (6, 7), and the forms of *S. mutans* isolated from children with caries have been found to be more virulent than those isolated from caries-free children (8). There is thus a clear need to better understand the genetic diversity of *S. mutans* in order to more fully elucidate its pathogenicity and its ability to be transmitted between humans.

S. mutans has previously been differentiated by serotyping (*c*, *e*, *f*, and *k*) on the basis of the composition of its cell surface rhamnose-glucose polymers (9, 10). However, this approach makes it difficult to separate strains of the same serotype. Genotypic methodologies, including ribotyping, arbitrary primed PCR (AP-PCR), chromosomal DNA restriction fragment polymorphism, repetitive extragenic palindromic PCR (rep-PCR), multilocus enzyme electrophoresis, and pulse-field gel electrophoresis (PFEG), have been used to subtype *S. mutans* (11–16). These phenotypic approaches, however, are time intensive and have difficulty differentiating between closely related bacterial strains (17). In 2007, Nakano et al. first reported the use of multilocus sequence typing (MLST) in order to more effectively and reliably discriminate between *S. mutans* strains (17). MLST strategies allow for an efficient analysis of the relationship between bacterial strains by sequencing defined conserved housekeeping genes and comparing this genotypic information between isolates (17). This approach can more reliably and reproducibly detect differences between closely related bacterial strains, as traditional gel-based typing strategies are based upon subjective similarity, whereas single-nucleotide mutations can be detected via MLST, thus allowing for the identification of a new sequence type (ST). Another MLST scheme was also reported by Do et al. in 2010 (18).

In recent years, a core genome MLST (cgMLST) strategy has been used to differentiate between bacterial strains with even greater reliability than traditional MLST approaches, as cgMLST strategies incorporate more sequence data than the 5 to 8 housekeeping genes typically sequenced for MLST (19, 20). Such cgMLST approaches have been successfully used to type many pathogens, including *Mycoplasma synoviae*, *Brucella* spp., *Yersinia*, *Listeria monocytogenes*, *Staphylococcus capitis*, *Staphylococcus argenteus*, and *Staphylococcus aureus* (21–27). While such cgMLST strategies have been well documented, they have not been well studied with respect to *S. mutans*. As such, in the present study, we defined a *S. mutans* cgMLST scheme and compared the resolution between MLST and cgMLST schemes.

RESULTS

Candidate cgMLST target gene set. We began by identifying genes that were well-suited to incorporation into our cgMLST scheme by filtering the *S. mutans* UA159 (GenBank accession number NC_004350.2) reference genome based upon minimum gene length, gene overlap, and start/stop codon positioning. This led to the identification of 1,811 genes, and this number was further winnowed to 1,797 genes following the removal of plasmid and transposon genes. These genes were then retained as candidate targets for cgMLST analyses.

Candidate genomes. We downloaded 199 total *S. mutans* genomes from the National Center for Biotechnology Information (NCBI) database for screening purposes. Of these genomes, 33 that lacked the eight traditional MLST gene segments were eliminated from further analysis, as were 5 genomes in which single copy gene numbers were <75%. The remaining 161 genomes were then used as candidates to screen our candidate cgMLST target genes. We additionally extracted the sequences corresponding to the eight housekeeping genes traditionally used for MLST-based bacterial typing, and we compared these sequences to the *S. mutans* MLST database. We then constructed a minimum spanning tree based upon the allelic profiles of these 161 genomes, the 161 genomes distributed in the whole tree, revealing them to be well distributed throughout the tree (Fig. 1; see also Table S1 in the supplemental material).
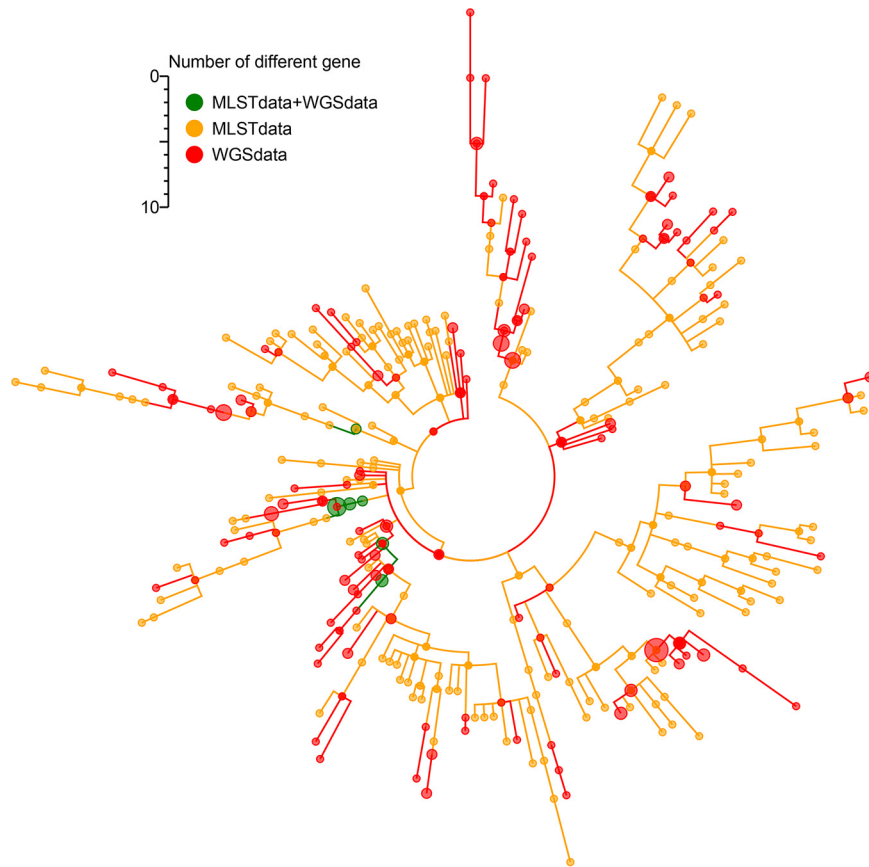
FIG 1 A representation of 161 candidate *S. mutans* genomes based upon the *S. mutans* multilocus sequence typing (MLST) database. Sequences corresponding to the eight housekeeping genes that comprise the allelic profile of the Nakano MLST scheme were extracted from these 161 genomes and were queried against the *S. mutans* Nakano MLST database. In the resultant minimum spanning tree, these 161 genomes (whole-genome sequencing [WGS] data) were found to be distributed throughout the MLST data, indicating that this 161-genome data set corresponds to a satisfactory representation of the *S. mutans* population.

cgMLST target gene definition, evaluation, and potential functional characteristics. Using these 161 genomes, we next sought to better define the list of target genes suitable for inclusion in our cgMLST scheme. Based on the criteria defined in Materials and Methods, we were able to identify 594 core genes (30.3% of the reference genome) suitable for incorporation into our cgMLST scheme (Fig. 2; see also Table S2 in the supplemental material). To preliminarily explore the potential functional characteristics of these target genes, gene ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways in which they were enriched were identified using the Database for Annotation, Visualization and Integrated Discovery (DAVID) software (see Fig. S1 in the supplemental material). We found that the target genes were associated with translation and ATP binding capacity. A total of 8 enriched pathways were identified via KEGG analysis as being significantly enriched for these genes ($P < 0.05$), including the pathways for carbon metabolism, glycolysis, biosynthesis of secondary metabolites crucial to acid production, cell growth, and antagonism of other oral streptococci in *S. mutans*.

Comparison between MLST and cgMLST schemes. Based on the MLST and cgMLST schemes, the sequenced isolates were found to belong to 35 distinct STs and 68 distinct core genome sequence types (cgSTs), respectively. According to the MLST scheme, a total of 4 STs (ST234, ST249, ST265, and ST269) were associated with a single isolate, accounting for 11.43% of the 35 STs. However, there were 56 cgSTs (82.35%) that were associated with a single isolate according to the cgMLST scheme, with this
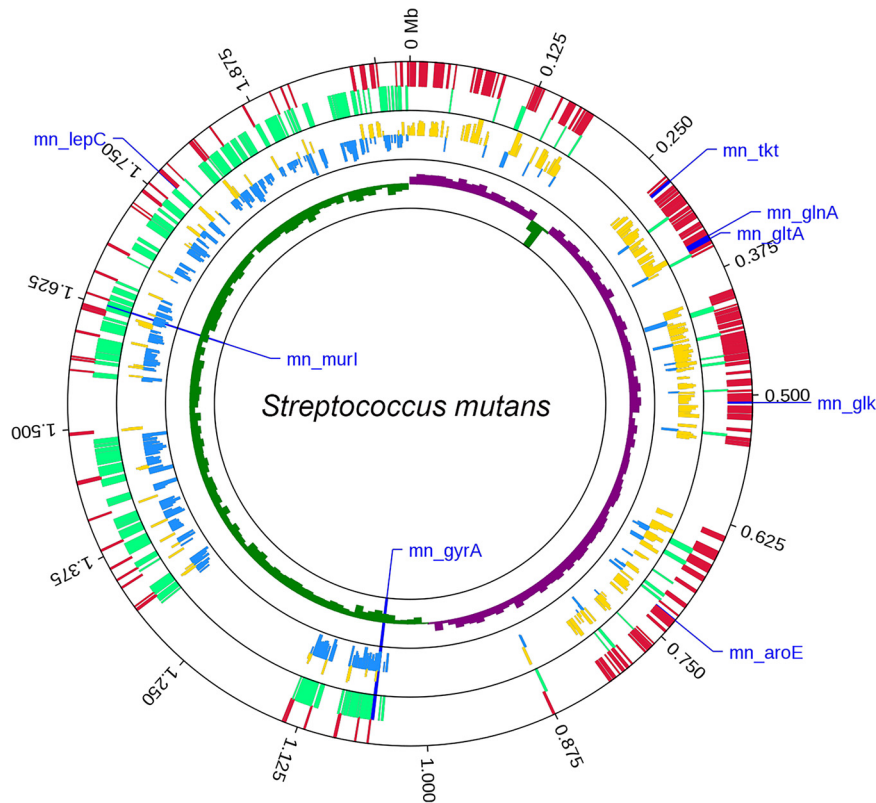
FIG 2 Distribution of the 594 cgMLST target genes within the *S. mutans* UA159 reference genome. The eight reference genes used for standard MLST genes are noted. Target genes are shown in the outer circle, with green and red corresponding to those genes on the negative and positive strands, respectively. The allele numbers for cgMLST target genes are shown in the middle circle, with blue and yellow corresponding to genes on the negative and positive strands, respectively. The GC skew value [the ratio of (G − C)/(G + C)] is also shown, with green and violet corresponding to these skew values on the negative and positive strands, respectively. The *S. mutans* UA159 genome (GenBank assembly number GCA_000007465.2) served as a reference for this analysis.

frequency being significantly higher than that in the MLST scheme ($P < 0.001$). Although some of these isolates were still found to be identical using this cgMLST scheme, 58.06% of all MLST profiles that contained ≥ 2 isolates were further differentiated by our cgMLST scheme (18 out of 31 STs). For example, the SMB6, SMB37, SMB79, and SMB80 isolates shared an identical ST (ST236) based upon the MLST scheme, whereas these ST236 isolates were separated into different cgSTs (cgST56, cgST29, cgST68, and cgST67 for SMB6, SMB37, SMB79, and SMB80, respectively) using our *S. mutans* cgMLST scheme. Each child provided two isolates. The two isolates from 70% of these children (28 out of the 40) were distinguished using our cgMLST scheme, with this rate being significantly higher than that achieved when using the MLST scheme, which yield a 7.5% discrimination rate ($P < 0.001$).

The topologies of the phylogenetic trees developed using Hasegawa-Kishino-Yano (HKY) and general time-reversible (GTR) models were highly congruent in the cgMLST scheme (Fig. 3). Both of the trees indicated that these isolates may have originated from SMB63 and SMB64. Strong bootstrapping support was observed in both trees. However, the topologies of the two trees developed when using the MLST scheme were relatively lower and they exhibited reduced bootstrap support (see Fig. S2 in the supplemental material). The topologies of the cgMLST scheme-based phylogenetic tree and a single-nucleotide polymorphism (SNP)-based phylogenetic tree generated under the HKY model were more congruent, whereas the topological structural consistency of the phylogenetic trees based upon the cgMLST scheme and the SNP-based approach generated under the HKY model was relatively poor (Fig. 4).

mSphere

A   HKY model (cgMLST)

B   GTR model (cgMLST)

Tree scale: 0.001

Tree scale: 0.001
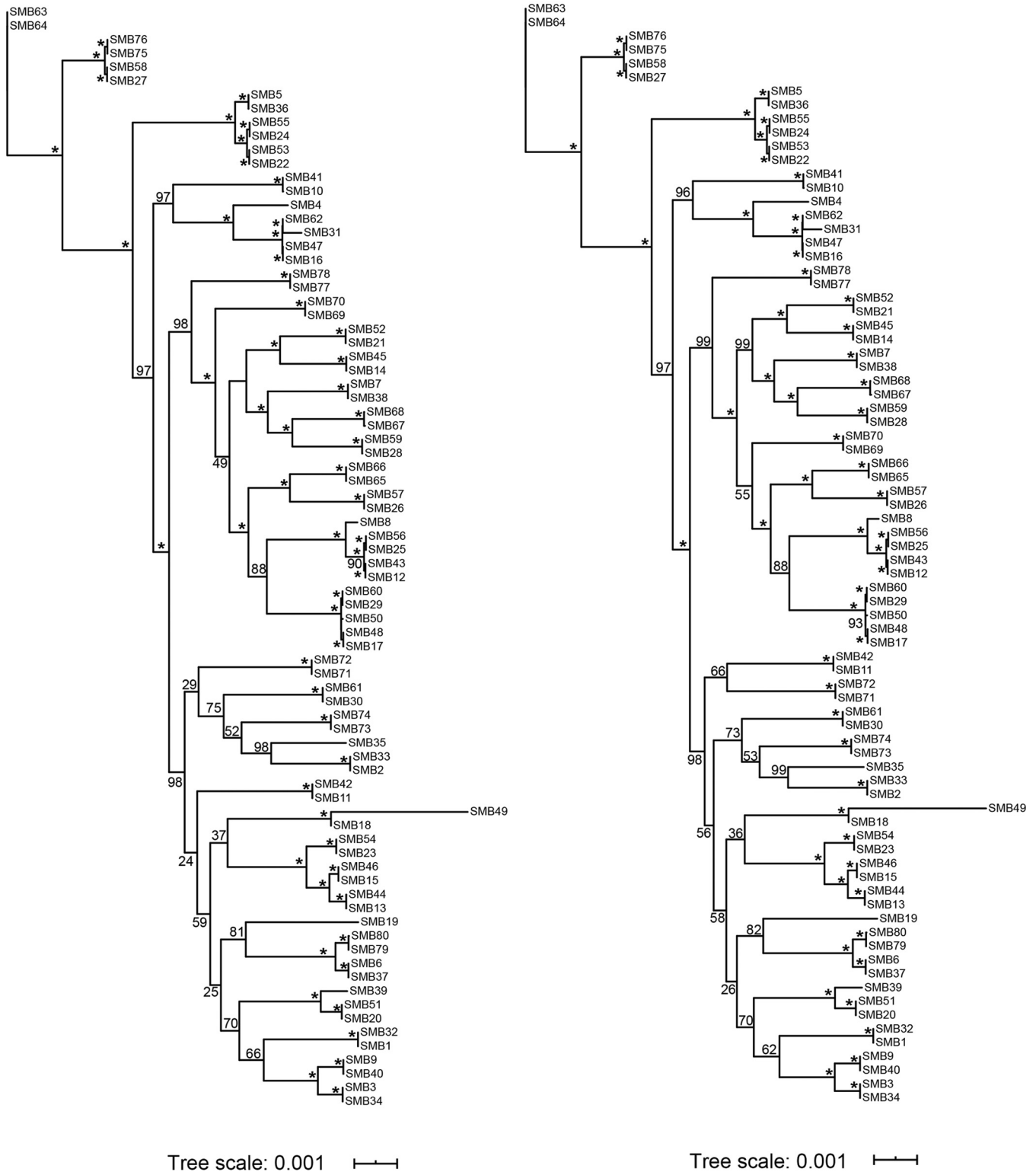
FIG 3 Phylogenetic trees corresponding to the 80 sequenced samples constructed according to the cgMLST scheme. Trees were generated using the maximum likelihood method. (A) Tree constructed under the HKY model and (B) tree constructed with the GTR model. Numbers on lines indicate bootstrap values determined for 1,000 replicates. Scale bar indicates the proportion of nucleotide substitutions. An asterisk (*) indicates 100%.

SNP-based phylogenetic tree
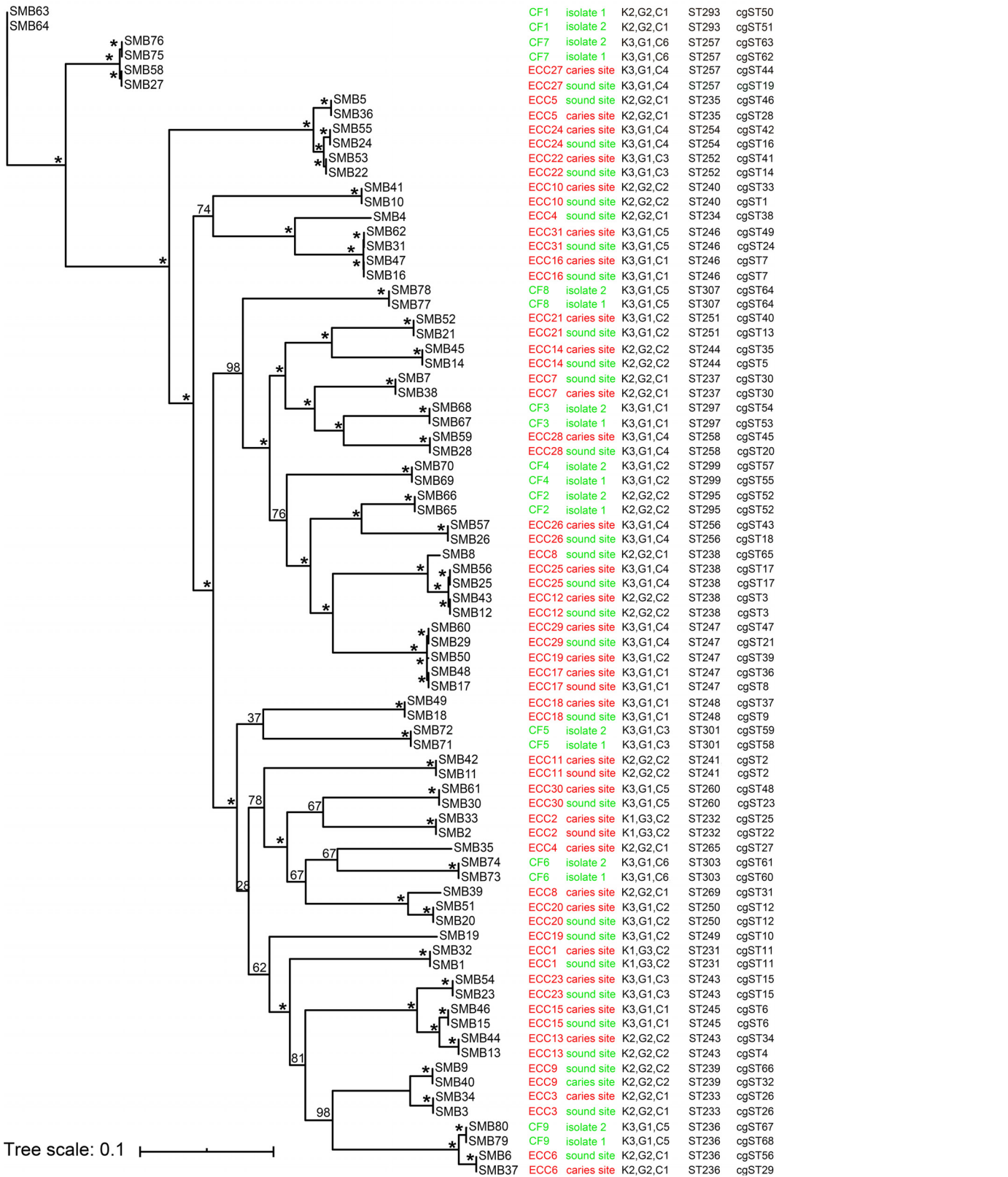
FIG 4 Single-nucleotide polymorphism (SNP)-based phylogenetic trees generated using the maximum-likelihood method under an HKY model. Numbers on lines are bootstrap values determined for 1,000 replicates. An asterisk (*) corresponds to

S. mutans epidemiological analysis. Child-to-child transmission of *S. mutans* between different kindergartens and between different classes in the same kindergarten was found to have occurred in Bengbu, China, according to the MLST scheme (Fig. 5). Among the 40 pairs of *S. mutans* isolates obtained from a given child, 37 pairs of isolates shared the same ST according to this MLST scheme, suggesting them to be epidemiologically related isolates. Among these 37 pairs of isolates, there were 12, 20, and 8 pairs of isolates that shared the same cgST, core genome SNPs, and total SNPs, respectively. More than 92% of epidemiologically related pairs of isolates were differentiated based upon 7 differences in core genome SNPs, 10 differences in total SNPs, and 7 alleles in cgMLST scheme. Minimum spanning trees constructed based on MLST, cgMLST, and core genome SNP analyses of these 80 isolates are shown in Fig. 6. A minimum spanning tree constructed based on the resultant cgMLST analysis using 145 isolates from 10 countries showed that no close epidemiological relationships among isolates were observed among different countries (≤ 7 allele differences) (Fig. 7). For details regarding the number of allele differences among these isolates, see Table S3 in the supplemental material.

DISCUSSION

In this study, we observed the distribution of 161 candidate *S. mutans* genomes throughout the tree, consistent with these genomes being representative of the diversity of *S. mutans* populations (Fig. 1). In prior studies, the proportion of core genes from the reference genome has ranged from 13.3% to 59.33% (28–31). In the present analysis, we retained 594 core target genes in our cgMLST scheme, corresponding to 30.3% of the reference genome. Differences in the proportion of candidate genomes included in each study influence the number and thus the percentage of core genes. For example, a total of 634 cgMLST (13.3% of the reference genome) targets were found based on 167 genomes for *Klebsiella pneumoniae*, whereas the number of targets rose to 1143 (24.0%) when the number of genomes increased to 907 (30, 31). In addition, the genome length may also influence the number and the percentage of core genes, with longer genomes including a greater proportion of target genes in the genome. For example, the median total length of the *Listeria monocytogenes*, *Mycoplasma gallisepticum*, and *S. mutans* genomes retrieved from the NCBI database (14 February 2020) were 2.97 Mb, 0.96 Mb, and 1.96 Mb, respectively. In contrast, the median total length of the *K. pneumoniae* genome is 5.59 Mb, and its core target genes account for just 24.0% of its reference genome. Therefore, the size of the genome is not the sole factor that determines the number of target genes. In the MLST scheme, housekeeping genes were evenly spaced throughout the reference genome, as this is critical for such schemes (18). Similarly, the target genes in this study were distributed throughout the genome (Fig. 2), confirming the validity and utility of this core gene set (30).

GO annotation and KEGG analyses revealed that these target genes were involved in 8 pathways and 5 molecular functions ($P < 0.05$) that may be related to cell growth under different conditions and to cariogenicity. For example, in *S. mutans*, the L-lactate dehydrogenase, enzyme encoded by *ldh* (SMU_1115) and activated by fructose-1,6-diphosphate (FDP), can catalyze the conversion of pyruvate to lactic acid and decrease the pH of the growth medium during glycolysis (32–34). The low-pH environment, in turn, leads to the demineralization of the tooth surface. Mutacin IV is an effective bacteriocin produced by *S. mutans* that can antagonize the growth of several oral streptococcal species (35). Mutacin IV is a core secondary metabolite according to the

FIG 4 Legend (Continued)

100%. CF, caries free; ECC, early childhood caries. Each child provided two isolates. CF1 to CF9 represent the identifiers (IDs) for CF children. ECC1 to ECC31 represent the IDs of children with ECC. CF children provided two isolates from the surfaces of healthy teeth, while ECC children provided one isolate from a sound site and one isolate from a caries site. CF children and their isolates are marked in green, whereas ECC children and their isolates from caries sites are marked in red and ECC child isolates from sound sites are marked in green. K, G, and C correspond to kindergarten, grade, and class, respectively.
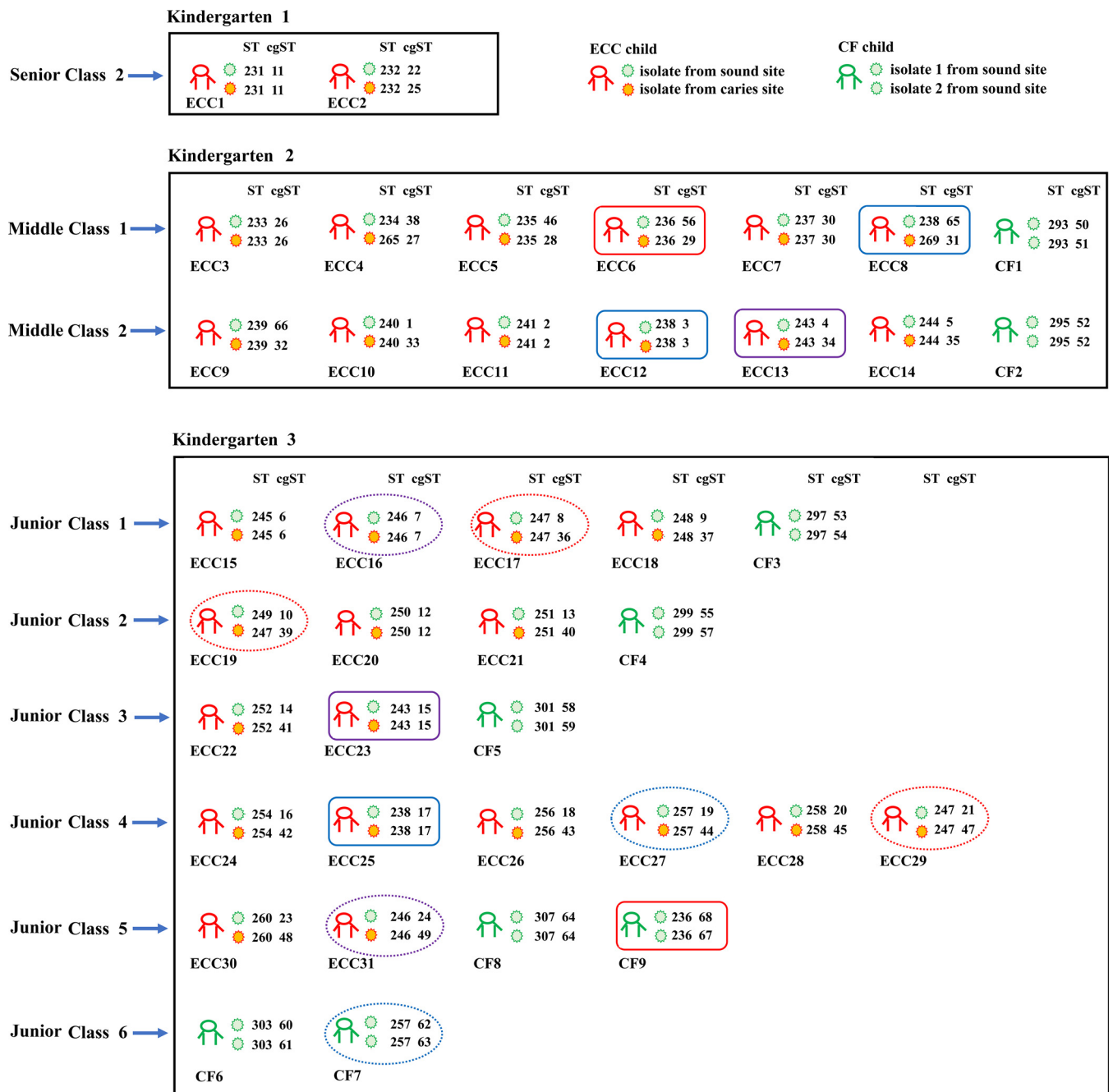
FIG 5 Epidemiological analysis of the 80 *S. mutans* isolates from Bengbu, China. CF children and isolates are marked in green, whereas ECC children and the isolates from caries site are marked in red and isolates from the sound sites in these ECC children are marked in green. Square boxes indicate transmission between kindergartens, and ovals indicate transmission between different classes within a given kindergarten. Isolates with identical STs according to the MLST scheme are marked in the same color using square boxes or ovals.

antiSMASH analysis presented in our previous study (36). SepM, a membrane-associated polypeptide encoded by *sepM* (SMU_518) in *S. mutans*, is a prerequisite for the synthesis of mutacin IV (37).

As shown in a previous study, 47 *S. mutans* isolates were analyzed with MLST scheme resulting in 29 STs, compared with 22 rep-PCR genotypes (38). Seven subtypes were further differentiated by MLST. In the present study, more cgSTs ($n = 68$) were defined based upon our cgMLST scheme relative to the 35 STs defined by this prior MLST scheme, and 82.35% of cgSTs were associated with a single isolate among these 68 STs, which is a significantly higher percentage than that under the MLST scheme. In
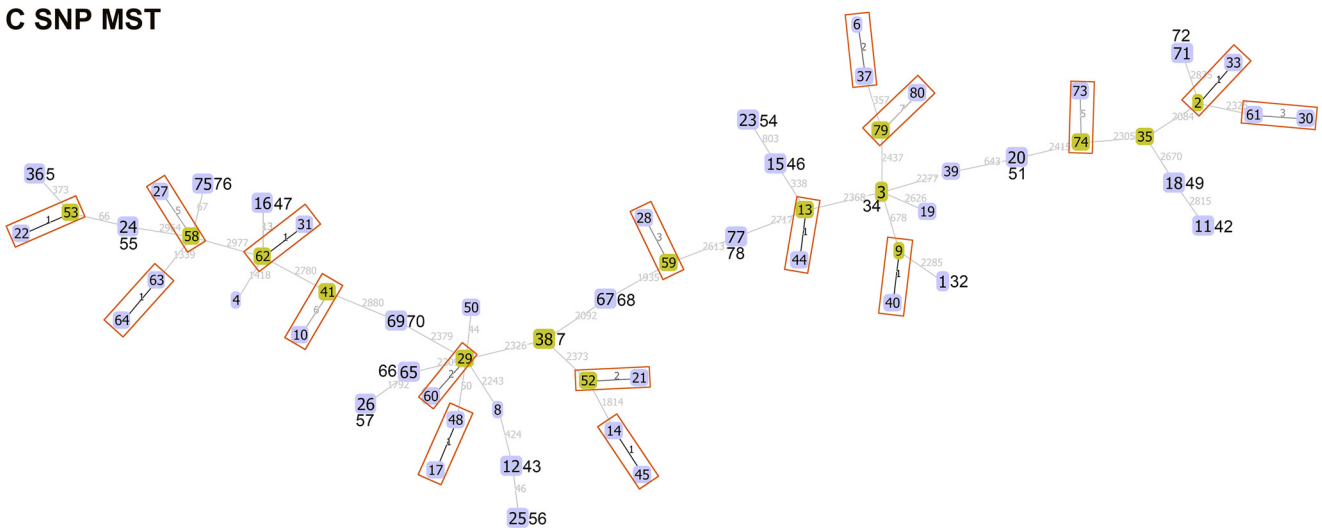
A MLST MST
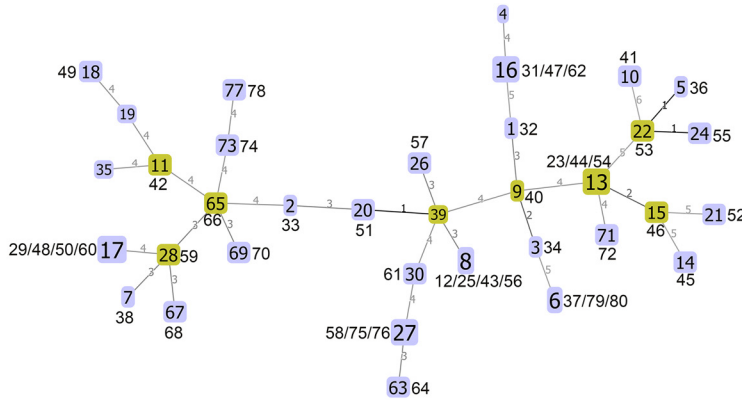
B cgMLST MST
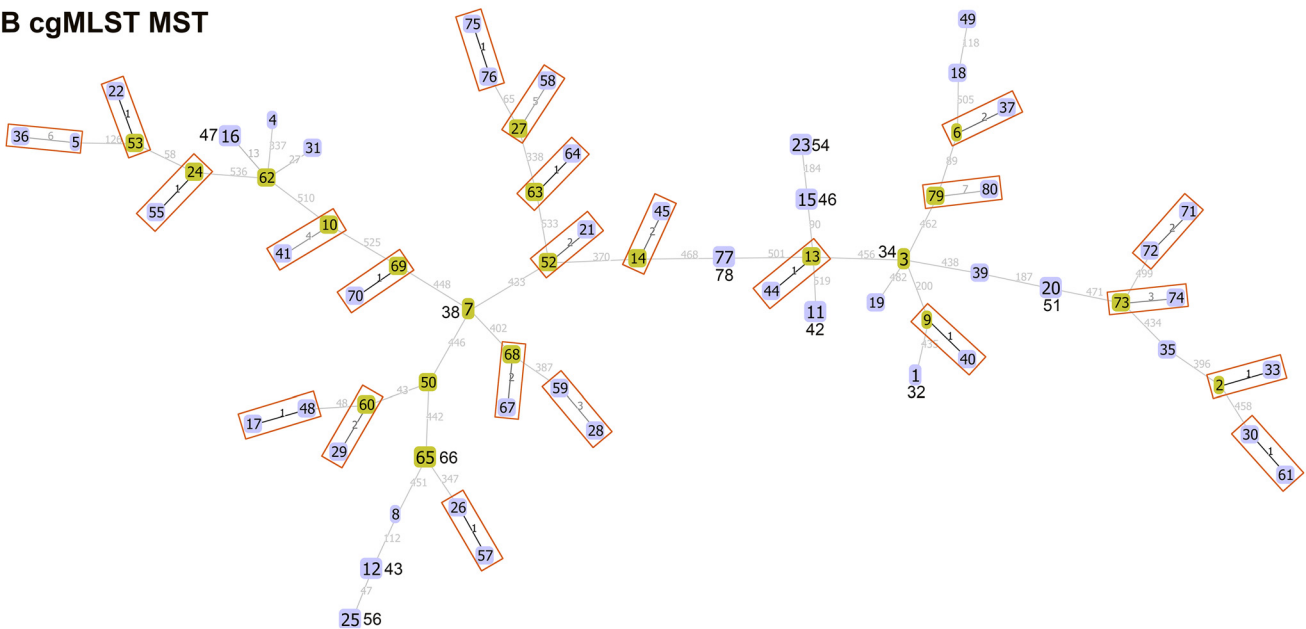
C SNP MST

FIG 6 Minimum spanning trees (MST) of the 80 *S. mutans* isolates sequenced in this study. (A) MST based on MLST, (B) MST based on cgMLST, and (C) MST based on core genome SNPs. Numbers in boxes correspond to the SMB ID (SMB1 to SMB80). A yellow background indicates samples in nodes, and blue backgrounds indicate samples in branches. Red squares represent the samples with identical STs according to the MLST scheme that were further distinguished by cgMLST or by core genome SNPs. Numbers on lines correspond to the number of different alleles in MLST and cgMLST and the number of SNP differences in the core genome.
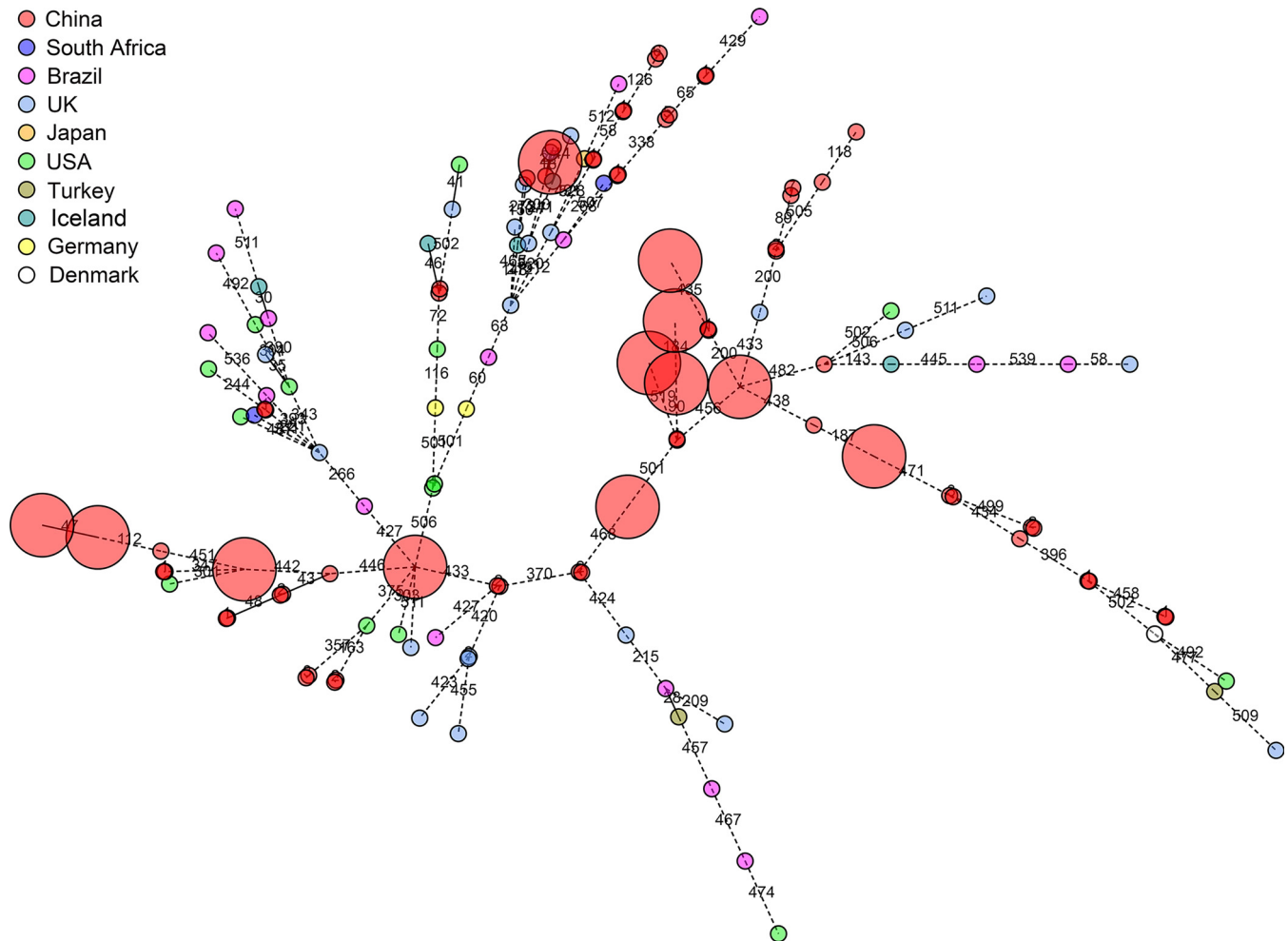
FIG 7 Minimum spanning tree analysis of the cgMLST profiles of 145 *S. mutans* isolates from 10 countries. The 80 different *S. mutans* isolates from China were sequenced specifically for the present study, while the whole genome sequences of the other 65 international isolates were downloaded from the NCBI database. Circles correspond to the allelic profiles of individual strains based upon analyses of cgMLST target gene sequences. Numbers on lines correspond to the number of target genes for which allelic differences were detected. Circle colors are used to differentiate the country of origin of the isolates. The core genome sequence types (cgSTs) at the nodes were removed for easy visualization. Isolates with identical cgSTs that originated from China are shown with large red circles.

addition, 58.06% of the MLST profiles associated with ≥2 isolates were further differentiated using our cgMLST scheme (18/31 STs). Each child in the present study provided two isolates, and in 70% of children, these two isolates were distinguished by our cgMLST scheme, with this rate being significantly higher than that achieved with the MLST scheme (7.5%). The topologies of the phylogenetic trees generated under HKY and GTR models were highly congruent and exhibited strong bootstrap support when using our cgMLST scheme, in contrast to the trees derived from the MLST scheme. These data suggest that this *S. mutans* cgMLST scheme is of higher reproducibility than the MLST scheme in question. The consistency of the topological structure of the phylogenetic tree constructed using our cgMLST scheme and SNP methods was better than that of the phylogenetic tree constructed using the MLST scheme and SNP methods. This indicates the higher reliability of our cgMLST scheme relative to that of this MLST scheme.

Reported patterns of *S. mutans* transmission include mother-to-child transmission, father-to-child transmission, intrafamilial transmission, and child-to-child transmission (39–41). In previous studies of *S. mutans*, evidence of transmission among children and their family members was defined via MLST based on identical ST assignments (17, 40). In this study, child-to-child transmission of *S. mutans* between different kindergartens

and between different classes in the same kindergarten was observed in Bengbu, China, according to the MLST scheme. However, based upon our cgMLST scheme, 25 pairs of isolates out of 37 pairs of isolates that had shared identical ST assignments were differentiated by cgST, with our cgMLST scheme having resulted in a larger genetic distance between isolates than the MLST scheme for the majority of isolate comparisons. This emphasized the higher discriminatory power of our cgMLST scheme relative to that of the MLST scheme, suggesting that the differences in the cgMLST target genes observed pairs of isolates from a given child may be the result of intrahost and/or intraniche evolution. In addition, this indicates that our cgMLST scheme is superior to the MLST scheme for the molecular typing of closely epidemiologically related isolates. In the absence of a gold standard molecular typing method with perfect discriminatory power, epidemiological link data were used to define related and unrelated isolates, with maximal contrast between related and unrelated isolates being sought. Epidemiologically related isolates were defined as being obtained from the same patient, belonging to the same ST, and cultured within a time window of 30 days (42). The threshold for cluster types was identified when 92% of all likely epidemiologically related pairs of isolates were differentiated (43). We utilized these criteria as a reference to identify the threshold for SNP and allele differences in cgSTs in the current study. We found that 92%, 100%, and 97% of epidemiologically related pairs of isolates differed by 7 alleles in our cgMLST scheme, 7 core genome SNPs, and 10 total SNPs, respectively. We combined 65 previously published sequences and 80 freshly prepared *S. mutans* sequences to analyze the diverse population structure of *S. mutans* (44, 45), and we found that there are 37 pairs of isolates that shared ≤7 allele differences and that were thus considered to be a clonally related based on our cgMLST scheme. Among these, 1 pair of isolates originated from the United Kingdom, 1 pair of isolates originated from the United States, and the remaining 35 pairs of isolates originated from China. No close epidemiological links among isolates were observed among different countries, with the fewest allele differences among countries being observed for isolates from China and the United Kingdom (22 alleles). These differences thus reveal that the genetic diversity of *S. mutans* is complex.

Although our study provides a new molecular typing method for the epidemiologic analysis of *S. mutans*, it does have some limitations. More than half of the STs were differentiated by cgST, but there were still 12 pairs of isolates were indistinguishable via this approach. Given that the number of candidate genomes influences the percentage of core genes in a cgMLST scheme, the use of more samples to prepare a cgMLST scheme for *S. mutans* may improve this scheme and make it more practical, particularly as a means of differentiating between closely related isolates. Moreover, the small number of samples of the *S. mutans* isolates from several countries used in this study (1 from Denmark, 2 from Germany, 1 from Japan, 2 from South Africa, and 2 from Turkey) may interfere with the results of our epidemiological analysis.

Conclusion. In summary, our cgMLST scheme represents a powerful and effective tool for the epidemiological typing of *S. mutans* isolates. By leveraging this tool, it will be possible to better understand *S. mutans* population structures and dynamics, thus allowing more effective development of public health strategies centered around this bacterium.

MATERIALS AND METHODS

Construction of a primitive cgMLST target gene set. For this study, the *S. mutans* UA159 (GenBank assembly accession number GCA_000007465.2) reference genome was used to guide analysis and candidate target gene identification. Candidate genes for the cgMLST scheme were selected by discarding genes that met the following criteria using balstall 2.2.26: genes that were <50 bp, genes lacking a start or stop codon; genes with multiple stop codons or with nonsense mutations resulting in premature stop codon positioning, the shorter of any pairs of genes that overlapped by >4 bp, genes with high homology to the transposon_db TransposonPSI database (http://transposonpsi.sourceforge.net/; accessed September 2017) (identity > 50%, coverage > 70%), and/or genes with high homology with plasmid genomes (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/plasmid/; accessed April 2018) (identity > 90%, overlap > 95%) (30).

Candidate genome set construction. We identified 199 total *S. mutans* whole genomes that were available from the National Center for Biotechnology Information (NCBI) as of 1 December 2019. Genomes were filtered based on the following criteria via blastall 2.2.26: genomes with a contig number of ≥200 and a scaffold of <500 bp, genomes not containing all MLST core genes, and/or genomes having a single gene copy number of <75% (30). Both Nakano's and Do's MLST approaches demonstrate similar discrimination for *S. mutans* (46). However, only two candidate genomes were preserved after data filtering based on Do's MLST scheme. Therefore, only Nakano's MLST scheme was used in this study. Information regarding the resultant 161 candidate genomes is listed in Table S1 in the supplemental material.

cgMLST target gene identification and functional annotation. The following criteria were used to define the target genes used in our cgMLST scheme via blastall 2.2.26: keep all single copy genes identified by a BLAST search in both candidate gene set and genomes and filter the genes of the primitive gene set with default parameter of stop codon or genes that have internal stop codons in more than 20% of the candidate genomes (29, 47). Information pertaining to the resultant 594 cgMLST target genes is listed in Table S2 in the supplemental material. All candidate genomes were then typed based upon the cgMLST target genes, and these genomes were integrated with STs defined in the MLST database (29 August 2019) (https://pubmlst.org/oralstrep/) in order to conduct a minimum spanning tree analysis. We evaluated these cgMLST target genes based upon candidate genome coverage and the distribution of cgMLST target genes in the genome of *S. mutans* UA159. The Database for Annotation, Visualization and Integrated Discovery (DAVID 6.8) gene annotation tool (https://david.ncifcrf.gov/) was employed for functional annotation bioinformatics microarray analysis, including Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and gene ontology (GO) associated with target genes in this cgMLST scheme (48). All of the pathways or functional categories meeting the following cutoff criterion were extracted for analysis: $P < 0.05$.

Whole-genome sequencing and assembly. For this study, the whole genomes of 80 *S. mutans* isolates that were obtained from the supragingival dental plaque of children in Bengbu, China, were preserved in our laboratory and were sequenced and assembled (see Table S4 in the supplemental material). A TIANamp bacteria DNA kit (Tiangen Biotech, Beijing, China) was used for bacterial genomic DNA (gDNA) isolation based on provided directions, after which 1 μg of DNA from each sample was used with a NEBNext Ultra DNA library prep kit for Illumina (NEB, USA) in order to prepare libraries for sequencing. Barcodes were added to each sample during this process. Samples were then sonicated into 350-bp fragments, followed by end polishing, poly(A) tail processing, ligation to full-length Illumina adaptors, and additional PCR amplification. The AMPure XP system was then used to extract the amplified libraries, which were subjected to quality control analysis and quantification using an Agilent 2100 Bioanalyzer and via real-time PCR. An Illumina NovaSeq 6000 PE150 instrument was then used to sequence these 80 whole genomes at Novogene Bioinformatics Technology (Beijing, China), and the SOAPdenovo software was used for sequence assembly.

Comparison between MLST and cgMLST. A total of 80 sequenced samples were used to assess the resolution of this *S. mutans* cgMLST scheme by comparing the cgMLST-ST (cgST) and Nakano MLST-ST (ST). The numbers of STs and cgSTs were analyzed using a Pearson Chi-square test by SPSS 20.0 software (IBM, Armonk, NY, USA). A P value of <0.05 was considered statistically significant. Maximum likelihood phylogenetic trees were generated based on 594 cgMLST core genes and 8 MLST housekeeping genes in all 80 samples via TreeBeST under Hasegawa-Kishino-Yano (HKY) and general time-reversible (GTR) models (49–51). The reliability of the branching orders was evaluated via 1,000 bootstrap replicates (52). The Phylogenetic trees were annotated using Interactive Tree Of Life (iTOL) v5 (https://itol.embl.de/) (53). The allele profiles and cgSTs of the 199 genomes downloaded from the NCBI database and of the 80 isolates sequenced in this study are listed in Tables S5 and S6 in the supplemental material.

Whole-genome sequence analysis for SNP-based phylogeny. Each sample was compared with the reference sequence to identify the potential SNP sites using MUMmer 3.23 (54); 100-bp sequences on both sides of the reference sequence SNP sites were extracted, and then the extracted reference sequence and the assembly results were compared to verify the SNP sites. If the alignment length was less than 101 bp, it was considered to be an untrustworthy SNP and was removed. Tandem Repeats Finder 4.07b and RepeatMasker 4.0.5 were used to evaluate the repeated sequence area of the reference sequence and to filter the SNPs located in the repeated area. Finally, a reliable SNP was obtained. SNP-based phylogenetic trees were built using TreeBeST under HKY and GTR models. Confidence was inferred by running 1,000 bootstrap replicates.

cgMLST-based epidemiological analysis of *S. mutans*. A minimum spanning tree was constructed based on our cgMLST scheme using 145 *S. mutans* isolates from 10 countries (China, $n = 80$; Brazil, $n = 16$; Denmark, $n = 1$; Germany, $n = 2$; Iceland, $n = 5$; Japan, $n = 1$; South Africa, $n = 2$; Turkey, $n = 2$, United Kingdom, $n = 22$, and Unite States, $n = 14$) via Perl 5.18.2 with a maximum length of 50 and opacity of 0.5. The whole-genome sequences of the 65 isolates from other countries are listed in Table S7 in the supplemental material.

Data availability. All assembled reads of the 80 *S. mutans* clinical isolates were submitted to the National Center for Biotechnology Information (NCBI) under the BioProject accession numbers PRJNA598422 to PRJNA598424, PRJNA598427, PRJNA598791, PRJNA598793 to PRJNA598796, PRJNA598882, PRJNA598884, PRJNA598886, PRJNA598888, PRJNA598889, PRJNA600562, PRJNA600581, PRJNA600582, PRJNA600585, PRJNA600586, PRJNA600589, PRJNA600592 to PRJNA600599, PRJNA600601, PRJNA600602, PRJNA600684, PRJNA600687, PRJNA600688, PRJNA600690 to PRJNA600692, PRJNA600694 to PRJNA600697, PRJNA600699 to PRJNA600701, PRJNA600703 to PRJNA600706, PRJNA600708 to PRJNA600710, PRJNA601102, PRJNA601537 to PRJNA601539,

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

FIG S1, TIFF file, 1.4 MB.
FIG S2, TIFF file, 1.1 MB.
TABLE S1, XLSX file, 0.01 MB.
TABLE S2, XLSX file, 0.02 MB.
TABLE S3, XLSX file, 0.02 MB.
TABLE S4, XLSX file, 0.01 MB.
TABLE S5, XLSX file, 0.6 MB.
TABLE S6, XLSX file, 0.02 MB.
TABLE S7, XLSX file, 0.01 MB.

REFERENCES

1. GBD 2016 Disease and Injury Incidence and Prevalence Collaborators. 2017. Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. Lancet 390:1211–1259. https://doi.org/10.1016/S0140-6736(17)32154-2.
2. Radnaabaatar M, Kim YE, Go DS, Jung Y, Jung J, Yoon SJ. 2019. Burden of dental caries and periodontal disease in South Korea: an analysis using the national health insurance claims database. Community Dent Oral Epidemiol 47:513–519. https://doi.org/10.1111/cdoe.12493.
3. Kastenbom L, Falsen A, Larsson P, Sunnegårdh-Grönberg K, Davidson T. 2019. Costs and health-related quality of life in relation to caries. BMC Oral Health 19:187. https://doi.org/10.1186/s12903-019-0874-6.
4. Baker JL, Faustoferri RC, Quivey RG, Jr. 2017. Acid-adaptive mechanisms of *Streptococcus mutans*—the more we know, the more we don't. Mol Oral Microbiol 32:107–117. https://doi.org/10.1111/omi.12162.
5. Liu SS, Zhu WH, Zhi QH, Liu J, Wang Y, Lin HC. 2017. Analysis of sucrose-induced small RNAs in *Streptococcus mutans* in the presence of different sucrose concentrations. Appl Microbiol Biotechnol 101:5739–5748. https://doi.org/10.1007/s00253-017-8346-x.
6. Bottner A, He RY, Sarbu A, Nainar SMH, Dufour D, Gong SG, Levesque CM. 2020. *Streptococcus mutans* isolated from children with severe-early childhood caries form higher levels of persisters. Arch Oral Biol 110:104601. https://doi.org/10.1016/j.archoralbio.2019.104601.
7. Agnello M, Marques J, Cen L, Mittermuller B, Huang A, Chaichanasakul Tran N, Shi W, He X, Schroth RJ. 2017. Microbiome associated with severe caries in Canadian First Nations children. J Dent Res 96:1378–1385. https://doi.org/10.1177/0022034517718819.
8. Zhao W, Li W, Lin J, Chen Z, Yu D. 2014. Effect of sucrose concentration on sucrose-dependent adhesion and glucosyltransferase expression of S. mutans in children with severe early-childhood caries (S-ECC). Nutrients 6:3572–3586. https://doi.org/10.3390/nu6093572.
9. Shibata Y, Ozaki K, Seki M, Kawato T, Tanaka H, Nakano Y, Yamashita Y. 2003. Analysis of loci required for determination of serotype antigenicity in *Streptococcus mutans* and its clinical utilization. J Clin Microbiol 41:4107–4112. https://doi.org/10.1128/jcm.41.9.4107-4112.2003.
10. Nakano K, Nomura R, Nakagawa I, Hamada S, Ooshima T. 2004. Dem-onstration of *Streptococcus mutans* with a cell wall polysaccharide specific to a new serotype, k, in the human oral cavity. J Clin Microbiol 42:198–202. https://doi.org/10.1128/jcm.42.1.198-202.2004.
11. Caufield PW, Walker TM. 1989. Genetic diversity within *Streptococcus mutans* evident from chromosomal DNA restriction fragment polymorphisms. J Clin Microbiol 27:274–278. https://doi.org/10.1128/JCM.27.2.274-278.1989.
12. Gronroos L, Saarela M, Matto J, Tanner-Salo U, Vuorela A, Alaluusua S. 1998. Mutacin production by *Streptococcus mutans* may promote transmission of bacteria from mother to child. Infect Immun 66:2595–2600. https://doi.org/10.1128/IAI.66.6.2595-2600.1998.
13. Li Y, Caufield PW, Emanuelsson IR, Thornqvist E. 2001. Differentiation of *Streptococcus mutans* and *Streptococcus sobrinus* via genotypic and phenotypic profiles from three different populations. Oral Microbiol Immunol 16:16–23. https://doi.org/10.1034/j.1399-302x.2001.160103.x.
14. Saarela M, Alaluusua S, Takei T, Asikainen S. 1993. Genetic diversity within isolates of mutans streptococci recognized by an rRNA gene probe. J Clin Microbiol 31:584–587. https://doi.org/10.1128/JCM.31.3.584-587.1993.
15. Moser SA, Mitchell SC, Ruby JD, Momeni S, Osgood RC, Whiddon J, Childers NK. 2010. Repetitive extragenic palindromic PCR for study of *Streptococcus mutans* diversity and transmission in human populations. J Clin Microbiol 48:599–602. https://doi.org/10.1128/JCM.01828-09.
16. Jiang Q, Yu M, Min Z, Yi A, Chen D, Zhang Q. 2012. AP-PCR detection of *Streptococcus mutans* and *Streptococcus sobrinus* in caries-free and caries-active subjects. Mol Cell Biochem 365:159–164. https://doi.org/10.1007/s11010-012-1255-5.
17. Nakano K, Lapirattanakul J, Nomura R, Nemoto H, Alaluusua S, Gronroos L, Vaara M, Hamada S, Ooshima T, Nakagawa I. 2007. *Streptococcus mutans* clonal variation revealed by multilocus sequence typing. J Clin Microbiol 45:2616–2625. https://doi.org/10.1128/JCM.02343-06.
18. Do T, Gilbert SC, Clark D, Ali F, Fatturi Parolo CC, Maltz M, Russell RR, Holbrook P, Wade WG, Beighton D. 2010. Generation of diversity in *Streptococcus mutans* genes demonstrated by MLST. PLoS One 5:e9073. https://doi.org/10.1371/journal.pone.0009073.
19. Kimura B. 2018. Will the emergence of core genome MLST end the role

of in silico MLST? Food Microbiol 75:28–36. https://doi.org/10.1016/j.fm.2017.09.003.

20. Patiño LH, Camargo M, Muñoz M, Ríos-Chaparro DI, Patarroyo MA, Ramírez JD. 2018. Unveiling the multilocus sequence typing (MLST) schemes and core genome phylogenies for genotyping *Chlamydia trachomatis*. Front Microbiol 9:1854. https://doi.org/10.3389/fmicb.2018.01854.

21. Savin C, Criscuolo A, Guglielmini J, Le Guern AS, Carniel E, Pizarro-Cerda J, Brisse S. 2019. Genus-wide *Yersinia* core-genome multilocus sequence typing for species identification and strain characterization. Microb Genom 5:e000301. https://doi.org/10.1099/mgen.0.000301.

22. Ghanem M, El-Gazzar M. 2018. Development of *Mycoplasma synoviae* (MS) core genome multilocus sequence typing (cgMLST) scheme. Vet Microbiol 218:84–89. https://doi.org/10.1016/j.vetmic.2018.03.021.

23. Sankarasubramanian J, Vishnu US, Gunasekaran P, Rajendhran J. 2019. Development and evaluation of a core genome multilocus sequence typing (cgMLST) scheme for *Brucella spp*. Infect Genet Evol 67:38–43. https://doi.org/10.1016/j.meegid.2018.10.021.

24. Chen Y, Gonzalez-Escalona N, Hammack TS, Allard MW, Strain EA, Brown EW. 2016. Core genome multilocus sequence typing for identification of globally distributed clonal groups and differentiation of outbreak strains of *Listeria monocytogenes*. Appl Environ Microbiol 82:6258–6272. https://doi.org/10.1128/AEM.01532-16.

25. Stenmark B, Hellmark B, Soderquist B. 2019. Genomic analysis of *Staphylococcus capitis* isolated from blood cultures in neonates at a neonatal intensive care unit in Sweden. Eur J Clin Microbiol Infect Dis 38:2069–2075. https://doi.org/10.1007/s10096-019-03647-3.

26. Giske CG, Dyrkell F, Arnellos D, Vestberg N, Hermansson Panna S, Froding I, Ullberg M, Fang H. 2019. Transmission events and antimicrobial susceptibilities of methicillin-resistant *Staphylococcus argenteus* in Stockholm. Clin Microbiol Infect 25:1289.e5–1289.e8. https://doi.org/10.1016/j.cmi.2019.06.003.

27. Chen Y, Hong J, Chen Y, Wang H, Yu Y, Qu T. 2020. Characterization of a community-acquired methicillin-resistant sequence type 338 *Staphylococcus aureus* strain containing a staphylococcal cassette chromosome *mec* type V$_T$. Int J Infect Dis 90:181–187. https://doi.org/10.1016/j.ijid.2019.10.034.

28. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, Harmsen D, Mellmann A. 2015. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Listeria monocytogenes*. J Clin Microbiol 53:2869–2876. https://doi.org/10.1128/JCM.01193-15.

29. Ghanem M, Wang L, Zhang Y, Edwards S, Lu A, Ley D, El-Gazzar M. 2017. Core genome multilocus sequence typing: a standardized approach for molecular typing of *Mycoplasma gallisepticum*. J Clin Microbiol 56:e01145-17. https://doi.org/10.1128/JCM.01145-17.

30. Zhou H, Liu W, Qin T, Liu C, Ren H. 2017. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Klebsiella pneumoniae*. Front Microbiol 8:371. https://doi.org/10.3389/fmicb.2017.00371.

31. Bialek-Davenet S, Criscuolo A, Ailloud F, Passet V, Jones L, Delannoy-Vieillard A-S, Garin B, Le Hello S, Arlet G, Nicolas-Chanoine M-H, Decré D, Brisse S. 2014. Genomic definition of hypervirulent and multidrug-resistant *Klebsiella pneumoniae* clonal groups. Emerg Infect Dis 20:1812–1820. https://doi.org/10.3201/eid2011.140206.

32. Wu J, Fan Y, Wang X, Jiang X, Zou J, Huang R. 2020. Effects of the natural compound, oxyresveratrol, on the growth of *Streptococcus mutans*, and on biofilm formation, acid production, and virulence gene expression. Eur J Oral Sci 128:18–26. https://doi.org/10.1111/eos.12667.

33. Brown AT, Wittenberger CL. 1972. Fructose-1,6-diphosphate-dependent lactate dehydrogenase from a cariogenic streptococcus: purification and regulatory properties. J Bacteriol 110:604–615. https://doi.org/10.1128/JB.110.2.604-615.1972.

34. Song L, Wang W, Conrads G, Rheinberg A, Sztajer H, Reck M, Wagner-Döbler I, Zeng A-P. 2013. Genetic variability of *mutans* streptococci revealed by wide whole-genome sequencing. BMC Genomics 14:430. https://doi.org/10.1186/1471-2164-14-430.

35. Hossain MS, Biswas I. 2012. SMU.152 acts as an immunity protein for mutacin IV. J Bacteriol 194:3486–3494. https://doi.org/10.1128/JB.00194-12.

36. Liu S, Li H, Guo Z, Guan J, Sun Y, Zhang K. 2019. Insight into the effect of small RNA srn225147 on mutacin IV in *Streptococcus mutans*. Indian J Microbiol 59:445–450. https://doi.org/10.1007/s12088-019-00820-2.

37. Hossain MS, Biswas I. 2012. An extracelluar protease, SepM, generates

38. functional competence-stimulating peptide in *Streptococcus mutans* UA159. J Bacteriol 194:5886–5896. https://doi.org/10.1128/JB.01381-12.

39. Momeni SS, Whiddon J, Moser SA, Cheon K, Ruby JD, Childers NK. 2013. Comparative genotyping of *Streptococcus mutans* by repetitive extragenic palindromic polymerase chain reaction and multilocus sequence typing. Mol Oral Microbiol 28:18–27. https://doi.org/10.1111/omi.12002.

39. Lapirattanakul J, Nakano K, Nomura R, Hamada S, Nakagawa I, Ooshima T. 2008. Demonstration of mother-to-child transmission of *Streptococcus mutans* using multilocus sequence typing. Caries Res 42:466–474. https://doi.org/10.1159/000170588.

40. Momeni SS, Whiddon J, Moser SA, Childers NK. 2018. Transmission patterns of *Streptococcus mutans* demonstrated by a combined rep-PCR and MLST approach. Clin Oral Invest 22:2847–2858. https://doi.org/10.1007/s00784-018-2371-8.

41. Momeni SS, Whiddon J, Cheon K, Ghazal T, Moser SA, Childers NK. 2016. genetic diversity and evidence for transmission of *Streptococcus mutans* by DiversiLab rep-PCR. J Microbiol Methods 128:108–117. https://doi.org/10.1016/j.mimet.2016.07.010.

42. Kluytmans-van den Bergh MF, Rossen JW, Bruijning-Verhagen PC, Bonten MJ, Friedrich AW, Vandenbroucke-Grauls CM, Willems RJ, Kluytmans JA. 2016. Whole-genome multilocus sequence typing of extended-spectrum-beta-lactamase-producing *Enterobacteriaceae*. J Clin Microbiol 54:2919–2927. https://doi.org/10.1128/JCM.01648-16.

43. de Been M, Pinholt M, Top J, Bletz S, Mellmann A, van Schaik W, Brouwer E, Rogers M, Kraat Y, Bonten M, Corander J, Westh H, Harmsen D, Willems RJ. 2015. Core genome multilocus sequence typing scheme for high-resolution typing of *Enterococcus faecium*. J Clin Microbiol 53:3788–3797. https://doi.org/10.1128/JCM.01946-15.

44. Cornejo OE, Lefebure 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. 2013. Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. Mol Biol Evol 30:881–893. https://doi.org/10.1093/molbev/mss278.

45. Argimon S, Konganti K, Chen H, Alekseyenko AV, Brown S, Caufield PW. 2014. Comparative genomics of oral isolates of *Streptococcus mutans* by in silico genome subtraction does not reveal accessory DNA associated with severe early childhood caries. Infect Genet Evol 21:269–278. https://doi.org/10.1016/j.meegid.2013.11.003.

46. Momeni SS, Whiddon J, Cheon K, Moser SA, Childers NK. 2015. Assessment of two multilocus sequence typing (MLST) schemes available for *Streptococcus mutans*. Arch Oral Biol 60:1769–1776. https://doi.org/10.1016/j.archoralbio.2015.09.012.

47. Janowicz A, De Massis F, Ancora M, Camma C, Patavino C, Battisti A, Prior K, Harmsen D, Scholz H, Zilli K, Sacchini L, Di Giannatale E, Garofolo G. 2018. Core genome multilocus sequence typing and single nucleotide polymorphism analysis in the epidemiology of *Brucella melitensis* infections. J Clin Microbiol 56:e00517-18. https://doi.org/10.1128/JCM.00517-18.

48. Huang da W, Sherman BT, Lempicki RA. 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37:1–13. https://doi.org/10.1093/nar/gkn923.

49. Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52:696–704. https://doi.org/10.1080/10635150390235520.

50. Xiao B, Sun YF, Lian B, Chen TM. 2016. Complete genome sequence and comparative genome analysis of the *Paenibacillus mucilaginosus* K02. Microb Pathog 93:194–203. https://doi.org/10.1016/j.micpath.2016.01.016.

51. Chung YS, Choi JY, Yoo MS, Seong JH, Choi BS, Kang C. 2019. Phylogenetic transmission clusters among newly diagnosed antiretroviral drug-naive patients with human immunodeficiency virus-1 in Korea: a study from 1999 to 2012. PLoS One 14:e0217817. https://doi.org/10.1371/journal.pone.0217817.

52. Ljubin-Sternak S, Meštrović T, Ivković-Jureković I, Kolarić B, Slović A, Forčić D, Tot T, Mijač M, Vraneš J. 2019. The emerging role of rhinoviruses in lower respiratory tract infections in children—clinical and molecular epidemiological study from Croatia, 2017–2019. Front Microbiol 10:2737. https://doi.org/10.3389/fmicb.2019.02737.

53. Letunic I, Bork P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Res 47:W256–W259. https://doi.org/10.1093/nar/gkz239.

54. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12. https://doi.org/10.1186/gb-2004-5-2-r12.