AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

Novel Multilocus Sequence Typing and Global Sequence Clustering Schemes for Characterizing the Population Diversity of *Streptococcus mitis*

Akuzike Kalizang'oma,[a,b] Brenda Kwambana-Adams,[a,b] Jia Mun Chan,[a] Aishwarya Viswanath,[c] Andrea Gori,[a] Damien Richard,[d] Keith A. Jolley,[e] John Lees,[f] David Goldblatt,[g] Sandra Beleza,[h] Stephen D. Bentley,[i] Robert S. Heyderman,[a] Chrispin Chaguza[a,i,j,k,l]

[a]NIHR Mucosal Pathogens Research Unit, Division of Infection and Immunity, University College London, London, United Kingdom

[b]Malawi-Liverpool-Wellcome Programme, Blantyre, Malawi

[c]UCL Division of Medicine, University College London, London, United Kingdom

[d]UCL Genetics Institute, University College London, London, United Kingdom

[e]Department of Zoology, University of Oxford, Oxford, United Kingdom

[f]Pathogen Informatics and Modelling, European Bioinformatics Institute, Hinxton, United Kingdom

[g]University College London, Great Ormond Street Institute of Child Health, London, United Kingdom

[h]University of Leicester, Department of Genetics and Genome Biology, Leicester, United Kingdom

[i]Parasites and Microbes, Wellcome Sanger Institute, Hinxton, United Kingdom

[j]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, Yale University, New Haven, Connecticut, USA

[k]Department of Clinical Infection, Microbiology and Immunology, University of Liverpool, Liverpool, United Kingdom

[l]Yale Institute for Global Health, Yale University, New Haven, Connecticut, USA

Robert S. Heyderman, and Chrispin Chaguza contributed equally to this study.

ABSTRACT *Streptococcus mitis* is a common oral commensal and an opportunistic pathogen that causes bacteremia and infective endocarditis; however, the species has received little attention compared to other pathogenic streptococcal species. Effective and easy-to-use molecular typing tools are essential for understanding bacterial population diversity and biology, but schemes specific for *S. mitis* are not currently available. We therefore developed a multilocus sequence typing (MLST) scheme and defined sequence clusters or lineages of *S. mitis* using a comprehensive global data set of 322 genomes (148 publicly available and 174 newly sequenced). We used internal 450-bp sequence fragments of seven housekeeping genes (*accA*, *gki*, *hom*, *oppC*, *patB*, *rlmN*, and *tsf*) to define the MLST scheme and derived the global *S. mitis* sequence clusters using the PopPUNK clustering algorithm. We identified an initial set of 259 sequence types (STs) and 258 global sequence clusters. The schemes showed high concordance (100%), capturing extensive *S. mitis* diversity with strains assigned to multiple unique STs and global sequence clusters. The tools also identified extensive within- and between-host *S. mitis* genetic diversity among isolates sampled from a cohort of healthy individuals, together with potential transmission events, supported by both phylogeny and pairwise single nucleotide polymorphism (SNP) distances. Our novel molecular typing and strain clustering schemes for *S. mitis* allow for the integration of new strain data, are electronically portable at the PubMLST database (https://pubmlst.org/smitis), and offer a standardized approach to understanding the population structure of *S. mitis*. These robust tools will enable new insights into the epidemiology of *S. mitis* colonization, disease and transmission.

KEYWORDS *Streptococcus mitis*, multilocus sequence typing, PopPUNK sequence clustering, population diversity, within-host diversity, transmission

*S*treptococcus mitis is an abundant oral commensal that can escape the oral niche to cause invasive disease (1, 2). *S. mitis* is a dominant cause of infective endocarditis

Address correspondence to Akuzike Kalizang'oma, akuzike.kalizang'oma.18@ucl.ac.uk, or Chrispin Chaguza, c.chaguza@ucl.ac.uk.

(1), a serious multisystem disease associated with high mortality that approaches 25 to 30% at 1 year postdiagnosis despite optimal treatment (3, 4). Also causing bacteremia among immunosuppressed patients undergoing chemotherapy (5–7), *S. mitis* infections are often associated with antimicrobial resistance (AMR) (2, 8).

However, understanding *S. mitis* naso-oropharyngeal carriage, transmission, and disease remains challenging due to the unreliable species identification methods frequently used, which fail to either differentiate *S. mitis* from other viridans group streptococci or provide a clear picture of population structure (9). For example, early studies investigated *S. mitis* biovar 1 genetic diversity using ribotyping, which uses restriction endonuclease to digest DNA followed by separation of the digested DNA fragments by electrophoresis on agarose gel (10–12). Other nonculture methods were also used to define genotypes based on sequencing of a limited number of genes such as the single glucose-6-phosphate dehydrogenase gene (13). Both nonculture methods suggested extensive diversity of *S. mitis* between and within individuals (10, 12, 13); however, accurate identification to the species level is uncertain as *Streptococcus oralis* isolates have often been misidentified as *S. mitis* biovar 1 (14).

Several pioneering *S. mitis* studies that relied on traditional microbiology and biochemistry methods (14–17) may have misidentified *S. mitis* (18), and previous studies have highlighted the challenges in accurately differentiating *S. mitis* and *S. oralis* based on the analysis of 16S rRNA gene amplicon sequencing data (19). More recently, matrix-assisted laser desorption ionization–time of flight mass spectrometry (MALDI-TOF MS) has also shown low precision in differentiating *S. mitis* from other closely related species (20, 21). A combination of the culture-based methods and whole-genome sequencing (WGS) offers an unambiguous method for identifying bacterial species and defining population structure with high resolution (22) and has distinguished clinical strains of *S. mitis* from other mitis group streptococci with high accuracy (23).

Although the application of WGS for analysis of population structure has provided insights into the genomic epidemiology for numerous bacterial species (24–27), standard bioinformatics tools and molecular schemes for characterizing *S. mitis* genetic diversity lag behind other clinically relevant bacterial species. These tools and schemes include multilocus sequence typing (MLST), which provides a simple and portable method for identifying bacterial clones, typically using seven housekeeping genes (28, 29). To date, MLST schemes have been defined for 130 bacterial species (29), including 12 *Streptococcus* species, such as *Streptococcus pneumoniae* (29, 30).

To complement MLST and other molecular typing approaches, newer methods, such as PopPUNK, have been developed, which takes into account variation in both core and accessory genomes to population structure (31). The PopPUNK algorithm is applicable to different species, able to distinguish very similar strains at a high resolution, and generates reproducible sequence clusters (31). This is in contrast to previously widely used methods, such as Bayesian Analysis of Population Structure (BAPS) (32, 33), which did not maintain consistent cluster naming upon integration of new strain data, which resulted in ambiguous definition of bacterial clades between laboratories (31). The application of PopPUNK to study *S. pneumoniae* has revealed global pneumococcal lineages and facilitated surveillance of carriage and disease strains, AMR, and the impact of serotype-specific conjugate vaccines (34).

The availability of a dedicated MLST scheme and PopPUNK sequence clusters for *S. mitis* would provide a standard method to define its population structure and to understand its genetic diversity and epidemiology.

Here, we describe the development and application of MLST and PopPUNK sequence clustering schemes for *S. mitis* to a comprehensive data set of confirmed publicly available and newly sequenced *S. mitis* isolates obtained from multiple sources and clinical contexts. Overall, our findings showed high concordance of the methods and captured the extensive genetic diversity of the *S. mitis* species at both individual and population levels, highlighting the utility of these methods for understanding the population genetics and epidemiology of this neglected bacterial species.

MATERIALS AND METHODS

Genome selection and identification to the species level. A comprehensive data set of publicly available *S. mitis* genomes was obtained together with additional isolates sequenced from a carriage study (see below) to develop the *S. mitis* MLST and global genomic clustering schemes (see Table S1 in the supplemental material). Genome quality was determined using the quality assessment tool for genome assemblies (QUAST v.5.0.2) using default settings (35). Taxonomic classification of *S. mitis* was initially done using KRAKEN v.1.0 (22) against the MiniKraken DB_8GB database. Default settings were used, and the taxonomic labels were converted to reports using KRAKEN postprocessing scripts (https://ccb.jhu.edu/software/kraken/). Species-level identification was also confirmed using KRAKEN v.2.1.2 (36) against the minikraken2_v2_8GB_201904 database. Default settings were used, and species-level reports were generated on running the initial command. The online PathogenWatch Speciator tool for assigning species to an assembled genome was also used (https://pathogen.watch/), and the in-house species identification tool applied MASH to search a curated NCBI RefSeq database. Finally, GTDB-Tk v.2.1.0 (37) was used together with the reference data version r207 to classify the species. Default parameters were used following the GTDB-Tk manual. Genomic analyses were conducted using Wellcome Sanger Institute (WSI) and University College London (UCL) computing clusters.

Development of the *S. mitis* MLST scheme. To define the MLST scheme, *S. mitis* genomes were first annotated using Prokka v.1.13.4 (38), and pangenome analysis was then conducted using the annotated *S. mitis* genomes to obtain core genes present in 100% of the isolates using Panaroo v.1.2.8 (39). The core genes were screened to identify and select 7 housekeeping genes evolving under negative selection. GenomegaMap (40) was used to determine the evolutionary pressures on protein-coding regions (nonsynonymous/synonymous substitution [dN/dS] ratios) for core genes with names and known functions. Seven housekeeping genes evolving under negative selection (dN/dS ratio of <1) were selected for the MLST scheme, and 450-bp internal fragments for each gene were used to define gene alleles. The seven genes were also selected for consistency with most of the MLST schemes. The MLST genes were analyzed within the Molecular Evolutionary Genetics Analysis (MEGA) software v.10.0 (41) and visualized on the *S. mitis* B6 reference genome (GenBank accession no. GCA_000027165.1) using DNAPlotter v.18.2.0 (42). Nucleotide-BLAST v.2.10.1 was also used to rule out duplication of the seven housekeeping genes among confirmed *S. mitis* genomes. Unique alleles were assigned arbitrary numbers, and the sequence type (ST) was determined by the combination of alleles at the seven housekeeping gene loci (*accA*, *gki*, *hom*, *oppC*, *patB*, *rlmN*, and *tsf*). Simpson's diversity index (SDI) was used to measure the diversity of the alleles for each MLST gene using an online tool (http://www.comparingpartitions.info/) (43). The MLST software (https://github.com/tseemann/mlst) was then used together with the *S. mitis* scheme to define STs of *S. mitis* isolates (29), and the BIGSdb BURST analysis tool was used to group related STs (29).

Inference of population structure. An alignment of polymorphic sites was generated from the core genome alignment using Snp-Sites v.2.5.1 (44). The isolates were clustered into subpopulations or sequence clusters (SCs) using hierBAPS as part of the Bayesian analysis of population structure (BAPS) v.6.0 software (32, 33).

Establishing global *S. mitis* sequence clusters. *S. mitis* global sequence clusters were established by applying PopPUNK on the *S. mitis* genome data set. Core and accessory genomes were determined using default PopPUNK parameters (31); however, the maximum thresholds for core and accessory distances were increased due to extensive *S. mitis* diversity that was incorrectly identified as contamination for all genomes in the data set. The Bayesian-Gaussian mixture model was applied and refined using the core and accessory distances as it best suited the population structure of the data set.

Application of genomic tools to *S. mitis* carriage isolates. A prospective pilot carriage study was conducted among healthy adults in November 2019 to obtain *S. mitis* isolates that would test the robustness and discriminatory power of the developed genomic tools. Healthy adults above the age of 18 who were UCL students or staff were eligible. Current or recent antibiotic use within 2 weeks prior to sampling was an exclusion criterion. In total, 12 healthy adults were recruited.

Briefly, nasopharyngeal, oral, and oropharyngeal samples were obtained. Nasopharyngeal swab samples were taken with Copan FLOQswabs (Copan, USA) using a previously described technique (45). Oral swab samples were taken using Copan FLOQswabs, and to ensure adequate sampling of regions known to be colonized by *S. mitis* (12), the surface of the teeth, dorsum of the tongue, hard palate, left and right buccal mucosa, and maxillary and mandibular gingiva were sampled with the same swab. Both oral and nasopharyngeal swabs were stored in skim milk-tryptone-glucose-glycerol (STGG) medium. Cough samples were obtained from volunteers directly onto sterile Columbia blood agar (CBA) plates. Additional respiratory samples were collected using a facemask device with a soluble polyvinyl alcohol (PVA) inner strip as previously described (46), and PVA strips were dissolved in Todd-Hewitt broth with yeast extract (THY) culture medium prior to inoculation.

Microbiology and molecular screening. Volumes of 10 μL of STGG and THY were used to inoculate CBA plates, which were incubated together with the direct cough plates for 18 h in 5% CO_2 and at 37°C. Alpha-hemolytic colonies were tested for optochin resistance, and a maximum of 6 alpha-hemolytic colonies were subcultured. DNA was obtained from purified isolates using heat lysis, screened using the *S. mitis pheA* PCR primers, and visualized via electrophoresis as previously described (47) to obtain potential *S. mitis* isolates. *S. mitis* NCTC 12261 (PHE, United Kingdom) was used as the positive control.

DNA extraction and whole-genome sequencing. Extracted genomic DNA for whole-genome sequencing (WGS) was obtained from pure overnight plate cultures that were *pheA* PCR positive using the Qiagen DNeasy blood and tissue kit and following the manufacturer's instructions (Qiagen, Germany). DNA quantification using Qubit fluorometric quantification (Thermo Fisher, United Kingdom), genomic DNA library preparation, and WGS were carried out by the UCL Pathogen Genomics Unit (PGU). The Illumina NextSeq

platform (Illumina, San Diego, CA, USA) was used for WGS, which generated paired-end sequence reads of 150 bp in length and 50 to 100× coverage.

Genome assembly. Illumina sequencing reads were checked for quality using FastQC (https://www .bioinformatics.babraham.ac.uk/projects/fastqc/). Sequenced read adapter trimming was done using Trimmomatic v.0.39 (48), and a phred score of at least 33 per read was used as the minimum quality score threshold. *De novo* genome assembly was performed using SPAdes v.3.12 (49), and genome quality was determined using the quality assessment tool for genome assemblies (QUAST) v.5.0.2 (35).

Sequence mapping and pairwise SNP distances. Genetic diversity of *S. mitis* isolated within and between the volunteers was investigated through mapping of sequence reads, single nucleotide polymorphism (SNP) distances calculated between isolates, and phylogenetic reconstruction. Snippy v.4.6.0 (https://github.com/tseemann/snippy) was used to map confirmed *S. mitis* sequence reads to the *S. mitis* B6 reference genome (GenBank accession no. GCA_000027165.1) and obtain a core genome SNP alignment, which was used to determine genetic diversity. Pairwise SNP distances were also calculated between the *S. mitis* strains using snp-dist v.0.7.0 (https://github.com/tseemann/snp-dists) and the *S. mitis* SNP alignment. The pairwise SNP distance matrix was visualized as a heat map in R v.2.11.1 (50).

Phylogeny and population diversity. The *S. mitis* SNP alignment of polymorphic sites was then used to construct maximum likelihood phylogenies using fasttree v.2.1.10 (51). The generalized time-reversible model of nucleotide evolution was used to generate the phylogenies, which were then visualized and annotated using the online Interactive Tree of Life (iToL) software v.3.0 (52). Isolates were then clustered into subpopulations using hierBAPS as part of the Bayesian analysis of population structure (BAPS) v.6.0 software (32), STs (29), and PopPUNK global sequence clusters (31), which were then visualized in microreact v.5.93.0 (53).

Antimicrobial resistance genotyping. AMR genes were identified among the *S. mitis* isolates using ARIBA v.2.14.6 (54), and the ResFinder database was used as a reference for AMR genes (55).

Experimental validation of the *S. mitis* MLST scheme. Amplification primers for the seven MLST genes (*accA*, *gki*, *hom*, *oppC*, *patB*, *rlmN*, and *tsf*) were designed using Primer3 (56). *In silico* PCR using the ipcress tool v.0.8.3 was first used to predict amplification for all *S. mitis* genomes, closely related species, and non-*Streptococcus* species. Default settings were used for *in silico* PCR amplifications (https://www .ebi.ac.uk/about/vertebrate-genomics/software/ipcress-manual). Thirty (*n* = 30) *S. mitis* isolates were then selected for experimental validation of the MLST scheme, where isolates were selected based on the rarity of alleles and the isolation condition. *S. mitis* NCTC 12261 (SK142; type strain) was used as a positive control (57). Selected isolates are listed in Table S1.

The 7 MLST internal fragments were amplified from genomic DNA using a commercially available PCR master mix per the manufacturer's instructions (OneTaq Quick Load 2× master mix; NEB). The primers are listed in Table S2, and the PCR cycling conditions used are as follows: 1 cycle of denaturation at 94°C for 30 s, followed by 35 cycles of amplification at 94°C for 15 s, 54°C for 15 s, and 68°C for 30 s, and then 1 cycle of final extension at 68°C for 5 min. Further optimization of the annealing temperature may be required to ensure that only a single DNA fragment is sequenced. PCR products were purified and Sanger sequenced in both directions by LGC Genomics. The paired sequences were merged using EMBOSS merger (58) and visualized in MEGAX, and PubMLST (https://pubmlst.org/smitis) was used to define alleles and STs (29).

Ethics. Ethical approval from UCL Research Ethics Committee (UCL REC) was obtained for the prospective pilot carriage study (UCL REC no. 15101/001). Informed written consent was obtained from all volunteers at the time of recruitment.

Data availability. Genome sequences generated in this project are available under BioProject no. PRJEB55310. Publicly available genomes used in this project are available under BioProject no. PRJNA480039, PRJEB42564, PRJEB42963, and PRJEB53188. Accession numbers for all genomes used are also listed in Table S1. MLST nucleotide sequences generated from Sanger sequencing are available under accession no. OP792763 to OP792980.

RESULTS

Genome selection and identification to the species level. Initial taxonomic screening of the genome assemblies using KRAKEN v.1.0 determined that all 322 genomes (100%) were *S. mitis*. Additional screening using KRAKEN v.2.1.2 also identified *S. mitis* among 322 genomes (100%) (see Tables S3 to S6 in the supplemental material). Identification to the species level using the MASH approach in PathogenWatch determined 299 out of 322 genomes (93%) were *S. mitis* (Table S7), while 21 out of 322 (7%) were unclassified *Streptococcus* species. GTDB-Tk determined 212 out of 322 (66%) were *S. mitis* (Table S8), while 48 out of 322 (15%) were unclassified *Streptococcus* species. The taxonomic screening tools use different approaches and reference databases (22, 36, 37); therefore, variation in the results was expected (Table S9). We proceeded with genomes identified using KRAKEN v.1.0 and v.2.1.2 due to consistency and to maximize capturing *S. mitis* genetic diversity.

Molecular typing of *S. mitis* using a novel MLST scheme. We used 322 *S. mitis* genomes to develop an initial MLST scheme for *S. mitis*, where assembly quality was established using QUAST (Tables S10 and S11). Among the total 322 genomes, 158 (49.1%) were from carriage, 138 (42.9%) were from infective endocarditis, 13 (4.0%) were
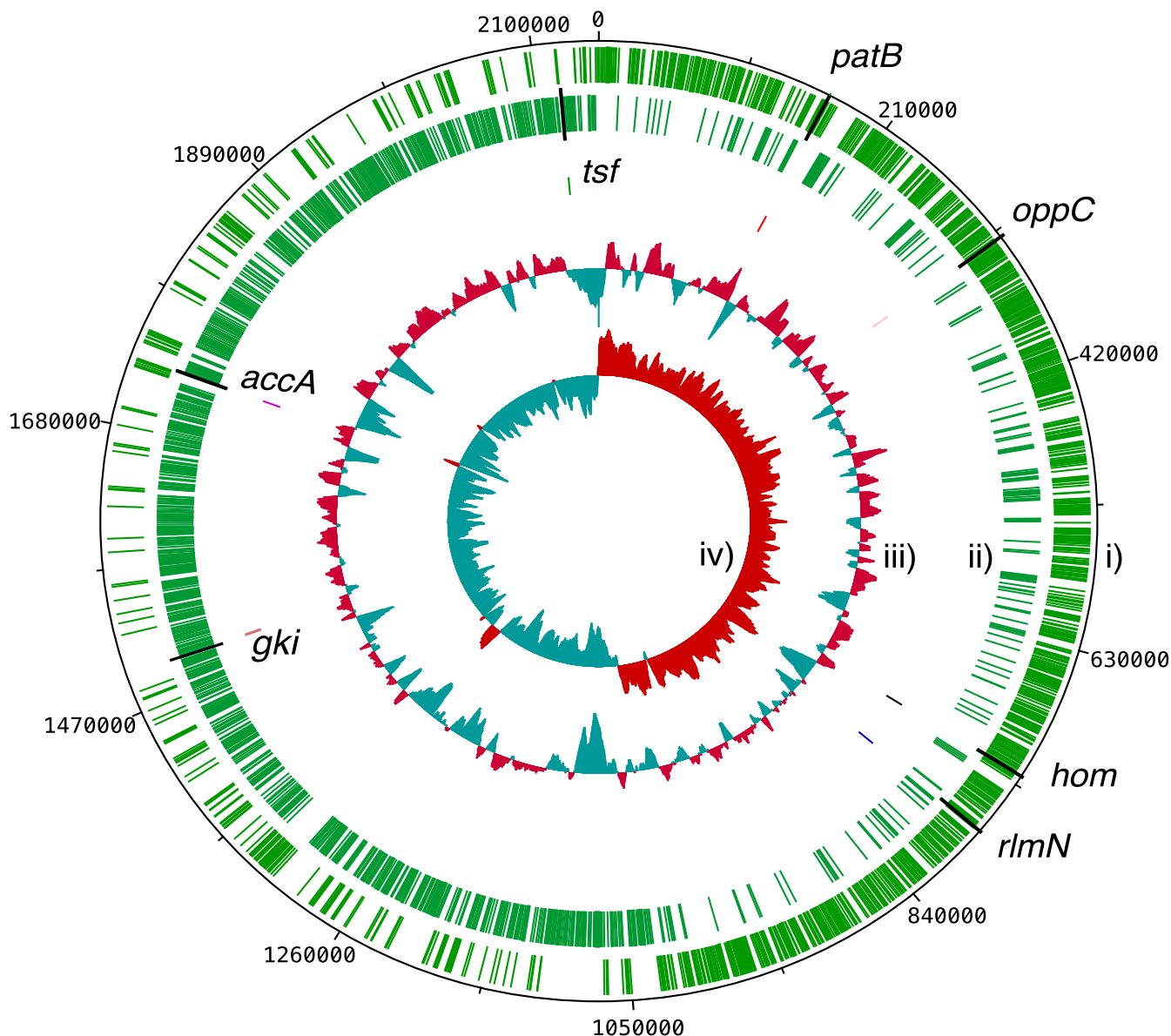
FIG 1 Location of the selected seven MLST genes on the *S. mitis* B6 reference genome (GenBank accession no. GCA_000027165.1). The plot shows *S. mitis* B6 as a circular genome, and the genome size in base pairs is shown on the outermost numbered region. The tracks from the outside represent the (i) forward and (ii) reverse coding sequence (CDS), (iii) GC content, and (iv) GC skew. Above average and below average values for GC content and GC skew are shown in red and blue, respectively. The locations of the MLST genes are indicated on the forward or reverse CDS.

from bacteremia, 12 (3.7%) were from unknown sources, and 1 (0.3%) was from pneumonia (Table S1). The genomes included clinically relevant isolates to obtain a scheme that can be applied in the clinical setting. From the total number of 1,183 core *S. mitis* genes obtained through the pangenome analysis, seven housekeeping genes with established names, known functions, and evolving under negative selection were selected to define the *S. mitis* MLST scheme: acetyl coenzyme A (acetyl-CoA) carboxylase carboxyl transferase subunit alpha (*accA*), glucose kinase (*gki*), homoserine dehydrogenase (*hom*), ABC transporter permease (*oppC*), pyridoxal phosphate-dependent aminotransferase (*patB*), rRNA large subunit methyltransferase N (*rlmN*), and translation elongation factor Ts (*tsf*). These genes have important roles in lipid metabolism (*accA*), amino acid biosynthesis (*hom*), transmembrane transport (*oppC* and *patB*), protein biosynthesis (*tsf*), glucose metabolism (*gki*), and rRNA and tRNA processing (*rlmN*) (59). The locations of the genes on the *S. mitis* B6 reference genome (GenBank accession no. GCA_000027165.1) are shown in Fig. 1.

TABLE 1 Genetic characterization of the seven housekeeping genes selected for the *S. mitis* MLST scheme

| Gene	Gene length (bp)	GC content (%)	*dN/dS* ratio (ω)	No. of alleles	Simpson's diversity	95% confidence interval
accA	768	43	0.0766	176	0.987	0.982–0.992
gki	924	45	0.0511	174	0.983	0.977–0.989
hom	1,287	42	0.0655	176	0.984	0.978–0.990
oppC	927	35	0.0364	142	0.966	0.956–0.975
patB	1,206	41	0.1116	195	0.991	0.987–0.995
rlmN	1,104	40	0.0246	184	0.987	0.981–0.992
tsf	1,041	41	0.0711	144	0.981	0.974–0.987

The MLST genes were characterized to determine gene length, GC content, average evolutionary pressures on protein-coding regions (*dN/dS*), number of alleles identified, and allele diversity (Table 1). The average *dN/dS* ratios for the genes were less than 1, indicating purifying selection making them ideal for use in the typing scheme. Allele diversity calculated using Simpson's diversity index (SDI) revealed extensive diversity among the alleles as all values were ~1, which suggested that genetic variation in these genes would be sufficient to capture the genetic relatedness of *S. mitis* isolates.

Evolutionary pressures on protein-coding regions (*dN/dS* ratios) for the seven housekeeping genes (*accA*, *gki*, *hom*, *oppC*, *patB*, *rlmN*, and *tsf*) were calculated along the entire gene length, including the 450-bp internal fragments used for the scheme, which largely indicated a lack of positive or neutral selection (Fig. 2A). The MLST scheme was then applied to the data set of 322 *S. mitis* isolates to identify similar and diverse STs, and a total of 259 different STs were resolved. The BIGSdb BURST analysis tool identified 11 groups of closely related STs; however, the founding (ancestral) genotype of the clonal complex could not be predicted due to the low frequency of STs per group. Figure 2B shows the total number of unique STs and clonal groups. MLST allows for the integration of new genomic data; therefore, as additional *S. mitis* isolates are analyzed more STs and clonal complexes will be identified. The *S. mitis* MLST scheme is available on PubMLST (https://pubmlst.org/smitis) (29).

Defining *S. mitis* global sequence clusters. To establish the global *S. mitis* sequence clusters, a pangenome analysis was conducted using the *S. mitis* data set, and totals of 1,183 core and 9,315 accessory genes were identified. Bayesian Analysis of Population Structure (BAPS) was initially applied to the *S. mitis* core genome SNP alignment to investigate the population structure of the species. However, BAPS generates inconsistent sequence cluster names and the method could only resolve 6 sequence clusters despite the highly diverse *S. mitis* data set (Fig. S1); therefore, BAPS was not adopted. In contrast, PopPUNK defines clusters with consistent names when used with the same standard database; therefore, we applied PopPUNK to established global *S. mitis* sequence clusters.

The population of *S. mitis* is represented by a network where a node represents an individual isolate using PopPUNK (Fig. 3A), and nodes that are closer to the origin represent within-strain relationships (Fig. 3B). Comparison of core and accessory genetic distances among *S. mitis* genomes in the data set revealed that isolates were highly diverse as shown by clustering based on greater core and accessory genetic distances (Fig. 3A and B), with a few highly similar isolates that were likely from the same individuals sampled in the small cohort with lesser core and accessory genetic distances. A large cluster of 23 isolates from global cluster 1 were from the United States (NCBI repository); however, isolate metadata were unavailable; therefore, it was unknown whether the isolates were from the same individual or individuals living within close proximity to each other. Overall, PopPUNK identified 258 global sequence clusters, and the genetic distances based on the accessory genomes are shown in Fig. 3C. The PopPUNK *S. mitis* reference database for inferring *S. mitis* global sequence clusters is publicly available (https://poppunk.net/).

High degree of within- and between-host *S. mitis* genetic diversity among carriage isolates confirmed by MLST. To demonstrate the ability of the genomic tools to resolve *S. mitis* genetic diversity, we applied the tools through the prospective pilot carriage study that sought to assess *S. mitis* genotypes that are carried among a cohort
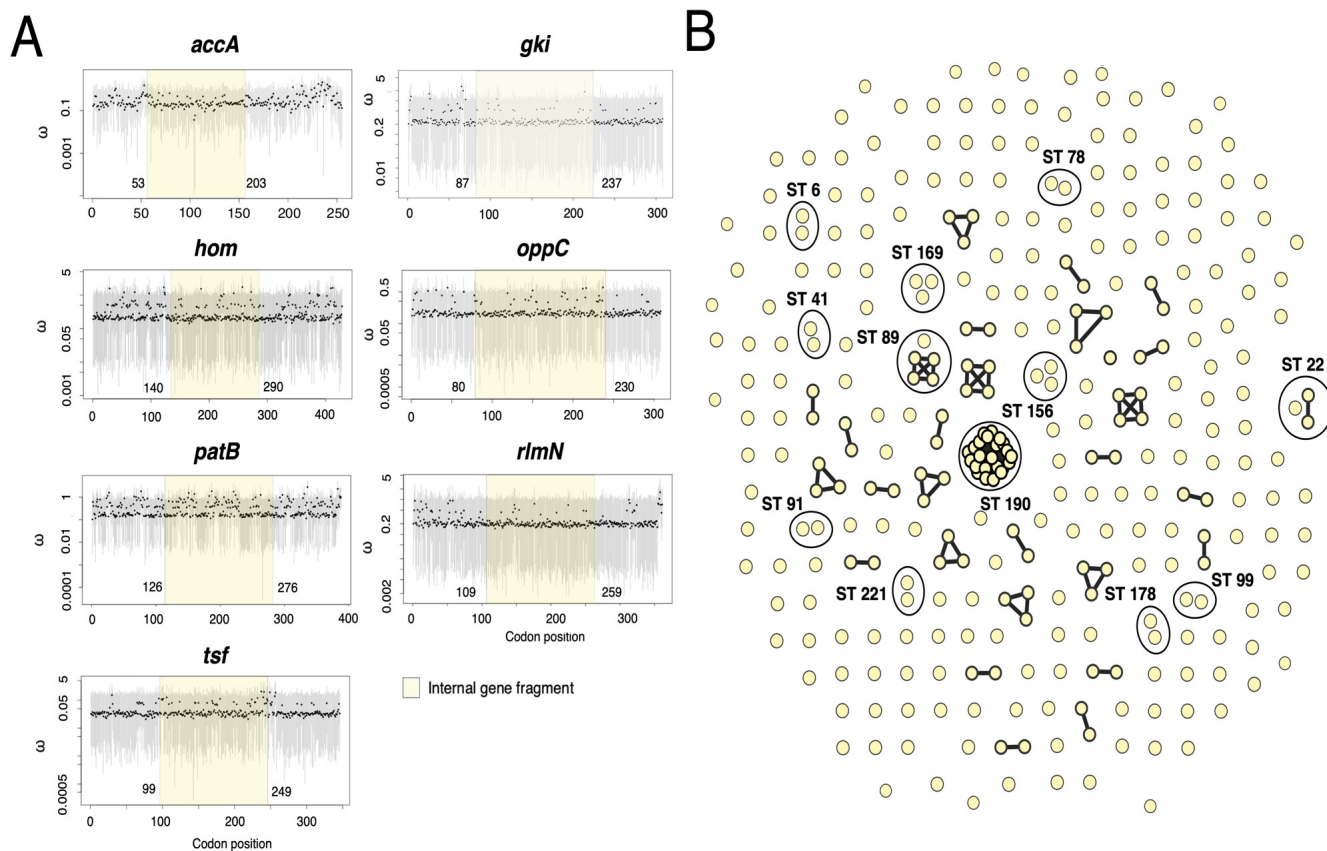
FIG 2 Development and application of the novel *S. mitis* MLST scheme. (A) Estimates of the *dN/dS* (ω) ratios among seven *S. mitis* housekeeping genes. *dN/dS* point estimates determined by GenomegaMap are shown by black points, and 95% credibility intervals are shown by gray bars for codon positions along each gene. The internal fragments selected to define alleles are highlighted. (B) Sequence types and clonal groups identified using the novel *S. mitis* MLST scheme. Unlinked nodes represent unique STs, while linked nodes belong to the same ST. Isolates assigned to the same clonal group are shown by ellipses, which have been assigned the name of one ST found within the group.

of 12 healthy volunteers (age range, 18 to 60 years). We obtained a total of 49 confirmed *S. mitis* isolates from the volunteers and sequenced a median of 4 genomes per individual (range, 1 to 7). Sequencing quality is shown in Fig. S2, and contamination was checked among the assemblies as previously described (Tables S5 and S6). Since the calculation of the pairwise SNP analysis required 2 or more isolates, we excluded volunteers 7 and 11 as they had only one sequenced *S. mitis* isolate each. The median number of SNPs between pairs of isolates among the 10 volunteers with more than one *S. mitis* isolate ranged from 92 to 58,466 bp (Fig. 4A). We considered isolates with an SNP distance of <50 bp to be the same strain, and therefore 7 out of 10 volunteers each had one or more isolates that were potentially the same strain. Apart from volunteer 1, who had isolates that were all closely related based on SNP distance (13 to 146 SNPs), the majority of volunteers with more than one isolate (9 out of 10) had diverse isolates with median pairwise SNP distances of more than 50,000 bp, which suggested high within-host genetic diversity.

We next constructed a maximum likelihood phylogenetic tree and pairwise SNP distance matrix from the whole-genome sequence alignment of the isolates to investigate within- and between-host genetic diversity (Fig. 4B). We applied the devised *S. mitis* MLST scheme to determine STs for the *S. mitis* isolates. In total, we found 31 unique STs among the 49 *S. mitis* isolates obtained from carriage and shedding, which suggested a highly genetically diverse *S. mitis* population among asymptomatic carriers. Each sampled individual carried a median of three unique STs, but the number of STs ranged from 1 to 5. The STs determined by the *S. mitis* MLST scheme were supported by pairwise SNP distances as few pairwise SNP distances between isolates were observed for the same ST (Fig. 4A). For example, the four *S. mitis* STs isolated from
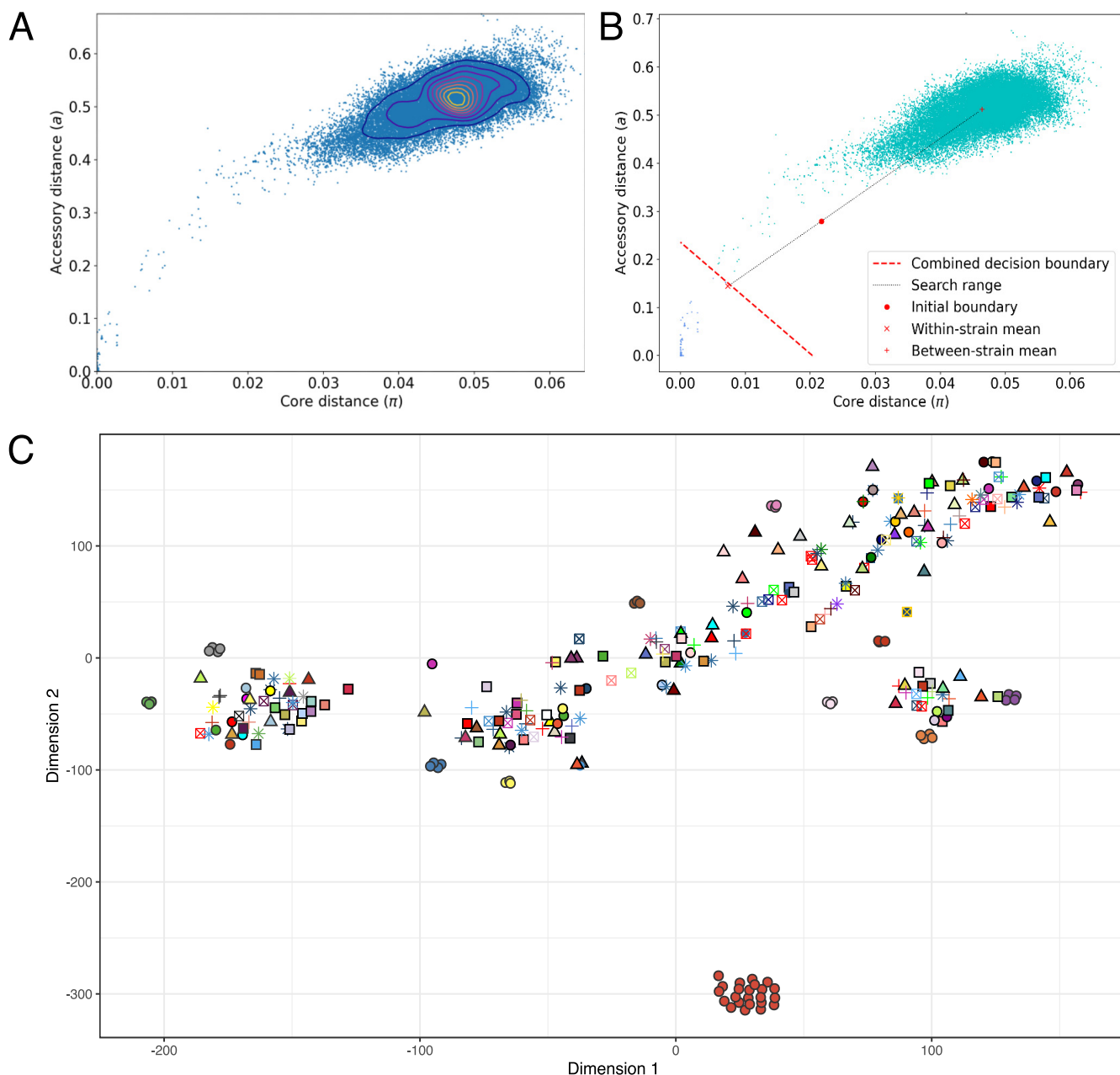
FIG 3 Global sequence cluster analysis of *S. mitis* genomes using PopPUNK. (A) Scatter plot based on core and accessory distances among all 322 *S. mitis* genomes in the database using PopPUNK. A node represents an individual isolate. (B) PopPUNK model fitting and refinement using the two-dimensional Gaussian mixture model. The threshold or combined decision boundary for defining within-strain relationships is then refined using a network score in order to generate a sparse but highly clustered network. (C) Accessory distances of *S. mitis* genomes produced using t-SNE (71). Isolates of the same color and shape that group together demonstrate the sequence clustering algorithm of PopPUNK using the global *S. mitis* data set.

volunteer 1 were all ST89, with a median pairwise SNP distance of 92 bp. A high degree of within-host genetic diversity was also supported by MLST and pairwise SNP distances, both methods that demonstrate the presence of multiple *S. mitis* lineages within the same host. For example, MLST identified multiple STs in volunteer 2, namely, ST126, ST61, ST234, and ST197, which had a mean pairwise SNP distance of 43,888 bp (Fig. 4B). The *S. mitis* MLST scheme also identified the same ST in different individuals. These STs included ST216 isolated from volunteers 4, 10, and 11 and ST27 isolated from volunteers 5 and 4. We found pairwise SNP distances of 21 to 23 bp and 11 to 20 bp for the ST216 and ST27 isolates sampled from different individuals, respectively, which suggested potential recent direct or indirect transmission events.
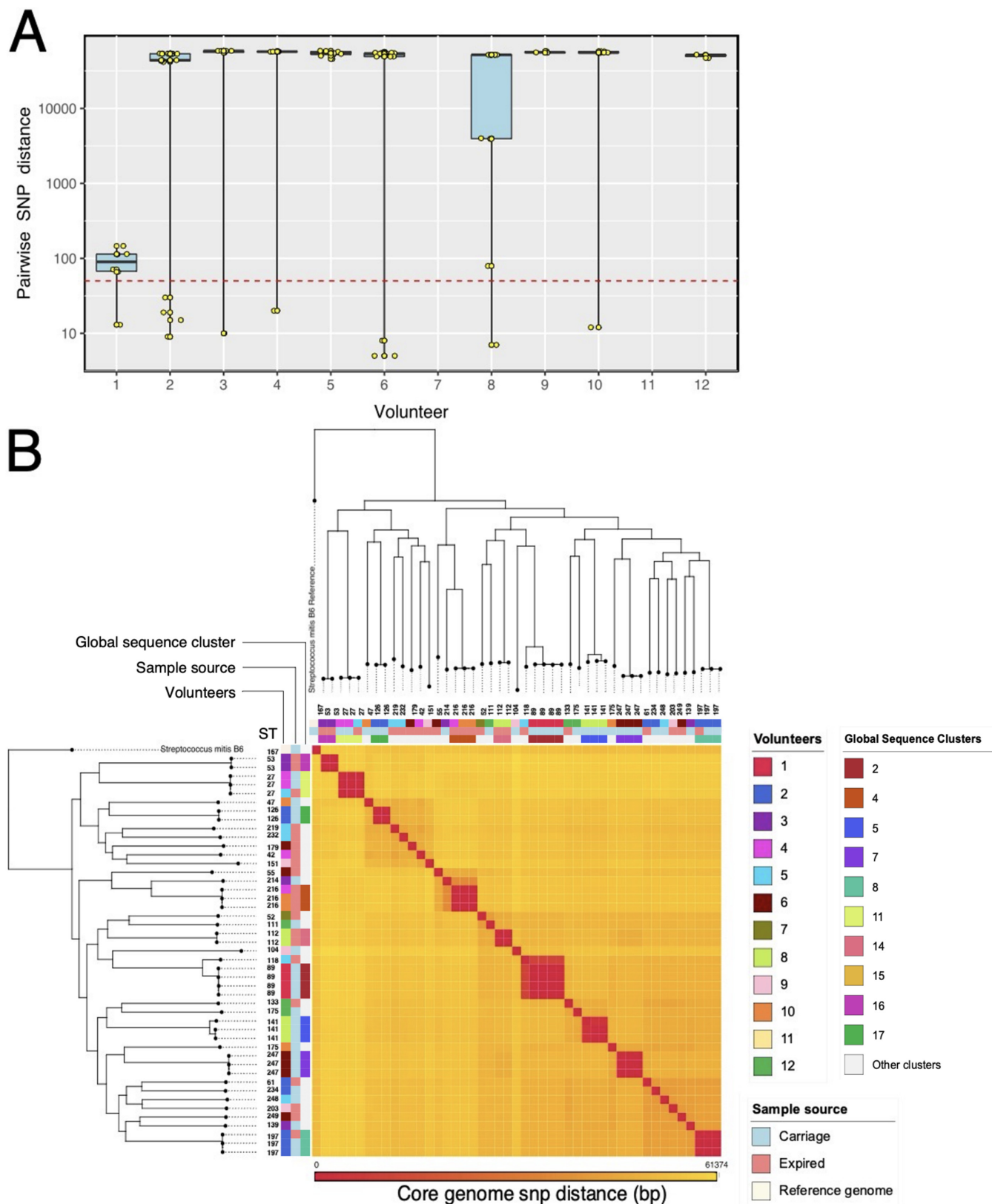
FIG 4 Within- and between-host genetic diversity of *S. mitis* isolates using pairwise SNP distances and MLST. (A) Pairwise SNP distances of *S. mitis* isolates obtained from carriage and expired respiratory secretions. The strip charts and box plots show the number of SNPs calculated between isolates from 10 volunteers with 2 or more sequenced *S. mitis* isolates. The *y* axis is shown in log$_{10}$ scale for clarity. The red dotted horizontal line marks the threshold for what we considered the same strain (<50 bp). (B) Phylogeny and heat map of SNP differences between *S. mitis* isolates. Pairwise SNP distances were

High concordance of global sequence clusters, sequence types, and AMR profiles. We investigated the genetic diversity of the sampled *S. mitis* isolates by constructing a phylogenetic tree and clustering the isolates into genetically distinct subpopulations. The maximum likelihood phylogenetic tree generated from mapped reads showed that the 49 *S. mitis* isolates obtained from the healthy individuals clustered into 31 global sequence clusters corresponding to 31 STs (Fig. 5), which illustrated a high concordance between the two typing schemes. We then determined antimicrobial resistance profiles to further support the observations of specific STs and sequence clusters. We identified a total of 4 acquired AMR gene classes among the 49 *S. mitis* isolates through mapping sequenced reads against reference genes from the ResFinder AMR database. Identified AMR genes included those that confer resistance to chloramphenicol, macrolide, and tetracycline antibiotics. The AMR genes were also concordant among isolates of the same ST and sequence cluster. These findings suggested that the MLST and global sequence clustering schemes were robust, and their integration in WGS analysis pipelines would ensure standardized molecular typing of *S. mitis* isolates to identify similar groups of isolates at local and global scales.

Validation of the *S. mitis* MLST scheme. We validated the *S. mitis* MLST scheme using *in silico* and conventional PCR methods. *In silico* amplification for the 7 MLST genes was observed for all 322 *S. mitis* isolates using the designed primers (Table S2). Among the 19 different species screened, *in silico* amplification for all 7 MLST genes was observed for *S. mitis*, *S. pneumoniae*, and *S. pseudopneumoniae* reference genomes (Table S12), which was supported by conventional PCR (Fig. S3). While biochemical and molecular testing can contribute to differentiating *S. pneumoniae* from other closely related species (45, 60), we recommend the use of whole-genome sequencing and bioinformatic species-level determination to accurately identify *S. mitis* due to the high genetic similarity with other closely related mitis group species.

We carried out Sanger sequencing for the amplified PCR products of 30 *S. mitis* isolates obtained from carriage and disease, and we compared the alleles obtained from Illumina and Sanger sequencing. Among the 30 isolates, there was 100% concordance for *accA*, *gki*, *hom*, *oppC*, *patB*, *rlmN*, *tsf*, and the corresponding STs (Tables S13 to S19). The Sanger sequencing results therefore demonstrate that the rare alleles defined using the Illumina platform are unlikely sequencing errors (Table S20).

Epidemiological application of *S. mitis* typing schemes. Finally, we investigated the potential epidemiological application of the schemes by analyzing global invasive and carriage *S. mitis* isolates. The global ML phylogeny showed that the *S. mitis* isolates that clustered into 259 STs and 258 global sequence clusters were distributed across multiple geographical regions (Fig. S1). Generally, no clear geographical clustering of STs or global sequence clusters was observed, likely due to the extensive *S. mitis* diversity and the limited data set. Among the invasive isolates, we identified two pairs of infective endocarditis isolates, all from the United Kingdom, belonging to the same sequence type and global sequence cluster, namely, ST30-GSC28 and ST36-GSC27 (highlighted in Fig. S1). However, we did not have patient-level epidemiological data; therefore, it was unknown whether these isolates were from the same individual or are common lineages in the U.K. population.

DISCUSSION

Here, we present a novel MLST scheme and PopPUNK sequence clusters for *S. mitis*. Our robust quality control procedure using the bioinformatic taxonomic classification tool (22) ensured that the genomes analyzed belonged to *S. mitis*. Therefore, the MLST and sequence clustering definitions are based on a well-curated data set. The developed strain

FIG 4 Legend (Continued)

interpreted as a heat map using the R statistical program with heat map clustering methods. Maximum likelihood SNP phylogenies across the top and the left of the heat map indicate the relatedness of the isolates. Isolate ST, volunteer origin, sampling source, and global sequence cluster are adjacent to phylogeny. SNP distances ranged from 0 to 61,374 bp and correspond to a color gradient ranging from red (lowest distance value) to yellow (highest distance value).
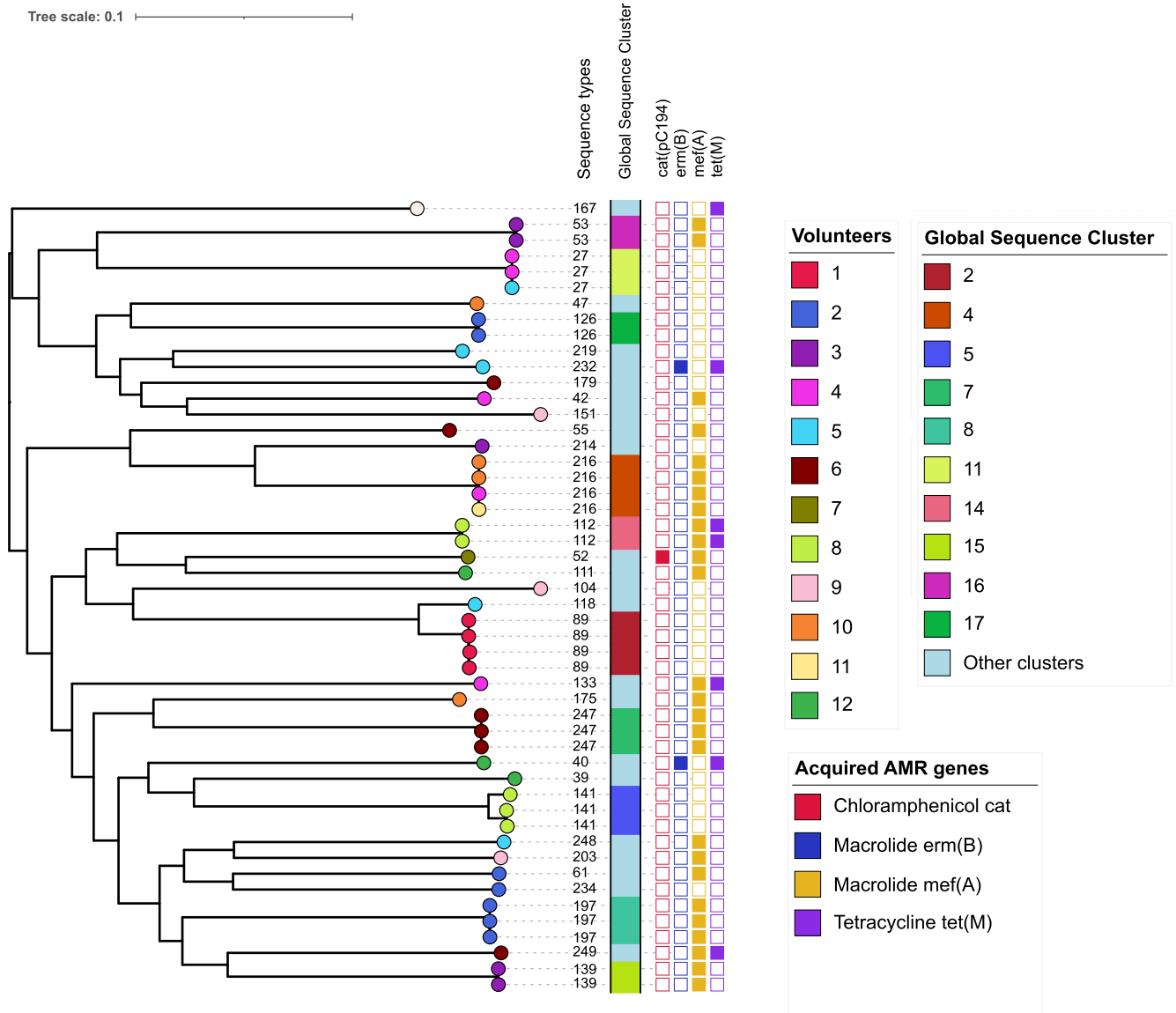
FIG 5 Maximum likelihood core genome phylogeny of *S. mitis* isolates obtained from healthy volunteers. The ML phylogeny of *S. mitis* isolates was constructed using SNPs from a mapping alignment of sequencing reads to *S. mitis* B6, and the tree was visualized in iTOL. The ML phylogeny demonstrates genetic similarity and diversity among the *S. mitis* isolates. The colored tips of the phylogeny show the volunteers, and other isolate metadata, namely, ST, global sequence cluster, and antimicrobial resistance genes, are shown. The tree was rooted at the midpoint of the branch separating the most genetically diverse genomes.

typing methods showed high discriminatory power when applied to *S. mitis* isolates sampled from a cohort of healthy individuals. Crucially, through their integration with the widely used and publicly available MLST and global cluster databases, these genomic schemes pave the way for easily defining new *S. mitis* strain types (29, 31). Our validated *S. mitis* MLST PCR primers can also be used together with Sanger sequencing in resource-limited settings to investigate the relatedness of clinical *S. mitis* strains. The tools developed here will facilitate *S. mitis* genomic surveillance as done for other species, such as *S. pneumoniae* (34) and *Salmonella enterica* (61).

  S. mitis is a highly genetically diverse species, sometimes seen as a species complex (13, 62, 63). Therefore, distinguishing different *S. mitis* strains is critical to understanding its epidemiology, population structure, and evolution within hosts and at the population level. MLST has been able to characterize the considerable population diversity of *Helicobacter pylori* (64–66), a similarly highly diverse species. Our MLST and PopPUNK global sequence clustering schemes adequately captured the genetic diversity of the species with high

resolution. We show that the *S. mitis* MLST and sequence clustering schemes were highly concordant, and they inferred unique STs and global sequence clusters that highlighted the remarkable genetic diversity of the species. Although our analysis was based on the largest collection available to date, the high genetic diversity revealed in this study of 322 *S. mitis* genomes represents only the tip of the iceberg.

To assess the utility of the genomic tools in resolving the genetic relatedness of *S. mitis* isolates, our MLST and global clustering approaches were applied to carriage isolates obtained from 12 healthy individuals. Our results showed a high degree of within- and between-host genetic diversity of *S. mitis* isolates, with a range of 1 to 5 distinct STs being found within individuals. These results suggest that carriage of multiple *S. mitis* lineages is common in healthy individuals, and the observed high within-host genetic diversity is similar to that seen in other pathogens, such as *H. pylori* (67–69). Therefore, our findings emphasize that sampling a single *S. mitis* isolate from an individual is not sufficient to capture the within-host diversity of the species. The sequence types defined based on the MLST scheme were supported by pairwise SNP distances between strains and phylogeny, and all the isolates that differed by <150 bp clustered together in the phylogeny, shared the same genotypic AMR profiles, and belonged to the same ST. Although the presence of multiple distinct *S. mitis* lineages among individuals has previously been described using 1 to 4 housekeeping genes (13, 62, 70), the use of a larger number of housekeeping genes in our scheme will provide a higher resolution to detect more cocolonization events. We have also shown that the MLST and PopPUNK schemes can potentially be used together with core genome SNP comparisons to obtain a high discriminatory power that may be sufficient to study the transmission of *S. mitis* isolates. Three individuals from the prospective pilot carriage study carried ST261 strains belonging to global cluster 4, while ST27 strains belonging to global cluster 11 were identified in 2 individuals. The pairwise SNP distances for the ST261 and ST27 strains ranged between 21 and 23 bp and 11 and 20 bp, respectively, which provided further support for potential transmission, particularly considering the high genetic diversity of the *S. mitis* species.

Overall, the *S. mitis* MLST and PopPUNK global sequence clustering schemes developed here offer a robust and standardized approach for molecular typing to identify genetically similar strains at local and global contexts. Both tools are highly portable and publicly available for immediate use and allow for the integration of new genomic data as they become available, making them particularly useful for *S. mitis* surveillance to better understand disease, carriage, and transmission. Our genomic tools can be integrated into WGS analysis pipelines to investigate population diversity of the *S. mitis* species.

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.
SUPPLEMENTAL FILE 1, PDF file, 1.1 MB.
SUPPLEMENTAL FILE 2, XLSX file, 0.5 MB.

ACKNOWLEDGMENTS

REFERENCES

1. Chamat-Hedemand S, Dahl A, Østergaard L, Arpi M, Fosbøl E, Boel J, Oestergaard LB, Lauridsen TK, Gislason G, Torp-Pedersen C, Bruun NE. 2020. Prevalence of infective endocarditis in streptococcal bloodstream infections is dependent on streptococcal species. Circulation 142:720–730. https://doi.org/10.1161/CIRCULATIONAHA.120.046723.

2. Husain E, Whitehead S, Castell A, Thomas EE, Speert DP. 2005. Viridans streptococci bacteremia in children with malignancy: relevance of species identification and penicillin susceptibility. Pediatr Infect Dis J 24:563–566. https://doi.org/10.1097/01.inf.0000164708.21464.03.

3. Prendergast BD. 2006. The changing face of infective endocarditis. Heart 92:879–885. https://doi.org/10.1136/hrt.2005.067256.

4. Murdoch DR, Corey GR, Hoen B, Miró JM, Fowler VG, Bayer AS, Karchmer AW, Olaison L, Pappas PA, Moreillon P, Chambers ST, Chu VH, Falcó V, Holland DJ, Jones P, Klein JL, Raymond NJ, Read KM, Tripodi MF, Utili R, Wang A, Woods CW, Cabell CH, International Collaboration on Endocarditis-Prospective Cohort Study (ICE-PCS) Investigators. 2009. Clinical presentation, etiology and outcome of infective endocarditis in the 21st Century: the International Collaboration on Endocarditis-Prospective Cohort Study. Arch Intern Med 169:463–473. https://doi.org/10.1001/archinternmed.2008.603.

5. Shelburne SA, Sahasrabhojane P, Saldana M, Yao H, Su X, Horstmann N, Thompson E, Flores AR. 2014. Streptococcus mitis strains causing severe clinical disease in cancer patients. Emerg Infect Dis 20:762–771. https://doi.org/10.3201/eid2005.130953.

6. Sahasrabhojane P, Galloway-Peña J, Velazquez L, Saldaña M, Horstmann N, Tarrand J, Shelburne SA. 2014. Species-level assessment of the molecular basis of fluoroquinolone resistance among viridans group streptococci causing bacteraemia in cancer patients. Int J Antimicrob Agents 43:558–562. https://doi.org/10.1016/j.ijantimicag.2014.01.031.

7. Lyytikäinen O, Rautio M, Carlson P, Anttila V-J, Vuento R, Sarkkinen H, Kostiala A, Väisänen M-L, Kanervo A, Ruutu P. 2004. Nosocomial bloodstream infections due to viridans streptococci in haematological and non-haematological patients: species distribution and antimicrobial resistance. J Antimicrob Chemother 53:631–634. https://doi.org/10.1093/jac/dkh159.

8. Renneberg J, Niemann LL, Gutschik E. 1997. Antimicrobial susceptibility of 278 streptococcal blood isolates to seven antimicrobial agents. J Antimicrob Chemother 39:135–140. https://doi.org/10.1093/oxfordjournals.jac.a020858.

9. Coykendall AL. 1989. Classification and identification of the viridans streptococci. Clin Microbiol Rev 2:315–328. https://doi.org/10.1128/CMR.2.3.315.

10. Fitzsimmons S, Evans M, Pearce C, Sheridan MJ, Wientzen R, Bowden G, Cole MF. 1996. Clonal diversity of Streptococcus mitis biovar 1 isolates from the oral cavity of human neonates. Clin Diagn Lab Immunol 3:517–522. https://doi.org/10.1128/cdli.3.5.517-522.1996.

11. Hohwy J, Reinholdt J, Kilian M. 2001. Population dynamics of Streptococcus mitis in its natural habitat. Infect Immun 69:6055–6063. https://doi.org/10.1128/IAI.69.10.6055-6063.2001.

12. Kirchherr JL, Bowden GH, Richmond DA, Sheridan MJ, Wirth KA, Cole MF. 2005. Clonal diversity and turnover of Streptococcus mitis bv. 1 on shedding and nonshedding oral surfaces of human infants during the first year of life. Clin Diagn Lab Immunol 12:1184–1190. https://doi.org/10.1128/CDLI.12.10.1184-1190.2005.

13. Bek-Thomsen M, Tettelin H, Hance I, Nelson KE, Kilian M. 2008. Population diversity and dynamics of Streptococcus mitis, Streptococcus oralis, and Streptococcus infantis in the upper respiratory tracts of adults, determined by a nonculture strategy. Infect Immun 76:1889–1896. https://doi.org/10.1128/IAI.01511-07.

14. Pearce C, Bowden GH, Evans M, Fitzsimmons SP, Johnson J, Sheridan MJ, Wientzen R, Cole MF. 1995. Identification of pioneer viridans streptococci in the oral cavity of human neonates. J Med Microbiol 42:67–72. https://doi.org/10.1099/00222615-42-1-67.

15. Smith DJ, Anderson JM, King WF, van Houte J, Taubman MA. 1993. Oral streptococcal colonization of infants. Oral Microbiol Immunol 8:1–4. https://doi.org/10.1111/j.1399-302x.1993.tb00535.x.

16. Lucas VS, Beighton D, Roberts GJ. 2000. Composition of the oral streptococcal flora in healthy children. J Dent 28:45–50. https://doi.org/10.1016/s0300-5712(99)00048-2.

17. Könönen E. 2000. Development of oral bacterial flora in young children. Ann Med 32:107–112. https://doi.org/10.3109/07853890009011759.

18. Friedrichs C, Rodloff AC, Chhatwal GS, Schellenberger W, Eschrich K. 2007. Rapid identification of viridans streptococci by mass spectrometric discrimination. J Clin Microbiol 45:2392–2397. https://doi.org/10.1128/JCM.00556-07.

19. Chen CC, Teng LJ, Chang TC. 2004. Identification of clinically relevant viridans group streptococci by sequence analysis of the 16S-23S ribosomal DNA spacer region. J Clin Microbiol 42:2651–2657. https://doi.org/10.1128/JCM.42.6.2651-2657.2004.

20. Isaksson J, Rasmussen M, Nilson B, Stadler LS, Kurland S, Olaison L, Ek E, Herrmann B. 2015. Comparison of species identification of endocarditis associated viridans streptococci using rnpB genotyping and 2 MALDI-TOF systems. Diagn Microbiol Infect Dis 81:240–245. https://doi.org/10.1016/j.diagmicrobio.2014.12.007.

21. Lee M, Chung H-S, Moon H-W, Lee SH, Lee K. 2015. Comparative evaluation of two matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS) systems, Vitek MS and Microflex LT, for the identification of Gram-positive cocci routinely isolated in clinical microbiology laboratories. J Microbiol Methods 113:13–15. https://doi.org/10.1016/j.mimet.2015.03.020.

22. Wood DE, Salzberg SL. 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol 15:R46. https://doi.org/10.1186/gb-2014-15-3-r46.

23. Rasmussen LH, Dargis R, Højholt K, Christensen JJ, Skovgaard O, Justesen US, Rosenvinge FS, Moser C, Lukjancenko O, Rasmussen S, Nielsen XC. 2016. Whole genome sequencing as a tool for phylogenetic analysis of clinical strains of Mitis group streptococci. Eur J Clin Microbiol Infect Dis 35:1615–1625. https://doi.org/10.1007/s10096-016-2700-2.

24. Walker TM, Ip CL, Harrell RH, Evans JT, Kapatai G, Dedicoat MJ, Eyre DW, Wilson DJ, Hawkey PM, Crook DW, Parkhill J, Harris D, Walker AS, Bowden R, Monk P, Smith EG, Peto TE. 2013. Whole-genome sequencing to delineate Mycobacterium tuberculosis outbreaks: a retrospective observational study. Lancet Infect Dis 13:137–146. https://doi.org/10.1016/S1473-3099(12)70277-3.

25. Croucher NJ, Finkelstein JA, Pelton SI, Mitchell PK, Lee GM, Parkhill J, Bentley SD, Hanage WP, Lipsitch M. 2013. Population genomics of post-vaccine changes in pneumococcal epidemiology. Nat Genet 45:656–663. https://doi.org/10.1038/ng.2625.

26. Vaughn EL, Vo QT, Vostok J, Stiles T, Lang A, Brown CM, Klevens RM, Madoff L. 2020. Linking epidemiology and whole-genome sequencing to investigate Salmonella outbreak, Massachusetts, USA, 2018. Emerg Infect Dis 26:1538–1541. https://doi.org/10.3201/eid2607.200048.

27. Greig DR, Schaefer U, Octavia S, Hunter E, Chattaway MA, Dallman TJ, Jenkins C. 2018. Evaluation of whole-genome sequencing for identification and typing of Vibrio cholerae. J Clin Microbiol 56:e00831-18. https://doi.org/10.1128/JCM.00831-18.

28. Maiden MCJ, Bygraves JA, Feil E, Morelli G, Russell JE, Urwin R, Zhang Q, Zhou J, Zurth K, Caugant DA, Feavers IM, Achtman M, Spratt BG. 1998. Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc Natl Acad Sci U S A 95:3140–3145. https://doi.org/10.1073/pnas.95.6.3140.

29. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.

30. Enright MC, Spratt BG. 1998. A multilocus sequence typing scheme for Streptococcus pneumoniae: identification of clones associated with serious invasive disease. Microbiology (Reading) 144:3049–3060. https://doi.org/10.1099/00221287-144-11-3049.

31. Lees JA, Harris SR, Tonkin-Hill G, Gladstone RA, Lo SW, Weiser JN, Corander J, Bentley SD, Croucher NJ. 2019. Fast and flexible bacterial genomic epidemiology with PopPUNK. Genome Res 29:304–316. https://doi.org/10.1101/gr.241455.118.

32. Cheng L, Connor TR, Sirén J, Aanensen DM, Corander J. 2013. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. Mol Biol Evol 30:1224–1228. https://doi.org/10.1093/molbev/mst028.

33. Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J. 2018. RhierBAPS: an R implementation of the population clustering algorithm hierBAPS. Wellcome Open Res 3:93. https://doi.org/10.12688/wellcomeopenres.14694.1.

34. Gladstone RA, Lo SW, Lees JA, Croucher NJ, van Tonder AJ, Corander J, Page AJ, Marttinen P, Bentley LJ, Ochoa TJ, Ho PL, Du Plessis M, Cornick JE, Kwambana-Adams B, Benisty R, Nzenze SA, Madhi SA, Hawkins PA, Everett DB, Antonio M, Dagan R, Klugman KP, von Gottberg A, McGee L, Breiman RF, Bentley SD, Global Pneumococcal Sequencing Consortium. 2019. International genomic definition of pneumococcal lineages, to contextualise disease, antibiotic resistance and vaccine impact. EBioMedicine 43:338–346. https://doi.org/10.1016/j.ebiom.2019.04.021.

35. Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013. QUAST: quality assessment tool for genome assemblies. Bioinformatics 29:1072–1075. https://doi.org/10.1093/bioinformatics/btt086.

36. Wood DE, Lu J, Langmead B. 2019. Improved metagenomic analysis with Kraken 2. Genome Biol 20:257. https://doi.org/10.1186/s13059-019-1891-0.

37. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2019. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36:1925–1927. https://doi.org/10.1093/bioinformatics/btz848.

38. Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics 30:2068–2069. https://doi.org/10.1093/bioinformatics/btu153.

39. Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C, Floto RA, Frost SDW, Corander J, Bentley SD, Parkhill J. 2020. Producing polished prokaryotic pangenomes with the Panaroo pipeline. Genome Biol 21:180. https://doi.org/10.1186/s13059-020-02090-4.

40. Wilson DJ, CRyPTIC Consortium. 2020. GenomegaMap: within-species genome-wide dN/dS estimation from over 10,000 genomes. Mol Biol Evol 37:2450–2460. https://doi.org/10.1093/molbev/msaa069.

41. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. 2018. MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096.

42. Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J. 2009. DNAPlotter: circular and linear interactive genome visualization. Bioinformatics 25:119–120. https://doi.org/10.1093/bioinformatics/btn578.

43. Hunter PR, Gaston MA. 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 26:2465–2466. https://doi.org/10.1128/jcm.26.11.2465-2466.1988.

44. Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, Keane JA, Harris SR. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microb Genom 2:e000056. https://doi.org/10.1099/mgen.0.000056.

45. Satzke C, Turner P, Virolainen-Julkunen A, Adrian PV, Antonio M, Hare KM, Henao-Restrepo AM, Leach AJ, Klugman KP, Porter BD, Sá-Leão R, Scott JA, Nohynek H, O'Brien KL, WHO Pneumococcal Carriage Working Group. 2013. Standard method for detecting upper respiratory carriage of Streptococcus pneumoniae: updated recommendations from the World Health Organization Pneumococcal Carriage Working Group. Vaccine 32:165–179. https://doi.org/10.1016/j.vaccine.2013.08.062.

46. Williams CML, Cheah ESG, Malkin J, Patel H, Otu J, Mlaga K, Sutherland JS, Antonio M, Perera N, Woltmann G, Haldar P, Garton NJ, Barer MR. 2014. Face mask sampling for the detection of Mycobacterium tuberculosis in expelled aerosols. PLoS One 9:e104921. https://doi.org/10.1371/journal.pone.0104921.

47. Park HK, Dang HT, Myung SC, Kim W. 2012. Identification of a pheA gene associated with Streptococcus mitis by using suppression subtractive hybridization. Appl Environ Microbiol 78:3004–3009. https://doi.org/10.1128/AEM.07510-11.

48. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

49. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.

50. R Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria.

51. Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One 5:e9490. https://doi.org/10.1371/journal.pone.0009490.

52. Letunic I, Bork P. 2016. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. Nucleic Acids Res 44:W242–W245. https://doi.org/10.1093/nar/gkw290.

53. Argimón S, Abudahab K, Goater RJE, Fedosejev A, Bhai J, Glasner C, Feil EJ, Holden MTG, Yeats CA, Grundmann H, Spratt BG, Aanensen DM. 2016. Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. Microb Genom 2:e000093. https://doi.org/10.1099/mgen.0.000093.

54. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA, Harris SR. 2017. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genom https://doi.org/10.1099/mgen.0.000131.

55. Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, Lund O, Aarestrup FM, Larsen MV. 2012. Identification of acquired antimicrobial resistance genes. J Antimicrob Chemother 67:2640–2644. https://doi.org/10.1093/jac/dks261.

56. Untergasser A, Cutcutache I, Koressaar T, Ye J, Faircloth BC, Remm M, Rozen SG. 2012. Primer3—new capabilities and interfaces. Nucleic Acids Res 40:e115. https://doi.org/10.1093/nar/gks596.

57. Kilian M, Tettelin H. 2019. Identification of virulence-associated properties by comparative genome analysis of Streptococcus pneumoniae, S. pseudopneumoniae, S. mitis, three S. oralis subspecies, and S. infantis. mBio 10:e01985-19. https://doi.org/10.1128/mBio.01985-19.

58. Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277. https://doi.org/10.1016/s0168-9525(00)02024-2.

59. UniProt Consortium. 2021. UniProt: the Universal Protein Knowledgebase in 2021. Nucleic Acids Res 49:D480–D489. https://doi.org/10.1093/nar/gkaa1100.

60. Tavares DA, Handem S, Carvalho RJ, Paulo AC, de Lencastre H, Hinds J, Sá-Leão R. 2019. Identification of Streptococcus pneumoniae by a real-time PCR assay targeting SP2020. Sci Rep 9:3285. https://doi.org/10.1038/s41598-019-39791-1.

61. Perez-Sepulveda BM, Heavens D, Pulford CV, Predeus AV, Low R, Webster H, Dykes GF, Schudoma C, Rowe W, Lipscombe J, Watkins C, Kumwenda B, Shearer N, Costigan K, Baker KS, Feasey NA, Hinton JCD, Hall N, Perez-Sepulveda BM, Heavens D, Pulford CV, Acuña MT, Antic D, Antonio M, Baker KS, Bernal J, Bolaños H, Chattaway M, Cheesbrough J, Chirambo A, Costigan K, Darboe S, Díaz P, Donado P, Duarte C, Duarte F, Everett D, Fanning S, Feasey NA, Feglo P, Ferreira AM, Floyd R, Gavilán RG, Gordon MA, Hall N, Hernandes RT, Hernández-Mora G, Hinton JCD, Hurley D, Kasumba IN, 10KSG consortium., et al. 2021. An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes. Genome Biol 22:349. https://doi.org/10.1186/s13059-021-02536-3.

62. Kilian M, Poulsen K, Blomqvist T, Håvarstein LS, Bek-Thomsen M, Tettelin H, Sørensen UBS. 2008. Evolution of Streptococcus pneumoniae and its close commensal relatives. PLoS One 3:e2683. https://doi.org/10.1371/journal.pone.0002683.

63. Kilian M, Riley DR, Jensen A, Brüggemann H, Tettelin H. 2014. Parallel evolution of Streptococcus pneumoniae and Streptococcus mitis to pathogenic and mutualistic lifestyles. mBio 5:e01490-14. https://doi.org/10.1128/mBio.01490-14.

64. Muñoz-Ramírez ZY, Mendez-Tenorio A, Kato I, Bravo MM, Rizzato C, Thorell K, Torres R, Aviles-Jimenez F, Camorlinga M, Canzian F, Torres J. 2017. Whole genome sequence and phylogenetic analysis show Helicobacter pylori strains from Latin America have followed a unique evolution pathway. Front Cell Infect Microbiol 7:50. https://doi.org/10.3389/fcimb.2017.00050.

65. Achtman M, Azuma T, Berg DE, Ito Y, Morelli G, Pan ZJ, Suerbaum S, Thompson SA, van der Ende A, van Doorn LJ. 1999. Recombination and clonal groupings within Helicobacter pylori from different geographical regions. Mol Microbiol 32:459–470. https://doi.org/10.1046/j.1365-2958.1999.01382.x.

66. Jiang X, Xu Z, Zhang T, Li Y, Li W, Tan H. 2021. Whole-genome-based Helicobacter pylori geographic surveillance: a visualized and expandable webtool. Front Microbiol 12:687259. https://doi.org/10.3389/fmicb.2021.687259.

67. Ailloud F, Didelot X, Woltemate S, Pfaffinger G, Overmann J, Bader RC, Schulz C, Malfertheiner P, Suerbaum S. 2019. Within-host evolution of Helicobacter pylori shaped by niche-specific adaptation, intragastric migrations and selective sweeps. Nat Commun 10:2273. https://doi.org/10.1038/s41467-019-10050-1.

68. Israel DA, Salama N, Krishna U, Rieger UM, Atherton JC, Falkow S, Peek RM. 2001. Helicobacter pylori genetic diversity within the gastric niche of a single human host. Proc Natl Acad Sci U S A 98:14625–14630. https://doi.org/10.1073/pnas.251551698.

69. Jackson LK, Potter B, Schneider S, Fitzgibbon M, Blair K, Farah H, Krishna U, Bedford T, Peek RM, Jr, Salama NR. 2020. Helicobacter pylori diversification during chronic infection within a single host generates sub-populations with distinct phenotypes. PLoS Pathog 16:e1008686. https://doi.org/10.1371/journal.ppat.1008686.

70. Whatmore AM, Efstratiou A, Pickerill AP, Broughton K, Woodard G, Sturgeon D, George R, Dowson CG. 2000. Genetic relationships between clinical isolates of Streptococcus pneumoniae, Streptococcus oralis, and Streptococcus mitis: characterization of "atypical" pneumococci and organisms allied to S. mitis harboring S. pneumoniae virulence factor-encoding genes. Infect Immun 68:1374–1382. https://doi.org/10.1128/IAI.68.3.1374-1382.2000.

71. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. 2011. Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830.