Genome Biology and Evolution

SMBE

Phylogenomic and MALDI-TOF MS Analysis of *Streptococcus sinensis* HKU4^T Reveals a Distinct Phylogenetic Clade in the Genus *Streptococcus*

Jade L.L. Teng[1,2,3,4,†], Yi Huang[1,†], Herman Tse[1,2,3,4], Jonathan H.K. Chen[1], Ying Tang[1], Susanna K.P. Lau[1,2,3,4,*], and Patrick C.Y. Woo[1,2,3,4,*]

[1]Department of Microbiology, The University of Hong Kong, Hong Kong, China

[2]Research Centre of Infection and Immunology, The University of Hong Kong, Hong Kong, China

[3]State Key Laboratory of Emerging Infectious Diseases, The University of Hong Kong, Hong Kong, China

[4]Carol Yu Centre for Infection, The University of Hong Kong, Hong Kong, China

*Corresponding author: E-mail: skplau@hku.hk; pcywoo@hku.hk.

†These authors contributed equally to this work.

Accepted: October 13, 2014

Data deposition: This Whole Genome Shotgun project has been deposited at DDBJ/EMBL/GenBank under the accession JPEN00000000. The version described in this paper is version JPEN01000000.

Abstract

Streptococcus sinensis is a recently discovered human pathogen isolated from blood cultures of patients with infective endocarditis. Its phylogenetic position, as well as those of its closely related species, remains inconclusive when single genes were used for phylogenetic analysis. For example, *S. sinensis* branched out from members of the anginosus, mitis, and sanguinis groups in the 16S ribosomal RNA gene phylogenetic tree, but it was clustered with members of the anginosus and sanguinis groups when *groEL* gene sequences used for analysis. In this study, we sequenced the draft genome of *S. sinensis* and used a polyphasic approach, including concatenated genes, whole genomes, and matrix-assisted laser desorption ionization-time of flight mass spectrometry to analyze the phylogeny of *S. sinensis*. The size of the *S. sinensis* draft genome is 2.06 Mb, with GC content of 42.2%. Phylogenetic analysis using 50 concatenated genes or whole genomes revealed that *S. sinensis* formed a distinct cluster with *Streptococcus oligofermentans* and *Streptococcus cristatus*, and these three streptococci were clustered with the "sanguinis group." As for phylogenetic analysis using hierarchical cluster analysis of the mass spectra of streptococci, *S. sinensis* also formed a distinct cluster with *S. oligofermentans* and *S. cristatus*, but these three streptococci were clustered with the "mitis group." On the basis of the findings, we propose a novel group, named "sinensis group," to include *S. sinensis*, *S. oligofermentans,* and *S. cristatus*, in the *Streptococcus* genus. Our study also illustrates the power of phylogenomic analyses for resolving ambiguities in bacterial taxonomy.

Key words: *Streptococcus sinensis*, genome, phylogenomic, "sinensis group", Illumina.

Background

The genus *Streptococcus* currently comprises more than 90 species with some of them being important human pathogens causing significant morbidity and mortality globally. Traditional phenotypic classification of *Streptococcus* relies on their Lancefield group antigens and hemolytic properties, which divides the genus into two major groups, pyogenic and viridans groups (Sherman 1937). As a result of the widespread use of polymerase chain reaction and DNA sequencing in the last two decades, genotypic methods such as amplification and sequencing of universal gene targets represent an advanced method for bacterial classification and identification. Among the various studied gene targets, 16S ribosomal RNA (rRNA) gene has been the most widely used. Classification of *Streptococcus* based on the 16S rRNA gene sequences divided the genus into six major groups, namely anginosus, mitis, salivarius, mutans, bovis, and pyogenic groups (Kawamura et al. 1995).

Subsequently, Facklam (2002) suggested the addition of sanguinis group to the existing groups based on the phenotypic properties, resulting in seven major groups. However, some studies still showed that 16S rRNA gene or other housekeeping genes failed to provide sufficient resolution and to delineate *Streptococcus* species into the same taxonomic groupings (Tapp et al. 2003; Glazunova et al. 2010; Park et al. 2010; Zbinden et al. 2011). Apart from these phenotypic and genotypic identification methods, a recently emerged technique, matrix-assisted laser desorption ionization time-of-flight mass spectrometry (MALDI-TOF MS), has been used for identification and cluster analysis of many bacterial species. The technique examines the profile of proteins detected directly from intact bacteria and involves pattern analysis of the mass spectra obtained within a mass range of 2,000–20,000 Da from whole cell protein fingerprinting using mathematical tools (Hsieh et al. 2008). Several studies have shown that MALDI-TOF MS is useful for identification and cluster analysis of different groups of medically important *Streptococcus* species (Dubois et al. 2013; Karpanoja et al. 2014).

In 2002, we reported the discovery of a novel species, *Streptococcus sinensis*, isolated from multiple blood cultures of a 42-year-old Chinese woman with infective endocarditis complicating her chronic rheumatic heart disease in Hong Kong (Woo et al. 2002). Two years later, we reported two other cases of infective endocarditis caused by *S. sinensis* with Lancefield F (Woo et al. 2004). Subsequently, additional cases reported from France and Italy suggested that the bacterium is an emerging pathogen of global importance (Uckay et al. 2007; Faibis et al. 2008). The oral cavity is the natural reservoir of *S. sinensis* (Woo et al. 2008). Phenotypically, the bacterium possessed characteristics from members of anginosus, sanguinis, and mitis groups, as different commercial kits gave different identities to different strains of *S. sinensis* (Woo et al. 2006). Genotypically, phylogenetic analysis using 16S rRNA gene sequences showed that *S. sinensis* was equally close to both the anginosus and mitis groups but was more closely related to sanguinis group when using GroEL sequences. Because of these inconclusive results, the phylogenetic position of *S. sinensis* remains uncertain.

In this study, we attempted to determine the exact phylogenetic position of *S. sinensis* using phylogenomic approach. We sequenced the first genome of *S. sinensis* type strain HKU4^T and performed comparative genomic studies utilizing publicly available *Streptococcus* genomes. Phylogenomic analysis using the whole-genome sequences revealed a new phylogenetic clade, including *S. sinensis, Streptococcus oligofermentans,* and *Streptococcus cristatus*, which we propose it as "sinensis group." Intergenomic distance analysis, phylogenetic analysis using concatenated sequences, and cluster analysis of MALDI-TOF MS were also performed.

Genome Sequencing of *S. sinensis* Type Strain HKU4^T

The de novo assembly using 26.81 million paired-end reads generated 116 contigs with lengths ranging from 212 to 125,948 bases, giving a total genome size of 2.06 Mb with the average GC content of 42.2%. All contigs generated were submitted to the RAST version 4.0 (Rapid Annotation using Subsystem Technology) annotation server, resulting in 1,992 protein-coding sequences (CDSs), 4 rRNA operons, and 47 transfer RNA-encoding genes (supplementary table S1, Supplementary Material online). To facilitate the subsequent genomic analysis, 86 genome sequences of other *Streptococcus* species and that of *Enterococcus faecalis* were also submitted to RAST for annotation. Among the 86 genome sequences, 25 complete genome sequences and 5 draft genome sequences (*Streptococcus australis, Streptococcus criceti, S. cristatus, Streptococcus dentisani,* and *Streptococcus tigurinus* because only one partial genome sequence was available for each species) were used for subsystem classification (table 1). Each CDS in annotated genomes was grouped into different RAST subsystems based on the predicted functional role. Among the 1,992 CDSs in the *S. sinensis* HKU4^T genome, 854 CDSs can be categorized into RAST subsystems (fig. 1*A*), in which 133 CDSs were classified into more than one category. Overall, majority of CDSs were classified into subsystems of carbohydrates (148 CDSs, 7.4%), protein metabolism (118 CDSs, 5.9%), DNA metabolism (111 CDSs, 5.6%), and amino acid and derivatives (110 CDSs, 5.5%) (fig. 1*A*). The remaining 1,138 (57.1%) CDSs could not be classified into any subsystems, in which 488 (24.5%) CDSs were only annotated as hypothetical proteins. Consistent to the results of previous studies (Olson et al. 2013), when we compared the distribution of CDSs in each subsystem of the *S. sinensis* genome with those of other *Streptococcus* genomes, all of them have a similar percentage of their genome dedicated to each subsystem (fig. 1*B*).

Phylogenomic Analyses of *S. sinensis* and Other *Streptococcus* Species Reveals a Distinct Clade Comprising *S. sinensis, S. oligofermentans,* and *S. cristatus*

Phylogenetic analyses using sequences of single gene loci, 16S rRNA and *groEL*, extracted from 87 *Streptococcus* genomes showed that *S. sinensis* closely clustered with *S. oligofermentans* and *S. cristatus,* respectively (fig. 2*A*–*C*). However, the topology of these trees was not concordant, in which *S. sinensis* branched out from members of anginosus, mitis, and sanguinis groups when using 16S rRNA gene sequences (fig. 2*A*) but clustered with members of anginosus and sanguinis groups when using *groEL* sequences for analyses (fig. 2*B* and *C*). This inconclusive result suggested that

Table 1

Genome Sequence Details and the Corresponding DDH Value with *Streptococcus sinensis* HKU4[T]

Species	Strain	Status	GenBank Accession No.	DDH (Formula 1)
"Sinensis" group				
Streptococcus sinensis	HKU4[T]	Draft	JPEN00000000[a,b]	—
Streptococcus oligofermentans	AS 1.3089	Finished	NC_021175.1[a,b]	45.6
Streptococcus cristatus	ATCC 51100	Draft	AEVC01000001–AEVC01000031[a,b]	44.3
Sanguinis group				
Streptococcus sanguinis	SK36	Finished	NC_009009[a,b]	28.9
Streptococcus sanguinis	SK678	Draft	AEXA01000001–AEXA01000010	—
Streptococcus gordonii	CH1	Finished	NC_009785[a,b]	23.2
Anginosus group				
Streptococcus intermedius	C270	Finished	NC_022237.1[b]	17.1
Streptococcus intermedius	JTH08	Finished	NC_018073.1[b]	16.7
Streptococcus intermedius	B196	Finished	NC_022246.1[a,b]	16.7
Streptococcus intermedius	F0413	Draft	AFXO01000001–AFXO01000013	—
Streptococcus constellatus subsp. *pharyngis*	C232	Finished	NC_022236.1[a,b]	16.8
Streptococcus constellatus subsp. *pharyngis*	C818	Finished	NC_022245.1[b]	16.8
Streptococcus constellatus subsp. *pharyngis*	C1050	Finished	NC_022238.1[b]	16.7
Streptococcus constellatus subsp. *pharyngis*	SK1060	Draft	BASX01000001–BASX01000066	—
Streptococcus anginosus	C1051	Finished	NC_022244.1[a,b]	16.2
Streptococcus anginosus	C238	Finished	NC_022239.1[b]	15.6
Streptococcus anginosus	SK1138	Draft	ALJO01000001–ALJO01000013	—
Streptococcus anginosus subsp. *whileyi*	CCUG 39159	Draft	AICP01000001–AICP01000083	—
Mitis group				
Streptococcus oralis	Uo5	Finished	NC_015291.1[a,b]	16.1
Streptococcus oralis	SK610	Draft	AJKQ01000001–AJKQ01000031	—
Streptococcus mitis	B6	Finished	NC_013853[a,b]	15.4
Streptococcus pneumoniae	ATCC 700669	Finished	NC_011900[a,b]	15.4
Streptococcus pneumoniae	D39	Finished	NC_008533[b]	15.4
Streptococcus pneumoniae	JJA	Finished	NC_012466[b]	15.4
Streptococcus pneumoniae	Taiwan19F-14	Finished	NC_012469[b]	15.4
Streptococcus pneumoniae	TIGR4	Finished	NC_003028[b]	15.4
Streptococcus pneumoniae	G54	Finished	NC_011072[b]	15.3
Streptococcus pneumoniae	P1031	Finished	NC_012467[b]	15.3
Streptococcus pneumoniae	R6	Finished	NC_003098[b]	15.3
Streptococcus pneumoniae	70585	Finished	NC_012468[b]	15.2
Streptococcus pneumoniae	CGSP14	Finished	NC_010582[b]	15.2
Streptococcus pneumoniae	Hungary19A-6	Finished	NC_010380[b]	15.0
Streptococcus parasanguinis	FW213	Finished	NC_017905.1[b]	15.2
Streptococcus parasanguinis	ATCC 15912	Finished	NC_015678.1[a,b]	15.1
Streptococcus pseudopneumoniae	IS7493	Finished	NC_015875.1[a,b]	14.8
Streptococcus infantis	ATCC 700779	Draft	AEVD01000001–AEVD01000031	—
Streptococcus infantis	SK970	Draft	AFUT01000001–AFUT01000009	—
Streptococcus peroris	ATCC 700780	Draft	AEVF01000001–AEVF01000017	—
Streptococcus tigurinus	1366	Draft	AORX01000001–AORX01000014[a]	—
Streptococcus dentisani	7746	Draft	CAUJ01000001–CAUJ01000008[a]	—
Streptococcus australis	ATCC 700641	Draft	AEQR01000001–AEQR01000027[a]	—
Pyogenic group				
Streptococcus agalactiae	NEM316	Finished	NC_004368[b]	13.5
Streptococcus agalactiae	09mas018883	Finished	NC_021485.1[a,b]	13.2
Streptococcus agalactiae	2603 V/R	Finished	NC_004116[b]	13.2
Streptococcus agalactiae	A909	Finished	NC_007432[b]	13.2
Streptococcus dysgalactiae subsp. *equisimilis*	GGS_124	Finished	NC_012891[a,b]	13.3
Streptococcus pyogenes	M1 GAS	Finished	NC_002737[a,b]	13.3
Streptococcus pyogenes	Manfredo	Finished	NC_009332[b]	13.3
Streptococcus pyogenes	MGAS10270	Finished	NC_008022[b]	13.3

(continued)

Table 1 Continued

Species	Strain	Status	GenBank Accession No.	DDH (Formula 1)
Streptococcus pyogenes	MGAS10394	Finished	NC_006086[b]	13.3
Streptococcus pyogenes	MGAS10750	Finished	NC_008024[b]	13.3
Streptococcus pyogenes	MGAS2096	Finished	NC_008023[b]	13.3
Streptococcus pyogenes	MGAS315	Finished	NC_004070[b]	13.3
Streptococcus pyogenes	MGAS5005	Finished	NC_007297[b]	13.3
Streptococcus pyogenes	MGAS6180	Finished	NC_007296[b]	13.3
Streptococcus pyogenes	MGAS8232	Finished	NC_003485[b]	13.3
Streptococcus pyogenes	MGAS9429	Finished	NC_008021[b]	13.3
Streptococcus pyogenes	NZ131	Finished	NC_011375[b]	13.3
Streptococcus pyogenes	SSI-1	Finished	NC_004606[b]	13.3
Streptococcus uberis	0140J	Finished	NC_012004[a,b]	13.2
Streptococcus parauberis	KCTC 11537	Finished	NC_015558.1[a,b]	13.2
Streptococcus iniae	SF1	Finished	NC_021314.1[a,b]	13.1
Streptococcus equi subsp. *zooepidemicus*	MGCS10565	Finished	NC_011134[a,b]	13.1
Streptococcus equi subsp. *equi*	4047	Finished	NC_012471[b]	13.0
Streptococcus equi subsp. *zooepidemicus*	H70	Finished	NC_012470[b]	13.0
Streptococcus criceti	HS-6	Draft	AEUV02000001–AEUV02000002[a]	—
Bovis group				
Streptococcus gallolyticus	UCN34	Finished	NC_013798[a,b]	13.3
Streptococcus lutetiensis	33	Finished	NC_021900.1[a,b]	13.3
Streptococcus pasteurianus	ATCC 43144	Finished	NC_015600.1[a,b]	13.3
Streptococcus equinus	ATCC 9812	Draft	AEVB01000001–AEVB01000057	—
Streptococcus infantarius	ATCC BAA-102	Draft	ABJK02000001–ABJK02000022	—
Mutans group				
Streptococcus mutans	NN2025	Finished	NC_013928[a,b]	13.3
Streptococcus mutans	GS-5	Finished	NC_018089[b]	13.2
Streptococcus mutans	LJ23	Finished	NC_017768.1[b]	13.2
Streptococcus mutans	UA159	Finished	NC_004350[b]	13.2
Salivarius group				
Streptococcus thermophilus	CNRZ1066	Finished	NC_006449[b]	13.8
Streptococcus thermophilus	LMD-9	Finished	NC_008532[b]	13.8
Streptococcus thermophilus	LMG 18311	Finished	NC_006448[b]	13.8
Streptococcus thermophilus	JIM 8232	Finished	NC_017581.1[a,b]	13.7
Streptococcus salivarius	JIM8777	Finished	NC_017595.1[b]	13.7
Streptococcus salivarius	CCHSS3	Finished	FR873481[a,b]	13.6
Streptococcus salivarius	SK126	Draft	ACLO01000001–ACLO01000101	—
No group				
Streptococcus suis	BM407	Finished	NC_012926[b]	13.7
Streptococcus suis	P1/7	Finished	NC_012925[b]	13.7
Streptococcus suis	98HAH33	Finished	NC_009443[b]	13.6
Streptococcus suis	05ZYH33	Finished	NC_009442[a,b]	13.6
Streptococcus suis	SC84	Finished	NC_012924[b]	13.6
Outgroup				
Enterococcus faecalis	V583	Finished	NC_004668.1[b]	12.6

[a]Genome sequence submitted for subsystem classification using RAST 4.0.
[b]Genome sequence used to construct the whole genome tree.

phylogenetic analysis using single gene target has not added much to our understanding on the phylogeny of *S. sinensis*, as they only represent a very small portion of the bacterial genome.

In view of this problem, we attempted to use a phylogenomic approach, which based on the whole bacterial genome, to resolve the taxonomic ambiguity of *S. sinensis*. In bacterial taxonomy, whole genomic DNA–DNA hybridization (DDH) has been used to determine the genetic distance between two strains and 70% DDH was proposed as a criterion for delineating species (Wayne 1988). However, this method is not widely used because it is tedious and results depend on experimental conditions. Therefore, it is often difficult to compare results between different laboratories objectively. Recent advancement of genome sequencing calls for bioinformatics methods to replace the wet-lab DDH by in silico

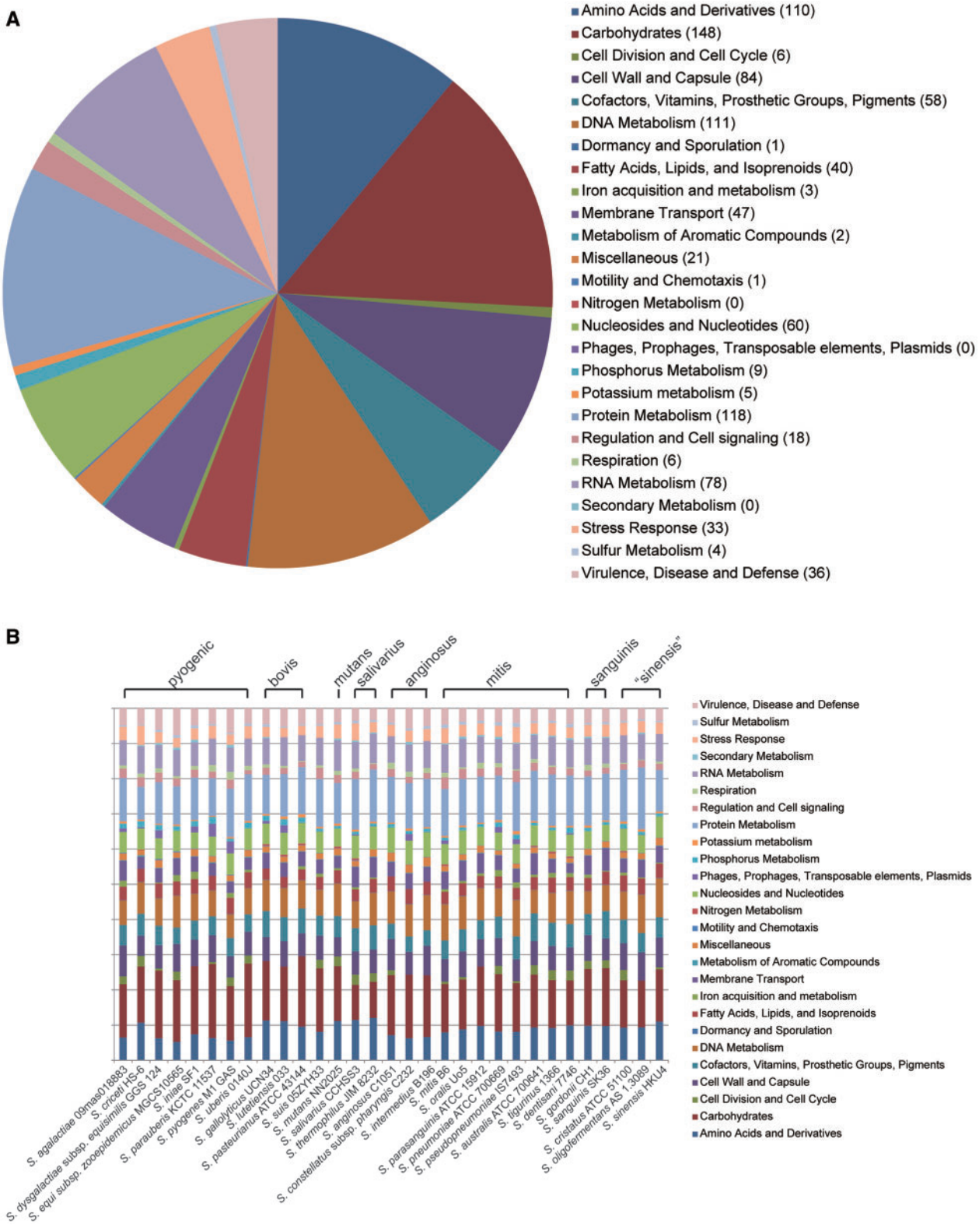

Fig. 1.—Distributions of predicted coding sequence function in the annotated genomes according to RAST subsystems. In (*A*), the number of CDSs of *S. sinensis* HKU4^T in different subsystems is indicated in bracket. In (*B*), a total of 31 genomes of *Streptococcus* species, including *S. sinensis* and representatives from all major groups, were analyzed. Each column indicates the number of CDSs of each *Streptococcus* species in different subsystems showing in different color.

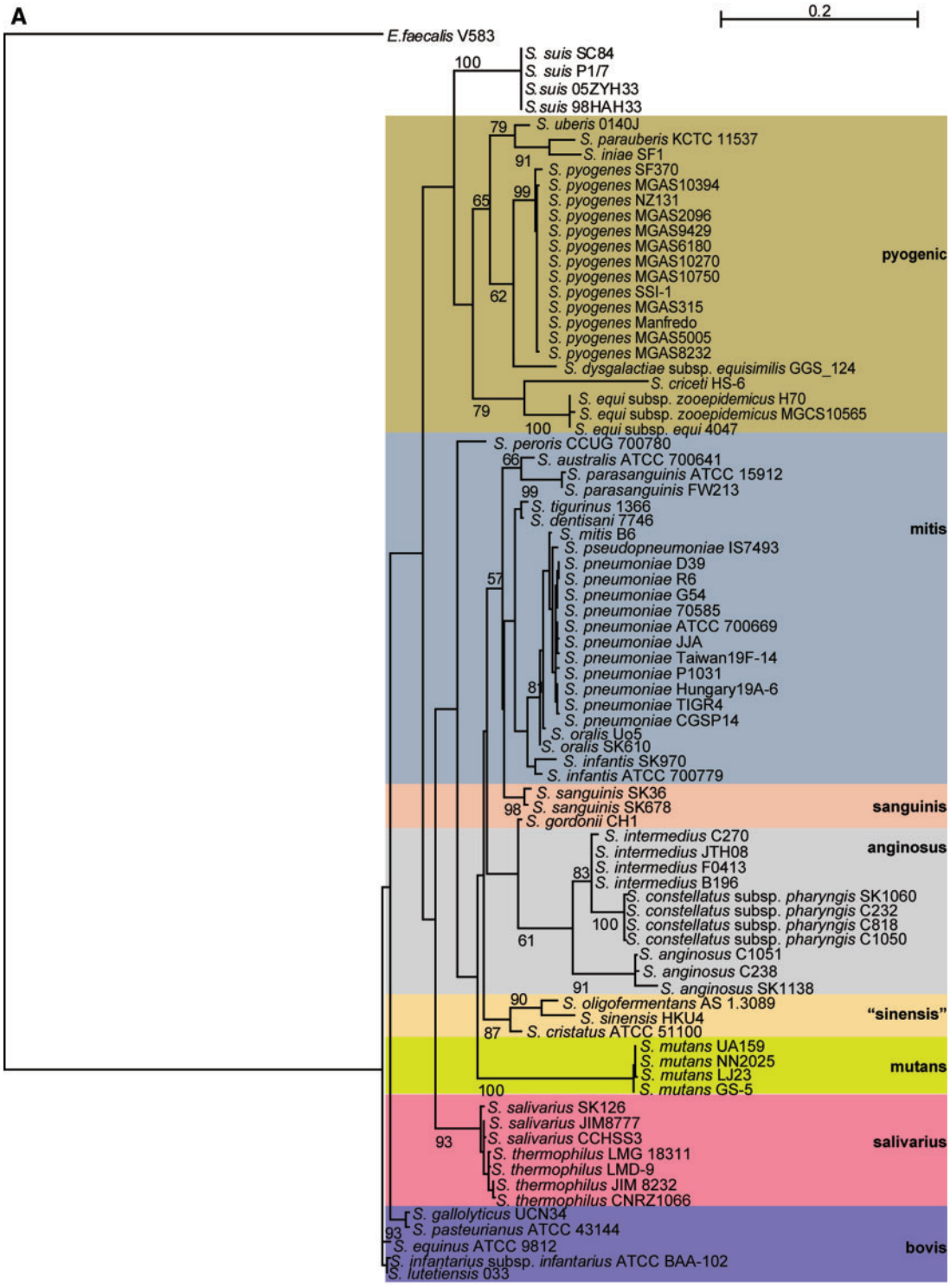

Fig. 2.—Phylogenetic relationship among *Streptococcus* strains. Three phylogenetic trees were constructed, each using a different genetic locus for analysis. (*A*) 16S rRNA. (*B*) *groEL*. (*C*) GroEL. The trees were constructed by maximum-likelihood method using RAxML (version 7.3) and *E. faecalis* V583 as the root. A total of 1,583 nucleotide positions of the 16S rRNA gene, 1,709 nucleotide positions and 565 deduced amino acid positions of *groEL* from 88 genomes were included for analyses. Bootstrap values were calculated from 1,000 replicates. The scale bar corresponds to the mean number of nucleotide/amino acid substitutions per site on the respective branch. Names and accession numbers are given as cited in GenBank in table 1.

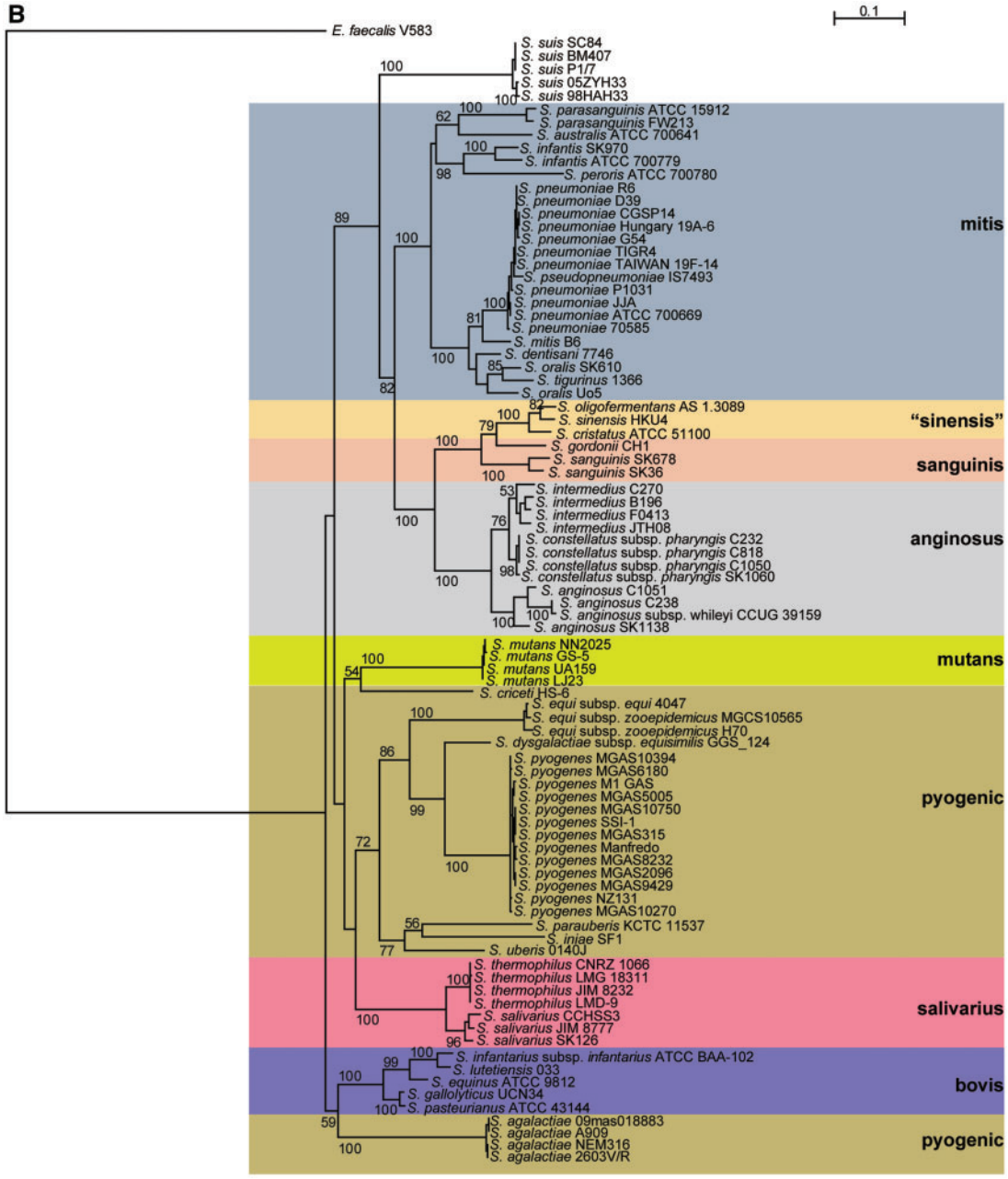

FIG. 2.—Continued.

genome-to-genome comparison. Among various sophisticated methods studied, a digital DDH method, GGDC 2.0, was shown to yield very good correlation with wet-lab DDH (Auch et al. 2010). In this study, with the availability of *Streptococcus* genome sequences, we were able to use GGDC 2.0 for intergenomic distance estimation, which allowed genome sequence comparison between two *Streptococcus* strains. The results showed that *S. sinensis* HKU4^T shared 45.6% and 44.3% nucleotide identities to the genome sequence of *S. oligofermentans* AS 1.3089 (GenBank accession number NC_021175) and *S. cristatus* ATCC 51100 (GenBank accession number AEVC01000001–AEVC01000031) but only 23.2–28.9%, 15.6–17.1%, and 14.8–16.1% nucleotide identities to those from members of sanguinis (3 genomes), anginosus (12 genomes), and mitis (23 genomes) groups, respectively (table 1). The present DDH value between *S. sinensis* and *S. oligofermentans* was quite different from the results obtained in the previous study, in which the wet-lab DDH value between *S. sinensis* and *S. oligofermentans* was only 15%. We speculate that the

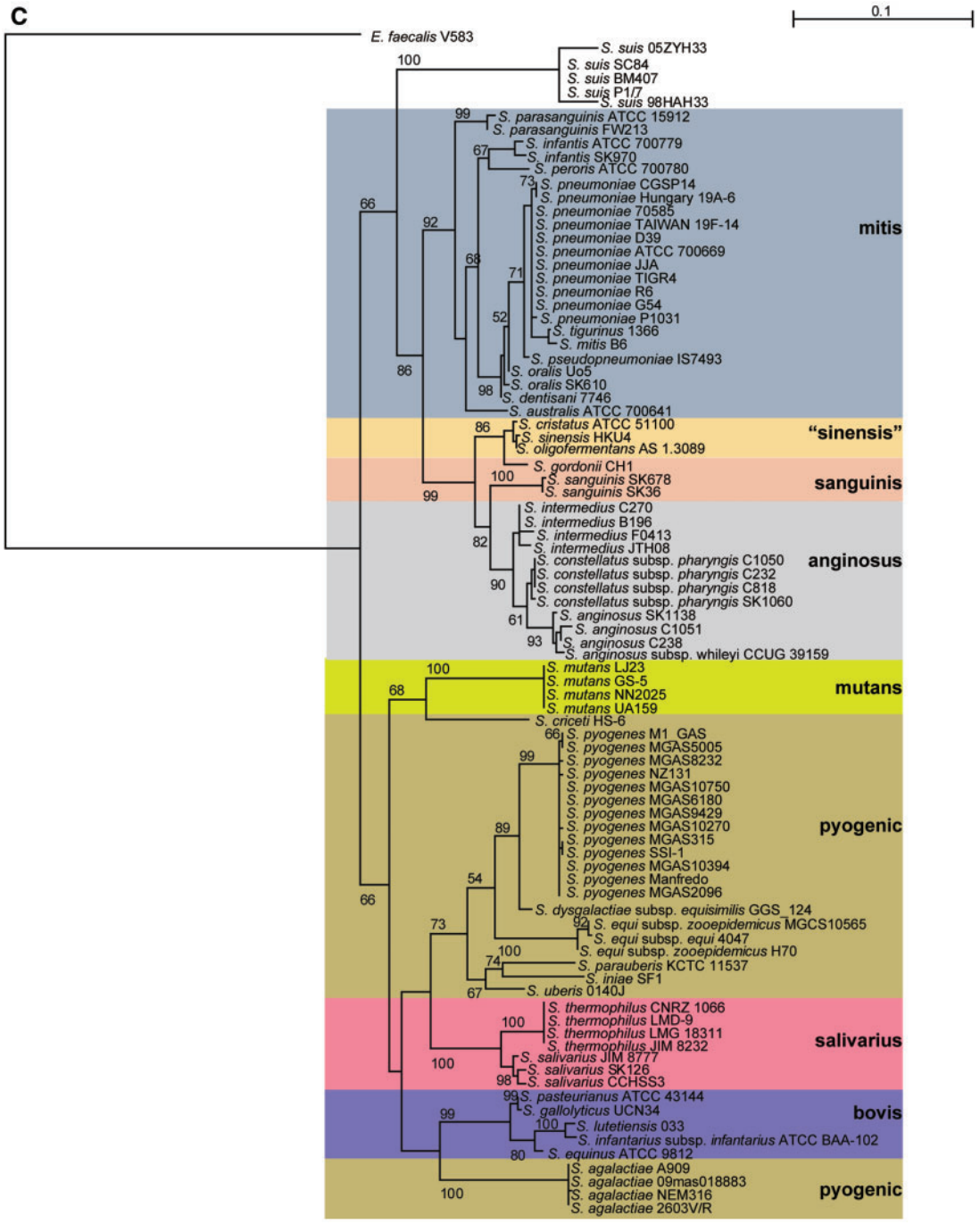

Fig. 2.—Continued.

discrepancy may due to the experimental error as DDH method is well-known not easily be made reproducible. Nevertheless, the present results using the entire genome sequences showed that *S. sinensis* was more closely related to *S. oligofermentans* and *S. cristatus* than members of sanguinis, anginosus, and mitis groups.

To further verify the phylogenetic position of *S. sinensis* among the genus *Streptococcus*, phylogenomic analysis utilizing the draft genome sequences of *S. sinensis* HKU4[T] and *S. cristatus* ATCC51100 as well as the available complete genome sequences of 69 *Streptococcus* species was performed and the results also supported that *S. sinensis* closely clustered with *S. oligofermentans* and *S. cristatus*, forming a unique clade (fig. 3*A*). Although this clade, comprising *S. sinensis, S. oligofermentans,* and *S. cristatus*, also clustered with members of sanguinis group, inclusion of this clade into

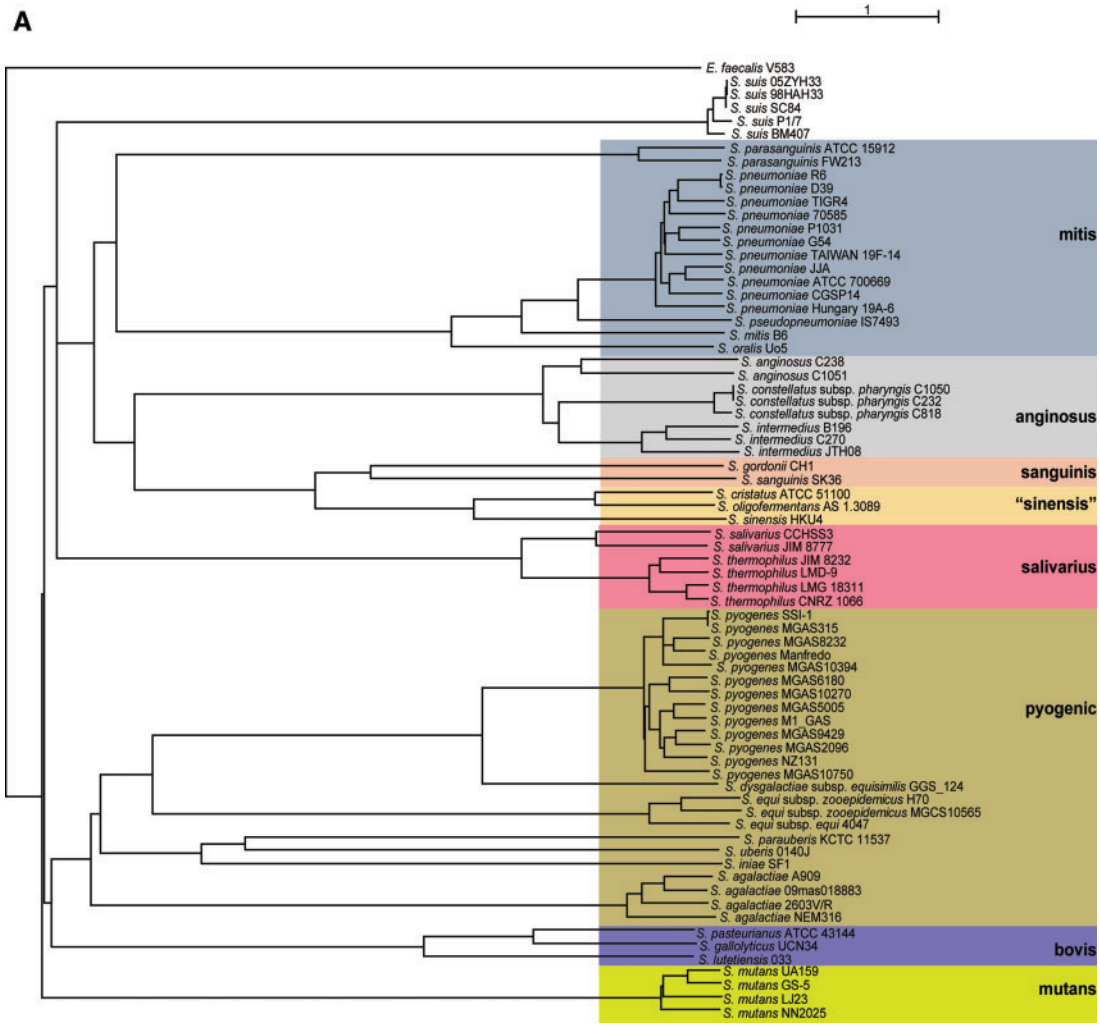

Fig. 3.—Phylogenetic tree constructed using draft genome sequences and concatenated nucleotide sequences of 50 ribosomal protein genes of *S. sinensis* HKU4^T. In (*A*), the tree based on draft genome sequences was constructed by neighbor-joining method using GGDC distance (formula 1) and *E. faecalis* V583 as the root. In (*B*), the tree based on 50 ribosomal protein genes, showing the relationship of *S. sinensis* HKU4^T to other *Streptococcus* species, was constructed by maximum-likelihood method using RAxML (version 7.3) and *E. faecalis* V583 as the root. Bootstrap values were calculated from 1,000 replicates. The scale bar corresponds to the mean number of nucleotide substitutions per site on the respective branch. The gene names and accession numbers are given as cited in GenBank in table 1.

the sanguinis group does not appear justified as they were connected by a relatively long branch and shared relatively low DDH value with *S. sinensis*. To confirm this result, we performed phylogenetic analysis using concatenated sequences of 50 ribosomal protein genes retrieved from 87 *Streptococcus* genomes including *S. sinensis*, representing 35 different species, and from one *E. faecalis* genome. These 50 ribosomal protein genes were chosen because they have been shown to be useful for phylogenetic delineation of bacterial species in previous studies (Jolley et al. 2012; Maiden et al. 2013) and their names were listed in supplementary table S2, Supplementary Material online. Results showed that the topologies of both trees were concordant and able to recover members of the seven taxonomic groups as described in previous studies (fig. 3*A* and *B*) (Kawamura et al. 1995; Facklam 2002). Consistently, the tree based on the concatenated sequences also revealed that *S. sinensis*, together with *S. oligofermentans* and *S. cristatus*, forming a distinct phylogenetic clade with a bootstrap value of 100%. It is worth noting that this unique clade also fell within the sanguinis group but the grouping was weakly supported by a low bootstrap value of 42%, meaning that it was not often clustered with members of sanguinis group (fig. 3*B*).

The phylogenetic position of *S. sinensis* has been considered controversial due to its simultaneous possession of

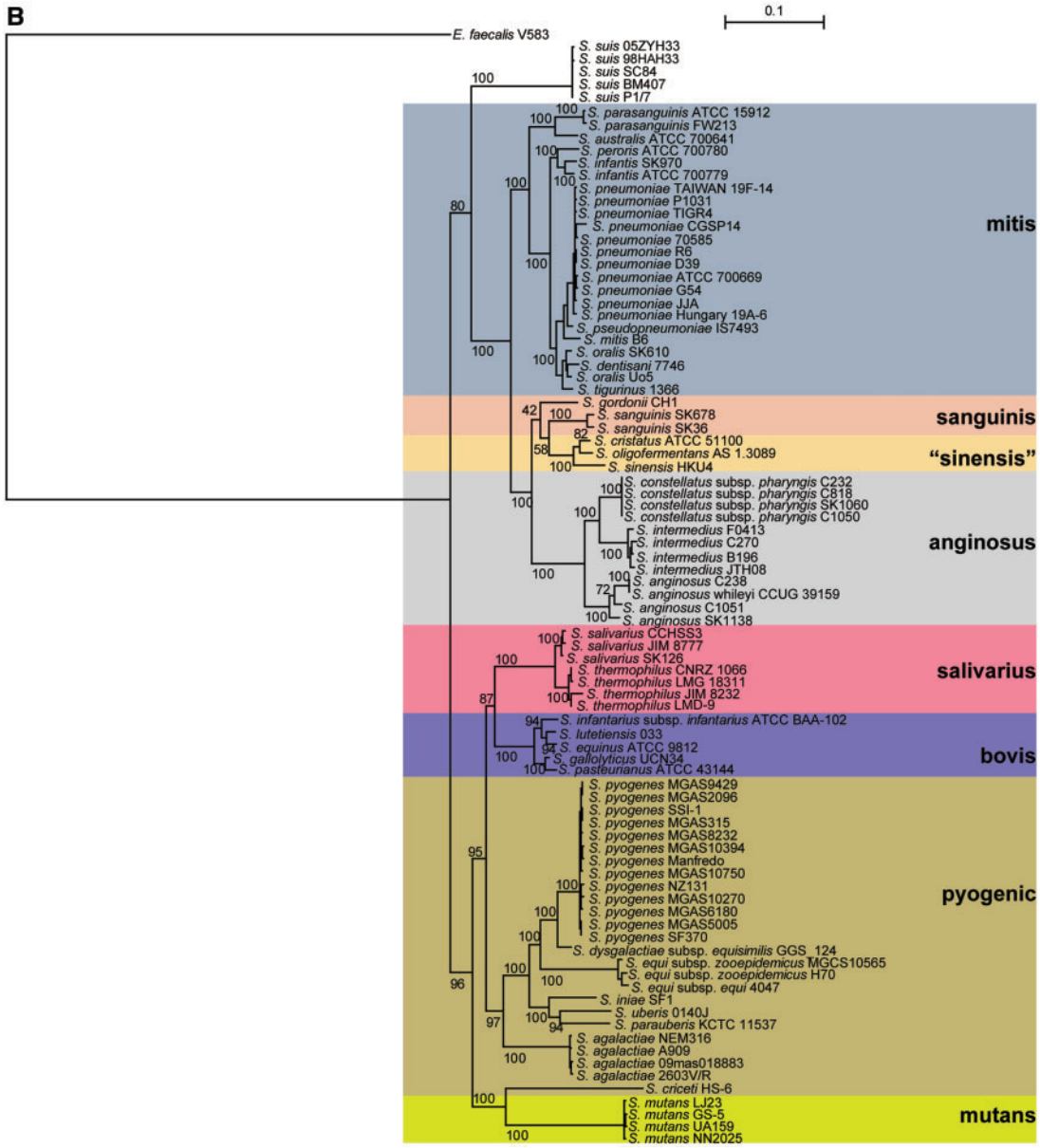

Fig. 3.—Continued.

phenotypic characteristics from members of mitis, sanguinis, and anginosus groups (Woo et al. 2006). Genotypic methods such as 16S rRNA or GroEL sequence analyses also failed to reveal its exact taxonomic group. In this study, the genome data, offering superior discriminatory power than any phenotypic and genotypic methods, resolve the taxonomic ambiguity of *S. sinensis*. Based on these genomic evidences, we suggest the three species, *S. sinensis*, *S. oligofermentans,* and *S. cristatus*, forming a well-supported clade, might best be considered as a separate group that we tentatively propose it as the sinensis group.

MALDI-TOF MS Analysis Supports the Newly Proposed Sinensis Group

Dendrogram generated from hierarchical cluster analysis of MALDI-TOF MS showed that the MALDI-TOF MS spectrum of *S. sinensis* HKU4^T and 28 nonduplicated *Streptococcus* species correctly recovered members of salivarius, mutans, bovis and pyogenic groups, and *Streptococcus suis* (fig. 4). *Streptococcus sinensis* consistently formed a distinct cluster with *S. oligofermentans* and *S. cristatus*, but these three streptococci were clustered with members belonging to "mitis

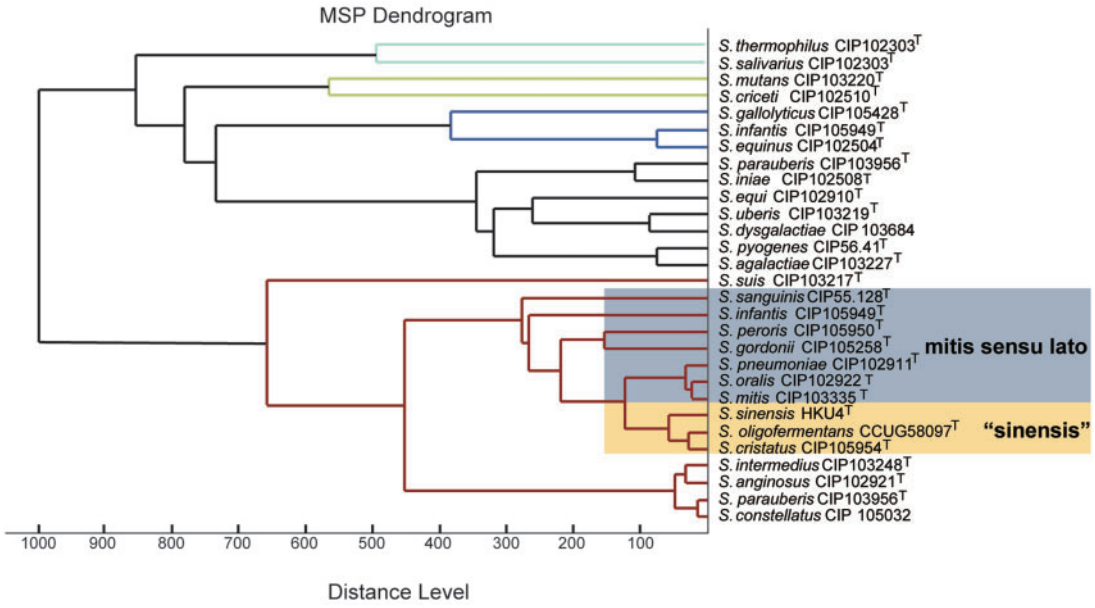

Fig. 4.—Dendrogram generated from hierarchical cluster analysis of MALDI-TOF MS spectra of *S. sinensis* HKU4^T and 28 isolates of other *Streptococcus* species. Distances are displayed in relative units.

group" (fig. 4). Notably, in contrast to the previous phylogenomic analysis using 50 concatenated genes or whole genomes, these three streptococci were clustered with "sanguinis group" (fig. 3*A* and *B*). On the basis of all these findings, these three streptococci should not be included into either sanguinis group or mitis group, but they should exist as a novel group, sinensis group in the genus *Streptococcus*.

Comparative Genomic Analyses of the Three Genomes of Sinensis Group

Based on BLAST (Basic Local Alignment Search Tool) analysis using 50% nucleotide identity as a threshold, we compared the genome sequence of *S. sinensis* HKU4^T with those of *S. oligofermentans* and *S. cristatus* and found that 1,241 CDSs were shared among the three genomes (fig. 5). There were 222 CDSs only shared with *S. oligofermentans* and 63 CDSs only shared with *S. cristatus* (fig. 5). A total of 466 CDSs were uniquely found in *S. sinensis* HKU4^T genome (fig. 5). Among these 466 unique CDSs, only 90 CDSs could be classified into RAST subsystems according to their predicted functional roles. For the remaining 376 CDSs, which did not belong to any subsystem, 256 CDSs were only annotated as hypothetical proteins. More in-depth analyses on these unique CDSs, especially those annotated as hypothetical proteins, may provide insights about the differences in metabolic capacity between *S. sinensis* and two other members of sinensis group.

Potential Virulence Factors in *S. sinensis* HKU4^T

Previous studies showed that *S. sinensis* has been recovered from patients with infective endocarditis globally (Woo et al. 2002, 2004; Uckay et al. 2007; Faibis et al. 2008), suggesting that the bacteria may possess virulence factors to adhere and to colonize heart valves and to cause endocardial damage. The draft genome of *S. sinensis* HKU4^T contains homologs of several virulence genes that are important for the pathogenesis of infective endocarditis. These include platelet activating factor, clumping factor B, collagen-binding protein, laminin-binding protein, and fibronectin/fibrinogen-binding protein which have been implicated in the pathogenesis of infective endocarditis by promoting bacterial adherence to endothelial tissues and triggering inflammatory responses (supplementary table S3, Supplementary Material online) (Abranches et al. 2011; Que and Moreillon 2011). Notably, two other members of the sinensis group, *S. oligofermentans* and *S. cristatus*, have also been isolated from patients with infective endocarditis (Matthys et al. 2006; Matta et al. 2009). Detailed comparative genomic studies on the *S. sinensis* genome and genomes of members of the sinensis group and those of the mitis group, including *Streptococcus mitis, S. tigurinus,* and *Streptococcus oralis*, may shed light on the ecology and biology of *S. sinensis*

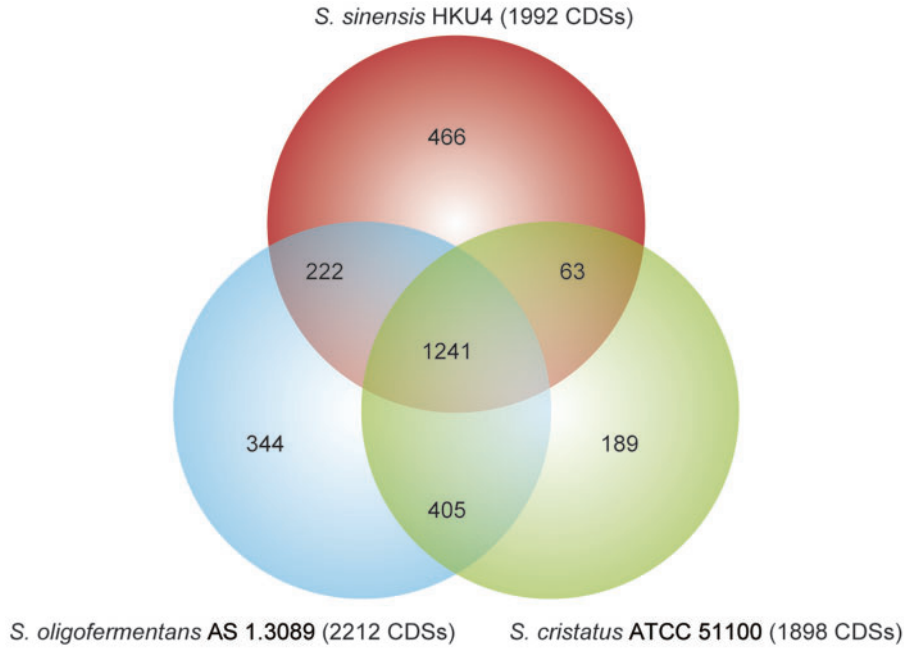

Fig. 5.—Comparative genomic analysis among three members of sinensis group, including *S. sinensis, S. oligofermentans,* and *S. cristatus*. The total number of CDSs per genome is given as indicated. The overlapping sections indicate shared numbers of CDSs among genomes.

as well as pathogenesis of infective endocarditis caused by members belonging to this new phylogenetic clade.

Conclusions

In this study, we sequenced the first draft genome of *S. sinensis* type strain HKU4^T and unambiguously determined the phylogenetic position of *S. sinensis* using phylogenomic approach. Phylogenomic and MALDI-TOF MS analysis revealed a distinct phylogenetic clade in the genus *Streptococcus*, which we proposed it as sinensis group, currently comprising three species, *S. sinensis, S. oligofermentans,* and *S. cristatus*. The present draft genome sequence has also allowed rapid exploration of potential virulence genes in *S. sinensis*. Our findings also illustrate the power of phylogenomic analyses for resolving ambiguities in bacterial taxonomy.

Materials and Methods

Bacterial Strains

A total of 29 nonduplicated *Streptococcus* strains, including *S. sinensis* HKU4^T, were included in the MALDI-TOF MS analysis (supplementary table S4, Supplementary Material online). *Streptococcus sinensis* HKU4^T was isolated from blood culture of a Chinese patient with infective endocarditis in Hong Kong, whereas the remaining strains were either purchased from the Collection of Institut Pasteur or Culture Collection, University of Göteborg (CCUG) (Woo et al. 2002).

Genome Sequencing, Assembly, and Annotation of *S. sinensis* Type Strain HKU4^T

The draft genome sequence of *S. sinensis* HKU4^T was determined by high-throughput sequencing with Illumina Hi-Seq 2500. Genomic DNA was extracted from overnight cultures (37 °C) grown on blood agars by genomic DNA purification kit (QIAgen, Hilden, Germany) as described previously (Woo et al. 2009; Tse et al. 2010). It was sequenced by 151-bp paired-end reads with mean library size of 350 bp. Sequencing errors were corrected by k-mer frequency spectrum analysis using SOAPec (http://soap.genomics.org.cn/about.html, last accessed October 22, 2014). De novo assembly was performed using SOAPdenovo2 (http://soap.genomics.org.cn/soapdenovo.html, last accessed October 22, 2014). Prediction of protein-coding regions and automatic functional annotation was performed using Glimmer3 and RAST server version 4.0 (Delcher et al. 2007; Aziz et al. 2008). Intergenomic distance between *S. sinensis* HKU4^T and other *Streptococcus* species, including representatives from seven major groups, was calculated using GGDC 2.0 (http://ggdc.dsmz.de/distcalc2.php, last accessed October 22, 2014) (Auch et al. 2010). *Streptococcus* species included for the distance calculation was shown in table 1.

Genome Sequence Data

Details of the 88 genome sequences used in this study are shown in table 1. Despite the 18 draft genome sequences, including one from *S. sinensis* HKU4^T which was sequenced to near completion as part of this study, the remaining 70 genome sequences were complete and downloaded from National Center for Biotechnology Information. Nucleotide sequences of 50 ribosomal protein genes (supplementary table S2, Supplementary Material online) and one copy of 16S rRNA and *groEL* genes were retrieved, respectively, from all 88 genomes (table 1).

Phylogenetic Characterization

The tree based on entire genome sequences was constructed by neighbor-joining method using GGDC distance (formula 1: length of all high-scoring segment pairs divided by total genome length) and *E. faecalis* as the root. The trees based on single gene loci, 16S rRNA and *groEL* gene (nucleotide and amino acid), were aligned by Geneious 7.1.5 (Biomatters Limited) and constructed, respectively, by maximum-likelihood method using RAxML (version 7.3). The tree based on the concatenated sequences of 50 ribosomal protein genes from 88 *Streptococcus* species, including 35 nonduplicated species, was constructed by maximum-likelihood method using RAxML (version 7.3). A total of 20,921 nucleotide positions were included in the analysis. Nucleotide sequences of corresponding homologs in *E. faecalis* were used as the outgroups where appropriate.

Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry

MALDI-TOF MS was performed as previously described with slight modifications (Tang et al. 2013; Lau et al. 2014). Twenty-nine nonduplicated *Streptococcus* strains, including *S. sinensis* HKU4^T, were grown on sheep blood agar at 37 °C with 5% CO_2 for 48 h. All isolates were analyzed by the ethanol formic acid extraction method using the same experimental conditions suggested by the Bruker system. Briefly, 1–3 colonies were suspended into 100 µl of HPLC grade water (Fluka, St Louis, MO). Then, 300 µl of absolute ethanol was added and incubated for 5 min at room temperature. Cell pellet was then air dried after centrifugation. The pellet was resuspended in 30 µl each of formic acid (Fluka) and acetonitrile (Sigma Aldrich, St Louis, MO). Each bacterial extract was spotted onto six spots of the MSP96 target plate after centrifugation. Samples were processed in the Bruker MicroFlex LT mass spectrometer (Bruker Daltonics, Bremen, Germany) with 1 µl of α-cyano 4-hydroxycinnimic acid matrix solution (Sigma Aldrich). Spectra were obtained with an accelerating voltage of 20 kV in linear mode and analyzed within m/z range 3,000–15,000 Da. Spectra were analyzed with MALDI Biotyper 3.1 and Reference Library V.3.1.2.0 (Bruker Daltonics). A mass spectrum profile (MSP) based on 24 separate determinations was created. The representative MSPs were then selected for hierarchical cluster analysis using MALDI Biotyper 3.1 (Bruker Daltonics) with default parameters (Ketterlinus et al. 2005), where distance values are relative and are always normalized to a maximum value of 1,000.

Supplementary Material

Supplementary tables S1–S4 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was partly supported by the Strategic Research Theme Fund and the Small Project Funding Scheme, The University of Hong Kong. The authors thank members of the Centre for Genomic Sciences, The University of Hong Kong, for their technical support in genome sequencing.

Literature Cited

Abranches J, et al. 2011. The collagen-binding protein Cnm is required for *Streptococcus mutans* adherence to and intracellular invasion of human coronary artery endothelial cells. Infect Immun. 79: 2277–2284.

Auch AF, Klenk HP, Goker M. 2010. Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Stand Genomic Sci. 2:142–148.

Aziz RK, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Delcher AL, Bratke KA, Powers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23: 673–679.

Dubois D, Segonds C, Prere MF, Marty N, Oswald E. 2013. Identification of clinical *Streptococcus pneumoniae* isolates among other alpha and nonhemolytic streptococci by use of the Vitek MS matrix-assisted laser desorption ionization-time of flight mass spectrometry system. J Clin Microbiol. 51:1861–1867.

Facklam R. 2002. What happened to the streptococci: overview of taxonomic and nomenclature changes. Clin Microbiol Rev. 15:613–630.

Faibis F, et al. 2008. *Streptococcus sinensis*: an emerging agent of infective endocarditis. J Med Microbiol. 57:528–531.

Glazunova OO, Raoult D, Roux V. 2010. Partial *recN* gene sequencing: a new tool for identification and phylogeny within the genus *Streptococcus*. Int J Syst Evol Microbiol. 60:2140–2148.

Hsieh SY, et al. 2008. Highly efficient classification and identification of human pathogenic bacteria by MALDI-TOF MS. Mol Cell Proteomics. 7:448–456.

Jolley KA, et al. 2012. Ribosomal multilocus sequence typing: universal characterization of bacteria from domain to strain. Microbiology 158:1005–1015.

Karpanoja P, Harju I, Rantakokko-Jalava K, Haanpera M, Sarkkinen H. 2014. Evaluation of two matrix-assisted laser desorption ionization-time of flight mass spectrometry systems for identification of viridans group streptococci. Eur J Clin Microbiol Infect Dis. 33:779–788.

Kawamura Y, Hou XG, Sultana F, Miura H, Ezaki T. 1995. Determination of 16S rRNA sequences of *Streptococcus mitis* and *Streptococcus gordonii* and phylogenetic relationships among members of the genus *Streptococcus*. Int J Syst Bacteriol. 45:406–408.

Ketterlinus R, Hsieh SY, Teng SH, Lee H, Pusch W. 2005. Fishing for biomarkers: analyzing mass spectrometry data with the new ClinProTools software. Biotechniques Suppl:37–40.
Lau SK, et al. 2014. Matrix-assisted laser desorption ionisation time-of-flight mass spectrometry for identification of clinically significant bacteria that are difficult to identify in clinical laboratories. J Clin Pathol. 67:361–366.
Maiden MC, et al. 2013. MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 11:728–736.
Matta M, et al. 2009. First case of *Streptococcus oligofermentans* endocarditis determined based on *sodA* gene sequences after amplification directly from valvular samples. J Clin Microbiol. 47:855–856.
Matthys C, et al. 2006. *Streptococcus cristatus* isolated from a resected heart valve and blood cultures: case reports and application of phenotypic and genotypic techniques for identification. Acta Clin Belg. 61: 196–200.
Olson AB, et al. 2013. Phylogenetic relationship and virulence inference of *Streptococcus anginosus* group: curated annotation and whole-genome comparative analysis support distinct species designation. BMC Genomics 14:895.
Park HK, Yoon JW, Shin JW, Kim JY, Kim W. 2010. *rpoA* is a useful gene for identification and classification of *Streptococcus pneumoniae* from the closely related viridans group streptococci. FEMS Microbiol Lett. 305:58–64.
Que YA, Moreillon P. 2011. Infective endocarditis. Nat Rev Cardiol. 8: 322–336.
Sherman JM. 1937. The streptococci. Bacteriol Rev. 1:3–97.
Tang BS, et al. 2013. Matrix-assisted laser desorption ionisation-time of flight mass spectrometry for rapid identification of *Laribacter hongkongensis*. J Clin Pathol. 66:1081–1083.
Tapp J, Thollesson M, Herrmann B. 2003. Phylogenetic relationships and genotyping of the genus *Streptococcus* by sequence determination of the RNase P RNA gene, *rnpB*. Int J Syst Evol Microbiol. 53: 1861–1871.
Tse H, et al. 2010. Complete genome sequence of *Staphylococcus lugdunensis* strain HKU09-01. J Bacteriol. 192:1471–1472.
Uckay I, et al. 2007. *Streptococcus sinensis* endocarditis outside Hong Kong. Emerg Infect Dis. 13:1250–1252.
Wayne LG. 1988. International Committee on Systematic Bacteriology: announcement of the report of the ad hoc Committee on Reconciliation of Approaches to Bacterial Systematics. Zentralbl Bakteriol Mikrobiol Hyg A. 268:433–434.
Woo PC, et al. 2002. *Streptococcus sinensis* sp. nov., a novel species isolated from a patient with infective endocarditis. J Clin Microbiol. 40: 805–810.
Woo PC, et al. 2004. *Streptococcus sinensis* may react with Lancefield group F antiserum. J Med Microbiol. 53:1083–1088.
Woo PC, et al. 2008. The oral cavity as a natural reservoir for *Streptococcus sinensis*. Clin Microbiol Infect. 14:1075–1079.
Woo PC, et al. 2009. The complete genome and proteome of *Laribacter hongkongensis* reveal potential mechanisms for adaptations to different temperatures and habitats. PLoS Genet. 5: e1000416.
Woo PC, Teng JL, Lau SK, Yuen KY. 2006. Clinical, phenotypic, and genotypic evidence for *Streptococcus sinensis* as the common ancestor of anginosus and mitis groups of streptococci. Med Hypotheses. 66: 345–351.
Zbinden A, Kohler N, Bloemberg GV. 2011. *recA*-based PCR assay for accurate differentiation of *Streptococcus pneumoniae* from other viridans streptococci. J Clin Microbiol. 49: 523–527.

Associate editor: B. Venkatesh