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

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

<sup>1</sup>Department of Microbiology, The University of Hong Kong, Hong Kong, China
<sup>2</sup>Research Centre of Infection and Immunology, The University of Hong Kong, Hong Kong, China
<sup>3</sup>State Key Laboratory of Emerging Infectious Diseases, The University of Hong Kong, Hong Kong, China
<sup>4</sup>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).
which we propose it as “sinensis group.” Intergenomic analysis using the whole-genome sequences of S. sinensis, S. oligofermentans, revealed a new phylogenetic clade, including S. cristatus. Phylogenomic analysis using the whole-genome sequences and cluster analysis of MALDI-TOF MS were also performed.

We performed comparative 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 dentisianus, 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 HKU4T genome, 854 CDSs can be categorized into RAST subsystems (fig. 1A), 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. 1A). 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. 1B).

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. 2A–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. 2A) but clustered with members of anginosus and sanguinis groups when using groEL sequences for analyses (fig. 2B and C). This inconclusive result suggested that
| Species | Strain | Status | GenBank Accession No. | DDH (Formula 1) |
|---------|--------|--------|----------------------|-----------------|
| **“Sinensis” group** | | | | |
| *Streptococcus sinensis* | HKU4⁷ | Draft | JPEN000000000⁶,⁷ | — |
| *Streptococcus olsogfermentans* | AS 1.3089 | Finished | NC_021175.1⁶,⁷ | 45.6 |
| *Streptococcus cristatus* | ATCC S1100 | Draft | AEVC01000001–AEVC01000031⁶,⁷ | 44.3 |
| **Sanguinis group** | | | | |
| *Streptococcus sanguinis* | SK36 | Finished | NC_009009⁶,⁷ | 28.9 |
| *Streptococcus sanguinis* | SK678 | Draft | AEVA10000001–AEVA1000100010⁶,⁷ | — |
| *Streptococcus gordonii* | CH1 | Finished | NC_009785⁶,⁷ | 23.2 |
| **Anginosus group** | | | | |
| *Streptococcus intermedius* | C270 | Finished | NC_022237.1⁶ | 17.1 |
| *Streptococcus intermedius* | JTH08 | Finished | NC_018073.1⁶ | 16.7 |
| *Streptococcus intermedius* | B196 | Finished | NC_022461.1⁶ | 16.7 |
| *Streptococcus intermedius* | F0413 | Draft | AFXO10000001–AFXO10000013 | — |
| *Streptococcus constellatus subsp. pharyngis* | C232 | Finished | NC_022236.1⁶,⁷ | 16.8 |
| *Streptococcus constellatus subsp. pharyngis* | C818 | Finished | NC_022245.1⁵ | 16.8 |
| *Streptococcus constellatus subsp. pharyngis* | C1050 | Finished | NC_022238.1⁵ | 16.7 |
| *Streptococcus constellatus subsp. pharyngis* | SK1060 | Draft | BASX01000001–BASX01000066 | — |
| *Streptococcus anginosus* | C1051 | Finished | NC_022441.1⁶,⁷ | 16.2 |
| *Streptococcus anginosus* | C238 | Finished | NC_022239.1⁵ | 15.6 |
| *Streptococcus anginosus* | SK1138 | Draft | ALJO10000001–ALJO10000013 | — |
| *Streptococcus anginosus subsp. whiteyi* | CCUG 39159 | Draft | AICP01000001–AICP01000083 | — |
| **Mitis group** | | | | |
| *Streptococcus oralis* | Uo5 | Finished | NC_015291.1⁶,⁷ | 16.1 |
| *Streptococcus oralis* | SK610 | Draft | AJHQ01000001–AJHQ010000031 | — |
| *Streptococcus mitis* | B6 | Finished | NC_013853.1⁷ | 15.4 |
| *Streptococcus pneumoniae* | ATCC 700669 | Finished | NC_011900.1⁷ | 15.4 |
| *Streptococcus pneumoniae* | D39 | Finished | NC_008533.1⁶ | 15.4 |
| *Streptococcus pneumoniae* | JJA | Finished | NC_012466.1⁶ | 15.4 |
| *Streptococcus pneumoniae* | Taiwanese19F-14 | Finished | NC_012469.1⁶ | 15.4 |
| *Streptococcus pneumoniae* | TIGR4 | Finished | NC_003028.1⁶ | 15.4 |
| *Streptococcus pneumoniae* | G54 | Finished | NC_011072.1⁶ | 15.3 |
| *Streptococcus pneumoniae* | P1031 | Finished | NC_012467.1⁶ | 15.3 |
| *Streptococcus pneumoniae* | R6 | Finished | NC_003098.1⁶ | 15.3 |
| *Streptococcus pneumoniae* | 70585 | Finished | NC_012468.1⁶ | 15.2 |
| *Streptococcus pneumoniae* | CGSP14 | Finished | NC_010582.1⁶ | 15.2 |
| *Streptococcus pneumoniae* | Hungary19A-6 | Finished | NC_010380.1⁶ | 15.0 |
| *Streptococcus parasanguinis* | FW213 | Finished | NC_017905.1⁷ | 15.2 |
| *Streptococcus parasanguinis* | ATCC 15912 | Finished | NC_015678.1⁶,⁷ | 15.1 |
| *Streptococcus pseudopneumoniae* | IS7493 | Finished | NC_015875.1⁶,⁷ | 14.8 |
| *Streptococcus infantis* | ATCC 700779 | Draft | AEVD01000001–AEVD0100000021 | — |
| *Streptococcus infantis* | SK970 | Draft | AFUT01000001–AFUT01000009 | — |
| *Streptococcus peroris* | ATCC 700780 | Draft | AEVF01000001–AEVF01000017 | — |
| *Streptococcus tigirinus* | 1366 | Draft | AORX01000001–AORX01000014⁶ | — |
| *Streptococcus dentisani* | 7746 | Draft | CAUJ10000001–CAUJ10000008⁶ | — |
| *Streptococcus australis* | ATCC 700641 | Draft | AEQR01000001–AEQR01000027⁶ | — |
| **Pyogenic group** | | | | |
| *Streptococcus agalactiae* | NEM316 | Finished | NC_004368.1⁶ | 13.5 |
| *Streptococcus agalactiae* | 09mas018883 | Finished | NC_021485.1⁶,⁷ | 13.2 |
| *Streptococcus agalactiae* | 2603 V/R | Finished | NC_004116.1⁶ | 13.2 |
| *Streptococcus agalactiae* | A909 | Finished | NC_007432.1⁶ | 13.2 |
| *Streptococcus dysgalactiae subsp. equisimilis* | GGS_124 | Finished | NC_012891.1⁶,⁷ | 13.3 |
| *Streptococcus pyogenes* | M1 GAS | Finished | NC_022737.1⁶,⁷ | 13.3 |
| *Streptococcus pyogenes* | Manfredo | Finished | NC_009332.1⁶ | 13.3 |
| *Streptococcus pyogenes* | MGAS10270 | Finished | NC_008022.1⁶ | 13.3 |

(continued)
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

### Table 1 Continued

| Species | Strain | Status | GenBank Accession No. | DDH (Formula 1) |
|---------|--------|--------|-----------------------|-----------------|
| *Streptococcus pyogenes* | MGAS10394 | Finished | NC_006086 | 13.3 |
| *Streptococcus pyogenes* | MGAS10750 | Finished | NC_008024 | 13.3 |
| *Streptococcus pyogenes* | MGAS2096 | Finished | NC_008023 | 13.3 |
| *Streptococcus pyogenes* | MGAS3115 | Finished | NC_004070 | 13.3 |
| *Streptococcus pyogenes* | MGAS5005 | Finished | NC_007297 | 13.3 |
| *Streptococcus pyogenes* | MGAS6180 | Finished | NC_007296 | 13.3 |
| *Streptococcus pyogenes* | MGAS8232 | Finished | NC_003485 | 13.3 |
| *Streptococcus pyogenes* | MGAS9429 | Finished | NC_008021 | 13.3 |
| *Streptococcus pyogenes* | NZ131 | Finished | NC_011375 | 13.3 |
| *Streptococcus pyogenes* | SSI-1 | Finished | NC_004606 | 13.3 |
| *Streptococcus uberis* | 0140J | Finished | NC_012004 | 13.2 |
| *Streptococcus parauberis* | KCTC 11537 | Finished | NC_015558 | 13.2 |
| *Streptococcus iniae* | SF1 | Finished | NC_021314 | 13.1 |
| *Streptococcus equi* subsp. *zooepidemicus* | MGCS10565 | Finished | NC_011134 | 13.1 |
| *Streptococcus equi* subsp. *equi* | 4047 | Finished | NC_012471 | 13.0 |
| *Streptococcus equi* subsp. *zooepidemicus* | H70 | Finished | NC_012470 | 13.0 |
| *Streptococcus cricteri* | HS-6 | Draft | AEUV02000001 | — |
| *Bovis group* | *Streptococcus galolyticus* | UCN34 | Finished | NC_013798 | 13.3 |
| *Streptococcus lutteniensis* | 33 | Finished | NC_021900 | 13.3 |
| *Streptococcus pasteurianus* | ATCC 43144 | Finished | NC_015600 | 13.3 |
| *Streptococcus equinus* | ATCC 9812 | Draft | AEVB01000001 | — |
| *Streptococcus infantarius* | ATCC BAA-102 | Draft | ABJ02000001 | — |
| *Mutans group* | *Streptococcus mutans* | NN2025 | Finished | NC_013928 | 13.3 |
| *Streptococcus mutans* | GS-5 | Finished | NC_018089 | 13.2 |
| *Streptococcus mutans* | LJ23 | Finished | NC_017768 | 13.2 |
| *Streptococcus mutans* | UA159 | Finished | NC_004350 | 13.2 |
| *Salivarius group* | *Streptococcus thermophilus* | CNRZ1066 | Finished | NC_006449 | 13.8 |
| *Streptococcus thermophilus* | LMD-9 | Finished | NC_008532 | 13.8 |
| *Streptococcus thermophilus* | LMG 18311 | Finished | NC_006448 | 13.8 |
| *Streptococcus thermophilus* | JIM 8232 | Finished | NC_017581 | 13.7 |
| *Streptococcus salivarius* | JIM8777 | Finished | NC_017595 | 13.7 |
| *Streptococcus salivarius* | CCHSS3 | Finished | FR83481 | 13.6 |
| *Streptococcus salivarius* | SK126 | Draft | ACL00100001 | — |
| *No group* | *Streptococcus suis* | BM407 | Finished | NC_012926 | 13.7 |
| *Streptococcus suis* | P1/7 | Finished | NC_012925 | 13.7 |
| *Streptococcus suis* | 90HAAH33 | Finished | NC_009443 | 13.6 |
| *Streptococcus suis* | 05ZHY33 | Finished | NC_009442 | 13.6 |
| *Streptococcus suis* | SC84 | Finished | NC_012924 | 13.6 |
| *Outgroup* | *Enterococcus faecalis* | V583 | Finished | NC_004668 | 12.6 |

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

*Streptococcus sinensis* HKU4T Genome Sequence

**Streptococcus sinensis**
**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* 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.
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 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. cristaus ATCC 51100 (GenBank accession number AEVC01000001–AEVC0100031) 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
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 HKU4T 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. 3A). Although this clade, comprising S. sinensis, S. oligofermentans, and S. cristatus, also clustered with members of sanguinis group, inclusion of this clade into...
the sanguinis group does not appear justified as they were connected by a relatively long branch and shared relatively low DDH value with \textit{S. sinensis}. To confirm this result, we performed phylogenetic analysis using concatenated sequences of 50 ribosomal protein genes retrieved from 87 \textit{Streptococcus} genomes including \textit{S. sinensis}, representing 35 different species, and from one \textit{E. faecalis} genome. These 50 ribosomal protein genes were chosen because they have been shown to be useful for phylogenetic delination 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. 3A and B) (Kawamura et al. 1995; Facklam 2002). Consistently, the tree based on the concatenated sequences also revealed that \textit{S. sinensis}, together with \textit{S. oligofermentans} and \textit{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. 3B).

The phylogenetic position of \textit{S. sinensis} has been considered controversial due to its simultaneous possession of...
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 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. 3.—Continued.**
“sanguinis 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. 3A 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*.
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 \textit{S. sinensis} type strain HKU4$^T$ and unambiguously determined the phylogenetic position of \textit{S. sinensis} using phylogenomic approach. Phylogenomic and MALDI-TOF MS analysis revealed a distinct phylogenetic clade in the genus \textit{Streptococcus}, which we proposed it as sinensis group, currently comprising three species, \textit{S. sinensis}, \textit{S. oligofermentans}, and \textit{S. cristatus}. The present draft genome sequence has also allowed rapid exploration of potential virulence genes in \textit{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 \textit{Streptococcus} strains, including \textit{S. sinensis} HKU4$^T$, were included in the MALDI-TOF MS analysis (supplementary table S4, Supplementary Material online). \textit{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 \textit{S. sinensis} Type Strain HKU4$^T$

The draft genome sequence of \textit{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 \textit{S. sinensis} HKU4$^T$ and other \textit{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). \textit{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
Abrahantes 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 AE, 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, Poxers EC, Salzberg SL. 2007. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics 23:673–679.
Dubos D, Segonds C, Prete 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-ization-time of flight mass spectrometry system. J Clin Microbiol. 51:1861–1867.
Facklam R. 2002. What happened to the streptococci: overview of taxonomy and nomenclature changes. Clin Microbiol Rev. 15:613–630.
Faihls F, et al. 2008. Streptococcus sinensis: an emerging agent of infective endocarditis. J Med Microbiol. 57:529–531.
Glazunova OQ, Raoul 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 gordoni 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 *Lanibacter 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, *mrP*. 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 *Lanibacter 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