Novel Molecular Method for Identification of *Streptococcus pneumoniae* Applicable to Clinical Microbiology and 16S rRNA Sequence-Based Microbiome Studies

Christian F. P. Scholz, Knud Poulsen, and Mogens Kilian
Department of Biomedicine, Faculty of Health Sciences, Aarhus University, Aarhus, Denmark

The close phylogenetic relationship of the important pathogen *Streptococcus pneumoniae* and several species of commensal streptococci, particularly *Streptococcus mitis* and *Streptococcus pseudopneumoniae*, and the recently demonstrated sharing of genes and phenotypic traits previously considered specific for *S. pneumoniae* hamper the exact identification of *S. pneumoniae*. Based on sequence analysis of 16S rRNA genes of a collection of 634 streptococcal strains, identified by multilocus sequence analysis, we detected a cytosine at position 203 present in all 440 strains of *S. pneumoniae* but replaced by an adenosine residue in all strains representing other species of mitis group streptococci. The *S. pneumoniae*-specific sequence signature could be demonstrated by sequence analysis or indirectly by restriction endonuclease digestion of a PCR amplicon covering the site. The *S. pneumoniae*-specific signature offers an inexpensive means for validation of the identity of clinical isolates and should be used as an integrated marker in the annotation procedure employed in 16S rRNA-based molecular studies of complex human microbiotas. This may avoid frequent misidentifications such as those we demonstrate to have occurred in previous reports and in reference sequence databases.

Exact differentiation of the important pathogen *Streptococcus pneumoniae* and its close commensal relatives is a prerequisite for full appreciation of their clinical significance, pathogenic properties, and epidemiology both in routine clinical microbiology and in research. Although a number of phenotypic traits were previously assumed to be specific markers of *S. pneumoniae*, recent reports demonstrated that such properties, including op-toxin susceptibility, capsule production, and bile solubility, are also present in a substantial proportion of isolates of the commensal *Streptococcus mitis* (2, 3, 6, 19, 23, 40). Moreover, the realization that clinical isolates of *S. pneumoniae*, e.g., many isolates from conjunctivitis, do not express a capsule (hence the term “atypical pneumococci”) further emphasizes this problem (4, 7). Similar problems apply to molecular identification methods. Because the full 16S rRNA gene sequences of *S. pneumoniae, Streptococcus pseudopneumoniae, S. mitis,* and *Streptococcus oralis* are more than 99% identical (21), attempts to identify clinical isolates by BLAST searches with an unidentified 16S rRNA gene sequence against sequences in recognized databases are often unsuccessful or may provide misleading results. This is becoming an increasing problem due to the frequent and automated use of this procedure in clinical microbiology and in the annotation process employed in attempts to map complex microbiotas, such as in the Human Microbiome Project. Likewise, the suggested identification of clinical isolates of *S. pneumoniae* by detection of genes encoding regulatory proteins of capsule biosynthesis, e.g., *cpsA*/weg, or putative virulence proteins such as autolysin (*lytA*), pneumolysin (*ply*), the surface protein PspA (*pspA*) (1, 30) is bound to result in a number of misidentifications due to the presence of homologs of these genes in a proportion of commensal streptococci (19, 20, 23, 27, 38, 40).

The biological explanation for the described problems is the phylogenetic history of these streptococcal species. In spite of striking differences in pathogenic potential, *S. pneumoniae* and *S. mitis, S. pseudopneumoniae, Streptococcus oralis,* and *Streptococcus in-
Molecular Identification of Streptococcus pneumoniae

To-Go PCR beads (Amersham Pharmacia Biotech, Uppsala, Sweden). The temperature profile for the PCR was an initial denaturation at 94°C for 5 min, followed by 30 cycles at 94°C for 1 min, 60°C for 1 min, and 72°C for 2 min, followed by a final extension at 72°C for 8 min. Amplicons were purified using Wizard minicolumns (Promega, Madison, WI). Sequencing of both strands of the amplified fragments was achieved with the same primers at GATC Biotech.

Sequencing of each of the four individual copies of the 16S rRNA gene in the genomes of selected strains was performed as described previously (23). Briefly, genome fragments each containing one copy of the 16S rRNA gene were separated by pulsed-field gel electrophoresis (PFGE) of genomic DNA prepared from bacteria harvested from a Todd-Hewitt broth culture and cleaved in the agar plugs with 1 U of the intron-specific endonuclease I-Ceu1 (New England BioLabs, Hertfordshire, United Kingdom), which recognizes a sequence unique to 23S rRNA genes (39). The individual fragments were purified from the gel using a Qiagen gel extraction kit (Qiagen) and subjected to amplification and sequencing of the 16S rRNA gene as described above.

Sequence comparisons and MLSA. All sequences were aligned in MEGA version 5.05. Housekeeping gene sequences were then trimmed to match the templates from http://viridans.emlsa.net, concatenated, and subjected to phylogenetic analysis using the Minimum evolution algorithm (Jones-Taylor-Thornton model) with the pairwise deletion option. Identification of strains was done against their clustering relative to designated type strains of relevant Streptococcus species.

Restriction endonuclease analysis. A 286-bp fragment of the 16S rRNA gene (positions 81 to 366 in the S. pneumoniae strain R6 16S rDNA gene) was amplified using the primers Pn-16S-up (ATGAGTTGCGAACGCGGATTGA) and Pn-16S-do (ATTTCCGAAGATTCCTACGGGCT). The PCR was performed on approximately 0.5 ng genomic DNA in a 25-μl reaction volume containing 10 pmol of each primer, using PuRe-Taq Ready-To-Go PCR beads (GE Healthcare). PCR parameters were 94°C for 5 min, 30 cycles of 94°C for 1 min, 60°C for 1 min, and 72°C for 2 min, and then an extension at 72°C for 8 min. The PCR products were analyzed by agarose gel electrophoresis. Digestion with the restriction endonuclease BshHKA1 (New England BioLabs) was performed in a 50-μl volume containing 20 μl PCR product, 3 μl 10× NEBuffer 4, 5 μl bovine serum albumin (BSA) (10 mg/ml), and 20 U enzyme. After incubation at 65°C for 1 h, the DNA fragments were analyzed by agarose gel electrophoresis.

RESULTS

Identity of strains. Based on the phylogenetic analysis of concatenated housekeeping genes, 440 strains were identified as S. pneumoniae, 60 as S. mitis, 41 as S. oralis, 17 as S. gordonii, 25 as S. infantis, 10 as belonging to the so-called “S. mitis biovar 2” (which constitutes a distinct, yet-to-be described species closely related to S. oralis), and five as S. pseudopneumoniae. Type strains and single additional representatives of the remaining species of the mitis group formed distinct lineages in the tree. A total of 28 strains formed separate clusters unrelated to recognized species. A tree based on a selection of strains (to facilitate visibility of individual lineages) is shown in Fig. 1.

Comparison of the 16S rRNA gene sequences showed that all 440 strains verified as S. pneumoniae based on the MLSA analysis had a cytosine (C) at position 203 (relative to the annotated 16S rRNA gene) based on the MLSA analysis of the 16S rRNA gene as described above.
tions, all 440 *S. pneumoniae* strains showed a fully conserved stretch of 15 nucleotides flanking the signature C at position 203. The sequence and exceptions are shown in Table 1.

A BLASTN search of the NCBI database identifies 17 16S rRNA gene sequences of strains designated *S. pneumoniae* that deviated from the mentioned patterns: strains NCTC 7466 (accession no. AJ001248), HK P116 (accession no. AY525791.1), SI P25 (accession no. AY525793.1), HK P55 (accession no. AY525790.1), Kor 145 (accession no. AY525788.1), SI P7 (accession no. AY525789.1), HER 1055 (accession no. AY525790.1), 578 (accession no. GU326241.1), 101/87 (accession no. GU326242.1), Ejeff (accession no. GU907525.1), PTO6 (accession no. GU561400.1), CCARM 4033 (accession no. GU045397.1), ToTT (accession no. FJ823147.1), CCRI-15796 (accession no. EU156780.1), and CCRI-14740 (accession no. EU156780.1). Strain NCTC 7466 is the same as strain D39 and its rough variant R6, both of which have been genome sequenced (accession numbers NC008533 and NC003098). All copies of the 16S rRNA sequences retrieved from these two genomes show the *S. pneumoniae*-characteristic cytosine at position 203 and do not confirm the 16S rRNA sequence under accession number AJ001248. Five sequences (HK P116, SI P25, HK P55, Kor 145, and SI P7) originated in a Korean study of “phenotypically and genotypically aberrant *Streptococcus pneumoniae* strains” isolated from Asian countries (24). Analysis of the deposited partial sequences of the glucose-6-phosphate dehydrogenase (*gdh*) genes of these strains supports that the identification was incorrect and that they are *S. mitis* strains. Strain HER 1055 is referred to and validated as *S. mitis* in the publication from which sequence accession no. AJ617796.1 originates (35) but is incorrectly listed as *S. pneumoniae* in the NCBI database. Accordingly, 9 out of 10 other sequences of strain HER 1055 in the NCBI database are labeled *S. mitis*. The strain corresponding to the sequence CCARM 4033 (accession no. GU045397.1) was identified based on single-locus sequencing.

![Phylogenetic tree based on concatenated sequences of seven housekeeping gene sequences, constructed with the Minimum evolution algorithm and used as the "gold standard" for identification of the 634 mitis group streptococcal strains included in the study. The clusters were named according to the clustering of designated type strains of the respective species.](https://jcm.asm.org/)

### FIG 1

### TABLE 1

Sequences (positions 193 to 209) flanking the *S. pneumoniae*-specific cytosine at position 203 in the 16S rRNA genes of mitis group *Streptococcus* species

<table>
<thead>
<tr>
<th>Species</th>
<th>Allele or strain(s)</th>
<th>Sequence&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>S. pneumoniae</em></td>
<td>Major allele</td>
<td>AAAAGGTGCACTGCAT</td>
<td>Includes type strain</td>
</tr>
<tr>
<td></td>
<td>EU NPO1</td>
<td>AAAAGGTGCACTGCAT</td>
<td>Accession no. AGQH10000017</td>
</tr>
<tr>
<td></td>
<td>VN O27</td>
<td>AAAAGGTGGAATGCAT</td>
<td>Accession no. AY525791.1</td>
</tr>
<tr>
<td></td>
<td>06-01-003MEF</td>
<td>AAAAGGTGCACTGCAT</td>
<td>Accession no. AEGF00000000.1</td>
</tr>
<tr>
<td><em>S. mitis</em></td>
<td>Major allele</td>
<td>AAAAGGTGCAATGCAT</td>
<td>Includes type strain</td>
</tr>
<tr>
<td></td>
<td>A1435,SK626,SK636, SK674 SK624 SK629</td>
<td>AAAAGGTGCAATGCAT</td>
<td></td>
</tr>
<tr>
<td><em>S. pseudopneumoniae</em></td>
<td>Major allele</td>
<td>AAAAGGTGCAATGCAT</td>
<td>Includes type strain</td>
</tr>
<tr>
<td></td>
<td>SK664</td>
<td>AAAAGGTGCAATGCAT</td>
<td></td>
</tr>
<tr>
<td><em>S. oralis</em></td>
<td>Major allele</td>
<td>AAAAGGTGCAATGCAT</td>
<td>Includes type strain</td>
</tr>
<tr>
<td></td>
<td>SK660</td>
<td>AAAAGGTGCAATGCAT</td>
<td></td>
</tr>
<tr>
<td><em>S. mitis 2</em></td>
<td>Major allele</td>
<td>AAAAGGTGCAATGCAT</td>
<td>Includes type strain</td>
</tr>
<tr>
<td></td>
<td>SK79, SK429</td>
<td>AAAAGGTGCAATGCAT</td>
<td></td>
</tr>
<tr>
<td><em>S. oligofermentans</em></td>
<td>Major allele</td>
<td>AAAAGGTGCAATGCAT</td>
<td>Includes type strain</td>
</tr>
<tr>
<td></td>
<td>SK674</td>
<td>AAAAGGTGCAATGCAT</td>
<td></td>
</tr>
<tr>
<td><em>S. infantis</em></td>
<td>Major allele</td>
<td>AAAAGGTGCAATGCAC</td>
<td>Includes type strain</td>
</tr>
<tr>
<td></td>
<td>SK660</td>
<td>AAAAGGTGCAATGCAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SK385,48291</td>
<td>AAAAGGTGCAATGCAC</td>
<td></td>
</tr>
<tr>
<td></td>
<td>SK655</td>
<td>AAAAGGTGCAATGCAC</td>
<td></td>
</tr>
<tr>
<td><em>S. peroris</em></td>
<td>Major allele</td>
<td>AAAAGGTGCAATGCAC</td>
<td>Includes type strain</td>
</tr>
<tr>
<td><em>S. sanguinis</em></td>
<td>GAAAGGGAATTTGCAC</td>
<td>GAAAGGGAATTTGCAC</td>
<td>Type strain</td>
</tr>
<tr>
<td><em>S. gordonii</em></td>
<td>GAAAGGGAATTTGCAC</td>
<td>GAAAGGGAATTTGCAC</td>
<td>Type strain</td>
</tr>
<tr>
<td><em>S. australis</em></td>
<td>GAAAGGGAATTTGCAC</td>
<td>GAAAGGGAATTTGCAC</td>
<td>Type strain</td>
</tr>
<tr>
<td><em>S. parasanguinis</em></td>
<td>GAAAGGGAATTTGCAC</td>
<td>GAAAGGGAATTTGCAC</td>
<td>Type strain</td>
</tr>
<tr>
<td><em>S. cristatus</em></td>
<td>GAAAGGGAATTTGCAC</td>
<td>GAAAGGGAATTTGCAC</td>
<td>Type strain</td>
</tr>
<tr>
<td><em>S. sieneus</em></td>
<td>GAAAGGGAATTTGCAC</td>
<td>GAAAGGGAATTTGCAC</td>
<td>Type strain</td>
</tr>
<tr>
<td><em>S. lactis</em></td>
<td>GAAAGGGAATTTGCAC</td>
<td>GAAAGGGAATTTGCAC</td>
<td>Type strain</td>
</tr>
</tbody>
</table>

<sup>a</sup> The nucleotide at position 203 is in bold.
(rpoA) (31), which according to other reports is insufficient (17). Among the remaining sequences, the identity of strain 101/87 is discussed in several studies, one of which concludes that the strain “may represent either an atypical or incorrectly classified organism” (40). Another, more recent study concludes that strain 101/87 “is as distant (or close) from pneumococci as it is from S. mitis” (35). There are no publications corresponding to the remaining sequences that deviate from the pattern, and the validity of the identifications therefore cannot be further analyzed.

A search of human microbiome data at NCBI and the RDP-II database identified the following sequences generated from uncultured material: clone 4V4 (accession no. AM157442.1) from an analysis of bacteria in breast milk (26) and ATC H22 37 (GU425360.1) from a comprehensive mapping of the human oral microbiome (10). Both sequences were annotated as S. pneumoniae, which according to our findings is incorrect. Contrasting examples were also disclosed among sequences generated in the human oral microbiome study (10). The sequences CMC176, CMC177, CMC178, CMC179, CMC183, CMC186, CMC188, CMC189, CMC192, CMC193, CMC195, CMC196, CMC199, CMC784, CMC785, CMC794, CMC792, CMC791, CMC787, CMC786, CMC785, CMC781, CMC778, CMC796, CMC782, and CMC783 were annotated as S. mitis but, according to the characteristic cytosine at position 203, belong to S. pneumoniae. A search in the RDP-II database did not reveal any strains annotated as S. infantis, S. oralis, S. peroris, S. pseudopneumoniae, S. sanguinis, S. gordonii, S. parasanguinis, or S. cistatus that had the pneumococcal signature cytosine.

**Detection of the signature C by restriction enzyme analysis.** The restriction endonuclease BsiHKAI (New England BioLabs) cleaves the sequence 5’…GWGCW ↓ C…3’ (W = A or T; C corresponds to the signature cytosine at position 203) and thus recognizes the sequence that includes the C at position 203 in all 440 confirmed strains of S. pneumoniae except strain 06-01-003MEF (Table 1). To examine whether restriction fragment length pattern analysis would be useful for validation of identification of S. pneumoniae, we digested PCR amplicons of the relevant 16S rRNA gene sequence in 114 validated strains of S. pneumoniae and 57 strains of S. mitis, S. oralis, S. pseudopneumoniae, and S. infantis. With one exception (S. pneumoniae clinical isolate 183/95), all strains of S. pneumoniae showed the expected banding pattern, while the PCR amplicons of all the strains representing other species remained uncleaved (Fig. 2). To analyze why the 16S rRNA gene amplicon of this strain was only partially cleaved by the endonuclease BsiHKAI, the amplicon was sequenced. In both directions, the sequence showed clear evidence of a mixture of T and C at position 201, suggesting different alleles in the four rRNA operons in the genome. Only the sequence with C at this position is recognized by the restriction endonuclease.

**DISCUSSION**

Frequent homologous recombination within and between Streptococcus species renders identification based on sequences of single loci problematic (5, 14, 15, 17, 18, 23). With this background, it is remarkable that the cytosine at position 203 in the 16S rRNA gene of S. pneumoniae appears to be fully conserved in the population of S. pneumoniae while strains of all other mitis group streptococci have an adenosine at this position. The evidence supporting this conclusion is based on our analysis of 440 strains of S. pneumoniae and 194 strains representing other species, all identified by multilocus sequence analysis and with the inclusion of relevant designated type strains. This distinguishing sequence signature that is directly detectable by sequence analysis or, alternatively, by restriction fragment length polymorphism (RFLP) analysis of a PCR amplicon of the 16S rRNA gene provides a simple and reliable means of identification or confirmation of S. pneumoniae isolates.

Our results demonstrate several examples of published studies in which strains of S. mitis were misidentified as S. pneumoniae, resulting in incorrect conclusions (24), and of published and unpublished sequences in databases incorrectly labeled as S. pneumoniae. This is a source of confusion and may result in further misidentifications when such sequences are used as reference in BLASTN analyses of sequences generated from unidentified bacterial isolates. One notable example is the 16S rRNA gene sequence of S. pneumoniae strain NCTC 7466 (accession no. AJ001248.1), which shows an adenosine at position 203. Analysis of all four 16S rRNA gene sequences in the genomes of both strain D99, which is the original designation for NCTC 7466, and strain R6, which is an acapsular derivative of the same strain, fail to confirm this. All sequences in these genomes show the S. pneumoniae-specific cytosine at position 203 and are in full agreement with our conclusions.

Automated BLAST analysis of 16S rRNA gene sequences has become a widely used method for molecular mapping of complex microbiotas, e.g., in the Human Microbiome Project, which often involves thousands or millions of sequences. In such studies, the species assignment of sequences is usually done automatically according to their closest fit in the databases. However, due to the relative conservation of 16S rRNA sequences among streptococci, particularly in S. pneumoniae, S. mitis, S. oralis, and S. infantis, separation to species level is often not achievable or may result in misidentifications. This study revealed several examples of misidentified sequences generated and annotated as part of studies of the human oral microbiome and of bacteria present in human breast milk (10, 26). As distinction between S. pneumoniae and commensal streptococci is important for understanding their ecology and relationships with the host, we recommend that the S.
pneumoniae-specific signature be used as an integrated part of the
annotation procedure employed in studies where these species are
relevant.

The conserved C corresponding to position 203 in the 16S rRNA sequence of S. pneumoniae strain R6 is located in helix 11 in the equivalent Escherichia coli 16S rRNA sequence (accession no. J01695) (37). Conservation of this nucleotide, which is unique to S. pneumoniae among mitis group streptococci, is intriguing and unexpected in a species that is known to exchange DNA with its close relatives. The bacterial ribosome is a very complex structure and consists of 23S, 16S, and 5S rRNAs and about 50 proteins. The start codon of mRNA molecules becomes positioned in the ribo-
some by base pairing between the 3′ end of 16S rRNA and the
Shine-Dalgarno sequence just upstream of the mRNA start codon, although the precise mechanism of translation initiation is still unclear (36). We speculate that an interaction between a pneumo-
coccus-specific ribosomal protein and helix 11 in the 16S rRNA molecule is crucial and thereby precludes mutations in this region without simultaneous changes in the protein.

In conclusion, this study identified a remarkable conservation of a cytosine residue at position 203 in the 16S rRNA gene of S. pneumoniae, which in all other species of mitis group streptococci is replaced by an adenosine. Detection of this S. pneumoniae-spec-
fic sequence signature by sequence analysis or, indirectly, by RFLP analysis using the restriction endonuclease BsiHKAI is an easy and reliable means of identification of S. pneumoniae in clinical
microbiology and as an integrated part the annotation procedure
in high-throughput molecular analyses of complex microbio-

ACKNOWLEDGMENT

This study was supported by grant 10-083748 from the Danish Research
Council for Health and Disease.

REFERENCES

identification of Streptococcus pneumoniae and description of Streptococ-
crob. 44:4163–4171.
genetic structure of nontypeable pneumococcal strains isolated from con-
5. Bishop CJ, et al. 2009. Assigning strains to bacterial species via the inter-
net. BMC Biol. 7:3.
7. Carvalho MGS, Steigerwalt AG, Thompson T, Jackson D, Facklam RR.
2003. Confirmation of nontypeable Streptococcus pneumoniae-like organ-
evolution and spread of a major class of mosaic pbp2x in Strepto-
512.
new tool for identification and phylogeny within the genus Streptococcus.
tionships among the causative agents of meningoitis and gladders, Burk-
holderia pseudomallei and Burkholderia mallei. J. Clin. Microbiol. 41:
2068–2079.
and interspecies genomic variation of Streptococcus pneumoniae. Infect.
nogenic bacteria. BMC Biol. 7:36.
17. Hoshino T, Fujisawa T, Kilian M. 2005. Use of phylogenetic and phe-
virulence genes in streptococci of the mitis group. J. Clin. Microbiol. 48:
2762–2769.
gene-technological and conventional methods in the identification of
tion of 16S rRNA Sequences of Streptococcus mitis and Streptococcus gar-
donii and phylogenetic relationships among members of the genus Strept-
approaches to the identification of the mitis group within the genus Strept-
taxonomic studies: a case study using 10 housekeeping genes in the genus
Ensifer (including former Sinorhizobium). Int. J. Syst. Evol. Microbiol. 58:
208–214.
rial diversity of breast milk among healthy women. Res. Microbiol. 158:
31–37.
cation of clinically relevant nonhemolytic streptococci on the basis of
sequence analysis of 16S-23S intergenic spacer region and partial gdh
30. Park HK, et al. 2010. Identification of the cpxA gene as a specific marker
for identification and classification of Streptococcus pneumoniae from the
closely related viridans group streptococci. FEMS Microbiol. Lett. 305:
58–64.
32. Pop M, Phillippy A, Delcher AL, Salzberg SL. 2004. Comparative gen-
33. Postic D, Garnier M, Baranton G. 2007. Multilocus sequence analysis of

Downloaded from http://jcm.asm.org on July 8, 2017 by guest