A molecular-capsular-type prediction system for 90 *Streptococcus pneumoniae* serotypes using partial *cpsA–cpsB* sequencing and *wzy*- or *wzx*-specific PCR

Fanrong Kong,1,5,6 Weizhen Wang,2 Jiang Tao,3 Lei Wang,3,5 Quan Wang,3 Archcna Sabananthan4 and Gwendolyn L. Gilbert1,6,7

1Centre for Infectious Diseases and Microbiology (CIDM), Institute of Clinical Pathology and Medical Research (ICPMR), Westmead, NSW 2145, Australia
2Wuhan First Hospital, Hubei Province, Wuhan 430022, PR China
3TEDA School of Biological Sciences and Biotechnology and Tianjin State Laboratory of Microbial Functional Genomics, Nankai University, TEDA College, Tianjin 300457, PR China
4School of Biotechnology and Biomolecular Sciences, The University of New South Wales, NSW 2052, Australia
5Tianjin Biochip Technology Corporation, TEDA, Tianjin 300457, PR China
6Department of Medicine, The University of Sydney, NSW 2066, Australia
7National Centre for Immunisation Research and Surveillance of Vaccine Preventable Diseases, Children's Hospital at Westmead, NSW 2145, Australia

In a previous study, a molecular capsular type (MCT) prediction system for 51 *Streptococcus pneumoniae* serotypes was developed based on a combination of partial *cpsA–cpsB* sequencing and serotype(s)/group(s)-specific PCR. In this study, another 169 *S. pneumoniae* isolates were added to the existing database of 427 isolates, including representatives of all 39 serotypes not previously studied. In addition to the authors' own limited sequence data for all 90 serotypes, *cpsA–cpsB* sequence data published by the *S. pneumoniae* capsular loci-sequencing group (http://www.sanger.ac.uk/Projects/S_pneumoniae/CPS/) at the Sanger Institute or available from GenBank were incorporated into the database. All serotypes, except 25A, were represented by at least two isolates. The number of sequence types identified was 138, of which 110 corresponded to single conventional serotypes (CSs); of these, 57 were represented by two or more isolates. Twenty-six sequence types were shared by between two and four CSs. To resolve these shared *cpsA–cpsB* sequence types and increase the discriminatory power of our system, the genes encoding the capsular polysaccharide flippase (*wzx*) and polymerase (*wzy*) were annotated and 24 new serotype(s)/group(s)-specific PCRs targeting *wzy* and two targeting *wzx* were designed. Using both *cpsA–cpsB* sequencing and *wzx/wzy* PCR, MCT correctly predicted the CSs of 516 (73 %) and the serogroup of an additional 155 (22 %) of the 708 isolates evaluated. For 5 % of isolates, MCT could not distinguish between members of five serotype pairs (37 isolates) containing members of different serogroups. Although further study of the relationship between MCT and CS is needed, this system now allows serotype or serogroup identification of 95 % of *S. pneumoniae* isolates.

Abbreviations: CS, conventional serotype/typing; MCT, molecular capsular type/typing.

The GenBank/EMBL/DDBJ accession numbers for the new partial *cpsA (wzg)–cpsB (wzh)* genes are AY508586–AY508641, AY621659, AY621660 and AY661448–AY661457.

Three phylogenetic trees, a schematic representation of related *wzx* genes and five tables of data are available as supplementary material in JMM Online.
INTRODUCTION

The importance of infections with *Streptococcus pneumoniae* is being increasingly recognized (Gillespie, 1999). The use of molecular tools to speed the process of detection, culture identification and typing of *S. pneumoniae* has been widely accepted (Gillespie, 1999).

*S. pneumoniae* comprises at least 90 serotypes (Henrichsen, 1995) distinguished by capsular polysaccharide antigens (Garcia et al., 2000). Capsule production in *S. pneumoniae* is largely controlled by the capsular polysaccharide synthesis (*cps*) gene cluster (Garcia et al., 2000). While serotyping and antibiotic-susceptibility testing remain the primary methods for characterizing pneumococci, molecular typing can add greater discrimination and complementary information (Hall, 1998). A molecular capsular typing (MCT) system for *S. pneumoniae* will be of greatest value when it can fully replace serotyping and allow monitoring of capsule evolution (Lawrence et al., 2000, 2003; Enright & Spratt, 1998). In a previous study, we developed a MCT prediction system for 51 *S. pneumoniae* serotypes based on sequencing and serotype(s)/group(s)-specific PCR, targeting the *cps* gene cluster (Kong & Gilbert, 2003). In this study, we aimed to extend the system to allow prediction of all 90 *S. pneumoniae* serotypes.

METHODS

Pneumococcal clinical isolates. This study was based on 169 well-characterized *S. pneumoniae* isolates that represented 63 serotypes, including all 39 that were not included in our previous study (which involved 427 isolates belonging to 51 serotypes) (Kong & Gilbert, 2003). Seventy-two isolates were obtained from the National Institute for the Control of Pharmaceutical and Biological Products, Beijing, PR China (Zhang et al., 1990), three from the Microbiological Diagnostic Unit, Public Health Laboratory, Department of Microbiology and Immunology, University of Melbourne, Victoria and 10 from the Statens Serum Institut, Denmark. Conventional serotyping (CS) was performed by donor laboratories and the serotypes were known at the time of receipt. A subset of well-characterized isolates, the serotypes of which were unknown at the time of receipt, was used to evaluate our genotyping system. It included 35 isolates provided by the Royal College of Pathologists of Australasia, Quality Assurance Program Pty Limited, New South Wales, Australia and 49 by the Department of Microbiology, Children’s Hospital at Westmead.

Isolates were retrieved from storage by subculture on blood agar plates (Columbia II agar base supplemented with 5% horse blood) and incubated overnight at 37°C in 5% CO2.

Annotation and analysis of *wzx* and *wzy*. Analysis of homology and protein hydrophobicity was performed to annotate the *wzx* and *wzy* genes in *S. pneumoniae* *cps* gene clusters. BLAST and PSI-BLAST (Altschul et al., 1997) were used to search GenBank and Pfam protein motif databases (Bateman et al., 2002) for possible gene functions. The TMHMM v2.0 analysis program (http://www.cbs.dtu.dk/services/TMHMM-2.0/) was used to identify potential transmembrane segments from amino acid sequences (Chen et al., 2003). Sequence alignment and comparison were done using the program CLUSTAL_W (Thompson et al., 1994).

Phylogenetic trees. The phylogenetic trees for partial *cpsA–cpsB*, *wzx* and *wzy* were generated by the neighbour-joining method using program MEGA (Kumar et al., 1994).

RESULTS AND DISCUSSION

CS results

CS was repeated on 23 isolates, at the Department of Microbiology, Children’s Hospital at Westmead, because of apparent inconsistencies between or sharing of sequence types and serotypes. After careful repetition by two different operators, isolates previously identified as serotypes 42 and 41F (Kong & Gilbert, 2003) were reassigned to serotypes 31 and 41A, respectively; serotypes of three additional isolates were also corrected. The serotypes of 15 isolates (including all those belonging to serotypes 27, 28F and 16A, and one each of serotypes 6A, 38 and 25F) were confirmed as previously defined. The final results are shown in Table S2 (available as supplementary data in JMM Online).

Extension of partial *cpsA–cpsB* sequence database

Partial *cpsA–cpsB* sequencing primers. The sequencing primers *cpsS1–cpsA3* produced amplicons from all isolates studied in this and our previous study, except for three belonging to rare serotypes, 25F, 25A and 38, and five that were non-serotypable (Kong & Gilbert, 2003). Two additional primer pairs, *cpsS1–cpsA1* and *cpsS3–cpsA2*, formed amplicons from isolates belonging to serotypes 25F, 25A and 38 and two non-serotypable isolates (the other three non-serotypable isolates were not studied further).

Sequence type nomenclature. Sequence types were generally named according to the corresponding serotype, with a
suffixed representing the source of the isolate in which they were first identified. The name given where a sequence type was shared by multiple serotypes includes all (two to four) of the serotypes in ascending numerical order (e.g. 15B-15C-22F-22A) (Henrichsen, 1995). One or two representative sequences of each sequence type were deposited into GenBank (see Tables S2 and S3, available as supplementary data in JMM Online, for sequence type nomenclature and corresponding GenBank accession numbers).

**Phylogenetic tree based on partial **_cpsA–cpsB_** genes.** A tree based on partial _cpsA–cpsB_ sequences of all isolates studied is shown in Fig. S1 (available as supplementary data in JMM Online). The tree shows similarities between sequence types – analogous to a multilocus sequence type tree (Enright & Spratt, 1998) – rather than true phylogenetic relationships. Adding the sequence of an isolate of unknown serotype to the _cpsA–cpsB_ tree, which contains all the sequences in our database, will allow us to infer its most likely serotype; the predictive accuracy will increase as the number of sequences increases (see Table S2, available as supplementary data in JMM Online).

**Sequence type vs. mutation.** Based on our sequence type definition, heterogeneity at one or more sites defines a new sequence type (Kong & Gilbert, 2003), which is consistent with the widely accepted principle for definition of a multilocus sequence type (Enright & Spratt, 1998, 1999). This strategy does not allow us to distinguish significant evolutionary mutations from ‘accidental’ point mutations, but it is a consistent, unambiguous basis for sequence type nomenclature (see Tables S2 and S3, available as supplementary data in JMM Online).

**Extension and ongoing evaluation of our sequence type database.** Our database has been extended to 90 serotypes, but its development will continue as new sequence types or even serotypes are identified. In future, more isolates of each serotype will be examined as they become available and the results added to our database. It may be necessary to modify the database when additional sequence types are identified and discrepancies are resolved (for example, our data differ from _cps_ gene cluster sequences reported by the Sanger Institute for serotypes 3, 25A and 28F).

Progress so far demonstrates that it is possible to generate an accessible _cpsA–cpsB_ sequence database for practical use by _S. pneumoniae_ serotyping reference laboratories (McEllistrem et al., 2004). In general, the more serotypes, sequence types and isolates of each that are included in the database, the greater the accuracy of serotype prediction using sequence data. The rationale for making our database available at this stage is to allow others to use and contribute to further evaluation of the effectiveness of the serotype prediction system.

The results of our preliminary evaluation of 84 selected well-characterized isolates, representing 46 serotypes, are shown in Table S4 (available as supplementary data in JMM Online). _cpsA–cpsB_ sequencing alone correctly characterized the serotype of 41 isolates and the serogroup of 13. Six isolates belonged to one of four new sequence types, which have been added to our database (Table S2, available as supplementary data in JMM Online). Serotype(s)/group(s)-specific PCR allowed correct serotype (18) or serogroup (seven) identification of another 25 isolates (including those belonging to new sequence types). The remaining five isolates belonged to one of three pairs of serotypes, individual members of which can be rapidly distinguished using serotype-specific antisera (see Table S4, available as supplementary data in JMM Online).

**Other potential uses of the _cpsA–cpsB_ database.** It has been recently reported that serotype 14 variants of the France 9V–3 clone from Baltimore, Maryland, can be differentiated by _cpsB_ gene sequence variation (McEllistrem et al., 2004). In addition, sequence variation in _cpsB_ is related to _S. pneumoniae_ strain virulence (Morona et al., 2004). Incorporation of these sequence variants and related epidemiological and virulence data into the _cpsA–cpsB_ database would allow them to be easily recognized by other researchers.

**Are shared sequence types plausible?** In order to explain the sharing of _cpsA–cpsB_ sequence types by more than one serotype, we studied their antigenic formulae (Henrichsen, 1995). Among the 24 shared _cps_ sequence types (genotypes), the majority involved closely related serotypes (or phenotypes). However, four (2-41A, 10A-17A, 10A-23F, 13-20) were shared between apparently unrelated serotypes (no antigenic cross-reactions) and three (11A-11D-18F, 15B-15C-22F-22A, 17F-35B-35C-42) between both cross-reacting and non-cross-reacting serotypes (Henrichsen, 1995) (see Table S3, available as supplementary data in JMM Online). The latter probably can be explained by recombination events (Coffey et al., 1998, 1999).

_S. pneumoniae_ is characterized by high-frequency recombination within the _cps_ gene cluster, including _wzx_, leading to serotype ‘switching’ among isolates within genetic lineages defined by relationships between the more conserved housekeeping genes (Coffey et al., 1998; Jiang et al., 2001). Although _wzx_ sequences are usually highly variable (Samuel & Reeves, 2003), we found that those of 24 serotypes share high-level (72–100%) homology. We found three main recombination sites within these 24 _wzx_ sequences (base positions 395, 775 and 1150) using PhylPro 1.0 (Weiller, 1998), which generated the diagrammatic representation of polymorphic sites and hypothetical recombination events as shown in Fig. S2 (available as supplementary data in JMM Online). These regions of high-level similarity in _wzx_ suggest recent recombination.

**Are _wzx_ and _wzy_ helpful?**

In our previous study, we showed that _wzx_- and _wzy_-based PCRs increase the accuracy of _cpsA–cpsB_ sequence-based serotype prediction (Kong & Gilbert, 2003; Rubin & Rizvi, 2004). Therefore, in order to extend our serotype-prediction...
Based on previous studies, both wzy and wzy should be serotype-specific (Jiang et al., 2001; Kong & Gilbert, 2003), but the present study suggests that this is not straightforward. Our analysis showed that, for most serotypes, wzy is shorter but more heterogeneous than wzy (see Tables S3 and S5, available as supplementary data in JMM Online). These observations, as well as evidence of wzy recombination events (see above and Fig. S2, available as supplementary data in JMM Online), suggest that wzy is a more suitable target for serotype(s)/group(s)-specific PCR for all 90 serotypes except serotype 3, which lacks these genes and so will need to be identified on the basis of other serotype 3-specific cps genes (Kong & Gilbert, 2003).

To increase the predictive accuracy of our system, we designed 26 serotype(s)/group(s)-specific PCRs targeting wzy and two targeting wzy, in addition to those developed in our previous study. The sensitivities and specificities of the 26 new PCR primer pairs were assessed, initially, for the corresponding shared sequence types and then with a reference set of all 90 serotypes (see Table S1, available as supplementary data in JMM Online for primer pair specificity). All primer pairs amplified isolates belonging to corresponding serotypes (see Tables S1–S4, available as supplementary data in JMM Online) and did not amplify unrelated serotypes. As shown in our previous study (Kong & Gilbert, 2003), partial wzy and wzy sequencing can distinguish serotypes 7B and 7C from 40, 10F from 10C, 11F/11B (identical) from 11C, 12A/46 (identical) from 12F/12B/44 (identical), 35A from 35C/42 (identical) and 35F from 47F. However, they cannot resolve individual serotypes within some isolates of sequence types 25F-38 and 6A-6B.

**Comprehensive MCT results.** The final MCT results for 596 isolates (427 previously studied and 169 new isolates) (Kong & Gilbert, 2003) and 112 previously published cps sequences are shown in Table S2 (available as supplementary data in JMM Online). Our database now includes all 90 S. pneumoniae serotypes and 140 partial cpsA–cpsB sequence types (including two non-serotypable strains). We have at least two isolates or sequences of 89 serotypes. We did not use the cps sequence published by Sanger Institute for serotype 25A, which was reported to be identical to that of serotype 29 and differs from our partial cpsA–cpsB sequence for the same serotype 25A strain (supplied to us and the Sanger Institute by Statens Serum Institut). Our partial sequence results for serotypes 38, 25F and 25A also show that serotype 25A cps is not identical to 29.

Most (110) sequence types correspond to a single serotype and, of these, 57 are represented by two or more isolates. For 516 of 708 (73 %) isolates and published sequences, CS and MCT are identical. MCT of another 155 (22 %) isolates identified the correct serogroup. For the remaining 37 (5 %) isolates, MCT could not distinguish between members of five pairs of CSs which shared the same sequence type (7B/40, 12A/46, 25F/38, 35F/47F, 35C/42; see Table S2, available as supplementary data in JMM Online). Two antisera would be required to identify individual members of these groups.

**Relationship between the partial cpsA–cpsB, wzy and wzy trees**

In the partial cpsA–cpsB tree (see Fig. S1, available as supplementary data in JMM Online), and as suggested in Table S2 (available as supplementary data in JMM Online), some serotypes are clustered because they share the same or very similar sequences. In the wzy and wzy trees these show similar relationships (see Figs S3 and S4, available as supplementary data in JMM Online), but there are differences between the trees. For example, serotypes 17A, 34 and 10B, which are closely related in the cpsA–cpsB tree (see Fig. S1), are only distantly related in both the wzy and the wzy trees (see Figs S3 and S4), suggesting that they have different evolutionary histories. This illustrates the potential risk of using a single gene or even one gene cluster to infer phylogenetic relationships (Trzcinski et al., 2003). It also implies that, for a final accurate MCT prediction system, the combination of different cps genes may increase the predictive accuracy. Based on our study, we recommend the combined use of both cpsA–cpsB and wzy, at least.

**PCR or microarray?**

In future, microarray (genechip or equivalent)-based technology should be a practical solution for MCT prediction of all 90 pneumococcal serotypes (Magee et al., 2001). As a prototype, we have developed a practical multiplex PCR and reverse line blot hybridization assay (van den Brule et al., 2002; Wang et al., 2004) to identify the 23 serotypes included in the polysaccharide vaccine. This assay showed very promising results in preliminary evaluation, using the 90-serotype reference panel and a small number of clinical isolates (data not shown). The results will be reported separately after systemic evaluation of a large number of clinical isolates. We are also trying to develop a genechip microarray to identify all the 90 serotypes. Meanwhile, we will use the cpsA–cpsB sequencing and selected wzy/wzy PCR strategy we previously described (Kong & Gilbert, 2003), for which we now have 26 additional primer sets, to resolve shared sequence types (Rubin & Rizvi, 2004).
The relationship between cps gene clusters and CSs

Because \textit{cpsA}--\textit{cpsB}, \textit{wzx} and \textit{wzy} PCR/sequencing cannot resolve all serotypes, we studied selected whole \textit{cps} gene cluster sequences, especially for serotypes in which \textit{wzx} and \textit{wzy} were very similar (see Table S3, available as supplementary data in JMM Online). Nevertheless, some serotypes remain unresolved, either because the heterogeneity between their \textit{cps} sequences was minor and inconsistent (e.g. serotypes 6A and 6B) (Kong & Gilbert, 2003) or because the serotype-specific gene was located outside the \textit{cps} gene cluster (e.g. serotype 37) (Llull \textit{et al.}, 2001). We cannot rule out the possibility that some rare serotypes have arisen as a result of aberrant gene replication and expression – such as serotypes 15B and 15C (van Selm \textit{et al.}, 2003) – or as an isolated accidental event (Waite \textit{et al.}, 2001, 2003), without a consistent molecular basis, which could explain their rarity (Henrichsen, 1995).

Benefits from the Sanger Institute \textit{S. pneumoniae} capsular loci project

In addition to several other completed and continuing \textit{S. pneumoniae} genomic projects (Hoskins \textit{et al.}, 2001; Tettelin \textit{et al.}, 2001), the \textit{S. pneumoniae} capsular loci project at the Sanger Institute (http://www.sanger.ac.uk/Projects/S_pneumoniae/CP5/) has been an invaluable resource in development of our MCT prediction system. Without it, our annotation of \textit{wzx} and \textit{wzy} and development of many of our serotype(s)/group(s)-specific PCR would have been impossible. By making available whole \textit{cps} gene cluster sequences, it also helped us to understand relationships between different serotypes that share the same sequence. Integration of the Sanger Institute \textit{S. pneumoniae} \textit{cps} cluster sequences into our database, allowed us to determine the subtypes to which they belong and examine them within the context of a larger \textit{S. pneumoniae} population. However, discrepancies between our results and those of the Sanger Institute for several \textit{cps} sequences, including serotypes 3, 25A and 28F, need to be resolved by repeat sequencing.

Conclusion

In this study, we have extended our previous MCT prediction system to 90 serotypes. The combination of \textit{cps} sequence data from the \textit{S. pneumoniae} capsular loci project and other known mechanisms (Llull \textit{et al.}, 2001; Waite \textit{et al.}, 2001) cannot fully account for all conventional serotype differences. Therefore, it is too early to fully replace CS with MCT (Hall, 1998). However, MCT can be used as an objective alternative to identify serotypes of \textdegree 73 \% of isolates and serogroups of another 22 \%, and limit the identification of the remainder to two serotypes. It can resolve discrepancies in CS and identify non-serotypable isolates. Moreover, it will allow development of rapid and relatively inexpensive typing systems (such as reverse line blot or, in future, genchips) for surveillance of distribution and prevalence of serotypes/groups and other important characteristics, such as anti-biotic resistance and virulence markers (Magee \textit{et al.}, 2001). However, unresolved controversies between CS and MCT deserve further study to improve our understanding of CS and the accuracy of the MCT system.

ACKNOWLEDGEMENTS

We are grateful for sequence data from and constructive dialogue with Dr Stephen Bentley of the \textit{S. pneumoniae} Capsular Loci Sequencing Group at the Sanger Institute (http://ftp.sanger.ac.uk/pub/pathogens/spn/) and Dr Margit Staum Kaltoft, Statens Serum Institut, Denmark. We thank Dr Michael Watson (Children’s Hospital at Westmead), Professor Geoffrey Hogg and Jenny Davis (Microbiological Diagnostic Unit) and Dr Margit Staum Kaltoft for providing isolates. This study was funded, in part, by the National Centre for Immunization Research and Surveillance of Vaccine Preventable Disease, Children’s Hospital at Westmead, NSW and a Chinese government 863 Program Grant (2002AA220501) to Lei Wang. We thank Mitchell Brown and Marianne Gail Stewart for serotyping selected isolates, and Mark Wheeler and Ilya Henner for sequencing.

REFERENCES


