Evolutionary Genetics of the Capsular Locus of Serogroup 6 Pneumococci

Angeliki Mavroidi,1 Daniel Godoy,1 David M. Aanensen,1 D. Ashley Robinson,2† Susan K. Hollingshead,3 and Brian G. Spratt1,*

Department of Infectious Disease Epidemiology, Imperial College London, St. Mary’s Campus, London,1 and Department of Biology and Biochemistry, University of Bath, Bath,2 United Kingdom, and Department of Microbiology, University of Alabama at Birmingham, Birmingham, Alabama3

Received 18 July 2004/Accepted 6 September 2004

The evolution of the capsular biosynthetic (cps) locus of serogroup 6 Streptococcus pneumoniae was investigated by analyzing sequence variation within three serotype-specific cps genes from 102 serotype 6A and 6B isolates. Sequence variation within these cps genes was related to the genetic relatedness of the isolates, determined by multilocus sequence typing, and to the inferred patterns of recent evolutionary descent, explored using the eBURST algorithm. The serotype-specific cps genes had a low percent G+C content compared to the flanking cps genes and the rest of the genome and may have been acquired by horizontal gene transfer from distant related species (2, 10). Although it appears that the serotype-specific regions of the cps locus have been introduced into pneumococci on multiple occasions from unknown sources, the number of times this may have happened, and the time scale over which such events occurred, are obscure. Thus, we do not know if some serotypes are very much younger than others and whether new pneumococcal serotypes arise at intervals of a few decades, a few centuries, or many millennia. It is also not known why there are so many (90) different capsular types in pneumococci (12) or whether the emergence of new serotypes is positively selected for by the host immune system and, if so, the intensity of this selection. Similarly, we know little about the evolutionary relationships among the capsular loci of the immunologically cross-reactive serotypes within a serogroup. Are the serological similarities of the serotypes within a single serogroup a reflection of their divergence from a recent common ancestral serotype, or does convergence occur, in which the introduction of very different cps sequences has resulted in immunologically similar capsular polysaccharides? The answers to many of these questions should start to emerge now that the sequences of the complete cps loci of all 90 serotypes are becoming available (1, 7, 11, 13, 14, 17, 20–25, 31, 32; http://www.sanger.ac.uk/Projects/S_pneumoniae/CPS/).

The evolution of individual serotypes can be addressed by exploring the patterns of sequence diversity among isolates of a single serotype and relating these to the levels of genetic relatedness and patterns of evolutionary descent of the isolates. The evolution of the cps locus is likely to be complicated. Many of the cps genes are common to several serotypes, so a history of recombination among these homologs may confound attempts to discern ancestry. However, one or more genes within the serotype-specific region of the cps locus often have no significant nucleotide sequence similarity with the cps genes of other pneumococcal serotypes or serogroups or with any other genes in the sequence databases, although their gene products may have homology with enzymes involved in polysaccharide biosynthesis in other species (18). Unlike the flanking cps genes, in which sequence variation can arise by point mutation or by homologous recombinational replacements from the homologs in other serotypes, the accumulation of variation within a serotype-specific gene can occur initially only by point mutation and, once some variation has accumulated, by recombination between different isolates of the same serotype.

Serotype-specific cps genes that have no significant sequence similarity to genes in other pneumococcal serotypes or serogroups can therefore be used to study the evolution of a serotype or of the multiple serotypes within a multitypic serogroup. In this paper, we explore the origins of the two serotypes (6A and 6B) within pneumococcal serogroup 6 by analyzing pat-
terms of sequence variation within the central cps region and relating the sequence variation within these genes to the genetic relatedness and inferred patterns of recent evolutionary descent of the isolates. The structures of the capsular polysaccharides of serotype 6A and 6B are identical, except for the linkage between the 1-rhamnopyranosyl and d-ribitol residues in the tetrasaccharide repeat unit, which is 1→3 in serotype 6A but 1→4 in serotype 6B (15, 16, 27). In addition to showing surprisingly frequent interconversion between serotypes 6A and 6B among closely related isolates and the spread of serogroup 6 sequences to distantly related pneumococcal lineages, we demonstrate that the difference between serotypes 6A and 6B correlates with a single nonsynonymous substitution in the putative rhamnosaferase gene (wciP).

MATERIALS AND METHODS

Bacterial isolates. A group of 56 serogroup 6 isolates were from the study by Robinson et al. (28), selected to cover the diversity of serotype 6A and 6B isolates among those clonally complex in which isolates of both serotypes were observed and all available examples of serogroup 6 isolates that were identical, or very closely related, in genotypy by multilocus sequence typing (MLST) but which differed at one MLST locus (6A versus 6B). We used an MLST scheme that characterized loci different from those in the standard pneumococcal MLST scheme, and the selected isolates from the study were therefore retyped using the standard MLST scheme of Enright and Spratt (9). The matrix of pairwise differences in the allelic profiles of all isolates of serotypes 6A and 6B within the public MLST database (http://www.pubmlst.org/spneumoniae/mlst.net), and of the additional isolates we selected and characterized by MLST from the study of Robinson et al. (28), was used to construct a dendrogram (data not shown). The implied genetic relatedness of these isolates was used to select additional isolates from our collection at Imperial College that further covered the diversity of serotype 6A and 6B isolates. In this study, we analyzed all isolates of serogroup 6 in the Imperial College collection that were identical by MLST or differed at only one MLST locus (single-locus variants [SLVs]) but whose serotypes varied. The final set of 102 serogroup 6 isolates chosen for further study included 43 isolates of serotype 6A and 59 isolates of serotype 6B.

Characterization of isolates. Chromosomal DNA from each isolate was prepared, and MLST was carried out using the seven pneumococcal housekeeping loci and primers described previously (8). For each locus, sequences were obtained on both DNA strands using an ABI3700 DNA sequencer (Applied Biosystems, Warrington, United Kingdom). The conditions used for amplification of the PCR products were a denaturation step at 95°C for 5 min and 30 cycles of 95°C for 1 min, 58°C for 30 s, and 72°C for 1 min, followed by a final extension at 72°C for 10 min in a PTC-200 Thermal Cycler (MJ Research Inc., Waltham, Mass.). Sequencing was carried out as for the MLST gene fragments. The presence or absence of the indel was established by the size of the intergenic fragment amplified using primers at the end of wcif (WCI-up, 5′-CATGATGTGATAATGC-3′) and the start of wcio (WCI-down, 5′-CTCTCTGCCATTGC-3′) and of the additional gene, from the neighbor-joining method using MEGA version 3.0 (19). Variable nucleotides and amino acids within the cps sequences and translated products were displayed using MEGA version 3.0.

One serotype 6B isolate (genBank accession number AF246897) displayed a divergent sequence compared with the other serogroup 6 isolates, and in this isolate there was an ~300-bp indel between the wcio and wcif genes. The presence or absence of the indel was established by the size of the intergenic fragment amplified using primers at the end of wcif (WCI-up, 5′-ATTGGTGT TACTCTCC-3′) and the start of wcio (WCI-down, 5′-CTCCATTGCCAGT ATTGC-3′). The PCRs were performed as described above. A predicted 958-bp fragment was obtained for isolates without the indel, and a 1,267-bp fragment was obtained for those with the indel.

The structures of the cps loci of serotype 6A and 6B isolates, the similarities between the sequences, and the percent G+C content were visualized using the Artemis Comparison Tool developed by Kim Rutherford at the Sanger Institute (http://www.sanger.ac.uk/Projects/S_pneumoniae/CPS/), were aligned using the ClustalW algorithm within MEGA version 3.0 (19). Primers were designed from these aligned sequences that allowed internal fragments of the three central cps genes of serotype 6A and 6B isolates to be amplified from each isolate by PCR and sequenced. The primers used both for PCR amplification and for sequencing were wcif-up, 5′-ATGTTGAGGAGATTGTGAC-3′, and wcif-down, 5′-AG CATGATGTGATAATGC-3′, for the wcif gene; S6-wysi-F, 5′-CCTAAAG TGAGGGAATTTCC-3′, and S6-wysi-R, 5′-CTCCATTACAAAGCTGAT G-3′, for the wciP gene; and S6-wzi-F, 5′-TTGAAATGGAATTCAATGG-3′, and S6-wzi-R, 5′-GGGAGCCTGCATTAGTA-3′, for the wzi gene. The PCRs were performed in a total volume of 50 μL, consisting of 2 μL of primer, 200 μM each deoxynucleoside triphosphates, and 2.5 U of Taq DNA polymerase in 1 X PCR buffer (QIAGEN Ltd., Crawley, United Kingdom).

RESULTS

Genetic relatedness of serotype 6A and 6B isolates. Fig. 1 shows a UPGMA dendrogram based on the pairwise differences in the MLST allelic profiles of the 102 serotype 6A and 6B isolates selected for further study. There were 78 different STs. Many of the isolates were relatively distantly related to each other, but there were several clusters of isolates with closely related STs. Four STs included isolates of both serotypes 6A and 6B, and there were several further cases in which isolates of both serotypes were present in STs that were SLVs, differing at only a single MLST locus (Fig. 1).

The central cps region of serogroup 6 isolates. Figure 2 shows the structures and percent G+C contents of the cps loci of serotype 6A and two serotype 6B isolates. The central region of the cps locus has a low-average percent G+C content (28.5%) compared to that of the flanking genes and to that typical of the pneumococcal genome (39.7%). Within the central region of low percent G+C are the wcio, wcif, wcip, wzi, and wxz sequences. The wcif, wcip, wzi, and wxz sequences of the serotype 6A and 6B isolates showed >90% nucleotide se-
FIG. 1. Relatedness of serogroup 6 isolates of *S. pneumoniae*. A UPGMA tree was constructed from the matrix of pairwise differences in the allelic profiles of the 102 isolates. The serotype of each isolate is shown, followed by the isolate name, ST number, and *cps* profile. The presence of the indel between the *wciN* and *wciO* genes is also indicated. STs that include isolates of both serotypes 6A and 6B are indicated by asterisks.
FIG. 2. Regions of serotype 6A and 6B S. pneumoniae CPS and a serotype 6B isolate with a class 2 CPS sequence from GenBank (AF24897) were displayed using the Artemis Comparison Tool (http://www.sanger.ac.uk/Software/ACT/). The average, maximum, and minimum percent G/C contents along the locus and the locations of the cps genes are shown. The 300-bp indel between the wciN and wciO genes is indicated by a striped box. Blocks of red color indicate sequence homology between pairs of sequences.

Sequences of the wzy gene. The wzy gene is homologous to rhamnosyl transferases in other species, and since the serogroup 6 capsule contains rhamnose, it has been assigned this function. A 645-bp internal region of the gene was sequenced, and 11 distinct alleles were distinguished among the 102 isolates (Fig. 3). The alleles in serotype 6A isolates and in the majority of serogroup 6B isolates were relatively similar. However, two divergent alleles (wzyP8 and wzyP12), which differed from each other at only a single nonsynonymous site, were found in 14 serotype 6B isolates (Fig. 1). These divergent alleles differed at ~3% of sites from the alleles in the majority of serotype 6B isolates. A single serotype 6A isolate (ACH-C2) had an allele that was a perfect mosaic (wzyP7), in which the front 537 bp were identical to an allele found in two serotype 19A and 19F pneumococci, where the similarity was ~82%. In contrast, the upstream wzy, wzh, wzd, wze, wcha, and wciN genes and the downstream rhamnose biosynthetic genes (mllA-mllD) of the serotype 6A and 6B cps loci had high levels of sequence similarity with genes within the cps loci of pneumococci of several other serogroups.

Sequences of the wciP gene. The wciP product is homologous to rhamnosyl transferases in other species, and since the serogroup 6 capsule contains rhamnose, it has been assigned this function. A 492-bp internal fragment of the gene was sequenced from the 102 isolates. There were eight alleles, four of which were present in both serotype 6A and 6B isolates (Fig. 3). Two alleles were more divergent (wciP7 and wciP9), and with one exception, these were found only in serotype 6B isolates. The exception was wciP7, which was also present in the serotype 6A isolate ACH-C2. One allele (wciP10) had a 6-bp deletion.

Sequences of the wzx gene. The wzx gene product has been assigned by homology as a polysaccharide transporter or flipase. A 477-bp internal fragment of the gene was sequenced from the 102 isolates. There were eight alleles, four of which were present in both serotype 6A and 6B isolates (Fig. 3). Two of the alleles (wzx-6 and -7) were divergent, and a further allele was a perfect mosaic (wzx-5), the front 228 bp being identical to the divergent alleles wzx-6 and -7, whereas the rest of the fragment was identical to wzx-4. The divergent alleles were all present in
FIG. 3. Allelic variation in the cps genes. (A) Neighbor-joining trees showing the relatedness among the alleles at wciP, wzy, and wzx. The scales represent a genetic distance of 0.5%. (B) The polymorphic nucleotide sites are shown for all alleles of the three cps genes. The nucleotides are numbered in vertical format. The nucleotide at each variable site is shown for the first allele, and only the nucleotides that differ from those in this sequence are shown. The dashes show the deletion in wzy-10. Alleles marked by an asterisk were not found in this study and are from published sequences of the cps genes deposited at GenBank (wciP10 in AF298581, wzy-8 and wzx-9 in AY078347, and wzx-10 in AF316640). (C) The polymorphic amino acids in the translated nucleotide sequences are shown. Dashes show the deletion in Wzy-10.
serotype 6B isolates, except the serotype 6A isolate ACH-C2, which possessed *wzx-6* (Fig. 1).

**Sequences of the central region of the \( \text{cps} \) locus of serotype 6A and 6B isolates.** The sequences of the \( \text{wciP} \), \( \text{wzy} \), and \( \text{wzx} \) fragments were joined end to end (concatenated). Twenty-one different sequences (\( \text{cps} \) profiles) were present among the 102 serotype 6A and 6B isolates. The most prevalent \( \text{cps} \) profiles were 2-1-1 among the serotype 6A isolates (23 out of 43) and 4-2-2 and 8-7-7 among the serotype 6B isolates (31 and 9 out of 59). The polymorphic sites within the concatenated sequences (1,614 bp), and a neighbor-joining tree illustrating the relatedness among these sequences, are shown in Fig. 4.

The majority of the sequences were very similar, and there was clustering of sequences according to the serotypes of the isolates. Within the main cluster of sequences (Fig. 4C) there was only one sequence (\( \text{cps} \) profile 3-1-1, present within two...
isolates, APH-10 and AAU-19) that clustered anomalously; these isolates were serotype 6B, but their cps region clustered within sequences restricted to serotype 6A isolates. Inspection of the sequence showed that this serotype 6B sequence differed at only one nucleotide site (a nonsynonymous substitution in \( wcIP \)) from the most prevalent sequence among the serotype 6A isolates (\( cps \) profile 2-1-1).

Twelve serotype 6B isolates had very closely related sequences (\( cps \) profiles 8-7-7, 12-7-7, 8-7-6, and 8-9-7) that were >5% divergent from those of the other serotype 6B isolates (Fig. 4A and C). These 12 isolates all had divergent alleles at \( wcIP \), \( wzy \), and \( wzx \), indicating that their \( cps \) regions were likely to be divergent throughout the whole central \( cps \) region. Nine of these serotype 6B isolates had the same sequence (\( cps \) profile 8-7-7), and the other three each had a different sequence, but each differed from the prevalent class 2 sequence at only a single nucleotide site. We refer to these divergent sequences as class 2 sequences to distinguish them from the class 1 sequences in the majority of serotype 6B isolates and in all of the serotype 6A isolates we examined. One further sequence (\( cps \) profile 8-7-5), present in two serotype 6B isolates (APO-445 and APO-446), clustered on the tree between the class 1 and 2 sequences and was a perfect mosaic (Fig. 4C); the sequence throughout the \( wcIP \) and \( wzy \) fragments, and the first half of \( wzx \), was identical to that in the predominant class 2 sequence of serotype 6B isolates, whereas the rest of the \( wzx \) fragment was identical to a class 1 sequence in several serotype 6B isolates (Fig. 4A).

One serotype 6A isolate (ACH-C2) also had a sequence (\( cps \) profile 7-7-6) that clustered on the tree between the class 1 and 2 sequences and was a mosaic (Fig. 4C). As noted earlier, the first part of the \( wcIP \) fragment in ACH-C2 was identical to that in some of the serotype 6A isolates, whereas the sequence of the rest of this fragment, and of the entire \( wzx \) and \( wzy \) fragments, was identical to the class 2 sequences in a serotype 6B isolate (AIS-C31).

The average pairwise diversity among the eight distinct concatenated sequences from serotype 6A isolates (excluding the mosaic sequence) was 0.4%; the sequences differed on average at 6.2 nucleotide sites. The class 1 serotype 6B sequences were similarly uniform (0.3% average pairwise diversity; differences at an average of 4.5 sites), and those of class 2 serotype 6B isolates (excluding the mosaic sequence) were extremely uniform (0.1% diversity), with the three unique class 2 sequences each differing from the predominant class 2 sequence at only a single nucleotide site. The average divergence between the serotype 6A sequences and the class 1 serotype 6B sequences was only 0.6%, whereas the divergence between both the serotype 6A and the class 1 serotype 6B sequences and the class 2 serotype 6B sequences was 5.4%.

**Distribution of the **\( wcIP \)-**intergenic indel.** The presence or absence of the indel between the \( wcIP \) and \( wcIO \) genes was examined in all serotype 6A and 6B isolates by PCR. The indel was restricted to those isolates with class 2 \( cps \) sequences (all of which were serotype 6B) and to the one serotype 6A isolate (ACH-C2) and two serotype 6B isolates (APO-445 and APO-446) that had mosaic \( cps \) sequences (Fig. 1). Only two of the serotype 6B isolates with class 2 \( cps \) sequences lacked this indel (ASA-20 and AIS-C31) (Fig. 1).

**Horizontal spread of the serotype 6A and 6B **\( cps \) **genes into divergent pneumococcal lineages.** Figure 1 shows the considerable diversity of genotypes among the 102 serogroup 6 isolates. The majority of the serogroup 6 isolates that were distantly related to all other serogroup 6 isolates were divergent lineages, as they were also not closely related to isolates of other serogroups in the MLST database. A history of horizontal transfer was apparent from the presence of the same \( cps \) profiles in several distantly related genetic backgrounds. Although many \( cps \) profiles were found only in isolates of a single ST or were restricted to isolates that were closely related by MLST, 7 of 21 (33%) \( cps \) profiles were present in pneumococci that were distantly related and differed at ≥6 of the 7 MLST loci (Fig. 1). This was particularly true of the \( cps \) profiles 2-1-1, 4-2-2, and 8-7-7 that predominated, respectively, among serotype 6A and serotype 6B isolates with class 1 and 2 sequences. The lack of any sequence variation in the three \( cps \) gene fragments in divergent isolates that had sequence differences at all (or most) MLST loci is most readily explained by the recent horizontal spread of the serotype 6A and 6B \( cps \) locus through the pneumococcal population.

In a few cases, serogroup 6 isolates were similar in genotype to isolates of other serotypes. For example, STs 947 and 948 were SLVs of each other and were both serotype 6B and had the same \( cps \) profile (5-4-1), but they were not closely related to any other serogroup 6 isolates in the MLST database. However, using eBURST on the entire MLST database showed that they were within a large clonal complex whose predicted founder (9) was ST429, where almost all the other isolates were serogroup 23 (data not shown). These serotype 6B isolates therefore appear to have arisen from a serogroup 23 isolate by the replacement of its \( cps \) region with that from a serotype 6B isolate. Similarly, ST460 and ST529 (serotype 6A) were SLVs of each other and had the same \( cps \) profile (1-1-1) but were within a clonal complex whose predicted founder (ST97), and most SLVs, were serotype 10A, suggesting the introduction of serotype 6A \( cps \) genes into isolates of clonal complexes descended from founding genotypes expressing a different capsular serotype.

**Variation in **\( cps \) **profiles within clonal complexes.** The relatedness of the serogroup 6 isolates inferred from the dendrogram showed several examples of isolates with very similar genotypes that differed in serotype, and there were four cases where serogroup 6 isolates of the same ST differed in serotype (Fig. 1). This suggested recent changes between serotypes 6A and 6B occurring within individual clones or clonal complexes. Two of the examples where isolates of the same ST differed in serotype could be attributed to recombination, as the \( cps \) profiles of the serotype 6A and 6B isolates within both ST1094 and ST315 were completely different, whereas those in the other two STs (ST361 and ST473) differed at only a single nonsynonymous site in \( wcIP \), and the differences were more likely to be due to point mutations (see below).

The eBURST algorithm was used to identify nonoverlapping groups of related STs (clonal complexes) and to predict the founding ST of each clonal complex and the patterns of evolutionary descent of all STs in the clonal complex from the predicted founding ST (9). The \( cps \) profiles and the sequences

Downloaded from http://jb.asm.org on October 15, 2017 by guest
at the \textit{wciP}, \textit{wzy}, and \textit{wzx} genes were then mapped onto the predicted patterns of descent of the isolates to explore the extents and mechanisms of serotype change among closely related isolates. eBURST identified nine clonal complexes among the 78 STs that were resolved among the 102 serogroup 6 isolates (Fig. 5), and there were 29 singleton STs that differed in allelic profile from all other STs at two or more of the seven MLST loci. Five clonal complexes included only two STs, and in all but one case, the isolates of each ST (which are SLVs of each other) had identical serotypes and \textit{cps} profiles. The other pair of STs (ST1288 and ST1289) included one serotype 6A isolate with the \textit{cps} profile 2-1-8 and a serotype 6B isolate with a divergent \textit{class 2 cps} profile (8-7-7). These isolates are therefore very closely related in overall genotype, differing by MLST at only a single locus, but their \textit{cps} regions are very different, implying a change from serotype 6A to 6B (or vice versa) due to recombination at the \textit{cps} locus.

One clonal complex included three STs, two of which were serotype 6B, and the single isolate of each of these STs shared the same class 2 \textit{cps} profile (8-7-7). The third ST (ST315) was represented by one serotype 6B isolate with the \textit{cps} profile 12-7-7 which differed at a single nonsynonymous substitution in \textit{wciP} and by a serotype 6B isolate with a divergent class 2 \textit{cps} profile (11-1-4). In the MLST database, there were an additional 18 isolates of ST315, all but one of which was serotype 6B, and 12 of the 13 SLVs of ST315 were also serotype 6B.

The founder of this clonal complex was predicted to be ST315 (100% bootstrap support) and was almost certainly serotype 6B, and the serotype 6A isolate of this ST presumably arose by recombination at the \textit{cps} locus. There were four STs (seven isolates) within a small clonal complex whose predicted founder was ST473 (Fig. 5). Applying eBURST to the whole MLST database, combined with the additional isolates characterized here, confirmed that ST473 was the predicted founder (bootstrap support, 99%) and identified five further SLVs of ST473, all of which included only serogroup 6 isolates. Three of the four isolates of ST473, and those of two of its three SLVs, had the same \textit{cps} profile (2-1-1) and were all serotype 6A (as were all isolates of the additional SLVs of ST473 in the MLST database). The ancestral state for this clonal complex was therefore almost certainly serotype 6A with the \textit{cps} profile 2-1-1. However, the fourth isolate of ST473 (AAU-19) was serotype 6B, and its central \textit{cps} region differed from that of the serotype 6A isolates by only a single nonsynonymous substitution in \textit{wciP}, resulting in a change from the 2-1-1 to the 3-1-1 \textit{cps} profile. The serotype of AAU-19 was rechecked and confirmed to be 6B. AAU-19 was from Australia, as was one of the serotype 6A isolates of ST473 with the ancestral \textit{cps} profile 2-1-1. The isolate of the third SLV of ST473 (ST399) was also serotype 6B but had a completely different \textit{cps} profile (4-2-2). Since the \textit{cps} regions of this serotype 6B isolate differed from those of the other isolates in this clonal complex at multiple sites and at each of the three se-
quenced *cps* loci, it was considered to have arisen by a recombinational replacement at the *cps* locus, resulting in a change from serotype 6A to 6B.

The ST490 clonal complex included 11 STs and 21 isolates (Fig. 5). ST490 was identified as the founding ST of the ST490 clonal complex (bootstrap support, 97%) by using eBURST on the 102 isolates combined with all other isolates in the MLST database. Using the combined data set, there were a further 10 STs compared to Fig. 5, and with a single exception, all isolates were serogroup 6. The eight isolates of ST490 among the 102 serogroup 6 isolates, as well as two of the SLVs of ST490, were serotype 6A and had the same *cps* profile (2-1-1). However, in this clonal complex there were five STs represented by serotype 6B isolates that had four different *cps* profiles that included both class 1 and class 2 sequences and a mosaic sequence. All four of these serotype 6B *cps* profiles differed at multiple nucleotide sites from each other and from the 2-1-1 profile found in all of the serotype 6A isolates within this clonal complex. Although it is difficult to provide a single evolutionary path from ST490 to all of its assumed descendants that is consistent with the changes in the *cps* profiles, it is apparent that changes from serotype 6A to serotype 6B (or from one serotype 6B *cps* profile to another) have occurred on at least four occasions within this one clonal complex.

The largest clonal complex identified by eBURST among the 102 isolates included 28 isolates and 21 STs; the great majority of them were serotype 6B and had the *cps* profile 4-2-2. By applying eBURST to the whole MLST database, ST176 was the predicted founder (bootstrap support, 92%) of this large clonal complex. Among the isolates of the clonal complex, variation in serotype and in *cps* profile was apparent within only one lineage descended from ST171, an SLV of ST176 (Fig. 5). ST361 was predicted to be descended from ST171, and two of the three isolates of this ST were serotype 6A (*cps* profile 2-1-1), and a descendant SLV of ST361 (ST1095) was also serotype 6A and had the same *cps* profile (Fig. 5). The other isolate of ST361 (APH-10) was serotype 6B, with the *cps* profile 3-1-1, which differed by only a single nonsynonymous substitution in *wciP* from the *cps* profile of the serotype 6A isolates of ST361. This substitution in the *wciP* gene of APH-10 (and in AUA-19 within the ST473 clonal complex, which has the same unusual *cps* profile) is believed to determine whether an isolate is serotype 6A or 6B (see Discussion). The most parsimonious explanation for the origin of APH-10 is that the ancestral *cps* profile of this clonal complex (4-2-2) changed by recombination to 2-1-1, resulting in a change from serotype 6B to 6A, and that the single substitution in *wciP* changed a serotype 6A isolate of ST361 back to serotype 6B, changing its *cps* profile to 3-1-1. Repeat serotyping of APH-10 confirmed that it was serotype 6B rather than 6A. The serotype 6B isolate APH-10, and serotype 6A isolates of this ST with the *cps* profile 2-1-1, were recovered in the Philippines.

**DISCUSSION**

The central *cps* region provides the most reliable information about the evolution of serogroups. This region in serotype 6A and 6B isolates has a lower percent G+C content than the flanking genes, which suggests that it has been introduced relatively recently, on an evolutionary timescale, from an unknown foreign source. The presence in pneumococci of many completely nonhomologous polysaccharide polymerases and glycosyl transferases in the serotype-specific region of the *cps* loci of the 90 different pneumococcal serotypes or serogroups also supports the view that these genes have been introduced on multiple occasions by horizontal gene transfer.

The sequence variation within the central *cps* region of all serotype 6A and 6B isolates was sufficiently low (<6%) to indicate that the sequences are all derived from a single recent common ancestral sequence, although this ancestral sequence may not necessarily have been present within a pneumococcus. The most likely scenario is that an ancestral serogroup 6 *cps* sequence was introduced into the pneumococcus from another species and that this ancestral sequence diverged, through random genetic drift or under selection for antigenic variation imposed by the host immune system. This resulted in two subfamilies of *cps* sequences that produced capsular polysaccharide structures that differed slightly in their immunochimistry and their rhamnose-ribitol linkages and which we now recognize as serotypes 6A and 6B. Although this scenario explains the high levels of sequence similarity within and between most of the serotype 6A and 6B isolates, it does not explain the origins of the more divergent class 2 *cps* sequences. The divergence between the class 1 and class 2 sequences is substantial (5.4%) and was greater than that found in typical housekeeping genes of the pneumococcus (8). Furthermore, the two classes of serotype 6B *cps* sequences are far more divergent from each other than class 1 serotype 6B sequences are from serotype 6A sequences.

Our favored scenario is that the class 2 sequence appeared by an independent introduction from a similar but unknown source. The class 2 *cps* sequences were more uniform than those of class 1; there were four class 2 *cps* profiles, and three of these differed from the predominant (presumably ancestral) sequence (the 8-7-7 *cps* profile) at only a single nucleotide site. The low level of sequence diversity among class 2 sequences, and the presence of perfect mosaics between class 1 and 2 sequences, is consistent with the recent appearance of the class 2 sequence in the pneumococcus and its recent recombination with class 1 sequences. Subsequently, the class 2 sequence has spread horizontally within the pneumococcal population and is now found in several serotype 6B isolates that appear from MLST to be only distantly related.

The introduction of genes encoding a structurally novel capsular polysaccharide from a distantly related species should be favored by natural selection, since it produces an antigenic variant of the pneumococcus against which there is no existing natural immunity. A more difficult question is whether immune selection drives the divergence of a single serotype into two related serotypes or favors the observed mutational or recombinational events that appear to have relatively frequently changed the serotype of isolates from 6A to 6B or vice versa. Selection for such events would require that the natural antibody response against the capsule of one of these serotypes is not fully protective against the subtly different capsule of the other serotype or that serotypes 6A and 6B differ in some other way that affects transmission to new hosts. There are no robust quantitative measures of the degree of cross protection from natural immunity between serotypes 6A and 6B, but studies of immunity induced by the conjugate vaccines, which include
serotype 6B but not serotype 6A capsular polysaccharide, indicate a significantly less effective antibody response against serotype 6A isolates than against 6B isolates (30). Also, a study of hybridoma antibodies elicited by serotype 6B showed that while half of the antibodies bound with equal avidity to, and were able to opsonize, both 6A and 6B, the other half bound to serotype 6B alone and failed to opsonize serotype 6A isolates (29). Antibodies elicited by one serotype may provide reduced protection against acquisition or carriage of the other. Natural selection could therefore favor switches of serotype between 6A and 6B or the original divergence that appears to have split the ancestral serotype 6 into serotypes 6A and 6B.

There was considerable evidence for recombination at the cps locus. In addition to evidence for the dissemination of the serotype 6A and 6B cps loci into distantly related lineages, resulting in changes of serogroup by recombination, a phenomenon well documented elsewhere (3–6, 26), there were a surprising number of examples of recombination between the cps regions of serotypes 6A and 6B leading to changes from one of the serotypes to the other. Many of the recombinational events that have spread the serotype 6A and 6B capsule between strains appear to be relatively recent; a third of the serogroup 6 cps profiles were found in very different genotypes, indicating that no sequence changes within the three cps genes had occurred since the horizontal-transfer events. Similarly, the mosaic cps sequences appear to have arisen recently, as no nucleotide differences have occurred in either the putative donor or recipient sequences since their formation. These results therefore provide supporting molecular evidence for changes between serotype 6A and 6B within clonal complexes that were proposed by Robinson et al. (28).

We analyzed only the central cps region and cannot address the typical size of the cps region that is replaced when isolates change by recombination from serotype 6A to 6B or when the serogroup 6 cps locus spreads horizontally into other lineages. Recombinational crossover points in pneumococci that have changed serotype have been identified previously, and changes of serotype by recombination appear to often include most or all of the cps region (4, 5, 25). The majority of the recombinational replacements that introduced the central 2 sequences into other lineages must have been relatively large (at least 4.5 kb), since in all but two isolates the characteristic indel between wciN and wciO has been cotransferred along with wciP, wzy, and wzw.

Comparisons of the published complete sequences of the cps loci of serotype 6A and 6B isolates identify those nonsynonymous substitutions in the serotype-specific region of the locus that correlate with an isolate being serotype 6A or 6B. Within this region (wciN-wzw) of the serotype 6A and 6B isolates there were several nonsynonymous substitutions that correlated with serotype. Our examination of sequence variation in a much larger set of diverse serotype 6A and 6B isolates established that only one nonsynonymous substitution in wciP correlated perfectly with serotype; isolates with serine at residue 195 of WciP were serotype 6A, whereas those with asparagine were serotype 6B (Fig. 4B). The importance of wciP in determining serotype was also supported by examination of the class 2 cps sequences that, with one exception, were found in serotype 6B isolates and that also have asparagine at position 195. The one class 2 cps region present in a serotype 6A isolate (ACH-C2) was a mosaic that appears to have arisen by a recombinational replacement that introduced the front half of wciP (introducing the serine residue into WciP) from a serotype 6A isolate.

The most convincing evidence for the key role of residue 195 comes from the wciP3 allele, which was found in two serotype 6B isolates and was identical to the wciP2 allele of serotype 6A isolates except at one nucleotide site, which changes residue 195. The two isolates (AAU-19 and APH110) were distantly related and appear to have independently changed from serotype 6A to 6B by the same single-nucleotide change in wciP. In both cases, the unusual serotype 6B isolates were identical by MLST to multiple isolates of serotype 6A and (except for the single-nucleotide change) had the same wciP-wzy-wzw sequence. Based on these observations, we conclude that these isolates independently changed from serotype 6A to 6B by a point mutation within wciP rather than by recombination.

There is therefore strong evidence that residue 195 within WciP determines the serotype within serogroup 6, although experimental evidence is required to establish conclusively that mutagenesis of this single nucleotide in wciP changes the serotype. A key role for wciP in determining whether an isolate is serotype 6A or 6B is consistent with the functional assignment of its gene product (rhamnosyl transferase), since their capsular polysaccharides differ only in the nature of the chemical linkage of rhamnose to ribitol (15, 16, 27), which is catalyzed by a rhamnosyl transferase.

ACKNOWLEDGMENTS

We thank Moon Nahm at the University of Alabama for providing monoclonal antibodies Hyp6BM1 and Hyp6BMs and Pan Cox for serotyping and shipping of strains. We are grateful to Angela Bruegmann and Dai Griffith at the John Radcliffe Hospital, Oxford, United Kingdom, for confirming the serotypes of some of the isolates and Stephen Bentley and Julian Parkhill at the Sanger Institute, Cambridge, United Kingdom, for providing annotated sequences of the serotype 6A and 6B cps loci. We thank all those who contributed isolates to this study and to the MLST database and Jay Butler and Karen Rudolph for providing replacements for isolates lost in a freezer accident. This work was supported by the Wellcome Trust. B.G.S. is a Principal Research Fellow of the Wellcome Trust.

REFERENCES


