JOURNAL OF BACTERIOLOGY, Nov. 2007, p. 7841–7855 Vol. 189, No. 21
0021-9193/07/$08.00+0 doi:10.1128/JB.00836-07

Genetic Relatedness of the *Streptococcus pneumoniae* Capsular Biosynthetic Loci▽†

Angeliki Mavroidi,[1]‡ David M. Aanensen,[1]‡ Daniel Godoy,[1] Ian C. Skovsted,[2] Margit S. Kaltoft,[2] Peter R. Reeves,[3] Stephen D. Bentley,[4] and Brian G. Spratt[1]*

Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom[1]; Statens Serum Institut, Copenhagen, Denmark[2]; School of Molecular and Microbial Biosciences, University of Sydney, Sydney, Australia[3]; and Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, United Kingdom[4]

Received 30 May 2007/Accepted 21 August 2007

Streptococcus pneumoniae **(the pneumococcus) produces 1 of 91 capsular polysaccharides (CPS) that define the serotype. The *cps* loci of 88 pneumococcal serotypes whose CPS is synthesized by the Wzy-dependent pathway were compared with each other and with additional streptococcal polysaccharide biosynthetic loci and were clustered according to the proportion of shared homology groups (HGs), weighted for the sequence similarities between the genes encoding the shared HGs. The *cps* loci of the 88 pneumococcal serotypes were distributed into eight major clusters and 21 subclusters. All serotypes within the same serogroup fell into the same major cluster, but in six cases, serotypes within the same serogroup were in different subclusters and, conversely, nine subclusters included completely different serotypes. The closely related *cps* loci within a subcluster were compared to the known CPS structures to relate gene content to structure. The *Streptococcus oralis* and *Streptococcus mitis* polysaccharide biosynthetic loci clustered within the pneumococcal *cps* loci and were in a subcluster that also included the *cps* locus of pneumococcal serotype 21, whereas the *Streptococcus agalactiae cps* loci formed a single cluster that was not closely related to any of the pneumococcal *cps* clusters.**

Many important bacterial pathogens produce a polysaccharide capsule that is believed to contribute to virulence by aiding bacterial survival in blood (47). Capsular polysaccharides (CPS) are surface exposed and immunogenic and provide a target for the host immune response. Consequently, there has been selection for mechanisms by which encapsulated bacteria evade the host immune system, and in most encapsulated species, this has been achieved by the generation of antigenic diversity, such that strains of the pathogen may express one of a number of different CPS (40).

In *Streptococcus pneumoniae*, 90 immunochemically distinct CPS types have been identified and sera that recognize the differences between these capsules provide a serological typing scheme, resolving pneumococci into individual serotypes or into serogroups, which include multiple immunologically related serotypes (16, 17). The serotyping scheme for pneumococci has been updated many times, and the 90 currently recognized serotypes probably represent a high proportion of the total capsule diversity in the species, although a new variant within serogroup 6 (serotype 6C) has recently been reported (33).

The biochemical structures of 54 capsular types are known (22), and the sequences of the CPS biosynthetic (*cps*) loci of several serotypes have been reported over the last decade (15, 18, 26, 51), but the sequences of the *cps* loci of all pneumococcal serotypes have only recently become available, allowing relationships among serology, genetics, and structure to be explored (5).

The CPS of 88 serotypes are known to be synthesized by the Wzy-dependent pathway, and the *cps* loci encoding this pathway in these serotypes are located at the same chromosomal location between *dexB* and *aliA* (5, 18, 36, 51); the capsules of the other two serotypes (3 and 37) are synthesized by the synthase pathway and are not considered further in this paper. The Wzy-dependent pathway is also found in several other streptococcal species (25). For example, it is found in CPS biosynthesis of *Streptococcus agalactiae* (10) and *Streptococcus suis* (39), in receptor polysaccharide synthesis (RPS) in the viridans group streptococci, *Streptococcus gordonii*, *Streptococcus mitis*, and *Streptococcus oralis* (12, 48–50) and exopolysaccharide synthesis in *Streptococcus thermophilus* (7, 8).

The evolution of these streptococcal loci is almost certainly very complex, with a long history of gene capture and loss and genetic rearrangements, and it is probably unrealistic to expect to be able to untangle their evolutionary history. Rather than take a phylogenetic approach, which seems inappropriate, we have explored the relatedness of the pneumococcal *cps* loci (and other streptococcal polysaccharide biosynthetic loci) by cluster analysis, incorporating similarities in both gene content and nucleotide sequence, and used this as a framework to relate differences in the *cps* loci of closely related serotypes to differences in their CPS structures.

MATERIALS AND METHODS

Polysaccharide biosynthetic sequences. A total of 116 streptococcal polysaccharide biosynthetic sequences were used in this study; those from strains of each of the 88 pneumococcal serotypes that are known to use the Wzy-dependent pathway (accession numbers CR931632, CR931633, CR931635, CR931637 to CR931708, and CR931710 to CR931722; 5), 17 pneumococcal *cps* loci sequenced

* Corresponding author. Mailing address: Department of Infectious Disease Epidemiology, Imperial College London, Room G22, Old Medical School Building, St. Mary's Hospital, Norfolk Place, London W2 1PG, United Kingdom. Phone: 44 (0)20 7594 3398. Fax: 44 (0)20 7402 3927. E-mail: b.spratt@imperial.ac.uk.

† Supplemental material for this article may be found at http://jb.asm.org/.

‡ A.M. and D.M.A. contributed equally to this work.

▽ Published ahead of print on 31 August 2007.

by other groups (see Table S1 in the supplemental material), including the serotype 4 locus extracted from the TIGR4 genome (42), and 11 *S. agalactiae* or viridans group streptococcal sequences. The latter included sequences of the *cps* loci of *S. agalactiae* serotypes Ia and II to VIII (accession numbers AB028896, AY375362, AF163833, AF355776, AF349539, AF337958, AY376403, and AY375363, respectively; 10, 11) and the RPS loci of *S. oralis* strains 34 (AB181234; 50) and J22 (AB181235; 49) and of *S. mitis* NCTC12261.

The RPS locus of *S. mitis* strain NCTC12261 (nucleotides 215361 to 240666) was extracted from the genome sequence available at the J. Craig Venter Institute website (www.tigr.org). This sequence is not completely finished and includes some unresolved nucleotides; therefore, only the region between the *wzg* (*cpsA*) and *rmlD* (*rfbD*) genes was included. It should be noted that GenBank accession no. AB181235 was initially submitted as the sequence of the RPS biosynthetic locus of strain *S. mitis* J22, but this strain has subsequently been recharacterized as *S. oralis* (49). *S. pneumoniae* serotype 6C (33) could not be included in the cluster analysis, as its *cps* sequence was unavailable.

Bioinformatic methods and cluster analysis. TribeMCL (14) was run on the data set of 116 polysaccharide biosynthetic sequences with a TBLASTX cutoff of $1e^{-50}$ to assign the gene products to homology groups (HGs). The *cps* gene products were classified into Pfam families based on hidden Markov model profiles by using the Pfam database (4; http://www.sanger.ac.uk/Software/Pfam/). The pairwise nucleotide sequence similarities among *cps* genes whose products were assigned to the same HG were calculated by the Needleman-Wunsch algorithm by using the EMBOSS program NEEDLE (32).

Visual representation of the alignments using nucleotide similarities (BLASTN) and similarities of the gene products (TBLASTX) of the polysaccharide biosynthetic loci were performed by ACT version 6 and WebACT, an online version of the Artemis comparison tool (2, 9). The nucleotide differences between similar *cps* loci were calculated by the EMBOSS program Diffseq (MRC Rosalind Franklin Centre for Genomics Research, Hinxton, United Kingdom) and are presented elsewhere (see Table S1 in the supplemental material).

For clustering of the *cps* sequences, a modification of the method described by Tekaia et al. (41) was applied. A total of 99 polysaccharide biosynthetic loci that encode capsules synthesized by the Wzy-dependent pathway were included in the cluster analysis, i.e., 88 *S. pneumoniae cps* loci (5) and 11 other streptococcal sequences. A profile of each *cps* locus (gene content) was produced according to the presence or absence of each of the 254 HGs. The *dexB*, *aliA*, *aliB*, transposase, and group II intron genes were not included in the profiles. The products of the four *cps* genes *wzg*, *wzh*, *wzd*, and *wze* of all pneumococcal *cps* loci and those of the *S. mitis*, *S. oralis*, and *S. agalactiae* loci each fall into a single HG, and these genes also were not included in the *cps* profiles used for clustering.

Each polysaccharide biosynthetic locus was compared with all other loci, calculating the similarity scores in sequence-weighted gene content, and a 99-by-99 data table (T) resulting from the pairwise comparisons of the 99 polysaccharide biosynthetic loci was produced, where the similarity score (Tij) for each pair of biosynthetic loci is the sum of the percentage of nucleotide sequence similarities of the HGs shared between loci i and j divided by the total number of HGs present in locus j. It should be noted that Tij is normalized because it is divided by the total number of HGs in locus j (and is 100% for comparisons of each *cps* locus with itself), and the data table is not symmetrical since Tij is different from Tji if the loci i and j differ in the number of HGs. The data table was used as the input for cluster analysis by the XL-STAT software version 7.5.3 (Addinsoft, New York, NY). A distance (or dissimilarity) matrix was produced by the Pearson distance by transformation of Pearson's coefficient into dissimilarity values in a range of 0 to 1 (rows were clustered according to the columns). The distance matrix was used for cluster analysis by the unweighted-pair group method using arithmetic averages (19).

Biochemical structures and antigenic formulas. The CPS structures for 52 pneumococcal serotypes that use the Wzy-dependent pathway have been adopted from a report by Kamerling (22), incorporating the revised CPS structures of serotypes 15B, 17F, and 33F (20, 21, 24), and their symbolic representations were previously presented by Bentley et al. (5). The biochemical structures of the *S. agalactiae* CPS and the cell wall polysaccharides of *S. oralis* 34 (type 1Gn RPS) and *S. oralis* J22 (type 2G RPS) were from reports by Cieslewicz et al. (10) and Cisar et al. (12), respectively. The serotypes that react with each of the pneumococcal factor (typing) sera were obtained from data reported by Henrichsen (17).

RESULTS

Assignment of HGs. The products of the four *cps* genes (*wzg*, *wzh*, *wzd*, and *wze*) are relatively conserved in sequence in all 88 pneumococcal serotypes, but there are multiple highly divergent or nonhomologous groups of flippases, polysaccharide polymerases, initial transferases (ITs), and other *cps* gene products, including a large number of different groups of glycosyltransferases (GTs). Previously, we have classified the gene products of the pneumococcal *cps* loci into those that share significant homology (HGs) by using the program TribeMCL (5). Genes that encode proteins within the same HG have been given the same name, except for those that encode the polymerases and flippases, where the widely used generic gene names (*wzy* and *wzx*, respectively) have been retained (5, 38).

In this study, the gene products encoded within 28 additional streptococcal biosynthetic loci were added to those of the 88 pneumococcal *cps* loci encoding capsules synthesized by the Wzy-dependent pathway (see Materials and Methods) and TribeMCL was rerun on this combined set to allow these additional proteins to be assigned to pneumococcal HGs or to novel HGs. Table 1 shows the HGs that were common to both pneumococcal and other streptococcal polysaccharide biosynthetic loci. In all cases, the *cps* loci from different *S. pneumoniae* strains of the same serotype had identical HG profiles (data not shown) and almost identical sequences (see Table S1 in the supplemental material); therefore, only the pneumococcal sequences of Bentley et al. (5) were used in all further analyses.

Using the TBLASTX cutoff of $1e^{-50}$, the average amino acid sequence similarities of proteins assigned by TribeMCL to the same HG ranged from 42 to 100%, and 83% of the HGs included proteins with average sequence similarities of >70%. Members of the same HG should therefore correspond to proteins (and thus genes) with broadly similar functions.

Predicted function of *cps* gene products. A detailed analysis of the predicted functions and specificities of the products of the genes in the pneumococcal *cps* loci is presented elsewhere (1), and we have used these functional assignments (and gene names) when discussing *cps* loci that were shown to be similar by cluster analysis. The assignment of the initial sugars of the repeat units and the specificity of the GTs and other transferases are also discussed in detail elsewhere, together with the basis for these assignments (1). In many cases, the comparison of closely related *cps* loci and their CPS structures provides strength to these assignments of GT specificity (discussed below).

Cluster analysis. Figure 1 shows the clustering of the *cps* loci, based on a measurement of the extent of sharing of HGs and the degree of sequence similarity of the genes encoding the shared HGs as described in Materials and Methods. This approach was used as shared HGs could be encoded by genes that are almost identical, or that are divergent, in nucleotide sequence, and adjusting for sequence similarity results in tighter clustering when the genes encoding the shared HGs have very similar sequences than when they have less similar sequences. Clustering was used to identify those *cps* loci that possess a high level of similarity in sequence-weighted gene content, which forms a useful framework for discussing the relationships among the *cps* loci of the different pneumococcal serotypes.

The *cps* loci of the 88 pneumococcal serotypes fell into eight clusters at a distance of 0.3 (Fig. 1). The *cps* loci of pneumococcal serotypes within the same serogroup were invariably in

TABLE 1. Products of other streptococcal polysaccharide biosynthetic loci that are within pneumococcal HGs

Name in *S. pneumoniae*[a]	Organism	Name in other species
WchA	*S. agalactiae* Ia	CpsIaE
	S. agalactiae II	CDS
	S. agalactiae III	CpsE
	S. agalactiae IV	CpsE
	S. agalactiae V	CpsE
	S. agalactiae VI	CpsE
	S. agalactiae VII	CDS
	S. agalactiae VIII	CDS
	S. mitis NCTC12261	CpsE
	S. oralis 34	WchA
	S. oralis J22	WchA
Glf	*S. mitis* NCTC12261	Glf
	S. oralis 34	Glf
	S. *oralis* J22	Glf
RmlB	*S. mitis* NCTC12261	RfbB
	S. oralis 34	RmlB
	S. oralis J22	RmlB
RmlD	*S. mitis* NCTC12261	RfbD
	S. oralis 34	RmlD
	S. oralis J22	RmlD
RmlA	*S. mitis* NCTC12261	RfbA
	S. oralis 34	RmlA
	S. oralis J22	RmlA
RmlC	*S. mitis* NCTC12261	SMT0231
	S. oralis 34	RmlC
	S. oralis J22	RmlC
Wzx-1	*S. mitis* NCTC12261	SMT0226
	S. oralis 34	Wzx
	S. oralis J22	Wzx
WchF	*S. agalactiae* VIII	CDS
	S. mitis NCTC12261	SMT0220
	S. oralis 34	WchF
	S. oralis J22	WchF
WchJ	*S. agalactiae* Ia	CpsIaF
	S. agalactiae II	CDS
	S. agalactiae III	CpsF
	S. agalactiae IV	CpsF
	S. agalactiae V	CpsF
	S. agalactiae VI	CpsF
	S. agalactiae VII	CDS
WciF	*S. oralis* 34	WefD
	S. oralis J22	WefG
WcrH	*S. mitis* NCTC12261	Eps6N
	S. oralis 34	WefE
	S. oralis J22	WefE
WcwA	*S. mitis* NCTC12261	SMT0221
	S. oralis 34	WefA
	S. oralis J22	WefA
WcwC	*S. mitis* NCTC12261	Eps9H
WchV	*S. agalactiae* VIII	CDS
Wzy-19	*S. agalactiae* VIII	CDS
WcwK	*S. mitis* NCTC12261	CpsY
	S. oralis 34	WefH
	S. oralis J22	WefF
Wzy-32	*S. oralis* 34	Wzy
	S. oralis J22	Wzy
WcwF	*S. mitis* NCTC12261	SMT0223

[a] Gene products that fall into the same HG as pneumococcal Wzg, Wzh, Wzd, and Wze were present in all of the *S. agalactiae* and viridans group streptococci listed.

the same cluster. In most cases, each cluster included *cps* loci of several different serogroups but the *cps* locus of serotype 8 was sufficiently different from those of the other *cps* loci to cluster alone, as were the *cps* loci of the serotypes within both serogroups 9 and 11. A ninth cluster included all of the *cps* loci of *S. agalactiae* but no pneumococcal *cps* loci. In contrast, the *S. mitis* and two *S. oralis* RPS loci clustered among pneumococcal *cps* loci within cluster 2 (Fig. 1).

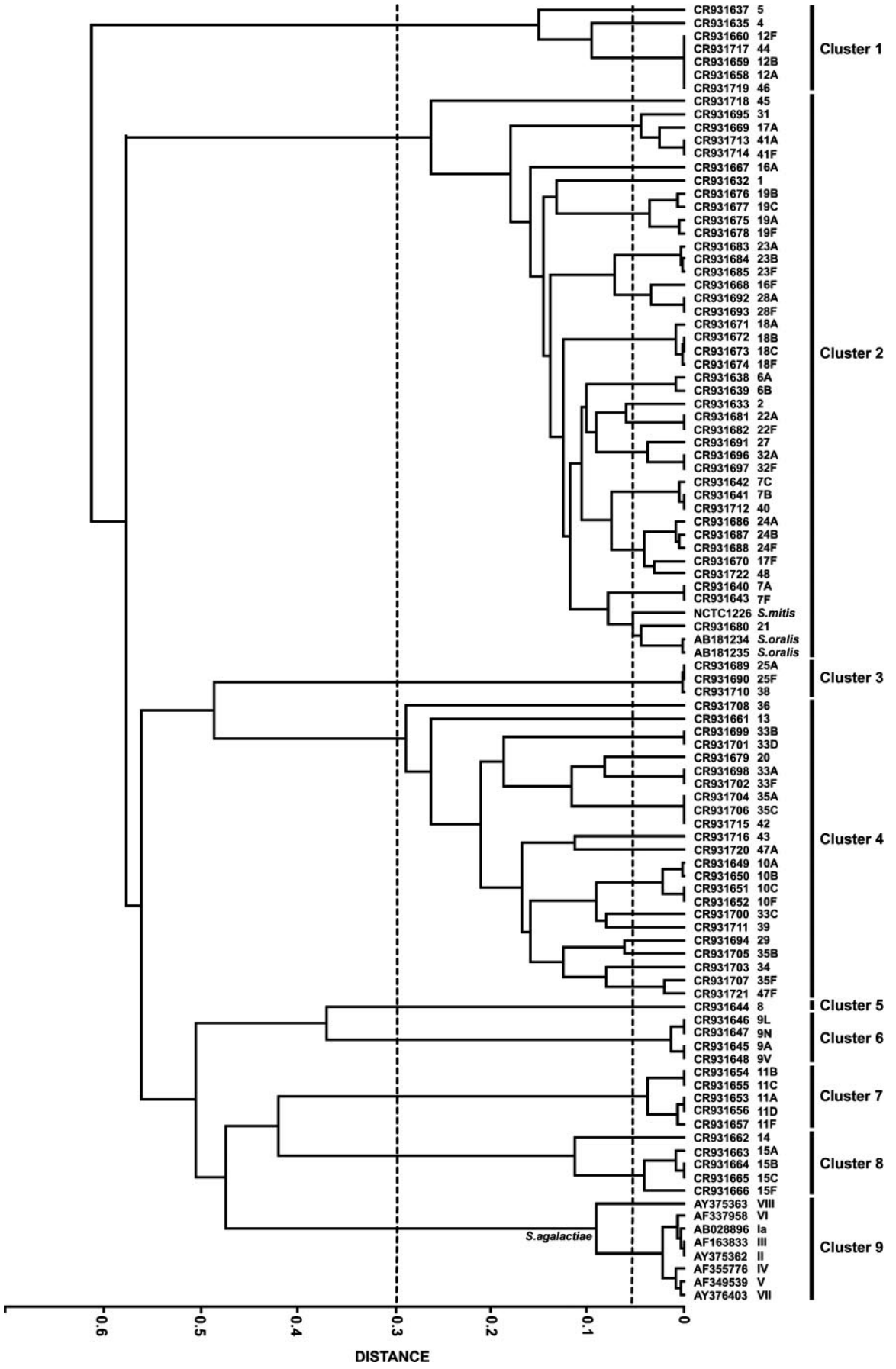
Subclusters were defined at a distance of 0.05 to identify very closely related *cps* loci (Fig. 1). All *S. pneumoniae* serotypes within the same serogroup fell into the same subcluster, except for the serotypes within serogroups 7, 16, 17, 33, 35, and 47, which were split into two or three different subclusters. Cluster analysis identified nine pneumococcal subclusters that included the *cps* loci of completely different serotypes (Fig. 1), and in three cases these loci were fully syntenic (synteny is defined here as the same genes being present in the same order). Thus, the *cps* loci of serotypes 44 and 46 were syntenic with those of serogroup 12 (Fig. 2), as shown previously (5), and those of serotypes 35A, 35C, and 42 (see Fig. 4) were syntenic, as were those of serotypes 7B, 7C, and 40 (see Fig. 3). For seven of these nine subclusters, there were factor sera that cross-reacted with the different serotypes within the subclusters (Table 2).

The similarities between the *cps* loci of different serotypes were sometimes greater than that between serotypes within the same serogroup (Fig. 1). For example, serotype 40 *cps* was more similar in sequence-weighted gene content to serotype 7B than 7B was to 7C (similarity scores of 97.5% and 90.3%, respectively) and was much more similar to the latter two serotypes, falling into the same subcluster, than these were to serotypes 7F and 7A (Fig. 1).

The serotypes within each major cluster are considered in the next sections, focusing on those that are sufficiently similar to be placed within the same subcluster (Fig. 1). The *cps* loci of those serotypes that are not within a subcluster are shown elsewhere (see Fig. S1 in the supplemental material).

Cluster 1. The *cps* loci of serogroup 12 and serotypes 44 and 46 form a subcluster and, ignoring transposon-related genes, are syntenic, presumably having diverged recently from a common ancestral *cps* locus (Fig. 2). The structures of serotypes 12F and 12A are known and differ only in the initial sugar and a side branch. As there is synteny, these differences in repeat unit structure are presumably due to the amino acid sequence variation in the IT WciI and the GT (WcxB) that is predicted to make this side branch (1, 5; see below). The structures of types 12B, 44, and 46 are unknown, but the synteny and immunological cross-reactivity between the serotypes in this subcluster (Table 2) suggest structural similarities with the CPS of serotypes 12F and 12A.

Serotype 4 and 5 *cps* loci are also members of cluster 1, but they are less closely related to the other members of the cluster (Fig. 1). Compared with serogroup 12 and types 44 and 46, the *cps* locus of type 4 shares the *wciI*, *wciJ*, *mnaA*, and *fnlA-C* genes and type 5 shares the *wciI*, *wciJ*, and *fnlA-C* genes, which explains why these serotypes cluster together (Fig. 2; see Fig. S1 in the supplemental material). The type 45 *cps* locus also shares the *wciI*, *wciJ*, *wcxB*, and *fnlA-C* genes but their sequences are distantly related to those of cluster 1 (data not


CR931637 5
CR931635 4
CR931660 12F
CR931717 44
CR931659 12B
CR931658 12A
CR931719 46
CR931718 45
CR931695 31
CR931669 17A
CR931713 41A
CR931714 41F
CR931667 16A
CR931632 1
CR931676 19B
CR931677 19C
CR931675 19A
CR931678 19F
CR931683 23A
CR931684 23B
CR931685 23F
CR931668 16F
CR931692 28A
CR931693 28F
CR931671 18A
CR931672 18B
CR931673 18C
CR931674 18F
CR931638 6A
CR931639 6B
CR931633 2
CR931681 22A
CR931682 22F
CR931691 27
CR931696 32A
CR931697 32F
CR931642 7C
CR931641 7B
CR931712 40
CR931686 24A
CR931687 24B
CR931688 24F
CR931670 17F
CR931722 48
CR931640 7A
CR931643 7F
NCTC1226 S.mitis
CR931680 21
AB181234 S.oralis
AB181235 S.oralis
CR931689 25A
CR931690 25F
CR931710 38
CR931708 36
CR931661 13
CR931699 33B
CR931701 33D
CR931679 20
CR931698 33A
CR931702 33F
CR931704 35A
CR931706 35C
CR931715 42
CR931716 43
CR931720 47A
CR931649 10A
CR931650 10B
CR931651 10C
CR931652 10F
CR931700 33C
CR931711 39
CR931694 29
CR931705 35B
CR931703 34
CR931707 35F
CR931721 47F
CR931644 8
CR931646 9L
CR931647 9N
CR931645 9A
CR931648 9V
CR931654 11B
CR931655 11C
CR931653 11A
CR931656 11D
CR931657 11F
CR931662 14
CR931663 15A
CR931664 15B
CR931665 15C
CR931666 15F
AY375363 VIII
AF337958 VI
AB028896 Ia
AF163833 III
AY375362 II
AF355776 IV
AF349539 V
AY376403 VII
S.agalactiae
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
0.6
0.5
0.4
0.3
0.2
0.1
0
DISTANCE

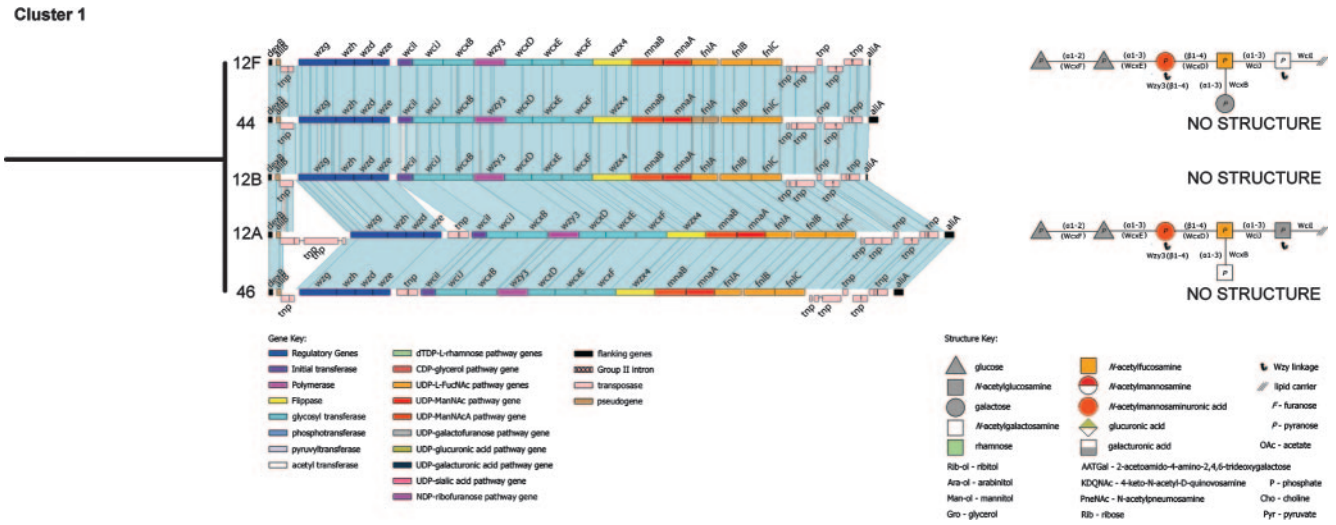


FIG. 2. Comparisons of those *cps* loci within cluster 1 assigned to the same subcluster. The relatedness of the *cps* loci in the subcluster is from Fig. 1. The *cps* loci and ACT comparisons based on the amino acid sequence similarities of the products of the *cps* loci are shown, along with the CPS repeat unit structures, where known. CPS structures are represented as in the report by Bentley et al. (5), with the initial sugar that is attached to the lipid carrier (///) on the right. The linkage between repeat units catalyzed by the Wzy polymerase is between the two sugars marked by arrows. The genes that are predicted to catalyze the linkages in the repeat units are shown on the structure. The basis of these gene assignments is discussed in detail by Aanensen et al. (1), and uncertain GT assignments are presented in parentheses. The color keys for the functional classes of genes in the *cps* loci and for the sugars (and other constituents) in the CPS repeat units are shown at the bottom.

shown) and it is placed in cluster 2 (Fig. 1; see Fig. S1 in the supplemental material).

The L-Fuc*p*NAc biosynthetic pathway genes (*fnlA-C*) are present in the *cps* loci of cluster 1, and L-Fuc*p*NAc is known to be present in the CPS of types 4, 5, 12A, and 12F, as well as the constituents of type 46. The type 44 *cps* locus has *fnlC* frameshifted; therefore, L-Fuc*p*NAc is not expected in its CPS structure. The type 5 CPS contains, in addition to L-Fuc*p*NAc, two sugars (Pne*p*NAc and KQDNAc) which are intermediates in the L-Fuc*p*NAc biosynthetic pathway (30) and within pneumococci are uniquely present in the type 5 CPS. The products of the *fnlA*, *fnlB*, and *fnlC* genes in type 5 are 76%, 98%, and 99% similar to the other members of the cluster. Therefore, the sequence divergence of FnlA may be implicated in the above difference in the L-Fuc*p*NAc pathway which results in these unusual sugars in the CPS of serotype 5 (1).

The Man*p*NAcA biosynthetic pathway genes *mnaA* and *mnaB* are both present in the *cps* loci of serogroup 12 and serotypes 44 and 46; therefore, Man*p*NAcA is present in the CPS structures of types 12A and 12F (and is predicted also to be present in types 12B, 44, and 46). In serotype 4, only *mnaA* is present and accordingly the type 4 CPS contains Man*p*NAc β-1,4 linked to L-Fuc*p*NAc. In the serotype 5 *cps* locus, neither of these genes is present and consequently the above sugars are absent from the type 5 CPS.

The IT gene *wciI* is present in the *cps* loci of cluster 1 and also in the *cps* loci of types 25A, 25F, 38, and 45, although the sequences of the latter are very divergent from those in cluster 1 (data not shown). WciI appears to have different specificities, as the initial sugar in cluster 1 varies (D-Glc*p*NAc in serotype 12A, D-Gal*p*NAc in serotypes 4 and 12F, and KQDNAc in serotype 5), although the sequences of WciI within this cluster are very similar (70 to 100% pairwise identity; average, 81.4%). This highlights the difficulties in predicting the enzymatic specificities of transferases from their amino acid sequence similarity.

The GT genes *wcxD*, *wcxE*, and *wcxF* are present only in serogroup 12 and types 44 and 46, and their products are predicted to form the linkages common to these serotypes. Inspection of the two known structures, types 12A and 12F, suggests that these GTs catalyze the formation of the common tetrasaccharide element D-Glc*p*-(α1-2)-D-Glc*p*-(α1-3)-D-Man-*p*NAcA-(β1-4)-L-Fuc*p*NAc. The GT gene *wciJ* is present in the *cps* loci of all serotypes of cluster 1 (and in type 45), and L-Fuc*p*NAc is α-1,3 linked to the repeat units but to different initial sugars (see above). Therefore, we have tentatively assigned *wciJ* as the putative α-1,3-L-Fuc*p*NAc transferase gene. As in the case of the IT of cluster 1, the specificities of the linkage made by similar members (average pairwise identity, 83.9%) of the same HG (WciJ) appear to vary. The GT gene *wcxB* is present in serogroup 12 and types 44 and 46 (and in type 45) and is suggested to encode the GT catalyzing the α-1,3 linkage of D-Gal*p*NAc to α-L-Fuc*p*NAc in the CPS of type 12A and of D-Gal*p* to α-L-Fuc*p*NAc in type 12F (and type 45).

FIG. 1. Cluster analysis of streptococcal polysaccharide biosynthetic loci. The relatedness of the loci was determined from the extent of sharing of HGs, adjusted for the percentage DNA sequence similarity between the genes encoding the shared HGs, as described in Materials and Methods. Clusters are defined by the dotted line at a distance of 0.3, and subclusters are defined by the dotted line at 0.05. Serotypes of *S. pneumoniae* (and *S. agalactiae*) or species of viridans group streptococci are shown at the end of each branch of the tree.

TABLE 2. Serotypes that react with each pneumococcal typing serum

Serum[a]	Serotype(s)
1	1
2	2
3	3
4	4
5	5
6a	6A, 6B, 33D
6b	6A
6c	6B
7a	7F, 7A, 7B, 7C
7b	7F, 7A
7c	7A
7d	7B, 7C
7e	7B
7f	7C
7g	7C, 20, 40
7h	7B, 7C, 19B, 19C, 24F, 24B, 40
8	8
9a	9A, 9L, 9N, 9V
9b	9L, 9N
9c	9A, 9L, 9V
9d	9A, 9V
9e	9N, 36
9f	9L
9g	9V
10a	10F, 10A, 10B, 10C
10b	10F, 10B, 10C
10c	10A, 10B, 10C
10d	10A, 10B, 39
10e	10B
10f	10C
11a	11F, 11A, 11B, 11C, 11D
11b	11F, 11B, 11C, 11D
11c	11A, 11C, 11D
11d	11A, 11C, 16F
11e	11F, 11A, 11D
11f	11B, 11C
11g	11F, 11B
12a	12F, 12A, 12B
12b	12F, 12B, 44
12c	12A, 12B, 46
12d	12F, 12A, 44
12e	12B
13	13
13b	13, 29
14	14
15a	15F, 15A, 15B, 15C, 23A
15b	15F, 15B
15c	15F, 15A
15d	15A, 15B, 15C
15e	15B, 15C
15f	15F
15g	15A
15h	15B
16a	16F, 16A
16b	16F, 28F
16c	16A
17a	17F, 17A
17b	17F
17c	17A
18a	18F, 18A, 18B, 18C
18b	18F, 18A, 18B, 18C, 23F
18c	18F, 18C
18d	18A
18e	18B, 18C
18f	18F
18g	18B
19a	19F, 19A, 19B, 19C
19b	19F
19c	19A, 19B, 19C
19d	19F, 19A
19e	19B
19f	19C
20	20
20b	20, 31, 33A, 35A, 35C, 42
21	21
22a	22F, 22A
22b	22F
22c	22A
23a	23F, 23A, 23B
23b	23F, 23B
23c	23A
23d	23B, 28F, 28A
24a	24F, 24A, 24B
24b	24F, 24B
24c	24A
24c	24F, 24A
24e	24B
25a	25F, 25A
25b	25F, 38
25c	25A
27	27
27b	27, 32F, 32A
28a	28F, 28A
28b	28F
28c	28
29	29
29b	29, 35B
31	31
32a	32F, 32A
32b	32A
33a	33F, 33A, 33B, 33C, 33D
33b	33F, 33A
33c	33B, 33C, 33D
33d	33F, 33A, 33B, 33D
33e	33C
33f	33B, 33D
34	34
34b	34, 35F
35a	35F, 35A, 35B, 35C, 47F
35b	35F, 47F
35c	35A, 35B, 35C, 42
36	36
37	37
38	38
38a	25A, 38
39	39
40	40
41a	41F, 41A
41b	41F
42	42
42a	42, 35C
43	43
43b	43, 47A
44	44
44b	44, 46
45	45
46	46
47a	47F, 47A
48	48

[a] Type or factor serum.

Cluster 2. The *cps* loci of 40 pneumococcal serotypes and the *S. oralis* and *S. mitis* polysaccharide biosynthetic loci (the latter two are discussed in a separate section below) are within this large cluster, and all *cps* loci within the cluster contain the rhamnose biosynthesis genes (*rml*). All of the available CPS structures of these serotypes contain rhamnose (except for types 1 and 24B, due to the presence of frameshift mutations within the *rmlC* and *rmlD* genes, respectively), and there is a perfect correlation between the presence of the linkage L-Rha*p*-(β1-4)-β-D-Glc*p* (49) in the repeat units where the *cps* locus possesses the putative rhamnosyltransferase gene *wchF*. It is therefore predicted that WchF is the GT that catalyzes the latter linkage. The other *cps* loci in this cluster have the *rml* genes present but in association with rhamnosyltransferase genes other than *wchF*, and where possible we have assigned these transferases (1).

Serotype 1 is the only case where there is no candidate within the *cps* locus for the IT gene. The serotype 1 repeat unit contains AAT-Gal*p*, which is a component of pneumococcal teichoic acid, and this unusual sugar is also present in *Bacteroides fragilis* CPS A, the *Shigella sonnei* form I antigen, and the *Plesiomonas shigelloides* serotype O17 antigen (1). In these latter species, an IT for AAT-Gal*p* has been identified and there is a homolog in the pneumococcal chromosome (13). It seems likely that this chromosomal IT synthesizes lipid-linked AAT-Gal*p*, which can be used for repeat unit synthesis in serotype 1, as further discussed by Aanensen et al. (1).

The serotypes within serogroups 6, 18, 19, 22, 23, 24, 28, 32, and 41 each fall within the same subcluster, suggesting that they are each descended from a common ancestral *cps* locus, whereas the *cps* loci of serogroups 7, 16, and 17 are split into different subclusters (Fig. 1). The *cps* loci of serotypes within serogroups 6, 22, 28, 32, and 41 are in each case syntenic, and differences in structure should be due to amino acid variation in one or more of the *cps* gene products (Fig. 3).

Serogroup 6. CPS of serotypes 6A and 6B differ only in the rhamnose linkage to Rib-ol (α-1,3 in 6A or α-1,4 in 6B), which correlates with a single nucleotide polymorphism in the rhamnosyltransferase gene *wciP* (27). The structure of the repeat unit of serotype 6C has been reported (33) and is most similar to that of 6A, containing the Rha*p*(α1-3)D-Rib-ol linkage. Serotype 6C differs from type 6A only in the presence of a Glc*p* residue as the second sugar, as opposed to Gal*p*, presumably due to amino acid differences in the GT, WciN (1).

Serogroup 7. The *cps* loci form two syntenic pairs (7F-7A and 7B-7C) that cluster apart from each other (Fig. 3), and the CPS structures of 7F, 7A, and 7B are known (22). Type 7A lacks only the side branch D-Gal*p*-(β1-2)-α-D-Gal*p* compared with type 7F, and this difference can be attributed to a frameshift mutation in the GT gene *wcwD*. The CPS of serotype 7B has little in common with those of types 7F and 7A. The CPS structures of types 7A, 7B, and 7F possess the L-Rha*p*-(β1-4)-β-D-Glc*p* linkage attributed to the presence of WchF, but L-Rha*p* is acetylated in types 7A and 7F, presumably due to the presence of the *wcwC* acetyltransferase gene in their *cps* loci. Additionally, the D-Glc*p*NAc-(α1-2)-α-L-Rha*p* linkage is present as a side branch in types 7A and 7F and in the main chain in type 7B and has been suggested to be the epitope common to all serotypes of serogroup 7 recognized by factor serum 7a (22), but there is no common GT as a candidate for forming the latter linkage. As mentioned previously, types 7B, 7C, and 40 constitute a separate subcluster, reflected in the serology (Table 2), with factor serum 7h cross-reacting with 7B, 7C, and 40, but not 7A or 7F, and 7g reacting with 7C and 40 (the CPS structures of types 7C and 40 are not known).

Serogroup 17. Serotype 17F and 17A *cps* loci are in different subclusters (Fig. 1) but share the IT gene *wchA*, three genes in the central *cps* region (the putative GT genes *wcrT* and *wcrV* and the acetyltransferase gene *wcrU*), and the polymerase gene *wzy-16*, although the order and the nucleotide sequences of the latter four shared genes differ substantially. The products of these shared genes should account for the synthesis of the common structural elements D-Gal*p*-(α1-3)-L-Rha*p*2Ac-(β1-4)-α-L-Rha*p* in the main chain of the repeat unit, the side branch D-Gal*p*-(β1-4)-L-Rha*p*2Ac (21), and the polymerase linkage D-Glc*p*(β1-3)D-Gal*p*, which are common to both serotypes 17F and 17A. WcrV contains two functional GT domains and is suggested to catalyze two transferase reactions (1). Except for these commonalities, serogroup 17 *cps* loci differ considerably and consequently type 17A clusters with serogroup 41 and type 31, whereas type 17F clusters with serogroup 24 and type 48 (Fig. 3).

Serogroup 18. Serotype 18F, 18A, 18B, and 18C *cps* loci form a separate subcluster; types 18B and 18C are syntenic, whereas type 18F has an extra acetyltransferase gene (*wcxM*) and type 18A lacks the acetyltransferase gene *wciX* (Fig. 3). As in serogroup 15, the O-acetylation state correlates with the presence of an intact acetyltransferase gene (*wciX*) in the *cps* loci of types 18F and 18C and accordingly the last Glc*p* residue of the repeat unit is acetylated, whereas it has a frameshift mutation in type 18B and is absent in type 18A. Type 18F additionally has the rhamnose residue acetylated due to the presence of the extra acetyltransferase gene *wcxM*.

Apart from the differences in the acetylation patterns, the CPS of types 18B, 18C, and 18F are identical; type 18A differs only in the third residue of the repeat unit (α-D-Glc*p*NAc instead of α-D-Glc*p*). Besides the rhamnosyltransferase gene *wchF*, serogroup 18 *cps* loci possess three shared GT genes (*wciU*, *wciV*, and *wciW*). WciV has homology to β-1,4-D-Gal transferases and is likely to catalyze the linkage of β-1,4-D-Gal*p* to α-D-Glc*p*NAc in type 18A but to α-D-Glc*p* in types 18B, 18C, and 18F. WciW possesses the same Pfam domain (PF05704) as WchN, which has been suggested to form the D-Gal*p*-(α1-2)-β-D-Gal*p* linkage in serogroup 15 (5). Serogroup 18 CPS structures do not have the latter linkage but do have a D-Glc*p*-(α1-2)-β-D-Gal*p* linkage. Thus, we propose that WciW functions as the α-1,2 D-Glc*p* transferase in serogroup 18. WciU would therefore, by a process of elimination, function as the putative D-Glc*p*NAc-(α1-3)-β-L-Rha*p* GT in type 18A or the D-Glc*p*-(α1-3)-β-L-Rha*p* GT in types 18F, 18B, and 18C. Serogroup 18 CPS contains glycerol-1-phosphate (Gro-1P), and accordingly, the sugar phosphate transferase *wciY* and the Gro-1P biosynthesis (*gct*) genes are present in their *cps* loci (5, 18).

Serogroup 19. A comparative analysis of serotype 19F, 19A, 19B, and 19C *cps* loci and the putative biosynthetic pathways has been presented previously (29). The *cps* loci of all four serotypes are in the same subcluster (Fig. 3), and those of types 19F and 19A are syntenic and their CPS structures differ only in the polymerization linkage between D-Glc*p* and α-L-Rha*p*

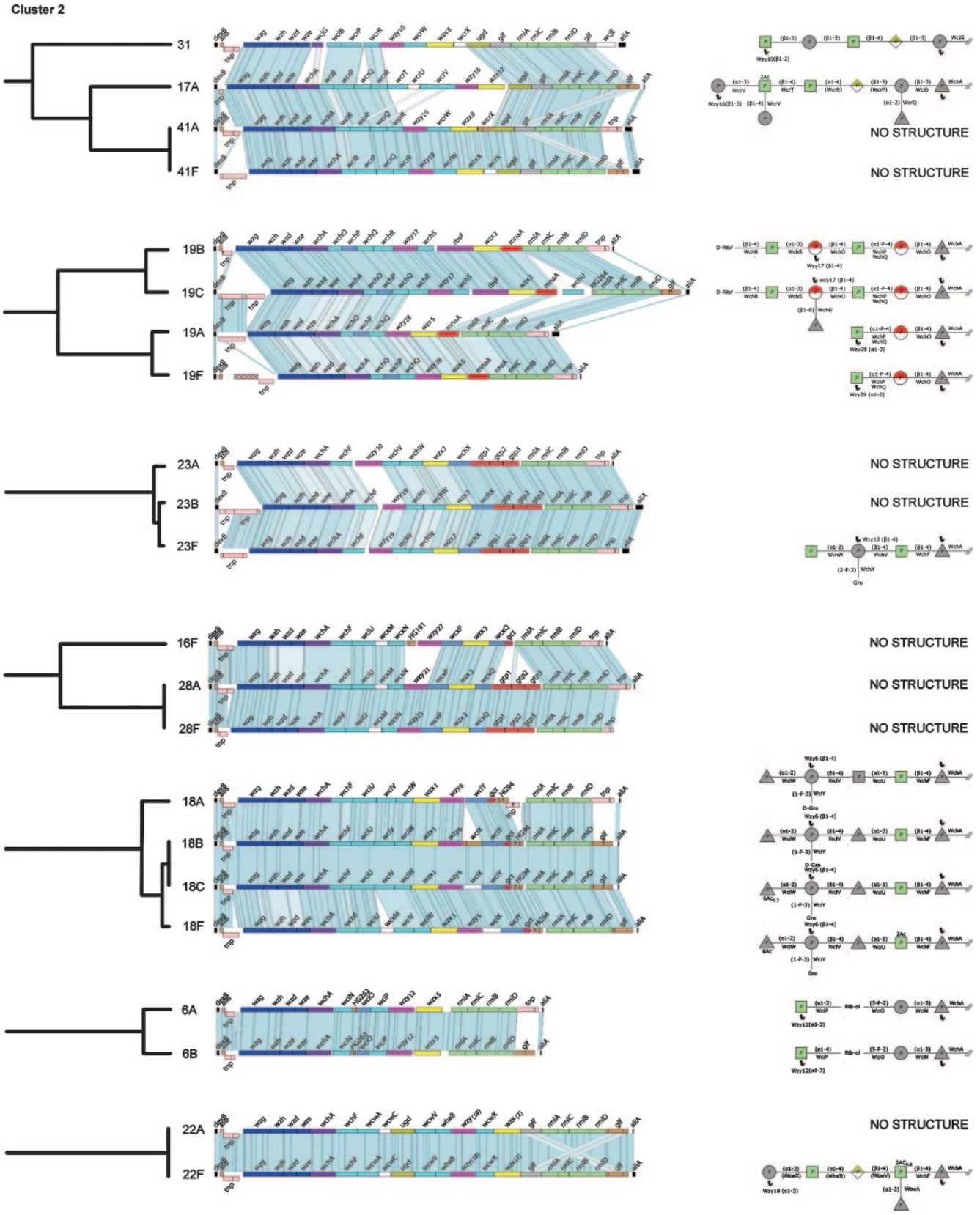


FIG. 3. Comparisons of those *cps* loci within cluster 2 assigned to the same subcluster. Details are as in Fig. 2.

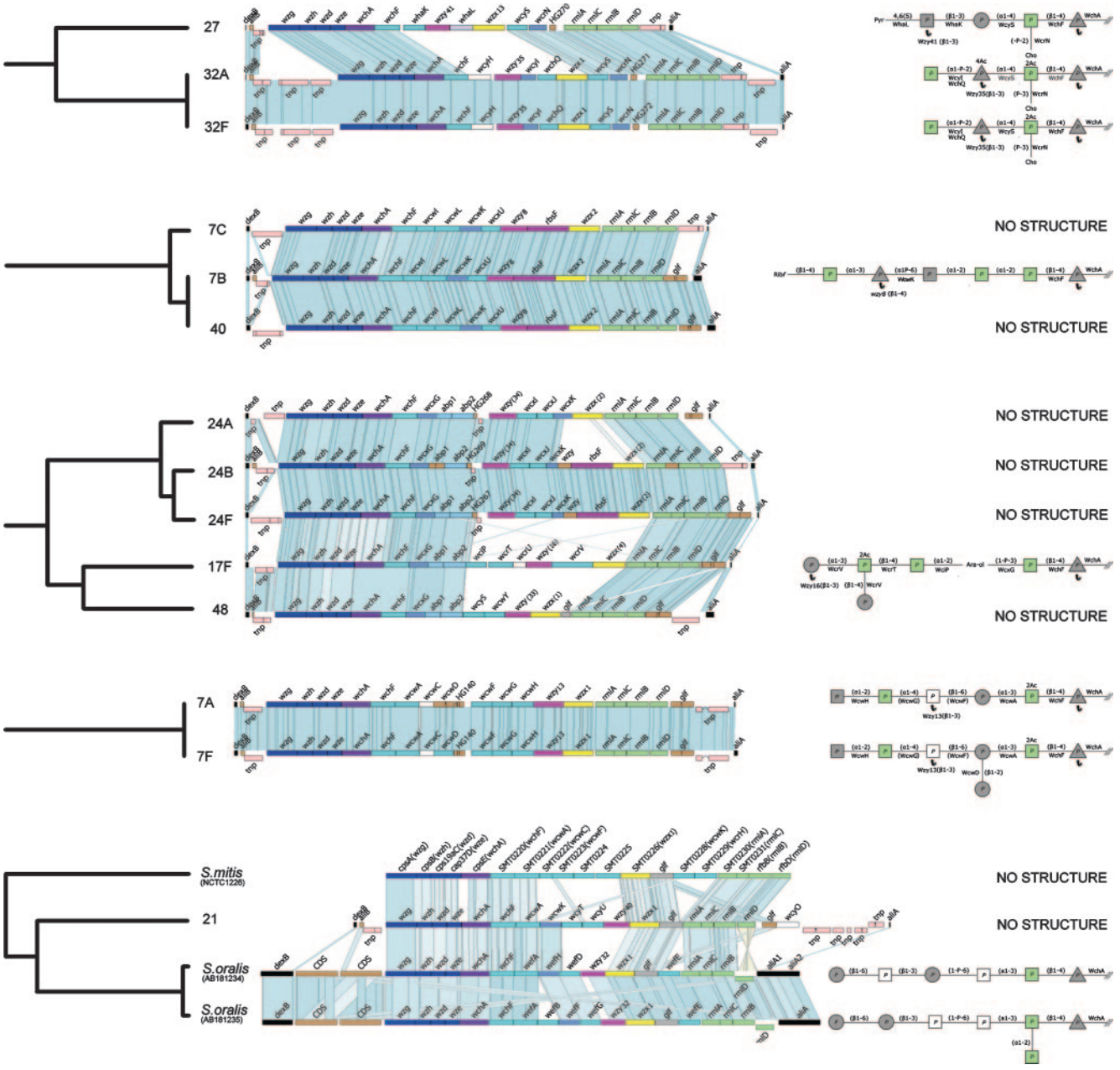


FIG. 3—*Continued*.

(α-1,2 in type 19F and α-1,3 in type 19A). Thus, it has been proposed that the differences between the products of the *wzy* genes should account for the different polymerization linkages (29). The products of the *wzy-28* genes of types 19F and 19A fall into the same HG, but they display 22% sequence divergence, which supports this view. The products of the *wzy* and *wzx* genes of types 19A and 19F fall into different HGs than those of types 19B and 19C.

Serogroup 22. The *cps* loci of serotypes 22F and 22A are syntenic (Fig. 3). The chemical differences between the CPS of these serotypes are not known since only the former structure is available.

Serogroup 23. The *cps* loci of serotypes 23F, 23A, and 23B are in the same subcluster (Fig. 3) and differ only in their polymerase genes; *wzy-19* in types 23F and 23B is replaced by *wzy-30* in type 23A, suggesting a different polymerase linkage (only the type 23F CPS structure is known). Type 23F CPS contains only one D-Glc*p* residue linked with a β-1,4 bond to β-D-Gal*p*. The IT gene *wchA* is present in the type 23F *cps* locus, and as this transfers Glc*p* to the lipid carrier, we are confident that the polymerization linkage is D-Glc*p*-(β1-4)-β-D-Gal*p*. Type 23F CPS contains glycerol-2-phosphate (Gro-2P), and as the *wchX* and *gtp1* and *-3* genes are also present in the *cps* loci of types 23A and 23B, it is assumed that they also contain Gro-2P (28).

Serogroup 24. Serotype 24F, 24A, and 24B *cps* loci form a subcluster with those of types 17F and 48, the latter two serotypes being less related as they differ in gene content (Fig. 3). The *cps* loci of types 24B and 24F are syntenic and, compared with the type 24A *cps* locus, have an extra fragment of a polymerase gene and the putative ribofuranose biosynthetic gene (*rbsF*) that also are present in types 7B, 7C, and 40. In the type 24B *cps* locus, the *abp1* and *rmlC* genes are frameshifted; therefore, no arabinitol or rhamnose is expected in the CPS. The type 24A *cps* locus lacks the *rbsF* gene (and the *wzy* fragment), and therefore no ribofuranose is expected in the CPS. The sugar phosphate transferase gene *wcxG* and the arabinitol biosynthetic pathway genes (*abp1* and *-2*) are only found in the *cps* loci of serogroup 24 and types 17F and 48.

Serogroup 28. Serotype 28F and 28A and type 16F *cps* loci constitute a subcluster (Fig. 3), and their CPS structures (none of these are known) show immunological cross-reactions, factor serum 16b reacting with both serotypes 28F and 16F (Table 2). The type 16F *cps* locus is very similar to those of serotypes 28F and 28A, but it lacks the *gtp1* and *-3* genes and has an additional gene (*gct*) and a gene fragment (*HG191*). In addition, the *wzy* gene product of type 16F falls into a different HG. In contrast, the type 16A *cps* locus has relatively little in common with that of serotype 16F and is in a different subcluster (Fig. 1).

Serogroup 32. Serotype 32F, 32A, and 27 *cps* loci constitute a subcluster (Fig. 3), and the extensive similarity is reflected in the commonalities of their CPS structures and immunological cross-reaction with factor serum 27b (Table 2). The CPS structures of serogroup 32 differ only in the acetylation pattern (type 32A has an extra acetyl group at the α-D-Glc*p* residue), presumably due to sequence differences in the acetyltransferase WcyH (5% sequence divergence between types 32A and 32F). The *cps* locus of type 27 shows extensive similarity at the 5′ and 3′ regions, but the central *cps* region differs considerably. Following assignment of WchF to the catalysis of the L-Rhap(β1-4)-β-D-Glc*p* linkage of the repeat units, the GT WcyS is common and presumably catalyzes the D-Gal*p*-(α1-4)-β-L-Rha*p* linkage in type 27 CPS but the D-Glc*p*-(α1-4)-β-L-Rha*p* linkage in serogroup 32 CPS. The putative sugar phosphate transferase WcrN should catalyze the incorporation of choline via a P-2 linkage in type 27 CPS or a P-3 linkage in serogroup 32 CPS (1).

Serogroup 41. Serotype 41F, 41A, 17A, and 31 *cps* loci constitute a subcluster (Fig. 3). The *cps* loci of serogroup 41 differ only in the 3′ end, where type 41F has a second, nonfunctional, copy of *glf*, whereas type 41A has a defective transposase gene instead. The latter serotype also has a stop codon in the putative acetyltransferase gene *wcrX*; thus, there should be differences in the acetylation patterns (neither of the CPS structures is known). Serogroup 41 and type 31 *cps* loci differ in their IT genes (*wchA* in serogroup 41 and *wcjG* in serotype 31), and type 31 has a second copy of *glf*, which appears to be functional, and an extra acetyltransferase gene (*wcjE*) at the 3′ end, but it lacks the putative GT gene *wcrQ*.

The type 17A *cps* locus is syntenic with that of type 41F at the 5′ and 3′ ends. The formation of the similar trisaccharide backbone L-Rha*p*-(α1-4)-β-D-Glc*p*A-(β1-3)-β-D-Gal*f* of serotype 17A and L-Rha*p*-(β1-4)-D-Glc*p*A-(β1-3)-β-D-Gal*f* of serotype 31 (differing only in the anomeric status of L-Rha*p*) should be attributed to the presence of the GT genes *wcrP* and *wcrR* that are common to both *cps* loci. In all serotypes where *wciB* is present in the *cps* loci, there is a Gal*f* residue β-1,3 linked in the repeat units of the available CPS structures and we have assigned WciB as the GT that makes the β-1,3 linkage of D-Gal*f* to β-D-Glc*p* in type 17A (1). *wciB* is also present in serotype 31 and could be assigned to the Gal*f*-(β1-3)-β-L-Rha*p* linkage; however, we are not confident in the assignment of linkages in serotype 31, in the absence of more recent nuclear magnetic resonance data for the CPS structure (3).

Cluster 3. Serotypes 25A, 25F, and 38 are assigned to cluster 3, and all fall within the same subcluster (Fig. 1 and 4). Their *cps* loci are very similar and have the first four *cps* regulatory genes interrupted by a transposase gene and rearranged in order. The *cps* loci of types 25A and 25F are syntenic, and type 38 differs only in the presence of the GT WcyV instead of the truncated WcyE in the former types. None of the three structures are available, but the synteny and immunological cross-reactivity (factor sera 25b and 38a cross-react with CPS of types 25F and 38 and types 25A and 38, respectively; Table 2) suggest structural similarities and recent common ancestry.

Cluster 4. Cluster 4 includes the *cps* loci of 23 pneumococcal serotypes (Fig. 1 and 4). The presence of certain GT genes (for example, *wciB*, *wcrC*, *wciF*, *wcrD*, and *wciE*) and the commonalities in structure at the 3′ end of the *cps* loci of several of these serotypes appear to account for their inclusion in the same cluster. The cross-reactions of some typing factor sera with several serotypes of cluster 4 also reflect some structural commonalities, for example, 13b with types 13 and 29, 34b with types 34 and 35F, and 35a with types 35A, 35B, 35C, 35F, and 47F (Table 2). The *cps* loci of types 13 and 36 are not very closely related to any of the others in the cluster. The *cps* loci of the serotypes within serogroup 10 cluster together, but those of serogroups 33, 35, and 47 do not (Fig. 1). The *cps* loci of some serotypes in cluster 4 are much more similar to those of a serotype within another serogroup than the latter are to the other members of that serogroup, for example, types 35A, 35C, and 42 or 35F and 47F (Fig. 1).

Serogroup 10. Serotype 10F, 10A, 10B, and 10C *cps* loci are in the same subcluster (Fig. 4) and are divided into two syntenic pairs, types 10A and 10B and types 10C and 10F, but there are structures only for types 10A and 10F. The GT genes and the ribitol phosphate transferase gene that direct the synthesis of each linkage in the common backbone of their repeat units, and the different side branches, are discussed by Aanensen et al. (1). The putative acetyltransferase gene *wciG* is present in type 10C-10F loci but absent from type 10A-10B *cps* loci; nevertheless, type 10F CPS has not been reported to be acetylated.

Serogroup 33. There are also two syntenic pairs of *cps* loci in serogroup 33 (types 33A-33F and 33B-33D), but in this case the pairs are in different subclusters (Fig. 1 and 4), and the available CPS structures of types 33F and 33B are also quite different. The type 33C *cps* locus differs considerably from both of the above pairs and clusters separately from both of the above syntenic pairs (Fig. 1; see Fig. S1 in the supplemental material). The *cps* locus of type 33F, compared with the type 33A locus, differs only in the acetyltransferase gene (*wcjE*), which is frameshifted in type 33F but intact in 33A. The antigenic formulas of types 33F and 33A differ in an extra reaction

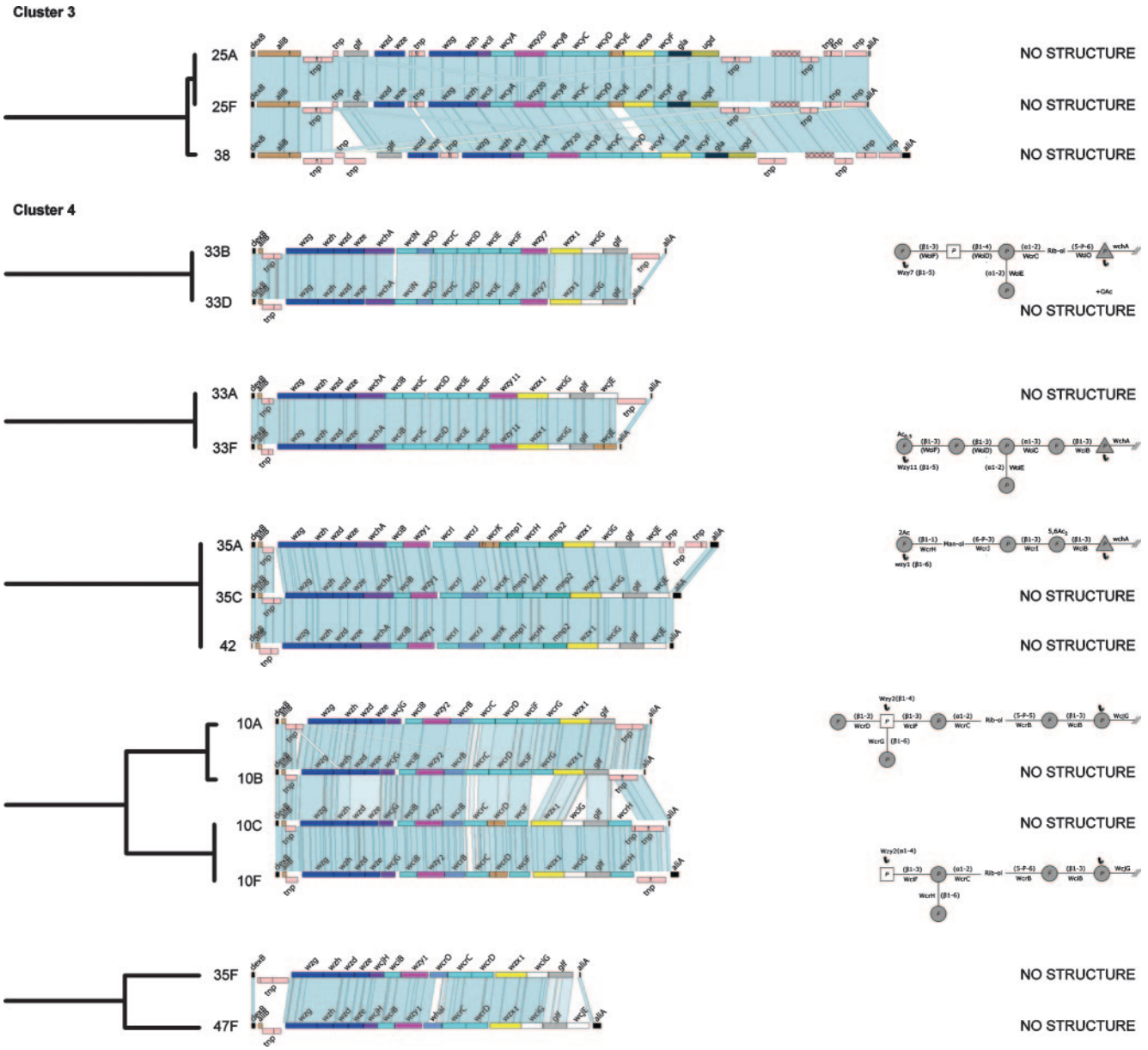


FIG. 4. Comparisons of those *cps* loci within clusters 3 and 4 assigned to the same subcluster. Details are as in Fig. 2.

of type 33A with factor serum 20b, and this serum might recognize a difference in acetylation (Table 2) (18). Type 33D reacts additionally with factor serum 6a compared to type 33B, but the difference in structure is not known as the 33D CPS structure is unavailable.

Serogroup 35. Serogroup 35, like serogroup 33, is unusual as it includes serotypes that fall into three different subclusters (Fig. 1); type 35A, 35C, and 42 *cps* loci (only the CPS structure of type 35A is known) are in the same subcluster (Fig. 4), whereas type 35F clusters apart and is very similar to serotype 47F (both CPS structures are unknown), and similarly, type 35B clusters apart and is most similar to type 29 (both structures are known). Although the serotypes within serogroup 35 fall into three subclusters, commonalities in their structures are reflected by the reactivity of factor serum 35a with all four serotypes (Table 2). The differences in the gene complements within serogroup 35 and the close similarity of some of the serotypes to unrelated serotypes are also reflected by serological cross-reactivity patterns (see below).

Except for the presence of defective transposase genes in the *cps* locus of type 35A, upstream of *aliA*, types 35A and 35C are syntenic with the type 42 *cps* locus (Fig. 4), and accordingly factor serum 35c cross-reacts with types 35A, 35C, and 42, 20b cross-reacts with types 35A and 35C, and 42a cross-reacts with types 35C and 42 (Table 2). Type 35A has a frameshift mutation in the GT gene *wcrK*, and this difference could explain the loss of reactivity with factor serum 42a.

Type 35B *cps* locus is also similar to 35A and 35C at the 5′ end (and cross-reacts with factor serum 35c) but resembles the type 29 locus in the central region (sharing *wcrJ*, *wcrM*, and *wcrH*) (Fig. 4;

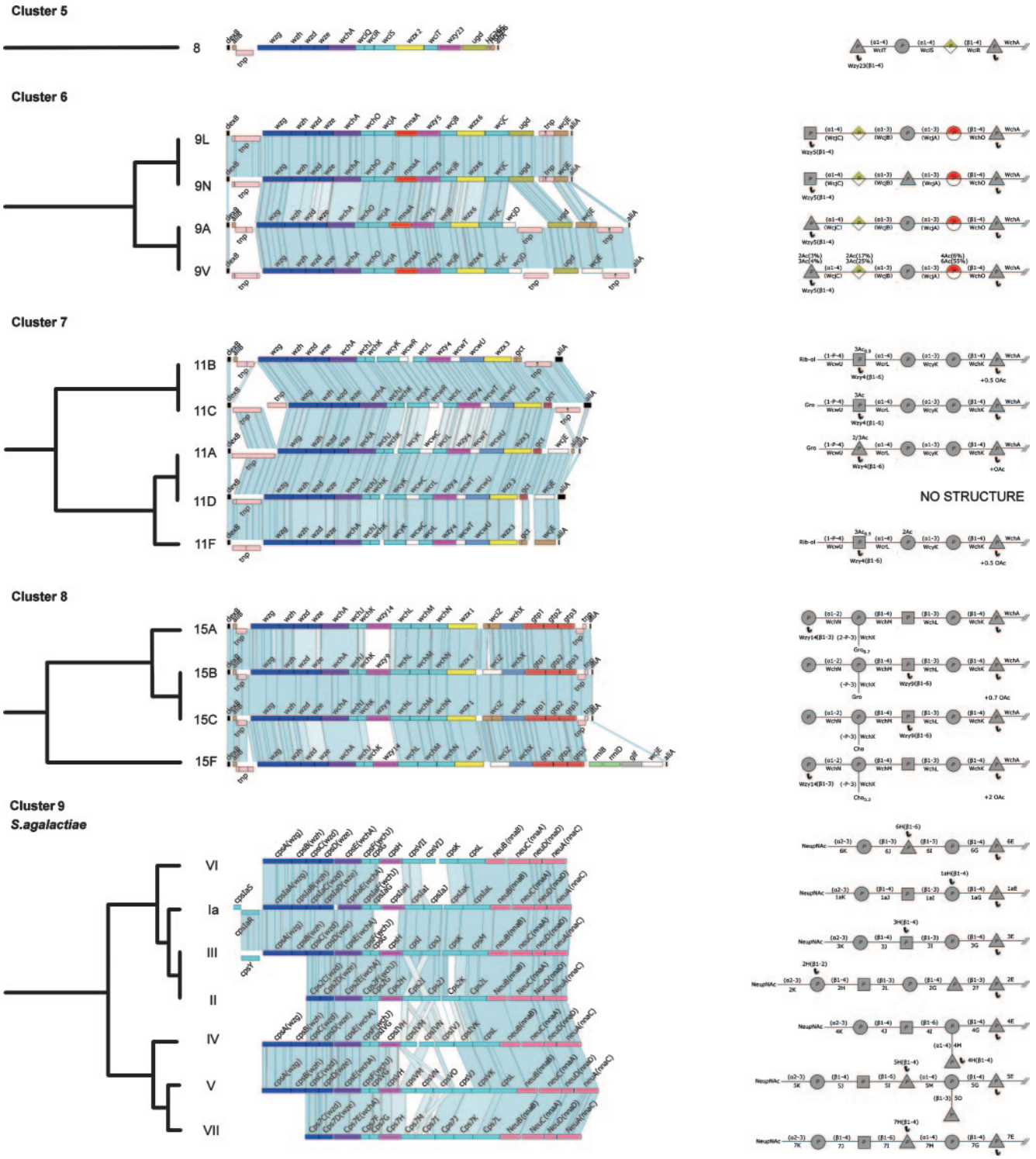


FIG. 5. Comparisons of those *cps* loci within clusters 5 to 9 assigned to the same subcluster. Details are as in Fig. 2.

see Fig. S1 in the supplemental material). The cross-reaction of types 35B and 29 with factor serum 29b is also reflected by the commonalities in their repeat units; they differ only in the initial sugar (Gal*p* transferred by WcjH and Glc*p* transferred by WchA, respectively), and the acetylation pattern of type 35B, due to the presence of an extra acetyltransferase gene (*wciG*).

Type 35F is similar at the 5′ and 3′ ends to the other serotypes within serogroup 35 but clusters with type 47F (Fig. 4), due to similarities in the central *cps* region, and accordingly, both 35F and 47F cross-react with factor sera 35a and 35b (Table 2); the CPS structures of both types are unknown. Types 35F and 47F differ only in a sugar phosphate transferase gene (type 35F has the

wcrO gene replaced by the *whaI* gene in type 47F) and the presence of an extra acetyltransferase gene (*wcjE*) in type 47F.

Serogroup 47. As mentioned above, type 47F is more similar to type 35F than to 47A, which is in a different subcluster (Fig. 1). The central *cps* region of type 47A is similar to type 43, reflected by the cross-reaction of both serotypes with factor serum 43b; the CPS structures of both types are unknown.

Cluster 5. Serotype 8 is the only member of this cluster (Fig. 1), and the biochemical functions of the IT WchA and the GT WciS and a biosynthetic pathway have previously been suggested (31, 34, 37). WciS is an α-1,4 galactosyltransferase, WciQ and WciR are homologs of the N-terminal and C-terminal halves, respectively, of the characterized glucuronosyl β-1,4 transferase SpsK of *Sphingomonas* (35) and are suggested to catalyze the formation of the D-Glc*p*A-(β1-4)-β-D-Glc*p* linkage, and by a process of elimination, WciT is suggested to catalyze the D-Glc*p*-(α1-4)-α-D-Gal*p* linkage.

Cluster 6. The *cps* loci of the four serotypes within serogroup 9 fall into the same subcluster (Fig. 5), and they form two syntenic pairs (9A-9V and 9L-9N). A putative model of CPS biosynthesis has been suggested previously (5, 45).

Cluster 7. The five serotypes within serogroup 11 have very similar *cps* loci and fall into the same subcluster (Fig. 5). There are two syntenic groups of serogroup 11 *cps* loci, 11F-11A-11D and 11B-11C, differing only in their acetyltransferase genes, and presumably they have diverged from a recent common ancestral *cps* locus. All five serotypes possess the acetyltransferase gene *wcwT*, but types 11F, 11A, and 11D have two extra acetyltransferase genes (*wcjE* and *wcwC*) whereas types 11B and 11C have one extra acetyltransferase gene (*wcwR*). WcwU has been assigned as a glycerol phosphate transferase, and the presence of Gro-1P correlates with an intact *gct* gene in types 11A and 11C; *gct* is frameshifted in types 11F and 11B, and Rib-ol is present in the CPS instead of Gro.

The *wchJ* and *wchK* genes are present in the serogroup 11 *cps* loci. These genes are also present in serotype 14 *cps*, where biochemical studies have shown that WchK is a β-1,4-galactosyltransferase that catalyzes the linkage D-Gal*p*-(β1-4)-β-D-Glc*p*, whereas WchJ is suggested to act as an enhancer of WchK activity (23). It is therefore likely that the homologous gene products in serogroup 11 are responsible for the D-Gal*p*-(β1-4)-β-D-Glc*p* linkage found in their CPS. WcrL possesses the same Pfam domain (PF04488) as WciT of type 8, which, as mentioned above, has been suggested to catalyze the D-Glc*p*-(α1-4)-α-D-Gal*p* linkage (37). The last sugar of the repeat unit is D-Glc*p* in type 11A but D-Glc*p*NAc in types 11B, 11C, and 11F (the type 11D structure is unknown), linked in all cases via an α-1,4 bond to α-D-Gal*p*. Thus, we suggest that the last sugar of the repeat unit in serogroup 11 is transferred by WcrL and differences in amino acid sequence probably affect the specificity for the donor sugar. WcyK would therefore, by a process of elimination, catalyze the linkage D-Gal*p*-(α1-3)-β-D-Gal*p* in serogroup 11 CPS.

Cluster 8. Cluster 8 includes serogroup 15 and serotype 14, and the relatedness of their *cps* loci has been presented previously (5, 23, 44). The differences in *cps* profile and the sequences of the shared genes result in the serotype 14 *cps* locus being in a different subcluster from those of serogroup 15 (Fig. 1 and 5).

Cluster 9. Cluster 9 includes only the *cps* loci of the eight *S. agalactiae* serotypes (Fig. 1), with type VIII being the most divergent member of the cluster (see Fig. S1 in the supplemental material), as has been suggested previously (10). Only a partial sequence was available for the *cps* locus of *S. agalactiae* type Ib (46), and therefore it was not included in our study. All of these loci, except that of type VIII, fall into the same subcluster (Fig. 5). These loci are not very similar to any of the pneumococcal *cps* loci (Fig. 1), although some of the gene products of the *cps* loci of the two species fall into the same HGs, including *wzg*, *wzh*, *wzd*, *wze*, the IT gene *wchA*, and the GT gene *wchJ* (Table 1). It has been suggested that *S. agalactiae* type VIII and *S. pneumoniae* type 23F could have exchanged DNA segments due to the commonalities of their CPS structures and the sequence similarities in the central *cps* region (10, 28), and this observation was confirmed by our results. Thus, the products of the *wchF*, *wchV*, and *wzy-19* (*HG168*) genes present in the type 23F *cps* locus each fall into the same HG as the corresponding *S. agalactiae* type VIII gene products and the linkages catalyzed are also thought to be similar.

Relatedness of *S. mitis and S. oralis* RPS loci to the pneumococcal *cps* loci. The *wzg*, *wzh*, *wzd*, and *wze* genes are present in the *S. oralis* and *S. mitis* RPS biosynthetic loci (Fig. 3). They also possess the rhamnose biosynthetic genes (*rml*), although the direction of transcription of *rmlD* is reversed compared to that of the pneumococcal *cps* loci (48) and the rhamnosyltransferase gene *wchF*. The *glf* gene is also present in *S. mitis* and *S. oralis*, as are the IT gene *wchA* (the most common IT gene in pneumococci), the *wzx-1* (*HG7*) flippase gene (present in 40 pneumococcal serotypes), and several GT genes that also are found in pneumococcal *cps* loci (Table 1).

The *S. mitis* and *S. oralis* sequences differ somewhat in the central region (Fig. 3), and for example, the Wzy polymerase of *S. oralis* 34 is in the same HG as that of pneumococcal serotype 36 whereas that of *S. mitis* is unique. Overall, the RPS loci of *S. mitis* and *S. oralis* are most similar to the *cps* loci of pneumococcal serotypes 7A, 7F, and 21 (Fig. 1 and 3), and they fall in the same subcluster as the pneumococcal serotype 21 locus. The two *S. oralis* sequences are the most similar to the pneumococcal serotype 21 *cps* locus (Fig. 3) and have similar HG profiles, although the sequences of the genes encoding the shared HGs differ considerably.

DISCUSSION

Clustering of pneumococcal *cps* loci on the basis of shared gene content (shared HGs), adjusted for the sequence similarity among the shared genes, provides a useful way of identifying those *cps* loci that are the most similar. The assignment of serotypes to subclusters by using a cutoff of 0.05 identifies *cps* loci that are sufficiently similar to predict that they have diversified relatively recently from a common ancestor. In some cases, these serotypes are completely syntenic and differences in CPS structure must be due to the ability of variant forms of the same IT or GTs to catalyze different reactions, but in other serotypes in the same subcluster the *cps* loci are syntenic except for a single *cps* gene that is nonhomologous (for example, types 15A and 15B, types 23A and 23B, serogroup 25 and type 38, and types 35F and 47F). The mechanism that gave rise to this phenomenon is not understood, although gene acquisition or replacement by some illegitimate recombination event is the most likely candidate. In effect, the alternative nonhomologous genes constitute a form of polymorphism and strains of these

serotypes should interconvert by recombination in the flanking common genes, as has been shown to occur for the interconversion of serotypes 6A and 6B, which differ by a single amino acid polymorphism (27).

Cluster analysis also provides a framework for correlating *cps* gene content, and gene function, of very similar *cps* loci with the known CPS structures and the immunochemical similarities inferred from cross-reactivity with factor typing sera. A single difference in gene content between *cps* loci, that correlates with a single difference in the CPS structure, provides a strong prediction of the specific reaction catalyzed by the products of these alternative genes, and in many cases prediction can be extended to situations where more that one gene differs between very similar *cps* loci, but in most cases predictions must be made by using comparative methods (1). The uncertainties in predicting the precise functions of HGs (particularly the specificity of the transferases) have been discussed elsewhere (1). However, there is little option but to predict function as extremely few of the gene products involved in pneumococcal CPS biosynthesis have been the subject of any biochemical study. We stress that many of our assignments of specific linkages to gene products are tentative (the basis for these assignments is presented by Aanensen et al. [1]), but we expect most to be correct. We provide them as they form a basis for further biochemical or structural work to firmly establish the linkages catalyzed by individual GTs.

The similarities between the genes of the *S. pneumoniae cps* loci and the polysaccharide biosynthetic loci of other streptococci have been reported previously (7, 25, 43, 51), and TribeMCL proved a useful tool to identify these homologies (Table 1). Cluster analysis allowed us to recognize the similarities between the *S. mitis* and *S. oralis* RPS loci and those of several pneumococcal serotypes, notably serotype 21, suggesting a history of interspecies exchange of polysaccharide biosynthetic loci.

In general, the immunological subdivision of pneumococci into serotypes is supported by the structures of the *cps* loci. Thus, all serotypes in the same serogroup were in the same cluster and in many cases were in the same subcluster. In those cases where serotypes in the same serogroup are syntenic, it is assumed that selection imposed by the host immune response has led to the accumulation of nonsynonymous substitutions (or in some case frameshift mutations) in one or more of the shared genes, which leads to a difference in their CPS structure, which is recognized by the typing sera. There are, however, several instances in which serotypes in the same serogroup have dissimilar *cps* loci and, conversely, where the *cps* loci of completely different serotypes are very similar. In these cases, a classification based on genetic similarity would be quite different from that which has been developed by serology. For example, serotypes 44 and 46 could be reassigned as serotypes within serogroup 12, and in other cases serotypes within the same serogroup could more logically be placed in different serogroups.

However, lumping different serotypes together, or splitting serogroups, would cause confusion and there are no strong arguments for realigning serotypes and serogroups based on genetic similarities or differences between *cps* loci. Indeed, in the context of pneumococcal conjugate vaccines (6), it is the extent of immunological cross-reactivity and the resulting presence or absence of cross-protection between serotypes that matter, and this is more likely to be captured by the current serotyping scheme than by genetic similarities and differences between *cps* loci. The serological cross-reactivity between serotypes has been defined in the rabbit, and cross-reactivity between serotypes may be different for antibodies produced in humans, but the genetic similarities between some completely different serotypes and common reactions with some typing sera suggest that there may be situations where antibodies induced in humans by CPS of one serotype might provide some cross-protection against CPS of completely different serotypes.

Conjugate vaccines that include the CPS from a limited number of serotypes (6) are likely to be the only effective way of protecting against pneumococcal disease until vaccines are developed that use antigens which can protect against disease caused by all pneumococcal strains. The availability of the sequences of the *cps* loci of all 90 serotypes and the analysis of the relatedness of the *cps* loci of the 88 serotypes that use the Wzy-dependent pathway will provide the basis to develop novel molecular serotyping methods to monitor changes in the serotypes of pneumococci following the introduction of conjugate vaccines. More generally, the clustering approach we used here can be applied to look at the relatedness of other genetic elements that share variable numbers of genes with various levels of sequence similarity.

ACKNOWLEDGMENTS

This work was funded by the Wellcome Trust. B.G.S. is a Wellcome Trust Principal Research Fellow.

REFERENCES

1. **Aanensen, D. M., A. Mavroidi, S. D. Bentley, P. R. Reeves, and B. G. Spratt.** 2007. Predicted functions and linkage specificities of the products of the *Streptococcus pneumoniae* capsular biosynthetic loci. J. Bacteriol. **189:**7856–7876.
2. **Abbott, J. C., D. M. Aanensen, K. Rutherford, S. Butcher, and B. G. Spratt.** 2005. WebACT—an online companion for the Artemis comparison tool. Bioinformatics **21:**3665–3666.
3. **Batavyal, L., and N. Roy.** 1983. Structure of the capsular polysaccharide of *Diplococcus pneumoniae* type 31. Carbohydr. Res. **119:**300–302.
4. **Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats, and S. R. Eddy.** 2004. The Pfam protein families database. Nucleic Acids Res. **32:**D138–141.
5. **Bentley, S. D., D. M. Aanensen, A. Mavroidi, D. Saunders, E. Rabbinowitsch, M. Collins, K. Donohoe, D. Harris, L. Murphy, M. A. Quail, G. Samuel, I. C. Skovsted, M. S. Kaltoft, B. Barrell, P. R. Reeves, J. Parkhill, and B. G. Spratt.** 2006. Genetic analysis of the capsular biosynthetic locus from all 90 pneumococcal serotypes. PLoS Genet. **2:**e31.
6. **Bogaert, D., P. W. Hermans, P. V. Adrian, H. C. Rumke, and R. de Groot.** 2004. Pneumococcal vaccines: an update on current strategies. Vaccine **22:**2209–2220.
7. **Bourgoin, F., A. Pluvinet, B. Gintz, B. Decaris, and G. Guedon.** 1999. Are horizontal transfers involved in the evolution of the *Streptococcus thermophilus* exopolysaccharide synthesis loci? Gene **233:**151–161.
8. **Broadbent, J. R., D. J. McMahon, D. L. Welker, C. J. Oberg, and S. Moineau.** 2003. Biochemistry, genetics, and applications of exopolysaccharide production in *Streptococcus thermophilus*: a review. J. Dairy Sci. **86:**407–423.
9. **Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell, and J. Parkhill.** 2005. ACT: the Artemis comparison tool. Bioinformatics **21:**3422–3423.
10. **Cieslewicz, M. J., D. Chaffin, G. Glusman, D. Kasper, A. Madan, S. Rodrigues, J. Fahey, M. R. Wessels, and C. E. Rubens.** 2005. Structural and genetic diversity of group B streptococcus capsular polysaccharides. Infect. Immun. **73:**3096–3103.
11. **Cieslewicz, M. J., D. L. Kasper, Y. Wang, and M. R. Wessels.** 2001. Functional analysis in type Ia group B Streptococcus of a cluster of genes involved in extracellular polysaccharide production by diverse species of streptococci. J. Biol. Chem. **276:**139–146.
12. **Cisar, J. O., A. L. Sandberg, G. P. Reddy, C. Abeygunawardana, and C. A.**

Bush. 1997. Structural and antigenic types of cell wall polysaccharides from viridans group streptococci with receptors for oral actinomyces and streptococcal lectins. Infect. Immun. **65:**5035–5041.
13. **Coyne, M. J., A. O. Tzianabos, B. C. Mallory, V. J. Carey, D. L. Kasper, and L. E. Comstock.** 2001. Polysaccharide biosynthesis locus required for virulence of *Bacteroides fragilis*. Infect. Immun. **69:**4342–4350.
14. **Enright, A. J., S. Van Dongen, and C. A. Ouzounis.** 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. **30:**1575–1584.
15. **García, E., D. Llull, R. Muñoz, M. Mollerach, and R. López.** 2000. Current trends in capsular polysaccharide biosynthesis of *Streptococcus pneumoniae*. Res. Microbiol. **151:**429–435.
16. **Heidelberger, M.** 1983. Precipitating cross-reactions among pneumococcal types. Infect. Immun. **41:**1234–1244.
17. **Henrichsen, J.** 1995. Six newly recognized types of *Streptococcus pneumoniae*. J. Clin. Microbiol. **33:**2759–2762.
18. **Jiang, S. M., L. Wang, and P. R. Reeves.** 2001. Molecular characterization of *Streptococcus pneumoniae* type 4, 6B, 8, and 18C capsular polysaccharide gene clusters. Infect. Immun. **69:**1244–1255.
19. **Jobson, J. D.** 1992. Applied multivariate data analysis. Volume II. Categorical and multivariate methods, p. 483–568. Springer-Verlag, New York, NY.
20. **Jones, C., and X. Lemercinier.** 2005. Full NMR assignment and revised structure for the capsular polysaccharide from *Streptococcus pneumoniae* type 15B. Carbohydr. Res. **340:**403–409.
21. **Jones, C., C. Whitley, and X. Lemercinier.** 2000. Full assignment of the proton and carbon NMR spectra and revised structure for the capsular polysaccharide from *Streptococcus pneumoniae* type 17F. Carbohydr. Res. **325:**192–201.
22. **Kamerling, J. P.** 2000. Pneumococcal polysaccharides: a chemical view, p. 81–114. *In* A. Tomasz (ed.), *Streptococcus pneumoniae*: molecular biology and mechanisms of disease. Mary Ann Liebert Inc., Larchmont, NY.
23. **Kolkman, M. A., B. A. van der Zeijst, and P. J. Nuijten.** 1997. Functional analysis of glycosyltransferases encoded by the capsular polysaccharide biosynthesis locus of *Streptococcus pneumoniae* serotype 14. J. Biol. Chem. **272:**19502–19508.
24. **Lemercinier, X., and C. Jones.** 2006. Full assignment of the ^{1}H and ^{13}C spectra and revision of the O-acetylation site of the capsular polysaccharide of *Streptococcus pneumoniae* type 33F, a component of the current pneumococcal polysaccharide vaccine. Carbohydr. Res. **341:**68–74.
25. **Llull, D., R. López, and E. García.** 2001. Genetic bases and medical relevance of capsular polysaccharide biosynthesis in pathogenic streptococci. Curr. Mol. Med. **1:**475–491.
26. **López, R., and E. García.** 2004. Recent trends on the molecular biology of pneumococcal capsules, lytic enzymes, and bacteriophage. FEMS Microbiol. Rev. **28:**553–580.
27. **Mavroidi, A., D. Godoy, D. M. Aanensen, D. A. Robinson, S. K. Hollingshead, and B. G. Spratt.** 2004. Evolutionary genetics of the capsular locus of serogroup 6 pneumococci. J. Bacteriol. **186:**8181–8192.
28. **Morona, J. K., D. C. Miller, T. J. Coffey, C. J. Vindurampulle, B. G. Spratt, R. Morona, and J. C. Paton.** 1999. Molecular and genetic characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* type 23F. Microbiology **145:**781–789.
29. **Morona, J. K., R. Morona, and J. C. Paton.** 1999. Comparative genetics of capsular polysaccharide biosynthesis in *Streptococcus pneumoniae* types belonging to serogroup 19. J. Bacteriol. **181:**5355–5364.
30. **Mulrooney, E. F., K. K. Poon, D. J. McNally, J. R. Brisson, and J. S. Lam.** 2005. Biosynthesis of UDP-*N*-acetyl-L-fucosamine, a precursor to the biosynthesis of lipopolysaccharide in *Pseudomonas aeruginosa* serotype O11. J. Biol. Chem. **280:**19535–19542.
31. **Muñoz, R., M. Mollerach, R. López, and E. García.** 1999. Characterization of the type 8 capsular gene cluster of *Streptococcus pneumoniae*. J. Bacteriol. **181:**6214–6219.
32. **Needleman, S. B., and C. D. Wunsch.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48:**443–453.
33. **Park, I. H., D. G. Pritchard, R. Cartee, A. Brandao, M. C. Brandileone, and M. H. Nahm.** 2007. Discovery of a new capsular serotype (6C) within serogroup 6 of *Streptococcus pneumoniae*. J. Clin. Microbiol. **45:**1225–1233.
34. **Pelosi, L., M. Boumedienne, N. Saksouk, J Geiselmann, and R. A. Geremia.** 2005. The glucosyl-1-phosphate transferase WchA (Cap8E) primes the capsular polysaccharide repeat unit biosynthesis of *Streptococcus pneumoniae* serotype 8. Biochem. Biophys. Res. Commun. **327:**857–865.
35. **Pollock, T. J., W. A. van Workum, L. Thorne, M. J. Mikolajczak, M. Yamazaki, J. W. Kijne, and R. W. Armentrout.** 1998. Assignment of biochemical functions to glycosyl transferase genes which are essential for biosynthesis of exopolysaccharides in *Sphingomonas* strain S88 and *Rhizobium leguminosarum*. J. Bacteriol. **180:**586–593.
36. **Reeves, P. R., M. Hobbs, M. A. Valvano, M. Skurnik, C. Whitfield, D. Coplin, N. Kido, J. Klena, D. Maskell, C. R. Raetz, and P. D. Rick.** 1996. Bacterial polysaccharide synthesis and gene nomenclature. Trends Microbiol. **4:**495–503.
37. **Saksouk, N., L. Pelosi, P. Colin-Morel, M. Boumedienne, P. L. Abdian, and R. A. Geremia.** 2005. The capsular polysaccharide biosynthesis of *Streptococcus pneumoniae* serotype 8: functional identification of the glycosyltransferase WciS (Cap8H). Biochem. J. **389:**63–72.
38. **Samuel, G., and P. Reeves.** 2003. Biosynthesis of O-antigens: genes and pathways involved in nucleotide sugar precursor synthesis and O-antigen assembly. Carbohydr. Res. **338:**2503–2519.
39. **Smith, H. E., V. Veenbergen, J. van der Velde, M. Damman, H. J. Wisselink, and M. A. Smits.** 1999. The *cps* genes of *Streptococcus suis* serotypes 1, 2, and 9: development of rapid serotype-specific PCR assays. J. Clin. Microbiol. **37:**3146–3152.
40. **Spratt, B. G., W. P. Hanage, and A. B. Bruegemann.** 2004. Evolutionary and population biology of *Streptococcus pneumoniae*, p. 119–135. *In* E. I. Tuomanen (ed.), The pneumococcus. ASM Press, Washington, D.C.
41. **Tekaia, F., A. Lazcano, and B. Dujon.** 1999. The genomic tree as revealed from whole proteome comparisons. Genome Res. **9:**550–557.
42. **Tettelin, H., K. E. Nelson, I. T. Paulsen, J. A. Eisen, T. D. Read, S. Peterson, J. Heidelberg, R. T. DeBoy, D. H. Haft, R. J. Dodson, A. S. Durkin, M. Gwinn, J. F. Kolonay, W. C. Nelson, J. D. Peterson, L. A. Umayam, O. White, S. L. Salzberg, M. R. Lewis, D. Radune, E. Holtzapple, H. Khouri, A. M. Wolf, T. R. Utterback, C. L. Hansen, L. A. McDonald, T. V. Feldblyum, S. Angiuoli, T. Dickinson, E. K. Hickey, I. E. Holt, B. J. Loftus, F. Yang, H. O. Smith, J. C. Venter, B. A. Dougherty, D. A. Morrison, S. K. Hollingshead, and C. M. Fraser.** 2001. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science **293:**498–506.
43. **van Kranenburg, R., H. R. Vos, I. I. van Swam, M. Kleerebezem, and W. M. de Vos.** 1999. Functional analysis of glycosyltransferase genes from *Lactococcus lactis* and other gram-positive cocci: complementation, expression, and diversity. J. Bacteriol. **181:**6347–6353.
44. **van Selm, S., L. M. van Cann, M. A. Kolkman, B. A. van der Zeijst, and J. P. van Putten.** 2003. Genetic basis for the structural difference between *Streptococcus pneumoniae* serotype 15B and 15C capsular polysaccharides. Infect. Immun. **71:**6192–6198.
45. **van Selm, S., M. A. Kolkman, B. A. van der Zeijst, K. A. Zwaagstra, W. Gaastra, and J. P. van Putten.** 2002. Organization and characterization of the capsule biosynthesis locus of *Streptococcus pneumoniae* serotype 9V. Microbiology **148:**1747–1755.
46. **Watanabe, M., K. Miyake, K. Yanae, Y. Kataoka, S. Koizumi, T. Endo, A. Ozaki, and S. Iijima.** 2002. Molecular characterization of a novel beta1,3-galactosyltransferase for capsular polysaccharide synthesis by *Streptococcus agalactiae* type Ib. J. Biochem. (Tokyo) **131:**183–191.
47. **Watson, D. A., D. M. Musher, and J. Verhoef.** 1995. Pneumococcal virulence factors and host immune responses to them. Eur. J. Clin. Microbiol. Infect. Dis. **14:**479–490.
48. **Xu, D. Q., J. Thompson, and J. O. Cisar.** 2003. Genetic loci for coaggregation receptor polysaccharide biosynthesis in *Streptococcus gordonii* 38. J. Bacteriol. **185:**5419–5430.
49. **Yoshida, Y., S. Ganguly, C. A. Bush, and J. O. Cisar.** 2005. Carbohydrate engineering of the recognition motifs in streptococcal co-aggregation receptor polysaccharides. Mol. Microbiol. **58:**244–256.
50. **Yoshida, Y., S. Ganguly, C. A. Bush, and J. O. Cisar.** 2006. Molecular basis of L-rhamnose branch formation in streptococcal coaggregation receptor polysaccharides. J. Bacteriol. **188:**4125–4130.
51. **Yother, J.** 2004. Capsules, p. 30–48. *In* E. I. Tuomanen (ed.), The pneumococcus. ASM Press, Washington, D.C.