Evolution of the Capsular Regulatory Genes in *Streptococcus pneumoniae*

Sirkka-Liisa Varvio,1,2 Kari Auranen,1,2 Elja Arjas,1,3 and P. Helena Mäkelä2

1Department of Mathematics and Statistics, University of Helsinki, and 2Department of Vaccination and Immune Protection, 3National Institute for Health and Welfare, Helsinki, Finland

The major pneumococcal virulence determinant is its capsule, and pneumococcal epidemiology is based on 91 capsular serotypes, each corresponding to the structure of the capsular polysaccharide determined by the type-specific capsular genome. Here, we provide the beginnings of an approach to intertwine serotype epidemiology, capsular regulatory gene characteristics on the basis of existing sequence information, and the reanalysis of published epidemiological data. We present an approach to explain epidemiological characteristics of serotypes on the basis of genetic differences in their capsular regulatory genes. The part of the capsular genome that regulates capsular expression falls into 2 highly divergent sequence clans: the ancestral pneumococcal capsular regulatory gene sequences (present in 49 serotypes) and laterally transferred sequences (present in 32 serotypes). Our survey of epidemiological data showed a tendency of the ancestral type of the capsular regulatory genome to be associated with carriage and the laterally transferred sequences to be associated with invasive disease isolates. The regulatory gene region shows mosaic structures that have signatures of recent recombination events, reminiscent of structures known from antibiotic resistance genes.

The pneumococcus (*Streptococcus pneumoniae*) is notorious for its success as a human parasite. Since its discovery in Pasteur’s time 130 years ago, it has kept its position as a major pathogen and a commensal in the human nasopharyngeal microflora. At any one time, 30%–90% of healthy children <5 years old are colonized by pneumococci, and the annual child death toll due to pneumococcal invasive diseases is estimated to be at least 1 million [1]. The major virulence factor of the pneumococcus is its capsule, of which 91 antigenically distinct serotypes are known at present. Remarkably, many serotypes coexist in the pneumococcal population at any time and geographical location. Serotypes isolated from invasive diseases have a frequency distribution different from that for carriage isolates [2–4].

The capsule consists of a layer of capsular polysaccharide (CPS) that helps the bacteria escape from phagocytic killing and allows their survival in invasive pneumococcal disease (IPD) [5, 6]. This traditional view has recently been complemented by the realization that the capsule also plays important—although complex—roles during the early phase of colonization of the upper airways. There it is needed to facilitate the transit of the bacteria through the viscous nasal mucus to reach their site of sojourn in association with the endothelial cells of the nasopharyngeal mucosa [7]. At the same time, the shielding function of the capsule becomes an obstacle to the adherence of the bacteria to their cellular receptors, and “transparent” colony morphology variants with reduced amount of CPS adhere better than do “opaque” variants [6, 8, 9]. Obviously, careful regulation of CPS expression in response to changing environmental clues is a prerequisite for the pneumococcus in association with the human host.

The pneumococcus has highly developed transfor-
Capsular Regulatory Genes in S. pneumoniae

RESULTS

Formation of 2 major clans by the capsular regulatory genes. The first capsular regulatory gene, wsg, of S. pneumoniae comprised 1 cluster of related sequences. In contrast, the 3 subsequent genes, wzh, wzd, and wze, exhibited deeply dichotomous phylogenetic patterns, with 2 distinct and statistically highly supported sequence clusters (figure 1A). For convenience, we call these sequence clusters the red and the blue clans. The same red-blue split applied to each of the 3 genes and extended further to wchA (the first type-specific gene), which codes for the initial glycosyl transferase in 66 serotypes but not beyond it (figure 1A). Many serogroups (clusters of antigenically related serotypes) were red-blue dichotomized (figure 1B). Three serotypes (25F, 25A, and 38) were exceptional with respect to both the location [18] and sequence of the regulatory genes, indicating their derivation from a nonstreptococcal gene pool (figure 1C).

The ancestral blues and laterally transferred reds. The analysis of the DNA sequence motifs revealed a structure consistent with recombination events that have replaced blocks of blue-type sequences with the corresponding sequences of the red type (figure 2). We surmise that in this recombination process the blue sequence represents the ancestral form of the capsular regulatory genome, because its percentage of GC content corresponds to the average 39.7% for S. pneumoniae [24] and to the 38% for the wsg gene. The red-sequence blocks have a percentage of GC content ~5 percentage points lower, and we thus conclude that they were derived by LGT from a different gene pool.

We exploited European Molecular Biology Laboratory/GenBank data accumulated from several studies of pneumococcal capsular genes and genomes and compared different sequences from the same serotypes to confirm that they were virtually identical. Streptococcal capsular regulatory sequences were retrieved by means of BLASTn and BLASTx, using published pneumococcal sequences for queries. The accession numbers for pneumococcal sequences are CR931632–CR931722, EF538714–EF538718, AF532632–AF532715, AY508586–AY508644, AF532619–AF532715, AY508586–AY508644, AY621659, AY621660, AY661448–AY661457, Z83335, AF026471, AF316639, AE00562, AY336008, AF316641, AJ239004, AF402095, X85787, AF094575, U09239, AF030373, AF057294, AJ006986, and AF106132–AF106138. The complete capsular sequences have been determined for at least 2 independent isolates of serotypes 1, 2, 4, 5, 6A, 6B, 8, 9V, 14, 18C, 19F, 19A, 19B, 23F, and 33F, and partial regulatory gene sequences (the genes wsg and wzh) have been determined for 90 serotypes (6–10 isolates per serotype). The accession numbers for other streptococcal sequences that were used as references in phylogenetic analyses were as follows: for Streptococcus oralis, AB289547; for Streptococcus mitis, AB181235; for Streptococcus gordoni, AF147914; for Streptococcus thermophilus, AF448502, CP000023, CP000419, AF454499, AF454495, AF454500, AF454498, AF454496, AY057915, AF454501, and AF454502; for Streptococcus suis, AF118389; for Streptococcus iniae, AY904444; and for Streptococcus agalactiae, AB028896, AB050723, AY375362, AF163833, AY355776, AF349539, AY3759785, AY376404, and AY375363.

Sequence alignments were performed using ClustalX [20] and were curated manually. Phylogenetic analyses were performed using MEGA (version 4) [21] and MrBayes (version 3.1) [22] software. Nucleotide diversities, percentage of guanine-cytosine (GC) content, and codon bias indices were calculated using DnaSP4 software [23].

METHODS

We exploited European Molecular Biology Laboratory/GenBank data accumulated from several studies of pneumococcal capsular genes and genomes and compared different sequences from the same serotypes to confirm that they were virtually identical. Streptococcal capsular regulatory sequences were retrieved by means of BLASTn and BLASTx, using published pneumococcal sequences for queries. The accession numbers for pneumococcal sequences are CR931632–CR931722, EF538714–EF538718, AF532632–AF532715, AY508586–AY508644, AY621659, AY621660, AY661448–AY661457, Z83335, AF026471, AF316639, AE00562, AY336008, AF316641, AJ239004, AF402095, X85787, AF094575, U09239, AF030373, AF057294, AJ006986, and AF106132–AF106138. The complete capsular sequences have been determined for at least 2 independent isolates of serotypes 1, 2, 4, 5, 6A, 6B, 8, 9V, 14, 18C, 19F, 19A, 19B, 23F, and 33F, and partial regulatory gene sequences (the genes wsg and wzh) have been determined for 90 serotypes (6–10 isolates per serotype). The accession numbers for other streptococcal sequences that were used as references in phylogenetic analyses were as follows: for Streptococcus oralis, AB289547; for Streptococcus mitis, AB181235; for Streptococcus gordoni, AF147914; for Streptococcus thermophilus, AF448502, CP000023, CP000419, AF454499, AF454495, AF454500, AF454498, AF454496, AY057915, AF454501, and AF454502; for Streptococcus suis, AF118389; for Streptococcus iniae, AY904444; and for Streptococcus agalactiae, AB028896, AB050723, AY375362, AF163833, AY355776, AF349539, AY3759785, AY376404, and AY375363.

Sequence alignments were performed using ClustalX [20] and were curated manually. Phylogenetic analyses were performed using MEGA (version 4) [21] and MrBayes (version 3.1) [22] software. Nucleotide diversities, percentage of guanine-cytosine (GC) content, and codon bias indices were calculated using DnaSP4 software [23].
Figure 1. Phylogenetic trees (minimum-evolution, neighbor-joining phylograms) of pneumococcal capsular genes and red and blue serotype characteristics. A, Four core capsular genes (wzg, wzh, wzd, and wze), including 2 examples of serotype-specific glycosyltransferase genes (wchA and wchF) and 2 other type-specific genes (mnaA and rmlACBD) (this analysis includes 4 concatenated rml genes). The red and blue coloring is also used in the mnaA and wchF genes to show that, in these cases, the clustering pattern differs from the red-blue classification, defined through capsular regulatory genes. B, Pneumococcal serotypes categorized as blue and red on the basis of their capsular regulatory genes. Serotypes shown in gray have an ambiguous classification (figure 2). C, Phylogram of the wze sequences (excluding the ambiguous serotypes 10A, 10B, and 18F) as a representation of patterns on the species level. All 8 Streptococcus species that (on the basis of sequence databases) have capsular synthesis through wzg, wzh, wzd, and wze are included here, and they represent 3 of the 6 official taxonomic groups. S. mitis, S. oralis, S. gordonii, and S. pneumoniae are from the mitis group, S. thermophilus is from the salivarius group, S. suis has equivocal taxonomic grouping, and S. agalactiae (group B streptococci) and S. iniae are from the pyogenic group. Bayesian and minimum-evolution phylogenies have identical topologies, the former with 1.0 credibility for all species-level branches and the latter with 0.8–1.0 bootstrapping confidence (10,000 iterations).

Several red integration points were identified in the wzh gene sequence (figure 2), with the predominant point near the 5’ end of the gene. Some serotypes showed red-blue mosaic structures, 2 serotypes (4 and 45) had a red integration point at wzd, and 1 serotype (18F) had a red integration point at the 3’ end of wze. In 3 serotypes (29, 39, and 43), a short red integration motif at wzh was accompanied by a nonred (and nonblue) streptococcal sequence type. In 1 serotype (20), the red sequence spanning from wzh to wze was continued by a nonred (and nonblue) sequence at wchA. In serotypes 10A and 10B, the wzd and wze genes were mosaics between blue and S. mitis– and/or S. oralis–like sequences.

Differential association of the reds and blues with carriage and IPD. The reds, even if newcomers in the pneumococcal genome, have an established standing within the species. Of the 91 serotypes, 32 had a red and 49 a blue capsular regulatory genome type (figures 1B and 2). We reviewed a large amount of epidemiological data on the distribution of serotypes (see Hausdorff et al [2] for a recent review of serotype distributions) among healthy carriers and among individuals with IPD. We
assigned each serotype to the blue or red category according to figure 1B. The data from large studies (at least 150 isolates) are shown in table 1. Despite the heterogeneity of these data sets, there was a consistent predominance of the blues among isolates from healthy carriers (two-thirds blue) and, correspondingly, a predominance of the reds among isolates from those with IPD (two-thirds red). Remarkably, these proportions held for different populations and continents and were thus not confined to the prevalence of some particular serotypes. The same applied to other studies not fulfilling the criteria for inclusion in table 1.

One of the data sets [3] included both invasive and carriage isolates from the same population and was thus suitable for calculation of an empirical odds ratio to compare the probability of invasive disease due to the red propensity. An odds ratio of 3.7 was found, with a 95% confidence interval of 2.4–5.5 \(([a]/[b])\), where \(a\) is the number of invasive red, \(b\) is the number of carried red, \(c\) is the number of invasive blue, and \(d\) is the number of carried blue.

In all, the traditional naming of certain serotypes as “primary pathogens” or “primarily invasive” [3, 4] gets one putative explanation through the capsular regulatory genes. Serotypes 1, 5, and 7F—all red—are very seldom found carried. Of the common serotypes (~10) named “carried,” only 19F is red; all others are blue. Note that of the 91 serotypes that have been characterized as different types (on the basis of antigenic reactions), only ~30 actually circulate. Characterization of the occasionally found (ie, rare) serotypes is not possible.

**DISCUSSION**

Our work is intended as an initiative to develop an understanding of the pneumococcal capsular evolution, previously regarded as elusive [18, 37]. We show that the 3 shared regulatory genes split most serotypes into 2 highly divergent clans, the red clan and the blue clan. There is no obvious pattern of preferential association between either regulatory type and any identifiable structural feature of the type-specific part of the CPS. Specifically, an examination of serogroups (groups of antigenically and structurally related serotypes) demonstrates...
Table 1. Classification of Serotypes as Red or Blue on the Basis of Their Capsular Regulatory Genes

<table>
<thead>
<tr>
<th>Isolate type, country, age group</th>
<th>Serotype regulatory genes, %</th>
<th>Serotyped isolates, no.</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carriage isolates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England, 0–5 years</td>
<td>58.0 35.8 5.1 1.1 ...</td>
<td>350</td>
<td>[3]</td>
</tr>
<tr>
<td>Sweden, 1–6 years</td>
<td>58.4 37.3 1.6 1.2</td>
<td>246</td>
<td>[4]</td>
</tr>
<tr>
<td>Netherlands, 1–19 years</td>
<td>58.1 35.9 6.0 ...</td>
<td>472</td>
<td>[25]</td>
</tr>
<tr>
<td>Canada, preschool age</td>
<td>69.6 25.4 1.0 1.5 ...</td>
<td>518</td>
<td>[26]</td>
</tr>
<tr>
<td>Kenya, &lt;5 years</td>
<td>58.9 37.1 1.0 ...</td>
<td>207</td>
<td>[27]</td>
</tr>
<tr>
<td>Gambia, &lt;18 months</td>
<td>52.2 31.1 2.8 14.0 ...</td>
<td>237</td>
<td>[28]</td>
</tr>
<tr>
<td>Invasive isolates</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>England, 0–5 years</td>
<td>31.6 67.0 1.3 0.1 ...</td>
<td>150</td>
<td>[3]</td>
</tr>
<tr>
<td>Sweden, 1–6 years</td>
<td>40.3 52.7 5.1 1.1</td>
<td>246</td>
<td>[4]</td>
</tr>
<tr>
<td>Norway, all ages</td>
<td>34.8 58.2 2.2 ...</td>
<td>325</td>
<td>[29]</td>
</tr>
<tr>
<td>Germany, adults</td>
<td>36.5 52.8 9.7 1.0 ...</td>
<td>647</td>
<td>[30]</td>
</tr>
<tr>
<td>Denmark</td>
<td>&lt;2 years</td>
<td>297</td>
<td></td>
</tr>
<tr>
<td>Norway, all ages</td>
<td>32.9 63.3 3.8 ...</td>
<td>1820</td>
<td></td>
</tr>
<tr>
<td>Germany, adults</td>
<td>&gt;60 years</td>
<td>3048</td>
<td></td>
</tr>
<tr>
<td>Canada, &lt;5 years</td>
<td>38.0 55.0 ... ...</td>
<td>7.0 303</td>
<td>[32]</td>
</tr>
<tr>
<td>Latin America, &lt;6 years</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Argentina</td>
<td>18.4 65.9 0.7 ...</td>
<td>15.0 1006</td>
<td></td>
</tr>
<tr>
<td>Brazil</td>
<td>28.4 57.2 2.0 ...</td>
<td>12.4 1203</td>
<td></td>
</tr>
<tr>
<td>Colombia</td>
<td>29.0 58.8 1.0 ...</td>
<td>11.2 623</td>
<td></td>
</tr>
<tr>
<td>Uruguay</td>
<td>15.1 73.1 4.5 ...</td>
<td>7.3 352</td>
<td></td>
</tr>
<tr>
<td>Taiwan, all ages</td>
<td>33.0 57.0 ... ...</td>
<td>10.0 483</td>
<td>[34]</td>
</tr>
<tr>
<td>Australia, &lt;15 years</td>
<td>28.9 64.5 1.5 ...</td>
<td>5.0 698</td>
<td>[35]</td>
</tr>
<tr>
<td>United States, &lt;5 years</td>
<td>39.8 58.9 0.7 0.6 ...</td>
<td>855</td>
<td>[36]</td>
</tr>
</tbody>
</table>

**NOTE.** Serotype data were extracted from tables and/or figures in published epidemiological studies. Data on carriage isolates are from nasopharyngeal swabs obtained from healthy individuals. Data on invasive isolates are from normally sterile body sites, blood, cerebrospinal fluid, and joint, pleural, and peritoneal fluid. Studies including at least 150 serotyped isolates and in which at least 85% of the serotypes were reported in detail are included here. The category “other” includes pooled rare serotypes and isolates that were not typed or were not typeable. Serotype 3 (synthesis through a different pathway) and serotypes 25F, 25A, and 38 (not blue or red) are listed separately.

that both regulatory types can be found in association with closely related capsules (e.g., serotypes 9A and 9V in serogroup 9 have blue regulatory genes, whereas types 9L and 9N have red regulatory genes) (figure 1B).

Throughout the regulatory gene region and at wchA, the blue clan corresponds to the typical pneumococcal genomic content in its percentage of GC. Its sequence diversity is higher than that in the red clan, suggesting that the blues form an older clan. The reds have a percentage of GC content ~5 percentage points lower than that of the blues, consistent with an LGT origin. On the basis of existing sequence information from other species, which have homologues of the pneumococcal capsular genes, the source species of the reds remains unidentified. Phylogenetic inference suggests that the source is within the mitis group of streptococci (figure 1C); this group comprises 15 species [38], of which capsular gene sequence information exists only for *S. pneumoniae*, *S. mitis*, *S. oralis*, and *S. gordonii*. It is well known that *S. mitis* and *S. oralis* are LGT partners with *S. pneumoniae* at antibiotic resistance genes (especially *pbp2x*) [13, 39]. They are probably not the source species of the reds (figure 1C), because their capsular regulatory gene percentage of GC content is similar to that of the blues and of *S. pneumoniae*. However, *S. mitis* is known to be extremely variable genetically [13, 40, 41]. Capsular sequence information is available only from a few *S. mitis* strains, and thus the possibility that *S. mitis* is the transmitter of the reds cannot be ruled out.

An inspection of the current phylogenetic pattern suggests that the diversification of the reds, as well as that of the blues, are (or have been) concurrent processes. Both clans include major clusters of approximately equidistant sequences (figure 1C). The patchiness of red-blue recombination events (figure 2) suggests that new combinations can be expected to emerge. The limited current sequence data already shows the presence
of 2 forms in 1 serotype (19F) (figure 2). That many serotypes do have completely or nearly identical blue or red regulatory gene sequences (figure 1C) speaks for very recent recombination events in their history.

Our epidemiologic survey showed a tendency of the blue type of the capsular regulatory genome to be associated with carriage and the red with invasive disease isolates. This conclusion rests on the assumption of a tight association of the assigned (blue or red) regulatory and the type-specific capsular genomes. This assumption is warranted on the basis of the current sequence information for the capsular genomes (even if limited), because the sequences from different isolates of a given serotype have proved to be almost identical (with the exception of serotype 19F; see above). The existence of highly divergent sequences among capsular regulatory genes has been reported from a small set of serotypes [42, 43] but were not considered further (and were not even mentioned in publications citing Morona et al [42] or Jiang et al [43]). A clue about the red-blue grouping has thus existed but has been ignored.

Pneumococci employing the red type of capsular regulation seem better fit to cause IPD. In traditional parlance, they are more virulent for this outcome than the blues. A thick layer of capsule is associated with protection of the bacteria from phagocytosis, the critical host defense in IPD [6]. Accordingly, if the red type of capsular regulation would lead to an increase in CPS production compared with the blue type, this could provide the reds with a pivotal advantage in causing IPD and described regulation of the amount of CPS synthesis is a posttranslational tyrosine phosphoregulatory system, the most essential component of which is the regulatory gene wze [18, 19]—that is, the gene that unequivocally defines the blues and the reds. Whether the red type of regulation would mean increased expression of the red genes remains to be shown. The reds might be more efficiently translated because their codon bias is higher (on average, 0.5 in the reds and 0.4 in the blues at wze). Interestingly, in one study [44] expression of wze in serotypes causing invasive infection (most of them red according to our classification) was reported to be 2-fold higher than serotypes causing nasopharyngeal colonization (most of them blue). In studies comparing gene expression profiles of nasopharyngeal isolates with those of invasive isolates, significant differences in capsular gene expression have not, however, been found (eg, in Oggoni et al [45]).

At the 5′ end of wze, an A motif is present in all the reds but is not present in any of the blues (figure 3). This motif has been identified as one of a set of putative virulence factors in the full genome of serotype 4 [24]. Simple sequence repeats, which are prone to strand slippages, are known to lead to phase variation in bacteria (gene transcription on or off), including the pneumococcus [46, 47]. Although the described regulatory mechanism of capsular genes is posttranslational, the presence of the A motif in reds suggests that it could also induce phase variation here and thus provide the reds with an additional capability by acting as a transcription regulatory mechanism.

Only a small minority of the total population of pneumococci causes IPD. The majority is represented by the bacteria colonizing the upper respiratory mucosa of healthy carriers. These isolates are crucial for the survival and spread of the pneumococcus. Because of the complex functions of the capsule in this process, our present knowledge is not sufficient to suggest which type of capsular regulation provides an advantage in colonization, although the epidemiological data suggest that it is the blue type of regulation.

The pneumococcus is currently exposed to a profound change in environment, which will provide pressures for new responses. The powerful new conjugate vaccines selectively disadvantage the serotypes represented in the vaccine. Their effect is seen in the pneumococcal population in the nasopharynx of carriers as a decrease in the frequency of vaccine serotypes and a concomitant increase in nonvaccine serotypes [48]. Among the serotypes replacing the vaccine types, the so-called vaccine escape recombinants are a concrete example of evolution through recombination having replaced the original capsular genome coding for the vaccine serotype 4 with the capsular genome coding for the unrelated nonvaccine serotype 19A [49]. Sequence mapping of the capsular regulatory genome in further vaccine escape recombinants is an urgent need, to give us a better understanding of pneumococcal evolution under the current vaccination pressure. To date, the recombination break points have been identified upstream and downstream of the capsular genome; however, as yet no sequence information on the capsular genomes exists.

Acknowledgments

We thank Bob O’Hara for checking the language and Joose Vainola for help with artwork.
References


36. Jiang SM, Wang L, Reeves PM. Molecular characterization of Strept-


