Polymorphic Competence Peptides Do Not Restrict Recombination in \textit{Streptococcus pneumoniae}

Omar E. Cornejo,*1 Lesley McGee,2 and Daniel E. Rozen1,3

1Department of Biology, Emory University, Program in Population Biology, Ecology, and Evolution
2Huber Department of Global Health, Rollins School of Public Health, Emory University
3Faculty of Life Sciences, University of Manchester, Manchester, UK

*Corresponding author: E-mail: ocornejo@gmail.com.
Associate editor: John H. McDonald

Abstract

Understanding the factors that limit recombination in bacteria is critical in order to better understand and assess its effects on genetic variation and bacterial population genetic structure. Transformation in the naturally competent bacterium, \textit{Streptococcus pneumoniae}, is regulated by a polymorphic competence (com) apparatus. It has been suggested that polymorphic types, called pherotypes, generate and maintain subpopulation genetic structure within this species. We test predictions stemming from this hypothesis using a cosmopolitan sample of clinical pneumococcal isolates. We sequenced the locus encoding the peptide that induces competence (comC) to assign clones to each known pherotype class and then used multilocus sequence typing to determine whether there is significant genetic differentiation between pherotypes subgroups. We find two dominant pherotypes within our sample, and both are maintained at high frequencies (CSP1 74%, CSP2 26%). Our analyses fail to detect significant genetic differentiation between pherotype groups and find strong evidence, from a coalescent analysis, for interpherotype recombination. In addition, our analyses indicate that positive selection may account for the maintenance of the fixed polymorphism in this locus (comC). Altogether, these results fail to support the prediction that the polymorphism in the competence system acts to limit recombination within \textit{S. pneumoniae} populations. We discuss why this result is expected given the mechanism underlying transformation and outline a scenario to explain the evolution of polymorphism in the competence system.

Key words: \textit{Streptococcus pneumoniae}, ComC, genetic diversity, population genetic structure, recombination, positive selection.

Introduction

The degree to which bacterial species are influenced by recombination will depend upon its rate of occurrence and the processes underlying it (Selander and Levin 1980; Maynard-Smith et al. 1993; Feil et al. 1999, 2000; Spratt and Maiden 1999; Spratt 2004; Fraser et al. 2007). For example, although bacteria that are naturally transformable can promiscuously acquire free DNA present in the environment, often from phylogenetically disparate sources (Majewski et al. 2000; Cohan 2001, 2002), recombination mediated by conjugative plasmids or transducing phage is thought to be more restrictive. Consistent with this premise, naturally transformable species that undergo high rates of transformation-mediated recombination, such as \textit{Helicobacter pylori} and \textit{Streptococcus pneumoniae}, have populations that are loosely genetically structured (Feil et al. 2000; Falush et al. 2001), whereas species such as \textit{Escherichia coli}, \textit{Salmonella} ssp., or \textit{Rhizobium meliloti} that recombine primarily via conjugation and transduction are more clonally structured (Maynard-Smith et al. 1993). Despite the promiscuity of naturally transformable species, there are two well-known limits to interspecific recombination mediated by transformation. First, some species, like \textit{Haemophilus influenzae} and \textit{Neisseria gonorrhoeae}, require short sequence tags to identify homotypic fragments of DNA, and only tagged DNA is recombined into the bacterial chromosome (Sisco and Smith 1979; Graves et al. 1982). Second, high levels of sequence divergence across species, in otherwise homologous fragments of DNA, significantly reduce the frequency of interspecific recombination (Zawadzki et al. 1995; Fraser et al. 2007). This latter limitation has been shown to exert a strong effect on population structure and the differentiation and maintenance of species within the genus \textit{Streptococcus} (Fraser et al. 2007). However, the effects of these processes on limiting recombination within species, where sequence divergence is low and sequence tags are shared are expected to be small. This has fostered the general belief that, at least within species, the limits to homologous recombination mediated by transformation are minimal.

In \textit{S. pneumoniae}, the development of natural competence is regulated by the action of the two-component signaling system encoded by comC and comD, specifying a small peptide signal and its cognate receptor, respectively (Havarstein et al. 1996; Pestova et al. 1996; Cheng et al. 1997; Campbell et al. 1998). The competence stimulating peptide (CSP) is secreted into the extracellular environment, where it binds to membrane-bound comD, initiating a chain of events that culminates in the uptake and incorporation of free DNA (Tomasz 1965; Havarstein et al. 1995, 1996). An intriguing and unexplained aspect of
Materials and Methods

Clinical Isolates of S. pneumoniae
A geographically and serotypically diverse collection of clinical pneumococcal isolates from 2000 and 2001 was employed in this study (Yu et al. 2003).

Genetic Characterization
Eighty-eight clones isolated in 2000 and 2001 were characterized by multilocus sequence typing (MLST). For MLST, internal fragments of the araE, gdh, gki, recP, spi, ddl, and xpt genes were amplified by polymerase chain reaction (PCR) from chromosomal DNA, and the fragments were directly sequenced in both directions using the primers that were used for the initial amplification. The sequences (alleles) at each locus were compared with those on the publicly accessed MLST web site (www.mlst.net) and were assigned allele numbers if they corresponded to sequences already submitted to the MLST database; novel sequences were submitted for new allele numbers and deposited in the database. The allele numbers at the seven loci were compared with those at the MLST website, and sequence types (STs) were assigned. Allelic profiles that were not represented in the MLST database were submitted for assignment of new ST numbers and deposited in the database. The sequences of the loci for all isolates are available at the MLST repository www.mlst.net and can be accessed by searching by the ST types for each isolate provided in Supplementary Material online. Sequences of the ddl locus were excluded from our analyses because it has been shown to be under selection (Enright and Spratt 1999) due to linkage with penicillin-binding proteins.

The sequences of comC for the 88 isolates were obtained with the primers: FOR: 5′-CAATAACCGTCCCAAATCCA-3′, and REV: 5′-AAAAGTACA CTTTGGGAGA AAAA-3′, producing a fragment of approximately 400 bp. The conditions for amplification were 1× PCR buffer, 1.5 mm MgCl2, 0.2 mm dinucleotides mix, 2U Taq Polymerase, and 20 pmol of each primer, per 50-μl reaction. The PCR cycling parameters were as follows: an initial denaturation step at 95 °C for 2 min, 25 cycles of amplification performed as follows: denaturation at 94 °C for 30 s, annealing temperature at 56 °C for 30 s and extension temperature at 72 °C for 1.0 min, and finally completed with an extension at 72 °C for 5 min. The isolates were assigned to a given pherotype, by comparing their translated amino acid sequence with the types reported in Kilian et al. (2008). We also obtained the comC sequences for the 26 original clones of the PMEN collection (http://www.sph.emory.edu/PMEN/), in order to have an independent assessment of the frequency of the CSP types in S. pneumoniae. The DNA sequences for the comC locus obtained in this work were submitted to the GenBank as a popset under accession numbers: GQ892099–GQ892186.

Population Genetic Analyses
Standard population genetic analyses were performed for all isolates using DNAsp v.4.20 (Rozas et al. 2003). Haplotype diversity, nucleotide diversity (π), and its standard deviation (SD) were estimated for the sample as a whole and for the sample stratified by CSP type (overall estimates by
geographic region of origin are provided in Supplementary Material online).

**Analysis of the Population Structure Mediated by Pherotypes**

In order to perform the population genetic structure analyses, CSP pherotype groups/subpopulations were assigned according to the criterion described above.

It is known that inferences derived from $F_{st}$ statistics may be limited due to the reliance of this statistic on the often unmet assumptions of uniform effective population sizes and symmetric migration rates. These limitations can be problematic when populations have large effective sizes, as in bacteria, and are weakly structured (Bossart and Prowell 1998), or for populations in which the subpopulations differ significantly in size. To overcome these restrictions (Gonzalez et al. 2008) and to generate more reliable estimates of genetic differentiation, the assessment of genetic structure in the sample was performed under a coalescent framework that allows independent estimation of $\Theta$ (the mutation parameter, proportional to the effective population size) and a migration parameter (Hudson 1991; Nath and Griffiths 1993; Beerli and Felsenstein 1999 2001); the latter is informative of the levels of recombination (gene exchange) occurring between the two pherotypes. Because the subpopulations are defined as carriers of different competence peptidase/receptor alleles (CSP types), the number of migration events estimated this way is an indicator of the levels of recombination (gene exchange) occurring between the two subpopulations. Maximum likelihood estimates of $\Theta$ (the mutation rate per site) and $m$ (the migration rate per site per generation) for each population, and immigration rates ($M_{1-2}$ and $M_{2-1}$, with $M = m/\mu$ and $m$ is the rate of migration per generation) were obtained under a Markov Chain Monte Carlo model with importance sampling, employing 10 short chains (100,000 used trees of 1,000,000 sampled) and 4 long chains (500,000 used trees of 3,000,000 sampled), with adaptive heating of the chains as implemented in Migrate v3.0 (Beerli and Felsenstein 1999, 2001). Initial values for $\Theta$ and $M$ were obtained from $F_{st}$ estimations. In order to assess the significance of the levels of migration inferred, we fitted a null model in which migration is constrained to contribute less than mutation to the differentiation between subpopulations ($M = m/\mu < 1$). The estimates under the null model were obtained employing 10 short chains (100,000 used trees of 1,000,000 sampled) and 4 long chains (500,000 used trees of 3,000,000 sampled). The comparison of the two models was done by a likelihood ratio test with two degrees of freedom (df).

The relationships among the isolates were represented graphically by means of a haplotype network. The network was constructed with the concatenated MLST genetic sequences (excluding $ddl$), employing a median joining algorithm of the pairwise distances among haplotypes, as implemented in the program Network v4.5.0.0 (Bandelt et al. 1999).

**Analysis of Polymorphism and Assessment of Selection in the comC Locus**

We estimated the synonymous and nonsynonymous polymorphism of the comC locus and MLST loci within and between pherotypes. To facilitate the discussion of the results, we will refer to the difference between CSP subpopulations as “divergence” between subpopulations.

We assessed the neutrality of comC broadly and for each pherotype independently by performing the modified version of the Hudson–Kreitman–Aguade (HKA) test proposed by Innan (Hudson et al. 1987; Innan 2006). This test assesses if the ratio of polymorphism to divergence observed in comC is significantly different than would be expected when compared with several reference loci where variation is expected to accumulate neutrally, here provided by the MLST loci.

We performed this test in two ways: 1) assessing neutrality of the comC locus in populations bearing pherotype 1 (CSP1) and using pherotype 2 populations (CSP2) as an outgroup and 2) assessing neutrality of the comC locus in populations bearing pherotype 2, while using pherotype 1 populations as the outgroup. In all cases, to obtain the probability distribution for the time of “speciation” we set 10,000 as the number of acceptances in the rejection-sampling algorithm, following the suggestion by Innan (2006). In order to obtain the probability value for the modified HKA test, 10,000 replicates were performed. The statistic obtained is the probability that $r$, defined as the ratio of polymorphism to divergence, for the locus presumably under selection (comC in this case) is significantly different to the ratio of polymorphism to divergence in a set of neutrally evolving loci. The null hypothesis is that the ratio $r$ in the target locus falls within the distribution of $r$ values generated by the reference neutral loci.

Also, a McDonald–Kreitman (MK) test was performed on comC sequences to detect departures from neutrality. MK tests on comC and assessment of the number of fixed polymorphic substitutions in simulated data were performed in DNAsp v4.20.

Because of multiple testing, corrections were performed according to Benjamini and Yekutieli (2001) to maintain an overall significance of 0.05.

**Results**

**CSP Polymorphism**

MLST profiles were obtained from 88 geographically diverse clinical isolates. Initial analyses of genetic diversity for each gene and the average over genes are shown in table 1. In general, haplotype diversity is relatively high, as is nucleotide diversity, which ranges from 0.9% to 1% per site. Because genetic structure among these clones arising from geographic subdivision could compromise our efforts to detect CSP-mediated structure, we first determined if clones could be distinguished on the basis of sampling
site. No strong signature of population genetic structure consistent with geographic location was detected. Although there is variability in the nucleotide diversity among loci, there is consistency between groups of isolates bearing different types, with most of them showing slightly larger pairwise diversity for the gki locus in both subsets, and the CSP1 subset showing slightly larger genetic diversity.

As shown in table 2, of the 88 isolates, 65 (74%) encode comC amino acid sequences identical to pherotype type 1 (CSP1), whereas 23 (26%) encode comC amino acid sequences identical to pherotype type 2 (CSP2). Our estimates for the frequency of the two pherotypes in an independent clinical data set (the PMEN collection) are not different from those found in a large collection of non-clinical, carriage isolates (Bogaert et al. 2001), where 73% of the isolates carry CSP1 and 27% CSP2. In addition, these frequencies are consistent with geographic location was detected. Although there is variability in the nucleotide diversity among loci, there is consistency between groups of isolates bearing different types, with most of them showing slightly larger pairwise diversity for the gki locus in both subsets, and the CSP1 subset showing slightly larger genetic diversity.

As shown in table 2, of the 88 isolates, 65 (74%) encode comC amino acid sequences identical to pherotype type 1 (CSP1), whereas 23 (26%) encode comC amino acid sequences identical to pherotype type 2 (CSP2). Our estimates for the frequency of the two pherotypes in an independent clinical data set (the PMEN collection) are not different from those found in a large collection of non-clinical, carriage isolates (Bogaert et al. 2001), where 73% of the isolates carry CSP1 and 27% CSP2. In addition, these frequencies are consistent with geographic location was detected. Although there is variability in the nucleotide diversity among loci, there is consistency between groups of isolates bearing different types, with most of them showing slightly larger pairwise diversity for the gki locus in both subsets, and the CSP1 subset showing slightly larger genetic diversity.

As shown in table 2, of the 88 isolates, 65 (74%) encode comC amino acid sequences identical to pherotype type 1 (CSP1), whereas 23 (26%) encode comC amino acid sequences identical to pherotype type 2 (CSP2). Our estimates for the frequency of the two pherotypes in an independent clinical data set (the PMEN collection) are not different from those found in a large collection of non-clinical, carriage isolates (Bogaert et al. 2001), where 73% of the isolates carry CSP1 and 27% CSP2. In addition, these frequencies are consistent with geographic location was detected. Although there is variability in the nucleotide diversity among loci, there is consistency between groups of isolates bearing different types, with most of them showing slightly larger pairwise diversity for the gki locus in both subsets, and the CSP1 subset showing slightly larger genetic diversity.

As shown in table 2, of the 88 isolates, 65 (74%) encode comC amino acid sequences identical to pherotype type 1 (CSP1), whereas 23 (26%) encode comC amino acid sequences identical to pherotype type 2 (CSP2). Our estimates for the frequency of the two pherotypes in an independent clinical data set (the PMEN collection) are not different from those found in a large collection of non-clinical, carriage isolates (Bogaert et al. 2001), where 73% of the isolates carry CSP1 and 27% CSP2. In addition, these frequencies are consistent with geographic location was detected. Although there is variability in the nucleotide diversity among loci, there is consistency between groups of isolates bearing different types, with most of them showing slightly larger pairwise diversity for the gki locus in both subsets, and the CSP1 subset showing slightly larger genetic diversity.

As shown in table 2, of the 88 isolates, 65 (74%) encode comC amino acid sequences identical to pherotype type 1 (CSP1), whereas 23 (26%) encode comC amino acid sequences identical to pherotype type 2 (CSP2). Our estimates for the frequency of the two pherotypes in an independent clinical data set (the PMEN collection) are not different from those found in a large collection of non-clinical, carriage isolates (Bogaert et al. 2001), where 73% of the isolates carry CSP1 and 27% CSP2. In addition, these frequencies are consistent with geographic location was detected. Although there is variability in the nucleotide diversity among loci, there is consistency between groups of isolates bearing different types, with most of them showing slightly larger pairwise diversity for the gki locus in both subsets, and the CSP1 subset showing slightly larger genetic diversity.

As shown in table 2, of the 88 isolates, 65 (74%) encode comC amino acid sequences identical to pherotype type 1 (CSP1), whereas 23 (26%) encode comC amino acid sequences identical to pherotype type 2 (CSP2). Our estimates for the frequency of the two pherotypes in an independent clinical data set (the PMEN collection) are not different from those found in a large collection of non-clinical, carriage isolates (Bogaert et al. 2001), where 73% of the isolates carry CSP1 and 27% CSP2. In addition, these frequencies are consistent with geographic location was detected. Although there is variability in the nucleotide diversity among loci, there is consistency between groups of isolates bearing different types, with most of them showing slightly larger pairwise diversity for the gki locus in both subsets, and the CSP1 subset showing slightly larger genetic diversity.
haplotypes with CSP type. The size of each pie is relative to the frequency of that haplotype. One haplotype shows both pherotypes, and there is no clear clustering of the haplotypes with CSP type.

in which frequent migration, or gene exchange, is maintained between subpopulations with different pherotypes (table 3). The analysis of gene migration summarized over all six MLST loci suggests that gene migration contributes 300 and 1,600 times more than mutation to the variation in the CSP1 subpopulation and CSP2 subpopulation, respectively (see table 3). Based on the relative contribution of migration and mutation, and the genetic diversity estimated per subpopulations, it is expected that between 1.7 and 5.7 migrations or gene shuffle events (i.e., lateral transfers, Njm) per site per generation occur between populations with different pherotypes.

We previously identified significant differences in the frequency of the subpopulations carrying the type 1 and type 2 pherotypes and proposed that these frequencies are proportional to the effective subpopulation size of each subpopulation. If we reasonably assume that there are no differences in the mutation rate (μ) between subpopulations, this implies that differences in the estimates of the mutation parameter Θ (2Nmμ) between subpopulations will be due to differences in the effective population size (Ne). Accordingly, because the maximum likelihood estimate of the genetic diversity of the CSP2 subpopulation (Θ2) is 0.005 (table 3), it is expected that the estimate of genetic diversity of the CSP1 subpopulation (Θ1) was 2.8 times that of the CSP2 subpopulation or 0.014. As shown in table 3, the estimate of Θ1 is equal to this independent estimation (0.014) supporting the idea that 1) the frequency of the pherotypes is proportional to the effective population size of the respective subpopulations and 2) that CSP1 and CSP2 subpopulations have different effective population sizes (the confidence intervals [CIs] of the maximum likelihood estimates of Θ do not overlap) as we assumed.

**ComC Diversity and Evolution**

Table 4 shows the genetic diversity of the comC locus in *S. pneumoniae*. When compared with MLST loci, the comC locus shows significantly higher genetic diversity, corresponding to a 4-fold difference (t = 88.047, P < 0.00001), as well as significantly higher variance (Hartley test for equality of variance F = 68.0625, P < 0.00001). One simple explanation of this pattern is the possibility of a higher mutation rate at the comC locus. If this is so, it would be expected that polymorphism within comC types was roughly equal to the difference (divergence) between types, as is found for the neutrally evolving MLSTs (see table 5). Clearly, this expectation is not realized. Instead, the ratio of polymorphism to divergence in the MLST loci (close to 1) is at least 10 times larger than that obtained for the comC locus when the two CSP subpopulations are compared (table 5). The results of 2D-HKA tests on CSP1 and CSP2 populations show that there is significantly less polymorphism in the comC locus within pherotype subpopulations, given the amount of divergence between pherotypes, when compared with what would be expected if they were evolving neutrally (CSP-1: r = 0.089, P value < 0.02; CSP-2: r = 0.0078, P value < 0.001). This result is strongly indicative of nonneutral change in comC.

Considering the polymorphism in comC in more detail, we find that all amino acid differences at this locus are fixed between pherotypes, as opposed to the pattern found at MLST loci in which all the changes are polymorphic with no fixed differences (table 6). The MK test performed over the comC locus shows that this pattern of substitution deviates from what would be expected under neutrality (Fisher's exact test P value = 0.006993). Because we have already shown that there is significant gene exchange between CSP subpopulations, the results of this test are unlikely to be affected by population structure. Also compelling

**Table 3. Fitting to Migration Models: Full Model (Free Parameter) with Migration Estimated and Restricted Migration Model with Migration Restricted (Fixed to m/μ = 0.5).**

<table>
<thead>
<tr>
<th>Model</th>
<th>Log L</th>
<th>Θ1 (95% CI)</th>
<th>Θ2 (95% CI)</th>
<th>M1 → 1* (95% CI)</th>
<th>M2 → 2* (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full</td>
<td>−201.05***</td>
<td>0.014 (0.013–0.016)</td>
<td>0.005 (0.004–0.006)</td>
<td>328.4 (276–388)</td>
<td>1,600 (1,310–1,950)</td>
</tr>
<tr>
<td>Restricted</td>
<td>−212.13</td>
<td>0.014 (0.012–0.016)</td>
<td>0.0083 (0.006–0.011)</td>
<td>n.a</td>
<td>n.a</td>
</tr>
</tbody>
</table>

*Significant P value ≤ 0.001.

a Maximum likelihood value for the fitting.

b Θi are the mutation parameter estimates for population i (CSP1 = 1 and CSP2 = 2).

c The migration parameter from population i to population j (Mi → j = m/μ) and 95% CI and the log likelihood (Log L) are shown. n.a. no estimates are provided because in the null model migration is fixed.
the observation of the pattern of fixed versus polymorphic differences in a simulated neutrally evolving comC locus (assuming high genetic diversity, \( \pi = 0.041 \)), which resembles that of the MLST genes: That is, no fixed differences are observed between CSP subpopulations (see table 6).

We believe that, all together, these analyses provide strong evidence that comC is evolving nonneutrally in S. pneumoniae and specifically that it has been subject to positive selection.

**Discussion**

Bacterial population genetic structure is influenced to varying degrees by recombination (Maynard-Smith et al. 1993; Fraser et al. 2007). In the case of streptococci, interspecific recombination is known to be limited by high levels of genetic divergence across species (Fraser et al. 2007), a limit that is not believed to play an important role within species. Here, we test the hypothesis that there may be genetic mechanisms, other than sequence divergence, that limit in- 

Discussion

is the observation of the pattern of fixed versus polymorphic differences in a simulated neutrally evolving comC locus (assuming high genetic diversity, \( \pi = 0.041 \)), which resembles that of the MLST genes: That is, no fixed differences are observed between CSP subpopulations (see table 6).

We believe that, all together, these analyses provide strong evidence that comC is evolving nonneutrally in S. pneumoniae and specifically that it has been subject to positive selection.

**Discussion**

Bacterial population genetic structure is influenced to varying degrees by recombination (Maynard-Smith et al. 1993; Fraser et al. 2007). In the case of streptococci, interspecific recombination is known to be limited by high levels of genetic divergence across species (Fraser et al. 2007), a limit that is not believed to play an important role within species. Here, we test the hypothesis that there may be genetic mechanisms, other than sequence divergence, that limit intraspecific recombination and as a consequence influence the genetic substructure of S. pneumoniae populations.

The observation that strains of S. pneumoniae can only induce competence among cells that present the same type of competence peptide (CSP pherotype) led to the suggestion that this could be a factor maintaining genetic differentiation among subpopulations of cells (Havarstein et al. 1997; Tortosa and Dubnau 1999; Steinmoen et al. 2002; Claverys et al. 2007). Our results suggest that this is unlikely to be the case. If such a scenario were correct, higher levels of genetic differentiation should have been observed between populations expressing different pherotypes than within populations expressing the same pherotype. In contrast, we have shown that no significant genetic differentiation is observed between pneumococcal subpopulations bearing distinct pherotypes. A model with restricted migration between pherotypes does not fit the data as well as a model in which there is considerable exchange of genes between pherotype subpopulations (table 3). This result is consistent with a recently proposed alternative and opposite prediction (Johnsborg et al. 2008) that there should be no such pherotype specific differentiation owing to the fact that the induction of competence coincides with the targeted lysis of cells bearing nonmatching comD sequences and thus the opposite pherotype.

We believe that our results are consistent with the understood ecology of transformation in this species. Consider a scenario in which two populations of cells with different pherotypes coexist within the human nasopharynx, a situation likely to be common for pneumococci given their high rates of co-colonization and clonal turnover (Bogaert et al. 2004). Cell type 1 produces CSP1 and cell type 2 produces CSP2, recognized by their respective receptor types. Now imagine that a certain proportion of each cell type initiates the competence cascade and secretes its strain specific CSP. These peptides are recognized by receptors in cells belonging to the same pherotype group and this causes: 1) induction of competence in like-pherotype cells and 2) production of bacteriocins or toxins that lead to the lysis of cells of both pherotypes that have not entered the competent state (Steinmoen et al. 2002; Kreth et al. 2005; Guiral et al. 2006). According to this scenario, cells belonging to both pherotypes will simultaneously become competent, whereas the noncompetent cells of both pherotypes will lyse and release DNA. Because competent cells are not discriminating in their uptake of DNA, it is easy to envision that free available DNA is taken up by competent cells, without regard to its pherotype. This would cause nonspecific recombination among pherotypes, thus preventing pherotype specific genetic differentiation. Recently Claverys et al. (2006) have suggested that transformation driven by this process would prevent genetic homogenization within pherotypes and increase population wide genetic diversity. Although this may be a consequence of the scenario we outline, it fails to explain the existence of the polymorphic competence system itself.

**Table 4. Genetic Diversity for the comC Locus.**

<table>
<thead>
<tr>
<th>Locus</th>
<th>( n ) (Length, bp)</th>
<th>( S )</th>
<th>( \pi ) (SD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>88 (123)</td>
<td>14</td>
<td>0.041 (0.0033)</td>
</tr>
<tr>
<td>CSP 1</td>
<td>65 (123)</td>
<td>5</td>
<td>0.0077 (0.0006)</td>
</tr>
<tr>
<td>CSP 2</td>
<td>23 (123)</td>
<td>1</td>
<td>0.0007 (0.0006)</td>
</tr>
</tbody>
</table>

\( n \) corresponds to the number of isolates, length is the nucleotide sequence length of the gene in base pairs, \( S \) is the number of segregating or polymorphic sites, and \( \pi \) is the nucleotide diversity estimated as the average of the Jukes and Cantor pairwise distances. In all cases, SD corresponds to the standard deviations.

**Table 5. Genetic Polymorphism (Average Pairwise Differences, Poly) within Streptococcus pneumoniae CSP1 and CSP2 Subpopulations and Divergence (Average Pairwise Differences) between CSP Subpopulations.**

<table>
<thead>
<tr>
<th>Locus</th>
<th>Sequence Length</th>
<th>Poly within CSP1</th>
<th>Poly within CSP2</th>
<th>Divergence</th>
</tr>
</thead>
<tbody>
<tr>
<td>comC</td>
<td>123</td>
<td>0.94</td>
<td>0.08</td>
<td>10.55</td>
</tr>
<tr>
<td>aroE</td>
<td>405</td>
<td>1.90</td>
<td>1.70</td>
<td>1.96</td>
</tr>
<tr>
<td>gdh</td>
<td>460</td>
<td>5.32</td>
<td>3.58</td>
<td>4.92</td>
</tr>
<tr>
<td>gki</td>
<td>483</td>
<td>7.96</td>
<td>5.23</td>
<td>7.08</td>
</tr>
<tr>
<td>recP</td>
<td>450</td>
<td>2.42</td>
<td>2.34</td>
<td>2.44</td>
</tr>
<tr>
<td>spi</td>
<td>474</td>
<td>4.01</td>
<td>4.64</td>
<td>4.41</td>
</tr>
<tr>
<td>xpt</td>
<td>486</td>
<td>4.48</td>
<td>3.79</td>
<td>4.39</td>
</tr>
</tbody>
</table>

Sequence length is in base pairs.

**Table 6. Synonymous (Syn) and Nonsynonymous (NonSyn) Fixed and Polymorphic Differences between CSP1 and CSP2 Subpopulations (within Streptococcus pneumoniae) in the comC Locus and the MLST Loci.**

<table>
<thead>
<tr>
<th>Locus</th>
<th>Fixed Syn/Nonsyn</th>
<th>Polymorphic Syn/Nonsyn</th>
</tr>
</thead>
<tbody>
<tr>
<td>comC</td>
<td>2/8</td>
<td>6/0</td>
</tr>
<tr>
<td>Simul comC</td>
<td>0/0</td>
<td>12.8/4.1</td>
</tr>
<tr>
<td>aroE</td>
<td>0/0</td>
<td>4/7</td>
</tr>
<tr>
<td>gdh</td>
<td>0/0</td>
<td>28/7</td>
</tr>
<tr>
<td>gki</td>
<td>0/0</td>
<td>27/5</td>
</tr>
<tr>
<td>recP</td>
<td>0/0</td>
<td>10/3</td>
</tr>
<tr>
<td>spi</td>
<td>0/0</td>
<td>39/3</td>
</tr>
<tr>
<td>xpt</td>
<td>0/0</td>
<td>22/9</td>
</tr>
</tbody>
</table>

Simul comC refers to average values of a simulated data set with 500 replicates.
It could be argued that interpherotype recombination could be prevented if there is sufficient divergence among clones with distinct pherotypes. However, our results indicate that this precondition is not met (table 1) and moreover that pherotypes are insufficient to cause this differentiation to begin with (see table 3 and fig. 1).

It is important to mention caveats associated with the fact that our analyses were performed on a geographically diverse collection of clinical isolates. First, it is possible that local differentiation among pherotypes is present but that this is obscured at a regional scale because of increasing genetic variation in each group due to geographic subdivision. A second caveat derives from the fact that our clones are all clinical isolates rather than clones isolated from carriage (nondisease causing), the predominant pneumococcal lifestyle. If clones causing disease were not a representative sample of pherotypes, this would limit our ability to detect localized genetic structuring. However, our results (Rozen DE, unpublished data) indicating that pherotype frequencies in carriage isolates are indistinguishable from those found here suggest that this concern is unwarranted. Despite this, it remains possible that the degree of differentiation between pherotypes is distinct in clinical and carriage isolates. In future work, we intend to address both caveats using a more geographically and temporally localized sample of exclusively carriage isolates. Toward that end, it is noteworthy that a recent study has reported significant differentiation between pneumococcal clinical subpopulations carrying different pherotypes (Carriolo et al. 2009). A possible explanation for the differences in our results is the underlying assumptions of the population genetic analyses used. The statistics employed by (Carriolo et al. 2009) rely upon $F_{st}$ and similar statistics, and its associated assumption of equal subpopulation sizes, an assumption that our analyses reveal is clearly violated. Although this may be one cause of the different results, we are uncertain if this represents the only cause for the discrepancies in the studies. Clearly, further work will be necessary to reconcile the apparent differences.

Overall, our results indicate that polymorphism in the competence peptide does not maintain genetically differentiated subpopulations of pneumococci. It remains intriguing, however, that the polymorphism in the competence system exists and that the two dominant pherotypes are maintained at such high frequencies. It has been recently suggested that comC and comD sequences across streptococci display substitution patterns indicative of positive selection (Ichihara et al. 2006). However, this previous analysis did not explicitly consider the within species polymorphism in comC that is examined here. Within S. pneumoniae, it is possible that some form of balancing selection or frequency dependent selection maintains the comC polymorphism. A region under balancing selection is expected to exhibit higher genetic diversity than loci evolving neutrally (Charlesworth 2006; Kawabe et al. 2007). Consistent with this prediction, the estimated gene diversity of the comC locus is higher than that estimated for the MLST loci (see tables 1 and 4). Further exploration of the polymorphism at comC revealed that there is significantly less variation within pherotypes than would be expected if the locus was evolving in a neutral fashion, as evidenced by the pattern of polymorphism and differentiation in the MLST loci (results of the 2D-HKA test on CSP1 and CSP2 subpopulations). Particularly interesting are the results of our analyses suggesting that the pattern of fixed versus polymorphic differences found in comC is consistent with positive selection maintaining the differentiation between alleles in this locus.

The results of our analyses and experimental evidence showing the inability of CSP1 to induce competence on CSP2 backgrounds (comD2 receptors), and vice versa (Iannelli et al. 2005), suggest that the polymorphism in this locus is maintained by selection. We propose a scenario by which this may have arisen. If an inability to become competent is associated with deleterious fitness effects, then any loss of function or reduction in the efficiency of the function will be selected against. If this is so, any mutation in the signal peptide that induces competence (CSP), or its receptor, leading to reduction in the recognition of the two-component system with a concomitant effect on competence will have two possible fates: 1) It will be selected against and that mutant will be lost from the population or 2) Compensatory mutations in the receptor (or signal peptide) that restore the efficiency of the recognition will occur before the original mutant is lost and will be maintained in the population because the process itself has been restored (negative selection maintaining the competence phenotype). The first possible fate, which may be the dominant fate due to stabilizing selection, is not of direct interest because this would not result in the development of a polymorphic system. On the other hand, the second fate is relevant because the restoration of wild-type competence by a form of compensation will have generated the initial variation necessary to cause polymorphism. This scenario for the evolution of lock and key components, or more generally for coevolving residues that structurally interact, in which mispairing of the alleles for signal and receptor significantly reduce the efficiency of the competence, could account for the emergence and further maintenance of polymorphism in this system.

A similar hypothesis has been recently proposed for the evolution of highly polymorphic self-incompatibility systems in crucifers (Chookajorn et al. 2004). However, the reduced variability within two CSP types found here represents a substantial difference with the mechanism proposed by Chookajorn et al. (2004) according to which variation within alleles is necessary for the diversification of the locus. On the other hand, there are eight proposed CSP alleles reported in the literature (Kilian et al. 2008), only two of which (the same as those found here) have been assessed experimentally for functional differences. It is possible that some of the alternative alleles are intermediate forms with reduced recognition, which have not yet completely differentiated from well-established types. It is notable that these alternatives alleles were not detected in our collection of clinical isolates, which suggests they are
present in low frequency in the population. A more extensive survey should be performed in order to assess their contribution to the population genetic structure of *S. pneumoniae* and the evolution of the comC locus.

The hypothesis proposed in this work is consistent with both the polymorphism in comC loci, the low polymorphism within comC allele, and the lack of genetic differentiation between subpopulations of isolates presenting different pherotypes. Further experimental and sequence analysis necessary to adequately test this idea represents a central aim of our future work.

**Supplementary Material**

Supplementary materials are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We would like to thank Mark Jensen for insightful suggestions. Also, we would like to thank John H. McDonald and two helpful reviewers for their helpful comments on an earlier version of this manuscript. O.E.C. was supported by a grant from the US National Institutes of Health, AI40662 awarded to Bruce R. Levin. D.E.R. was supported by a grant from the BBSRC (BBF0020681).

**References**


