Estimating Recombinational Parameters in *Streptococcus pneumoniae* From Multilocus Sequence Typing Data

Edward J. Feil,* John Maynard Smith,† Mark C. Enright* and Brian G. Spratt*

**Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, Oxford OX1 3FY, United Kingdom and †School of Biological Sciences, University of Sussex, Falmer, Brighton BN1 9QG, United Kingdom*

Manuscript received September 14, 1999
Accepted for publication December 28, 1999

ABSTRACT

Multilocus sequence typing (MLST) is a highly discriminatory molecular typing method that defines isolates of bacterial pathogens using the sequences of ∼450-bp internal fragments of seven housekeeping genes. This technique has been applied to 575 isolates of *Streptococcus pneumoniae* and identifies a number of discrete clonal complexes. These clonal complexes are typically represented by a single group of isolates sharing identical alleles at all seven loci, plus single-locus variants that differ from this group at only one out of the seven loci. As MLST is highly discriminatory, the members of each clonal complex can be assumed to have a recent common ancestor, and the molecular events that give rise to the single-locus variants can be used to estimate the relative contributions of recombination and mutation to clonal divergence. By comparing the sequences of the variant alleles within each clonal complex with the allele typically found within that clonal complex, we estimate that recombination has generated new alleles at a frequency ∼10-fold higher than mutation, and that a single nucleotide site is ∼50 times more likely to change through recombination than mutation. We also demonstrate how to estimate the average length of recombinational replacements from MLST data.

RECOMBINATIONAL exchanges between isolates of bacterial species are commonly observed in genes whose products are subject to strong selection by the host immune system (Brunham *et al.* 1993) or by antibiotic usage (Spratt 1994), but the contribution of recombination to the accumulation of neutral genetic variation within bacterial populations is controversial (Maynard Smith *et al.* 1993; Guttman 1997; Spratt and Maiden 1999). Very low rates of recombinational exchanges between different isolates of bacterial species lead to high levels of linkage disequilibrium between alleles and a population structure that is characterized by the presence of independent clonal lineages, which diversify slowly by the accumulation of point mutations (Selander and Musser 1990). With an increasing contribution of recombinational exchanges, compared to point mutations, the levels of linkage disequilibrium between alleles are reduced, and the clones within the population become increasingly unstable as their integrity is disrupted by recombinational replacements (Maynard Smith *et al.* 1993; Spratt and Maiden 1999).

Evidence for a history of recombination within bacterial populations has been obtained by analyzing the sequences of housekeeping genes from different isolates (Dykhuizen and Green 1991; Feil *et al.* 1995, 1996; Zhou *et al.* 1997; Suerbaum *et al.* 1998; Holmes *et al.* 1999) or from low levels of linkage disequilibrium between the alleles present at different loci within a sample of the population (Istock *et al.* 1992; O'Rourke and Stevens 1993). However, although a method based on the patterns of descent has recently been described (Maynard Smith 1999), comparative estimates of rates of recombination in different species have proved difficult to obtain (Guttman 1997; Spratt and Maiden 1999). Consequently, there are very different views on the relative importance of recombinational exchanges and point mutations in the diversification of bacterial lineages.

Guttman and Dykhuizen (1994) proposed that the relative impact of recombination and mutation on the evolution of bacterial populations can be estimated by comparing the sequences of stretches of the chromosome in isolates that are very closely related. The identification of the recombinational and mutational events that have occurred since the recent common ancestor of these isolates should be relatively simple, compared to comparisons between distantly related isolates where ancient recombinational or mutational events are frequently obscured by more recent events. Using this approach, Guttman and Dykhuizen (1994) estimated (as an upper bound) that a single-nucleotide site in an *Escherichia coli* housekeeping gene is 50 times more likely to change by recombination than by mutation.

This approach has also been used to estimate the relative impact of recombination compared to point mutation in the diversification of *Neisseria meningitidis*

Corresponding author: Edward Feil, Wellcome Trust Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, South Parks Rd., Oxford OX1 3FY, United Kingdom. E-mail: ed.feil@ceid.ox.ac.uk

clones (Feil *et al.* 1999), using data obtained during the molecular characterization of isolates of this species by multilocus sequence typing (MLST). MLST is a new molecular typing method that is conceptually similar to multilocus enzyme electrophoresis, except that allelic variation within housekeeping genes is detected directly by sequencing ∼450-bp internal fragments of the genes, rather than indirectly through the differing electrophoretic mobilities of their gene products (Maiden *et al.* 1998; Enright and Spratt 1999a). The variation within these housekeeping genes is assumed to be largely neutral. Every different sequence at each of seven loci is assigned an allele number, and isolates are defined by the alleles at the seven loci (the allelic profile). Besides providing a powerful approach for the characterization of isolates of pathogens for epidemiological studies, MLST provides the sequences of fragments of seven housekeeping genes from hundreds of isolates and these data can be used to address aspects of the population and evolutionary biology of bacterial species (Spratt 1999).

The relative contributions of recombination and point mutation to the diversification of *N. meningitidis* clones has been estimated from MLST data by selecting the clusters of isolates that have identical allelic profiles (clones) and identifying the minor variants that differ from the typical allelic profile of these clones at only one of the seven housekeeping loci (Feil *et al.* 1999). The nucleotide sequence differences between the variant allele and the allele typically present in the clone can be used to infer whether the clonal diversification that has given rise to each variant has arisen by recombination (multiple-nucleotide differences) or point mutation (single-nucleotide difference). Meningococcal housekeeping genes are relatively variable and, in the study of Feil *et al.* (1999), the variant alleles differed from the typical alleles in the clone, either at a single site or at ≥5 sites. The assumption that the former class of variant alleles arose by point mutation, whereas the latter are due to recombinational imports, is therefore likely to be valid. However, housekeeping genes of many bacterial species are more uniform than those of *N. meningitidis*, and some recombinational exchanges in these species will introduce only a single-nucleotide difference and may be mistaken as point mutations. Furthermore, some variant alleles will differ at two sites, and the possibility that these represent two independent mutational events or a single mutational event altering two sites, rather than a recombinational event, has to be considered.

In this article we describe procedures that may be used to estimate from MLST data the true number of recombinational replacements and point mutations, and thus the ratio of recombination to mutation, in a relatively uniform species, *Streptococcus pneumoniae* (the pneumococcus). The pneumococcus is naturally transformable and the recombinational exchanges described in this analysis are assumed to have been mediated via this process. As transformation results in very localized exchanges (as discussed later), it is unlikely that single recombinational exchanges will affect more than one of the MLST loci. *S. pneumoniae* is the causative agent of a variety of diseases of man, ranging in severity from otitis media and sinusitis to pneumonia, septicemia, and meningitis (Feldman and Klugman 1997). Despite being such an important human pathogen, very little is known about the population structure of pneumococci. Our analysis of MLST data from 575 isolates of *S. pneumoniae* indicates that recombinational exchanges generate new alleles ∼10-fold more frequently than point mutations and that an individual nucleotide site within a pneumococcal housekeeping gene is ∼50 times more likely to change by recombination than by point mutation. We also demonstrate how MLST data can be used to estimate average replacement lengths and draw comparisons between *S. pneumoniae* and *N. meningitidis*.

MATERIALS AND METHODS

Bacterial isolates: The sequences of the seven housekeeping gene fragments used for MLST have been determined for 575 *S. pneumoniae* isolates. The 575 isolates included a collection of 380 isolates from invasive disease in eight countries (Enright and Spratt 1998; Enright *et al.* 1999), 74 penicillin-resistant isolates recovered from hospitals in Taiwan (Shi *et al.* 1998), and 66 isolates of the major Spanish penicillin-resistant and multiply antibiotic-resistant clones (Zhou *et al.* 2000). The allelic profiles of these isolates, associated epidemiological data, and the sequences of the alleles at the seven loci are available from the pneumococcal MLST database (isolates OX1-OX575; http: // mlst.zoo.ox.ac.uk).

Data analysis: The different sequences at each locus were assigned as alleles, thus allowing each of the 575 pneumococcal isolates to be defined by its allelic profile, corresponding to the alleles present at each of the seven loci (Enright and Spratt 1998). We defined a clonal complex as a group of at least two identical isolates, which we termed the consensus group, plus those isolates differing from the consensus group at only one locus out of the seven (single-locus variants; SLVs). As it is extremely unlikely that two isolates will possess identical alleles at six out of the seven loci by chance (see below), isolates that are identical in allelic profile and the SLVs can be assumed to have a recent common ancestor. We assume the isolates of the consensus group have the ancestral allelic profile and the SLVs are descended from isolates of the consensus group. Comparisons between consensus groups and associated SLVs allowed an estimate of how many of the SLVs have arisen through mutation and how many by recombination.

The consensus group was defined by the cluster of isolates that had the most common allelic profile within a clonal complex. Evidence to support the assumption that this group represents the ancestral type of the clonal complex is discussed in the next section. In three cases (clonal complexes 1, 2, and 18; Figure 2), SLVs, each represented by multiple isolates, were themselves associated with a number of other SLVs and thus defined a second consensus group within the clonal complex. For comparisons between consensus groups, the group represented by the largest number of isolates was assumed to be ancestral; the exception being clonal complex 1(a), which was assumed to be ancestral to clonal complex 1(b) on the basis that the latter group are all penicillin-resistant and probably arose from the former (which are all penicillin-susceptible).

The consensus groups and corresponding SLVs were nonoverlapping, except that one isolate was a SLV of two different consensus groups and this isolate was removed from the analysis. The identification of the consensus groups and SLVs was carried out using a program written in Visual Basic by E. J. Feil and was confirmed using the query software available on the MLST web site (http://mlst.zoo.ox.ac.uk). Comparisons of the sequences of variant alleles were carried out using "Sequence Output," which is also available from the MLST web site.

The single-nucleotide differences in those variant alleles that differed at only a single site from the allele found in the consensus group were carefully rechecked on both strands, using the original ABI377 sequencer traces. If there was any cause for doubt, the gene fragment was reamplified by PCR and was resequenced on both strands.

RESULTS AND DISCUSSION

The identification of clonal complexes: Of the 575 isolates, 410 were assigned to 28 clonal complexes, each containing at least one consensus group (the isolates with the predominant allelic profile) and one SLV. Of the other 165 isolates, 53 had allelic profiles that differed from those of all other isolates at two or more loci, and 112 were members of uniform clones that had no SLVs, or of clonal complexes that only contained double-locus variants. The relationships among the isolates within three representative clonal complexes are shown as a dendrogram in Figure 1. The isolates within each of the 28 clonal complexes are shown in Figure 2.

There was an average of ~37 alleles per locus and the pneumococcal MLST scheme can therefore resolve billions of allelic profiles. The range in the number of alleles per locus was 26–54, which implies that the loci were evolving at approximately the same rate, although there is evidence that *ddl* is evolving faster than the other loci because of hitchhiking effects (see below). Because of the large number of alleles per locus, pneumococcal isolates with identical allelic profiles are highly unlikely to occur by chance and can be assumed to be closely related by descent (Enright and Spratt 1998, 1999a; Spratt 1999). This assumption is supported by the observation that all isolates within 21 of the 28 clonal complexes were uniform in serogroup (Table 1). Serogroup variation within the other seven clonal complexes is likely to have arisen through recombinational exchanges at the capsular locus that determines serogroup (Coffey *et al.* 1991), and this has been demonstrated experimentally for clonal complexes 1a and 6 (Coffey *et al.* 1998, 1999), which correspond to the major penicillin-resistant Spanish serotype 9V and 23F clones, respectively (Crook and Spratt 1998).

Estimating the contributions of recombination and mutation to the emergence of variant alleles: A total of 98 SLVs were identified within the 28 clonal complexes (Figure 2). The variant alleles in the SLVs were assigned as descendant and were compared with the allele at the corresponding locus in the consensus group, which was assigned as the ancestral allele. A recombinational replacement is likely to introduce multiple nucleotide changes, whereas a point mutation will only result in a single change. The number of polymorphic sites between alleles therefore allows a provisional assignment of which descendant alleles have arisen through recombination and which by mutation (Figure 3).

Descendant alleles differing at three or more nucleotide sites from the ancestral allele can be assigned as the result of recombination, as it is unlikely that an isolate will accumulate multiple independent point mutations at a single locus and no changes at the other six loci. Thus, suppose that three mutations occur within the seven loci. The probability that all three will be in the same locus is $\sim(1/7)^2$ or 0.02. The possibility that alleles that differ at only two sites could be due to two independent mutations is more serious and is discussed later.

However, it cannot be assumed that a descendant allele that differs at a single nucleotide site has arisen by point mutation. This is because single-nucleotide polymorphisms may have arisen through recombinational exchanges between very similar alleles. Descendant alleles that differ at a single site will therefore represent a mixture of recombinational replacements and point mutations (Figure 4). To estimate the impact of recombination compared to point mutation, we need to be able to distinguish which of these alleles have arisen by mutation and which by recombination.

Estimating the true number of mutations: Point mutations can also be distinguished from recombinational exchanges on the basis of the presence of the resultant allele elsewhere in the MLST data set. A point mutation will almost certainly result in a novel allele that is not found elsewhere in the data set, but a recombinational exchange that replaces the entire sequenced fragment will result in an allele that must be present elsewhere in the pneumococcal gene pool, but which may or may not be present in the sample of the population that constitutes the data set. We therefore examined the occurrence of the variant alleles in the SLVs elsewhere in the data set. To remove the possibility that two isolates share the same variant allele by virtue of descent rather than by horizontal gene transfer, we made comparisons only between unrelated isolates, defined as those that differ at three or more loci.

In 17 of the 98 allelic comparisons the descendant allele differed from the ancestral allele at a single site; 11 of these single-nucleotide differences were synonymous. Nine (53%) of these 17 alleles were novel and 8 (47%) were shared (*i.e.*, they were found elsewhere in the data set). Of the 81 descendant alleles that differed at multiple (≥2) sites, 15 (19%) were novel and 66 (81%) were shared (Table 1; Figure 4). The greater percentage of

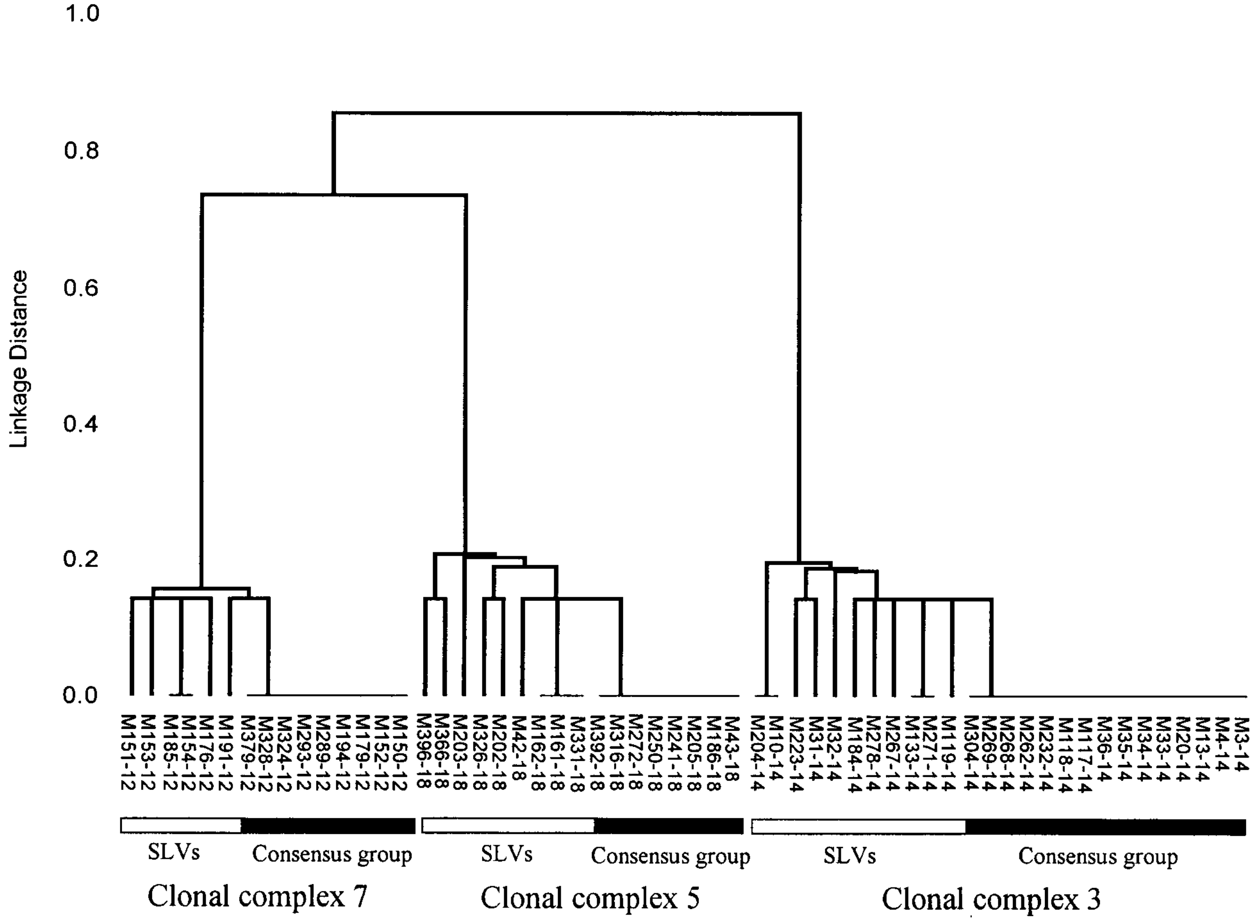

Figure 1.—Dendrogram showing the relatedness of isolates within three of the clonal complexes. A dendrogram was constructed by the unweighted pair-group method with arithmetic averages from the matrix of pairwise percentage differences between the allelic profiles of the 58 isolates assigned to clonal complexes 3, 5, and 7. The isolates within each clonal complex were identified as described in materials and methods. The serogroup of each isolate is shown, separated from the name of the isolate by a hyphen. The horizontal bars indicate the consensus groups and the corresponding SLVs. No inferences should be made about the genetic relationships between the clonal complexes.

novel descendant alleles that differed at a single site, compared to those that differed at multiple sites, was statistically significant using a 2 × 2 chi square test ($P < 0.01$). This difference is to be expected, as descendant alleles differing at single sites should include both point mutations (which are highly likely to result in a novel allele) and recombinational events (which may well result in a shared allele).

This observation also suggests that our assumptions about the directionality of events (*i.e.*, the assignment of alleles as either ancestral or descendant, based on the frequency of isolates within putative consensus groups) have generally been correct. Thus if we repeat the comparison, treating the ancestral alleles as if they were descendant, and vice versa, we find no difference between the proportion of novel alleles among the single-nucleotide changes (2/17) compared to multiple-nucleotide changes (8/81; $P > 0.1$).

The high proportion (81%) of descendant alleles differing at multiple sites (which we assume to have arisen via recombination) that are found elsewhere in the data set suggests that the large MLST data set contains the majority of alleles in the pneumococcal population and means that we can approximate the number of mutations as the number of novel descendant alleles that differ at a single site. However, there remain several potential sources of error that need to be considered. For example, it is possible that a point mutation may result in a shared allele by chance, or that some of the novel alleles that differ at a single site are due to short (part-fragment) recombinational exchanges resulting in novel mosaic alleles where at least one of the recombinational crossovers is within the sequenced region. Alternatively, a novel allele differing at a single site may be an imported allele that is not represented elsewhere in the data set. Furthermore, the presence of hypermutable sites would increase the probability that a point mutation will result in a shared allele and so be incorrectly assigned as a recombinational exchange. A discussion of these potential sources of error is given in the appendix.

An additional potential source of error, not discussed

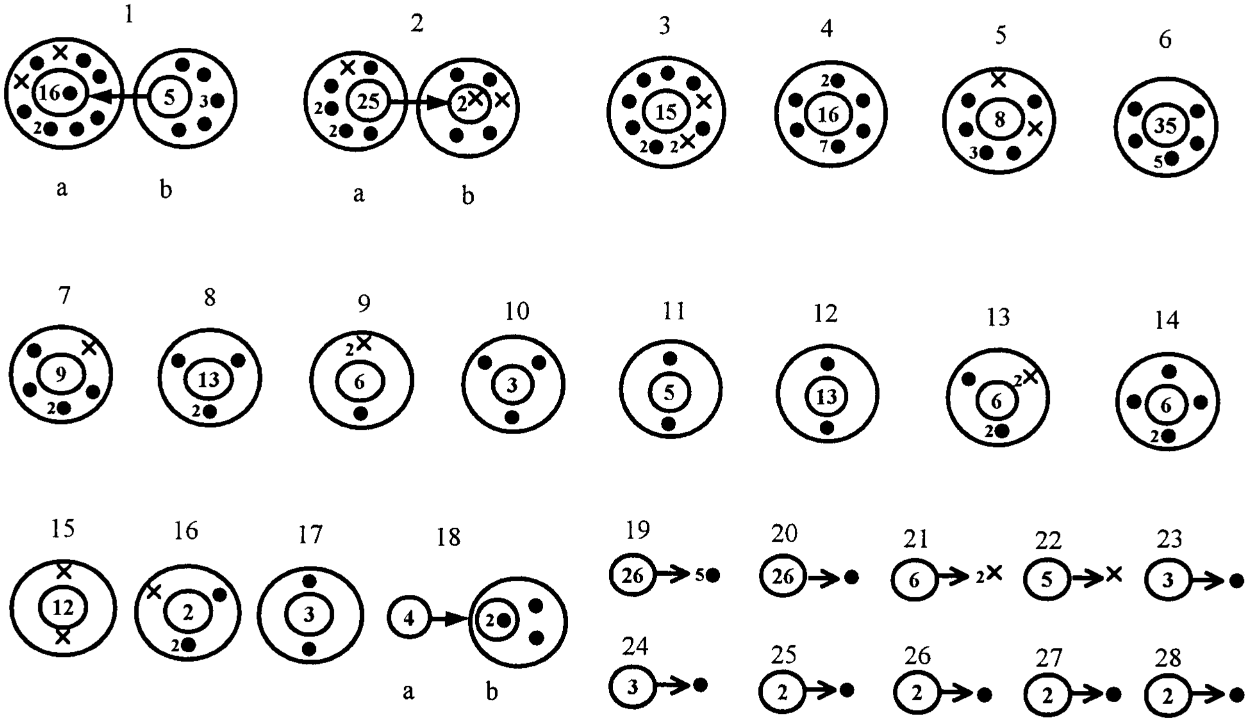

Figure 2.—The clonal complexes among the 575 isolates of *S. pneumoniae*. Each of the 28 clonal complexes is assigned a number, which is given above the clonal complex. Where two consensus groups were identified within a clonal complex, the labels a and b are given below the consensus groups. Consensus groups are represented as a small circle, and the number of isolates within the consensus group is given inside the circle. In cases where more than one SLV is associated with a consensus group (clonal complexes 1–18), the SLVs are enclosed in a larger circle around the consensus group. Each SLV is shown either as a dot (●) where the variant allele in the SLV differs at multiple (≥ 2) sites from the ancestral allele in the consensus group or as a cross (×) where the variant allele differs at only a single nucleotide from the ancestral clonal allele. Where a SLV is represented by multiple isolates, the number of identical isolates of the SLV is given as a prefix (*i.e.*, 2● refers to a SLV where the variant allele differs at multiple sites from the ancestral clonal allele and is represented by two identical isolates). In cases where only a single SLV is associated with a consensus group (clonal complexes 19–28), the relationship between the consensus group and the SLV is given by an open-headed arrow, which shows the assumed direction of descent. In the three cases where a clonal complex contains two consensus groups, the assumed direction of descent between the consensus groups is given by a closed-headed arrow.

in the appendix, is the possibility that some shared alleles that differ at a single-nucleotide site represent point mutations that have subsequently been exported to unrelated isolates. There are two reasons why this is unlikely to be a serious source of error. First, it is less parsimonious to assume that these alleles have been exported out of a clonal complex (which would require a point mutation followed by a recombination event), rather than imported into a clonal complex. Second, the eight shared alleles that differed at single sites were found in an average of >50 unrelated isolates representing a number of distantly related lineages. The widespread distribution of these alleles among distantly related lineages makes it likely that they predate the origin of the SLVs and have been imported rather than exported. A summary of the four most likely scenarios leading to the generation of a single-locus variant is given in Figure 4.

Estimating the true number of recombinational replacements: The possibility that a descendant allele that differs from its ancestral allele at two nucleotide sites has arisen by two independent point mutations has to be considered. There were 20 descendant alleles that differed at two nucleotide sites; 5 of these were novel (25%) and 15 were shared alleles (75%). As it is very unlikely that the same two mutations will have occurred independently within more than one isolate (the possibility of hypermutable sites can be ignored; see above and the appendix), the shared alleles can be assumed to have been generated through recombination. Although two independent mutations would almost certainly produce a novel allele, the percentage of novel alleles was not significantly different for descendant alleles that differed at two sites compared to those that differed at three or more sites. Most of these novel alleles are therefore likely to be imports of alleles that are not within the MLST database and the vast majority, if not all, of the variant alleles that differed at two nucleotide sites are probably due to recombination rather than two independent mutational events.

In the analysis of the *N. meningitidis* MLST data set there were no descendant alleles that differed from their ancestral alleles at between two, three, or four nucleotide sites (Feil *et al.* 1999). The absence of de-

TABLE 1

The consensus groups and associated SLVs of each clonal complex

Clonal complex	Sequence type of consensus group	No. of strains in consensus group	Serogroup(s) of strains in consensus group[a]	Serogroup(s) of SLVs[b]	Resistance to penicillin[c]	No. of SLVs differing at a single-nucleotide site[d]		No. of SLVs differing at multiple-nucleotide sites[d]	
						Novel	Shared	Novel	Shared
1a	ST156	16	9(11), 14(4), 23(1)	14(5), 9(4)	R	2	0	2	5
1b	ST162	5	19(3), 9(1), 14(1)	9(3), 19(1), [9(2),19(1)] + 1(a)	S	0	0	0	6
2a	ST9	25	14(24), 19(1)	14	S	0	2	2	3
2b	ST15	2	14	14	S	1	0	1	3
3	ST124	15	14	14	S	0	2	0	7
4	ST90	16	6	6	R	0	0	5(4)	1
5	ST113	8	18	18	S	2	0	2	3
6	ST81	35	23(25), 19(1)	23(4), 19(1)	R	0	0	0	5(4)
7	ST218	9	12	12	S	0	1	0	4
8	ST180	13	3	3	S	0	0	0	3
9	ST53	6	8	8(1), 11(1)	S	0	1	0	1
10	ST37	3	23	23	S	0	0	0	3
11	ST199	5	19(4), 15(1)	19(1), 14(1)	S	0	0	0	2
12	ST242	13	23(12), 19(1)	23	S	0	0	0	2
13	ST247	6	4	4	S	1	0	0	2
14	ST18	6	14	14	R	0	0	0	4(3)
15	ST138	12	6	6	S	0	2	0	0
16	ST173	2	23	23	R	1	0	2(1)	0
17	ST66	3	9(2), NK(1)	9(1), 14(1)	S	0	0	1	1
18a	ST77	4	19	19	S	0	0	0	1
18b	ST177	2	19	19	S	0	0	0	2
19	ST236	26	19	19	R	0	0	0	1
20	ST191	26	7	7	S	0	0	0	1
21	ST205	6	4(5), NT(1)	4	S	1	0	0	0
22	ST73	5	15	15	S	1	0	0	0
23	ST135	3	6	6	R	0	0	0	1
24	ST202	3	19	19	S	0	0	0	1
25	ST30	2	16	16	S	0	0	0	1
26	ST210	2	23	23	S	0	0	0	1
27	ST233	2	3	3	S	0	0	0	1
28	ST146	2	6	6	S	0	0	0	1
					Totals	9	8	15(13)	66(64)

[a] The numbers of isolates of each serogroup are shown in parentheses. NT, nontypable; NK, not known as the isolate is no longer available.

[b] The numbers of SLVs (not the number of isolates) of each serogroup are shown in parentheses. In one case, a single SLV was represented by multiple isolates of differing serogroups. This SLV, which is associated with consensus group 1b, is represented by three isolates, two of serogroup 19 and one of serogroup 9, and is shown in square brackets. + 1(a), the isolates in the consensus group of clonal complex 1a are also SLVs of the consensus group of clonal complex 1b.

[c] The consensus groups containing penicillin-resistant isolates [minimum inhibitory concentration (MIC) > 0.1 μg/ml] are designated R. Those in which all isolates were penicillin-susceptible are designated S.

[d] The numbers of SLVs corresponding to the four classes—multiple- or single-nucleotide differences, which may be either novel or shared. The numbers of SLVs in parentheses are those after the exclusion of the four highly diverged replacements at *ddl*. The ST number can be used to find additional information about the isolates from the MLST web site (http://mlst.zoo.ox.ac.uk).

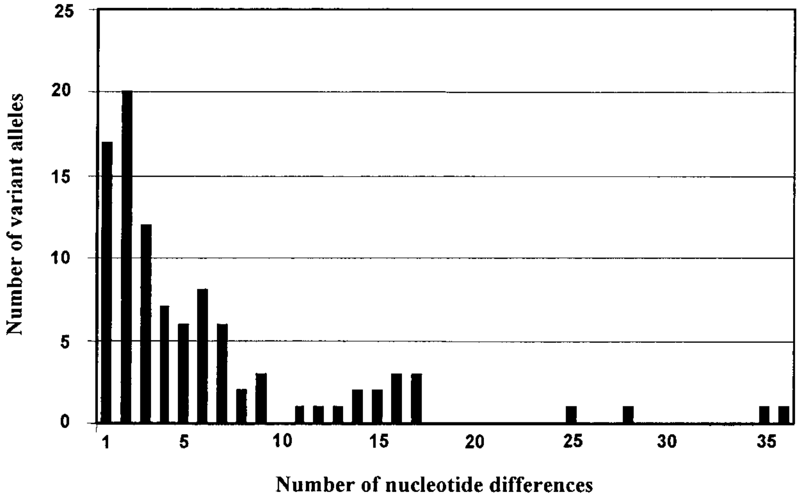

Figure 3.—Frequency distribution of the numbers of nucleotide differences between ancestral and descendant alleles. Approximately half of the variant alleles that differ at a single-nucleotide site are assumed to represent point mutations, the other half representing recombination between very similar parental sequences (see text). The four highly diverged replacements (differing at 25, 28, 35, and 36 nucleotide sites) are all within *ddl*, and, for reasons given in the text, these are excluded from the analysis.

scendant alleles that differ at two to four sites is almost certainly due to the higher levels of diversity within meningococcal genes (average of ~4.4%), compared to pneumococcal genes (average of ~1.1%), such that recombinational replacements in meningococci usually result in the introduction of at least five nucleotide differences within a 450-bp gene fragment. The absence of descendant alleles that differed at two sites suggests that two independent mutational events, or the introduction of two nucleotide differences by a DNA repair process, are rare events in meningococci. Although caution is clearly required when extrapolating from meningococci to pneumococci, this indirectly supports the suggestion that the descendant pneumococcal alleles that differ at two sites are due to recombinational events within a species of low sequence diversity. This view was supported further by simulating 1000 whole-fragment recombinational replacements between randomly chosen alleles at each of the seven pneumococcal loci. Approximately 18% of these simulated allelic replacements involved donor and recipient alleles that differed at only one or two nucleotide sites (data not shown).

Estimating the ratio of recombinational to mutational events: A crude estimate of the ratio of recombinational events compared to mutational events during clonal diversification can be obtained by assuming that the 81 descendant alleles that differ at multiple sites are due to recombination and the 17 that differ at a single site are due to point mutation. This provides a ratio per gene fragment of ~5:1, which, as outlined above, is likely to be a considerable underestimate of the significance of recombinational replacements, as some of the alleles that differ at a single site will be due to recombination, rather than mutation. A more realistic approach is to consider only the 9 novel alleles that differ at a single site to have arisen by mutation and the 81 alleles that differ at multiple sites, plus the 8 shared alleles that differ at a single site, to have arisen by recombination. On this assumption the estimated ratio of recombinational events to mutational events per gene fragment is ~10:1.

The per site ratio of recombination to mutation: A more meaningful comparative measure is the relative likelihood that an individual nucleotide site within a housekeeping gene will change by recombination compared to mutation (the per site *r*/*m* parameter). This can be calculated from the total number of nucleotide changes within the 89 alleles assumed to have arisen via recombination (592) compared to the number introduced by mutation (9). The resulting per site *r*/*m* parameter of ~66:1 implies that recombination changes an individual nucleotide site within a pneumococcal housekeeping gene ~66 times more frequently than point mutation.

From Figure 3 it is apparent that four of the recombinational replacements in the SLVs have introduced highly diverged sequences, and each of these replacements is within the *ddl* locus. These four replacements, which were all in SLVs of penicillin-resistant isolates, have introduced a total of 124 polymorphic sites, and their inclusion in the analysis markedly alters the per site *r*/*m* parameter in favor of recombination. The *ddl* gene is 783 bp downstream of the penicillin-binding protein 2b (PBP2b) gene. Interspecies recombinational replacements that result in reduced affinity of PBP2b for β-lactam antibiotics and increased resistance to these antibiotics sometimes extend into or through *ddl.* The divergent *ddl* alleles are therefore the result of hitchhiking with a gene at which rare interspecies replacements are strongly selected by antibiotic usage (Enright and Spratt 1999b). Excluding these four replacements reduced the per site *r*/*m* parameter to 52:1. There were 12 other SLVs where *ddl* was the altered locus. If we also excluded these SLVs, the estimated per site *r*/*m* parameter reduced to 45:1. This suggests that, excepting the very diverged replacements, the inclusion of *ddl* does not markedly alter the per site *r*/*m* parameter.

A timescale for clonal diversification: It is assumed

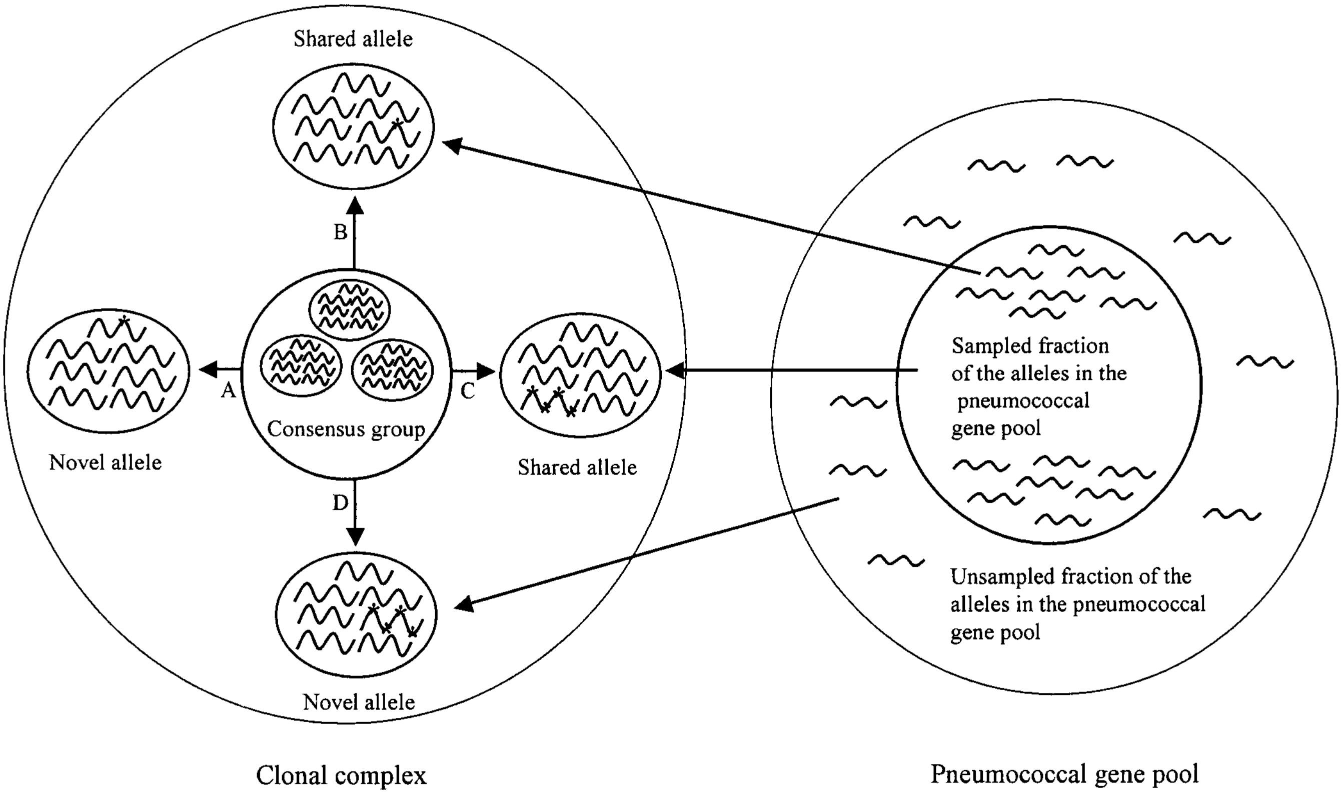
Shared allele
B
A
C
Consensus group
Novel allele
Shared allele
D
Novel allele
Sampled fraction of the alleles in the pneumococcal gene pool
Unsampled fraction of the alleles in the pneumococcal gene pool
Clonal complex
Pneumococcal gene pool

Figure 4.—Scenarios for the generation of SLVs from an isolate belonging to a consensus group. Each allele is shown by a wavy line. The members of the consensus group have identical alleles at all seven loci whereas the SLVs differ from this group at a single locus. The nucleotide differences between the variant allele in each SLV and the allele present in the consensus group are represented by asterisks. Single-nucleotide changes are represented by a single asterisk and multiple-nucleotide changes are represented by four asterisks. The four scenarios are as follows: (A) A point mutation generates a novel variant allele that differs from the consensus allele at a single-nucleotide site. This class is believed to account for ~50% of the variant alleles that differ at a single site. (B) A recombinational replacement with a donor allele that differs at only a single-nucleotide site generates a shared allele that differs at a single site, as the donor allele is present in unrelated isolates in the pneumococcal MLST database. This class is believed to account for ~50% of the variant alleles that differ at a single site. (C) A recombinational replacement with a donor allele that differs at multiple-nucleotide sites generates a shared allele, as the donor allele is present in unrelated isolates in the MLST database. This class accounts for ~81% of the variant alleles that differ at multiple sites. (D) A recombinational replacement with a donor allele that differs at multiple-nucleotide sites generates a novel allele, as the donor allele is not present elsewhere in the MLST database. This class accounts for ~19% of the variant alleles that differ at multiple sites. It is also possible that a novel variant allele that differs at a single site may arise through a recombinational replacement with an allele not represented in the database, or that novel alleles may arise by intrafragment replacements. However, we believe these classes of alleles are rare (see text and appendix).

that the variation within the clonal complexes has occurred recently, but it is not possible to estimate the age of most clonal complexes due to the absence of old isolates. However, it is possible to date the clonal complexes that consist exclusively of penicillin-resistant isolates, as these must have arisen since the introduction of penicillin in the 1940s and have probably arisen in the last 25 years (Crook and Spratt 1998). Seven of the clonal complexes contained isolates that were all penicillin-resistant (Table 1). The per site *r*/*m* parameter estimated for these clones was 53:1, a very similar value to that calculated for the data set as a whole. The rate of clonal diversification is thus sufficiently high that clones <50 years old are no longer uniform by MLST. It is possible that penicillin resistance has emerged independently more than once within a clone. However, molecular evidence suggests this is unlikely as the three highly variable penicillin-binding protein genes conferring resistance to penicillin are almost always completely uniform within the penicillin-resistant clones (Zhou *et al.* 2000).

Replacement lengths in pneumococci and meningococci: The per site *r*/*m* parameter is a product of three other parameters: the rate of recombination (*i.e.*, the rate at which the chromosome is affected by a recombinational event as compared to mutation), the average amount of diversity between the parental molecules, and the average size of recombinational replacements. Therefore, to compare the frequency of replacements, compared to mutation, in pneumococci and meningococci from the per site *r*/*m* parameter, it is necessary to estimate average replacement sizes.

The proportion of recombination events that result in replacement of only part of a gene fragment of known size can be used to estimate the average size of recombinational replacements as follows. Assume, for simplicity, that all replacements have the same length, w bp, and that the length of the gene fragment is d bp. A particular replacement is characterized by the (unknown) position of the upstream crossover point. The range of positions that affect a particular gene is $w + d$, and, within this range, a length of $2d$ results in a part-fragment replacement. Hence if r is the fraction of recombination events that result in a part-fragment replacement, $r = 2d/(w + d)$, or $w = d(2 - r)/r$.

Because estimates of replacement lengths are useful when drawing comparisons between different species, consider first the meningococcus (Feil *et al.* 1999). The occurrence of a part-fragment replacement was determined by two criteria: the nonrandom distribution of polymorphic nucleotides (the confinement of polymorphic sites to one end of the fragment) and the absence of the variant allele in unrelated isolates in the database. In the meningococcus, 18 clonal variants were identified that differed from the ancestral allele at five or more nucleotide sites (Feil *et al.* 1999), and we assigned 2 of these clonal variant alleles as part-fragment replacements. Hence, $d = 450$ and $r = \frac{1}{9}$, so the estimated length of a typical recombinational replacement in the meningococcus is 7.6 kb.

The pneumococcal data are harder to interpret, because the gene fragments contain fewer polymorphic sites, and the polymorphisms tend to be clustered (probably because of more ancient recombinational events). This makes it difficult to identify part-fragment replacements, so we confined our analysis to two of the most polymorphic gene loci, *ddl* and *gki*. Applying the same criteria, we concluded that 3/16 recombinants were part-fragment replacements at these two loci, and hence $w = {\sim}4.4$ kb. This is in good agreement with a recent experimental estimate that showed that 50% of replacements were >5 kb. We conclude that, in both genera, the typical size of a replacement is of the order of 5–10 kb.

Further comparisons between the pneumococcus and the meningococcus: Table 2 gives the estimated recombinational parameters for the meningococcus and the pneumococcus. Although these figures should be regarded only as approximate, we can begin to draw comparisons between the two species. Feil *et al.* (1999) recently observed a total of 18 recombinational events and five putative point mutations within meningococcal clonal complexes. However, one of the putative point mutations corresponded to a "shared" allele and, following the criteria used in this article, was reassigned from a possible recombinational exchange to a probable recombinational exchange. This changes the per site r/m parameter from >80:1 (Feil *et al.* 1999) to ~100:1 for the meningococcus, which is roughly double that observed within the pneumococcus. However, the estimated average ratio of recombinational events to mutational events (the ratio at which SLVs arise by recombination compared to mutation) in the meningococcus is now ~5:1 (19/4), or roughly half the ratio observed for the pneumococcus (Table 2). Assuming that the average size of recombinational replacements is similar between these two species (see above), the difference between the ratio of recombinational events compared to the per site r/m ratio is explained by the fact that the average sequence diversity between parental molecules is approximately fourfold higher in the meningococcus (4.4%) than the pneumococcus (1.1%). Thus, each replacement, on average, introduces fourfold as many polymorphisms in the meningococcus compared to the pneumococcus. This increased diversity results in a higher per site r/m ratio in the meningococcus, despite the frequency of recombinational events in the meningococcus being lower than that in the pneumococcus.

The high degree of diversity within meningococcal genes is due to the high frequency of interspecies recombinational exchanges with closely related species in housekeeping genes of this species (Zhou and Spratt 1992; Feil *et al.* 1995, 1996; Zhou *et al.* 1997). Interspecies replacements within housekeeping genes of *S. pneumoniae* appear to be much less common (Enright and Spratt 1998). The reason behind this difference is at present unclear, as commensal streptococcal species cohabit the nasopharynx alongside the pneumococcus,

TABLE 2

Recombinational parameters for S. pneumoniae and *N. meningitidis*

	Ratio of recombinational to mutational events per 450-bp gene fragment	Average size of recombinational replacements	Average sequence diversity between parental molecules (%)	Per site *r/m* parameter
S. pneumoniae	~10:1	5–10 kb	1.1	~50:1
N. meningitidis[a]	~5:1	5–10 kb	4.4	~100:1

The value of the meningococcal per site *r/m* parameter is slightly different from that given previously (>80:1) as in line with the criteria used here, we reassigned one SLV that differed as a single site as the result of recombination since the resulting allele was present within an unrelated isolate in the meningococcal MLST database.

[a] Feil *et al.* (1999).

and interspecific recombinational imports from these species are known to occur in genes under strong selection (Spratt 1994).

Concluding remarks: There is now convincing evidence that recombination in *S. pneumoniae*, *N. meningitidis*, and *E. coli* has more impact than point mutation on sequence diversification at neutral loci. However, it is unwise to generalize from such a limited number of species as the rate of recombination among bacterial populations probably varies greatly (Maynard Smith *et al.* 1993). This results in a range of populations from those with relatively stable clones (*e.g.*, the clone of *Salmonella enterica* that causes typhoid fever; Selander and Musser 1990) to those in which recombination is very common and clones cannot readily be discerned (*e.g.*, *N. gonorrhoeae* or *Helicobacter pylori*; O'Rourke and Stevens 1993; Suerbaum *et al.* 1998). Many species appear to occupy the middle ground, where recombination leads to a higher frequency of nucleotide substitution than point mutation, but is not sufficiently frequent to prevent the emergence of transient clones (Spratt and Maiden 1999).

The analysis presented here demonstrates how MLST data sets can be used to compare the relative importance of recombination compared to point mutation in bacterial species. The great advantage of using MLST data is that databases for new species will become available (*e.g.*, *Staphylococcus aureus* and *Streptococcus pyogenes* MLST databases are now available; http://mlst.zoo.ox.ac.uk), and the existing meningococcal and pneumococcal databases will expand, as the method is increasingly used for molecular typing (Maiden *et al.* 1998; Enright and Spratt 1999a; Spratt 1999), which will allow refinements of the existing estimates. Furthermore, any differences in the parameter within subpopulations of pneumococci and meningococci should eventually be discerned (*e.g.*, between clones with different epidemiological features or between invasive *vs.* carried isolates).

The approach described here, combined with the generation of large MLST data sets for an increasing number of species, should therefore provide a much improved view of the impact of recombination on bacterial population structure. This in turn should provide clues as to which environmental, ecological, or biological factors determine the differences in recombination rates among bacterial species.

We are very grateful to Jiaji Zhou and Tracey Coffey for their assistance in checking the data, to Man-Suen Chan for programming advice and maintenance of the MLST web site, and to Martin Maiden for comments on the manuscript. This work was supported by the Wellcome Trust. B.G.S. is a Wellcome Trust Principal Research Fellow.

LITERATURE CITED

Brunham, R. C., F. A. Plummer and R. S. Stephens, 1993 Bacterial antigenic variation, host immune response, and pathogen-host coevolution. Infect. Immun. **61:** 2273–2276.

Coffey, T. J., C. G. Dowson, M. Daniels, J. Zhou, C. Martin *et al.*, 1991 Horizontal gene transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of *Streptococcus pneumoniae*. Mol. Microbiol. **5:** 2255–2260.

Coffey, T. J., M. C. Enright, M. Daniels, J. K. Morona, R. Morona *et al.*, 1998 Recombinational exchanges at the capsular polysaccharide biosynthetic locus lead to frequent serotype changes among natural isolates of *Streptococcus pneumoniae*. Mol. Microbiol. **27:** 73–83.

Coffey, T. J., M. Daniels, M. C. Enright and B. G. Spratt, 1999 Serotype 14 variants of the Spanish penicillin-resistant serotype 9V clone of *Streptococcus pneumoniae* arose by large recombinational replacements of the *cpsA-pbp1a* region. Microbiology **145:** 2023–2031.

Crook, D. W. M., and B. G. Spratt, 1998 Multidrug resistance in *Streptococcus pneumoniae*. Brit. Med. Bull. **54:** 593–608.

Dykhuizen, D. E., and L. Green, 1991 Recombination in *Escherichia coli* and the definition of biological species. J. Bacteriol. **173:** 7257–7268.

Enright, M. C., and B. G. Spratt, 1998 A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology **144:** 3049–3060.

Enright, M. C., and B. G. Spratt, 1999a Multilocus sequence typing. Trends Microbiol. **7:** 482–487.

Enright, M. C., and B. G. Spratt, 1999b Extensive variation in the *ddl* gene of penicillin-resistant *Streptococcus pneumoniae* results from a hitchhiking effect driven by the penicillin-binding protein 2b gene. Mol. Biol. Evol. **16:** 1687–1695.

Enright, M. C., A. Fenoll, D. Griffiths and B. G. Spratt, 1999 The three major Spanish clones of penicillin-resistant *Streptococcus pneumoniae* are the most common clones recovered from recent cases of meningitis in Spain. J. Clin. Microbiol. **37:** 3210–3216.

Feil, E. J., G. Carpenter and B. G. Spratt, 1995 Electrophoretic variation in adenylate kinase of *Neisseria meningitidis* is due to inter- and intra-species recombination. Proc. Natl. Acad. Sci. USA **92:** 10535–10539.

Feil, E. J., J. Zhou, J. Maynard Smith and B. G. Spratt, 1996 A comparison of the nucleotide sequences of the *adk* and *recA* genes of pathogenic and commensal *Neisseria* species: evidence for extensive inter-species recombination within *adk*. J. Mol. Evol. **43:** 631–640.

Feil, E. J., M. C. J. Maiden, M. Achtman and B. G. Spratt, 1999 The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. Mol. Biol. Evol. **16:** 1496–1502.

Feldman, C., and K. P. Klugman, 1997 Pneumococcal infections. Curr. Opin. Infect. Dis. **10:** 109–115.

Guttman, D. S., 1997 Recombination and clonality in natural populations of *Escherichia coli*. Trends Ecol. Evol. **12:** 16–22.

Guttman, D. S., and D. E. Dykhuizen, 1994 Clonal divergence in *Escherichia coli* as a result of recombination, not mutation. Science **266:** 1380–1383.

Holmes, E. C., R. Urwin and M. C. J. Maiden, 1999 The influence of recombination on the population structure and evolution of the human pathogen *Neisseria meningitidis*. Mol. Biol. Evol. **16:** 741–749.

Istock, C. A., K. E. Duncan, N. Ferguson and X. Zhou, 1992 Sexuality in a natural population of bacteria—*Bacillus subtilis* challenges the clonal paradigm. Mol. Ecol. **1:** 95–103.

Maiden, M. C. J., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell *et al.*, 1998 Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms. Proc. Natl. Acad. Sci. USA **95:** 3140–3145.

Maynard Smith, J., 1999 The detection and measurement of recombination from sequence data. Genetics **153:** 1021–1027.

Maynard Smith, J., N. H. Smith, M. O'Rourke and B. G. Spratt, 1993 How clonal are bacteria? Proc. Natl. Acad. Sci. USA **90:** 4384–4388.

O'Rourke, M., and E. Stevens, 1993 Genetic structure of *Neisseria gonorrhoeae* populations: a non-clonal pathogen. J. Gen. Microbiol. **139:** 2603–2611.

Selander, R. K., and J. M. Musser, 1990 Population genetics of bacterial pathogenesis, pp. 11–36 in *Molecular Basis of Bacterial Infections*, edited by B. H. Iglewski and V. L. Clark. Academic Press, San Diego.

Shi, Z.-Y., M. C. Enright, P. Wilkinson, D. Griffiths and B. G. Spratt, 1998 Identification of three major clones of multiply antibiotic-resistant *Streptococcus pneumoniae* in Taiwanese hospitals by multilocus sequence typing. J. Clin Microbiol. **36:** 3514–3519.

Spratt, B. G., 1994 Resistance to antibiotics mediated by target alterations. Science **264:** 388–393.

Spratt, B. G., 1999 Multilocus sequence typing: molecular typing of bacterial pathogens in an era of rapid DNA sequencing and the Internet. Curr. Opin. Microbiol. **2:** 312–316.

Spratt, B. G., and M. C. J. Maiden, 1999 Bacterial population genetics, evolution and epidemiology. Philos. Trans. R. Soc. Lond. Ser. B **354:** 701–710.

Suerbaum, S., J. Maynard Smith, K. Bapumia, G. Morelli, N. H. Smith *et al.*, 1998 Free recombination within *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA **95:** 12619–12624.

Zhou, J., and B. G. Spratt, 1992 Sequence diversity within the *argF*, *fbp* and *recA* genes of natural isolates of *Neisseria meningitidis*. Interspecies recombination within the *argF* gene. Mol. Microbiol. **6:** 2135–2146.

Zhou, J., L. D. Bowler and B. G. Spratt, 1997 Interspecies recombination, and phylogenetic distortions, within the glutamine synthetase and shikimate dehydrogenase genes of *Neisseria meningitidis* and commensal *Neisseria* species. Mol. Microbiol. **23:** 799–812.

Zhou, J., M. C. Enright and B. G. Spratt, 2000 Identification of the major Spanish clones of penicillin-resistant pneumococci via the Internet using multilocus sequence typing. J. Clin. Microbiol. (in press).

Communicating editor: L. Partridge

APPENDIX: ESTIMATING THE NUMBER OF POINT MUTATIONS

There are 17 cases in which the altered allele in a SLV differs from the ancestral allele by a single nucleotide. Such single-nucleotide changes (SNCs) could be caused by (i) a point mutation; (ii) a whole fragment replacement; or (iii) a part-fragment replacement, altering a single nucleotide, even though the donor and recipient fragments differ at many sites. The problem is to estimate how many of the 17 cases were point mutations. Alleles that differ at a single site from an ancestral allele could be (i) "shared"; that is, identical to an allele in an unrelated isolate; (ii) "shared nucleotide"; that is, with a changed nucleotide identical to one present in the database (but not in a "shared allele"); or (iii) "new"; that is, with a nucleotide not present elsewhere in the database. The observed numbers were 8 "shared," 1 "shared nucleotide," and 8 "new." We start by estimating the proportions of these three types expected from point mutation.

Summing over all ancestral alleles, let

$n1$ represent the number of potential neutral mutations: that is, possible mutations that are neutral, favorable, or, if deleterious, whose effect on fitness is not so large as to prevent their being represented in the database;

$n2$ represent the number of alleles present in the database, in unrelated isolates, that differ at only one site from the ancestral allele; and

$n3$ represent the number of polymorphic sites in the database.

From the data, $n2 = 63$ and $n3 = 992$. We take $n1 = 2637$, the number of third sites. This is likely to be an underestimate, since some of the SNCs are nonsynonymous. However, the estimates below are not sensitive to the value chosen.

Then the proportions of the three types expected from point mutation are

Shared	Shared nucleotide	New
$n2/n1$	$(n3 - n2)/n1$	$(n1 - n3)/n1$
0.02	0.35	0.62

If, as an initial guess, we assume all new SNCs to have been produced by point mutation, the expected values are 0.26 shared, 4.5 shared nucleotide, and 8 new. Note that the expected number of shared nucleotide SNCs is greater than the observed number of 1.

Now consider the possibility that some new SNCs were produced by part-fragment replacements. These could only be generated by replacements from outside the database. Such replacements would also generate shared nucleotide SNCs. However, 66 out of 81 multiple nucleotide replacements, or 81%, were by alleles present elsewhere in the database. There is no reason why the proportion of part-fragment replacements from within the database should not be similar. Such replacements would necessarily be shared nucleotide SNCs, as would some fraction of replacements from outside the database. Hence part-fragment replacements are more likely to produce shared nucleotide than new SNCs. However, as pointed out above, the number of shared nucleotide SNCs expected (assuming most new SNCs to be caused by mutation) is already greater than the number observed. It follows that part-fragment replacements from outside the database have been rare or absent; they are ignored in what follows.

There remains the possibility that SNCs have been generated by whole-fragment replacements. Approximately 81% of these will be from within the database, and all these will be shared SNCs. The question is, however, whether the remaining 19% from outside the database contribute a significant number of new SNCs. This is unlikely. The probability, p_u, that a whole-fragment replacement from outside the database carries a unique nucleotide can be estimated by the proportion of the $n2$ potential SNCs within the database that carry a nucleotide not present elsewhere; that is, $p_u = 33/63$. In other words, the 19% of whole-fragment replacements will generate 50% new and 50% shared nucleotide SNCs. Hence the proportions of the three types generated by whole-gene replacements are

Shared	Shared nucleotide	New
0.81	0.095	0.095

Given the expected proportions generated by point mutation, and by whole fragment replacement, and the observed numbers of eight shared and eight new SNCs, the best fit to observation is

	Shared	Shared nucleotide	New	Total
Point mutation	0.2	1 (3.6)	7	8.2
Whole fragment replacement	7.8	0 (0.9)	1	8.8
Total observed	8	1	8	17

The numbers in parentheses are the number of shared nucleotide SNCs expected, given the observed numbers of shared and new SNCs. The observed number of 1 is lower than the expected number of 4.5. The difference is barely significant ($P = 0.05$), but it is suggestive; a contributory factor may be that we have underestimated *n*1, the number of potential neutral mutations. It is worth noting that there is no sign of hypermutable sites; if such sites existed, we would expect to see more and not fewer shared nucleotide SNCs.

These calculations are unavoidably very approximate. However, they suggest that (i) part-fragment replacements are rare or absent; and (ii) most new and shared nucleotide SNCs are the result of point mutation, and most shared SNCs are the result of whole-fragment replacements. If so, a reasonable estimate of the number of point mutations is nine.