JOURNAL OF BACTERIOLOGY, Sept. 2005, p. 6223–6230 Vol. 187, No. 17
0021-9193/05/$08.00+0 doi:10.1128/JB.187.17.6223–6230.2005

Using Multilocus Sequence Data To Define the Pneumococcus

William P. Hanage,[1]* Tarja Kaijalainen,[2] Elja Herva,[2] Annika Saukkoriipi,[2] Ritva Syrjänen,[3] and Brian G. Spratt[1]

Department of Infectious Disease Epidemiology, St. Mary's Hospital, Imperial College London, London W2 1PG, United Kingdom[1]; Department of Microbiology, National Public Health Institute, PL 310, 90101 Oulu, Finland[2]; and Department of Vaccines, National Public Health Institute (KTL), Mannerheimintie 166, 00100 Helsinki, Finland[3]

Received 18 March 2005/Accepted 2 June 2005

We investigated the genetic relationships between serotypeable pneumococci and nonserotypeable presumptive pneumococci using multilocus sequence typing (MLST) and partial sequencing of the pneumolysin gene (*ply*). Among 121 nonserotypeable presumptive pneumococci from Finland, we identified isolates of three classes: those with sequence types (STs) identical to those of serotypeable pneumococci, suggesting authentic pneumococci in which capsular expression had been downregulated or lost; isolates that clustered among serotypeable pneumococci on a tree based on the concatenated sequences of the MLST loci but which had STs that differed from those of serotypeable pneumococci in the MLST database; and a more diverse collection of isolates that did not cluster with serotypeable pneumococci. The latter isolates typically had sequences at all seven MLST loci that were 5 to 10% divergent from those of authentic pneumococci and also had distinct and divergent *ply* alleles. These isolates are proposed to be distinct from pneumococci but cannot be resolved from them by optochin susceptibility, bile solubility, or the presence of the *ply* gene. Complete resolution of pneumococci from the related but distinct population is problematic, as recombination between them was evident, and a few isolates of each population possessed alleles at one or occasionally more MLST loci from the other population. However, a tree based on the concatenated sequences of the MLST loci in most cases unambiguously distinguished whether a nonserotypeable isolate was or was not a pneumococcus, and the sequence of the *ply* gene fragment was found to be useful to resolve difficult cases.

Assigning bacteria to discrete populations, or species, can be problematic, since bacteria differ greatly in the extent and promiscuity of recombination (8). In some bacteria, homologous recombination appears to be restricted because of either inefficient mechanisms or vectors of genetic exchange or ecological factors that limit the extent to which genetically distinct strains meet each other in nature. In other bacteria, recombination is very frequent, and in some cases, homologous recombination occurs between bacteria that differ substantially (20 to 25%) in sequence and that are assigned to different but closely related named species. Attempts to define species typically determine the relationships among a set of isolates that are considered to represent each named species, using DNA-DNA hybridization or the sequences of rRNA or conserved genes.

Homologous recombination distorts the true relationships between isolates of closely related named species and can lead to inconsistent relationships among those species inferred from the sequences of different genes (6, 9). Consequently, defining species using single loci is inappropriate, particularly for those species where rates of recombination are high. The use of multilocus sequence-based approaches ensures that recombination at one locus is buffered by the more reliable indications of relatedness provided by the other loci. Furthermore, in defining any species, we must analyze populations of each candidate species and not just one or a few isolates (9).

* Corresponding author. Mailing address: Department of Infectious Disease Epidemiology, St. Mary's Hospital, Imperial College London, Norfolk Place, London W2 1PG, United Kingdom. Phone: (020) 75943622. Fax: (020) 75943693. E-mail: w.hanage@imperial.ac.uk.

Two general types of multilocus approaches can be considered as tools to distinguish related species. Microarrays can detect differences in gene repertoire among isolates (15) but suffer the serious disadvantage that genes that differ among isolates of a named species, or between related species, reflect the least stable part of the genome, rather than the core genome, which is likely to be the most phylogenetically informative (12). The alternative approach is to use the sequences of multiple housekeeping genes that are part of the core genome (22). These data are now widely available as isolates of many pathogens are characterized by sequencing internal fragments of seven housekeeping loci, a technique referred to as multilocus sequence typing (MLST) (13).

Previously, we have shown that MLST-based approaches are capable of discriminating named species among the human *Neisseria* (9). In this paper, we test the utility of this approach in defining the boundaries of the named species *Streptococcus pneumoniae* (the pneumococcus) and the extent to which it can be resolved from closely related isolates of uncertain taxonomic status that cocolonize the human nasopharynx. The confident identification of these members of the mitis group of alpha-hemolytic streptococci is fraught with difficulty. At present, presumptive members of the species *S. pneumoniae* are usually identified in clinical microbiology laboratories by colonial morphology when grown on blood agar and by optochin sensitivity. In the case of optochin-insensitive isolates that otherwise appear similar to pneumococci, bile solubility may also be used. Pneumococci are further characterized by serotyping using the Quellung reaction, and isolates that can be assigned to one of the 90 recognized pneumococcal sero-

types are considered unambiguously to be *S. pneumoniae*. After applying these tests, a number of nonserotypeable isolates that appear to be similar to pneumococci but are of uncertain taxonomic status remain.

The relationship between nontypeable pneumococcus-like isolates and genuine pneumococci has been considered by Whatmore et al. (23). Those authors point out the difficulty in resolving these issues by gene content. Genes previously thought to be limited to the pneumococcus, including those encoding the virulence factors pneumolysin (*ply*) and autolysin, have been found in isolates assigned to other named species such as *S. oralis* and *S. mitis* (14, 23).

This work considers the potential of MLST in defining the pneumococcus and in distinguishing it from closely related species. We used MLST to characterize a set of 121 nonserotypeable presumptive pneumococci from Finland and compared the sequences of a fragment of the *ply* gene with those of authentic serotypeable pneumococci. The sequence data clearly identify the nontypeable isolates as members of either the pneumococcal population or a clearly resolved and more diverse related population. We propose that these latter isolates should be recognized as distinct from pneumococci and demonstrate the utility of the sequence of a fragment of the pneumolysin gene for distinguishing nonserotypeable pneumococci from this related nonpneumococcal population.

MATERIALS AND METHODS

Bacterial isolates. A reference set of 39 serotypeable pneumococci was chosen to define the diversity found within the pneumococcus. To construct this data set, the entire MLST database (http://spneumoniae.mlst.net/) was divided into nonoverlapping groups of related isolates using the program eBURST (http://eburst.mlst.net/) (7), with the default setting for the group definition (six out of seven shared loci). The reference set includes examples of the founding genotypes of the major clonal complexes identified by eBURST and also isolates of some of the major internationally disseminated antibiotic-resistant clones.

For comparison with the pneumococcal reference set described above, 121 isolates of nonserotypeable presumptive pneumococci were obtained from the Finnish otitis media studies conducted in Finland to investigate pneumococcal disease and carriage (5, 10, 11). Details of these isolates are summarized in Table 1. Presumptive pneumococci were identified by colony morphology and sensitivity to optochin (6 μg; Biodisk PDM Diagnostic Disks, Sweden). All isolates discussed here were optochin sensitive in the first testing, but a few isolates were optochin resistant in later testing, as indicated in Table 1. All isolates were, on first inspection, not serotypeable. In some cases, however, subsequent testing revealed reactions with omniserum, and in these cases, serotype was determined by the Quellung reaction. Isolates were obtained from either middle ear fluid (MEF) of children with acute otitis media (AOM) (11) or nasopharyngeal (NP) swabs of healthy children or children with AOM (20). Two isolates were excluded during analysis (IOPR 4609 and IOPR 3386), as they failed to yield good-quality sequence at any MLST locus despite repeated attempts.

DNA isolation and sequencing. Genomic DNA was isolated using Qiamp DNA Mini kits (QIAGEN). Internal fragments of MLST loci were PCR amplified (*Taq* polymerase and 10× buffer; QIAGEN) with 50 nM deoxynucleoside triphosphates (Geneamp; Applied Biosystems, Foster City, Calif.) using the primers and PCR conditions described previously (4). The PCR products were precipitated with 20% polyethylene glycol 8000–2.5 M NaCl (Sigma, St. Louis, Mo.), and the fragments were sequenced on both strands using the same primers and Big Dye II terminators (Applied Biosystems). The products of the sequencing reactions were precipitated with 185 mM sodium acetate in 70% ethanol and were resuspended in 10 μl HiDi formamide (Applied Biosystems) and loaded onto an ABI Prism 3700 sequencer. Sequences were analyzed using STARS (obtainable from http://www.mlst.net), a modified Staden interface developed by Man-Suen Chan for use with MLST projects. Alleles were assigned by comparing the sequences to those in the pneumococcal MLST database, and those already present were assigned the relevant allele number. For those not found in the pneumococcal database, each unique allele from nontypeable isolates was initially given an alphabetic identifier to distinguish it from those alleles found in serotypeable pneumococci. All new alleles were sequenced at least twice on each DNA strand. The pneumolysin gene fragment was amplified and sequenced on both strands using the primers Ia and Ib as described previously by Toikka et al. (21) and the same conditions for PCR and sequencing used for the MLST loci. Sequences were trimmed to 282 bp in length, and each distinct sequence was assigned as a different *ply* allele. All disagreements between strands were resolved by manual inspection of trace files, and all *ply* alleles were verified by sequencing at least twice on each DNA strand.

Phylogenetics and population genetics. To illustrate differences between individual alleles at the MLST and *ply* loci, minimum evolution trees were constructed using all nucleotide differences and the Kimura 2 parameter distance correction in MEGA 2.1. The sequences of all loci except *ddl* (see below) were concatenated, maintaining the +1 reading frame, and trees were constructed from the concatenated 2,751-bp sequence using MrBayes 3.0b4 (16). A starting neighbor-joining tree was determined in PAUP*4.0beta v.10 (http://paup.csit.fsu.edu/) (19), with distances corrected using the HKY85 model. This was input as a starting tree into MrBayes, four Markov chain Monte Carlo chains were run with default heating parameters until convergence, and 10,000 trees were sampled from the posterior probability distribution. These were then used to produce a consensus tree. The choice of evolutionary model for MrBayes was made using MrModeltest 2.2 (http://www.ebc.uu.se/systzoo/staff/nylander.html) and corresponded to the general reversible model with rates of substitution being gamma distributed between sites, a proportion of which were invariant. Nucleotide diversities were estimated using DNAsp (17). Other population genetic analyses were performed using Arlequin v2.0 (http://lgb.unige.ch/arlequin/).

RESULTS

Clustering of multilocus genotypes defines a monophyletic group containing all serotypeable pneumococcal isolates. The sequences of the seven MLST loci from the pneumococcal reference set, and 121 nontypeable isolates of uncertain status, were determined. The combinations of alleles (i.e., the allelic profiles) for all isolates studied in this work are shown in Table 1. Figure 1 shows a phylogenetic tree constructed from concatenated sequences of the alleles at all loci except *ddl* (see below). The tree was constructed in MrBayes 3.04b using all nucleotide sites. All 39 strains of the pneumococcal reference set are descended from a single node and group with several of the nontypeable isolates, most of which were found to have sequence types (STs) already present in the MLST database. Support for this node was high (99%), and we designate isolates descended from it as group 1 in Fig. 1. A small, well-supported (99%) cluster is evident within this group (1b in Fig. 1). This includes the three most common STs found among the nontypeable isolates (STs 449, 448, and 344). Isolates of each of these STs have previously been submitted to the MLST database from other locations and were found to be nontypeable in all cases. The remaining two STs in this cluster were minor variants of these previously recognized STs.

The remainder of the nontypeable isolates form a highly diverse group distinct from group 1. The mean genetic distance (computed using all nucleotide positions with DNAsp) within this group of divergent nontypeable isolates is greater (86.19 nucleotide differences) than that within group 1 (28.28 nucleotide differences). The mean distance between group 1 and the other isolates was 146.78. Wright's F_{ST} (24), the proportion of genetic variation distributed among subpopulations relative to the total population (computed using Arlequin v.2.0), was 0.61, indicating significant differentiation between the two groups.

Relationships of alleles in nontypeable isolates to those found in the MLST database. The nontypeable isolates that clustered within group 1 were considered to be authentic pneu-

TABLE 1. Allelic profiles and STs of strains used in this work

Strain ID	Allele assignment									Serotype[b]	Optochin sensitivity[c]	Site of isolation[d]
	aroE	*gdh*	*gki*	*recP*	*spi*	*xpt*	*ddl*	ST[a]	*ply*[a]			
IOPR 5877	1	5	4	5	5	3	8	15	1		+	NP
IOKOR 818	5	12	29	12	9	39	18	100	1	33	+	NP
IOPR 1128	5	12	29	12	9	39	18	100	1	33	+	MEF
IOPR 1295	5	12	29	12	9	39	18	100	1	33	+	NP
IOPR 3524	5	12	29	12	9	39	18	100	1	33	+	NP
IOPR 5853	5	12	29	12	9	39	18	100	1	33	+	NP
IOPR 868	5	12	29	12	9	39	18	100	1	33	+	NP
IOPR 1387	7	5	8	5	10	6	14	138	2		+	NP
IOPR 3071	7	11	10	1	6	8	14	162	3	9V	+	NP
IOPR 3942	7	11	10	1	6	8	14	162	3	9V	+	NP
IOPR 3399	8	13	14	4	17	4	14	199	1		+	NP
IOPR 5878	15	8	8	18	15	1	31	235	7	20	+	NP
IOPR 1329	8	37	9	29	2	12	53	344	1		+	NP
IOPR 1734	8	37	9	29	2	12	53	344	1		+	MEF
IOPR 1746	8	37	9	29	2	12	53	344	1		+	NP
IOPR 2052	8	37	9	29	2	12	53	344	1		+	NP
IOPR 2257	8	37	9	29	2	12	53	344	1		+	NP
IOPR 2302	8	37	9	29	2	12	53	344	1		+	MEF
IOPR 4124	8	37	9	29	2	12	53	344	1		+	NP
IOPR 4125	8	37	9	29	2	12	53	344	1		+	NP
IOPR 4169	8	37	9	29	2	12	53	344	1		+	MEF
IOPR 4573	8	37	9	29	2	12	53	344	1		+	NP
IOPR 4848	8	37	9	29	2	12	53	344	1		+	NP
IOPR 6449	8	37	9	29	2	12	53	344	1		+	MEF
IOKOR 1051	8	5	2	27	2	11	71	448	4		+	NP
IOKOR 1858	8	5	2	27	2	11	71	448	4		+	NP
IOKOR 294	8	5	2	27	2	11	71	448	4		+	MEF
IOKOR 543	8	5	2	27	2	11	71	448	4		+	MEF
IOKOR 768	8	5	2	27	2	11	71	448	4		+	NP
IOPR 129	8	5	2	27	2	11	71	448	4		+	MEF
IOPR 1422	8	5	2	27	2	11	71	448	4		+	NP
IOPR 2740	8	5	2	27	2	11	71	448	4		+	MEF
IOPR 2851	8	5	2	27	2	11	71	448	4		+	MEF
IOPR 3595	8	5	2	27	2	11	71	448	4		+	NP
IOPR 48	8	5	2	27	2	11	71	448	4		+	MEF
IOPR 5580	8	5	2	27	2	11	71	448	4		+	MEF
IOPR 5944	8	5	2	27	2	11	71	448	4		+	NP
IOPR 5966	8	5	2	27	2	11	71	448	4		+	NP
IOPR 6073	8	5	2	27	2	11	71	448	4		+	MEF
IOPR 1496	8	37	9	29	2	47	5	449	1		+	MEF
IOPR 1586	8	37	9	29	2	47	5	449	1		+	NP
IOPR 1793	8	37	9	29	2	47	5	449	1		+	NP
IOPR 2368	8	37	9	29	2	47	5	449	4		+	NP
IOPR 2687	8	37	9	29	2	47	5	449	1		+	NP
IOPR 2866	8	37	9	29	2	47	5	449	1		+	NP
IOPR 3014	8	37	9	29	2	47	5	449	1		+	NP
IOPR 3065	8	37	9	29	2	47	5	449	1		+	NP
IOPR 3223	8	37	9	29	2	47	5	449	1		+	NP
IOPR 3933	8	37	9	29	2	47	5	449	1		+	MEF
IOPR 4131	8	37	9	29	2	47	5	449	1		+	NP
IOPR 4184	8	37	9	29	2	47	5	449	1		+	NP
IOPR 4293	8	37	9	29	2	47	5	449	1		+	NP
IOPR 4847	8	37	9	29	2	47	5	449	1		+	NP
IOPR 5027	8	37	9	29	2	47	5	449	1		+	NP
IOPR 5599	8	37	9	29	2	47	5	449	1		+	NP
IOPR 5745	8	37	9	29	2	47	5	449	1		+	NP
IOPR 5844	8	37	9	29	2	47	5	449	1		+	NP
IOKOR 1054	7	5	1	1	6	31	9	492	1	6B	+	NP
IOKOR 492	13	8	65	1	60	16	6	508	8		+	NP
IOKOR 609	13	8	65	1	60	16	6	508	8		+	NP
IOKOR 707	13	8	65	1	60	16	6	508	8		+	NP
IOKOR 314	1	1	4	1	18	16	17	520	1	22F	+	NP
IOKOR 328	1	1	4	1	18	16	17	520	1	22F	+	MEF
IOKOR 345	1	1	4	1	18	16	17	520	1	22F	+	MEF
IOPR 4036	1	1	4	1	18	16	17	520	1	22F	+	NP
IOPR 5268	8	37	9	29	2	47	59	1054	1		+	NP

Continued on following page

TABLE 1—*Continued*

Strain ID	Allele assignment									Serotype[b]	Optochin sensitivity[c]	Site of isolation[d]
	aroE	*gdh*	*gki*	*recP*	*spi*	*xpt*	*ddl*	ST[a]	*ply*[a]			
IOPR 2619	7	5	8	5	87	6	14	1055	2		+	NP
IOPR 4849	8	5	2	27	2	136	71	1229	4		+	MEF
IOPR 2837	2	5	29	16	42	3	146	1248	1	33	+	NP
IOPR 193	8	5	2	27	9	11	71	1290	4		+	MEF
IOKOR 22	A	A	B	B	B	H	B	NT1	NT1		+	NP
IOPR 1711	A	E	C	29	F	C	E	NT2	NT5		+	NP
IOKOR 98	A	E	E	D	D	F	B	NT3	NT7		+	NP
IOKOR 213	A	F	F	F	E	B	B	NT4	NT1		+	NP
IOPR 418	A	G	C	E	B	H	B	NT5	NT1		−	NP
IOPR 870	A	H	G	H	F	G	B	NT6	NT1		−	NP
IOPR 2017	A	J	D	H	B	R	G	NT7	NT4		+	NP
IOPR 5427	A	L	C	M	E	B	B	NT8	NT1		+	NP
IOKOR 898	A	O	C	E	F	B	G	NT9	NT9		+	NP
IOPR 6117	B	40	A	N	I	A	G	NT10	NT3		+	NP
IOKOR 50	A	B	C	C	C	H	B	NT11	NT11		+	NP
IOKOR 534	B	57	G	P	I	M	B	NT12	NT2		+	NP
IOKOR 56	B	C	A	A	A	Q	A	NT13	NT1		+	NP
IOKOR 8	B	C	A	A	A	Q	A	NT13	NT1		+	NP
IOPR 1791	B	J	H	J	F	D	F	NT14	NT1		+	NP
IOKOR 809	B	N	J	Q	J	L	B	NT15	NT8		+	NP
IOKOR 332	C	D	F	29	I	B	49	NT17	NT1		+	NP
IOKOR 731	C	F	D	R	K	C	H	NT18	NT6		+	MEF
IOKOR 220	C	G	B	G	F	J	B	NT19	NT1		+	NP
IOKOR 226	A	D	B	H	F	B	B	NT20	NT1		+	NP
IOKOR 290	A	D	B	H	F	B	B	NT20	NT1		+	NP
IOPR 1386	C	I	C	G	G	E	B	NT21	NT1		−	NP
IOKOR 254	D	57	D	I	E	I	D	NT22	NT2		+	NP
IOPR 2717	E	I	C	G	G	E	B	NT23	NT1		+	NP
IOPR 923	E	I	C	G	G	E	B	NT23	NT1		+	NP
IOKOR 106	A	D	C	E	B	H	C	NT24	NT1		+	NP
IOPR 5370	A	D	C	L	F	C	E	NT25	NT1		+	NP
IOKOR 484	5	M	J	O	6	G	B	NT26	NT5		+	NP
IOPR 3458	A	D	C	DD	B	H	B	NT27	NT1		+	NP
IOPR 2716	A	D	C	E	B	H	O	NT28	NT1		+	NP
IOKOR 769	A	D	D	B	B	H	B	NT29	NT1		+	NP
IOPR 3101	A	E	B	D	O	F	B	NT30	NT7		+	NP
IOPR 3074	A	I	H	W	E	B	L	NT31	NT10		−	NP
IOKOR 78	A	D	D	B	B	K	B	NT32	NT1		+	NP
IOPR 5052	A	O	P	B	F	B	G	NT33	NT9		+	NP
IOPR 2640	A	Y	4	1	M	B	B	NT34	Mixed		+	NP
IOKOR 1148	B	R	A	J	I	K	E	NT35	NT3		+	NP
IOKOR 1674	B	R	A	N	I	U	P	NT36	NT3		+	NP
IOPR 3582	B	R	A	N	I	U	P	NT36	NT3		+	NP
IOKOR 1096	A	D	C	CC	F	C	E	NT37	NT1		+	NP
IOKOR 937	A	D	C	CC	F	C	E	NT37	NT1		+	NP
IOKOR 545	A	D	C	B	M	B	B	NT38	NT12		+	NP
IOKOR 675	A	D	C	B	M	B	B	NT38	NT12		+	NP
IOKOR 869	A	D	C	B	B	H	B	NT39	NT13		+	NP
IOPR 1260	A	E	B	D	D	F	B	NT40	NT7		+	NP
IOKOR 1362	A	E	B	D	D	1	B	NT41	NT7		+	MEF
IOKOR 485	C	S	L	G	F	ND	B	ND	NT1		+	NP
IOKOR 810	C	S	L	G	F	ND	B	ND	NT1		+	NP
IOPR 2148	ND	K	I	K	X	R	K	ND	ND		−	NP

[a] NT is used as a prefix to distinguish between true pneumococci and the related isolates shown in Fig. 1 (groups 1 and 2). ND indicates that it was not possible to obtain a sequence for this locus from this isolate. One isolate gave a mixed sequence for *ply*.
[b] All isolates were nontypeable on first inspection, but where a serotype was found on subsequent analysis, this is indicated in parentheses.
[c] + indicates that the isolate was optochin sensitive; − indicates that it was resistant.
[d] NP, retrieved from nasopharyngeal carriage; MEF, obtained from middle ear fluid of children suffering from otitis media.

mococci that failed to express a capsule, and any novel alleles in these isolates were assigned allele numbers and added to the pneumococcal MLST database. For those nontypeable isolates that did not fall into group 1, each unique sequence was given an alphabetic allele identifier and retained in a separate database to prevent confusion with the pneumococcal alleles in the MLST database. Minimum evolution trees were constructed for each locus from the sequences of all known pneumococcal alleles in the MLST database together with the alleles from all nontypeable isolates (Fig. 2). In the case of the *aroE* locus, the sequences cluster into two groups that are highly divergent from one another but with relatively little diversity within each

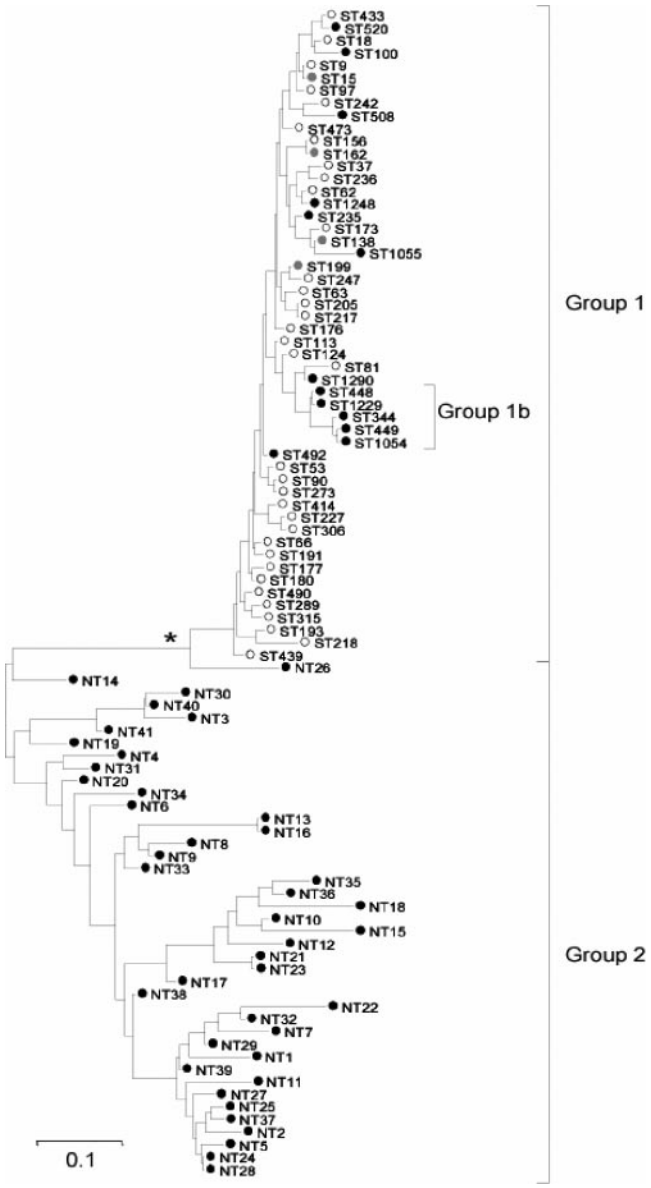
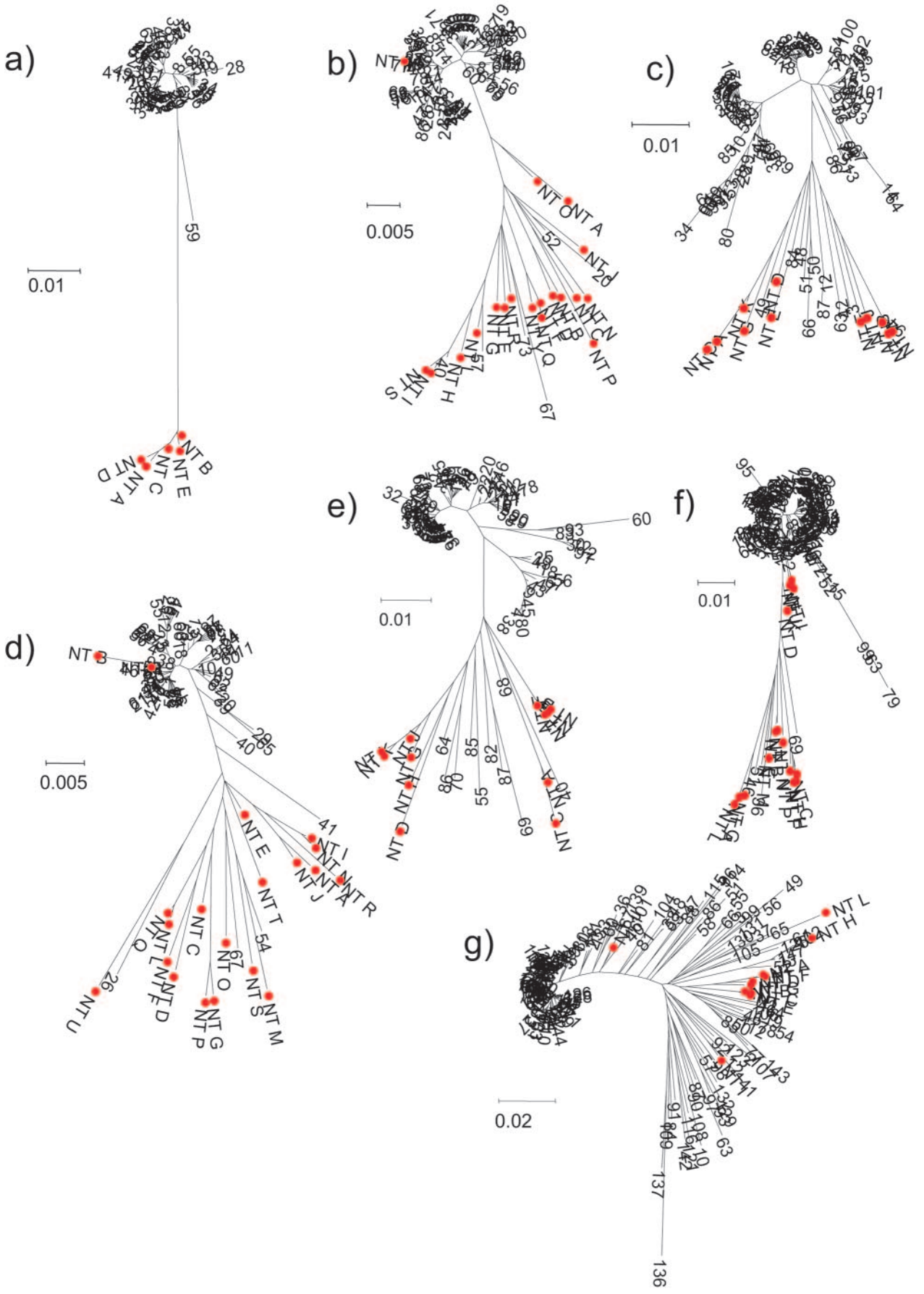


FIG. 1. Phylogenetic tree from concatenated sequences of MLST loci excepting *ddl*. The tree shows the relationships between the pneumococcal reference set (indicated by open circles) and nontypeable isolates (closed circles). Nontypeable isolates that had the same ST as members of the reference set are shown in gray. Group 1 contains the reference set of pneumococci and all nontypeable strains clustering with them. Group 1b is defined in the text. Group 2 contains the remaining nontypeable isolates, indicated with the prefix NT in Table 1. Trees were generated using MrBayes as described in Materials and Methods, using all nucleotide sites. The scale bar indicates substitutions per site. The asterisk marks the node that is considered to separate authentic pneumococci from group 2 isolates.

group. The mean diversity of *aroE* alleles in the MLST database is 2.4 nucleotide differences, in comparison to a mean of 4.1 nucleotide differences within the *aroE* alleles identified in group 2 strains (significantly greater [Student's *t* test; $P = 0.006$]). This is the situation for all other loci with the exception of *ddl*, with nucleotide diversity in group 2 strains being significantly greater among the alphabetic alleles (Student's *t* test; $P \ll 0.05$ for each) than those from typical pneumococci, as is apparent from the trees shown in Fig. 2, and diverging by >5% from the latter group of alleles. However, certain alleles present in serotypeable pneumococci in the MLST database clearly cluster with the alphabetic alleles (e.g., *recP* allele 26) and almost certainly represent instances of lateral transfer between the groups (presumably importation of alleles into pneumococci). The bidirectional nature of this is evident in Table 1 through the finding of alleles which are present, indeed common, in authentic pneumococci among the nontypeable isolates that cluster outside group 1.

The *ddl* gene is located close to the penicillin-binding protein 2b (*pbp2b*) gene, and penicillin-resistant pneumococci often have highly divergent *ddl* alleles, since the DNA fragment carrying the *pbp2b* gene that is imported from related species during the emergence of penicillin resistance frequently includes all or part of the flanking *ddl* gene (3). The tree derived from *ddl* sequences of authentic pneumococci therefore shows a cluster of similar alleles and a spectrum of increasingly divergent alleles due to the importation of divergent alleles and the generation of mosaic alleles where the recombinational junction between imported divergent sequence and the resident pneumococcal sequence is within the *ddl* gene. Unlike the other loci, the levels of diversity among *ddl* alleles within the pneumococcal MLST database were not significantly different from those found in group 2 strains (mean of 25.4 nucleotide differences in comparison with 24.9 nucleotide differences [Student's *t* test; $P > 0.05$]). As the locus is known to be subject to hitchhiking (3), *ddl* was excluded from the concatenation procedure described above.

Pneumolysin gene sequence. The *ply* gene, once considered to be a defining property of the pneumococcus, has recently been demonstrated to be present in isolates of related species (14, 23). We therefore sequenced a 282-bp fragment of the *ply* locus from all nontypeable isolates and from the pneumococcal reference set. The *ply* gene was found to be present in all but one isolate, which was highly divergent at the MLST loci (IOPR 2148 [Table 1]). Another isolate, IOPR 2640 (NT34), appeared to contain more than one *ply* sequence; sequencing of the amplified fragment from several DNA preparations from purified single colonies gave a mixed sequence, although this was not observed with the MLST loci. Interestingly, some of these were typical pneumococcal alleles, while others were divergent, suggesting that this strain has a history of interspecific recombination. Each unique *ply* sequence was assigned as a different allele following the same conventions described above for MLST genes (i.e., integers for alleles of isolates in group 1 and alphabetic identifiers for the remainder). A minimum evolution tree showing the relationships between the *ply* sequences is presented in Fig. 3, and the *ply* alleles assigned to individual isolates are shown in Table 1. Alleles from the reference pneumococcal set again form a distinct cluster, and all of the nontypeable isolates that fall into group 1 in Fig. 1 had *ply* sequences that either clustered with or were identical to the *ply* alleles from the pneumococcal reference set. The other nontypeable isolates had *ply* alleles that were distinct from those of the reference pneumococci and which clustered apart from them on the tree.



DISCUSSION

The identification of clearly defined groups of isolates that are similar to each other in genotype, but which are clearly distinguished from other related groups of similar genotypes and which may be assigned as species, is a central problem in microbiology (18). The species concept is particularly problematic for bacteria, and it may be unrealistic to expect to have a single satisfying concept of species that can encompass bacteria that are almost totally asexual through to ones in which localized homologous recombinational replacements are very frequent and may be relatively promiscuous. In this paper, we have examined one of the more highly recombinogenic species, the pneumococcus, and have attempted to discriminate it from isolates that are closely related and which colonize the same niche, the human nasopharynx, and to explore the extent to which these very similar bacteria can be resolved into distinct populations.

Single genes are clearly unsatisfactory for exploring these issues, and we have therefore used a multilocus approach. We have previously applied this approach to the human *Neisseria* species and found it to be capable of discriminating named species even in the presence of recombination (9). Here, we consider the related but distinct question of whether we can define the boundaries of a named species and distinguish it from other related species which may not currently be designated as such. Nontypeable presumptive pneumococci provide a useful source of isolates that are closely related to pneumococci and which are known to include authentic pneumococci that for various reasons may not express a capsule, in addition to isolates that are genetically distinct. Trees based on the concatenated sequences of the MLST loci (excluding *ddl*) clearly resolved the nontypeable presumptive pneumococci into two groups. Approximately 58% of the isolates clustered with serotypeable pneumococci, whereas the others were a separate and more diverse population. Based on these results, we propose that *S. pneumoniae* may be defined as all isolates falling into group 1 on a tree based on the concatenated sequences of the six MLST loci. The pneumococcal MLST website now contains a function that allows users to test whether isolates sequenced in their laboratory should be considered true pneumococci under this definition (1).

The presence of the pneumolysin gene in isolates closely related to pneumococci has been previously demonstrated (14, 23) and precludes using the presence of this gene as a means of distinguishing pneumococci from their closest relatives. However, the *ply* alleles in pneumococci were different from those in isolates that grouped outside the pneumococci in the tree constructed using concatenated MLST sequences. Precisely the same groups were obtained using the sequence of the pneumolysin gene fragment or the concatenated MLST loci. The *ply* sequences therefore provide a further means of identifying true pneumococci in difficult cases, although rigorous assignment of a nontypeable isolate as a pneumococcus should involve examination of the clustering obtained with both the concatenated MLST loci and the *ply* gene fragment.

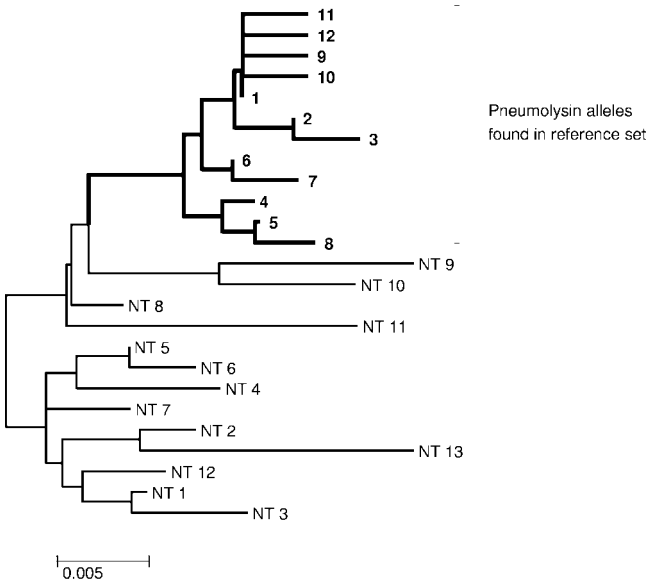


FIG. 3. Minimum evolution tree for the *ply* alleles. All alleles found in the pneumococcal reference set or other group 1 isolates (shown in boldface) were found to descend from a single node. The remainder were alleles from nontypeable isolates that clustered apart from group 1 in Fig. 1. The tree was generated using MEGA 2.1. All nucleotide differences were used in the analysis, and distances were corrected using the K2P model. The scale bar indicates substitutions per site.

Those isolates not identified as pneumococci by this approach are a much more highly diverse grouping than the pneumococci, as demonstrated by comparing the mean genetic distance within the two groups. This is even more striking if you consider that group 1 contains strains from a reference set specifically chosen to illustrate the diversity of the pneumococcal population, whereas the atypical isolates reported here were retrieved in longitudinal studies of carriage and AOM within a limited geographic area and are therefore unlikely to represent the full diversity of this population. The relationship of these organisms to the recently proposed species *S. pseudopneumoniae* (2) remains to be determined. It should be noted that two of the strains falling outside group 1 (IOKOR 731 and IOKOR 1362) were isolated from MEF in children suffering from AOM, suggesting that organisms of this group may harbor pathogenic potential in some disease contexts.

The approach described here clearly delineates the boundaries of the pneumococcal cluster in sequence space and, thanks to its multilocus nature, is resistant to limited shuffling of genetic information across this boundary. Unlike previous attempts to define bacterial species using conceptually similar

FIG. 2. Minimum evolution trees for the seven MLST loci. Trees were constructed using the sequences of all alleles from the pneumococcal MLST database and those from the nontypeable isolates. The latter are indicated by red markers. (a) *aroE*; (b) *gdh*; (c) *gki*; (d) *recP*; (e) *spi*; (f) *xpt*; (g) *ddl*. Trees were generated using MEGA 2.1. All nucleotide differences were used in the analysis, and distances were corrected using the K2P model. The scale bar indicates substitutions per site.

approaches (22), we tested the ability of our method to discriminate between the pneumococcus as a population and a large group of very closely related isolates that are indistinguishable by other methods. We are also impressed by the ability of this approach to resolve the pneumococcus with confidence, even in the presence of relatively high levels of recombination. However, it is likely that the pneumococcus is an example of a "fuzzy species," and further sampling of the fringes of the pneumococcal cluster may require us to update our definitions. This issue can only be resolved by further work. It should be noted that recombination means that these trees contain no useful information about the relationships within group 1 and group 2, but this is insufficient to obscure the differences between them. It remains to be seen if it is possible to define combinations of phenotypic characteristics shared by all members of group 1 that are not found in any members of the diverse group of related organisms this work has shown clustering apart from genuine pneumococci. The recombinogenic nature of these organisms may mean that attempts to do so are misguided.

This work raises several questions about the nature of the mechanism which generates and maintains these divisions. One possibility is effective reproductive isolation, in which strains mainly undergo recombination with isolates of the same named species. While interspecific recombination does occur, and renders a single-locus approach untenable, it is not common enough to prevent the emergence of those clusters in sequence space we refer to as species. In the case of clonal bacteria, the virtual absence of recombination will necessarily lead to clusters of related strains. What we do not know is what generates such effective isolation. We are also ignorant of to what degree recombination must be limited in order to achieve effective isolation and consequent speciation. To resolve these issues, further studies that combine theoretical and molecular approaches are required.

ACKNOWLEDGMENTS

This work was supported by a grant from the Wellcome Trust (grant number 030662) B.G.S. is a Wellcome Trust Principal Research Fellow. The Finnish otitis media studies were supported by Merck & Co., Aventis Pasteur, and Wyeth-Lederle Vaccines and Pediatrics.

REFERENCES

1. **Aanensen, D. M., and B. G. Spratt.** 2005. The Multilocus Sequence Typing network: mlst.net. Nucleic Acids Res. **33:**W728–W733.
2. **Arbique, J. C., C. Poyart, P. Trieu-Cuot, G. Quesne, G. C. Mda, A. G. Steigerwalt, R. E. Morey, D. Jackson, R. J. Davidson, and R. R. Facklam.** 2004. Accuracy of phenotypic and genotypic testing for identification of *Streptococcus pneumoniae* and description of *Streptococcus pseudopneumoniae* sp. nov. J. Clin. Microbiol. **42:**4686–4696.
3. **Enright, M. C., and B. G. Spratt.** 1999. 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.
4. **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.
5. **Eskola, J., T. Kilpi, A. Palmu, J. Jokinen, J. Haapakoski, E. Herva, A. Takala, H. Käyhty, P. Karma, R. Kohberger, G. Siber, and P. H. Mäkelä.** 2001. Efficacy of a pneumococcal conjugate vaccine against acute otitis media. N. Engl. J. Med. **344:**403–409.
6. **Feil, E. J., E. C. Holmes, D. E. Bessen, M. S. Chan, N. P. Day, M. C. Enright, R. Goldstein, D. W. Hood, A. Kalia, C. E. Moore, J. Zhou, and B. G. Spratt.** 2001. Recombination within natural populations of pathogenic bacteria: short-term empirical estimates and long-term phylogenetic consequences. Proc. Natl. Acad. Sci. USA **98:**182–187.
7. **Feil, E. J., B. C. Li, D. M. Aanensen, W. P. Hanage, and B. G. Spratt.** 2004. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. **186:**1518–1530.
8. **Feil, E. J., and B. G. Spratt.** 2001. Recombination and the population structures of bacterial pathogens. Annu. Rev. Microbiol. **55:**561–590.
9. **Hanage, W. P., C. Fraser, and B. G. Spratt.** 2005. Fuzzy species among recombinogenic bacteria. BMC Biol. **3:**6.
10. **Kilpi, T., H. Åhman, J. Jokinen, K. S. Lankinen, A. Palmu, H. Savolainen, M. Grönholm, M. Leinonen, T. Hovi, J. Eskola, H. Käyhty, N. Bohidar, J. C. Sadoff, and P. H. Mäkelä.** 2003. Protective efficacy of a second pneumococcal conjugate vaccine against pneumococcal acute otitis media in infants and children: randomized, controlled trial of a 7-valent pneumococcal polysaccharide-meningococcal outer membrane protein complex conjugate vaccine in 1666 children. Clin. Infect. Dis. **37:**1155–1164.
11. **Kilpi, T., E. Herva, T. Kaijalainen, R. Syrjänen, and A. K. Takala.** 2001. Bacteriology of acute otitis media in a cohort of Finnish children followed for the first two years of life. Pediatr. Infect. Dis. J. **20:**654–662.
12. **Lan, R., and P. R. Reeves.** 2001. When does a clone deserve a name? A perspective on bacterial species based on population genetics. Trends Microbiol. **9:**419–424.
13. **Maiden, M. C., J. A. Bygraves, E. Feil, G. Morelli, J. E. Russell, R. Urwin, Q. Zhang, J. Zhou, K. Zurth, D. A. Caugant, I. M. Feavers, M. Achtman, and B. G. Spratt.** 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.
14. **Neeleman, C., C. H. Klaassen, D. M. Klomberg, H. A. De Valk, and J. W. Mouton.** 2004. Pneumolysin is a key factor in misidentification of macrolide-resistant *Streptococcus pneumoniae* and is a putative virulence factor of *S. mitis* and other streptococci. J. Clin. Microbiol. **42:**4355–4357.
15. **Ochman, H., and S. R. Santos.** 2005. Exploring microbial microevolution with microarrays. Infect. Genet. Evol. **5:**103–108.
16. **Ronquist, F., and J. P. Huelsenbeck.** 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics **19:**1572–1574.
17. **Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas.** 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:**2496–2497.
18. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. Grimont, P. Kämpfer, M. C. Maiden, X. Nesme, R. Rosselló-Mora, J. Swings, H. G. Truper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. **52:**1043–1047.
19. **Swofford, D. L.** 2003. PAUP* v. 4beta; phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Mass.
20. **Syrjänen, R. K., T. M. Kilpi, T. H. Kaijalainen, E. E. Herva, and A. K. Takala.** 2001. Nasopharyngeal carriage of *Streptococcus pneumoniae* in Finnish children younger than 2 years old. J. Infect. Dis. **184:**451–459.
21. **Toikka, P., S. Nikkari, O. Ruuskanen, M. Leinonen, and J. Mertsola.** 1999. Pneumolysin PCR-based diagnosis of invasive pneumococcal infection in children. J. Clin. Microbiol. **37:**633–637.
22. **Wertz, J. E., C. Goldstone, D. M. Gordon, and M. A. Riley.** 2003. A molecular phylogeny of enteric bacteria and implications for a bacterial species concept. J. Evol. Biol. **16:**1236–1248.
23. **Whatmore, A. M., A. Efstratiou, A. P. Pickerill, K. Broughton, G. Woodard, D. Sturgeon, R. George, and C. G. Dowson.** 2000. Genetic relationships between clinical isolates of *Streptococcus pneumoniae*, *Streptococcus oralis*, and *Streptococcus mitis*: characterization of "atypical" pneumococci and organisms allied to *S. mitis* harboring *S. pneumoniae* virulence factor-encoding genes. Infect. Immun. **68:**1374–1382.
24. **Wright, S.** 1951. The genetical structure of populations. Ann. Eugenics **15:**323–354.