J Mol Evol (1998) 47:222–229

JOURNAL OF MOLECULAR EVOLUTION

The IS*1167* Insertion Sequence Is a Phylogenetically Informative Marker Among Isolates of Serotype 6B *Streptococcus pneumoniae*

D. Ashley Robinson,[1] Susan K. Hollingshead,[1] James M. Musser,[4] Alan J. Parkinson,[5] David E. Briles,[1–3] Marilyn J. Crain[1,2]

[1] Department of Microbiology, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[2] Department of Pediatrics, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[3] Department of Comparative Medicine, University of Alabama at Birmingham, Birmingham, AL 35294, USA
[4] Section of Molecular Pathobiology, Department of Pathology, Baylor College of Medicine, Houston, TX 77030, USA
[5] Arctic Investigations Program, National Center for Infectious Diseases, Centers for Disease Control and Prevention, Anchorage, AK 99501, USA

Received: 20 May 1997 / Accepted: 20 November 1997

Abstract. The phylogenetic utility of the IS*1167* insertion sequence was examined with restriction fragment length polymorphism (RFLP) analyses of a sample of 50, predominantly invasive, capsular serotype 6B *Streptococcus pneumoniae* isolates previously characterized by multilocus enzyme electrophoresis (MLEE). The strains represented a genetically diverse assemblage of 34 distinct clonotypes composed of 26 restriction fragment types and 23 multilocus enzyme types. All isolates carried the IS*1167* insertion sequence, with an average of 9.5 copies. The cross-classification of isolates based on RFLP and MLEE typing schemes was 81% concordant. Phylogenetic analyses demonstrated a significant ($P < 0.0001$) association between strains of a given RFLP lineage with those of a given MLEE lineage. A significant correlation ($P < 0.00004$) was also found between the proportion of restriction fragments shared by any given pair of isolates and their genetic distances estimated from the MLEE data. Parity between the two genetic markers implied that the sampled isolates were in linkage disequilibrium. The existence of nonrandom associations among genetic loci was confirmed by Monte Carlo analyses of the MLEE data. These studies, thus, demonstrated that invasive pneumococcal isolates of a single capsule type recovered on a regional scale can retain a largely clonal population structure over a period of 8 years. The ability to detect linkage disequilibrium and generate relatively congruent dendrograms based on distance and parsimony methods suggested that the restriction fragment data were robust to phylogenetic analysis.

Key words: Insertion sequences — IS*1167* — Multilocus enzyme electrophoresis — Molecular epidemiology — Population genetics — Restriction fragment length polymorphism — *Streptococcus pneumoniae*

Introduction

A variety of phenotypic and genetic markers has been used to distinguish among isolates of the human pathogen, *Streptococcus pneumoniae*. The emergence and geographic spread of multidrug-resistant clones (McDougal et al. 1992; Munoz et al. 1991, 1992; Sibold et al. 1992; Soares et al. 1993) have highlighted the need for increased surveillance of this pathogen using markers capable of tracking the clinically important strains. Virulence factors such as the polysaccharide capsule, pneumococcal surface protein A (PspA), and immunoglobin A1 (IgA1) protease display extensive serological variability (Crain et al. 1990; Lomholt 1995). The high rate at which virulence factors can accumulate variations sug-

Correspondence to: D.A. Robinson; *e-mail:* daro@uab.edu

gests that these markers should be useful for distinquishing between isolates that have been separated for short periods of time. Moreover, studies of PspA and IgA1 protease have provided evidence for genetic exchange among natural populations of pneumococci (Crain et al. 1990; Lomholt 1995). Markers that evolve slowly are preferred for addressing historical questions such as the spatial and temporal stability of a given strain. Multilocus enzyme electrophoresis (MLEE) has been used to index such genetic relationships and to assess the extent of genetic recombination among pneumococci. Studies using MLEE have provided additional evidence for the exchange of genes including penicillin resistance (Kell et al. 1993), polysaccharide capsule (Sibold et al. 1992), and IgA1 protease (Lomholt 1995).

Lomholt (1995) demonstrated that, among pneumococci of nine capsule types, genetic recombination was frequent enough to render the population effectively sexual. A population structure characterized by a high ratio of genetic recombination to epidemic spread, termed panmictic (Smith et al. 1993), would profoundly affect the ability of a given marker to reflect strain relationships accurately. Under conditions of frequent genetic exchange, markers can be mosaically distributed among otherwise distantly related isolates, thus obscuring clonal relationships. This condition, in turn, impacts the ability of epidemiologists to trace the origin and migration of clinically important strains. Several investigators have demonstrated the potential epidemiological utility of fingerprinting pneumococcal DNA (Lefevre et al. 1993, 1994; Hermans et al. 1995; Viering and Fine 1989), although the use of this technique for unraveling the complex population structure of pneumococci has not been explored.

The present study was designed to compare the performance of restriction fragment length polymorphism (RFLP) analysis of the recently described IS*1167* insertion sequence (Zhou et al. 1995; Zhou and Morrison 1995) for inferring phylogenetic relationships among a collection of pneumococcal strains previously characterized by MLEE (Munoz et al. 1992). An ancilliary goal of this survey was to examine the genetic structure of a serologically, geographically, and temporally restricted subset of *S. pneumoniae* isolates. Do pneumococci sampled on a limited scale also present evidence of statistically significant genetic recombination?

Among the approximately 90 recognized capsular serotypes, serogroup 6 pneumococci consistently rank with the most frequent strains causing disease (Sniadack et al. 1995). In several geographic areas including Alaska, Finland, Iceland, Spain, and Texas, 6B pneumococci are often represented by penicillin-resistant or multidrug-resistant clones (Munoz et al. 1992; Sibold et al. 1992; Soares et al. 1993; Versalovic et al. 1993). Serotype 6B isolates therefore provide a clinically relevant model for investigating the phylogenetic utility of different markers.

Materials and Methods

Bacterial Isolates. Strains of *S. pneumoniae* were provided by the Arctic Investigations Program (AIP) from a statewide surveillance of invasive diseases involving 23 hospitals and bacteriologic laboratories in Alaska. Fifty strains received by the AIP from 11 geographic locations across Alaska between January 1982 and June 1990 and serotyped as 6B pneumococci were selected for this analysis. The collection was a representative and nearly complete sample of invasive serotype 6B isolates recovered from this region and time period. Eighty percent of the strains were isolated during the latter half of the sampling period (1986–1990). Thirteen of the sampled isolates demonstrated reduced susceptibility to penicillin, with minimum inhibitory concentrations of either 0.120 or 0.250 μg/ml. The remaining isolates were fully susceptible to penicillin. Forty-eight strains were collected from normally sterile body sites such as blood and cerebrospinal fluid, and two strains were isolated from throat cultures. Bacteria were grown on blood agar plates or in Todd–Hewitt broth with 0.5% yeast extract overnight at 37°C in a candle jar. The capsular serotype was confirmed by slide agglutination using rabbit antisera (Statens Seruminstitut, Copenhagen, Denmark).

Southern Blotting and RFLP Analyses. An IS*1167* probe of *orf1* was constructed using the primers DAM013 and DAM061 of Zhou et al. (1995). The polymerase chain reaction (PCR) product was purified by GeneClean (Bio101 Inc., Vista, CA) and random prime-labeled with digoxigenin-11-dUTP using the Genius nonradioactive DNA labeling and detection kit as described by the manufacturer (Boehringer Mannheim, Indianapolis, IN). Genomic DNA was isolated, digested to completion with *Hin*dIII, electrophoresed on a 0.8% agarose gel, and pressure-transferred to a nylon membrane as described previously (Crain et al. 1996). The membranes were incubated for 2 h at 42°C in high-sodium dodecyl sulfate (SDS) prehybridization buffer consisting of 7% SDS, 50% deionized formamide, 5× standard saline citrate (SSC), 50 m*M* sodium phosphate (pH 7.0), 0.1% *N*-lauroylsarcosine, and 2% blocking solution (composed of 150 m*M* NaCl, 100 m*M* Tris–HCl, and 2% blocking reagent). Probe hybridization was conducted at 42°C for 18 h. All membranes were washed twice in a posthybridization solution containing 0.1% SDS plus 2× SSC for 15 min at 65°C and developed chemiluminescently using the Genius nonradioactive DNA labeling and detection kit as described by the manufacturer (Boehringer Mannheim). Since plasmids, which are rare in the pneumococcus (Zhou et al. 1995), were not observed in our genomic DNA preparations, it was assumed that all copies of IS*1167* resided in the bacterial chromosome. Isolates were assigned an RFLP type (RT) based on unique banding patterns. The genetic distance between each isolate or RT was calculated as $d_{ij} = 1 - [2n_{ij}/(n_{ij} + n_j)]$, where n_{ij} is the number of shared fragments, and n_i and n_j are the total number of fragments in isolate or RT i and j, respectively (Denny et al. 1988).

MLEE Analyses. Munoz et al. (1992) assayed these serotype 6B pneumococcal isolates for electrophoretically detectable variation at 14 enzyme-encoding loci including: nucleoside phosphorylase (NSP), leucine aminopeptidase (LAP), adenylate kinase (ADK), carbamylate kinase (CAK), phosphoglucose isomerase (PGI), α,β-naphthylproprionate (EST), phosphoglucomutase (PGM), lactate dehydrogenase (LDH), glucose-6-phosphate dehydrogenase (G6P1, G6P2), 6-phosphogluconate dehydrogenase (6PG), glyceraldehyde-3-phosphate dehydrogenase (G3P), glutamate dehydrogenase (GD2), and alcohol dehydrogenase (ADH). Electrophoretically distinquishable banding patterns were treated as different alleles, and unique combinations of

alleles were used to assign each isolate an electrophoretic type (ET). Genetic diversity at an enzyme locus was calculated as $h = 1 - \sum x_i^2[n/(n-1)]$, where x_i is the frequency of the ith allele for a given locus, n is the number of isolates or ETs, and $n/(n-1)$ is a correction for small sample size (Selander et al. 1986). Mean genetic diversity per locus, H, represents the average of h values for all loci. Genetic diversity is an estimate of the probability that two randomly selected isolates or ETs will have different alleles at a given locus (Selander et al. 1986). Allelic mismatches were used as the basis for estimating genetic distance and the extent of genetic recombination between strains. Nei's (1987) standard genetic distance was calculated as $d_{ij} = -\ln(1 - D)$, where D equals the proportion of mismatched loci between isolates or ETs i and j. A Monte Carlo procedure using 10,000 iterations was used to test whether the observed variance in allelic mismatches (V_o) differed significantly from that expected in linkage equilibrium (V_e). The null hypothesis of a random association was rejected if the calculated V_o/V_e ratio was outside the 95% Monte Carlo confidence interval. Linkage tests were accomplished with the computer programs LDV and LDD kindly provided by T. Nguyen (Souza et al. 1992).

Marker Comparisons. The index of diversity was used to assess the discriminatory power of the RFLP and MLEE typing schemes. This index provides an estimate of the probability that two randomly selected strains would differ and was calculated as $D = 1 - [(1/N(N-1))\sum n_i(n_i - 1)]$, where n is the number of strains belonging to the ith type, and N is the total number of strains (Hunter and Gaston 1988). The agreement between RFLP and MLEE typing systems was examined with three statistical tests. As calculated in previous stuides of *S. pneumoniae* (Versalovic et al. 1993) and *S. pyogenes* (Hasse et al. 1994), all possible pairs of isolates were cross-classified on the basis of matched or mismatched types. Each pair of strains was examined for matched or mismatched RTs and, concurrently, for matched or mismatched ETs. The resulting 2 × 2 table was evaluated with a chi-square statistic and the percentage of concordant cells. The significance of the chi-square statistic was not estimated because the pairs on which the analysis was based are not independent. The second test, examining the agreement between RFLP and MLEE typing schemes, was based on the phylogenetic classification of isolates, where the distribution of strains classified into RFLP lineages was compared with that from the MLEE lineages. To avoid bias from small expected frequencies, some phylogenetic lineages were combined and the resulting 2 × 2 table was examined with Fisher's exact test. The third test of marker congruence was based on a correlation analysis of genetic distances estimated from RFLP and MLEE data among all possible pairs of isolates. The comparison of genetic distances estimated from different markers has been called the *g* test, and may be an especially sensitive statistic because it quantitates the differences between genotypes rather than simply scoring differences as mismatches (Tibayrenc 1995). The MANTEL program of the R 3.0 computer package (Legendre and Vaudor 1991) was used to estimate the significance of Pearson product–moment correlation coefficients between genetic distance matrices by calculating the *r*-to-*t* transformation of Mantel (1967) and by comparing the observed correlation coefficient with a null distribution based on 10,000 random permutations. The Bonferroni technique was used to establish a significance level for rejecting the null hypothesis of no correlation and was calculated as $0.05/n$, where n equals the number of pairwise comparisons (Douglas and Endler 1982).

Phylogenetic Analyses. The NEIGHBOR program of the PHYLIP 3.54 computer package (Felsenstein 1989) was used to generate UPGMA dendrograms from pairwise matrices of genetic distances. Phylogenetic patterns in the RFLP data were further investigated with a cladistic analysis using the PAUP 3.1.1 computer program (Swofford 1993). Fragments were coded as present (1) or absent (0) and treated as unweighted and unordered data. A large number of RTs coupled with few informative fragments made the recovery of a single most-parsimonious tree unlikely. Therefore, a majority-rule consensus tree was obtained using the heuristic search option, random addition of taxa, MAXTREES set to 100, and TBR branch swapping parameters. The cladogram was statistically evaluated with 100 bootstrap iterations.

Results

Genetic Diversity

The IS*1167* insertion element was present in all 50 isolates. Based on restriction fragments of 42 different sizes, we observed 26 unique banding patterns (Figs. 1 and 2). The apparent number of fragments per pattern ranged from 6 to 15, with an average ± SD of 9.46 ± 1.96. The most frequently occurring pattern (RT 2) represented only 16% of the strains. Both the number of restriction fragments and the relative intensities of the bands in a given pattern were reproducible. The intensities of different bands within a pattern, however, were not necessarily equal. Phylogenetic analyses of the RFLP patterns using distance and cladistic methods revealed four major lineages (Figs. 3A,B). Although negligible bootstrap support was found for lineages C and D in the cladistic analysis, the strong topological congruence between the two IS*1167* dendrograms supported their distinction. Furthermore, constructing several UPGMA dendrograms with a randomized input order of genotypes did not change the relative branching order or the composition of the major lineages.

These serotype 6B isolates represented 23 distinct ETs. Electrophoretically detectable variation was observed at 10 enzyme-encoding loci, while 4 dehydrogenase loci were monomorphic (G6P2, 6PG, GD2, and ADH). The average number of alleles per locus was 2.2, and the average number of loci at which isolates differed was 4.4. The most common ET (ET 2) represented 44% of the strains. The mean genetic diversity, H, was 0.205 for isolates and 0.282 for ETs (Table 1). Six major multilocus enzyme lineages were suggested by UPGMA clustering of genetic distances (Fig. 3C). The relative branching order of the major lineages differed among several UPGMA dendrograms constructed with a randomized input order of genotypes. The composition of the major lineages, however, remained the same with the exception of ETs 10 and 23 which could not be reliably assigned to any particular major lineage. The chosen topology was one that most clearly illustrated the association with the strains of the IS*1167* dendrograms.

The RFLP analyses provided a more discriminatory data set than MLEE based on 14 scorable loci (Table 2). The probability that any two strains could be distinquished by their IS*1167* fingerprints was 94%, whereas the probability that two randomly selected isolates could be distinquished using MLEE was 81%. Combining the two markers offered only a slightly greater resolving

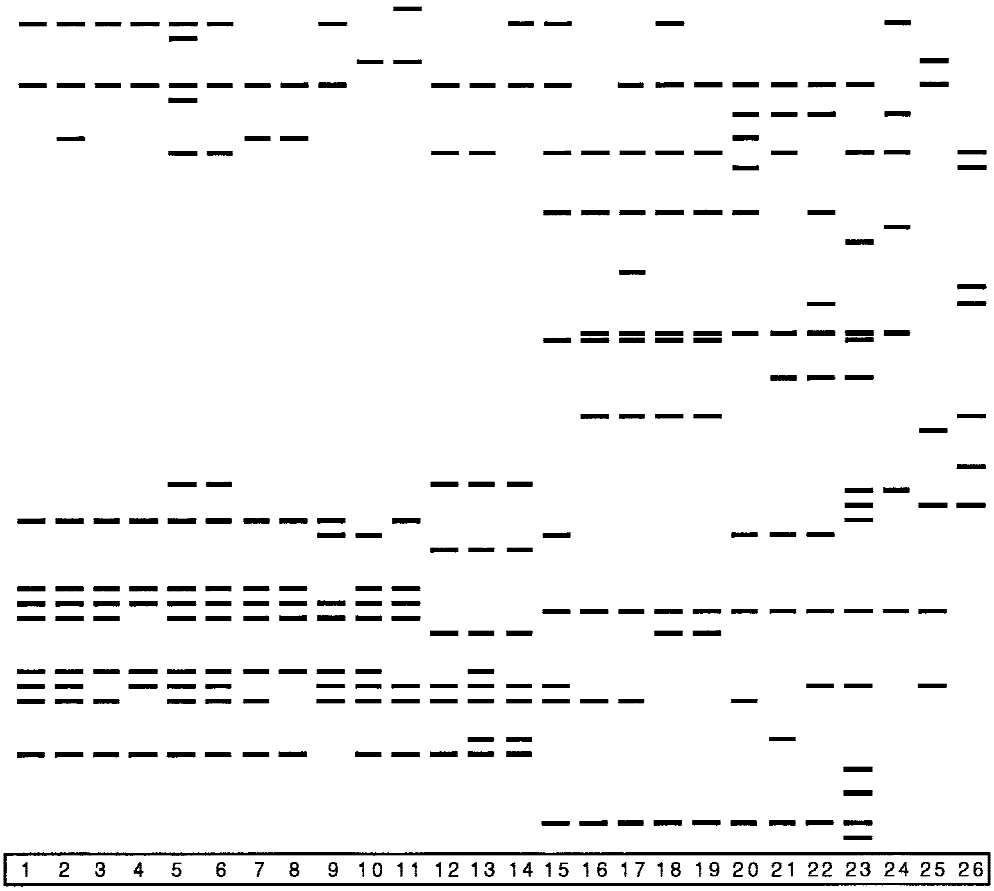

Fig. 1. Diagram of all IS*1167* RFLP patterns observed from a sample of 50 capsular serotype 6B pneumococcal isolates from Alaska.

power than the IS*1167* data alone (Table 2). These data suggest that the diversity detected by IS*1167* fingerprinting is at least comparable to electrophoretic variation at 14 metabolic enzymes.

Parity Between RFLP and MLEE Typing

Although the RFLP data were more diverse than the MLEE data, notable agreement was found between the classification of strains by these two markers. Cross-classifying the isolates based on matched or mismatched RTs and ETs demonstrated that the two typing schemes were 81% concordant (Table 3). Any given pair of strains distinquished by one marker would tend to be distinquished by the other marker. Note that ETs were more predictive of RTs than vice versa. Approximately 56% (40/72) of all pairs of isolates sharing identical ETs also had identical RTs, but only 17% (40/240) of the pairs with identical RTs shared identical ETs (Table 3). This result can be explained by the greater diversity detected by RFLP typing than by MLEE typing, since a marker that displays more variation provides a greater opportunity for a given pair of strains to differ. A contingency table classifying the strains according to the phylogenetic analyses further illustrated the concordance between the typing systems (Table 4). A significant association ($P < 0.0001$) was found between the isolates originating from a given RFLP lineage and those of a given MLEE lineage. Approximately 97% (35/36) of the strains classified into MLEE lineages A and B belonged to RFLP lineages A and B, whereas 79% (11/14) of the isolates in MLEE lineages C, D, E, and F belonged to RFLP lineages C and D (Table 4). The correlation between genetic distances estimated from RFLP and MLEE data among all possible pairs of strains was also highly significant ($r = 0.594$, $P < 0.00004$). These data all suggest that IS*1167* fingerprinting can produce results congruent with MLEE in a serologically, geographically, and temporally restricted sample of pneumococcal isolates.

A Clonal Population Structure

Agreement between independent genetic markers provides strong evidence that the markers are reflecting the same general patterns of relationship and, likewise, that the sampled isolates are a relatively clonal population (Tibayrenc 1995; Desjardins et al. 1995). Confirmation of a clonal hypothesis was sought using Monte Carlo simulations of the multilocus enzyme data. In general, clonally related isolates are expected to show nonrandom associations between metabolic loci, whereas panmictic

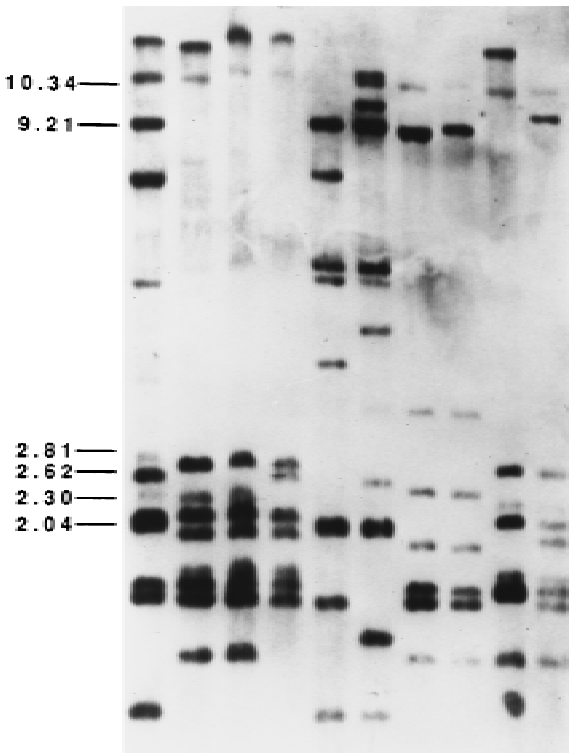

Fig. 2. Representative Southern blot of *Hin*dIII-digested genomic DNA using an IS*1167* probe. The hybridization pattern of laboratory strain RX1 was used as a standard **(left)**. Sizes are given in kilobase pairs. The RFLP patterns shown are (from left to right): 15, 1, 1, 9, 16, 21, 12, 12, 4, and 2.

populations undergoing frequent genetic recombination should reveal random associations. The Monte Carlo simulations indicated a significant difference ($P < 0.05$) between the observed variance in allelic mismatches with that expected from a chance assortment of alleles (Table 5). The association between pairs of loci with allele frequencies of at least 0.200 (Souza et al. 1992) was evaluated with Monte Carlo simulations using 1,000 iterations. These data provided evidence for limited gene flow because 61% of the two-locus comparisons revealed associations that could be due to random recombination (data not shown). Thus, while both the allelic (V_o/V_e test) and the correlative (g test) linkage tests indicated that gene flow had not been frequent enough to randomize all the genetic markers in the overall sample of isolates, the two-locus comparisons suggested that gene flow was occurring on a limited scale.

The allelic test further supported a hypothesis of clonality within phylogenetic lineages. The lack of correlation between genetic distances suggested, however, that there was some degree of intragroup gene flow (Table 5). We interpret this conflict between the allelic and the correlative analyses to be a result of the smaller sample sizes comprising the phylogenetic lineages. The powerful information required to reject the null hypothesis of random recombination may be absent from smaller samples.

Discussion

The presence of one or more copies of the IS*1167* insertion sequence in all 50 strains suggested that this polymorphic marker may be widely distributed among isolates of serotype 6B pneumococci. It was anticipated previously that IS*1167* would be a useful epidemiological marker because it is relatively stable during prolonged maintenance of laboratory cultures and may not experience the intense selective pressures associated with virulence factors (Zhou et al. 1995; Swiatlo et al. 1996). The significant associations between RFLP and MLEE patterns, coupled with the high resolving power of the former, suggested that IS*1167* fingerprinting is suitable for both phylogenetic and epidemiologic applications. Although these two typing schemes provided highly congruent data, they were not identical and therefore should be considered to provide complementary information. Caution has been recommended in using a single class of genetic markers to infer bacterial population structure (Denamur et al. 1993).

The examined isolates were highly diverse in this serologically, geographically, and temporally restricted sample. One remarkable finding of this survey was that the RFLP patterns generated from a single restriction enzyme, *Hin*dIII, would distinquish any given pair of strains on 94% of occasions. Insertion sequences are dynamic elements, where the rate of transposition is higher than that of deletion and nucleotide substitution (Egner and Berg 1981; Foster et al. 1981). Since polymorphism in multilocus enzyme electrophoretic patterns results mainly from mutation, the greater diversity detected by IS*1167* fingerprinting might have been expected. Multilocus enzyme variability was also notably extensive. A mean genetic diversity of 0.282 for ETs was only slightly lower than the 0.319 reported by Lomholt (1995) from a sample of 114 isolates of multiple capsular serotypes.

It should be noted that the variation detected by MLEE is a function of both the variation of individual enzymes and the number of enzymes examined. In this case the diversity detected using IS*1167* and a single restriction digest was greater than that offered by MLEE using 14 scorable enzymes. The inclusion of additional enzymes would be expected to detect more genetic variation than IS*1167* fingerprinting. It is entirely possible that some of the diversity observed in our sample of strains was the result of isolates from divergent genetic backgrounds acquiring serotype 6B capsule biosynthetic genes. Several studies have provided evidence that capsular serotypes can be switched in nature (Crain et al. 1990; Coffey et al. 1991; Kell et al. 1993; Lomholt 1995; Sibold et al. 1992).

Since pneumococci reproduce by binary fission, a clonal population structure would obviously be expected in extremely restricted samples. However, the extensive genetic diversity observed in our samples would not be expected if the population of isolates had been immedi-

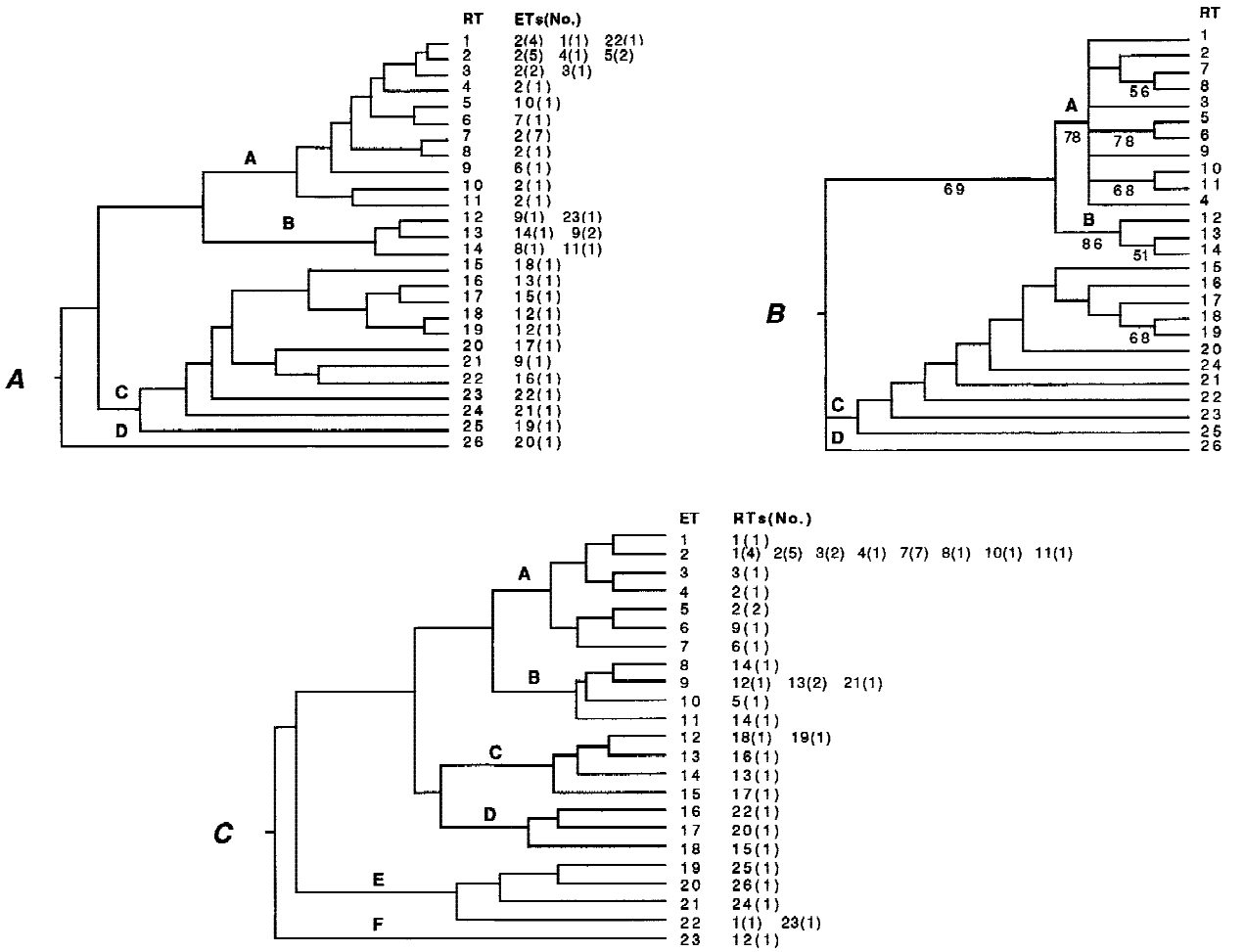

Fig. 3. Phylogenetic trees of pneumococcal genotypes based on RFLP and MLEE data. **A** and **B** represent RFLP trees generated using UPGMA and cladistic algorithms, respectively. Bootstrap values for nodes supported with at least 50% confidence are presented in the cladogram (tree length = 94, consistency index = 0.447, retention index = 0.713). **C** represents an MLEE tree constructed with the UPGMA algorithm. Major phylogenetic lineages are marked with *capital letters* (A–F). The identity and number of ETs found within a given RT are noted adjacent to dendrogram A, whereas the identity and number of RTs found within a given ET are listed next to dendrogram C.

ately derived from a common ancestor. Two explanations have previously been offered to account for the absence of statistically significant genetic recombination among bacterial isolates of certain species: (i) no mechanism for genetic exchange and (ii) spatial isolation (Smith et al. 1993). The analyses of linkage disequilibrium in this study provided evidence for limited levels of gene flow. Although genetic recombination was a relatively infrequent event, it was clearly an operative mechansim. Another explanation must therefore account for the observed clonal population structure.

One type of spatial isolation has been observed in *Rhizobium leguminosarum*, where the limited mobility of isolates presumably provides limited opportunity for distant strains to exchange genetic information (Souza et al. 1992). Strict geographic isolation among pneumococcal isolates would appear unlikely since genetically identical clones can span intercontinental distances (McDougal et al. 1992; Munoz et al. 1991, 1992; Sibold et al. 1992). A second type of spatial isolation might be represented by the partitioning of strains along epidemiologic boundaries. It should be noted that age-specific incidence rates of pneumococcal disease have been shown to vary geographically in Alaska (Davidson et al. 1994). For children less than 2 years of age the incidence of pneumococcal infection is over two times higher in western Alaska than in other regions of Alaska. Attempts to determine whether the phylogenetic lineages represented by these 50 pneumococci are isolated along geographic or epidemiologic boundaries, and whether these partitions can account for the variation in incidence rates, have been inconclusive, due largely to the small sample size (data not shown). We are presently addressing this question by applying IS*1167* fingerprinting to a larger sample of pneumococcal isolates, including serotype 6B carriage strains, from Alaska. It is possible that certain subsets within the host population are more susceptible to infection by subsets of the pneumococcal population. Genetic markers have been found in Native Alaskan children with invasive *Haemophilus influenzae* type b dis-

Table 1. Allele frequencies among 23 ETs and genetic diversity at 14 enzyme loci

	Frequency of allele					Genetic diversity (*h*)	
Locus[a]	1	2	3	4	5	Isolates	ETs
NSP	0.39	0.61				0.352	0.453
LAP	0.35	0.61	0.04			0.402	0.484
ADK	0.13	0.30	0.57			0.378	0.551
CAK	0.26	0.44	0.17	0.13		0.536	0.682
PGI	0.26	0.74				0.306	0.358
EST	0.28	0.04	0.55	0.09	0.04	0.363	0.597
PGM	0.22	0.74	0.04			0.230	0.377
LDH	0.96	0.04				0.058	0.042
G6P1	0.74	0.26				0.225	0.358
G6P2	1.00					0.000	0.000
6PG	1.00					0.000	0.000
G3P	0.96	0.04				0.020	0.042
GD2	1.00					0.000	0.000
ADH	1.00					0.000	0.000
Mean genetic diversity (*H*)						0.205	0.282

[a] Nucleoside phosphorylase (NSP), leucine aminopeptidase (LAP), adenylate kinase (ADK), carbamylate kinase (CAK), phosphoglucose isomerase (PGI), α,β-naphthylproprionate (EST), phosphoglucomutase (PGM), lactate dehydrogenase (LDH), glucose-6-phosphate dehydrogenase (G6P1, G6P2), 6-phosphogluconate dehydrogenase (6PG), glyceraldehyde-3-phosphate dehydrogenase (G3P), glutamate dehydrogenase (GD2), alcohol dehydrogenase (ADH).

Table 2. Resolving power of RFLP and MLEE typing schemes

Method	No. of types	Percentage of most frequent type	Index of diversity
RFLP	26	16	0.941
MLEE	23	44	0.807
Combined	34	14	0.967

Table 3. Cross-classification of all possible pairs of isolates based on matched or mismatched RTs and ETs

	ETs[a]		
RTs	Match	Mismatch	Sum
Match	40 (14.11)	200 (225.89)	240
Mismatch	32 (57.89)	953 (927.11)	985
Sum	72	1,153	1,225

[a] Observed (expected). $\chi^2 = 62.81$, 81% concordance.

Table 4. Associations of the phylogenetic classifications of isolates by RFLP and MLEE data[a]

	MLEE lineage[b]		
RFLP lineage	A+B	C+D+E+F	Sum
A+B	35 (27.36)	3 (10.64)	38
C+D	1 (8.64)	11 (3.36)	12
Sum	36	14	50

[a] Based on lineages depicted in Fig. 3.
[b] Observed (expected). Fisher's exact test, $P < 0.0001$.

Table 5. Tests of linkage disequilibrium based on the variance in allelic mismatches (V_o/V_e test) and correlations between genetic distances (*g* test)

			V_o/V_e test		*g* test	
RFLP lineage[a]	No. of isolates	No. of ETs	All isolates	ETs only	*r*	t_{Mantel}
A	31	9	3.363*	2.573*	−0.100	−0.640
B	7	5	2.505*	2.194*	−0.057	−0.288
C	11	10	1.647*	1.418*	0.362	2.087
A+B+C+D	50	23	2.720*	1.572*	0.594	6.248**

[a] Based on lineages depicted in Fig. 3.
* Reject null hypothesis of $V_o = V_e$ at $P < 0.05$.
** Reject null hypothesis of $r = 0$ at Bonferroni significance level.

ease (Petersen et al. 1985, 1987), an organism with many biological similarities to *S. pneumoniae*.

The natural capacity for genetic transformation provides pathogenic bacteria such as *S. pneumoniae* with an opportunity to exchange genes between distant lineages resulting in panmixia. The acquisition of fitness genes, such as those encoding novel penicillin-binding proteins, can lead to amplification and geographic spread of distinct clones followed by a period of linkage disequilibrium. The recovery of penicillin-resistant clones of pneumococci in different regions of the world attests to the reality of clonal expansion in this species (McDougal et al. 1992; Munoz et al. 1991, 1992; Sibold et al. 1992; Soares et al. 1993). This study demonstrated that while the population structure of the species as a whole may be characterized by a high ratio of genetic recombination to epidemic spread, clinically relevant subgroups such as the serotype 6B can retain a largely clonal structure over a period of 8 years, and their microevolution can be effectively followed using either RFLP or MLEE analyses.

Acknowledgments. These studies were supported in part by NIH Grants AI21548 and WHOV23/181/47 (to D.E.B.) and NIH Grants AI33119, AI37004, AI41168, and DA09238 (to J.M.M.). J.M.M. is an established investigator of the American Heart Association.

References

Coffey TJ, Dowson CG, Daniels M, Zhou J, Martin C, Spratt BG, Musser JM (1991) Horizontal transfer of multiple penicillin-binding protein genes, and capsular biosynthetic genes, in natural populations of *Streptococcus pneumoniae.* Mol Microbiol 5:2255–2260

Crain MJ, Waltman WD, Turner JS, Yother J, Talkington DF, McDaniel LS, Gray BM, Briles DE (1990) Pneumococcal surface protein A (PspA) is serologically highly variable and is expressed by

all clinically important capsular types of *Streptococcus pneumoniae*. Infect Immun 58:3293–3299

Crain MJ, Turner JS, Robinson DA, Coffey TJ, Brooks-Walter A, McDaniel LS, Briles DE (1996) Evidence for the simultaneous expression of two PspAs by a clone of capsular serotype 6B *Streptococcus pneumoniae*. Microb Pathogen 21:265–275

Davidson M, Parkinson AJ, Bulkow LR, Fitzgerald MA, Peters HV, Parks DJ (1994) The epidemiology of invasive pneumococcal disease in Alaska, 1986-1990—ethnic differences and opportunities for prevention. J Infect Dis 170:368–376

Denamur E, Picard B, Decoux G, Denis JB, Elion J (1993) The absence of correlation between allozyme and *rrn* RFLP analysis indicates a high gene flow rate within human clinical *Pseudomonas aeruginosa* isolates. FEMS Microbiol Lett 110:275–280

Denny TP, Gilmour MN, Selander RK (1988) Genetic diversity and relationships of two pathovars of *Pseudomonas syringae*. J Gen Microbiol 134:1949–1960

Desjardins P, Picard B, Kaltenbock B, Elion J, Denamur E (1995) Sex in *Escherichia coli* does not disrupt the clonal structure of the population: evidence from random amplified polymorphic DNA and restriction-fragment-length polymorphism. J Mol Evol 41:440–448

Douglas ME, Endler JA (1982) Quantitative matrix comparisons in ecological and evolutionary investigations. J Theor Biol 99:777–795

Egner C, Berg DE (1981) Excision of transposon Tn*5*. Proc Natl Acad Sci USA 78:459–463

Felsenstein J (1989) PHYLIP: phylogeny inference package (version 3.2). Cladistics 5:164–166

Foster TJ, Lundblad V, Hanley-Way S, Halling SM, Kleckner N (1981) Three Tn*10*-associated excision events: relationship to transposition and role of direct and inverted repeats. Cell 23:215–227

Hasse AM, Melder A, Mathews JD, Kemp DJ, Adams M (1994) Clonal diversity of *Streptococcus pyogenes* within some M-types revealed by multilocus enzyme electrophoresis. Epidemiol Infect 113:455–462

Hermans PWM, Sluijter M, Hoogenboezem T, Heersma H, van Belkum A, de Groot R (1995) Comparative study of five different DNA fingerprint techniques for molecular typing of *Streptococcus pneumoniae* strains. J Clin Microbiol 33:1606–1612

Hunter PR, Gaston MA (1988) Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J Clin Microbiol 26:2465–2466

Kell CM, Jordens JZ, Daniels M, Coffey TJ, Bates J, Paul J, Gilks C, Spratt BG (1993) Molecular epidemiology of penicillin-resistant pneumococci isolated in Nairobi, Kenya. Infect Immun 61:4382–4391

Lefevre JC, Faucon G, Sicard AM, Gasc AM (1993) DNA fingerprinting of *Streptococcus pneumoniae* strains by pulsed-field gel electrophoresis. J Clin Microbiol 31:2724–2728

Lefevre JC, Gasc AM, Lemozy J, Sicard AM, Faucon G (1994) Pulsed-field gel electrophoresis for molecular epidemiology of penicillin resistant *Streptococcus pneumoniae* strains. Pathol Biol 42:547–552.

Legendre P, Vaudor A (1991) The R package: multidimensional analysis, spatial analysis (version 3.0). Departemente de sciences biologiques, Universite de Montreal

Lomholt H (1995) Evidence for recombination and an antigenically diverse immunoglobin A1 protease among strains of *Streptococcus pneumoniae*. Infect Immun 63:4238–4243

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

McDougal LK, Facklam R, Reeves M, Hunter S, Swenson JM, Hill BC, Tenover FC (1992) Analysis of multiply antimicrobial-resistant isolates of *Streptococcus pneumoniae* from the United States. Antimicrob Agents Chemother 36:2176–2184

Munoz R, Coffey TJ, Daniels M, Dowson CG, Laible G, Casal J, Hakenbeck R, Jacobs M, Musser JM, Spratt BG, Tomasz A (1991) Intercontinental spread of a multiresistant clone of serotype 23F *Streptococcus pneumoniae*. J Infect Dis 164:302–306

Munoz R, Musser JM, Crain M, Briles DE, Marton A, Parkinson AJ, Sorenson U, Tomasz A (1992) Geographic distribution of penicillin-resistant clones of *Streptococcus pneumoniae*: characterization by penicillin-binding protein profile, surface protein A typing, and multilocus enzyme analysis. Clin Infect Dis 15:112–118

Nei M (1987) Molecular evolutionary genetics. Columbia University Press, New York, p 220

Petersen GM, Silimperi DR, Scott EM, Hall DB, Rotter JI, Ward JI (1985) Uridine monophosphate kinase 3: a genetic marker for susceptibility to *Haemophilus influenzae* type b disease. Lancet 2:417–419

Petersen GM, Silimperi DR, Rotter JI, et al. (1987) Genetic factors in *Haemophilus influenzae* type b disease susceptibility and antibody acquisition. J Pediatr 110:228–233

Selander RK, Caugant DA, Ochman H, Musser JM, Gilmour MN, Whittam TS (1986) Methods of multilocus enzyme electrophoresis for bacterial population genetics and systematics. Appl Environ Microbiol 51:873–884

Sibold C, Wang J, Henrichsen J, Hakenbeck R (1992) Genetic relationships of penicillin-susceptible and -resistant *Streptococcus pneumoniae* strains isolated on different continents. Infect Immun 60:4119–4126

Smith JM, Smith NH, O'Rourke M, Spratt BG (1993) How clonal are bacteria? Proc Natl Acad Sci USA 90:4384–4388

Sniadeck DH, et al. (1995) Potential interventions for the prevention of childhood pneumoniae: geographic and temporal differences in serotype and serogroup distribution of sterile site pneumococcal isolates from children—implications for vaccine strategies. Pediatr Infect Dis J 14:503–510

Soares S, Kristinsson KG, Musser JM, Tomasz A (1993) Evidence for the introduction of a multiresistant clone of serotype 6B *Streptococcus pneumoniae* from Spain to Iceland in the late 1980s. J Infect Dis 168:158–163.

Souza V, Nguyen TT, Hudson RR, Pinero D, Lenski RE (1992) Hierarchical analysis of linkage disequilibrium in *Rhizobium* populations: evidence for sex? Proc Natl Acad Sci USA 89:8389–8393

Swiatlo E, Crain MJ, McDaniel LS, Brooks-Walter A, Coffey TJ, Spratt BG, Morrison DA, Briles DE (1996) DNA polymorphisms and variant penicillin-binding proteins as evidence that relatively penicillin-resistant pneumococci in western Canada are clonally related. J Infect Dis 174:884–888

Swofford DL (1993) PAUP: phylogenetic analysis using parsimony (version 3.1.1). Natural History Survey, Champaign

Tibayrenc M (1995) Population genetics and strain typing of microorganisms: how to detect departures from panmixia without individualizing alleles and loci. CR Acad Sci Paris 318:135–139

Versalovic J, Kapur V, Mason EO, Shah U, Koeuth T, Lupski JR, Musser JM (1993) Penicillin-resistant *Streptococcus pneumoniae* strains recovered in Houston: identification and molecular characterization of multiple clones. J Infect Dis 167:850–856

Viering TP, Fine DP (1989) Genetic analysis of *Streptococcus pneumoniae* serotypes with the use of DNA fingerprinting. J Infect Dis 160:76–82

Zhou L, Morrison DA (1995) The new insertion sequence IS*1167* of *Streptococcus pneumoniae* is related to IS*1096* and a family of IS elements occurring widely among gram-positive bacteria. Dev Biol Stand 85:77–81

Zhou L, Hui FM, Morrison DA (1995) Characterization of IS*1167*, a new insertion sequence in *Streptococcus pneumoniae*. Plasmid 33: 127–138