Published in final edited form as:
Infect Genet Evol. 2011 December ; 11(8): 1929–1939. doi:10.1016/j.meegid.2011.08.016.

Population structure of hyperinvasive serotype 12F, clonal complex 218 *Streptococcus pneumoniae* revealed by multilocus *boxB* sequence typing

Alexey V. Rakov[a], **Kimiko Ubukata**[b], and **D. Ashley Robinson**[a,*]

[a]Department of Microbiology, University of Mississippi Medical Center, Jackson, MS

[b]Laboratory of Molecular Epidemiology for Infectious Agents, Kitasato University, Tokyo, Japan

Abstract

At least four outbreaks of invasive disease caused by serotype 12F, clonal complex 218 *Streptococcus pneumoniae* have occurred in the United States over the past two decades. We studied the population structure of this clonal complex using a sample of 203 outbreak and surveillance isolates that were collected over 22 years from 34 US states and eight other countries. Conventional multilocus sequence typing identified five types and distinguished a single outbreak from the others. To improve typing resolution, multilocus *boxB* sequence typing (MLBT) was developed from 10 variable *boxB* minisatellite loci. MLBT identified 86 types and distinguished between each of the four outbreaks. Diversity across *boxB* loci tended to be positively correlated with repeat array size and, overall, best fit the infinite alleles mutation model. Multilocus linkage disequilibrium was strong, but pairwise disequilibrium decreased with the physical distance between loci and was strongest in one large region of the chromosome, indicating recent recombinations. Two major clusters were identified in the sample, and they were differentiated geographically, as western and more easterly US clusters, and temporally, as clusters that predominated before and after the licensure of pneumococcal conjugate vaccines. The diversity and linkage disequilibrium within these two clusters also differed, suggesting different population dynamics. MLBT revealed hidden aspects of the population structure of these hyperinvasive pneumococci, and it may provide a useful adjunct tool for outbreak investigations, surveillance, and population genetics studies of other pneumococcal clonal complexes.

Keywords

Streptococcus pneumoniae; outbreaks; population structure; minisatellites

1. Introduction

Streptococcus pneumoniae is a major cause of life-threatening human diseases such as septicemia, community-acquired pneumonia, and meningitis (O’Brien et al., 2009). At least 93 capsular serotypes are known in this species, and polysaccharides from seven of the most

[*]Corresponding author. Postal address: Department of Microbiology, University of Mississippi Medical Center, 2500 North State Street, Jackson, MS 39216. Phone: (601) 984-1702. Fax: (601) 984-1708. darobinson@umc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

common serotypes are included in a pneumococcal conjugate vaccine that was licensed for use in children in the year 2000 in the United States. This vaccine has been highly effective in reducing the burden of invasive pneumococcal disease in children (Poehling et al., 2006; Whitney et al., 2003) and, indirectly, in adults (Isaacman et al., 2007). However, the serotype-specific immunity induced by this vaccine sets up an ecological opportunity for replacement of the vaccine serotypes by the numerous non-vaccine serotypes. This serotype replacement phenomenon was first demonstrated by the non-vaccine serotype 19A, which had significant post-vaccine increases in invasive disease and antibiotic resistance (Pai et al., 2005). A new conjugate vaccine that covers 13 serotypes, including serotype 19A, was deployed last year in the US.

The remaining non-vaccine serotypes include some of the most invasive pneumococci in the species (Sleeman et al., 2006). In particular, serotype 12F is not covered by any of the conjugate vaccine formulations, but it is hyperinvasive in multiple regions of the world as evidenced by its significantly more frequent isolation from people with invasive disease than from healthy carriers (Kronenberg et al., 2006; Michel et al., 2005; Saha et al., 1997;, Sandgren et al., 2004; Shouval et al., 2006; Sleeman et al., 2006). In the US, 12F invasive disease remained steady or decreased in most populations following vaccination (Hicks et al., 2007; Lexau et al., 2005; Whitney et al., 2003), but it significantly increased in certain at-risk populations (Weatherholtz et al., 2010). In the United Kingdom, where national immunization with the 7-valent conjugate vaccine began in the year 2006, a significant pre-vaccine decrease in 12F invasive disease has been halted and nominally reversed after introduction of the vaccine (Foster et al., 2011). Although the conjugate vaccines have not been implemented nationally in either Poland or Japan, 12F has significantly increased among meningitis cases in Poland (Skoczyńska et al., 2011) and, notably, it ranks first among serotypes causing invasive pneumococcal disease in adults in Japan (Chiba et al., 2010).

Serotype 12F causes outbreaks of invasive disease in human populations with identifiable risk factors. For example, in the US, 12F caused outbreaks in a Texas jail in 1989 (Hoge et al., 1994), in a Maryland child care center in 1992 (Cherian et al., 1994), in a California homeless shelter in 2004 (Lee et al., 2005), and in a rural region of Alaska in 2003–2004 (CDC, 2005). Serotype 12F has also caused outbreaks in Australia (Gratten et al., 1995). A previous study indicated that the Texas and Maryland outbreaks were caused by different clones of a major genetic lineage of 12F pneumococci (Robinson et al., 1999). These outbreak strains were indistinguishable based on IS*1167* and *boxA* fingerprinting procedures, but they differed based on polymorphisms in the pneumococcal surface protein A (Robinson et al., 1999). The genetic relationships between these older outbreak strains and strains from the more recent California and Alaska outbreaks have not been described.

Multilocus sequence typing (MLST) of housekeeping genes has become a tool of choice for unambiguously defining clones of pneumococci and for revealing a coarse outline of the species population structure (Enright and Spratt, 1998). However, to obtain a fine-scale outline of population structure, portable tools that are even more discriminating than MLST are needed. Next generation sequencing of pneumococcal genomes is one such tool (Croucher et al., 2011), but its application can be limited by the need for extensive bioinformatics infrastructure (Joseph and Read, 2010). Another tool involves the typing of multiple, rapidly-evolving loci that each contain a variable number of tandem repeats (MLVA) (Vergnaud and Pourcel, 2009). An MLVA scheme has been developed for *S. pneumoniae* that scores electrophoretically-determined size variation in tandem repeat arrays from 16 loci (Koeck et al., 2005). This MLVA scheme has been useful for confirming relatedness of pneumococcal outbreak isolates of serotypes 1 (Yaro et al., 2006) and 5 (Pichon et al., 2010). However, use of MLVA to further elucidate population structure can be limited by the independent evolution of the same repeat array sizes in different strains (i.e. size homoplasy) (Estoup et al., 2002). Dispersed throughout the pneumococcal genome is a tandem repeat of the minisatellite class, the 45 bp tandemly repeated *boxB* sequence, which is part of the BOX intergenic repetitive elements (Martin et al., 1992). Here, we use sequence-based typing of *boxB* minisatellites to provide a portable, discriminatory typing scheme that should be less prone to size homoplasy than electrophoretic schemes. This typing scheme requires the same laboratory infrastructure as that needed to produce conventional MLST data. The goal of this study was to obtain a fine-scale outline of the population structure of serotype 12F, clonal complex 218 *S. pneumoniae*. To accomplish this goal, we first developed and validated multilocus *boxB* sequence typing (MLBT) for pneumococcal outbreak investigations. Typing of a larger collection of surveillance isolates allowed us to characterize some diversity at *boxB* loci. We then used the new typing scheme to reveal hidden aspects of the population structure of these hyperinvasive pneumococci.

2. Materials and methods

2.1. Bacterial isolates

A total of 203 outbreak and surveillance isolates of serotype 12F, clonal complex 218 *S. pneumoniae* were included in this study. Four outbreaks were represented by a total of 28 isolates, including 9 isolates from the 1989 Texas jail outbreak (Hoge et al., 1994), 6 isolates from the 1992 Maryland child care center outbreak (Cherian et al., 1994), 5 isolates from the 2004 California homeless shelter outbreak (Lee et al., 2005), and 8 isolates from a newly identified 2006–2008 outbreak in a rural region of Alaska (Zulz et al., in preparation). Isolates from a 2003–2004 outbreak in another rural region of Alaska (CDC, 2005) were found to belong to a different clonal complex (CC1527) by MLST and were excluded from this study. Surveillance sources provided a total of 175 isolates, including 32 isolates from the Active Bacterial Core surveillance (ABCs), which is a component of the Centers for Disease Control and Prevention (CDC) Emerging Infections Program (http://www.cdc.gov/abcs/index.html), 97 isolates from the US and global Prospective Resistant Organism Tracking and Epidemiology for the Ketolide Telithromycin (PROTEKT) surveys (Farrell and Jenkins, 2004), 16 isolates from the Arctic Investigations Program (AIP) of the CDC, 28 isolates from Queensland Health Services and the Department of Pathology of the Royal Darwin Hospital, and 2 isolates from a survey of invasive pneumococcal disease in Japan (Chiba et al., 2010). The study sample included 163 isolates from 34 US states, and 40 isolates from 8 other countries. Dates of isolation spanned 22 years (1986–2008). Characteristics of all study isolates are listed in Table S1.

Long-term storage of isolates was at −80°C in a solution of Todd Hewitt broth with 0.5% yeast extract and 15% glycerol. Bacterial growth was done overnight on blood agar plates at 37°C in a candle jar. Bacterial genomic DNA was isolated with a DNeasy kit (Qiagen), according to the manufacturer's instructions. The presence of serotype 12F capsule genes were verified for each isolate with PCR, using the 12F-specific and *cpsA*-control primers of Pai et al. (2006). MLST was done for each isolate according to Enright and Spratt (1998). eBURST analysis of MLST data (Feil., 2004) was used to assign the isolates to clonal complex 218.

2.2. Identification and sequence typing of variable boxB loci

The draft genome sequence of serotype 12F strain CDC0288-04, which is a blood isolate from the 2004 California outbreak, was searched for copies of the BOX intergenic repetitive elements. These elements are modular, with a 59 bp *boxA* sequence and a 50 bp *boxC* sequence that often flank one or more copies of a 45 bp *boxB* sequence (Martin et al., 1992). Of the 112 identified BOX elements in this strain, 33 had an $AB_{\geq 2}C$ configuration with tandemly repeated *boxB* sequences. These 33 loci were screened for variation in a subset of six isolates, which we named the discovery isolates, that included one outbreak isolate from each of the Texas, Maryland, and California outbreaks along with a random surveillance isolate from each of these three US states. With these six discovery isolates, the 33 loci were amplified by PCR and sequenced on both DNA strands.

Loci were subsequently excluded from the multilocus *boxB* sequence typing (MLBT) scheme if they were i) problematic to amplify and/or sequence, ii) not variable among the discovery isolates, or iii) physically located within 10 kb of MLST loci or other candidate *boxB* loci based on coordinates from the finished genome sequence of serotype 4 strain TIGR4 (Tettelin et al., 2001). The 10 *boxB* loci and the primer sequences selected for MLBT are listed in Table S2. We note that all of these loci can be typed for 12F, CC1527 pneumococci (not shown) and that the majority of these loci are detected in genome-sequenced pneumococci of diverse backgrounds, suggesting typeability beyond the 12F, CC218 isolates studied here. Seven of these loci are also included in the MLVA scheme (Koeck et al., 2005), whereas three of these loci are new to this study. Thermal cycling parameters for the amplification of all loci include an initial denaturation step of 94°C for 5 min, followed by 30 cycles of 94°C for 30 s, 60°C for 30 s, and 72°C for 45 s, and a final elongation step of 72°C for 7 min.

The 10 *boxB* loci used for MLBT are identified by an uppercase "B" followed by a number; for example, locus B25. Each unique 45 bp *boxB* repeat sequence is identified by a lowercase "b" followed by a number; for example, repeat sequence b16. Each unique combination of *boxB* repeat sequences at a locus defines an allele, and alleles are identified based on their repeat profile (i.e. repeat array). For ease of viewing, the repeat profile need not include the lowercase "b" for each repeat; for example, at locus B25, allele 5-1-2 is defined by the repeats b5, b1, and b2 in tandem. As another example, at locus B27, allele 9-9-9-9-11 is defined by four copies of the repeat b9, followed by one copy of repeat b11. To verify that the entire *boxB* repeat array has been sequenced and to determine the correct orientation of the array, it is necessary to find the flanking sequences of *boxA* and *boxC* in the sequenced amplicons. These flanking sequences are not included as typing information even though they exhibit sequence variation. Finally, each unique combination of alleles across the 10 *boxB* loci defines a *boxB* sequence type (BT).

2.3. Stability tests

The three discovery isolates from outbreaks were subsequently used to conduct stability tests of the 10 *boxB* loci used for MLBT. For each of these isolates, individual colonies from blood agar plates were picked at random and passed onto fresh plates followed by overnight incubation at 37°C in a candle jar. Isolates were passaged in this fashion for seven days, then DNA was isolated and all 10 *boxB* loci were amplified by PCR and sequenced.

2.4. Measures of genetic diversity and differentiation

Genetic diversity was measured using Simpson's index (Hunter and Gaston, 1988) and the k_{e3} estimator of the effective number of types (Nielsen et al., 2003). 95% confidence intervals (CIs) for these diversity measures were calculated as described previously (Grundmann et al., 2001; Smyth et al., 2010). Simpson's index provides an estimate of the probability that two isolates picked at random from the population belong to different types, $1-\Sigma_{pi}{}^2$, and the effective number of types provides an estimate of the number of equally frequent types that will produce the observed diversity, $1/\Sigma_{pi}{}^2$, where p is the proportion of isolates of the *i*th type.

Genetic structuring of the population was investigated using the frequencies of two sequence clusters among various geographically- and temporally-defined subpopulations. Cluster frequencies were estimated directly as the proportion of isolates of a given cluster, p. Assuming a binomial sampling distribution, the variance of p was calculated as $p(1-p)/n$, where n is the number of isolates in the subpopulation. 95% CIs around these cluster frequencies were calculated as $p \pm 2\sqrt{\mathrm{V}}\mathrm{ar}(p)$. Differentiation of the subpopulations was also investigated using Jost's D (Jost, 2008), which partitions genetic diversity into within and between subpopulation components using a measure of diversity based on the effective number of types. Jost's D will be 0 when subpopulations are identical in cluster frequencies and 1 when subpopulations are completely different in cluster frequencies (i.e. completely differentiated). The SPADE program (Chao and Shen, 2003) was used to calculate the D estimator in equation 13 of Jost (Jost, 2008) and to calculate 95% CIs based on bootstrap resampling with 1,000 replicates. Geographic subpopulations were defined according to the four US census regions (http://www.census.gov/popest/geographic): west, midwest, south, northeast. Temporal subpopulations were defined according to pre- and post-licensure of the first pneumococcal conjugate vaccine in the US: 1986–1999, 2000–2008.

2.5. Measures of linkage disequilibrium

The standardized index of association, I_{AS}, was used as a measure of multilocus linkage disequilibrium (Haubold and Hudson, 2000). The null hypothesis of linkage equilibrium, I_{AS}=0, was tested with 1,000 Monte Carlo simulations using the LIAN v3.5 program (Haubold and Hudson, 2000). Two approaches were used to measure linkage disequilibrium between pairs of loci. First, the TASSEL v2.0 program (Bradbury et al., 2007) was used to calculate the classical measures D' (Lewontin et al., 1964) and r^2 (Hill and Robertson, 1966). These measures were originally designed to work with genetic systems of low diversity (i.e. only two alleles at a locus). TASSEL accounts for multiple alleles using an approach described by Farnir et al. (2000). Secondly, we calculated haplotype homozygosity as described by Sabatti and Risch (2002). Briefly, this approach compares the complement of Simpson's index of diversity, which estimates the probability that two random isolates are identical, $\Sigma_{pi}{}^2$, for individual loci and their two-locus haplotypes. With linkage equilibrium, the product of the identities of individual loci are expected to be equal to their two-locus haplotype identity. Linkage disequilibrium can be detected through either an excess or deficit of two-locus haplotype identity. For each pair of loci, A and B, we calculated

$$H = H_{AB} - H_A H_B$$

where H_A is the identity at locus A, H_B is the identity at locus B, and H_{AB} is the two-locus haplotype identity. H will be 0 in linkage equilibrium and greater or less than 0 in linkage disequilibrium. HR^2 is then calculated in a manner analogous to that of r^2 as

$$HR^2 = \frac{H^2}{H_A(1-H_A)H_B(1-H_B)}$$

We used r^2 and HR^2 as measures of the strength of the association between loci A and B. Both measures are known to be negatively correlated with the physical distance between loci in chromosomes that undergo recombination (Jakobsson et al., 2008; Meunier and Eyre-Walker, 2001). To test the significance of these negative correlations, the ZT v1.1 program (Bonnet and Van de Peer, 2002) was used to perform Mantel tests with 100,000 permutations. Only variable loci were included in each of the subsets of isolates that were used to study linkage disequilibrium.

2.6. Mutation models and phylogenetic analyses

The process of change in repeat array size can be described by two contrasting neutral mutation models: the infinite alleles model (IAM) and the stepwise mutation model (SMM). Under the IAM (Kimura and Crow, 1964), a repeat array is equally likely to change by a number of repeats, so mutations often result in new alleles and size homoplasy is rare. Under the SMM (Ohta and Kimura, 1973), each mutation results in the gain or loss of single repeats, so previously exisiting alleles are often recreated and size homoplasy is common. Between these two extremes is the two-phase model (TPM), which allows a defined amount of change under both the IAM and SMM (Di Rienzo et al., 1994). 1,000 coalescent simulations were done with the program BOTTLENECK v1.2 (Cornuet and Luikart, 1997) to test the standardized difference between the observed diversity (in the sense of Simpson's index) at *boxB* loci and that expected under the IAM, SMM, and TPM given the number of alleles and sample size. Note that locus diversity can be in excess or deficit to that expected under the different models. In equilibrium populations, roughly equal numbers of loci are expected to have diversity excesses and deficits, whereas growing populations and populations with recently imported alleles are expected to have diversity deficits given their (higher) number of alleles (Luikart et al., 1998).

The mutation model with the best fit to the observed diversity at *boxB* loci subsequently determined the best choices of genetic distance coefficients to be used in phylogenetic analyses. Trees were constructed with neighbor-joining analysis of *p*-distances (an IAM distance) from the combined MLST and MLBT alleles using the PAUP* 4.0b10 program (Swofford, 2003). Additional trees were constructed with neighbor-joining analysis of D_A-distances (another IAM distance) from MLBT alleles using the POPTREE2 program (Takezaki et al., 2010). The robustness of tree topologies were evaluated with 1,000 bootstrap replicates. The goeBURST algorithm (Francisco et al., 2009) was also used to cluster the isolates.

2.7. Other statistical analyses

Two-tailed, non-parametric tests for correlation and for differences in median were done using the InStat v3.1 program (GraphPad Software).

2.8. Nucleotide sequence accession numbers

Each of the 44 unique *boxB* repeat sequences were submitted to GenBank with accession numbers JF705882-JF705925.

3. Results

3.1. Epidemiological validation of MLBT

Based on MLST, the Texas, Maryland, and Alaska outbreak isolates were ST218, and the California outbreak isolates were ST220 (Table 1). These STs differ at one of the seven MLST loci, *spi*. Based on MLBT, the Texas outbreak isolates were BT4, the California outbreak isolates were BT2, the Alaska outbreak isolates were BT5, and the Maryland outbreak isolates were BT6 with the exception of a single BT15 isolate that had an additional b5 repeat in locus B34 (Table 1). The average number of pairwise *boxB* locus differences (±SD) was 0.056 (±0.232) when isolates from the same outbreak were compared versus 5.30 (±2.60) when isolates from the different outbreaks were compared. Thus, MLBT clearly identified isolates of the same outbreak and clearly distinguished between the different outbreaks.

The 28 outbreak isolates were compared with 175 surveillance isolates of the same serotype and clonal complex. BT4 and BT6 from the older Texas and Maryland outbreaks were unique in the sample (Table S1). The single BT15 isolate from the Maryland outbreak matched a surveillance isolate from Spain from 2006. BT2 from the California outbreak matched seven surveillance isolates from six US states collected between 1996 and 2005. BT5 from the Alaska outbreak matched four surveillance isolates from Alaska and California collected between 1995 and 2004. These results suggested hidden epidemiological links between outbreak and certain surveillance cases. The average number of pairwise *boxB* locus differences (±SD) was 5.45 (2.38) when outbreak isolates were compared with surveillance isolates. These results indicated that isolates from the different outbreaks were generally as different from each other as they were to surveillance isolates, with the noted exceptions.

A representative isolate of each unique BT from the outbreaks along with all surveillance isolates made a less-biased sample of 180 isolates for subsequent study of diversity and population structure. From this sample, MLST identified 5 STs, MLBT identified 86 BTs, and no additional discrimination was provided by combining data from the two typing schemes. Simpson's index of diversity (95% CI) was 0.624 (0.587, 0.660) for MLST and 0.973 (0.961, 0.984) for MLBT. The effective number of types (95% CI) was 2.66 (2.50, 2.82) for MLST and 36.45 (24.09, 48.81) for MLBT. By both measures of diversity, MLBT had significantly greater discriminatory ability than MLST.

The 10 *boxB* loci used for MLBT were stable during limited passage in the laboratory; isolates passaged on solid media for seven days did not produce any changes in these loci. Thus, stability tests done over longer periods of time or done with higher cell densities would be needed to observe *in vitro* changes in *boxB*. In addition, no differences were observed between the *boxB* sequences from the partial genome sequence of 12F strain CDC0288-04 and the sequences from our stock of this strain. Finally, some BTs were stable enough in natural populations to be collected from multiple continents over a period of a decade (e.g. BT15). Taken together, the above results validated the MLBT scheme for pneumococcal outbreak investigations.

3.2. Characteristics of boxB minisatellite diversity

An understanding of the diversity exhibited by molecular markers can inform the interpretative criteria used to assess epidemiological relationships. The number of alleles identified at different *boxB* loci ranged from 3 to 19 (Table 2). Across loci, the initial number of alleles ascertained among the six discovery isolates tended to be positively correlated with the final number of alleles found among all isolates (Spearman's r=0.622, P=0.060). Loci starting with 1 or 2 alleles among discovery isolates ended with 3 to 6 alleles among all isolates, whereas loci starting with 3 or 4 alleles among discovery isolates ended with 6 to 19 alleles among all isolates (Table 2). Moreover, maximum and average repeat array sizes tended to be positively correlated with diversity across loci (Spearman's r=0.796 and r=0.636, P=0.009 and P=0.054, respectively). Since Simpson's index and the effective number of types ranked the diversity of these loci in the same order, the above correlations were the same for both measures of diversity. Thus, loci with longer repeat arrays tended to yield more diversity.

A total of 44 unique *boxB* repeat sequences were identified (Table 3). Most loci contained private sets of repeat sequences: 38 of 44 repeats were unique to single loci and 6 of 44 repeats were found at multiple loci. Repeats b16, b5, and b7 were all found at multiple loci and were among the most similar to the consensus repeat (Table 3).

Concern about size homoplasy was one reason for sequencing these *boxB* loci rather than scoring their repeat array sizes electrophoretically as done for conventional MLVA. If repeat array sizes had been scored perfectly by electrophoresis (as determined from their sequences), 51 different alleles would have been identified. By sequencing, 100 different alleles were identified, indicating that nearly half of these alleles might have been misclassified by conventional MLVA.

To better characterize the processes underlying the genetic variation at these *boxB* loci, the observed diversity was compared with that expected under different mutation models. The infinite alleles model (IAM) best fit the observed diversity at seven *boxB* loci, whereas the two-phase model (TPM) best fit the diversity at two loci and the stepwise mutation model (SMM) best fit the diversity at one locus (Table 4). Of note, the SMM was rejected at the P<0.05 level for seven loci and at the highly stringent Bonferroni P<0.005 level for four loci. While the diversity at loci B31 and B4 best fit the IAM, all three models were a poor fit for these two loci (Table 4). The coalescent simulations assume an equilibrium population with no recombination, and a violation of these assumptions may impact model fit. Nonetheless, these results underscored the complicated dynamics of minisatellites. They also provided optimism for use of *boxB* sequences to unravel 12F, CC218 population structure because the overall best fit of their diversity to the IAM predicts lower levels of size homoplasy than would be the case had their diversity best fit the SMM.

3.3. A large region of linkage disequilibrium in the chromosome

The degree to which recombination has broken down allelic associations within bacterial chromosomes is a defining characteristic of bacterial population structure (Maynard Smith et al., 1993). Significant multilocus linkage disequilibrium was detected among the 13 variable MLST and MLBT loci when all 180 isolates were examined (I_{AS}=0.255, P<0.001) and when single isolates of each of the 86 unique haplotypes were examined (I_{AS}=0.228, P<.001). Such results are typically interpreted to indicate either a relatively clonal population that has rarely undergone recombination or a cryptically subdivided population where recombination is more frequent within subpopulations than between subpopulations (Maynard Smith et al., 1993). Since recombination needs to be about 20 times more common than mutation for I_A to reflect random associations of alleles at different loci (Maynard Smith, 1994), bacterial populations can undergo much recombination yet still reflect linkage disequilibrium by I_A.

This possibility was investigated by measuring the association between pairwise disequilibrium and the physical distance between loci in the chromosome. Based on TASSEL's approach for dealing with multiple alleles (Farnir et al., 2000), the classical measure r^2 was negatively correlated with the physical distance between loci (Mantel test, r=−0.238, P=0.044) (Fig. 1A), whereas the classical measure D' was not (Mantel test, r= −0.024, P=0.420). These same patterns were found when alleles at each locus were recoded to include only two allele classes, including a class with the most frequent allele and a class with all other alleles lumped together (not shown). The negative correlation observed between r^2 and the physical distance between loci is a hallmark of recombination (Meunier and Eyre-Walker, 2001).

We also used a more recently introduced measure of linkage disequilibrium called haplotype homozygosity (Sabatti and Risch, 2002). This measure treats multiple alleles in the same fashion as Simpson's index of diversity, by downweighting rare alleles. As shown in Fig. 1A, HR^2 was also negatively correlated with the physical distance between loci (Mantel test, r=−0.271, P=0.026). Surprisingly, 14 of 15 pairs of loci with the strongest associations (HR^2 >0.25) mapped to one large region of the chromosome (Fig. 1B). This pattern was also observed with r^2 and cannot be explained by more or less diverse loci localizing to this region; Mann-Whitney U-test of median Simpson's diversity for the nine loci between B39 and B12 versus the remaining four loci was 15.0, P=0.711. Thus, these results highlighted one large region of the chromosome as having unusual patterns of variation in this population.

3.4. Genetic structure of the population

A neighbor-joining analysis of MLST and MLBT alleles (Fig. 2), as well as a separate neighbor-joining analysis of MLBT alleles and a goeBURST analysis of combined data (not shown), identified two major sequence clusters. One major cluster had 60% bootstrap support and was mostly composed of ST218 isolates (Fig. 2, bottom). A robust subcluster, composed of ST3523 isolates from Australia, nested within the ST218 cluster with 96% bootstrap support. The other major cluster had 67% bootstrap support and was mostly composed of ST220 isolates, along with several intermediate ST218 isolates (Fig. 2, top). Branching between the two major clusters were two ST221 isolates from Hungary. Subsequent analyses were restricted to the clusters themselves, due to the poor bootstrap support for most of the fine-scale branching structure within the clusters.

Interestingly, the ST3523 isolates from Australia differed by an average (±SD) of 0.443 (±0.758) *boxB* loci, indicating that they were within the range of variation that defined 12F outbreaks. These isolates were collected from Queensland and the Northern Territory between 1994 and 2005 and they likely include representatives from a 1993–1994 outbreak in the Alice Springs region of the Northern Territory (Gratten et al., 1995). These results pointed to ongoing transmission of closely related strains in Australia.

Further study of the geographic distribution of isolates revealed that 34 of 40 non-US isolates belonged to the ST218 cluster. Moreover, significant differences were detected in the distribution of the ST218 and ST220 clusters among the four US census regions (Fig. 3A). In particular, western US isolates were predominantly from the ST218 cluster, whereas isolates from the three more easterly US regions were predominately from the ST220 cluster. An overall non-zero Jost's *D* (95% CI) of 0.184 (0.059, 0.309) confirmed the differentiation of the two sequence clusters according to the four US census regions and, in pairwise geographic comparisons, the western region was significantly different from each of the three more easterly regions. Significant differences were also detected in the temporal distribution of the ST218 and ST220 clusters (Fig. 3B). In particular, pre-conjugate vaccine US isolates were predominantly from the ST218 cluster, whereas post-conjugate vaccine US isolates were predominantly from the ST220 cluster. Again, a non-zero Jost's D (95% CI) of 0.325 (0.111, 0.539) confirmed the differentiation of the two sequence clusters with respect to vaccine licensure date. In summary, these results revealed that the 12F, CC218 population was structured in space and time.

Several other characteristics distinguished between the ST218 and ST220 sequence clusters. The ST220 cluster was significantly more diverse in haplotypes than the ST218 cluster, both by Simpson's index and by the effective number of types (Table 5). Multilocus linkage disequilibrium was not detected in the ST220 cluster but it was highly significant in the ST218 cluster, both for all isolates and for single examples of each haplotype (Table 5). Finally, the ST220 cluster had substantial deficits of *boxB* diversity compared to neutral expectations, whereas the *boxB* diversity of the ST218 cluster was more consistent with that of an equilibrium population (Table 5). These data strongly suggested that these two sequence clusters had different population dynamics.

4. Discussion

Epidemiological studies indicate that conjugate vaccines are having a major impact on pneumococcal populations; serotypes covered by the vaccines are facing extinction, whereas non-vaccine serotypes are facing opportunities for expansion. Serotype 12F is a hyperinvasive, non-vaccine serotype (Sleeman et al., 2006) with an ability to cause outbreaks of invasive disease (CDC, 2005; Cherian et al., 1994; Hoge et al., 1994; Lee et al., 2005). Even in human populations without recognized 12F outbreaks, 12F isolates have been reported to be overrepresented from invasive disease in comparison to asymptomatic colonization (Kronenberg et al., 2006; Michel et al., 2005; Saha et al., 1997; Sandgren et al., 2004; Shouval et al., 2006; Sleeman et al., 2006). Like most potential replacement serotypes, the population structure of serotype 12F pneumococci has been largely unexplored. The goal of this study was to reveal the population structure within a single MLST-defined clonal complex of 12F pneumococci that has been a persistent cause of outbreaks in the US over the past two decades.

Tools for studying genetic relationships of pathogenic bacteria have dramatically improved over the last decade. MLVA typing schemes can be highly discriminatory (Vergnaud and Pourcel, 2009), and one such scheme has been useful for confirming relatedness of pneumococcal outbreak isolates of serotypes 1 (Yaro et al., 2006) and 5 (Pichon et al., 2010). However, size homoplasy in MLVA data is not ideal for certain population genetics applications, though this shortcoming might be somewhat offset by typing large numbers of loci (Estoup et al., 2002). Here, 10 *boxB* minisatellite loci were incorporated into a multilocus *boxB* sequence typing (MLBT) scheme that proved to have multiple uses. Sequences from these minisatellites distinguished each of the most recent US 12F, CC218 outbreaks from each other, yet they remained stable during limited passage in the laboratory; thus, we conclude that these minisatellites meet the standards of a valid molecular epidemiological marker (Riley, 2004).

The higher levels of diversity detected by MLBT compared to MLST were possible consequences of both our ascertainment procedure of selecting variable *boxB* loci and higher mutation rates of some *boxB* loci. It has been suggested previously that microsatellites (i.e. tandem repeats of ~10 bp or shorter) from human populations may evolve so rapidly that they can overcome ascertainment bias in diversity measures (Rogers and Jorde, 1996). Like their eukaryotic counterparts, microsatellites from bacterial populations can be of different mutation rates and they generally display a positive correlation between repeat array size and diversity (Farlow et al., 2002; Vogler et al., 2006). Minisatellites (i.e. tandem repeats of ~10–100 bp) are subject to complicated dynamics and need to be evaluated on a locus-by-locus basis (Supply et al., 2000). Our results showed that the amount of *boxB* minisatellite diversity among discovery isolates tended to be positively correlated with diversity in the final sample, and they confirmed the expected relationship of more repeats yielding more diversity. Potential biological explanations for these observations include a higher likelihood of DNA polymerase slippage or recombination for longer repeat arrays.

We observed that several *boxB* repeat sequences similar to the consensus sequence occurred at multiple loci. It is possible that these repeats are more mobilizable, or they are older repeats that represent the founders of different *boxB* loci, or they originated independently at different loci. These particular repeats might also provide BOX elements with optimal structure/function. It is known that the repeat array configuration of BOX elements can impact expression of downstream genes, with longer *boxB* repeat arrays reducing expression (Knutsen et al., 2006). Some BOX elements that are extensions to the 5' or 3' regions of operons appear to be transcribed (Croucher et al., 2011). Thus, *boxB* polymorphisms might not be entirely neutral from an evolutionary viewpoint.

Multilocus linkage disequilibrium was strong, indicating a relatively clonal population or a cryptically subdivided population (Maynard Smith et al., 1993). However, using r^2 and HR^2, a negative correlation was observed between pairwise linkage disequilibrium and the physical distance between loci, whereas no such correlation was observed for *D*'. These results are consistent with the evidence that multilocus measures of linkage disequilibrium do not reveal the full extent of recombination in bacterial populations (Maynard Smith, 1994) and that r^2 is a more sensitive test for recombination than *D*' (Meunier and Eyre-

Walker, 2001). These results also indicated that the newer measure, HR^2, provides a useful test for recombination in bacterial populations.

Another interesting observation was that most loci with the strongest pairwise linkage disequilibrium mapped to one large region of the chromosome. Our previous work had pointed to recombinations that included the *pspA* locus in 12F outbreak strains (Robinson et al., 1999), and we note that *pspA* maps to this region of the chromosome (Fig. 1B). Recombination in pneumococci may tend to introduce multiple fragments of 13 to 28 kb of donor DNA into the recipient's chromosome *in vivo* (Hiller et al., 2010), so there could be several recombinations throughout this large region. One possible explanation for the existence of linkage disequilibrium despite the previous evidence for recombination in this region involves selective sweeps at one or more loci in this region. If the 12F polysaccharide itself is a primary determinant for the hyperinvasive character of these strains, a selective sweep centered on the capsule locus would be feasible; the capsule locus also maps to this region of the chromosome (Fig. 1B). However, selective sweeps will not likely explain the pairwise linkage disequilibrium throughout the entire region because it is not accompanied by lower diversity throughout. Genome sequencing can be used to test the hypothesis that this region of the chromosome exhibits unusual patterns of variation among these pneumococci.

The east-west geographic structuring of 12F, CC218 pneumococci in the US was unexpected. It was reported previously that different clonal complexes of hyperinvasive serotype 1 had very different global geographic distributions (Brueggemann and Spratt, 2003) but, to our knowledge, our study is the first to demonstrate geographic differentiation within a single pneumococcal clonal complex or within a single continent. These phylogeographic patterns may be detectable because of the hyperinvasive character of these serotypes, which may involve a more transient carriage relative to less virulent serotypes (Sleeman et al., 2006). The temporal structuring of the population was less surprising. Nominal increases in the number of clonal complexes expressing 12F polysaccharide were reported in the US after licensure of the conjugate vaccine (Beall et al., 2006). However, our results are unique in that they describe a significant temporal shift in the frequency of sequence clusters within the predominant 12F clonal complex in the US. Taken together, these results indicate that the MLBT scheme developed here might be useful in identifying emerging sequence clusters in other pneumococcal clonal complexes.

Highlights

> Hyperinvasive pneumococci not targeted by conjugate vaccines should be monitored

> We developed a minisatellite sequence typing method to study 12F, CC218 pneumococci

> Hidden population structure was revealed among these hyperinvasive pneumococci

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank Shilpa Reddy for technical assistance. We also gratefully acknowledge Ellen Lee (New York City Department of Health and Mental Hygiene), Denise Murphy (Queensland Health Services), Gary Lum (Royal Darwin Hospital), Karen Rudolph (CDC, Arctic Investigations Program), Bernie Beall and the *Streptococcus*

Laboratory (CDC, Respiratory Diseases Branch) as well as the clinicians, microbiologists, and investigators of the ABCs program, Sanofi-Aventis, and David Farrell (JMI Laboratories) for providing bacterial isolates. This work was supported by NIH grant AI079345 (to D.A.R.).

References

Beall B, McEllistrem MC, Gertz RE Jr, Wedel S, Boxrud DJ, Gonzalez AL, Medina MJ, Pai R, Thompson TA, Harrison LH, McGee L, Whitney CG. Active Bacterial Core Surveillance Team. Pre- and postvaccination clonal compositions of invasive pneumococcal serotypes for isolates collected in the United States in 1999, 2001, and 2002. J. Clin. Microbiol. 2006; 44:999–1017. [PubMed: 16517889]

Bonnet E, Van de Peer Y. zt: a software tool for simple and partial Mantel tests. J. Statistical Software. 2002; 7:1–12.

Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007; 23:2633–2635. [PubMed: 17586829]

Brueggemann AB, Spratt BG. Geographic distribution and clonal diversity of *Streptococcus pneumoniae* serotype 1 isolates. J. Clin. Microbiol. 2003; 41:4966–4970. [PubMed: 14605125]

CDC. Outbreak of invasive pneumococcal disease - Alaska, 2003–2004. MMWR Morb. Mortal. Wkly. Rep. 2005; 54:72–75. [PubMed: 15674187]

Chao A, Shen T-J. Program SPADE (Species Prediction and Diversity Estimation). 2003

Cherian T, Steinhoff MC, Harrison LH, Rohn D, McDougal LK, Dick J. A cluster of invasive pneumococcal disease in young children in child care. JAMA. 1994; 271:695–697. [PubMed: 8309033]

Chiba N, Morozumi M, Sunaoshi K, Takahashi S, Takano M, Komori T, Sunakawa K, Ubukata K. IPD Surveillance Study Group. Serotype and antibiotic resistance of isolates from patients with invasive pneumococcal disease in Japan. Epidemiol. Infect. 2010; 138:61–68. [PubMed: 19538821]

Cornuet JM, Luikart G. Description and power analysis of two tests for detecting recent population bottlenecks from allele frequency data. Genetics. 1997; 144:2001–2014. [PubMed: 8978083]

Croucher NJ, Harris SR, Fraser C, Quail MA, Burton J, van der Linden M, McGee L, von Gottberg A, Song JH, Ko KS, Pichon B, Baker S, Parry CM, Lambertsen LM, Shahinas D, Pillai DR, Mitchell TJ, Dougan G, Tomasz A, Klugman KP, Parkhill J, Hanage WP, Bentley SD. Rapid pneumococcal evolution in response to clinical interventions. Science. 2011; 331:430–434. [PubMed: 21273480]

Croucher NJ, Vernikos GS, Parkhill J, Bentley SD. Identification, variation and transcription of pneumococcal repeat sequences. BMC Genomics. 2011; 12:120–133. [PubMed: 21333003]

Di Rienzo A, Peterson AC, Garza JC, Valdes AM, Slatkin M, Freimer NB. Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA. 1994; 91:3166–3170. [PubMed: 8159720]

Enright MC, Spratt BG. A multilocus sequence typing scheme for *Streptococcus pneumoniae*: identification of clones associated with serious invasive disease. Microbiology. 1998; 144:3049–3060. [PubMed: 9846740]

Estoup A, Jarne P, Cornuet JM. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics analysis. Mol. Ecol. 2002; 11:1591–1604. [PubMed: 12207711]

Farlow J, Postic D, Smith KL, Jay Z, Baranton G, Keim P. Strain typing of *Borrelia burgdorferi*, *Borrelia afzelii*, and *Borrelia garinii* by using multiple-locus variable-number tandem repeat analysis. J. Clin. Microbiol. 2002; 40:4612–4618. [PubMed: 12454161]

Farnir F, Coppieters W, Arranz JJ, Berzi P, Cambisano N, Grisart B, Karim L, Marcq F, Moreau L, Mni M, Nezer C, Simon P, Vanmanshoven P, Wagenaar D, Georges M. Extensive genome-wide linkage disequilibrium in cattle. Genome Res. 2000; 10:220–227. [PubMed: 10673279]

Farrell DJ, Jenkins SG. Distribution across the USA of macrolide resistance and macrolide resistance mechanisms among *Streptococcus pneumoniae* isolates collected from patients with respiratory tract infections: PROTEKT US 2001–2002. J. Antimicrob. Chemother. 2004; 54 Suppl.:i17–i22. [PubMed: 15265832]

Feil EJ, Li BC, Aanensen DM, Hanage WP, Spratt BG. eBURST: inferring patterns of evolutionary descent among clusters of related bacterial genotypes from multilocus sequence typing data. J. Bacteriol. 2004; 186:1518–1530. [PubMed: 14973027]

Foster D, Walker AS, Paul J, Griffiths D, Knox K, Peto TE, Crook DW. Oxford Invasive Pneumococcal Surveillance Group. Reduction in invasive pneumococcal disease following implementation of the conjugate vaccine in the Oxfordshire region, England. J. Med. Microbiol. 2011; 60:91–97. [PubMed: 20864548]

Francisco AP, Bugalho M, Ramirez M, Carriço JA. Global optimal eBURST analysis of multilocus typing data using a graphic matroid approach. BMC Bioinformatics. 2009; 10:152. [PubMed: 19450271]

Gratten M, Morey F, Dixon J, Torzillo P, Erlich J. Invasive type 12F pneumococcal disease in central Australia. Commun. Dis. Intell. 1995; 19:470–472.

Grundmann H, Hori S, Tanner G. Determining confidence intervals when measuring genetic diversity and the discriminatory abilities of typing methods for microorganisms. J. Clin. Microbiol. 2001; 39:4190–4192. [PubMed: 11682558]

Haubold B, Hudson RR. LIAN 3.0: detecting linkage disequilibrium in multilocus data. Linkage Analysis. Bioinformatics. 2000; 16:847–848. [PubMed: 11108709]

Hicks LA, Harrison LH, Flannery B, Hadler JL, Schaffner W, Craig AS, Jackson D, Thomas A, Beall B, Lynfield R, Reingold A, Farley MM, Whitney CG. Incidence of pneumococcal disease due to non-pneumococcal conjugate vaccine (PCV7) serotypes in the United States during the era of widespread PCV7 vaccination, 1998–2004. J. Infect. Dis. 2007; 196:1346–1354. [PubMed: 17922399]

Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genet. Res. 1966; 8:269–294. [PubMed: 5980116]

Hiller NL, Ahmed A, Powell E, Martin DP, Eutsey R, Earl J, Janto B, Boissy RJ, Hogg J, Barbadora K, Sampath R, Lonergan S, Post JC, Hu FZ, Ehrlich GD. Generation of genic diversity among *Streptococcus pneumoniae* strains via horizontal gene transfer during a chronic polyclonal pediatric infection. PLoS Pathog. 2010; 6 e1001108.

Hoge CW, Reichler MR, Dominguez EA, Bremer JC, Mastro TD, Hendricks KA, Musher DM, Elliott JA, Facklam RR, Breiman RF. An epidemic of pneumococcal disease in an overcrowded, inadequately ventilated jail. N. Engl. J. Med. 1994; 331:643–648. [PubMed: 8052273]

Hunter PR, Gaston MA. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. 1988; 26:2465–2466. [PubMed: 3069867]

Isaacman DJ, Fletcher MA, Fritzell B, Ciuryla V, Schranz J. Indirect effects associated with widespread vaccination of infants with heptavalent pneumococcal conjugate vaccine (PCV7; Prevnar). Vaccine. 2007; 25:2420–2427. [PubMed: 17049677]

Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB. Genotype, haplotype and copy-number variation in worldwide human populations. Nature. 2008; 451:998–1003. [PubMed: 18288195]

Joseph SJ, Read TD. Bacterial population genomics and infectious disease diagnostics. Trends Biotechnol. 2010; 28:611–618. [PubMed: 20961641]

Jost L. G_{ST} and its relatives do not measure differentiation. Mol. Ecol. 2008; 17:4015–4026. [PubMed: 19238703]

Kimura M, Crow JF. The number of alleles that can be maintained in a finite population. Genetics. 1964; 49:725–738. [PubMed: 14156929]

Knutsen E, Johnsborg O, Quentin Y, Claverys JP, Håvarstein LS. BOX elements modulate gene expression in *Streptococcus pneumoniae*: impact on the fine-tuning of competence development. J. Bacteriol. 2006; 188:8307–8312. [PubMed: 16997972]

Koeck JL, Njanpop-Lafourcade BM, Cade S, Varon E, Sangare L, Valjevac S, Vergnaud G, Pourcel C. Evaluation and selection of tandem repeat loci for *Streptococcus pneumoniae* MLVA strain typing. BMC Microbiol. 2005; 5:66. [PubMed: 16287512]

Kronenberg A, Zucs P, Droz S, Mühlemann K. Distribution and invasiveness of *Streptococcus pneumoniae* serotypes in Switzerland, a country with low antibiotic selection pressure, from 2001 to 2004. J. Clin. Microbiol. 2006; 44:2032–2038. [PubMed: 16757594]

Lee, EH.; Hosea, S.; Schulman, E.; Bellomy, A.; Jackson, D.; Glass, N.; Nguyen, D.; Sekhar, J.; Kimura, A.; Feikin, D. *Streptococcus pneumoniae* serotype 12F outbreak in a homeless population - California, 2004; 54th Annual Epidemic Intelligence Service (EIS) Conference; 2005. p. 79Abstract

Lewontin RC. The interaction of selection and linkage. I. General considerations; Heterotic models. Genetics. 1964; 49:49–67. [PubMed: 17248194]

Lexau CA, Lynfield R, Danila R, Pilishvili T, Facklam R, Farley MM, Harrison LH, Schaffner W, Reingold A, Bennett NM, Hadler J, Cieslak PR, Whitney CG. Active Bacterial Core Surveillance Team. Changing epidemiology of invasive pneumococcal disease among older adults in the era of pediatric pneumococcal conjugate vaccine. JAMA. 2005; 294:2043–2051. [PubMed: 16249418]

Luikart G, Allendorf FW, Cornuet JM, Sherwin WB. Distortion of allele frequency distributions provides a test for recent population bottlenecks. J. Hered. 1998; 89:238–248. [PubMed: 9656466]

Martin B, Humbert O, Camara M, Guenzi E, Walker J, Mitchell T, Andrew P, Prudhomme M, Alloing G, Hakenbeck R, Morrison DA, Boulnois GJ, Claverys J-P. A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res. 1992; 20:3479–3483. [PubMed: 1630918]

Maynard Smith J, Smith NH, O'Rourke M, Spratt BG. How clonal are bacteria? Proc. Natl. Acad. Sci. USA. 1993; 90:4384–4388. [PubMed: 8506277]

Maynard Smith J. Estimating the minimum rate of genetic transformation in bacteria. J. Evol. Biol. 1994; 7:525–534.

Meunier J, Eyre-Walker A. The correlation between linkage disequilibrium and distance: implications for recombination in hominid mitochondria. Mol. Biol. Evol. 2001; 18:2132–2135. [PubMed: 11606711]

Michel N, Watson M, Baumann F, Perolat P, Garin B. Distribution of *Streptococcus pneumoniae* serotypes responsible for penicillin resistance and the potential role of new conjugate vaccines in New Caledonia. J. Clin. Microbiol. 2005; 43:6060–6063. [PubMed: 16333099]

Nielsen R, Tarpy DR, Reeve HK. Estimating effective paternity number in social insects and the effective number of alleles in a population. Mol. Ecol. 2003; 12:3157–3164. [PubMed: 14629394]

O’Brien KL, Wolfson LJ, Watt JP, Henkle E, Deloria-Knoll M, McCall N, Lee E, Mulholland K, Levine OS, Cherian T. Hib and Pneumococcal Global Burden of Disease Study Team. Burden of disease caused by *Streptococcus pneumoniae* in children younger than 5 years: global estimates. Lancet. 2009; 374:893–902. [PubMed: 19748398]

Ohta T, Kimura M. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. 1973; 22:201–204. [PubMed: 4777279]

Pai R, Moore MR, Pilishvili T, Gertz RE, Whitney CG, Beall B. Active Bacterial Core Surveillance Team. Postvaccine genetic structure of *Streptococcus pneumoniae* serotype 19A from children in the United States. J. Infect. Dis. 2005; 192:1988–1995. [PubMed: 16267772]

Pai R, Gertz RE, Beall B. Sequential multiplex PCR approach for determining capsular serotypes of *Streptococcus pneumoniae* isolates. J. Clin. Microbiol. 2006; 44:124–131. [PubMed: 16390959]

Pichon B, Moyce L, Sheppard C, Slack M, Turbitt D, Pebody R, Spencer DA, Edwards J, Krahé D, George R. Molecular typing of pneumococci for investigation of linked cases of invasive pneumococcal disease. J. Clin. Microbiol. 2010; 48:1926–1928. [PubMed: 20164267]

Poehling KA, Talbot TR, Griffin MR, Craig AS, Whitney CG, Zell E, Lexau CA, Thomas AR, Harrison LH, Reingold AL, Hadler JL, Farley MM, Anderson BJ, Schaffner W. Invasive pneumococcal disease among infants before and after introduction of pneumococcal conjugate vaccine. JAMA. 2006; 295:1668–1674. [PubMed: 16609088]

Riley, LW. Molecular epidemiology of infectious diseases: principles and practices. Washington, DC: ASM Press; 2004.

Robinson DA, Turner JS, Facklam RR, Parkinson AJ, Breiman RF, Gratten M, Steinhoff MC, Hollingshead SK, Briles DE, Crain MJ. Molecular characterization of a globally distributed lineage of serotype 12F *Streptococcus pneumoniae* causing invasive disease. J. Infect. Dis. 1999; 179:414–422. [PubMed: 9878026]

Rogers AR, Jorde LB. Ascertainment bias in estimates of average heterozygosity. Am. J. Hum. Genet. 1996; 58:1033–1041. [PubMed: 8651264]

Sabatti C, Risch N. Homozygosity and linkage disequilibrium. Genetics. 2002; 160:1707–1719. [PubMed: 11973323]

Saha SK, Rikitomi N, Biswas D, Watanabe K, Ruhulamin M, Ahmed K, Hanif M, Matsumoto K, Sack RB, Nagatake T. Serotypes of *Streptococcus pneumoniae* causing invasive childhood infections in Bangladesh, 1992 to 1995. J. Clin. Microbiol. 1997; 35:785–787. [PubMed: 9041437]

Sandgren A, Sjostrom K, Olsson-Liljequist B, Christensson B, Samuelsson A, Kronvall G, Henriques-Normark B. Effect of clonal and serotype-specific properties on the invasive capacity of *Streptococcus pneumoniae*. J. Infect. Dis. 2004; 189:785–796. [PubMed: 14976594]

Shouval DS, Greenberg D, Givon-Lavi N, Porat N, Dagan R. Site-specific disease potential of individual *Streptococcus pneumoniae* serotypes in pediatric invasive disease, acute otitis media and acute conjunctivitis. Pediatr. Infect. Dis. J. 2006; 25:602–607. [PubMed: 16804429]

Skoczyńska A, Sadowy E, Bojarska K, Strzelecki J, Kuch A, Gołębiewska A, Waśko I, Foryś M, van der Linden M, Hryniewicz W. Participants of a laboratory-based surveillance of community acquired invasive bacterial infections (BINet). The current status of invasive pneumococcal disease in Poland. Vaccine. 2011; 29:2199–2205. [PubMed: 20943207]

Sleeman KL, Griffiths D, Shackley F, Diggle L, Gupta S, Maiden MC, Moxon ER, Crook DW, Peto TE. Capsular serotype-specific attack rates and duration of carriage of *Streptococcus pneumoniae* in a population of children. J. Infect. Dis. 2006; 194:682–688. [PubMed: 16897668]

Smyth DS, McDougal LK, Gran FW, Manoharan A, Enright MC, Song JH, de Lencastre H, Robinson DA. Population structure of a hybrid clonal group of methicillin-resistant *Staphylococcus aureus*, ST239-MRSA-III. PLoS ONE. 2010; 5:e8582. [PubMed: 20062529]

Supply P, Mazars E, Lesjean S, Vincent V, Gicquel B, Locht C. Variable human minisatellite-like regions in the *Mycobacterium tuberculosis* genome. Mol. Microbiol. 2000; 36:762–771. [PubMed: 10844663]

Swofford, DL. Version 4.0b10. Sunderland, MA: Sinauer Associates; 2003. PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods).

Takezaki N, Nei M, Tamura K. POPTREE2: Software for constructing population trees from allele frequency data and computing other population statistics with Windows interface. Mol. Biol. Evol. 2010; 27:747–252. [PubMed: 20022889]

Tettelin H, Nelson KE, Paulsen IT, Eisen JA, Read TD, Peterson S, Heidelberg J, DeBoy RT, Haft DH, Dodson RJ, Durkin AS, Gwinn M, Kolonay JF, Nelson WC, Peterson JD, Umayam LA, White O, Salzberg SL, Lewis MR, Radune D, Holtzapple E, Khouri H, Wolf AM, Utterback TR, Hansen CL, McDonald LA, Feldblyum TV, Angiuoli S, Dickinson T, Hickey EK, Holt IE, Loftus BJ, Yang F, Smith HO, Venter JC, Dougherty BA, Morrison DA, Hollingshead SK, Fraser CM. Complete genome sequence of a virulent isolate of *Streptococcus pneumoniae*. Science. 2001; 293:498–506. [PubMed: 11463916]

Vergnaud G, Pourcel C. Multiple locus variable number of tandem repeats analysis. Methods Mol. Biol. 2009; 551:141–158. [PubMed: 19521873]

Vogler AJ, Keys C, Nemoto Y, Colman RE, Jay Z, Keim P. Effect of repeat copy number on variable-number tandem repeat mutations in *Escherichia coli* O157:H7. J. Bacteriol. 2006; 188:4253–4263. [PubMed: 16740932]

Weatherholtz R, Millar EV, Moulton LH, Reid R, Rudolph K, Santosham M, O'Brien KL. Invasive pneumococcal disease a decade after pneumococcal conjugate vaccine use in an American Indian population at high risk for disease. Clin. Infect. Dis. 2010; 50:1238–1246. [PubMed: 20367225]

Whitney CG, Farley MM, Hadler J, Harrison LH, Bennett NM, Lynfield R, Reingold A, Cieslak PR, Pilishvili T, Jackson D, Facklam RR, Jorgensen JH, Schuchat A. Active Bacterial Core Surveillance of the Emerging Infections Program Network. Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine. N. Engl. J. Med. 2003; 348:1737–1746. [PubMed: 12724479]

Yaro S, Lourd M, Traoré Y, Njanpop-Lafourcade BM, Sawadogo A, Sangare L, Hien A, Ouedraogo MS, Sanou O, Parent du Châtelet I, Koeck JL, Gessner BD. Epidemiological and molecular characteristics of a highly lethal pneumococcal meningitis epidemic in Burkina Faso. Clin. Infect. Dis. 2006; 43:693–700. [PubMed: 16912941]

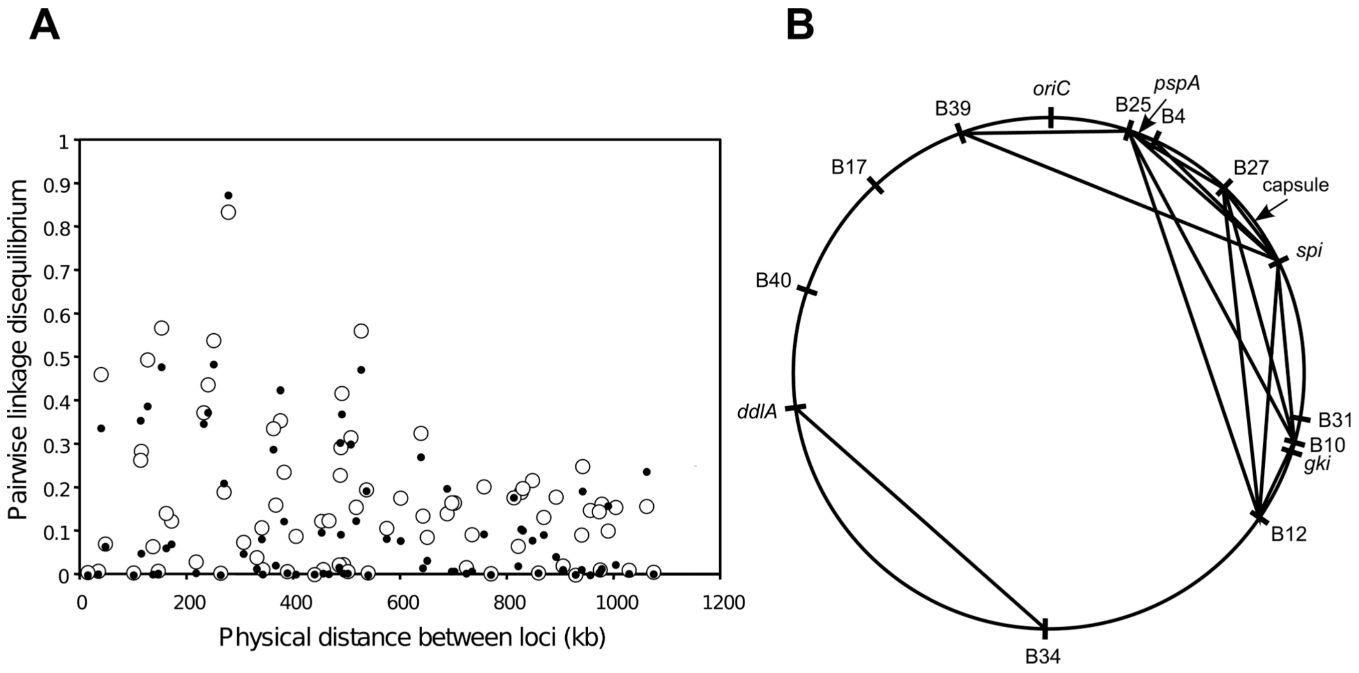

Fig. 1.

Analysis of pairwise linkage disequilibrium. A, Relationship between pairwise linkage disequilibrium using r^2 (open circles) and HR^2 (black dots) and the physical distance between loci in the chromosome. B, Chromosomal positions of the 13 variable MLST and MLBT loci, plus *pspA* and capsule loci. Coordinates based on TIGR4 genome sequence. Lines connect the 15 pairs of loci with the strongest pairwise disequilibrium (HR^2>0.25).

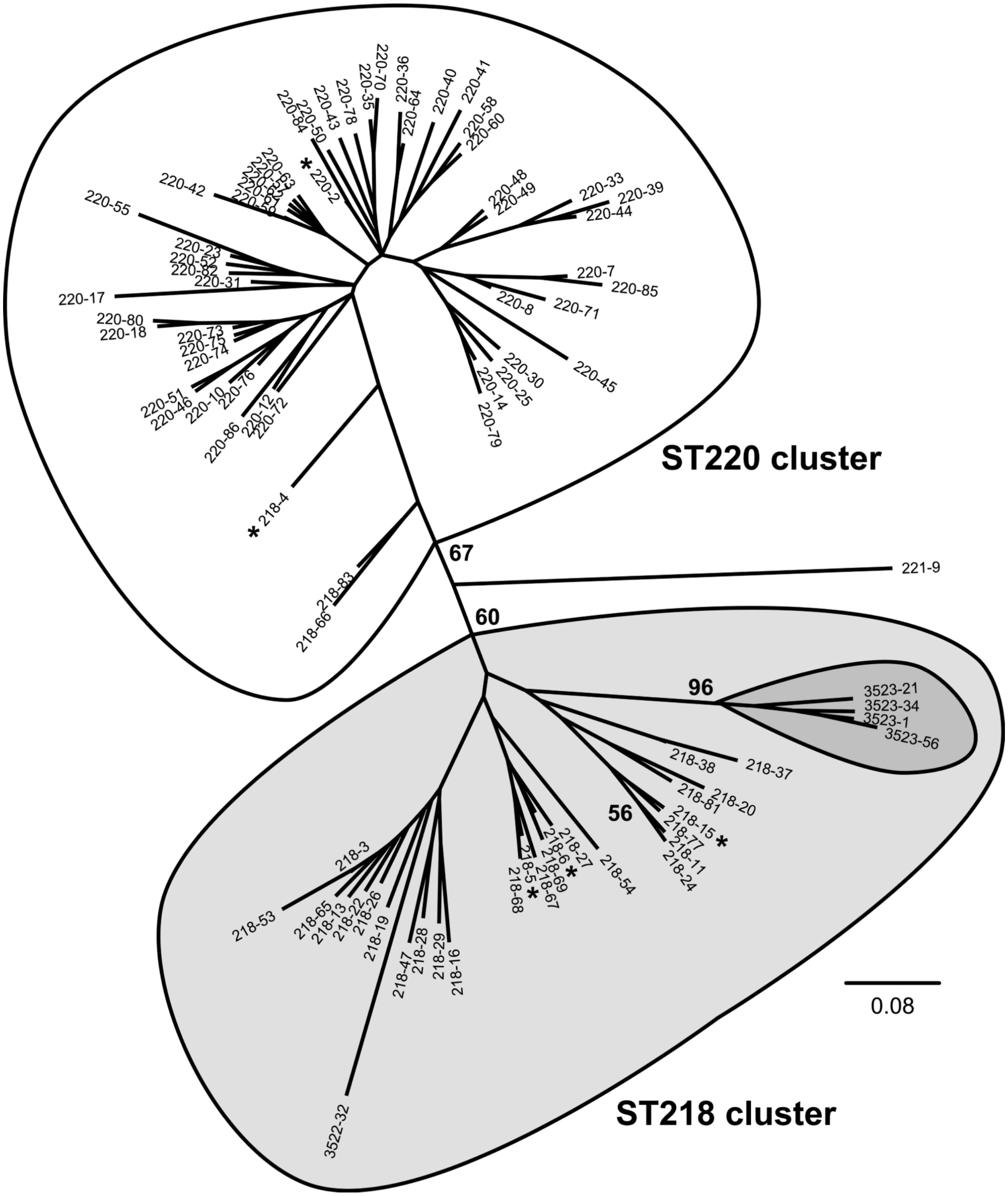

Fig. 2.

Sequence clusters within serotype 12F, clonal complex 218 pneumococci. Tree is based on neighbor-joining analysis of *p*-distances from the alleles of the 13 variable MLST and MLBT loci. The 86 haplotypes are named according to ST-BT. Outbreak haplotypes are identified by asterisks. Numbers on branches are bootstrap proportions >50%. Two major sequence clusters and one subcluster are highlighted.

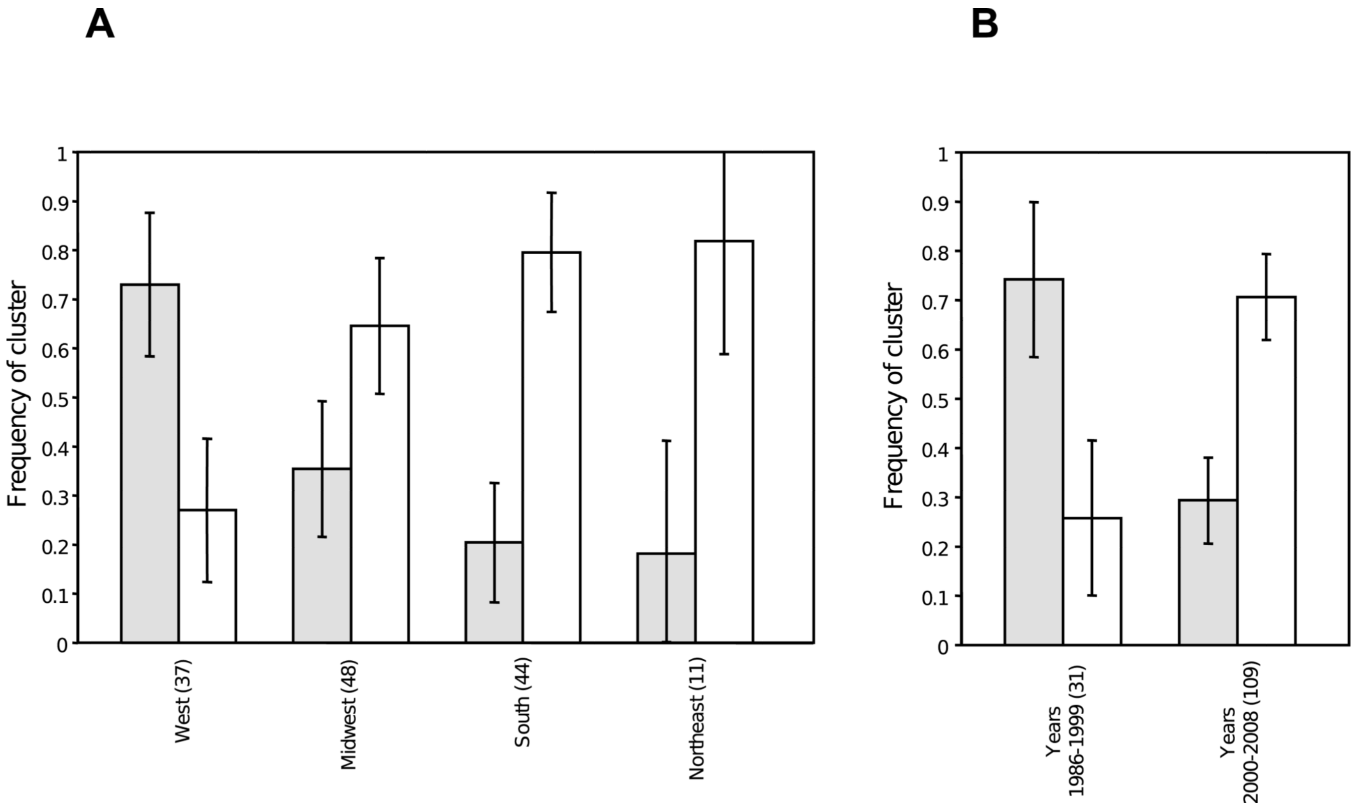

Fig. 3.

Frequencies of two sequence clusters among various geographically- and temporally-defined subpopulations. The ST218 (grey) and ST220 (white) clusters are defined in Figure 2. A, Frequencies among the four US census regions. B, Frequencies according to pre- and post-licensure of the first pneumococcal conjugate vaccine in the US. Black bars indicate 95% CIs. Numbers in parentheses indicate number of isolates in each subpopulation.

Table 1

Genetic characterization of outbreak isolates by MLST and MLBT.

					***boxB* repeat profile at the following loci**[b]									
Outbreak	**No. of isolates**	**MLST type**	**MLST allelic profile**[a]	***boxB* type**	**B25**	**B27**	**B31**	**B34**	**B40**	**B17**	**B39**	**B4**	**B10**	**B12**
Texas 1989	9	ST218	10-20-14-1-6-1-29	BT4	5-1-2	9–10	5–12	5–16	7-17-5	20-19-10-21	22-22-23-14	71-67-71-71-71-71-72	77-78-79	82
California 2004	5	ST220	10-20-14-1-9-1-29	BT2	5-1-2	9–10	5–12	5–16	7-7-17-5	20-19-10-21	22-22-23-14	71-67-71-71-71-71-72	77-78-79	26-26-81-30-82
Maryland 1992	5	ST218	10-20-14-1-6-1-29	BT6	4	9-9-9-9-11	5–12	5–16	7	20-19-10-21	22-23-14	5-99-72	78–79	82-81-82-82
	1	ST218	10-20-14-1-6-1-29	BT15	4	9-9-9-9-11	5–12	5-5-16	7	20-19-10-21	22-23-14	5-99-72	78–79	82-81-82-82
Alaska 2006–2008	8	ST218	10-20-14-1-6-1-29	BT5	4	9-9-9-9-11	5–12	5–16	7	20-19-10-21	22–14	5-99-72	78–79	82-81-82-82

[a] Alleles at the seven housekeeping loci used for conventional MLST: *aro-gdh-gki-rec-spi-xpt-ddl*

[b] Alleles at the 10 *boxB* loci used for MLBT. Numbers correspond to *boxB* repeat sequences as defined in Table 3.

Table 2

Diversity at the 10 *boxB* loci used for MLBT.

	No. of alleles		No. of repeats		Measures of diversity (95% CI)[a]	
Locus	Among discovery isolates	Among all isolates	Range	Average per isolate[a]	Simpson's index	Effective no. of types
B25	2	3	1, 3, 4	2.02	0.514 (0.499, 0.529)	2.06 (1.99, 2.12)
B27	4	6	2–5	2.77	0.682 (0.629, 0.735)	3.15 (2.94, 3.36)
B31	1	6	1–3	1.99	0.139 (0.069, 0.209)	1.16 (0.95, 1.37)
B34	2	6	2–4	2.42	0.470 (0.386, 0.553)	1.88 (1.67, 2.10)
B40	4	17	1–7	4.17	0.846 (0.822, 0.870)	6.50 (5.43, 7.57)
B17	2	5	2–5	3.84	0.341 (0.262, 0.419)	1.52 (1.36, 1.67)
B39	3	16	1–7	3.12	0.754 (0.712, 0.796)	4.07 (3.09, 5.04)
B4	3	19	1–9, 11, 12	4.87	0.704 (0.649, 0.758)	3.37 (2.10, 4.64)
B10	2	6	2–4	2.51	0.634 (0.594, 0.674)	2.73 (2.52, 2.94)
B12	3	16	1–6	3.83	0.766 (0.727, 0.805)	4.27 (3.29, 5.24)

[a] Calculations based on the subset of 180 isolates

Table 3

Alignment of all 44 unique *boxB* repeat sequences among study isolates.

Repeat	Sequence[a]	SNPs from consensus	Occurrence of repeat at the following loci
b16	CTGACTTCGTCAGTTCTATCTACAACCTCAAAACAGTGTTTTGAG	0	B34,B39
	....................C........................		
b5		1	B25,B31,B34,B40,B39,B4
	.....................G.......................		
b7	..................................G..........	1	B25,B40,B39
b67	.....................................A.......	1	B4
	.C..................C........................		
b62		1	B39
	......A.........C............................		
b18	......A.............C........................	2	B40
b12	.......T..........................G..........	2	B31
	........A...........C........................		
b15		2	B31
	........A............G.......................		
b113	...............TC............................	2	B4
b99	................C...C........................	2	B4
	................C...........T................		
b72	....................C.....T..................	2	B4
b26	....................C.............G..........	2	B12
	....................C................A.......		
b8		2	B25
	....................C.......................G.		
b2	.....................G...............A.......	2	B25
b34	.....................G................C......	2	B39
	.....................G....................T..		
b71		2	B4
	.....................G.......................T		
b22	....G...C.	2	B40,B39
b24	...............TC.................T..........	2	B39
	................C...C.......T................		
b25		2	B40
	................C...........T............A...		
b11	....................C.T...T..................	2	B27
b9	....................C..............CC........	2	B27
	.....................G..........G.....C......		
b17		2	B40
	.....................G....................T.A		
b21	.................................ACA.........	2	B17
	...............TC...............G.....C......		
b82		3	B12
	...............TC................AC..........		
b1	...............TC..................CA.........	3	B25
b4	................C...C.......T......A.........	3	B25
	..................C..............ACA.........		
b14		3	B31,B39
	....................C.T...T..........A.......		
b107	...............TC................ACA.........	3	B4
b10	...............T................G.G..AC......	3	B27,B17
	...............T...T............G.G..AC......		
b35		3	B27
	...............T................G.TA.AC......		
b19	......C........T................G.TA.AC......	3	B17
b81	.......T.......T................G.TA.AC......	4	B12
	...............T...T............G.TA.AC......		
b112	* **** ****** * *** * *** *	4	B12
b29		4	B12

Repeat	Sequence[a]	SNPs from consensus	Occurrence of repeat at the following loci
b3		4	B25
b20		4	B17
b23		4	B39
b30		5	B12
b108		5	B10
b79		6	B10
b78		6	B10
b77		7	B10
b111		7	B10
b110		7	B10

[a] Dots indicate identity with the top sequence.

[*] indicates conserved site.

Table 4

Fit of three mutation models to the diversity at the 10 *boxB* loci used for MLBT.

	IAM		TPM		SMM		
Locus	**DH/sd**	***P***	**DH/sd**	***P***	**DH/sd**	***P***	**Best fit model**
B25	1.145	0.160	0.893	0.191	0.546	0.330	SMM
B27	0.911	0.193	0.477	0.371	−0.507	0.246	TPM
B31	−2.466	0.018[*]	−4.439	0.000[**]	−8.120	0.000[**]	IAM
B34	−0.361	0.304	−1.451	0.093	−3.311	0.014[*]	IAM
B40	0.406	0.413	−1.109	0.136	−2.532	0.013[*]	IAM
B17	−0.792	0.229	−1.706	0.081	−3.497	0.011[*]	IAM
B39	−0.808	0.181	−3.306	0.010[*]	−8.430	0.000[**]	IAM
B4	−2.576	0.026[*]	−7.089	0.000[**]	−13.088	0.001[**]	IAM
B10	0.630	0.316	0.052	0.447	−1.113	0.132	TPM
B12	−0.618	0.205	−3.332	0.008[*]	−8.050	0.000[**]	IAM

Note: Infinite alleles model (IAM), stepwise mutation model (SMM), two-phase model (TPM). DH/sd is the standardized difference between the observed and expected diversity, given the number of alleles (Table 2) and the sample size (*n*=180) at each locus, under each mutation model. Positive values indicate diversity excess and negative values indicate diversity deficit. Each model was tested with 1,000 coalescent simulations using BOTTLENECK v1.2.

[*] *P*<0.05,

[**] *P*<0.005 (Bonferroni level)

Table 5

Some genetic characteristics of the ST218 and ST220 sequence clusters.

Characteristic	ST218 cluster	ST220 cluster
No. isolates	89	89
No. haplotypes	31	54
Diversity of haplotypes (95% CI)		
Simpson's index	0.908 (0.865, 0.947)	0.980 (0.969, 0.991)
Effective no. of types	10.8 (5.5, 16.2)	49.0 (44.4, 53.5)
Multilocus linkage disequilibrium		
I_{AS} all isolates	0.191, *P*<0.001	0.016, *P*=0.097
I_{AS} haplotypes only	0.075, *P*<0.001	−0.004, *P*=0.568
boxB diversity deficit under the IAM	2 of 9 loci	8 of 8 loci