Identification and Evolution of an IS6110 Low-Copy-Number
*Mycobacterium tuberculosis* Cluster

Barun Mathema,1 Pablo J. Bifani,1 Jeffrey Driscoll,2 Lauren Steinlein,3 Natalia Kurepina,1 Soraya L. Moghazeh,1 Elena Shashkina,1 Salvatore A. Marras,1 Shannon Campbell,4 Bonita Mangura,5 Kenneth Shilkret,6 Jack T. Crawford,3 Soraya L. Moghazeh,1 Elena Shashkina,1 Barun Mathema,1 Pablo J. Bifani,1 Jeffrey Driscoll,2 Richard Frothingham,7 and Barry N. Kreiswirth1

1Public Health Research Institute Tuberculosis Center, New York, and 2Wadsworth Center, New York State Department of Health, Albany, New York; 3Centers for Disease Control and Prevention, Atlanta, Georgia; 4Isosceles Information Solutions, Manotick, Canada; 5New Jersey Medical School National Tuberculosis Center, Newark, and 6New Jersey Department of Health and Senior Services, Communicable Disease Service, Trenton; 7Durham Veterans Affairs Medical Center, Durham, North Carolina

A cohort of 56 patients infected with related strains of *Mycobacterium tuberculosis*, the S75 group, was identified in a New Jersey population–based study of all isolates with a low number of copies of the insertion element IS6110. Genotyping was combined with surveillance data to identify the S75 group and to elucidate its recent evolution. The S75 group had similar demographic and geographic characteristics. Seventeen persons (30%) were linked epidemiologically. The S75 group was segregated from other low-copy-number isolates on the basis of several independent molecular methods. This group included 3 IS6110 genotype variants: BE, H6, and C28, containing 1, 2, and 3 IS6110 insertions, respectively. IS6110 insertion site mapping and comparative sequence analysis strongly suggest a stepwise acquisition of IS6110 elements from BE to H6 to C28. S75 represents a locally produced strain cluster that has recently evolved. The combination of multiple molecular tools with traditional epidemiology provides novel insights into dissemination, local transmission, and evolution of *M. tuberculosis*.

The development and implementation of genotyping techniques have transformed epidemiologic investigations of disease caused by *Mycobacterium tuberculosis*. The most widely used and accepted method of generating genotypes is based on the IS6110 Southern blot hybridization technique, which is simple and reproducible and has been standardized among molecular biology laboratories [1]. Although conclusions are dependent on the epidemiologic setting, in general, matching IS6110 restriction fragment length polymorphism (RFLP) patterns indicate possibly recent transmission, whereas different patterns suggest reactivation of latent infection or no link with an index case or cluster.

Although genotyping of *M. tuberculosis* has proven to be reliable and robust in the evaluation of transmission dynamics of *M. tuberculosis*, limitations persist. One such limitation is the interpretation of molecular data to draw epidemiologic conclusions. This is particularly important when investigating clustering in isolated geographic areas or in regions of high incidence of tuberculosis; under these conditions, genetic clustering is not always synonymous with recent infection, because infection may involve a number of different transmission pathways [2–4]. Under such circumstances, the use of multiple and independent genetic markers in combination with surveillance data is recommended. Likewise, conventional field epidemiology involving contact tracing often fails to identify patterns of transmission, particularly when high incidence and marginalized social groups are involved [5].

Because tuberculosis in immunocompetent persons commonly has a long latency from time of exposure, epidemiologic links over time become less clear, unless cases are restricted to a small geographic area or are part of an outbreak. However, when molecular methods are applied with appropriate discretion, the validity and accuracy of conventional surveillance methods can be greatly enhanced. A number of studies have used molecular tools to identify patients involved in recent transmission that were not linked by specific source cases but rather by geographic aggregation [4, 6], demographic and social profiles, or unique molecular characteristics [7–9]. Some direct links were established for each of these reports, reinforcing the combined molecular and epidemiologic approach.

A second limitation is that the IS6110-based RFLP typing method is poor at discriminating *M. tuberculosis* isolates with <6 IS6110 insertions. Isolates with low numbers of copies of IS6110 (<6) and identical RFLP patterns were shown by means of secondary genotyping methods to be genetically distinct [10–12]. Consequently, studies involving low-copy-number isolates commonly use additional genotyping techniques in conjunction with IS6110 RFLP typing to better describe relatedness among clinical isolates [13, 14].
In the present study, we describe how these 2 limitations were overcome in investigating the transmission of a large low-copy-number cluster. As part of an ongoing population-based molecular epidemiologic study of *M. tuberculosis*, we sought to elucidate relationships among our low-copy-number *M. tuberculosis* isolates via extensive molecular characterization. The combined strength of molecular techniques was used to guide traditional epidemiologic investigation, uncovering linkages within a patient population and revealing the evolution of this mycobacterial strain cluster.

**Methods**

**Study population.** This population-based study included all incident pulmonary culture–positive cases of tuberculosis reported to the New Jersey Department of Health and Senior Services between January 1996 and March 2000 as a part of the Centers for Diseases Control and Prevention National Tuberculosis Genotyping and Surveillance Network. The genotyping was done at the Public Health Research Institute (PHRI) Tuberculosis Center, New York City. During the study period, 2224 (78%) of 2838 counted active tuberculosis surveillance network. The genotyping was done in accordance with standard methods, described elsewhere [1], for both the 5’ and 3’ fragments of the IS6110 genetic element. The hybridization patterns were compared on a Sun Sparc5 Workstation (Sun Microsystems) with Whole Band Analyzer software, version 3.4 (BioImage). In brief, the nomenclature used for classifying IS6110 RFLP patterns was as follows: 2 isolates with an identical IS6110 banding pattern were assigned the same arbitrary letter code (e.g., H, P, or AB), which started with the first observed cluster, strain A. IS6110 patterns that were similar but not identical were denoted by the addition of a number (e.g., H2, BE1, or W4).

Purified chromosomal DNA recovered from the 381 isolates and used in the IS6110 RFLP analysis was also restricted with Alu I and was hybridized with a PGRS-specific probe, as described elsewhere [15, 16]. For both IS6110 and PGRS analysis, patterns were compared for relatedness and subsequent subgrouping. For both IS6110 and PGRS Southern blot hybridization experiments, strain H37Rv (ATCC 25177) was used as a control.

**Spacer oligonucleotide genotyping (spoligotyping).** The direct repeat (DR) region of each isolate was amplified by polymerase chain reaction (PCR) and was used as a probe against a spoligotyping membrane, which was prepared as described elsewhere [17–19]. The nomenclature in this study followed that assigned by the Centers for Disease Control and Prevention (CDC), in which isolates are labeled with an arbitrary number (e.g., 0034, 0075, or 0125) as well as the international octal code [20]. Spoligotyping was done at the Wadsworth Center (New York State Department of Health), where a database has been developed from >3000 individual *M. tuberculosis* isolates.

**Determination of IS6110 flanking regions.** To identify the location of the IS6110 insertion in the chromosome of studied strains, the modification of the conventional inverse PCR method was used. The *Pvu* II (New England BioLabs) fragments of interest were separated on 1% SeaPlaque (FMC BioProducts) x1 TAE, excised from the gel, purified by use of Gene Clean (Bio101), circularized with the Perfectly Blunt Cloning Kit (Novagen), and linearized with Scal (New England BioLabs). Primers 9917 and IS53 were used to amplify the sequence flanking the 3’ end of the IS6110 copy. The sequence was determined with the CEQ 2000 Dye Terminator Cycle Sequencing Kit (Beckman Instruments), primer 9917, and the CEQ 2000 capillary sequencer [21].

PCR-amplified fragments of defined sequences flanking IS6110 were used as probes for Southern hybridization in insertion site mapping, to confirm the IS6110 insertion loci in the chromosome of strains studied [22]. The amplicons were used to rehybridize the *Pvu* II-digested chromosomal DNA membranes used for the routine IS6110 Southern blot hybridization analysis. Table 1 lists primers and H37Rv chromosomal loci [23].

Sites of IS6110 insertions were also confirmed by PCR amplification of the flanking regions with one of the primers described above (dnaA-F, dnaN-R, E-F, E-R, D-F, or D-R) in combination with 1 of 2 IS6110 primers (S1 or S2) [24].

**Principal genetic grouping.** Determination of principal genetic grouping is based on polymorphisms at *katG* codon 463 (CTG or CGG) and at *gyrA* codon 95 (ACC to AGC) [25]. These polymorphisms were identified by PCR, using 2 pairs of molecular beacons [26, 27]. For each codon, 2 molecular beacons labeled with either fluorescein or tetrachlorofluorescein were designed and synthesized to hybridize specifically to one of the polymorphisms [25]. A protocol on the synthesis of molecular beacons is available on the Web at http://www.molecular-beacons.org.

**Variable number tandem repeat (VNTR) loci.** The VNTR loci ETR-A to ETR-E were characterized, as described elsewhere [28]. Results from each of the 5 loci were combined to form a 5-digit allele profile. VNTR analyses were done at both the PHRI Tuberculosis Center and the Durham, NC, VA Hospital.

**Epidemiologic analysis.** Demographic and clinical data were obtained from the tuberculosis surveillance system of the New Jersey Department of Health and Senior Services. This included data from the medical charts, antimycobacterial susceptibility profiles, patient interviews, and contact investigation reports, which were used to evaluate epidemiologic links among the patients. Routine contact investigations were conducted for all cases of proven or suspected pulmonary tuberculosis.

Analysis was done with SAS software (version 8; SAS Institute). Fisher’s exact and $x^2$ tests were used, as appropriate, to compare the proportions of categorical variables between groups.

**Results**

**New Jersey study population.** Of the 1764 isolates from New Jersey tuberculosis patients that were genotyped, 381 (22%;
165 genotypes were identified. There were 140 isolates, rep-

identified. Among the 381 isolates, 20 had IS

6110 copies (figure 1), 56 (15%) pansusceptible

isolates were further segregated on the basis of their unique spoli-
gotype, CDC0075 (octal code, 777776407760601), and were la-
abled group S75 (figure 2). Spoligopattern 0075 was found to

be unique when compared with the Wadsworth database of

>3000 isolates, which includes samples from New Jersey, New

York, and several additional northeastern states. This group

comprised 3 IS6110 patterns arbitrarily labeled BE (n = 41 iso-
lates), H6 (n = 13 isolates), and C28 (n = 2 isolates), corre-
sponding to 1, 2, and 3 IS6110 copies, respectively (figure 1).

The PGRS analysis of Alu-I-digested chromosomal DNA re-
vealed identical patterns for most (95%) strains, and the remain-
ing 3 had closely related patterns (1 H6 and 2 C28 strains [data not shown]). The remaining 325 low-copy-number isolates rep-
resented an array of diverse genotypes and, in contrast to group S75, displayed diverse PGRS and spoligotype patterns. All mem-
ers of the S75 group had an identical VNTR allelic profile, 22433.

In contrast to the distinct spoligopattern, this VNTR profile is not
unique to New Jersey, with a worldwide distribution accounting
for 6% of M. tuberculosis isolates in an international collection
[29] and 7% of clinical isolates in the United Kingdom [30].

IS6110 insertion site mapping data indicated that all strains in
this group shared an identical insertion in the E region of the
chromosome. The second IS copy (for isolates with 2 and 3

Table 1. Primers and chromosomal locations of IS6110 insertions found in S75 Mycobacterium tuberculosis isolates.

<table>
<thead>
<tr>
<th>Locus name, gene (Rv)b</th>
<th>Primer designations</th>
<th>Primers</th>
<th>Location of IS6110 on H37Rv map; corresponding cosmids; and chromosomal sequence flanking IS6110</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. mmpS1 (Rv0403c)</td>
<td>E-F 5'-GGTAATTGATGCTCGAGCCGCT-3'</td>
<td>IS6110 9917 5'-GCCGTCGACTCGGAGCTTG-3'</td>
<td>IS6110c, segment 20/162; MTCY04D9;</td>
<td>All S75 isolates</td>
</tr>
<tr>
<td></td>
<td>E-R 5'-ATCCCGTGATGATGATGCTACG-3'</td>
<td>ISS3 5'-CCCGACGCGCTCCGCGGCGT-3'</td>
<td>IS6110c, segment 1/162; MTV029;</td>
<td>All H6 or C28 isolates</td>
</tr>
<tr>
<td>A. dnaA-dnaN (Rv0001–Rv0002f)</td>
<td>dna-A-F 5'-CCACAGGTCACACAGTCA-3'</td>
<td>IS6110 9917 5'-GCCGACGCGCTCCGCGGCGT-3'</td>
<td>IS6110c, segment 155/162; MTV025;</td>
<td>C28 isolates but not</td>
</tr>
<tr>
<td></td>
<td>dna-R 5'-CAACTCTGTTCGAGCCGCG-3'</td>
<td></td>
<td></td>
<td>BE or H6</td>
</tr>
<tr>
<td>D (Rv3734c)</td>
<td>D-F 5'-CGAACGTGTAACCTCGATGC-3'</td>
<td></td>
<td></td>
<td>No insertions in S75, but</td>
</tr>
<tr>
<td></td>
<td>D-R 5'-GGTAATGGGTGTAATGGGTGCAT-3'</td>
<td></td>
<td></td>
<td>most low-copy-no.</td>
</tr>
<tr>
<td>DR, direct repeat (Rv2813–Rv2816c)</td>
<td>H-F 5'-GACGACGCGTTGCGTG-3'</td>
<td>IS6110 9917 5'-GCCGTCGACTCGGAGCTTG-3'</td>
<td>IS6110c, segment 123/162; MTCY16B7</td>
<td>isolates have IS6110 insertion here</td>
</tr>
<tr>
<td></td>
<td>H-R 5'-TCCTGGTTGCTGTCGAC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From [23].

b Underlined sequences identify the 3-bp direct repeats generated as a result of the IS6110 insertion event.

f IS6110 insertion in the intergenic region between Rv0001 and Rv0002.

b Rv designations correspond to direct repeat flanking regions.

from 381 persons) were determined to have a low number of copies of IS6110 insertions and were subject to further analysis. On the basis of IS6110 genotyping (results with both 5' and 3' fragments were consistent), 80 different genotypes were identified. Among the 381 isolates, 20 had IS6110 hybridization patterns unlike any others in the New Jersey collection or the PHRI Tuberculosis Center archive (unique patterns). There were 247 isolates grouped into 8 large clusters (representing ≥10 clinical isolates). In addition, 65 isolates were grouped into 13 strain types, with 3–9 patients each, and 20 isolates into 10 genotypes, with 2 patients each. Twenty-nine strains were unique to the New Jersey collection but not to the PHRI Tuberculosis Center archive.

Table 2. Primers and chromosomal locations of IS6110 insertions found in S75 Mycobacterium tuberculosis isolates.

<table>
<thead>
<tr>
<th>Locus name, gene (Rv)b</th>
<th>Primer designations</th>
<th>Primers</th>
<th>Location of IS6110 on H37Rv map; corresponding cosmids; and chromosomal sequence flanking IS6110</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>E. mmpS1 (Rv0403c)</td>
<td>E-F 5'-GGTAATTGATGCTCGAGCCGCT-3'</td>
<td>IS6110 9917 5'-GCCGTCGACTCGGAGCTTG-3'</td>
<td>IS6110c, segment 20/162; MTCY04D9;</td>
<td>All S75 isolates</td>
</tr>
<tr>
<td></td>
<td>E-R 5'-ATCCCGTGATGATGATGCTACG-3'</td>
<td>ISS3 5'-CCCGACGCGCTCCGCGGCGT-3'</td>
<td>IS6110c, segment 1/162; MTV029;</td>
<td>All H6 or C28 isolates</td>
</tr>
<tr>
<td>A. dnaA-dnaN (Rv0001–Rv0002f)</td>
<td>dna-A-F 5'-CCACAGGTCACACAGTCA-3'</td>
<td>IS6110 9917 5'-GCCGACGCGCTCCGCGGCGT-3'</td>
<td>IS6110c, segment 155/162; MTV025;</td>
<td>C28 isolates but not</td>
</tr>
<tr>
<td></td>
<td>dna-R 5'-CAACTCTGTTCGAGCCGCG-3'</td>
<td></td>
<td></td>
<td>BE or H6</td>
</tr>
<tr>
<td>D (Rv3734c)</td>
<td>D-F 5'-CGAACGTGTAACCTCGATGC-3'</td>
<td></td>
<td></td>
<td>No insertions in S75, but</td>
</tr>
<tr>
<td></td>
<td>D-R 5'-GGTAATGGGTGTAATGGGTGCAT-3'</td>
<td></td>
<td></td>
<td>most low-copy-no.</td>
</tr>
<tr>
<td>DR, direct repeat (Rv2813–Rv2816c)</td>
<td>H-F 5'-GACGACGCGTTGCGTG-3'</td>
<td>IS6110 9917 5'-GCCGTCGACTCGGAGCTTG-3'</td>
<td>IS6110c, segment 123/162; MTCY16B7</td>
<td>isolates have IS6110 insertion here</td>
</tr>
<tr>
<td></td>
<td>H-R 5'-TCCTGGTTGCTGTCGAC-3'</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

From [23].

b Underlined sequences identify the 3-bp direct repeats generated as a result of the IS6110 insertion event.

f IS6110 insertion in the intergenic region between Rv0001 and Rv0002.

b Rv designations correspond to direct repeat flanking regions.

from 381 persons) were determined to have a low number of copies of IS6110 insertions and were subject to further analysis. On the basis of IS6110 genotyping (results with both 5' and 3' fragments were consistent), 80 different genotypes were identified. Among the 381 isolates, 20 had IS6110 hybridization patterns unlike any others in the New Jersey collection or the PHRI Tuberculosis Center archive (unique patterns). There were 247 isolates grouped into 8 large clusters (representing ≥10 clinical isolates). In addition, 65 isolates were grouped into 13 strain types, with 3–9 patients each, and 20 isolates into 10 genotypes, with 2 patients each. Twenty-nine strains were unique to the New Jersey collection but not to the PHRI Tuberculosis Center archive.

Spoligotyping data for the 381 isolates identified a total of 109 different patterns. Seventy-seven spoligopatterns (20%) were unique in this New Jersey collection; 242 isolates were grouped into 10 large spoligotype clusters (≥10 clinical isolates), 36 isolates into 9 spoligotype patterns with 3–9 patients each, and 26 isolates into 13 spoligotypes with 2 patients each. When both spoligotyping and IS6110 data were combined, 165 genotypes were identified. There were 140 isolates, representing 7 genotypes, with a cluster size of ≥10 patients; 79 isolates, representing 14 genotypes, were in groups of 3–9 patients; 36 samples, representing 18 genotypes, had a cluster size of 2; and 126 types were unique in the collection.

The demographic characteristics of the 381 study patients, compared with those of all others from whom data were collected during the study period (n = 1383), were not significantly different in terms of sex (P = .276), age (P = .780), or proportions of Asians (P = .227) and whites (P = .807). In addition, no significant difference was observed between groups in the proportion of patients with a history of incarceration (P = .984) or alcohol abuse (P = .106).
bands) was inserted in region A (the dnaA–dnaN intergenic region). The third IS copy for strain C28 was found in region D (Rv3734c) of the chromosome (figure 3). Insertion site mapping analysis of the S75 strains indicated no IS6110 insertion in the DR region of the chromosome. Sequencing of regions A and D in BE and of region D in H6 revealed no disruption, compared with wild-type sequences in sequenced strains H37Rv and CDC1551, in which IS6110 elements are not present in regions A and D.

Principal genetic grouping further supported relatedness of the 56 isolates in the S75 group. All isolates grouped to S75 had KatG codon 463 sequence CGG (Arg) and gyrA codon 95 sequence ACC (Thr), placing them into principal genetic group 2.

Epidemiologic analysis of the S75 group. The basic demographic characteristics of the 56 patients in the S75 group are summarized in table 2. The median age in this group was 42 years (range, 7 months to 88 years). All 56 patients in this cluster were born in the United States, and 51 (91%) were non-Hispanic black. Thirty-one (55%) were male. Among the 47 patients for whom HIV serology was known, 53% (25) were seropositive. Of the 25 patients with HIV coinfection, 18 reported injection drug use. Including HIV seropositivity, 66% of patients in this cluster had at least 1 known risk factor for disease (e.g., alcohol abuse, injection drug use, history of incarceration, or homelessness).

Table 2 describes proportional differences between demographic characteristics of patients in the S75 group (n = 56) and those of patients with all other isolates reported in New Jersey (n = 1708). The groups were similar in sex (P = .302) and age distribution (P = .129) and in proportion of patients with a history of incarceration (P = .395). The S75 isolates were recovered mainly from patients who were US born (P < .001) and resided in Essex County, NJ (P < .001). S75 group isolates were more likely to come from patients of non-Hispanic black origin (P < .001) and those with history of alcohol abuse (P < .001), positive HIV serology (P = .02), injection drug use (P < .001), and recent history of homelessness (P < .001). In contrast, patients in the S75 group were less likely to be of Asian (P < .001), Hispanic (P = .002), or non-Hispanic white origins (P < .001), compared with all other patients with tuberculosis in New Jersey.

Within the S75 cluster, 13 patients with IS6110 genotype H6 were very homogeneous in their demographic characteristics. Ten of the 13 were HIV seropositive, and, of these, 9 reported injection drug use. All of them were of non-Hispanic black origin, and 10 resided in the city of Newark. They were predominantly male (12/13), and 6 reported alcohol abuse.

We were able to epidemiologically link 17 (30%) of the 56 cases in the S75 group. Among the 17 linked cases, a large extended family comprised 7 members and 1 close acquaintance. Seven additional patients were linked in 2 groups of 2 and 1 group of 3 cases. All epidemiologic links were consistent with molecular types. There were 9 patients who, although they had no identifiable epidemiologic links, all resided in close proximity to each other (within 500 m) in Newark (figure 4). This group comprised 2 S75 variants, BE and H6.

Figure 4 illustrates the geographic aggregation of patients in the S75 group during the study period. Eighty-nine percent (n = 50) of patients in the S75 cluster were from Essex County, and, of this group, 62% (n = 35) were from Newark and an additional 29% (n = 16) were from neighboring cities. The remaining 9% (n = 5) were from adjacent or close-by counties. Furthermore, S75 group cases accounted for 38% and 11% of
all cases due to low-copy-number isolates \( (n = 133) \) and of all tuberculosis cases \( (n = 437) \), respectively, reported in Essex County. Likewise, S75 cases accounted for 43\% \( (n = 81) \) and 13\% \( (n = 260) \) of all cases in Newark and Essex County, respectively. In contrast to the S75 group, other low-copy-number isolates were from more geographically diverse regions.

**Discussion**

*Mycobacterium tuberculosis* strains having \( \leq4–6 \) IS6110 insertions have been defined as low-copy-number strains [12–14, 31]. Historically, low-copy-number isolates were excluded from genetic cluster analysis because of the inconsistency in interpretation
we also identified 20 strains in the New Jersey data set and C28 (comprises 3 IS6110 insertions, respectively) to suggest true case clustering of low-copy-number isolates [14, 31, 32]. A recent population-based study of low-copy-number isolates named BE have a common IS6110 hybridizing band size, the 41 BE isolates in the S75 group have a unique IS6110 chromosomal insertion site. All 20 BE isolates outside the S75 group did not share epidemiologic features or PGRS or VNTR profiles (18 and 13 different patterns, respectively) with S75 members. Besides the 15 isolates belonging to the S75 group, no additional isolates in the data set were designated H6 or C28.

In sharp contrast to previously reported low-copy-number investigations, the S75 cluster could be further distinguished on the basis of 2 unique molecular characteristics. First, unlike most low-copy-number Mycobacterium tuberculosis isolates reported, the S75 cluster does not have an IS6110 insertion in the DR locus, reported earlier as an IS6110 preferential insertion site [24]. The DR locus is believed by some to have been the target site for the first IS6110 insertion in the early evolution of human tuberculosis [24, 33, 34]. Given the DR spacer deletions in cluster S75 (see figure 2), it is possible that an IS6110 element was lost at this site during the course of evolution. A second distinguishing property of these isolates is that they all belong to principal genetic group 2. This is in contrast to other isolates with a single IS6110 copy, which are invariably members of genetic group 1 but not 2 or 3 [25]. In support of this observation, the 20 BE isolates not part of the S75 group were all members of principal genetic group 1.

In this investigation, isolates containing 1 or 2 additional IS6110 insertions (H6 and C28, respectively; figure 1) were grouped together with the BE isolates, on the basis of their unique spoligotype, and together they define the S75 cluster. IS6110 insertion site mapping was used to locate the exact chromosomal position of the IS6110 element in the M. tuberculosis genome [22]. All S75 strains share the same insertion in the E region, and H6 and C28 have a second insertion in the A region of the chromosome (figure 3). BE and H6 were found to have wild-type sequences in the A and D regions, compared with H37Rv and CDC1551 genomes; that is, no remnants of IS6110 or of target site duplication indicative of a previous insertion were detected. However, because no chromosomal rearrangements were seen in regions A and D, it is possible, although unlikely, that there was either a precise excision of IS6110 that included the usual 3-bp duplication or no target duplication at the time of IS6110 insertion [24]. The most parsimonious interpretation of the data, taken together with other molecular techniques, is to propose the stepwise acquisition of IS6110 from BE to H6 to C28. The finding that no isolates with the H6 or C28 IS6110 RFLP patterns have a spoligopattern different from 0075 lends support to the notion that the IS6110 copy—transposition event, and the evolution from BE to H6 and subsequently to C28, were the result of 2 single genetic events. Given that the stability of IS6110 genotypes has an estimated half-life of 3–4 years, one can elucidate the evolutionary time

| Table 2. Demographic comparison of S75 strains from New Jersey with all other Mycobacterium tuberculosis isolates. |
|------------------|------------------|------------------|
| Characteristic   | S75 cluster (n = 56) | All other isolates (n = 1708) | P* |
| Age, years       | 0.129             |                               |     |
| < 50             | 41 (73)           | 1081 (63)                   |     |
| > 50             | 15 (27)           | 627 (37)                    |     |
| Median (interquartile range) | 41 (17) | 42 (28)                     |     |
| Race/ethnicityc  |                  |                               |     |
| Non-Hispanic white | 1 (2)            | 332 (19)                    | <0.001 |
| Non-Hispanic black | 51 (91)          | 590 (35)                    | <0.001 |
| Hispanic         | 4 (7)             | 390 (23)                    | 0.002  |
| Asian            | 0                 | 396 (23)                    | <0.001 |
| HIV serologyd    |                  |                               | 0.02   |
| Positive         | 25 (45)           | 288 (17)                    |     |
| Negative         | 22 (39)           | 523 (31)                    |     |
| Unknown          | 9 (16)            | 897 (53)                    |     |
| History of incarceration |      |                               | 0.395  |
| Yes              | 10 (18)           | 65 (4)                      |     |
| No               | 46 (82)           | 1643 (96)                   |     |
| Injection drug use |                |                               | <0.001 |
| Yes              | 20 (36)           | 109 (6)                     |     |
| No               | 36 (64)           | 1599 (94)                   |     |
| Alcohol abuse    |                  |                               | <0.001 |
| Yes              | 22 (39)           | 246 (14)                    |     |
| No               | 34 (61)           | 1462 (86)                   |     |
| Birthplace       |                  |                               | <0.001 |
| United States    | 56 (100)          | 780 (46)                    |     |
| Other            | 0                 | 928 (54)                    |     |
| Residence in Essex County |      |                               | <0.001 |
| Yes              | 50 (89)           | 390 (23)                    |     |
| No               | 6 (11)            | 1318 (77)                   |     |

NOTE. Data are no. (%) of patients, except for median age. HIV, human immunodeficiency virus.

*a*χ² or Fisher’s exact test (2-tailed) was used, as appropriate.

*b*Range between 25th and 75th percentiles.

*c*P value for comparison of each race/ethnic group with all other race/ethnic groups.

*d*Unknown HIV status was not included in P value.

when correlated with epidemiologic data. Increasingly, however, a number of reports have used additional molecular techniques to suggest true case clustering of low-copy-number isolates [14, 31, 32]. A recent population-based study of low-copy-number isolates clustered by spoligotyping documented an epidemiologic relationship among 15% of cases in their reported cohort [13].

In the present investigation, multiple molecular methods were used in conjunction with routine surveillance data to segregate a large (56 cases), epidemiologically relevant cluster (S75 group) from a total of 381 low-copy-number isolates reported in New Jersey between January 1996 and March 2000. The S75 group comprises 3 IS6110 RFLP patterns—BE (n = 41), H6 (n = 13), and C28 (n = 2)—with 1, 2, and 3 IS6110 insertions, respectively. We also identified 20 strains in the New Jersey data set with the BE IS6110 RFLP pattern, but these were distinguished on the basis of their having 15 different spoligopatterns other than 0075. This study also revealed that, even though all 61 isolates named BE have a common IS6110 hybridizing band size, the 41 BE isolates in the S75 group have a unique IS6110 chromosomal insertion site. All 20 BE isolates outside the S75 group did not share epidemiologic features or PGRS or VNTR profiles (18 and 13 different patterns, respectively) with S75 members. Besides the 15 isolates belonging to the S75 group, no additional isolates in the data set were designated H6 or C28.

In sharp contrast to previously reported low-copy-number investigations, the S75 cluster could be further distinguished on the basis of 2 unique molecular characteristics. First, unlike most low-copy-number Mycobacterium tuberculosis isolates reported, the S75 cluster does not have an IS6110 insertion in the DR locus, reported earlier as an IS6110 preferential insertion site [24]. The DR locus is believed by some to have been the target site for the first IS6110 insertion in the early evolution of human tuberculosis [24, 33, 34]. Given the DR spacer deletions in cluster S75 (see figure 2), it is possible that an IS6110 element was lost at this site during the course of evolution. A second distinguishing property of these isolates is that they all belong to principal genetic group 2. This is in contrast to other isolates with a single IS6110 copy, which are invariably members of genetic group 1 but not 2 or 3 [25]. In support of this observation, the 20 BE isolates not part of the S75 group were all members of principal genetic group 1.

In this investigation, isolates containing 1 or 2 additional IS6110 insertions (H6 and C28, respectively; figure 1) were grouped together with the BE isolates, on the basis of their unique spoligotype, and together they define the S75 cluster. IS6110 insertion site mapping was used to locate the exact chromosomal position of the IS6110 element in the M. tuberculosis genome [22]. All S75 strains share the same insertion in the E region, and H6 and C28 have a second insertion in the A region of the chromosome (figure 3). BE and H6 were found to have wild-type sequences in the A and D regions, compared with H37Rv and CDC1551 genomes; that is, no remnants of IS6110 or of target site duplication indicative of a previous insertion were detected. However, because no chromosomal rearrangements were seen in regions A and D, it is possible, although unlikely, that there was either a precise excision of IS6110 that included the usual 3-bp duplication or no target duplication at the time of IS6110 insertion [24]. The most parsimonious interpretation of the data, taken together with other molecular techniques, is to propose the stepwise acquisition of IS6110 from BE to H6 to C28. The finding that no isolates with the H6 or C28 IS6110 RFLP patterns have a spoligopattern different from 0075 lends support to the notion that the IS6110 copy—transposition event, and the evolution from BE to H6 and subsequently to C28, were the result of 2 single genetic events. Given that the stability of IS6110 genotypes has an estimated half-life of 3–4 years, one can elucidate the evolutionary time
frame of this cluster to be ~1 decade [35]. It is likely that IS6110 stability may vary in different genetic backgrounds, but the recent epidemiology associated with the S75 cluster in this study, as shown by the demographic homogeneity of the H6 subgroup, is consistent with an estimated 10-year time span.

The presence of epidemiologic characteristics such as geographic aggregation (figure 4), absence of foreign-born patients, and high proportion of specific demographic characteristics (table 2) among S75 members strongly suggests a locally produced cluster. Examination of this cohort for epidemiologic links by use of contact tracing information from the New Jersey Department of Health and Senior Services and review of medical charts identified matches for 30% (17/56 cases). All epidemiologically linked cases shared identical IS6110 patterns. There were 9 additional patients with no apparent links who reside near each other. A recent report by Acevedo-Garcia [36] suggests that a ZIP code–level risk factor for tuberculosis may be relevant to explain a neighborhood-level association among 9 cases. During the course of this study, 4 cases with the S75 molecular profile were reported in New York City. These patients shared demographic characteristics similar to those reported in this investigation. Furthermore, from conventional contact trac-

Figure 4. Locations of tuberculosis cases with S75 group isolates and all other IS6110 low-copy-no. isolates in New Jersey, January 2001. Data sources: US Environmental Protection Agency Region II and Public Health Research Institute Tuberculosis Center; prepared by Isosceles Information Solutions, Manotick, Canada.
Inferences. The sequential acquisition of IS 6110 isolates with suspected case cluster. Similar to a recent study involving isolates that differ from those reported in this investigation. Higher capture of cases in New Jersey could have identified more links and possibly the missing association between the IS 6110 subgroups within this cluster. The majority of cases were reported from private clinics, and the remaining isolates were not viable. This lack of isolates may introduce sampling bias, because the group not captured in this study may represent patients with demographic features that differ from those reported in this investigation.

In sum, by using a number of genetic markers in conjunction with surveillance data, we have identified a large, previously unsuspected case cluster. Similar to a recent study involving isolates with >18 IS 6110 insertions [7], we demonstrate here the usefulness of grouping low-copy-number strains with similar but not identical IS 6110 RFLP patterns to draw epidemiologic inferences. The sequential acquisition of IS 6110 has highlighted previously unrecognized patient links and has shed light on the evolutionary dissemination of this clone. Therefore, routine genotyping of incident M. tuberculosis isolates, although resource intensive, if applied with discretion, will aid in better understanding M. tuberculosis epidemiology and may make for more effective control programs.

Acknowledgments

We thank M. Wolman, B. Nivin, and A. Ravikovich for assistance with the patient and fingerprint databases. The authors would like to thank B. Sard Salim, S. Lutwick, W. Eisner, and B. Shopsin for their critical reading of the manuscript and their comments and suggestions.

References

24. Fang Z, Morrison N, Watt B, Doig C, Forbes KJ. IS 6110 transposition and


