JOURNAL OF CLINICAL MICROBIOLOGY, June 2000, p. 2081–2086 Vol. 38, No. 6
0095-1137/00/$04.00+0

False Molecular Clusters due to Nonrandom Association of IS*6110* with *Mycobacterium tuberculosis*

S. H. GILLESPIE,* A. DICKENS, AND T. D. MCHUGH

Department of Medical Microbiology, Royal Free and University College Medical School, University College London, Royal Free Campus, London NW3 2PF, United Kingdom

Received 8 November 1999/Returned for modification 1 February 2000/Accepted 4 March 2000

We sought to determine whether nonrandom association of IS*6110* with *Mycobacterium tuberculosis* could result in false-positive clustering in unselected collections of isolates. We typed 196 strains of *M. tuberculosis* from an unselected community-based study in northern Tanzania by IS*6110* and polymorphic GC-rich repetitive-sequence (PGRS) methodologies. The strains were analyzed by Gelcompar computer software. Analysis of 13 out of 25 groups showed that isolates with identical IS*6110* and PGRS patterns were likely to be the same strain. Some IS*6110* groups containing strains with identical PGRS patterns had similar IS*6110* patterns that differed only by movement of the element. Isolates assigned to a single group (i.e., group 11) on the basis of sharing an identical IS*6110* fingerprint pattern did not share identical PGRS fingerprint patterns. Six out of the nine bands in these isolates were in hot-spot locations, as previously defined. This indicates that nonrandom association may result in false-positive clustering in unselected community-based studies. Only strains with identical PGRS and IS*6110* patterns are likely to be recently transmitted.

The incidence of tuberculosis is rising throughout the world, prompting detailed investigation of the epidemiology of this disease. Typing techniques based on the polymorphism of insertion sequence position in the genome (IS*6110*) and repetitive sequences (e.g., polymorphic GC-rich repetitive sequence [PGRS]) have provided molecular methods to distinguish strains. There have been many reports of their application in routine clinical practice (3, 8, 15).

Mycobacterial typing has moved beyond the investigation of point source outbreaks into wider community-based studies. IS*6110* typing has proved valuable in this context when a focused question is being posed. Thus, it has been possible to follow the spread of particular isolates from intravenous drug abuser and alcoholic groups into the general population (6, 10). It has also been useful to follow the spread of drug-resistant clones, such as strain W, in New York and throughout the United States (19). Similarly, it has been possible to follow the spread of the susceptible strain C in the same population (9). These studies were successful, as they sought to identify one or more identical isolates within a larger unselected group of isolates.

Several large genotyping studies have been initiated in major cities, such as London (P. D. Butcher, H. C. Maguire, A. Pearson, S. H. Gillespie, J. W. Dale, and D. K. Banerjee, Proc. 17th Annu. Meet. Eur. Soc. Mycobacteriol., abstr. P163, 1996), Paris (12), New York (19), and San Francisco (2). Networks have been established to pool the results of smaller studies into databases covering the United States and the European Union. The databases have allowed the transmission of some strains to be followed across state and national borders as described above to answer focused epidemiological questions (1, 16). It has been hoped that retrospective interrogation of these databases would provide insight into the epidemiology of tuberculosis on a wider scale. This approach raises questions about the definition and significance of molecular clustering. Typing of *Mycobacterium tuberculosis* using only IS*6110* is dependent on the assumption that integration of the insertion sequence into the genome is random and that the discrimination of the technique increases in proportion to copy number. We have shown previously that there are hot spots for integration (17) and therefore insertion is, in fact, a nonrandom event. It is possible that nonrandom insertion might confound the algorithms for cluster analysis. These observations make no assumptions with regard to evolutionary associations between IS*6110* and the PGRS region.

This study was designed to define a practical approach for the interpretation of IS*6110* and PGRS cluster analyses. To do this, we investigated an unselected collection of sequential routine isolates from Northern Tanzania, among which there were no known epidemiological clusters, by both IS*6110* and PGRS typing. The terms "strain" and "isolate" are often used without precision. In this study, isolate refers to a mycobacterial culture derived from a single patient on a single occasion. Strain defines one or more isolates that are thought to be isogenic. A practical aim of this study is to offer a paradigm for the classification of strains using established molecular techniques, thus providing a rationale for further epidemiological investigation.

MATERIALS AND METHODS

Bacterial isolates. Single *M. tuberculosis* isolates were prospectively collected from all culture-positive patients diagnosed by the National Tuberculosis and Leprosy Control Programme Reference Laboratory at Kibongoto Hospital over the 6-month period from April to September 1995. Speciation was confirmed by standard microbiological techniques. The isolates were maintained on Löwenstein-Jenson slopes at 37°C for a minimum of 4 weeks and subsequently transported to the Department of Medical Microbiology, Royal Free & University College Medical School, London, United Kingdom.

DNA extraction. Bacteria were harvested from the Löwenstein-Jensen slopes, heat killed, and incubated with lysozyme (1 h; 37°C) followed by digestion with 50 μg of proteinase K (Sigma, Poole, United Kingdom) in 10% sodium dodecyl sulfate for 10 min at 65°C. A further incubation with 1% (wt/vol) cetyltrimethylammonium bromide in ≥0.5 M NaCl for 10 min at 65°C was followed by partition using chloroform-isoamyl alcohol (24:1 [vol/vol]).

IS*6110* typing. Genomic DNA was digested with *Pvu*II restriction endonuclease and separated by agarose gel electrophoresis and Southern hybridization according to the International Standard Typing method for *M. tuberculosis* (21). The probe used was a PCR amplimer derived from reactions using *M. tubercu-*

* Corresponding author. Mailing address: Department of Medical Microbiology, Royal Free & University College Medical School, University College London, Royal Free Campus, Rowland Hill St., London NW3 2PF, United Kingdom. Phone: (44)-171-794-0500. Fax: (44)-171-794-0433. E-mail: stepheng@rfhsm.ac.uk.

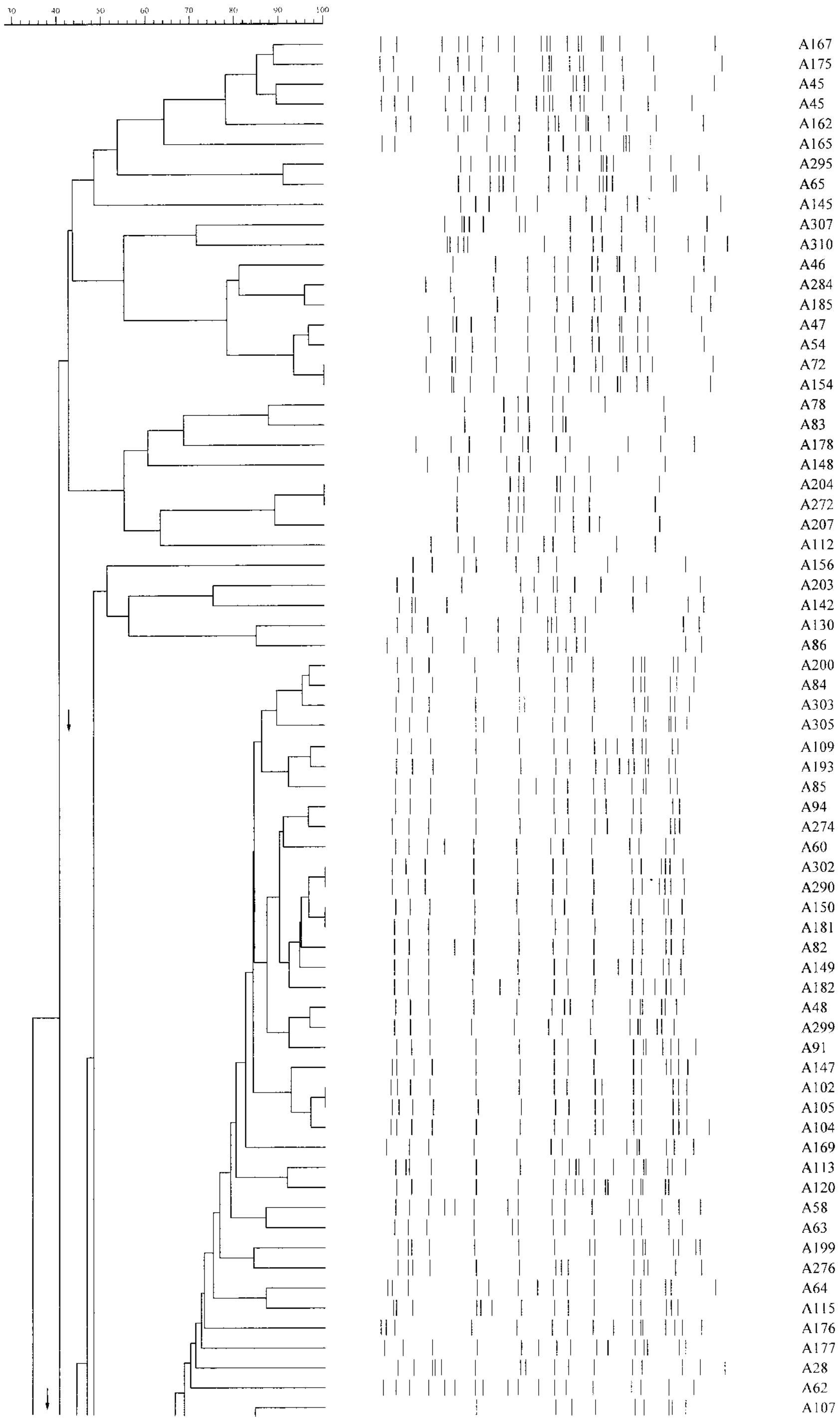

FIG. 1. Master dendrogram of relationships between IS*6110* profiles of 141 isolates of *Mycobacterium tuberculosis* as calculated by the Dice coefficient.

losis strain H37Rv DNA as a template and the following primer set: INS1, 5′ CGT GAG GGC ATC GAG GTG GC 3′, and INS2, 5′ GCG TAG GCG TCG GTG ACA AA 3′ (18). Hybridization was performed at 50°C, and the final wash was 0.5× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate) at 50°C.

PGRS typing. All available isolates were submitted to PGRS analysis. Genomic DNA was digested with *Alu*I restriction endonuclease and separated by agarose gel electrophoresis followed by Southern hybridization. The probe used was an oligonucleotide consisting of two copies of the PGRS consensus repeat (5) sequence, 3′ GGC GGC AAC GGC GGC AAC GGC GGC GGC GGC AAC GGC GGC AAC GGC GGC 5′. Hybridization was performed at 40°C, and the final wash was 2× SSC at room temperature.

Detection. The probes were labeled and detected by chemiluminescent procedures as recommended by the manufacturers (Gene-Star and Amersham Pharmacia Biotech).

Cluster analysis. Comparison of DNA fingerprints was done with the GelCompar version 4.0 package (Applied Maths, Kourtrai, Belgium). Cluster analysis was performed by calculation of the Dice coefficient. Similarity, defined by the Dice coefficient, was calculated using the parameter settings at 0.8% band position tolerance for IS*6110* typing and 1.2% band position tolerance for PGRS typing. A cluster was defined as a series of isolates with 100% identity, and a group was defined as a series of isolates with less than 100% but greater than 90% similarity by either technique.

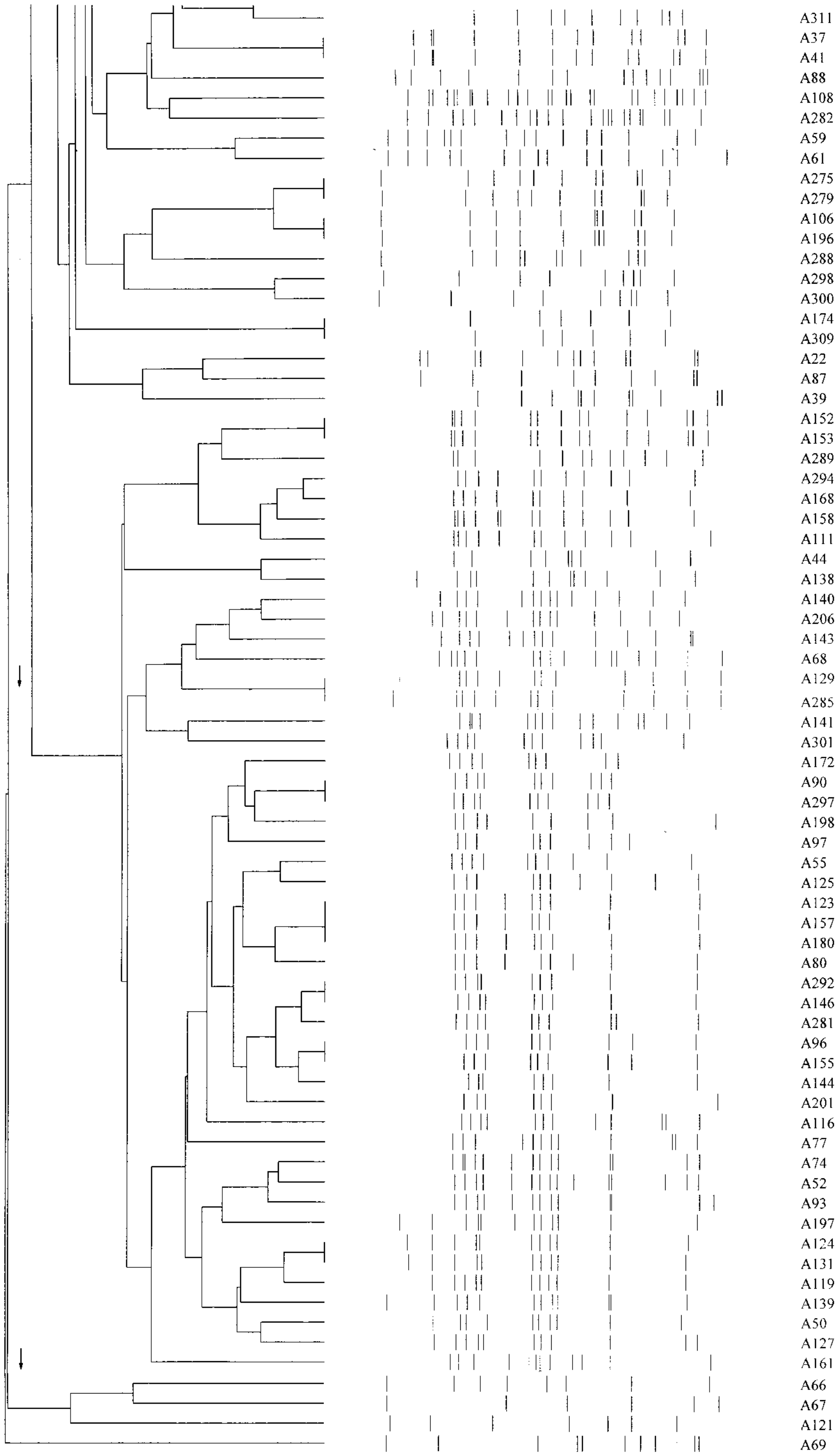

FIG. 1—*Continued.*

RESULTS

IS*6110* analysis of 141 isolates of *M. tuberculosis* with six or more bands revealed 25 groups (66 isolates) at 90% identity, of which 4 were robust at 100% identity (Fig. 1). Cultures were available for further analysis of 13 of these groups (35 isolates) by PGRS (Table 1). For the remaining groups, one or more cultures were no longer viable and so were not available for DNA extraction.

Isolates in 7 of the 13 different IS*6110* groups shared the same PGRS pattern and thus maintained the close relationships defined by IS*6110* typing (Table 1).

Group 11 contained three isolates clustered at 100% by IS*6110* analysis, but the PGRS analysis showed only 84% iden-

TABLE 1. Comparison of isolates by IS*6110* and PGRS typing

Group	Isolates	% Similarity by IS6110	% Similarity by PGRS
1	A84 A303	98	100
2	A94 A274	98	100
3	A48 A299	98	100
4	A82 A150 A181 A182 A290 A302	92	96
5	A102 A104 A105 A147	92	100
6	A120 A113	92	96
7	A37 A41	100	100
8	A47 A72 A154	93	91
9	A90 A297	100	100
10	A281 A292	94	92
11	A123 A157 A180	100	84
12	A119 A124 A131	91	93
13	A129 A285	100	100

tity. Isolates A180 and A157 were 88% similar; A180 had one more band than A157, and there were marked band position differences. A180 had the same number of bands as A123 (13 bands) but had different positional changes (Fig. 2).

In groups 6 and 10, both the IS*6110* pattern and the PGRS pattern showed marked variation. Group 6 contained two isolates with 92% similarity by IS*6110* analysis; isolate A113 had one more band than A120, and there were further band position differences between these isolates. By PGRS analysis, these isolates were 96% similar, with isolate A113 containing one extra band. Group 10 showed 94% similarity between two isolates by IS*6110* analysis. Isolate A281 contained one extra band. PGRS analysis revealed a similarity of 92%; the isolates had the same number of bands but had two positional changes.

Group 4 contained six isolates with 92% similarity by IS*6110* analysis and 96% similarity by PGRS analysis (Fig. 3).

DISCUSSION

In this unselected community-based study, we were seeking to identify molecular relationships between strains that had not been detected in clinical practice. IS*6110* typing of 141 isolates with six or more bands yielded 25 groups, of which 13 were available for further analysis. Current practice would lead to detailed epidemiological investigation of patients in these groups and the attribution of epidemiological significance to these results.

For isolates in group 11, the IS*6110* patterns are identical but the PGRS patterns show considerable variation, indicating that these are likely to be different strains. Of the nine IS*6110* band positions in the group 11 pattern, six are located in hot spots for integration that we have defined previously (17). We therefore conclude that this IS*6110* similarity has arisen through a chance association. In a large genotype study of an unselected population, the bias resulting from hot spots for integration represents a risk that may serve to compromise the epidemiological investigation. Such a bias is already recognized by many groups, as low-copy-number isolates are routinely excluded from analysis. The majority of low-copy-number band positions are at hot spots for integration (4, 7, 14, 17). Where the PGRS genotypes are different but the IS*6110* genotypes are similar, this represents a false cluster caused by nonrandom association which should not be investigated further.

Introduction of a second sensitive typing technique enabled the validation of each of the groups. Three IS*6110* groups were robust when tested by PGRS analysis. Clusters 7, 9, and 13 were identical by both methods and must be considered to be composed of one strain each. Further investigation is required

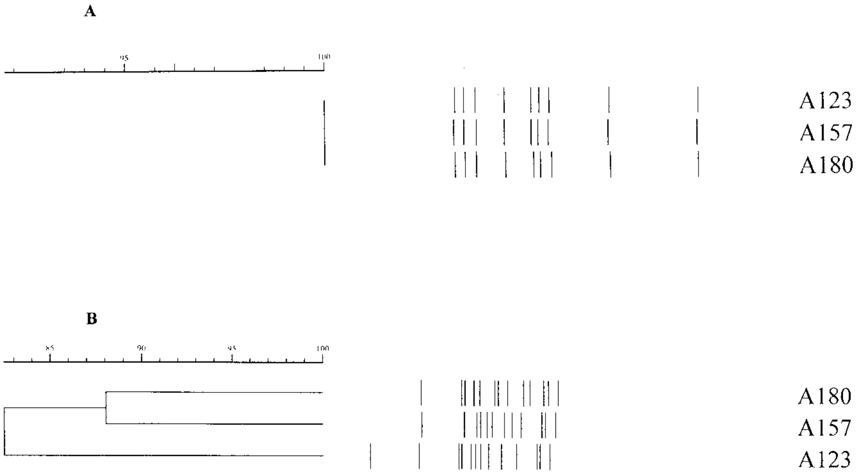

FIG. 2. IS*6110* (A) and PGRS (B) analyses of group 11 demonstrating the disparity in the calculated dendrograms. The IS*6110* profiles are identical (A), but the PGRS profiles demonstrate that isolates A180 and A157 were 88% similar, that A180 had one more band than A157, and that there were marked band position differences. A180 had the same number of bands as A123 (13 bands) but had different positional changes.

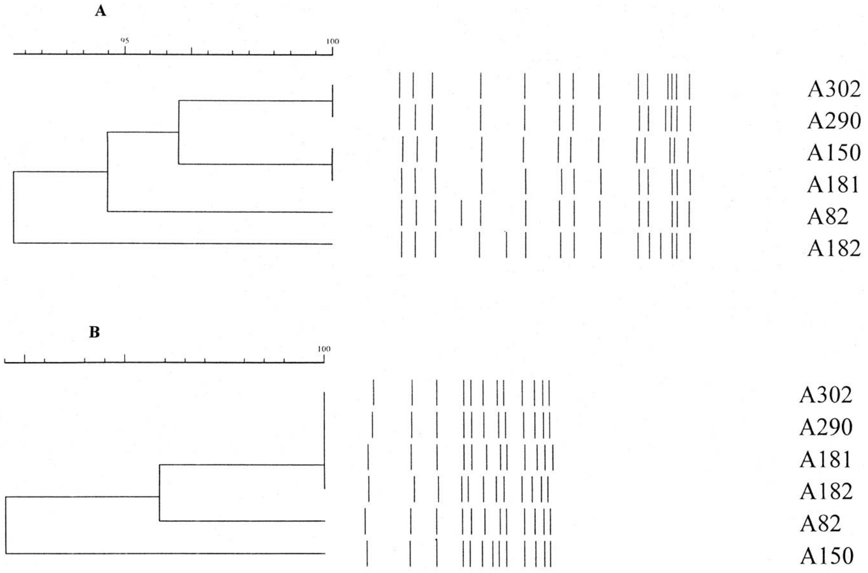

FIG. 3. IS*6110* (A) and PGRS (B) analyses of group 4 demonstrating the disparity in the calculated dendrograms. Isolates A302 and A290 were identical by both IS*6110* and PGRS analyses. Isolates A150 and A181 showed 100% identity by IS*6110* analysis but 96% similarity by PGRS analysis. These two pairs form a subgroup by IS*6110* typing at 96% similarity, with isolates A302 and A290 having one more band than A150 and A181. By PGRS analysis, group 4 divides into one subgroup and two isolates; the subgroup contained four isolates with 100% identity, A302, A290, A181, and A182, and A82 is 98% similar with band position changes. A150 has one additional band. Isolates A302 and A290 were identical by both IS*6110* and PGRS analyses. In the same group, isolates A150 and A181 showed 100% identity by IS*6110* analysis but 96% similarity by PGRS analysis.

in such cases to plot the epidemiological links between them. Clusters 1, 2, 3, and 5 were identical by PGRS typing, but the IS*6110* patterns showed slight variation. This variation was due to the addition or deletion of bands and suggested that the PGRS clustering was valid and the difference in IS*6110* was due to recent movement of the element. Evidence presented by Yeh et al. indicates that the rate of change of PGRS is significantly lower than that of IS*6110* (22). These authors also note that PGRS and IS*6110* instability are independent of each other. Our data support the view that IS*6110* typing provides the more stringent test for recent epidemiological links between cases. Groups 6 and 10 had a calculated similarity by both methods. By previously published criteria for IS*6110* typing, the members of these groups would have been classified as related (11). In group 6 there are four changes in IS*6110* band position, indicating that if these isolates were related, a considerable time had elapsed since they had diverged, and this is confirmed by the difference in PGRS pattern. In group 10 there are differences in the position and number of bands, also suggesting that any relationship is distant. This conclusion is confirmed by the differences in the PGRS types. Although these isolates may be related, they have diverged a considerable time previously and are therefore unlikely to be linked epidemiologically, and thus they are not worthy of an investment of epidemiological resources to identify links.

We propose that only isolates with identical PGRS and IS*6110* patterns can be considered a strain and can indicate recent transmission. Subsequent epidemiological investigation is essential to demonstrate the nature of any relationship. Where the PGRS genotypes are identical and the IS*6110* patterns differ, these are likely to be related isolates in which movement of the element has occurred. The relationship of such isolates in an unselected population would be uncertain, as they have had the opportunity to diverge from a common origin. These more distant relationships may be of significance in a community-based study and could reveal important epidemiological associations, as seen in our previous study (11). Thus, such isolates might merit further detailed investigation to identify epidemiological relationships in a community-based or unselected population study but could probably be excluded from an outbreak investigation. A large community-based study is under way in London which will address this issue (Butcher et al., Proc. 17th Annu. Meet. Eur. Soc. Mycobacteriol.). An exception to the exclusion of such isolates from an outbreak investigation may be if the outbreak occurs in an enclosed community and takes place over a number years with multiple infection events. An example of such a community is provided by the outbreak described in Portugal (13), in which a single strain of *M. tuberculosis* circulated in the human immunodeficiency virus-positive community over a number of years. Minor differences in the IS*6110* patterns were observed, and these were interpreted as transpositional events within a single strain. Epidemiological data corroborated this conclusion.

The data presented here are in accordance with the observation of Salamon et al. (20) that dendrogram clustering methods are at risk of failing to reconstruct the true relationships between isolates. Thus, molecular clustering cannot be accepted uncritically; nonidentical isolates may have their associations obscured as a result of the movement of elements, and conversely, identical clusters may be false due to nonrandom association of insertion elements.

ACKNOWLEDGMENTS

We gratefully acknowledge the help of N. Sam, A. Ramsey, A. O. S. Saruni, and G. M. Kisyombe for their assistance in collecting isolates.

This study was funded by grants from Special Trustees of the Royal Free Hospital held by S. H. Gillespie.

REFERENCES

1. **Bifani, P. J., B. B. Plikaytis, V. Kapur, K. Stockbauer, X. Pan, M. L. Lutfey, S. L. Moghazeh, W. Eisner, T. M. Daniel, M. H. Kaplan, J. T. Crawford, J. M. Musser, and B. N. Kreiswirth.** 1996. Origin and interstate spread of a New York City multidrug-resistant Mycobacterium tuberculosis clone family. JAMA **275:**452–457.

2. **Bradford, W. Z., J. Koehler, H. El-Hajj, P. C. Hopewell, A. L. Reingold, C. B. Agasino, M. D. Cave, S. Rane, Z. Yang, C. M. Crane, and P. M. Small.** 1998. Dissemination of *Mycobacterium tuberculosis* across the San Francisco Bay area. J. Infect. Dis. **177:**1104–1107.
3. **Breathnach, A. S., A. de Ruiter, G. M. C. Holdsworth, N. T. Bateman, D. G. O'Sullivan, P. J. Rees, D. Snashall, H. J. Milburn, B. S. Peters, J. Watson, F. A. Drobniewski, and G. L. French.** 1998. An outbreak of multi-drug-resistant tuberculosis in a London teaching hospital. J. Hosp. Infect. **39:**111–117.
4. **Cole, S. T., R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry III, F. Tekaia, K. Badcock, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. Davies, K. Devlin, T. Feltwell, S. Gentles, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, B. G. Barrell, et al.** 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. Nature **393:**537–544.
5. **Doran, T. J., A. L. Hodgson, J. K. Davies, and A. J. Radford.** 1993. Characterisation of a highly repeated DNA sequence from *Mycobacterium bovis*. FEMS Microbiol. Lett. **111:**147–152.
6. **Dwyer, B., K. Jackson, K. Raios, A. Sievers, E. Wilshire, and B. Ross.** 1993. DNA restriction fragment analysis to define an extended cluster of tuberculosis in homeless men and their associates. J. Infect. Dis. **167:**490–494.
7. **Fang, Z., and K. J. Forbes.** 1997. A *Mycobacterium tuberculosis* IS*6110* preferential locus (ipl) for insertion into the genome. J. Clin. Microbiol. **35:**479–481.
8. **Frieden, T. R., C. L. Woodley, J. T. Crawford, D. Lew, and S. M. Dooley.** 1996. The molecular epidemiology of tuberculosis in New York City: the importance of nosocomial transmission and laboratory error. Tubercle Lung Dis. **77:**407–413.
9. **Friedman, C. R., G. C. Quinn, B. N. Kreiswirth, D. C. Perlman, N. Salomon, N. Schluger, M. Lutfey, J. Berger, N. Poltoratskaia, and L. W. Riley.** 1997. Widespread dissemination of a drug-susceptible strain of *Mycobacterium tuberculosis*. J. Infect. Dis. **176:**478–484.
10. **Genewein, A., A. Telenti, C. Bernasconi, C. Mordasini, S. Weiss, A. M. Maurer, H. L. Rieder, K. Schopfer, and T. Bodmer.** 1993. Molecular approach to identifying route of transmission of tuberculosis in the community. Lancet **342:**841–844.
11. **Gillespie, S. H., N. Kennedy, F. I. Ngowi, N. G. Fumokong, S. Al-Maamary, and J. W. Dale.** 1995. Restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolated from patients with pulmonary tuberculosis in Northern Tanzania. Trans. R. Soc. Trop. Med. Hyg. **89:**335–338.
12. **Gutiérrez, M. C., V. Vincent, D. Aubert, J. Bizet, O. Gaillot, L. Lebrun, C. Le Pendeven, M. P. Le Pennec, D. Mathieu, C. Offredo, B. Pangon, and C. Pierre-Audigier.** 1998. Molecular fingerprinting of *Mycobacterium tuberculosis* and risk factors for tuberculosis transmission in Paris, France, and surrounding area. J. Clin. Microbiol. **36:**486–492.
13. **Hannan, M. M., A. Hayward, H. Peres, S. H. Gillespie, T. D. McHugh, M. Nelson, A. Hawkins, and B. Gazzard.** 1997. Investigation of the high prevalence of the multidrug resistant strain 'Cabral' of *M. tuberculosis* in a Lisbon Community Hospital. Trans. R. Soc. Trop. Med. Hyg. **91:**509.
14. **Hermans, P. W. M., D. van Soolingen, E. M. Bik, P. E. de Haas, J. W. Dale, and J. D. van Embden.** 1991. Insertion element IS*987* from *Mycobacterium bovis* BCG is located in a hotspot integration region for insertion elements in *Mycobacterium tuberculosis* complex strains. Infect. Immun. **59:**2695–2705.
15. **Jereb, J. A., D. R. Burwen, S. W. Dooley, W. H. Haas, J. T. Crawford, L. J. Geiter, M. B. Edmond, J. N. Dowling, R. Shapiro, A. W. Pasculle, et al.** 1993. Nosocomial outbreak of tuberculosis in a renal transplant unit: application of a new technique for restriction fragment length polymorphism analysis of *Mycobacterium tuberculosis* isolates. J. Infect. Dis. **168:**1219–1224.
16. **Kiers, A., A. P. Drost, D. van Soolingen, and J. Veen.** 1997. Use of DNA fingerprinting in international source case finding during a large outbreak of tuberculosis in The Netherlands. Int. J. Tuberc. Lung. Dis. **1:**239–245.
17. **McHugh, T. D., and S. H. Gillespie.** 1998. Nonrandom association of IS*6110* and *Mycobacterium tuberculosis*: implications for molecular epidemiological studies. J. Clin. Microbiol. **36:**1410–1413.
18. **McHugh, T. D., L. E. Newport, and S. H. Gillespie.** 1997. IS*6110* homologs are present in multiple copies in mycobacteria other than tuberculosis-causing mycobacteria. J. Clin. Microbiol. **35:**1769–1771.
19. **Moss, A. R., D. Alland, E. Telzak, D. Hewlett, Jr., V. Sharp, P. Chiliade, V. La Bombardi, D. Kabus, B. Hanna, L. Palumbo, K. Brudney, A. Weltman, K. Stoekle, K. Chirgwin, M. Simberkoff, S. Moghazeh, W. Eisner, M. Lutfey, and B. Kreiswirth.** 1997. A city-wide outbreak of a multiple-drug-resistant strain of *Mycobacterium tuberculosis* in New York. Int. J. Tuberc. Lung Dis. **1:**115–121.
20. **Salamon, H., M. R. Segal, A. Ponce de Leon, and P. M. Small.** 1998. Accommodating error analysis in comparison and clustering of molecular fingerprints. Emerg. Infect. Dis. **4:**159–168.
21. **van Embden, J. D., M. D. Cave, J. T. Crawford, J. W. Dale, K. D. Eisenach, B. Gicquel, P. Hermans, C. Martin, R. McAdam, T. M. Shinnick, et al.** 1993. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: recommendations for a standardized methodology. J. Clin. Microbiol. **31:** 406–409.
22. **Yeh, R. W., A. Ponce de Leon, C. B. Agasino, J. A. Hahn, C. L. Daley, P. C. Hopewell, and P. M. Small.** 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J. Infect. Dis. **177:**1107–1111.