# *Mycobacterium tuberculosis* lineage: a naming of the parts

T.D. McHugh[1*], S.L. Batt[1], R.J. Shorten[1], R.D. Gosling[1], L. Uiso[2] & S.H. Gillespie[1]

Running title: *M. tuberculosis* lineage

1. Centre for Medical Microbiology
Department of Infection
University College London
Royal Free Campus
Rowland Hill Street
Hampstead
London
NW3 2PF

2. Kibong'oto National Tuberculosis and Leprosy Hospital,
Sanya Juu
Tanzania

* Corresponding Author
t.mchugh@rfc.ucl.ac.uk
Tel. + 44 (0) 20 7472 6402
Fax + 44 (0) 20 7794 0433

**Abstract**

There have been many reports of groups of related *Mycobacterium tuberculosis* strains described variously as lineages, families or clades. There is no objective definition of these groupings making it impossible to define relationships between those groups with biological advantages. Here we describe two groups of related strains obtained from an epidemiological study in Tanzania which we define as the Kilimanjaro and Meru lineages on the basis of IS*6110* restriction fragment length polymorphism (RFLP), polymorphic GC rich sequence (PGRS) RFLP and mycobacterial interspersed repeat unit (MIRU) typing. We investigated the concordance between each of the typing techniques and the dispersal of the typing profiles from a core pattern. The Meru lineage is more dispersed than the Kilimanjaro lineage and we speculate that the Meru lineage is older.

We suggest that this approach provides an objective definition that proves robust in this epidemiological study. Such a framework will permit associations between a lineage and clinical or bacterial phenomenon to be tested objectively. This definition will also enable new putative lineages to be objectively tested.

## Introduction

Until relatively recently *M. tuberculosis* was thought to be a highly homogeneous species with differences in disease presentation and complications being due to difference in host response (1). The organism has proved itself very adaptable as demonstrated by the ability of mycobacterium to be transmitted and by its ability to adapt to new environments (8). Study of the virulence of *M. tuberculosis* has been handicapped by the paucity of tools to differentiate the organism into different types.

This situation has been transformed by the description of a number of methods of subdividing isolates of the genus including IS*6110* RFLP, spoligotyping, PGRS typing, MIRU and deletion analysis (10, 11, 13, 17, 19). These techniques were first applied to epidemiological studies and outbreak investigations (7, 12). When applied to very large collections of strains, those strains with similarities have been identified. For example a group of strains has been identified by IS*6110* and spoligotyping and designated the Beijing family (3). This is a group of strains of considerable importance as it includes the organisms implicated in the "strain W" outbreaks in the United States (15). Also it has been suggested that Beijing strains may be associated with an enhanced febrile response in patients on treatment and multiple drug resistance may be more common in strains of this family (18).

It is generally accepted that 100% identity by IS*6110* type is found between strains that are related and may be defined as a 'cluster' (9). Clustering is used as a surrogate marker for recent transmission, even when the direct relationships between the patients infected have not been established. For strains that are more distantly related

this 100% rule is likely to be broken. Recent analysis of the evolutionary relationships between strains of *M. tuberculosis,* using deletion analysis, has been able to root studies of the molecular epidemiological associations of isolates in the evolutionary tree for this organism (4). Analysis of sequential samples suggests that the mean time between IS*6110* transposition events is 0.5 - 5 years (20). Thus, the speed of the molecular clock for deletions is likely to be at least an order of magnitude slower than that for the molecular markers used in epidemiological studies. Different research groups have variously applied different degrees of similarity as defined by the Dice coefficient of between 40-95 % calling these 'families', 'groups' or 'clades'(5, 14, 16). There are no agreed definitions of what constitutes a significant collection of isolates or indeed what it should be called. In this study we have adopted the term lineage.

It is clear that an objective definition of a lineage, or rules whereby a lineage can be identified and assessed, is required. To do this we studied two groups of related strains obtained in an epidemiological study in Tanzania which we typed by IS*6110* RFLP, PGRS RFLP and MIRU PCR to determine the anatomy of a lineage, and to assist in the proposition of rules for lineage definition.

## 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 Kibong'oto Hospital over the 6 month period April

- September 1995. Speciation was confirmed by standard microbiological techniques. Isolates were maintained on Löwenstein-Jenson (LJ) 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 (6).

**Clinical/Epidemiological data.** The following data was collated for each isolate; age, sex, district of domicile, TB smear status, HIV status. For analysis of this data, the Kruskal-Wallis test was used for non-parametric continuous data, i.e. age, and categorical data was compared using the Chi square statistic.

**Molecular analysis.** We have previously reported the molecular analysis of these isolates (6) by IS*6110* and PGRS typing. In brief, isolates of *M. tuberculosis* were genetically fingerprinted using IS*6110* RFLP typing using the international standard protocol (19). All patterns were entered by one researcher (SB) onto a database using Bionumerics software (Applied Maths, Koutrai, Belgium). All available isolates were submitted to PGRS analysis. Genomic DNA was digested with *Alu* I restriction endonuclease and a Southern Blot probed using an oligonucleotide consisting of two copies of the PGRS consensus repeat (6).

MIRU ~~VNTR~~ typing was performed using the technique described by Supply (17). PCR mixtures were prepared as follows, using the HotStartTaq DNA polymerase kit (Qiagen, Crawley, West Sussex, UK). A final volume of 50 µl containing 1 U of DNA polymerase, 10 µl of Q solution, 0.2 mM (each) dATP, dCTP, dGTP, and dTTP, 5 µl of x10 PCR buffer, 0.4 µM (each) primer, 1 µl DNA, 25.8 µl of water and a final $MgCl_2$ concentration of 2.5mM. The PCR reactions were carried out using a

OmniGene thermocycler (Hybaid, Ashford, Great Britain), starting with a denaturing step of 15 min at 95°C, followed by 40 cycles of 1 min at 94°C, 1 min at 59°C, and 1 min 30 s at 72°C, followed by a final incubation at 72°C for 10 min. PCR products were sized using an 11cm, 2% agarose electrophoresis with 20 bp Super Ladder-low and 100 bp Super Ladder-low (Gensura, San Diego, California, USA).

**Cluster analysis.** Comparison of DNA fingerprints was performed using the Bionumerics Edition 3.0 package (Applied Maths, Kourtrai, Belgium). Cluster analysis of profiles was performed by calculation of the Dice coefficient; optimization was set at 1% and position tolerance at 1.2%. A cluster was defined as a series of isolates with 100% identity, a putative lineage was identified as a series of isolates with 70% or greater similarity by IS*6110* RFLP pattern.

On the basis of IS*6110* type, 2 putative lineages were identified. A putative lineage was defined as series of isolates, over represented in the population (greater than 10% of the total, with no evidence of recent transmission), and greater than 70% similarity by the Dice coefficient on IS*6110* typing. DNA for each putative lineage was submitted to MIRU~~-VNTR~~ typing.

**Results**

A total of 246 sequential isolates of *Mycobacterium tuberculosis* from the National Tuberculosis Control Program of Tanzania's Zonal TB laboratory in Moshi, Kilimanjaro Region were collected in 1995 and typed.

RFLP patterns of 219 patients were obtained and patterns from 195 were entered into Bionumerics 3.0 software and a dendrogram was drawn: twelve isolates had unreadable RFLP patterns and were excluded and 12 were duplicate samples (Figure 1). Excluded isolates were not significantly different in any respect when compared to those included in the dendrogram. Fifty two isolates with 4 copies or fewer were designated 'low copy isolates' and excluded from further molecular analysis. Two groups of high copy number isolates with RFLP patterns that had a similarity of 70% or greater using the Dice coefficient were seen: the largest consisted of 33 out of the 195 (16.9%) isolates designated the Kilimanjaro lineage and the second represented 31 (15.9%) isolates designated the Meru lineage.

We examined those Kilimanjaro and Meru lineage isolates still available further using 2 other typing methods to determine how each of the strains in the lineage were related to each other (Table 1). For the Kilimanjaro lineage for 27/32 (84%) isolates had a similarity of 70% using PGRS. Twenty isolates were available for MIRU typing had an overall similarity of 83%; 15/20 (75%) were identical. Of the 5 isolates with varying MIRU patterns all varied at one locus only, 2/5 by 1 repeat unit, 2/5 by 3 repeat units and the remaining isolate by 4 repeats. The Meru lineage showed a similar level of concordance between IS*6110* and PGRS with 22/26 of the isolates grouping with a similarity of 70% or greater by PGRS. On MIRU typing the 19 isolates available for testing had a similarity of 84% and 8/19 (42%) were identical. The discordant Meru MIRU images were the result of changes at up to 3 separate loci/strain and 1-3 repeat units per locus.

Using the IS*6110* data, spatial diagrams were used to map a hypothetical path for changes in the IS*6110* profile for both the Kilimanjaro and Meru lineages (Figures 2 & 3). For each lineage a 'core' IS*6110* pattern was identified and successive putative transposition events were followed (addition or loss of a band) although for the Meru lineage it was necessary to identify a hypothetical core pattern (a; Figure 3). The MIRU data were then superimposed. Comparison of the diagrams shows that Kilimanjaro lineage is least divergent with a maximum of 5 transpositional changes from its core pattern (branch A - H; Figure 2). Whereas, the Meru lineage has a maximum of 7 transpositional changes from its putative core profile (branch a - w; Figure 3) and importantly many of the links or nodes have not been identified (eg a - h). The comparison in the divergence of IS*6110* patterns between the Kilimanjaro and Meru lineages is in agreement with the MIRU data; the Kilimanjaro lineage showed variation in at locus in two patterns, whereas the Meru lineage differed at multiple loci.

The epidemiological data was examined to seek clinical correlates with lineage. Out of the 195 isolates we had clinical data on 166. Of the 29 isolates with missing data, 1 was in the Kilimajaro lineage, 5 were in the Meru lineage and 23 were not grouped. There were no significant differences between the groups with respect to age, sex, region of domicile, HIV status or smear result (Table 2).

**Discussion**

There is circularity in many of the discussions of lineages, clades and families. They are defined as similar on the basis of a single test, for example the Beijing family is

defined by spoligotyping (3), or in a restricted geographical setting (2). To break the circle we decided to arbitrarily define a lineage at 70% similarity by the Dice coefficient using IS*6110* RFLP typing and then to test this definition using other typing methods. We reasoned that if our putative lineages were truly related then they would be robust, when tested by another typing technique. In the two lineages in our collection of strains collected in Northern Tanzania, which we have named the Kilimanjaro and Meru lineages, IS*6110* typing has been confirmed by typing with alternative methods: PGRS and MIRU. The clinical epidemiological data confirmed that these isolates were not the result of direct transmission or found in a specific patient population.

Thus, our data provides important evidence about which methods are most valuable in defining new lineages. When we studied the Kilimanjaro lineage at 70% similarity we identified 33 related strains. When 20 of these strains were retested with MIRU all but 5 had an identical MIRU number. Moreover, when the divergent strains were included, the similarity of the group was 84% suggesting that the divergent patterns emerged from the majority MIRU type for the Kilimanjaro lineage. In the same way, all of the differences in the IS*6110* profiles of strains included in the Kilimanjaro lineage followed a pattern that was predictable; it was possible to track the changes, either a gain of a band at a new site or the loss of a band (see Figures 2 & 3). Undoubtedly, given the limited scale of this study we have not identified all of the possible IS*6110* types for the Kilimanjaro lineage and there are other strains yet to be identified. It is reassuring that the lineage demonstrated by IS*6110* is also a lineage by MIRU typing. We found that PGRS typing did confirm the associations that defined a lineage but also included unrelated isolates, for example at the 70% level 38 isolates

were grouped with Kilimanjaro lineage isolates, but only 28 of these were defined as Kilimanjaro by IS*6110*. PGRS although technically straight forward presents problems for interpretation, thus we would suggest that it lacks the discrimination necessary to define a biologically relevant lineage.

In the Kilimanjaro lineage, 2 strains had an identical IS*6110* pattern which we designated the core pattern. Using this group of organisms as a root, all of the other strains included in the lineage could be linked and the differences identified as single or double changes in the IS*6110* banding pattern, reflecting transpositional events. Using this same strategy differences were tracked in the Meru lineage, which appeared more dispersed. Although the figures identified type strains for each lineage and changes are 'tracked' these designations are purely arbitrary. We are unable to identify any true ancestor for these lineages and it is illogical to look for one as strains change and adapt in their interaction with a range of human hosts (8).

Although, addition of MIRU data confirmed that the Meru lineage is more dispersed than the Kilimanjaro lineage. This may be because the Meru lineage is older than the Kilimanjaro lineage. For each of the lineages we have been able to track changes from the 'core' IS*6110* pattern to all of the other strains included within our definition supporting the idea that the strains are related.

We can see clearly that both of these groups of organisms are over represented in this community and this raises the question why? Our study can throw little light on this question although there was a trend to older age in patients infected with Kilimanjaro lineage strains. This may indicate that this strain was introduced into the community

some decades ago and spread widely in the child population of the day and now these strains are re-emerging as re-activation tuberculosis develops.

Over-representation of a lineage may suggest that it possesses a biological advantage and study of such strains may help us to understand the characteristics that make *M. tuberculosis* such an effective pathogen. Any particular type could be over-represented in a collection of strains in several different ways. Some of these may have no bearing on the pathogenicity of an organism. For example strains taken from an outbreak are an obvious example where there presence in a collection does not necessarily represent strains with a biological advantage. Their presence is because of direct transmission (a characteristic of all strains) not because the strains have spread widely in a community (a characteristic of strains with an enhanced ability for transmission). Thus, when a set of cultures is being examined for the presence of a lineage, strains that are identical should only be included in the calculation if there is no evidence of direct transmission.

With this experience we believe that some rules may be proposed for establishing a lineage. We propose that in defining a lineage it is necessary to use an initial typing system with a discriminatory ability at least as good as or similar to MIRU. In this retrospective analysis we already had an IS*6110* database and so confirmed associations with MIRU. If IS*6110* typing is used as a preliminary screen identical strains should be included in the lineage if they are not linked directly i.e., part of outbreaks or close relatives. Non-identical strains can also be included using a lower cut-off, provided the secondary typing such as MIRU confirms similarity. Strains that differ by more than one step could be provisionally included in the lineage until

further information is obtained. When MIRU is used as the preliminary screen the lower discrimination of this technique should be considered. In our study we did not perform MIRU on all of the strains and confirm similarity by IS*6110* and thus cannot be certain of the correct approach in this circumstance. However, our data does indicate that a lineage defined by MIRU should have an identical number and strains should be included if the IS*6110* confirms this. Now we know that it is likely that many different IS*6110* patterns will be obtained, but in members of the lineage it will be possible to see how one strain pattern could have developed into that of another member of the lineage, this trackability is illustrated in Figures 3 and 4.

Our study does not investigate the use of spoligotyping as a method of lineage definition but the literature contains a considerable amount of evidence to suggest how this method could be used. The Beijing family is defined on the basis of spoligotyping as isolates containing spacers 35 to 43, or a subset of these spacers (3). When IS*6110* typing is performed on isolates with this spoligotype a wide range of IS*6110* types are found and we found it impossible to track the changes as described in this paper. This can be explained if the Beijing lineage is older than the Kilimanjaro and Meru lineages and the molecular clock for spoligotyping is sufficiently slow to not show many changes. An alternative explanation is that the spoligotype of the Beijing lineage has arisen more than once and thus the lineage contains different IS*6110* types. Thus studies of this lineage must use carefully defined collections of isolates to ensure reproducibility.

The approach we have adopted in this study provides a robust framework in which it is possible to test the hypothesis that the associations of *M. tuberculosis* isolates

defined by molecular typing methods do have a biological significance which is manifest in the clinical outcome of disease. The definitions that we propose would ensure that only strains that are related in a defined way are included in lineage studies.

**Acknowledgements**

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

**References**

1. Bellamy R., Hill AV. Genetic susceptibility to mycobacteria and other infectious pathogens in humans. *Curr. Opin. Immunol*. 1998; **10:** 483-487.
2. Bhanu NV, van Soolingen D, van Embden JD, *et al*. Predominace of a novel Mycobacterium tuberculosis genotype in the Delhi region of India. *Tubercul*. 2002; **82:**105-112.
3. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN. Global dissemination of the Mycobacterium tuberculosis W-Beijing family strains. *Trends Microbiol*. 2002; **10:** 45-52.
4. Brosch R, Pym AS, Gordon SV, Cole ST. The evolution of mycobacterial pathogenicity: clues from comparative genomics. *Trends Microbiol.* 2001; **9:** 452-8.
5. Dale JW, Al-Ghusein H, Al-Hashmi S, *et al*. Evolutionary relationships amongst strains of *Mycobacterium tuberculosis* with few copies of IS*6110*. *J. Bact.* 2003; **185:** 2555-2562.
6. Gillespie SH, Dickens A, McHugh TD. False molecular clusters due to non-random association of IS*6110* with *Mycobacterium tuberculosis*. *J. Clin.*

*Microbiol* 2000; **38:** 2081-2086.

7. Gillespie SH, Kennedy N, Ngowi FI, *et al.* Restriction fragment length polymorphism analysis of Mycobacterium tuberculosis isolated from patients with pulmonary tuberculosis in northern Tanzania. Trans *R Soc Trop Med Hyg*. 1995; **89:** 335-338.
8. Gillespie SH, Billington OJ, Breathnach A, McHugh TD. The fitness of a multiple drug resistant *Mycobacterium tuberculosis* strain changes after transmission between hosts during and outbreak. *Micro. Drug Res* 2002;. **8:** 273 - 280.
9. Glynn JR, Bauer J, de Boer AS, *et al*. Interpreting DNA fingerprint clusters of Mycobacterium tuberculosis. European Concerted Action on Molecular Epidemiology and Control of Tuberculosis. *Int J Tuberc Lung Dis.* 1999; **3:** 1055-60.
10. Kamerbeek J, Schouls L, Kolk A, *et al*. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *J. Clin. Microbiol.* 1997; **35:** 907-914.
11. Kato-Maeda M, Rhee JT, Gingeras TR, *et al*. Comparing genomes within the species Mycobacterium tuberculosis. *Genome Res.* 2001; **11:** 547-554.
12. Maguire H, Dale JW, McHugh TD, *et al*. Molecular epidemiology of tuberculosis in London 1995 to 1997 demonstrating low rate of active transmission. *Thorax* 2002; **57:** 617-622.
13. McHugh TD, Dickens A, Gillespie SH. False molecular clusters due to non-random association of IS*6110* with *Mycobacterium tuberculosis*. *J. Clin. Microbiol.* 2000; **38:** 2081-2086.
14. McHugh T D, Gillespie SH. Nonrandom association of IS*6110* and *Mycobacterium tuberculosis*: implications for molecular epidemiological studies.

*J. Clin. Microbiol*. 1998; **36:** 1410-1413.

15. Moss AR, Alland D, Telzak E, *et al*. A city-wide outbreak of a multiple-drug-resistant strain of Mycobacterium tuberculosis in New York. *Int J Tuberc Lung Dis*. 1997; **1:** 115-121.
16. Sola C, Filliol I, Legrand E, *et al*. Genotyping of the Mycobacterium tuberculosis complex using MIRUs: association with VNTR and spoligotyping for molecular epidemiology and evolutionary genetics. *Infect Genet Evol.* 2003; **3:** 125-33.
17. Supply P, Lesjean S, Savine E, *et al*.. Automated high-through-put genotyping for study of global epidemiology of *Mycobacterium tuberculosis* based on mycobacterial interspersed repetitive units. *J. Clin. Microbiol.* 2001; **39**: 3563-3571.
18. van Crevel R, Nelwan RH, de Lenne W, *et al*. Mycobacterium tuberculosis Beijing genotype strains associated with febrile response to treatment. *Emerg Infect Dis.* 2001; **7:** 880-3.
19. van Embden J DA, Cave MD, Crawford JT, et al.. Strain identification of *Mycobacterium tuberculosis* by DNA fingerprinting: Recommendations for a standardized methodology. *J. Clin. Microbiol*. 1993; **31:** 406-409.
20. Yeh RW, Ponce de Leon A, Agasino CB, *et al*. Stability of Mycobacterium tuberculosis DNA genotypes. *J Infect Dis*. 1998; **177:** 1107-11.

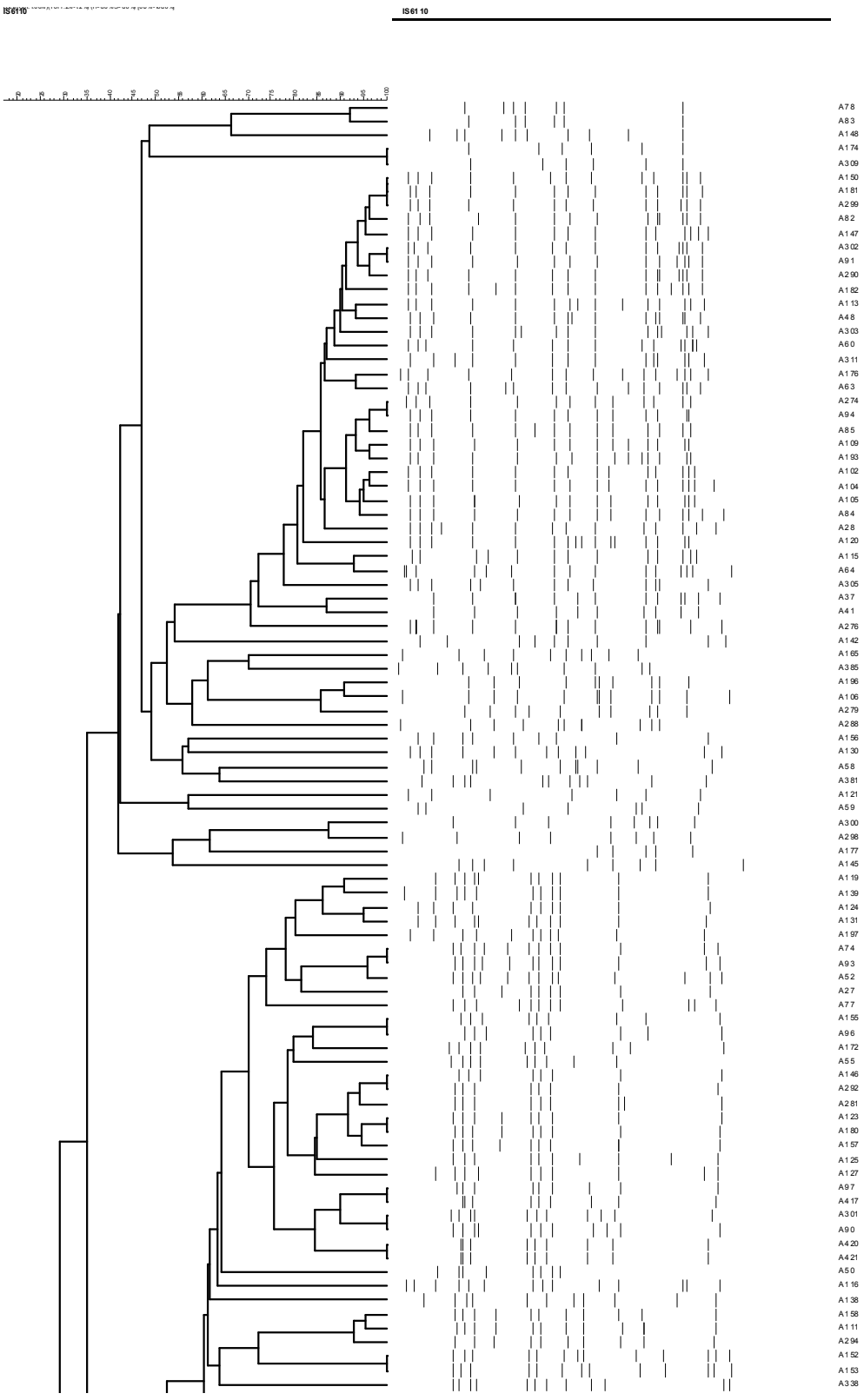
**Figure Legend**

Figure 1: Master dendrogram of relationships between IS*6110* profiles of 171 isolates of *M. tuberculosis* as calculated by the Dice coefficient.

Figure 2: Schematic representation of Kilimanjaro lineage. 28 patterns designated A - B1 represent 33 isolates in the lineage. Pattern A is designated the 'core pattern' and has two isolates.
← denotes the addition of a single band, ⇇ denotes the addition of one or two bands giving a double band, ○ denotes the loss of a band, ■ denotes a an extra transpositional event where the intermediate pattern is not present (designated a 'minor node'), ● denotes divergence from one pattern to two (designated a 'major node'), * denotes the number of strains represented by each pattern.

Figure 3: Schematic representation of Meru lineage. 22 patterns designated b - w represent the 31 isolates in the Meru lineage. Pattern a represents the hypothetical core pattern. Key as Figure 2.

**Figure 1**


IS6110
IS6110
A78
A83
A148
A174
A309
A150
A181
A299
A82
A147
A302
A91
A290
A182
A113
A48
A303
A60
A311
A176
A63
A274
A94
A85
A109
A193
A102
A104
A105
A84
A28
A120
A115
A64
A305
A37
A41
A276
A142
A165
A385
A196
A106
A279
A288
A156
A130
A58
A381
A121
A59
A300
A298
A177
A145
A119
A139
A124
A131
A197
A74
A93
A52
A27
A77
A155
A96
A172
A55
A146
A292
A281
A123
A180
A157
A125
A127
A97
A417
A301
A90
A420
A421
A50
A116
A138
A158
A111
A294
A152
A153
A338

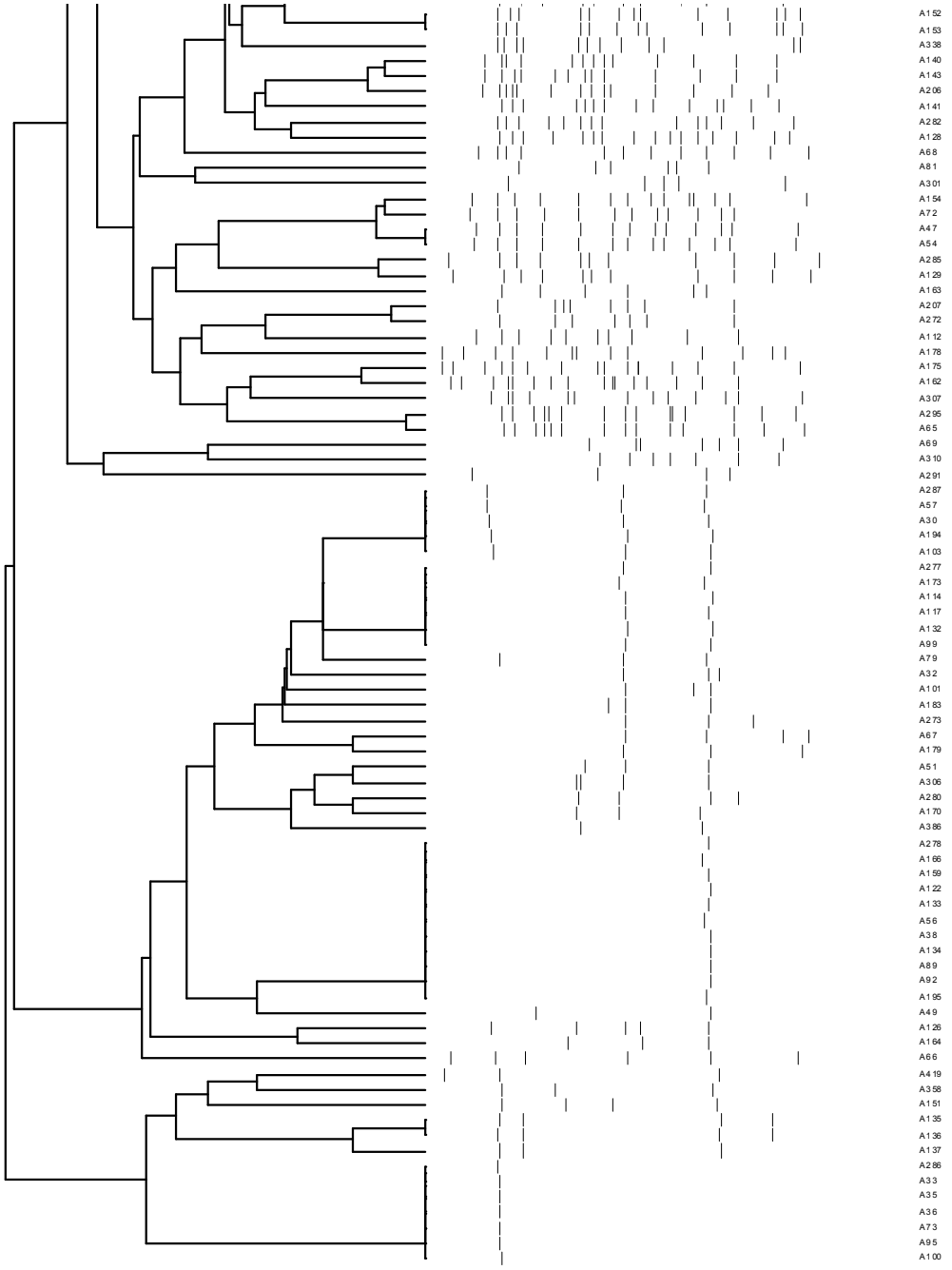



**Figure 1 (continued)**

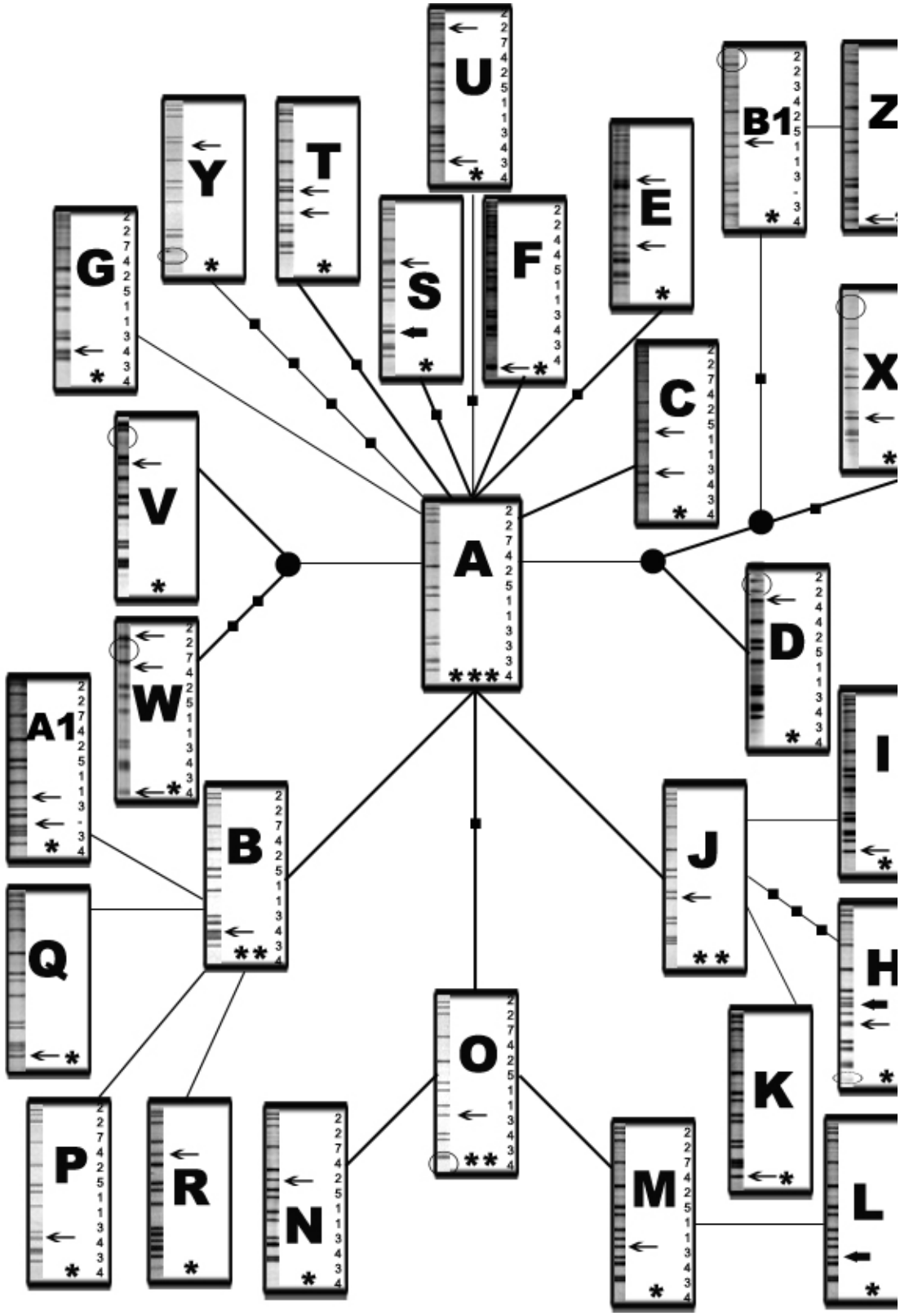



**Figure 2**

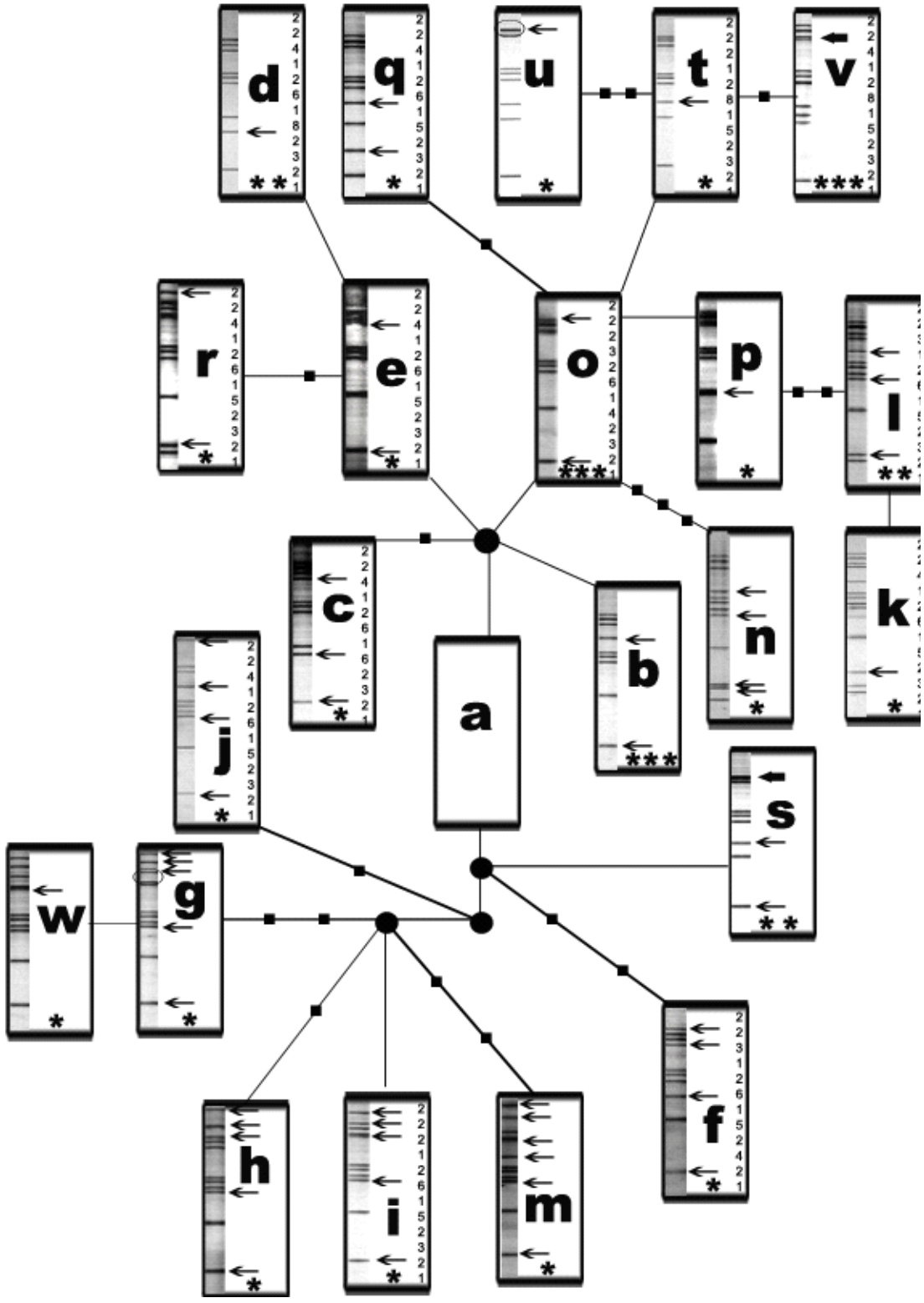



**Figure 3**Table 2: Comparison of the principle characteristics of members of lineages and non members for which clinical data was available.

| | Total Number | of isolates in a | analysis | Kilimanjaro | lineage |
|---|---|---|---|---|---|
| | Meru lineage | Not grouped | Significance | | |

| | | | | | |
|---|---|---|---|---|---|
| Number (%) | 166 (100) | 33 (19.9) | 26 (15.7) | 107 (64.5) | |
| Median age in years (range) | 33.0 (15-70) | 165 | 40.0 (19-70) | | 37.5 (19-60) |
| Male Sex (%) | 166 | 25 (75.8) | 19 (73.1) | 77 (72.0) | Not significant ** |
| HIV positive (%) | 162 | 14 (42.4) | 10 (40.0) | 38 (36.5) | Not significant ** |
| Smear positive (%) | 161 | 24 (81.3) | 20 (76.9) | 85 (70.9) | Not significant ** |

Not significant*

* Kruskal- Wallis
** Chi squared

Table 1: Summary of the total number of isolates tested by each test and concordance of results.

| Lineage | Number in lineage: 70% similarity by IS*6110* (% all isolates)<br>Number that correspond @ 70% for PGRS (% of IS*6110*)<br>Number MIRU performed<br>Number that correspond @ 100% by MIRU | | % similarity by Dice coefficient | |
|---|---|---|---|---|
| Kilimanjaro (75) | 33 (16.9) | 27/32 (84) | 20 | 83 15 |
| Meru (42) | 31 (15.9) | 22/26 (85) | 19 | 84 8 |