Available online at www.sciencedirect.com
ScienceDirect
Infection, Genetics and Evolution 8 (2008) 182–190
Infection, Genetics and Evolution
www.elsevier.com/locate/meegid

Interpreting genotype cluster sizes of *Mycobacterium tuberculosis* isolates typed with IS*6110* and spoligotyping

Fabio Luciani [a,*], Andrew R. Francis [b], Mark M. Tanaka [a,c]

[a] *School of Biotechnology and Biomolecular Sciences, University of New South Wales, NSW 2052, Australia*
[b] *School of Computing and Mathematics, University of Western Sydney, Australia*
[c] *Evolution & Ecology Research Centre, University of New South Wales, Australia*

Received 10 September 2007; received in revised form 12 December 2007; accepted 12 December 2007
Available online 1 February 2008

Abstract

Molecular techniques such as IS*6110*-RFLP typing and spacer oligonucleotide typing (spoligotyping) have aided in understanding the transmission patterns of *Mycobacterium tuberculosis*. The degree of clustering of isolates on the basis of genotypes is informative of the extent of transmission in a given geographic area. We analyzed 130 published data sets of *M. tuberculosis* isolates, each representing a sample of bacterial isolates from a specific geographic region, typed with either or both of the IS*6110*-RFLP and spoligotyping methods. We explored common features and differences among these samples. Using population models, we found that the presence of large clusters (typically associated with recent transmission) as well as a large number of singletons (genotypes found exactly once in the data set) is consistent with an expanding infectious population. We also estimated the mutation rate of spoligotype patterns relative to IS*6110* patterns and found the former rate to be about 10–26% of the latter. This study illustrates the utility of examining the full distribution of genotype cluster sizes from a given region, in the light of population genetic models.

Keywords: Tuberculosis; Population genetics; Genetic diversity; Evolution; Molecular epidemiology; Mutation; Genotype; IS*6110*; Spoligotype

1. Introduction

Tuberculosis remains a major global infectious disease, causing around two million deaths each year. While traditional epidemiological methods such as contact tracing are central to attempts to contain tuberculosis, these are increasingly complemented by use of data from molecular typing techniques (Small et al., 1994; Kodmon et al., 2006; Lari et al., 2005; Quitugua et al., 2002; Chan-Yeung et al., 2003; Van Soolingen et al., 1999; Mathema et al., 2006). DNA fingerprinting methods have enabled the classification of isolates into distinct strains, and thus the characterization of genetic diversity of *Mycobacterium tuberculosis* in outbreaks (Nguyen et al., 2004; Van Soolingen, 2001). The most commonly applied molecular typing techniques are IS*6110*-restriction fragment length polymorphism (IS*6110*-RFLP) (Eisenach et al., 1988), and spacer oligonucleotide typing (spoligotyping) (Kamerbeek et al., 1997). The former uses variability produced by movement of the insertion sequence IS*6110*, while the latter identifies the presence or absence of 43 DNA spacer sequences located between repetitive units at the direct repeat (DR) locus. The IS*6110*-RFLP and spoligotyping methods are sufficiently discriminating to separate unrelated isolates, and yet the resulting patterns are stable enough to allow closely related isolates to be grouped (Niemann et al., 1999; Van Soolingen, 2001). The use of these typing techniques to study the epidemiology of tuberculosis has now become widespread, with over 150 papers appearing in 2006 alone that mention IS*6110* or spoligotyping (Fig. 1).

Clusters of identical or highly similar genotypes are widely interpreted as resulting from recent transmission caused by a single case. On the other hand, genotypes appearing uniquely within a data set, named singletons from here on, are generally considered to have arisen from migration or recent reactivation of remotely acquired infections. In addition to transmission rates, several other factors influence the patterns of clustering.

* Corresponding author. Tel.: +61 2 9385 3701; fax: +61 2 9385 1483.
E-mail address: luciani@unsw.edu.au (F. Luciani).

1567-1348/$ – see front matter
doi:10.1016/j.meegid.2007.12.004

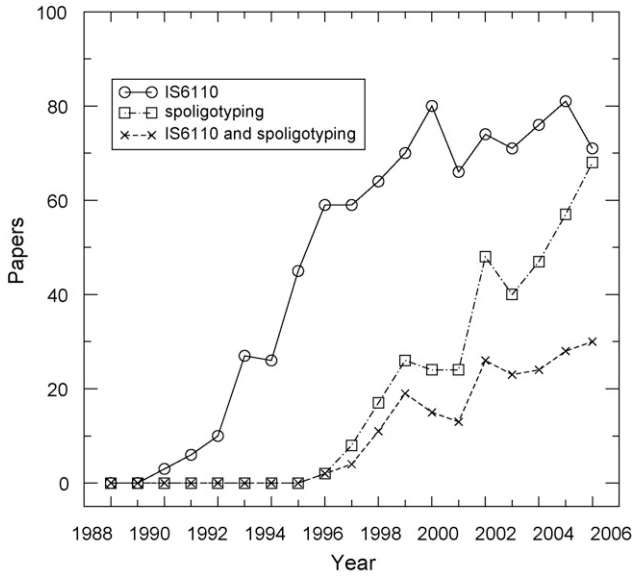

Fig. 1. Number of publications concerning IS*6110*-RFLP and spoligotyping schemes in molecular epidemiology of tuberculosis. Data were obtained from PubMed (http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed).

These include sampling and the mutation rate of the molecular marker used for fingerprinting (Tanaka and Francis, 2005). To form conclusions about transmission from such data, it is therefore necessary to develop an understanding of generic features of distributions of cluster sizes, and the mutation rates of different markers.

The infinite alleles model (IAM), developed in population genetics, provides a theoretical framework for analyzing the distribution of the number of copies of alleles in a sample. The IAM describes the balance between mutation and random genetic drift under the assumption of selective neutrality (Ewens, 2004; Hubbell, 2001). By viewing molecular patterns resulting from genotyping as alleles of a genetic locus, the IAM can be applied as a baseline model to study bacterial variation. To make this application, we equate within-host substitution – that is, the combination of both mutation and fixation within the host – with mutation of genotypes.

Here, we analyzed data extracted from 130 molecular epidemiological studies of tuberculosis published in the biomedical literature. With the exception of one set of globally sampled isolates, each of these data sets represents a sample of bacterial isolates collected from an outbreak in a specific geographic area, and typed with either or both of the IS*6110*-RFLP and spoligotyping methods. We explore common features and differences between the data sets by applying the infinite alleles model. With this model, we obtain a measure of the genetic diversity in each data set, and use this measure to provide estimates of the mutation rate of spoligotype patterns. Finally, we discuss the epidemiological implications of our findings.

2. Analysis of data using population genetic models

Of the 130 data sets forming the basis of this study, 64 report the cluster size of every genotype represented in the data. Others include less specific information, for instance, many report only the total number of genotypes and the total number of isolates (see Supplementary material for more information about the data sets included in this study). Those data sets that use both IS*6110*-RFLP and spoligotyping in their studies often use spoligotyping to further discriminate the clusters corresponding to low copy number IS*6110* fingerprints. For such data, we considered only one marker at a time to define the clusters, in order to legitimately apply the IAM. In this section, we investigate the suitability of the infinite alleles model to explain the observed configurations of clusters by comparing common features of the data with the prediction based on the IAM (Section 2.2), as well as by using some established statistical tests (Section 2.3).

2.1. The infinite alleles model and genetic diversity

The infinite alleles model is a well-studied population genetic model describing the evolution of a population undergoing mutation and genetic drift (Kimura and Crow, 1964; Ewens, 2004). The model assumes a constant population size, that each mutation event results in a genotype never encountered before and that all genotypes are selectively equivalent. In this case, alleles correspond to the genotypes in our data. Although a pattern that already exists in the population can possibly be generated again through a new mutation event, such events are likely to be relatively rare. While in the case of IS*6110*, the probability of regenerating an identical pattern would be extremely low, for spoligotypes, this possibility is higher because there are fewer possible patterns. However, the frequency of convergence (homoplasy) is still likely to be low, because mutation occurs through deletion of spacers, and the regeneration of a pre-existing pattern requires a suitable parental genotype to exist and a specific deletion event to occur in this potential parent. Under this model, the distribution of genotype frequencies at equilibrium has been well studied (Ewens, 2004; Kimura, 1983). A major advance within the IAM framework was the analytical treatment of the effects of sampling (Ewens, 2004), which is useful in the context of tuberculosis data sets, where sample size is an important consideration.

The fundamental parameter of this model is $\theta = 2N_e\mu$, where μ is the rate at which new genotypes arise per individual per generation, and N_e is the effective population size. N_e is a key parameter in population genetics representing the size of an ideal population in which genetic drift operates at the same rate as in an actual population. N_e is a measure of how readily a population maintains genetic variation (see Ewens, 2004 for more details). The parameter θ is a measure of genetic diversity in the population. A maximum likelihood estimator $\hat{\theta}$ for the parameter θ can be obtained from data using only the sample size n and the number g of distinct genotypes in the sample (see Ewens, 2004 for details). Note that cluster sizes are not needed to compute $\hat{\theta}$.

Fig. 2 shows the $\hat{\theta}$ values estimated from the 130 data sets (76 using IS*6110*-RFLP and 54 using spoligotyping). The $\hat{\theta}$ values obtained from the IS*6110* and spoligotype data sets form

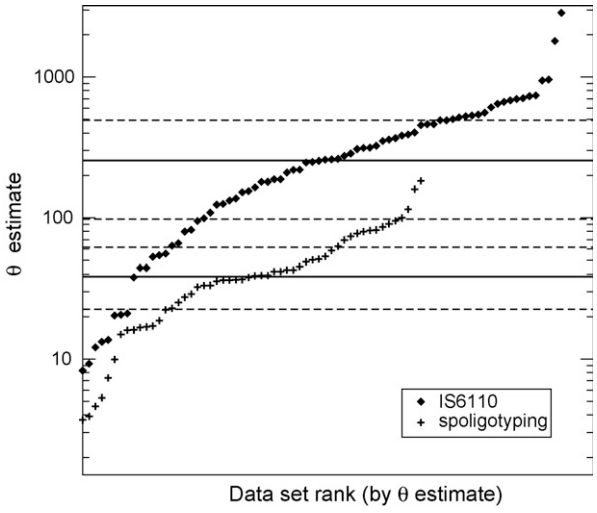

Fig. 2. Estimated values $\hat{\theta}$ of the parameter θ from the 130 data sets (76 with IS*6110* and 54 with spoligotyping) considered in this work, ranked by $\hat{\theta}$. Diamonds represent the IS*6110* subset, and "+" the spoligotype data sets. Continuous horizontal lines represent the median values. The corresponding upper and lower dashed lines represent the first and the third quartiles, respectively.

two overlapping but distinct distributions. The $\hat{\theta}$ values from the IS*6110* data have higher median and mean values (255.9 and 349.5, respectively) than those arising from the spoligotype data (38.42 and 47.11, respectively). Because the parameter θ reflects the genetic diversity of a population (Ewens, 2004; Hubbell, 2001), this result confirms that the IS*6110* typing scheme is more effective than spoligotyping in discriminating *M. tuberculosis* strains (Kremer et al., 1999).

2.2. Features of cluster size distributions: large clusters and singletons

A casual glance at a selection of data sets reporting cluster sizes for individual genotypes reveals that most data sets contain both a large number of genotypes represented by a single isolate (singletons), and a number of genotypes with significantly larger clusters than most. Natural questions include whether the IAM explains these features and whether they are useful indicators of the extent of tuberculosis transmission.

We examine these features quantitatively. Let $\alpha(j)$ be the number of genotypes having cluster size j. Conspicuously large cluster sizes (e.g., 15 isolates of a given genotype) appear in many data sets, particularly in those using spoligotyping. These large cluster sizes are almost always represented by a single genotype (e.g., in the IS*6110* data from Iran, $\alpha(15) = 1$). Define L to be the proportion of genotypes having unique cluster sizes in the sample. That is,

$$L = \frac{1}{g}\sum_{j=1}^{m}\alpha(j)\delta_{\alpha(j),1} \tag{1}$$

where m is the largest cluster size in the sample, g is the number of distinct genotypes and $\delta_{\alpha(j),1}$ is the Kronecker delta function, which equals 1 when $\alpha(j) = 1$ and 0 otherwise. This quantity L is used as a proxy for the proportion of large cluster sizes (the tail of the distributions $\alpha(j)$; see Fig. 3 and comments below). We computed the statistics $\alpha(1)$ and L for each of the 64 data sets that provide the full distribution of cluster sizes. The number of singletons, $\alpha(1)$, is often higher in data sets typed with the IS*6110*-RFLP scheme than in spoligotyping data.

Fig. 3 shows the distributions $\alpha(j)$ for several data sets. The number of genotypes $\alpha(j)$ generally decreases with the cluster size j. This general pattern of decline is similar across data from different geographic regions and even across data using different markers. The theoretical distribution conditioned on the sample size n, according to the IAM, is (Watterson, 1974)

$$E[\alpha(j|n)] = \frac{\theta}{j}\frac{\binom{\theta+n-j-1}{n-j}}{\binom{\theta+n-1}{n}}. \tag{2}$$

We can fit this function to the data shown in Fig. 3 by estimating θ as described in Section 2.1. This distribution exhibits the general pattern of decline, but does not always describe the large number of singletons and the presence of large cluster sizes. Indeed, the definition of L (Eq. (1)) is intended to capture the presence of large cluster sizes missed by the IAM (Fig. 3). From Eq. (2), the expected number of singletons is $E[\alpha(1)] = \theta n/(\theta + n - 1)$. This relationship between the number of singletons and the θ values is evident in Fig. 4.

To explore further whether the IAM can account for these statistics, we simulated samples under the assumptions of the model and compared simulated statistics with the observed values. We used the algorithm of (Hubbell, 2001, p. 291) to simulate distributions of genotypes in samples according to the IAM. This elegant algorithm uses the fact that in sampling a sequence of n genotypes from a population that follows the IAM, the probability that the $(j+1)$ th genotype is new to the sample is $\theta/(\theta + j)$ (see Ewens, 2004). Such a sample can be obtained by drawing n random numbers from a uniform distribution, and sequentially comparing each with the probability given above. For each of the 64 observed data sets, we generated 10,000 simulated samples (data sets) under the IAM, using its estimated $\hat{\theta}$ value (described in Section 2.1). Fig. 4 shows the relationship between $\hat{\theta}$ and each of the two statistics of interest: L and $\alpha(1)$. The observed statistics are shown with the average simulated values along with intervals indicating the central 95% of simulated values. This figure demonstrates that while the IAM is usually able to explain the data with respect to these two statistics, close inspection reveals that the IAM tends to underestimate the statistics. In extreme cases, the observed values are far outside the 95% range of simulated values.

Fig. 4 also shows that L decreases with $\hat{\theta}$ while $\alpha(1)$ increases with $\hat{\theta}$. Together, these two trends imply that L and $\alpha(1)$ are inversely related. The relationships among $\hat{\theta}$, L and $\alpha(1)$ are explained by noting that $\hat{\theta}$ is a measure of genetic diversity in the population. When diversity is high (perhaps due to a high mutation rate or a large effective population size) the

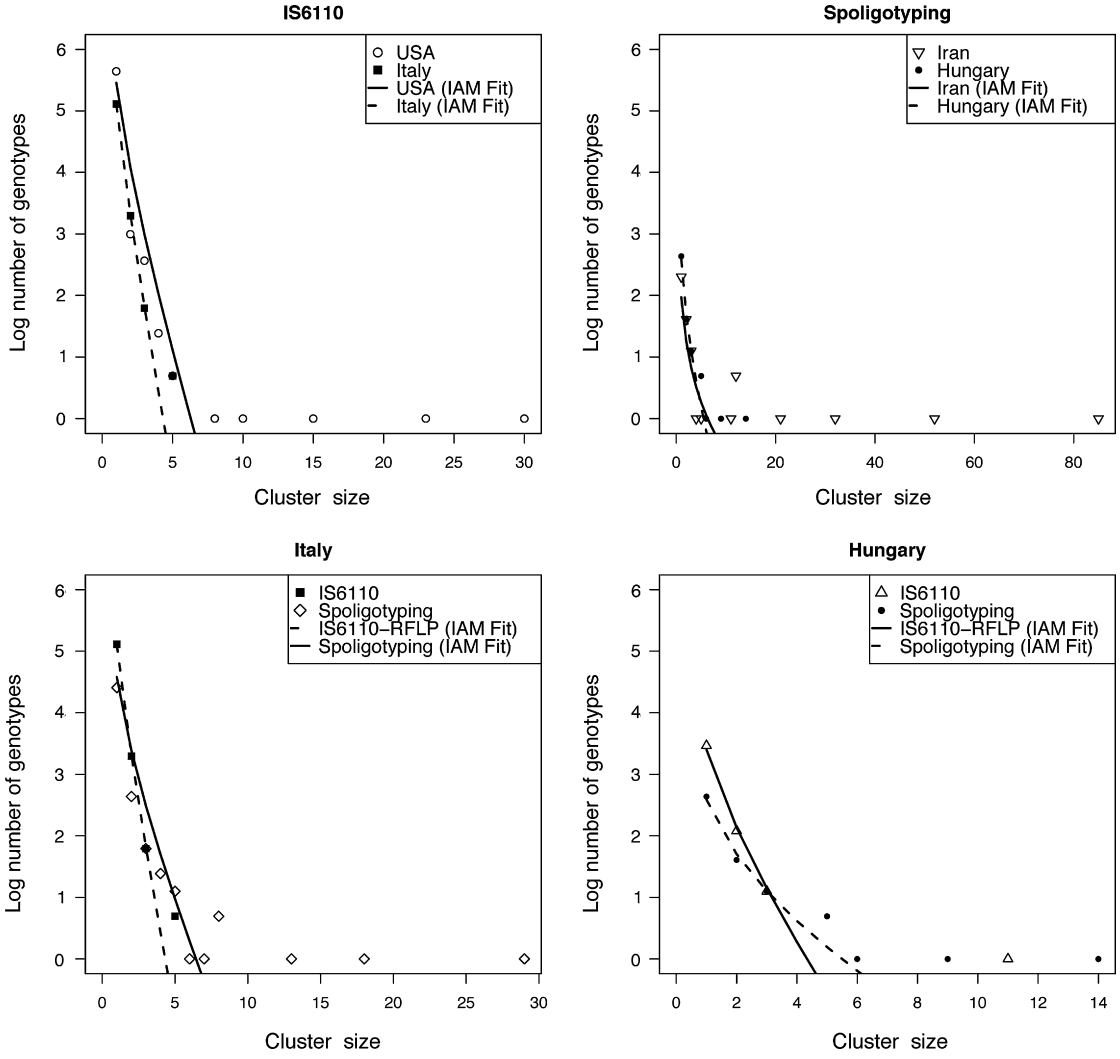

Fig. 3. Plot of the natural logarithm of the number of genotypes against cluster size, for data from different geographic areas and using different genotype markers. In each panel, the symbols represent observed data while the curves represent the expected $\alpha(j)$ under the IAM. Top left panel, two genotype distributions typed with the IS*6110* marker; data from Italy (Lari et al., 2005) and the United States (Small et al., 1994). Top right panel, two cluster size distributions obtained with the spoligotyping scheme; data from Iran (Farnia et al., 2006) and Hungary (Kodmon et al., 2006). Bottom left panel, two cluster size distributions from the same sample of isolates genotyped with both IS*6110* and spoligotyping; data from Italy (Lari et al., 2005). Bottom right panel, a sample of isolates collected in Hungary (Kodmon et al., 2006) typed with both methods.

number of distinct genotypes, and therefore the number of singletons, is also high. The scenario of high diversity would also produce a scarcity of large clusters because mutation events would break clusters down. Inversely, when diversity is low, large clusters can appear.

2.3. Testing the infinite alleles model

While the above comparison of the actual data with simulations based on the IAM gives some information about whether a data set might be from a population satisfying the assumptions of the IAM, formal statistical tests are available. We applied two standard tests, which use different aspects of the data, to assess whether the data are consistent with the IAM. The Watterson–Ewens test (Ewens, 2004) and the exact test described in Slatkin (1996a) have both been implemented and made available on the internet (Slatkin, 1996b). Rejection of the IAM may be due to the departure from selective neutrality of genotypes or other assumptions of the IAM such as a constant population size.

We tested the IAM for each of the 64 data sets (those providing the full distribution of cluster sizes). We applied a method to control the false discovery rate (Benjamini and Hochberg, 1995) for the 64 data sets within each of the two test types, using a significance level of 0.05. The IAM was deemed to have failed if either or both of the two tests resulted in rejection. That is, the null hypothesis was not rejected (the IAM was accepted) if the model was not rejected by either test, because each test accounts for a different aspect of samples under the IAM. Interestingly, the two tests disagreed in only three out of the 64 data sets. In all three cases Slatkin's exact test rejected the IAM. The full results of both tests are reported in Supplementary material.

The IAM failed in 53 (82.8%) of the 64 genotype distributions. Among the data sets for which the IAM failed,

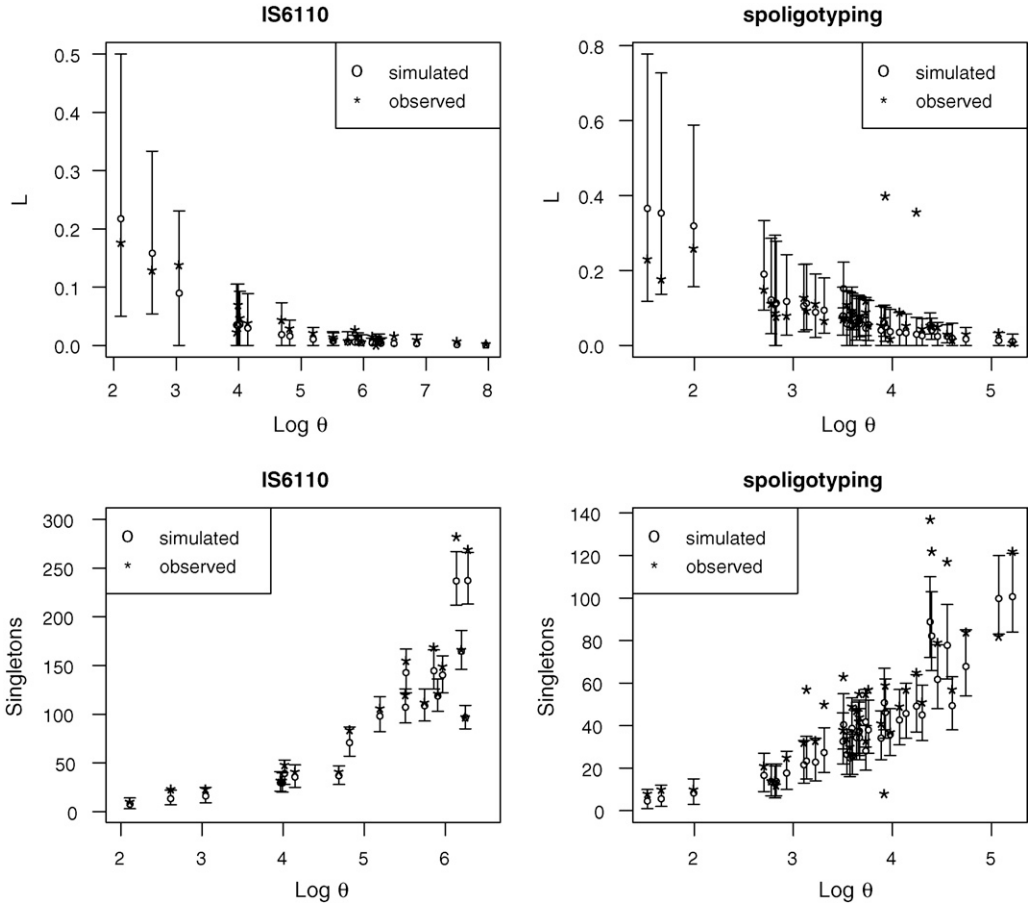

Fig. 4. Comparison of IAM simulations with the 64 data sets for which the complete genotype frequency distribution for each cluster size is available. Upper panels: plots of observed and IAM simulated values of the proportion of unique cluster sizes, L, against the natural logarithm of $\hat{\theta}$ values. Bottom panels: plots of the observed and simulated values of the number of singletons, $\alpha(1)$. Left panels: data sets typed with IS*6110* (23 data sets). Right panels: data sets typed with spoligotyping (41 data sets). Error bars indicate the central 95% of simulated values. The bottom left panel excludes two data sets, the United States (Driver et al., 2006)(observed 1535, simulations 1321.4) and Mexico (Quitugua et al., 2002)(observed 506, simulations 386.4).

the two statistics (number of singletons, $\alpha(1)$ and the large-cluster proportion L) tend to be underestimated by the model, and sometimes very strongly so. In contrast, among all data sets for which the IAM was accepted, the observed values of these statistics lay within the central 95% intervals of the simulations. For example, the IAM failed for the spoligotyping data from Poland (Sajduda et al., 2004) and India (Suresh et al., 2006) which exhibited a strong discrepancy between observed and simulated values. In contrast, the IAM was not rejected for the IS*6110* data set from Italy (Tuscany) (Lari et al., 2005) and the spoligotyping data set from Hungary (Kodmon et al., 2006)(see also Fig. 2), and both L and $\alpha(1)$ accord with the expectations of the IAM. There is general agreement between the formal approach of this section and the examination of the two statistics in Section 2.2. Namely, the data sets for which the IAM does not fail have $\alpha(1)$ and L values within the central 95% interval of the simulated values. Of the 53 data sets for which the IAM is rejected, 24 (45.3%) data sets have $\alpha(1)$ values outside the simulated central 95% interval, and 10 (18.9%) have L values outside this range. Eight of the aforementioned 53 data sets (15%) have both $\alpha(1)$ and L values outside the central 95% interval of the simulated values. Of the 34 instances for which a statistic took a value outside the central 95% interval, 31 were above this central interval, confirming the observation that under the IAM the proposed statistics are almost always underestimated.

2.4. An alternative model with expanding population size

In this section, we further investigated cluster size distributions using a model that does not constrain the population size to be constant (Tanaka et al., 2006a). We focused on one population from the United States (San Francisco, CA) (Small et al., 1994) for which the IAM was rejected. This data set is of particular interest because the population parameters under a model of growing size have been estimated using approximate Bayesian computation (Tanaka et al., 2006a). Those estimates indicate that the data from this outbreak can be explained by a growing population of infectious individuals.

The alternative process we used is a stochastic model of transmission and mutation, which we will refer to as the birth–death-mutation (BDM) model. This model is an extension of the linear birth–death process in that it includes mutation and

tracks the number of infectious individuals of different genotypes. This model, like the IAM, assumes selective neutrality of genotypes and that each mutation event produces a new genotype. The population can grow from new infections, diminish in size due to death or recovery and produce new genotypes through mutation (Tanaka et al., 2006a).

We used the parameters previously estimated to compare samples simulated under the BDM to those of the IAM. Table 1 gives the observed distribution of cluster sizes and the simulated distributions of the IAM and BDM models. We ran the IAM simulation 10,000 times to estimate the expected cluster sizes $\bar{\alpha}_{\mathrm{IAM}}(j)$, and the BDM simulation 3,000 times to estimate the corresponding expected cluster size distribution $\bar{\alpha}_{\mathrm{BDM}}(j)$ (using only the 2127 simulations that did not result in extinction of the infectious population). We also provide intervals covering the central 95% of simulated cluster size frequencies. The BDM successfully describes the observed distribution $\alpha_{\mathrm{OBS}}(j)$ with the exception of the cluster size $j = 2$ (in this case, only 5 out of the 2,127 simulations gave rise to a value less than or equal to the observed value of $\alpha_{\mathrm{OBS}}(2) = 20$). By contrast, the IAM-generated distribution reveals a strong departure from observations. In particular, the large clusters and singletons observed in the data were not generated from the simulations with the IAM.

Finally, we note that the analysis of this section represents an initial qualitative examination rather than a formal comparison of the two models. The BDM has two more parameters than the IAM and is thus expected to fit the data better. A more thorough analysis would require treating both and perhaps other models within the same statistical framework, which is beyond the scope of this paper.

Table 1
Observed cluster sizes from Small et al. (1994) and simulated cluster size frequencies according to the IAM using 10,000 runs and BDM using 2127 runs

Size (j)	$\alpha_{\mathrm{OBS}}(j)$ [a]	$\bar{\alpha}_{\mathrm{IAM}}(j)$ [b]	IAM 95% [c]	$\bar{\alpha}_{\mathrm{BDM}}(j)$ [d]	BDM 95% [e]
1	282	236.7	[211, 262]	266.4	[241,293]
2	20	55.4	[43, 69]	35.7	[25, 47]
3	13	19.0	[12, 27]	10.4	[4, 17]
4	4	7.7	[3, 13]	4.4	[1, 9]
5	2	3.2	[0, 7]	2.3	[0, 5]
6	0	1.5	[0, 4]	1.4	[0, 4]
7	0	0.71	[0, 3]	0.87	[0, 3]
8	1	0.35	[0, 2]	0.63	[0, 3]
9	0	0.20	[0, 1]	0.46	[0, 2]
10	1	0.10	[0, 1]	0.36	[0, 2]
11	0	0.05	[0, 1]	0.25	[0, 1]
12	0	0.02	[0, 0]	0.23	[0, 1]
13	0	0.01	[0, 0]	0.17	[0, 1]
14	0	0.01	[0, 0]	0.15	[0, 1]
15	1	0.004	[0, 0]	0.13	[0, 1]
16	0	0.002	[0, 0]	0.11	[0, 1]
17	0	0.0005	[0, 0]	0.08	[0, 1]
18	0	0	[0, 0]	0.08	[0, 1]
19	0	0	[0, 0]	0.06	[0, 1]
20	0	0	[0, 0]	0.06	[0, 1]
21	0	0	[0, 0]	0.06	[0, 1]
22	0	0	[0, 0]	0.05	[0, 1]
23	1	0	[0, 0]	0.06	[0, 1]
24	0	0	[0, 0]	0.04	[0, 1]
25	0	0	[0, 0]	0.04	[0, 1]
26	0	0	[0, 0]	0.04	[0, 1]
27	0	0	[0, 0]	0.04	[0, 1]
28	0	0	[0, 0]	0.03	[0, 1]
29	0	0	[0, 0]	0.03	[0, 1]
30	1	0	[0, 0]	0.03	[0, 1]
>30	0	0	[0, 0]	0.42	[0, 1]

[a] Observed cluster size distribution.
[b] Mean simulated frequencies under the IAM.
[c] Interval covering the central 95% of values obtained in the simulations using the IAM.
[d] Mean simulated frequencies under the birth–death-mutation model.
[e] Interval covering the central 95% of values obtained in the simulations using the birth–death-mutation model.

3. Estimates of the relative mutation rates

One by-product of the estimation of θ for each data set is that by considering these values for each marker (as in Fig. 3), an estimate of the ratio of the mutation rates of IS*6110* and spoligotype markers can be computed. As noted above, the parameter θ is proportional to the mutation rate μ and to the effective population number N_e. The latter is unrelated to the molecular typing methods and depends on the particular history and other properties of the population. To estimate the ratio of the two mutation rates, we assume that the distributions of N_e values (which are unknown) arising from the two subsets of data defined with IS*6110* and spoligotype markers have identical mean or median values. We then take the ratio of the means and medians.

Although differences in sample size distributions between IS*6110* and spoligotype data sets may conceivably affect estimates of θ and therefore the estimate of the relative mutation rates, an examination of the sample sizes removes this concern. Because the sample size distributions for the two markers are not normal, a Wilcoxon rank sum test was used to assess whether the distributions of the sample sizes differ. This test returned a p-value of 0.1336, indicating that the distributions of sample sizes do not affect the comparison of the θ estimates between IS*6110* and spoligotype-based studies.

We took several approaches to estimate the relative mutation rate of spoligotypes with respect to IS*6110*, summarized in Table 2. First, we used all 130 data sets. In the second approach we formed a subset of the data sets in which each geographic region was represented at most once for each kind of marker (70 in total). Third, we considered those data sets for which both IS*6110* and spoligotyping were applied to the same sample, and each geographic region was represented at most once (16 samples giving 32 data sets). For these data sets we take the mean of the ratios of $\hat{\theta}$ rather than the ratio of the means, because the effective population size N_e is the same within each pair of samples. Hence, this procedure does not make any assumptions about the values of N_e. In the fourth method we considered the 11 data sets for which the IAM was not rejected (see Section 2.3). The fifth method applied restrictions from both the second and fourth methods (each geographic region chosen at most once per marker and IAM not rejected) to consider a subset of six data sets. The estimates based on the most restricted subsets in Table 2 are perhaps the most accurate

Table 2
Estimates of the relative mutation rate of spoligotypes compared to that of IS*6110*, obtained from the θ estimates

Data source of estimates	Data sets	Ratio of means	Ratio of medians
All data sets	130	0.135	0.150
Each region represented at most once for each marker	70	0.186	0.205
All in which both IS*6110*-RFLP and spoligotyping schemes were applied and each region was represented at most once	16	0.255 [a]	0.265 [a]
All in which IAM not rejected	11	0.169	0.202
Each region represented at most once for each marker + IAM not rejected	6	0.118	0.102

[a] These estimates were calculated using the mean or median of the ratios of the pair of $\hat{\theta}$ values for each data set. See text for details.

(methods three, four and five). The values from method four and five are close to the value obtained in Tanaka and Francis (2005) (13.5%) which was based on a single global sample of genotypes. We conclude that the mutation rate of spoligotypes is likely to be around 10% to 26% of the mutation rate of IS*6110*.

4. Discussion

This study offers a new perspective, based on population genetic models, for the interpretation of molecular epidemiological data. We investigated 130 published data sets of *M. tuberculosis* isolates that were genotyped with IS*6110*-RFLP and spoligotyping methods, and identified common features and differences between the data sets. The pattern of decline in frequency against cluster size is remarkably similar across data sets from different regions, as well as across data sets typed with the two different markers. This decline is described well by the infinite alleles model (Fig. 4). By examining the full distributions of cluster sizes in a given data set, we observe that the number of singletons and the weight of large clusters are often distinguishing features of individual data sets. The application of population models shows that these two statistics are informative of the epidemic state of a population. In particular, the presence of many singletons and large clusters may signal a growing population. The result about the information carried by singletons may be counter-intuitive but can be understood as follows. For any sample, the IAM specifies an effective population size and an expected number of singletons. If the population is growing, its effective population size is lower than the actual population size near the time of sampling. Thus, since many new genotypes will have been created in recent history, the observed number of singletons is likely to be greater than the expected number under the IAM.

We have raised the possibility of using tests of the infinite alleles model to discern whether an infectious population is expanding. When the IAM is rejected, there are multiple plausible explanations. For instance, rejection is often taken to indicate the presence of natural selection, by which some strains are favored over others. We cannot discount this possibility, and indeed a method has recently been proposed for detecting inter-strain differences in transmission using spoligotype data (Tanaka and Francis, 2006). It may be tempting to automatically attribute differences in cluster size in a given data set to differences in transmissibility of strains. However, the acceptance of the IAM in some data (11 out of 64 in this study) demonstrate that there is often no support for such an untested inference. Even when the IAM is rejected, population growth rather than selection may be at work. As we have shown in Section 2.4, a model with an expanding infectious population is able to explain the San Francisco data without requiring selection. That is, the range of observed cluster sizes could be the result of the interplay between mutation and drift in a population. Hence, a large cluster may simply be due to having arisen early and persisted to the present time. We conclude that the three features together: the failure of the IAM, a large number of singletons and the presence of large cluster sizes suggest – but do not imply – an expanding population.

The extent of recent tuberculosis transmission is often quantified using closely related statistics measuring the proportion of clustered cases (Alland et al., 1994; Small et al., 1994; also known as the recent transmission indices or RTIs). Although we do not intend our singleton and large cluster statistics to replace the RTI, we remark on the relationship between these statistics. As noticed in Tanaka and Francis (2005), the RTI_n index can be expressed as $1 - \alpha(1)/n$ where $\alpha(1)$ is the number of singletons (unique genotypes) and n is the sample size. Therefore, given a fixed sample size n, the ordering of the data sets according to the RTI_n is the reverse of the ordering based on the number of singletons $\alpha(1)$. This reverse ordering appears to contradict the notion that both statistics indicate the extent of transmission. However, the number of singletons needs to be considered in the context of the IAM and the sample size n. One way to use the number of singletons for the purposes of quantifying transmission would be to standardize it using the expectation and variance of $\alpha(1)$ under the IAM. This may be an interesting area to explore further, although summary statistics may be insufficient to measure the degree of transmission with accuracy (Tanaka et al., 2006b). Ultimately, statistical approaches based on explicit population models are likely to be more informative about the transmission dynamics underlying a particular data set (Sisson et al., 2007).

We remark that geographic regions with high rates of tuberculosis transmission should be those with low diversity. Quantitatively, this implies that $\hat{\theta}$ should be negatively correlated with the incidence of tuberculosis — an implication confirmed using incidence data from (Dye et al., 1999; World Health Organization, 2005). In the case of IS*6110*, this

correlation is statistically significant (Kendall's $\tau = -0.245$; $p - \text{value} = 0.00206$; $\text{CI} = (-0.090, -0.399)$). We cannot explain the lack of strong negative correlation in the case of spoligotype data, but it may be due to the relatively low diversity of spoligotypes stemming from a lower underlying mutation rate. Thus, the inverse of $\hat{\theta}$ is itself another measure of the degree of transmission, although its effectiveness for this purpose is unknown.

A further outcome of our comparative study has been the estimation of the relative mutation rate of spoligotype patterns compared to IS*6110*. Note that here, ''mutation'' refers to the within-host substitution of genotypes (the combination of both mutation and fixation within the host). The arguments of Section 3 yield an estimate of the ratio to be around 10% to 26%. Taking a mutation rate associated with IS*6110* of 0.2 per case per year (Tanaka et al., 2006a, 2004), the mutation rate of spoligotypes is approximately 0.020 to 0.052 per case per year. Because the mutation rates of markers are a crucial factor in determining the configuration of cluster sizes in a population (and sample), it is essential to have some knowledge of these rates. In particular, estimates of mutation rates can be used in models for simulating molecular epidemiological data.

A summary of the epidemiological implications of our analysis is as follows.

1. Our study represents the application of tools from population genetics to molecular epidemiology data, in a way not previously done. This provides a different perspective, enabling the potential development of new methods.
2. We outline an approach to the analysis of molecular data obtained using spoligotypes or IS*6110* applied to tuberculosis data:

   Molecular epidemiological data can be analyzed by testing the IAM.

   (a) If the IAM is not rejected, then the infectious population is probably not expanding.

   (b) If the IAM is rejected, a likely cause is population expansion. Alternatively, there could be selective differences among strains.

      i. To measure the rate of population growth, one can estimate growth parameters explicitly through statistical methods, such as computational Bayesian analysis (Tanaka et al., 2006a).

      ii. Testing for selection is a more delicate problem that cannot be easily resolved. However, in the case of data obtained using spoligotyping, differences between strains can be detected using the method of Tanaka and Francis (2006). Such differences may indicate selection. It would be helpful to develop methods to differentiate between selection and population growth using molecular epidemiological data of this kind.

Acknowledgments

We thank Josephine Reyes for helpful discussions and comments on the manuscript. This work was supported by the Australian Research Council through the Discovery scheme (DP0556732) and by the Faculty of Science, University of New South Wales.

Appendix A. Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.meegid.2007.12.004.

References

Alland, D., Kalkut, G.E., Moss, A.R., McAdam, R.A., Hahn, J.A., Bosworth, W., Drucker, E., Bloom, B.R., 1994. Transmission of tuberculosis in New York City. An analysis by DNA fingerprinting and conventional epidemiologic methods. N. Engl. J. Med. 330, 1710–1716.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Statist. Soc. B 57, 289–300.

Chan-Yeung, M., Tam, C.M., Wong, H., Leung, C.C., Wang, J., Yew, W.W., Lam, C.W., Kam, K.M., 2003. Molecular and conventional epidemiology of tuberculosis in Hong Kong: a population-based prospective study. J. Clin. Microbiol. 41, 2706–2708.

Driver, C.R., Macaraig, M., McElroy, P.D., Clark, C., Munsiff, S.S., Kreiswirth, B., Driscoll, J., Zhao, B., 2006. Which patients' factors predict the rate of growth of *Mycobacterium tuberculosis* clusters in an urban community? Am. J. Epidemiol. 164, 21–31.

Dye, C., Scheele, S., Dolin, P., Pathania, V., Raviglione, M.C., 1999. Consensus statement. Global burden of tuberculosis: estimated incidence, prevalence, and mortality by country. WHO Global Surveillance and Monitoring Project. JAMA 282, 677–686.

Eisenach, K.D., Crawford, J.T., Bates, J.H., 1988. Repetitive DNA sequences as probes for *Mycobacterium tuberculosis*. J. Clin. Microbiol. 26, 2240–2245.

Ewens, W.J., 2004. Mathematical Population Genetics, second ed., vol. 1., Springer.

Farnia, P., Masjedi, M.R., Mirsaeidi, M., Mohammadi, F., Jallaledin-Ghanavi, Vincent, V., Bahadori, M., Velayati, A.A., 2006. Prevalence of Haarlem I and Beijing types of *Mycobacterium tuberculosis* strains in Iranian and Afghan MDR-TB patients. J. Infect. 53, 331–336.

Hubbell, S., 2001. The Unified Neutral Theory of Biodiversity and Biogeography. Princeton University Press, Princeton and Oxford.

Kamerbeek, J., Schouls, L., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., Bunschoten, A., Molhuizen, H., Shaw, R., Goyal, M., Van Embden, J., 1997. Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. J. Clin. Microbiol. 35, 907–914.

Kimura, M., 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kimura, M., Crow, J.F., 1964. The number of alleles that can be maintained in a finite population. Genetics 49, 725–738.

Kodmon, C., Niemann, S., Lukacs, J., Sor, E., David, S., Somoskovi, A., 2006. Molecular epidemiology of drug-resistant tuberculosis in Hungary. J. Clin. Microbiol. 44, 4258–4261.

Kremer, K., Van Soolingen, D., Frothingham, R., Haas, W.H., Hermans, P.W., Martin, C., Palittapongarnpim, P., Plikaytis, B.B., Riley, L.W., Yakrus, M.A., Musser, J.M., Van Embden, J.D., 1999. Comparison of methods based on different molecular epidemiological markers for typing of *Mycobacterium tuberculosis* complex strains: interlaboratory study of discriminatory power and reproducibility. J. Clin. Microbiol. 37, 2607–2618.

Lari, N., Rindi, L., Sola, C., Bonanni, D., Rastogi, N., Tortoli, E., Garzelli, C., 2005. Genetic diversity, determined on the basis of katG463 and gyrA95 polymorphisms, spoligotyping, and IS6110 typing, of *Mycobacterium tuberculosis* complex isolates from Italy. J. Clin. Microbiol. 43, 1617–1624.

Mathema, B., Kurepina, N.E., Bifani, P.J., Kreiswirth, B.N., 2006. Molecular epidemiology of tuberculosis: current insights. Clin. Microbiol. Rev. 19, 658–685.

Nguyen, L.N., Gilbert, G.L., Marks, G.B., 2004. Molecular epidemiology of tuberculosis and recent developments in understanding the epidemiology of tuberculosis. Respirology 9, 313–319.

Niemann, S., Richter, E., Rusch-Gerdes, S., 1999. Stability of *Mycobacterium tuberculosis* IS6110 restriction fragment length polymorphism patterns and spoligotypes determined by analyzing serial isolates from patients with drug-resistant tuberculosis. J. Clin. Microbiol. 37, 409–412.

Quitugua, T.N., Seaworth, B.J., Weis, S.E., Taylor, J.P., Gillette, J.S., Rosas, I.I., Jost, K.C., Magee, D.M., Cox Jr., R.A., 2002. Transmission of drug-resistant tuberculosis in Texas and Mexico. J. Clin. Microbiol. 40, 2716–2724.

Sajduda, A., Brzostek, A., Poplawska, M., Rastogi, N., Sola, C., Augustynowicz-Kopec, E., Zwolska, Z., Dziadek, J., Portaels, F., 2004. Molecular epidemiology of drug-resistant *Mycobacterium tuberculosis* strains isolated from patients with pulmonary tuberculosis in Poland: a 1-year study. Int. J. Tuberc. Lung. Dis. 8, 1448–1457.

Sisson, S.A., Fan, Y., Tanaka, M.M., 2007. Sequential Monte Carlo without likelihoods. Proc. Natl. Acad. Sci. U.S.A. 104, 1760–1765.

Slatkin, M., 1996a. A correction to the exact test based on the Ewens sampling distribution. Genet. Res. 68, 259–260.

Slatkin, M., 1996b. Ewens exact program. Available from: URL: http://ib.berkeley.edu/labs/slatkin/software.html; accessed 22-December-2006.

Small, P.M., Hopewell, P.C., Singh, S.P., Paz, A., Parsonnet, J., Ruston, D.C., Schecter, G.F., Daley, C.L., Schoolnik, G.K., 1994. The epidemiology of tuberculosis in San Francisco: a population-based study using conventional and molecular methods. N. Engl. J. Med. 330, 1703–1709.

Suresh, N., Singh, U.B., Arora, J., Pant, H., Seth, P., Sola, C., Rastogi, N., Samantaray, J.C., Pande, J.N., 2006. rpob gene sequencing and spoligotyping of multidrug-resistant *Mycobacterium tuberculosis* isolates from India. Infect. Genet. Evol. 6, 474–483.

Tanaka, M.M., Francis, A.R., 2005. Methods of quantifying and visualising outbreaks of tuberculosis using genotypic information. Infect. Genet. Evol. 5 (1), 35–43.

Tanaka, M.M., Francis, A.R., 2006. Detecting emerging strains of tuberculosis by using spoligotypes. Proc. Natl. Acad. Sci. U.S.A. 103, 15266–15271.

Tanaka, M.M., Francis, A.R., Luciani, F., Sisson, S.A., 2006a. Using approximate Bayesian computation to estimate tuberculosis transmission parameters from genotype data. Genetics 173, 1511–1520.

Tanaka, M.M., Phong, R., Francis, A.R., 2006b. An evaluation of indices for quantifying tuberculosis transmission using genotypes of pathogen isolates. BMC Infect. Dis. 6, 92.

Tanaka, M.M., Rosenberg, N.A., Small, P.M., 2004. The control of copy number of IS*6110* in *Mycobacterium tuberculosis*. Mol. Biol. Evol. 21 (12), 2195–2201.

Van Soolingen, D., 2001. Molecular epidemiology of tuberculosis and other mycobacterial infections: main methodologies and achievements. J. Intern. Med. 249, 1–26.

Van Soolingen, D., Borgdorff, M.W., De Haas, P.E., Sebek, M.M., Veen, J., Dessens, M., Kremer, K., Van Embden, J.D., 1999. Molecular epidemiology of tuberculosis in the Netherlands: a nationwide study from 1993 through 1997. J. Infect. Dis. 180, 726–736.

Watterson, G.A., 1974. Models for the logarithmic species abundance distributions. Theor. Popul. Biol. 6, 217–250.

World Health Organization, 2005. Global tuberculosis control: surveillance, planning, financing. WHO report 2006. Tech. Rep. (WHO/HTM/TB/2006.362)., Geneva, available from: URL: http://www.who.int/tb/publications/2006/en/index.html.