Detecting emerging strains of tuberculosis by using spoligotypes

Mark M. Tanaka* **and Andrew R. Francis**[†‡]

*School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney NSW 2052, Australia; and [†]School of Computing and Mathematics, University of Western Sydney, Sydney NSW 1797, Australia

Edited by Burton H. Singer, Princeton University, Princeton, NJ, and approved August 21, 2006 (received for review April 18, 2006)

The W-Beijing strain of tuberculosis has been identified in many molecular epidemiological studies as being particularly prevalent. This identification has been made possible through the development of a number of genotyping technologies including spoligotyping. Highly prevalent genotypes associated with outbreaks, such as the W-Beijing strain, are implicitly regarded as fast spreading. Here we present a quantitative method to identify ''emerging'' strains, those that are spreading faster than the background rate inferred from spoligotype data. The approach uses information about the mutation process specific to spoligotypes, combined with a model of both transmission and mutation. The core principle is that if two comparable strains have the same number of isolates, then the strain with fewer inferred mutation events must have spread faster if the mutation process is common. Applying this method to four different data sets, we find not only the W-Beijing strain, but also a number of other strains, to be emerging in this sense. Importantly, the strains that are identified as emerging are not simply those with the largest number of cases. The use of this method should facilitate the targeting of individual genotypes in intervention programs.

mutation | transmission rate | Beijing strain | infectious disease | molecular marker

A broad goal of the development of effective tools for genotyping the bacteria or viruses causing infectious diseases has been the classification of isolates into distinct types. In the case of *Mycobacterium tuberculosis*, one of the outcomes of this development has been the identification of a particularly aggressive strain known as the Beijing or W-strain (1, 2). These genotypic data, however, can also be used to verify chains of transmission and to make inferences about population level transmission patterns. For example, the occurrence of large clusters of identical genotypes in a sample is thought to be indicative of recent tuberculosis transmission (3, 4), and in this context, the size of a genotype cluster carries some information about the rate of transmission associated with the genotype. One use of such information is to study possible risk factors for infection, such as HIV status, by correlating them with the extent to which these data form clusters (3, 4).

An unusually large cluster in a sample may indicate a rapidly spreading strain; however, it may simply indicate the age of the genotype (5). For instance, strains that have been present in a population for a long time may have accumulated a large number of cases despite having a slow transmission rate. One way to access information about the age of a strain is to consider the number of mutation events identified in the history of that strain. That is, an old genotype has had ample time to generate many mutants, which should be manifested in the sample. We observe that on average, if two strains have the same cluster size and the same mutation rate, then the one with more observed mutation events is older, and correspondingly, because the clusters are of the same size, the one with fewer observed mutation events has spread faster.

The features of data obtained from spoligotyping (spacer oligonucleotide typing) technology are key to making our analysis possible. Spoligotyping is a reliable and informative technology for characterizing the genetic structure of tuberculosis populations (6). Spoligotype patterns are produced by hybridization of sample DNA to oligonucleotides based on well characterized specific sequences at the direct repeat locus (7). Each genotype is represented by a binary string of length 43. Variation at this locus results from deletions of adjacent blocks of repetitive units (8). (Examples of spoligotype patterns and their mutational relationships are given in Fig. 1). For our purposes, spoligotyping has the following advantages: first, it allows isolates to be placed into unambiguous genotypic classes; second, genotypes related by a single mutation (deletion) event can be identified; and third, the direction of the mutation event can be determined. These advantages permit, for each genotype, the identification of other genotypes in the sample that could have arisen from it by a single deletion event. The number of such descendent genotypes is the inferred number of mutations from the parent genotype.

Defining a "strain" to be a set of organisms with the same spoligotype, we designate as "emerging" a strain that is spreading significantly faster than the background transmission rate. In this article, we propose a method of detecting emerging strains by invoking the principles outlined above, by constructing a model of disease transmission and marker mutation, and by using false discovery rate analysis to correct for multiple statistical tests. Our approach is an advance on examining cluster sizes, which, to our knowledge, is the only available basis for comparing the transmissibility associated with different genotypes. We then apply our method to four published data sets of spoligotypes from *M. tuberculosis* and *Mycobacterium bovis* isolates.

Overview of the Model and Methods

The underlying model of transmission is that the growth of the number of cases is exponential and deterministic. The mutation model has such features as: (*i*) the mutation rate of a given genotype is proportional to the number of spacer units present in the direct repeat region, and (*ii*) each mutation event gives rise to a new genotype, the infinite alleles (IA) assumption. We assume that a constant proportion of the infectious population is sampled across all of the strains. This model is used to obtain an estimate of a parameter related to the ratio of mutation rate to transmission rate, based on the entire data set. We then test the hypothesis that this parameter computed for a given strain is the same as that for the data set. Testing is done by computing the probability of observing the same number of deletion events from the associated genotype as actually observed or fewer (the

Author contributions: M.M.T. and A.R.F. designed research, performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviation: IA, infinite allele(s).

[‡] To whom correspondence should be addressed at: School of Computing and Mathematics, University of Western Sydney, Locked Bag 1797, Penrith South DC NSW 1797, Australia. E-mail: a.francis@uws.edu.au.

www.pnas.org/cgi/doi/10.1073/pnas.0603130103

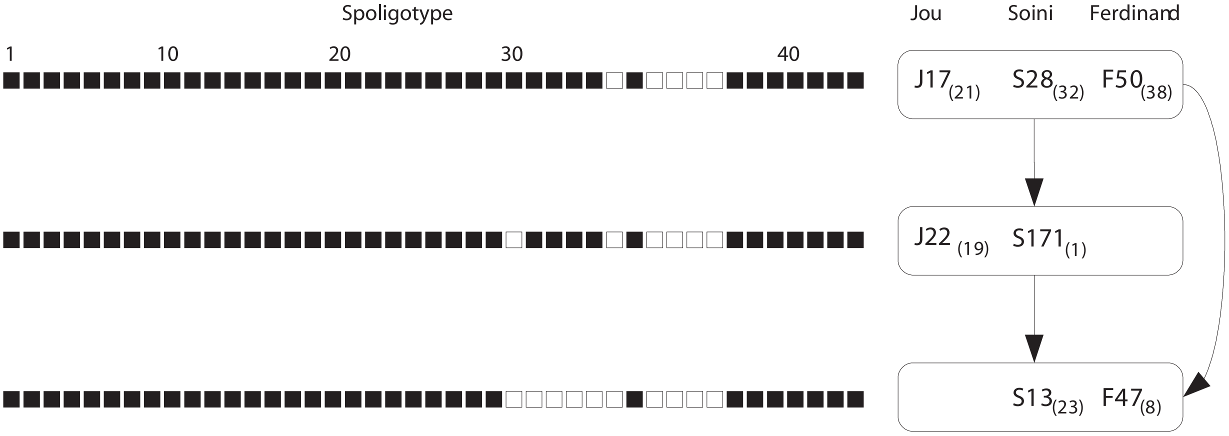

Fig. 1. Spoligotypes related to strain S171/J22 in the *M. tuberculosis* data sets of Soini *et al.* (S13, S28, and S171) (9), Jou *et al.* (J17and J22) (10), and Ferdinand *et al.* (F47 and F50) (11). Each filled square represents the presence of one of the 43 spacers at the direct repeat locus. The numbers in parentheses refer to the sizes of clusters in the relevant data sets. Arrows indicate possible deletion events such that derived genotypes are at the head and parental genotypes are at the tail of each arrow. Note that relationships with genotypes other than these three are not shown here (some are shown in Fig. 3). In this set of related strains, only J22 was identified as emerging.

p value). The *q* value of a strain, derived from the *p* value and preserving the same ordering on the strains, represents the minimum false discovery rate across multiple hypothesis tests for the strain to be regarded as significant. More details of the model and methods are provided in *Methods*.

Application of the Model to Genotype Data

We analyzed three data sets of *M. tuberculosis* spoligotypes and one of *M. bovis* spoligotypes: Soini *et al.* (9), Jou *et al.* (10), Ferdinand *et al.* (11), and Aranaz *et al.* (12). We will refer to these data sets by the name of the first author. The studies cover a variety of regions: Texas, Taiwan, Madagascar, and Spain, respectively. Soini is a very large data set (over 1,200 isolates); Jou and Ferdinand are recent data from geographically distant populations; and Aranaz is an interesting contrast concerning a different bacterial species (*M. bovis*) in different hosts (including cattle and goats).

For each genotype in each data set we obtained a *q* value, and those whose *q* values are <0.8 are reported in Table 1. This value was chosen to capture as many plausibly emerging strains as possible. To test the stability of the results under differing assumptions about the completeness of sampling, we used a wide range of values for the sampling fraction f. Across the board, changes in the assumed sampling fraction had little effect on the outcomes of the study in terms of which strains were detected. Minor changes in ordering were observed, and the number of strains detected dropped with the sampling fraction.

The Texas data set (Soini) is the largest data set we analyzed, with 1,283 isolates consisting of 225 genotypes. The analysis in Table 1 reveals five strains that exhibit an elevated transmission rate. Notably, although the W-Beijing strain S1 is one of these strains, other strains (S12, S3, S24, and S214) are also found to be growing rapidly. It is remarkable that all these strains have *q* values <0.04, regardless of sampling fraction, and that none of the other 220 strains in the data have *q* values <0.8. The data set from Taiwan (Jou), with 421 isolates and 113 genotypes, has fewer emerging strains. However, again we see that the W-Beijing strain J9 is spreading faster than the background rate (genotypes from the Jou data set are labeled in this article by a J, together with the order in which they appear in figure 1 of ref.

Table 1. Emerging strains detected in the analysis

	f							
	0.95		0.5		0.1		0.01	
Data set	Strain	*q* value	Strain	*q* value	Strain	*q* value	Strain	*q* value
Soini	S12	0.00001	S12	0.00012	S1	0.00040	S1	0.00043
	S1	0.00005	S1	0.00014	S12	0.00040	S12	0.00058
	S3	0.00045	S24	0.00257	S24	0.00734	S24	0.00853
	S24	0.00045	S3	0.00401	S3	0.01584	S3	0.02272
	S214	0.00669	S214	0.01638	S214	0.03227	S214	0.03782
Jou	J9 (S1)	0.03428	J9	0.11067	J9	0.26617	J9	0.35067
	J20	0.04641	J20	0.33131				
	J22 (S171)	0.04641	J22	0.33131				
	J64 (S7)	0.78511						
Ferdinand	F109	0.55080						
	F86	0.55080						
	F21	0.72113						

The genotypes under each data set heading are ordered by *q* value. Where genotypes in the table from the Jou or Ferdinand data sets also appear in the Soini data set, the corresponding labels from Soini are given in parentheses. Genotypes from the Jou data set are labeled in this paper by a J, together with the order in which they appear in figure 1 of ref. 10.

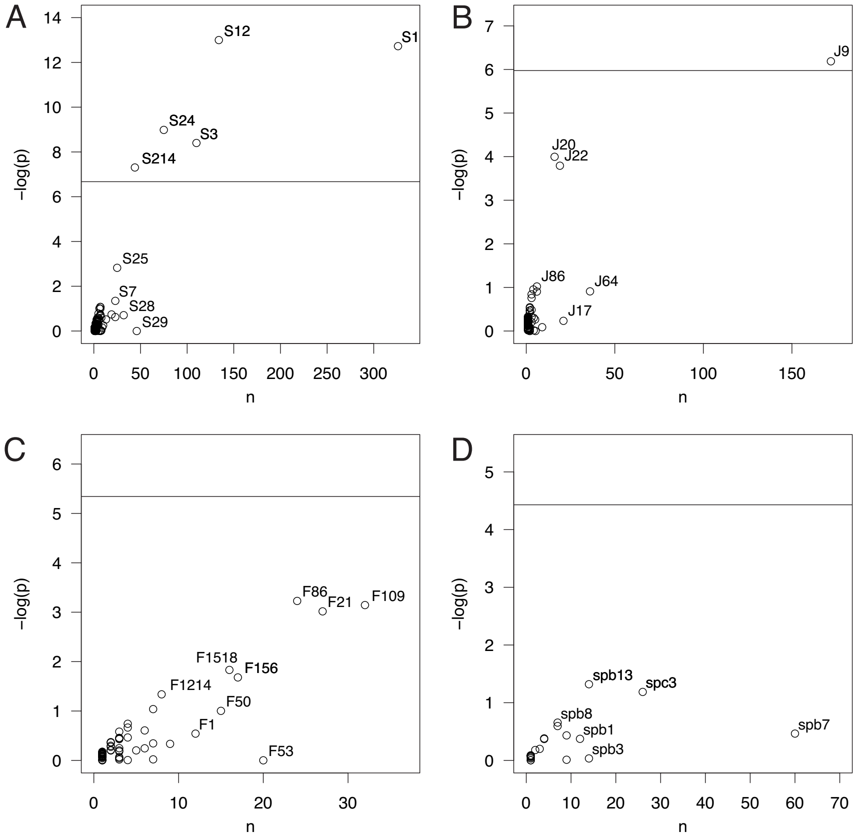

Fig. 2. Negative log p values plotted against cluster size n. The four data sets used are Soini (*A*), Jou (*B*), Ferdinand (*C*), and Aranaz (*D*). Note that the W-Beijing strain appears here in *A*, *B*, and *C* as S1, J9, and F1, respectively. The value used for sampling fraction was $f = 0.1$. The horizontal line represents the threshold under the Dunn–Sidak correction for multiple testing (each strain has been tested individually), with a permissive significance level of $\alpha = 0.25$, allowing a comparison with the q value results (see Table 1).

10. As with the Soini data, it is noteworthy that this is not the only such strain. Strains J22, J20, and J64 also appear (although J64 occurs only when $f = 0.95$, and with a high q value). The data set from Madagascar (Ferdinand), with 301 isolates and 92 genotypes, reveals only strains with q value <0.8 when the sample is assumed to be almost complete, and these strains all have a high probability of being false discoveries. Finally, the *M. bovis* data from Spain (Aranaz), with 182 isolates consisting of 24 genotypes, do not exhibit any emerging strains. This outcome may be because of barriers to transmission caused by farming or species-related subdivisions of the host population.

Discussion

The present study aims to identify strains within a given outbreak that are spreading considerably faster than the background of that particular data set. The key step to enabling this identification is the incorporation of the mutation process, which adds information about the transmission rate beyond that provided by the cluster size. The advantage of using this additional information can be seen by plotting the negative log p values (or negative log q values) against cluster sizes. Fig. 2 shows this relationship using p values (because they are more clearly separated than the q values). Although a correlation between p values and cluster sizes is apparent, there are clear exceptions. For instance, strain J64 in the Jou data has 36 isolates, but is ranked below J22 and J20, which have cluster sizes 19 and 16, respectively. In general, large cluster size could be attributable to the age of the strain rather than rapid transmission. In the case of the W-Beijing strain, which is known to be evolutionarily old (9, 13), our analysis suggests that age is not the only factor that explains the prevalence of this strain. That is, the W-Beijing strain has the largest cluster in both the Soini and Jou data sets, and it is also associated with very low p values and q values. Although the W-Beijing strain is present in the Ferdinand data (F1), it is not detected as an emerging strain in this case, which could be either because the conditions (biological or otherwise) that make it an emerging strain in some parts of the world are not present in Madagascar, or because it only recently entered Madagascar and has not yet generated a large number of cases.

The goal of identifying epidemiologically important strains has been addressed for other diseases. For example, research into the influenza virus has been able to identify codons in the hemagglutinin HA1 gene that predict the evolutionary success of a strain by using genetic data without detailed epidemiological or immunological information (14, 15). Broadly, the question addressed in this article of which strains will emerge and the information used (genetic data alone) are similar to those of the influenza studies. However, the two pathogens and the markers used to identify them differ. Human influenza is commonly typed by using sequences of a particular gene, whereas *M. tuberculosis* is generally typed by using markers that produce allelic data, giving "clusters" of identical genotypes. They also differ in the process and rate of mutation. Instead of predicting the evolution of the pathogen, as in studies of influenza (14, 15), our method identifies tuberculosis strains that are currently spreading significantly faster than the others.

In theory, our method can be applied to other pathogens having appropriate properties. The pathogen must be clonal (or the genetic marker must not undergo recombination among different strains), and it must be possible to infer the direction of evolution of the marker locus. The mutation rate of the marker must be high enough to generate diversity in the course of an epidemic, but low enough to allow clusters of the same type to be observed. In the case of *M. tuberculosis*, spoligotypes are well suited to the application of this method. Although most pathogens do not satisfy the above conditions, it may be possible to use the principles presented here to develop similar methods to study other diseases.

This study represents an attempt to construct a method of identifying emerging strains of tuberculosis. Improvements should become possible as understanding of the relevant biological processes develops. For instance, if more empirical information about mutation at the spoligotype locus is generated, the mutation model could be appropriately refined. We have modeled mutation rate as a simple linear function of copy number. However, a different function, possibly nonlinear or involving other parameters, may be more appropriate. Future efforts might also focus on creating a more detailed model of sampling. Addressing this challenge would probably require the development of an alternative statistical approach.

The transmission model used in this study also has its limitations. First, the exponential growth model does not strictly hold in a population where the disease has become endemic. It does, however, model the early stages of an epidemic and the growth of those strains that have recently emerged. When the growth dynamics are slower than exponential, the exponential model is conservative in that it underestimates the number of cases in the history of each strain. Thus, this procedure underestimates the expected number of mutations, making it harder for the observed number of mutations from a given genotype to be significantly low. This effect is desirable because any false discoveries produced by the procedure are then less likely to be because of inaccuracies of the growth model. Second, it may be possible to relax the deterministic assumptions of the transmission model. Whereas a stochastic model such as a birth–death process would be more realistic, it has the problems of requiring appropriate parameterization and estimation procedures. If these problems are overcome, the result should be a more conservative procedure. This conservatism is because the stochasticity would lead to variance in the number of cases in the history of a strain, which in turn increases the variance of D (modeling the outdegree). Consequently, it would be harder to reject the null hypothesis.

An inherent property of our approach to detecting emerging strains is that any determination of emerging strains based on the q values is specific to the context of the particular data set. The q values for genotypes in different data sets do have the same meaning in that they give the probability that a genotype represents a false discovery. However, this approach does not allow a comparison of the actual rates of transmission across different data sets. For instance, a nonemerging strain in a data set of rapidly spreading strains may have a higher transmission rate than one identified as emerging among slowly spreading strains. In other words, the measure of emergence is not absolute, but relative to the data set. With this in mind, it may be informative to examine the behavior of a strain of interest in a variety of data sets, as illustrated in Fig. 1.

In terms of the applications of the method to the published data, the reasons why strains differ with respect to transmission rates are not straightforward. Although the differences may be due to genetic variation among strains, other possibilities should be considered. Any factors correlated with an emerging strain (bacterial or host genes, environmental factors, or a combination of such factors) could be responsible for the difference. That is, the association of a given genotype with rapid spread could be because of biological properties of the strain, or, for example, the configuration of social networks underlying the transmission of this strain. The consistent appearance of a particular strain as emerging in different populations would diminish the force of non-pathogen-related explanations. Indeed, multiple drug resistance has been associated with the W-Beijing strain and may be a pathogen-related factor contributing to the rapid spread of this strain in more than one region (Fig. 2). See ref. 2 for discussion on the possible factors contributing to the prevalence of the W-Beijing strain. Regardless of the explanations for emergence, the identification of genotypes associated with fast transmission may be useful in directing strategies for control.

Methods

The Relationship Between Copy Number and Mutation Rate. The analysis requires a set of spoligotypes of mycobacterial isolates, together with information about the mutation process of the marker. Spoligotypes undergo mutation through deletion events that remove sequences corresponding to contiguous blocks in the spoligotype pattern (8). We will assume that the overall mutation rate for spoligotypes is proportional to copy number, where "copy" is used to mean the number of unique spacer sequences present in the spoligotype. That is, we assume the mutation rate is $c\mu$, where c is the copy number at the direct repeat locus and μ is the underlying rate per copy. The intuition here is that a genotype with a high copy number has more opportunity for deletion events to occur compared with a genotype with low copy number.

To justify the assumption that the mutation rate of a genotype is proportional to its copy number, we examined the possible association between copy number and outdegree of each cluster, averaged over a large sample (1,000) of IA forests (described below). For this purpose we examined the data of Soini *et al.* (9), the largest of the data sets used in this study (see *Application of the Model to Genotype Data*). Applying the Kendall tau test, a nonparametric correlation analysis, the tau value was 0.428, which was highly significant, with a p value of $<10^{-15}$.

The Transmission and Mutation Model. We assume that the number of cases of the bacterial infection increases exponentially and deterministically in a population. Mutation events at the marker locus are assumed to occur stochastically because mutation events are rare relative to transmission events. In the following, we will consider the growth dynamics of a single strain. The data from an outbreak represents a sample from the total population; let f be the proportion of the infectious population sampled. We assume that the sampling proportion is the same across all strains. Let ρ be the transmission rate per infectious case of the strain in question (that is, the number of new infectious cases per individual infected with this strain per unit time).

Noting that the mutation process removes cases from a strain, the number of individuals in a single strain as a function of time from its origination is given by $e^{(\rho-c\mu)t}$. This model assumes that each new mutation event produces a new genotype, and so there is no entry into any established strain through mutation (the IA assumption).

Let H be the total number of cases in the history of a particular strain from the time of its origin (including those in the present population that have not been sampled). Let T be the time since the strain arose through mutation. Let N be the number of cases of the strain in the sample. Then

$$H = \int_0^T e^{(\rho-c\mu)t} dt = \frac{(N-f)}{f(\rho-c\mu)}, \qquad [1]$$

because $N = fe^{(\rho-c\mu)T}$.

For a given strain whose genotype has copy number c, let D be the number of mutation events experienced by the strain in its history (i.e., from time 0 to T). Assume $D \sim \text{Pois}(\lambda)$, where

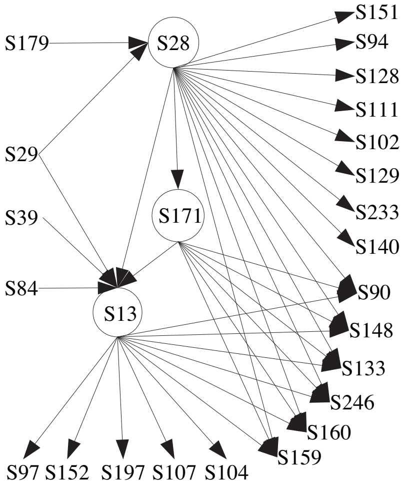

Fig. 3. An example of part of the cluster graph representing the data in ref. 9. The figure shows all edges to and from the three circled vertices (genotypes) S28, S171, and S13 but omits other edges and vertices not connected to these three clusters. The spoligotypes of S28, S171, and S13 are those shown in Fig. 1.

$$\lambda = c\mu H = \frac{c(N-f)}{f(\zeta - c)}, \qquad [2]$$

with $\zeta = \rho/\mu$, the compound parameter describing the rate of transmission relative to mutation. Because λ depends on c and μ, different genotypes may experience mutation events at different rates. The random variable D models the number of other genotypes in a sample to which the genotype in question could have evolved through a single deletion event. Note that this model assumes that sampling has no effect on the observed degree associated with a genotype because we assume a "deterministic" sampling scheme whereby a fixed fraction (f) of cases are sampled.

Cluster Graphs of the Data and IA Forests. Now consider a data set consisting of g genotypes. For each genotype i, let n_i be its cluster size (consider this to be an instance of the above described N, and let its copy number be c_i. We need to obtain the number of other genotypes in the history of the epidemic that i mutated to through a single mutation event. First, define the cluster graph of the spoligotype data to be the graph whose vertices are the genotypes in the sample and whose directed edges represent possible single-step mutation events (which are inferred from the pattern of deletions among the spoligotypes) (16). In Fig. 3, we show an example of part of a cluster graph using data from ref. 9. Under the IA assumption, a given genotype can have, at most, a single inbound edge. Therefore, it is necessary to consider collections of trees (forests) consisting of an appropriate subset of edges of the cluster graph. We call such forests, whose vertices are the genotypes and which conform with the IA assumption, IA forests. There can be many IA forests consistent with the data.

For a given IA forest F, let d_i be the number of other genotypes in F that are connected to i through a single mutation event (the outbound degree of genotype i). The quantity d_i is taken to be an observation of the random variable D described above, whose parameter for a particular genotype i is

$$\lambda_i = \frac{c_i(n_i - f)}{f(\zeta - c_i)}. \qquad [3]$$

The *p* Values and *q* Values of Strains. The next step is to find the maximum likelihood estimate of the compound parameter $\zeta = \rho/\mu$ under the model and for a given IA forest and value of f. The likelihood is given by

$$\mathrm{Lik}(\zeta) = \prod_{i=1}^{g} e^{-\lambda_i} \frac{\lambda_i^{d_i}}{d_i!}. \qquad [4]$$

Note that whereas n_i, c_i, and d_i are obtained from the data, the sampling proportion f is fixed and not estimated. This is because any prior information about f should be used to set its value. Not estimating f also allows alternative assumptions about f to be examined. The maximum likelihood estimate $\hat{\zeta}$ of ζ is obtained by numerically maximizing the logarithm of the likelihood function with respect to ζ. Because $\mathrm{Lik}(\zeta)$ is a product over all strains, $\hat{\zeta}$ is an estimate of ζ based on all strains in the data.

For a given genotype, we can now compute the probabilities of observing at most d_i for D . These probabilities are the p values arising from testing the null hypothesis that the observed d_i is consistent with the model given $\hat{\zeta}$. The alternative hypothesis is that the given strain has a higher ζ than the $\hat{\zeta}$ derived from the whole data set. Then, for a given forest F,

$$P(D_i \le d_i|F) = \sum_{j=0}^{d_i} \left(\frac{c_i(n_i - f)}{f(\hat{\zeta} - c_i)} \right)^j \frac{1}{j!} \exp\left(-\frac{c_i(n_i - f)}{f(\hat{\zeta} - c_i)} \right), \qquad [5]$$

and, summing over the set of all possible forests,

$$P(D_i \le d_i) = \sum_F P(D_i \le d_i|F)P(F). \qquad [6]$$

There are often many genotypes in a cluster graph with multiple inbound edges, resulting in a large number of possible IA forests. For instance, if the cluster graph contains 30 genotypes that each have two inbound edges, then there would be over one billion possible IA forests. In fact, it is common for genotypes to have more than two inbound edges. In practice, we sample random IA forests from the set of possible IA forests by randomly selecting a single inbound edge for each genotype with multiple inbound edges. Assuming that each IA forest is equally probable, we approximate the above p value by averaging $P(D_i \le d_i|F)$ across a large random sample of IA forests (1,000 for each data set).

This procedure gives rise to g tests that must be corrected for multiple testing. Because we are interested in identifying potentially important strains, our strategy is to control the false discovery rate, which is the proportion of significant tests that are falsely rejected (17, 18), instead of a more traditional approach controlling the familywise error rate. Using the methods of Storey (18, 19) and the QVALUE software package of Dabney and Storey (http://faculty.washington.edu/jstorey/qvalue), we estimate the q value for each strain. The q value of a strain is the positive false discovery rate for the set of hypothesis tests when the significance level is set to be the p value of the corresponding strain. A relatively low p value (or q value) indicates that the strain has spread too fast for many mutation events to have occurred. Another interpretation of the q value is that it is the Bayesian posterior probability of making an error when regarding a strain to be significant (18).

We thank Michael Lachmann and Susan Ptak for their suggestions on the development of the models and Susan Ptak, Noah Rosenberg, and the reviewers for their comments on the manuscript. This work was supported in part by Discovery Grant DP0556732 from the Australian Research Council.

1. Kremer K, Glynn JR, Lillebaek T, Niemann S, Kurepina NE, Kreiswirth BN, Bifani PJ, van Soolingen D (2004) *J Clin Microbiol* 42:4040–4049.
2. Bifani PJ, Mathema B, Kurepina NE, Kreiswirth BN (2002) *Trends Microbiol* 10:45–52.
3. Small PM, Hopewell PC, Singh SP, Paz A, Parsonnet J, Ruston DC, Schecter GF, Daley CL, Schoolnik GK (1994) *N Engl J Med* 330:1703–1709.
4. Alland D, Kalkut G, Moss A, McAdam R, Hahn J, Bosworth W, Drucker E, Bloom B (1994) *N Engl J Med* 330:1710–1716.
5. Kimura M, Ohta T (1973) *Genetics* 75:199–212.
6. Kremer K, van Soolingen D, Frothingham R, Haas WH, Hermans PW, Martin C, Palittapongarnpim P, Plikaytis BB, Riley LW, Yakrus MA, *et al.* (1999) *J Clin Microbiol* 37:2607–2618.
7. 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) *J Clin Microbiol* 35:907–914.
8. Fang Z, Morrison N, Watt B, Doig C, Forbes KJ (1998) *J Bacteriol* 180:2102–2109.
9. Soini H, Pan X, Amin A, Graviss EA, Siddiqui A, Musser JM (2000) *J Clin Microbiol* 38:669–676.
10. Jou R, Chiang, CY, Huang WL (2005) *J Clin Microbiol* 43:95–100.
11. Ferdinand S, Sola C, Chanteau S, Ramarokoto H, Rasolonavalona T, Rasolofo-Razanamparany V, Rastogi N (2005) *Infect Genet Evol* 5:340–348.
12. Aranaz A, Liebana E, Mateos A, Dominguez L, Vidal D, Domingo M, Gonzolez O, Rodriguezferri EF, Bunschoten AE, van Embden JDA, Cousins D (1996) *J Clin Microbiol* 34:2734–2740.
13. Sreevatsan S, Pan X, Stockbauer KE, Connell ND, Kreiswirth BN, Whittam TS, Musser JM (1997) *Proc Natl Acad Sci USA* 94:9869–9874.
14. Bush RM, Bender CA, Subbarao K, Cox NJ, Fitch WM (1999) *Science* 286:1921–1925.
15. Plotkin JB, Dushoff J, Levin SA (2002) *Proc Natl Acad Sci USA* 99:6263–6268.
16. Tanaka MM, Francis AR (2005) *Infect Genet Evol* 5:35–43.
17. Benjamini Y, Hochberg Y (1995) *J R Stat Soc B* 57:289–300.
18. Storey JD (2003) *Ann Stat* 31:2013–2035.
19. Storey JD (2002) *J R Stat Soc B* 64:479–498.