Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 63 (2003) 347–363
Theoretical Population Biology
http://www.elsevier.com/locate/ytpbi

Estimating change rates of genetic markers using serial samples: applications to the transposon IS*6110* in *Mycobacterium tuberculosis*

Noah A. Rosenberg,[a,*] Anthony G. Tsolaki,[b] and Mark M. Tanaka[c]

[a] *Program in Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, 1042 W. 36th Place-DRB 289, Los Angeles, CA 90089, USA*
[b] *Stanford Center for Tuberculosis Research, Lab S143, Stanford University Medical Center, 300 Pasteur Drive, Stanford, CA 94305, USA*
[c] *School of Biotechnology and Biomolecular Sciences, University of New South Wales, Sydney, NSW 2052, Australia*

Received 25 May 2002

Abstract

In infectious disease epidemiology, it is useful to know how quickly genetic markers of pathogenic agents evolve while inside hosts. We propose a modular framework with which these genotype change rates can be estimated. The estimation scheme requires a model of the underlying process of genetic change, a detection scheme that filters this process into observable quantities, and a monitoring scheme that describes the timing of observations. We study a linear "birth–shift–death" model for change in transposable element genotypes, obtaining maximum-likelihood estimators for various detection and monitoring schemes. The method is applied to serial genotypes of the transposon IS*6110* in *Mycobacterium tuberculosis*. The estimated birth rate of 0.0161 (events per copy of the transposon per year) and death rate of 0.0108 are both significantly larger than the estimated shift rate of 0.0018. The sum of these estimates, which corresponds to a "half-life" of 2.4 years for a typical strain that has 10 copies of the element, substantially exceeds a previous estimate of 0.0135 *total* changes per copy per year. We consider experimental design issues that enable the precision of estimates to be improved. We also discuss extensions to other markers and implications for molecular epidemiology.

Keywords: Birth-and-death process; Insertion sequence; Mutation rate; Pathogen; Pulsed-field gel electrophoresis; Substitution rate; Transposition

1. Introduction

Mutation serves as the ultimate source of polymorphism, and thus, the rate at which it occurs affects all aspects of genetic variation. Rates of genetic change have been studied in many systems. Because genetic change processes occur on time scales that depend on organismal generation times, and because generation times vary greatly across organisms, diverse statistical techniques are used for estimating change rates. For each situation in which genetic change rates are desired, the statistical approach must compensate for the fact that the genetic history of a group of organisms cannot usually be known completely. In humans or in macroscopic laboratory organisms, it may be possible to restrict attention to a limited set of individuals, for which genetic history is known over one or a few generations. For example, mutation rates are sometimes obtained by counting the number of mutations that have occurred in simple pedigrees (Mukai and Cockerham, 1977; Heyer et al., 1997).

For genealogies that are too complex for direct counting, a historical model together with a snapshot of the genetic status of a population can be applied. This approach is perhaps best represented by the Luria–Delbrück method of bacterial mutation rate estimation, which utilizes the variation in mutant frequency across independent experimental runs of evolution (Luria and Delbrück, 1943; Kepler and Oprea, 2001). If multiple runs of evolution cannot be performed, the historical modeling approach can employ coalescent models in place of the Luria–Delbrück bifurcation scheme, obtaining rate estimates with data from the unique realization of evolutionary history (Rosenberg and Nordborg, 2002).

A third estimation method involves genotypic comparisons of individuals of two separate species, and normalization of the number of interspecific genotypic

*Corresponding author. Fax: +213-740-2437.
E-mail address: noahr@usc.edu (N.A. Rosenberg).

0040-5809/03/$ - see front matter © 2003 Elsevier Science (USA). All rights reserved.
doi:10.1016/S0040-5809(03)00010-8

differences using nucleotide evolution models, species divergence dates from the fossil record, and generation times (Kimura, 1983; Li, 1997). This strategy is most appropriate when mutation occurs over time scales that are large enough for within-species differences to be small compared to between-species differences.

Here we consider a qualitatively different approach for estimating rates of genetic change, suitable to pathogens in host–pathogen systems. In this scheme, serially sampled pathogens from many host individuals are genotyped. Genetic change occurs in each host's pathogen population during periods between genotyping events. Because the complete history of the pathogen population is not observed, a model of how this change occurs is required in order to estimate change rates.

Traditional methods are difficult to apply in this situation: first, the pathogen has a complex genealogy so that mutations cannot be counted directly. Second, because the rate of change of the pathogen *within hosts* is of interest, methods that require laboratory cultures are not suitable (moreover, culturing is difficult for some species). Last, because markers may change rapidly with respect to evolutionary time, the correlation of genotypes in different species may quickly disappear, rendering species comparison methods ineffective.

In any situation where populations of a species that have short generation times live in a host that has a long generation time, so that populations in different hosts are isolated from each other, mutation rates might be estimated from sequential genotype data. In principle, we might similarly discuss other systems that have this property. Here, we use the language of host–pathogen interactions, because a primary application is to bacterial and other pathogens in human and animal hosts. As is described in Sections 9 and 10, the results can be applied to molecular epidemiology, in which the spatiotemporal pattern of pathogen genotypes among a collection of patients is used to infer properties of an epidemic.

We have previously (Tanaka and Rosenberg, 2001) considered a special case, estimating genotype change rates for transposons using pathogens sampled from patients at two points in time. We developed a maximum-likelihood estimator for a one-parameter model, assuming that the rate at which a transposon genotype changes was proportional to the number of copies of the transposon initially present in the genome of the pathogen. With genotypes of paired *Mycobacterium tuberculosis* isolates taken from 56 patients, we estimated the change rate of the transposon IS*6110* to be 0.0135 changes per copy of the transposon per year (this estimate corrects a numerical error in the fourth decimal place), corresponding to a "half-life" of 5.1 years for a typical 10-copy strain.

In this article, we develop maximum-likelihood estimators of genetic change rates for a general class of scenarios in which pathogen genotypes are monitored over time. Serial samples have been previously used to estimate genetic parameters, for example, with DNA sequences from human immunodeficiency virus (Rodrigo and Felsenstein, 1999; Fu, 2001; Drummond et al., 2002) and from ancient penguin bones (Lambert et al., 2002). Our work differs in that we consider different markers, namely transposons, and because we assume that pathogen populations are monomorphic. We estimate rates of substitution of genotypes in absolute time, rather than rates of mutations per cell division or per generation; we discuss, however, the possibility of equating these rates using the argument of Kimura (1968). We treat the change in pathogen population structure and size inside the host as a "black box." Future work might model this within-host process explicitly; the process will depend on the pathogen under consideration.

For a given genetic marker in a specific organism, our estimation procedure requires that several decisions be made. First, an appropriate stochastic process model that describes the biologic changes in the genetic marker must be selected. Second, a detection scheme, which describes how the experimentally detectable quantities under the available technology relate to the underlying process of genetic change, must be chosen. Third, the investigator must decide on the frequency at which observations are taken. Last, the size of the sample must be selected. Investigators can often choose technologies, monitoring frequencies, and sample sizes in such a way as to reduce the variance of the estimated rate.

In Sections 2–4, we describe components of the design of rate estimation schemes. Sections 5 and 6 develop the estimation framework in the case that the genetic marker is a transposable element. In Sections 7–9, we show applications to the estimation of change rates of the element IS*6110* in *M. tuberculosis*. Finally, we discuss in Section 10 experimental design issues and implications. Our presentation is meant to be modular, in that if any of the components are modified (underlying process, detection scheme, and observation frequency), the estimation procedure could be correspondingly modified without much effort.

2. The process of genetic change: transposable elements

Transposable elements or transposons are sequences that are capable of moving to new locations in a genome. It is convenient to conceptualize three types of changes that can be experienced by a given copy of a transposon. (1) It can be duplicated, with the new copy moving to a new genomic location; (2) it can be shifted to a different position; or (3) it can be excised and lost from the genome. We term these types of events *births*,

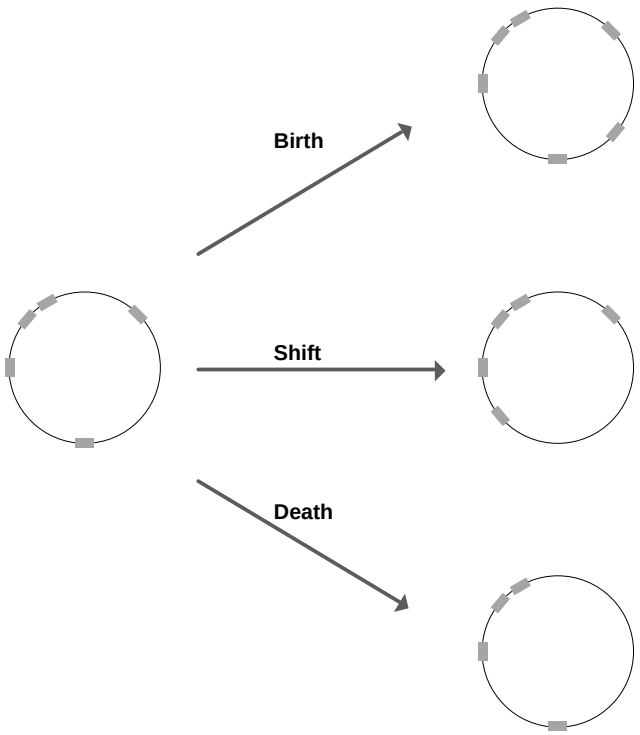

Fig. 1. Three types of transposition that can happen to a genome, here represented by a circle. Birth—a copy of the transposon is added; shift—a copy is moved to another location; death—a copy is lost. Transposons are represented by rectangles.

shifts, and *deaths*, respectively (Fig. 1). Note that some types of transposon might be able to experience only one or two of these types of changes, as has been assumed in some previous models (Sawyer and Hartl, 1986, for example).

To model this "birth–shift–death" process, let $K(t)$ be the number of copies of a particular transposable element that a specific cell has at time t. For simplicity, we assume that copies of the transposon change independently of each other. Instantaneous birth, shift and death rates for the cell are then proportional to the number of copies of the element, and consequently, $\{K(t)\}$ is a linear birth–shift–death process in continuous time. Owing to the fact that transposition need not happen only during cell division, we assume that the birth, shift, and death rates for a cell with a fixed number of elements are constant over the lifetime of the cell. This time-homogeneity assumption, which leads to an exponentially distributed waiting time until the occurrence of an event, ignores the possibility that transposition rates might vary with transcription and gene expression levels over the cell cycle. For mathematical tractability, however, using homogeneous rates is preferable to modeling the rates as functions of cell cycle stages.

For a cell with a single element, we denote instantaneous birth, shift and death rates by λ, γ, and μ, respectively, and measure them in events per copy per unit time. If the process lasts long enough, it is possible that all copies of the transposon might be excised from the genome. Here we assume *nonextinction*, or that at least one copy of the transposon is always present.

The above assumptions are natural for a single cell. Additionally, however, if all copy number variants are selectively neutral, the rate at which neutral allelic variants become fixed in the entire pathogen population (the substitution rate) equals the neutral mutation rate, independent of population size and structure (Kimura, 1968, 1983). Intuitively, the equality of substitution and mutation rates follows from the fact that a potential increase in substitution rate due to an increase in mutation rate is exactly counteracted by a decrease in the probability that any particular mutant will become fixed in the population. Thus, if neutrality can be assumed, with mutation rates linear in copy number, the rate at which a k-copy strain within the host is replaced equals the cellular mutation rate, or $k(\lambda+\gamma+\mu)$, regardless of within-host population processes. Because three types of mutation are possible—births, shifts, and deaths—within-host substitution rates for strains resulting from births, shifts, and deaths, respectively, are $k\lambda$, $k\gamma$, and $k\mu$. If any transposon variants are selected or linked to selected alleles, as is likely to occur, especially in non-recombining pathogens, owing to pressures induced by host immune systems or by drugs, explicit modeling of within-host population genetics is necessary to determine the relationship between the cellular birth–shift–death process and the population-level process. As a first approximation we assume neutrality.

To translate the cellular birth–shift–death process into a population-level birth–shift–death process, we additionally assume that only one strain is present within each host at appreciable frequency. This is justifiable, as pathogen isolates from patients often show only a single genotype. For example, in one study of 1277 *M. tuberculosis* isolates (de Boer et al., 2000), 92.6% of isolates typed for the transposon IS*6110* were monomorphic. We also assume that sufficiently many pathogens are in each isolate that their genotype is representative of that of the whole pathogen population.

Henceforth, we treat the substitution process of the whole pathogen population as a linear birth–shift–death process with substitution rates equal to the cellular transposition rates. This step involves an additional assumption, as the Kimura (1968) argument justifies equating only the substitution and cellular transposition *rates*, and not the substitution and cellular transposition *stochastic processes*. Although the distribution of the waiting time until substitution depends on within-host mutation and population processes, and need not be exponentially distributed even for simple neutral population-genetic models of the pathogen (Kelly, 1979; Watterson, 1982), as a first approximation we assume an exponential waiting time, analogously to the corre-

sponding assumption for the classical molecular substitution clock (Kimura, 1983, p. 69).

3. Detection schemes

The biological process of transposition is observed through a genotyping technology. With the ideal technology, births, shifts, and deaths are all detectable. For example, if the genome of a cell could be sequenced at two points closely separated in time, the times and types of genetic changes would be immediately apparent. Of course, simpler technologies can also produce full knowledge of the genetic arrangement. We say that such a technology has *complete resolution* (Fig. 2).

With some technologies, not all types of changes are detectable. The genotyping procedure might only allow determination of changes in copy number, or complex genotypes might be summarized by recording copy numbers. Under this scheme, births and deaths are detectable. Because shifts change only positions of copies, and not the number of copies, however, they are not detected. This level of resolution could be produced if the number of copies of the transposon corresponds to a number of gel bands. For example, suppose that a restriction site is located inside the transposable element. Then the number of copies of the element equals the number of bands that the restriction-digested genome produces minus the corresponding number of bands for a zero-copy strain. We denote the situation in which the detectable quantity is the number of copies of the transposon *copy number resolution* (Fig. 2).

Other technologies that imperfectly view the transposition process might be imagined. We also consider *change resolution* (Fig. 2), in which it is possible to detect a modification in transposon genotype, but impossible to determine the type of change that occurred. In this scheme, births, deaths, and shifts are grouped as *changes*. This situation might result from limited record-keeping: for example, past studies might have noted whether changes occurred in gel-banding patterns, but might not have recorded details of the patterns. Note that unlike copy number resolution, change resolution detects shifts if they occur. Thus, detectable events under change resolution are not necessarily subsumed by those seen with copy number resolution.

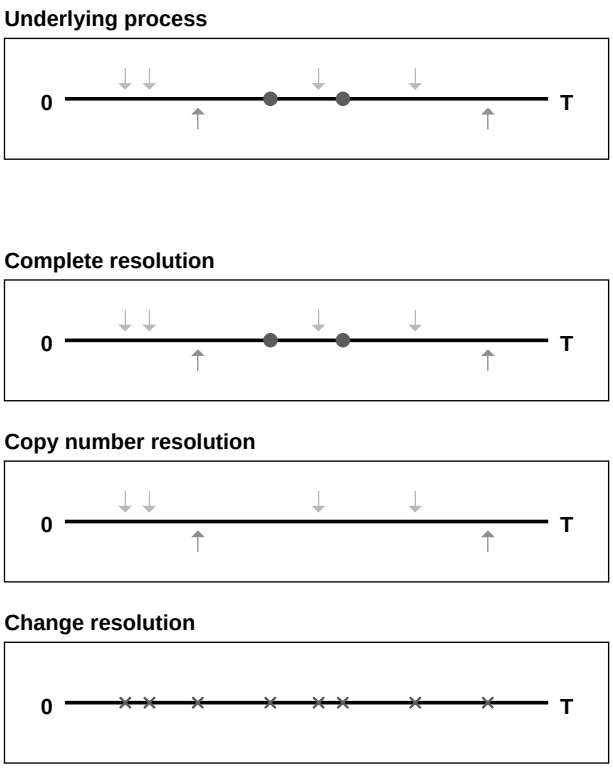

Fig. 2. Detection schemes for a birth–shift–death process. Time proceeds from 0 to T. During the time interval, births (arrows pointing upward), shifts (circles), and deaths (arrows pointing downward) occur. With *complete resolution*, all types of changes are detectable. With *copy number resolution*, shifts are not detectable. With *change resolution*, all three types of events are detected only as "changes" (crosses).

4. Monitoring schemes

Genotypes are monitored over time, and we distinguish three classes of "monitoring schemes." The simplest is that in which the process is *continuously* observed (Fig. 3): the exact state of the genotype is known at all points in time. Although this scheme is impractical, it provides results that help in the consideration of more useful schemes. Also, as the expense of genotyping decreases, it may be possible to genotype often enough to approximate continuous monitoring.

A second scheme is that in which the process is *frequently* observed (Fig. 3). We define "frequently" to mean that the process is observed sufficiently often that the probability that more than one change could have happened between observations is negligible. Thus, the definition of "frequent" varies with the rate of the process. This scheme is analogous to the parsimony assumption of phylogenetics—if initial and final states differ in only a single way, it is assumed that only one change occurred. "Frequent monitoring" leads to a straightforward estimation procedure, and we show in Section 8 that the scheme applies to observed data.

We also consider the situation in which the process is *infrequently* observed (Fig. 3). In this scheme, more than one change might happen between two observations. The frequent and infrequent schemes each have two notable special cases. If all genotyping intervals have equal lengths, then monitoring is *equidistant*. A second special case is that in which the process is only observed

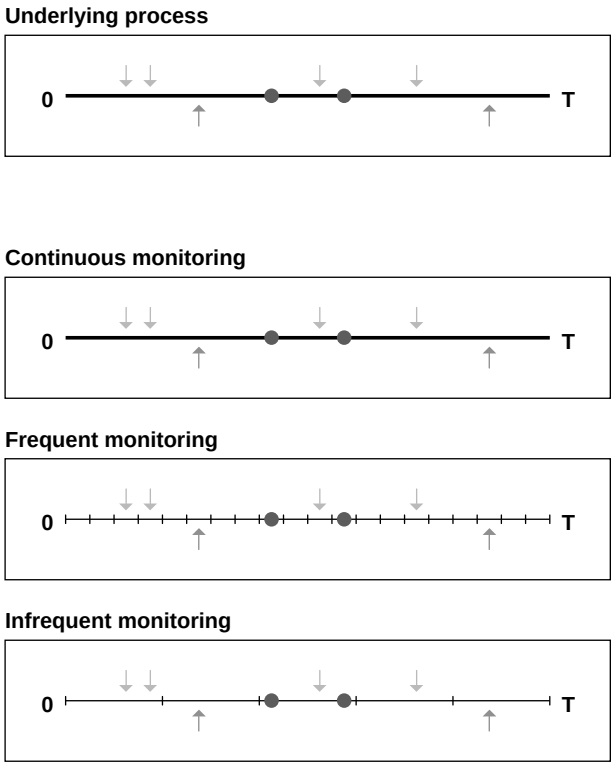

Fig. 3. Monitoring schemes for a birth–shift–death process, for which all types of events are detectable. Time proceeds from 0 to T. During the time interval, births (arrows pointing upward), shifts (circles), and deaths (arrows pointing downward) occur. A vertical line at a point in time indicates that an observation is taken at that time point. With *continuous monitoring*, the process is observed throughout the interval. With *frequent monitoring*, observations are made often enough that at most one change happens between observations. With *infrequent monitoring*, observations are rare, and several genotypic changes can occur between two observations. For frequent and infrequent monitoring schemes, the case in which observations occur equidistantly is shown.

twice—at a beginning and at an ending time point. Of course, other schemes, such as those that treat observation times as random variables, can also be envisioned. Those schemes are beyond the scope of this article.

5. Estimating the rate of change

Detection and monitoring designs can in principle be combined with models of the underlying biological process to obtain maximum-likelihood estimators of parameters. Table 1 classifies and presents likelihood functions for various schemes, some of which are discussed below.

5.1. Estimation: continuous monitoring

Here we adapt previous work on estimation for a continuously monitored linear birth–death process (Darwin, 1956; Reynolds, 1973; Keiding, 1975) to our situation of a linear "birth–shift–death" process. The problem described in this section is case 1 in Table 1. We derive the likelihood when only one host is being studied, multiplying individual host likelihoods to obtain the overall likelihood for many independent hosts. If all hosts are observed continuously over the same interval, it is equivalent to consider a single host with initial copy number equal to the sum of the copy numbers of all the hosts.

Let θ denote the overall change rate, that is, the sum of the birth rate (λ), shift rate (γ), and death rate (μ): $\theta = \lambda + \gamma + \mu$. Recall that rates at which births, shifts, and deaths occur are proportional to the number of copies of the transposable element. For the birth–shift–death process, if k copies are present in a strain, the waiting time until a change occurs has an exponential distribution with probability density function

$$p(t) = k\theta e^{-k\theta t}, \quad t > 0. \tag{1}$$

For the three types of change—births, shifts, and deaths—the waiting times until the first changes of those types are also exponentially distributed, with parameters $k\lambda$, $k\gamma$, and $k\mu$, respectively. Using well-known properties of exponentially distributed random variables, regardless of the time of an event, the probability that the event is a birth equals λ/θ. Similarly, the probability is γ/θ that it is a shift and μ/θ that it is a death.

Suppose that initial copy number is k_0, that events occur at times $t_1, t_2, \ldots, t_n$, that the type of change represented by the ith event is C_i (where C_i can equal *birth*, *shift*, or *death*), and that the number of elements present immediately after the ith event is k_i. Denote the beginning of an interval over which the process is observed by t_0 and the end of the interval by t_{n+1} (we then allow C_0 and C_{n+1} to take the value *no change*). Observed data in this scheme consist of the ordered triples $E_i = (C_i, k_i, t_i)$, $i = 0, 1, 2, \ldots, n+1$.

For $i = 0, \ldots, n-1$, the probability density for a birth at time t_{i+1} (that is, the likelihood of the parameters given E_{i+1}, with $E_{i+1} = (birth, k_i + 1, t_{i+1})$) is

$$L(\lambda, \gamma, \mu | E_{i+1}) = k_i \lambda e^{-k_i\theta(t_{i+1} - t_i)}. \tag{2}$$

Corresponding expressions for shifts and deaths are $k_i\gamma \exp[-k_i\theta(t_{i+1} - t_i)]$ and $k_i\mu \exp[-k_i\theta(t_{i+1} - t_i)]$, respectively. The probability that no change occurs in the last interval from t_n to t_{n+1} is $\exp[-k_n\theta(t_{n+1} - t_n)]$. Multiplying over independent intervals,

$$\begin{aligned} L(\lambda, \gamma, \mu | E_0, E_1, \ldots, E_{n+1}) &= \lambda^b \gamma^f \mu^d k_0 k_1 k_2 \ldots k_{n-1} \\ &\quad \times \exp\left[-\theta \sum_{i=0}^{n} k_i (t_{i+1} - t_i)\right], \end{aligned} \tag{3}$$

Table 1
Likelihood functions and maximum-likelihood estimators of three types of transposition rates

Case	Underlying process	Level of resolution	Monitoring scheme	Definition of θ	Likelihood	Maximum-likelihood estimators
1	Birth–shift–death	Complete	Continuous	$\lambda+\gamma+\mu$	$\lambda^b\gamma^f\mu^d k_0k_1k_2\ldots k_{n-1}\exp[-\theta\sum_{i=0}^{n}k_i(t_{i+1}-t_i)]$	$\hat{\lambda}=\frac{b}{s}$, $\hat{\gamma}=\frac{f}{s}$, $\hat{\mu}=\frac{d}{s}$
2	Birth–shift–death	Copy number	Continuous	$\lambda+\mu$	$\lambda^b\mu^d k_0k_1k_2\ldots k_{n-1}\exp[-\theta\sum_{i=0}^{n}k_i(t_{i+1}-t_i)]$	$\hat{\lambda}=\frac{b}{s}$, $\hat{\mu}=\frac{d}{s}$
3	Birth–shift–death	Change	Continuous	$\lambda+\gamma+\mu$	$\theta^c k_0k_1k_2\ldots k_{n-1}\exp[-\theta\sum_{i=0}^{n}k_i(t_{i+1}-t_i)]$	$\hat{\theta}=\frac{c}{s}$
4	Birth–shift	Complete	Continuous	$\lambda+\gamma$	$\lambda^b\gamma^f k_0k_1k_2\ldots k_{n-1}\exp[-\theta\sum_{i=0}^{n}k_i(t_{i+1}-t_i)]$	$\hat{\lambda}=\frac{b}{s}$, $\hat{\gamma}=\frac{f}{s}$
5	Birth–death	Complete	Continuous	$\lambda+\mu$	Same as case 2	Same as case 2
6	Shift–death	Complete	Continuous	$\gamma+\mu$	$\gamma^f\mu^d k_0k_1k_2\ldots k_{n-1}\exp[-\theta\sum_{i=0}^{n}k_i(t_{i+1}-t_i)]$	$\hat{\gamma}=\frac{f}{s}$, $\hat{\mu}=\frac{d}{s}$
7	Birth	Complete	Continuous	λ	$\lambda^b k_0k_1k_2\ldots k_{n-1}\exp[-\theta\sum_{i=0}^{n}k_i(t_{i+1}-t_i)]$	$\hat{\lambda}=\frac{b}{s}$
8	Shift	Complete	Continuous	γ	$\gamma^f k_0^f\exp[-\theta k_0t]$	$\hat{\gamma}=\frac{f}{s}$
9	Shift	Copy number	Continuous	γ	No detectable changes	Not possible
10	Shift	Change	Continuous	γ	Same as cases 3 and 8	Same as cases 3 and 8
11	Death	Complete	Continuous	μ	$\mu^d k_0k_1k_2\ldots k_{n-1}\exp[-\theta\sum_{i=0}^{n}k_i(t_{i+1}-t_i)]$	$\hat{\mu}=\frac{d}{s}$
12	Birth–shift–death	Complete	Frequent	$\lambda+\gamma+\mu$	$\prod_{\{i:G_i=\mathscr{B}\}}\frac{\lambda}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{F}\}}\frac{\gamma}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{D}\}}\frac{\mu}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
13	Birth–shift–death	Copy number	Frequent	$\lambda+\mu$	$\prod_{\{i:G_i=\mathscr{B}\}}\frac{\lambda}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{D}\}}\frac{\mu}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
14	Birth–shift–death	Change	Frequent	$\lambda+\gamma+\mu$	$\prod_{\{i:G_i=\mathscr{C}\}}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
15	Birth–shift	Complete	Frequent	$\lambda+\gamma$	$\prod_{\{i:G_i=\mathscr{B}\}}\frac{\lambda}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{F}\}}\frac{\gamma}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
16	Birth–death	Complete	Frequent	$\lambda+\mu$	Same as case 13	Numerical
17	Shift–death	Complete	Frequent	$\gamma+\mu$	$\prod_{\{i:G_i=\mathscr{F}\}}\frac{\gamma}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{D}\}}\frac{\mu}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
18	Birth	Complete	Frequent	λ	$\prod_{\{i:G_i=\mathscr{B}\}}\frac{\lambda}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
19	Shift	Complete	Frequent	γ	$\prod_{\{i:G_i=\mathscr{F}\}}\frac{\gamma}{\theta}(1-e^{-k^1\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k^1\theta u_i}$	Numerical
20	Shift	Copy number	Frequent	γ	No detectable changes	Not possible
21	Shift	Change	Frequent	γ	Same as cases 14 and 19	Numerical
22	Death	Complete	Frequent	μ	$\prod_{\{i:G_i=\mathscr{D}\}}\frac{\mu}{\theta}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
23	Birth–shift–death	Change	Infrequent	$\lambda+\gamma+\mu$	$\prod_{\{i:G_i=\mathscr{C}\}}(1-e^{-k_i\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_i\theta u_i}$	Numerical
24	Birth–shift	Change	Infrequent	$\lambda+\gamma$	Same as case 23	Numerical
25	Birth–death	Change	infrequent	$\lambda+\mu$	Same as case 23	Numerical
26	Shift–death	Change	Infrequent	$\gamma+\mu$	Same as case 23	Numerical
27	Birth	Change	Infrequent	λ	Same as case 23	Numerical
28	Shift	Complete	Infrequent	γ	$\prod_{\{i:G_i=\mathscr{F}\}}\frac{\gamma}{\theta}(1-e^{-k_1\theta u_i})\prod_{\{i:G_i=\mathscr{N}\}}e^{-k_1\theta u_i}$	Numerical
29	Shift	Copy number	Infrequent	γ	No detectable changes	Not possible
30	Shift	Change	Infrequent	γ	Same as cases 23 and 28	Numerical
31	Death	Change	Infrequent	μ	Same as case 23	Numerical

Birth, shift, and death rates are λ, γ, and μ, respectively. Observed total numbers of births, shifts, deaths, and total changes are b, f, d, and c, respectively, and the total time experienced in the interval is s. For continuous monitoring, the copy number at the time t_i of the ith event becomes k_i immediately after the event, the total length of the time interval is t, and the t_i are measured since time 0. For frequent and infrequent monitoring k_i is the copy number at the start of the ith interval, and the values of u_i denote times between events. The symbols $\mathscr{B}$, $\mathscr{F}$, $\mathscr{D}$, $\mathscr{C}$, and $\mathscr{N}$ refer to the words *birth*, *shift*, *death*, *change*, and *no change*, respectively. For continuous monitoring one host with many intervals is assumed, and likelihoods for many individuals are obtained by multiplying across independent hosts; for frequent and infrequent monitoring the likelihood assumes multiple intervals, which may derive from one or many hosts. For completeness, case 3 is shown, although it is perhaps not sensible, as discerning k_i for $i\geqslant 1$ requires a higher level of resolution; also, cases 9, 20, and 29 are abstract scenarios that do not allow any parameters to be estimated.

Table 2
Likelihood of the parameters given initial copy number k and a time interval of length u, and partial derivatives of components of the log-likelihood

Type of event (G)	Likelihood $L(\lambda,\gamma,\mu\|G,k,u)$	$\frac{\partial}{\partial\lambda}\ln[L(\lambda,\gamma,\mu\|G,k,u)]$	$\frac{\partial}{\partial\gamma}\ln[L(\lambda,\gamma,\mu\|G,k,u)]$	$\frac{\partial}{\partial\mu}\ln[L(\lambda,\gamma,\mu\|G,k,u)]$
Birth	$\frac{\lambda}{\theta}(1-e^{-k\theta u})$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}+\frac{1}{\lambda}-\frac{1}{\theta}$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}-\frac{1}{\theta}$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}-\frac{1}{\theta}$
Shift	$\frac{\gamma}{\theta}(1-e^{-k\theta u})$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}-\frac{1}{\theta}$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}+\frac{1}{\gamma}-\frac{1}{\theta}$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}-\frac{1}{\theta}$
Death	$\frac{\mu}{\theta}(1-e^{-k\theta u})$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}-\frac{1}{\theta}$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}-\frac{1}{\theta}$	$\frac{kue^{-k\theta u}}{1-e^{-k\theta u}}+\frac{1}{\mu}-\frac{1}{\theta}$
No change	$e^{-k\theta u}$	$-ku$	$-ku$	$-ku$

where b, f, and d denote the total numbers of births, shifts, and deaths that occur, respectively.

The total time experienced by all copies during the interval, or $\sum_{i=0}^{n} k_i(t_{i+1}-t_i)$, is denoted by s. Maximizing (3), we obtain the maximum-likelihood estimators

$$\hat{\lambda}=\frac{b}{s},\quad \hat{\gamma}=\frac{f}{s},\quad \hat{\mu}=\frac{d}{s}. \tag{4}$$

If independent hosts 1 to m are indexed by j and L is computed using (3), the overall likelihood is

$$L_{total}(\lambda,\gamma,\mu|\text{data})=\prod_{j=1}^{m} L(\lambda,\gamma,\mu|\text{data}_j). \tag{5}$$

The estimators in (4) still apply, except that b, f, d, and s now refer to total numbers of births, shifts, and deaths, and the total time experienced by all copies in all hosts.

From this situation of an underlying linear birth–shift–death process, completely detected and continuously monitored, other cases can be studied (Table 1, cases 2–11). For example, likelihoods can be derived for situations in which the process includes only one or two of the types of events (Keiding, 1974, 1975; Basawa and Prakasa Rao, 1980), and likelihoods under change resolution can be obtained from (3) by grouping all changes into a single category.

5.2. Estimation: frequent monitoring

Because continuous monitoring would require an organism to be connected to a "genotyping machine," a continuous sampling scheme cannot describe any currently realistic observation system. If genotyping occurs often compared to the rates of a process, however, continuous monitoring is approximated by assuming that changes occur exactly at the times of observation.

Alternatively, we can take note that a change occurred in the interval, and not associate that change with a particular time point during the interval. This is our approach when we say that a process is "frequently monitored." The problem considered in this section is that of case 12 in Table 1, the estimation of the rates of a frequently monitored linear birth–shift–death process that is observed completely. Again, we assume a single host, and the likelihood for many hosts is obtained by taking the product of individual host likelihoods.

"Frequent monitoring" assumes that at most one event happens during an observation interval. Suppose we observe a process $m+1$ times, with the m intervals separating observations having lengths u_1, $u_2,\ldots,$ u_m. For all values of j from 1 to m, let $G_j \in \{birth, shift, death, no\ change\}$ equal the type of event that occurred during the jth interval. Observed data under frequent monitoring consist of ordered triples $E_j=(G_j,k_j,u_j)$, where k_j is the copy number at the start of the jth interval (j ranges from 1 to m).

The probability that an event occurs during a time interval of length u is the cumulative density function obtained by integrating (1): $P(u)=\int_0^u k\theta e^{-k\theta t}\,dt = 1-e^{-k\theta u}$. As before, given that an event has taken place, the probability that it is a birth is λ/θ. For a shift, this probability is γ/θ, and for a death, it is μ/θ.

We computed in (2) the density function of the time at which a birth occurs. By integrating this density function, we can compute the probability that a birth occurs during an interval of length u. That is, if E equals $(birth,k,u)$, the likelihood of the parameters is

$$L(\lambda,\gamma,\mu|E)=\int_0^u k\lambda e^{-k\theta t}dt=\frac{\lambda}{\theta}(1-e^{-k\theta u}). \tag{6}$$

Similar expressions give the likelihoods if a shift or a death occurs in the interval in place of a birth (Table 2).

By multiplying across independent intervals, with L given in Table 2, we obtain the likelihood of the parameters given the observations (G_j,k_j,u_j), $j=1,2,\ldots,m$.

$$L_{total}(\lambda,\gamma,\mu)=\prod_{j=1}^{m} L(\lambda,\gamma,\mu|G_j,k_j,u_j). \tag{7}$$

It is straightforward to differentiate the logarithm of likelihood (7), and to numerically find maximum-likelihood estimates of the three parameters. The component derivatives are shown in Table 2. For one interval, the derivative $\partial \ln L/\partial\lambda$ is as follows; the

overall derivative is obtained by summing across intervals.

$$\begin{aligned}\frac{\partial}{\partial\lambda}\ln L(\lambda,\gamma,\mu|G,k,u) &= \chi(G=birth)\frac{\partial}{\partial\lambda}\ln L(\lambda,\gamma,\mu|birth,k,u)\\ &+ \chi(G=shift)\frac{\partial}{\partial\lambda}\ln L(\lambda,\gamma,\mu|shift,k,u)\\ &+ \chi(G=death)\frac{\partial}{\partial\lambda}\ln L(\lambda,\gamma,\mu|death,k,u)\\ &+ \chi(G=no\ change)\\ &\times\frac{\partial}{\partial\lambda}\ln L(\lambda,\gamma,\mu|no\ change,k,u),\end{aligned} \tag{8}$$

where $\chi(G=g)=1$ if $G=g$, and is zero otherwise. Derivatives with respect to γ and μ are analogous. As before, this situation of an underlying linear birth–shift–death process, completely detected and frequently monitored, allows likelihoods to be derived for other processes and detection schemes (Table 1, cases 12–22).

5.3. Estimation: infrequent monitoring

Under infrequent monitoring, many events might occur between observations. For example, a change from 4 to 5 copies can follow one of infinitely many paths, such as 4-3-2-3-4-5 or 4-5-6-5. Except in special cases (Kendall, 1949; Basawa and Prakasa Rao, 1980), calculation of likelihood functions becomes more difficult.

For many applications, infrequent monitoring may be less applicable than other schemes. The amount of genetic change might be small over the length of time for which genotypes can be monitored, so that intervals rarely contain multiple changes. Our earlier method (Tanaka and Rosenberg, 2001) can be interpreted as the case of change resolution and infrequent monitoring, for any process that is linear in copy number (Table 1, case 23). The idea was that in an interval of length u, the probability of no change is $\exp(-k\theta u)$, and the probability of one or more changes is $1-\exp(-k\theta u)$. Thus, the likelihood function is a product of $\exp(-k\theta u)$ and $1-\exp(-k\theta u)$ terms, depending on whether or not at least one change occurred in the interval.

The likelihood function in Tanaka and Rosenberg (2001) can also be seen as the likelihood for a shift process with complete resolution (where all changes are counted as if they were shifts). For a shift-only process, likelihoods for complete resolution and change resolution are equivalent (see Table 1). We will not consider infrequent monitoring further; recently developed numerical schemes (Golinelli, 2000) may in the future be applied to inference in the full infrequently monitored birth–shift–death process.

6. Variance of estimates

Variances of estimates are of interest partly because they may depend on quantities that can be controlled (for example, the length of time between observations, the total number of observations). Thus, investigators might select values of these quantities to reduce the variances.

6.1. Variance: continuous monitoring

As was noted earlier, if all hosts have equal sampling intervals, a large initial copy number is equivalent to a large number of hosts, each with small copy number. Thus, asymptotic results for a continuously monitored birth–death process (Keiding, 1975), assuming a large initial copy number in a single host, can be treated as large-sample results. We follow previous work to obtain the variances, incorporating the shift process.

For a single host with large copy number, the approximate variance–covariance matrix is given by the inverse of the expected information matrix I (Elandt-Johnson, 1971, p. 306). This information matrix for case 1 is obtained using derivatives of the likelihood L in (3).

$$I(\lambda,\gamma,\mu)=-\mathbb{E}\begin{pmatrix}\frac{\partial^2\ln L}{\partial\lambda^2} & \frac{\partial^2\ln L}{\partial\lambda\,\partial\gamma} & \frac{\partial^2\ln L}{\partial\lambda\,\partial\mu}\\ \frac{\partial^2\ln L}{\partial\gamma\,\partial\lambda} & \frac{\partial^2\ln L}{\partial\gamma^2} & \frac{\partial^2\ln L}{\partial\gamma\,\partial\mu}\\ \frac{\partial^2\ln L}{\partial\mu\,\partial\lambda} & \frac{\partial^2\ln L}{\partial\mu\,\partial\gamma} & \frac{\partial^2\ln L}{\partial\mu^2}\end{pmatrix} \tag{9}$$

$$=\begin{pmatrix}\frac{\mathbb{E}[B]}{\lambda^2} & 0 & 0\\ 0 & \frac{\mathbb{E}[F]}{\gamma^2} & 0\\ 0 & 0 & \frac{\mathbb{E}[D]}{\mu^2}\end{pmatrix}. \tag{10}$$

Here, componentwise expectation is written with $\mathbb{E}$ in front of the matrix. $\mathbb{E}[B]$, $\mathbb{E}[F]$, and $\mathbb{E}[D]$ denote expected numbers of births, shifts and deaths that occur in an interval of length t. If S is the total amount of time experienced by all copies of the element, then $\mathbb{E}[B]=\lambda\mathbb{E}[S]$, $\mathbb{E}[F]=\gamma\mathbb{E}[S]$, and $\mathbb{E}[D]=\mu\mathbb{E}[S]$. Because shifts do not affect copy number, $\mathbb{E}[S]$ is the same as in the linear birth–death process (Puri, 1968):

$$\mathbb{E}[S]=\frac{e^{(\lambda-\mu)t}-1}{\lambda-\mu}k_0. \tag{11}$$

Inverting the matrix in (10) and noting that covariances are zero, we can apply (11) and the estimators from (4) to obtain variance estimates:

$$\widehat{\mathrm{Var}}\begin{pmatrix}\hat\lambda\\ \hat\gamma\\ \hat\mu\end{pmatrix}=\begin{pmatrix}b\\ f\\ d\end{pmatrix}\frac{b-d}{k_0s^2[e^{(b-d)t/s}-1]}, \tag{12}$$

where the variance operator is applied componentwise; b, f, and d are the observed numbers of births, shifts,

Table 3
Second partial derivatives of components of the log-likelihood

Type of event (G)	$\frac{\partial^2}{\partial\lambda^2}\ln[L(\lambda,\gamma,\mu\|G,k,u)]$	$\frac{\partial^2}{\partial\gamma^2}\ln[L(\lambda,\gamma,\mu\|G,k,u)]$	$\frac{\partial^2}{\partial\mu^2}\ln[L(\lambda,\gamma,\mu\|G,k,u)]$	$\frac{\partial^2}{\partial\lambda\,\partial\gamma}\ln[L(\lambda,\gamma,\mu\|G,k,u)]$
Birth	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}-\frac{1}{\lambda^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$
Shift	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}-\frac{1}{\gamma^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$
Death	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}-\frac{1}{\mu^2}+\frac{1}{\theta^2}$	$\frac{-(ku)^2e^{-k\theta u}}{(1-e^{-k\theta u})^2}+\frac{1}{\theta^2}$
No change	0	0	0	0

$(\partial^2/\partial\lambda\,\partial\mu)\ln[L(\lambda,\gamma,\mu|G,k,u)]$ and $(\partial^2/\partial\gamma\,\partial\mu)\ln[L(\lambda,\gamma,\mu|G,k,u)]$ both have the same values as $(\partial^2/\partial\lambda\,\partial\gamma)\ln[L(\lambda,\gamma,\mu|G,k,u)]$.

and deaths, respectively, and s is the total time experienced by all copies of the element.

Because of the multiplicative nature of the likelihood function across independent individuals, information matrices are additive across hosts. Thus, the variance–covariance matrix for the estimates is the inverse of the sum across individuals of information matrices. If I_j is the information matrix for the jth host, then

$$\widehat{\boldsymbol{I}_{total}}(\hat\lambda,\hat\gamma,\hat\mu)=\sum_{j=1}^{m}\widehat{\boldsymbol{I}_j}(\hat\lambda,\hat\gamma,\hat\mu), \tag{13}$$

and the variance–covariance matrix is obtained by inverting (13). Variances for cases 2–11 can be obtained as special cases, or by analogous reasoning.

6.2. Variance: frequent monitoring

Unlike the continuous case, in the frequent case (case 12) there is no simple way to treat many individuals as a single individual with large copy number. If individuals are observed many times, the different intervals for a given host are independent under the assumptions of the model. Thus, the information matrix for many hosts is obtained as the sum of the information matrices of the individuals, or more simply, as the sum of the information matrices for independent intervals.

Consider an individual of initial copy number k observed for a single interval of length u, with event G occurring during the interval. We require expectations of second derivatives of the log-likelihood. For example,

$$\begin{aligned}
&-\mathbb{E}\left[\frac{\partial^2}{\partial\lambda\,\partial\gamma}\ln L(\lambda,\gamma,\mu|G,k,u)\right]\\
&\quad=\Pr(birth,k,u)\frac{\partial^2}{\partial\lambda\,\partial\gamma}\ln L(\lambda,\gamma,\mu|birth,k,u)\\
&\qquad+\Pr(shift,k,u)\frac{\partial^2}{\partial\lambda\,\partial\gamma}\ln L(\lambda,\gamma,\mu|shift,k,u)\\
&\qquad+\Pr(death,k,u)\frac{\partial^2}{\partial\lambda\,\partial\gamma}\ln L(\lambda,\gamma,\mu|death,k,u)\\
&\qquad+\Pr(no\ \ change,k,u)\\
&\qquad\times\frac{\partial^2}{\partial\lambda\,\partial\gamma}\ln L(\lambda,\gamma,\mu|no\ \ change,k,u).
\end{aligned}\tag{14}$$

The required probabilities are obtained from Table 2: for example, $\Pr(birth,k,u)$, or $L(\lambda,\gamma,\mu|birth,k,u)$, equals $\lambda\theta^{-1}(1-e^{-k\theta u})$. The second derivatives (Table 3) are computed from first derivatives in Table 2. Inserting quantities from Tables 2 and 3 into (14),

$$\begin{aligned}
&-\mathbb{E}\left[\frac{\partial^2}{\partial\lambda\,\partial\gamma}\ln L(\lambda,\gamma,\mu|G,k,u)\right]\\
&\quad=\frac{(ku)^2e^{-k\theta u}}{1-e^{-k\theta u}}-\frac{1-e^{-k\theta u}}{\theta^2}.
\end{aligned}\tag{15}$$

Abbreviating the right-hand side of (15) by A, calculating the remaining derivatives, and using the same form for the expected information matrix as in (9), we obtain

$$I(\lambda,\gamma,\mu)=\begin{pmatrix}A+z/\lambda & A & A\\ A & A+z/\gamma & A\\ A & A & A+z/\mu\end{pmatrix}, \tag{16}$$

where $z=(1-e^{-k\theta u})/\theta$.

To estimate the variance–covariance matrix, the maximum-likelihood estimates $\hat\lambda$, $\hat\gamma$, and $\hat\mu$ obtained from maximizing (7) are inserted into the estimated information matrix for the data set (that is, the following)

$$\widehat{\boldsymbol{I}_{total}}(\hat\lambda,\hat\gamma,\hat\mu)=\sum_{j=1}^{m}\widehat{\boldsymbol{I}_j}(\hat\lambda,\hat\gamma,\hat\mu). \tag{17}$$

Here m is the total number of intervals, and $\hat I_j$ is the estimated information for each interval, obtained by inserting maximum-likelihood estimates into (16). Finally, the estimated variance–covariance matrix is $\widehat{\boldsymbol{I}_{total}}^{-1}$. Variances in cases 12–22 are analogously obtained.

Table 4
Summary statistics for three data sets

	Niemann	SF	Niemann + SF
Number of intervals	56	247	303
Number of intervals with births	4	14	18
Number of intervals with shifts	0	2	2
Number of intervals with deaths	1	11	12
Number of intervals with no change	51	220	271
Number of patients	56	204	260
Mean initial copy number	10.77	10.01	10.15
Standard deviation of initial copy number distribution	3.06	5.36	5.02
Mean interval length (days)	237	115	137
Standard deviation of interval length distribution (days)	227	144	168
Total length of all time intervals (days)	13,250	28,274	41,524

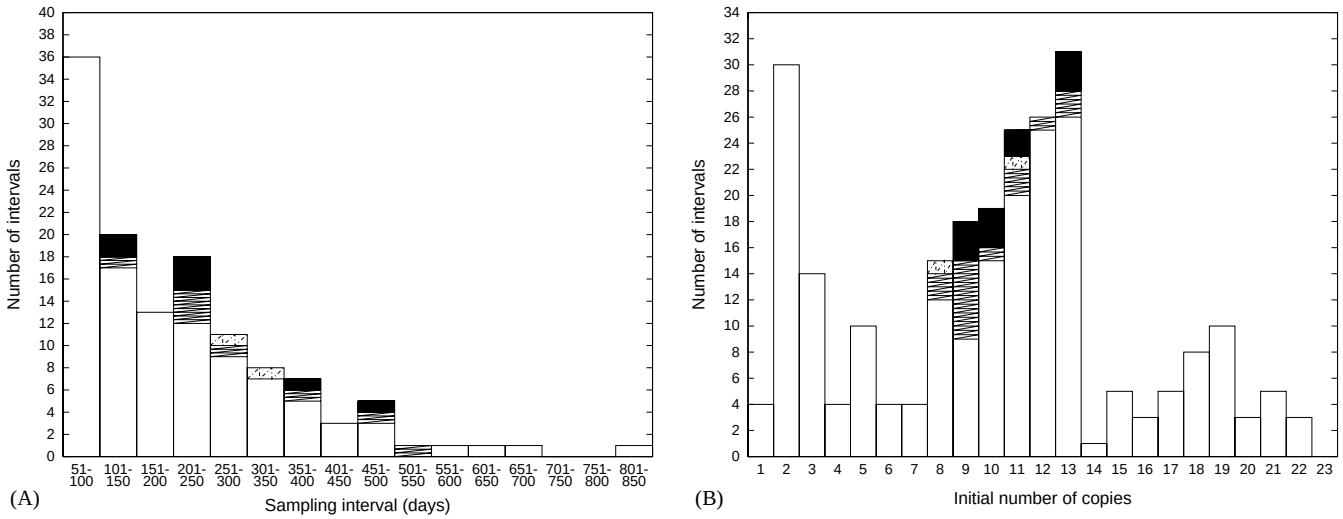

Fig. 4. Observations in the SF data set. (A) Distribution of observations as a function of the length of the sampling interval. One hundred and twenty one intervals of length 50 days or less are not shown; of these, 111 showed no change, six included births, and four included deaths. (B) Distribution of observations as a function of initial copy number. The patterns are the same for both plots: shaded—deaths; dot-dashed—shifts; zigzag—births; empty boxes—no change.

Unlike the continuous case, in which observations used for inferences of each of the parameters do not depend on the other parameters, covariances in the frequent case are nonzero. As soon as one event occurs in an interval, other events are precluded. Thus, parameter estimates should be negatively correlated with each other.

7. IS6110 data

We consider two data sets, "Niemann" and "SF," as well as the combined data "Niemann + SF" (Table 4). Both data sets include serial genotypes of the transposon IS*6110* in *M. tuberculosis*, a frequently used marker in tuberculosis molecular epidemiology.

The "Niemann" data set (Niemann et al., 1999) includes 56 time intervals from patients in Germany. For each interval, genotyping occurred both at its beginning and at its end. Each interval corresponds to a different patient, each of whom was infected with multiple-drug resistant *M. tuberculosis*. Initial IS*6110* copy numbers range from 3 to 17. Interval lengths range from 1 to 772 days, and 34 of 56 intervals were longer than 100 days. Histograms of the initial copy number distribution and the time interval distribution are shown in Fig. 1 of Niemann et al. (1999). In five of 56 intervals, changes were observed. Four patients gained one copy of IS*6110*, and one patient lost a copy. It is unclear whether Niemann et al. found shifts and did not report them, searched for shifts and did not find them, or simply did not assay for shifts.

The "SF" data set (Fig. 4) derives from a database of IS*6110* genotypes of *M. tuberculosis* patients from San Francisco. New genotypes are continually added to the database, maintained by the laboratory of Peter Small, and findings are periodically reported (Small et al., 1994; Yeh et al., 1998; Jasmer et al., 1999; Rhee et al., 1999).

As of April 2001, the database included genotypes for 320 independent time intervals from 260 patients. Because the time of genotyping was recorded by date and not by time of day, we discarded intervals for which the start and end were on the same day, leaving 291 intervals for 237 patients. During 44 of these intervals, a "complex" change occurred, involving multiple copies of the element. Many of these changes are likely attributable to reinfection, that is, replacement of an individual's strain with a newly contracted alternate strain; others might represent within-host polymorphisms. We discarded these intervals, only considering the 247 intervals of paired isolates from 204 patients in which initial and final genotypes were identical or differed only by a single birth, shift, or death. Intervals were distributed among patients as follows: 172 patients were typed over one interval, 26 over two intervals, four over three intervals, and one patient for each of five and six intervals.

The combined "Niemann + SF" data set includes 303 intervals. Owing to ambiguity about shifts in the Niemann data, inferences about shifts for the combined data should be interpreted cautiously. The individuals in the Niemann data had drug resistance, while the SF patients included a mixture of drug-resistant and drug-susceptible cases. However, Niemann et al. (1999) did not observe differences in copy number distribution or change rates that related to drug resistance.

8. Simulations

For the data in Section 7, sampling generally occurs often enough that each interval contains no more than one genotype change. Of the patients in the Niemann data set that experienced change, all had either one more or one less transposon copy (Table 4), suggesting that no patient experienced more than one change in copy number over the sampling interval. Thus, for the marker IS*6110*, it seems acceptable to assume that at most one event occurs during an interval, and to use the "frequent monitoring" theory of Sections 5.2 and 6.2.

We tested this argument by simulating birth–shift–death processes, using various distributions for initial copy numbers and sampling intervals, with a range of birth, shift, and death rates, and a range of sample sizes. In each simulation, for each individual, initial copy number and sampling interval were simulated independently. Events were simulated according to the linear birth–shift–death process in Section 2. Times at which events occurred and the types of these events were recorded.

For each simulated data set, maximum-likelihood parameter estimates were obtained under the assumption of continuous monitoring (5), and under the assumption of frequent monitoring (7). An underlying linear birth–shift–death process was assumed, so that likelihoods were computed using cases 1 and 12 in Table 1. Under continuous monitoring, exact times and types of events were taken into account to compute the estimates. With frequent monitoring, only the types of events were used; in case more than one event occurred during the sampling interval, events subsequent to the first were ignored when estimates were obtained. If the copy number for an individual reached zero, the individual was not allowed to undergo additional genotype changes.

Initial copy number and sampling interval distributions were taken from the Niemann + SF data set. For change rates that might be expected, we found that the frequent monitoring assumption was quite reasonable.

Fig. 5A shows that the birth rate estimator under continuous monitoring (4) is very accurate. Because events are rare enough that relatively few individuals experience multiple changes, the maximum-likelihood estimator under the frequent monitoring assumption also performs well (Fig. 5B). The continuous monitoring estimator is slightly more accurate, and has smaller variance, than the frequent monitoring estimator. Both estimators have larger variances as change rates increase.

Similarly, Fig. 6A and B demonstrate that both estimators are fairly accurate for small sample sizes. As with Fig. 5, comparison of Fig. 6A and B indicates that the continuous monitoring estimator gives values that are both closer to the true value and more precise than those produced by the frequent monitoring estimator. Lastly, Fig. 7A and B demonstrate that for time intervals that might be characteristic of epidemiological studies, estimates under frequent monitoring are close to the true value, and have reasonably small variance.

Thus, for analysis of the IS*6110* data, it seems acceptable to assume that the birth–shift–death process is monitored frequently. However, this assumption will not apply to markers with rates and sampling intervals large enough that multiple events happen per interval.

9. Estimation of change rate of IS6110

For these three data sets, we estimated rates assuming frequent monitoring (Table 5). The likelihood we maximized was (7), employing case 12 in Table 1 (case 13 for the Niemann data, which had no shifts). We also computed standard normal 95% confidence intervals for each parameter, treating estimates of the different parameters as independent (Table 5). To obtain confidence intervals, variances were computed by inverting the matrix in (17), employing likelihood functions from Table 1 in calculating the information

matrix. The resulting confidence set is a cube; because estimates of birth, shift, and death rates for IS*6110* are

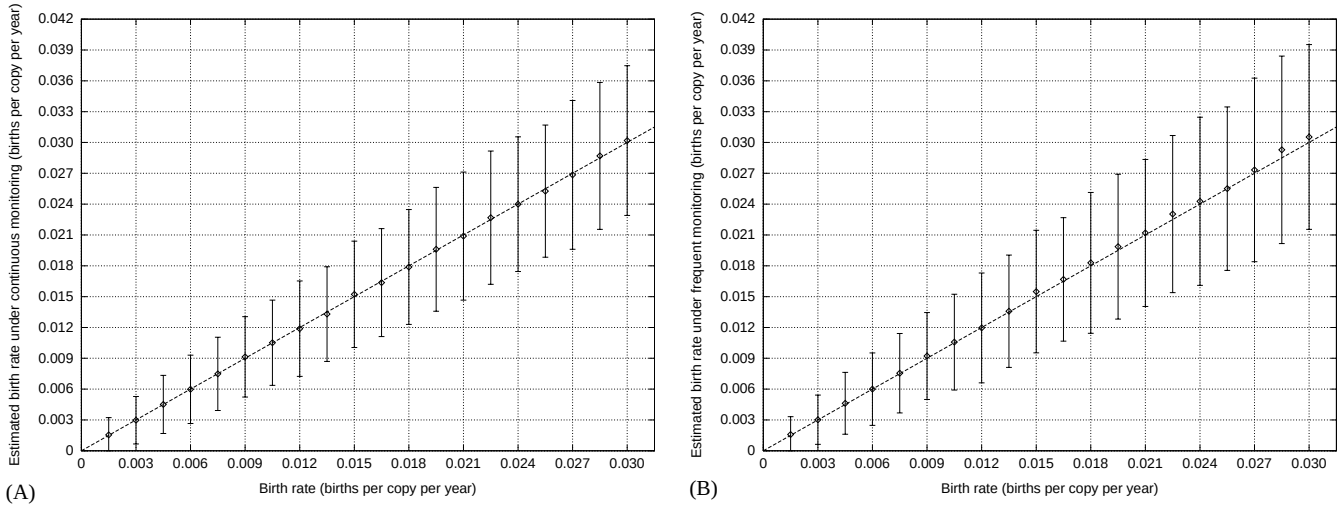

Fig. 5. Mean estimated birth rates as a function of true birth rate in replicate simulations. (A) Rate estimated assuming continuous monitoring. (B) Rate estimated assuming frequent monitoring. Each point is based on 1000 replicate simulations using a sample of size 150. Initial copy numbers and sampling intervals were simulated based on the empirical distributions in the Niemann + SF data set. Sampling intervals were simulated independently of initial copy numbers. Birth, shift, and death rates were assumed equal. Error bars denote standard deviations of the distribution of 1000 estimates.

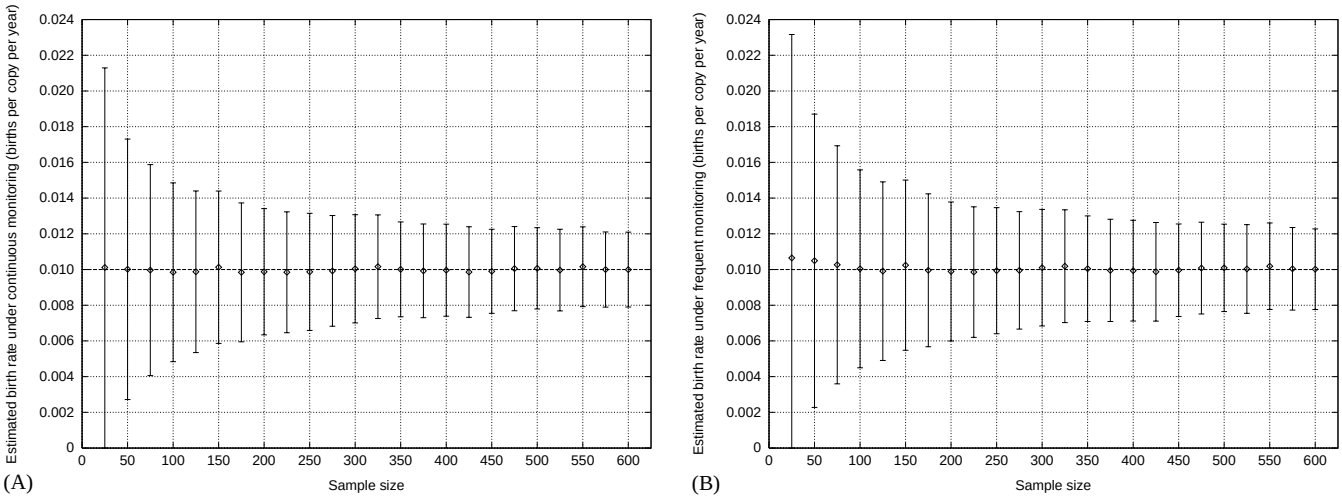

Fig. 6. Mean estimated birth rates as a function of sample size in replicate simulations. (A) Rate estimated assuming continuous monitoring. (B) Rate estimated assuming frequent monitoring. Each point is based on 1000 replicate simulations using birth, shift, and death rates equal to 0.01 events per copy per year. Initial copy numbers and sampling intervals were simulated based on the empirical distributions in the Niemann + SF data set. Sampling intervals were simulated independently of initial copy numbers. Error bars denote standard deviations of the distribution of 1000 estimates.

estimates are only weakly dependent, however, this set is only slightly conservative. To compute the overall change rate and its confidence intervals, we used the one-parameter model of case 14.

Because sampling intervals were measured in days (Fig. 4), estimates were obtained first in events per transposon copy per day. If replacement of the dominant strain within a patient occurs exclusively through genetic drift, then as discussed in Section 2, in fact estimates of corresponding rates at the cellular level. Converting our Niemann + SF estimates from events per element per day to events per element per generation, using the 1 day generation time of *M. tuberculosis*, cellular birth, shift, and death rates are 4.42×10^{-5}, 4.91×10^{-6}, and 2.94×10^{-5}, respectively (top of Table 5). These per-generation estimates should be regarded with caution, as they depend heavily on the assumption of selective neutrality.

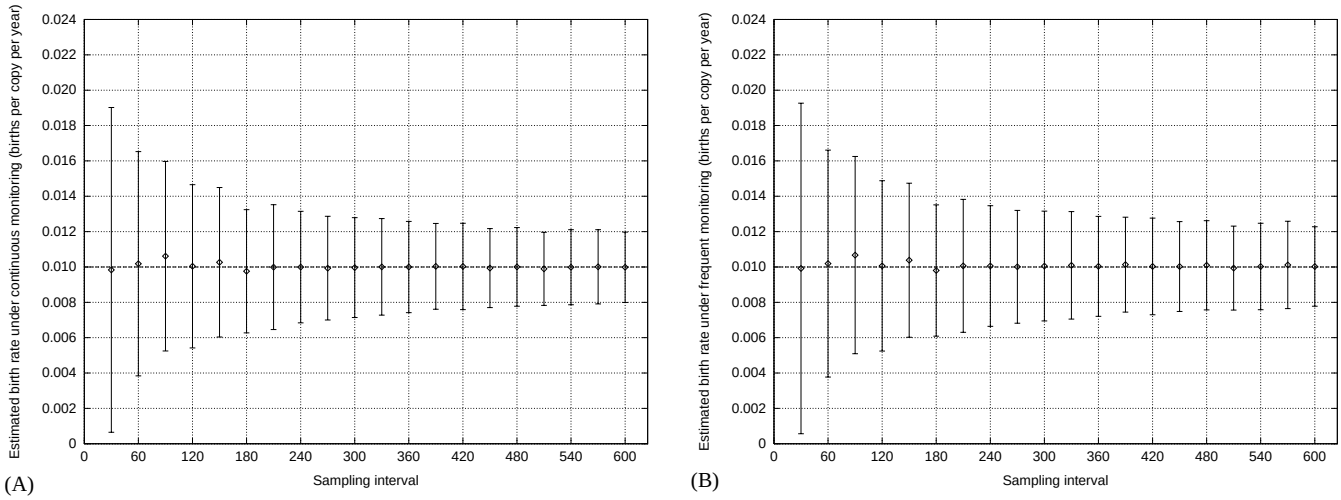

Fig. 7. Mean estimated birth rates as a function of sampling interval in replicate simulations. (A) Rate estimated assuming continuous monitoring. (B) Rate estimated assuming frequent monitoring. Each point is based on 1000 replicate simulations using birth, shift, and death rates equal to 0.01 events per copy per year. Initial copy numbers were simulated based on the empirical distribution in the Niemann + SF data set. The sample size was fixed at 150. Error bars denote standard deviations of the distribution of 1000 estimates.

Table 5
Estimated birth, shift, and death rates

	Niemann		SF		Niemann + SF	
Events per copy per generation						
Birth rate	2.95×10^{-5}	$(6.13 \times 10^{-7}, 5.84 \times 10^{-5})$	5.14×10^{-5}	$(2.34 \times 10^{-5}, 7.94 \times 10^{-5})$	4.42×10^{-5}	$(2.31 \times 10^{-5}, 6.52 \times 10^{-5})$
Shift rate	—		7.34×10^{-6}	$[0, 1.79 \times 10^{-5})$	4.91×10^{-6}	$[0, 1.19 \times 10^{-5})$
Death rate	7.38×10^{-6}	$[0, 2.18 \times 10^{-5})$	4.04×10^{-5}	$(1.56 \times 10^{-5}, 6.52 \times 10^{-5})$	2.94×10^{-5}	$(1.23 \times 10^{-5}, 4.66 \times 10^{-5})$
Overall rate	3.69×10^{-5}	$(4.57 \times 10^{-6}, 6.92 \times 10^{-5})$	9.91×10^{-5}	$(6.01 \times 10^{-5}, 1.38 \times 10^{-4})$	7.85×10^{-5}	$(5.03 \times 10^{-5}, 1.07 \times 10^{-4})$
Events per copy per year						
Birth rate	0.0108	$(2.239 \times 10^{-4}, 0.0213)$	0.0188	(0.0085, 0.0290)	0.0161	$(8.40 \times 10^{-3}, 0.0238)$
Shift rate	—		2.68×10^{-3}	$[0, 6.54 \times 10^{-3})$	1.79×10^{-3}	$[0, 4.35 \times 10^{-3})$
Death rate	2.70×10^{-3}	$[0, 7.97 \times 10^{-3})$	0.0147	$(5.68 \times 10^{-3}, 0.0238)$	0.0108	$(4.47 \times 10^{-3}, 0.0170)$
Overall rate	0.0135	$(1.67 \times 10^{-3}, 0.0252)$	0.0362	(0.0219, 0.0504)	0.0287	(0.0184, 0.0390)
Half-life of one-copy strain in years (*divide by k for k-copy strain*)						
Birth process	64.3	(32.5, 3095.4)	36.9	(23.9, 81.2)	43.0	(29.1, 82.2)
Shift process	—		258.6	$(106.1, \infty)$	386.8	$(159.3, \infty)$
Death process	257.1	$(87.0, \infty)$	47.0	(29.1, 122.0)	64.5	(40.7, 154.9)
Overall process	51.4	(27.4, 415.4)	19.2	(13.7, 31.6)	24.2	(17.8, 37.7)

95% confidence intervals around estimates are shown in parentheses. Negative numbers were rounded to zero. Confidence limits for half-life units were obtained by transforming corresponding limits for the other units according to (18). In the Niemann data set, no shifts were observed and the shift rate was not estimated.

The conversion of 1 year per 365.25 days enables conversion into units of events per copy per year for rates of genotype change (middle of Table 5). Finally, if viewed as the decay of a configuration of elements in the genome, estimates can be converted into "half-life" units (de Boer et al., 1999). Because we assume exponential waiting times until "decay," the conversion is straightforward. For an exponential process with rate σ, the half-life is

$$t_{1/2} = \sigma^{-1} \ln 2. \tag{18}$$

Under the linear birth–shift–death process, waiting times until births, shifts, and deaths are exponentially distributed, as is the waiting time until the first event of any type. For a k-copy strain the half-life of a pattern by the birth process, for example, is obtained by inserting $k\hat{\lambda}$ in place of σ in (18). Because "decay" by shift or death may precede the first decay by the birth process, the half-life for decay by any of the individual processes is more difficult to interpret than the half-life for the overall process. This overall half-life is obtained by inserting the overall estimated

rate $k(\hat{\lambda}+\hat{\gamma}+\hat{\mu})$ in place of σ in (18). The Niemann + SF half-life estimate of 24.2 years for a one-copy strain corresponds to 2.4 years when divided by a typical copy number of 10.

The estimated shift rate was significantly smaller than both the birth rate and the death rate (95% confidence intervals did not overlap). In both the Niemann and SF data sets, the estimated birth rate was larger than the estimated death rate, though not significantly. The estimated overall change rate for the Niemann data set was 0.0135, as was obtained previously (Tanaka and Rosenberg, 2001). The overall change rate was higher for the SF data, equaling 0.0362, and it was 0.0287 for the combined data. This rate would have been even larger had "complex changes" been included; in addition to birth, shift, and death parameters, the overall rate would then have included a component for a rate of complex changes.

10. Discussion

10.1. Markers for molecular epidemiology

Pathogen genotypes at marker loci and their frequencies assist in studying transmission processes and biological causes of disease. Of crucial importance to inference of epidemiologic relationships from genetic relationships is knowledge of rates of genetic change of the pathogen (Yeh et al., 1998; Tanaka and Rosenberg, 2001, for example). A marker that changes rapidly compared to the time scale of an epidemic will obscure links in transmission chains, producing divergent genotypes among causally related cases. On the other hand, a slowly changing marker may provide little information about recent transmission, as marker genotypes may be similar solely because of distant common ancestry.

Various epidemiological settings might apply the following protocol: (a) determine the time scale on which a phenomenon of interest is occurring; (b) select genetic markers whose mutation rates are appropriate to the time scale of the phenomenon; (c) genotype individuals; (d) make inferences about the phenomenon.

The presence of step (b) makes it useful to accurately estimate the rate of change. As the relationship between the time scales of epidemiological phenomena and the informativeness of markers has been little explored, it will be important to quantitatively determine how change rates affect epidemiological inference. An approach to this problem might consider the proportion of correct inferences made about epidemiological clusters based on clustering of similar genotypes, assuming both an epidemic transmission model and a genetic change model.

10.2. Optimum experimental design

Because it is desirable to maximize precision estimates of genotype change rates, investigators may make several decisions in advance that will lead to reduced variance in estimates. These include (a) increasing the number of time points at which pathogens are isolated from each individual during a fixed time interval; (b) optimally selecting lengths of the intervals between these time points; (c) increasing the sample size; (d) using only individuals whose pathogens initially have a high copy number.

As is suggested in (a), it is more informative to observe pathogen genotypes for each individual at more than two points in time. One way to view this increase is that the monitoring scheme more closely approximates continuous monitoring scheme as the number of intervals is increased. In simulations, however, precision was only slightly improved when the exact timing of events was known (Figs. 5A, 6A and 7A), compared to when it was only known that an event had occurred inside the interval (Figs. 5B, 6B and 7B). Thus, subdividing a fixed time interval is not likely to greatly increase precision. Note that if a fixed time interval is subdivided, the optimal division depends on change rates, and is not in general the one that produces subintervals of equal size (Becker and Kersting, 1983, for example). For example, if the birth rate exceeds the death rate, then more changes occur later in the interval and the optimal division uses smaller subintervals later in the fixed interval.

As was studied by Tanaka and Rosenberg (2001), the optimal choice of interval lengths (b) is a concern if change resolution and infrequent monitoring are used. If the interval is long, then all individuals will change during the interval, and it will not be possible to estimate the rate. If the interval is small, then no individuals will change and the estimate will be zero. Thus, an optimal intermediate length must exist. For frequent monitoring, however, intervals are by definition short enough that at most one event usually occurs per interval. The variance minimization approach in Tanaka and Rosenberg (2001) is thus inappropriate in the frequent case.

Increasing the sample size (c) decreases the variance in the usual manner, assuming that there are identical distributions of copy number and sampling intervals for newly sampled and previously sampled individuals. As subdividing intervals is inefficient and optimal interval choice might require impractically large sampling intervals, this option might be the simplest way to increase the precision of estimates. Fig. 6 shows how the estimates become more precise as sample size is increased, so that precision might be chosen in advance and an appropriate sample size obtained by comparison with these or similar graphs.

If high-copy strains do indeed change more rapidly than low-copy strains, high-copy strains provide more information about unknown rates during a fixed length of time. Thus, it may be useful to genotype initial samples soon after they have been obtained, rather than waiting until the end of the sampling interval to genotype initial and final samples together. Hosts with low-copy strains might be eliminated early in studies, and sampling intervals might be chosen to be shorter for high-copy strains. Of course, these decisions are not recommended when the goal is to study population dynamics of the marker (Tanaka et al., 2000, for example) rather than to estimate rates: if high copy numbers are preferentially included, samples will no longer be random.

Precision in the estimates must be balanced by additional costs of genotyping more individuals, genotyping the same individuals at additional time points, and the additional time it will take before an estimate can be obtained. Simulations such as those in Section 8 enable precision of estimates to be predicted in advance of studies, so that the desired precision can be taken into account when designing a study. For optimal design, because there are multiple parameters, the criterion of obtaining the minimum variance estimator as in Tanaka and Rosenberg (2001) must be generalized to minimizing a function of the variances and covariances of the estimates of the parameters (Atkinson and Donev, 1992). Caution is warranted in choosing optimal designs based on the linear birth–shift–death process model, a model that might be superseded upon direct tests of its accuracy.

It is often of interest to have a rough estimate of genotype change rates before performing a longer study. As we have seen, actual rates of the process affect the way in which those very rates should be estimated. Thus, it is useful to obtain preliminary estimates. For this purpose, with similar markers to IS*6110*, Figs. 5–7 suggest that reasonable precision can be obtained using at least 100 intervals of at least 60–90 days. This type of pilot study will be unreliable in case the assumption of time-homogeneity of the process is violated. The degree to which time-homogeneity holds will depend on the within-host process; thus, testing for time-homogeneity of the change rate should provide insight into temporal variation in selection inside the host environment.

In general, we find it acceptable to assume frequent monitoring. If many intervals are sampled for each host, it is reasonable to assume that the change happened at the sampling point and to use continuous monitoring.

10.3. Other types of markers

The modularity of our approach enables applications to other types of markers. In place of the linear birth–shift–death process model for transposons, appropriate genotype evolution models might be inserted for markers of interest. For DNA sequences, nucleotide substitution models might be used (Jukes and Cantor, 1969; Li, 1997, for example). The probability of an observed DNA sequence at time t given the initial DNA sequence and the sampling interval is then easily computed.

A more closely related application is to pulsed-field gel electrophoresis (PFGE) markers. In PFGE, the pathogen genome is digested with a restriction enzyme, and the pathogen's genotype is the resulting pattern of genomic fragment lengths on a gel (Tenover et al., 1995). The underlying processes that cause genotype profiles to change are mutations that produce a new restriction site, and a consequent new band on the gel (births); random insertions and deletions that do not affect restriction sites, consequently changing the size and hence position of a single band (shifts); and mutations that excise restriction sites or change them so that they are no longer restriction sites (deaths). The production rate of restriction sites should be a genomic constant, independent of the number of existing sites. Similarly, as shifts occur any time a major insertion or deletion happens anywhere in the genome, the shift rate should also be a genomic constant. The death rate, however, should be linear in the number of restriction sites, as sites decay independently. Thus, a reasonable model for PFGE is a linear death process with constant birth and shift rates.

The same detection schemes as those described in Section 3 apply. Complete resolution provides the whole band pattern, including births (one band is replaced by two bands representing smaller fragments), shifts (one band is moved), and deaths (two bands are replaced by one band representing a larger fragment). For PFGE, the estimation theory here could be adapted by replacing the linear birth–shift–death process model with a linear death process model including constant births and shifts. If deaths only are considered, models here and in Tanaka and Rosenberg (2001) directly apply to PFGE.

10.4. IS6110 and M. tuberculosis

We have assumed that each element has the same rate of change. If we were to consider the positions of the elements, it could be tested if certain regions of the *M. tuberculosis* genome enable more rapid transposition than other regions. Unlike many other transposons, IS*6110* does not include a promoter region and must be located near transcription initiation sites in the genome in order to transpose. Thus, the transposition rate of a given copy surely depends on its proximity to promoters.

Even if transposition rates across the genome are homogeneous, a nonlinear model might more accurately describe the process, so that change rates are not linear

in copy number. Under such a model, the birth, shift and death rates would be the corresponding rates for strains with exactly a single copy of the element, and additional parameters would describe departures from the independent action of separate copies of the transposon.

The SF data produce change rate estimates rates that are higher than those based on Niemann et al. (1999). One possible cause of this discrepancy might be inhomogeneity of the process in time. The Niemann data set generally has long intervals and the SF data set has many short intervals in which changes are observed. The Niemann data set may represent populations at a later stage of infection, during which changes might occur more slowly (Warren et al., 2002). Alternatively, there could be strain heterogeneity (van Soolingen et al., 1999; Tanaka et al., 2000): the strains of the Niemann data set may have less active transposases compared to the SF strains. Systematic biases in laboratory procedures may also be responsible, or changes in genotyping procedures that have occurred over time. Comparisons of transposition rates as a function of neutral mutation rates in different strains (or whether strains are "mutators"), geographic origins of strains, positions of elements, genetic susceptibilities of patients, and clinical variables such as whether patients also have AIDS may be of interest.

Given the high change rates inferred for IS*6110* compared to previous studies (de Boer et al., 1999; Tanaka and Rosenberg, 2001; Warren et al., 2002), researchers who use IS*6110* genotypes to classify individuals into epidemiological clusters should be careful not to underestimate the composition of these groups. In some cases, transmission among patients may take long enough that the marker will have changed in the intervening period, so that *M. tuberculosis* patients with similar but nonidentical IS*6110* patterns may be causally connected.

Finally, we have ignored complex changes because of uncertainty about their meaning. This has allowed us to ignore within-host polymorphism and to justify the equality of cellular transposition rates and substitution rates within hosts. If complex changes could be distinguished into within-host polymorphisms, reinfection with alternate strains, and coinfection with multiple strains of different origins, a more sophisticated coalescent-based model that accounted for within-host polymorphism would be more appropriate. Such a model would better accommodate the selective sweeps that likely occur within hosts, and that have been ignored here by assuming all variants are neutral.

10.5. Conclusions

We have proposed a general framework for estimation of genotype change rates in pathogens, and we have developed the method for transposable element markers. Using serial samples, we have estimated the birth, shift, and death rates for IS*6110* in *M. tuberculosis*, and we have discussed how better estimates of the parameters might be obtained. Extensions to other markers, including pulsed-field gel electrophoresis, are also possible. Future work might test for nonlinearity of change rates as functions of copy number, study heterogeneity of change rates in time, and identify determinants of variation of estimated rates across data sets.

Acknowledgments

Elizabeth Fair and Melvin Javonillo provided assistance with the SF data set. We thank Marcus Feldman, Margaret Hannan, Dmitri Petrov, David Siegmund, and Peter Small for discussions, and Simon Tavaré for comments on an earlier version of the manuscript. NAR was supported by an NSF Postdoctoral Fellowship in Biological Informatics. The research was supported by NIH Grants (AI 34238, GM28428, and GM 58897).

References

Atkinson, A.C., Donev, A.N., 1992. Optimum Experimental Designs. Clarendon Press, Oxford.

Basawa, I.V., Prakasa Rao, B.L.S., 1980. Statistical Inference for Stochastic Processes. Academic Press, London.

Becker, G., Kersting, G., 1983. Design problems for the pure birth process. Adv. Appl. Probab. 15, 255–273.

Darwin, J.H., 1956. The behaviour of an estimator for a simple birth and death process. Biometrika 43, 23–31.

de Boer, A.S., Borgdorff, M.W., de Haas, P.E.W., Nagelkerke, N.J.D., van Embden, J.D.A., van Soolingen, D., 1999. Analysis of rate of change of IS*6110* RFLP patterns of *Mycobacterium tuberculosis* based on serial patient isolates. J. Infect. Dis. 180, 1238–1244.

de Boer, A.S., Kremer, K., Borgdorff, M.W., de Haas, P.E.W., Heersma, H.F., van Soolingen, D., 2000. Genetic heterogeneity in *Mycobacterium tuberculosis* isolates reflected in IS*6110* restriction fragment length polymorphism patterns as low-intensity bands. J. Clin. Microbiol. 38, 4478–4484.

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., Solomon, W., 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics 161, 1307–1320.

Elandt-Johnson, R.C., 1971. Probability Models and Statistical Methods in Genetics. Wiley, New York.

Fu, Y.-X., 2001. Estimating mutation rate and generation time from longitudinal samples of DNA sequences. Mol. Biol. Evol. 18, 620–626.

Golinelli, D., 2000. Bayesian inference in hidden stochastic population processes. Ph.D. Thesis, University of Washington, Seattle.

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E., de Knijff, P., 1997. Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum. Mol. Genet. 6, 799–803.

Jasmer, R.M., Hahn, J.A., Small, P.M., Daley, C.L., Behr, M.A., Moss, A.R., Creasman, J.M., Schecter, G.F., Paz, E.A., Hopewell,

P.C., 1999. A molecular epidemiologic analysis of tuberculosis trends in San Francisco, 1991–1997. Ann. Intern. Med. 130, 971–978.

Jukes, T.H., Cantor, C.R., 1969. Evolution of protein molecules. In: Munro, H.N. (Ed.), Mammalian Protein Metabolism. Academic Press, New York, pp. 21–132.

Keiding, N., 1974. Estimation in the birth process. Biometrika 61, 71–80.

Keiding, N., 1975. Maximum likelihood estimation in the birth-and-death process. Ann. Stat. 3, 363–372.

Kelly, F.P., 1979. Reversibility and Stochastic Networks. Wiley, Chichester, UK.

Kendall, D.G., 1949. Stochastic processes and population growth. J. R. Statist. Soc. Ser. B 11, 230–264.

Kepler, T.B., Oprea, M., 2001. Improved inference of mutation rates: I. An integral representation for the Luria–Delbrück distribution. Theor. Popul. Biol. 59, 41–48.

Kimura, M., 1968. Evolutionary rate at the molecular level. Nature 217, 624–626.

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

Lambert, D.M., Ritchie, P.A., Millar, C.D., Holland, B., Drummond, A.J., Baroni, C., 2002. Rates of evolution in ancient DNA from Adélie penguins. Science 295, 2270–2273.

Li, W.-H., 1997. Molecular Evolution. Sinauer, Sunderland, MA.

Luria, S., Delbrück, M., 1943. Mutations of bacteria from virus sensitivity to virus resistance. Genetics 28, 491–511.

Mukai, T., Cockerham, C.C., 1977. Spontaneous mutation rates at enzyme loci in *Drosophila melanogaster*. Proc. Natl Acad. Sci. USA 74, 2514–2517.

Niemann, S., Richter, E., Rüsch-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.

Puri, P.S., 1968. Some further results on the birth-and-death process and its integral. Proc. Cambridge Philos. Soc. 64, 141–154.

Reynolds, J.F., 1973. On estimating the parameters of a birth–death process. Aust. J. Stat. 15, 35–43.

Rhee, J.T., Piatek, A.S., Small, P.M., Harris, L.M., Chaparro, S.V., Kramer, F.R., Alland, D., 1999. Molecular epidemiologic evaluation of transmissibility and virulence of *Mycobacterium tuberculosis*. J. Clin. Microbiol. 37, 1764–1770.

Rodrigo, A.G., Felsenstein, J., 1999. Coalescent approaches to HIV population genetics. In: Crandall, K.A. (Ed.), The Evolution of HIV. Johns Hopkins University Press, Baltimore, pp. 233–272.

Rosenberg, N.A., Nordborg, M., 2002. Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3, 380–390.

Sawyer, S., Hartl, D., 1986. Distribution of transposable elements in prokaryotes. Theor. Popul. Biol. 30, 1–16.

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.

Tanaka, M.M., Rosenberg, N.A., 2001. Optimal estimation of transposition rates of insertion sequences for molecular epidemiology. Stat. Med. 20, 2409–2420.

Tanaka, M.M., Small, P.M., Salamon, H., Feldman, M.W., 2000. The dynamics of repeated elements: applications to the epidemiology of tuberculosis. Proc. Natl Acad. Sci. USA 97, 3532–3537.

Tenover, F.C., Arbeit, R.D., Goering, R.V., Mickelsen, P.A., Murray, B.E., Persing, D.H., Swaminathan, B., 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. 33, 2233–2239.

van Soolingen, D., de Boer, A.S., Alito, A., Morcillo, N., 1999. Authors' reply. J. Clin. Microbiol. 37, 3078–3079.

Warren, R.M., van der Spuy, G.D., Richardson, M., Beyers, N., Borgdorff, M.W., Behr, M.A., van Helden, P.D., 2002. Calculation of the stability of the IS6110 banding pattern in patients with persistent *Mycobacterium tuberculosis* disease. J. Clin. Microbiol. 40, 1705–1708.

Watterson, G.A., 1982. Substitution times for mutant nucleotides. J. Appl. Probab. 19a, 59–70.

Yeh, R.W., Ponce de Leon, A., Agasino, C.B., Hahn, J.A., Daley, C.L., Hopewell, P.C., Small, P.M., 1998. Stability of *Mycobacterium tuberculosis* DNA genotypes. J. Infect. Dis. 177, 1107–1111.