Methods for Estimating Gene Frequencies and Detecting Selection in Bacterial Populations

Bruce Rannala, Wei-Gang Qiu and Daniel E. Dykhuizen

Department of Ecology and Evolution, State University of New York, Stony Brook, New York 11754-5245

Manuscript received September 3, 1999
Accepted for publication February 23, 2000

ABSTRACT

Recent breakthroughs in molecular technology, most significantly the polymerase chain reaction (PCR) and *in situ* hybridization, have allowed the detection of genetic variation in bacterial communities without prior cultivation. These methods often produce data in the form of the presence or absence of alleles or genotypes, however, rather than counts of alleles. Using relative allele frequencies from presence-absence data as estimates of population allele frequencies tends to underestimate the frequencies of common alleles and overestimate those of rare ones, potentially biasing the results of a test of neutrality in favor of balancing selection. In this study, a maximum-likelihood estimator (MLE) of bacterial allele frequencies designed for use with presence-absence data is derived using an explicit stochastic model of the host infection (or bacterial sampling) process. The performance of the MLE is evaluated using computer simulation and a method is presented for evaluating the fit of estimated allele frequencies to the neutral infinite alleles model (IAM). The methods are applied to estimate allele frequencies at two outer surface protein loci (*ospA* and *ospC*) of the Lyme disease spirochete, *Borrelia burgdorferi*, infecting local populations of deer ticks (*Ixodes scapularis*) and to test the fit to a neutral IAM.

EXTENSIVE variations in allele frequencies and molecular (DNA and protein) sequence polymorphisms pervade the majority of natural populations. Although genetic variation at its ultimate level can now be quickly and accurately resolved by DNA sequencing, much of our ecological and evolutionary understanding of natural populations continues to be based on the results of allele frequency analyses. Strong natural selection, for example, can distort the allele frequency distribution at a locus resulting in departures from Hardy-Weinberg equilibrium and from the Ewens sampling distribution (Ewens 1972; Manly 1985), which describes the samples of alleles expected when mutations follow an infinitely many neutral alleles model (IAM). Classical tests of neutrality such as the Ewens-Watterson test have increased our understanding of the patterns of selection among regions of the genome and among populations.

Methods of allele frequency analysis developed for use in studies of animal and plant populations are, in principle, applicable to bacterial species as well. Unlike animal or plant populations, however, it has only recently become technically feasible to perform *in situ* sampling to examine the genetic variability of bacteria in their natural (uncultivated) environment. Classical studies of bacterial population genetics (see review by Selander *et al.* 1994) have therefore focused on a few model organisms that can be readily cultured. It is well known that most (90% or more) of the genetic diversity within bacterial communities is lost through cultivation, however, and nutrient enrichment methods are highly selective for the growth of only a small number of species (or genotypes of a species; see Table 1 in Amann *et al.* 1995).

With the advent of new molecular techniques, such as PCR and *in situ* hybridization (Amann *et al.* 1995; von Wintzingerode *et al.* 1997), a previously unknown level of microbial diversity has been uncovered in environmental and clinical microflora that challenges traditional views on microbial ecology and pathogenicity (see recent reviews by Hugenholtz *et al.* 1998; Relman 1999). Using molecular typing methods it is now also possible to survey the genetic variation of a bacterial species in its natural environment. Thus far, however, formal statistical methods have been lacking for estimating allele frequencies using the kinds of data obtained in studies of microbial diversity by molecular methods. In this article, we attempt to redress this problem by developing some new statistical methods for analyzing these emerging data.

In recent population genetic studies of *Borrelia burgdorferi*, the spirochete that causes Lyme disease, Borrelia genes were directly amplified from infected ticks (*Ixodes scapularis*) using a nested PCR design (Guttman *et al.* 1996). The amplified genes were subsequently surveyed for sequence variations using cold single-strand conformation polymorphism (SSCP) and individual ticks were commonly found to be infected with multiple strains of

Corresponding author: Bruce Rannala, Department of Ecology and Evolution, State University of New York, Stony Brook, NY 11794-5245. E-mail: rannala@life.bio.sunysb.edu

Borrelia (Guttman *et al.* 1996; Qiu *et al.* 1997; Wang *et al.* 1999). In this work, the number of bands of alleles was counted directly from electrophoretic gels to estimate the frequencies of various SSCP alleles in a local Borrelia population (Qiu *et al.* 1997; Wang 1999). This method of direct counting can produce biased estimates of the population allele frequencies, however, tending to underestimate the frequencies of common alleles and to overestimate those of rare alleles. The result is an inferred frequency distribution that appears more even than the actual frequency distribution (see Methods and Materials in Qiu *et al.* 1997).

In this study, bacterial allele frequencies are estimated more accurately by deriving an explicit maximum-likelihood estimator (MLE) that takes account of the transmission process of bacteria to hosts and the host sampling process. Our approach was to formulate the sampling distribution of the presence-absence data as a function of the population allele frequencies. Allele frequencies could then be estimated directly using analytical maximum-likelihood techniques. The method should be generally applicable to problems involving the estimation of population allele frequencies from presence-absence data; such data sets are becoming increasingly common, particularly in studies of bacterial genetic diversity using molecular techniques. If one is studying gene frequencies in populations of free-living microbes, rather than parasitic ones, the model can still be used but is instead a model of the microbial substrate sampling process. For example, samples of equal amounts of soil might be collected from different regions and the microbes in each sample genetically characterized; in this case, the samples are equivalent to hosts and the number of infecting microbes is the population of bacteria in each sample.

THEORY

Estimation of allele frequencies in microbial populations: Consider a sample of n hosts infected with a particular species of parasitic microbe. Let Y_i be the number of microbes that infect the ith host, where $1 \le i \le n$. If the rate of infection of hosts by the microbe is low then we can model the distribution of the number of infecting microbes per host as Poisson with parameter λ, where this is the expected number of infecting microbes per host. If we consider only the infected hosts, the distribution of $Y_i \ge 1$ is

$$\Pr(Y_i|Y_i \ge 1, \lambda) = \frac{e^{-\lambda}\lambda^{Y_i}}{(1 - e^{-\lambda})Y_i!}. \qquad (1)$$

Let $\mathbf{p} = \{p_j\}$ be a vector of the allele frequencies in the microbial population, where p_j is the frequency of allele j, there are k alleles in total, and $\Sigma_{j=1}^{k} p_j = 1$. If host i is infected by Y_i microbes, the probability that $0 \le y_{ij} \le k$ of these carry allele j for all $1 \le j \le k$ is specified by the multinomial distribution

$$\Pr(\mathbf{y}_i|Y_i, Y_i \ge 1, \mathbf{p}) = \binom{Y_i}{y_{i1}y_{i2}\ldots y_{ik}}\prod_{j=1}^{k} p_j^{y_{ij}}, \qquad (2)$$

where $\mathbf{y}_i = \{y_{ij}\}$ is a vector of the number of copies of each allele among the Y_i microbes infecting individual i and $Y_i = \Sigma_{j=1}^{k} y_{ij}$. A well-known result that is useful to consider in the context of this problem is the genesis of a multinomial distribution as the joint distribution of the number of copies of each of k independent Poisson random variables, the jth of which has the parameter λp_j, conditioned on their sum. If y_{ij} is now the number of copies of the jth type in the ith replicate population formed by this process, the probability distribution is

$$\Pr(\mathbf{y}_i|Y_i, Y_i \ge 1, \mathbf{p}, \lambda) = \frac{Y_i!}{\lambda^{Y_i}e^{-\lambda}}\prod_{j=1}^{k}\frac{(\lambda p_j)^{y_{ij}}e^{-\lambda p_j}}{y_{ij}!}. \qquad (3)$$

Note that (3) algebraically reduces to (2) in this way:

$$\begin{aligned}\frac{Y_i!}{\lambda^{Y_i}e^{-\lambda}}\prod_{j=1}^{k}\frac{(\lambda p_j)^{y_{ij}}e^{-\lambda p_j}}{y_{ij}!} &= \left(\frac{\lambda^{\Sigma_{j=1}^{k} y_{ij}}e^{-\lambda\Sigma_{j=1}^{k} p_j}}{e^{-\lambda}\lambda^{Y_i}}\right) \\ &\quad \times \frac{Y_i!}{\prod_{j=1}^{k} y_{ij}!}\prod_{j=1}^{k} p_j^{y_{ij}} \\ &= \binom{Y_i}{y_{i1}y_{i2}\ldots y_{ik}}\prod_{j=1}^{k} p_j^{y_{ij}}.\end{aligned}$$

The joint probability distribution of $\mathbf{y}_i$ and Y_i is then

$$\begin{aligned}\Pr(\mathbf{y}_i, Y_i|\mathbf{p}, \lambda, Y_i \ge 1) &= \Pr(\mathbf{y}_i|Y_i, \mathbf{p}) \times \Pr(Y_i|Y_i \ge 1) \\ &= \frac{1}{(1 - e^{-\lambda})}\prod_{j=1}^{k}\frac{(\lambda p_j)^{y_{ij}}e^{-\lambda p_j}}{y_{ij}!}. \qquad (4)\end{aligned}$$

If the number of microbes of each allelic type that infect the ith host was directly observed, (4) would be the likelihood of the observations and the MLEs of the allele frequencies among microbes would be just the frequencies observed in the hosts. In most cases, however, the observations are actually the numbers of hosts infected with one or more microbes of each allelic type. It is then natural to represent the observations on the ith infected host as the binary vector $\mathbf{X}_i = \{X_{i1}, X_{i2}, \ldots, X_{ik}\}$, where X_{ij} indicates whether the jth microbial allele is observed in the sample of microbes from the ith infected host. Accordingly, we define

$$X_{ij} = \begin{cases} 1 & \text{if allele } j \text{ is present} \\ 0 & \text{otherwise.}\end{cases}$$

The observations are then given by the matrix $\mathbf{X} = \{X_{ij}\}$. The probability that one or more microbes of allelic type j infect the ith host is

$$\Pr(y_{ij} \ge 1) = \sum_{l=1}^{\infty}\frac{(\lambda p_j)^l e^{-\lambda p_j}}{l!} = (1 - e^{-\lambda p_j}), \qquad (5)$$

and the probability of observing no microbes of allelic

type j is $e^{-\lambda p_j}$. Because Equation 4 is a product of independent Poisson random variables, the probability of $\mathbf{X}_i$ is

$$\Pr(\mathbf{X}_i|\lambda, \mathbf{p}) = \frac{1}{(1 - e^{-\lambda})}\prod_{j=1}^{k}\{X_{ij}(1 - e^{-\lambda p_j}) + (1 - X_{ij})e^{-\lambda p_j}\}. \tag{6}$$

Each infected host is an independent observation from this process and the probability of $\mathbf{X}$ is then

$$\Pr(\mathbf{X}|\lambda, \mathbf{p}) = \frac{1}{(1 - e^{-\lambda})^n}\prod_{i=1}^{n}\prod_{j=1}^{k}\{X_{ij}(1 - e^{-\lambda p_j}) + (1 - X_{ij})e^{-\lambda p_j}\}. \tag{7}$$

From (7) we obtain the log-likelihood of the observed data as

$$\ell = -n\log(1 - e^{-\lambda}) \sum_{i=1}^{n}\sum_{j=1}^{k}\log[X_{ij}(1 - e^{-\lambda p_j}) + (1 - X_{ij})e^{-\lambda p_j}]. \tag{8}$$

The log-likelihood then simplifies to

$$\ell = -n\log(1 - e^{-\lambda}) + \sum_{j=1}^{k}\{z_j\log(1 - e^{-\lambda p_j}) - (n - z_j)\lambda p_j\}, \tag{9}$$

where z_j is the number of hosts sampled for which the microbes display allele j. To maximize the likelihood with respect to parameter p_j, we differentiate the log-likelihood, set this partial derivative to equal zero, and solve for p_j. The derivative of l taken with respect to p_j is

$$\frac{\partial\ell}{\partial p_j} = \frac{z_j\lambda e^{-\lambda p_j}}{1 - e^{-\lambda p_j}} - (n - z_j)\lambda. \tag{10}$$

Setting Equation (10) equal to zero and solving for p_j gives the MLE,

$$\hat{p}_j = -\frac{1}{\lambda}\log\left(\frac{n - z_j}{n}\right). \tag{11}$$

Because the allele frequencies are constrained to sum to 1 the MLE of λ is

$$\hat{\lambda} = -\sum_{j=1}^{k}\log\left(\frac{n - z_j}{n}\right). \tag{12}$$

The estimator of p_j given by Equation 11 does not take account of the uncertainty in the parameter λ. In most cases, λ is unknown and is estimated from the data using Equation 12. A more rigorous statistical approach would be to specify a prior probability density for λ and to then integrate over this prior. The marginal likelihood could then be used to estimate p_j. Alternatively, one could maximize the likelihood of the parameters $\mathbf{p}$ and λ jointly.

It is informative to consider the mathematical properties of Equation 11 for alleles in high, low, or uniform frequency. If we let $q_i = z_i/n$ be the frequency of hosts infected with microbes carrying allele i, then we can write the MLE of p_i as

$$\hat{p}_i = \frac{\log(1 - q_i)}{\sum_{j=1}^{k}\log(1 - q_j)}.$$

Applying a Taylor series expansion to represent logarithms in the numerator and denominator as polynomials, we can rewrite the above equation as

$$\hat{p}_i = \frac{q_i + \tfrac{1}{2}q_i^2 + \tfrac{1}{3}q_i^3 + \cdots}{1 + \tfrac{1}{2}\sum_{j=1}^{k} q_j^2 + \tfrac{1}{3}\sum_{j=1}^{k} q_j^3 + \cdots}.$$

One can see from this representation that, if the sample (presence-absence) allele frequencies are perfectly uniform so that $q_i = 1/k$, then the above equation reduces to $\hat{p}_i = q_i = 1/k$ and the sample (presence-absence) allele frequencies are MLEs of the population allele frequencies. In other words, the correction provided by the MLE will have little effect when population allele frequencies are very uniform. If we let $S = \tfrac{1}{2}\sum_{j=1}^{k}q_j^2 + \tfrac{1}{3}\sum_{j=1}^{k} q_j^3 + \ldots$, then for alleles in high frequency,

$$\hat{p}_i \approx \frac{q_i + S}{1 + S},$$

which is strictly greater than the uncorrected estimate q_i if $q_i < 1$, and for alleles in low frequency,

$$\hat{p}_i \approx \frac{q_i}{1 + S},$$

which is strictly less than than the uncorrected estimate q_i if $q_i > 0$. The MLEs will then down-weight population allele frequency estimates for alleles that are in low frequency in the sample and up-weight the estimates for those alleles in high frequency as intuition suggests they should.

Test of the neutral infinite allele model: Ewens (1972) showed that, under the infinitely many neutral alleles model of mutation and drift, the observed number of alleles, k, in a sample of size ξ from a haploid population of effective size N_e and with mutation rate μ is a sufficient statistic for estimating the parameter $\theta = 2N_e\mu$. Ewens provided the following implicit formula, which can be solved numerically to estimate θ for given values of k and ξ (*i.e.*, his Equation 5):

$$k = \frac{\hat{\theta}}{\hat{\theta}} + \frac{\hat{\theta}}{\hat{\theta} + 1} + \cdots + \frac{\hat{\theta}}{\hat{\theta} + 2\xi - 1}. \tag{13}$$

In the case of the bacterial samples described above, however, ξ, which is the total number of bacteria sampled, is unknown because the numbers of bacteria infecting each tick are unobserved random variables. The total sample size for the samples of bacteria from ticks can be written as

$$\xi = \sum_{j=1}^{n} Y_i. \tag{14}$$

Since the number of bacteria infecting each tick is assumed to follow an independent Poisson process with common infection rate parameter λ, the distribution of

the above convolution is also Poisson but with parameter $n\lambda$. If we estimate $\hat{\lambda}$ using the techniques outlined above, an approximate estimate of θ can be obtained by substituting $\hat{\lambda}n$ in place of ξ in (13) and solving for θ.

Using this estimate of θ, we can compare the expected allele frequencies under the neutral IAM with the MLEs of the allele frequencies for the bacteria to examine the fit of these data to the neutral IAM. The population frequency distribution of alleles is given by the Poisson-Dirichlet distribution (see Griffiths 1979). The marginal expectation of the frequency $\alpha_{(r)}$ of the rth most common allele is

$$E(\alpha_{(r)}) = \int_0^\infty \frac{(\theta E_1(y))^{r-1}}{(r-1)!} \exp\{-(y + \theta E_1(y))\} dy, \qquad (15)$$

where

$$E_1(y) = \int_y^\infty \frac{e^{-x}}{x} dx \qquad (16)$$

is the familiar exponential integral. Equation 15 can be easily evaluated using numerical methods. To graphically evaluate the similarity of the bacterial allele frequencies to those expected under the neutral IAM, we plotted the MLE estimates of allele frequencies *vs.* those expected under the neutral IAM with $\hat{\theta}$ estimated by using the observed number of alleles and the expected sample size estimated as $n\hat{\lambda}$ and numerically solving Equation 13 above to obtain the MLE of θ.

MONTE CARLO SIMULATION STUDY

Monte Carlo simulations were used to examine the bias and variance of the maximum-likelihood estimators of λ and **p**. From a hypothetical population of bacteria with k different alleles (frequencies of which are $p_1, p_2, \ldots, p_k$, respectively), we simulated the sampling of n infected hosts. The number of bacterial lineages infecting each host is assumed to be Poisson distributed with a mean of λ and the observations are the presence and absence of individual alleles in each infected host. From these n independent simulated host samples, MLEs of the bacterial population allele frequencies, **p**, and λ, were calculated using Equations 11 and 12. The process of sampling n infected hosts from the bacterial population was simulated 1000 times and the bias and variance were calculated for each estimator.

The results shown in Figure 1 are based on simulated samples from a hypothetical bacterial population with $k = 3$ distinct alleles with sample sizes of $n = 100$, 500, 1000, and 5000 simulated infected hosts. Deviations (in percentage) of the estimated values of the allele frequencies from their actual values (the bias; Figure 1, A and B) and the variance of the estimated allele frequencies (Figure 1, C and D) are plotted against actual frequency of one of the alleles (frequencies of the remaining two alleles are kept equal). It can be seen from Figure 1, A and B, that the relative bias of the estimated allele frequencies falls between -15 and $+10\%$ under the various ranges of allele frequencies and λ values examined. The accuracy of the gene frequency estimates is improved by increasing the sample size, n, when λ is large relative to the number of alleles (Figure 1B). However, when λ is small (*e.g.*, $\lambda = 2$), the accuracy is not improved by sampling more hosts (Figure 1A). As expected, the variance of all estimates is reduced when the sample size is increased (*e.g.*, Figure 1, C and D).

The evenness of the gene frequency distribution has a strong effect on the bias and variance of the allele frequency estimates. The best estimates (meaning those with smallest bias and variance) of allele frequencies are achieved when the actual allele frequencies are evenly distributed, even when the sample size is small (see Figure 1, A and B, at the point $p = 0.3333$, where the distribution is even). Estimates of λ (the average number of bacterial lineages infecting a host) are more biased and have higher variance than estimates of allele frequencies, especially when λ is small. For example, the method overestimates λ regardless of sample size, n, and the shape of gene frequency distribution, **p**, when the true value of λ is 2.0 (the estimates ranged from 2.5 to 3.6, data not shown).

EXAMPLE: LYME DISEASE

We have applied the method to two published studies of local populations of *B. burgdorferi*, the bacterial agent of Lyme disease. Lyme disease is transmitted mainly by Ixodes ticks and is the most prevalent vector-borne disease in the United States (CDC 1997). In endemic regions of Lyme disease, the ticks, vertebrate hosts, and patients are often infected with multiple genospecies or strains of *B. burgdorferi* (Demaerschalck *et al.* 1995; Pichon *et al.* 1995; Guttman *et al.* 1996). Population genetic analyses of *B. burgdorferi* suggested that the high level of local genetic diversity observed in this species may be maintained by frequency-dependent selection mediated by host immune responses to spirochete infection (Qiu *et al.* 1997; Wang *et al.* 1999). One piece of evidence for diversifying selection came from the consistently significant results of Ewens-Watterson tests of frequency distributions of SSCP alleles sampled from local populations of *B. burgdorferi* (Qiu *et al.* 1997; Wang *et al.* 1999).

In light of our study, the method of gene frequency estimation used in previous studies of Borrelia needs to be reevaluated. This is because in the earlier studies frequencies of SSCP alleles were estimated by directly counting the number of bands (distinctive for each allele) observed on electrophoresis gels. Since the presence of a pair of SSCP bands on a gel indicates the presence of a particular allele in a tick regardless of the actual number of microbes infecting the tick, the direct counting method tends to underestimate the frequen-

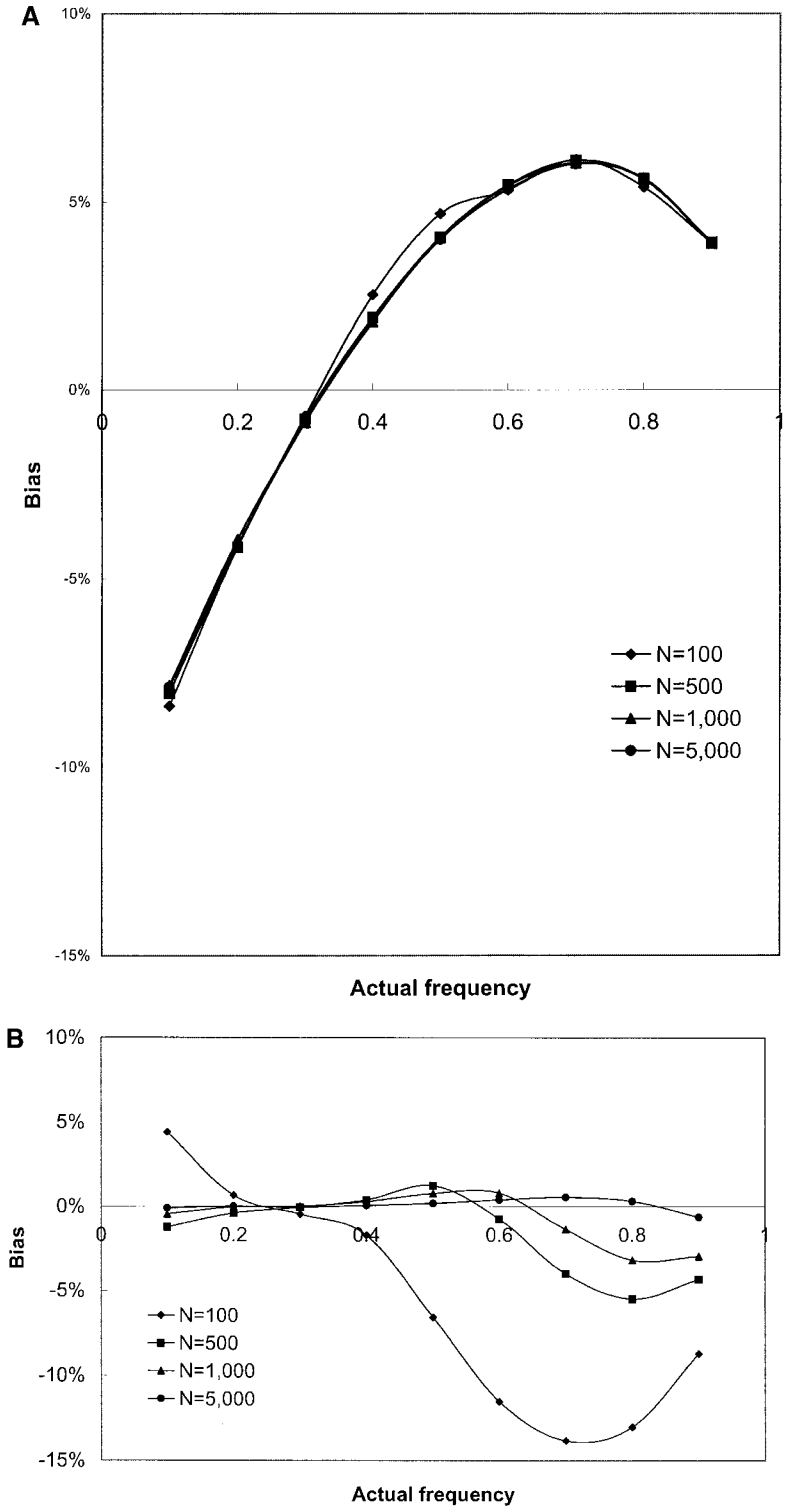


Figure 1.—Results of simulation study of the performance of maximum-likelihood estimators of gene frequencies (**p**) and number of infecting bacteria (λ). A hypothetical microbial population with $k = 3$ distinct alleles infecting a host population was simulated. The frequency of one allele (p_1) was successively set to be 0.1, 0.2, . . . , 0.9 while frequencies of the remaining two alleles (p_2 and p_3) were kept equal (these frequencies sum to unity). Samples of infected hosts of size $N = 100$, 500, 1000, and 5000 were simulated and the frequency of each allele was estimated using Equations 13 and 14. The bias of ($\hat{p}_1$) and the variance of $\hat{\mathbf{p}}$ were obtained from 1000 replicate simulations of the sampling process. The standardized bias of $\hat{p}_1$ was calculated using the formula $(\hat{p}_1 - p_1)/(p_1)$, where p_1 is the (known) true value of p_1 used in the simulations (A and B). As well, the variance of $\hat{\mathbf{p}}$ was calculated for the simulated data sets (C and D) and these values are plotted against p_1. The individual parts are the bias of $\hat{p}_1$ (A and B) and variance of $\hat{\mathbf{p}}$ (C and D) when the average number of microbial lineages infecting hosts is either low (λ = 2, A and C) or high (λ = 10, B and D).

cies of common alleles and overestimate those of rare ones. Allele frequencies estimated by using the direct counting method would thus tend to appear more uniform (having lower values of homozygosity, F) and would bias the results of the Ewens-Watterson test in favor of balancing selection.

Molecular data from surveys of two local populations of *B. burgdorferi* (Qiu *et al.* 1997; Wang *et al.* 1999) are reproduced in Table 1 and allele frequencies on two outer surface protein loci were reestimated using the maximum-likelihood method. It can be seen that in each population the corrected frequencies are less evenly distributed than the uncorrected ones (Table 1). Nonetheless, the allele frequency distributions in both

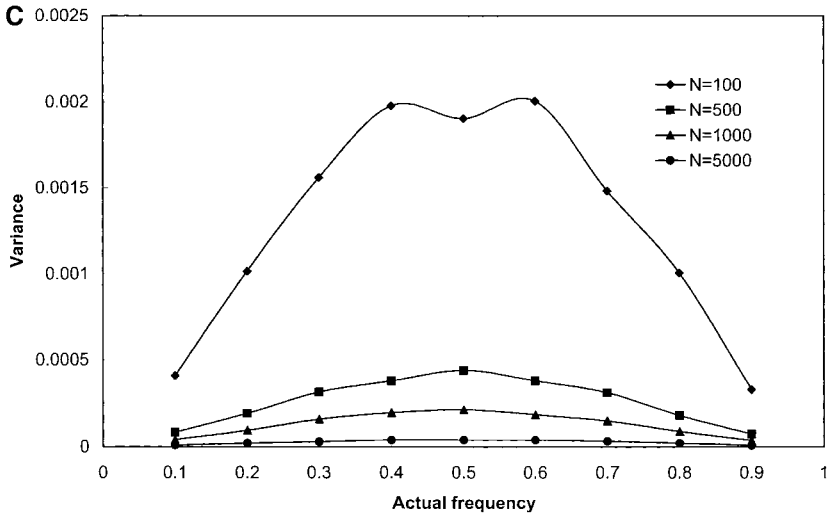
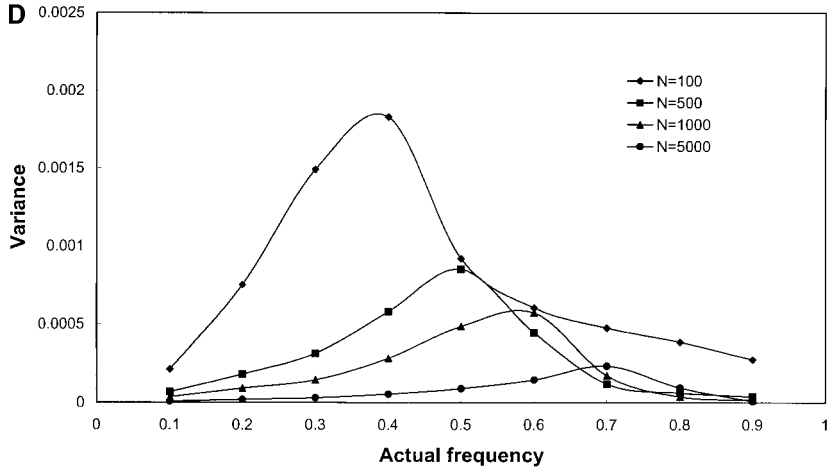


Figure 1.—*Continued.*



populations are still more even than expected under a neutral evolutionary model (compare the first and second columns of the histograms shown in Figure 2). A more uniform frequency distribution than expected under neutrality can be due to either evolutionary factors like balancing selection (Qiu *et al.* 1997; Wang *et al.* 1999) or recent population growth. To construct the expected frequencies under neutrality that are given in the second column of the histograms of Figure 2, the population parameter $\theta = 2N_e\mu$ was estimated as described above using $\hat{\lambda}n$ as the approximate sample size. Estimates of θ obtained in this way were quite large for both loci examined, $\theta = 0.43$ for *ospA* and $\theta = 2.48$ for *ospC*. If we assume that the mutation rate is on the order of $\mu \approx 10^{-8}$, this suggests that the corresponding estimates of the effective population size will be 1.2×10^8 and 2.2×10^7, respectively. Either a large population of hosts or potential subdivision of bacterial populations among hosts (and regions) could account for this large effective population size.

One possible source of error for estimates of θ are the estimates of λ used in the calculation; these might be too large due to the upward bias of the MLE of λ. To investigate whether such bias could account for the deviation of estimated bacterial gene frequencies from those expected under neutrality, we also calculated the expected frequencies using estimates of λ that were one-fifth as large as the MLEs. These expected frequencies are shown in the third column of Figure 2 and remain very different from the observed frequencies (column 1). These results suggest that the deviation of gene frequencies from the neutral expectation that we observed are not an artifact of bias in estimates of θ.

DISCUSSION

In this study, a maximum-likelihood method for estimating gene frequencies in a bacterial population based on the presence and absence of alleles in infected hosts is developed. The method is particularly useful for application to the data generated by the increasingly common studies of bacterial populations that use non-culture-based molecular detection techniques. Additionally, a graphical method based on the neutral distribution of allele frequencies has been developed to test for the presence of natural selection (or a recent popula-

TABLE 1

Estimated allele frequency distributions of outer surface protein (Osp) alleles of *B. burgdorferi*

Population	Gene locus (allele type)	Allele	Total no. of infected ticks (n)	No. of ticks infected with jth allele (z_j)	Uncorrected estimates	Maximum-likelihood estimates	
						Allele frequencies [p_j, SE(p_j)]	Average no. of bacterial lineages per host [λ, SE(λ)]
Adult 95 and 96[a]	*ospA* (four mobility classes)	1	367	46	0.11	0.09 (0.01)	1.496 (0.005)
		2		166	0.37	0.40 (0.02)	
		3		112	0.25	0.24 (0.02)	
		4		120	0.27	0.27 (0.02)	
SI94[b]	*ospC* (11 major groups)	A	40	12	0.16	0.17 (0.04)	2.09 (0.05)
		B		12	0.16	0.17 (0.04)	
		C		11	0.15	0.16 (0.04)	
		D		9	0.12	0.12 (0.04)	
		E		4	0.05	0.05 (0.02)	
		F		6	0.08	0.08 (0.03)	
		G		5	0.07	0.06 (0.03)	
		H		7	0.10	0.09 (0.03)	
		I		1	0.01	0.01 (0.03)	
		J		2	0.03	0.03 (0.01)	
		K		5	0.07	0.06 (0.02)	

[a] Qiu *et al.* (1997).
[b] Wang *et al.* (1999).

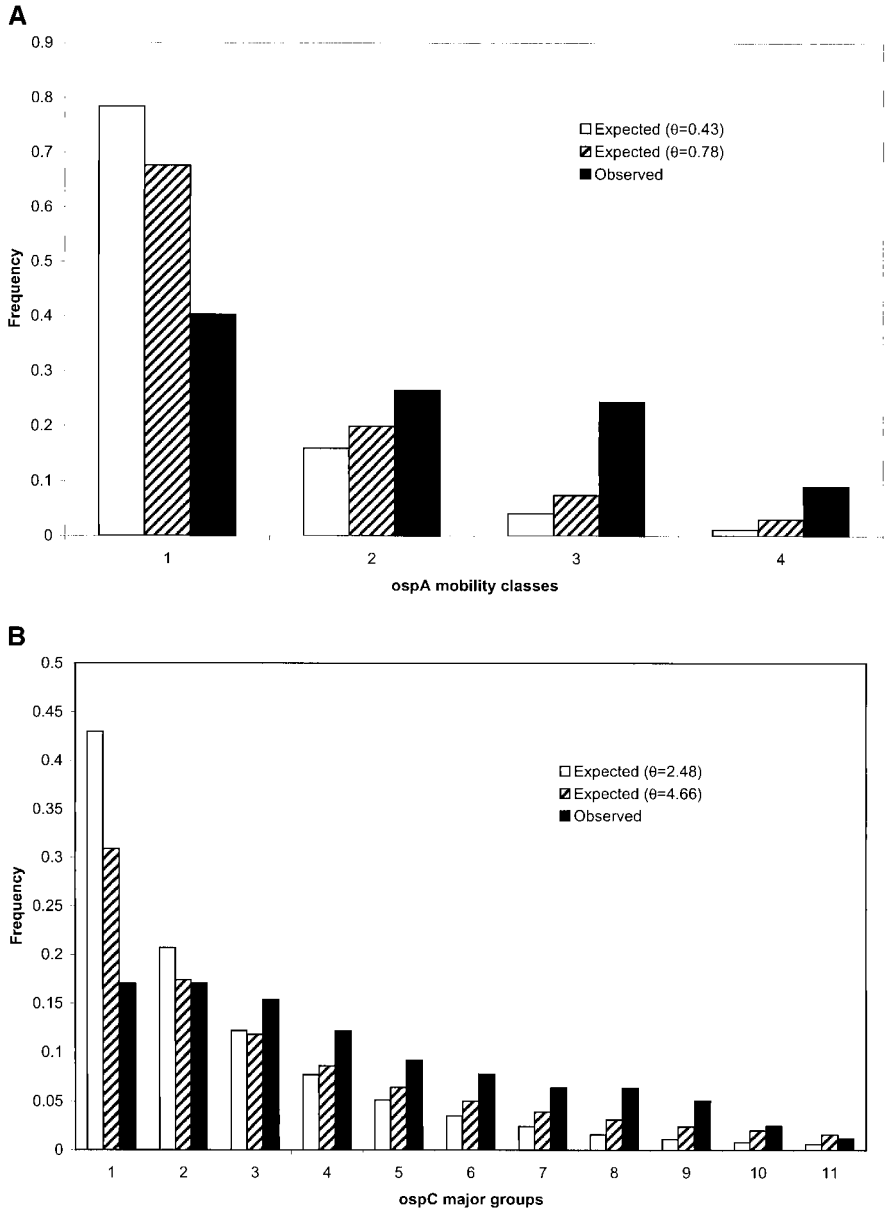


Figure 2.—Comparison of estimated allele frequencies with the expected frequencies under neutrality for outer surface protein (*osp*) alleles in natural populations of *B. burgdorferi*. (A) Frequencies of four SSCP mobility classes of *ospA* in a population of 367 infected adult *I. scapularis* ticks. The maximum-likelihood estimate of the average number of bacterial lineages infecting a tick, $\hat{\lambda}$, is $\sim$1.50, which gives a total sample size of 549 *osp*A lineages. $\hat{\theta}$ was obtained using Equation 13 on the basis of either the estimate $\hat{\lambda}$ (in which case $\hat{\theta} = 0.426$) or one-fifth the value of $\hat{\lambda}$ (in which case $\hat{\theta} = 0.776$). Expected allele frequencies under neutrality in both cases were obtained using Equation 15. (B) Frequencies of 11 major groups of alleles at *osp*C in a population of 40 infected ticks. For these data, the estimate of λ is $\hat{\lambda} = 2.09$, which gives an estimate of the sample size to be 84 *osp*C lineages. Using this estimate of $\hat{\lambda}$, the estimate of θ was $\hat{\theta} = 2.48$. If instead a value of $1/5\ \hat{\lambda}$ was used, we obtained $\hat{\theta} = 4.66$. Expected allele frequencies under neutrality in both cases were obtained using Equation 15.

tion expansion), using bacterial gene frequencies. This method provides a test of neutrality for bacterial genes assayed using SSCP that is analogous to the classical test based on the Ewens (1972) sampling distribution. Both tests are sensitive to recent population expansions, which can cause a rejection of the neutral hypothesis even in the absence of selection.

Model of microbial sampling: The basic probability model that we have developed (and used to derive our estimators) in this article assumes that hosts are collected and examined for the presence (or absence) of microbes carrying particular alleles. Implicit in the model is an assumption that hosts are infected, at random, from a "source" population of microbes. Our methods estimate the allele frequencies in this source population from which microbes are sampled during the host infection process. In the case of *B. burgdorferi*, the source population is the population of microbes infecting the definitive (vertebrate) hosts. For other microbes, the source population about which inferences are being made will differ. For example, if one is instead studying populations of free-living soil bacteria (rather than a parasite), collects *n* independent soil samples of equal mass (rather than *n* hosts), and then determines the alleles present in each sample, an application of our method will provide estimates of the allele frequencies for the population of soil microbes over the entire region of sampling. Our model also assumes that no mutations occur within the sampled hosts; this assumption is reasonable for short-lived hosts harboring small populations of bacteria ($< \sim 10^6$ bacteria per host). Finally, the model assumes that there is no loss of bacterial lineages (due to genetic drift, for example) within hosts following infection. If these assumptions are satisfied, it is not necessary to explicitly model the within-host population dynamics. The approach could be extended

to allow more complex within-host microbial dynamics, mutation, and drift.

Prospects for new statistical methods: The methods developed in this article represent a reasonable first solution to this problem. The maximum-likelihood method that we have used to estimate allele frequencies has some potential disadvantages, however. The most significant disadvantage of the likelihood approach is that it provides only point estimates of the allele frequencies. Often estimated allele frequencies will subsequently be used to test other hypotheses (in this study, for example, we used them to test for neutrality of alleles). A technically superior approach would be to calculate the Bayesian posterior probability density of the allele frequencies (see, for example, Rannala and Mountain 1997). This posterior density could then be used in subsequent hypothesis tests involving the allele frequencies and would take account of the fact that the allele frequencies are uncertain (they have been estimated from the data). A numerical Bayesian approach (using Markov chain Monte Carlo methods, for example) would also allow the method to be more easily extended to allow for increasingly complex models of the host infection process, etc.

Bacterial clonality: Bacteria are asexually reproducing organisms and the rate of recombination between genotypes is much lower than in sexually reproducing organisms. One consequence of an asexual mode of reproduction and clonal population structure is that bacterial genotypes or strains are relatively stable genetic identities (for recent reviews of bacterial clonality see Selander *et al.* 1994; Maynard Smith 1995; Guttman 1997). Population genetic (evolutionary) processes (*e.g.*, genetic drift and natural selection) in bacterial species are thus best described using units of individual bacterial lineages rather than the individual cells constituting a bacterial population. A bacterial lineage is defined here as bacterial cells asexually propagated (*e.g.*, by binary fission) from a single ancestral bacterial cell. Different bacterial lineages, like different individuals in plant or animal populations, can be of the same as well as of different genotypes or strains. Therefore, individual cells in a bacterial population are best considered to be a mixture of genotypically identical (or different) lineages rather than individual organisms as in an animal or plant population.

Our analysis of *B. burgdorferi* supports the idea that the number of bacterial lineages infecting a host is usually small. Presumably this is because the hosts are often initially infected with a small number of bacterial cells and an even smaller number of bacterial lineages. For example, the average number of Borrelia lineages infecting a tick (*i.e.*, the parameter λ) was estimated to be $\sim$1.5 as determined by *ospA* SSCP types in one population and 2.1 as determined by *ospC* SSCP types in another population (Table 1). These low estimates of the number of infecting lineages lend support to the clonal view of bacterial population genetics outlined above.

A second important consequence of bacterial clonality is the high degree of linkage disequilibrium that exists among various loci (Dykhuizen and Green 1991). As a result of linkage disequilibrium, natural selection at one locus can have a genome-wide effect, causing population genetic dynamics such as selective sweeps (Cohan 1994). For highly clonal bacterial species such as *B. burgdorferi* (Dykhuizen *et al.* 1993), it is therefore not surprising that balancing selection appears to manifest itself at multiple loci (*e.g.*, *ospA* and *ospC*; see Wang *et al.* 1999). A test of neutrality applied to one gene, therefore, can potentially detect selection acting anywhere in the bacterial genome.

To summarize, population genetic studies of bacterial species differ from those of plant or animal species in at least two important aspects. First, gene frequencies in a bacterial population most accurately reflect the relative abundance of bacterial strains existing in a population; they are poorer measures of relative frequencies of individual bacterial cells that differ in their genotypes. Second, the genome-wide population genetic dynamics of bacterial species can often be approximated by population genetic dynamics at a single locus (*e.g.*, changes of allele frequencies due to natural selection) due to the extensive linkage disequilibrium among loci across the bacterial genome.

***In situ* quantitative methods:** Over the past decade, various molecular methods aimed at quantifying *in situ* cellular abundance of bacteria (such as quantitative PCR and quantitative hybridization) have been developed (Orlando *et al.* 1998). Because these techniques can directly quantify the number of gene copies in biological samples, it is tempting to use molecular quantitative techniques to estimate gene frequencies in bacterial populations experimentally. However, as discussed above, the total number of cells that make up each bacterial strain in a population may not be as relevant a population genetic parameter as the number of independent bacterial lineages that infect a host. It is therefore neither necessary nor proper to estimate bacterial gene frequencies from, for example, the number of gene copies giving rise to the individual SSCP bands, using molecular quantification methods. Using the present methods for estimating gene frequencies in bacterial populations, it is only necessary to obtain information on the presence and absence of alleles in infected hosts using the much simpler qualitative molecular detection techniques. Apart from these theoretical considerations, molecular quantitative methods also suffer from experimental complications such as preferential primer or probe annealing to templates of certain haplotype sequences, natural variations in DNA or RNA copy numbers in a genome, and formation of chimera molecules in PCR amplifications (Amann *et al.* 1995; von Wintzingerode *et al.* 1997).

Effects of sample size and interaction among genotypes: The results of our Monte Carlo computer simulations suggest that the accuracy and variance of maxi-

mum-likelihood estimates are influenced mainly by two factors: the sample size, *n*, and the average number of lineages infecting a host, λ. Increasing the sample size generally improves the accuracy and reduces the variance of the maximum-likelihood estimates. However, when hosts are infected by a small number of bacterial lineages, the accuracy of gene frequency estimates is relatively insensitive to changes in sample size (Figure 1A). In the Borrelia example, estimates of the number of ospA and ospC SSCP lineages in infected ticks are both found to be low (1.5 and 2.1, respectively). In this case, the gene frequency estimates tend to be very reliable even if sample sizes are small (see standard errors in Table 1). However, caution should be used in interpreting the estimated values of λ because simulations suggest that these may overestimate the true values by as much as 80%. It is also important to note that the maximum-likelihood method presented in this article is based on the assumption that infection of hosts by bacterial lineages is a Poisson process (see theory) and that there are no interactions (*e.g.*, attraction or repulsion) among coexisting strains in a host. The use of these methods may not be justified when nonrandom associations are found among different bacterial strains or species.

Program availability: The program BFREQ (written in C++) is available to estimate allele frequencies and perform the Monte Carlo simulations described in this article. A *Mathematica* (version 2.2, Wolfram 1992) package is also available for calculating expected allele frequencies under neutrality based on the Poisson-Dirichlet distribution (Griffiths 1979) and for estimating the parameter θ by solving Equation 13. The computer programs can be obtained from http://allele.bio.sunysb.edu.

The authors benefited from discussions with Ing-Nang Wang and suggestions from Simon Tavaré. This study was supported by a grant from the National Institute of Allergy and Infectious Diseases (RO1AI33454) to Dr. Benjamin J. Luft of Department of Medicine, State University of New York at Stony Brook and by cooperative agreement number U50/CCU206608 and U50/CCU210518 from the Centers for Disease Control and Prevention to Dr. Benjamin J. Luft and Dr. Edward M. Bosler of Department of Medicine, State University of New York at Stony Brook, respectively. B. Rannala was supported by National Institutes of Health Grant HG-01988-01. This is contribution 1063 from Graduate Studies in Ecology and Evolution, State University of New York at Stony Brook.

LITERATURE CITED

Amann, R. I., W. Ludwig and K. H. Schleifer, 1995 Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59:** 143–169.

CDC, 1997 Lyme disease—United States, 1996. Morb. Mortal. Wkly. Rep. **46:** 531–535.

Cohan, F. M., 1994 Genetic exchange and evolutionary divergence in prokaryotes. Trends Ecol. Evol. **9:** 175–180.

Demaerschalck, I., A. B. Messaoud, M. D. Kesel, B. Hoyois, Y. Lobet *et al.*, 1995 Simultaneous presence of different *Borrelia burgdorferi* genospecies in biological fluids of Lyme disease patients. J. Clin. Microbiol. **33:** 602–608.

Dykhuizen, D. E., and L. Green, 1991 Recombination in *Escherichia coli* and definition of biological species. J. Bacteriol. **173:** 7257–7268.

Dykhuizen, D. E., D. S. Polin, J. J. Dunn, B. Wilske, V. Preac-Mursic *et al.*, 1993 *Borrelia burgdorferi* is clonal: implications for taxonomy and vaccine development. Proc. Natl. Acad. Sci. USA **90:** 10163–10167.

Ewens, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

Griffiths, R. C., 1979 On the distribution of allele frequencies in a diffusion model. Theor. Popul. Biol. **15:** 140–158.

Guttman, D. S., 1997 Recombination and clonality in natural populations of *Escherichia coli*. Trends Ecol. Evol. **12:** 16–22.

Guttman, D. S., P. W. Wang, I. Wang, E. Bosler, B. J. Luft *et al.*, 1996 Multiple infection of *Ixodes scapularis* ticks by *Borrelia burdorferi* as revealed by single-strand conformation polymorphism analysis. J. Clin. Microbiol. **34:** 652–656.

Hugenholtz, P., B. M. Goebel and N. R. Pace, 1998 Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. **180:** 4765–4774.

Manly, B. F. J., 1985 *The Statistics of Natural Selection on Animal Populations*. Chapman, New York.

Maynard Smith, J., 1995 Do bacteria have population genetics? pp. 1–12 in *Population Genetics of Bacteria*, edited by S. Baumberg, J. P. W. Young, E. M. H. Wellington and J. R. Saunders. Cambridge University Press, Cambridge, United Kingdom.

Orlando, C., P. Pinzani and M. Pazzagli, 1998 Developments in quantitative PCR. Clin. Chem. Lab. Med. **36:** 255–269.

Pichon, B., E. Godfroid, B. Hoyois, A. Bollen, F. Rodhain *et al.*, 1995 Simultaneous infection of *Ixodes ricinus* nymphs by two *Borrelia burgdorferi* sensu lata species: possible implications for clinical manifestations. Emerg. Infect. Dis. **1:** 89–90.

Qiu, W., E. M. Bosler, H. J. Campbell, G. D. Ugine, I.-N. Wang *et al.*, 1997 A population genetic study of *Borrelia burgdorferi* sensu stricto from eastern Long Island, New York, suggested frequency-dependent selection, gene flow and host adaptation. Hereditas **127:** 203–216.

Rannala, B., and J. L. Mountain, 1997 Detecting immigrants by using multilocus genotypes. Proc. Natl. Acad. Sci. USA **94:** 9197–9201.

Relman, D. A., 1999 The search for unrecognized pathogens. Science **284:** 1308–1310.

Selander, R. K., J. Li, E. F. Boyd, F. S. Wang and K. Nelson, 1994 DNA sequence analysis of the genetic structure of populations of *Salmonella enterica* and *Escherichia coli*, pp. 17–49 in *Bacterial Diversity and Systematics*, edited by F. G. Priest, A. Ramos-Cormenzana and B. J. Tindall. Plenum, New York.

von Wintzingerode, F., U. B. Gobel and E. Stackebrandt, 1997 Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. FEMS Microbiol. Rev. **21:** 213–229.

Wang, I., D. E. Dykhuizen, W. Qiu, J. J. Dunn, E. M. Bosler *et al.* 1999 Genetic diversity of ospC in a local population of *Borrelia burgdorferi* sensu stricto. Genetics **151:** 15–30.

Wolfram Research, Inc., 1992 *Mathematica*, Version 2.2. Wolfram Research, Inc., Champaign, IL.

Communicating editor: A. G. Clark